

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

Latent Models for Cross-Covariance

Jacob A. Wegelin

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2001

Program Authorized to Offer Degree: Statistics

UMI Number: 3014048



UMI Microform 3014048

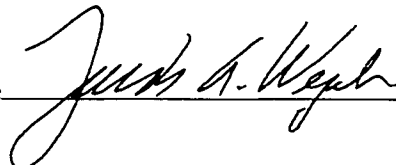
Copyright 2001 by Bell & Howell Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature



Date

07 June 2001

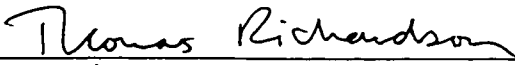
University of Washington
Graduate School

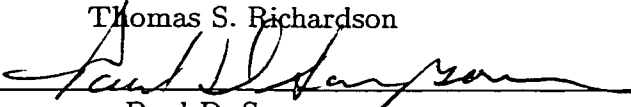
This is to certify that I have examined this copy of a doctoral dissertation by

Jacob A. Wegelin

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.


Co-Chairs of Supervisory Committee:

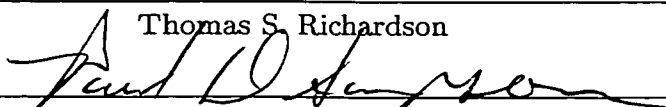


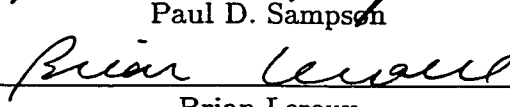
Thomas S. Richardson


Paul D. Sampson

Reading Committee:



Thomas S. Richardson


Paul D. Sampson


Brian Leroux

Date: 07 June 2001

University of Washington

Abstract

Latent Models for Cross-Covariance

by Jacob A. Wegelin

Co-Chairs of Supervisory Committee:

Associate Professor Thomas S. Richardson
Statistics

Professor Paul D. Sampson
Statistics

Cross-covariance problems arise in the analysis of multivariate data that can be divided naturally into two blocks of variables, \mathbf{X} and \mathbf{Y} , observed on the same units. In a cross-covariance problem we are interested, not in the within-block covariances, but in the way the \mathbf{Y} s vary with the \mathbf{X} s.

In the current work several approaches to the cross-covariance problem are discussed, including Reduced-Rank Regression (RRR), Canonical Correlation Analysis (CCA), Partial Least Squares (PLS, also called Projection to Latent Structures), Structural Equation Models (SEM), and Graphical Markov Models (GMM).

A family of latent models for cross-covariance, called **paired latent models**, is specified. It is shown that the set of covariance matrices which can be modeled under the rank- r paired latent model is the same as those which can be modeled under rank- r Reduced-Rank Regression. The degree to which the parameters of the rank-one paired latent model are underidentified is precisely characterized, and a natural convention is proposed which makes the model identifiable. This result has implications for the estimation of correlation between the latent variables.

It is shown that symmetric and asymmetric versions of the paired latent model are

covariance equivalent, and that this equivalence fails when the within-block covariance is constrained to be diagonal.

TABLE OF CONTENTS

List of Figures	iii
Chapter 1: Problem definition, terminology, and notation	1
1.1 Introduction	1
1.2 Path diagrams	9
1.3 M-separation	12
1.4 Classes of latent models for cross-covariance	13
Chapter 2: Current frameworks for cross-covariance problems	16
2.1 Canonical correlation analysis (CCA)	16
2.2 Structural Equation Models	17
2.3 Partial least squares (PLS)	20
2.4 Conclusion	31
Chapter 3: Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case	33
3.1 Abstract	33
3.2 Framework	34
3.3 Background and Overview	35
3.4 Framework for Two-Block PLS	39
3.5 PLS-W2A	40
3.6 PLS-SVD	48
3.7 Rank and Orthogonality in PLS-W2A and PLS-SVD	49
3.8 Herman Wold's Original PLS Algorithm	49
3.9 PLS2 and PLS1	59

3.10	PLS-W2A, PLS-SVD, PLS2: Summary of differences and similarities	61
3.11	Canonical Correlation Analysis (CCA), alias Mode B PLS	61
3.12	The Power Method in PLS	64
3.13	Proofs of PLS-W2A Properties	68
3.14	Proof of Convergence of the Power Method, Lemma 3.12.1	75
3.15	S-PLUS code for PLS-W2A	78
Chapter 4:	Rank-one latent models for cross-covariance	83
4.1	Introduction	83
4.2	Model specification	83
4.3	Maps between spaces of models	87
4.4	Examples	94
4.5	Discussion	103
4.6	Appendix	107
Chapter 5:	Rank-r latent models for cross-covariance	118
5.1	Model specification	118
5.2	Maps between spaces of models	121
5.3	Example	126
5.4	Discussion	127
Chapter 6:	Related equivalence results	129
Chapter 7:	Future Work	135
7.1	Future Work	135
Bibliography		139
Appendix A:	Equivalence of symmetric and asymmetric constraint models of rank R	146

LIST OF FIGURES

1.1	Path diagram representing the symmetric cross-diagonal parameterization of a rank R latent model. Error terms are displayed in this figure. The same model is displayed without error terms in Figure 1.2 on page 11. Path diagrams are introduced in Section 1.2. The cross-diagonal parameterization is introduced in Section 1.1.	10
1.2	Path diagram representing the symmetric cross-diagonal parameterization of a rank R latent model. This represents exactly the same model as is represented by Figure 1.1. The only difference between the path diagrams is that error terms are not displayed in this one. Error terms are always understood, however. Path diagrams are introduced in Section 1.2. The cross-diagonal parameterization is introduced in Section 1.1.	11
1.3	Path diagram representing an unrestricted rank R latent model, as described in Section 1.4.	14
1.4	Path diagram representing a within-block-diagonal rank R latent model, as described in Section 1.4.	15
1.5	Path diagram representing a double-diagonal rank R latent model, as described in Section 1.4.	15
2.1	A misspecified model, as discussed on page 25. The analyst believes that the data are properly divided into two blocks, $\mathbf{X}_{.1}, \dots, \mathbf{X}_{.5}$ and $\mathbf{Y}_{.1}, \dots, \mathbf{Y}_{.4}$, and that the cross-covariance is of rank one. A direct dependency exists however between $\mathbf{X}_{.5}$ and $\mathbf{Y}_{.4}$, making the rank-one model inappropriate.	25
2.2	Leave-one-out salience plot for a correctly specified model, as discussed in Section 2.3.1.	27

2.3	Leave-one-out salience plot for a misspecified model, as discussed on page 26.	28
3.1	Path diagram for a study of fetal alcohol exposure and IQ. Path diagrams are introduced in Section 1.2. Enclosed in ellipses are latent variables ξ and ω . The \mathbf{X} and \mathbf{Y} variables, enclosed in rectangles, are observed or indicator variables. The double-headed arrow between ξ and ω indicates a non-zero correlation. The single-headed arrows from the latent variables to the indicators indicate there are non-zero coefficients for the latent variables in the equations for the indicators. The lack of an arrow, or edge, between ξ and $\mathbf{Y}_{.1}$ means that any dependence between these variables can occur only through the other variables. The fact that the removal of ω from the diagram would result in the lack of a path between ξ and $\mathbf{Y}_{.1}$ means that ξ and $\mathbf{Y}_{.1}$ are conditionally independent given ω . The doubleheaded arrows between the \mathbf{X} variables means that they are not conditionally independent given ξ . Similarly the doubleheaded arrows between the \mathbf{Y} variables means that they are not conditionally independent given ω .	36
3.2	Example of a diagram by which the analyst might postulate inner and outer relations in specifying a model in the context of Herman Wold's original PLS algorithm. This is discussed in Section 3.8.2 on page 52.	57
3.3	A path diagram for a two-block PLS model. As noted in Section 3.8.5, when Wold's general PLS algorithm is applied in Mode A to the two-block case, the coefficients are identical to those computed by PLS-W2A. As noted on page 56, the direction of the arrows between ξ and ω has no effect on the values of the coefficients computed by PLS-W2A.	57
4.1	Path diagram of a paired latent correlation model. Paired latent correlation models are defined on page 85, Section 4.2.2.	85
4.2	Path diagram of a rank-one single latent model, discussed on page 86, Section 4.2.3.	86

- 4.3 The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ (the decreasing function) and $\Sigma_{\zeta\zeta}(\alpha)$ in the single latent parameterization of the matrix at line (4.22), page 94. The feasible values for α lie in the closed interval $\left[\sqrt{\frac{203}{90}}, \sqrt{7}\right] \approx [1.50, 2.65]$. These are the values of α for which the least eigenvalues of both $\Sigma_{\epsilon\epsilon}(\alpha)$ and $\Sigma_{\zeta\zeta}(\alpha)$ are nonnegative. 97
- 4.4 Feasible ρ and α for the paired-latent correlation parameterization of the rank-constraint distribution specified by (4.22). Feasible values are in the shaded region. The right boundary of the feasible set corresponds to the single latent model. The (curved) left boundary is the line $\rho\alpha = \alpha_{\min}$. The minimum feasible correlation is $\rho_{\min} \equiv \frac{\alpha_{\min}}{\alpha_{\max}}$ 98
- 4.5 The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ and $\Sigma_{\zeta\zeta}(\alpha)$ for the singular matrix at line (4.23), page 99, where there are two \mathbf{X} -variables and two \mathbf{Y} -variables. The decreasing function is the least eigenvalue of $\Sigma_{\epsilon\epsilon}(\alpha)$. Since $\sqrt{2}$ is the only point where both curves equal or exceed zero, this is the only feasible value for α 99
- 4.6 The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ (the nonincreasing function) and $\Sigma_{\zeta\zeta}(\alpha)$ for the matrix (4.24), page 100. 101
- 4.7 The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ (the decreasing function) and $\Sigma_{\zeta\zeta}(\alpha)$ for the matrix at line (4.25), page 102. As we would expect, since this matrix fails to be positive semidefinite there is no α , or scale for the \mathbf{X} -salience vector, by which the single latent model can parameterize it. This may be seen by the fact that there is no value of α for which both curves are greater than or equal to zero. 102

4.8	Feasible set for the SVD parameterization of the constraint model (4.22) on page 94. The SVD paired latent model is defined in Section 4.6.3. A bijection exists between this feasible set and the feasible set plotted in Figure 4.4 on page 98, representing the paired latent correlation parameterization of the same constraint model. Level sets of ρ lie on the curves $\psi = \frac{d^2}{\rho^2 \phi}$. Feasible values for ρ are in the closed interval $\left[\frac{\sqrt{290}}{30}, 1 \right]$. At $\rho = \frac{\sqrt{290}}{30} \approx 0.568$, the level set is the point $\left(7, \frac{225}{29} \right)$, that is, the upper right corner of the broken rectangle. A selection of level sets for ρ are plotted as solid curves. Feasible values for ϕ and ψ lie inside the broken rectangle and on or to the right of the $\rho = 1$ curve.	117
5.1	Path diagram representing a symmetric rank- r paired latent model. This model is defined in Section 5.1.2. In the parameterization guaranteed by Theorem 5.2.1, $\rho_1 = \dots = \rho_r = 1$ since $\xi_k \equiv \omega_k$ for all k	120
5.2	Path diagram representing a symmetric rank- r single latent model.	121
6.1	Path diagrams corresponding to two-block latent variable models. Under (I) the dashed edges are present; under (II) they are absent. Under (I) all models are covariance equivalent over \mathbf{X} and \mathbf{Y}	129

ACKNOWLEDGMENTS

Thanks are due to a long sequence of encouragers and helpers, above whom loom my parents. They are the first and the last.

Paul Sampson encouraged me to return to the graduate program when I had been away for six years. He, Anne Streissguth, Fred Bookstein, and Helen Barr welcomed me into the field of behavioral teratology. They introduced me to a study and a dataset that became the prime motivation for this dissertation. Paul provided financial support for fifteen months. He read my work incisively and diligently. Through the struggles to come, I was always able to rely on his kindness.

I met Thomas Richardson outside an empty, sealed elevator shaft which is a permanent fixture in Padelford Hall at the University of Washington. Walking past the shaft one day, I noticed a small slip of paper, which read, "This elevator is for the use of University of Washington faculty only." Thomas heard my laughter, came out of his office, and joined in.

I told Thomas about the behavioral teratology work to which Paul had introduced me, and soon he had pointed out a substantial unsolved theoretical problem embedded in that context. He proved to be a diligent and effective mentor. The metaphors of midwife and of a gardener carefully tending a plant both apply. Conversations with Thomas almost always digress into areas far removed from the business at hand. In fact they often start there, and only gradually work their way back to business. It is in part this quality which has made the conversations highly pleasurable. Thomas claims to be cynical, but his cynicism is of a high, ultimately constructive kind.

Werner Stuetzle, Brian Leroux, and Michael Perlman each provided valuable insights. Werner, with great good will, kept the Department running and contributed a vivid, practical point of view. I thank him for this. Brian Leroux kept surprising me with his calm and his optimism. Even more surprisingly, he consented without hesitation to join the read-

ing committee. The presence of Michael, with his knowledge of multivariate statistics and graphical models, was both reassuring and stimulating. These men were a great pleasure to work with.

David Ragozin and Asa Packer of the Mathematics Department contributed essential insights into linear algebra. I thank Vic Klee, K. Bruce Erickson, and Ron Irving, also in the Math Department. Along with Edwin Hewitt, they got me my NSF graduate fellowship. They encouraged me to return to the program and continued to talk with me after I did return.

I depended (as do all Stat students) on Kristin Sprague for her knowledge of how the University and the Department work and for her almost pastoral interest in how each of us students was doing.

I have depended on the kindness and forbearance of my fellow students and of many others in Padelford Hall. Cheryl Bissett, Kent Kalnasy, and Reuel Riley were anchors. I thank Connie Sugatan and Cheryl for moral support when I was distraught about the health of a family member. Judy McPhee showed me the house finches which nested every year in the feeder hanging outside her office window.

Thanks to Martha Tucker and Rebecca Wynkoop of the Mathematics Research Library. Their enthusiasm made library research a treat. Richard Fairfield, Michael O'Connell, and their colleagues kept the computer system working and endured my frequent emails to "help@stat."

On the first day of my last year in college, as I sat in a course in Real Analysis, a burly man with a trace of gray in his black hair turned to me and extended his hand. He was Robert Henry Tarr, Jr., machinist. He made an immense difference in my life that year. We labored side by side like mules plowing a field. I would beat my head against a problem for hours, while Bob researched it in the library, a tactic which had never occurred to me. Bob demonstrated repeatedly how useful this skill is. He also taught me out of his vast reservoir of people skills, and he gave me a course in friendship.

We sat together in Edwin Hewitt's class. It was Dr. Hewitt's last year before retirement.

He never missed a day, in spite of the fact that he fell sick during winter quarter. He always told us, "I'll be there if I'm not in jail." He went to bat with the NSF for me and told me I simply had to get an advanced degree. After his first stroke, when I encountered him walking on the UW campus with a cane, barely able to speak, he asked me what I was up to. I was dubious about finishing the Ph.D. I was more interested in my acting class. He stammered, "GET IT!"

Dr. Hewitt was a macho man. I hope and pray that he is looking down, if he can spare the time, from Valhalla, where departed heroes get everything they want.

DEDICATION

To M&D.

Chapter 1

PROBLEM DEFINITION, TERMINOLOGY, AND NOTATION

1.1 Introduction

Suppose we have matrices \mathbf{X} and \mathbf{Y} , respectively $N \times p$ and $N \times q$, where the columns correspond to variables and the rows to observations or units. We call these matrices **blocks**. Suppose further that we are primarily interested, not in the within-block covariances $\mathbf{X}^T \mathbf{X}$ and $\mathbf{Y}^T \mathbf{Y}$, but in the **cross-covariance** $\mathbf{X}^T \mathbf{Y}$. This is the **cross-covariance problem**.

An example from behavioral teratology. In a study of the relationship between fetal alcohol exposure and neurobehavioral deficits reported by Sampson et al. [SSBB89] and by Streissguth et al. [SBSB93a], \mathbf{X} has thirteen columns, each corresponding to a different measure of the mother's reported alcohol consumption during pregnancy. \mathbf{Y} has eleven columns, each corresponding to a different IQ subtest. The researchers are not primarily interested either in the relationships between the different measures of the mother's alcohol intake or in the relationships between the different IQ subtests. They are interested in the relationship between alcohol intake and IQ. Neither of these phenomena can be measured directly, and they may possess more than one dimension of interest.

Models for cross-covariance. The models which are the focus of the current work have the following properties in common:

- It is only

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \quad (1.1)$$

which is being modeled. Means are assumed to be zero, and higher moments are ignored.

- The models place no restriction on the within-block population covariances, Σ_{XX} and Σ_{YY} .

The parameters of a **constraint model** are Σ_{XX} , Σ_{XY} , and Σ_{YY} . In such a model the only constraint is a rank constraint on the population cross-covariance, Σ_{XY} . (Provided Σ is nonsingular, this is equivalent to placing a rank constraint on the off-diagonal block of the inverse covariance matrix. For a detailed proof see Proposition A.0.3 on page 151.) This constraint gives us immediately a way to define a sequence of nested models. Let R be the maximum rank of Σ_{XY} which is allowed by a model; then each value of R from 0 to $\text{rank}(\mathbf{X}^T \mathbf{Y})$ specifies a distinct model. The model with highest rank will fit Σ exactly. The covariances Σ which can be modeled by a constraint model are equivalent to those that can be modeled by reduced-rank regression (T. W. Anderson [And99], Reinsel and Velu [RV98]). This fact is reviewed in Appendix A.

Latent models for cross-covariance. Hidden or latent variables, together with a set of linear relationships between these and the data, may be used to model cross-covariance. In this context the data are called **observed variables** or **indicators**. Each row, \mathbf{X}_n and \mathbf{Y}_n , corresponds to a unit of observation. These are assumed to be independent, identically distributed (iid) draws from a population specified by the model. A model specifies s explanatory latent variables for the \mathbf{X} -block and t explanatory latent variables for the \mathbf{Y} -block. The scores on the explanatory latent variables are denoted by the $N \times s$ matrix Ξ for the \mathbf{X} block and the $N \times t$ matrix Ω for the \mathbf{Y} block. Thus the explanatory latent variable scores for \mathbf{X}_n are Ξ_n and the explanatory latent variable scores for \mathbf{Y}_n are Ω_n . The matrices of noise, \mathbf{E} for \mathbf{X} and \mathbf{Z} for \mathbf{Y} , are respectively $N \times p$ and $N \times q$.

It will be convenient to use lowercase letters to denote the values for individual obser-

vations, seen as column vectors. The lowercase notation is:

$$\begin{aligned} \mathbf{x}_n &= \mathbf{X}_{n\cdot}^T, \\ \mathbf{y}_n &= \mathbf{Y}_{n\cdot}^T, \\ \boldsymbol{\xi}_n &= \boldsymbol{\Xi}_{n\cdot}^T, \\ \boldsymbol{\omega}_n &= \boldsymbol{\Omega}_{n\cdot}^T, \\ \boldsymbol{\epsilon}_n &= \mathbf{E}_{n\cdot}^T, \\ \boldsymbol{\zeta}_n &= \mathbf{Z}_{n\cdot}^T. \end{aligned}$$

Although $\boldsymbol{\epsilon}_n$ and $\boldsymbol{\zeta}_n$ are latent, it will be convenient to call them simply “noise,” and to reserve the term “latent variables” for the explanatory latent variables $\boldsymbol{\xi}_n$ and $\boldsymbol{\omega}_n$.

The covariance of the latent variables is

$$\text{Cov} \begin{bmatrix} \boldsymbol{\xi}_n \\ \boldsymbol{\omega}_n \end{bmatrix} = \text{Cov} \begin{bmatrix} \boldsymbol{\Xi}_{n\cdot}^T \\ \boldsymbol{\Omega}_{n\cdot}^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi} & \mathbf{G} \\ \mathbf{G} & \boldsymbol{\Psi} \end{bmatrix}, \quad (1.2)$$

and the dispersion matrices for the noise are respectively $\boldsymbol{\Sigma}_{\epsilon\epsilon}$ and $\boldsymbol{\Sigma}_{\zeta\zeta}$.

Let \mathbf{A} be $p \times s$ and \mathbf{B} $q \times t$. Specify the data by

$$\begin{aligned} \mathbf{x}_n &= \mathbf{A}\boldsymbol{\xi}_n + \boldsymbol{\epsilon}_n, \\ \mathbf{y}_n &= \mathbf{B}\boldsymbol{\omega}_n + \boldsymbol{\zeta}_n, \end{aligned} \quad (1.3)$$

equivalently

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\Xi}\mathbf{A}^T + \mathbf{E}, \\ \mathbf{Y} &= \boldsymbol{\Omega}\mathbf{B}^T + \mathbf{Z}. \end{aligned}$$

The family of latent models defined thus far is very general. The models considered in the current work will be subject to the following two additional constraints:

- The number of latent variables for the \mathbf{X} block is the same as the number of latent variables for the \mathbf{Y} block. In this case we speak of pairs of latent variables, each pair containing one latent variable for the \mathbf{X} block and one for the \mathbf{Y} block. We let R denote the number of pairs of latent variables ($s = t = R$), so that $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and \mathbf{G} are all $R \times R$.

- The noise for any block has zero covariance with the noise for the other block, and with the latent variables for either block. Under this assumption we may write

$$\text{Cov} \begin{bmatrix} \xi_n \\ \omega_n \\ \epsilon_n \\ \zeta_n \end{bmatrix} = \text{Cov} \begin{bmatrix} \Xi_n^T \\ \Omega_n^T \\ \mathbf{E}_n^T \\ \mathbf{Z}_n^T \end{bmatrix} = \begin{bmatrix} \Phi & \mathbf{G} & 0 & 0 \\ \mathbf{G} & \Psi & 0 & 0 \\ 0 & 0 & \Sigma_{\epsilon\epsilon} & 0 \\ 0 & 0 & 0 & \Sigma_{\zeta\zeta} \end{bmatrix} \quad (1.4)$$

These constraints imply

$$\begin{aligned} \Sigma_{XX} &= \mathbf{A}\Phi\mathbf{A}^T + \Sigma_{\epsilon\epsilon} , \\ \Sigma_{YY} &= \mathbf{B}\Psi\mathbf{B}^T + \Sigma_{\zeta\zeta} , \\ \Sigma_{XY} &= \mathbf{A}\mathbf{G}\mathbf{B}^T . \end{aligned} \quad (1.5)$$

The resulting model is called a **paired latent model**. It is by design that Σ_{XY} has a simpler form than the other blocks. This is the block for which we seek a parsimonious, easily-interpreted model.

It will often be convenient to assume that $R \leq p$ and $R \leq q$, and that \mathbf{A} , \mathbf{B} , and \mathbf{G} are all of full rank. Under such an assumption we call R the **rank** of the model.

Cross-diagonal models. If a paired latent model for cross-covariance is constrained to have

$$\begin{aligned} \mathbf{G} &= \text{diag}(g_1, \dots, g_R) , \\ g_1 &\geq g_2 \geq \dots \geq g_R \geq 0 , \end{aligned}$$

we say that it is a **cross-diagonal latent model**. The parameters of this model are interpretable in a straightforward way. The diagonality of \mathbf{G} implies

$$\begin{aligned} \text{Cov}((\mathbf{x}_n)_i, (\omega_n)_r) &\equiv \text{Cov}(\mathbf{X}_{ni}, \Omega_{nr}) = g_r \mathbf{A}_{ir} \\ \text{and} \\ \text{Cov}((\mathbf{y}_n)_j, (\xi_n)_r) &\equiv \text{Cov}(\mathbf{Y}_{nj}, \Xi_{nr}) = g_r \mathbf{B}_{jr} . \end{aligned}$$

In this case, the values \mathbf{A}_{ir} and \mathbf{B}_{jr} are called **saliences**.

SVD latent models. The parameters of the cross-diagonal latent models are not identifiable. We may constrain this class further by requiring

$$\begin{aligned} \mathbf{A}^T \mathbf{A} &= \mathbf{I}_p \\ \mathbf{B}^T \mathbf{B} &= \mathbf{I}_q . \end{aligned}$$

Under these conditions the parameters \mathbf{A} , \mathbf{B} , and \mathbf{G} constitute a singular value decomposition of Σ_{XY} . Since this decomposition is unique up to sign, the constraints guarantee that \mathbf{A} , \mathbf{B} , and \mathbf{G} are identifiable up to sign. Cross-diagonal latent models which satisfy this constraint are called **svd latent models**.

In keeping with familiar svd notation (for instance, the function `svd()` in S-PLUS [Mat96]), the cross-diagonal parameters of the svd model will often be called \mathbf{U} , \mathbf{V} , and \mathbf{D} in the current work.

Behavioral teratology, revisited. Recall the example introduced on page 1 in Section 1.1. A class of methods exists for estimating saliences and for computing composite scores on hypothesized latent variables which uses no probability theory. These methods, called Partial Least Squares (PLS), are discussed in Section 2.3, Chapter 3, and in Wegelin [Weg00]. Using a PLS method, the researchers estimated saliences for a latent model with rank $R = 1$. In addition, they computed estimated composite scores $\hat{\Xi}_{n,1}$ and $\hat{\Omega}_{n,1}$ for each subject in the study. Since the model was of rank one, $\hat{\Xi}$, $\hat{\Omega}$, $\hat{\mathbf{A}}$, and $\hat{\mathbf{B}}$ each consisted of a single column.

The composite score for the \mathbf{Y} -block, $\hat{\Omega}_{\cdot,1}$, summarized the 11 IQ subtests, and the composite score for the \mathbf{X} -block, $\hat{\Xi}_{\cdot,1}$, summarized the mother's drinking during pregnancy. These vectors of scores were optimal in the sense that

$$\left| \text{Cov} \left(\hat{\Xi}_{\cdot,1}, \hat{\Omega}_{\cdot,1} \right) \right| = \max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1} \left| \text{Cov} (\mathbf{X}\mathbf{a}, \mathbf{Y}\mathbf{b}) \right| .$$

Another optimality criterion is that of canonical correlation analysis (CCA). The two optimality criteria are contrasted in Section 3.11, and the researchers' reasons for using PLS rather than CCA are given.

The i th estimated \mathbf{X} -salience $\hat{\mathbf{A}}_{i,1}$ was proportional to the covariance of the i th measure of maternal drinking and the explanatory latent variable for IQ. Thus it measured the

degree of association of the i th measure of maternal drinking with the IQ explanatory latent variable, relative to the other measures of maternal drinking.

Similarly the j th estimated Y-salience $\hat{B}_{j,1}$, proportional to the covariance of the j th IQ subtest and the explanatory latent variable for maternal drinking, measured the degree of association between the j th IQ subtest and this latent variable, relative to the other IQ tests.

Asymmetric models. The latent model specified thus far is symmetric in the following sense.

1. y_n is not explicitly a function of x_n , nor vice versa; rather
2. The indicators are functions of the latents: x_n of ξ_n , y_n of ω_n .

In many applications it makes more sense to think of y_n as a function of x_n , because of what we may believe about the process which caused the phenomenon we are attempting to measure. For instance, in the example from behavioral teratology mentioned on pages 1 and 5, we might believe that IQ is a function of alcohol exposure. This suggests that we might specify the data by the following asymmetric latent model:

$$\begin{aligned}
 x_n &= \mu_x + \epsilon_n, \\
 \xi_n &= A^T x_n + (\eta_\xi)_n, \\
 \omega_n &= G \xi_n + (\eta_\omega)_n, \text{ and} \\
 y_n &= B \omega_n + \zeta_n,
 \end{aligned}$$

where

μ_x is a p -dimensional mean vector for the X-block,
 ϵ_n is p -dimensional random vector of errors for the X-block,
 $(\eta_\xi)_n$ is an R -dimensional random vector of errors for ξ_n ,
 A is a $p \times R$ constant matrix linking ξ_n to x_n ,
 $(\eta_\omega)_n$ is an R -dimensional random vector of errors for ω_n ,
 G is an $R \times R$ constant symmetric matrix linking ω_n to ξ_n ,
 ζ_n is a q -dimensional random vector of errors for the Y-block,
 B is a $q \times R$ constant matrix linking y_n to ω_n , and
 $\epsilon_n, (\eta_\xi)_n, (\eta_\omega)_n$, and ζ_n are mutually independent.

In this case

$$\begin{aligned}
 y_n &= B\omega_n + \zeta_n \\
 &= B(G\xi_n + (\eta_\omega)_n) + \zeta_n \\
 &= B\left(G\left[A^T x_n + (\eta_\xi)_n\right] + (\eta_\omega)_n\right) + \zeta_n \\
 &= B\left(G\left[A^T \mu_X + A^T \epsilon_n + (\eta_\xi)_n\right] + (\eta_\omega)_n\right) + \zeta_n \\
 &= BGA^T \mu_X + BGA^T \epsilon_n + BG(\eta_\xi)_n + B(\eta_\omega)_n + \zeta_n.
 \end{aligned}$$

Let

$$\Sigma_{\epsilon\epsilon}, \Sigma_{\eta_\xi\eta_\xi}, \Sigma_{\eta_\omega\eta_\omega}, \Sigma_{\zeta\zeta}$$

be the variances of

$$\epsilon_n, (\eta_\xi)_n, (\eta_\omega)_n, \zeta_n$$

respectively. Then the covariance of the data is given by

$$\begin{aligned}
 \Sigma_{XX} &= \Sigma_{\epsilon\epsilon}, \\
 \Sigma_{YY} &= BGA^T \Sigma_{\epsilon\epsilon} A G B^T + B G \Sigma_{\eta_\xi\eta_\xi} G B^T + B \Sigma_{\eta_\omega\eta_\omega} B^T + \Sigma_{\zeta\zeta}, \\
 \Sigma_{XY} &= \Sigma_{\epsilon\epsilon} A G B^T.
 \end{aligned} \tag{1.6}$$

Thus in the asymmetric model the error covariance for the X block is part both of the cross-covariance and of the covariance of the Ys. Nevertheless we shall see in Chapter 6

that the symmetric and asymmetric models are covariance equivalent over the indicators. That is, for each set of parameters of the symmetric model there is a set of parameters of the asymmetric model which induces the same covariance over \mathbf{X} and \mathbf{Y} , and vice versa. Consequently $\Sigma_{\epsilon\epsilon}$ must be interpreted differently in the two models.

Normal distribution. A multivariate normal distribution will be assumed in part of the current work. Latent models for cross-covariance are defined first without the specification of any density, however, and some work will be done in the more general setting.

Factor models. The paired latent model is not a traditional factor model. This is because factor models are usually assumed to have diagonal within-block error covariance, or “uncorrelated errors of measurement.” Recall that one of the defining characteristics of the paired latent model for cross-covariance is *unconstrained* within-block error covariance.

Traditional factor analysis is also called exploratory factor analysis (EFA). In confirmatory factor analysis (CFA), the requirement that the within-block error variance be diagonal is relaxed. Thus the paired latent model for cross-covariance does fit this broader definition of a factor model. As long as we keep this broader definition in mind, we can call the saliences of the paired latent model by the more familiar term **factor loadings**.

Bollen discusses both exploratory and confirmatory factor analysis ([Bol89], Chapter seven). Mardia, Kent and Bibby [MKB79] and Chatfield and Collins [CC80] each devote a chapter to factor analysis. The reader is also referred to Harman [Har76].

Notational convention. Subscripts “ n ” indicating that an observation belongs to a sample of size N have been used up to this point. In the sequel the “ n ” subscript may be dropped when this can be done without confusion. Thus, for instance, rather than $\mathbf{X}_{n,i}$, indicating the n th realization of the i th indicator of the \mathbf{X} -block, the notation \mathbf{X}_i may be used. This represents the i th indicator as a random variable.

Zero partial correlation. The notion of a path diagram is introduced in Section 1.2, and that of m -separation in Section 1.3. On page 13 a result is stated which pertains to zero partial correlations. The following property of the paired latent model for cross-covariance

follows from that result. For any i and any j ,

$$\left. \begin{aligned} \text{Cor}(X_i, Y_j | \xi_1, \dots, \xi_R) &= \text{Cor}(X_i, Y_j | \omega_1, \dots, \omega_R) \\ &= \text{Cor}(X_i, Y_j | \xi_1, \dots, \xi_R, \omega_1, \dots, \omega_R) \\ &= 0. \end{aligned} \right\} \quad (1.7)$$

In particular, if the data have a joint multivariate normal distribution, X_i and Y_j are conditionally independent given either $\{\xi_1, \dots, \xi_R\}$, $\{\omega_1, \dots, \omega_R\}$, or the union of these two sets.

1.2 Path diagrams

Path diagrams are a convenient way to display a set of hypothesized relationships between observed and latent variables. An example of a path diagram may be seen in Figure 1.1 on page 10. Path diagrams exist within a larger class of objects called **graphs**, in which variables are represented by vertices, and edges are drawn between the vertices. Many different conventions exist, each defining a different kind of graph. The conventions used in the current work will now be stated.

Vertices (equivalently, nodes) in a path diagram represent random variables, either observed or hidden (latent). When a distinction between observed and hidden variables needs to be made, observed variables are enclosed in rectangles and hidden variables in ellipses.

Edges in the current work will be either directed (\leftarrow or \rightarrow) or bidirected (\leftrightarrow). These may also be called, respectively, singleheaded and doubleheaded arrows. Two nodes will share at most one edge.

Path diagrams are translated into the familiar notation of covariance matrices and linear relationships, and thus into a set of parameters, in the following manner.

- Any node at which one or more directed edges points is a linear function of the nodes on the tail ends of the directed edges, plus error. Thus to each edge corresponds a linear coefficient.
- If two nodes share a bidirected edge, the covariance between their errors is not constrained to equal zero. Thus an explicit expression for the error covariance is included

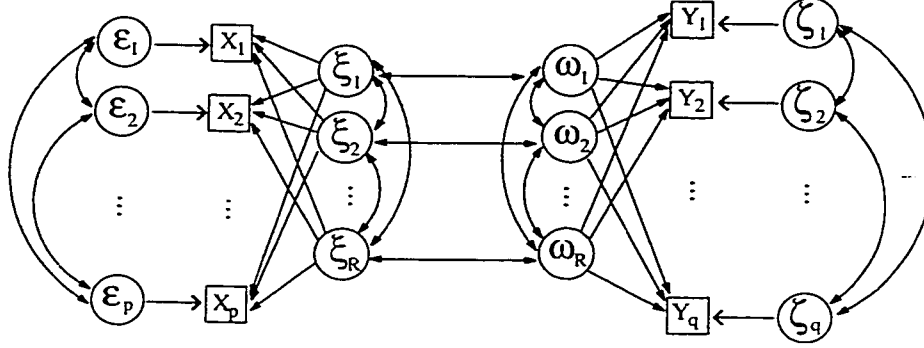


Figure 1.1: Path diagram representing the symmetric cross-diagonal parameterization of a rank R latent model. Error terms are displayed in this figure. The same model is displayed without error terms in Figure 1.2 on page 11. Path diagrams are introduced in Section 1.2. The cross-diagonal parameterization is introduced in Section 1.1.

in the specification of the model.

- Errors may be displayed in path diagrams as latent variables, or may be omitted. Each variable other than an error is assumed to have an error term, whether the error term is displayed or not.

Path diagrams for the symmetric cross-diagonal latent model may be found in Figures 1.1 on page 10 (with error terms displayed) and 1.2 on page 11 (error terms not displayed). Both path diagrams make the following specifications.

- $X_{n,i}$ is a linear function of the following $R + 1$ latent variables,

$$\Xi_{n,i}, \dots, \Xi_{n,R}, \text{ and } \epsilon_{n,i},$$

and $Y_{n,j}$ is a linear function of

$$\Omega_{n,j}, \dots, \Omega_{n,R}, \text{ and } \zeta_{n,j}.$$

This relationship is specified in Equation 1.3 on page 3.

- For any n and any r , the pair of latent variables $(\Xi_{n,r}, \Omega_{n,r})$ has unrestricted covariance, but for $r \neq s$

$$\text{Cov}(\Xi_{n,r}, \Omega_{n,s}) = 0.$$

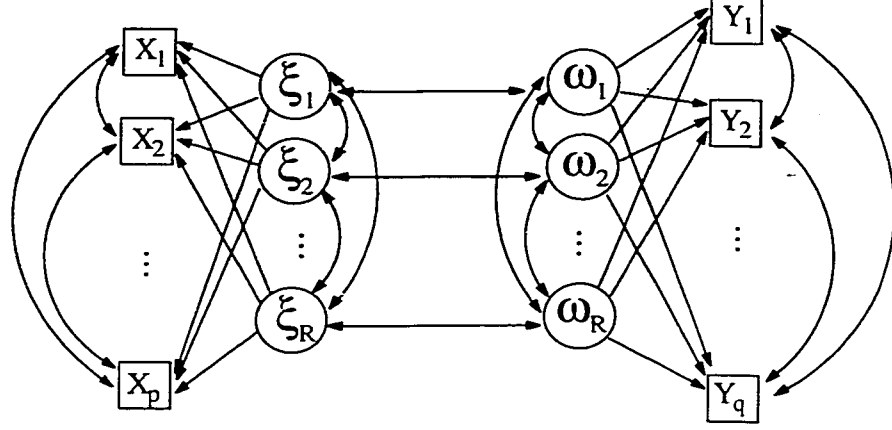


Figure 1.2: Path diagram representing the symmetric cross-diagonal parameterization of a rank R latent model. This represents exactly the same model as is represented by Figure 1.1. The only difference between the path diagrams is that error terms are not displayed in this one. Error terms are always understood, however. Path diagrams are introduced in Section 1.2. The cross-diagonal parameterization is introduced in Section 1.1.

This is equivalent to constraining \mathbf{G} to be diagonal in Section 1.1, but placing no constraints on the diagonal values other than that the entire covariance matrix of the latent variables, $\begin{bmatrix} \Phi & \mathbf{G} \\ \mathbf{G} & \Psi \end{bmatrix}$, be positive semidefinite.

- The within-block covariance of explanatory latent variables is unrestricted. That is, for any n , r , and s , $\text{Cov}(\Xi_{n,r}, \Xi_{n,s})$ is unrestricted, and $\text{Cov}(\Omega_{n,r}, \Omega_{n,s})$ is unrestricted. This is equivalent to placing no constraints on Φ and Ψ other than that they be positive semidefinite, and again that the entire covariance matrix of the latent variables be positive semidefinite.
- The within-block noise is unrestricted. That is, for any n , i , and i' , $\text{Cov}(\mathbf{E}_{n,i}, \mathbf{E}_{n,i'})$ is unrestricted, and for any n , j , and j' , $\text{Cov}(\mathbf{Z}_{n,j}, \mathbf{Z}_{n,j'})$ is unrestricted. This is equivalent to placing no constraint on $\Sigma_{\epsilon\epsilon}$ and $\Sigma_{\zeta\zeta}$ other than that they be positive semidefinite.

1.3 *M-separation*

Zero-partial-correlation relationships $\text{Cor}(X, Y|Z) = 0$, where X and Y are variables and Z is a set of variables, will be of interest in the current work. It may require extensive calculation to determine, however, from a variance-covariance matrix Σ , whether a given zero-partial-correlation relationship obtains. Fortunately theorems are available, in the context of path diagrams, which often make a question regarding zero partial correlation easy to answer. Theorems are also available, in the same context, which will enable us to prove that certain models, represented by different path diagrams, induce the same set of covariance matrices.

In the current section we consider a fixed path diagram \mathcal{G} , and a covariance matrix $\Sigma[\mathcal{G}]$ over the variables represented by the vertices of \mathcal{G} . Since different parameter values can be assigned to the model represented by a path diagram, potentially many different $\Sigma[\mathcal{G}]$ s may be associated with a fixed \mathcal{G} . Note that a zero-partial-correlation relationship is a property of $\Sigma[\mathcal{G}]$.

We now introduce some definitions, necessary to make use of the theorems mentioned above. A sequence of vertices joined by edges, in which no vertex is repeated, is called a **path**. A path of the form $\alpha \rightarrow \dots \rightarrow \beta$, on which every edge is of the form \rightarrow , with all the arrowheads pointing toward β , is a **directed path from α to β** .

A directed cycle is a directed path from α to β , together with a directed edge from β to α :

$$\alpha \rightarrow \dots \rightarrow \beta \rightarrow \alpha$$

In the current work all path diagrams considered will be free of directed cycles.

A vertex α is said to be an **ancestor** of a vertex β either if there is a directed path from α to β , or if $\alpha = \beta$.

A non-endpoint vertex ζ is a **collider on a path** if the edges before and after ζ have

arrowheads at ζ . That is, ζ looks like one of

$$\begin{aligned} &\rightarrow \zeta \leftarrow, \\ &\leftrightarrow \zeta \leftrightarrow, \\ &\leftrightarrow \zeta \leftarrow, \text{ or} \\ &\rightarrow \zeta \leftrightarrow. \end{aligned}$$

Otherwise it is a **non-collider**.

Let Z be a set of vertices in \mathcal{G} . If α is an ancestor of any vertex in Z , then α is said to be an ancestor of Z . A path between vertices α and β is **m-connecting in \mathcal{G} given Z** if

1. Every non-collider on the path is not in Z , and
2. Every collider on the path is an ancestor of Z .

If there is no m-connecting path in \mathcal{G} between α and β given Z , the vertices are **m-separated in \mathcal{G} given Z** .

