Response surface methodology revisited

Published in:
Proceedings of the 2002 Winter Simulation Conference

Publication date:
2002

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.
RESPONSE SURFACE METHODOLOGY REVISITED

Ebru Angün
Jack P.C. Kleijnen
Department of Information Management/Center for Economic Research (CentER)
School of Economics and Business Administration
Tilburg University
5000 LE Tilburg, THE NETHERLANDS

Dick Den Hertog
Gül Gürkan
Department of Econometrics and Operations Research /Center for Economic Research (CentER)
School of Economics and Business Administration
Tilburg University
5000 LE Tilburg, THE NETHERLANDS

1 INTRODUCTION

Response Surface Methodology (RSM) searches for the input combination that optimizes the simulation output. RSM treats the simulation model as a black box. Moreover, this paper assumes that simulation requires much computer time. In the first stages of its search, RSM locally fits first-order polynomials. Next, classic RSM uses steepest descent (SD); unfortunately, SD is scale dependent. Therefore, Part 1 of this paper derives scale independent ‘adapted’ SD (ASD) accounting for covariances between components of the local gradient. Monte Carlo experiments show that ASD indeed gives a better search direction than SD. Part 2 considers multiple outputs, optimizing a stochastic objective function under stochastic and deterministic constraints. This part uses interior point methods and binary search, to derive a scale independent search direction and several step sizes in that direction. Monte Carlo examples demonstrate that a neighborhood of the true optimum can indeed be reached, in a few simulation runs.
independent search direction – inspired by interior point methods - and several step sizes - inspired by binary search. This search direction namely scaled and projected SD, is a generalization of classic RSM’s SD. We then combine these search direction and step sizes into an iterative heuristic. Notice that if there are no binding constraints at the optimum, then classic RSM combined with ASD might suffice.

The remainder of this paper is organized as follows. For the unconstrained problem, §2 derives ASD, its mathematical properties, and its interpretation. §3 compares the search directions SD and ASD, by means of Monte Carlo experiments. §4 derives a novel heuristic combining a search direction and a step size procedure for constrained problems. §5 studies the performance of the novel heuristic by means of Monte Carlo experiments. §6 gives conclusions.

Note that this paper summarizes two separate papers, namely Kleijnen, Den Hertog, and Angün (2002), and Angün et al. (2002), which give all mathematical proofs and additional experimental results.

2 ADAPTED STEEPEST DESCENT

RSM uses the following approximation:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j d_j + e$$

where \( y \) denotes the predictor of the expected simulation output; \( e \) denotes the noise consisting of intrinsic noise caused by the simulation’s pseudo-random numbers (PRN) plus lack of fit. RSM assumes white noise; that is, \( e \) is normally, identically, and independently distributed with zero mean \( \mu_e \) and constant variance \( \sigma_e^2 \).

The OLS estimator of the \( q ? k + 1 \) parameters \( \beta ? \beta_0, \ldots, \beta_k \) in (1) is

$$\hat{\beta} = (XX)^{-1} X w$$

with \( X \): \( N \times q \) matrix of explanatory variables including the ‘dummy’ variable with constant value 1; \( X \) is assumed to have linearly independent columns

\[ \begin{align*}
N &= \sum_{i=1}^{m_1} n_i \\
m_i &= \text{number of replicates at input combination } i, \text{ with } m_i \neq 0 \\
n &= \text{number of different, simulated input combinations with } n \neq 0 \n\end{align*} \]

\( w \): vector with \( N \) simulation outputs \( w_{i,r} \), 

\[ \text{The noise in (1) may be estimated through the mean squared residual (MSR):} \]

$$\hat{s}_e^2 = \frac{\sum_{i=1}^{N} \sum_{r=1}^{m_i} (w_{i,r} - \hat{y}_i)^2}{(N-q)}$$

where \( \hat{y}_i \) follows from (1) and (2):

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j d_{i,j}.$$
Angün, Den Hertog, Gürkan, and Kleijnen

We mention the following mathematical properties and interpretations of ASD.

