Response modeling methodology (RMM)—maximum likelihood estimation procedures

Haim Shore*

Faculty of Engineering Science, Department of Industrial Engineering, Ben-Gurion University of the Negev
P.O.B. 653, Beer-Sheva 84105, Israel

Available online 20 August 2004

Abstract

Response modeling methodology (RMM) is a new approach for empirical modeling. ML estimation procedures for the RMM model are developed. For relational modeling, the RMM model is estimated in two phases. In the first phase, the structure of the linear predictor (LP) is determined and its parameters estimated. This is accomplished by combining canonical correlation analysis with linear regression analysis. The former procedure is used to estimate coefficients in a Taylor series approximation to an unspecified response transformation. Canonical scores are then used in the latter procedure as response values in order to estimate coefficients of the LP. In the second phase, the parameters of the RMM model are estimated via ML, given the LP estimated earlier. For modeling random variation, it is assumed that the LP is constant and a new simple percentile-based estimating procedure is developed. The new estimation procedures are demonstrated for some published data.

Keywords: Canonical correlation analysis; Generalized linear models; Response modeling methodology

1. Introduction

Response modeling methodology (RMM) is a new approach for empirical modeling recently introduced (Shore, 2002a, b, 2003, 2004a, b, c). Two features distinguish RMM
from current general approaches for empirical relational modeling (like generalized linear models, GLM, or data transformational approaches).

First, the RMM model has originated in relational models developed over the years in various engineering and scientific disciplines, and shown to include these models as special cases (Shore, 2004a). By contrast, existing general approaches for relational modeling have originated in models of random variation. For example, linear regression originated in the normal distribution, and the exponential family of distributions served as a platform to develop GLM.

Secondly, fitting an RMM model to given data determines not only the model’s parameters, but also the model structure, namely, the non-linear relationship between the response and the linear predictor (a linear combination of effects that transmit systematic variation to the response, henceforth denoted LP). This is to be compared with GLM, where both the link function, which determines the structure of the non-linear model, and the response distribution, are specified prior to data analysis.

For LP constant, the component of systematic variation vanishes, and the RMM model provides the quantile function of the error distribution. This function expresses the response quantile in terms of two, possibly correlated, normal errors. The allied density function (d.f.) and cumulative distribution function (CDF) were derived in Shore (2002b). The RMM error distribution has been shown (Shore, 2004b) to include as special cases some known statistical distributions, transformations and approximations like the inverse Box–Cox transformation, Johnson systems of distributions, Tukey’s g- and h-systems of distributions and the Cauchy distribution. Furthermore, the RMM error distribution provides a good platform for empirical modeling of random variation (Shore, 2000, 2004b). Such modeling capability is needed, for example, in statistical process control or process capability analyses, applied in non-normal environments.

An initial approximate ML estimating procedure for the RMM model has been developed in Shore (2002a). This procedure alternates between weighted non-linear least-squares, where the RMM “structural” parameters are estimated, and linear regression analysis, where LP is estimated. The procedure is approximate in the sense that some approximating assumptions are made to allow transitions between the two parts of the analysis. More efficient and rigorous estimation procedures are needed. In this paper we develop these procedures. They are divided into two: estimation procedures for RMM used for relational modeling, and estimation procedures for RMM used for modeling random variation.

To model systematic variation (relational modeling), estimation comprises two phases. In the first phase, the LP is constructed and estimated independently of the model used to relate the LP to the response (be it RMM or otherwise). This is accomplished in a newly developed procedure that combines canonical correlation analysis (CCA) with linear regression analysis. In the second phase, the parameters of the RMM model are estimated, given the LP from the earlier phase.

Applying RMM to model random variation, we assume that the LP is constant (no systematic variation is transmitted to the response). The resulting RMM error distribution is used as a platform to model data originating in an unknown distribution. Percentile-based and moment-based matching procedures have been developed in Shore (2002b). Here we develop a new simple percentile-based estimation procedure. Two real-life problems demonstrate application of the new procedures for relational modeling (the Economist Big Mac
Parity Index) and to model random variation (the distribution of Intra-Galactic Velocities for galaxy A1775).

In Section 2 we provide a brief review of the RMM model. In Section 3, the two phases of the estimation procedure for relational modeling are detailed. In Section 4 we describe the new percentile-based estimating procedure for modeling random variation. The new procedures are applied in Section 5 to two numerical examples (detailed above). Section 6 concludes with some summarizing observations.

2. The RMM model

Presentation of the general model in this section pursues references cited earlier. Only the main expressions associated with the new RMM are given here. The interested reader is advised to consult the above references for the technical details of the axiomatic derivation of the model.

The RMM model assumes that the relationship between the LP and the response is monotone convex (or concave). Expressing the response, \( Y \), in terms of the LP (\( \eta \)), two possibly correlated random normal errors, \( \varepsilon_1 \) and \( \varepsilon_2 \), and the model’s parameters, we obtain for the RMM model

\[
Y = \exp\left\{ ((x / \lambda) [(\eta + \varepsilon_1)^\lambda - 1] + \mu_2 + \varepsilon_2 \right\},
\]

or

\[
W = \log(Y) = (x / \lambda) [(\eta + \varepsilon_1)^\lambda - 1] + \mu_2 + \varepsilon_2,
\]

where \( \{x, \lambda, \mu_2\} \) are real-valued parameters, and the errors, \( \varepsilon_1 \) and \( \varepsilon_2 \), are normally distributed with zero mean and standard deviations \( \sigma_{\varepsilon_1} \) and \( \sigma_{\varepsilon_2} \), respectively.

Various error structures may be assumed that may produce variations of the model. Since \( \varepsilon_1 \) is the random error associated with the LP (\( \eta \)), we may assume that \( \varepsilon_1 \ll \eta \), and write [refer to (1)]:

\[
(\eta + \varepsilon_1)^\lambda = \eta^\lambda (1 + \varepsilon_1 / \eta)^\lambda \cong \eta^\lambda \exp(\lambda \varepsilon_1 / \eta).
\]

Similarly, for reasons detailed in the axiomatic derivation of the model (refer to the afore-cited references), we may assume that \( \varepsilon_2 \ll 1 \), and re-write

\[
\exp(\varepsilon_2) \cong 1 + \varepsilon_2.
\]

These two possible expressions for each error define altogether four possible error-structures, some of which may prove to be more useful than others. Furthermore, some of these error-structures often appear in engineering and scientific models (Shore, 2004a). We will introduce several variations of the basic RMM model (1) that result from these different definitions of the error-structure shortly.

Expressing the errors in terms of independent standard normal variates, \( Z_1 \) and \( Z_2 \) (the independence assumption will later be relaxed), we may write approximately, based on a Taylor series expansion, for the mean and for the variance of \( Y \) in (1) and for the mean
of \( W = \log(Y) \), respectively:

\[
E(Y) \cong \exp[\{ (\alpha / \lambda) \eta \hat{\lambda} - 1 \} + \mu_2 + (\sigma_{\varepsilon_2}^2 / 2)] \\
\times \{ 1 + (\sigma_{\varepsilon_1}^2 / 2)[\eta^{2(\lambda - 1)} \sigma^2 + \eta^{\lambda - 2} \sigma(\lambda - 1)] \}.
\]

\[
V(Y) = V[f_1(\eta, \varepsilon_1; \theta_1)]E[\{ f_2(\varepsilon_2; \theta_2) \}^2] + V[f_2(\varepsilon_2)]E[\{ f_1(\eta; \varepsilon_1) \}^2]
\]

and

\[
E(W) = E[(\alpha / \lambda)[(\eta + \varepsilon_1)^{\hat{\lambda}} - 1]] + \mu_2 \\
\cong (\alpha / \lambda)[(\eta^{\hat{\lambda}} - 1] + (1 / 2)(\alpha)(\lambda - 1)(\sigma_{\varepsilon_1}^2) \eta^{\lambda - 2} + \mu_2.
\]