We now are able to state the following fact. It is proved as Theorem 1 in Spirtes et al. [SRM⁺ar].¹ Consider a path diagram \mathcal{G} , and any covariance matrix $\Sigma[\mathcal{G}]$ over the variables represented by the vertices of \mathcal{G} , induced by any set of parameters for the model represented by \mathcal{G} . When vertices α and β are m-separated in \mathcal{G} given Z , we necessarily have $\text{Cor}(\alpha, \beta | Z) = 0$. The partial correlation result stated at 1.7 on page 9 follows from this fact.

For a more extensive treatment of graphical models with bidirected edges, the reader is referred to Richardson and Spirtes [RS00] and to Spirtes et al. [SRM⁺ar].

1.4 *Classes of latent models for cross-covariance*

In Section 1.1 and in Figure 1.1 the symmetric rank R cross-diagonal parameterization was presented. In this section three more classes of symmetric rank R latent models for cross-covariance will be described and diagrammed. The four models differ only in the constraints

¹In the reference cited, the term **d-separation** is used rather than m-separation. In the path diagrams considered in the current work, m-separation and d-separation are equivalent.

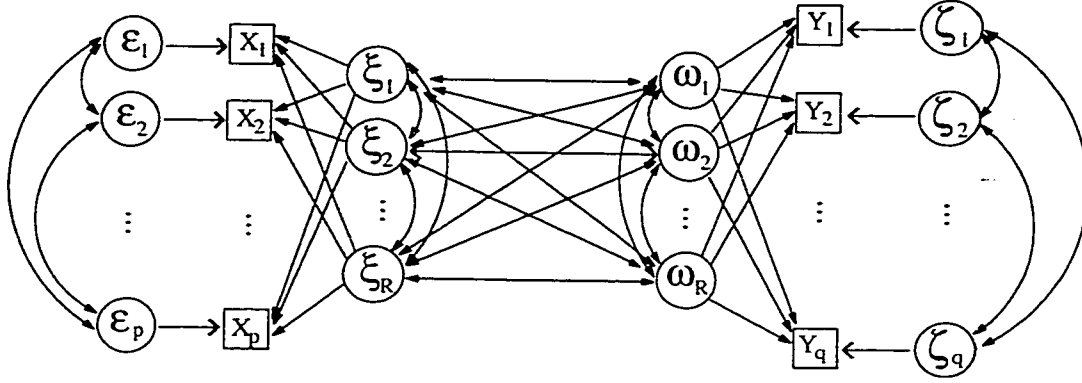


Figure 1.3: Path diagram representing an unrestricted rank R latent model, as described in Section 1.4.

they place on the covariance matrix of the latent variables,

$$\begin{bmatrix} \Phi & G \\ G & \Psi \end{bmatrix} \quad (1.8)$$

The models are as follows.

- A The unrestricted rank R model, Figure 1.3 on page 14, places no constraint on Φ , Ψ , or G other than that Φ and Ψ must be positive semidefinite, and that (1.8) must be positive semidefinite.
- B The cross-diagonal rank R model, Figure 1.1 on page 10, adds the constraint that G must be diagonal.
- C The within-block diagonal rank R model, Figure 1.4 on page 15, leaves G as in the unrestricted model, and constrains Φ and Ψ to be diagonal.
- D The double-diagonal rank R model, Figure 1.5 on page 1.5, combines the constraints of the cross-diagonal and within-block-diagonal models, as its name suggests.

By definition we have $B \subset A$, $C \subset A$, $D = B \cap C$.

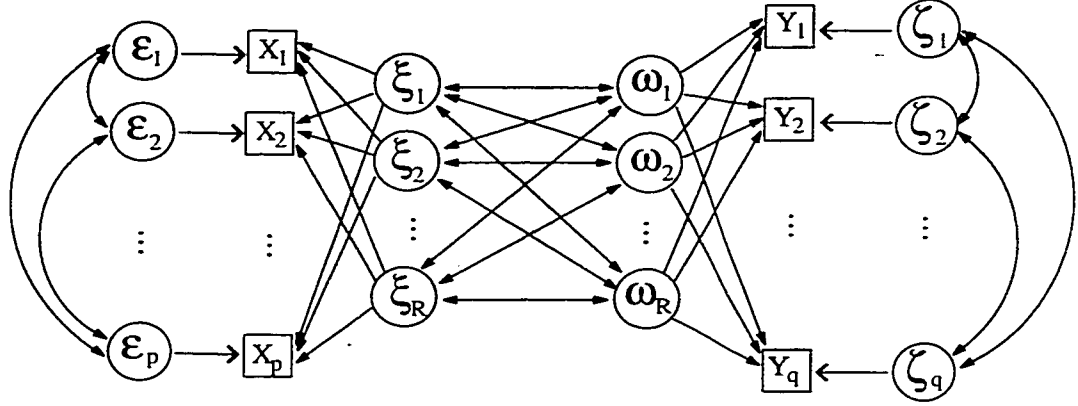


Figure 1.4: Path diagram representing a within-block-diagonal rank R latent model, as described in Section 1.4.

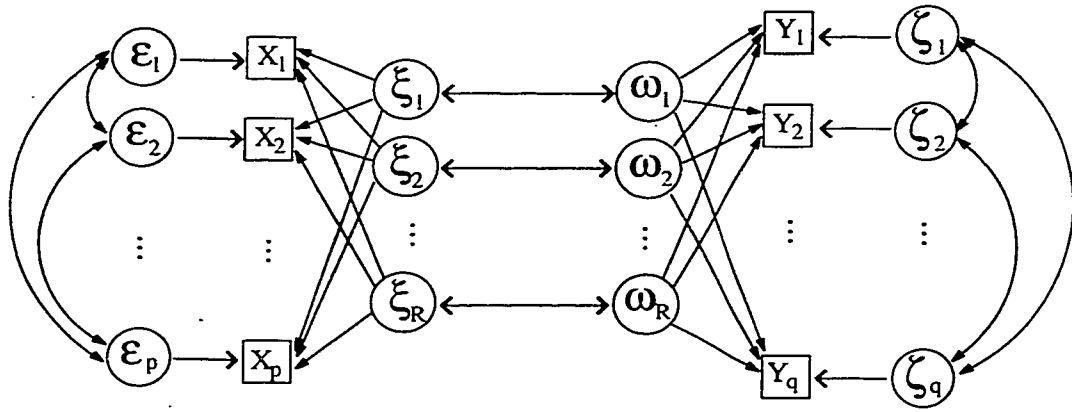


Figure 1.5: Path diagram representing a double-diagonal rank R latent model, as described in Section 1.4.

Chapter 2

CURRENT FRAMEWORKS FOR CROSS-COVARIANCE PROBLEMS

In this chapter, three frameworks are discussed which are applicable to two-block problems. These are Canonical correlation analysis (CCA), Structural equation models (SEMs), and Partial Least Squares or Projection to Latent Structures (PLS). PLS has been used successfully on cross-covariance problems. CCA and SEMs may superficially appear to be suited to cross-covariance problems. The reasons why they are not so suited are summarized.

2.1 Canonical correlation analysis (CCA)

CCA is a well-known method for the analysis of two-block data. Harold Hotelling published the seminal paper in 1936 [Hot36]. In CCA, a sequence of up to $\min(p, q)$ paired N -vectors (ξ_r, ω_r) are computed. These vectors are linear combinations of the data, \mathbf{X} and \mathbf{Y} respectively, and have been viewed by some as scores on latent variables. Let \mathbf{a}_r and \mathbf{b}_r be the coefficients which define the linear combinations. They are computed as follows. Let $\mathbf{X}^{(1)} \leftarrow \mathbf{X}$, $\mathbf{Y}^{(1)} \leftarrow \mathbf{Y}$. Then

$$(\mathbf{a}_1, \mathbf{b}_1) \leftarrow \operatorname{argmax}_{\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^q} \operatorname{Cor}(\mathbf{X}^{(1)}\mathbf{a}, \mathbf{Y}^{(1)}\mathbf{b}) .$$

When the r th pair of coefficients are computed, the following vectors, satisfying the criterion of maximal correlation, are also computed:

$$\begin{aligned} \xi_r &\leftarrow \mathbf{X}^{(r)}\mathbf{a}_r , \\ \omega_r &\leftarrow \mathbf{Y}^{(r)}\mathbf{b}_r . \end{aligned}$$

Rank-one projections onto these vectors are computed:

$$\begin{aligned} \hat{\mathbf{X}}^{(r)}(\xi_r) &\leftarrow \xi_r (\xi_r^T \xi_r)^{-1} \xi_r^T \mathbf{X}^{(r)} , \\ \hat{\mathbf{Y}}^{(r)}(\omega_r) &\leftarrow \omega_r (\omega_r^T \omega_r)^{-1} \omega_r^T \mathbf{Y}^{(r)} , \end{aligned}$$

and residual matrices are computed by subtraction of the rank-one projections:

$$\begin{aligned}\mathbf{X}^{(r+1)} &\leftarrow \mathbf{X}^{(r)} - \widehat{\mathbf{X}}^{(r)}(\xi_r) , \\ \mathbf{Y}^{(r+1)} &\leftarrow \mathbf{Y}^{(r)} - \widehat{\mathbf{Y}}^{(r)}(\omega_r) .\end{aligned}$$

In general, \mathbf{a}_r and \mathbf{b}_r are chosen to maximize correlation as follows:

$$(\mathbf{a}_r, \mathbf{b}_r) \leftarrow \operatorname{argmax}_{\mathbf{a} \in \mathbb{R}^p, \mathbf{b} \in \mathbb{R}^q} \operatorname{Cor} \left(\mathbf{X}^{(r)} \mathbf{a}, \mathbf{Y}^{(r)} \mathbf{b} \right) .$$

Although ξ_r and ω_r are sometimes viewed as scores on latent variables, the coefficients \mathbf{a}_r and \mathbf{b}_r which link them to the data possess no interpretation in terms of the parameters of the paired latent model introduced in Section 1.1. In particular, these coefficients are not consistent for the saliences of the paired latent model. To see this, consider the case where there is just one \mathbf{Y} variable ($q = 1$), and recall that in this case CCA is equivalent to multiple regression of \mathbf{Y} on \mathbf{X} . Let \mathbf{u} signify the vector of saliences for the \mathbf{X} -block under the paired latent model. Under this model, we have the following expression for the $p \times 1$ cross-covariance:

$$\Sigma_{XY} = d\mathbf{u} ,$$

so that the salience vector \mathbf{u} is proportional to the cross-covariance. The CCA coefficients, on the other hand, in this case are

$$\mathbf{a} = \Sigma_{XX}^{-1} \Sigma_{XY}$$

by the familiar regression formula. Thus the canonical correlation coefficients are consistent for the paired latent model saliences only when the within-block dispersion matrix is a scalar multiple of the identity.

CCA is discussed in Chapter 3 and in Wegelin [Weg00], in the context of PLS, the method which will be introduced in Section 2.3.

2.2 Structural Equation Models

The latent models for cross-covariance introduced in Section 1.1 belong to a large class called Structural Equation Models (SEMs). SEMs can have any number of latent variables and any

set of linear relationships between the latent variables. The paired latent model, introduced in Section 1.1, is therefore a small subset of these models. Even if we consider only those SEMs with two latent variables and one block of indicators for each latent variable, however, the latent model of the current work forms a small subset of this class. This fact is due to the constraints which define this model, specified in Section 1.1, page 3. We noted on page 8, Section 1.1 that

- After adjusting for (or conditioning on) the explanatory latent variables, the partial correlation of the X block with the Y block is zero. Under the assumption of joint multivariate normality, this means that indicators in different blocks are conditionally independent given the latents.
- On the other hand, indicators in the same block are correlated, even after adjustment for the explanatory latent variables.

Structural equation models typically do not have this property. For instance, Bollen [Bol89] contains more than 60 path diagrams, several of which are of the two-block, two-latent-variable variety. None of these has the covariance structure of the current model class.

Although the class of SEMs is much bigger than the current model class, the kind of SEM encountered most frequently in the literature is in an important sense more restricted than this class, because constraints are placed on the within-block covariance. This can be seen by an examination of the models in Bollen's book. In many of these, indicators within a block are assumed to have independent errors. When correlation is allowed within a block, typically it is of a kind such that

- some indicators of a block are constrained to be independent of each other, in spite of the correlation allowed between other indicators in the same block, and/or
- some direct dependency exists between the blocks—that is, dependency which cannot be removed by conditioning on the latent variables.

For the kind of problem which motivates the current work, i.e., for a cross-covariance problem, it is unrealistic to place constraints on the within-block covariance. Take for

instance the example from behavioral teratology. There is no reason to believe that a set of latent variables which account for the covariance of the 13 measures of maternal drinking with the 11 IQ subtests will also account for the covariance of the 13 measures of maternal drinking with each other.

Software packages exist for performing maximum-likelihood estimation in general structural equation models. Three of these are LISREL, EQS, and AMOS. We saw in Section 1.1 that cross-covariance problems are characterized by a focus on $\mathbf{X}^T \mathbf{Y}$ and a lack of interest in the within-block covariance. The latent model class of interest in the current work is specified with this in mind. No constraint was placed on within-block covariance, and the model was so constructed that the \mathbf{X} and \mathbf{Y} blocks are conditionally independent of each other given the latent variables. A method for data analysis in the context of this class should exploit this special covariance structure. LISREL, EQS, and AMOS, being general tools, do not.

LISREL, EQS, and AMOS work in the following manner. The user specifies a set of linear relations between the observed variables (the columns of our \mathbf{X} and \mathbf{Y}) and a set of hypothesized latent variables. With EQS the user also specifies a starting point for iteration. The software generates a likelihood, and then uses a standard optimization routine to seek a local maximum. EQS uses “a modified Gauss-Newton method” (Bentler [Ben89] page 228).

LISREL and EQS are not guaranteed to converge. Convergence depends on the starting value. The experience of the author has been that it is extremely easy to furnish a dataset and a starting value for which EQS will not converge. Attempts to apply EQS to two datasets suited to a two-block PLS analysis, one consisting of simulated data and one derived from standard multivariate datasets from the S-PLUS libraries, both led to convergence difficulties.

In their convergence difficulties, LISREL and EQS differ radically from PLS, a method which will be introduced in Section 2.3. These difficulties are not surprising, in view of the fact that these software packages were not designed specifically for the two-block case, but rather for a much more general class of problems.

For further discussion of SEMs the reader is referred to Steiger [Ste01].

2.3 Partial least squares (PLS)

Partial Least Squares (PLS) is a class of methods for analyzing data which naturally can be divided into multiple blocks of variables observed on the same units. PLS is not based on a probability model. In the words of its inventor, Herman Wold, PLS is “distribution-free” [Wol85]. Given any two-block dataset as described in Section 1.1, PLS methods exist which can express the dataset in terms of pairs of latent variables, linear coefficients linking the data to the latent variables, and errors. Coefficients computed by PLS are numerically stable in the presence of partial or complete collinearity. In the example from behavioral teratology on page 5, Section 1.1, a PLS method was used to estimate latent scores and saliences.

PLS methods are discussed in detail in Chapter 3. The PLS methods of greatest interest in the current context will be referred to as PLS-W2A and PLS-SVD. Precise statements of the algorithms are found in Chapter 3. The method used in the example from behavioral teratology was PLS-SVD.

An analogy exists between paired latent models for cross-covariance on the one hand and two-block PLS methods such as PLS-W2A and PLS-SVD on the other. Both model cross-covariance by latent variables. The equations by which PLS-SVD and PLS-W2A express the data are exactly those which appear in the statement of a latent-variable model, Equations (1.3) on page 3. PLS-SVD is consistent for the cross-covariance parameters, \mathbf{A} , \mathbf{B} , and \mathbf{G} , of the “svd” variant of the paired latent model. (Recall that this variant is defined by adding, to the definition of the paired latent model, the additional constraints that \mathbf{G} is a diagonal matrix and that $\mathbf{A}^T \mathbf{A} = \mathbf{I}$ and $\mathbf{B}^T \mathbf{B} = \mathbf{I}$.) This is a consequence of the continuity of the singular value decomposition.

PLS is not an estimator of the entire parameter set in the paired latent model. PLS is an empirical method, of course, not a statistical model, and thus the analogy is not complete. Some methods which originated apart from a statistical model, however, turn out to be method-of-moments estimators for statistical models. For instance, the sample mean and variance are moment estimators for the parameters of the multivariate normal

distribution. PLS does not stand in an analogous relationship to the paired latent model. This is because the coefficients computed by PLS pertain only to $\text{Cov}(\mathbf{X}, \mathbf{Y})$, not to $\text{Var}(\mathbf{X})$ or $\text{Var}(\mathbf{Y})$.

Although PLS is not a statistical model, it is sometimes said that the vectors of latent scores for a block “model” an interesting subspace of the column space of that block. This is merely another way to say that these vectors form a basis for that subspace. This sense of the term “model” is different from the statistical sense. The following example serves to underline this. First observe that, in a full-rank PLS analysis, as long as $\mathbf{X}^T \mathbf{Y}$ is of full rank the block with the minimum number of variables is “modeled” without error. That is, the vectors of latent scores for that block form a complete basis for the column space of that block. For instance, if $p \leq q$, $\mathbf{X}^T \mathbf{Y}$ is of full rank, and the analyst, using either PLS-SVD or PLS-W2A, computes p pairs of latent score vectors, then \mathbf{X} can be exactly recovered (up to rounding error) from the p vectors of latent scores and the p salience vectors for the \mathbf{X} block. Let us return, however, to the statistical sense of the word “model.” There is no reason that a paired latent model with $p \leq q$ and Σ_{XY} of full rank must specify a fully deterministic relationship between the random vector \mathbf{x}_n and its latent variable ξ_n . This would, of course, mean an identically zero error, $\epsilon_n \equiv 0$.

Since PLS is not an estimator for the paired latent model, in particular we cannot use PLS to provide starting points for an iterative maximum-likelihood algorithm designed for structural equation modeling. Suppose we take, as estimates of within-block variance of the latents Φ and Ψ , the sample variances and covariances of the PLS latent scores. These values may fail even to be feasible, in that they may yield a matrix $\begin{bmatrix} \Phi & \mathbf{D} \\ \mathbf{D} & \Psi \end{bmatrix}$ which fails to be positive semidefinite. The reason for this is explored on page 23. Finding a feasible value is not trivial. In the current work, Chapters 4 and 5, algorithms will be presented by which estimates of within-block parameters can be obtained, starting from a SVD decomposition of $\mathbf{X}^T \mathbf{Y}$ such as is used in the PLS-SVD algorithm.

We might shift our attention from the sample dispersion to the population dispersion. Then we might ask ourselves whether, given a population dispersion for the indicators, a PLS method might be used to find feasible parameters for the paired latent model. This

is a continuation of the foregoing theme, and the answer again is no. The matrices of latent scores and errors computed by PLS methods are not guaranteed to satisfy constraints analogous to the population constraints of the latent model. The error variables for a given block are guaranteed to be uncorrelated with the latent variables for that block. That is, $\hat{\mathbf{E}} \perp \hat{\Xi}$ and $\hat{\mathbf{Z}} \perp \hat{\Omega}$. The errors will not, however, in general be uncorrelated with the matrix of latent variable scores for the other block, or with the error matrix for the other block. That is, in general

$$\begin{aligned}\hat{\mathbf{E}}^T \hat{\Omega} &\neq \mathbf{0} \\ \hat{\mathbf{Z}}^T \hat{\Xi} &\neq \mathbf{0} \\ \hat{\mathbf{E}}^T \hat{\mathbf{Z}} &\neq \mathbf{0} .\end{aligned}\tag{2.1}$$

An example with $N = 3$ and $p = q = 2$ is in Section 2.3.2. As a consequence of (2.1), the coefficients linking \mathbf{X} and \mathbf{Y} to the latent variables do not provide a decomposition of the cross-covariance.

PLS and likelihood. Since PLS is not linked to a probability model, no unified set of techniques exists for performing inference on the coefficients computed by the algorithm. Some researchers (for instance, Sampson et al. [SSBB89] page 482 and Streissguth et al. [SBSB93a] page 83) advocate use of the bootstrap for performing inference. Cross-validation has been used for model selection [HHMT97]. Others have proposed approaches based on a multivariate normal likelihood (for instance Höskuldsson [H88] and Holcomb et al. [HHMT97]). In spite of the assumption of multivariate normality that underlies some of these approaches, however, none of them explores the full potential of a parametric, likelihood-based framework.

The lack of principled methods for model selection and verification in the context of PLS analysis is a serious shortcoming. In a properly specified model, PLS saliences are stable, even when one or more indicators is removed from the analysis and saliences are recomputed. If the model has been incorrectly specified, however, saliences will not in general have this property. Thus a misspecified model can produce misleading results. The effect of misspecification on saliences computed by PLS is examined further in Section 2.3.1.

PLS and correlation. PLS underestimates the between-block, within-pair correlation of the latent variables. Consider, for instance, the SVD model when rank $R = 1$. (This model is defined at page 5, Section 1.1.) The population correlation is

$$\rho = \frac{d}{\sqrt{\phi\psi}}.$$

The PLS estimate of correlation is computed from the vectors of latent scores. Let \mathbf{u} and \mathbf{v} be the population saliences for the \mathbf{X} - and \mathbf{Y} -blocks, d the covariance of the latents. (Recall that in the SVD model we use \mathbf{u} and \mathbf{v} for the population saliences, as stated on page 5.) Let their PLS estimates be $\hat{\mathbf{u}}$, $\hat{\mathbf{v}}$, and \hat{d} . Since these come from the singular value decomposition, they are known to be numerically stable. They are also consistent, as argued on page 2.3. The PLS estimate of the variance of the latent for the \mathbf{X} -block is

$$\begin{aligned} \hat{\phi} &= \frac{1}{N} (\mathbf{X}\hat{\mathbf{u}})^T (\mathbf{X}\hat{\mathbf{u}}) \\ &\xrightarrow{N \rightarrow \infty} \mathbf{u}^T \Sigma_{XX} \mathbf{u} \\ &= \mathbf{u}^T (\phi \mathbf{u} \mathbf{u}^T + \Sigma_{\epsilon\epsilon}) \mathbf{u} \\ &= \phi + \mathbf{u}^T \Sigma_{\epsilon\epsilon} \mathbf{u}; \end{aligned} \tag{2.2}$$

similarly $\hat{\psi} \rightarrow \psi + \mathbf{v}^T \Sigma_{\zeta\zeta} \mathbf{v}$ as N approaches infinity. There is no reason for the second term in these expressions to be close to zero. The PLS estimate of correlation between the latents is

$$\begin{aligned} \hat{\rho} &= \frac{\hat{d}}{\sqrt{\hat{\phi}\hat{\psi}}} \\ &\xrightarrow{N \rightarrow \infty} \frac{d}{\sqrt{(\phi + \mathbf{u}^T \Sigma_{\epsilon\epsilon} \mathbf{u})(\psi + \mathbf{v}^T \Sigma_{\zeta\zeta} \mathbf{v})}} \\ &\ll \frac{d}{\sqrt{\phi\psi}} \\ &= \rho. \end{aligned} \tag{2.3}$$

Spearman's correction for attenuation. Readers will note that the bias toward zero of $\hat{\rho}$ in (2.3) is nothing but the “attenuation” first discussed by Spearman [Spe04]. This may be seen even more explicitly in the fact that we have the following PLS-SVD estimates

of the latent scores for the n th observation:

$$\widehat{\xi}_n = \mathbf{x}_n^T \widehat{\mathbf{u}} = \xi_n + \epsilon_n^T \widehat{\mathbf{u}} \quad \text{and}$$

$$\widehat{\omega}_n = \boldsymbol{\omega}_n^T \widehat{\mathbf{v}} = \omega_n + \zeta_n^T \widehat{\mathbf{v}} .$$

Corrections for attenuation, resulting in **disattenuated** correlations, have been used since Spearman's seminal article. The original formula is

$$\text{Cor}(T_x, T_y) = \frac{\text{Cor}(X, Y)}{\sqrt{\text{Cor}(X, X')\text{Cor}(Y, Y')}} .$$

In this formula, T_x and T_y are true, unobservable scores. What are observed are X , X' , Y , and Y' , where

$$X = T_x + E , \quad X' = T_x + E' ,$$

$$Y = T_y + Z , \quad Y' = T_y + Z' ;$$

E, E', Z, Z' are independent of each other,

$$\text{Var}(E) = \text{Var}(E') , \text{ and } \quad \text{Var}(Z) = \text{Var}(Z') .$$

The variables X and X' are called **parallel measurements**. The quantities $\text{Cor}(X, X')^2$ and $\text{Cor}(Y, Y')^2$ are “reliabilities,” or squared correlations between observed scores and true scores.

Corrections for attenuation have been used since Spearman's seminal article, yet they are controversial and fraught with difficulty. Since reliabilities themselves must be estimated, disattenuated “correlations” exceeding unity are common. In the rank-one paired latent model, however, explicit disattenuation methods, with their attendant estimation of reliabilities, are unnecessary. It will be seen in Section 4.3.3 that ρ can only be identified up to a lower bound on its absolute value; a correlation of 1 or -1 is always feasible. This is true even if the population dispersion of the indicators, $\boldsymbol{\Sigma}$, is known. The lower bound on ρ is obtained directly from the dispersion of the indicators, without computation of latent scores.

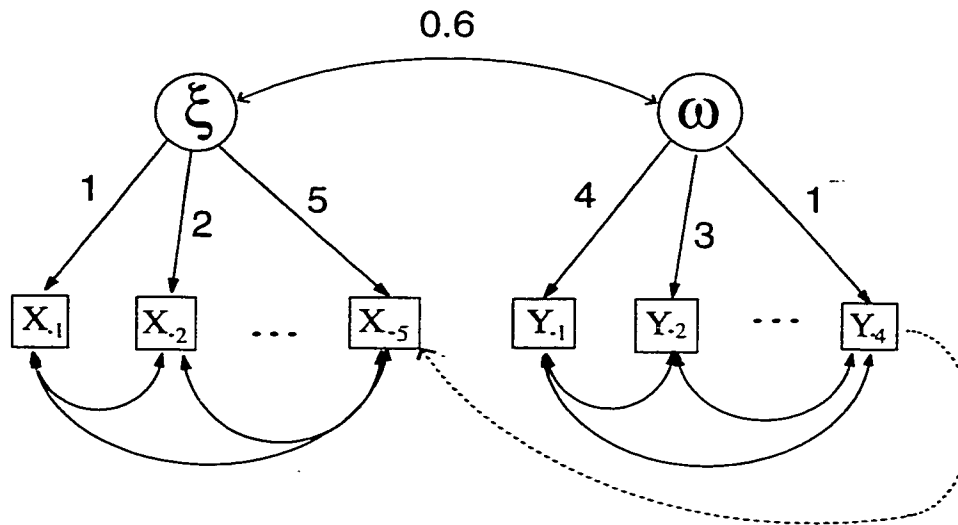


Figure 2.1: A misspecified model, as discussed on page 25. The analyst believes that the data are properly divided into two blocks, X_1, \dots, X_5 and Y_1, \dots, Y_4 , and that the cross-covariance is of rank one. A direct dependency exists however between X_5 and Y_4 , making the rank-one model inappropriate.

There is an extensive literature on attenuation and on corrections for attenuation. Spearman's seminal article [Spe04] was reprinted in 1987 [Spe87]. Attenuation is mentioned by Kendall and Stuart [KS67], page 327, and by Fisher and van Belle [FvB93], page 385. Lord and Novick provide a mathematical justification for the correction for attenuation [LN68]. The current exposition is obtained from that reference. Muchinsky reviews the issues and controversies surrounding disattenuation, including alternate formulas [Muc96]. Zimmerman and Williams use simulation to investigate the properties of the disattenuated correlation under various conditions [ZW97].

2.3.1 An Example of Misspecification

It was stated in Section 2.3 that the application of PLS under a misspecified model can produce misleading results. In the current section we look at an example. Consider the covariance structure specified in Figure 2.1 on page 25. Without the relationship indicated by the broken line, the data satisfy a rank $R = 1$ paired latent model. Suppose however

that the relationship indicated by the broken line is included. Then

$$\mathbf{X}_{n\cdot}^T = \mathbf{u}\xi_n + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ v_4 \end{bmatrix} \omega_n + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} (\zeta_n)_4 + \epsilon_n,$$

and

$$\begin{aligned} \text{Cov}(\mathbf{X}_{n\cdot}^T, \mathbf{Y}_{n\cdot}^T) &= d\mathbf{u}\mathbf{v}^T + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ v_4 \end{bmatrix} \mathbf{v}^T + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \text{Cov}((\zeta_n)_4, \zeta_n) \\ &= d\mathbf{u}\mathbf{v}^T + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} (v_4\mathbf{v}^T + (\Sigma_{\zeta\zeta})_4), \end{aligned}$$

where $(\zeta_n)_4$ is the fourth component of ζ_n , and $\Sigma_{\zeta\zeta} = \text{Var}(\zeta_n)$. The rank-one PLS model is no longer appropriate, since this is a rank-two matrix.

Graphical Evaluation of Saliency Stability. In the misspecified model we cannot expect that saliencies will be stable. An example was simulated to demonstrate this. Two thousand observations were simulated from the properly specified model discussed on page 25. This dataset was used as input to the two-block Mode A PLS algorithm with R , the number of pairs of latent variables, equal to one, and scaled saliencies were computed for the full set of indicators. (The two-block, Mode A PLS algorithm is discussed in detail in Chapter 3. In the $R = 1$ case, the saliencies computed by this algorithm are simply the scaled left and right singular vectors of $\mathbf{X}^T\mathbf{Y}$ corresponding to the largest singular value.) Then one by one each of the indicators was removed from the dataset, leaving the others in,

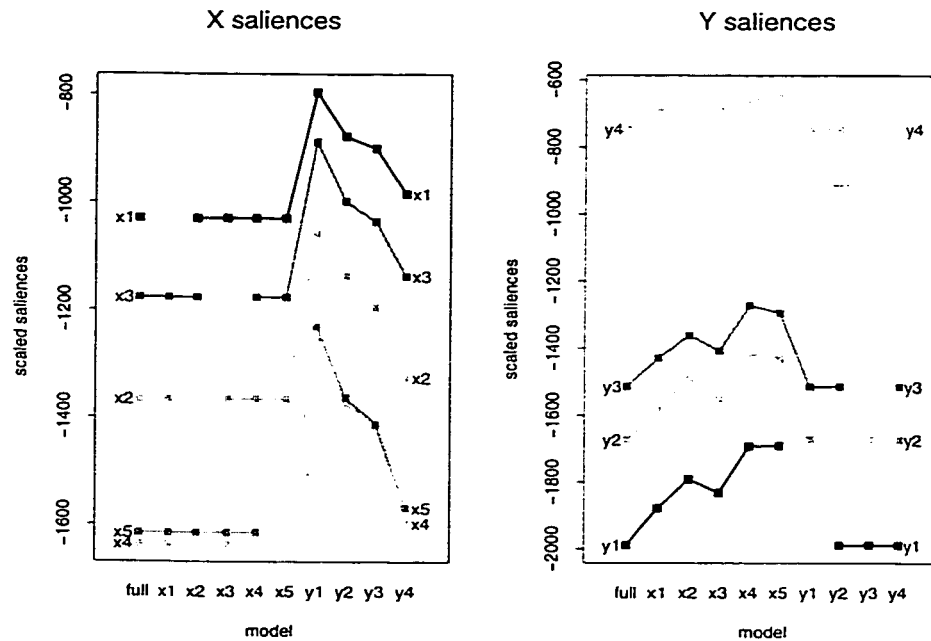


Figure 2.2: Leave-one-out salience plot for a correctly specified model, as discussed in Section 2.3.1.

and the scaled saliences were recomputed for the remaining indicators. Finally the scaled saliences were plotted as a function of “model.” “Model” here means which indicator has been removed before computation of PLS saliences. The “full” model is the model when all indicators are included. The results for the properly-specified model may be seen in Figure 2.2 on page 27. When indicators for the **Y** block are removed, saliences for the **X** block change, but retain the ordering of the full model. Similarly when indicators for the **X** block are removed, saliences for the **Y** block change but retain the ordering of the full model. The movement of saliences for a given block is almost imperceptible when indicators for that block are removed.

When the model is incorrectly specified, however, removal of an indicator can change the order of the saliences. This can be seen in Figure 2.3 on page 28. With **X**₅ in the model, the salience for **Y**₄ is less than the saliences for **Y**₁ and **Y**₃. When **X**₅ is removed, the salience for **Y**₄ becomes greater than these saliences.

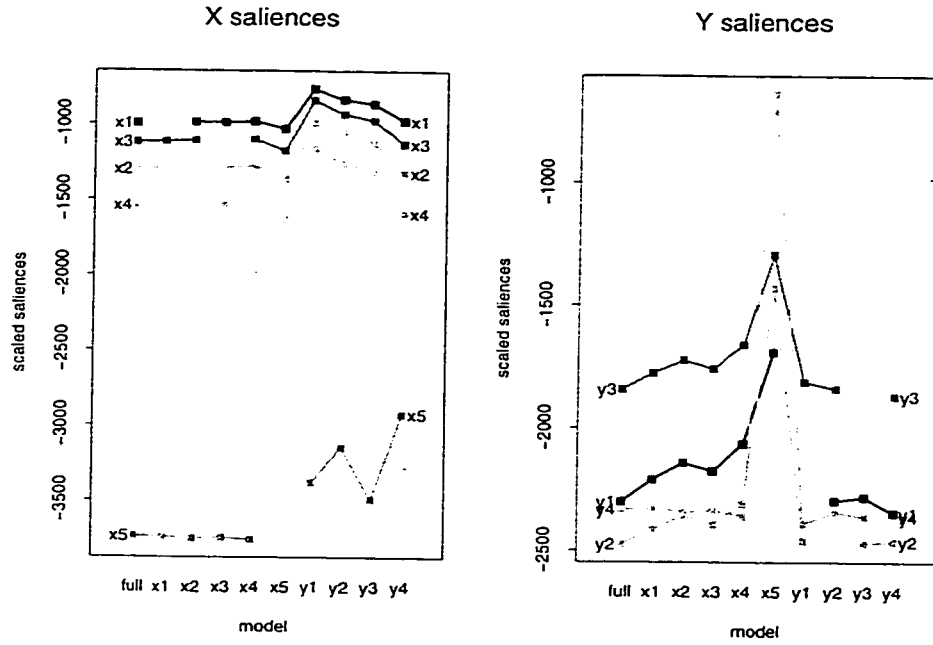


Figure 2.3: Leave-one-out salience plot for a misspecified model, as discussed on page 26.

2.3.2 An example where the analogy between PLS and the latent-variable model breaks down

In Section 2.3 it was claimed that “the matrices of latent scores and errors computed by PLS methods are not guaranteed to satisfy constraints analogous to the population constraints of the latent model.” In this section a counterexample is given which proves the claim.

Suppose

$$\mathbf{X} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} -3 & -1 \\ 3 & 5 \\ 0 & -2 \end{bmatrix}.$$

Both matrices are of full rank, but the first column of \mathbf{Y} is orthogonal to both columns of \mathbf{X} . Thus the cross-product is of rank one:

$$\mathbf{X}^T \mathbf{Y} = \begin{bmatrix} 0 & -2 \\ 0 & 2 \end{bmatrix} = \hat{\mathbf{U}} \hat{\mathbf{D}} \hat{\mathbf{V}}^T$$

where

$$\hat{\mathbf{U}} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix},$$

$$\hat{\mathbf{D}} = \begin{bmatrix} 2\sqrt{2} \end{bmatrix}$$

and

$$\hat{\mathbf{V}} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

This means that the PLS latent score matrices will be of rank one:

$$\hat{\Xi} = \mathbf{X}\hat{\mathbf{U}} = \begin{bmatrix} -1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix}$$

and

$$\hat{\Omega} = \mathbf{Y}\hat{\mathbf{V}} = \begin{bmatrix} 1 \\ -5 \\ 2 \end{bmatrix},$$

so that the residuals will be nonzero. Regressing \mathbf{X} on $\hat{\Xi}$ we obtain the following coefficients and residuals for \mathbf{X} :

$$\hat{\mathbf{\Gamma}} = \begin{bmatrix} 0 \\ -\sqrt{2} \end{bmatrix}, \quad \hat{\mathbf{E}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}.$$

Regressing \mathbf{Y} on $\hat{\Omega}$ we obtain the following coefficients and residuals for \mathbf{Y} :

$$\hat{\Delta} = \begin{bmatrix} -0.6 \\ -1 \end{bmatrix}, \quad \hat{\mathbf{Z}} = \begin{bmatrix} -2.4 & 0 \\ 0 & 0 \\ 1.2 & 0 \end{bmatrix}.$$

We confirm that the decomposition is correct:

$$\begin{aligned}
 \hat{\Xi}\hat{\Gamma}^T + \hat{\mathbf{E}} &= \begin{bmatrix} -1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix} \begin{bmatrix} 0 & -\sqrt{2} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \\
 &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \\
 &= \mathbf{X}
 \end{aligned}$$

and

$$\begin{aligned}
 \hat{\Omega}\hat{\Delta}^T + \hat{\mathbf{Z}} &= \begin{bmatrix} 1 \\ -5 \\ 2 \end{bmatrix} \begin{bmatrix} -0.6 & -1 \end{bmatrix} + \begin{bmatrix} -2.4 & 0 \\ 0 & 0 \\ 1.2 & 0 \end{bmatrix} \\
 &= \begin{bmatrix} -0.6 & -1 \\ 3 & 5 \\ -1.2 & -2 \end{bmatrix} + \begin{bmatrix} -2.4 & 0 \\ 0 & 0 \\ 1.2 & 0 \end{bmatrix} \\
 &= \mathbf{Y} .
 \end{aligned}$$

The residuals for the \mathbf{X} block are not orthogonal to the latent variable scores for the \mathbf{Y} block, the residuals for the \mathbf{Y} block are not orthogonal to the latent variable scores for the \mathbf{X} block, and the residuals for the the two blocks are not orthogonal to each other:

$$\hat{\mathbf{E}}^T \hat{\Omega} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$\hat{\Xi}^T \hat{\mathbf{Z}} = \begin{bmatrix} \frac{12}{5\sqrt{2}} & 0 \end{bmatrix}$$

$$\hat{\mathbf{E}}^T \hat{\mathbf{Z}} = \begin{bmatrix} 1.2 & 0 \\ 1.2 & 0 \end{bmatrix}$$

2.4 Conclusion

Three classes of methods are currently in use which can conceivably be applied to cross-covariance problems. Of these, canonical correlation analysis (CCA) and a class of methods designed for structural equation models (SEMs) are unsuited for cross-covariance problems. Partial Least Squares, alias Projection to Latent Structures (PLS), has been used successfully on cross-covariance problems. PLS has hitherto not been linked to a statistical model, however.

In the current work a class of latent-variable models, the paired latent models, is specified. This class has the property that the rank-one PLS saliences are consistent for the subset of the parameters of the rank-one paired latent model which govern the relationship between the latent variables and the indicators. PLS is not a way to obtain values for the full parameter set, however. In particular PLS does not compute values for the parameters which govern the within-block covariance. Also it does not provide a consistent estimate for the within-block correlation ρ_k between the k th pair of latent variables. In the current work a method is presented by which any distribution in the rank- r reduced-rank-regression model (equivalently, a rank- r constraint model, as defined on page 2 in Section 1.1) can be parameterized by a rank- r paired latent model. Since it starts with the singular value decomposition of Σ_{XY} , this method may be seen as an extension of PLS. Furthermore, by linking PLS to the paired latent model we obtain a consistent estimate for ρ_1 .

We end this chapter by noting a recent article in the chemometric literature. Burnham et al., writing in the *Journal of Chemometrics*, say: "...to the best of our knowledge, there is no research that has dealt with PLS from the standpoint of a parameter estimation method for a statistical model. If PLS could be derived as a method arising from the application of a reasonable statistical parameter estimation technique to a believable statistical model for the data, this would lend some strength to the argument that it is a good choice of parameter estimation method for such data" (page 50 of [BMV99]). Burnham et al. then express two-block asymmetric PLS as maximum-likelihood estimation for a family of statistical models, but their family contains no latent variables. Their analogue to the latent variables ξ and

ω of the current work is a fixed parameter vector \mathbf{T} , satisfying

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} ,$$

$$\mathbf{Y} = \mathbf{TQ} + \mathbf{F} .$$

Thus, as the number of observations increases so does the number of parameters. The parameter set in the current work, on the other hand, does not change with the number of observations. Instead, pairs of latent variables are postulated, and a method is presented by which the parameters governing their distribution may be estimated. Since the number of parameters remains constant as the number of observations increases, there is the possibility that standard asymptotic methods may be used to gain insight into this model family.

Chapter 3

SURVEY OF PARTIAL LEAST SQUARES (PLS) METHODS, WITH EMPHASIS ON THE TWO-BLOCK CASE

3.1 *Abstract*

Partial Least Squares (PLS; the acronym has also been explained as “Projection to Latent Structures” [BEE98]) is a class of techniques for modeling the association between blocks of observed variables by means of latent variables. Originated by Herman Wold in the 1970’s, PLS is important in many scientific disciplines, including psychology, economics, chemistry, medicine and the pharmaceutical sciences, and process modelling (Rannar et al. [RLGW94]).