The first term in (6a) means that the ASD path starts from the point with minimal predict or variance. The second term means that the classic SD direction \( \hat{\beta}_0 \) (second term’s last factor) is adjusted for the covariance matrix of \( \hat{\beta}_0 \); see \( C \) in (4). The step size \( ? \) is quantified in (6b).

Kleijnen et al. (2002) proves that ASD is scale independent.

In case of large signal/noise ratios \( \hat{\beta}_0 \) var / \( \beta_{jj} \), the denominator under the square root in (6b) is negative so (6) does not give a finite solution for \( d^+ \). Indeed, if the noise is negligible, we have a deterministic problem, which our technique is not meant to address (many other researchers - including Conn et al. (2000) - study optimization of deterministic simulation models).

In case of a small signal/noise ratio, no step is taken. Kleijnen et al (2002) further discusses two subcases: (i) the signal is small; (ii) the noise is big.

3 COMPARISON OF ASD AND SD THROUGH MONTE CARLO EXPERIMENTS

To compare the ASD and SD search directions, we perform Monte Carlo experiments. The Monte Carlo method is an efficient and effective way to estimate the behavior of search techniques applied to random simulations; see Kleijnen et al. (2002).

We limit the example to two inputs, so \( k = 2 \). We generate the simulation output \( w \) through a second-order polynomial with white noise:

\[
w = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_{1,1} d_1^2 + \beta_{2,2} d_2^2 + \beta_{1,2} d_1 d_2 + e.
\]  

The response surface (7) holds for the global area, for which we take the unit square: ?1 ? d_1 ? 1 and ?1 ? d_2 ? 1. (We have already seen that RSM fits first-order polynomials locally.)

In the local area we use a one-at-a-time design, because this design is non-orthogonal - and in practice designs in the original inputs are not orthogonal (see Kleijnen et al. (2002)). The specific local area is in the lower corner of Figure 1. To enable the computation of the MSR, we simulate one input combination twice: \( m_1 = 2 \).

In case of large signal/noise ratios \( \hat{\beta}_0 \) var / \( \beta_{jj} \), the denominator under the square root in (6b) is negative so (6) does not give a finite solution for \( d^+ \). Indeed, if the noise is negligible, we have a deterministic problem, which our technique is not meant to address (many other researchers - including Conn et al. (2000) - study optimization of deterministic simulation models).

In case of a small signal/noise ratio, no step is taken. Kleijnen et al (2002) further discusses two subcases: (i) the signal is small; (ii) the noise is big.

Figure 1: Tilted Ellipsoid Contours \( E[w|d_1, d_2] \)

with Global and Local Experimental Areas

We consider a specific case of (7): \( \beta_0 = \beta_1 = \beta_2 = 0, \beta_{1,2} = 2, \beta_{1,1} = -2, \beta_{2,2} = -1 \), so the contour functions (for example, iso-cost curves) form ellipsoids tilted relative to the \( d_1 \) and \( d_2 \) axes. Hence, (7) has as its true optimum \( \beta^* = 0, 0 \). After fitting a first-order polynomial, we estimate the SD and ASD paths starting from \( d^*_0 = (0.85, -0.95) \), explained above (4).

In this Monte Carlo experiment we know the truly optimal search direction, namely the vector (say) \( g \) that starts at \( d^*_0 \) and ends at the true optimum \( \beta^* = 0, 0 \). So we compute the angle (say) \( \hat{\beta} \) between the true search direction \( g \) and the estimated search direction \( p \):

\[
\hat{\beta} = \arccos \frac{g \cdot p}{|g||p|}.
\]  

Obviously, the smaller \( \hat{\beta} \) is, the better the search technique performs.

