Multivariate Statistical Analysis Applied to an IL6 Signal Transduction Model in Hepatocytes

McArdle, A., Kruger, U., & Hahn, J. (2009). Multivariate Statistical Analysis Applied to an IL6 Signal Transduction Model in Hepatocytes. Statistics in Medicine, 28 (19)(19), 2401-2434. DOI: 10.1002/sim.3621

Published in:
Statistics in Medicine

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:2401–2434
Published online 22 June 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3621

Multivariate statistical analysis applied to an IL6 signal transduction model in hepatocytes

Alison McArdle[1,*,†], Uwe Kruger[2,‡,§] and Juergen Hahn[3]

[1]*School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT9 5AH, U.K.*
[2]*Department of Electrical Engineering, The Petroleum Institute, P.O. Box 2533, Abu Dhabi, U.A.E.*
[3]*Department of Chemical Engineering, Texas A&M University, College Station, TX 77843-3122, U.S.A.*

SUMMARY

This paper introduces the application of linear multivariate statistical techniques, including partial least squares (PLS), canonical correlation analysis (CCA) and reduced rank regression (RRR), into the area of Systems Biology. This new approach aims to extract the important proteins embedded in complex signal transduction pathway models.

The analysis is performed on a model of intracellular signalling along the janus-associated kinases/signal transducers and transcription factors (JAK/STAT) and mitogen activated protein kinases (MAPK) signal transduction pathways in interleukin-6 (IL6) stimulated hepatocytes, which produce signal transducer and activator of transcription factor 3 (STAT3).

A region of redundancy within the MAPK pathway that does not affect the STAT3 transcription was identified using CCA. This is the core finding of this analysis and cannot be obtained by inspecting the model by eye. In addition, RRR was found to isolate terms that do not significantly contribute to changes in protein concentrations, while the application of PLS does not provide such a detailed picture by virtue of its construction.

This analysis has a similar objective to conventional model reduction techniques with the advantage of maintaining the meaning of the states prior to and after the reduction process. A significant model reduction is performed, with a marginal loss in accuracy, offering a more concise model while maintaining the main influencing factors on the STAT3 transcription.

The findings offer a deeper understanding of the reaction terms involved, confirm the relevance of several proteins to the production of Acute Phase Proteins and complement existing findings regarding cross-talk between the two signalling pathways. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: multivariate statistical analysis; IL6 signal transduction; model reduction; canonical correlation analysis

*Correspondence to: Alison McArdle, Electrical Engineering, Queen's University Belfast, Ashby Building, Stranmillis Road, Belfast, BT9 5AH, U.K.
†E-mail: amcardle02@qub.ac.uk
‡Correspondence to: Uwe Kruger, Department of Electrical Engineering, The Petroleum Institute, P.O. Box 2533, Abu Dhabi, U.A.E.
§E-mail: ukruger@pi.ac.ae

Received 13 June 2008
Accepted 9 April 2009

1. INTRODUCTION

The rapid progression and popularization of biological research at a systems level is due in part to the availability of comprehensive data sets and the potential scope for accessing hidden information. Systems Biology is transforming the approach to medical diagnostics through the focus on the intra- and extra-cellular communication between the cells of multicellular organisms [1–3]. By producing mathematical models of these complex signal transduction pathways, through the construction of component balances for the relevant proteins, the changes in cytoplasmic components and the resulting initiation and regulation of protein transcription in the nucleus can be studied.

Research on the molecular mechanisms of signal transduction is a very important topic that has attracted significant interest from biologists, bioengineers and biotechnologists [4–7]. Although considerable progress in the identification of the molecular components involved in cell functions has been made over the past decades, the resulting dynamic models are highly complex and it is not possible to substantiate if each aspect of the model is correct. Reducing models can help as it allows us to focus on essential aspects that can be verified.

This work introduces the use of multivariate statistical analysis concepts to Systems Biology for the purpose of simplifying signal transduction models. The application of multivariate data analysis tools, such as Partial Least Squares (PLS), Canonical Correlation Analysis (CCA) and Reduced Rank Regression (RRR) can play a crucial role in providing a detailed analysis of the model in order to extract important information and underlying relationships between the variables that may otherwise go undetected [8], or may only be acquired through expensive and tedious trials in a laboratory. More precisely, the use of these tools allows for the extraction and isolation of dominantly contributing terms from those that describe marginal and therefore negligible information encapsulated in the predictor and response variable sets [9].

As a benchmark study, this work analyses a recently proposed model of signal transduction pathways in hepatic cells when stimulated by interleukin-6 (IL6) [10]. From the associated medical literature, it is known that IL6 represents one of the principal factors involved in the regulation of most Acute Phase Proteins (APPs) [11]. These are a product of the Acute Phase Response which is a beneficial short-term response to a tissue trauma, injury or infection in mammals [12].

Cytokines, such as IL6 are produced to stimulate complex intracellular signalling resulting in the up and/or down regulation of specific plasma proteins, namely the APPs. A deeper understanding of the pathways and mechanisms involved can lead to the prevention or mediation of the problems that can occur under prolonged exposure to these elevated levels of plasma proteins [13–15].

This work shows how the underlying mechanistic model, describing the components involved in the signal transduction initiated by IL6, can be translated into a form that is linear-in-parameters so as the complex model can be represented by linear predictor/response interrelationships. This then allows for the application of conventional multivariate data analysis tools.

The RRR estimator, pioneered by Anderson [16], is a projection method to analyse multivariate data sets to produce the most accurate regression model with as few linearly independent projection directions within the predictor space as possible. In contrast, CCA [17] and PLS [18] are techniques that produce projections of the observations within the predictor and response space that maximize a correlation and covariance criterion, respectively, and therefore analyse the underlying interrelationships completely in the projection, or latent variable, spaces.

To determine the degree of contribution of the terms within the ordinary differential equations of the IL6 model to the prediction of the derivatives, all three of these techniques have been considered. A comparison of the three techniques is made by examining the response and predictor

DOI: 10.1002/sim

variable weights and residuals. To test the results, the terms identified as having a negligible impact on the model are removed from the original IL6 signal transduction model and a measure of the impact on the model accuracy is used to identify the viability of the various results. It is important to note that multivariate analysis techniques have not received significant attention in the field of systems biology as analysis tools for the study of such pathway models.

Previous work on analysing such models include sensitivity analysis [19], which provides an insight into the importance of the parameters on the concentration of the transcription factor and fuzzy modelling [20], which generates a linguistic model to help to describe the dynamic behaviour of the model. RRR, CCA and PLS on the other hand will analyse the contribution of individual reaction terms to the dynamical changes exhibited by protein concentrations in the model. With these results it is possible to gain further insight into the relevance of the reaction terms and simplify the existing model. As demonstrated in this article, the extracted information can then identify parts of the model that may warrant further model refinements.

The paper is divided into the following sections. Section 2 introduces preliminary information on the algorithms and the selection criteria for choosing the number of latent variables. This is followed in Section 3 by a brief overview of the IL6 signal transduction model being analysed, at which point this paper corrects an error in the presentation of the IL6 pathway model from a previous publication [10]. Section 4 provides an overview of the steps involved in the investigation before expanding on the generation of the data in Section 5 and a presentation of the results from each technique in Section 6. Section 7 begins the interpretation and in-depth discussion of the results with the validation of the results via the model reduction. This is then followed by a concluding summary in Section 8.

2. PRELIMINARIES

This section briefly reviews PLS, CCA and RRR. The multivariate analysis is based on a predictor variable set $\mathbf{x}\in\mathbb{R}^N$, a response variable set $\mathbf{y}\in\mathbb{R}^M$ and a total of K observations for each variable set that are stored as row vectors in $\mathbf{X}\in\mathbb{R}^{K\times N}$ (predictor matrix) and $\mathbf{Y}\in\mathbb{R}^{K\times M}$ (response matrix). Each multivariate technique is designed to establish a linear regression model for $\mathbf{Y}=\mathbf{XB}+\mathbf{E}$, where $\mathbf{B}\in\mathbb{R}^{N\times M}$ is the regression matrix and $\mathbf{E}\in\mathbb{R}^{K\times M}$ is a residual matrix.

2.1. Partial least squares

Partial least squares maximizes the covariance between projections of the predictor and response variables onto one-dimensional subspaces [21, 22]. With $\mathbf{X}_k$ and $\mathbf{Y}_k$ referring to the kth predictor and response matrix, respectively, after $k-1$ deflation steps have been performed, the kth pair of weight vectors, $\mathbf{w}_k\in\mathbb{R}^N$ and $\mathbf{v}_k\in\mathbb{R}^M$, and score vectors, $\mathbf{t}_k\in\mathbb{R}^K$ and $\mathbf{u}_k\in\mathbb{R}^K$, are determined by maximizing the following cost function:

$$J_k=\mathbf{t}_k^{\mathrm{T}}\mathbf{u}_k=\mathbf{w}_k^{\mathrm{T}}\mathbf{X}_k^{\mathrm{T}}\mathbf{Y}_k\mathbf{v}_k \tag{1}$$

which is subject to the following constraints:

$$C_{1,\mathrm{PLS}}^{(k)}=\|\mathbf{w}_k\|_2^2-1=0,\quad C_{2,\mathrm{PLS}}^{(k)}=\|\mathbf{v}_k\|_2^2-1=0 \tag{2}$$

where $\|\cdot\|_2^2$ is the norm of a vector.

DOI: 10.1002/sim

Table I. Deflation procedure for PLS.

Step	Equation	Description
1	$\mathbf{p}_k = \frac{\mathbf{X}_k^{\mathrm{T}}\mathbf{t}_k}{\mathbf{t}_k^{\mathrm{T}}\mathbf{t}_k}$	Determine predictor matrix loading vector
2	$\mathbf{X}_{k+1} = \mathbf{X}_k - \hat{\mathbf{X}}_k = \mathbf{X}_k - \mathbf{t}_k\mathbf{p}_k^{\mathrm{T}}$	Deflate to produce predictor matrix for the next iteration

Defining λ_k as an eigenvalue, the solution of this cost function is given by the largest eigenvalue of the following eigenvector-eigenvalue problem:

$$\mathbf{X}_k^{\mathrm{T}}\mathbf{Y}_k\mathbf{Y}_k^{\mathrm{T}}\mathbf{X}_k\mathbf{w}_k = \lambda_k\mathbf{w}_k, \quad \mathbf{Y}_k^{\mathrm{T}}\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}}\mathbf{Y}_k\mathbf{v}_k = \lambda_k\mathbf{v}_k \tag{3}$$

The vectors $\mathbf{w}_k, \mathbf{v}_k, \mathbf{t}_k$ and $\mathbf{u}_k$ are found by the iterative power method and the subsequent vectors are determined using a deflation procedure that involves the subtraction of score vectors from the predictor and response matrices to produce $\mathbf{X}_{k+1}$. This is done by determining loading vectors $\mathbf{p}_k$ and $\mathbf{q}_k$, which represent the contribution of the t-score vector to the predictor and response matrices, respectively. It should be noted that the subscript k implies that $\mathbf{X}_1 = \mathbf{X}$. The required regression steps are listed in Table I. The estimated regression matrix is given by

$$\widehat{\mathbf{B}} = \mathbf{W}[\mathbf{P}^{\mathrm{T}}\mathbf{W}]^{-1}\mathbf{Q}^{\mathrm{T}} \tag{4}$$

where $\mathbf{W}$, $\mathbf{P}$ and $\mathbf{Q}$ are matrices storing the retained $n \leqslant N$ vectors $\mathbf{w}_k, \mathbf{p}_k$ and $\mathbf{q}_k$, respectively.

2.2. Canonical correlation analysis

Canonical correlation analysis involves finding two sets of canonical variates, $\mathbf{w}_k \in \mathbb{R}^N$ and $\mathbf{v}_k \in \mathbb{R}^M$, so as the correlation between the projections $\mathbf{t}_k = \mathbf{X}\mathbf{w}_k \in \mathbb{R}^K$ and $\mathbf{u}_k = \mathbf{Y}\mathbf{v}_k \in \mathbb{R}^K$ is maximized.

The correlation coefficient, $r_k = \mathbf{t}_k^{\mathrm{T}}\mathbf{u}_k$, gives the function to be maximized as

$$r_k = \mathbf{w}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v}_k \tag{5}$$

and is subject to the following constraints:

$$C_{1,\mathrm{CCA}}^{(k)} = \mathbf{w}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}_k - 1 = 0, \quad C_{2,\mathrm{CCA}}^{(k)} = \mathbf{v}_k^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{v}_k - 1 = 0 \tag{6}$$

as well as

$$\mathbf{C}_{1,\mathrm{CCA}}^{(k)} = \begin{bmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \mathbf{w}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_{k-1}^{\mathrm{T}} \end{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}_k = \mathbf{0}$$

DOI: 10.1002/sim

$$\mathbf{C}_{2,\mathrm{CCA}}^{(k)} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_{k-1}^{\mathrm{T}} \end{bmatrix} \mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{v}_k = \mathbf{0}$$
$$\mathbf{C}_{3,\mathrm{CCA}}^{(k)} = \begin{bmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \mathbf{w}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_{k-1}^{\mathrm{T}} \end{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v}_k = \mathbf{0} \tag{7}$$

The eigenvectors associated with the kth largest eigenvalue λ_k represent the solution of this constrained optimization problem:

$$[[\mathbf{X}^{\mathrm{T}}\mathbf{X}]^{\dagger}\mathbf{X}^{\mathrm{T}}\mathbf{Y}[\mathbf{Y}^{\mathrm{T}}\mathbf{Y}]^{\dagger}\mathbf{Y}^{\mathrm{T}}\mathbf{X} - \lambda_k]\mathbf{w}_k = \mathbf{0} \tag{8}$$

$$[[\mathbf{Y}^{\mathrm{T}}\mathbf{Y}]^{\dagger}\mathbf{Y}^{\mathrm{T}}\mathbf{X}[\mathbf{X}^{\mathrm{T}}\mathbf{X}]^{\dagger}\mathbf{X}^{\mathrm{T}}\mathbf{Y} - \lambda_k]\mathbf{v}_k = \mathbf{0} \tag{9}$$

where † represents the generalized inverse.