PLS has many variants. The algorithm can be run in two modes, called A and B. It can be applied to data that are divided into two or more blocks. The general algorithm due to Wold can be followed, or it can be modified.

Wold stated his general algorithm in terms different from those customarily used by statisticians. In the current work the algorithm is placed into more familiar terminology and notation, and the two-block case is discussed. Wold’s two-block Mode A PLS (PLS-W2A) is stated. Its properties, and the properties of the coefficients it computes, are examined in detail. In particular PLS-W2A is shown to be a special case of Wold’s general algorithm. Another two-block Mode A variant, PLS-SVD, is shown to depart from Wold. PLS-SVD has been used for both modeling (Sampson et al. [SSBB89], Bookstein et al. [BSSB96]) and for prediction (Tishler and Lipovetsky [TL00]). PLS-SVD is also called Robust Canonical Analysis, Intercorrelations Analysis, and Canonical Covariance (Tishler et al. [TDSL96]). The article by Tishler et al. does not refer to the work of Sampson and Bookstein regarding PLS. Differences between the coefficients computed by PLS-W2A and PLS-SVD are discussed.

PLS-W2A and PLS-SVD are contrasted with another two-block variant, PLS2, which appears in the chemometric literature and is used for prediction. PLS1, another variant which appears in the chemometric literature, is shown to be a special case of PLS2 with one of the blocks consisting of a single variable. Canonical Analysis, also called Canonical Correlation (CCA), an older approach to the modeling of association between blocks of data, is equivalent to Mode B of Wold's algorithm. The properties of CCA are contrasted with those of PLS-W2A and PLS-SVD. Finally, an inner loop which appears in some statements of the algorithm is nothing but the well-known power method for computing the singular vectors of a matrix. A detailed proof of this fact is presented.

3.2 Framework

Suppose we have two matrices, \mathbf{X} and \mathbf{Y} , respectively $N \times p$ and $N \times q$, where the columns correspond to variables and the rows to observations. Two-block Mode A Partial Least Squares is a class of techniques for modeling $\mathbf{X}^T \mathbf{Y}$, the cross-covariance of \mathbf{X} and \mathbf{Y} , by means of latent variables.

Partial Least Squares (PLS) is an important tool in many scientific fields, including psychology, economics, chemistry, medicine and the pharmaceutical sciences and process modelling. (Rannar et al. [RLGW94] p. 111.) This widespread use is explained by the fact that PLS has many attractive properties. The coefficients computed in a PLS analysis are well-defined and easy to interpret. PLS is especially useful when the columns of \mathbf{X} or of \mathbf{Y} are collinear or nearly collinear, or when there are more variables than observations ($p > N$ or $q > N$), since few other methods are available in such a case. The PLS algorithms which are used in the "two-block, mode A" case (to be defined below) are numerically stable. Provided the singular values of $\mathbf{X}^T \mathbf{Y}$ are distinct, these algorithms are guaranteed to converge. As we have seen in Chapter 2, PLS compares favorably with other techniques which might be used for modeling association between two blocks of variables, such as canonical analysis, multiple regression, and the software packages LISREL and EQS.

An Example from Behavioral Teratology In a study of the relationship between fetal alcohol exposure and neurobehavioral deficits reported by Sampson et al. [SSBB89] and by

Streissguth et al. [SBSB93a], the $X_{.i}$ are 13 different measures of the mother's alcohol intake during pregnancy, and the $Y_{.j}$ are 11 IQ subtests.

Path Diagrams Path diagrams are useful for displaying graphically a set of hypothesized relationships between variables [SRM⁺ar]. In particular, they are useful for diagramming the set of assumptions which justifies the application of a PLS analysis to a specific dataset. A path diagram for the example from behavioral teratology may be found in Figure 3.1 on page 36. Observed, or **indicator**, variables are enclosed in rectangles, latent variables in ellipses. A double-headed arrow between two variables indicates a non-zero correlation between their errors. A single-headed arrow from one variable to another indicates that an equation is hypothesized for the first variable, and that there is a non-zero coefficient for the second variable in this equation. The entire set of independence and conditional independence relationships between the variables represented by the vertices of a path diagram can be determined from a graph. For details the reader is referred to Section 1.2.

Several different kinds of path diagrams appear in the literature, following different conventions. No attempt will be made to survey them. It should be noted, however, that the path diagrams that accompany Herman Wold's general PLS algorithm, an example of which is displayed in Figure 3.2 on page 57, do not follow the convention stated here. The reason for this difference is that the path diagram discussed in this section specifies all dependent relationships, or nonzero correlations, between the variables in question. H. Wold's path diagrams, on the other hand, at least those which accompany his 1985 article [Wol85], specify only those relationships which will be translated into instructions for the algorithm which he defines in that article.

3.3 *Background and Overview*

Many variants of PLS appear in the literature. Herman Wold used the term first, in the context of structural equation models [Wol75]. A survey and history of PLS methods may be found in Geladi [Gel88]. Wold's method is not a single algorithm but a class of algorithms, encompassing arbitrary numbers of blocks of indicators with their associated latent variables, arbitrary linear relationships between the latent variables, and two **Modes**

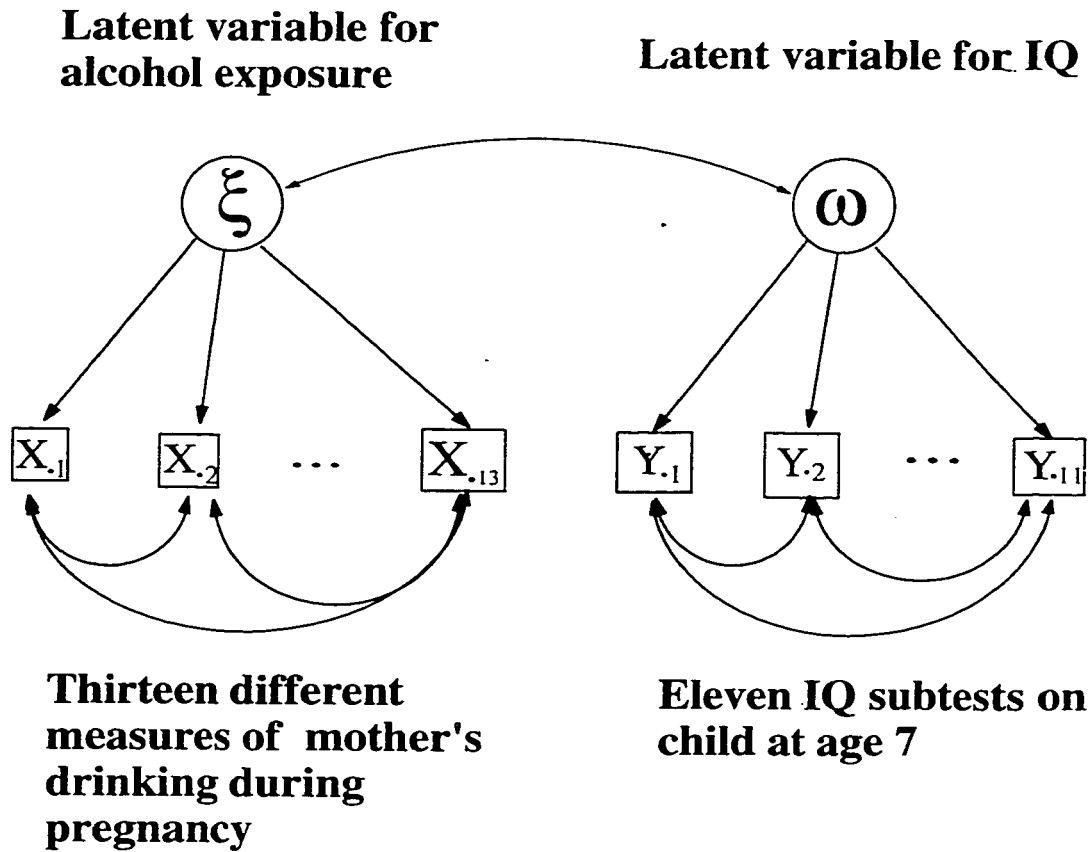


Figure 3.1: Path diagram for a study of fetal alcohol exposure and IQ. Path diagrams are introduced in Section 1.2. Enclosed in ellipses are latent variables ξ and ω . The X and Y variables, enclosed in rectangles, are observed or **indicator** variables. The double-headed arrow between ξ and ω indicates a non-zero correlation. The single-headed arrows from the latent variables to the indicators indicate there are non-zero coefficients for the latent variables in the equations for the indicators. The lack of an arrow, or edge, between ξ and $Y_{.1}$ means that any dependence between these variables can occur only through the other variables. The fact that the removal of ω from the diagram would result in the lack of a path between ξ and $Y_{.1}$ means that ξ and $Y_{.1}$ are conditionally independent given ω . The doubleheaded arrows between the X variables means that they are not conditionally independent given ξ . Similarly the doubleheaded arrows between the Y variables means that they are not conditionally independent given ω .

in which computation can be performed, Mode A or Mode B [Wol85]. The algorithms can be computed in either of the Modes or in a combination of the two. Coefficients computed by an algorithm run completely in Mode B are interpreted in a fundamentally different way from coefficients computed by an algorithm run completely in Mode A.

The current work focuses on what is called, in the context of Herman Wold's work, "Two-Block Mode A Partial Least Squares." Here it will be referred to as **PLS-W2A** (Wold's Two-Block, Mode A PLS). This is first described in detail, and its properties are discussed. Then Wold's original, general class of algorithms is described in detail. Then several more algorithms in the PLS class, and their differences from PLS-W2A, are discussed. The algorithms are placed, as much as possible, into a common notation to facilitate comparison. The following points are made.

- Wold's algorithm, applied to two blocks, but in Mode B rather than Mode A, is equivalent to canonical correlation analysis (CCA). CCA coefficients are interpreted in a fundamentally different way from PLS-W2A coefficients. Although CCA belongs to the class of PLS algorithms, for historical reasons the term PLS has come to be associated with Mode A algorithms. Thus two-block Mode B PLS will be referred to in the current work as CCA, not as PLS. CCA, and its differences from PLS-W2A, are discussed in Section 3.11.
- In some contexts, Mode A algorithms are reported without use of the term "PLS." For instance, Tishler et al. use the terms "Intercorrelations Analysis," "Canonical Covariance," and "Robust Canonical Analysis" [TDSL96] [TL00] to refer to PLS-SVD: Section 3.6.
- PLS-SVD has been used by Sampson, Bookstein, Streissguth et al. in the study of behavioral teratology [BSSB96]. In this application it is used for modeling, not for prediction, and the PLS-SVD coefficients are interpreted in a way similar to those of PLS-W2A: Section 3.6.
- PLS-SVD can also be used for prediction, as shown by Tishler and Lipovetsky: Section

3.6.

- In both PLS-SVD and PLS-W2A, there is nothing in the algorithms themselves to prevent the number of pairs of vectors of latent scores computed from exceeding the rank of $\mathbf{X}^T \mathbf{Y}$. Although this may not occur in practice, this fact should be noted in a rigorous study of the algorithms. In PLS-SVD, if $S \equiv \text{rank}(\mathbf{X}^T \mathbf{Y})$, it is only the first S pairs of latent score vectors which have nonzero covariance: Section 3.7.
- PLS2 and PLS1, variants of two-block Mode A PLS which appear in the chemometric literature, are used for prediction: Section 3.9.
- PLS1, the special case of PLS when the \mathbf{Y} block consists of a single column, is a regularization technique in the same class as ridge regression: Section 3.9.
- PLS-W2A, PLS-SVD, PLS2, and PLS1 are equivalent when just one pair of latent variable scores (ξ and ω) is computed. The differences occur in cases when the outer loop exceeds one iteration: Section 3.10.

In addition we shall deal with the following more technical issues.

- We shall confirm that PLS-W2A is a special case of Herman Wold's original algorithm: Section 3.8.5.
- Wold's algorithm can use as input either the raw data (\mathbf{X} and \mathbf{Y} in the two-block case) or what Wold calls product data ($\mathbf{X}^T \mathbf{Y}$ in the two-block case), and the result is the same up to roundoff error: Section 3.8.4.
- Frequently in the literature the computation of the first pair of singular vectors of the current cross-product matrix (Step 3 on page 41, in the PLS-W2A algorithm) is not stated in terms of singular vectors; no mention is made of singular vectors or of the singular value decomposition. Instead this computation is stated as an explicit inner loop. We shall see that an *a priori* starting value exists for this inner loop such that

it is guaranteed to converge, and to yield the first pair of singular vectors, as in Step 3 of the PLS-W2A algorithm: Section 3.12.

3.4 Framework for Two-Block PLS

Let \mathbf{X} and \mathbf{Y} be data matrices, respectively $N \times p$ and $N \times q$. We use PLS-W2A when

- we think that the \mathbf{X} variables (the columns of \mathbf{X}) serve as “indicators” for one or more latent variables, say ξ_r , and the \mathbf{Y} variables serve as “indicators” for the same number of latent variables, ω_r ,
- we are primarily interested, not in the covariance matrix of the \mathbf{X} variables or the covariance matrix of the \mathbf{Y} variables, but in the cross-covariance of \mathbf{X} and \mathbf{Y} , and
- we wish to model the cross-covariance by pairs of composite scores on hypothesized latent variables, say

$$(\xi_1, \omega_1), \dots, (\xi_R, \omega_R).$$

- we wish the set $\{\xi_1, \dots, \xi_R\}$ to be orthogonal, and we wish the set $\{\omega_1, \dots, \omega_R\}$ to be orthogonal.

Note that each distinct value of R corresponds to a different model. Thus a decision regarding the number of pairs of latent variables to compute constitutes the selection of a model. The value of R has been called the **rank**. Nothing in the PLS-W2A or PLS-SVD algorithms themselves prevents the number of pairs of latent variables from exceeding the rank of $\mathbf{X}^T \mathbf{Y}$, however, as we shall see in Sections 3.5.2 and 3.7.

On the absence of “hat” notation in this chapter. A difference exists between the notation in this chapters and in the other chapters of this thesis. In the rest of this thesis, the “hat” notation, such as $\hat{\mathbf{u}}$ and $\hat{\xi}_1$, may be used to represent estimates of parameters and latent variable scores. This chapter is not concerned with statistical models, however. All quantities represent either data, as in the case of \mathbf{X} and \mathbf{Y} , or quantities computed from the data, for instance ξ_1 and ω_1 .

3.5 PLS-W2A

The data are as in Section 3.3. PLS-W2A computes a sequence of pairs of vectors of latent scores

$$(\xi_1, \omega_1), \dots, (\xi_R, \omega_R)$$

such that the sets

$$\{\xi_1, \dots, \xi_R\} \quad \text{and} \quad \{\omega_1, \dots, \omega_R\}$$

are orthogonal. For any value of r between 1 and R , the sets

$$\{\xi_1, \dots, \xi_r\} \quad \text{and} \quad \{\omega_1, \dots, \omega_r\}$$

span the “most interesting” subspaces of the ranges (column spaces) of \mathbf{X} and of \mathbf{Y} . The subspaces are “most interesting” not from the point of view of accounting for $\mathbf{X}^T \mathbf{X}$ or $\mathbf{Y}^T \mathbf{Y}$, but from the point of view of accounting for $\mathbf{X}^T \mathbf{Y}$. We have

$$d_1 \equiv \text{Cov}(\xi_1, \omega_1) = \max_{\|\mathbf{u}\|=\|\mathbf{v}\|=1} \text{Cov}(\mathbf{X}\mathbf{u}, \mathbf{Y}\mathbf{v}). \quad (3.1)$$

Let \mathbf{u}_1 and \mathbf{v}_1 be the vectors of coefficients which maximize (3.1). Then by a well-known property of the singular value decomposition we know that $d_1 \mathbf{u}_1 \mathbf{v}_1^T$ is the best rank-one approximation of $\mathbf{X}^T \mathbf{Y}$ in the least-squares sense (Harville [Har97] page 556).

To obtain the subsequent pairs of latent vector scores,

$$(\xi_2, \omega_2), \dots, (\xi_R, \omega_R),$$

a sequence of residual matrices

$$\left\{ \left(\mathbf{X}^{(r)}, \mathbf{Y}^{(r)} \right) : r = 1, \dots, R \right\}$$

is computed by subtracting, at each step, from the current versions of \mathbf{X} and \mathbf{Y} , rank-one approximations based on the latent vector scores already computed, and repeating the optimization in (3.1) with \mathbf{X} and \mathbf{Y} replaced by $\mathbf{X}^{(r)}$ and $\mathbf{Y}^{(r)}$. Details are in Section 3.5.1.

In addition to latent variable scores, PLS-W2A yields vectors of coefficients \mathbf{u}_r and \mathbf{v}_r , $r = 1, \dots, R$. These coefficients, called **salience**s, are interpretable in relation to the residual matrices. We have

$$\begin{aligned} \mathbf{u}_r &= \begin{bmatrix} u_{1r} \\ \vdots \\ u_{I_r} \end{bmatrix} \propto \begin{bmatrix} \text{Cov}(\mathbf{X}_{\cdot 1}^{(r)}, \omega_r) \\ \vdots \\ \text{Cov}(\mathbf{X}_{\cdot I_r}^{(r)}, \omega_r) \end{bmatrix} \\ \mathbf{v}_r &= \begin{bmatrix} v_{1r} \\ \vdots \\ v_{J_r} \end{bmatrix} \propto \begin{bmatrix} \text{Cov}(\mathbf{Y}_{\cdot 1}^{(r)}, \xi_r) \\ \vdots \\ \text{Cov}(\mathbf{Y}_{\cdot J_r}^{(r)}, \xi_r) \end{bmatrix} \end{aligned}$$

The scalar $u_{i,r}$ is a measure of the importance of $\mathbf{X}_{\cdot i}^{(r)}$ in relation to the latent variable for the $\mathbf{Y}^{(r)}$ variables. The scalar $v_{j,r}$ is interpreted in an entirely symmetric manner.

3.5.1 PLS-W2A Algorithm

In order that the coefficients may be readily interpreted, we usually assume that the columns of \mathbf{X} and \mathbf{Y} have been centered. In addition, we assume that within each block, the variables are measured in the same units. (In practice, the variables may have no intrinsic units. In this case they may be simply standardized so that each column has unit norm.)

1. $r \leftarrow 1$.
2. $\mathbf{X}^{(1)} \leftarrow \mathbf{X}$,
 $\mathbf{Y}^{(1)} \leftarrow \mathbf{Y}$.
3. Compute the first pair of singular vectors of $(\mathbf{X}^{(r)})^T \mathbf{Y}^{(r)}$. These are the singular vectors associated with the largest singular value. Let \mathbf{u}_r and \mathbf{v}_r be respectively the left and right vectors of this pair.

The following convention makes this step unambiguous:

$$\begin{aligned} \mathbf{u}_r^T \mathbf{u}_r &= 1, \quad \mathbf{v}_r^T \mathbf{v}_r = 1, \text{ and} \\ (\mathbf{u}_r)_i &> 0, \quad \text{where } i = \text{argmax} |(\mathbf{u}_r)_i|. \end{aligned}$$

$$4. \quad \xi_r \leftarrow \mathbf{X}^{(r)} \mathbf{u}_r,$$

$$\omega_r \leftarrow \mathbf{Y}^{(r)} \mathbf{v}_r.$$

5. Regress $\mathbf{X}^{(r)}$ on ξ_r , $\mathbf{Y}^{(r)}$ on ω_r , obtaining rank-one approximations of the data matrices:

$$\widehat{\mathbf{X}}^{(r)}(\xi_r) = \xi_r (\xi_r^T \xi_r)^{-1} \xi_r^T \mathbf{X}^{(r)}$$

$$\widehat{\mathbf{Y}}^{(r)}(\omega_r) = \omega_r (\omega_r^T \omega_r)^{-1} \omega_r^T \mathbf{Y}^{(r)}$$

It will be useful to write the regression coefficients explicitly:

$$\gamma_r^T = (\xi_r^T \xi_r)^{-1} \xi_r^T \mathbf{X}^{(r)}$$

$$\delta_r^T = (\omega_r^T \omega_r)^{-1} \omega_r^T \mathbf{Y}^{(r)}$$

The vectors γ_r and δ_r will form the columns of matrices Γ and Δ . Using this notation, we have

$$\widehat{\mathbf{X}}^{(r)}(\xi_r) = \xi_r \gamma_r^T$$

$$\widehat{\mathbf{Y}}^{(r)}(\omega_r) = \omega_r \delta_r^T$$

6. Subtract the rank-one approximations to obtain remainder matrices:

$$\mathbf{X}^{(r+1)} \leftarrow \mathbf{X}^{(r)} - \widehat{\mathbf{X}}^{(r)}(\xi_r) = \mathbf{X}^{(r)} - \xi_r \gamma_r^T$$

$$\mathbf{Y}^{(r+1)} \leftarrow \mathbf{Y}^{(r)} - \widehat{\mathbf{Y}}^{(r)}(\omega_r) = \mathbf{Y}^{(r)} - \omega_r \delta_r^T$$

7. If

- $\left(\mathbf{X}^{(r+1)}\right)^T \mathbf{Y}^{(r+1)} = \mathbf{0}$, or

- If a decision has been made that the model's dimension should not exceed the current value of r ,

(a) $R \leftarrow r$. This is the rank or dimension of the PLS model.

(b) Halt and exit.

Else continue.

8. $r \leftarrow r + 1$

9. Go to Step 3.

Inner and Outer Loops. For purposes of comparison with Herman Wold's original PLS algorithm and with other PLS variants, we should note that the above algorithm contains two loops, one nested within the other.

- The outer loop begins at Step 3 and ends at Step 9. This loop is indexed by r . The number of times we iterate through this loop is equal to the dimension or rank of the PLS model we are calculating.
- The computation of singular vectors at Step 3 is a concise statement of a procedure which appears as an inner loop in Herman Wold's general algorithm (Step 7 on page 54). As we have noted on page 3.1, this inner loop is simply the power method for computing singular vectors. Thus we may think of the inner loop as being implicit at Step 3 of the PLS-W2A algorithm. Frequently in the literature the inner loop is stated explicitly.

3.5.2 Properties of PLS-W2A

We have decomposed \mathbf{X} and \mathbf{Y} by expressing each as the sum of mutually orthogonal rank-one matrices, plus a residual:

$$\left. \begin{aligned}
 \mathbf{X} &= \widehat{\mathbf{X}}^{(1)}(\xi_1) + \dots + \widehat{\mathbf{X}}^{(R)}(\xi_R) + \mathbf{X}^{(R+1)} \\
 &= \xi_1 \gamma_1^T + \dots + \xi_R \gamma_R^T + \mathbf{X}^{(R+1)} \\
 &= \Xi \Gamma^T + \mathbf{X}^{(R+1)} \\
 \text{and} \\
 \mathbf{Y} &= \widehat{\mathbf{Y}}^{(1)}(\omega_1) + \dots + \widehat{\mathbf{Y}}^{(R)}(\omega_R) + \mathbf{Y}^{(R+1)} \\
 &= \omega_1 \delta_1^T + \dots + \omega_R \delta_R^T + \mathbf{Y}^{(R+1)} \\
 &= \Omega \Delta^T + \mathbf{Y}^{(R+1)}
 \end{aligned} \right\} \quad (3.2)$$

where possibly $\mathbf{X}^{(R+1)} = \mathbf{0}$ or $\mathbf{Y}^{(R+1)} = \mathbf{0}$. The decomposition is unique, provided we agree on a convention such as is stated in Step 3.

The rank-one approximations of \mathbf{X} are mutually orthogonal, and the rank-one approximations of \mathbf{Y} are mutually orthogonal:

$$r \neq s \Rightarrow \left\{ \begin{array}{l} \left(\hat{\mathbf{X}}^{(r)}(\xi_r) \right)^T \hat{\mathbf{X}}^{(s)}(\xi_s) = \gamma_r \xi_r^T \xi_s \gamma_s^T \\ \quad = 0 \\ \text{and} \\ \left(\hat{\mathbf{Y}}^{(r)}(\omega_r) \right)^T \hat{\mathbf{Y}}^{(s)}(\omega_s) = \delta_r \omega_r^T \omega_s \delta_s^T \\ \quad = 0 \end{array} \right\}, \quad (3.3)$$

by the orthogonality of the latent scores, so that $\Xi^T \Xi$ and $\Omega^T \Omega$ are diagonal. A well-known property of linear regression gives us

$$r < s \Rightarrow \left\{ \begin{array}{l} \left(\hat{\mathbf{X}}^{(r)}(\xi_r) \right)^T \mathbf{X}^{(s)} = 0 \\ \left(\hat{\mathbf{Y}}^{(r)}(\omega_r) \right)^T \mathbf{Y}^{(s)} = 0 \end{array} \right\}, \quad (3.4)$$

that is, the rank-one approximations are orthogonal to the residual. In general neither Γ nor Δ is orthogonal, and in general $\Xi^T \Omega$ is not diagonal. This may be seen in the example on pages 45 to 46.

If the algorithm continues long enough, the condition $\left(\mathbf{X}^{(r+1)} \right)^T \mathbf{Y}^{(r+1)} = \mathbf{0}$ in step 7 will eventually be satisfied, so that the cross-product of the residual matrices equals zero:

$$\left(\mathbf{X}^{(R+1)} \right)^T \mathbf{Y}^{(R+1)} = \mathbf{0}. \quad (3.5)$$

This is because the operation by which the matrices are updated at each iteration guarantees that

$$\begin{aligned} \text{rank} \left(\mathbf{X}^{(r+1)} \right) &\leq \text{rank} \left(\mathbf{X}^{(r)} \right) - 1 \quad \text{and} \\ \text{rank} \left(\mathbf{Y}^{(r+1)} \right) &\leq \text{rank} \left(\mathbf{Y}^{(r)} \right) - 1. \end{aligned}$$

The algorithm cannot continue beyond $\min(\text{rank}(\mathbf{X}), \text{rank}(\mathbf{Y}))$ iterations because at that point at least one of $\mathbf{X}^{(r+1)}$ and $\mathbf{Y}^{(r+1)}$ would equal zero. It does not however follow that

either of the residual matrices must be zero. Consider for instance the centered matrices

$$\mathbf{X} = \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}.$$

Here $\mathbf{X}^T \mathbf{Y} = 0$, so that the algorithm has nothing to do. The residual matrices equal \mathbf{X} and \mathbf{Y} .

The number of pairs of vectors of latent scores, R , can exceed $\text{rank}(\mathbf{X}^T \mathbf{Y})$. For instance, let

$$\mathbf{X} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

and

$$\mathbf{Y} = \begin{bmatrix} -3 & -1 \\ 3 & 5 \\ 0 & -2 \end{bmatrix},$$

so that

$$\mathbf{X}^T \mathbf{Y} = \begin{bmatrix} 0 & -2 \\ 0 & 2 \end{bmatrix},$$

of rank one. Although an example could be constructed with centered matrices, for ease of exposition we apply the algorithm to uncentered matrices. After the first iteration, we have

$$\mathbf{X}^{(2)} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix},$$

$$\mathbf{Y}^{(2)} = \begin{bmatrix} -2.4 & 0 \\ 0 & 0 \\ 1.2 & 0 \end{bmatrix},$$

and

$$\left(\mathbf{X}^{(2)}\right)^T \mathbf{Y}^{(2)} = \begin{bmatrix} 1.2 & 0 \\ 1.2 & 0 \end{bmatrix},$$

also of rank one. An additional iteration yields

$$\mathbf{\Gamma} = \begin{bmatrix} 0 & -\frac{1}{\sqrt{2}} \\ -\sqrt{2} & -\frac{1}{\sqrt{2}} \end{bmatrix},$$

$$\mathbf{\Xi} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & 0 \\ 0 & -\sqrt{2} \end{bmatrix}, \text{ so that } \mathbf{\Xi}^T \mathbf{\Xi} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},$$

$$\mathbf{\Delta} = \begin{bmatrix} -0.6 & -1 \\ -1 & 0 \end{bmatrix},$$

$$\mathbf{\Omega} = \begin{bmatrix} 1 & 2.4 \\ -5 & 0 \\ 2 & -1.2 \end{bmatrix}, \text{ so that } \mathbf{\Omega}^T \mathbf{\Omega} = \begin{bmatrix} 30 & 0 \\ 0 & 7.2 \end{bmatrix},$$

and the residual matrices are zero. The reader may check that $\mathbf{\Xi} \mathbf{\Gamma}^T = \mathbf{X}$ and $\mathbf{\Omega} \mathbf{\Delta}^T = \mathbf{Y}$. In spite of the fact that $\text{rank}(\mathbf{X}^T \mathbf{Y}) = 1$ and $\text{rank}(\mathbf{\Xi}^T \mathbf{\Omega}) = 1$, both pairs of latent scores have nonzero covariance:

$$\mathbf{\Xi}^T \mathbf{\Omega} = \begin{bmatrix} 2\sqrt{2} & -1.2\sqrt{2} \\ -2\sqrt{2} & 1.2\sqrt{2} \end{bmatrix}.$$

This example also serves to demonstrate that $\mathbf{\Gamma}^T \mathbf{\Gamma}$ is not necessarily orthogonal, and $\mathbf{\Xi}^T \mathbf{\Omega}$ is not necessarily diagonal.

The latent scores ξ_r and ω_r in PLS-W2A are chosen in an optimal way. We have

$$\left. \begin{aligned} |\text{Cov}(\xi_r, \omega_r)| &= |\text{Cov}(\mathbf{X}^{(r)} \mathbf{u}_r, \mathbf{Y}^{(r)} \mathbf{v}_r)| \\ &= d_r \\ &= \max_{\|\mathbf{a}\|=\|\mathbf{b}\|=1} |\text{Cov}(\mathbf{X}^{(r)} \mathbf{a}, \mathbf{Y}^{(r)} \mathbf{b})|. \end{aligned} \right\} \quad (3.6)$$

This is a well-known fact about the singular value decomposition. For didactic purposes a proof is given on page 69 in Section 3.13.

Proofs and counterexamples for the remaining properties are in Section 3.13.

$$\left. \begin{aligned} (\mathbf{u}_r)_i &= (1/d_r) \left(\mathbf{X}_{\cdot i}^{(r)} \right)^T \boldsymbol{\omega}_r = (N/d_r) \text{Cov} \left(\mathbf{X}_{\cdot i}^{(r)}, \boldsymbol{\omega}_r \right) \\ (\mathbf{v}_r)_j &= (1/d_r) \left(\mathbf{Y}_{\cdot j}^{(r)} \right)^T \boldsymbol{\xi}_r = (N/d_r) \text{Cov} \left(\mathbf{Y}_{\cdot j}^{(r)}, \boldsymbol{\xi}_r \right) \end{aligned} \right\} \quad (3.7)$$

$$\left. \begin{aligned} \mathbf{u}_r &\text{ is an eigenvector of } \left(\mathbf{X}^{(r)} \right)^T \mathbf{Y}^{(r)} \left(\mathbf{Y}^{(r)} \right)^T \mathbf{X}^{(r)} \\ \mathbf{v}_r &\text{ is an eigenvector of } \left(\mathbf{Y}^{(r)} \right)^T \mathbf{X}^{(r)} \left(\mathbf{X}^{(r)} \right)^T \mathbf{Y}^{(r)} \end{aligned} \right\} \quad (3.8)$$

$$\left. \begin{aligned} \boldsymbol{\xi}_r &\text{ is an eigenvector of } \mathbf{X}^{(r)} \left(\mathbf{X}^{(r)} \right)^T \mathbf{Y}^{(r)} \left(\mathbf{Y}^{(r)} \right)^T \\ \boldsymbol{\omega}_r &\text{ is an eigenvector of } \mathbf{Y}^{(r)} \left(\mathbf{Y}^{(r)} \right)^T \mathbf{X}^{(r)} \left(\mathbf{X}^{(r)} \right)^T \end{aligned} \right\} \quad (3.9)$$

$$\left. \begin{aligned} &\text{The } \mathbf{u}_r \text{ are mutually orthogonal and} \\ &\text{the } \mathbf{v}_r \text{ are mutually orthogonal,} \\ &\text{but they are not necessarily singular} \\ &\text{vectors of } \mathbf{X}^T \mathbf{Y}. \end{aligned} \right\} \quad (3.10)$$

$$\left. \begin{aligned} &\text{The } \boldsymbol{\xi}_r \text{ are mutually orthogonal, and hence} \\ &\text{form an orthogonal basis for an } R\text{-dimensional} \\ &\text{subspace of the column space of } \mathbf{X}. \\ &\text{The } \boldsymbol{\omega}_r \text{ are mutually orthogonal, and hence} \\ &\text{form an orthogonal basis for an } R\text{-dimensional} \\ &\text{subspace of the column space of } \mathbf{Y}. \end{aligned} \right\} \quad (3.11)$$

$$\text{For } r < s, \quad \left\{ \begin{aligned} \mathbf{u}_r &\perp \boldsymbol{\gamma}_s, \\ \mathbf{v}_r &\perp \boldsymbol{\delta}_s. \end{aligned} \right\} \quad (3.12)$$

$$\text{For } r < s, \quad \left\{ \begin{aligned} \mathbf{X}^{(s)} \mathbf{u}_r &= \mathbf{0}, \\ \mathbf{Y}^{(s)} \mathbf{v}_r &= \mathbf{0}. \end{aligned} \right\} \quad (3.13)$$

3.6 PLS-SVD

Sampson et al. [SSBB89] and Streissguth et al. [SBSB93a] report an analysis, in the context of behavioral teratology, using a PLS variant which is here called PLS-SVD. In this context the method is used for modeling, not prediction.

Tishler et al. use the same method in the context of industrial management, where the goal again is modeling rather than prediction [TDSL96]. In this paper the method is called “Intercorrelations Analysis” or “Canonical Covariance.” Tishler and Lipovetsky demonstrate that the method can be used for prediction [TL00]. To emphasize differences between this method and CCA, they call the former “Robust Canonical Analysis.”

For $r = 1$ this method is identical to PLS-W2A, but differs for $r > 1$. At the end of the outer loop, when r is incremented, PLS-W2A subtracts rank-one estimates of the data matrices $\mathbf{X}^{(r)}$ and $\mathbf{Y}^{(r)}$ to obtain $\mathbf{X}^{(r+1)}$ and $\mathbf{Y}^{(r+1)}$. From these updates a new cross-product matrix is computed. Sampson, Streissguth et al., on the other hand, do not update \mathbf{X} and \mathbf{Y} . Instead they subtract a rank-one approximation directly from $\mathbf{X}^T \mathbf{Y}$.

For instance, for $r = 2$, the updated value in PLS-SVD is $\mathbf{X}^T \mathbf{Y} - d_1 \mathbf{u}_1 \mathbf{v}_1^T$, equivalently $\sum_{r=2}^R d_r \mathbf{u}_r \mathbf{v}_r^T$. Thus the \mathbf{u}_r and \mathbf{v}_r in PLS-SVD are simply the columns of \mathbf{U} and \mathbf{V} in the singular value decomposition of $\mathbf{X}^T \mathbf{Y}$.

In PLS-SVD the singular value decomposition only needs to be computed once, on the original cross-product matrix $\mathbf{X}^T \mathbf{Y}$. By contrast, in PLS-W2A singular vectors must be computed at each iteration of the outer loop. On the other hand, in PLS-W2A only the first pair of singular vectors needs to be computed at each of these iterations.

After the first iteration of the outer loop, the two algorithms yield different answers. The vectors of latent scores computed by PLS-SVD, ξ^r and ω^r , are not in general orthogonal. In addition the updated version of the cross-product matrix with which each algorithm begins the outer loop differs in the two algorithms for $r > 1$. For instance, for $r = 2$ it is not generally true that $(\mathbf{X}^{(2)})^T \mathbf{Y}^{(2)} = \mathbf{X}^T \mathbf{Y} - d_1 \mathbf{u}_1 \mathbf{v}_1^T$.

3.7 Rank and Orthogonality in PLS-W2A and PLS-SVD

Both PLS-SVD and PLS-W2A can compute $R = \min(p, q)$ pairs of latent variable scores, even when $R > \text{rank}(\mathbf{X}^T \mathbf{Y})$. The latent scores for the \mathbf{X} block form the $N \times R$ matrix Ξ , and the latent scores for the \mathbf{Y} block form the $N \times R$ matrix Ω . The properties of $[\Xi|\Omega]$ differ in the two algorithms, however, in the following way.

- In PLS-SVD, in general neither $\Xi^T \Xi$ nor $\Omega^T \Omega$ is diagonal, but $\Xi^T \Omega$ is diagonal.
- In PLS-W2A, in general $\Xi^T \Omega$ is not diagonal, but $\Xi^T \Xi$ and $\Omega^T \Omega$ are diagonal.

In particular, if $R > \text{rank}(\mathbf{X}^T \mathbf{Y})$, the two algorithms differ as follows.

- In PLS-SVD, the last $R - \text{rank}(\mathbf{X}^T \mathbf{Y})$ pairs of latent scores have zero covariance, so that all of $\mathbf{X}^T \mathbf{Y}$ is accounted for by the first $(\text{rank}(\mathbf{X}^T \mathbf{Y}))$ pairs of latent scores. This follows from the fact that all the latent scores for a given block in PLS-SVD are defined from the singular value decomposition $\mathbf{X}^T \mathbf{Y} = \mathbf{U} \mathbf{D} \mathbf{V}^T$:

$$\Xi \equiv \mathbf{X} \mathbf{U}$$

$$\Omega \equiv \mathbf{Y} \mathbf{V},$$

so that

$$\Xi^T \Omega = \mathbf{D}.$$

- In PLS-W2A, it is possible for all R pairs of latent scores to have nonzero covariance, even though the rank of $\mathbf{X}^T \mathbf{Y}$ is less than R . The example on pages 45 through 46 in Section 3.5.2 demonstrates this.

3.8 Herman Wold's Original PLS Algorithm

We noted in Section 3.3 that Herman Wold originated the term “Partial Least Squares.” His algorithm is more general than PLS-W2A, being used to model linear relationships between an arbitrary number of blocks of variables. In Wold’s setup, rather than \mathbf{X} and \mathbf{Y}

we have a set of blocks of variables, or matrices, \mathbf{X}^a , $a = 1, \dots, A$. These are the observed variables. To each block \mathbf{X}^a is linked a single latent variable ξ^a . The analyst specifies a set of inner relations, that is, linear relations between the ξ^a . When there are more than two blocks ($A > 2$), hence more than two latent variables, many different sets of inner relations are possible, i.e., many different models. The analyst must specify a particular set of inner and outer relations before starting Wold's algorithm.

Once the set of inner and outer relations has been specified, the algorithm is started. It takes as input the data, \mathbf{X}^a , $a = 1, \dots, A$, and the model specification. At convergence (provided it converges) it yields a set of coefficients defining linear relationships. These relationships are of two kinds.

- Linear relationships between latent variables are called **inner relations**.
- Linear relationships between an observed variable \mathbf{X}_j^a and the latent variable for the same block, ξ^a , are called **outer relations**.

The specification of the model for Wold's algorithm consists of two components:

- A set of outer and inner relations. This is determined in advance, as stated above.
- A decision regarding the dimension R of the model. It will be seen that the algorithm contains two loops, one nested within the other. The dimension of the model is determined by the number of times the analyst chooses to iterate through the outer loop.

Note that the specification of the model does not include a *probabilistic* model or a likelihood. In Wold's words, PLS is "distribution-free."

For further information beyond what is presented here, the reader is referred to Wold [Wol85] and [Wol82].

3.8.1 Framework and Notation.

- k indexes the inner loop, r the outer loop.

- \mathbf{X}^a , $a = 1, \dots, A$, are data matrices, or blocks, of dimension $N \times p_a$.
- $\mathbf{X}^{a(r)}$ is the r th remainder matrix of \mathbf{X}^a computed at the end of the $(r - 1)$ st outer loop. $\mathbf{X}^{a(1)} = \mathbf{X}^a$.
- $\xi^a = (\xi_1^a, \dots, \xi_N^a)^T$ are vectors of latent variable scores associated with the blocks.
- \mathbf{Z}^a is a diagonal matrix, defined as follows. For each column $\mathbf{X}_{\cdot j}^a$ of \mathbf{X}^a the analyst decides in advance the sign of $\text{Cor}(\mathbf{X}_{\cdot j}^a, \xi^a)$. These postulated signs form the diagonal of \mathbf{Z}^a .
- The analyst postulates a set of linear relations in the form of a path diagram. An example may be seen in Figure 3.2 on page 57. This is discussed further in Section 3.8.2.
- For any latent variable ξ^a , another latent variable ξ^b is said to be **adjoint** to ξ^a if an edge connects ξ^a and ξ^b . G^a is the set of indices of latent variables adjoint to ξ^a , and $|G^a|$ is its cardinality.
- H^a is the set of all $b \in G^a$ such that the arrow points from ξ^b to ξ^a .
- For each $b \in G^a$, the analyst decides in advance s_{ab} , the sign of $\text{Cor}(\xi^a, \xi^b)$.
- Formulas containing the symbol \approx signify linear models. For instance, the notation $\mathbf{X} \approx \xi^a (\mathbf{v}^a)^T$ is shorthand for the linear model

$$\mathbf{X} = \xi^a (\mathbf{v}^a)^T + \epsilon.$$

Although Wold includes errors in his model statements, neither explicit notation for errors nor explicit distributional assumptions are needed for a statement of his algorithm. For reasons of economy such notation is not introduced here. When models are fit, regression coefficients are estimated using the standard formulas of ordinary least squares.