To estimate the distribution of \( \hat{\beta} \) defined in (8), we take 100 macro-replications. Figure 2 shows a bundle of 100 \( p \)'s around \( g \) when \( s = 0.1 \). In each macro-replicate, we apply SD and ASD to the same I/O data \( \tilde{w}, d_1, d_2 \). We characterize the resulting empirical \( \hat{\beta} \) distribution through several statistics, namely its average, standard deviation, and specific quantiles; see Table 1.
In RSM, there have been several approaches to multiresponse optimization. Khuri (1996) surveys most of these approaches (including desirability functions, generalized distances, and dual responses). Angün et al. (2002) discusses drawbacks of these approaches. To overcome these drawbacks, we propose the following alternative based on mathematical programming.

We select one of the responses as the objective and the remaining (say) \( z \neq 1 \) responses as constraints. The SD search would soon hit the boundary of the feasible area formed by these constraints, and would then creep along this boundary. Instead, our search starts in the interior of the feasible area and avoids the boundary; see Barnes (1986) on Karmarkar’s algorithm for linear programming.

Note that our approach has the additional advantage of avoiding areas in which the simulation model is not valid and may even crash.

Formally, our problem becomes:

\[
\begin{align*}
\text{minimize} & \quad E[w_0'] d^T d \\
\text{subject to} & \quad E[w_h'] d^T d \leq a_h \quad \text{for } h \neq 1, ..., z \neq 1 \\
& \quad l \leq d \leq u
\end{align*}
\]

where \( d \) is the vector of simulation inputs, \( l \) and \( u \) the deterministic lower and upper bounds on \( d \), \( a_h \) the right-hand-side value for the \( h \)th constraint, and \( w_h' d \leq a_h \) for \( h \neq 1, ..., z \neq 1 \) is the response \( h \).

Note that probabilities (for example, service percentages) can be formulated as expected values of indicator functions. Further, the multiple simulation responses are correlated, since they are estimated through the same PRN fed into the same simulation model.

As in Part 1, we locally fit a first-order polynomial – but now for each response; see (1). However, the noise is now multi-variate normal – still with zero means but now with covariance matrix \( S \). Yet, since the same design is used for all \( z \) responses, the GLS estimator reduces to the OLS estimator; see Ruud (2000, p. 703).

Therefore we still use (2), but to the symbol \( \hat{\beta} \) we add the subscript \( h \). Further, for \( h = 1, ..., z \neq 1 \) we estimate \( S \) through the analogue of (3):

\[
\hat{\beta}_{0;h} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{(w_k - \hat{y}_{k,h})^2}{\Sigma_{0;h}^{1/2}} e^{-\frac{(w_k - \hat{y}_{k,h})^2}{2\Sigma_{0;h}}} \mathrm{d}w_k
\]

(10)

We introduce \( \hat{\beta}_{h} = \hat{\beta}_{1;h}, ..., \hat{\beta}_{z;h} \), where \( \hat{\beta}_{h} = \hat{\beta}_{1;h}, ..., \hat{\beta}_{z;h} \) denotes the vector of OLS estimates \( \hat{\beta}_{0;h} \) (excluding the intercept \( \hat{\beta}_{0;h} \)) for the \( h \)th response. Adding slack vectors \( s, r, \) and \( v \), we obtain
where \( b_0 \) denotes the vector of OLS estimates \( \hat{\beta}_{0,0} \)
(excluding the intercept \( \hat{\beta}_{0,0} \)) for \( w_0 \), and \( c \) is the
vector with components \( c_h \cdot a_h \cdot \hat{\beta}_{0,h} \cdot h = 1, \ldots, z-1 \).
Through (11) we obtain a local linear approximation for
(9). Then using ideas from interior point methods - more
specifically the affine scaling method - Angün et al. (2002)
derives the following search direction:

\[
p \cdot (B^T \bar{S}^{-2} B + \bar{R}^{-2} + \bar{V}^{-2})^{-1} b_0
\]

where \( \bar{S}, \bar{R}, \) and \( \bar{V} \) are diagonal matrices with the current
estimated slack vectors \( \bar{S}, \bar{R}, \bar{V} \cdot 0 \) on the diagonal.