(5)

where \( f_1 \) and \( f_2 \) are two functions of \( \varepsilon_1 \) and \( \varepsilon_2 \), respectively, the product of which results in the quantile function given in (1). Specifically,

\[
f_1(\eta, \varepsilon_1; \theta_1) = \exp[(\alpha / \lambda)[(\eta + \varepsilon_1)^{\hat{\lambda}} - 1]] = \exp[(\alpha / \lambda)[(\eta + \sigma_{\varepsilon_1} Z_1)^{\hat{\lambda}} - 1]]
\]

and

\[
f_2(\varepsilon_2; \theta_2) = \exp[\mu_2 + \varepsilon_2] = \exp[\mu_2 + \sigma_{\varepsilon_2} Z_2].
\]

(6)

with \( \theta_1 = \{ \alpha, \lambda, \sigma_{\varepsilon_1} \} \) and \( \theta_2 = \{ \mu_2, \sigma_{\varepsilon_2} \} \) as vectors of parameters, associated with \( f_1 \) and \( f_2 \), respectively. The approximate expressions in (5) were derived by expanding \( f_1 \) in terms of \( Z_1 \) into a Taylor series around zero, and taking expectation of the first four terms. Note that since both \( f_1 \) and \( f_2 \) are exponential functions, once the means are known there is little difficulty in finding the means of \( (f_1)^2 \) and \( (f_2)^2 \), required to calculate the variance. Thus, the \( r \)th non-central moment (moment about zero) of \( f_1 \) is obtained by replacing \( \alpha \) in the expression for the mean by \( r \alpha \), namely [refer to (5)]:

\[
E[[f_1(\eta, \varepsilon_1; \theta_1)]^r] \\
\cong \exp[(r \alpha / \lambda)[\eta^{\hat{\lambda}} - 1][1 + (\sigma_{\varepsilon_1}^2 / 2)[\eta^{2(\lambda - 1)}(r \alpha)^2 + \eta^{\lambda - 2}(r \alpha)(\lambda - 1)]]
\]

(7)

Examining the expressions for the means in (5), we realize that they are versatile in their capability to model variously-shaped link functions. In particular, they provide the link functions common within the GLM framework. Thus, if we ignore the term that contains \( \sigma_{\varepsilon_1}^2 \) in (5) (assume \( \sigma_{\varepsilon_1}^2 = 0 \)), we obtain for \( \alpha = 0 \) and \( \lambda = 1 \) a linear link function. The parameters \( \lambda = 0 \) and \( \alpha = -1 \) result in a reciprocal relationship between the LP and the mean, which is the canonical link for the gamma (exponential) error distribution. A value of \( \lambda = 1 \) would result in a log link function, typical to the Poisson distribution.

These special cases are indeed various “discrete steps” in “The ladder of fundamental uniformly convex functions” (Shore, 2002b, 2004a, b, c). The RMM model renders the discrete steps in the “Ladder” into mere points on a continuous conveyor that may take the modeler from a linear relationship to a power relationship, on to an exponential relationship and on to an exponential-power relationship. In fact, this cycle may be repeated indefinitely by introducing additional parameters. The basic pattern of “linear-power-exponential”, represented by the first three “steps” of the ladder, can be repeated at will if enough additional parameters are introduced into the model (two additional parameters are needed for each successive cycle, refer to the afore-cited references for details).
The RMM model may thus provide any desirable degree of convexity that seems to best fit the data. This “continuous spectrum” of possible models of monotone convexity, provided by RMM, is a unique feature that proves to be particularly valuable in fitting to engineering and scientific data. Furthermore, the final structure of the model, unlike GLM, is determined solely by the data on hand.

Let us next describe the RMM error distribution. The development herewith pursues Shore (2002b), and the reader is referred to this source for further detail.

First we discard the assumption that the two errors are independent, and assume that they derive from a bi-variate normal distribution. Specifically, writing the errors as: \( \varepsilon_1 = \sigma_{\varepsilon_1} Z_1 \)
\( \varepsilon_2 = \sigma_{\varepsilon_2} Z_2 \), we assume that \( Z_1 \) and \( Z_2 \) are distributed according to a bi-variate standard normal distribution with correlation \( \rho \). The conditional distribution of \( Z_2 \), given \( Z_1 = z_1 \), is then normal with mean \( \rho z_1 \) and variance \( 1 - \rho^2 \), and we may write for \( Z_2 \) in (1):
\[
Z_2 | (Z_1 = z_1) = \rho z_1 + (1 - \rho^2)^{1/2} Z,
\]
where \( Z \) is a standard normal variable, independent of \( Z_1 \). Introducing from (8a) into (1) and (2), we obtain, respectively,
\[
Y = \exp\{ (\alpha/\lambda) [ (\eta + \sigma_{\varepsilon_1} Z_1)^{1/2} - 1 ] + \mu_2 + \sigma_{\varepsilon_2} [ \rho Z_1 + (1 - \rho^2)^{1/2} Z_2 ] \}
\]
and
\[
W = \log (Y) = (\alpha/\lambda) [ (\eta + \sigma_{\varepsilon_1} Z_1)^{1/2} - 1 ] + \mu_2 + \sigma_{\varepsilon_2} [ \rho Z_1 + (1 - \rho^2)^{1/2} Z_2 ].
\]
where for consistency in notation we have replaced \( Z \) with \( Z_2 \), namely, \( Z_2 \) is assumed now to be a standard normal variable which is uncorrelated with (independent of) \( Z_1 \).

Equivalently, by conditioning \( Z_1 \) on a given value of \( Z_2 = z_2 \), we have
\[
Z_1 | (Z_2 = z_2) = \rho z_2 + (1 - \rho^2)^{1/2} Z,
\]
and we may write, instead of (10),
\[
W = (\alpha/\lambda) [ (\eta + \sigma_{\varepsilon_1} [ \rho Z_2 + (1 - \rho^2)^{1/2} Z_1 ])^{1/2} - 1 ] + \mu_2 + \sigma_{\varepsilon_2} Z_2.
\]
Like in (10), \( Z_1 \) and \( Z_2 \) are again uncorrelated (independent).

For \( \rho = \pm 1 \) and LP constant (arbitrarily assumed \( \eta = 1 \)), (9)–(11) reduce to the origin “Inverse Normalizing Transformation” (INT), developed and numerically demonstrated in Shore (2000, 2001a, b). These expressions then describe the quantile of a non-negative (possibly skewed) response in terms of the corresponding quantile of the standard normal distribution. In other words, the RMM model becomes a convenient platform for modeling random variation, as will be demonstrated in Section 4.

Referring to the variations in the definition of the errors, given in (3) and (4), we obtain, respectively, from (1):
\[
Y = (M) \exp\{ (\alpha/\lambda) [ \exp[\lambda \log(\eta) + \lambda \varepsilon_1 / \eta] - 1 ] + \varepsilon_2 \}
\]
and
\[
Y = (M) \exp\{ (\alpha/\lambda) [ (\eta + \varepsilon_1)^{1/2} - 1 ] (1 + \varepsilon_2) \}
\]
where \( M = \exp(\mu_2) \).
The expressions developed in this section will later serve to derive maximum likelihood estimates for the RMM model.

For \( \varepsilon_1 \) and \( \varepsilon_2 \) correlated, these equations become, respectively [refer to (8a)]:

\[
Y = (M) \exp\{ (\alpha/\lambda) [\exp\{ \lambda \log(\eta) + \lambda \sigma_1 Z_1/\eta \} - 1] + \sigma_2 \rho Z_1 + (1 - \rho^2)^{(1/2)} Z_2 \} \tag{14}
\]

and

\[
Y = (M) \exp\{ (\alpha/\lambda) [(\eta + \sigma_1 Z_1)\hat{\lambda} - 1] \{1 + \sigma_2 \rho Z_1 + (1 - \rho^2)^{(1/2)} Z_2 \}, \tag{15}
\]

with \( Z_1 \) and \( Z_2 \) being independent.

For \( \rho = \pm 1 \) and \( \eta = 1 \), \( M \) is the median of \( Y (Z_1 = 0) \), and expressions (14)–(15) become again INTs. Note, that due to the two definitions of the errors [given in (3) and (4)] and the two equivalent presentations in (8a) and (8b), the RMM model may be introduced in altogether eight different variations. The RMM models in (10), (11), (14) and (15) are just four variations out of these eight possible variations, which define different error structures for the RMM model. We will henceforth concentrate mainly on (9) and (14).