- Inner relations take the form

$$\xi^a \approx \beta_0^a \mathbf{1} + \sum_{b \in H^a} \beta_b^a \xi^b.$$

For each block a and each iteration r of the outer loop the set of coefficients $\beta_{r0}^a \cup \{\beta_{rb}^a : b \in H^a\}$ specifies the inner relation. Inner relations are signified in the path diagram by arrows between the ξ^a .

- Outer relations take the form

$$\mathbf{X}_{\cdot j}^a \approx \gamma_{j0}^a \mathbf{1} + \gamma_j^a \xi^a.$$

For a given iteration r of the outer loop and for each observed variable $\mathbf{X}_{\cdot j}^{a(r)}$ the pair $(\gamma_{rj0}^a, \gamma_{rj}^a)$ specifies the outer relation. Outer relations are signified in the path diagram by arrows between ξ^a and the variables $\mathbf{X}_{\cdot j}^a$.

“Each indicator is linear in its LV [latent variable]” (Wold [Wol85] page 583), in spite of the fact that, in the path diagram, some arrows point *from* the indicators *to* the latent variables. See, for example, Figure 3.2 on page 57.

- $\mathbf{v}^a = (v_1^a, \dots, v_{p_a}^a)$ is a vector of coefficients such that $\mathbf{X} \approx \xi^a (\mathbf{v}^a)^T$.
- $\widehat{\xi}_r^{a(k)}$ denotes the estimate of ξ^a at the k th iteration of the inner loop and the r th iteration of the outer loop and $\widehat{\xi}_r^a = (\widehat{\xi}_{r1}^a, \dots, \widehat{\xi}_{rN}^a)^T$ the estimate at convergence of the inner loop.
- $\widehat{\mathbf{v}}_r^{a(k)} = (\widehat{v}_{r1}^{a(k)}, \dots, \widehat{v}_{rp_a}^{a(k)})^T$ denotes the estimate of \mathbf{v}^a at the k th iteration of the inner loop and the r th iteration of the outer loop and $\widehat{\mathbf{v}}_r^a = (\widehat{v}_{r1}^a, \dots, \widehat{v}_{rp_a}^a)^T$ the estimate at convergence of the inner loop.

3.8.2 Path Diagram

As stated above, the analyst must specify a model before starting Wold’s algorithm, and this is done by drawing a path diagram. An example of such a path diagram may be found in Figure 3.2 on page 57. This diagram is based on Wold [Wol85].

The three arrowheads pointing at ξ^6 in the ellipse in the lower right-hand corner of the figure signify the following inner relation:

$$\xi^6 \approx \beta_0^6 \mathbf{1} + \beta_4^6 \xi^4 + \beta_3^6 \xi^3 + \beta_5^6 \xi^5.$$

The three arrowheads pointing from ξ^6 to the three indicator variables in the \mathbf{X}^6 block signify the following outer relations:

$$\mathbf{X}_{.1}^6 \approx \gamma_{1,0}^6 \mathbf{1} + \gamma_1^6 \xi^6,$$

$$\mathbf{X}_{.2}^6 \approx \gamma_{2,0}^6 \mathbf{1} + \gamma_2^6 \xi^6,$$

$$\mathbf{X}_{.3}^6 \approx \gamma_{3,0}^6 \mathbf{1} + \gamma_3^6 \xi^6.$$

The rest of the path diagram is interpreted in a similar manner. As stated in Section 1.2, this path diagram does not follow the convention used elsewhere in this paper and specified in Section 1.2. The indicators for a given latent variable may fail to be conditionally independent given that latent variable; nevertheless, the diagram contains no doubleheaded arrows between indicators.

3.8.3 The Algorithm

1. $r \leftarrow 1$.

2. For all a , $\mathbf{X}^{a(1)} \leftarrow \mathbf{X}^a$.

3. Beginning of outer loop:

For all a , center and scale $\mathbf{X}^{a(r)}$.

4. Set $k \leftarrow 0$.

5. Assign arbitrary starting values $\widehat{\mathbf{v}}_r^{a(0)}$. For instance, set $\widehat{\mathbf{v}}_r^{a(0)} \leftarrow \frac{1}{p_a} \mathbf{1}$.

6. $\widehat{\mathbf{v}}_r^{a(0)} \leftarrow \widehat{\mathbf{v}}_r^{a(0)} \mathbf{Z}^a / \|\widehat{\mathbf{v}}_r^{a(0)}\|$ (rescale or normalize).

7. Inner loop.

Estimate $\widehat{\xi}_r^a$ and $\widehat{\mathbf{v}}_r^a$ iteratively, as follows.

Repeat

- (a) $k \leftarrow k + 1$.
- (b) For all a , $\widehat{\xi}_r^{a(k)} \leftarrow \mathbf{X}^{a(r)} \mathbf{Z}^a \widehat{\mathbf{v}}_r^{a(k-1)}$. (Recall that \mathbf{Z}^a is defined on page 51.)
- (c) For all a , compute the $N \times 1$ sign-weighted sum $(\omega^a)^{(k)} \leftarrow \sum_{b \in G^a} s_{ab} \widehat{\xi}_r^{b(k)}$.
- (d) For all a , estimate $\widehat{\mathbf{v}}_r^{a(k)}$ by one of the following modes.
 - Mode A: Compute $\widehat{\mathbf{v}}_r^{a(k)}$ by fitting p_a simple linear models. That is, for $j = 1, \dots, p_a$, fit the model $\mathbf{X}_{\cdot j}^{a(r)} \approx v_{rj}^a (\omega^a)^{(k)}$.
 - Mode B: Compute $\widehat{\mathbf{v}}_r^{a(k)}$ by performing multiple regression. That is, fit the linear model $(\omega^a)^{(k)} \approx \mathbf{X}^{a(r)} \mathbf{v}_r^a$.
- (e) $\widehat{\mathbf{v}}_r^{a(k)} \leftarrow \widehat{\mathbf{v}}_r^{a(k)} / \|\widehat{\mathbf{v}}_r^{a(k)}\|$.

until for all a , $\widehat{\mathbf{v}}_r^{a(k)}$ has converged.

8. Estimate the inner relations, up to intercept, by regressing ξ_r^a on the latent variables which have arrows pointing to ξ^a . That is, for each a fit the linear model

$$\widehat{\xi}_r^a \approx \sum_{b \in H^a} \beta_b^a \widehat{\xi}_r^b.$$

9. Estimate the outer relations, up to intercept, by regressing each column of $\mathbf{X}^{a(r)}$ on the latent variable for this block. That is, for each $\mathbf{X}_{\cdot j}^{a(r)}$ fit the simple linear model

$$\mathbf{X}_{\cdot j}^{a(r)} \approx \gamma_j^a \widehat{\xi}_r^a.$$

As noted on page 52, in the outer relations the indicators are seen as functions of the latent variables, not vice versa, regardless of the direction the arrows point in the path diagram.

10. Estimate the intercepts. For each a ,

- (a) $\overline{\mathbf{X}^{a(r)}} \leftarrow \frac{1}{N} \left(\mathbf{X}^{a(r)} \right)^T \mathbf{1}$, a $p_a \times 1$ vector of columnwise means.
- (b) $\overline{\xi_r^a} \leftarrow \overline{\mathbf{X}^{a(r)}}^T \widehat{\mathbf{v}}_r^a$, a scalar.
- (c) $\beta_{r0}^a \leftarrow \overline{\xi_r^a} - \sum_{b \in H^a} \beta_{rb}^a \overline{\xi_r^b}$, the scalar intercept for the inner relation.
- (d) $\gamma_{r0}^a \equiv (\gamma_{r10}^a, \dots, \gamma_{rp_a0}^a)^T \leftarrow \overline{\mathbf{X}^{a(r)}} - \gamma_r^a \overline{\xi_r^a}$, the vector of intercepts for the outer relations for the a th block.

11. At this point the analyst makes a decision, or acts on a decision already made, regarding the dimension R of the model. Recall the discussion of model specification on page 50. If r , the number of times we have iterated through the outer loop, does not yet equal the model dimension on which the analyst has decided, we do the following:

For each a update $\mathbf{X}^{a(r)}$ to equal the residuals of the outer relations:

$$\begin{aligned} \mathbf{X}^{a(r+1)} &\leftarrow \mathbf{X}^{a(r)} - \left(\mathbf{1} (\gamma_{r0}^a)^T + \widehat{\xi}_r^a (\gamma_r^a)^T \right) \\ r &\leftarrow r + 1 \end{aligned}$$

and return to the beginning of the outer loop at Step 3.

3.8.4 Raw Data Versus Product Data.

The blocks \mathbf{X}^a are, in Wold's terminology, "the **raw data** of the model. The **product data** are the means and the product moments ..." ([Wol85] page 584). Wold states that the algorithm can be run with product data input and will yield parameter estimates that are the same, up to rounding error, as when raw data are used. Without raw data, of course, we cannot obtain estimates $\widehat{\xi}_r^a$ of the latent variable case values, and we cannot update $\mathbf{X}^{a(r)}$ in Step 11. It is possible, however, to run the inner loop on product data alone, and to compute the $\widehat{\xi}_r^a$ after convergence of the inner loop.

PLS-W2A, described in Section 3.5, uses product data input. Section 3.8.5 demonstrates equivalence of the raw data and product data approaches for the two-block case.

3.8.5 PLS-W2A is a Special Case of Wold

Under certain conditions the inner loop of Wold's general algorithm yields the same coefficient values as the "singular value decomposition" step of PLS-W2A (step 3 on page 41). These conditions are:

- There are two blocks ($A = 2$).
- Estimation is performed in Mode A as in line (7d).

This is true for any iteration r of the outer loop and for any $\mathbf{X}^{1(r)}$ and $\mathbf{X}^{2(r)}$. Thus PLS-W2A is a special case of Wold's general algorithm. When we prove this, we will obtain as a corollary the fact that, in the two-block Mode A case, the raw-data approach yields identical values for saliences and latent scores as the product-data approach.

To see the result, let us first translate Wold's notation into notation for the two-block case, compatible with our statement of PLS in Section 3.5.1. A table of equivalent notation may be found in Table 3.1.

A path diagram such as could be used for the two-block Mode A case may be found in Figure 3.3 on page 57.

It should be noted that the direction of the arrows between the ξ variables and the ω variables in a path diagram for the two-block Mode A case has no influence on the coefficients computed by PLS-W2A. The directions of the arrows in a path diagram for Wold's general algorithm influence the kind of inner relations which are computed at the end of each iteration of the outer loop. In the case of PLS-W2A, or any two-block PLS method, the sole question would be whether we regress ω^r on ξ^r or vice versa. The explicit specification and estimation of inner relations is not part of PLS-W2A, however.

At initialization (step 5), we have

$$\begin{aligned}\hat{\mathbf{u}}_r^{(0)} &\leftarrow \frac{1}{I} \mathbf{1} \\ \hat{\mathbf{v}}_r^{(0)} &\leftarrow \frac{1}{J} \mathbf{1}.\end{aligned}$$

At step (7b) we have

$$\xi_r^{(k)} \leftarrow \mathbf{X}^{(r)} \mathbf{u}_r^{(k-1)}$$

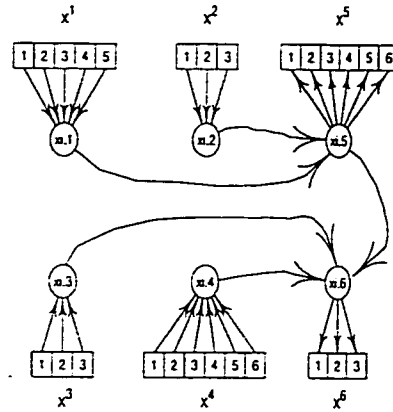


Figure 3.2: Example of a diagram by which the analyst might postulate inner and outer relations in specifying a model in the context of Herman Wold's original PLS algorithm. This is discussed in Section 3.8.2 on page 52.

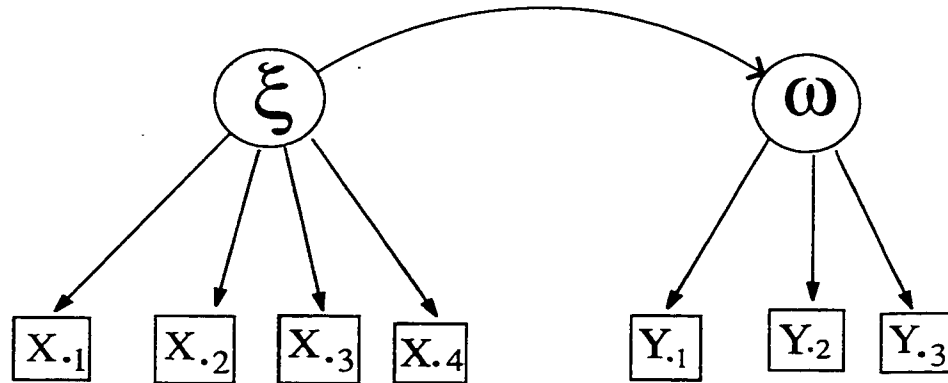


Figure 3.3: A path diagram for a two-block PLS model. As noted in Section 3.8.5, when Wold's general PLS algorithm is applied in Mode A to the two-block case, the coefficients are identical to those computed by PLS-W2A. As noted on page 56, the direction of the arrows between ξ and ω has no effect on the values of the coefficients computed by PLS-W2A.

Table 3.1: Translation of the notation of Herman Wold's general algorithm, described in Section 3.8, into two-block notation compatible with PLS as described in 3.5.1.

General	Two Blocks
p_1	I
p_2	J
\mathbf{X}^1	\mathbf{X}
\mathbf{X}^2	\mathbf{Y}
$\xi^1, \widehat{\xi}_r^{(k)}$	$\xi, \widehat{\xi}_r^{(k)}$
$\xi^2, \widehat{\xi}_r^{(k)}$	$\omega, \widehat{\omega}_r^{(k)}$
\mathbf{v}^1	\mathbf{u}
\mathbf{v}^2	\mathbf{v}
\mathbf{Z}^a	\mathbf{I}
G^1 , the latent variables adjoint to ξ^1	ω , the only latent variable besides ξ
s_{12}, s_{21}	1 (that is, we postulate $d > 0$)
$(\omega^1)^{(k)}$	$\widehat{\omega}_r^{(k)}$
G^2 , the latent variables adjoint to ξ^2	ξ , the only latent variable besides ω
$(\omega^2)^{(k)}$	$\widehat{\xi}_r^{(k)}$

$$\omega_r^{(k)} \leftarrow \mathbf{Y}^{(r)} \mathbf{v}_r^{(k-1)}.$$

At step (7d), for each column $\mathbf{X}_j^{(r)}$ of $\mathbf{X}^{(r)}$ we compute the regression coefficient of $\mathbf{X}_j^{(r)}$ on $\omega_r^{(k)}$, and we similarly regress the columns of $\mathbf{Y}^{(r)}$ on $\xi_r^{(k)}$. But since the columns of $\mathbf{X}^{(r)}$ and of $\mathbf{Y}^{(r)}$ have been centered, the regression coefficients are proportional to inner products. That is,

$$\begin{aligned}\widehat{\mathbf{u}}_r^{(k)} &\propto \left(\mathbf{X}^{(r)}\right)^T \widehat{\omega}_r^{(k)} \\ \widehat{\mathbf{v}}_r^{(k)} &\propto \left(\mathbf{Y}^{(r)}\right)^T \widehat{\xi}_r^{(k)}.\end{aligned}$$

But this means

$$\begin{aligned}\widehat{\mathbf{u}}_r^{(k)} &\propto \left(\mathbf{X}^{(r)}\right)^T \widehat{\omega}_r^{(k)} \\ &\propto \left(\mathbf{X}^{(r)}\right)^T \mathbf{Y}^{(r)} \widehat{\mathbf{v}}_r^{(k-1)} \\ &\propto \left(\mathbf{X}^{(r)}\right)^T \mathbf{Y}^{(r)} \left(\mathbf{Y}^{(r)}\right)^T \widehat{\xi}_r^{(k-2)} \\ &\propto \left(\mathbf{X}^{(r)}\right)^T \mathbf{Y}^{(r)} \left(\mathbf{Y}^{(r)}\right)^T \mathbf{X}^{(r)} \widehat{\mathbf{u}}_r^{(k-2)}.\end{aligned}$$

So at convergence—provided the inner loop does converge—we have

$$\widehat{\mathbf{u}}_r \propto \left(\mathbf{X}^{(r)}\right)^T \mathbf{Y}^{(r)} \left(\mathbf{Y}^{(r)}\right)^T \mathbf{X}^{(r)} \widehat{\mathbf{u}}_r.$$

Thus $\widehat{\mathbf{u}}_r$ is a left singular vector of $\left(\mathbf{X}^{(r)}\right)^T \mathbf{Y}^{(r)}$, and similarly $\widehat{\mathbf{v}}_r$ is a right singular vector. Conditions for convergence are given in Lemma 3.12.1 in Section 3.12.1.

3.9 PLS2 and PLS1

The variants of PLS that appear most frequently in the chemometric literature are called PLS2 and PLS1, and differ from PLS-W2A in a way that shall be seen shortly. My sources for PLS2 are Höskuldsson [H88] and Holcomb et al. [HHMT97]. PLS1 is the case of PLS2 where \mathbf{Y} consists of a single column (Geladi [Gel88] page 237), but has been examined in its own right.

Like PLS-SVD, PLS2 differs from PLS-W2A only when r is incremented. PLS-W2A and PLS-SVD are symmetric in the blocks. In PLS-W2A both blocks are updated by the

subtraction of a rank-one estimate based on what Wold calls the outer relation and what Höskuldsson calls the loadings. In PLS-SVD the cross-product matrix $\mathbf{X}^T \mathbf{Y}$ is updated. In PLS2, on the other hand, \mathbf{Y} is updated by subtracting an estimate based on ξ ; the latent variable score estimate for the \mathbf{X} block, and the inner relation between ξ and ω . That is, the following additional step is inserted:

$$\beta_r \leftarrow (\xi_r^T \xi_r)^{-1} \xi_r^T \omega_r$$

and $\mathbf{Y}^{(r)}$ is updated as follows:

$$\mathbf{Y}^{(r+1)} \leftarrow \mathbf{Y}^{(r)} - \beta_r \xi_r \omega_r^T.$$

This difference is crucial, and is most easily seen in PLS1, the case where \mathbf{Y} consists of a single column. For any PLS algorithm, let R be the maximum number of times that we could iterate through the outer loop, disregarding any consideration of model selection in the choice of R . Now consider the case where \mathbf{Y} consists of a single column. In PLS-W2A or in PLS-SVD we compute once the $I \times 1$ vector of inner products $\mathbf{X}^T \mathbf{Y}$ and have nothing more to do. For instance, in PLS-W2A, the space spanned by \mathbf{Y} is a line through the origin in \mathbb{R}^J . A rank-one approximation of \mathbf{Y} is equal to \mathbf{Y} ; subtract it and you have nothing left for the second iteration, for $r = 2$. In PLS1, on the other hand, R is limited only by the rank of $\mathbf{X}^T \mathbf{X}$.

PLS1, the case of PLS2 where \mathbf{Y} is a single column, is a regularization technique, in the same class as ridge regression and principal component regression. At each iteration of the outer loop (each value of r), PLS1 finds a direction in the space of regression coefficients orthogonal to the previous regression coefficients. Regression here is of the single response variable \mathbf{Y} on the predictor variables \mathbf{X} . An analyst will typically use a regularization technique when the problem is inherently asymmetric: There is a large, possibly redundant set of predictor (or carrier or “independent”) variables from which a single response (or dependent) variable is to be predicted. For more information on PLS1, see Sardy [Sar98], Frank and Friedman [FF93], and Helland [Hel88]. For the relationship between PLS1 and other regularization techniques, see Sardy.

3.10 PLS-W2A, PLS-SVD, PLS2: Summary of differences and similarities

We have seen that PLS-W2A, PLS-SVD and PLS2 differ from each other when the outer loops exceeds one iteration, but that in the first iteration they are equivalent.

The equivalence is seen in the fact that the first iteration of the outer loop in each algorithm involves the computation of the first singular vectors of $\mathbf{X}^T \mathbf{Y}$; that is, the singular vectors corresponding to the greatest singular value.

The difference between PLS-W2A and PLS-SVD is discussed in Section 3.7. In addition, recall that \mathbf{u}_r and \mathbf{v}_r in PLS-SVD are by definition singular vectors of $\mathbf{X}^T \mathbf{Y}$. In PLS-W2A, on the other hand, this is not true in general. This is stated as property (3.10) on page 47, and proved on page 73. Instead \mathbf{u}_r and \mathbf{v}_r are the singular vectors corresponding to the greatest singular value of $(\mathbf{X}^{(r)})^T \mathbf{Y}^{(r)}$.

The difference between PLS2 on the one hand and PLS-W2A and PLS-SVD on the other is the difference between symmetric and asymmetric treatments, and is discussed in Section 3.9.

3.11 Canonical Correlation Analysis (CCA), alias Mode B PLS

The difference between Mode A and Mode B PLS lies in the way the coefficients are updated which relate an indicator variable to its latent variable. (In Wold's general algorithm this is Step 7d on page 54.) In Mode A the coefficients are computed by a set of simple linear models, one for each coefficient. In Mode B they are computed by one multiple regression model for each block of indicators.

Two-block Mode B PLS is equivalent to Canonical Correlation Analysis (CCA). CCA, however, has been known since 1936, when Harold Hotelling published the seminal paper [Hot36], long before Herman Wold originated PLS (1975 [Wol75]). For this reason CCA will be called by its customary name, and the term "PLS," unless otherwise specified, will be used to refer to Mode A.

CCA, like PLS-W2A, is a method for analyzing association between two blocks of variables. In addition to being older than PLS it is better known. Both PLS-W2A and CCA compute orthogonal sets of latent scores which are linear combinations of the \mathbf{X} and \mathbf{Y}

variables. In both methods, the latent scores come in pairs—one for \mathbf{X} , one for \mathbf{Y} —and these pairs maximize a criterion which measures association.

Although the two methods may appear superficially to be similar, they differ fundamentally, both in their numerical properties and in the interpretation of the coefficients which they yield.

PLS coefficients are computed by the singular value decomposition, which is known to be numerically stable. Stability is not affected by the relationship between the number of variables ($I + J$) and the number of observations (N). CCA, on the other hand, involves the computation of two inverses, that of $\mathbf{X}^T \mathbf{X}$ and that of $\mathbf{Y}^T \mathbf{Y}$. Thus when the number of variables exceeds the number of observations the canonical correlation coefficients are not uniquely defined. Similarly, we run into problems if $\mathbf{X}^T \mathbf{X}$ or $\mathbf{Y}^T \mathbf{Y}$ is ill-conditioned, even if $I + J < N$.

We may work around the numerical difficulties of CCA by applying a penalty to $\mathbf{X}^T \mathbf{X}$ and $\mathbf{Y}^T \mathbf{Y}$, for instance a ridge penalty (Vinod 1976 [Vin76]). This however does not alter the fact that CCA coefficients are not interpretable in the same way as PLS saliences. A PLS salience for a given indicator variable—say, u_i , the salience for $\mathbf{X}_{\cdot i}$ —is proportional to the sample covariance of the indicator variable and the computed vector of latent scores for the other block:

$$u_i \propto \text{Cov}(\mathbf{X}_{\cdot i}, \boldsymbol{\omega}).$$

If another indicator, say $\mathbf{X}_{\cdot(i+1)}$, is removed from or added to the analysis, this change has little effect on u_i , provided the model has been correctly specified. This property, in fact, can be used to test whether the model has been incorrectly specified (see page 26).

CCA coefficients, on the other hand, are analogous to multiple regression coefficients. In the case where \mathbf{Y} consists of a single column, they are identical to multiple regression coefficients. The coefficient for $\mathbf{X}_{\cdot i}$, say u_i , means something different each time another indicator, say $\mathbf{X}_{\cdot(i+1)}$, is added to or removed from the analysis.

Although CCA is older than PLS, and better known among statisticians, inference about CCA coefficients and latent scores is not much easier than inference about the analogous values computed by PLS. “[T]he distribution theory of canonical correlations and canonical

vectors is complicated even in the null case . . . The important case from the practical point of view is the non-null case, and the distribution theory associated with it is almost intractable” (Kshirsagar 1972 [Ksh72], page 278). In particular, the CCA coefficients are not known to maximize a likelihood.

Kettenring [Ket82] gives an introduction to canonical analysis, and an example of its application to a dataset. Although Kettenring’s article is an exposition of canonical analysis, not an argument against its use, his examples demonstrate both the difficulty in interpreting the coefficients, and their instability. For instance, he repeats an analysis several times using “various combinations of” the indicator variables, “[t]o gain insight into the relative importance of the variables” (page 361). By contrast, PLS saliences themselves measure the relative importance of the variables, and as stated above, the salience of a variable remains nearly constant when other variables in the same block are added or removed, provided the model is correctly specified.

Whereas Kettenring’s article focuses on canonical analysis, Sampson et al. [SSBB89] apply both PLS and canonical analysis to the problem in behavioral teratology introduced in Section 1.1.

The algorithm used by Sampson et al., PLS-SVD, differs from PLS-W2A, as discussed in Section 3.6. Nevertheless the differences which Sampson et al. point out between canonical analysis and PLS exist regardless of whether we use PLS-SVD or PLS-W2A. PLS gives results that are clearer and easier to interpret.

In addition, Sampson et al. give a qualitative argument why PLS is preferable in their particular application (pages 481f):

Canonical correlation may be explained in terms of multiple regression: canonical variable coefficients α_i and β_j would be computed as *multiple* regression coefficients rather than the *simple* regression coefficients . . . in the PLS solution. However, the idea of multiple regression is inappropriate here. We should not compute the regression of any IQ item, or any weighted combination of items, on the alcohol variables taken as thirteen separate predictors. This is because we cannot imagine the partial effect on an IQ subtest of changing one alcohol

variable while holding constant the values of all other alcohol scores. The alcohol scores covary jointly as different aspects of a single underlying exposure scale ...and so we cannot hold other alcohol scores unchanged when we vary one of them.

The differences between PLS-W2A and CCA are listed in greater detail in Table 3.2 on page 65.

CCA is a powerful tool, and some of its numerical shortcomings can be overcome, for instance by the use of penalties on the sample dispersion matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{Y}^T\mathbf{Y}$. In some cases, both PLS and CCA might be applied to the same problem. The question whether a penalized version of CCA would perform better than PLS in some cases is open.

Methodological research must begin, however, with a clear awareness of the fundamental differences between the two methods. As for data analysis, the researcher would do well to follow the example of Sampson et al., thinking carefully about the scientific question which is to be answered, and then choose the method which best fits the situation.

3.12 The Power Method in PLS

We have seen that Wold's general algorithm contains an inner loop. We have seen that in the two-block, Mode A case, this inner loop reduces to the extraction of the first pair of singular vectors of the cross-product matrix $\mathbf{X}^T\mathbf{Y}$ —the pair which correspond to the largest singular value.

It turns out that all the PLS algorithms discussed in this work contain an inner loop, either implicit or explicit. This inner loop is implicit, for instance, in the statement of the PLS-W2A algorithm in Section 3.5.1. The inner loop is often stated explicitly in the literature, however. For instance, Sampson et al. [SSBB89] and Streissguth et al. [SBSB93a] state the inner loop explicitly, as does Höskuldsson [H88].

In this section we take a closer look at the explicit form of the inner loop in the two-block Mode A case. This is well-known as the power method for computing eigenvectors (Stewart [Ste73], page 340). The power method is stated, and a lemma giving conditions under which the power method is guaranteed to converge.

Table 3.2: Contrast between Canonical Two-Block Mode A Partial Least Squares (PLS) and two-block Mode B PLS, which is better known as Canonical Correlation Analysis (CCA). The index r , used in the discussion of PLS algorithms in Sections 3.5 and 3.3, is left out of the table. The contrast holds for any value of r .

	PLS-W2A	CCA
Objective	$ \text{Cov}(\mathbf{Xu}, \mathbf{Yv}) $	$ \text{Cor}(\mathbf{Xu}, \mathbf{Yv}) $
Feasible set	$\ \mathbf{u}\ = \ \mathbf{v}\ = 1$	all \mathbf{u}, \mathbf{v}
Solution is eigenvector of	$\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}$ $\mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Y}$	$(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{X}$ $(\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$
Objective sensitive to scale of columns of \mathbf{X}, \mathbf{Y} ?	yes	no
If \mathbf{X} or \mathbf{Y} are nearly collinear	stable	unstable
Interpretation of \mathbf{u} and \mathbf{v}	$u_i \propto \text{Cov}(\mathbf{X}_{\cdot i}, \mathbf{Yv})$ $v_j \propto \text{Cov}(\mathbf{Y}_{\cdot j}, \mathbf{Xu})$	No straightforward interpretation
If \mathbf{Y} is a single column	$u_i \propto$ simple regression coefficient	$u_i \propto$ multiple regression coefficient
If $p > n$ or $q > n$	\mathbf{u} and \mathbf{v} remain uniquely defined	\mathbf{u} and \mathbf{v} no longer uniquely defined
Effect of adding $\mathbf{X}_{\cdot(i+1)}$ nearly collinear with $\mathbf{X}_{\cdot i}$	u_i changes little	u_i can change a great deal
Easiest to interpret when columns of a block ...	all measure the same underlying latent variable	are orthogonal

Although the inner loop of PLS-W2A requires only the greatest singular value and the associated pair of singular vectors, for the sake of completeness the algorithm stated here yields the entire singular value decomposition. One reason for the inclusion of the more general algorithm is that PLS-SVD, described in Section 3.6, uses the entire singular value decomposition.

The lemma is proved for the case where $\mathbf{C} \equiv \mathbf{X}^T \mathbf{Y}$ has distinct eigenvalues.

The algorithm as applied to a single matrix \mathbf{C} corresponds to the “product-data-input” version of PLS, mentioned in Section 3.8.4 on page 55. First we shall state this version, and give a proof of convergence. Then we shall see the raw-data version. The proof that the raw-data algorithm converges is a corollary of the proof for product data.

3.12.1 The Algorithm

Superscripts on vectors (\mathbf{u}^k and \mathbf{v}^k) are indices of iteration in the inner loop (1d), whereas subscripts (\mathbf{u}_r and \mathbf{v}_r) are indices of the outer loop, and denote columns of \mathbf{U} and of \mathbf{V} . Superscripts on scalars (d^k , etc.) indicate exponents, following conventional notation.

Let \mathbf{C} be an $I \times J$ matrix. Set $r \leftarrow 0$.

1. Repeat

(a) Set $r \leftarrow r + 1$.

(b) Choose $\mathbf{u}^0 \in \mathbb{R}^I$.

(c) Set $k \leftarrow 0$.

(d) Repeat

$$k \leftarrow k + 1$$

$$\mathbf{v}^k \leftarrow \mathbf{C}^T \mathbf{u}^{k-1}$$

$$\mathbf{v}^k \leftarrow \mathbf{v}^k / \|\mathbf{v}^k\|$$

$$\mathbf{u}^k \leftarrow \mathbf{C} \mathbf{v}^k$$

$$\mathbf{u}^k \leftarrow \mathbf{u}^k / \|\mathbf{u}^k\|$$

until $\|\mathbf{u}^k - \mathbf{u}^{k-1}\|$ is less than some convergence criterion.

(e) Save

$$\begin{aligned} \mathbf{u}_r &\leftarrow \mathbf{u}^k \\ \mathbf{v}_r &\leftarrow \mathbf{v}^k \\ d_r &\leftarrow (\mathbf{u}_r)^T \mathbf{C} \mathbf{v}_r. \end{aligned}$$

(f) Set $\mathbf{C} \leftarrow \mathbf{C} - d_r \mathbf{u}_r \mathbf{v}_r^T$.

until $\|\mathbf{C}\|$ is less than some criterion.

2. Reorder the d_r so that

$$d_1 > d_2 > \dots > d_R,$$

and reorder the \mathbf{u}_r and \mathbf{v}_r accordingly.

3.12.2 Conditions for Convergence

Lemma 3.12.1 *Let \mathbf{C} have distinct singular values. If at step (1b) \mathbf{u}^0 is chosen so as not to be in the left nullspace of \mathbf{C} , the algorithm stated in Section 3.12.1 yields the singular value decomposition. If \mathbf{u}^0 is chosen to equal $\frac{1}{J}\mathbf{C}\mathbf{1}$, the mean of the columns of \mathbf{C} , then the singular vectors computed by the algorithm will be computed in the descending order of their corresponding singular values. Thus step (2) will be unnecessary.*

Proof. See Section 3.14.

3.12.3 Raw Data Input

1. Initialize $\mathbf{u}^0 \in \mathbb{R}^I$ and $\mathbf{v}^0 \in \mathbb{R}^J$ by

$$\begin{aligned} \mathbf{u}^0 &\leftarrow \frac{1}{I}\mathbf{1}, \\ \mathbf{v}^0 &\leftarrow \frac{1}{J}\mathbf{1}. \end{aligned}$$

2. Repeat until convergence:

(a) $k \leftarrow k + 1$

(b) Update latent scores.

$$\begin{aligned}\xi^k &\leftarrow \mathbf{X}\mathbf{u}^{k-1} \\ \omega^k &\leftarrow \mathbf{Y}\mathbf{v}^{k-1}\end{aligned}$$

(c) Update saliences.

$$\begin{aligned}\mathbf{u}^k &\leftarrow \mathbf{X}^T \omega^k \left((\omega^k)^T \omega^k \right)^{-1} \\ \mathbf{u}^k &\leftarrow \mathbf{u}^k / \|\mathbf{u}^k\| \\ \mathbf{v}^k &\leftarrow \mathbf{Y}^T \xi^k \left((\xi^k)^T \xi^k \right)^{-1} \\ \mathbf{v}^k &\leftarrow \mathbf{v}^k / \|\mathbf{v}^k\|.\end{aligned}$$

Proof of convergence is in 3.14.

3.13 Proofs of PLS-W2A Properties

These properties were stated in Section 3.5.2.

For the proof of Property (3.7) we dispense with the r index. Recall that at any iteration of the outer loop (any value of r), \mathbf{u}_r and \mathbf{v}_r are the first pair of singular vectors of $(\mathbf{X}^{(r)})^T \mathbf{Y}^{(r)}$.

$$\begin{aligned}\begin{bmatrix} \mathbf{X}_1^T \omega \\ \vdots \\ \mathbf{X}_I^T \omega \end{bmatrix} &= \begin{bmatrix} \mathbf{X}_1^T \mathbf{Y} \mathbf{v} \\ \vdots \\ \mathbf{X}_I^T \mathbf{Y} \mathbf{v} \end{bmatrix} \\ &= \mathbf{X}^T \mathbf{Y} \mathbf{v} \\ &= \mathbf{U} \mathbf{D} \mathbf{V}^T \mathbf{v} \text{ by the SVD} \\ &= d_1 \mathbf{u}_1.\end{aligned}$$

Thus

$$u_i = (1/d) \mathbf{X}_i^T \omega.$$

The result for v_j holds by symmetry. This proves Property (3.7).

Property (3.8) follows from the definition of the singular value decomposition.

$$\begin{aligned}
 \mathbf{X}^{(r)} \left(\mathbf{X}^{(r)} \right)^T \mathbf{Y}^{(r)} \left(\mathbf{Y}^{(r)} \right)^T \boldsymbol{\xi}_r &= \mathbf{X}^{(r)} \left(\mathbf{X}^{(r)} \right)^T \mathbf{Y}^{(r)} \left(\mathbf{Y}^{(r)} \right)^T \mathbf{X}^{(r)} \mathbf{u}_r \\
 &\propto \mathbf{X}^{(r)} \mathbf{u}_r && \text{by (3.8)} \\
 &= \boldsymbol{\xi}_r && \text{by definition,}
 \end{aligned}$$

which proves the first half of (3.9). The second follows by symmetry.

The proof of Property (3.6) on page 46 follows from the fact that \mathbf{u}_r and \mathbf{v}_r are the first pair of singular vectors of $\left(\mathbf{X}^{(r)} \right)^T \mathbf{Y}^{(r)}$. The property holds for any conformable \mathbf{X} , \mathbf{Y} , and their first pair of singular vectors. We drop the r subscript in this proof, since it is a distraction.

$$\begin{aligned}
 (n-1)\text{Cov}(\mathbf{X}\mathbf{a}, \mathbf{Y}\mathbf{b}) &= \mathbf{a}^T \mathbf{X}^T \mathbf{Y} \mathbf{b} \\
 &= \mathbf{a}^T \mathbf{U} \mathbf{D} \mathbf{V}^T \mathbf{b} && \text{by the singular value decomposition} \\
 &= \boldsymbol{\alpha}^T \mathbf{D} \boldsymbol{\beta} \\
 &= \sum_{i=1}^R d_i \alpha_i \beta_i && (3.14)
 \end{aligned}$$

where R is the number of nonzero singular values, and where

$$\begin{aligned}
 (\alpha_1, \dots, \alpha_R) &= \boldsymbol{\alpha}^T \\
 &= \mathbf{a}^T \mathbf{U}, \text{ and} \\
 (\beta_1, \dots, \beta_R)^T &= \boldsymbol{\beta} \\
 &= \mathbf{V}^T \mathbf{b}.
 \end{aligned}$$

Note that

$$\begin{aligned}
 \|\boldsymbol{\alpha}\| &= \boldsymbol{\alpha}^T \boldsymbol{\alpha} \\
 &= \mathbf{a}^T \mathbf{U}^T \mathbf{U} \mathbf{a} \\
 &= \mathbf{a}^T \mathbf{a} && \text{by the orthogonality of } \mathbf{U} \\
 &= 1, \text{ and similarly} && (3.15) \\
 \|\boldsymbol{\beta}\| &= 1.
 \end{aligned}$$

By (3.14),

$$\begin{aligned}
 (n-1) |\mathbf{Cov}(\mathbf{Xa}, \mathbf{Yb})| &= \left| \sum_{i=1}^R d_i \alpha_i \beta_i \right| \\
 &\leq \left(\sum_{i=1}^R d_i \alpha_i^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^R d_i \beta_i^2 \right)^{\frac{1}{2}} \quad (3.16)
 \end{aligned}$$

by Cauchy-Schwarz. We can maximize each term of (3.16) separately, in α and β respectively, noting that their domains are compact. Since $d_1 \geq d_2 \geq \dots d_R > 0$,

$$\begin{aligned}
 \sum_{i=1}^R d_i \alpha_i^2 &\leq d_1 \sum_{i=1}^R \alpha_i^2 \\
 &= d_1 \quad \text{by (3.15), and similarly} \\
 \sum_{i=1}^R d_i \beta_i^2 &\leq d_1.
 \end{aligned}$$

Thus

$$(n-1) |\mathbf{Cov}(\mathbf{Xa}, \mathbf{Yb})| \leq d_1.$$

This value is attained when \mathbf{a} and \mathbf{b} are the first left and right singular vectors of $\mathbf{X}^T \mathbf{Y}$.

This completes the proof of Property 3.6.

The proofs of Properties 3.10 through 3.12 are adapted from Höskuldsson [H88].

$$\begin{aligned}
 \mathbf{X}^{(r+1)} &= \mathbf{X}^{(r)} - \xi_r (\xi_r^T \xi_r)^{-1} \xi_r^T \mathbf{X}^{(r)} \\
 &= \mathbf{A}^{(r)} \mathbf{X}^{(r)}, \quad \text{where} \\
 \mathbf{A}^{(r)} &= \mathbf{I} - \xi_r (\xi_r^T \xi_r)^{-1} \xi_r^T.
 \end{aligned} \quad (3.17)$$

We may then write

$$\begin{aligned}
 \mathbf{X}^{(s)} &= \mathbf{A}^{(s-1)} \mathbf{X}^{(s-1)} \\
 &= \mathbf{A}^{(s-1)} \mathbf{A}^{(s-2)} \mathbf{X}^{(s-2)}
 \end{aligned}$$

and in general for $1 \leq r < s$,

$$\begin{aligned}
 \mathbf{X}^{(s)} &= \mathbf{A}^{(s-1)} \dots \mathbf{A}^{(r+1)} \mathbf{A}^{(r)} \mathbf{X}^{(r)} \\
 &= \mathbf{ZA}^{(r)} \mathbf{X}^{(r)}
 \end{aligned} \quad (3.18)$$

for some matrix \mathbf{Z} .