Based on the work by Golub [23] on computationally efficient and numerically stable solutions for CCA, the steps in Table II can be applied to obtain the $n \leqslant \min\{M, N\}$ pairs of weight vectors, $\mathbf{w}_k$ and $\mathbf{v}_k$, stored in $\mathbf{W}$ and $\mathbf{V}$, respectively. The estimation of the regression matrix $\mathbf{B}$ is given by:

$$\widehat{\mathbf{B}} = \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{S}_{XY} \tag{10}$$

Table II. Steps to compute weight vectors for CCA and RRR.

Step	Equation	Description
1	$\mathbf{\Sigma}_{XX} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$	Cross product matrix for predictor set
2	(CCA) $\mathbf{\Sigma}_{YY} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ (RRR) $\mathbf{\Sigma}_{YY} = \mathbf{I}$	Cross product matrix for response set
3	$\mathbf{\Sigma}_{XY} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$	Cross product matrix for predictor and response sets
4	$\mathbf{\Sigma}_{XX} = \mathbb{U}_X\mathbb{L}_X\mathbb{V}_X^{\mathrm{T}}$ $\mathbf{\Sigma}_{YY} = \mathbb{U}_Y\mathbb{L}_Y\mathbb{V}_Y^{\mathrm{T}}$	Singular value decomposition of $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ (CCA only)
5	$\mathbf{\Sigma}_{XX}^{-1/2} = \mathbb{U}_X\mathbb{L}_X^{-1/2}\mathbb{V}_X^{\mathrm{T}}$	Square root of inverse of $\mathbf{\Sigma}_{XX}$
6	$\mathbf{\Sigma}_{YY}^{-1/2} = \mathbb{U}_Y\mathbb{L}_Y^{-1/2}\mathbb{V}_Y^{\mathrm{T}}$	Square root of inverse of $\mathbf{\Sigma}_{YY}$ (CCA only)
7	$\mathbf{S} = \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2}$	Set-up matrix product
8	$\mathbf{S} = \mathbb{U}\mathbf{R}\mathbb{V}^{\mathrm{T}}$	Singular value decomposition of $\mathbf{S}$
9	$\mathbf{W} = \mathbf{\Sigma}_{XX}^{-1/2}\mathbb{U}$	Compute predictor weight vectors
10	$\mathbf{V} = \mathbf{\Sigma}_{YY}^{-1/2}\mathbb{V}$	Compute response weight vectors

DOI: 10.1002/sim

2.3. Reduced rank regression

Reduced rank regression directly determines score vectors, $\mathbf{t}_k$ and $\mathbf{u}_k \in \mathbb{R}^K$, as a linear combination of the predictor and response set, respectively, such that the response set is predicted with maximum accuracy:

$$\mathbf{E}_k = \mathbf{Y} - \mathbf{t}_k \mathbf{t}_k^{\mathrm{T}} \mathbf{Y} \tag{11}$$

where $\mathbf{E}_k \in \mathbb{R}^{K \times M}$ is the residual matrix, $\mathbf{t}_k = \mathbf{X}\mathbf{w}_k$ is a score vector and $\mathbf{w}_k \in \mathbb{R}^N$ is a weight vector. This produces the following cost function:

$$\|\mathbf{E}_k^{\mathrm{T}} \mathbf{E}_k\|_2^2 = \|\mathbf{Y}^{\mathrm{T}}[\mathbf{I} - \mathbf{t}_k \mathbf{t}_k^{\mathrm{T}}]\mathbf{Y}\|_2^2 \tag{12}$$

which can alternatively be formulated to be:

$$J_k = \mathbf{w}_k \mathbf{X}^{\mathrm{T}} \mathbf{Y} \mathbf{v}_k \tag{13}$$

where $\mathbf{v}_k \in \mathbb{R}^M$. Equation (13) is subject to the following constraints:

$$C_{1,\mathrm{RRR}}^{(k)} = \mathbf{w}_k^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{w}_k - 1 = 0, \qquad C_{2,\mathrm{RRR}}^{(k)} = \|\mathbf{v}_k\|_2^2 - 1 = 0 \tag{14}$$

as well as:

$$\begin{aligned}
\mathbf{C}_{1,\mathrm{RRR}}^{(k)} &= \begin{bmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \mathbf{w}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_{k-1}^{\mathrm{T}} \end{bmatrix} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{w}_k = \mathbf{0} \\
\mathbf{C}_{2,\mathrm{RRR}}^{(k)} &= \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_{k-1}^{\mathrm{T}} \end{bmatrix} \mathbf{v}_k = \mathbf{0} \\
\mathbf{C}_{3,\mathrm{RRR}}^{(k)} &= \begin{bmatrix} \mathbf{w}_1^{\mathrm{T}} \\ \mathbf{w}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{w}_{k-1}^{\mathrm{T}} \end{bmatrix} \mathbf{X}^{\mathrm{T}} \mathbf{Y} \mathbf{v}_k = \mathbf{0}
\end{aligned} \tag{15}$$

The solution of this constrained optimization problem for the kth pair of weight vectors is given by:

$$[[\mathbf{X}^{\mathrm{T}}\mathbf{X}]^{\dagger} \mathbf{X}^{\mathrm{T}} \mathbf{Y} \mathbf{Y}^{\mathrm{T}} \mathbf{X} - \lambda_k \mathbf{I}]\mathbf{w}_k = \mathbf{0} \tag{16}$$

$$[\mathbf{Y}^{\mathrm{T}} \mathbf{X} [\mathbf{X}^{\mathrm{T}}\mathbf{X}]^{\dagger} \mathbf{X}^{\mathrm{T}} \mathbf{Y} - \lambda_k \mathbf{I}]\mathbf{v}_k = \mathbf{0} \tag{17}$$

DOI: 10.1002/sim

The pairs of weight vectors stored in **W** and **V** can be calculated using the steps shown in Table II. These steps are the same as used for CCA with the exception that the cross product matrix for the response data are set equal to the identity matrix. The estimated regression matrix is that of equation (10) by inserting the matrix **W** of the RRR solution.

2.4. Selection of the number of retained latent variables

This subsection briefly reviews the techniques for determining the number of retained latent variables (LVs). Over the past decades, many statistical and heuristical approaches have been proposed for determining this number.

One of the most popular techniques to determine the number of 'meaningful' components is the broken stick model, which was first presented in 1957 by MacArthur in his study of the structure of animal communities [24]. This technique involves comparing the data with a stick of unit length on which $n-1$ points are randomly selected from a uniform distribution. The stick is then broken at these points and the lengths of the n resulting segments are proportional to the n principal components of the data sets under investigation.

Other techniques include the Kaiser–Guttman test, log-eigenvalue diagram (LEV), cross validation, Velicer's partial correlation procedure, Cattell's SCREE test, bootstrapping techniques, cumulative percentage of total variance and Bartlett's test for equality of eigenvalues. An extensive comparison of the tests frequently utilized for this open problem can be found in [25, 26].

As discussed in Section 5, the analysis performed in this paper is based on a theoretically perfect model that does not include any residuals, that is $\mathbf{E}=\mathbf{0}$. Therefore, taking one less sample will produce the same model, since the rows in the predictor and response data sets must be linearly dependent. Given these circumstances, the use of these different techniques will not offer any advantage. With this in mind, the number of dominant latent variables that will be retained for each technique is decided by inspecting the cumulative variance contribution to the response matrix and also the residual error.

2.5. Notes on the PLS deflation procedure

The analysis of the data from the signalling pathway model in Sections 6 and 7 is based on the weight vectors and the residuals of the predictor and response variables sets. It is therefore important that the weight vectors, in particular, are comparable. Reference [22] showed that it is sufficient to deflate one of the data matrices, i.e. the predictor or response variables only. In this work, only the predictor matrix has been deflated. This, in turn, implies that for PLS the v-weight vectors determine the u-score vectors directly from the undeflated response matrix for each of the three multivariate methods.

This is, however, different for the predictor matrix, where CCA and RRR determine the t-score vectors from the undeflated predictor matrix, while the PLS requires a deflation procedure. In order to accommodate this procedure into the associated weight vectors, reference [22] outlined that r-weight vectors can be iteratively computed:

$$\mathbf{r}_k=[\mathbf{I}-\mathbf{r}_{k-1}\mathbf{p}_{k-1}^{\mathrm{T}}]\mathbf{w}_k=\left[\prod_{i=1}^{k-1}[\mathbf{I}-\mathbf{w}_i\mathbf{p}_i^{\mathrm{T}}]\right]\mathbf{w}_k \tag{18}$$

DOI: 10.1002/sim

Given that the CCA and RRR algorithms can alternatively be computed iteratively utilizing the deflation procedure shown in Table I, the analysis in equation (18) yields that $\mathbf{w}_k = \mathbf{r}_k$ and hence the following theorem.

Theorem 1
If iterative algorithms for CCA and RRR are used instead of the batch ones, described in Table II, the t-score vectors calculated from the original predictor matrix, $\mathbf{Z}$, are identical to those computed from the deflated predictor matrix, $\mathbf{X}_k$, that is $\mathbf{t}_k = \mathbf{X}_k\mathbf{w}_k = \mathbf{X}\mathbf{w}_k$.

A proof of Theorem 1 is given in the Appendix and shows that $\mathbf{p}_i^{\mathrm{T}}\mathbf{w}_k = 0 \ \forall i < k$.

This therefore suggests that the deflation procedure yields two different kth weight vectors of the predictor matrix for the PLS algorithm, while that of the CCA and RRR algorithm are identical. For completeness and a rigorous comparison, we include both the w- and the r-weight vectors in our analysis.

3. IL6 SIGNAL TRANSDUCTION IN HEPATOCYTES

The signalling pathway model to be analysed describes the signal transduction in hepatocytes when stimulated by IL6. This model integrates signalling through the JAK/STAT and MAPK pathways and consists of 68 nonlinear ordinary differential equations (ODEs), which can be represented by:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = f(\mathbf{x}, \mathbf{p}, u) \tag{19}$$

The equations represent concentration balances of individual proteins and protein complexes, and are derived according to the law of mass action or Michaelis–Menten kinetics. The parameters, $\mathbf{p} \in \mathbb{R}^{94}$, represent the reaction constants, the states, $\mathbf{x} \in \mathbb{R}^{68}$, are the concentrations of the proteins in the pathway and the input, $u \in \mathbb{R}$, is the stimulating concentration of IL6.

The JAK/STAT and MAPK pathways are highly complex intracellular signal transduction pathways involving many protein components that regulate the activity in the cytoplasm and nucleus in response to extracellular stimuli on the plasma membrane.

The JAK/STAT pathway is considered to be one of the most important signalling pathways downstream of cytokine receptors [27] and can be activated by a wide variety of cytokines and growth factor signals [28]. The MAPK cascade is another important pathway widely involved in eukaryotic signal transduction [29]. Both pathways interact together leading to diverse responses involving gene expression, cell proliferation, mitogenesis, differentiation and stress response in mammalian cells.

The model under investigation describes both of these pathways in parallel. When the cell is initially stimulated by IL6 via the membrane receptors, communication begins along both pathways with the JAK/STAT pathway leading to the production of dimerized STAT3 in the cytoplasm, $(\mathrm{STAT3C}^*)_2$, which can then translocate to the nucleus where it is responsible for the transcription and translation of important Acute Phase Proteins. The production of the nuclear STAT3 dimer, $(\mathrm{STAT3N}^*)_2$, is of extreme importance for the regulation of the Acute Phase Response [30]. A detailed description of the model can be found in [10].

DOI: 10.1002/sim

3.1. Correction to reference

The model described in [10] contains a typographical error in an equation that can be found in Appendix II, page 859 of the publication and which is shown here as equation (20). It should be as presented in equation (21).

$$\frac{dx_{16}}{dt} = k_{f9}x_{14}x_8 - k_{r9}x_{16} - k_{10}x_{16} - k_{f34}x_{16} + k_{r34}x_{39} \tag{20}$$

$$\frac{dx_{16}}{dt} = k_{f9}x_{15}x_8 - k_{r9}x_{16} - k_{10}x_{16} - k_{f34}x_{16} + k_{r34}x_{39} \tag{21}$$

It is also worth noting that in order to replicate the results published in [10], the model equations for dx_1 and dx_3, which also appear in Appendix II, page 859, must be set equal to zero. This is due to the fact that the work was based on the common assumption that, due to the large number of receptors on the cell surface, any variation in this number can be considered insignificant.