Next we develop the d.f. and the CDF associated with the responses, as given by (9) and (10). First, note that the conditional distribution of \( W|z_1 \) has normal distribution with mean and standard deviation given by, respectively:

\[
\mu(W|z_1, \theta) = (\alpha/\lambda) [(\eta + \sigma_1 z_1)\hat{\lambda} - 1] + \mu_2 + \sigma_2 \rho z_1 \tag{16}
\]

and

\[
\sigma(W|z_1, \theta) = \sigma_2 (1 - \rho^2)^{(1/2)}, \tag{17}
\]

where \( \theta = \{ \alpha, \lambda, \mu_2, \rho, \sigma_1, \sigma_2 \} \). The marginal (unconditional) d.f. of \( W \), and the respective CDF, may now be easily derived. The marginal (unconditional) d.f. of \( W \) is

\[
f_W(w; \theta) = \int_{-\infty}^{\infty} f_{W|z_1}(w|\varepsilon_1, \theta) f_{\varepsilon_1}(\varepsilon_1) d\varepsilon_1
\]

\[
= \int_{-\infty}^{\infty} [\sigma_2 (1 - \rho^2)^{(1/2)}]^{-1} \times \phi\{ [w - \mu(W|z_1, \theta)]/[\sigma(W|z_1, \theta)] \} \phi(z_1) dz_1, \tag{18}
\]

and \( \phi(.) \) is the standard normal d.f. The marginal CDF of \( W \) is

\[
F_W(w; \theta) = \int_{-\infty}^{\infty} \Phi\{ [w - \mu(W|z_1, \theta)]/[\sigma(W|z_1, \theta)] \} \phi(z_1) dz_1, \tag{19}
\]

and \( \Phi(.) \) is the standard normal CDF. The d.f. of \( Y \), denote it by \( f_Y(y) \), is, in terms of (18):

\[
f_Y(y) = (1/y) f_W[\log(y); \theta]. \tag{20}
\]

Similarly the distribution function, \( F_Y(y) \), is

\[
F_Y(y) = F_W[\log(y); \theta]. \tag{21}
\]

The expressions developed in this section will later serve to derive maximum likelihood estimates for the RMM model.
3. RMM estimation for relational modeling

3.1. A brief overview

In estimating the parameters of the RMM model, we have to distinguish between three sets of parameters:

I. Parameters associated with the LP:
\[ \eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k. \]  
(22)

II. Parameters associated with the non-linear structure of the model, \( \{\alpha, \lambda, \mu_2\} \).

III. Parameters associated with the error terms, \( \{\rho, \sigma_{\varepsilon 1}, \sigma_{\varepsilon 2}\} \).

The suggested estimation approach separates estimation of the three sets of parameters into three different procedures. The procedure to estimate the linear predictor is conducted once, in a combined analysis that involves canonical correlation analysis (CCA) and step-wise linear regression. The two procedures to estimate Set II and Set III are conducted sequentially in an iterative fashion. The structural parameters that belong to Set II are estimated via weighted non-linear least-squares (NL-LS). The error parameters that belong to Set III are estimated via maximization of an approximate log-likelihood function. However, the latter does not require numerical integration, as may be deduced from (18). Rather, it is based on expanding into a Taylor series the conditional d.f. of \( W \) and then taking expected value to remove the conditioning.

We will refer to estimation of Set I parameters as Phase 1 and to the estimation of the parameters in Set II and Set III as Phase 2. Let us elaborate in some detail on each of these phases. In the first phase, a non-linear transformation is applied to the response (not its mean!) so that the transformed response is equal to the LP plus a random error. However, the needed transformation is not specified (in contrast to other transformational approaches). Instead, we approximate the unknown (unspecified) transformation via a Taylor series expansion, and then identify the coefficients of the Taylor expansion via CCA (The reader is referred to Appendix A, where a brief introduction to CCA is provided). As implemented here, CCA attempts to maximize the correlation between a linear combination of the terms in the above Taylor expansion and a linear combination of effects that may potentially transmit variation to the response (the initial LP). Several CCA analyses can be attempted, having Taylor expansions for the transformed response that differ in the number of terms included. We will shortly address the problem of how to determine the number of terms in the Taylor expansion. The canonical scores, obtained from the selected CCA, and calculated individually from the “fitted” Taylor expansion for each observation, are then used as response values in stepwise linear regression analysis, when the final LP is estimated.

In the second phase of the analysis, the RMM model is perceived as a non-linear model with one regressor (the LP), and ML procedures are implemented to estimate the non-linear coefficients of the model (Set II) and the parameters associated with the errors (Set III).

Accordingly, the estimation procedures are introduced in two separate sections: an estimation procedure for the LP (Phase 1, Section 3.2), and estimation procedures for the “structural” parameters of the RMM model and the errors (Phase 2, Section 3.3).
3.2. Phase 1—estimating the linear predictor

Let $Y$ be a response (assumed to be expressed in standardized units so that it has zero mean and unit variance) and let $\{X_1, X_2, X_3, \ldots, X_k\}$ be a set of $k$ predictors (factors or parameter-free functions of factors) that may potentially transmit statistically significant variation to $Y$. Assume that preliminary examination of the data suggests that the relationship between $Y$ and the LP does not justify using linear regression analysis (the relationship is non-linear, the assumption of a constant error variance is violated by the data, or, more generally, no known normalizing transformation can revolve the normal scenario).

Let $G(Y)$ denote the (unspecified) transformation of the response that achieves additivity of effects: $G(Y) = \eta + \epsilon$, where $\epsilon$ is the error. Note that $G(Y)$ is not a link function (which relates to a transformation of the mean in the framework of GLM).

Developing $G(Y)$ into a Taylor series expansion around zero (the response mean), and keeping the first $p + 1$ terms, we obtain

$$G(Y) \cong G(0) + G_1(0)Y + G_2(0)Y^2/2 + \cdots + G_p(0)Y^p/p!,$$  

(23)

where $G_j(0)$ is the $j$th derivative, calculated around $Y = 0$. We will later discuss considerations in determining $p$. Define:

$$Y_{Tj} = Y^j/j! \quad (j = 0, 1, \ldots, p).$$  

(24)

Then $G(Y)$ may be expressed as

$$G(Y) \cong B_0 + B_1YT_1 + \cdots + B_pYT_p,$$  

(25)

with coefficients $\{B_j\}$, which will be estimated from the data.

Let the LP be as in (22). We wish to determine the parameters $\{B_j\} \ (j = 0, 1, \ldots, p)$ so that $G(Y)$, as represented by (25), may later serve as the response in a stepwise linear regression procedure that will determine the coefficients in the LP (22). In other words, our model is

$$(\text{Transformed response, approximated via Taylor series expansion}) = (\text{LP + error}),$$

or

$$B_0 + B_1YT_1 + \cdots + B_pYT_p = \eta + \epsilon = \beta_0 + \beta_1X_1 + \cdots + \beta_kX_k + \epsilon.$$  

(26)

The suggested estimating procedure comprises two stages. In the first stage, we implement CCA to determine the coefficients $\{\beta_j\} \ (i = 0, 1, \ldots, k)$ of an initial LP and the final coefficients of the “fitted” Taylor approximation, $\{B_j\} \ (j = 0, 1, \ldots, p)$. This stage of the analysis is expected to deliver canonical variables that maximally correlate $G(Y)$ [as approximated by (23)] with $\eta$ (22). If more than one canonical correlation is significantly different from zero, this stage may be conducted iteratively in order to maximally extract in one pair of canonical variables the significant co-variation between $G(Y)$ and the LP. In the second stage, we use canonical scores, obtained in the previous stage for the LHS of (26), as response values in stepwise multiple linear regression analysis to determine the final structure of the LP.
This estimating approach is supposed to deliver the structure of the LP (up to a linear transformation) that maximally correlates with an unspecified non-linear transformation of the response. The latter, being unspecified, is represented by a Taylor expansion that has been “fitted” (estimated) to maximally correlate with a linear combination of the effects. We have referred to the latter as “the initial LP” because it is not necessarily identical to the “final LP”, obtained in the linear regression analysis of the second stage. We will elaborate on this in more detail in the following subsections, when the various stages of Phase 1 are expounded in detail.