Side calculation:

$$\begin{aligned}
 \mathbf{A}^{(r)}\mathbf{X}^{(r)}\mathbf{u}_r &= 0, \text{ since} \\
 \mathbf{A}^{(r)}\mathbf{X}^{(r)}\mathbf{u}_r &= \mathbf{A}^{(r)}\boldsymbol{\xi}_r \quad \text{by definition of } \boldsymbol{\xi}_r \\
 &= \boldsymbol{\xi}_r - \boldsymbol{\xi}_r (\boldsymbol{\xi}_r^T \boldsymbol{\xi}_r)^{-1} \boldsymbol{\xi}_r^T \boldsymbol{\xi}_r \\
 &= \boldsymbol{\xi}_r - \boldsymbol{\xi}_r.
 \end{aligned} \tag{3.19}$$

Putting (3.18) and (3.19) together, we obtain

$$\begin{aligned}
 \mathbf{X}^{(s)}\mathbf{u}_r &= \mathbf{Z}\mathbf{A}^{(r)}\mathbf{X}^{(r)}\mathbf{u}_r \\
 &= 0,
 \end{aligned} \tag{3.20}$$

which is Property (3.13). Now since \mathbf{u}_s is a singular vector,

$$\begin{aligned}
 \mathbf{u}_s^T &\propto \mathbf{u}_s^T \left(\mathbf{X}^{(s)} \right)^T \mathbf{Y}^{(s)} \left(\mathbf{Y}^{(s)} \right)^T \mathbf{X}^{(s)}, \text{ and so} \\
 \mathbf{u}_s^T \mathbf{u}_r &\propto \mathbf{u}_s^T \left(\mathbf{X}^{(s)} \right)^T \mathbf{Y}^{(s)} \left(\mathbf{Y}^{(s)} \right)^T \mathbf{X}^{(s)} \mathbf{u}_r \\
 &= 0,
 \end{aligned}$$

by (3.20). The example on pages 73 to 75 shows that the \mathbf{u}_r are not necessarily singular vectors of $\mathbf{X}^T \mathbf{Y}$. By symmetry the \mathbf{v}_r are also not necessarily singular vectors of $\mathbf{X}^T \mathbf{Y}$. This completes the proof of Property (3.10).

To prove Property (3.11), first we prove that $\boldsymbol{\xi}_r^T \mathbf{X}^{(s)} = 0$ for $r < s$.

$$\mathbf{X}^{(s)} = \mathbf{X}^{(s-1)} - \boldsymbol{\xi}_{s-1} (\boldsymbol{\xi}_{s-1}^T \boldsymbol{\xi}_{s-1})^{-1} \boldsymbol{\xi}_{s-1}^T \mathbf{X}^{(s-1)}$$

(but $\boldsymbol{\xi}_{s-1} = \mathbf{X}^{(s-1)}\mathbf{u}_{s-1}$, so that)

$$\begin{aligned}
 &= \mathbf{X}^{(s-1)} - \mathbf{X}^{(s-1)}\mathbf{u}_{s-1} (\boldsymbol{\xi}_{s-1}^T \boldsymbol{\xi}_{s-1})^{-1} \boldsymbol{\xi}_{s-1}^T \mathbf{X}^{(s-1)} \\
 &= \mathbf{X}^{(s-1)} \left[\mathbf{I} - \mathbf{u}_{s-1} (\boldsymbol{\xi}_{s-1}^T \boldsymbol{\xi}_{s-1})^{-1} \boldsymbol{\xi}_{s-1}^T \mathbf{X}^{(s-1)} \right].
 \end{aligned}$$

Let

$$\mathbf{B}^{(r)} = \mathbf{I} - \mathbf{u}_r (\boldsymbol{\xi}_r^T \boldsymbol{\xi}_r)^{-1} \boldsymbol{\xi}_r^T \mathbf{X}^{(r)}, \tag{3.21}$$

so that

$$\begin{aligned}
\mathbf{X}^{(s)} &= \mathbf{X}^{(s-1)} \mathbf{B}^{(s-1)} \\
&= \mathbf{X}^{(s-2)} \mathbf{B}^{(s-2)} \mathbf{B}^{(s-1)} \\
&= \dots \\
&= \mathbf{X}^{(r+1)} \mathbf{B}^{(r+1)} \dots \mathbf{B}^{(s-1)} \\
&= \mathbf{X}^{(r+1)} \mathbf{Z}
\end{aligned} \tag{3.22}$$

for some matrix \mathbf{Z} . Multiplying this equality on the left by ξ_r^T and plugging in (3.17) we obtain

$$\begin{aligned}
\xi_r^T \mathbf{X}^{(s)} &= \xi_r^T \mathbf{A}^{(r)} \mathbf{X}^{(r)} \mathbf{Z} \\
&= \xi_r^T \left[\mathbf{I} - \xi_r (\xi_r^T \xi_r)^{-1} \xi_r^T \right] \mathbf{X}^{(r)} \mathbf{Z} \\
&= \left[\xi_r^T - \xi_r^T \xi_r (\xi_r^T \xi_r)^{-1} \xi_r^T \right] \mathbf{X}^{(r)} \mathbf{Z} \\
&= 0.
\end{aligned} \tag{3.23}$$

Thus

$$\begin{aligned}
\xi_r^T \xi_s &= \xi_r^T \mathbf{X}^{(s)} \mathbf{u}_s \\
&= 0.
\end{aligned} \tag{3.24}$$

So for $r = 1, \dots, R$ we have ξ_r orthogonal. By construction, each ξ_r is in the column space of \mathbf{X} . So the ξ_r are an orthogonal basis for an R -dimensional subspace of the column space of \mathbf{X} . The result for ω_r follows by symmetry. This completes the proof of Property (3.11).

Recall that γ_s is computed by regressing each column of $\mathbf{X}^{(s)}$ on ξ_s . That is,

$$\begin{aligned}
\gamma_s^T &= (\xi_s^T \xi_s)^{-1} \xi_s^T \mathbf{X}^{(s)} \quad , \text{ or} \\
\gamma_s &\propto \left(\mathbf{X}^{(s)} \right)^T \xi_s
\end{aligned}$$

so that

$$\begin{aligned}
\mathbf{u}_r^T \gamma_s &\propto \mathbf{u}_r^T \left(\mathbf{X}^{(s)} \right)^T \xi_s \\
&= 0,
\end{aligned} \tag{3.25}$$

since by (3.20) the product of the first two terms is zero. This proves Property (3.12).

Property (3.10): To see that the \mathbf{u}_r are not necessarily singular vectors of $\mathbf{X}^T \mathbf{Y}$, consider the following example:

$$\mathbf{X} = \begin{bmatrix} 2.88 & -0.35 & -0.07 & 0.27 \\ -1.03 & -0.13 & -1.01 & -0.45 \\ 0.91 & -0.97 & 1.08 & -1.48 \\ 0.79 & -0.69 & -0.32 & 1.42 \\ 0.62 & -1.11 & 0.65 & 0.13 \end{bmatrix},$$

$$\mathbf{Y} = \begin{bmatrix} 2.742 & 2.966 & 0.195 & 3.061 & 2.119 \\ -2.385 & -6.058 & -1.357 & -1.829 & -2.618 \\ -0.421 & -3.631 & 0.252 & -0.703 & -0.799 \\ 0.989 & 4.085 & 0.279 & 1.698 & -0.149 \\ 0.251 & 0.84 & 0.865 & 0.545 & -0.922 \end{bmatrix}.$$

Then

$$\begin{aligned} \mathbf{X}^T \mathbf{Y} &= \begin{bmatrix} 10.90733 & 15.22556 & 2.94534 & 11.73914 & 7.38282 \\ -1.2023 & -0.47954 & -1.28894 & -1.92824 & 1.49995 \\ 1.6089 & 1.22828 & 2.10205 & 0.68467 & 1.08131 \\ 3.87368 & 14.8107 & 0.79897 & 5.17197 & 2.60131 \end{bmatrix} \\ &= \mathbf{A} \mathbf{G} \mathbf{B}^T \end{aligned}$$

by the singular value decomposition, where \mathbf{G} is diagonal, $\mathbf{A}^T \mathbf{A} = \mathbf{I}_5$, and $\mathbf{B}^T \mathbf{B} = \mathbf{I}_4$. In fact

$$\mathbf{A} = \begin{bmatrix} -0.825620 & 0.519257 & 0.161596 & 0.150365 \\ 0.051425 & -0.164196 & 0.966686 & -0.189508 \\ -0.084740 & 0.192279 & -0.152159 & -0.965762 \\ -0.555451 & -0.816358 & -0.127485 & -0.093710 \end{bmatrix}$$

after rounding. Recall that \mathbf{A} , the matrix of left singular vectors, satisfies

$$\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{A} \propto \mathbf{A}$$

by definition. We may confirm this by computing

$$\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} = \begin{bmatrix} 551.775992 & -35.773533 & 58.461860 & 350.025429 \\ -35.773533 & 9.304810 & -4.931103 & -18.860437 \\ 58.461860 & -4.931103 & 10.153849 & 32.457440 \\ 350.025429 & -18.860437 & 32.457440 & 268.516672 \end{bmatrix}$$

and then computing the componentwise quotient, defined by

$$[\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{A} / \mathbf{A}]_{ij} \equiv [\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{A}]_{ij} / \mathbf{A}_{ij}.$$

This is computed in S-PLUS [Mat96] by the single command

$$(\text{t}(\mathbf{X}) \%*\% \mathbf{Y} \%*\% \text{t}(\mathbf{Y}) \%*\% \mathbf{X} \%*\% \mathbf{A}) / \mathbf{A} .$$

Each column of this matrix consists of a single value (up to rounding error), repeated four times:

$$\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{A} / \mathbf{A} = \begin{bmatrix} 795.49 & 34.44 & 6.59 & 3.23 \\ 795.48 & 34.44 & 6.59 & 3.23 \\ 795.49 & 34.44 & 6.59 & 3.23 \\ 795.49 & 34.44 & 6.59 & 3.23 \end{bmatrix}.$$

Applying PLS-W2A to \mathbf{X} and \mathbf{Y} we obtain, after rounding,

$$\begin{aligned} \mathbf{U} &\equiv [\mathbf{u}_1 | \mathbf{u}_2 | \mathbf{u}_3 | \mathbf{u}_4] \\ &= \begin{bmatrix} -0.825620 & -0.524052 & -0.068777 & -0.197461 \\ 0.051425 & 0.000730 & -0.990437 & 0.128024 \\ -0.084740 & -0.243436 & 0.119356 & 0.958808 \\ -0.555451 & 0.816154 & -0.007675 & 0.159081 \end{bmatrix}. \end{aligned}$$

These coefficients were computed by the S-PLUS function `PLS2blockModeA` (Section 3.15). We see that $\mathbf{U}_{\cdot 1} = \mathbf{A}_{\cdot 1}$, but that the subsequent columns are not equal. Since singular vectors are unique up to permutation and scaling, it follows that not all the columns of \mathbf{U} are singular vectors of $\mathbf{X}^T \mathbf{Y}$. We may see this more dramatically by examining $\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{U} / \mathbf{U}$. If the columns of \mathbf{U} were singular vectors of $\mathbf{X}^T \mathbf{Y}$, we would observe the same pattern for

$\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{U} / \mathbf{U}$ as we observed for $\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{A} / \mathbf{A}$. We observe this only in the first column, however:

$$\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{U} / \mathbf{U} = \begin{bmatrix} 795.49 & 33.86 & -25.78 & 9.11 \\ 795.48 & 6248.45 & 7.27 & 4.11 \\ 795.49 & 27.20 & 15.30 & 2.84 \\ 795.49 & 34.07 & 466.51 & 14.49 \end{bmatrix}.$$

Similar results are obtained if the columns of \mathbf{X} and \mathbf{Y} are centered before the two algorithms are applied to them.

3.14 Proof of Convergence of the Power Method, Lemma 3.12.1

The lemma is stated on page 67.

Let R be the rank of \mathbf{C} . Because of the symmetry of the singular value decomposition, we may assume without loss of generality that $I \geq J$. Let

$$\mathbf{C} = \mathbf{U} \mathbf{D} \mathbf{V}^T$$

be the singular value decomposition, where \mathbf{U} , \mathbf{D} , and \mathbf{V} have been augmented. That is, \mathbf{U} is $I \times I$, \mathbf{D} is $I \times I$, and \mathbf{V} is $I \times J$, with

$$\mathbf{D} = \text{diag}(d_1, \dots, d_R, d_{R+1}, \dots, d_I)$$

and

$$d_1 > d_2 > \dots > d_R > 0 = d_{R+1} = \dots = d_I. \quad (3.26)$$

We have strict inequality in (3.26) because we assumed that the singular values were distinct.

First we shall see that the loop at (1d) yields a pair of singular vectors. We may write

$$\mathbf{u}^0 = \alpha_1 \mathbf{u}_1 + \dots + \alpha_R \mathbf{u}_R + \alpha_{R+1} \mathbf{u}_{R+1} + \dots + \alpha_I \mathbf{u}_I, \quad (3.27)$$

for some set of coefficients α_r . Because \mathbf{u}^0 is not in the left nullspace of \mathbf{C} , we have $|\alpha_r| > 0$ for at least one $r \leq R$. Then, still in the loop at (1d), we have

$$\mathbf{v}^1 \propto \mathbf{C}^T \mathbf{u}^0 \quad (3.28)$$

$$\begin{aligned}
&= \sum_{r=1}^I \alpha_r \mathbf{V} \mathbf{D} \mathbf{U}^T \mathbf{u}_r \\
&= \sum_{r=1}^I \alpha_r d_r \mathbf{v}_r
\end{aligned} \tag{3.29}$$

$$= \sum_{r=1}^R \alpha_r d_r \mathbf{v}_r. \tag{3.30}$$

In going from (3.29) to (3.30) we drop the last $(I - R)$ terms because $0 = d_{R+1} = \dots = d_I$. Next we plug the value for \mathbf{v}^1 obtained at (3.30) to obtain

$$\begin{aligned}
\mathbf{u}^1 &\propto \mathbf{C} \mathbf{v}^1 \\
&= \sum_{r=1}^R \alpha_r d_r \mathbf{U} \mathbf{D} \mathbf{V}^T \mathbf{v}_r \\
&= \sum_{r=1}^R \alpha_r d_r^2 \mathbf{u}_r.
\end{aligned}$$

More generally, the value of the estimate of \mathbf{u} obtained at the k th iteration of the (inner) loop at (1d) contains the scalar d_r raised to the $(2k)$ power:

$$\mathbf{u}^k \propto \sum_{r=1}^R \alpha_r d_r^{2k} \mathbf{u}_r. \tag{3.31}$$

We have noted that at least one of the coefficients in (3.27) must be nonzero. Let us assign the indices r_1, \dots, r_M to these nonzero coefficients. Then (3.31) becomes

$$\mathbf{u}^k \propto \sum_{m=1}^M \alpha_{r_m} d_{r_m}^{2k} \mathbf{u}_{r_m}.$$

That is, at the k th iteration of the (inner) loop at (1d) there is a nonzero constant γ_k such that

$$\mathbf{u}^k = \gamma_k \sum_{m=1}^M \alpha_{r_m} d_{r_m}^{2k} \mathbf{u}_{r_m}.$$

For each m and each k , define the constant $\beta_m^k \equiv \gamma_k \alpha_{r_m} d_{r_m}^{2k}$, so that we can write

$$\mathbf{u}^k = \sum_{m=1}^M \beta_m^k \mathbf{u}_{r_m}. \tag{3.32}$$

We have expressed \mathbf{u}^k , the estimate of \mathbf{u} at an arbitrary iteration k of the inner loop, as a linear combination of the M left singular vectors $\mathbf{u}_{r_1}, \dots, \mathbf{u}_{r_M}$. This subset of the left singular vectors of \mathbf{C} does not change as the inner loop iterates. We shall now show that, as k approaches infinity, the first coefficient β_1^k goes to one and the other coefficients go to zero.

Recall that \mathbf{u}^k has been scaled to have unit norm. Recall furthermore that the \mathbf{u}_r have unit norm by definition, and form an orthogonal set, so that on the right side a vector has been expressed as a linear combination of orthonormal basis vectors. Using this information, let us take the norm of both sides of (3.32) to obtain

$$\begin{aligned} 1 &= \left(\sum_{m=1}^M (\beta_m^k)^2 \right)^{\frac{1}{2}} \\ &= \gamma_k \left(\sum_{m=1}^M \alpha_{r_m}^2 d_{r_m}^{4k} \right)^{\frac{1}{2}}. \end{aligned} \quad (3.33)$$

Let us divide β_m^k by the expression for one obtained in (3.33) to obtain

$$\beta_m^k = \frac{\beta_m^k}{\left(\sum_{m=1}^M (\beta_m^k)^2 \right)^{\frac{1}{2}}} = \frac{\alpha_{r_m} d_{r_m}^{2k}}{\left(\sum_{m=1}^M \alpha_{r_m}^2 d_{r_m}^{4k} \right)^{\frac{1}{2}}}. \quad (3.34)$$

Inverting and squaring (3.34), and plugging in $m = 1$, we have

$$\left(\frac{1}{\beta_1^k} \right)^2 = \frac{\alpha_{r_1}^2 d_{r_1}^{4k}}{\alpha_{r_1}^2 d_{r_1}^{4k}} + \frac{\alpha_{r_2}^2 d_{r_2}^{4k}}{\alpha_{r_1}^2 d_{r_1}^{4k}} + \dots + \frac{\alpha_{r_M}^2 d_{r_M}^{4k}}{\alpha_{r_1}^2 d_{r_1}^{4k}}.$$

The first term is identically 1. Because $d_{r_1} > d_{r_m}$ for $m > 1$, the subsequent terms go to 0 as $k \rightarrow \infty$. Thus $\lim_{k \rightarrow \infty} \beta_1^k = 1$. But in view of (3.33), we must have $\lim_{k \rightarrow \infty} \beta_m^k = 0$ for $m > 1$. Thus $\lim_{k \rightarrow \infty} \mathbf{u}^k = \mathbf{u}_{r_1}$, where \mathbf{u}_{r_1} is the left singular vector of \mathbf{C} , among those not orthogonal to \mathbf{u}^0 , corresponding to the largest singular value.

It follows from construction that $\lim_{k \rightarrow \infty} \mathbf{v}^k = \mathbf{v}_{r_1}$, and that the value of d computed in (1e) is d_{r_1} .

Since $\mathbf{C} = \sum_{r=1}^R d_r \mathbf{u}_r \mathbf{v}_r^T$, resetting the value of \mathbf{C} in (1f) simply removes one of the terms from this sum. After the R th iteration of the outer loop, the new value of \mathbf{C} will be zero. Thus both the inner and outer loops will halt.

Setting \mathbf{u}^0 to the mean of the columns of \mathbf{C} guarantees that $|\alpha_1| > 0$ in Equation 3.27 on page 75. Thus at each iteration of the outer loop the pair of singular vectors computed will be those associated with the largest singular value for the current \mathbf{C} . \square

To see that the raw data method of 3.12.3 converges, note that

$$\begin{aligned} \mathbf{v}^k &\propto \mathbf{Y}^T \boldsymbol{\xi}^k \\ &\propto \mathbf{Y}^T \mathbf{X} \mathbf{u}^{k-1} \\ &= \mathbf{C}^T \mathbf{u}^{k-1}. \end{aligned}$$

This is the general form of line (3.28) on page 75, and the proof proceeds as before.

3.15 *S-PLUS code for PLS-W2A*

```
"PLS2blockModeA"<-
function(X = structure(.Data = c(288, -103, 91, 79, 62, -35, -13,
  -97, -69, -111, -7, -101, 108, -32, 65, 27, -45, -148,
  142, 13), .Dim = c(5, 4), .Dimnames = list(NULL, paste(
  "x", 1:4, sep = "")))/100, Y = structure(.Data = c(2742,
  -2385
  , -421, 989, 251, 2966, -6058, -3631, 4085, 840, 195,
  -1357, 252, 279, 865, 3061, -1829, -703, 1698, 545, 2119,
  -2618, -799, -149, -922), .Dim = c(5, 5), .Dimnames =
  list(NULL, paste("y", 1:5, sep = "")))/1000, R = NULL,
  center = T, returnData = F, tolerance = 1e-08)
{
#
# Two-block Mode A PLS according to Wold, using what Wold calls
# "product data input."
# Decomposes X and Y by computing Xi, Omega, Gamma, Delta, E, and
# Z such that
# (1) X = Xi %*% t(Gamma) + E
```



```

# (2)  $Y = \Omega t(\Delta) + Z$ 
# (3)  $t(X_i) E = 0$ 
# (4)  $t(\Omega) Z = 0$ 
# (5)  $t(X_i)$   $X_i$  is diagonal
# (6)  $t(\Omega)$   $\Omega$  is diagonal
#
# In addition the following numbers are output:
# integer R: the number of columns in  $X_i$  and in  $\Omega$ 
# vector d, of length R:
# inner products of the pairs of latent scores  $t(X_i[,r])$ 
#  $\Omega[,r]$  for  $r=1, \dots, R$  .
#
# Optionally the following can be output (if returnData=T):
# rankXtY: rank of  $t(X)$   $Y$ 
# Matrices U and V: the coefficients used to compute  $X_i$ 
# and  $\Omega$ 
# data X and Y
#
#
  assign("kitn", function(...)
    cat(..., "\n"), frame = 1)
  assign("euclideanNorm", function(x)
    sqrt(sum(x^2)), frame = 1)
  if(is.matrix(X$X) && is.matrix(X$Y)) {
# X and Y are passed in a single list.
    thelist <- X
    X <- thelist$X
    Y <- thelist$Y
  }
  if(center) {

```

```

    if(any(abs(apply(X, 2, mean)) > tolerance)) {
      themax <- max(abs(apply(X, 2, mean)))
      cat("X was not centered.  max(abs(apply(X, 2, mean)))=",
        themax, ". Centering X now.\n")
      X <- scale(X, scale = F)
    }
    else kitn("X was centered already.")
    if(any(abs(apply(Y, 2, mean)) > tolerance)) {
      themax <- max(abs(apply(Y, 2, mean)))
      cat("Y was not centered.  max(abs(apply(Y, 2, mean)))=",
        themax, ". Centering Y now.\n")
      Y <- scale(Y, scale = F)
    }
    else kitn("Y was centered already.")
  }
  else kitn("No attempt to center the data.")
  svdStuff <- function(X, Y, tolerance)
  {
    A <- t(X) %*% Y
    thesvd <- svd(A)
    Rank <- sum(abs(thesvd$d) > tolerance)
    if(Rank < 1) {
      thesvd <- NULL
      return(Rank, thesvd)
    }
    dimnames(thesvd$u) <- list(dimnames(X)[[2]],
      NULL)
    dimnames(thesvd$v) <- list(dimnames(Y)[[2]],
      NULL)
    return(Rank, thesvd)
  }

```

```

}
rankXtY <- svdStuff(X, Y, tolerance)$Rank
Xr <- X
Yr <- Y
Xi <- NULL
Omega <- NULL
U <- NULL
V <- NULL
Gamma <- NULL
Delta <- NULL
d <- NULL
r <- 1
current <- svdStuff(Xr, Yr, tolerance)
while(current$Rank > 0) {
  cat("r=", r, " ")
  thesvd <- current$thesvd
  ur <- thesvd$u[, 1, drop = F]
  vr <- thesvd$v[, 1, drop = F]
  xir <- Xr %*% ur
  omegar <- Yr %*% vr
  gammarT <- as.numeric(t(xir) %*% xir)^(-1) * t(
    xir) %*% Xr
  deltarT <- as.numeric(t(omegar) %*% omegar)^(-1) *
    t(omegar) %*% Yr
  U <- cbind(U, ur)
  V <- cbind(V, vr)
  Xi <- cbind(Xi, xir)
  Omega <- cbind(Omega, omegar)
  Gamma <- cbind(Gamma, t(gammarT))
  Delta <- cbind(Delta, t(deltarT))
}

```

```

    d <- c(d, thesvd$d[1])
    Xr <- Xr - xir %*% gammarT
    Yr <- Yr - omegar %*% deltarT
    r <- r + 1    #
#
# We either halt when t(Xr) %*% Yr has rank 0 (is zero),
# or when we've computed as many pairs of
# latent variables as the user specified.
#
#
    if((!is.null(R)) && (r > R))
        break
    current <- svdStuff(Xr, Yr, tolerance)
}
kitn()
E <- Xr
Z <- Yr
R <- length(d)
if(returnData)
    invisible(return(rankXtY, X, Y, R, d, U, V, Xi,
        Omega, Gamma, Delta, E, Z))
invisible(return(R, d, Xi, Omega, Gamma, Delta, E, Z))
}

```

Chapter 4

RANK-ONE LATENT MODELS FOR CROSS-COVARIANCE

4.1 *Introduction*

A class of Gaussian latent-variable models for cross-covariance is specified, and the set of distributions over the observed variables to which they correspond is precisely characterized. In this class the observed variables, or **indicators**, are divided into two blocks, \mathbf{X} and \mathbf{Y} . A pair of latent variables is postulated, one for each block, ξ for \mathbf{X} and ω for \mathbf{Y} . The indicators are linear functions of their respective latent variables plus error, and errors for the \mathbf{X} block are uncorrelated with those of the \mathbf{Y} block. This latent-variable model differs from the well-known exploratory factor model in that the within-block covariances of the errors are unconstrained.

Any variance-covariance matrix over the indicators with $\text{rank}(\Sigma_{XY}) = 1$ can be fit exactly by the latent-variable model. Although the model is underidentified, the linear coefficient vectors \mathbf{a} , linking ξ to \mathbf{X} , and \mathbf{b} , linking ω to \mathbf{Y} , are identified up to sign and scale. $\text{Cor}(\xi, \omega) = 1$ is always feasible, and $|\text{Cor}(\xi, \omega)|$ is bounded below. When $|\text{Cor}(\xi, \omega)|$ attains its minimum, the scales of \mathbf{a} and \mathbf{b} are maximized and within-block errors are minimized. Subject to the constraint that $|\text{Cor}(\xi, \omega)|$ is at its minimum, the model is identified up to sign.

4.2 *Model specification*

Basic terms are introduced which will be used to state the result.

4.2.1 *Rank-one constraint models*

Let p be the number of \mathbf{X} -variables and q the number of \mathbf{Y} -variables. A **rank-one symmetric constraint model** (equivalently, a rank-one reduced-rank-regression model) is the

set of $(p + q) \times (p + q)$ positive semidefinite matrices satisfying a rank constraint on the cross-covariance matrix:

$$\left. \begin{aligned} \Sigma &= \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}, \\ \text{where } \Sigma_{XY} &\text{ is } p \times q \text{ of unit rank.} \end{aligned} \right\} \quad (4.1)$$

4.2.2 Paired latent correlation models

A **rank-one symmetric paired latent correlation model** is the set of distributions over the latent variables ξ and ω , the observed variables \mathbf{X} and \mathbf{Y} , and the errors ϵ and ζ , specified as follows.

$$\left. \begin{aligned} \mathbf{x} &= \mathbf{a}\xi + \epsilon, \\ \mathbf{y} &= \mathbf{b}\omega + \zeta, \end{aligned} \right\} \text{where} \\ \text{Var} \begin{bmatrix} \xi \\ \omega \end{bmatrix} &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \\ \text{Var}(\epsilon) &= \Sigma_{\epsilon\epsilon}, \\ \text{Var}(\zeta) &= \Sigma_{\zeta\zeta}, \\ \epsilon \perp\!\!\!\perp \begin{bmatrix} \xi \\ \omega \end{bmatrix}, \quad \epsilon \perp\!\!\!\perp \zeta, \quad \begin{bmatrix} \xi \\ \omega \end{bmatrix} \perp\!\!\!\perp \zeta, \\ \mathbf{a} \in \mathbb{R}^p, \quad \mathbf{b} \in \mathbb{R}^q. \end{aligned} \right\} \quad (4.2)$$

Thus the parameters are ρ , \mathbf{a} , \mathbf{b} , $\Sigma_{\epsilon\epsilon}$, and $\Sigma_{\zeta\zeta}$, subject to the constraints that $|\rho| \leq 1$ and that $\Sigma_{\epsilon\epsilon}$ and $\Sigma_{\zeta\zeta}$ must be positive semidefinite. The observed variables \mathbf{X} and \mathbf{Y} are called **indicators**, and the vectors \mathbf{a} and \mathbf{b} are called **saliences** or **loadings**.

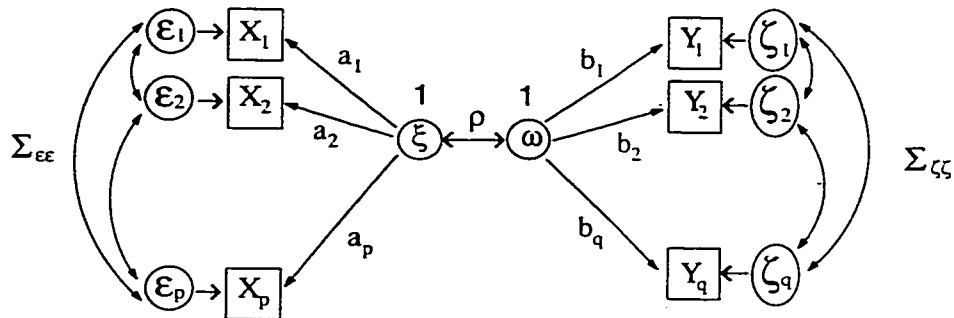


Figure 4.1: Path diagram of a paired latent correlation model. Paired latent correlation models are defined on page 85, Section 4.2.2.

A path diagram for a paired latent model may be seen in Figure 4.1 on page 85.

Lack of identifiability. The paired latent correlation model is underidentified. That is, in general there may be an infinite number of values of the full parameter set $\{\rho, \mathbf{a}, \mathbf{b}, \Sigma_{\epsilon\epsilon}, \Sigma_{\zeta\zeta}\}$ which induce the same distribution in the constraint model. This fact will be demonstrated in the proof of Theorem 4.3.1. We shall precisely characterize the degree of non-identifiability, however, and suggest a natural convention which makes the model identifiable.

4.2.3 Single latent models

A **rank-one symmetric single latent model** is equivalent to a paired latent model where $\xi \equiv \omega$. It is the set of distributions over the latent variable η , the errors ϵ and ζ , and the

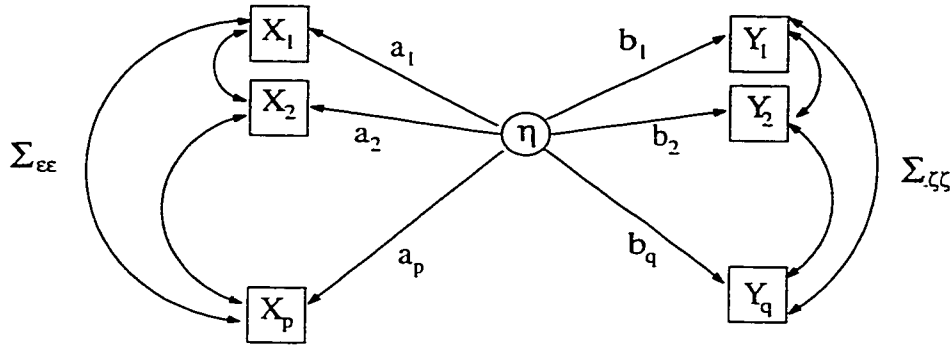


Figure 4.2: Path diagram of a rank-one single latent model, discussed on page 86, Section 4.2.3.

observed variables \mathbf{X} and \mathbf{Y} , specified as follows.

$$\left. \begin{aligned} \mathbf{x} &= \mathbf{a}\eta + \boldsymbol{\epsilon}, \\ \mathbf{y} &= \mathbf{b}\eta + \boldsymbol{\zeta}, \end{aligned} \right\} \text{where}$$

$$\text{Var}(\eta) = 1,$$

$$\text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\epsilon\epsilon}, \quad p \times p,$$

$$\text{Var}(\boldsymbol{\zeta}) = \boldsymbol{\Sigma}_{\zeta\zeta}, \quad q \times q,$$

$$\boldsymbol{\epsilon} \perp\!\!\!\perp \eta, \quad \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\zeta}, \quad \eta \perp\!\!\!\perp \boldsymbol{\zeta},$$

$$\mathbf{a} \in \mathbb{R}^p, \quad \mathbf{b} \in \mathbb{R}^q.$$

Thus the parameters of a symmetric single latent model are $\boldsymbol{\Sigma}_{\epsilon\epsilon}$, $\boldsymbol{\Sigma}_{\zeta\zeta}$, \mathbf{a} and \mathbf{b} , where $\boldsymbol{\Sigma}_{\epsilon\epsilon}$ and $\boldsymbol{\Sigma}_{\zeta\zeta}$ must be positive semidefinite. A path diagram is seen in Figure 4.2 on page 86.

4.3 Maps between spaces of models

Every set of parameter values in a rank-one paired latent correlation model induces a distribution in the rank-one constraint model as follows:

$$\left. \begin{aligned} \Sigma_{XX} &= \mathbf{a}\mathbf{a}^T + \Sigma_{\epsilon\epsilon}, \\ \Sigma_{YY} &= \mathbf{b}\mathbf{b}^T + \Sigma_{\zeta\zeta}, \\ \Sigma_{XY} &= \mathbf{a}\mathbf{b}^T \rho. \end{aligned} \right\} \quad (4.3)$$

The equations (4.3) define a map from the space of symmetric rank-one paired latent correlation model parameterizations into the space of rank-one constraint model distributions. The existence of such a map immediately raises the question whether every distribution in the rank-one constraint model can be obtained by a set of parameter values in a paired latent correlation model—i.e., is the map onto. If such a set of parameters may be found for a given distribution in the constraint model, we shall say that this set **parameterizes** or is a **paired latent parameterization** of the distribution over the constraint model.

The answer to the question in the previous paragraph is yes. Every rank-one constraint model can be parameterized by a symmetric paired latent correlation model. We show this by first proving a stronger result, i.e., that any rank-one constraint model can be parameterized by a symmetric single latent model. The result regarding paired latent correlation models is then obtained as a corollary.

4.3.1 A theorem regarding single latent models

We now state and prove the main result.

Theorem 4.3.1 *For each distribution within the rank-one constraint model there is a non-void class of parameter values in the symmetric single latent model which induce this distribution.*

Proof. We use two lemmas, stated and proved in Section 4.6.1. Decompose Σ as follows:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{a}\mathbf{a}^T & \mathbf{a}\mathbf{b}^T \\ \mathbf{b}\mathbf{a}^T & \mathbf{b}\mathbf{b}^T \end{bmatrix}, \quad (4.4)$$

$$(4.5)$$

$$\mathbf{E} = \begin{bmatrix} \Sigma_{\epsilon\epsilon} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\zeta\zeta} \end{bmatrix},$$

so that

$$\Sigma = \mathbf{Q} + \mathbf{E}.$$

Given a covariance Σ as in (4.1), that is, a distribution under the rank-one constraint model, we seek \mathbf{a} , \mathbf{b} , $\Sigma_{\epsilon\epsilon}$, and $\Sigma_{\zeta\zeta}$ such that

$$\left. \begin{aligned} \Sigma_{XX} &= \mathbf{a}\mathbf{a}^T + \Sigma_{\epsilon\epsilon}, \quad \Sigma_{\epsilon\epsilon} \text{ positive semidefinite,} \\ \Sigma_{YY} &= \mathbf{b}\mathbf{b}^T + \Sigma_{\zeta\zeta}, \quad \Sigma_{\zeta\zeta} \text{ positive semidefinite,} \end{aligned} \right\}, \quad (4.6)$$

$$\text{and } \Sigma_{XY} = \mathbf{a}\mathbf{b}^T. \quad (4.7)$$

Since Σ_{XY} has rank one, by the singular value decomposition we can always find \mathbf{a} and \mathbf{b} satisfying (4.7). The two vectors are only determined up to sign and scale, however, since for any $\delta \neq 0$,

$$\Sigma_{XY} = \mathbf{a}\mathbf{b}^T \Rightarrow \Sigma_{XY} = (\delta\mathbf{a}) \left(\frac{\mathbf{b}^T}{\delta} \right).$$

The scale and sign of \mathbf{a} constitute the only degree of freedom, or lack of identifiability, in the map from the constraint model to the single latent model. This is because the direction of \mathbf{a} is determined by (4.7). Once the sign and scale of \mathbf{a} are determined, then \mathbf{b} is determined by (4.7), and $\Sigma_{\epsilon\epsilon}$ and $\Sigma_{\zeta\zeta}$ are determined by (4.6).

Let us express the single degree of freedom in this model formally. Define \mathbf{u} and \mathbf{v} according to the convention of the singular value decomposition. That is, let

$$\Sigma_{XY} = \mathbf{u}\mathbf{v}^T d, \quad \|\mathbf{u}\| = \|\mathbf{v}\| = 1, \quad (4.8)$$

where $\|\cdot\|$ represents the Euclidean norm. Furthermore let us assume that a sign convention has been adopted, so that the lack of identifiability consists only in the scale of \mathbf{a} . For $0 < \alpha$, let

$$\mathbf{a}(\alpha) \equiv \alpha\mathbf{u}, \quad \mathbf{b}(\alpha) \equiv \frac{\mathbf{v}d}{\alpha}. \quad (4.9)$$

For future reference we note that

$$\|\mathbf{a}(\alpha)\| = \alpha, \quad \|\mathbf{b}(\alpha)\| = \frac{d}{\alpha}. \quad (4.10)$$

Thus $\mathbf{a}(\alpha)$ and $\mathbf{b}(\alpha)$ satisfy $\Sigma_{XY} = \mathbf{a}(\alpha) [\mathbf{b}(\alpha)]^T$. To show that a latent parameterization exists it suffices to show that, if Σ is positive semidefinite, a value of α can always be found such that the values determined by

$$\left. \begin{aligned} \Sigma_{\epsilon\epsilon}(\alpha) &\equiv \Sigma_{XX} - \mathbf{a}(\alpha) [\mathbf{a}(\alpha)]^T = \Sigma_{XX} - \alpha^2 \mathbf{u} [\mathbf{u}]^T \\ \Sigma_{\zeta\zeta}(\alpha) &\equiv \Sigma_{YY} - \mathbf{b}(\alpha) [\mathbf{b}(\alpha)]^T = \Sigma_{YY} - \frac{\mathbf{v}\mathbf{v}^T d^2}{\alpha^2} \end{aligned} \right\} \quad (4.11)$$

are positive semidefinite. Define $f : (0, \infty) \mapsto \mathbb{R}$ and $g : (0, \infty) \mapsto \mathbb{R}$ by

$$\begin{aligned} f(\alpha) &= \min \{\text{eigenvalues of } \Sigma_{\epsilon\epsilon}(\alpha)\}, \\ g(\alpha) &= \min \{\text{eigenvalues of } \Sigma_{\zeta\zeta}(\alpha)\}. \end{aligned} \quad (4.12)$$

It may be shown that these functions are continuous (Theorem 6.3.2, page 365 of Horn and Johnson [HJ85]). By Parts 1 and 3 of Lemma 4.6.1:

- f is monotone nonincreasing and goes to $-\infty$ as $\alpha \rightarrow \infty$;
- g is monotone nondecreasing and goes to $-\infty$ as $\alpha \downarrow 0$.

Let

$$\begin{aligned} \mathcal{F} &= \{\alpha : f(\alpha) < 0\}, \quad \text{and} \\ \mathcal{G} &= \{\alpha : g(\alpha) < 0\}. \end{aligned}$$

By the continuity of f and g these sets are open, but by monotonicity they are in fact intervals:

$$\begin{aligned} \mathcal{F} &= (\alpha_1, \infty) \quad \text{and} \\ \mathcal{G} &= (0, \alpha_2). \end{aligned}$$

The closed set $\mathbb{R} \setminus (\mathcal{F} \cup \mathcal{G})$ is the set of feasible α values. By Lemma 4.6.4, this set is nonvoid; that is, we must have $\alpha_2 \leq \alpha_1$. Since this is the case, let us call them respectively α_{\min} and α_{\max} . The feasible set of values for α is

$$[\alpha_{\min}, \alpha_{\max}], \quad (4.13)$$

and we note for future reference:

$$\begin{aligned}\alpha_{\min} &= \min \left\{ \alpha : \Sigma_{YY} - \frac{\mathbf{v}\mathbf{v}^T d^2}{\alpha^2} \text{ is positive semidefinite} \right\} , \\ \alpha_{\max} &= \max \left\{ \alpha : \Sigma_{XX} - \alpha^2 \mathbf{u}\mathbf{u}^T \text{ is positive semidefinite} \right\} .\end{aligned}\tag{4.14}$$

These follow from the definitions at (4.11) and (4.12).

The fact that $\mathcal{R} \setminus (\mathcal{F} \cup \mathcal{G})$ is nonvoid means the following: In equations (4.6) and (4.7) on page 88 there is at least one scale of the salience vector \mathbf{a} such that both $\Sigma_{\epsilon\epsilon}$ and $\Sigma_{\zeta\zeta}$ are positive semidefinite. Thus there is a single-latent parameterization of any rank-one constraint model. \square

Examples of constraint models and their parameterizations by the single latent model are presented in Section 4.4, starting on page 94.