4. INTRODUCTION OF THE ANALYSIS APPROACH

This section details the steps involved in utilizing PLS, CCA and RRR for the analysis of complex cell models. Although this is a generic approach that can be applied to linear-in-parameter models, its application is demonstrated here for a model describing signal transduction pathways for hepatic cells when stimulated by IL6. A thorough description of this model may be found in Reference [10].

The steps of this analysis include (i) obtaining data using the available mechanistic model, (ii) establishing a linear-in-parameter regression model, (iii) scaling the recorded data appropriately, (iv) applying PLS, CCA and RRR to analyse the linear-in-parameter model, (v) identifying variable clusters and marginally contributing terms of the rate equations and (vi) simplifying the model by removing such terms, and validating the performance of the reduced model with the original one to estimate the impact of the isolated variable clusters or terms. A more specific breakdown of this analysis is given below.

1. Simulating the model.
   - Input the 68 ODEs into MATLAB, with all initial conditions and kinetic constants initialized to the values depicted in the original published model [10].
   - Design an input signal (IL6 concentration) to properly excite the system, as per the recommendations in reference [31]. This involves a sequence of step inputs of varying duration and magnitude.
   - For each step input:
     - Integrate the system of differential equations across the duration of the specifically designed input signal to produce a dynamic profile for each of the 68 state variables (proteins).
     - Update the initial conditions by equating them to the final state variable values from the previous input condition.
2. Producing a linear-in-parameter model ($\mathbf{Y} = \mathbf{ZB} + \mathbf{E}$).

DOI: 10.1002/sim

- Three of the ODEs were calculated using Michaelis–Menten kinetics, and as a result contain rational fractions. They must be re-written in a linear parametric representation.
- For each step input the resulting data from the dynamic profiles is substituted into these modified 68 ODEs to calculate a value for the response data set.
- **Y** is equated to the derivative terms (i.e. left-hand side of ODEs). Each row of this response matrix includes the values of each of the 68 derivative terms and each column stores the consecutive values of a particular derivative term.
- The right-hand side of the equations is made up of individual proteins/mechanisms as well as interactions between these variables. The 94 terms/cross-products that occur (e.g. x_1u, x_2x_5, x_5) are used to construct the data set **Z**, where each row vector, $\mathbf{z}^{\mathrm{T}}$, is a function of the state variables, **x**, that is $\mathbf{z}^{\mathrm{T}} = \mathbf{f}(\mathbf{x})$. The setup of these matrices is presented in equation (23).
- It is important that each row of the response matrix represents the derivative terms that correspond to the 94 terms of the predictor matrix, i.e. the simulated time for 94 cross-product terms must match the simulated time for the 68 derivative terms.
- The regression matrix **B** contains the kinetic coefficients that precede the terms of the predictor data set.

3. Scaling.
   - The derivative and cross-product terms are scaled to ensure that they each have a comparable variance.
   - This is a common and essential practice in multivariate data analysis for guaranteeing that

     (i) each term has the same chance of contributing to the predictor/response structure and
     (ii) that no term has a dominant variance, which would render the analysis meaningless.

4. Apply multivariate statistical data techniques.
   - Apply PLS, CCA and RRR to the predictor and response matrix pair as discussed in Sections 2.1–2.3.
   - The resulting predictor and response weights are scaled to be of unit length.
5. Data analysis.
   - Estimate the dominant number of latent variables by inspecting the residual variance and cumulative variance contribution to the response matrix for each technique.
   - Plot the values of the weight vectors as well as the residuals of the predictor and response variables for each method by only retaining the selected number of dominant latent variables.
   - The variable clusters and significant terms are identified as those associated with the largest weights or the smallest residuals.
   - The negligible terms are identified as those associated with the smallest weights and/or the largest residuals.
6. Interpretation and model reduction.
   - The results from the three techniques are compared and analyzed.
   - The variables identified as negligible are replaced in the model by a constant value.
   - ODEs with negligible derivative terms are also removed.
   - The accuracy of the reduced model is compared with the original model on the basis of the crucial protein $(\mathrm{STAT3N}^*)_2$.

DOI: 10.1002/sim

This analysis is based on the assumption that the model accurately describes the cell mechanisms and dynamics and that the initial conditions for the states are suitably defined. We would like to note that if any of these conditions be violated the interpretation of the results may not be representative. It is worth noting at this point that the model under examination in this paper has been developed from previously published literature on the MAPK and JAK/STAT pathway that contains western blot analysis data [32–34].

Given that the generation of the data and the subsequent scaling are pivotal to this analysis scheme, the next section provides a detailed guideline of the steps required to reproduce this analysis of the hepatocyte model studied in this article.

5. DATA GENERATION

This section details the determination of the state sequences and their respective derivative terms. The simplification of the Michaelis–Menten kinetics is explained and the construction of the predictor and response variables is presented. The input sequence of the stimulating IL6 concentration is described and is followed by a detailed discussion on the scaling of the data sets.

The objective is to generate the data in such a way as to have a linear-in-parameter model, which will then allow for the application of the multivariate data techniques.

5.1. Determination of state sequences

The state sequences are obtained by initially solving the set of 68 nonlinear ODEs. This is performed using the ode15s routine in MATLAB. As stated in Section 4, this involves setting the initial conditions and kinetic constants to those published in the original model [10]. For consecutive step inputs, the values for each state variable are then substituted into the ODEs, which are represented by equation (19), to provide a series of values for the rate of change of these variables.

These derivative terms (i.e. the 68 terms from the left-hand side of the equations) are used to construct the response matrix $\mathbf{Y} \in \mathbb{R}^{K \times M}$, where $M = 68$ is the number of derivative terms and K is the number of observations.

The predictor matrix $\mathbf{Z} \in \mathbb{R}^{K \times N}$ is constructed from each of the cross product terms from the ODEs (i.e. the 94 different combinations of the state variables), where $N = 94$ refers to the number of cross product terms.

This can be illustrated by examining the first three ODEs of the model, which describe the changes in concentration of gp80, IL6-gp80 and gp130, respectively, and are detailed in equation (22). This data is then converted to a matrix-vector representation, as shown in equation (23).

$$\begin{aligned} \mathrm{d}x_1 &= -k_{f0}x_1u + k_{r0}x_2 \\ \mathrm{d}x_2 &= k_{f0}x_1u - k_{r0}x_2 + k_{r2}x_6 - k_{f2}x_2x_5 \\ \mathrm{d}x_3 &= -k_{f1}x_3x_4 + k_{r1}x_5 \end{aligned} \tag{22}$$

DOI: 10.1002/sim

$$[\mathbf{dx}_1 \;\; \mathbf{dx}_2 \;\; \mathbf{dx}_3 \;\; \cdots] = \begin{bmatrix} (\mathbf{x}_1 \circ \mathbf{u})^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ (\mathbf{x}_2 \circ \mathbf{x}_5)^{\mathrm{T}} \\ (\mathbf{x}_3 \circ \mathbf{x}_4)^{\mathrm{T}} \\ \mathbf{x}_5^{\mathrm{T}} \\ \mathbf{x}_6^{\mathrm{T}} \\ \vdots \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -k_{f0} & k_{f0} & 0 & \cdots \\ k_{r0} & -k_{r0} & 0 & \cdots \\ 0 & -k_{f2} & 0 & \cdots \\ 0 & 0 & -k_{f1} & \cdots \\ 0 & 0 & k_{r1} & \cdots \\ 0 & k_{r2} & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{23}$$

where the response data are given by the 68 consecutive derivative terms,

$$\mathbf{Y} = [\mathbf{dx}_1 \;\; \mathbf{dx}_2 \;\; \mathbf{dx}_3 \;\; \cdots] \tag{24}$$

and the predictor data are defined as the 94 cross product terms, as shown in the following equation:

$$\mathbf{Z} = [\mathbf{x}_1 \circ \mathbf{u} \;\; \mathbf{x}_2 \;\; \mathbf{x}_2 \circ \mathbf{x}_5 \;\; \mathbf{x}_3 \circ \mathbf{x}_4 \;\; \mathbf{x}_5 \;\; \mathbf{x}_6 \;\; \cdots] \tag{25}$$

In the above equations, the symbol $\circ$ relates to an element wise operation, where the elements of the resultant vector are the product of the elements of the individual vectors. For example, each element in $\mathbf{x}_2 \circ \mathbf{x}_5$ is the product of the elements of $\mathbf{x}_2$ and $\mathbf{x}_5$ stored at the same position.

The matrices $\mathbf{Z}$ and $\mathbf{Y}$ can store the generated data of 65 out of the 68 ODEs. However, the remaining 3 ODEs represent Michaelis–Menten kinetics and therefore involve fractions rather than products of the state variables only. The next subsection shows how to reformulate these 3 ODEs to produce a linear-in-parameter representation of the recorded data.

5.2. Simplification of Michaelis–Menten kinetics

Of the 68 ODEs, 3 contain rational fraction functions derived from Michaelis–Menten kinetics. These need to be rewritten in a linear parametric representation. Equation (26) shows an example of how this simplification is performed.

$$\begin{aligned} \frac{\mathrm{d}x_{25}}{\mathrm{d}t} &= \frac{k_{18a}x_{20}}{k_{18b}+x_{20}} - k_{19}x_{25} \\ \frac{\mathrm{d}x_{25}}{\mathrm{d}t} + \frac{x_{20}}{k_{18b}}\frac{\mathrm{d}x_{25}}{\mathrm{d}t} &= \frac{k_{18a}}{k_{18b}}x_{20} - k_{19}x_{25} - \frac{k_{19}}{k_{18b}}x_{20}x_{25} \\ \frac{\mathrm{d}\widetilde{x}_{25}}{\mathrm{d}t} &= \left(1 + \frac{x_{20}}{k_{18b}}\right)\frac{\mathrm{d}x_{25}}{\mathrm{d}t} \end{aligned} \tag{26}$$

Three stages are detailed in this equation. First, a cross multiplication is performed on the original fraction to produce a linear parametric representation of the ODE. The derivative term can then be calculated using only the right-hand side of this equation, before being scaled up by the factor from the left-hand side. This multiplication is shown as the final step and produces the resultant derivative term denoted by $(\mathrm{d}\widetilde{x}_{25}/\mathrm{d}t)$.

DOI: 10.1002/sim

As a result of this simplification, there will be some exclusion in the terms of the response matrix leading to the following estimation of the linear regression model:

$$\mathbf{Y} = \mathbf{ZB} \tag{27}$$

where $\mathbf{Z} \in \mathbb{R}^{K \times N}$ is the predictor matrix, $\mathbf{Y} \in \mathbb{R}^{K \times M}$ is the response matrix and $\mathbf{B} \in \mathbb{R}^{M \times N}$ is the regression matrix that consists of the first-order rate constants and Michaelis constants.

5.3. Input signal

In order to properly excite the system (model), the input consists of a series of steps [31]. The range for the IL6 concentration is selected based on the values previously and currently being used in experiments, which is 0.00383–0.383 nM. It is expected that this range of concentrations represents different levels of cell stimulation up to the level of cell saturation. By using an input succession of 60 steps from 0.001–0.4 nM a realistic range for the input concentration is maintained.

From the dynamic profiles of the state variables it is evident that the most dominant responses arise within the first 2 h. There is very little response after 8 h but slight changes can still be observed 100 h after the initial stimulation. These long-term changes seen in the model have no biological significance and are not seen in experimental results. The simulated model, as is the case with the real system, has no true steady state so. As such, for this analysis, it is assumed that a complete steady state is achieved within 100 h, with the most significant results occurring within the first 8 h.

Therefore, when selecting the time to apply each step input, the first 55 steps focus on the range 0.1–15 h, which corresponds to 0.2 times the shortest response [0.5 h] and 1.5 times the longest response [10 h] and the final 5 steps run for 100 h. This ensures that each variable can reach the assumed steady state in addition to being properly excited.

In addition, to ensure there was enough variation on the input signal, randomly generated mean-centered noise drawn from a normal distribution was added. This signal is presented in Figure 1 along with a plot that shows an expanded view of the first 55 steps. The frequency of changes in

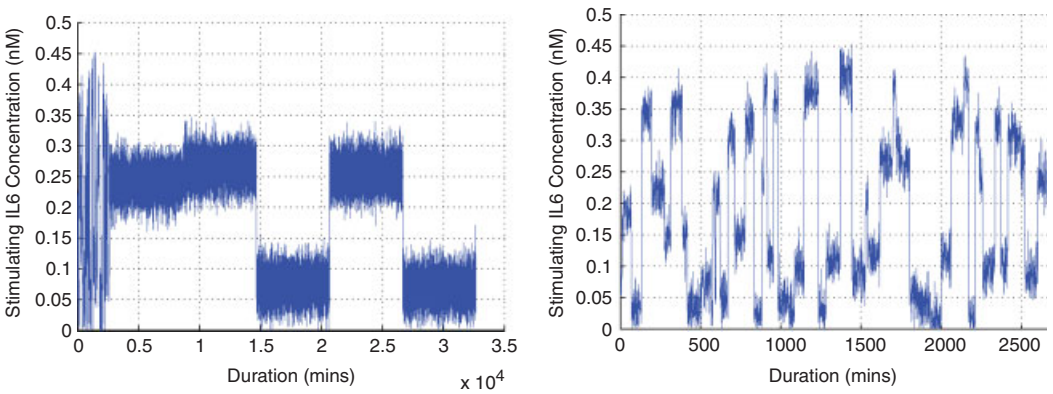

Figure 1. Input signal for the stimulating concentration of IL6 displaying all 60 steps and the first 55 steps, respectively.