Stage I-Approximating a transformed response via a Taylor series expansion and estimating the coefficients via CCA

At this stage we derive the coefficients of the approximate \( G(Y) \) (25) that maximally correlate with \( \eta \) (22). The first decision to be made is how many terms to keep in the Taylor expansion of \( G(Y) \). Two opposing considerations play part in that decision. On the one hand, we wish to increase the number of terms in order to enhance the accuracy of the approximation. On the other hand, adding to the set of variables used in CCA may add noise which would obscure the true relationship between the two sets being correlated.

Examining a large sample of numerical examples, it seems that no strict rules may be specified, and several CCA should be run to determine the best composition of the Taylor approximation to \( G(Y) \). A major consideration is to have the first canonical correlation (that associated with the first eigenvalue) be highly significant while having the second canonical correlation (and those that follow) be highly insignificant (a \( \chi^2 \) sequential test is conducted to provide these significance levels; refer to the Appendix for further detail). Successive CCAs may therefore be run with an increasing number of power terms, and the best CCA be selected. The associated canonical scores may then be used as response values for the second stage of the analysis, when the final structure of the LP is decided. Note, that if a single first-degree term is selected for the Taylor approximation \( (p = 1) \), this implies that no response transformation is required.

Three observations need to be addressed at this point of the analysis. First, all potentially significant predictors may be included in the initial LP, used for the CCA. This seems like adding unnecessary sources of noise. However, if a certain predictor fails to transmit systematic variation to the response its weight in the canonical variable, which represents an initial estimate of the LP, will be negligible. Consequently, its influence in determining the final canonical scores, used as response values later on, will also be negligible. Conversely, we may decide to include in the initial estimate of the LP (in the first stage) only potentially significant main effects. This will not prevent inclusion of other effects (like interactions) in the stepwise linear regression analysis of the second stage. We will demonstrate this in the first numerical example of Section 5, where only main effects of potentially significant factors are included in the initial LP. The final LP (derived in the second stage) is derived based on stepwise regression, where candidate predictors to be included in the LP may comprise other than main effects, like interactions or second degree predictors. The LP thus derived will, in most cases, be surprisingly similar (up to a linear transformation) to the LP obtained via GLM.
Secondly, in most statistical packages, the raw canonical scores (the canonical scores calculated from the un-standardized variables) are standardized to have zero mean and unit variance. Correspondingly, \( G(Y) \), as approximated by the final canonical scores, will have zero mean and unit variance.

Finally, the number of canonical variates (or canonical roots, eigenvalues) is restricted by the number of terms in the response Taylor expansion. More specifically, the number of canonical correlations computed by CCA will be \( \min\{k, p\} \). For example, if expansion to degree four is employed \( (p=4) \), and the number of predictors in the initial LP is seven \( (k=7) \), then at most four significant canonical variables would be derived. As we have indicated and discussed earlier, a smaller number of terms may still deliver good approximation to the unknown (unspecified) response transformation and at the same time reduce the amount of noise contained in the derived canonical variables.

**Stage II—Stepwise linear regression analysis with canonical scores as response values**

In the second stage of the analysis, we determine the final structure of the LP. As response values, we use the standardized canonical scores, saved from the CCA of the earlier stage. Stepwise linear regression is applied, where all potentially significant predictors, including those that have not participated in Stage I, are candidates for inclusion in the LP. As an initial (preliminary) model, we may use a full factorial of second degree, or a full response surface model (namely, a quadratic model with all second degree interactions).

Once the LP has been derived in a stepwise fashion, we may perform diagnostics checking as commonly practiced in application of stepwise linear regression analysis. Having being satisfied with the linear model, we may proceed to the second phase of the modeling effort, where a non-linear RMM model, relating the LP to the response, is estimated.

**3.3. Phase 2—estimating the RMM model**

The sample-values of the LP, derived in the first phase of the analysis, now serve as input for a single-regressor RMM model. We first need to address the location and scale of the LP. As we may realize from (5), calculating the mean requires that the LP is raised to the power of \( \lambda \). When an ML estimate for the latter is searched, it is desirable that the LP retains only non-negative values. We therefore suggest that a change of location be applied by adding the minimal (negative) sample value of the LP to all sample values. It can easily be realized that this simple translation would not undermine the validity of the analysis by which the LP has been derived in Phase 1. To realize that note that two linear transformations applied, respectively, to two correlated variables, do not alter their correlation (provided the linearly transformed variables co-vary in the same relative direction as the untransformed variables). Consequently, adding a constant to all values of the LP would not alter the significance of the results obtained in the stepwise linear regression analysis of Phase 1. In particular, the regression multiple-correlation would not be altered. In the following, we assume that all values of the LP are non-negative.

Suppose that we have a sample of \( n \) observations, \( \{w_i, \eta_i\} \), where \( \eta_i \) is the value of the LP for the \( i \)th observation, and \( w_i \) is the log-transformed response-value for observation \( i \).
Since the LP has already been separately estimated in Phase 1, we need to estimate now
two sets of parameters (as detailed in Section 3.1):

Set II: Parameters associated with the non-linear structure of the model, \( \{z, \lambda, \mu_2\} \)

Set III: Parameters associated with the error terms, \( \{\rho, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}\} \).

The estimation procedure of Phase 2 is accordingly divided into two stages:

- Estimating the “non-linear” parameters (Set II) via weighted NL-LS. Weights used are
  reciprocal values of variance estimates, calculated from an updated estimate of the d.f.
  (18), or from the updated estimate of the quantile function (to be detailed shortly).

- Estimating the “error parameters” (Set III) via the log-likelihood function.

Estimation is implemented by iteratively alternating between these two stages, where for
each successive stage recently updated estimates from the previous stage are used. The two
stages will now be described in detail.

**Estimating the RMM “non-linear” parameters \( \{z, \lambda, \mu_2\} \)**

A weighted non-linear least-squares (NL-LS) would minimize

\[
S = \sum_{i=1}^{n} \left[ w_i - \mu_i(W) \right]^2 / [\sigma_i(W)]^2,
\]

where \( \mu_i(W) \) and \( \sigma_i(W) \) are the mean and the standard deviation associated with observation
\( w_i \). Myers et al. (2002) quote the work of Wedderburn (1974), which suggests that use of
weighted least-squares produces asymptotic properties that are quite similar to those of
maximum likelihood estimators. Thus, if good estimates of the variances of individual
observations are available, minimizing (27) would provide nearly ML estimates for the
parameters of the RMM model. Note again that since the LP has already been estimated,
only the RMM parameters, \( \{z, \lambda, \mu_2\} \), are estimated at this stage of the analysis.

Let us develop expressions for the mean and the standard deviation in (27). From (10),
taking expected value of \( W \), we obtain

\[
E(W) = \mu(W) = E[(z/\lambda)((\eta + \sigma_{\varepsilon_1}Z_1)^2 - 1)] + \mu_2,
\]

since \( E(Z_1) = E(Z_2) = 0 \). Eq. (28) implies that the mean does not include the parameters
\( \{\sigma_{\varepsilon_2}, \rho\} \). Therefore, we may use for \( \mu(W) \) in (27) the same approximate expression given
in (5) (the latter, we recall, has been derived assuming that the errors are independent).

Since the LP has already been estimated, only the four parameters: \( \{z, \lambda, \mu_2, \sigma_{\varepsilon_1}\} \) need be
estimated by minimization of (27). We prefer to estimate all parameters associated with the
errors, \( \{\rho, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}\} \), by a separate ML estimation routine that will be developed shortly.