Corollary 4.3.2 *Each constraint model can be parameterized by at least one paired latent model.*

Proof. Let η be the latent variable of the single latent model. Let ξ and ω be the latent variables in the paired latent model, and let $\xi \equiv \omega \equiv \eta$. \square

4.3.2 Practical considerations

The proof of Theorem 4.3.1 suggests that the task of finding a single-latent parameterization of a covariance matrix (4.1) in the rank-one constraint model might be broken into the following two steps: First find a decomposition

$$\Sigma_{XY} = \mathbf{a}\mathbf{b}^T ;\tag{4.15}$$

then estimate

$$\begin{aligned}\alpha_{\max} &\equiv \max \left\{ \alpha : \Sigma_{XX} - \alpha^2 \mathbf{a}\mathbf{a}^T \text{ is positive semidefinite} \right\} \\ \alpha_{\min} &\equiv \min \left\{ \alpha : \Sigma_{YY} - \mathbf{b}\mathbf{b}^T / \alpha^2 \text{ is positive semidefinite} \right\} .\end{aligned}\tag{4.16}$$

The decomposition of Σ_{XY} . If Σ is known, the decomposition (4.15) is exact, and can be found directly. For instance, set \mathbf{a} equal to the first nonzero column of Σ_{XY} , and determine \mathbf{b} by

$$\begin{aligned} \mathbf{b}_j &= \frac{(\Sigma_{XY})_{ij}}{\mathbf{a}_i}, \text{ where} \\ i &= \min\{k : \mathbf{a}_k \neq 0\}. \end{aligned}$$

When Σ is estimated by the sample covariance matrix \mathbf{S} , in most cases we will have $\text{rank}(\mathbf{S}_{XY}) > 1$ and (4.15) will be an approximation. Then standard singular value decomposition software could be used. For instance, let \mathbf{u} and \mathbf{v} be the first pair of singular vectors, and define \mathbf{a} and \mathbf{b} as in (4.9) on page 88.

Estimating α_{\min} and α_{\max} . From (4.16) we have

$$\begin{aligned} \alpha_{\max} &= \max\{\alpha : \text{least eigenvalue of } \Sigma_{XX} - \alpha^2 \mathbf{a}\mathbf{a}^T \text{ is nonnegative}\} \\ &= \max\{\alpha : \text{least eigenvalue of } \Sigma_{XX} - \alpha^2 \mathbf{a}\mathbf{a}^T \text{ is zero}\} \end{aligned} \quad (4.17)$$

$$= \max\{\alpha : |\Sigma_{XX} - \alpha^2 \mathbf{a}\mathbf{a}^T| = 0\}, \quad (4.18)$$

where the equality at (4.17) follows by the continuity of the least eigenvalue. By Propositions 4.6.5 and 4.6.6 on page 111, to find α satisfying the expression inside the brackets at (4.18) we may consider the eigenvalue problem

$$|\Sigma_{XX}^{-1} \mathbf{a}\mathbf{a}^T - \lambda \mathbf{I}| = 0,$$

provided the inverse exists. Since $\Sigma_{XX}^{-1} \mathbf{a}\mathbf{a}^T$ is a rank-one matrix there will be a single nonzero eigenvalue, say λ . But since $\Sigma_{XX}^{-1} \mathbf{a}\mathbf{a}^T$ is not necessarily symmetric, we have not yet confirmed that the eigenvalue will be positive or even real. Let us therefore convert the equation inside the brackets at (4.18) to a symmetric, positive semidefinite eigenvalue problem. Let \mathbf{R} be the inverse of a symmetric positive definite square root of Σ_{XX} . Since $|\mathbf{R}| > 0$, we may left-multiply and right-multiply both sides of the equation by $|\mathbf{R}|$ without changing the problem, to obtain

$$|\mathbf{I} - \alpha^2 \mathbf{R}\mathbf{a}\mathbf{a}^T \mathbf{R}| = 0. \quad (4.19)$$

By Proposition 4.6.5, or by multiplying both sides of (4.19) by $\left(-\frac{1}{\alpha^2}\right)^n$, we obtain the symmetric positive semidefinite eigenvalue problem

$$\left| \mathbf{R} \mathbf{a} \mathbf{a}^T \mathbf{R} - \frac{1}{\alpha^2} \mathbf{I} \right| = 0 .$$

An example of an eigenvalue approach to finding α_{\min} and α_{\max} is presented in Section 4.4.

The equation inside the brackets at (4.18) expresses a **generalized eigenvalue problem**. The problem of computing generalized eigenvalues has been studied extensively. Computational methods exist which are preferable to the procedures presented above. A brief discussion of the generalized eigenvalue problem, an introduction to computational issues, and a bibliography, may be found in Golub and van Loan [GVL96], pages 375–390 and 461ff.

4.3.3 Parameterization of paired latent correlation models

The proof of Theorem 4.3.1 guarantees at least one paired-latent parameterization of any rank-one constraint model. In this parameterization the latent variables (ξ for the \mathbf{X} block, ω for the \mathbf{Y} block) have unit variance and unit covariance, hence unit correlation. In the current section the complete set of paired latent parameterizations for a given rank-one constraint model will be characterized.

A distribution in the rank-one constraint model is mapped to an equivalence class of parameter values in the symmetric paired latent correlation model as follows. Consider the following set:

$$\{(\rho, \alpha) : |\rho| \leq 1, \quad \alpha_{\min}^2 \leq \alpha^2 \rho^2, \quad \alpha \leq \alpha_{\max}\} , \quad (4.20)$$

where α_{\min} and α_{\max} are defined by (4.14) in the proof of Theorem 4.3.1. When $\alpha_{\min} = \alpha_{\max}$, this set is a singleton; otherwise it is a continuous closed region. Let

$$\boldsymbol{\theta} \equiv (\rho, \mathbf{a}, \mathbf{b}, \Sigma_{\epsilon\epsilon}, \Sigma_{\zeta\zeta})$$

denote the parameter vector for the paired latent correlation model. Each point in (4.20)

determines a value of θ as follows.

$$\left. \begin{aligned} \mathbf{a} &= \alpha \mathbf{u} , \\ \mathbf{b} &= \frac{\mathbf{v}}{\alpha \rho} d , \\ \Sigma_{\epsilon\epsilon} &= \Sigma_{XX} - \mathbf{a}\mathbf{a}^T , \\ \Sigma_{\zeta\zeta} &= \Sigma_{YY} - \mathbf{b}\mathbf{b}^T . \end{aligned} \right\} \quad (4.21)$$

When the map (4.21) is applied to any value outside the set (4.20), the resulting values do not define feasible parameters for the paired latent correlation model. This will be shown in the current section. Consequently (4.20) is called the **feasible set** for the rank-one paired latent correlation model. An example of a feasible set may be seen in Figure 4.4 on page 98.

We have already seen in Corollary (4.3.2) that $\text{Cor}(\xi, \omega) = 1$ is always feasible. Observing that

$$\alpha^2 \leq \alpha_{\max}^2 \text{ and } \alpha_{\min}^2 \leq \rho^2 \alpha^2 \Rightarrow |\rho| \geq \frac{\alpha_{\min}}{\alpha_{\max}}$$

we see that the constraints at (4.20) entail a lower bound on the correlation, and we define

$$\rho_{\min} = \frac{\alpha_{\min}}{\alpha_{\max}} .$$

To justify the term “feasible set” for (4.20), it suffices to show

1. that the parameter values defined above recover Σ (that this parameterization maps into the constraint model with which we started), and
2. that the constraints at (4.20) are necessary and sufficient for the parameters to be feasible—that is, for $\Sigma_{\epsilon\epsilon}$, $\Sigma_{\zeta\zeta}$, and $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ to be positive semidefinite.

Condition (1), the fact that the parameterization recovers the constraint model, follows from (4.21) and from the fact that

$$\begin{aligned}\text{Cov}(\mathbf{x}, \mathbf{y}) &= \mathbf{a}\mathbf{b}^T \rho \quad \text{by (4.2)} \\ &= \mathbf{u}\mathbf{v}^T d \quad \text{by (4.21)} \\ &= \Sigma_{XY} \quad \text{by (4.8)}.\end{aligned}$$

As for Condition (2), the feasibility of the set, the constraint $|\rho| \leq 1$ is necessary and sufficient for $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ to be positive semidefinite. Since

$$\Sigma_{\epsilon\epsilon} = \Sigma_{XX} - \mathbf{a}\mathbf{a}^T = \Sigma_{XX} - \alpha^2 \mathbf{u}\mathbf{u}^T,$$

by (4.14) on page 90 a necessary and sufficient condition for $\Sigma_{\epsilon\epsilon}$ to be positive semidefinite is $\alpha \leq \alpha_{\max}$. By the definitions of $\Sigma_{\zeta\zeta}$ and \mathbf{b} at (4.21),

$$\Sigma_{\zeta\zeta} = \Sigma_{YY} - \mathbf{b}\mathbf{b}^T = \Sigma_{YY} - \frac{\mathbf{v}\mathbf{v}^T}{\rho^2 \alpha^2} d^2.$$

By (4.14) on page 90, the greatest real number t such that $\Sigma_{YY} - t\mathbf{v}\mathbf{v}^T$ is positive semidefinite is $\frac{d^2}{\alpha_{\min}^2}$. Thus

$$\Sigma_{\zeta\zeta} \text{ is positive semidefinite} \Leftrightarrow \frac{d^2}{\rho^2 \alpha^2} \leq \frac{d^2}{\alpha_{\min}^2} \Leftrightarrow \alpha_{\min}^2 \leq \rho^2 \alpha^2.$$

The parameterization has been shown to be correct.

4.4 Examples

Parameterization of a 5×5 positive definite matrix in the rank-one constraint model. Consider the following symmetric positive definite matrix.

$$\Sigma = \begin{bmatrix} 7 & 0 & 0 & 1 & 0.5 \\ 0 & 7 & 0 & 2 & 1 \\ 0 & 0 & 7 & 3 & 1.5 \\ 1 & 2 & 3 & 9 & 0 \\ 0.5 & 1 & 1.5 & 0 & 5 \end{bmatrix} \quad (4.22)$$

Let $p = 3$, $q = 2$. We shall first parameterize (4.22) using a simple approach which parallels the proof of Theorem 4.3.1. Then we shall parameterize it again, using the generalized eigenvalue approach of Section 4.3.2, and verify that the approaches give identical results.

Choosing the convention that both d and the component of \mathbf{u} with greatest absolute value shall be positive, we obtain

$$\mathbf{a}(\alpha) = \frac{\alpha}{\sqrt{14}} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{b}(\alpha) = \frac{\sqrt{14}}{2\alpha} \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad d = \sqrt{\frac{35}{2}},$$

$$\Sigma_{\epsilon\epsilon}(\alpha) = \frac{1}{14} \begin{bmatrix} 98 - \alpha^2 & -2\alpha^2 & -3\alpha^2 \\ -2\alpha^2 & 98 - 4\alpha^2 & -6\alpha^2 \\ -3\alpha^2 & -6\alpha^2 & 98 - 9\alpha^2 \end{bmatrix},$$

$$\Sigma_{\zeta\zeta}(\alpha) = \begin{bmatrix} 9 - \frac{14}{\alpha^2} & -\frac{7}{\alpha^2} \\ -\frac{7}{\alpha^2} & 5 - \frac{7}{2\alpha^2} \end{bmatrix},$$

$$\det \Sigma_{\epsilon\epsilon}(\alpha) = 343 - 49\alpha^2, \quad \det \Sigma_{\zeta\zeta}(\alpha) = 45 - \frac{203}{2\alpha^2},$$

$$[\alpha_{\min}, \alpha_{\max}] = \left[\sqrt{\frac{203}{90}}, \sqrt{7} \right] \approx [1.50, 2.65].$$

The curves of least eigenvalues are plotted in Figure 4.3 on page 97. The minimum feasible correlation is $\rho_{\min} \equiv \frac{\alpha_{\min}}{\alpha_{\max}} = \frac{1}{30}\sqrt{290} \approx 0.57$. The feasible set for the paired latent correlation model is displayed in Figure 4.4 on page 98.

In Section 4.3.2 it was pointed out that the problem of finding α_{\min} and α_{\max} is a generalized eigenvalue problem. We shall now follow the derivation in Section 4.3.2 and verify that its solution is identical to that derived above. Following (4.15) on page 90, let us take $\mathbf{a}(1)$ and $\mathbf{b}(1)$ as our decomposition of the cross-covariance. To find α_{\max} we then

solve $\left| \Sigma_{XX}^{-1} \mathbf{a}(1) \mathbf{a}(1)^T - \frac{1}{\alpha^2} \mathbf{I} \right| = 0$, where

$$\Sigma_{XX}^{-1} \mathbf{a}(1) \mathbf{a}(1)^T = \frac{1}{98} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}.$$

The sole nonzero eigenvalue is $\frac{1}{7}$; thus $\alpha_{\max}^2 = 7$, as expected. Similarly we note that α_{\min} solves $|\Sigma_{YY}^{-1} \mathbf{b}(1) \mathbf{b}(1)^T - \alpha^2 \mathbf{I}| = 0$, where

$$\Sigma_{YY}^{-1} \mathbf{b}(1) \mathbf{b}(1)^T = \begin{bmatrix} \frac{14}{9} & \frac{7}{9} \\ \frac{7}{5} & \frac{7}{10} \end{bmatrix}.$$

The single positive eigenvalue is $\frac{203}{90}$, our solution for α_{\min}^2 , as expected.

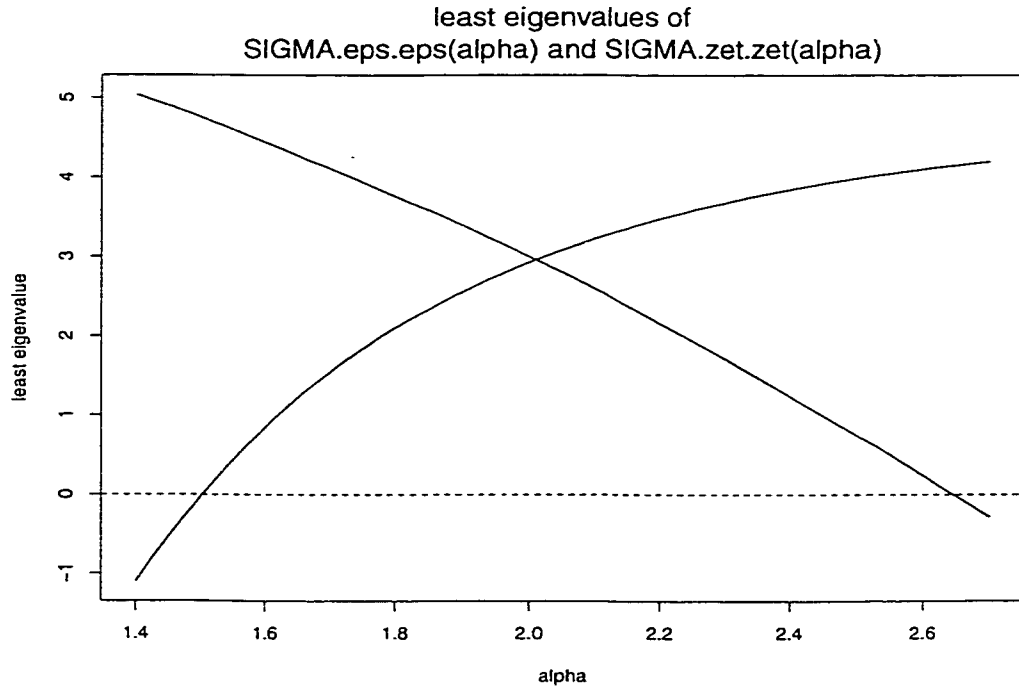


Figure 4.3: The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ (the decreasing function) and $\Sigma_{\zeta\zeta}(\alpha)$ in the single latent parameterization of the matrix at line (4.22), page 94. The feasible values for α lie in the closed interval $\left[\sqrt{\frac{203}{90}}, \sqrt{7}\right] \approx [1.50, 2.65]$. These are the values of α for which the least eigenvalues of both $\Sigma_{\epsilon\epsilon}(\alpha)$ and $\Sigma_{\zeta\zeta}(\alpha)$ are nonnegative.

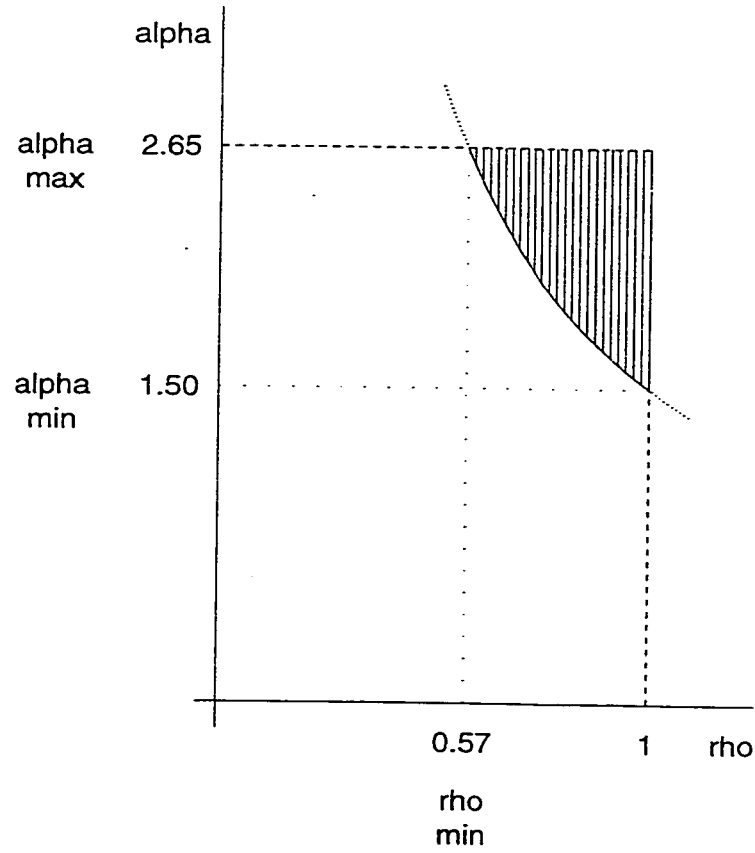


Figure 4.4: Feasible ρ and α for the paired-latent correlation parameterization of the rank-constraint distribution specified by (4.22). Feasible values are in the shaded region. The right boundary of the feasible set corresponds to the single latent model. The (curved) left boundary is the line $\rho\alpha = \alpha_{\min}$. The minimum feasible correlation is $\rho_{\min} \equiv \frac{\alpha_{\min}}{\alpha_{\max}}$.

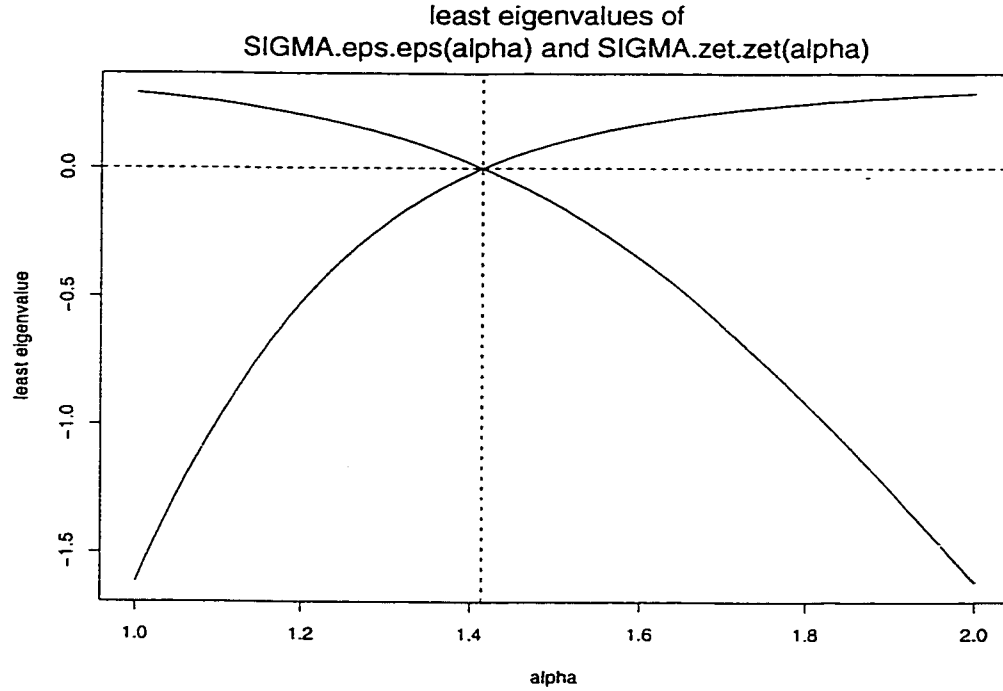


Figure 4.5: The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ and $\Sigma_{\zeta\zeta}(\alpha)$ for the singular matrix at line (4.23), page 99, where there are two \mathbf{X} -variables and two \mathbf{Y} -variables. The decreasing function is the least eigenvalue of $\Sigma_{\epsilon\epsilon}(\alpha)$. Since $\sqrt{2}$ is the only point where both curves equal or exceed zero, this is the only feasible value for α .

A singular matrix in the rank-one constraint model for which the feasible set consists of a single point. The following matrix is singular. For purposes of comparison with the matrix at (4.24) on page 100 we note that its eigenvalues are $\{4.56, 1, 0.44, 0\}$.

$$\Sigma = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (4.23)$$

Let $p = q = 2$. Then the only value of α which parameterizes the single latent model is $\sqrt{2}$. Consequently, in the paired latent parameterization the only feasible correlation is unity. The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ and $\Sigma_{\zeta\zeta}(\alpha)$ are plotted in Figure 4.5, page 99.

A singular matrix in the rank-one constraint model for which the feasible set is infinite. The example at (4.23) notwithstanding, a matrix in the rank-one constraint model may be singular and still admit an infinite number of paired-latent parameterizations. The following degenerate case illustrates this. The matrix at (4.24) has eigenvalues 4.56, 1, 0.44, and 0, as does the matrix at (4.23).

$$\Sigma = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (4.24)$$

Again let $p = q = 2$. The matrix (4.24) represents a degenerate distribution, since the two \mathbf{Y} variables are perfectly correlated. The fact that there are infinitely many feasible parameterizations follows from the fact that $\mathbf{v}\mathbf{v}^T$ is proportional to Σ_{YY} . The feasible set is $[\sqrt{2}, \sqrt{3}]$. The value $\alpha = \sqrt{2}$ entails zero error for the \mathbf{Y} -block, so that each \mathbf{Y} variable measures the latent $\boldsymbol{\eta}$ exactly. All feasible values of α , however, entail a singular error covariance for the \mathbf{Y} -block. For all values of α , whether feasible or not, $\Sigma_{\zeta\zeta}(\alpha) = \begin{bmatrix} \delta & \delta \\ \delta & \delta \end{bmatrix}$ for some $\delta \in \mathbb{R}$. When $\sqrt{2} \leq \alpha$, so that $\delta \geq 0$, the least eigenvalue is 0. For $\alpha < \sqrt{2}$, hence $\delta < 0$, $\Sigma_{\zeta\zeta}(\alpha)$ is not a covariance and the least eigenvalue is strictly increasing in α . The least eigenvalues are plotted against α in Figure 4.6, page 101.

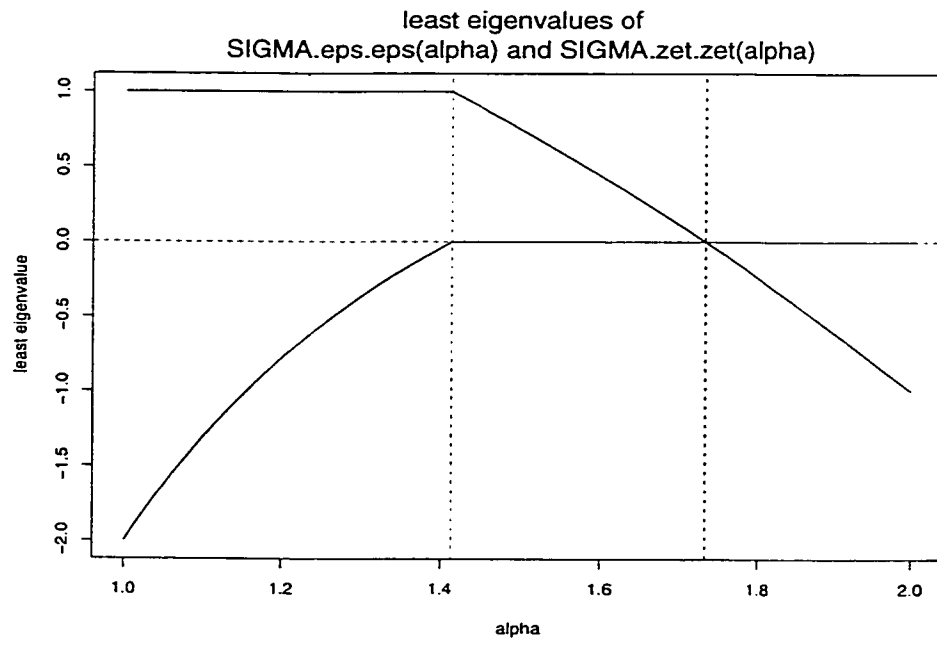


Figure 4.6: The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ (the nonincreasing function) and $\Sigma_{\zeta\zeta}(\alpha)$ for the matrix (4.24), page 100.

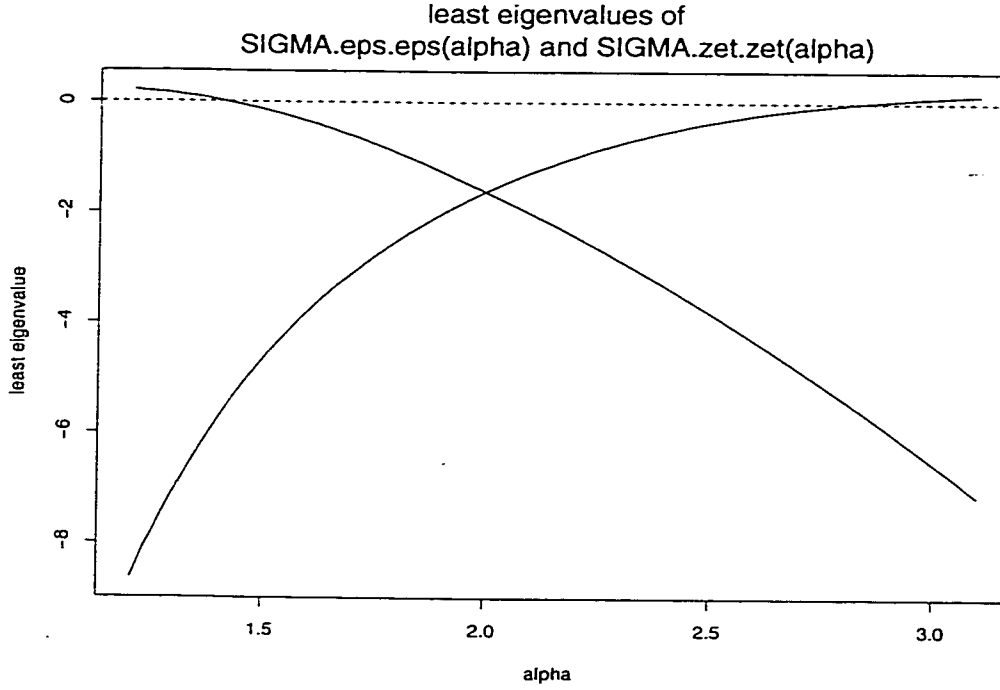


Figure 4.7: The least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ (the decreasing function) and $\Sigma_{\zeta\zeta}(\alpha)$ for the matrix at line (4.25), page 102. As we would expect, since this matrix fails to be positive semidefinite there is no α , or scale for the \mathbf{X} -salience vector, by which the single latent model can parameterize it. This may be seen by the fact that there is no value of α for which both curves are greater than or equal to zero.

A matrix which cannot be parameterized. The following matrix,

$$\Sigma = \begin{bmatrix} 2 & 1 & 2 & 2 \\ 1 & 1 & 2 & 2 \\ 2 & 2 & 2 & 1 \\ 2 & 2 & 1 & 1 \end{bmatrix}, \quad (4.25)$$

is not a variance; that is, it fails to be positive semidefinite. Its least eigenvalue is -1.62 . The curves of least eigenvalues of $\Sigma_{\epsilon\epsilon}(\alpha)$ and $\Sigma_{\zeta\zeta}(\alpha)$ are plotted in Figure 4.7 on page 102, under the assumption that $p = q = 2$.

4.5 Discussion

Likelihood and the equivalence of model spaces. Three spaces of covariance matrices over the observed variables \mathbf{X} and \mathbf{Y} are of interest in the current work. They are:

1. Those corresponding to the rank-constraint model.
2. Those induced by the symmetric paired latent model.
3. Those induced by the symmetric single latent model.

It follows from definitions and from Equations (4.3) that $\text{Set } 3 \subset \text{Set } 2 \subset \text{Set } 1$. Theorem 4.3.1, however, implies that $\text{Set } 1 \subset \text{Set } 3$. Hence $\text{Set } 1 = \text{Set } 2 = \text{Set } 3$, a fact which we state as the following corollary.

Corollary 4.5.1 *The sets of covariance matrices over the observed variables induced by the symmetric paired latent correlation model and the symmetric single latent model are equal to the set of covariance matrices belonging to the rank-one constraint model.*

Thus all single and paired latent parameterizations within an equivalence class have the same likelihood under the multivariate normal model for $(\mathbf{X}^T, \mathbf{Y}^T)^T$, and consequently there is no way using only data to distinguish between the three models. Furthermore it may be shown that the rank-one constraint model is covariance equivalent to reduced-rank regression (RRR). This fact is reviewed in Appendix A. Since maximum-likelihood estimation procedures are available for RRR (Anderson [And51] [And80] [And99]), the problems of maximum-likelihood estimation for the paired and single symmetric latent models are solved, at least when the covariance matrix is invertible.

Within-block error tradeoffs in the single latent model. The feasible sets introduced at (4.13) and (4.20) characterize the degree to which a single latent or paired latent model is not identified.

Let us first consider the single latent model. If $\alpha_{\min} = \alpha_{\max}$ there is only one parameterization, say α^* . Since f and g , the functions used to define α_{\min} and α_{\max} , are continuous,

$f(\alpha^*) = g(\alpha^*) = 0$ and the unique parameterization yields singular within-block error variances for both blocks.

If the joint variance-covariance Σ is strictly positive definite, $\alpha_{\min} < \alpha_{\max}$. The converse is not true, however, because f or g may fail to be strictly monotone. We have seen an example at (4.24) on page 100.

When $\alpha_{\min} < \alpha_{\max}$, the likelihood for the single latent model gives no information on what value of α within the feasible set should be chosen. The choice of α does however entail a decision as to which block of indicators is interpreted as measuring the latent variable more precisely. This is because of the way the within-block error variances vary with α in (4.11). The variances of the **X**-block errors decrease linearly in α^2 , those of the **Y**-block in $\frac{1}{\alpha^2}$. Thus when a constraint distribution permits a within-block error covariance to be nonsingular, that error covariance can be made singular only by simultaneously

- reducing the variances of all the errors in the block, and
- increasing the variances of all the errors in the other block.

Thus α may be considered a “tradeoff parameter.” When $\alpha = \alpha_{\min}$, $\Sigma_{\zeta\zeta}$ is singular and consequently **Y** measures the latent η as closely as possible. At the same time the errors of the **X**-block, $\Sigma_{\epsilon\epsilon}$, have the greatest variance permitted by the constraint model. When $\alpha = \alpha_{\max}$ the reverse is true.

Correlation, identifiability, and tradeoffs in the paired latent correlation model.

When $\alpha_{\min} = \alpha_{\max}$ the paired latent correlation model has a unique parameterization, just as the single latent model has. When $\alpha_{\min} < \alpha_{\max}$ this model, like the single latent model, is not identified. In this case, however, the feasible set lies in the plane, and we consequently have two tradeoffs, not one as with the single latent model.

The choice of the quantity ρ within the feasible interval $[\rho_{\min}, 1]$ entails a tradeoff between error variance on one hand and correlation on the other. When $\rho = \rho_{\min}$ the error variances for both blocks are at their minimum, and in fact the covariance matrices $\Sigma_{\epsilon\epsilon}$ and $\Sigma_{\zeta\zeta}$ are singular. When $\rho = 1$ the feasible values for α are exactly those for the single latent model,

and the tradeoff described for that model applies. In particular, at least one of the error variances may be nonsingular.

Recall that $\rho_{\min} \equiv \frac{\alpha_{\min}}{\alpha_{\max}}$ is a constant determined by the constraint model, that is, by the joint population variance-covariance matrix Σ . When this quantity is less than one we may choose to have latent variables perfectly correlated but poorly measured ($\rho = 1$); or latent variables measured with minimal error, in fact with a singular error distribution, but poorly correlated with each other ($\rho = \rho_{\min}$); or anything between these two extremes. The former entails an additional choice as to which block shall have greater error, as in the single latent model. If the latter choice, ($\rho = \rho_{\min}$), were adopted as a convention, and a sign convention were also adopted, the model would be identifiable.

Singular value decomposition. In two-block rank-one Mode A Partial Least Squares (PLS), an application of the singular value decomposition to the sample cross-covariance matrix, empirical saliences \mathbf{u} and \mathbf{v} are computed. These PLS saliences are scaled sample covariances between indicators and paired latent-variable scores, one for the \mathbf{X} -block and one for the \mathbf{Y} -block. The scores are computed as linear combinations of the variables for their respective blocks.

Although the PLS procedure depends on no statistical model, it is closely related to the family of paired latent models. This relationship is seen most easily when an equivalent paired latent model, the SVD paired latent model, is considered rather than the paired latent correlation model. The SVD paired latent model is defined and discussed in Section 4.6.3.

Provided the two largest singular values of Σ_{XY} are distinct and a sign convention has been adopted, it can easily be shown that the PLS saliences are consistent for the saliences or loadings of the SVD paired latent model. That is, as the number of observations approaches infinity, the values of the PLS saliences approach the loadings of the SVD paired latent model with probability one. This fact follows from the continuity of the singular value decomposition (Theorem 6.3.2, page 365 of Horn and Johnson [HJ85]). Note that prior to Theorem 4.3.1 it was not known whether every population covariance matrix over $(\mathbf{X}^T, \mathbf{Y}^T)^T$ with $\text{rank}(\Sigma_{XY}) = 1$ could be interpreted as having arisen from a paired latent

model. This has now been shown.

In PLS applications, vectors of scores on paired latent variables are computed as linear combinations of the indicators. The correlation between these variables has been estimated from the vectors of latent scores. This estimate of correlation is subject to the attenuation discussed by Spearman [Spe87]. A traditional correction for attenuation however is not necessary.¹ Instead the lower bound for correlation shown in the current work may be used.

An example of two-block Mode A PLS, and of the interpretation of PLS saliences, may be seen in Streissguth et al. [SBSB93b].

Factor models. Bollen ([Bol89], pages 227ff.) distinguishes between confirmatory and exploratory factor models. He states that, in exploratory factor models, within-block errors or “measurement errors” are uncorrelated. In confirmatory factor models, on the other hand, these errors may be correlated. Thus single and paired latent models are closely related to confirmatory factor models. When they are identified they satisfy Bollen’s definition, and may be considered confirmatory factor models.

A practical difference exists, however, between the current approach and the manner in which confirmatory factor models have customarily been treated. Although general statements of the confirmatory factor model family often place no *a priori* constraints on the within-block error covariance, few if any specific confirmatory factor models can be found in the literature with unconstrained within-block covariance. Reasons for this are both that such a model would be underidentified, and that it could be difficult to fit.

The current work deals with both difficulties. The degree to which the model is underidentified has been characterized. The model has been shown to be identified under the convention that $\rho = \rho_{\min}$. In addition the problem of fitting the model by maximum likelihood has been transformed into the well-studied problem of fitting a reduced-rank regression model.

¹There is an extensive literature on attenuation and on corrections for attenuation. Spearman’s seminal article [Spe04] was reprinted in 1987 [Spe87]. Attenuation is mentioned by Kendall and Stuart [KS67], page 327, and by Fisher and van Belle [FvB93], page 385. Lord and Novick provide a mathematical justification for the correction for attenuation [LN68]. Muchinsky reviews the issues and controversies surrounding disattenuation, including alternate formulas [Muc96]. Zimmerman and Williams use simulation to investigate the properties of the disattenuated correlation under various conditions [ZW97].

4.6 Appendix

4.6.1 Lemmas

To prove Theorem 4.3.1 we require the following lemmas.

Lemma 4.6.1 *Let \mathbf{A} and \mathbf{C} be symmetric matrices of the same dimension, \mathbf{C} positive semidefinite. Let $h : [0, \infty) \mapsto \mathbb{R}$ be defined by*

$$h(\alpha) = \text{the smallest eigenvalue of } (\mathbf{A} - \alpha\mathbf{C}) .$$

Then

1. *The function h is monotone nonincreasing. If \mathbf{C} is strictly positive definite, the function is strictly monotone decreasing.*
2. $\lim_{\alpha \downarrow 0} h(\alpha) = h(0)$.
3. *If \mathbf{C} has at least one positive eigenvalue, $\lim_{\alpha \uparrow \infty} h(\alpha) = -\infty$.*

Proof. Let $\mathbf{z}(\alpha)$ be the eigenvector belonging to the smallest eigenvalue of $(\mathbf{A} - \alpha\mathbf{C})$, without loss of generality let $\|\mathbf{z}(\alpha)\| = 1$, and recall that, with this convention, $\mathbf{z}(\alpha)^T (\mathbf{A} - \alpha\mathbf{C}) \mathbf{z}(\alpha)$ equals the smallest eigenvalue.

Part 1. Let $\alpha < \beta$.

$$\begin{aligned} h(\alpha) &= \mathbf{z}(\alpha)^T (\mathbf{A} - \alpha\mathbf{C}) \mathbf{z}(\alpha) \\ &= \mathbf{z}(\alpha)^T \mathbf{A} \mathbf{z}(\alpha) - \alpha \mathbf{z}(\alpha)^T \mathbf{C} \mathbf{z}(\alpha) \\ &\geq \mathbf{z}(\alpha)^T \mathbf{A} \mathbf{z}(\alpha) - \beta \mathbf{z}(\alpha)^T \mathbf{C} \mathbf{z}(\alpha) \end{aligned} \tag{4.26}$$

$$\begin{aligned} &= \mathbf{z}(\alpha)^T (\mathbf{A} - \beta\mathbf{C}) \mathbf{z}(\alpha) \\ &\geq \mathbf{z}(\beta)^T (\mathbf{A} - \beta\mathbf{C}) \mathbf{z}(\beta) \\ &= h(\beta) . \end{aligned} \tag{4.27}$$

At line (4.26) the inequality is not strict because possibly $\mathbf{z}(\alpha)^T \mathbf{C} \mathbf{z}(\alpha) = 0$. If \mathbf{C} is strictly positive definite, the inequality at this line is strict and hence h is strictly decreasing. The inequality at line (4.27) occurs because $\mathbf{z}(\beta)$, by definition, minimizes the quadratic form.

Part 2. This is a consequence of a well-known theorem regarding the eigenvalues of a diagonalizable matrix under perturbation. See, for example, Horn and Johnson [HJ85], Theorem 6.3.2, page 365. An ad-hoc proof of the continuity of h at 0 is included here as an exercise.

$$h(\alpha) \equiv \min_{\|x\|=1} (x^T A x - \alpha x^T C x) \quad (4.28)$$

$$\geq \min_{\|x\|=1} x^T A x - \max_{\|y\|=1} (\alpha y^T C y) , \quad (4.29)$$

since in line (4.28) the feasible set is a subset of the feasible set in line (4.29). Letting $\alpha \downarrow 0$ we see that the limit is bounded below by $h(0)$.

$$\begin{aligned} h(\alpha) &= z(\alpha)^T (A - \alpha C) z(\alpha) \\ &\leq z(0)^T (A - \alpha C) z(0) \\ &= h(0) - \alpha z(0)^T C z(0) . \end{aligned}$$

Thus the limit is also bounded above by $h(0)$.

Part 3. Let y be an eigenvector corresponding to a positive eigenvalue of C .

$$\min_{\|x\|=1} x^T (A - \alpha C) x \leq y^T A y - \alpha y^T C y .$$

The first term is constant. The second term approaches negative infinity as α approaches infinity.

Lemma 4.6.2 *Let x be a p -vector, y a q -vector. Let U be $p \times R$, V $q \times R$, let $W = \begin{bmatrix} U \\ V \end{bmatrix}$, $Q = WW^T$. Let the entries in these matrices and vectors be real, and consider the quadratic form*

$$Z(t) = \begin{bmatrix} x^T & ty^T \end{bmatrix} Q \begin{bmatrix} x \\ ty \end{bmatrix} .$$

Then there is a real t such that $Z(t) = 0$ if and only if one or more of the following conditions holds:

$$\begin{aligned} x^T U &= 0 , \\ y^T V &= 0 , \\ x^T U &\propto y^T V . \end{aligned}$$

Furthermore the real solution, if it exists, is unique.