First, we compute the maximum step size assuming that
the local approximation (11) holds globally:

\[
\gamma_{\text{max}} \cdot \max\{0, \min\{\gamma_1, \gamma_2, \gamma_3\}\}
\]

where

\[
\gamma_1 \cdot \min\{c_h \cdot b_h^T d / p \cdot b_h : h = 1, \ldots, z-1\}, \gamma_2 \cdot \min\{a_j \cdot d_j^T / p_j : j = 1, \ldots, k\}, \gamma_3 \cdot \min\{a_j \cdot d_j^T / p_j : j = 1, \ldots, k\}
\]

To increase the probability of staying within the interior of
the feasible region, we take only 80% of \( \gamma_{\text{max}} \) as our
maximum step size.

The subsequent step sizes are inspired by binary
search, as follows. We systematically halve the current
step size along the search direction. At each step, we select
as the best point the one with the minimum value for the
simulation objective \( w_0 \), provided it is feasible. We stop
the search in a particular direction after a user-specified
number of iterations (say) \( G \cdot 3 \). For details see Angün et
al. (2002).

For all these steps we use common random numbers
(CRN), in order to better test whether the objective
improves. Moreover, we test whether the other \( z \cdot 1 \)
responses remain within the feasible area; we test the slack
vector \( s \), introduced in (11). These \( z \) tests use ratios
instead of absolute differences, to avoid scale dependence.

A statistical complication is that these ratios may not
have finite moments. Therefore we test their medians (not
their means). For these tests we use Monte Carlo sampling,
which takes negligible computer time compared with the
expensive simulation runs. This Monte Carlo takes (say)
\( K \cdot 1000 \) samples from the assumed distributions with
means and variances estimated through the simulation; in
the numerical example of the next section we assume \( z \)
normal distributions ignoring correlations between these
responses. From these Monte Carlo samples we compute
slack ratios.

We formulate pessimistic null-hypotheses: that is, we
‘accept’ an input combination only if it gives a significantly
lower objective value and all its \( z \cdot 1 \) slacks imply a ‘feasible’ solution. Actually, our interior point
method implies that a new slack value is a percentage –
say, 20% - of the old slack value; our pessimistic
hypotheses make our acceptable area smaller than the
original feasible area in (9).

After we have run \( G \) simulations along the search
path, we find a ‘best’ solution - so far. Now we wish to re-
estimate the search direction \( p \) defined in (12). Therefore
we again use a resolution-3 design in the \( k \) factors. We
still use CRN; actually, we take the same seeds as we used
for the very first local exploration. We save one
(expensive) run by using the best combination found so far,
as one of the combinations for the design.

We stop the whole search when either the computer
budget is exhausted or the search returns to an old
combination twice. When the search returns to an old
combination for the first time, we use a new set of seeds.

5 Monte Carlo Experiments for
Multiple RSM

Like in Part 1 (§3), we study the novel procedure -
explained in §4 - by means of a Monte Carlo example. As
in §3, we assume globally valid functions quadratic in two
inputs, but now we consider three responses; moreover, we
add deterministic ‘box’ constraints for the two inputs:
minimize \[ \mathbb{E} \left[ (d_1 ? 1)^2 ? (d_2 ? 5)^2 ? 4d_1 d_2 ? e_0 \right], \]
subject to \[ \mathbb{E} \left[ (d_1 ? 3)^2 ? d_2^2 ? d_1 d_2 ? e_1 \right] \leq 4, \]
\[ \mathbb{E} \left[ d_1^2 ? 3(d_2 ? 1.061)^2 ? e_2 \right] \leq 9, \]
\[ 0 ? d_1 ? 3, ? 2 ? d_2 ? 1 \]

where the noise has \( s_0 \sim 1, s_1 \sim 0.15, s_2 \sim 0.4, \) and correlations \( \rho_{0,1} \sim 0.6, \rho_{0,2} \sim 0.3, \rho_{1,2} \sim 0.1 \).

It is easy to derive the analytical solution as \( \hat{d}^* \sim (1.24, 0.52) \) with a mean objective value of 22.96 approximately.