This implies that in minimizing (27), we will estimate only the RMM parameters \( \{z, \lambda, \mu_2\} \),
and \( \sigma_{\varepsilon_1} \) will be estimated in the pursuing stage. Since (28) includes \( \sigma_{\varepsilon_1} \), an initial estimate
to use in (27) (for the first iteration only!) may be taken from the residual standard deviation
of the linear regression analysis of Phase 1. Alternatively, we may include \( \sigma_{\varepsilon_1} \), in the first
iteration only, as an additional parameter to be estimated via minimization of \( S(27) \).
Let us next relate to the determination of the weights in (27). Assume first that all parameters are known (or that updated estimates of all parameters are available). Then the variances of individual observations may be derived numerically from the quantile function in (10). Let us re-write the latter as

\[ W = Q_1(Z_1; \omega_1) + Q_2(Z_2; \omega_2), \]  

(29)

where

\[ Q_1(Z_1; \omega_1) = (\lambda/\lambda)(\eta + \sigma_{\varepsilon 1}Z_1)^2 - 1 + \mu_2 + \sigma_{\varepsilon 2}\rho Z_1 \]

and

\[ Q_2(Z_2; \omega_2) = \sigma_{\varepsilon 2}(1 - \rho^2)(1/2)Z_2, \]  

(30)

with, respectively, parameter vectors

\[ \omega_1 = \{\lambda, \mu_2, \sigma_{\varepsilon 1}, \sigma_{\varepsilon 2}\rho\} \]

and

\[ \omega_2 = \sigma_{\varepsilon 2}(1 - \rho^2)(1/2). \]  

(31)

Note that \(\omega_1\) contains five parameters while \(\omega_2\) has only one parameter.

Since \(Z_1\) and \(Z_2\) are independent random variables so are \(Q_1\) and \(Q_2\), and for the \(r\)th non-central moment of \(W\) we obtain:

\[ E(W^r) = \mu_r(W) = \sum_{j=0}^{r} \binom{r}{j} M_j(Q_1)M_{r-j}(Q_2), \]  

(32)

where \(M_j(Q_1)\) and \(M_{r-j}(Q_2)\) are the non-central \(j\)th and \((r-j)\)th non-central moments of \(Q_1\) and \(Q_2\), respectively, namely:

\[ M_k(Q_1) = \int_{-\infty}^{\infty} \{((\lambda/\lambda)(\eta + \sigma_{\varepsilon 1}z)^2 - 1] + \mu_2 + \sigma_{\varepsilon 2}\rho z\}^k \phi(z) \, dz, \]  

(33)

and

\[ M_k(Q_2) = \int_{-\infty}^{\infty} [\sigma_{\varepsilon 2}(1 - \rho^2)(1/2)]^k z^k \phi(z) \, dz = [\sigma_{\varepsilon 2}(1 - \rho^2)(1/2)]^k \mu_k(Z), \]  

(34)

where \(\mu_k(Z)\) is the \(k\)th moment of the standard normal variable.

The variance of \(W\) for individual observations may be estimated from (29)–(34), and then inserted into the minimization routine (27). Alternatively, the variances may be calculated numerically from the exact d.f. of \(W\) (18), or the approximate d.f. (35), which eliminates one numerical integration.

Various diagnostics checking, routinely available in any statistical package that provides non-linear regression, may be used in any iteration of the estimating procedure in order to judge the quality of the derived estimates. For example, correlations between estimates or estimates’ confidence intervals may be obtained. We will demonstrate this in the numerical examples of Section 5.
Estimating the RMM “error parameters” \{\rho, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}\}

Having obtained estimates for the RMM parameters, \{x, \lambda, \mu_2\}, we may proceed to estimate the parameters associated with the errors (\rho, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}). We estimate these via an ML routine, based on the log-likelihood function of the RMM error distribution.

In Section 2 we realized that the expression for the d.f. of \(W\) (18) requires numerical integration. This may render application of an MLE procedure a numerically prohibitive undertaking. We therefore opt for a Taylor expansion which would eliminate the need for integration. The result is a relatively simple expression for the d.f. of \(W\). This expression is based on the expected values of the first four terms in a Taylor expansion of the conditional density \(f_{W|e_1}(w|e_1, \theta)\). We obtain, approximately (refer to (18)):

\[
\begin{align*}
   f_W(w; \theta) & = \int_{-\infty}^{\infty} f_{W|e_1}(w|e_1, \theta) f_{e_1}(e_1) \, de_1 \\
   & \cong [2\pi(1 - \rho^2)\sigma_{\varepsilon_2}^2]^{-1/2} \exp[-(1/2)Z^2] \\
   & \times \{1 + (1/2)\sigma_{\varepsilon_1}^2[1 - \rho^2(1/2)\sigma_{\varepsilon_2}]^{-2} \\
   & \times \{((\eta_i)_{i=1}^{\lambda - 1} + \rho\sigma_{\varepsilon_2}/\sigma_{\varepsilon_1})^2(Z^2 - 1) \\
   & + \alpha\eta_i^{\lambda - 2}(\lambda - 1)(1 - \rho^2(1/2)\sigma_{\varepsilon_2}Z)\},
\end{align*}
\]

where \(\theta = \{x, \lambda, \mu_2, \rho, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}\}\) and

\[
Z = \{w - (x/\lambda)((\eta_i)_{i=1}^{\lambda - 1} - \mu_2)/[(1 - \rho^2)\sigma_{\varepsilon_2}^2]^{1/2}. \tag{36}\n\]

Accordingly, the log-likelihood function is

\[
\begin{align*}
   \text{LL}(\theta|w) & = \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} f_{W|e_1}(w_i|e_1, \theta) f_{e_1}(e_1) \, de_1 \\
   & \cong - (n/2) \log[2\pi(1 - \rho^2)\sigma_{\varepsilon_2}^2] + \sum_{i=1}^{n} [-(1/2)Z_i^2] \\
   & + \sum_{i=1}^{n} \log[1 + (1/2)\sigma_{\varepsilon_1}^2(1 - \rho^2(1/2)\sigma_{\varepsilon_2})^{-2}X \\
   & \times \{((\eta_i)_{i=1}^{\lambda - 1} + \rho\sigma_{\varepsilon_2}/\sigma_{\varepsilon_1})^2(Z_i^2 - 1) \\
   & + \alpha\eta_i^{\lambda - 2}(\lambda - 1)(1 - \rho^2(1/2)\sigma_{\varepsilon_2}Z_i)\}, \tag{37}\n\end{align*}
\]

where for the \(i\)th observation:

\[
Z_i = \{w_i - (x/\lambda)((\eta_i)_{i=1}^{\lambda - 1} - \mu_2)/[(1 - \rho^2)\sigma_{\varepsilon_2}^2]^{1/2}. \tag{38}\n\]

Maximization of (37) with respect to the error parameters would provide ML estimates for \{\rho, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}\}.
Summary of the estimation procedure (Phase 2)

We assume that values of the LP for individual observations are known (estimated) from Phase 1 of the analysis (Section 3.2):

I. Minimize $S$ (27) to obtain initial estimates for $\{\alpha, \lambda, \mu_2\}$. Use un-weighted NL-LS for this step. As an estimate for $\sigma_{\epsilon 1}^2$, use the residual variance, obtained from the stepwise linear regression of Phase 1. Alternatively, estimate $\{\alpha, \lambda, \mu_2, \sigma_{\epsilon 1}\}$ in the first iteration.

II. For the updated estimates of $\{\alpha, \lambda, \mu_2\}$, maximize the log-likelihood function (37) to find estimates for $\{\rho, \sigma_{\epsilon 1}, \sigma_{\epsilon 2}\}$.

III. Calculate variances for individual observations [use either (32) or (35)]. Use reciprocal values of the derived variances as weights for the next iteration (Step IV);

IV. Derive updated estimates for $\{\alpha, \lambda, \mu_2\}$, using weighted NL-LS to minimize $S$ (27). Use diagnostic checks embedded in the NL-LS procedure.

V. Test for convergence of the solution (various criteria may be used, for example, significant reduction in the residual variance). If convergence has not yet occurred go to step II. Otherwise end the estimation procedure.