Proof.

$$\begin{aligned}
 Z(t) &= \begin{bmatrix} \mathbf{x}^T, t\mathbf{y}^T \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} [\mathbf{U}^T, \mathbf{V}^T] \begin{bmatrix} \mathbf{x} \\ t\mathbf{y} \end{bmatrix} \\
 &= (\mathbf{x}^T \mathbf{U} + t\mathbf{y}^T \mathbf{V}) (\mathbf{U}^T \mathbf{x} + t\mathbf{V}^T \mathbf{y}) \\
 &= (\mathbf{x}^T \mathbf{U} \mathbf{U}^T \mathbf{x}) + 2t (\mathbf{x}^T \mathbf{U} \mathbf{V}^T \mathbf{y}) + t^2 (\mathbf{y}^T \mathbf{V} \mathbf{V}^T \mathbf{y}) .
 \end{aligned}$$

This is quadratic in t . Let $\mathbf{z} = \mathbf{x}^T \mathbf{U}$ and $\mathbf{w} = \mathbf{y}^T \mathbf{V}$, and let θ be the angle between \mathbf{z} and \mathbf{w} . Then $Z(t)$ is of the form

$$Z(t) = at^2 + bt + c ,$$

where

$$\begin{aligned}
 a &= \mathbf{y}^T \mathbf{V} \mathbf{V}^T \mathbf{y} = \mathbf{w}^T \mathbf{w} = \|\mathbf{w}\|^2 , \\
 b &= 2\mathbf{x}^T \mathbf{U} \mathbf{V}^T \mathbf{y} = 2\mathbf{z}^T \mathbf{w} = 2 \|\mathbf{z}\| \|\mathbf{w}\| \cos \theta , \text{ and} \\
 c &= \mathbf{x}^T \mathbf{U} \mathbf{U}^T \mathbf{x} = \mathbf{z}^T \mathbf{z} = \|\mathbf{z}\|^2 .
 \end{aligned}$$

If the discriminant is nonnegative, there is a real root t such that $Z(t) = 0$; if the discriminant is zero, the root is unique.

$$\begin{aligned}
 \frac{b^2 - 4ac}{4} &= \|\mathbf{z}\|^2 \|\mathbf{w}\|^2 \cos^2 \theta - \|\mathbf{z}\|^2 \|\mathbf{w}\|^2 \\
 &= \|\mathbf{z}\|^2 \|\mathbf{w}\|^2 (\cos^2 \theta - 1) .
 \end{aligned}$$

This value is real and nonpositive. It is zero if and only if at least one of the conditions holds which are stated in the lemma. \square

Corollary 4.6.3 *If $R = 1$, there is a unique t such that $Z(t) = 0$.*

Proof. Apply Lemma 4.6.2, and notice that in this case $\mathbf{x}^T \mathbf{U}$ and $\mathbf{y}^T \mathbf{V}$ are scalars. \square

Lemma 4.6.4 *Let*

$$\Sigma = \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} ,$$

where Σ is symmetric positive semidefinite, A and B are respectively $p \times p$ and $q \times q$, and C is of rank one. Let u and v be p - and q -vectors satisfying

$$C = uv^T.$$

Define

$$A^* = A - uu^T,$$

$$B^* = B - vv^T.$$

Then at least one of A^* and B^* is positive semidefinite. Furthermore, if Σ is positive definite, at least one of A^* and B^* is positive definite.

Proof. Let x and y be the eigenvectors of A^* and B^* corresponding to their smallest eigenvalues, δ and ϵ , so that

$$A^*x = \delta x \quad \text{and} \quad B^*y = \epsilon y.$$

Let

$$w = \begin{bmatrix} u \\ v \end{bmatrix}, \quad Q = ww^T, \quad E = \begin{bmatrix} A^* & 0 \\ 0 & B^* \end{bmatrix},$$

so that

$$\Sigma = Q + E.$$

For real t , consider the following quadratic form:

$$\begin{aligned} [x^T, ty^T] \Sigma \begin{bmatrix} x \\ ty \end{bmatrix} &= [x^T, ty^T] Q \begin{bmatrix} x \\ ty \end{bmatrix} + x^T A^* x + t^2 y^T B^* y \\ &= [x^T, ty^T] Q \begin{bmatrix} x \\ ty \end{bmatrix} + \delta + t^2 \epsilon. \end{aligned}$$

By Corollary 4.6.3, there is a unique real t such that the first term is zero. Thus

$$\max(\delta, \epsilon) < 0 \Rightarrow \Sigma \text{ has a negative eigenvalue, and}$$

$$\max(\delta, \epsilon) \leq 0 \Rightarrow \Sigma \text{ has a nonpositive eigenvalue.}$$

By the contrapositive, it follows that, if Σ is strictly positive definite, then at least one of \mathbf{A}^* and \mathbf{B}^* is strictly positive definite; and if Σ is positive semidefinite, then at least one of \mathbf{A}^* and \mathbf{B}^* is positive semidefinite. \square

4.6.2 Facts related to generalized eigenvalues

The following facts are used in Section 4.3.2.

Proposition 4.6.5 *Let \mathbf{A} and \mathbf{C} be $n \times n$ matrices. If $\lambda \neq 0$, then*

$$|\mathbf{A} - \lambda \mathbf{C}| = 0 \Leftrightarrow \left| \mathbf{C} - \frac{1}{\lambda} \mathbf{A} \right| = 0 .$$

Proof. Multiply both sides of the expression on the left by $\left(-\frac{1}{\lambda}\right)^n$. \square

Proposition 4.6.6 *Let \mathbf{A} be nonsingular. Then any generalized eigenvalue λ satisfying $|\mathbf{C} - \lambda \mathbf{A}| = 0$ is an eigenvalue of $\mathbf{A}^{-1}\mathbf{C}$.*

Proof. For some nonzero \mathbf{x} we have

$$\begin{aligned} \mathbf{0} &= \mathbf{C}\mathbf{x} - \lambda \mathbf{A}\mathbf{x} \\ &= \mathbf{A}^{-1}\mathbf{C}\mathbf{x} - \lambda \mathbf{x} . \end{aligned}$$

\square

4.6.3 SVD paired latent models

The rank-one **SVD paired latent model** is equivalent to the rank-one paired latent correlation model specified at (4.2). The rank-one SVD paired latent model is the set of distributions over the latent variables ξ and ω , the observed variables \mathbf{X} and \mathbf{Y} , and the

errors ϵ and ζ , specified as follows.

$$\left. \begin{aligned} \mathbf{x} &= \mathbf{u}\xi + \epsilon, \\ \mathbf{y} &= \mathbf{v}\omega + \zeta, \\ \text{Var} \begin{bmatrix} \xi \\ \omega \end{bmatrix} &= \begin{bmatrix} \phi & d \\ d & \psi \end{bmatrix}, \\ \text{Var}(\epsilon) &= \Sigma_{\epsilon\epsilon}, \\ \text{Var}(\zeta) &= \Sigma_{\zeta\zeta}, \\ \epsilon \perp\!\!\!\perp \begin{bmatrix} \xi \\ \omega \end{bmatrix}, \quad \epsilon \perp\!\!\!\perp \zeta, \quad \begin{bmatrix} \xi \\ \omega \end{bmatrix} \perp\!\!\!\perp \zeta, \\ \mathbf{u} \in \mathbb{R}^p, \quad \mathbf{v} \in \mathbb{R}^q, \quad \|\mathbf{u}\| = \|\mathbf{v}\| = 1. \end{aligned} \right\} \text{where} \quad (4.30)$$

Thus the parameters of the SVD paired latent model are $\phi, \psi, d, \mathbf{u}, \mathbf{v}, \Sigma_{\epsilon\epsilon}$, and $\Sigma_{\zeta\zeta}$, subject to the following constraints:

$$\|\mathbf{u}\| = \|\mathbf{v}\| = 1, \quad (4.31)$$

$$\Sigma_{\epsilon\epsilon} \text{ positive semidefinite}, \quad (4.32)$$

$$\Sigma_{\zeta\zeta} \text{ positive semidefinite}, \quad (4.33)$$

$$\phi\psi \geq d^2. \quad (4.34)$$

The parameters of the SVD paired latent model may be partitioned into those which govern cross-covariance,

$$\mathbf{u}, \mathbf{v}, d,$$

and those which govern within-block covariance,

$$\phi, \psi, \Sigma_{\epsilon\epsilon}, \Sigma_{\zeta\zeta}.$$

Such a partition is not possible in the paired latent correlation model. Note that d is a covariance, not a correlation. The correlation between the latents, ρ , is a function of the parameters:

$$\rho = \frac{d}{\sqrt{\phi\psi}} . \quad (4.35)$$

A rank-one constraint model is mapped to an SVD paired latent model as follows. Using the singular value decomposition, determine d , \mathbf{u} , and \mathbf{v} by

$$\Sigma_{XY} = d\mathbf{u}\mathbf{v}^T , \quad (4.36)$$

where, as we have noted, a sign convention is necessary to make these parameters identifiable. Determine the constants ϕ_{\max} and ψ_{\max} by

$$\phi_{\max} \equiv \max \{ \phi : (\Sigma_{XX} - \phi\mathbf{u}\mathbf{u}^T) \text{ is positive semidefinite} \} , \quad (4.37)$$

$$\psi_{\max} \equiv \max \{ \psi : (\Sigma_{YY} - \psi\mathbf{v}\mathbf{v}^T) \text{ is positive semidefinite} \} .$$

By Lemma 4.6.1, ϕ_{\max} and ψ_{\max} exist. For future reference we recall line (4.14), page 90, noting thereby that

$$\phi_{\max} = \alpha_{\max}^2 , \quad \psi_{\max} = \frac{d^2}{\alpha_{\min}^2} . \quad (4.38)$$

Then each point in the two-dimensional feasible set,

$$\{(\phi, \psi) : \phi \leq \phi_{\max} , \psi \leq \psi_{\max} , \phi\psi \geq d^2\} , \quad (4.39)$$

represents a SVD parameterization of the rank constraint model. That this set is nonempty is a consequence of Theorem 4.3.1. For any (ϕ, ψ) in the feasible set, the remaining parameters are defined by

$$\begin{aligned} \Sigma_{\epsilon\epsilon}(\phi) &= \Sigma_{XX} - \phi\mathbf{u}\mathbf{u}^T , \\ \Sigma_{\zeta\zeta}(\psi) &= \Sigma_{YY} - \psi\mathbf{v}\mathbf{v}^T . \end{aligned} \quad (4.40)$$

To see that this parameterization is correct, it suffices to note that the following requirements are met.

1. The parameter values defined recover Σ —that is, this parameterization maps into the constraint model with which we started. This follows immediately from (4.36) and (4.40).
2. The constraints at (4.39) are necessary and sufficient for the parameters to be feasible—that is, for $\Sigma_{\epsilon\epsilon}$, $\Sigma_{\zeta\zeta}$, and $\begin{bmatrix} \phi & d \\ d & \psi \end{bmatrix}$ to be positive semidefinite. This follows immediately from the definitions of ϕ_{\max} and ψ_{\max} , and the condition $\phi\psi > d^2$ in the definition of the feasible set.

The values of ϕ and ψ corresponding to the single latent model form the lower left boundary of the feasible set, which is the intersection of the curve

$$\psi = \frac{d^2}{\phi}$$

with the rectangle $[0, \phi_{\max}] \times [0, \psi_{\max}]$. For constraint models such that $\alpha_{\max} = \alpha_{\min}$, the feasible set degenerates to the single point $(\phi_{\max}, \psi_{\max})$.

Define ϕ_{\min} and ψ_{\min} by

$$\phi_{\min} \equiv \frac{d^2}{\psi_{\max}}, \quad \psi_{\min} \equiv \frac{d^2}{\phi_{\max}}. \quad (4.41)$$

The constraints

$$\phi \geq \phi_{\min}, \quad \psi \geq \psi_{\min}$$

are implicit in the definition of the feasible set, since both ϕ and ψ have maximum values, and $\phi\psi \geq d^2$. At any point where ϕ or ψ attains its maximum the covariance of the latents is singular. The minimum value for ϕ is attained only when ψ attains its maximum, and vice versa.

The value $(\phi_{\max}, \psi_{\max})$ entails a nonsingular variance for $(\xi, \omega)^T$ whenever the feasible set is not degenerate. In all cases, however, $(\phi_{\max}, \psi_{\max})$ entails singular within-block error variances for both blocks. The minimum feasible correlation is attained at this point, and defined by

$$\rho_{\min} \equiv \frac{d}{\sqrt{\phi_{\max}\psi_{\max}}}.$$

Bijection between parameterizations. It was stated on page 111 that the SVD paired latent model and the paired latent correlation model are equivalent. This may be seen by the fact that a bijection exists between the feasible sets for the two parameterizations, as defined at (4.20) and at (4.39). Given a point (α, ρ) in the feasible set for the paired latent correlation model, we obtain a point in the feasible set for the SVD paired latent model by

$$\phi = \alpha^2, \quad \psi = \frac{d^2}{\alpha^2 \rho^2}.$$

Given a point (ϕ, ψ) in the feasible set for the SVD paired latent model, we obtain a point in the feasible set for the paired latent correlation model by

$$\alpha = \sqrt{\phi}, \quad \rho = \frac{d}{\sqrt{\phi\psi}}.$$

Example. Let us return to the constraint model (4.22) on page 94, and compare its SVD-model feasible set with its paired-correlation-model feasible set. Recall that

$$d = \sqrt{\frac{35}{2}}, \quad \alpha_{\min} = \frac{1}{30\sqrt{2030}}, \quad \alpha_{\max} = \sqrt{7}.$$

Comparing (4.14) on page 90 with (4.37) on page 113 we see that the constraints defining the SVD-model feasible set are obtained by

$$\phi_{\max} = \alpha_{\max}^2 = 7,$$

$$\psi_{\max} = \frac{d^2}{\alpha_{\min}^2} = \frac{225}{29} \approx 7.76.$$

Then we also have

$$\phi_{\min} = \frac{d^2}{\psi_{\max}} = \frac{203}{90} \approx 2.26,$$

$$\psi_{\min} = \frac{d^2}{\phi_{\max}} = \frac{5}{2} = 2.5,$$

$$\rho_{\min} = \frac{d}{\sqrt{\phi_{\max}\psi_{\max}}} = \frac{\sqrt{290}}{30} \approx 0.568.$$

The feasible set is plotted in Figure 4.8 on page 117. The point where $\rho = \rho_{\min}$ and consequently $\Sigma_{\epsilon\epsilon}(\phi)$ and $\Sigma_{\zeta\zeta}(\psi)$ are singular is in the upper right corner of the feasible set.

In Figure 4.4 on page 98, the level set for the paired latent correlation model, by contrast, $\rho = \rho_{\min}$ at the leftmost point of the feasible set. The points corresponding to the single latent model ($\rho = 1$) in this Figure are on the lower left curved boundary, whereas in Figure 4.4 they are on the right boundary.

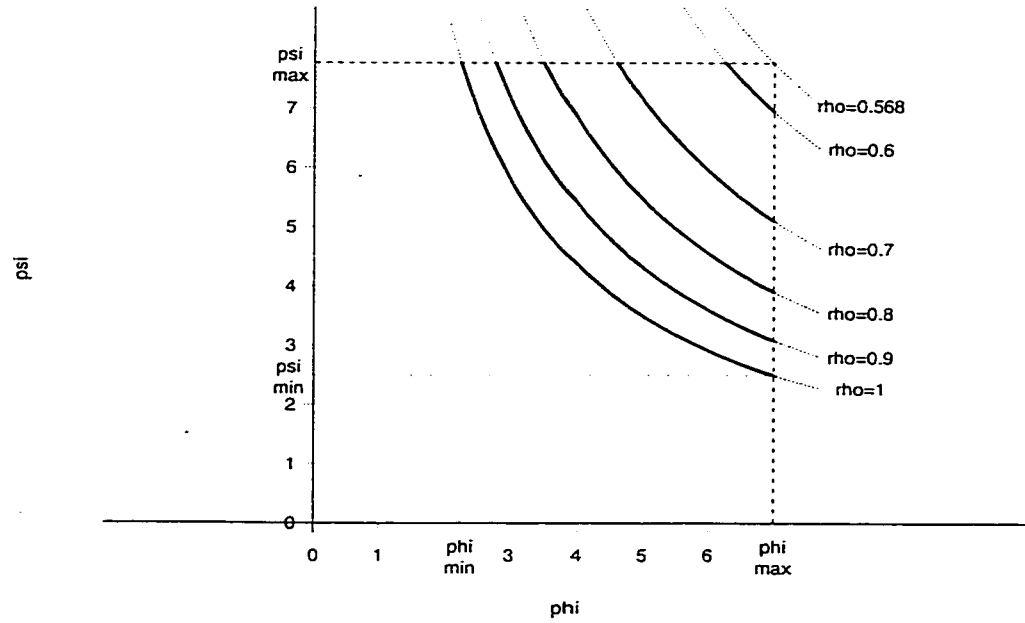


Figure 4.8: Feasible set for the SVD parameterization of the constraint model (4.22) on page 94. The SVD paired latent model is defined in Section 4.6.3. A bijection exists between this feasible set and the feasible set plotted in Figure 4.4 on page 98, representing the paired latent correlation parameterization of the same constraint model. Level sets of ρ lie on the curves $\psi = \frac{d^2}{\rho^2 \phi}$. Feasible values for ρ are in the closed interval $\left[\frac{\sqrt{290}}{30}, 1\right]$. At $\rho = \frac{\sqrt{290}}{30} \approx 0.568$, the level set is the point $\left(7, \frac{225}{29}\right)$, that is, the upper right corner of the broken rectangle. A selection of level sets for ρ are plotted as solid curves. Feasible values for ϕ and ψ lie inside the broken rectangle and on or to the right of the $\rho = 1$ curve.

Chapter 5

RANK- R LATENT MODELS FOR CROSS-COVARIANCE

Abstract. We specify a class of rank- r latent models for cross-covariance. We show by construction that any variance-covariance matrix for the observed variables induced by rank- r reduced-rank regression can be induced by a rank- r latent model.

This chapter parallels Chapter 4, with fewer examples. Theorem 5.2.1 is new.

5.1 *Model specification*

Basic terms are introduced which will be used to state the result.

5.1.1 *Rank- r constraint models*

Let p be the number of \mathbf{X} -variables and q the number of \mathbf{Y} -variables. The rank- r symmetric **constraint model** (equivalently, the rank- r reduced-rank-regression model) is the set of $(p + q) \times (p + q)$ positive semidefinite matrices satisfying a rank constraint on the cross-covariance matrix:

$$\left. \begin{aligned} \Sigma &= \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}, \\ \text{where } \Sigma_{XY} &\text{ is } p \times q \text{ of rank } r. \end{aligned} \right\} \quad (5.1)$$

5.1.2 *Rank- r paired latent models*

The rank- r symmetric **paired latent model** is the set of covariances over the latent r -vectors ξ and ω , the observed p -vector \mathbf{X} , the p -vector of errors ϵ , the observed q -vector \mathbf{Y} , and the q -vector of errors ζ , specified as follows.

$$\left. \begin{aligned}
 \mathbf{x} &= \mathbf{A}\boldsymbol{\xi} + \boldsymbol{\epsilon} \quad \text{and} \\
 \mathbf{y} &= \mathbf{B}\boldsymbol{\omega} + \boldsymbol{\zeta} \quad , \text{ where} \\
 \text{Var}(\boldsymbol{\xi}^T, \boldsymbol{\omega}^T)^T &= \begin{bmatrix} \mathbf{I}_r & \mathbf{R} \\ \mathbf{R} & \mathbf{I}_r \end{bmatrix} , \\
 \mathbf{R} &= \text{diag}(\rho_1, \dots, \rho_r) \quad , \\
 \text{Var}(\boldsymbol{\epsilon}) &= \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}}, \quad p \times p, \\
 \text{Var}(\boldsymbol{\zeta}) &= \boldsymbol{\Sigma}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}, \quad q \times q, \\
 \boldsymbol{\epsilon} \perp\!\!\!\perp (\boldsymbol{\xi}^T, \boldsymbol{\omega}^T)^T, \quad \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\zeta}, \quad (\boldsymbol{\xi}^T, \boldsymbol{\omega}^T)^T \perp\!\!\!\perp \boldsymbol{\zeta}, \\
 \mathbf{A} \in \mathbb{R}^{(p \times r)}, \quad \mathbf{B} \in \mathbb{R}^{(q \times r)} .
 \end{aligned} \right\} \quad (5.2)$$

Thus the parameters of the symmetric rank- r paired latent model are the correlations ρ_1, \dots, ρ_r and the matrices \mathbf{A} , \mathbf{B} , $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}$, subject to the feasibility constraints that $-1 \leq \rho_k \leq 1$ for all k and that $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}$ must be positive semidefinite. The observed variables \mathbf{X} and \mathbf{Y} are called **indicators**, and the columns of \mathbf{A} and \mathbf{B} are called **saliences** or **loadings**. A path diagram for this model may be seen in Figure 5.1 on page 120.

5.1.3 Rank- r single latent models

The rank- r symmetric **single latent model** is the set of distributions over the latent r -vector $\boldsymbol{\eta}$, the observed p -vector \mathbf{X} , the p -vector of errors $\boldsymbol{\epsilon}$, the observed q -vector \mathbf{Y} , and

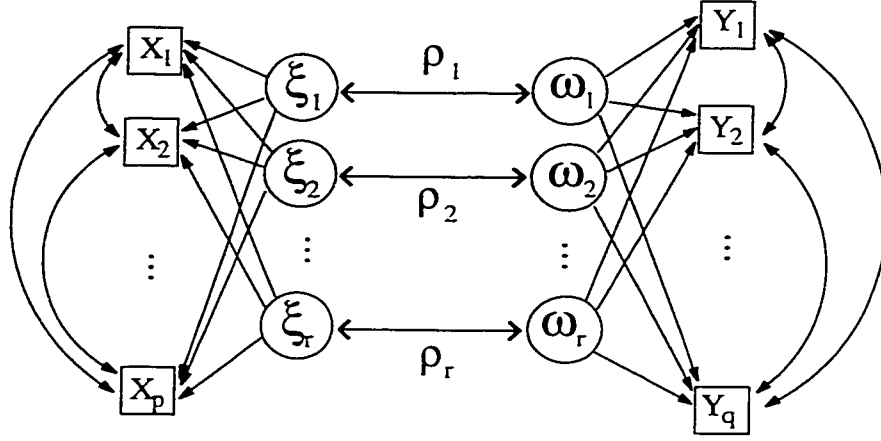


Figure 5.1: Path diagram representing a symmetric rank- r paired latent model. This model is defined in Section 5.1.2. In the parameterization guaranteed by Theorem 5.2.1, $\rho_1 = \dots = \rho_r = 1$ since $\xi_k \equiv \omega_k$ for all k .

the q -vector of errors ζ , specified as follows.

$$\left. \begin{aligned}
 \mathbf{x} &= \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\epsilon} \quad \text{and} \\
 \mathbf{y} &= \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}, \quad \text{where} \\
 \text{Var}(\boldsymbol{\eta}) &= \mathbf{I}_R, \\
 \text{Var}(\boldsymbol{\epsilon}) &= \boldsymbol{\Sigma}_{\epsilon\epsilon}, \quad p \times p, \\
 \text{Var}(\boldsymbol{\zeta}) &= \boldsymbol{\Sigma}_{\zeta\zeta}, \quad q \times q, \\
 \boldsymbol{\epsilon} &\perp\!\!\!\perp \boldsymbol{\eta}, \quad \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\zeta}, \quad \boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{\zeta}, \\
 \mathbf{A} &\in \mathbb{R}^{(p \times r)}, \quad \mathbf{B} \in \mathbb{R}^{(q \times r)}.
 \end{aligned} \right\} \quad (5.3)$$

Thus the parameters of the symmetric rank- r single latent model are the matrices \mathbf{A} , \mathbf{B} , $\boldsymbol{\Sigma}_{\epsilon\epsilon}$, and $\boldsymbol{\Sigma}_{\zeta\zeta}$, subject to the feasibility constraint that $\boldsymbol{\Sigma}_{\epsilon\epsilon}$ and $\boldsymbol{\Sigma}_{\zeta\zeta}$ must both be positive semidefinite. The reader will observe that the rank- r single latent model is a special case of

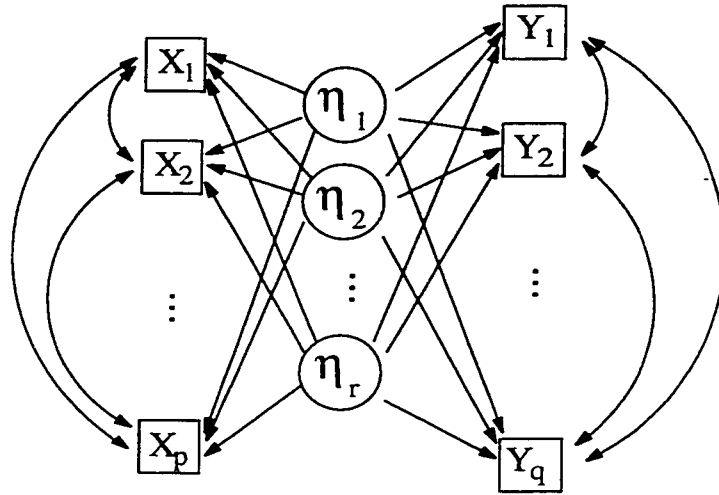


Figure 5.2: Path diagram representing a symmetric rank- r single latent model.

the rank- r paired latent model where $\xi \equiv \omega$. A path diagram for a symmetric rank- r single latent model may be seen in Figure 5.2 on page 121.

5.2 Maps between spaces of models

Every set of parameter values for the paired latent model induces a set of covariances over the observed variables as follows.

$$\left. \begin{aligned} \Sigma_{XX} &= \mathbf{A}\mathbf{A}^T + \Sigma_{\epsilon\epsilon}, \\ \Sigma_{YY} &= \mathbf{B}\mathbf{B}^T + \Sigma_{\zeta\zeta}, \\ \Sigma_{XY} &= \mathbf{A}\mathbf{R}\mathbf{B}^T. \end{aligned} \right\} \quad (5.4)$$

The equations (5.4) define a map from the space of rank- r paired latent models into the space of rank- r constraint models. The existence of such a map immediately raises the question whether every covariance in the rank- r constraint model can be obtained by a set of parameter values in the rank- r latent model—i.e., is the map onto. If such a set of parameter values exists, we say that it **parameterizes** or is a **latent parameterization** of the covariance matrix.

The answer to the question in the previous paragraph is yes. Every rank- r constraint

model can be parameterized by a rank- r symmetric paired latent model. We show this by first proving a stronger result, i.e., that any rank- r constraint model can be parameterized by a symmetric rank- r single latent model. The result concerning paired latent models is obtained as a corollary.

5.2.1 A theorem regarding rank- r single latent models

We now state and prove the main result.

Theorem 5.2.1 *For each covariance matrix (5.1) in the rank- r constraint model there is at least one set of parameter values in the rank- r symmetric single latent model which induces it.*

Proof. Let \mathbf{X} and \mathbf{Y} be matrices such that

$$\begin{aligned}\Sigma_{XX} &= \mathbf{X}^T \mathbf{X} , \\ \Sigma_{XY} &= \mathbf{X}^T \mathbf{Y} , \text{ and} \\ \Sigma_{YY} &= \mathbf{Y}^T \mathbf{Y} .\end{aligned}$$

Such matrices are guaranteed to exist. For instance they may be obtained by partitioning the symmetric positive semidefinite square root of Σ . Let r_x be the rank of \mathbf{X} and r_y the rank of \mathbf{Y} . Without loss of generality suppose $r_x \leq r_y$. It can be shown that there are matrices \mathbf{U} and \mathbf{V} such that \mathbf{U} is a basis for the range (the column space) of \mathbf{X} , \mathbf{V} is a basis for the range of \mathbf{Y} , $\mathbf{U}^T \mathbf{U} = \mathbf{I}_{r_x}$, $\mathbf{V}^T \mathbf{V} = \mathbf{I}_{r_y}$, and

$$\mathbf{U}^T \mathbf{V} = [\mathbf{D} | \mathbf{0}] , \quad (5.5)$$

where $\mathbf{0}$ is an $r_x \times (r_y - r_x)$ matrix of zeroes, absent if $r_x = r_y$, \mathbf{D} is an $r_x \times r_x$ diagonal matrix satisfying

$$\mathbf{D} = \text{diag}(\cos(\theta_1), \dots, \cos(\theta_r), 0, \dots, 0) , \quad (5.6)$$

the last $(r_x - r)$ diagonal entries of (5.6) are zero if $r_x > r$, and

$$\cos(\theta_1) \geq \dots \geq \cos(\theta_r) > 0 .$$

The columns of \mathbf{U} and \mathbf{V} are **principal vectors**, and the θ_k are **principal angles**. These facts are reported as Theorem 9.1 and Corollary 9.11 in Afriat [Afr57]. Golub and van Loan report an algorithm for computing \mathbf{U} and \mathbf{V} in the case when \mathbf{X} and \mathbf{Y} are of full rank (pages 603f [GVL96]). Björck and Golub [BG73] discuss numerical methods, including the case where \mathbf{X} and \mathbf{Y} are rank-deficient. In a statistical context the $\cos(\theta_k)$ are known as **canonical correlations** and the principal vectors as **canonical correlation variables** or **canonical variates**. Mardia, Kent and Bibby develop these concepts within a statistical context for the case where Σ has full rank (Chapter 10, pages 281–299 [MKB79]), as does T. W. Anderson [And99]. The `cancor()` function in S-PLUS [Mat96] may be used to compute canonical correlations. S-PLUS also computes two matrices, respectively $r_x \times r_x$ and $r_y \times r_y$, which may be used to compute \mathbf{U} and \mathbf{V} from \mathbf{X} and \mathbf{Y} provided \mathbf{X} and \mathbf{Y} have full rank.

Let n be the number of rows in \mathbf{X} . Then \mathbf{U} is $n \times r_x$ and \mathbf{V} is $n \times r_y$. Let \mathbf{E} be an $r_x \times p$ matrix and \mathbf{F} an $r_y \times q$ matrix such that

$$\mathbf{X} = \mathbf{U}\mathbf{E}, \quad \mathbf{Y} = \mathbf{V}\mathbf{F}. \quad (5.7)$$

Define the $p \times r_x$ matrix \mathbf{A} and the $q \times r_x$ matrix \mathbf{B} by

$$\begin{aligned} \mathbf{A} &= \mathbf{E}^T \sqrt{\mathbf{D}}, \\ \mathbf{B} &= \mathbf{F}^T \begin{bmatrix} \sqrt{\mathbf{D}} \\ \mathbf{0}^T \end{bmatrix}, \end{aligned}$$

where \mathbf{D} and $\mathbf{0}$ have the same value as in (5.5). Then by (5.5)

$$\begin{aligned} \mathbf{A}\mathbf{B}^T &= \mathbf{E}^T \mathbf{U}^T \mathbf{V} \mathbf{F} \\ &= \mathbf{X}^T \mathbf{Y}. \end{aligned}$$

Then

$$\begin{aligned} \Sigma_{XX} - \mathbf{A}\mathbf{A}^T &= \mathbf{X}^T \mathbf{X} - \mathbf{E}^T \mathbf{D} \mathbf{E} \\ &= \mathbf{E}^T \mathbf{U}^T \mathbf{U} \mathbf{E} - \mathbf{E}^T \mathbf{D} \mathbf{E} \\ &= \mathbf{E}^T (\mathbf{I}_{r_x} - \mathbf{D}) \mathbf{E} \\ &= \mathbf{E}^T \text{diag}(1 - \cos(\theta_1), \dots, 1 - \cos(\theta_r), 1, \dots, 1) \mathbf{E}, \end{aligned} \quad (5.8)$$

a positive semidefinite matrix. By a similar argument

$$\begin{aligned}\Sigma_{YY} - \mathbf{B}\mathbf{B}^T &= \mathbf{F}^T \left(\mathbf{I}_q - \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \mathbf{F} \\ &= \mathbf{F}^T \text{diag}(1 - \cos(\theta_1), \dots, 1 - \cos(\theta_r), 1, \dots, 1) \mathbf{F},\end{aligned}\quad (5.9)$$

also positive semidefinite. Define the $p \times p$ matrix $\Sigma_{\epsilon\epsilon}$ and the $q \times q$ matrix $\Sigma_{\zeta\zeta}$ by

$$\begin{aligned}\Sigma_{\epsilon\epsilon} &= \Sigma_{XX} - \mathbf{A}\mathbf{A}^T, \\ \Sigma_{\zeta\zeta} &= \Sigma_{YY} - \mathbf{B}\mathbf{B}^T.\end{aligned}$$

The values of \mathbf{A} , \mathbf{B} , $\Sigma_{\epsilon\epsilon}$, and $\Sigma_{\zeta\zeta}$ satisfy the definition of a rank- r single latent model, stated in (5.3), and they induce Σ . \square

Corollary 5.2.2 *The values of both $\Sigma_{\epsilon\epsilon}$ and $\Sigma_{\zeta\zeta}$, derived in the proof of Theorem 5.2.1, are strictly positive definite if and only if Σ is strictly positive definite.*

Proof. Σ is strictly positive definite if and only if the columns of the combined matrix $[\mathbf{X}|\mathbf{Y}]$ are linearly independent. This condition holds if and only if the following three conditions hold.

1. The columns of \mathbf{X} are linearly independent of the columns of \mathbf{Y} , so that the first principal angle satisfies $\cos(\theta_1) < 1$ (the first canonical correlation is less than one in absolute value). Note that this is the only way that

$$\text{diag}(1 - \cos(\theta_1), \dots, 1 - \cos(\theta_r), 1, \dots, 1)$$

can have full rank.

2. The following equivalent conditions hold.

- The columns of \mathbf{X} are linearly independent.
- $r_x = p$.
- $\text{rank}(\mathbf{E}) = p$.

3. The following equivalent conditions hold.

- The columns of \mathbf{Y} are linearly independent.
- $r_y = q$.
- $\text{rank}(\mathbf{F}) = q$.

Thus if Σ is strictly positive definite, both (5.9) and (5.8) are of full rank, that is, strictly positive definite.

Suppose on the other hand that (5.9) and (5.8) are of full rank. The matrix at (5.9) is $p \times p$, the product of a $p \times r_x$ matrix, an $r_x \times r_x$ matrix, and an $r_x \times p$ matrix. Since $r_x \leq p$, this matrix can be of full rank only if $r_x = p$. Furthermore it can be of full rank only if the middle matrix is of full rank, which requires $\rho_1 < 0$. Similarly if (5.8) is of full rank it follows that $r_y = q$ and $\rho_1 < 0$. Thus Σ is of full rank. \square

Remark on Corollary 5.2.2. Corollary 5.2.2 notwithstanding, for a given strictly positive definite covariance matrix in the constraint model there may be parameterizations, different from those derived in the proof of Theorem 5.2.1, with singular within-block covariance. For instance, Wegelin et al. [WRR01] show in the rank-one case that a parameterization is always possible in which the within-block-error covariance matrices are singular for both blocks.

Corollary 5.2.3 *Each rank- r constraint model can be parameterized by at least one rank- r paired latent model.*

Proof. Let η be the latent variable of the single latent model be given by Theorem 5.2.1, and let $\xi \equiv \omega \equiv \eta$.

Remark on Corollary 5.2.3. The correlations between ξ_k and ω_k , written ρ_k , are not to be confused with the canonical correlations which appear in the proof of Theorem 5.2.1. The correlation between the latents in the paired latent parameterization guaranteed by Theorem 5.2.1 is unity: $\text{Cor}(\xi_k, \omega_k) = 1$ for all k . The canonical correlation which appears

in the proof of Theorem 5.2.1, on the other hand, is only unity if Σ is singular. In the example on page 126, for instance, the canonical correlation is about 0.57.

5.3 Example

Consider the following symmetric strictly positive definite matrix, and let $p = 2$, $q = 3$. This is a distribution in the rank-one constraint model, and is identical to (4.22) on page 94, except that \mathbf{X} and \mathbf{Y} have been transposed.

$$\Sigma = \begin{bmatrix} 9 & 0 & 1 & 2 & 3 \\ 0 & 5 & 0.5 & 1 & 1.5 \\ 1 & 0.5 & 7 & 0 & 0 \\ 2 & 1 & 0 & 7 & 0 \\ 3 & 1.5 & 0 & 0 & 7 \end{bmatrix}.$$

Following the proof of Theorem 5.2.1, we obtain

$$\mathbf{U} = \begin{bmatrix} 0.79431 & 0.56228 \\ 0.5326 & -0.82666 \\ 0.07811 & -0.0058 \\ 0.15622 & -0.0116 \\ 0.23433 & -0.0174 \end{bmatrix},$$

$$\mathbf{V} = \begin{bmatrix} 0.2615 & 0 & 0 \\ 0.15284 & 0 & 0 \\ 0.25471 & -0.9443 & 0.19202 \\ 0.50941 & -0.02053 & -0.8449 \\ 0.76412 & 0.32845 & 0.49926 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} 0.56765 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathbf{U}_1 = \begin{bmatrix} 0.75342 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathbf{E} = \begin{bmatrix} 2.49136 & 1.24568 \\ 1.67126 & -1.85695 \end{bmatrix},$$

$$\mathbf{F} = \begin{bmatrix} 0.70711 & 1.41421 & 2.12132 \\ -2.49838 & -0.05431 & 0.869 \\ 0.50805 & -2.23541 & 1.32092 \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} 1.87705 & 0 \\ 0.93853 & 0 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0.53275 & 0 \\ 1.0655 & 0 \\ 1.59825 & 0 \end{bmatrix}.$$

We check that

$$\mathbf{AB}^T = \begin{bmatrix} 1 & 2 & 3 \\ 0.5 & 1 & 1.5 \end{bmatrix} = \Sigma_{XY}.$$

Then the within-block covariances are

$$\Sigma_{\epsilon\epsilon} = \begin{bmatrix} 5.47668 & -1.76166 \\ -1.76166 & 4.11917 \end{bmatrix},$$

full rank with least eigenvalue 2.9, and

$$\Sigma_{\zeta\zeta} = \begin{bmatrix} 6.71618 & -0.56765 & -0.85147 \\ -0.56765 & 5.86471 & -1.70294 \\ -0.85147 & -1.70294 & 4.44559 \end{bmatrix},$$

also full rank with least eigenvalue 3.03.

5.4 Discussion

Likelihood and the equivalence of model spaces. Three spaces of covariance matrices over the observed variables \mathbf{X} and \mathbf{Y} are of interest in the current work. They are:

1. Those corresponding to the rank- r rank-constraint, or reduced-rank-regression, model.
2. Those induced by the rank- r symmetric paired latent model.
3. Those induced by the rank- r symmetric single latent model.

It follows from definitions and from Equations (5.4) on page 121 that $\text{Set } 3 \subset \text{Set } 2 \subset \text{Set } 1$. Theorem 5.2.1, however, implies that $\text{Set } 1 \subset \text{Set } 3$. Hence $\text{Set } 1 = \text{Set } 2 = \text{Set } 3$, a fact which we state as the following corollary.

Corollary 5.4.1 *The sets of covariance matrices over the observed variables induced by the rank- r symmetric paired latent correlation model and the rank- r symmetric single latent model are equal to the set of covariance matrices belonging to the rank- r constraint model.*

Thus all single and paired latent parameterizations within an equivalence class have the same likelihood under the multivariate normal model for $(\mathbf{X}^T, \mathbf{Y}^T)^T$, and consequently there is no way using only data to distinguish between the three models. Furthermore it may be shown that the rank-one constraint model is covariance equivalent to reduced-rank regression (RRR). This fact is reviewed in Appendix A. Maximum-likelihood estimation procedures, and asymptotics, available for RRR (see Anderson [And51] [And80] [And99]) and Ryan et al. [RHC⁺92]) are thus available for the paired and single latent models.

Chapter 6

RELATED EQUIVALENCE RESULTS

In this chapter we extend the results described so far by considering a number of other latent models which relate the two blocks of observed variables.

These graphs are shown in Figure 6.1. (a) and (b) represent two path diagrams in which

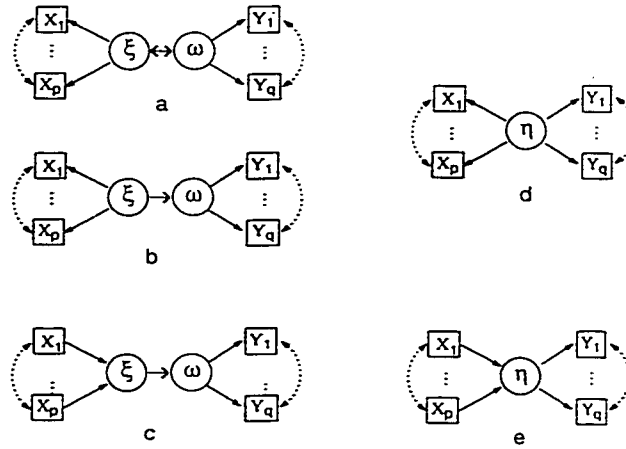


Figure 6.1: Path diagrams corresponding to two-block latent variable models. Under (I) the dashed edges are present; under (II) they are absent. Under (I) all models are covariance equivalent over \mathbf{X} and \mathbf{Y} .

the latent variables ξ and ω are parents of the observed variables. The only difference between the models is that (a) specifies that ξ and ω are correlated, while in (b) ξ is a parent of ω . The graph shown in Figure 6.1 (c) differs from that shown in (b) in that the \mathbf{X} variables are parents of ξ . The graph in (d) is analogous to (a) and (b) but the pair of latent variables ξ, ω are replaced with a single variable. Likewise (e) represents the single latent analogue to (c).

We consider the five models corresponding to these graphs, under two sets of conditions

on the error terms:

(I) $\text{Cov}(\epsilon_i, \zeta_j) = 0$, but $\text{Cov}(\epsilon_i, \epsilon_k)$ and $\text{Cov}(\zeta_j, \zeta_\ell)$ are unrestricted.

