Studies in Systems, Decision and Control 22

J. Christopher Westland

Structural Equation Models

From Paths to Networks

Second Edition

Springer

Studies in Systems, Decision and Control

Volume 22

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Systems Research Institute, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series "Studies in Systems, Decision and Control" (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control–quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/13304

J. Christopher Westland

Structural Equation Models

From Paths to Networks

Second edition

Springer

J. Christopher Westland
Information & Decision Systems
University of Illinois at Chicago
Chicago, IL, USA

ISSN 2198-4182 ISSN 2198-4190 (electronic)
Studies in Systems, Decision and Control
ISBN 978-3-030-12507-3 ISBN 978-3-030-12508-0 (eBook)
https://doi.org/10.1007/978-3-030-12508-0

Library of Congress Control Number: 2019931906

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword to the Second Edition of Structural Equation Models

Since publication of the first edition of *Structural Equation Models*, I have been fortunate to maintain an active dialog on structural equation modeling (SEM) with many of my colleagues around the world. I never cease to be surprised with the broad divergence of opinions and myriad applications of SEM methodologies. Statistical methods such as regression and ANOVA rely on datasets of objectively measured constructs. These fail to satisfy a widespread need among researchers to analyze data concerning relationships between hypothesized and unobservable constructs: aesthetics, perceptions, utilities, and other human and social constructs. Criticisms of SEM have arisen from its lack of fit statistics and indeed the lack of defensible sampling strategies. But my argument is that these problems can be repaired while retaining the desirable features of SEM.

SEM has been applied in both the natural and the social sciences, but it has proven particularly valuable in the social sciences, where researchers apply SEM approaches rather than more structured regression approaches by the inclusion of unobservable (or latent) constructs and by the use of computationally intensive iterative searches for coefficients that fit the data. The expansion of statistical analysis to encompass unmeasurable constructs using SEM, canonical correlation, Likert scale quantification, principal components, and factor analysis has vastly extended the scope and relevance of the social sciences over the past century. Subjects that were previously the realm of abstract argumentation have been transported into the mainstream of scientific research (see Allen and Seaman 2007; Altman and Royston 2000).

This new edition of this widely cited book surveys the full range of available structural equation modeling (SEM) methodologies. The book has been updated throughout to reflect the arrival of new software packages, which have made analysis much easier than in the past. Applications in a broad range of disciplines are discussed, particularly in the social sciences where many key concepts are not directly observable. This is the first book to present SEM's development in its proper historical context—essential to understanding the application, strengths, and weaknesses of each particular method. This book also surveys emerging approaches that complement SEM. They have been applied in diverse areas in

engineering, including neuroscience for accurate examination of the activity among neural regions during different behaviors. The partial least squares SEM method was contemporaneously developed with path analysis software (PLS) regression to address problems in chemistry and spectrography. They improve on predecessor path models that were widely used in genetic research in livestock and agriculture and environmental studies in the elicitation of ecological networks. SEM's ability to accommodate unobservable theory constructs through latent variables is of significant importance to social scientists. Latent variable theory and application are comprehensively explained, and methods are presented for extending their power, including guidelines for data preparation, sample size calculation, and the special treatment of Likert scale data. Tables of software, methodologies, and fit statistics provide a concise reference for any research program, helping assure that its conclusions are defensible and publishable.

Chicago, IL, USA
26 December 2018

J. Christopher Westland

Contents

Chapter 1
An Introduction to Structural Equation Models

The past two decades have witnessed a remarkable acceleration of interest in structural equation modeling (SEM) methods in many areas of research. In the social sciences, researchers often distinguish SEM approaches from more powerful systems of regression equation approaches by the inclusion of unobservable constructs (called latent variables in the SEM vernacular), and by the use of computationally intensive iterative searches for coefficients that fit the data. The expansion of statistical analysis to encompass unmeasurable constructs using SEM, canonical correlation, Likert scale quantification, principal components, and factor analysis has vastly extended the scope and relevance of the social sciences over the past century. Subjects that were previously the realm of abstract argumentation have been transported into the mainstream of scientific research (see Allen and Seaman 2007; Altman and Royston 2000).

Statistical methods to identify latent constructs underlying observations evolved in the 1930s. Principal component analysis (PCA), factor analysis, and other methods look for methods to reduce the dimensionality of a complex multicollinear dataset. Latent factors accounting for most of the similarity or distance of measurements could potentially be inferred from these factors. SEM methods grew out of efforts to infer additional structure between these latent constructs.