4. RMM estimation for modeling random variation

4.1. A brief overview

In Section 4 we develop new percentile-based estimating procedures for (10) and (14), and relate briefly to moment-based matching procedures. Some other estimation procedures may be found in Shore (2002b). Assuming that LP is constant (arbitrarily set to 1) and that $\rho = \pm 1$, (10) and (14) become expressions for the quantile of a skewed variate in terms of a single standard normal variable, $Z$. We obtain, respectively,

$$W = (\alpha/\lambda)[(1 + \sigma_{\epsilon 1} Z)^{\lambda} - 1] + \mu_2 + \sigma_{\epsilon 2} \rho Z$$

and

$$W = \log(Y) = (\alpha/\lambda)[\exp(\lambda \sigma_{\epsilon 1} Z) - 1] + \mu_2 + \sigma_{\epsilon 2} \rho Z.$$  \hspace{1cm} (40)

Note, that in fitting (39) to real data it is indistinguishable from its equivalent form, given by (11). Thus, with $\eta = 1$ and $\rho = \pm 1$, (11) becomes

$$W = (\alpha/\lambda)[(1 + \sigma_{\epsilon 1} \rho Z)^{\lambda} - 1] + \mu_2 + \sigma_{\epsilon 2} Z.$$  \hspace{1cm} (41)

From (39) and (41) we can only state that the two coefficients of $Z$, that appear either in (39) or (41), cannot both be negative. However, when (39) (or, equivalently, (41)) is fitted to a given distribution, and $\rho = -1$, either one coefficient or the other may be negative. This observation is confirmed when either of these quantile functions are fitted to actual skewed distributions (refer to Shore, 2000, for examples). A similar observation may be made with respect to (40) and its corresponding equivalent counterpart (not developed here).

In deciding whether to use (39) or (40) to model random variation, two considerations may lead us to choose the latter over the former. First, numerical search for $\lambda$ in (39) may
be troublesome if the expression in brackets become negative. Secondly, (40) has one less parameter to estimate. If we estimate $\mu_2$ via the sample median, than (40) has only three parameters that need estimating, evidently a desirable feature. On the other hand, if $\sigma_{x1}$ may be expected to be relatively large (in contradiction to the assumption which led to the derivation of (40) from the basic model in (39)), then the original RMM model (39) should be used. In general, goodness-of-fit criteria should guide us in selecting (39) or (40). In developing the percentile-based estimating procedure (Section 4.2), we use (40) as a model for random variation.

Written with re-defined parameters, (40) becomes

$$W = \log(Y) = \log(M) + B[\exp(CZ) - 1] + DZ.$$  \hfill (42)

Note again that $M$ may be estimated by the sample median. Let us first address a percentile-based matching procedure. Using (42) to model random variation, given sample data, we may employ the sample order statistics to estimate percentile values, identify the corresponding standard normal percentiles, and then integrate both sets in a NL-LS procedure to identify the parameters $\{M, B, C, D\}$. If we wish to estimate $M$ by the sample median of $Y$, this leaves us with only three parameters that need estimating. The percentile-based estimating procedure will be detailed shortly.

Next, we refer to moment-based matching procedures. In (32), a general expression for the non-central $r$th moment of $W$ is given. We may wish to use the first three moments of $W$ to identify the above three parameters (assuming that $\mu_2$ is estimated via the sample median). However, we have repeatedly expressed our reservation about using sample moments of high degree in moment-based matching procedures (refer for a recent reference to this problem in Shore, 2004d).

Alternatively, we may opt to use low degree moments (second degree at most), partial and complete, in a moment-based matching procedure. It has been repeatedly demonstrated by us that this alternative approach for moment-based estimation results in appreciably reduced mean-squared-errors (MSEs) for the fitted distribution (refer to references given in the afore-cited publication). Some such procedures have been developed in Shore (2000), and the reader is referred to this source for details. Accordingly, we confine ourselves in this section only to the new percentile-based matching procedure. This will be expounded in the next sub-section.

4.2. Percentile-based estimation procedure

This procedure requires identifying the CDF values, $\{P_i\}$, represented by given ordered sample-values (the order statistics) and then finding the corresponding standard normal quantiles. The sample of paired quantiles is then integrated into an NL-LS procedure to estimate the parameters in (42). To determine the $P_j$ value associated with a given order statistic, $y(j)$ (there are $j - 1$ smaller observations in the sample), most statistical packages use the simple approximating expression, based on what is known as the “Median Rank” (approximated by (43)), or by White’s plotting position (44):

$$P_j = (j - 0.3)/(n + 0.4)$$  \hfill (43)
\( P_j = \frac{(j - 3/8)}{(n + 1/4)}. \) \hspace{1cm} (44)

In most cases, the difference between the two approximations will be negligible (find details in Dodson, 1994). Once \( \{P_j\} \) are known, the corresponding standard normal quantiles are easily identified from the inverse standard normal CDF, \( \Phi^{-1}(P) \).

The complete procedure is now detailed:

- Arrange the observations in an ascending order (from smallest to largest)
- For the order statistic, \( y(j) \) \( (j = 1, 2, \ldots, n) \) calculate \( P_j \) from (44)
- Find the corresponding standard normal quantile \( Z_{P_j} = \Phi^{-1}(P_j) \)
- Generate the sample \( \{z_p, y_p\} \) for the NL-LS fitting, where \( z_p \) and \( y_p \) are standard normal and response quantiles, respectively, which correspond to \( P \)
- Apply NL-LS to the sample of quantile values and obtain estimates for \( \{B, C, D\} \) in (42). Use for \( M \) the sample median, or include \( M \) as an additional parameter that needs to be estimated in the NL-LS procedure.

A numerical example (for preparing the sample for a NL-LS procedure)

Suppose that we have a sample of \( n = 7 \) ordered observations: 2,5,6,13,17,22,30. Using (44), the corresponding \( P \) values are (median is estimated from the sample)

\[ P = \{0.08620, 0.2241, 0.3621, 0.5, 0.6379, 0.7759, 0.9138\} \]

The corresponding standard normal quantile values (\( z \)) are

\[ z = \{-1.3645, -0.7583, -0.3529, 0, 0.3529, 0.7583, 1.3645\} \]

The sample for the NL-LS fitting is

\[ \{(z_p, y_p)\} = \{(-1.3645, 2), (-0.7583, 5), (-0.3529, 6), (0, 13), (0.3529, 17), (0.7583, 22), (1.3645, 30)\}. \]

5. Numerical examples

In this section we present RMM-based analyses of two real-world problems. For both examples, data were collected from field observations (not from designed experiments). Example 1 relates to relational modeling and Example 2 to modeling of random variation.

5.1. Example 1: relational modeling—the Economist Big Mac parity index

This example is given in Cook and Weisberg (1999, henceforth CW), and is reproduced here with permission. The data may be found in the file \texttt{big-mac.lsp} at the web-site, associated with the above reference: \url{http://www.stat.umn.edu/arc/index.html}. 
Table 1
Example 1—the Big Mac data

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Mac (B-M) (response)</td>
<td>Minutes of labor required by an average worker to buy a BigMac and French fries</td>
</tr>
<tr>
<td>Bread (B)</td>
<td>Minutes of labor required to buy one kilogram of bread</td>
</tr>
<tr>
<td>BusFare (BF)</td>
<td>The lowest cost of a ten-kilometer bus, tram or subway ticket, in U.S. dollars</td>
</tr>
<tr>
<td>EngSal (ES)</td>
<td>The average annual salary of an electric engineer, in thousands of U.S. dollars</td>
</tr>
<tr>
<td>EngTax (ET)</td>
<td>The average tax rate paid by engineers</td>
</tr>
<tr>
<td>Service (S)</td>
<td>Annual cost of 19 services, primarily relevant to Europe and North America</td>
</tr>
<tr>
<td>TeachSal (TS)</td>
<td>The average annual salary of a primary school teacher, in thousands of U.S. dollars</td>
</tr>
<tr>
<td>TeachTax (TT)</td>
<td>The average tax rate paid by primary teachers</td>
</tr>
<tr>
<td>VacDays (VD)</td>
<td>Average days of vacation per year</td>
</tr>
<tr>
<td>WorkHrs (WH)</td>
<td>Average hours worked per year</td>
</tr>
<tr>
<td>City</td>
<td>Name of city</td>
</tr>
</tbody>
</table>

Table 2
Example 1—results from CCA with different number of terms in the Taylor approximation to \(G(Y)\) (the optimal choice is starred). \("YT_i"\) means terms up to and including \(YT_i\)

<table>
<thead>
<tr>
<th>Statistic</th>
<th>(&quot;YT_1&quot;)</th>
<th>(&quot;YT_2&quot;)</th>
<th>(&quot;YT_3&quot;)</th>
<th>(&quot;YT_4&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(R^2) (Canon. Corr. squared)</td>
<td>.5727</td>
<td>.8008</td>
<td>.8022</td>
<td>.8258</td>
</tr>
<tr>
<td>Significance</td>
<td>.000149</td>
<td>.00000</td>
<td>.00000</td>
<td>.00000</td>
</tr>
<tr>
<td>Sig. (1st root removed)</td>
<td>.2840</td>
<td>.02281</td>
<td>.07883</td>
<td></td>
</tr>
<tr>
<td>Sig. (2nd removed)</td>
<td>.2712</td>
<td>.5039</td>
<td></td>
<td>.9027</td>
</tr>
<tr>
<td>Sig. (3rd removed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The file includes economic data on 45 world cities from the period 1990–1991. The Big Mac hamburger is a simple commodity that is virtually identical worldwide. One might expect that the price of a Big Mac be the same everywhere, but obviously it is not. The Economist magazine publishes a Big Mac Parity Index, which compares the costs of a Big Mac in various places, as a measure of inefficiency in currency exchange. In this example, we use economic indicators that describe each city in order to model, via RMM, the price of a Big Mac in terms of the given indicators. The data in the file are described in Table 1.