(II) $\text{Cov}(\epsilon_i, \zeta_j) = 0$, and

$$\text{Cov}(\epsilon_i, \epsilon_k) = \text{Cov}(\zeta_j, \zeta_\ell) = 0 \text{ for } i \neq k, j \neq \ell.$$

Let \mathcal{N}_a^I denote the set of Gaussian distributions over \mathbf{X} and \mathbf{Y} given by graph (a) in Figure 6.1 under condition (I) on the errors, likewise for \mathcal{N}_a^{II} , \mathcal{N}_b^I , \mathcal{N}_b^{II} and so on. Corollary 4.5.1 thus shows that $\mathcal{N}_a^I = \mathcal{N}_d^I$. We extend these results further in the next theorem.

Theorem 6.0.2 *The following relations hold:*

$$\mathcal{N}_a^I = \mathcal{N}_b^I = \mathcal{N}_c^I = \mathcal{N}_d^I = \mathcal{N}_e^I$$

$$\mathcal{N}_a^{II} = \mathcal{N}_b^{II} \neq \mathcal{N}_c^{II} = \mathcal{N}_e^{II} \neq \mathcal{N}_d^{II} \neq \mathcal{N}_a^{II}$$

(The first and third inequalities require $p > 1$. The second also requires $q > 1$.)

In words: When the within-block errors are not restricted, all of the latent structures in Figure 6.1 are indistinguishable. When the errors are uncorrelated, on the other hand, the following conditions hold:

- We can distinguish structures in which ξ is a parent of the \mathbf{X} 's from those in which the \mathbf{X} 's are parents of ξ .
- When the \mathbf{X} 's are parents of ξ we cannot distinguish between models with one and two latent variables.
- When ξ is a parent of the \mathbf{X} 's we can distinguish models with two latent variables from those containing only one.

The existence of equivalent models containing different numbers of hidden variables is important for the purpose of interpretation. It highlights the danger of postulating the existence of variables for which there is no evidence in the data.

6.0.1 PROOFS OF EQUIVALENCE RESULTS

In order to prove the results in Theorem 6.0.2 we need several definitions. Following [RS00] we say that a path diagram, which may contain directed edges (\rightarrow) and bi-directed edges (\leftrightarrow) is *ancestral* if:

- (a) there are no directed cycles;
- (b) if there is an edge $x \leftrightarrow y$ then x is not an ancestor of y , (and vice versa);

where a vertex x is said to be an *ancestor* of y if either $x = y$ or there is a directed path from x to y . Conditions (a) and (b) may be summarized by saying that if x and y are joined by an edge and there is an arrowhead at x then x is *not* an ancestor of y ; this is the motivation for the term ‘ancestral’. (In [RS00] a more general version of this definition is given which applies to graphs containing undirected edges.)

A natural extension of Pearl’s d-separation criterion may be applied to graphs containing directed and bi-directed edges. A non-endpoint vertex v on a path is said to be a *collider* if two arrowheads meet at v , i.e. $\rightarrow v \leftarrow$, $\leftrightarrow v \leftrightarrow$, $\leftrightarrow v \leftarrow$ or $\rightarrow v \leftrightarrow$; all other non-endpoint vertices on a path are *non-colliders*. A path π between α and β is said to be *m-connecting given Z* if the following hold:

- (i) no non-collider on π is on Z ;
- (ii) every collider on π is an ancestor of a vertex in Z .

Two vertices α and β are said to be **m-separated given Z** if there is no path m-connecting α and β given Z . Disjoint sets of vertices A and B are said to be **m-separated given Z** if there is no pair α, β with $\alpha \in A$ and $\beta \in B$ such that α and β are m-connected given Z . (This an extension of the original definition of d-separation for DAGs in that the notions of ‘collider’ and ‘non-collider’ now include bi-directed edges.) Two graphs \mathcal{G}_1 and \mathcal{G}_2 are said to be *Markov equivalent* if for all disjoint sets A, B, Z (where Z may be empty), A and B are m-separated given Z in \mathcal{G}_1 if and only if A and B are m-separated given Z in

\mathcal{G}_2 . A distribution P is said to obey the *global Markov property with respect to graph \mathcal{G}* if $A \perp\!\!\!\perp B \mid Z$ in P whenever A is m-separated from B given Z in \mathcal{G} .

An ancestral graph is said to be *maximal* if for every pair of non-adjacent vertices α, β there exists some set Z such that α and β are m-separated given Z .

It is proved in [RS00] that the set of Gaussian distributions given by parameterizing the path diagram \mathcal{G} is exactly the set of Gaussian distributions that obey the global Markov property with respect to \mathcal{G} . More formally, we have:

Theorem 6.0.3 *If \mathcal{G} is a maximal ancestral graph then the following equality holds regarding Gaussian distributions:*

$$\begin{aligned} & \{N \mid N \text{ results from some assignment of} \\ & \quad \text{parameter values to } \mathcal{G}\} \\ & = \{N \mid N \text{ satisfies the global Markov property for } \mathcal{G}\}. \end{aligned}$$

See Theorem 8.14 in [RS00]. As an immediate Corollary we have:

Corollary 6.0.4 *If \mathcal{G}_1 and \mathcal{G}_2 are two Markov equivalent maximal ancestral graphs then they parameterize the same sets of Gaussian distributions.*

See Corollary 8.19 in [RS00]. These results do not generally hold for path diagrams which are not both maximal and ancestral.

The sets of distributions given by the models under (I) correspond to the path diagrams shown in Figure 6.1 in which there are bi-directed edges between all variables within the same block, thus $\mathbf{X}_i \leftrightarrow \mathbf{X}_k$ ($i \neq k$) and $\mathbf{Y}_j \leftrightarrow \mathbf{Y}_\ell$ ($j \neq \ell$).

6.0.2 PROOF OF THEOREM 6.0.2

We first show $\mathcal{N}_a^I = \mathcal{N}_b^I = \mathcal{N}_c^I$. Observe that in each of the graphs in Figure 6.1(a), (b) and (c), the following m-separation relations hold:

- (i) \mathbf{X}_i is m-separated from \mathbf{Y}_j by any non-empty subset of $\{\xi, \omega\}$;
- (ii) \mathbf{X}_i is m-separated from ω by ξ ;

(iii) \mathbf{Y}_j is m-separated from ξ by ω .

Further, when bi-directed edges are present between vertices within each block all other pairs of vertices are adjacent so there are no other m-separation relations. Consequently these graphs are Markov equivalent and maximal since there is a separating set for each pair of non-adjacent vertices. It then follows directly by Corollary 6.0.4 that these graphs parameterize the same sets of distributions over the set $\{\mathbf{X}, \mathbf{Y}, \omega, \xi\}$, hence they induce the same sets of distributions on the margin over $\{\mathbf{X}, \mathbf{Y}\}$.

The proof that $\mathcal{N}_d^I = \mathcal{N}_e^I$ is very similar. When bi-directed edges are present within each block the only pairs of non-adjacent vertices are \mathbf{X}_i and \mathbf{Y}_j which are m-separated by ξ . It then follows as before that these graphs are Markov equivalent and maximal and hence by Corollary 6.0.4 they parameterize the same sets of distributions over $\{\mathbf{X}, \mathbf{Y}, \xi\}$, and consequently over $\{\mathbf{X}, \mathbf{Y}\}$.

Since we have already shown $\mathcal{N}_a^I = \mathcal{N}_d^I$ in Corollary 4.5.1, the proof of equivalences concerning models with error structure given by (I) is complete. It remains to prove the results concerning models of type (II). These correspond to the path diagrams in Figure 6.1, without the dashed edges between vertices within the same block. Subsequent references to graphs in this figure will be to the graphs without these within-block edges.

First note that the m-separation relations given by (i), (ii), (iii) above continue to hold when there are no edges between vertices within each block. In graphs (a) and (b) we also have:

(iv) \mathbf{X}_i and \mathbf{X}_j are m-separated given ξ ;

(v) \mathbf{Y}_i and \mathbf{Y}_j are m-separated given ω .

Consequently these graphs are Markov equivalent and maximal. Hence $\mathcal{N}_a^{II} = \mathcal{N}_b^{II}$ by Corollary 6.0.4. In the path diagrams corresponding to (c) and (e), we have

(vi) \mathbf{X}_i and \mathbf{X}_j are m-separated by the empty set.

Consequently the variables in the \mathbf{X} block are marginally independent in $\mathcal{N}_c^{\text{II}} = \mathcal{N}_e^{\text{II}}$, while this is not so under $\mathcal{N}_a^{\text{II}}, \mathcal{N}_b^{\text{II}}, \mathcal{N}_d^{\text{II}}$. This establishes two of the inequalities. By direct calculation it may be seen that for any distribution in $\mathcal{N}_d^{\text{II}}$ it holds that

$$\begin{aligned} \text{Cov}(\mathbf{X}_i, \mathbf{X}_k) \text{Cov}(\mathbf{Y}_j, \mathbf{Y}_\ell) &= \\ \text{Cov}(\mathbf{X}_i, \mathbf{Y}_j) \text{Cov}(\mathbf{X}_k, \mathbf{Y}_\ell) \end{aligned}$$

while this does not hold for distributions in $\mathcal{N}_a^{\text{II}} = \mathcal{N}_b^{\text{II}}$. This establishes the third inequality. It only remains to show that $\mathcal{N}_c^{\text{II}} = \mathcal{N}_e^{\text{II}}$. First observe that the set of m-separation relations which hold among $\{\mathbf{X}, \mathbf{Y}, \omega\}$ in the graph in (c), i.e. (i), (ii), (iii) and (v), is identical to the set of relations holding among $\{\mathbf{X}, \mathbf{Y}, \eta\}$ in (e), i.e. (i), (ii), (iii) and

(vii) \mathbf{Y}_j and \mathbf{Y}_ℓ are m-separated by η ,

where η is substituted for ω . Consequently any marginal distribution over $\{\mathbf{X}, \mathbf{Y}, \xi\}$ that is obtained from the graph in (c) may also be parameterized by the graph in (e) after substituting η for ω . It then follows that $\mathcal{N}_c^{\text{II}} \subseteq \mathcal{N}_e^{\text{II}}$. To prove the opposite inclusion it is sufficient to observe that any distribution over $\{\mathbf{X}, \mathbf{Y}, \eta\}$ that is parameterized by the graph in (e) may be parameterized by the graph in (c) by setting $\omega = \xi + \epsilon_\omega$ and letting $\text{Var}(\epsilon_\omega) + \text{Var}(\epsilon_\xi) = \text{Var}(\epsilon_\eta)$. This completes the proof. \square

Chapter 7

FUTURE WORK

7.1 Future Work

In the remainder of this chapter some directions for extension of the current work are outlined.

Extension of results to rank $r > 1$. In Chapter 4, in the context of a rank-one latent model, the sets of parameter values are characterized which induce a given covariance matrix. This is done for both the single and paired rank-one latent models. Two constants, α_{\min} and α_{\max} , functions of the covariance Σ , are sufficient to characterize each set. Furthermore a natural convention is presented by which the rank-one paired latent model can be made identifiable. The convention is that the correlation between the latent variables attains its minimum feasible value, $\frac{\alpha_{\min}}{\alpha_{\max}}$.

It would be interesting to see whether the results for the rank-one case can be extended to rank r ; that is, answers to the following questions would be interesting.

- How is the feasible set characterized for rank r ? Is it always a “corner” in $\mathbb{R}^{(2r)}$, as is the case when $r = 1$? Is there any way to visualize it?
- Are there constants, analogous to α_{\min} and α_{\max} , which characterize the feasible set for rank r ?
- Is there a single point in the feasible set which minimizes a reasonable criterion related to the correlations between the r pairs of latent variables? In particular, can the r correlations be simultaneously minimized, or does the minimization of, for instance, $\text{Cor}(\xi_1, \omega_1)$ imply that $\text{Cor}(\xi_2, \omega_2)$ is not minimized? In other words, are these parameters variation independent?

Likelihood. (See the following references: [Ize75] [Hea80] [HS79] [CM79])

The current work suggests a way that maximum likelihood estimation could be performed for a rank- r single or paired latent model: First compute the maximum-likelihood estimate of the covariance Σ of the observed variables under rank- r reduced-rank regression, and then map from Σ to the parameter values of the appropriate latent model. This needs further exploration. In particular, issues of inference under the Gaussian model need to be explored. Note that Izenman [Ize75] and T. W. Anderson [And99] report an asymptotic variance for the reduced-rank estimator, and Ryan et al. [RHC⁺92] report a reduced-rank-regression analysis with hypothesis tests.

Degrees of freedom. In the cross-covariance problem we are not concerned with within-block covariance, only the relationship between \mathbf{X} and \mathbf{Y} . In the context of “projection to Latent Structures” (PLS), it is possible to take advantage of this fact in situations where the number of variables exceeds the number of units ($p + q > n$). Although sample within-block covariances are singular, it is possible to estimate the loadings (“saliences”). The PLS estimates are consistent for the loadings of the paired latent model; standard errors may be computed using the bootstrap, and models can be compared by cross-validation.

Situations exist where it is unrealistic to expect the number of units to exceed the number of variables, where a very large number of variables is in fact desirable. Examples may be found in chemistry and toxicology; see for example Sardy [Sar98]. Behavioral teratology is another field where this occurs. The following statement by Bookstein et al. is apropos: “[B]ehavioral teratology is best studied in breadth, not depth. There appears to be a great variety of ‘moderately good’ measurements, and a complete dearth of ‘very good’ measurements. *There is no gold standard for measuring alcohol-induced brain damage* across the first 7 years of human life; rather, the presence of alcohol damage is a truly latent variable, one developed more and more clearly by longer and longer series of outcomes studied more and more patiently” ([BSSB90] emphasis in original). The same statement could be applied to behavioral instruments in situations outside the realm of teratology.

Thus the following question deserves further exploration: What relationship between the the number of variables and the number of units is required for inference in latent models

for cross-covariance? Is there any means available for performing principled inference in latent models for cross-covariance which does not require $n > p + q$?

More than two blocks. The current work is concerned with the two-block case. Projection to Latent Structures (PLS), however, the data-analytic method which inspired the current work, can be applied to any number of blocks of indicators. Thus a natural extension to the current work would be to consider an arbitrary number of blocks. Canonical Correlation Analysis provides a precedent for such an extension (see Kettenring [Ket82]).

Generalized inverse of the covariance matrix. By Proposition A.0.3 on page 151 we know that a rank-constraint on Σ_{XY} is equivalent to a rank-constraint on the corresponding block of the inverse covariance matrix, provided the inverse exists. Can this fact be extended to generalized inverses?

Scientific applications. The original motivation for the current work was a study in behavioral teratology (Streissguth et al. [SBSB93a]). Two-block symmetric PLS analyses may be found in other fields, such as economic forecasting and modeling (Tishler and Lipovetsky [TL00]). Applications of the current results to a dataset from one or more of these fields would be interesting. In particular, the current work provides a lower bound on the correlation between hypothesized paired latent variables. This bound is computed directly from an estimate of Σ . An explicit correction for attenuation, as introduced by Spearman [Spe87], is not necessary. It would be interesting to explore the significance of this discovery in scientific applications.

Non-Gaussian likelihood. Only second moments have been considered in the current work. Thus the current work provides a way to perform maximum likelihood estimation for multivariate Gaussian models, but not for other distributions. Some questions for further investigation are:

- To what degree is the usefulness of the current result limited in applications to real data, in view of the fact that many datasets are not even approximately Gaussian?

- Is there a way to extend the current result to one or more non-Gaussian distributions?

Vector graphs. Graphical models are used in the current work. An extension to the existing graphical models system is proposed. The existing system was first developed by Sewall Wright (1923). In this system, each vertex in a graph corresponds to exactly one (univariate) variable in a family of probability distributions.

It is possible to define a class of graphs, called **vector graphs**, in which a vertex can correspond to a random vector. Anderson and Perlman have done this for acyclic directed graphs (ADGs or DAGs) [AP98]. The proposed work would deal with mixed graphs (Richardson and Spirtes [RS00]).

Mixed ancestral graphs in the existing (univariate) system could be called **scalar graphs**. Vector graphs are attractive for the following reasons.

- Some sets of vertices which must be represented individually in a scalar graph can be represented by a single vertex in a vector graph; consequently
- More information can be conveyed in a given amount of space by a vector graph than a scalar graph, and
- Vector graphs may be easier to read.

Vector graphs are natural and intuitive, and provide a notation as convenient for graphical models as is linear algebra for systems of linear equations. Work is required, however, to make the notion of a vector graph rigorous. The following work is proposed:

- Maps must be defined between the scalar and vector systems, with the result that for each vector graph a corresponding scalar graph may be constructed and vice versa.
- The maps must be defined in such a way that m-separation relationships can be read off vector graphs in exactly the same way as they can be read off scalar graphs.

BIBLIOGRAPHY

- [Afr57] S. N. Afriat. Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proceedings of the Cambridge Philosophical Society (Mathematical and Physical Sciences)*, 53(4):800–816, October 1957.
- [And51] T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of mathematical statistics*, 22(3):327–351, September 1951.
- [And80] T. W. Anderson. Correction to “estimating linear restrictions on regression coefficients for multivariate normal distributions”. *Annals of Statistics*, 8(6):1400, November 1980.
- [And99] T. W. Anderson. Asymptotic distribution of the reduced rank regression estimator under general conditions. *Annals of Statistics*, 27(4):1141–1154, 1999.
- [AP98] Steen A. Anderson and Michael D. Perlman. Normal linear regression models with recursive graphical markov structure. *Journal of Multivariate Analysis*, 66:133–187, 1998.
<http://www.stat.washington.edu/www/research/online/>.
- [BD77] Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics : basic ideas and selected topics*. Holden-Day Series in Probability and Statistics. Holden-Day, Oakland, California, 1977.
- [BEE98] Elisabet Bøstrom, Solveig Engen, and Ingvar Eide. Mutagenicity testing of organic extracts of diesel exhaust particles after spiking with polycyclic aromatic hydrocarbons (PAH). *Arch Toxicol*, 72:645–649, 1998.

- [Ben89] Peter M. Bentler. *EQS structural equations program manual*. BMDP Statistical Software, 1989.
- [BG73] Ake Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594, July 1973.
- [BMV99] Alison J. Burnham, John F. MacGregor, and Roman Viveros. A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *Journal of chemometrics*, 13:49–65, 1999.
- [Bol89] Kenneth A. Bollen. *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1989.
- [BSSB90] Fred L. Bookstein, Paul D. Sampson, Ann P. Streissguth, and Helen M. Barr. Measuring “dose” and “response” with multivariate data using partial least squares techniques. *Communications in statistics: theory and methods*, 19(3):765–804, 1990.
- [BSSB96] Fred L. Bookstein, Paul D. Sampson, Ann P. Streissguth, and Helen M. Barr. Exploiting redundant measurement of dose and developmental outcome: new methods from the behavioral teratology of alcohol. *Developmental psychology*, 32(3):404–415, 1996.
- [CC80] C. Chatfield and A. J. Collins. *Introduction to Multivariate Analysis*. Texts in Statistical Science. Chapman and Hall, 1980.
- [CM79] Rough-Jane Chou and Robb J. Muirhead. On some distribution problems in manova and discriminant analysis. *Journal of multivariate analysis*, 9:410–419, 1979.
- [FF93] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135, 1993.

- [FvB93] Lloyd D. Fisher and Gerald van Belle. *Biostatistics: a methodology for the health sciences*. John Wiley and Sons, Inc., New York, 1993.
- [Gel88] Paul Geladi. Notes on the history and nature of partial least squares (PLS) modelling. *Chemometrics*, 2(4):231–246, 1988.
- [GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins, third edition, 1996.
- [H88] Agnar Höskuldsson. PLS regression methods. *Chemometrics*, 2(3):211–228, June 1988.
- [Har76] Harry H. Harman. *Modern Factor Analysis*. The University of Chicago Press, third edition, 1976.
- [Har97] David A. Harville. *Matrix algebra from a statistician's perspective*. Springer, 1997.
- [Hea80] John D. Healy. Maximum likelihood estimation of a multivariate linear functional relationship. *Journal of Multivariate Analysis*, 10:243–251, 1980.
- [Hel88] Inge S. Helland. On the structure of partial least squares regression. *Commun. Statist. Simul.*, 17:581–607, 1988.
- [HHMT97] Tyler R. Holcomb, Håkan Hjalmarsson, Manfred Morari, and Matthew L. Tyler. Significance regression: A statistical approach to partial least squares. *Journal of Chemometrics*, 11(4):283–309, Jul-Aug 1997.
- [HJ85] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge, 1985.
- [Hot36] H. Hotelling. Relations between two sets of variables. *Biometrika*, 28:321–377, 1936.

- [HS79] Harold V. Henderson and S. R. Searle. Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *The Canadian Journal of Statistics*, 7(1):65–81, 1979.
- [Ize75] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264, 1975.
- [Ket82] J. R. Kettenring. Canonical analysis. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of the Statistical Sciences*, pages 354–365. Wiley, 1982.
- [KS67] Maurice G. Kendall and Alan Stuart. *The Advanced Theory of Statistics*, volume 2: Inference and Relationship. Hafner, second edition, 1967.
- [Ksh72] Anant M. Kshirsagar. *Multivariate Analysis*, volume 2 of *Statistics Textbooks and Monographs*. Marcel Dekker, New York, 1972.
- [LM86] Richard J. Larsen and Morris L. Marx. *An introduction to mathematical statistics and its applications*. Prentice-Hall, Englewood Cliffs, New Jersey, second edition, 1986.
- [LN68] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. The Addison-Wesley Series in Behavioral Science: Quantitative Methods. Addison-Wesley, Menlo Park, California, 1968.
- [Mat96] MathSoft, Inc. S-PLUS Version 3.4 Release 1 for Silicon Graphics Iris, IRIX 5.3, 1996. Software.
- [MKB79] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Probability and mathematical statistics, a series of monographs and textbooks. Academic Press, New York, 1979.
- [Muc96] Paul M. Muchinsky. The correction for attenuation. *Educational and Psychological Measurement*, 56(1):63–75, February 1996.

- [RHC⁺92] D.A.J. Ryan, J. J. Hubert, E. M. Carter, J. B. Sprague, and J. Parrott. A reduced-rank multivariate regression approach to aquatic joint toxicity experiments. *Biometrics*, 48:155–162, March 1992.
- [RLGW94] Stefan Rännar, Fredrik Lindgren, Paul Geladi, and Svante Wold. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: theory and algorithm. *Journal of Chemometrics*, 8(2):111–125, March-April 1994.
- [RS00] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. Technical Report 375, Department of Statistics, University of Washington, Seattle, 2000.
- [RV98] Gregory C. Reinsel and Raja P. Velu. *Multivariate reduced-rank regression: theory and applications*. Lecture Notes in Statistics. Springer, New York, 1998.
- [Sar98] Sylvain Sardy. *Regularization techniques for linear regression with a large set of carriers*. PhD thesis, University of Washington, Seattle, February 1998.
- [SBSB93a] Ann P. Streissguth, Fred L. Bookstein, Paul D. Sampson, and Helen M. Barr. *The enduring effects of prenatal alcohol exposure on child development: birth through seven years, a partial least squares solution*. Number 10 in International Academy for Research in Learning Disabilities Monograph Series. University of Michigan Press, Ann Arbor, 1993.
- [SBSB93b] Ann P. Streissguth, Fred L. Bookstein, Paul D. Sampson, and Helen M. Barr. *The enduring effects of prenatal alcohol exposure on child development: birth through seven years, a partial least squares solution*, chapter 4, Methods of latent variable modeling by partial least squares, pages 55–85. In *International Academy for Research in Learning Disabilities Monograph Series* [SBSB93a], 1993.

- [Spe04] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.
- [Spe87] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 100(3–4):441, 1987. Reprinted from original 1904 article in the same journal.
- [SRM⁺ar] Peter Spirtes, Thomas Richardson, Chris Meek, Richard Scheines, and Clark Glymour. Using path diagrams as a structural equation modelling tool. *Sociological Methods and Research*, to appear. <http://www.stat.washington.edu/tsr/papersfr.html>.
- [SSBB89] Paul D. Sampson, Ann P. Streissguth, Helen M. Barr, and Fred L. Bookstein. Neurobehavioral effects of prenatal alcohol: Part II. Partial Least Squares analysis. *Neurotoxicology and teratology*, 11(5):477–491, 1989.
- [Ste73] Gilbert W. Stewart. *Introduction to matrix computations*. Academic Press, New York, 1973.
- [Ste01] James H. Steiger. Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association*, 96(453):331–338, 2001.
- [TDSL96] A. Tishler, D. Dvir, A. Shenhar, and S. Lipovetsky. Identifying critical success factors in defense development projects: a multivariate analysis. *Technological forecasting and social change*, 51:151–171, 1996.
- [TL00] A. Tishler and S. Lipovetsky. Modelling and forecasting with robust canonical analysis: method and application. *Computers and operations research*, 27(3):217–232, March 2000.
- [Vin76] H. D. Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4:147–166, 1976.

- [Weg00] Jacob A. Wegelin. A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle, 2000.
- [Wol75] H. Wold. Path models with latent variables: the NIPALS approach. In H. M. Blalock et al., editors, *Quantitative sociology : international perspectives on mathematical and statistical modeling*, pages 307–357. Academic, 1975.
- [Wol82] H. Wold. Soft modeling: the basic design and some extensions. In K. G. Jöreskog and H. Wold, editors, *Systems under indirect observation : causality, structure, prediction, Part II*, number 139 in Contributions to economic analysis, chapter 1, pages 1–54. North-Holland, 1982. Proceedings of the Conference on Systems Under Indirect Observation, held Oct. 18-20, 1979, at Cartigny, Switzerland.
- [Wol85] H. Wold. Partial least squares. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of the Statistical Sciences*, pages 581–591. Wiley, 1985.
- [WRR01] Jacob A. Wegelin, Thomas S. Richardson, and David L. Ragozin. Rank-one latent models for cross-covariance. Technical Report 391, Department of Statistics, University of Washington, Seattle, 2001.
- [ZW97] Donald W. Zimmerman and Richard H. Williams. Properties of the Spearman correction for attenuation for normal and realistic non-normal distributions. *Applied Psychological Measurement*, 21(3):253–270, September 1997.

Appendix A

EQUIVALENCE OF SYMMETRIC AND ASYMMETRIC
CONSTRAINT MODELS OF RANK R

In this section two reduced-rank models for $\text{Cov} \begin{bmatrix} \mathbf{x}_n \\ \mathbf{y}_n \end{bmatrix}$ are reviewed, the symmetric and asymmetric constraint models, and are shown to parameterize the same family of covariances. (These were introduced in Section 1.1.) In the language of Spirtes et al. [SRM⁺ar], the models are covariance equivalent over the observed variables.

It follows that, if \mathbf{x}_n and \mathbf{y}_n have a joint multivariate normal distribution, the maximum-likelihood estimates of the within-block variances for the constraint models are the familiar statistics

$$\begin{aligned} \widehat{\Sigma_{XX}} &= \mathbf{S}_{XX} = (1/N)\mathbf{x}_n^T \mathbf{x}_n, \\ \widehat{\Sigma_{YY}} &= \mathbf{S}_{YY} = (1/N)\mathbf{y}_n^T \mathbf{y}_n. \end{aligned}$$

These results are known (T. W. Anderson [And99]). Proofs are included here for completeness.

Definition of the symmetric and asymmetric constraint models. To emphasize the fact that the models *a priori* represent different distributions, different names will initially be used for the variables governed by the two distributions.

The first model is called the symmetric constraint model. It is specified by

$$\mathbb{E} \begin{bmatrix} \mathbf{x}_n \\ \mathbf{y}_n \end{bmatrix} = \mathbf{0}, \quad \text{Var} \begin{bmatrix} \mathbf{x}_n \\ \mathbf{y}_n \end{bmatrix} = \Sigma,$$

where the covariance is partitioned as

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}.$$

Σ_{XX} and Σ_{YY} are arbitrary positive definite matrices of dimension $p \times p$ and $q \times q$ respectively, Σ_{XY} is of rank R , and the composite matrix Σ is positive definite. (By Proposition A.0.3 on page 151, an equivalent definition would place a rank constraint on the off-diagonal block of the inverse covariance matrix.)

The second model is called the asymmetric constraint model. It is specified by

$$\mathbf{w}_n = \mathbf{B}^T \mathbf{v}_n + \delta_n,$$

where \mathbf{v}_n and \mathbf{w}_n are observed random vectors of dimension p and q respectively, corresponding to \mathbf{x}_n and \mathbf{y}_n in the symmetric model, δ_n is error of dimension q , \mathbf{B} is a $p \times q$ matrix of rank R , $E(\mathbf{v}_n) = \mathbf{0}$, $\text{Var}(\mathbf{v}_n) = \Sigma_{VV}$, $E(\delta_n) = \mathbf{0}$, $\text{Var}(\delta_n) = \Sigma_{\delta\delta}$, $\text{Cov}(\mathbf{v}_n, \delta_n) = \mathbf{0}$, and Σ_{VV} and $\Sigma_{\delta\delta}$ are positive definite matrices of dimension respectively $p \times p$ and $q \times q$. The covariance over the indicators is

$$\Sigma_{\text{asym}} = \text{Var} \begin{bmatrix} \mathbf{v}_n \\ \mathbf{w}_n \end{bmatrix} = \begin{bmatrix} \Sigma_{VV} & \Sigma_{VV}\mathbf{B} \\ \mathbf{B}^T \Sigma_{VV} & \mathbf{B}^T \Sigma_{VV}\mathbf{B} + \Sigma_{\delta\delta} \end{bmatrix}.$$

Lemma A.0.1 *The symmetric and asymmetric models are two ways of parameterizing the same family of covariance matrices.*

Proof. It suffices to show a bijection between the space of symmetric constraint models and the space of asymmetric constraint models.

Suppose we have an asymmetric model. Then if we simply set $\Sigma = \Sigma_{\text{asym}}$, the reader can check that Σ satisfies the conditions of a symmetric constraint model.

Suppose on the other hand that we have a symmetric constraint model. Set

$$\begin{aligned} \Sigma_{VV} &= \Sigma_{XX} \\ \mathbf{B} &= \Sigma_{YX} \Sigma_{XX}^{-1} \\ \Sigma_{\delta\delta} &= \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}. \end{aligned}$$

Then Σ_{VV} is positive definite by definition, and \mathbf{B} has the necessary dimension and rank. $\Sigma_{\delta\delta}$ is positive definite, since it is the conditional variance of $\mathbf{y}_n | \mathbf{x}_n$.

The bijection has been demonstrated. Thus

$$\begin{bmatrix} \mathbf{x}_n \\ \mathbf{y}_n \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mathbf{v}_n \\ \mathbf{w}_n \end{bmatrix}$$

are indistinguishable, so that we may dispense with the $\mathbf{v}_n, \mathbf{w}_n$ notation, referring henceforth only to \mathbf{x}_n and \mathbf{y}_n .

Maximum-likelihood estimation of within-block covariance. Recall that we have the following equivalence of notation,

$$\begin{aligned}\mathbf{x}_n &\equiv \mathbf{X}_n^T \\ \mathbf{y}_n &\equiv \mathbf{Y}_n^T,\end{aligned}$$

so that \mathbf{x}_n and \mathbf{y}_n are the n th rows of the data matrices, seen as column vectors. Suppose that these rows are independent and identically distributed (iid) according to a multivariate normal distribution. It is natural to ask what the maximum-likelihood estimate is of the within-block covariance when Σ_{XY} is subject to a rank constraint.

Two cases are well-known. If $\Sigma_{XY} \equiv \mathbf{0}$, \mathbf{x}_n and \mathbf{y}_n are independent and their covariances are estimated independently by

$$\widehat{\Sigma_{XX}} = (1/N)\mathbf{X}^T\mathbf{X}, \quad \widehat{\Sigma_{YY}} = (1/N)\mathbf{Y}^T\mathbf{Y}.$$

On the other hand, if Σ_{XY} is unconstrained, we are simply estimating an unconstrained Σ . Then we use the familiar result that

$$\begin{bmatrix} \widehat{\Sigma_{XX}} & \widehat{\Sigma_{XY}} \\ \widehat{\Sigma_{YX}} & \widehat{\Sigma_{YY}} \end{bmatrix} = \widehat{\Sigma} = \mathbf{S} = (1/N) \begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Y} \\ \mathbf{Y}^T\mathbf{X} & \mathbf{Y}^T\mathbf{Y} \end{bmatrix}$$

(Mardia, Kent and Bibby [MKB79] page 104).

Thus at the two extremes of rank—cross-covariance constrained to rank zero, and cross-covariance unconstrained—we have the same maximum-likelihood estimates. When Σ_{XY} is constrained to rank $R < \min(p, q)$, however, we may naturally ask whether we do not need the \mathbf{y}_n s as well as the \mathbf{x}_n s to estimate Σ_{XX} . For instance, in the bivariate normal case, if we know ρ , the \mathbf{x}_n 's are not sufficient to estimate σ_x (Kendall and Stuart [KS67] pages 57–60). It turns out however that the consequences of this constraint are different from the consequences of a rank constraint on Σ_{XY} . This is stated formally as follows.

Corollary A.0.2 *Under the assumption of multivariate normality, the maximum-likelihood estimates of Σ_{XX} and Σ_{YY} are*

$$\begin{aligned}\widehat{\Sigma_{XX}} &= \mathbf{S}_{XX} = (1/N)\mathbf{X}^T\mathbf{X} , \\ \widehat{\Sigma_{YY}} &= \mathbf{S}_{YY} = (1/N)\mathbf{Y}^T\mathbf{Y} .\end{aligned}$$

Thus \mathbf{S}_{XX} is sufficient for Σ_{XX} , and \mathbf{S}_{YY} is sufficient for Σ_{YY} . This holds even when there is a rank constraint on Σ_{XY} .

Proof. First note that by symmetry a proof for Σ_{XX} will suffice.

One way to prove that $\mathbf{x}_1, \dots, \mathbf{x}_N$ are sufficient to estimate Σ_{XX} would be first to recall that the asymmetric and symmetric models produce identical covariance matrices for the observed variables. Then we could use the asymmetric parameterization, and notice that

$$\mathbf{y}_n|\mathbf{x}_n \sim \mathbf{N}(\mathbf{B}^T\mathbf{x}_n, \Sigma_{\delta\delta}) ,$$

a distribution which does not appear to depend on Σ_{XX} . This argument could be seen as skirting the issue, however, since, in the map from the symmetric to the asymmetric model, the parameter Σ_{XX} appears in the expressions for \mathbf{B} and $\Sigma_{\delta\delta}$. For this reason, sufficiency will be proven in the context of the symmetric constraint model. Once this is accomplished we will return to the argument which uses the asymmetric parameterization.

According to a familiar definition of sufficiency (Bickel and Doksum [BD77], Mardia, Kent and Bibby [MKB79], and Larsen and Marx [LM86]), $\mathbf{x}_1, \dots, \mathbf{x}_N$ are sufficient to estimate Σ_{XX} if and only if the conditional density of $\mathbf{y}_n|\mathbf{x}_n$ is not a function of Σ_{XX} . A naive application of this definition, however, could lead to an incorrect conclusion. Since Σ_{XX} appears in the expressions for the conditional mean and conditional variance,

$$\begin{aligned}\mathbb{E}(\mathbf{y}_n|\mathbf{x}_n) &= \Sigma_{YX}\Sigma_{XX}^{-1}\mathbf{x}_n \\ \text{Var}(\mathbf{y}_n|\mathbf{x}_n) &= \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} ,\end{aligned}$$

we might be tempted to conclude that the conditional density depends on Σ_{XX} and that the \mathbf{x}_n are not sufficient to estimate Σ_{XX} .

To see that this is false, first note that the conditional density depends on two parameters, a $q \times p$ matrix of rank R , say \mathbf{B} , and a $q \times q$ positive definite matrix, say $\Sigma_{\delta\delta}$. That these

parameters do not vary with Σ_{XX} can be shown by holding them constant and varying Σ_{XX} over its space of feasible values. Let $(\Sigma_{XX})_0$, $(\Sigma_{XY})_0$, and $(\Sigma_{YY})_0$ be one set of feasible parameter values, so that

$$\left. \begin{aligned} \mathbf{B}\mathbf{x}_n &= E(\mathbf{y}_n|\mathbf{x}_n) = (\Sigma_{YX})_0 (\Sigma_{XX})_0^{-1} \mathbf{x}_n, \\ \Sigma_{\delta\delta} &= \text{Var}(\mathbf{y}_n|\mathbf{x}_n) = (\Sigma_{YY})_0 - (\Sigma_{YX})_0 (\Sigma_{XX})_0^{-1} (\Sigma_{XY})_0. \end{aligned} \right\} \quad (\text{A.1})$$

Now fix \mathbf{x}_n , $\Sigma_{\delta\delta}$, and \mathbf{B} , and suppose that Σ_{XX} is an arbitrary positive definite matrix of dimension $p \times p$. Then we may set

$$\Sigma_{YX} \equiv \mathbf{B}\Sigma_{XX} = (\Sigma_{YX})_0 (\Sigma_{XX})_0^{-1} \Sigma_{XX}. \quad (\text{A.2})$$

Σ_{XY} has the same rank as $(\Sigma_{YX})_0$, so it is feasible. Set

$$\Sigma_{YY} = \Sigma_{\delta\delta} + \underbrace{\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}}_{\text{second term}}. \quad (\text{A.3})$$

$\Sigma_{\delta\delta}$ is positive definite, since it is a conditional variance. The second term is positive semidefinite, and consequently the entire right-hand side is positive definite.

Multiplying Equation A.2 on the right by Σ_{XX}^{-1} , we have

$$\Sigma_{YX}\Sigma_{XX}^{-1} = \mathbf{B} = (\Sigma_{YX})_0 (\Sigma_{XX})_0^{-1}. \quad (\text{A.4})$$

Subtracting the second term from both sides of Equation A.3 and recalling the value of $\Sigma_{\delta\delta}$ in Equation A.1, we obtain

$$\begin{aligned} \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} &= \Sigma_{\delta\delta} \\ &= (\Sigma_{YY})_0 - (\Sigma_{YX})_0 (\Sigma_{XX})_0^{-1} (\Sigma_{XY})_0. \end{aligned} \quad (\text{A.5})$$

Since Equations A.4 and A.5 hold for any feasible Σ_{XX} , we have shown that \mathbf{B} and $\Sigma_{\delta\delta}$ do not vary with Σ_{XX} . Thus, in spite of appearances, the conditional density of $\mathbf{y}_n|\mathbf{x}_n$ is not a function of Σ_{XX} . We have established that $\mathbf{x}_1, \dots, \mathbf{x}_N$ are sufficient to estimate Σ_{XX} .

Cartesian product. We are now ready to return to the argument which explicitly uses the asymmetric parameterization. We have established that when we map from a symmetric to an asymmetric parameterization, the parameters \mathbf{B} and $\Sigma_{\delta\delta}$ of the asymmetric

parameterization are not functions of Σ_{XX} . Thus we have a bijection between the following two spaces:

$$\left\{ \Sigma : \Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} > 0, \ (p+q) \times (p+q); \ \Sigma_{XY} \ p \times q \text{ of rank } R \right\}$$

and the Cartesian product

$$\{\Sigma_{XX} : \Sigma_{XX} > 0, \ p \times p\} \times \{\mathbf{B} : \ q \times p, \ \text{rank } R\} \times \{\Sigma_{\delta\delta} : \Sigma_{\delta\delta} > 0, \ q \times q\} .$$

The proof that the parameters \mathbf{B} and $\Sigma_{\delta\delta}$ do not vary with Σ_{XX} constitutes a proof that the second space is indeed a Cartesian product. To prove sufficiency, we are now free to return to our original argument. We work in the Cartesian product space, write

$$\mathbf{y}_n | \mathbf{x}_n \sim \mathbf{N}(\mathbf{B}\mathbf{x}_n, \Sigma_{\delta\delta}) ,$$

and observe that the density does not involve Σ_{XX} . This completes the proof of Corollary A.0.2. \square

We stated on page 147 that a rank- R constraint on Σ_{XY} is equivalent to a rank- R constraint on the off-diagonal block of the inverse covariance matrix. We now prove this fact.

Proposition A.0.3 *Let*

$$\Sigma = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$$

be square and invertible, where \mathbf{A} is $p \times p$, \mathbf{B} is $p \times q$, \mathbf{D} is $q \times q$, etc. Let the inverse be partitioned as

$$\Sigma^{-1} = \Lambda = \begin{bmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} \end{bmatrix}$$

where the dimensions of \mathbf{E} match those of \mathbf{A} , the dimensions of \mathbf{F} match those of \mathbf{B} , etc. Then $\text{rank}(\mathbf{F}) = \text{rank}(\mathbf{B})$.

Proof. Since $\Sigma\mathbf{A} = \mathbf{I}$, the upper right corner of the product equals zero; in other words

$$\mathbf{A}\mathbf{F} = \mathbf{B}\mathbf{H} \text{ , whence}$$

$$\mathbf{F} = \mathbf{A}^{-1}\mathbf{B}\mathbf{H} \text{ .}$$

Since Σ is invertible, so are both \mathbf{A} and \mathbf{H} . By a well-known property of the rank of products, $\text{rank}(\mathbf{F}) \leq \text{rank}(\mathbf{B}) = \min(p, q)$. But since \mathbf{A}^{-1} and \mathbf{H} are both invertible, equality must hold. □

VITA

Jacob A. Wegelin received a B.A. in Classics and a B.S. in Mathematics from the University of Washington in 1986, and an M.S. in Statistics from the University of Washington in 1989.