Many of the seminal studies on structural statistical models in economics took place in the Cowles Commission (then at the University of Chicago) in the 1940s and 1950s and later in the Chicago school of economics from the 1950s on. In a 1976 paper, Robert Lucas of the Chicago school argued that generic additive linear models such as those invoked in the panel regressions commonly used in econometrics lacked stability and robustness (Lucas 1992). He argued, in what has come to be known as the "Lucas critique," that empirical models are improved when constructs are policy invariant, i.e., structural, implying that they would be unlikely to change whenever the competitive environment or a particular policy changed. Lucas suggested that researchers need to model the "deep structural parameters" (relating to preferences, technology, and resource constraints) that are assumed

J. C. Westland, *Structural Equation Models*, Studies in Systems,
Decision and Control 22, https://doi.org/10.1007/978-3-030-12508-0_1

to govern individual behavior. Structural models in Lucas (1992) were intended to enable a positive research program for econometrics, allowing for prediction and real-world decisions. Policy-invariant structural models are constructed through analysis of the underlying dynamics of the construct relationships and behavior, and are based on a "theory" of how the real-world works. The "Lucas critique" promoted a priori theory building, and this has become common practice in structural equation modeling. It is now a standard practice to design the theorized causal structures in an SEM, whether the statistical method is PLS-PA, LISREL, or regression approach, prior to statistical estimation.

The products of SEM statistical analysis algorithms fall into three groups:

1. pairwise canonical correlations between pairs of prespecified latent variables computed from observable data (from the so-called partial least squares path analysis, or PLS-PA approaches);
2. multivariate canonical correlation matrices for prespecified networks of latent variables computed from observable data (from a group of computer intensive search algorithms originating with Karl Jöreskog); and
3. systems of regression approaches that fit data to networks of observable variables whose clusters are hypothesized to co-vary with latent constructs. Other methods of latent variable analysis are now emerging with the introduction of machine learning new social network analysis.

Many of the PLS-PA algorithms are variations on an incompletely documented software package described in Lohmöller (1988), Lohmöller (1989), and Lydtin et al. (1980) and we sometimes still see some of their old Fortran code inside a customized user interface wrapper. Fortunately Monecke and Leisch (2012) have incorporated Wold's mathematics in their thoroughly modern semPLS package for R. PLS-PA has a tendency to be confused with Wold's partial least squares regression—a problem Herman Wold tried unsuccessfully to correct. The path analysis PLS-PA commonly used in latent variable investigations is unrelated to Wold's (Wold 1966; Hill 1979) partial least squares regression methods, instead being a variation on Wold's (Wold 1966; Hill 1979) canonical correlation methods to elicit the correlations of latent variables.

Two different covariance structure algorithms are widely used: (1) LISREL (an acronym for linear structural relations) (Jöreskog and Van Thillo 1972; Jöreskog 1993; Jöreskog and Sörbom 1982; Jöreskog et al. 1979; Jöreskog 1970) and the AMOS (analysis of moment structures) (Fox 2006; McArdle and Epstein 1987; McArdle 1988). Variations on these algorithms have been implemented in EQS, TETRAD, and other packages.

Methods in systems of equations modeling and social network analytics are not as familiar in the social sciences as the first two methods, but offer comparatively more analytical power. Accessible and comprehensive tools for these additional approaches are covered in this book, as are research approaches to take advantage of the additional explanatory power that these approaches offer to social science research.

The breadth of application of SEM methods has been expanding, with SEM increasingly applied to exploratory, confirmatory, and predictive analysis through a variety of ad hoc topics and models. SEM is particularly useful in the social sciences where many if not most key concepts are not directly observable, and models that inherently estimate latent variables are desirable. Because many key concepts in the social sciences are inherent.

Methods in systems of equations modeling and social network analytics are not as familiar in the social sciences as the first two methods, but offer comparatively more analytical power. Accessible and comprehensive tools for these additional approaches are covered in this book, as are research approaches to take advantage of the additional explanatory power that these approaches offer to social science research.

The breadth of application of SEM methods has been expanding, with SEM increasingly applied to exploratory, confirmatory, and predictive analysis through a variety of ad hoc topics and models. SEM is particularly useful in the social sciences where many if not most key concepts are not directly observable, and models that inherently estimate latent variables are desirable. Because many key concepts in the social sciences are inherent.