We will now apply the RMM estimation procedure that would lead us to the final model. Starting with Phase 1, we attempt several CCAs, starting with \(p = 1\) (only \(YT_1\) included in the Taylor expansion) and concluding with \(p = 4\). The results are displayed in Table 2. Although several measures for the quality of the results may be employed to deter-
mine the final structure of the Taylor approximation to the transformed response, we use two indicators: the increase in the canonical correlation as we add terms to the expansion, and the significance level of the first root extracted, relative to the significance level obtained when this root is removed. We wish the significance level for the first root to be small and for the rest of the roots (with the first one removed) to be high. It is obvious from Table 2 that selecting terms up to and including $YT_2$ (namely a quadratic equation) would deliver the best Taylor approximation for the transformed response. The associated expression for the canonical scores, obtained from the “fitted” approximation for $G(Y)$, is

$$G(Y) = -1.6320YT_1 + 1.0624YT_2.$$  \hfill (45)

The canonical scores calculated from (45) for individual observations are now used in stepwise multiple linear regression analysis. The initial model for generating potential effects to be included in the analysis is factorial with second degree.

Diagnostic checking of the resulting model shows that out of the 45 cities that participate in the analysis, seven cities have studentized residuals that exceed about $\pm 2$. These are (order number, name, studentized residual):

- (#2, Athens, 1.94), (#6, Buenos Aires, $-2.32$), (#22, Luxembourg, $-2.08$), (#23, Madrid, $-2.05$), (#25, Mexico City, 2.0), (#35, Sao Paulo, $-2.10$), (#42, Tokyo, 1.98).

(A negative residual implies that the response is smaller than the model suggests.) Excluding the above from pursuing analyses, we obtain the final model (significance level appears below each effect, in brackets):

$$G(Y) = LP = -0.003876(B) + 0.7000(BF) + 0.01900(B - 26.4)(BF - 1.0095)$$
$$+ 0.002987(S) - 0.007796(S - 245.5)(BF - 1.0095)$$
$$+ 0.03222(TS).$$

$$p < 0.0001 \quad (p < 0.0005) \quad (p < 0.0001) \quad (p < 0.0001)$$

Note, that the intercept is zeroed since it was non-significant. To assess properly the influence of the various predictors (which are measured on different scales) we present also the LP with scaled estimates (scaling is performed by subtracting the mean and dividing by half the range in the data). We obtain

$$G(y) = 0.5306 - 0.4108(B) + 0.8995(BF) + 2.588(B - 26.4)(BF - 1.0095)$$
$$- 0.5525(S) - 1.8533(S - 245.5)(BF - 1.0095) + 0.9230(TS).$$  \hfill (47)

One can learn from (47) that the interactions play important role in modeling Big Mac. Also, given the main effects and the interactions that are significant, this model makes sense on intuitive grounds. Observing goodness-of-fit statistics, the model has $R^2$-adjusted of 0.915 with residual standard deviation of 0.291. The model $F$-ratio is 66.2, which, with model-d.f. of $v_1 = 6$, and error-d.f. of $v_2 = 32$, is highly significant ($p$ of the order of $10^{-16}$).

We may now proceed to the second phase of the analysis. Adopting the model given by (2), and assuming that the errors are independent ($\rho = 0$), we first use NL-LS with the mean, $E(W)$, as given by (5). Adding 2.17 to the LP (46) in order to ensure non-negative values
Table 3
Example 1—final estimates of the “structural” parameters, \(\{x, \mu_2\}\), with standard errors and 95% confidence intervals

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E.</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x)</td>
<td>(-0.6172)</td>
<td>0.0283</td>
<td>([-0.674, -0.559])</td>
</tr>
<tr>
<td>(\mu_2)</td>
<td>4.385</td>
<td>0.0424</td>
<td>([4.30, 4.47])</td>
</tr>
</tbody>
</table>

(a justification for this was given earlier), we obtain, from a first run of NL-LS, estimates for \(\hat{\lambda}\) very near 1 (\(\hat{\lambda} = 0.987\)) and for \(\sigma_{x1}\) very near zero. We therefore assume: \(\hat{\lambda} = 1, \sigma_{x1} = 0\), and re-run NL-LS to obtain the following estimates:

\[ x = -0.6172, \quad \mu_2 = 4.385, \quad \sigma^2_{res} = 0.02691. \]

Some statistics relating to these estimates are given in Table 3. The above results imply that a log transformation of the response would deliver additive effects, as the later are modeled by (46). Therefore there is no need to proceed to the second stage of the analysis, and we may assume that the residual variance, \(\sigma^2_{res}\), obtained in the NL-LS procedure, estimate well the parameter \(\sigma^2_{\mu_2}\).

Fig. 1 displays a scatter plot of the relative errors \((n = 38)\) for the Big Mac example (RMM model). Horizontal axis is Big Mac in the original scale.

\[(B - M)^{-1/3} = \beta_0 + \beta_1(ES)^{1/3} + \beta_2(TS)^{1/3} + \text{error}, \tag{48}\]

where the error is expected to conform to the normal scenario, namely, linear regression analysis may be applied. Implementing linear regression, we realize that \((ES)^{1/3}\) is
non-significant \((p = 0.432)\). Furthermore, there are three cities with outlying studentized residuals: (#23, Madrid, \(-1.932\)), (#38, Stockholm, \(-2.56\)), (#39, Sydney, 2.27).

Re-running the model with a single predictor, \((TS)^{(1/3)}\), we realize that two additional cities have exceeding residuals: (#14, Helsinki, \(-2.089\)), (#25, Mexico City, \(-2.319\)). Note that two of the cities, (#23, Madrid), (#25, Mexico City) appear as outliers in both the RMM and in the CW models. The final model (with \(n = 40\)) is

\[
(B - M)^{(1/3)} = 0.1366 + 0.06571(TS)^{(1/3)}.
\]

This model has 62-adj. of 0.860, residual standard deviation of 0.0219 and \(F\)-ratio of 240.2 (with model-d.f. of \(v_1 = 1\), and error-d.f. of \(v_2 = 38\)). This \(F\) value is highly significant \((p < 0.0001)\). Fig. 2 displays a scatter plot of the relative errors, where the latter are calculated in the original scale of the response (Big-Mac). Comparing the average absolute value of the relative error obtained from the two models (including the 38 observations analyzed via the RMM), we realize that for the RMM model we obtain a value of 13.2\% vs. 20.0\% for the CW model. The issue of which model is more plausible, (47) or (49), is left open.

5.2. Example 2: modeling random variation—the intra-galactic velocities

This example appears in Karian and Dudewicz (2000, p. 208, henceforth KD), and is reproduced here with permission. In astronomy, the cluster named A1775 is believed to consist of two clusters that are in close proximity. Oegerle et al. (1995) gave velocity observations (in kilometers per second) from this galaxy. Some of these \((n = 51)\) are shown in an increasing order in Table 4 (read lines).