DOI: 10.1002/sim

the concentration of IL6 are made purely for analysis using this model and do not correspond to changes anticipated in vivo.

5.4. Scaling

It is common practise to scale variables prior to any analysis [35]. It is desirable for each set of data to have comparable variances so as no prior reason exists for any term to be selected for their significance by the various multivariate statistical data techniques.

5.4.1. Predictor data set. When considering the predictor set, the scaling is not straight-forward as the 68 state variables (proteins) are embedded in the 94 cross-product terms. These 94 terms cannot be scaled directly since, in most cases, the individual proteins appear in more than one of the cross-product terms. Therefore, further consideration was required when optimizing their variances.

A genetic algorithm (GA) was used to arrive at these optimal scaling factors. GA optimization techniques were first proposed by Holland in 1975 [36] and are now well recognisee for their application to optimization problems, due to their ability to locate a reasonable solution without an excessive computational cost [37].

The GA consisted of 300 chromosomes and was run for 2000 iterations, with upper and lower search limits placed on each state variable term to aid in focusing the search. The function to be minimized is given by

$$J = \sum_{i=1}^{N} (\text{var}\{z_i\} - \sigma^2)^2 \tag{28}$$

where $z_i = f_i(\boldsymbol{\xi}), \sigma$ is the calculated value of the standard deviation, which was obtained by the GA to be 0.637 for this example, and var$\{\cdot\}$ represents the variance of a variable. The scaling is therefore performed on the state sequences, $\mathbf{x}$, such that

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{68} \end{pmatrix} = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{68} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{68} \end{pmatrix} \tag{29}$$

where α_i represents the scaling factors and $\boldsymbol{\xi}$ is the scaled state variables. The chromosomes representing the GA cost function therefore include values of

$$\boldsymbol{\alpha}^{\mathrm{T}} = (\alpha_1 \;\; \alpha_2 \;\; \cdots \;\; \alpha_{68} \;\; \sigma) \tag{30}$$

It is important to note that equating the variance of the predictor variables to that of the response variables is a common practice in multivariate data analysis and ensures that the variables are equally influential. As shown in equation (29), the scaling factors are applied to the original state sequences, which, in turn, produce the entries of the predictor matrix $\mathbf{Z}$. The optimization function in equation (28) therefore ensures that the values of the variances within the variable set $\mathbf{z}$ are within a narrow range.

DOI: 10.1002/sim

5.4.2. Response data set. Since the response data set was constructed from the 68 derivative terms, an analytical solution was available to arrive at its scaling factors, β_i.

$$\begin{aligned} \mathrm{var}\{\mathbf{y}_i\} &= \mathrm{var}\left\{\frac{\mathrm{d}x_i}{\mathrm{d}t}\right\} \\ \mathrm{var}\left\{\beta_i \frac{\mathrm{d}x_i}{\mathrm{d}t}\right\} &= \beta_i^2 \mathrm{var}\left\{\frac{\mathrm{d}x_i}{\mathrm{d}t}\right\} = \sigma^2 \\ \beta_i &= \sqrt{\frac{\sigma^2}{\mathrm{var}\left\{\dfrac{\mathrm{d}x_i}{\mathrm{d}t}\right\}}} \end{aligned} \tag{31}$$

where x_i represents a chosen state variable i $(i = 1, 2, 3, \ldots, M)$ whose variance is then used to find the required scaling factors for all the other terms so that their variance equates to this same value.

As with the predictor data, these scaling factors are applied to the response data set and the scaled values are used in any subsequent analysis by the multivariate statistical data techniques.

6. ANALYSIS OF PATHWAY MODEL

This section describes how the latent variable information was used and it also details the resulting calculations with both the weight vectors and residuals of the response/predictor variables. All weight vectors have been scaled to be of unit length. The matrices used in this study did not present an ill-conditioned problem but the high degree of correlation between the predictor variables, which the subsequent analysis yields, favour the use of multivariate projection-based methods [38].

6.1. Choosing the dominant latent variables

Both the model error and the cumulative variance contribution to $\mathbf{Y}$ for each multivariate technique were considered when choosing the number of latent variables to retain. The chosen stopping point corresponds to the point at which the model error approaches a minimum with at least 98 per cent of the the response matrix reconstructed.

Figure 2 shows the cumulative variance of both the predictor ($\mathbf{Z}$) and response ($\mathbf{Y}$) data for each latent variable in the three different multivariate techniques and the residual error plots are presented in Figure 3.

The cumulative variance contribution to the predictor and response data are calculated as described in equations (32) and (33) and the residual error, $\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$ with $\hat{\mathbf{y}}$ being the prediction of $\mathbf{y}$ using the multivariate statistical models, is as defined in equation (34).

$$\text{Cumulative variance contribution to } \mathbf{z} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{N} (\sum_{k=1}^{n} t_{ik} p_{jk})^2}{\sum_{i=1}^{K} \sum_{j=1}^{N} x_{ij}^2} \tag{32}$$

$$\text{Cumulative variance contribution to } \mathbf{y} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{M} (\sum_{k=1}^{n} t_{ik} q_{jk})^2}{\sum_{i=1}^{K} \sum_{j=1}^{M} y_{ij}^2} \tag{33}$$

DOI: 10.1002/sim

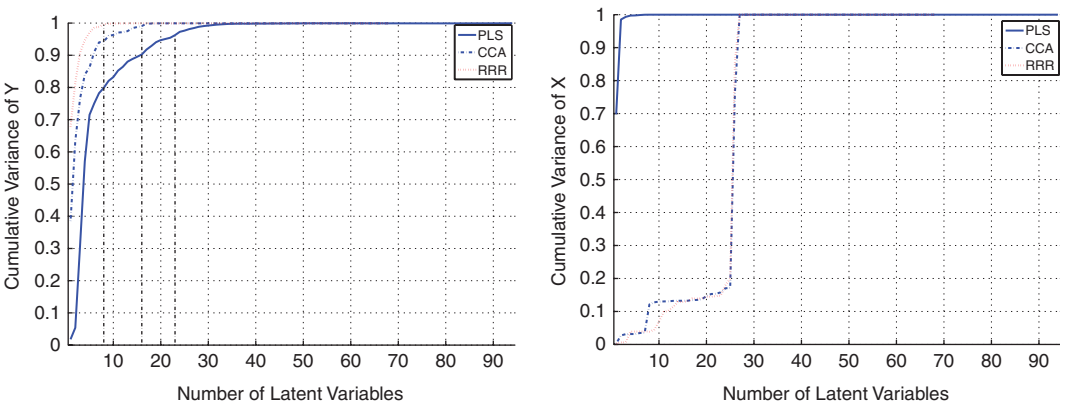

Figure 2. Cumulative variance contribution to the response and predictor matrices for each technique.

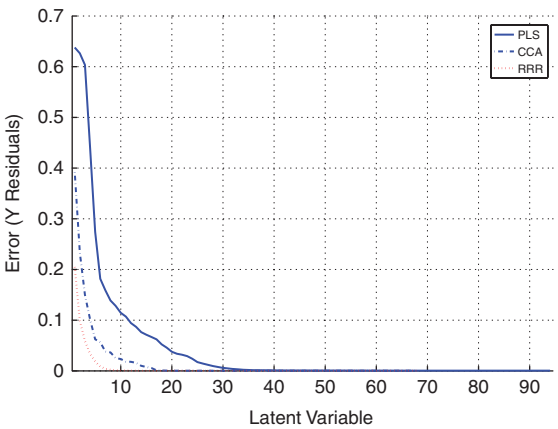

Figure 3. Model error for each multivariate technique.

$$\text{Scaled prediction error}\frac{\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2}{M\cdot K}=\frac{\sum_{i=1}^{K}\sum_{j=1}^{M}(y_{ij}-\sum_{k=1}^{n}t_{ik}q_{jk})^2}{M\cdot K} \tag{34}$$

From the graphs in Figures 2 and 3 it can be seen that a suitable stopping point for PLS, CCA and RRR is after 23, 16 and 8 LV's, respectively. The information associated with the subsequent latent variables is negligible and can be disregarded.

6.2. Weight vectors

The weight vectors reveal the interrelationships between both the predictor and response variables. Terms with a greater weight can be identified as those which cause a significant change in the state variables and therefore indicate that they play an important role in the model. Conversely, those terms that have a lower weight may be recognized as causing a lesser change and consequently

DOI: 10.1002/sim

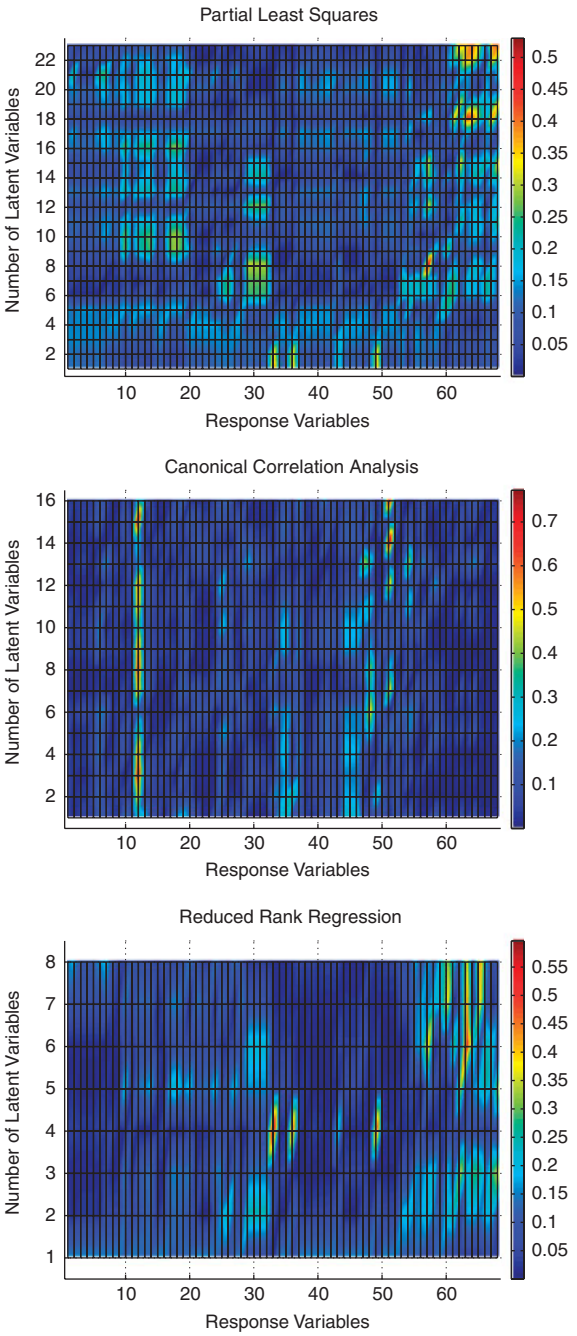

Figure 4. Surface plot showing the absolute value of the response variable weights for the dominant 23, 16 and 8 latent variables for PLS, CCA and RRR, respectively. The gradient alongside each image depicts the colour associated with the various weight magnitudes. The results from PLS, CCA and RRR show no correlation due to the differing criteria used by each technique.

DOI: 10.1002/sim

produce a minor and possibly negligible change in the system. By virtue of the construction of CCA and RRR, along with the fact that a well designed excitation of the model was conducted, it is not possible for a predictor term that was deemed negligible to show a substantial contribution to the response variables (RRR case), or for a response variable identified as having a negligible contribution to the underlying latent variable structure to be an important part of the model (CCA case).

6.2.1. V-weight vectors. Figure 4 shows an illustration of the absolute values for the scaled v-weights of the dominant latent variables for each of the three techniques investigated. The weight vectors used in these plots relate to the deflated variables. In the case of PLS this is described by equations (1) and (3).