1.1 Latent Constructs as Organizing Principles of Science in the Twentieth Century

In science, an idea is a hypothesis that gives structure to our observations. Ideas are latent constructs embellished with mechanisms to test, use, predict, and control their implementation. Three ideas revolutionized science in the twentieth century: the atom, the bit, and the gene.

The atom provided an organizing principle for twentieth century physics. Hypotheses about the atom date from the Greek philosopher Democritus, and steady advancements marshaled the evolution of chemistry out of alchemy. But it was Einstein's obsession with determining the size of an atom that indirectly motivated his groundbreaking papers on the photoelectric effect, Brownian motion and special relativity.

The gene is innately human. Its origins have seduced the attention of philosophers and politicians, more often than not, leading them astray. Genes are the unseen first cause of human and animal "phenotypes"—their observable, externalized consequences resulting from interaction of an organism's genotype with the environment. Phenotypes manifest themselves as morphology, skin color, strength, and numerous other characteristics. The word "gene" was coined by botanist Wilhelm Johannsen as a shortening of Darwin's pangene.

The search for the unobservable genes that would lead to various desirable or undesirable phenotypes has been a major factor in the history and philosophy of mankind. In the twenty-first century, the quest to master genetics has enlisted our

knowledge of atoms and bits as well. The physicist John Wheeler famously stated that "…all things physical are information-theoretic in origin," a sentiment that drives much of modern genetics, and brings us to the last of our twentieth century "ideas."

The bit, a portmanteau of "binary digit," arose from efforts to quantify and encode information, particularly in such devices as the Jacquard looms in the early 1800s. Attempts to improve bandwidth in telegraph lines in the mid-nineteenth century led to speculation that there existed some sort of fundamental measure of information: a bit. Bits were fundamental to Morse code, and the basis for Hartley's and Shannon's seminal work on information theory.

It was not originally the desire to make better men or women that spurred developments in the science of genetics; it was man's desire to improve domesticated crops and animals.

1.2 Path Analysis in Genetics

Though structural equation models today are usually associated with soft problems in the social sciences, they had their origin in the natural sciences—specifically biology. Europe's nineteenth century scholars were challenged to make sense of the diverse morphologies observed during an age of explorations, in Asia, Africa, and the Americas, as well as at home. In this period, new species of plants and animals were transplanted, domesticated, eaten, and bred at an unprecedented rate. An American ultimately provided one statistical tool that allowed scholars to build a science out of their diverse observations.

Seldom has a non-human animal been so thoroughly poked, observed, trained, and dissected as the domesticated dog. A member of the Canidae family, the dog is distantly related to coyotes, jackals, dingoes, foxes, and wolves. There is evidence of distinct dog breeds as early as five thousand years ago in drawings from ancient Egypt. The business of designing dogs for particular purposes began in earnest around the sixteenth century, and by the nineteenth century, clubs and competitions abounded for the naming and monitoring of breeds. There is a huge variation of sizes, shapes, temperaments, and abilities in modern dogs—much more so that in their homogeneous wolf ancestors. This has resulted from humans consciously influencing the genetics of dog populations through an involved network of interbreeding and active selection. But none of this was a science at the dawn of the twentieth century, despite enormous expenditures, and centuries of breeding and contests to create "the perfect dog." There was no theory (or perhaps too many competing but unsupported theories) about how particular characteristics arose in a particular sub-population of dogs. The sciences of evolution and genetics seldom spoke to each other before the twentieth century. The most influential biologists held the idea of blending inheritance, promoted in a particular form in Charles Darwin's theory of pangenesis— inheritance of tiny heredity particles called gemmules that could be transmitted from parent to offspring. In those days, the

work of the Augustinian friar and polymath Gregor Mendel was unknown, having been rejected and forgotten in the biology community when published in the 1860s. Mendel's sin was to introduce mathematics into a field that biologists felt should be a descriptive science, not an analytical one. Rediscovery of Mendel's writings in the early twentieth century led biologists towards the establishment of genetics as a science and basis for evolution and breeding. Geneticist, Sewall Wright, along with statisticians R.A. Fisher and J.B.S. Haldane, were responsible for the modern synthesis that brought genetics and evolution together. Wright's work brought quantitative genetics into animal and plant breeding, initiating the hybrid seed revolution that transformed US agriculture in the first half of the twentieth century. Wright actively mapped the breeding networks that created desirable hybrids—of particular significance to the dog breeders was Wright discovery of the inbreeding coefficient and of methods of computing it in pedigrees. The synthesis of statistical genetics into the evolution of populations required a new quantitative science with which to map the networks of influence, on random genetic drift, mutation, migration, selection, and so forth. Wright's quantitative study of influence networks evolved in the period 1918– 1921 into Wright's statistical method of path analysis—one of the first statistical methods using a graphical model, and one which is the subject of this book. Let's begin by reviewing the evolution of path analysis from the dark ages of nineteenth century evolution debates, through today's statistical methods, to emerging techniques for mapping the extensive networks of biological interactions important to genetics and biotechnology in the future.