We select the initial local area shown in the lower left corner of Figure 3. We run 100 macro-replicates; Figure 3 displays the macro-replicate that gives the median result for the objective; that is, 50% of the macro-replicates have worse objective values. In this figure we have \( \hat{d}^* \sim (1.46, 0.49) \) and an estimated objective of 25.30 approximately.

Figure 3: The “Average” (50th Quantile) of 100 Estimated Solutions

Table 2 summarizes the 100 macro-replicates, where Criterion 1 is the relative expected objective \( \mathbb{E} \left[ w_0 \hat{d}_1, d_2^* \right] \leq 22.96 \); Criterion 2 and 3 stand for the relative expected slacks \( \mathbb{E} \left[ w_1 \hat{d}_1, d_2^* \right] \leq 4 \) and \( \mathbb{E} \left[ w_2 \hat{d}_1, d_2^* \right] \leq 9 \) for the first and the second constraints. Ideally, Criterion 1 is zero; Criteria 2 and 3 are zero if the constraints are binding at the optimum. Our heuristic tends to end at a feasible combination: the table displays only positive quantiles for the Criteria 2 and 3. This feasibility is explained by our pessimistic null hypotheses (and our small significance level \( \alpha \sim 0.01 \).

Our conclusion is that the heuristic reaches the desired neighborhood of the real optimum in a relatively small number of simulation runs. Once the heuristic reaches this neighborhood, it usually stops at a feasible point.

CONCLUSIONS

In Part 1 of this paper we addressed the problem of searching for the simulation input combination that minimizes the output. RSM is a classic technique for tackling this problem, but it uses SD, which is scale dependent. Therefore we devised adapted SD (ASD), which corrects for the covariances of the estimated gradient components. ASD is scale independent. Our Monte Carlo experiments demonstrate that in general - ASD gives a better search direction than SD.

In Part 2, we account for multiple simulation responses. We use a mathematical programming approach; that is, we minimize one random objective under random and deterministic constraints. As in classic RSM, we locally fit functions linear in the simulation inputs. Next we apply interior point techniques to these local approximations, to estimate a search direction. This direction is scale independent. We take several steps into this direction, using binary search and statistical tests. Then we re-estimate these local linear functions, etc. Our Monte Carlo experiments demonstrate that our method indeed approaches the true optimum, in relatively few runs with the (expensive) simulation model.

REFERENCES


AUTHOR BIOGRAPHIES

EBRU ANGÜN is a Ph.D. student at the Department of Information Management at Tilburg University in the Netherlands since 2000. Her e-mail address is <M.E.Angun@uvt.nl>.

DICK DEN HERTOG is a Professor of Operations Research and Management Science at the Center for Economic Research (CentER), within the Faculty of Economics and Business Administration at Tilburg University, in the Netherlands. He received his Ph.D. (cum laude) in 1992. From 1992 until 1999 he was a consultant for optimization at CQM in Eindhoven. His research concerns deterministic and stochastic simulation-based optimization and nonlinear programming, with applications in logistics and production. His e-mail and web address are: <D.denHertog@uvt.nl> and <http://cwis.kub.nl/~few5/center/staff/hertog>.

GÜL GÜRKAN is an Associate Professor of Operations Research and Management Science at the Center for Economic Research (CentER), within the Faculty of Economics and Business Administration at Tilburg University, in the Netherlands. She received her Ph.D. in Industrial Engineering from the University of Wisconsin-Madison in 1996. Her research interests include simulation, mathematical programming, stochastic optimization, and equilibrium models with applications in logistics, production, telecommunications, economics, and finance. She is a member of INFORMS. Her e-mail and web address are: <ggurkan@uvt.nl> and <http://infolab.kub.nl/people/ggurkan/>.

JACK P.C. KLEIJNEN is a Professor of Simulation and Information Systems. His research concerns simulation, mathematical statistics, information systems, and logistics;
this research resulted in six books and nearly 200 articles. He has been a consultant for several organizations in the USA and Europe, and has served on many international editorial boards and scientific committees. He spent several years in the USA, at both universities and companies, and received a number of international fellowships and awards. His e-mail and web address are:
<kleijn@uvt.nl> and