It can be seen from these plots that each of the three techniques gives a completely different significance and hence ranking to the individual response terms in the model. This is due to the differing emphasis of each technique in producing a linear regression model. PLS and CCA maximize covariance and correlation, respectively, of the predictor and response variable projections and hence will produce different results. RRR will differ from these again since it determines score vectors that permit the most accurate prediction of the response data set. Step 2 in Table II shows

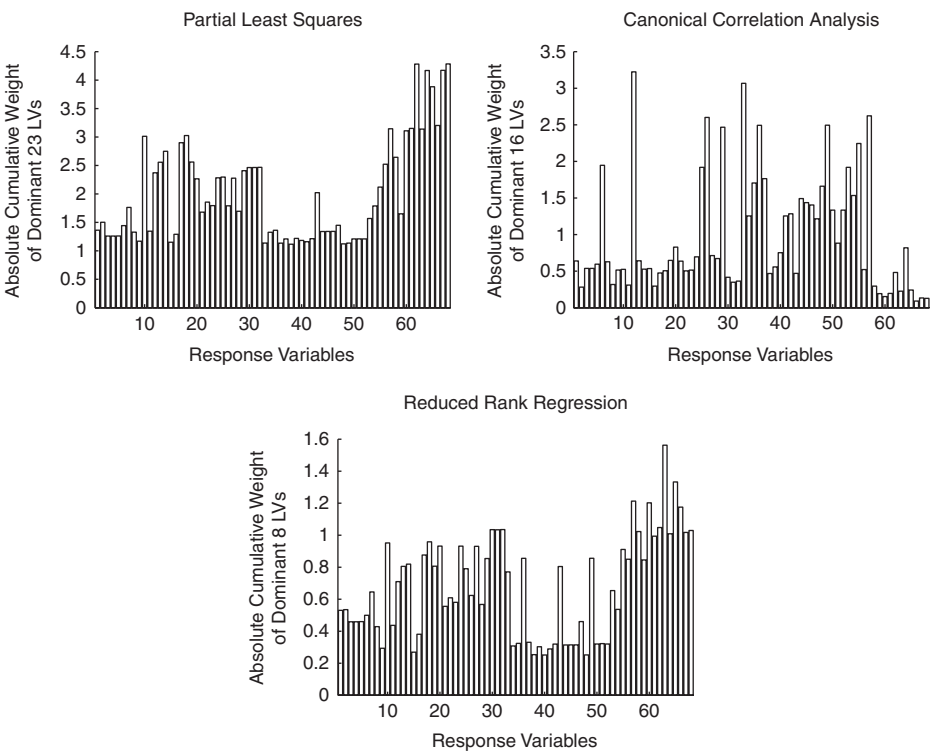

Figure 5. Bar charts corresponding to the cumulative sum of the dominant latent variable weights for each response variable for PLS, CCA and RRR, respectively.

DOI: 10.1002/sim

Table III. Response terms with smallest contributions to the model as depicted by PLS.

	Response term	Description	Relative weight
1	**dx38**	Grb2-SOS	0.26047
2	**dx48**	Raf-Ras-GTP	0.26194
3	**dx36**	Ras-GDP	0.26489
4	**dx49**	Ras-GTP*	0.26505
5	**dx33**	$(IL6\text{-}gp80\text{-}gp130\text{-}JAK^*)_2$-STAT3C-SHP2	0.26528
6	**dx15**	SHP2	0.26865
7	**dx41**	$(IL6\text{-}gp80\text{-}gp130\text{-}JAK^*)_2 - SHP2^*$-Grb2-SOS	0.27087
8	**dx9**	STAT3C	0.27297
9	**dx40**	$(IL6\text{-}gp80\text{-}gp130\text{-}JAK^*)_2 - SHP2^*$-Grb2	0.27612
10	**dx37**	Ras-GTP	0.28228

Table IV. Response terms with smallest contributions to the model as depicted by CCA.

	Response term	Description	Relative weight
1	**dx66**	Phosp3	0.02832
2	**dx68**	ERK-P-Phosp3	0.03979
3	**dx67**	ERK-PP-Phosp3	0.04145
4	**dx60**	MEP-P-Phosp2	0.04733
5	**dx59**	Phosp2	0.05981
6	**dx61**	ERK	0.06080
7	**dx63**	ERK-P	0.07021
8	**dx65**	ERK-PP	0.07504
9	**dx2**	IL6-gp80	0.08748
10	**dx16**	$(IL6\text{-}gp80\text{-}gp130\text{-}JAK^*)_2$-SHP2	0.09157

the approach of RRR, which unlike CCA, sets the response cross product matrix to the identity matrix and hence the computation of different v-weight vectors is an expected outcome.

Using the information from the surface plots in Figure 4 and summing the weights for each latent variable results in the bar charts shown in Figure 5. Again, the differing results from each technique are obvious but it is easier to see the relative significance of each term.

Tables III–V show the 10 terms that were identified by each technique as having the least significant contribution. If these response terms do not contribute greatly to the overall model then there should be little or no impact on the model if their associated kinetic equations were set to zero. The application of results to reduce the model is described in detail in Section 7.

6.2.2. W-weight vectors. Figure 6 shows the surface plot of the absolute values for the scaled w-weights, or predictor variable weights, assigned by each of the dominant latent variables for the three techniques investigated. In this case it can be seen that the results from CCA and RRR show a high degree of correlation, whereas the results from PLS differ significantly. The similarity between CCA and RRR can be explained by examining the similar steps involved in the computation of the w-weight vectors as shown in Table II. The first surface plot in Figure 6 displays the absolute values for the scaled r-weight vectors assigned for each of the dominant latent variables for PLS. These vectors represent the predictor variable weights calculated from the original undeflated predictor

DOI: 10.1002/sim

Table V. Response terms with smallest contributions to the model as depicted by RRR.

	Response term	Description	Relative weight
1	**dx40**	$(\text{IL6-gp80-gp130-JAK}^*)_2 - \text{SHP2}^*\text{-Grb2}$	0.16099
2	**dx48**	Raf-Ras-GTP	0.16126
3	**dx38**	Grb2-SOS	0.16232
4	**dx15**	SHP2	0.17226
5	**dx41**	$(\text{IL6-gp80-gp130-JAK}^*)_2 - \text{SHP2}^*\text{-Grb2-SOS}$	0.18511
6	**dx9**	STAT3C	0.18787
7	**dx39**	$(\text{IL6-gp80-gp130-JAK}^*)_2 - \text{SHP2}^*$	0.19387
8	**dx34**	Grb2	0.19715
9	**dx44**	$\text{SHP2}^*\text{-Grb2-SOS}$	0.20114
10	**dx45**	$\text{SHP2}^*\text{-Grb2}$	0.20143

matrix and differ from the w-weight vectors, which are calculated from the deflated predictor matrix. It can be seen that although there are variations between the results from both methods, there is a general agreement between them.

By summing the weights for each latent variable the 2D bar charts presented in Figure 7 are produced. The similarity between the results from CCA and RRR can be observed along with a pictorial representation of the relative significance of each predictor term.

It is expected that cross product terms assigned with a small weight do not contribute greatly to changes in the derivative terms and hence it may be possible to replace these state variables by constant values. Tables VI–VIII show the 10 predictor terms identified as having the smallest contributions by PLS (by examining both the undeflated and deflated predictor matrices), CCA and RRR, respectively.

6.3. Residuals of response and predictor matrix

An examination of the response and predictor variable residuals for PLS, CCA and RRR offers another tool to extract information from this model. Figure 8 shows how these values compare for each technique. Since the scaling of the variables has been performed first, as described in Section 4, these plots are proportional. The variance of the **Y** residuals follow a similar pattern for all algorithms, with a greater concurrence between CCA and RRR. There is some significant departure in the variance of the **Z** residuals. It was investigated if ranking these values from the smallest to the largest could provide another indication of the significance of the terms, where the terms with the highest residual variances are less significant for the model and hence may warrant a reduction. This method of ranking the terms in order of decreasing residual variance was also validated by ranking the sum of the square of each residual that produced an identical set of results.

Tables IX and X show the 10 predictor and response terms identified via this method as having the smallest impact. These will be discussed in detail in the following section.

6.4. Correlation of response variables

The power of multivariate data analysis relates to a high degree of correlation among the predictor variables [39, 40]. Specifically, for CCA it is desired to have a correlation in both variable sets, whereas for RRR and PLS a high degree of correlation within the predictor variable set is desirable

DOI: 10.1002/sim

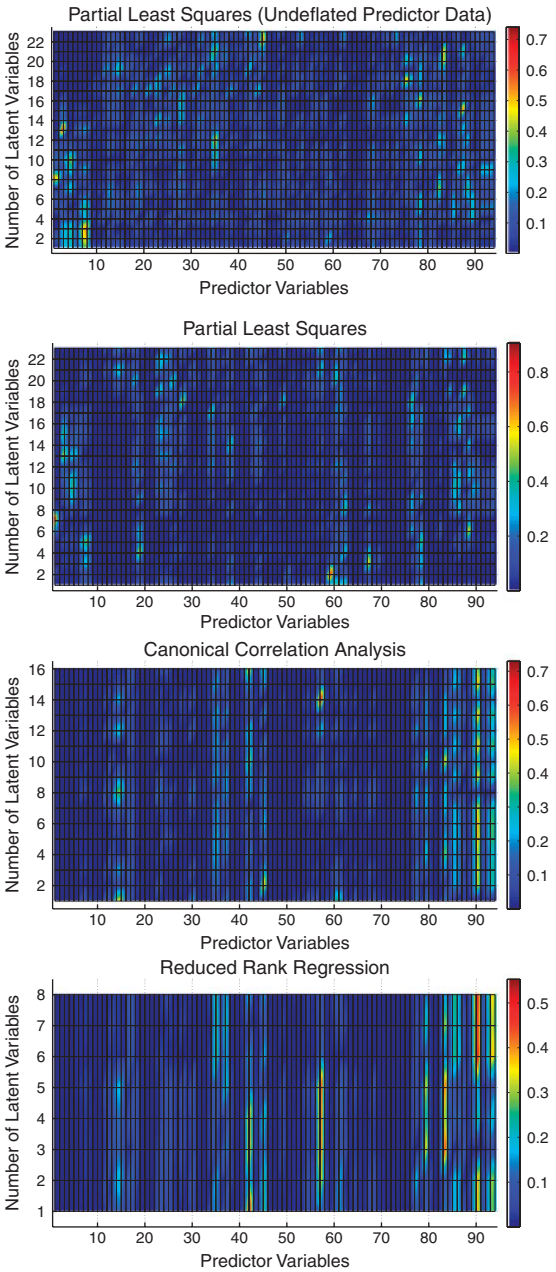

Figure 6. Surface plot showing the absolute value of the predictor variable weights for the dominant 23, 16 and 8 latent variables for PLS, CCA and RRR, respectively. An additional plot has been included for PLS, displaying the weights calculated from the original (undeflated) predictor matrix. The gradient along each image depicts the colour associated with the various weight magnitudes. The results from PLS are unique, while CCA and RRR have produced similar weights for each predictor variable.

DOI: 10.1002/sim

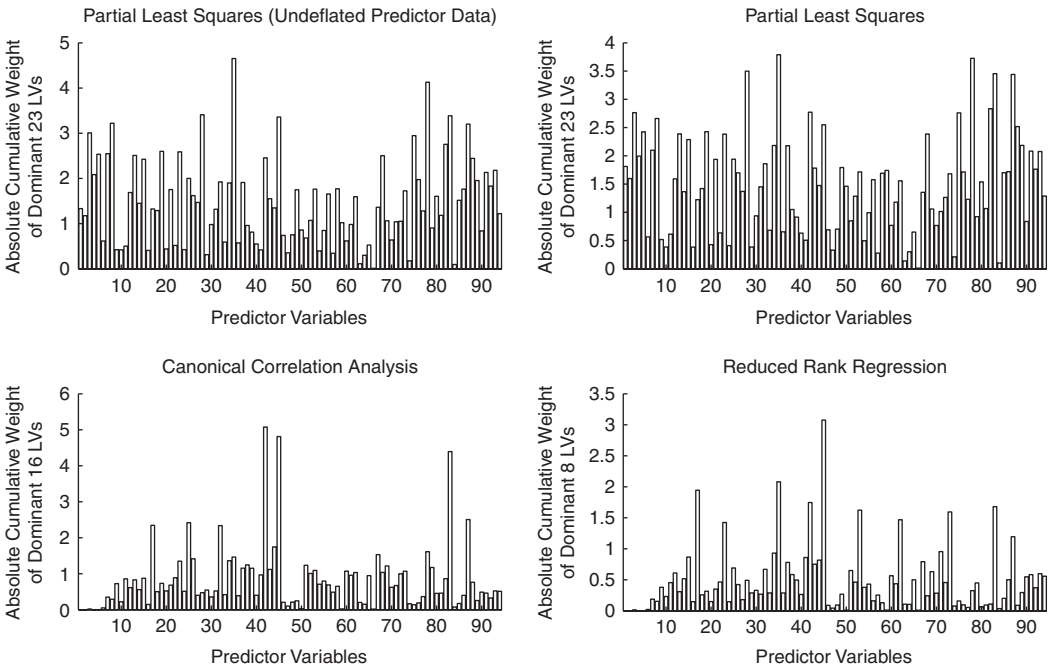

Figure 7. Bar charts corresponding to the cumulative sum of the dominant latent variable weights for each predictor variable for PLS, CCA and RRR, respectively. The results from both the original and deflated predictor terms are shown.

but correlation among the response variables is not essential. A correlation matrix, $\mathbf{R_{yy}}$, is presented in equation (35) showing a representation of the correlation among the response terms by using the first seven response variables.

$$\mathbf{R_{yy}} = \begin{bmatrix} 1.0000 & 0.9724 & 0.9996 & 0.9996 & 0.9996 & 0.9770 & 0.9622 \\ 0.9724 & 1.0000 & 0.9785 & 0.9785 & 0.9784 & 0.9995 & 0.9752 \\ 0.9996 & 0.9785 & 1.0000 & 1.0000 & 1.0000 & 0.9825 & 0.9665 \\ 0.9996 & 0.9785 & 1.0000 & 1.0000 & 1.0000 & 0.9825 & 0.9665 \\ 0.9996 & 0.9784 & 1.0000 & 1.0000 & 1.0000 & 0.9824 & 0.9665 \\ 0.9770 & 0.9995 & 0.9825 & 0.9825 & 0.9824 & 1.0000 & 0.9799 \\ 0.9622 & 0.9752 & 0.9665 & 0.9665 & 0.9665 & 0.9779 & 1.0000 \end{bmatrix} \quad (35)$$

It can be seen from this correlation matrix that the absolute values indicate a strong correlation between the response variables since $|\mathbf{r_{ij}}|>0.95$.