We first identify the CDF value associated with each order statistic and the corresponding standard normal quantiles. For example, for the fourth order statistic, 19026, we have \(P_4 = (4 - 3/8)/(51 + 1/4) = 0.07073\). The standard normal quantile is \(\Phi^{-1}(0.07073) = -1.47036\).

Attempting to fit (42) to the data, a first run of NL-LS shows that the parameters (BC) and \(D\) are very close in their absolute values. Following good empirical-modeling practices,
Table 4
Data for example 2

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>18499</td>
<td>18792</td>
<td>18933</td>
<td>19026</td>
<td>19111</td>
<td>19111</td>
<td>19130</td>
<td></td>
</tr>
<tr>
<td>19179</td>
<td>19404</td>
<td>19408</td>
<td>19595</td>
<td>19595</td>
<td>19619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19673</td>
<td>19740</td>
<td>19807</td>
<td>19866</td>
<td>20210</td>
<td>20210</td>
<td>20875</td>
<td></td>
</tr>
<tr>
<td>21911</td>
<td>22192</td>
<td>22193</td>
<td>22417</td>
<td>22417</td>
<td>22426</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22513</td>
<td>22625</td>
<td>22682</td>
<td>22738</td>
<td>22738</td>
<td>22744</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22779</td>
<td>22781</td>
<td>22809</td>
<td>22922</td>
<td>22922</td>
<td>23017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23059</td>
<td>23121</td>
<td>23220</td>
<td>23303</td>
<td>23303</td>
<td>23408</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23432</td>
<td>24909</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5
Example 2—estimates and 95% confidence intervals for the RMM parameters derived from NL-LS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Asymptotic SE</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>0.03657</td>
<td>0.004260</td>
<td>{0.02801, 0.04513}</td>
</tr>
<tr>
<td>$C_1$</td>
<td>-0.6838</td>
<td>0.1875</td>
<td>{−1.0606, −0.3070}</td>
</tr>
</tbody>
</table>

where a model is modified according to results obtained in previous analyses, we re-define from (42) a two-parameter model (apart from $M$):

$$\log(Y) = A_1 \{(1/C_1)[\exp(C_1 Z) - 1] + Z]\} + \log(M),$$

(50)

where $M$ is the median (22417). Note that a subscript was added to the parameters to distinguish them from the parameters used in the original model (42).

Fitting (50) to the data via NL-LS we obtain the results (estimates and confidence intervals) in Table 5. An analysis of variance from the NL-LS procedure delivers a model $F$ value of $(1.05) \times 10^6$. With 2 and 49 degrees-of-freedom for the mean-squared-errors of the model and the error, respectively, the model is highly significant. Fig. 3 displays the relative errors (in %). An error is defined as the difference between the quantile calculated from the fitted value (50) and the actual value. The mean absolute value of the relative errors is 0.39%. Calculating numerically the first two moments of $W$ from the fitted equation we obtain: calculated \{μ, σ^2\} : {10.0, 0.007439}; sample \{μ, σ^2\} : {9.970, 0.006929}.

The sample skewness and kurtosis values are not well preserved in the fitted model. This is to be expected given the small sample size and the well-known high MSE of direct sample estimates of skewness and kurtosis.

We wish to compare these results to the analysis given in KD. They conclude that the generalized beta distribution (GBD) may model well this dataset (therein, p. 208). The four-parameter d.f. and CDF of GBD is given by (therein, p. 119):

$$f(y) = (y - \beta_1)^{\beta_3} (\beta_1 + \beta_2 - y)^{\beta_4} / [\beta(\beta_3 + 1, \beta_4 + 1)\beta_2^{(\beta_3 + \beta_4 + 1)}],$$

$\beta_1 \leq y \leq \beta_1 + \beta_2$
Fig. 3. Example 2—relative errors (in %) of the fitted (50) vs. actual log of the intra-galactic velocities (RMM analysis).

Fig. 4. Example 2—relative errors (in %) of the fitted (51) vs. actual log of the intra-galactic velocities (KD model, based on GBD).

\[ F(y) = \int_{\beta_1}^{y} f(u) \, du, \quad (51) \]

where \( \beta \) is the beta function: \( \beta(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b) \).

For the above data, KD obtain the fitted GLB:

\[ \beta_1 = 18675.3862; \quad \beta_2 = 4880.6411; \quad \beta_3 = -0.4789; \quad \beta_4 = -0.6045. \]

Fig. 4 displays the relative errors (in %). An error is defined as the difference between the quantile, \( y_i \), calculated from solving: \( F(y_i) = P_i \), and the actual \( i \)th order statistic \( P_i \) is the CDF value for the \( i \)th order statistic, calculated from (44)). We realize that the fit is
about the same as that obtained via (50) (Fig. 3). The mean absolute value of the relative errors is 0.28% (vs. 0.39% for the RMM model).

6. Conclusions

In this paper we developed ML procedures to estimate the parameters of the RMM model. For relational modeling, the estimation methodology is divided into two phases. In the first phase, the structure of the LP is determined, independently of the non-linear model selected for the data. In Phase 2 of the analysis, ML estimation is conducted to determine the parameters of the non-linear model. For modeling random variation, both percentile-based and moment-based matching procedures were addressed, although in this paper we developed only a new percentile-based procedure. Two numerical examples have demonstrated the effectiveness of the RMM approach relative to current approaches.

There are two unique contributions to this paper. First, we have demonstrated that the LP may be derived independently of any assumptions regarding the link function or the error distribution. In fact, for the first example an LP was derived that seems more in line with the nature of the problem than the alternative solution, based on a transformational approach.

Secondly, we have shown that RMM is effective in modeling both systematic variation (relational modeling) and random variation. In both cases, the new approach is compatible with the most basic tenet of empirical modeling, namely, that available data determine the final structure of the model, and no assumptions are made (regarding properties of the final model) prior to statistical analysis of the data.

RMM provides data-based analysis which is uniform across a large spectrum of possible scenarios, and it has properties that seem to be rare in other currently available approaches for empirical modeling.

Acknowledgements

This research was partly supported by the Paul Ivanier Center of Robotics at Ben-Gurion University.

Appendix. Canonical correlation analysis—a brief review

In this appendix we deliver a brief background on canonical correlation analysis, as given by SAS user guide. Canonical correlation is a technique for analyzing the relationship between two sets of variables—each set can contain several variables. As such, canonical correlation is a variation on the concept of multiple regression and correlation analysis. In multiple regression and correlation, you examine the relationship between a linear combination of a set of X variables and a single Y variable. In canonical correlation analysis, you examine the relationship between a linear combination of the set of X variables with a linear combination of a set of Y variables. Simple and multiple correlations are special cases of canonical correlation in which one or both sets contain a single variable.
Consider the situation in which you have a set of $p \ X$ variables and $q \ Y$ variables. The canonical correlation procedure finds the linear combinations

\[
v_1 = a_1x_1 + a_2x_2 + \cdots + a_px_p,
\]

\[
w_1 = b_1y_1 + b_2y_2 + \cdots + b_qy_q,
\]

such that the correlation between the two canonical variables, $v_1$ and $w_1$, is maximized. This correlation between the two canonical variables is the first canonical correlation. The coefficients of the linear combinations are canonical coefficients or canonical weights. It is customary to normalize the canonical coefficients so that each canonical variable has a variance of 1. The procedure continues by finding a second set of canonical variables, uncorrelated with the first pair, which produces the second highest correlation coefficient. The process of constructing canonical variables continues until the number of pairs of canonical variables equals the number of variables in the smaller group. Each canonical variable is uncorrelated with all the other canonical variables of either set except for the one corresponding canonical variable in the opposite set. The canonical coefficients are not generally orthogonal, however, so the canonical variables do not represent jointly perpendicular directions through the space of the original variables. The first canonical correlation is at least as large as the multiple correlation between any variable and the opposite set of variables.

Finally, statistical tests are applied to test a series of hypotheses that each consecutive canonical correlation and all smaller canonical correlations are zero in the population. This implies that testing is conducted sequentially, where at each stage the largest canonical correlation is removed in order to test the statistical significance of all the rest (smaller) canonical correlations.

References

Dodson, B., 1994. Weibull Analysis, ASQ Press, Milwaukee, WI.