DOI: 10.1002/sim

Table VI. Predictor terms with smallest contributions to the model as depicted by PLS.

	Predictor term	Description	Relative weight
Undeflated predictor data			
1	**x41**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS	0.00203
2	**x57x59**	MEK-PP and Phosp2	0.02051
3	**x39**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*	0.02406
4	**x46**	SHP2*	0.03778
5	**x39x46**	(IL6-gp80-gp130-JAK*)$_2$-SHP2* and SHP2*	0.06423
6	**x16**	(Il6-gp80-gp130-JAK*)$_2$-SHP2	0.06718
7	**x37x41**	Ras-GTP and (IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS	0.07389
8	**x31**	(IL6-gp80-gp130-JAK*)$_2$-STAT3C-SOCS3	0.07596
9	**x35x40**	SOS and (IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2	0.08490
10	**x8x46**	(Il6-gp80-gp130-JAK*)$_2$ and SHP2*	0.08747
Deflated predictor data			
1	**x41**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS	0.00298
2	**x57x59**	MEK-PP and Phosp2	0.02705
3	**x39**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*	0.03675
4	**x46**	SHP2*	0.05569
5	**x37x41**	Ras-GTP and (IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS	0.07330
6	**x39x46**	(IL6-gp80-gp130-JAK*)$_2$-SHP2* and SHP2*	0.07962
7	**x31**	(IL6-gp80-gp130-JAK*)$_2$-STAT3C-SOCS3	0.08709
8	**x8x46**	(Il6-gp80-gp130-JAK*)$_2$ and SHP2*	0.10137
9	**x16**	(Il6-gp80-gp130-JAK*)$_2$-SHP2	0.10146
10	**x8x10**	(Il6-gp80-gp130-JAK*)$_2$ and (STAT3C*)	0.10183

Table VII. Predictor terms with smallest contributions to the model as depicted by CCA.

	Predictor term	Description	Relative weight
1	**x2**	IL6-gp80	0.00009
2	**x3x4**	gp130 and JAK	0.00040
3	**x5**	gp130-JAK	0.00048
4	**x41**	(IL6-gp80-gp130-JAK)*2-SHP2*-Grb2-SOS	0.00322
5	**x2x5**	IL6-gp80 and gp130-JAK	0.00382
6	**x38**	Grb2-SOS	0.00433
7	**x34x35**	Grb2 and SOS	0.00501
8	**x6**	Il6-gp80-gp130-JAK	0.00994
9	**x57x59**	MEK-PP and Phosp2	0.01471
10	**x31**	(IL6-gp80-gp130-JAK)*2-STAT3C-SOCS3	0.19790

7. INTERPRETATION

If the techniques were successful in identifying the negligible variables in the model then it follows that if these terms are then replaced by a constant value in the model, there should be little or no impact on the model accuracy. This section compares the three multivariate statistical techniques in their ability to identify both the dominant and negligible contributors to the production of (STAT3N*)$_2$. It also highlights the versatility of the techniques by recognizing their ability to reveal

DOI: 10.1002/sim

Table VIII. Predictor terms with smallest contributions to the model as depicted by RRR.

	Predictor term	Description	Relative weight
1	**x2**	IL6-gp80	0.00008
2	**x5**	gp130-JAK	0.00025
3	**x3x4**	gp130 and JAK	0.00029
4	**x41**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS	0.00301
5	**x34x35**	Grb2 and SOS	0.00336
6	**x38**	Grb2-SOS	0.00366
7	**x2x5**	IL6-gp80 and gp130-JAK	0.00482
8	**x6**	Il6-gp80-gp130-JAK	0.00675
9	**x57x59**	MEK-PP and Phosp2	0.01165
10	**x31**	(IL6-gp80-gp130-JAK)$*$2-STAT3C-SOCS3	0.19790

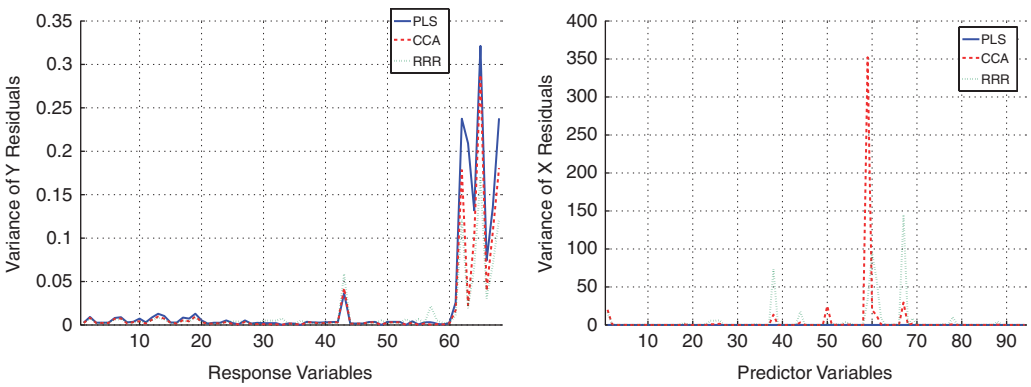

Figure 8. Variance of response and predictor variable residuals as depicted by PLS, CCA and RRR.

potential redundancy in the model, while conserving its original structure. This model reduction is then performed and the biological significance discussed.

7.1. Comparison of RRR, CCA and PLS

When examining the model from the response side, it was found that valid information was extracted from the ranking of both the v-weights and the residuals of the response variables. As Figure 4 highlights each technique suggests a different ranking for the v-weights but with CCA proving to be the only technique whose results can be directly used to perform a reduction on the IL6 signal transduction model. It is worth noting that the relative weights associated with the terms using CCA are significantly smaller in magnitude than those from the other techniques, which were less successful in this task (see Table IV). All techniques generated a similar ranking using the variance of the residuals and were equally successful in identifying further suitable terms for reduction in the model.

When examining the model from the predictor side, CCA and RRR showed a strong correlation. However, the specific ranking for the techniques meant that neither these techniques nor PLS

DOI: 10.1002/sim

Table IX. Predictor terms with smallest contributions to the model as depicted by the variances of the residuals for PLS, CCA and RRR.

		Predictor term	Description
PLS	1	**x29**	SOCS3
	2	**x37x41**	Ras-GTP and (IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS
	3	**x26**	mRNA-SOCS3C
	4	**x21x22**	STAT3N* and STAT3N
	5	**x38x46**	Grb2-SOS and SHP2*
	6	**x8x45**	(IL6-gp80-gp130-JAK*)$_2$ and SHP2*-Grb2
	7	**x50x51**	Phosp1 and Raf*
	8	**x15x31x46**	SHP2 and (IL6-gp80-gp130-JAK*)$_2$-STAT3C-SOCS3
	9	**x8x44**	(IL6-gp80-gp130-JAK*)$_2$ and SHP2*-Grb2-SOS
	10	**x9x10**	STAT3C and STAT3C*
CCA	1	**x38**	Grb2-SOS
	2	**x41x49**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS and Ras-GTP*
	3	**x34x35**	Grb2 and SOS
	4	**x38x39**	Grb2-SOS and (IL6-gp80-gp130-JAK*)$_2$-SHP2*
	5	**x21x23**	STAT3N* and PP2
	6	**x38x46**	Grb2-SOS and SHP2*
	7	**x28**	PP2-STAT3N*
	8	**x43**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS-Ras-GTP
	9	**x14**	STAT3C-STAT3C*
	10	**x51x53**	Raf* and MEK
RRR	1	**x41x49**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS and Ras-GTP*
	2	**x38x39**	Grb2-SOS and (IL6-gp80-gp130-JAK*)$_2$-SHP2*
	3	**x21x23**	STAT3N* and PP2
	4	**x38x46**	Grb2-SOS and SHP2*
	5	**x38**	Grb2-SOS
	6	**x28**	PP2 − STAT3N*
	7	**x51x53**	Raf* and MEK
	8	**x38x46x46**	Grb2-SOS and SHP2*
	9	**x43**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS-Ras-GTP
	10	**x14**	STAT3C-STAT3C*

successfully identified any state variable that could be removed/replaced by its steady-state value without affecting the model accuracy.

From an inspection of the residuals of the predictor terms it was found that RRR and CCA ranked terms similarly but on this occasion CCA was the only technique successful in identifying a further variable that could be removed from the model.

In summary, CCA was the most successful tool for the application to this model for identifying suitable predictor and response data to permit a model reduction. The balanced approach of CCA lends itself well to analyzing this signal transduction pathway model since the predictor–response relationships are projected into the latent variable space and reduction occurs on both sides of the problem. It is advantageous to point out that a well studied feature of CCA is the fact that it will not identify areas of redundancy in a non-redundant model [41].

In contrast, RRR is heavily focused on producing an accurate response data set. This approach produces weights for the predictor data, which are in close agreement with those from CCA and weights for the response data, which completely differ significantly.

DOI: 10.1002/sim

Table X. Response terms with smallest contributions to the model as depicted by residual response matrix from PLS, CCA and RRR.

	Response term		Description
PLS	1	**dx65**	ERK-PP
	2	**dx68**	ERK-P-Phosp3
	3	**dx62**	ERK-MEK-PP
	4	**dx63**	ERK-P
	5	**dx67**	ERK-PP-Phosp3
	6	**dx64**	ERK-P-MEK-PP
	7	**dx66**	Phosp3
	8	**dx43**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS-Ras-GTP
	9	**dx61**	ERK
	10	**dx19**	PP1-(STAT3C*)$_2$
CCA	1	**dx65**	ERK-PP
	2	**dx68**	ERK-P-Phosp3
	3	**dx62**	ERK-MEK-PP
	4	**dx67**	ERK-PP-Phosp3
	5	**dx64**	ERK-P-MEK-PP
	6	**dx43**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS-Ras-GTP
	7	**dx66**	Phosp3
	8	**dx63**	ERK-P
	9	**dx61**	ERK
	10	**dx19**	PP1-(STAT3C*)$_2$
RRR	1	**dx65**	ERK-PP
	2	**dx68**	ERK-P-Phosp3
	3	**dx62**	ERK-MEK-PP
	4	**dx67**	ERK-PP-Phosp3
	5	**dx64**	ERK-P-MEK-PP
	6	**dx43**	(IL6-gp80-gp130-JAK*)$_2$-SHP2*-Grb2-SOS-Ras-GTP
	7	**dx66**	Phosp3
	8	**dx57**	MEK-PP
	9	**dx63**	ERK-P
	10	**dx61**	ERK

PLS, on the other hand, only looks at covariance and is not aiming to produce a perfect model. This results in the deflation of the predictor data set with much fewer variables compared with the other techniques but it takes many LVs to offer the same deflation for the response data, as can be seen in Figure 2. If all of the LVs are included then the same results would be produced as CCA/RRR, but it still fails to identify any underlying latent variable structure and as a result none of the information from the response or predictor weights provided useful information for this application.

The comparison of the predictor/response results for PLS, CCA and RRR, made possible by examining both the variance of residuals and the marginal sum of weights, provides a comprehensive examination that permits the following conclusions for the analysis of the IL6 model.

- CCA is the most successful tool for comprehensively analysing the IL6 model.
- The most significant ranking of the predictor terms can be found by examining the variance of the residuals.

DOI: 10.1002/sim

- The most significant ranking of the response terms involves an analysis of both the variance of the residuals and the marginal sum of weights, with the latter offering the best result if used alone.
- The ranking produced by examining the variance of the response terms shows a high degree of correlation for PLS, CCA and RRR.
- The ranking produced by examining the marginal sum of weights of the response terms shows no correlation for PLS, CCA and RRR (see Figures 4 and 5 for a visual representation).
- The ranking produced by examining the variance of the predictor terms shows a high degree of correlation between CCA and RRR.
- The ranking produced by examining the marginal sum of weights of the predictor terms shows a high degree of correlation between CCA and RRR (see Figures 4 and 5 for a visual representation).
- The ranking produced by examining the marginal sum of weights of the predictor terms calculated from the original predictor matrix proved similar to that arrived at by analysing the deflated predictor matrix (see Table VI).

7.2. Model reduction

Using the results from the CCA v-weights allowed for the first eight derivative terms from Table IV to be set equal to zero in the model. Also, the two further terms identified from an examination of the response residuals, namely $\mathbf{dx}_{62}$ and $\mathbf{dx}_{64}$, which represent ERK-MEK-PP and ERK-P-MEK-PP, respectively, as shown in Table X, can also be set equal to zero in the model. These specific proteins have been identified for removal but this does not permit the removal of any compound of, or reaction containing them.

The suggested w-weight ranking did not identify any terms that could be replaced by their steady-state value but the ranking produced from the residuals suggested that Grb2-SOS could be replaced by its steady-state value of 33.1715 nM (see Table IX). Therefore, the reduced model consists of 58 ODE's (or 56 ODE's if it is assumed that $\mathbf{dx}_1$ and $\mathbf{dx}_3$ are set to zero as in the original model) and 89 predictor terms. The ODE's which were set to zero in the reduced model contained four predictor terms, which did not appear anywhere else. Therefore, these terms were also eliminated from the model by default. Table XI contains the description of these four predictor terms. A summary of the terms identified for the model reduction along with their associated ODEs and kinetic constants are presented in Tables XII and XIII. It can be seen from these tables that the kinetic constants are of varying magnitudes and hence could not have been identified without the use of a statistical analysis.

Table XI. Predictor terms eliminated from the model due to the removal of their respective ODE's.

Predictor term	Description
x63x66	ERK-P and Phosp3
x65x66	ERK-PP and Phosp3
x67	ERK-PP-Phosp3
x68	ERK-P-Phosp3

DOI: 10.1002/sim

Table XII. Summary of terms identified for model reduction along with their respective differential equations.

Term	Description	ODE
x38	Grb2-SOS	associated kinetic constant = $\mathbf{k_{f35}}$
dx59	Change in Phosp2	$\mathbf{k}_{49}\mathbf{x}_{58} - \mathbf{k_{f48}x_{57}x_{59}} + \mathbf{k_{r48}x_{58}} - \mathbf{k_{f50}x_{55}x_{59}} + \mathbf{k_{r50}x_{60}} + \mathbf{k}_{51}\mathbf{x}_{60}$
dx60	Change in MEP-P-Phosp2	$\mathbf{k_{f50}x_{55}x_{59}} - \mathbf{k_{r50}x_{60}} - \mathbf{k}_{51}\mathbf{x}_{60}$
dx61	Change in ERK	$-\mathbf{k_{f52}}x_{57}x_{61} + \mathbf{k_{r52}}x_{62} + k_{59}x_{68}$
dx62	Change in ERK-MEK-PP	$\mathbf{k_{f52}x_{57}x_{61}} - \mathbf{k_{r52}x_{62}} - \mathbf{k}_{53}\mathbf{x}_{62}$
dx63	Change in ERK-P	$\mathbf{k}_{53}\mathbf{x}_{62} - \mathbf{k_{f54}x_{57}x_{63}} + \mathbf{k_{r54}x_{64}} + \mathbf{k}_{57}\mathbf{x}_{67} - \mathbf{k_{f58}x_{63}x_{66}} + \mathbf{k_{r58}x_{68}}$
dx64	Change in ERK-P-MEK-PP	$\mathbf{k_{f54}x_{57}x_{63}} - \mathbf{k_{r54}x_{64}} - \mathbf{k}_{55}\mathbf{x}_{64}$
dx65	Change in ERK-PP	$\mathbf{k}_{55}\mathbf{x}_{64} - \mathbf{k_{f56}x_{65}x_{66}} + \mathbf{k_{r56}x_{67}}$
dx66	Change in Phosp3	$-\mathbf{k_{f56}x_{65}x_{66}} + \mathbf{k_{r56}x_{67}} + \mathbf{k_{r57}x_{67}} - \mathbf{k_{f58}x_{63}x_{66}} + \mathbf{k_{r58}x_{68}} + \mathbf{k}_{59}\mathbf{x}_{68}$
dx67	Change in ERK-PP-Phosp3	$\mathbf{k_{f56}x_{65}x_{66}} + \mathbf{k_{r56}x_{67}} + \mathbf{k}_{57}\mathbf{x}_{67}$
dx68	Change in ERK-P-Phosp3	$\mathbf{k_{f58}x_{63}x_{66}} - \mathbf{k_{r58}x_{68}} - \mathbf{k}_{59}\mathbf{x}_{68}$

Table XIII. Values of kinetic constants associated with terms identified for model reduction. First-order rate constants are in units s^{-1} and second-order rate constants are expressed in $nM^{-1}s^{-1}$.

Kinetic constants	
$\mathbf{k_{f35}}$=0.0015	
$\mathbf{k_{f48}} = 1.43 \times 10^{-2}$	$\mathbf{k_{r48}} = 0.8$
$\mathbf{k_{49}} = 0.058$	
$\mathbf{k_{f50}} = 2.7 \times 10^{-4}$	$\mathbf{k_{r50}} = 0.5$
$\mathbf{k_{51}} = 0.058$	
$\mathbf{k_{f52}} = 1.1 \times 10^{-4}$	$\mathbf{k_{r52}} = 0.033$
$\mathbf{k_{53}} = 16$	
$\mathbf{k_{f54}} = 1.1 \times 10^{-4}$	$\mathbf{k_{r54}} = 0.033$
$\mathbf{k_{55}} = 5.7$	
$\mathbf{k_{f56}} = 1.4 \times 10^{-2}$	$\mathbf{k_{r56}} = 0.6$
$\mathbf{k_{57}} = 0.27$	
$\mathbf{k_{f58}} = 5 \times 10^{-3}$	$\mathbf{k_{r58}} = 0.5$
$\mathbf{k_{59}} = 0.3$	

Applying these changes yields a reduced model with a departure from the original model of 0.9185 per cent. The accuracy is measured as described in equation (36) and involves calculating the departure of the dynamic profile of $(STAT3N^*)_2$ produced by the reduced model from that of the original model. Figure 9 shows the dynamic profile for $(STAT3N^*)_2$ in the original model along with a trace of the degree at which the reduced model departs from this.

$$\text{Departure from original model (per cent)} = \frac{\int_{t=0}^{T} |(\mathbf{STAT3N^*})_{2(\mathbf{original})} - (\mathbf{STAT3N^*})_{2(\mathbf{reduced})}| \mathrm{d}t}{\int_{t=0}^{T} (\mathbf{STAT3N^*})_{2(\mathbf{original})} \mathrm{d}t} \tag{36}$$

DOI: 10.1002/sim

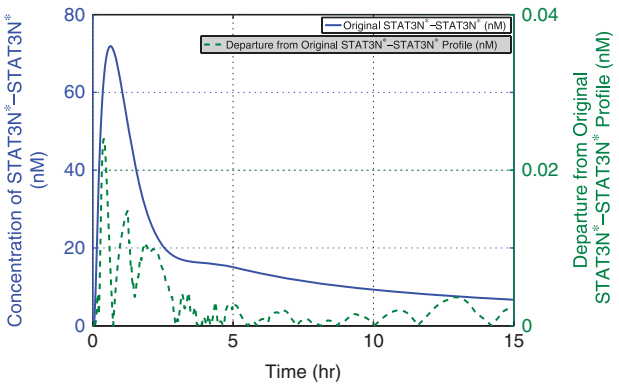

Figure 9. Dynamic profile of $(STAT3N^*)_{2(\mathbf{original})}$ and the departure from this profile for the reduced model.

where $\mathbf{t=0}$ is the time at which the simulation begins and $\mathbf{T=15\,h}$ is the time at which the simulation terminates.

If any further predictor or response terms are removed from the model, the accuracy of the reduced model is diminished by over 1000 fold. For this reason, any further reduction is not permitted. It is important to note that this cut-off point is driven by and specific to this model. If a different system is being studied, then this cut-off point should be re-examined.

A pictorial representation of this reduced model is presented in Figure 10.

7.3. Biological relevance

One cross-product term was identified as causing a small enough change in the state variables that it could be removed from the model without causing a significant impact. This mechanism, which was identified via the inspection of the response residuals is that of Grb2-SOS.

The response terms that were successfully removed from the model without significantly impacting the model accuracy are all components from the terminal end of the modelled MAPK pathway. This set of reactions that do not have a significant impact on the model accuracy starts to occur just after the point at which the phosphatase Phosp2 acts on MEK-PP to convert it to MEK-P.

Since the MAPK pathway does not lead to the production of the STAT3 nuclear dimer it is not surprising that a certain degree of redundancy was be identified in this pathway when measuring the accuracy of the reduced model against the dynamic profile of $(STAT3N^*)_{2(\mathbf{original})}$. However, the results do indicate that although the pathway does not directly produce this transcription factor, it plays a significant role in regulating it through cross-talk with the JAK/STAT pathway. If no cross-talk was evident then the entire pathway would have been identified as redundant. A detailed investigation into this cross-talk between these two pathways has been carried out [10].

8. CONCLUSIONS

In this work a kinetic model describing the signal transduction in hepatocytes stimulated by IL6 is analysed using PLS, CCA and RRR.

DOI: 10.1002/sim

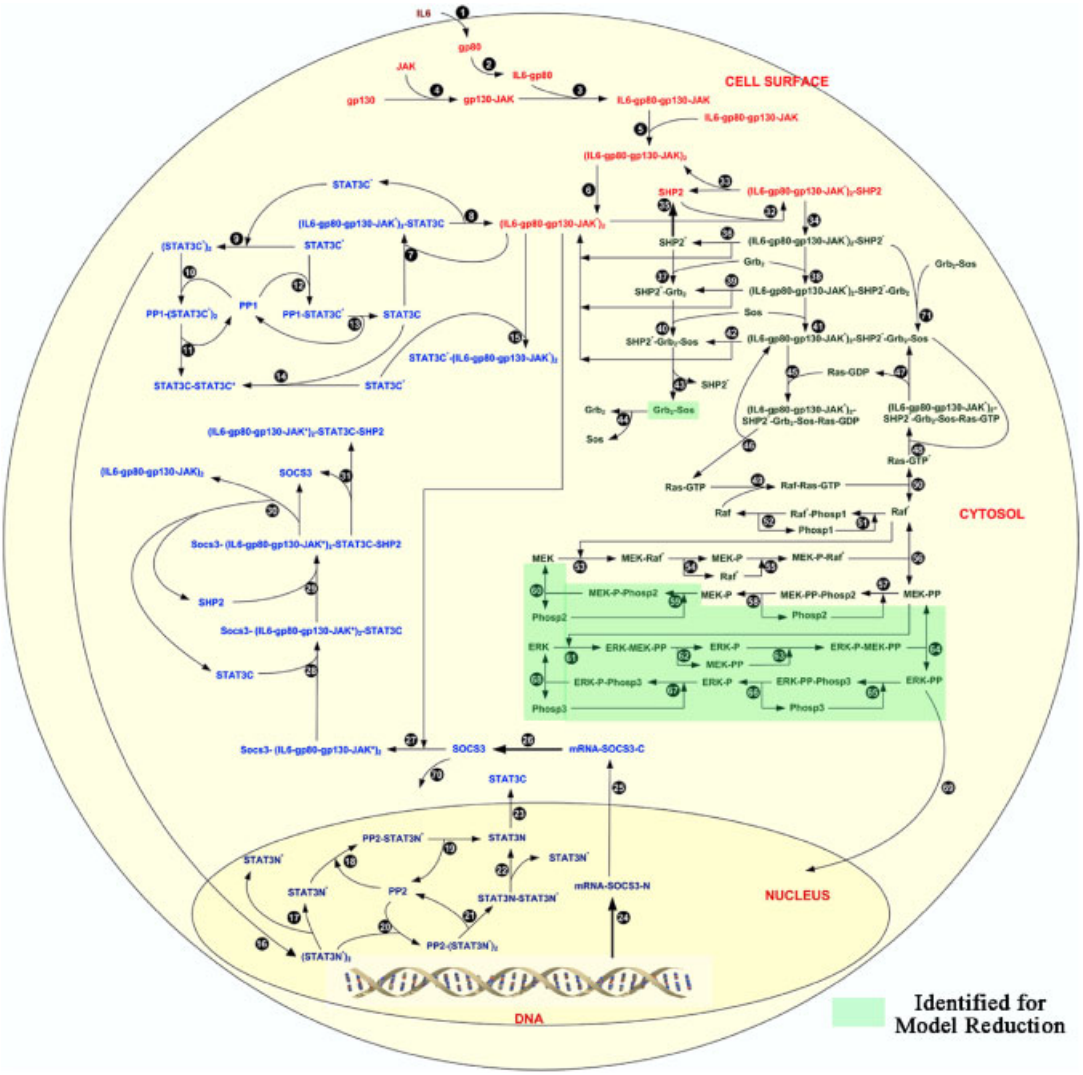

Figure 10. Pictorial representation of the steps involved in the reduced model.

CCA, which has found its application of analysing variable interrelationships in fields as diverse as biometrics [42, 43], economics [44], social sciences [45] and criminology [46], proved to be the most successful algorithm in identifying the least significant terms in the model for both the predictor and response variables, by using both the information from the weight vectors and from the residuals of both the predictor and response terms.

An analysis of the resulting weight vectors proved that PLS was unsuccessful in identifying any underlying latent variable structure and therefore produced an inaccurate ranking for the relative contribution of terms to the model.

DOI: 10.1002/sim

RRR, in contrast, failed to identify any derivative terms suitable for model reduction but its ranking of the w-weights (predictor weights) showed a high degree of correlation with the predictor rankings given by CCA. Despite the fact that in this analysis the specific ranking of the predictor terms from their associated w-weights did not lead to a model reduction, the Grb2-SOS protein, which was successfully removed and identified via an examination of the predictor residuals, was also recognized by both CCA and RRR among the 10 least significant contributors. As such, this suggests that both CCA and RRR could successfully be used to identify the cross-product terms that may not play a crucial role in the signalling pathway.

The variance of the residuals of the predictor data set was different for each technique, with CCA and RRR showing some correlation in the ranking of terms. Despite these similarities, CCA was the only technique, whose specific ranking identified a variable that could successfully be removed from the model. The response residuals generated by each technique were in agreement and offered further information towards the areas of the model which may warrant reduction.

Although this analysis has a similar objective to conventional model reduction techniques, the application of multivariate statistical tools is advantageous as it does not produce linear combinations of the physical state variables. More precisely, the model interpretation remains unchanged, which implies that the remaining states (after the reduction process) have the same meaning as those prior to the application of the reduction process.

A model reduction was performed using the results from CCA, along with the extra confidence presented by the RRR predictor weights/residuals and the CCA/RRR response residuals. The analysis suggests that there are areas on the MAPK pathway, which do not contribute to the resulting concentration of $(\mathrm{STAT3N}^*)_2$ and a simplified model is presented.

The identification of this region of the pathway suggests that it holds little or no significance to the regulation of the nuclear STAT3 dimer, or alternatively that it is currently incomplete. It has been suggested in the literature that there is a missing feedback loop on this pathway, in the same area, which was identified through this analysis [47]. If indeed this feedback loop does play a role in the IL6 signal transduction pathway and the model was refined to reflect this, then the importance of some of these reactions may prove different from the results found in this analysis.

APPENDIX A

A.1. Proof of theorem 1

Formulating iterative CCA and RRR algorithms requires the incorporation of a deflation procedure for the predictor matrices:

$$\begin{aligned}
\mathbf{Z_{k+1}} &= \mathbf{Z_k} - \mathbf{t_k p_k^T} \\
\mathbf{Z_{k+1}} &= \mathbf{Z_k} - \mathbf{Z_k w_k p_k^T} \\
\mathbf{Z_{k+1}} &= \mathbf{Z_k}[\mathbf{I} - \mathbf{w_k p_k^T}] \\
\mathbf{Z_{k+1}} &= \mathbf{Z_{k-1}}[\mathbf{I} - \mathbf{w_{k-1} p_{k-1}^T}][\mathbf{I} - \mathbf{w_k p_k^T}] \\
\mathbf{Z_{k+1}} &= \mathbf{Z}[\mathbf{I} - \mathbf{w_1 p_1^T}] \cdots [\mathbf{I} - \mathbf{w_k p_k^T}]
\end{aligned} \tag{A1}$$

DOI: 10.1002/sim

where $\mathbf{p_k} = \mathbf{Z_k^T t_k}$ represents the $\mathbf{k}$th loading vector of the predictor matrix. To show that $\mathbf{t_{k+1}} = \mathbf{Z_{k+1} w_{k+1}} = \mathbf{Z w_{k+1}}$ requires that $\mathbf{p_i^T w_{k+1}} = 0 \quad \forall \mathbf{i} < \mathbf{k}+1$:

$$\mathbf{Z_{k+1} w_{k+1}} = \mathbf{Z}[\mathbf{I} - \mathbf{w}_1 \mathbf{p}_1^{\mathbf{T}}] \cdots [\mathbf{I} - \mathbf{w_k p_k^T}] \mathbf{w_{k+1}} = \mathbf{Z w_{k+1}} \tag{A2}$$

Geometrically, it can be shown that $\mathbf{p_i^T w_j} = \boldsymbol{\delta}_{\mathbf{ij}}$, which follows from:

$$\mathbf{w_i^T p_{j=w_i^T} Z_j^T t_j} = \mathbf{t_i^T t_j} = \delta_{\mathbf{ij}} \tag{A3}$$

Equation (A2) is valid since both, CCA and RRR require the length of the t-score vector to be length one, as expressed in equations (6) and (7) (CCA) and equations (14) and (15) (RRR). To show that the t-score vectors are mutually orthonormal, which is required to guarantee that $\mathbf{p_i^T w_j} = \boldsymbol{\delta}_{\mathbf{ij}}$, we assuming that $\mathbf{i} < \mathbf{j}$. Analysing the deflation procedure and expressing $\mathbf{t_i^T t_j}$ as $\mathbf{t_i^T Z_j w_j}$ gives rise to:

$$\begin{aligned} \mathbf{t_i^T t_j} &= \mathbf{t_i^T}[\mathbf{I_K} - \mathbf{t_i t_i^T}] \mathbf{Z_i} \left[\prod_{\mathbf{k=i}}^{\mathbf{j-1}} [\mathbf{I_N} - \mathbf{w_k p_k^T}] \right] \mathbf{w_j} \\ \mathbf{t_i^T t_j} &= (\mathbf{t_i^T} - \mathbf{t_i^T}) \mathbf{Z_i} \left[\prod_{\mathbf{k=i}}^{\mathbf{j-1}} [\mathbf{I_N} - \mathbf{w_k p_k^T}] \right] \mathbf{w_j} = \mathbf{0} \quad \text{hence } \mathbf{t_i^T t_j} = \boldsymbol{\delta}_{\mathbf{ij}} \end{aligned} \tag{A4}$$

Note that in cases where $\mathbf{i} > \mathbf{j}$, the deflation procedure of $\mathbf{Z_i}$ can be expressed as $[\mathbf{I_K} - \mathbf{t_j t_j^T}] \mathbf{Z_j} \left[\prod_{\mathbf{k=j}}^{\mathbf{i-1}} [\mathbf{I} - \mathbf{w_k p_k^T}] \right]$ and applying the same procedure gives accordingly $\mathbf{t_j^T} - \mathbf{t_j^T} = \mathbf{0}$.

It should also be noted that the deflation procedure has the same effect upon the algorithm as imposing the constraints in equation (7) for CCA and equation (15) for RRR. It can also be shown that iterative CCA/RRR algorithms produce identical weight and score vectors to those obtained by those of the batch algorithm in Table II. Such a detailed analysis, however, is beyond the scope of this article.

REFERENCES

1. Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. *Science* 2004; **306**:640–643.
2. Kitano H. Systems biology: a brief overview. *Science* 2002; **295**:1662–1664.
3. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annual Review of Geneomics and Human Genetics* 2001; **2**:343–372.
4. Gadkar KG, Varner J, Doyle III FJ. Model identification of signal transduction networks from data using a state regulator problem. *Systems Biology* 2005; **2**(1):17–30.
5. Singh A, Jayaraman A, Hahn J. A case study representing signal transduction in liver cells as a feedback control problem. *Chemical Engineering Education* 2007; **41**(3):177–182.
6. Janes KA, Yaffe MB. Data-driven modelling of signal-transduction networks. *Nature Reviews*: *Molecular Cell Biology* 2006; **7**(11):820–828.
7. Kholodenko BN. Cell-signalling dynamics in time and space. *Nature Reviews*: *Molecular Cell Biology* 2006; **7**(3):165–176.
8. Jaqaman K, Danuser G. Linking data to models: data regression. *Nature Reviews*: *Molecular Cell Biology* 2006; **7**:813–819.
9. Burnham AJ, Viveros R, MacGregor JF. Frameworks for latent variable multivariate regression. *Journal of Chemometrics* 1996; **10**(1):31–45.
10. Singh A, Jayaraman A, Hahn J. Modeling regulatory mechanisms in IL-6 signal transduction in hepatocytes. *Biotechnology and Bioengineering* 2006; **95**(5):850–862.

DOI: 10.1002/sim

11. Kushner I, Mackiewicz A, Baumann H. *Acute Phase Proteins*: *Molecular Biology*, *Biochemistry and Clinical Applications*. CRC Press: Boca Raton, 1993.
12. Baumann H, Gauldie J. The acute phase response. *Immunology Today* 1994; **15**(2):74–80.
13. Cerra FB. The systemic septic response: concepts of pathogenesis. *Journal of Trauma* 1990; **30**(12 Suppl.): S169–S174.
14. Beal AL, Cerra FB. Multiple organ failure syndrome in the 1990s. Systemic inflammatory response and organ dysfunction. *Journal of the American Medical Association* 1994; **271**(3):226–233.
15. Pinilla JC, Hayes P, Laverty W, Arnold C, Laxdal V. The C-reactive protein to prealbumin ratio correlates with the severity of multiple organ dysfunction. *Surgery* 1998; **124**:799–806.
16. Anderson TW. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics* 1951; **22**(3):327–351.
17. Hotelling H. The most predictable criterion. *Journal of Educational Psychology* 1935; **26**:139–142.
18. Wold H. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, Krishnaiaah PR (ed.). Academic Press: New York, 1966; 391–420.
19. Chu Y, Jayaraman A, Hahn J. Parameter sensitivity analysis of IL-6 signaling pathways. *IET Systems Biology* 2007; **1**(6):342–352.
20. Chu Y, Hahn J. Selection of parameter sets and design of experiments for estimation of nonlinear dynamic systems. *Proceedings of the IFAC World Congress*, Seoul, Korea, 2008; 5545–5550, to appear.
21. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 1986; **185**:1–17.
22. Höskuldsson A. PLS regression methods. *Journal of Chemometrics* 1988; **2**:211–228.
23. Golub GH. Matrix decompositions and statistical calculations. *Statistical Computation*. Academic Press: New York, 1969.
24. MacArthur RH. On the relative abundance of bird species. *Proceedings of the National Academy of Science*, U.S.A., vol. 43, 1957; 293–295.
25. Jackson DA. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 1993; **74**(8):2204–2214.
26. Cangelosi R, Goriely A. Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct* 2007; **2**(2). DOI:10.1186/1745-6150-2-2.
27. Imada K, Leonard WJ. The JAK-STAT pathway. *Molecular Immunology* 2000; **37**:1–11.
28. Yamada S, Shiono S, Joo A, Yoshimura A. Control mechanism of JAK/STAT signal transduction pathway. *Federation of Biochemical Societies* (*FEBS*) *Letters* 2003; **534**:190–196.
29. Kholodenko BN. Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. *Federation of Biochemical Societies* (*FEBS*) *Letters* 2000; **267**:1583–1588.
30. Levy DE, Darnell Jr JE. STATS: transcriptional control and biological impact. *Nature Reviews*: *Molecular Cell Biology* 2002; **3**:651–662.
31. Hensen MA, Seborg DE (eds). *Nonlinear Process Control*. Prentice-Hall: NJ, U.S.A., 1997.
32. Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnology* 2002; **20**:370–375.
33. ten Hoeve J, Ibarra-Sanchez M, Fu Y, Zhu W, Tremblay M, David M, Shuai K. Identification of a nuclear Stat1 protein tyrosine phosphatase. *Molecular and Cellular Biology* 2002; **22**:5662–5668.
34. Lang R, Pauleau AL, Parganas E, Takahashi Y, Mages J, Ihle JN, Rutschman R, Murray PJ. SOCS3 regulates the plasticity of gp130 signaling. *Nature Immunology* 2003; **4**(6):546–550.
35. Jackson JE. *A User's Guide to Principal Components*. Wiley: New York, 1991.
36. Holland J. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, 1975.
37. Bäck T, Hammel U, Schwefel H. Evolutionary computation: comments on the history and current state. *IEEE Transactions on Evolutionary Computation* 1997; **1**(1):3–17.
38. Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalised inverses. *SIAM Journal of Scientific Statistical Computations* 1984; **5**(3):735–743.
39. Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. Academic Press: London, 1989.
40. Anderson TW. *An Introduction to Multivariate Statistical Analysis*. Wiley: New York, 2003.
41. Muller KE. Understanding canonical correlation through the general linear model and principal components. *The American Statistician* 1982; **36**(4):342–354.
42. Alonso A, Geys H, Molenberghs G, Kenward MG, Vangeneugden T. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. *Biometrics* 2004; **60**: 845–853.

DOI: 10.1002/sim

43. Kowalski J, Tu XM, Jia G, Perlis M, Frank E, Crits-Christoph P, Kupfer DJ. Generalized covariance-adjusted canonical correlation analysis with application to psychiatry. *Statistics in Medicine* 2003; **22**:595–610.
44. Dunn JW, Doeksen GA. Canonical correlation analysis of selected demographic and health personnel variables. *Southern Journal of Agricultural Economics* 1977; **9**(1):95–99.
45. Dijksterhuis G (ed.). *Multivariate Data Analysis in Sensory and Consumer Science*. Wiley-Blackwell: Oxford, U.K., 2008.
46. Holland TR, Levi M, Beckett GE. Associations between violent and nonviolent criminality: a canonical contingency-table analysis. *Multivariate Behavioral Research* 1981; **16**:237–241.
47. Huang Z, Chu Y, Senocak F, Jayaraman A, Hahn J. Model update of signal transduction pathways in hepatocytes based upon sensitivity analysis. *Proceedings of the Foundations of Systems Biology and Engineering* (*FOSBE*), 2007; 45–50.

DOI: 10.1002/sim