1.3 Sewall Wright's Path Analysis

Path analysis was developed in 1918 by geneticist Sewall Wright (1920, 1921, 1934), who used it to analyze the genetic makeup of offspring of laboratory animals (Fig. 1.1).

Early graphs were very descriptive, with pictures and stories attached. But gradually pictures of laboratory critters gave way to representative boxes and positive or negative correlations (Figs. 1.2 and 1.3).

Rensis Likert's work at the University of Michigan in the 1930s and 1940s saw path analysis directed towards social science research. Social scientists need to model many abstract and unobservable constructs—things like future intentions, happiness, customer satisfaction, and so forth. Though not directly observable, there typically exist numerous surrogates that can provide insight into such abstract (or latent) constructs—these observable surrogates are called "indicators" of the latent variable. Further innovation in path models evolved around Hermann Wold's extensions of Hotelling's seminal work in principal component analysis (PCA). Wold began promoting the principal components as representations of abstract (latent) constructs. Latent abstractions proved useful in the evolving fields of psychometrics and sociological surveys, and were widely adopted in the 1950s and 1960s (Hotelling 1936; Wold 1966). Path diagrams evolved once again, to

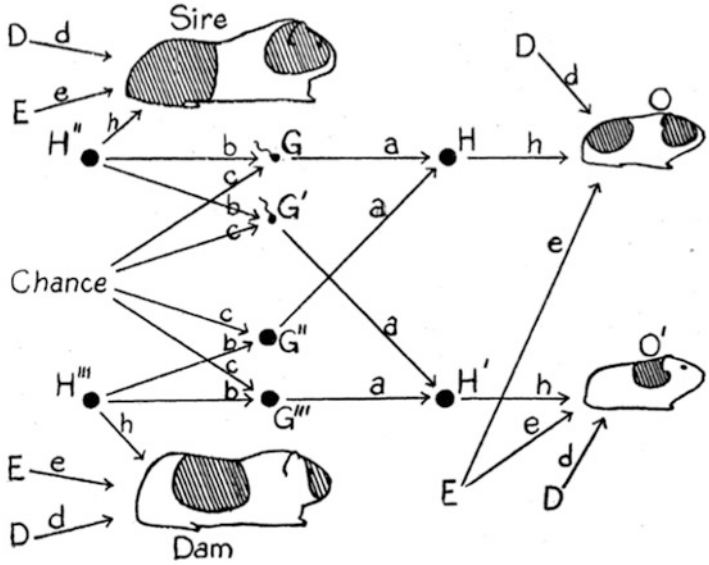

Fig. 1.1 Relations between litter mates and their parents (H represents (latent) genetic components, other capital letters are (manifest) environmental factors, and lowercase letters are path coefficients)

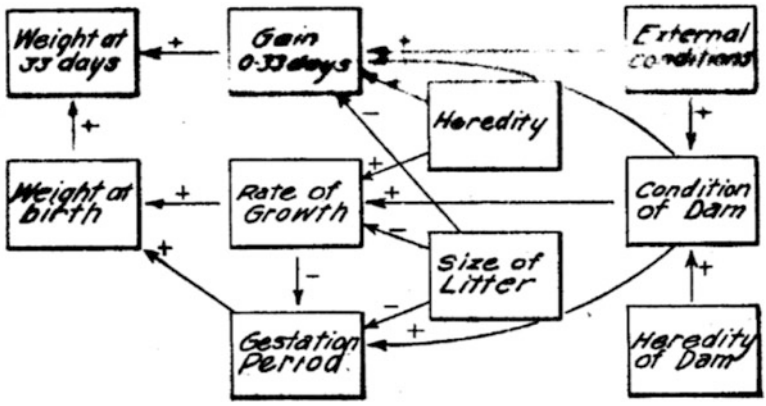

Fig. 1.2 Interrelationship between the factors that determine weight of guinea pigs at birth and at weaning

incorporate Wold's conceptualization of latent constructs as the first component from a principal component analysis. Wold called the network model of latent variables the "structural model" or sometimes the "inner" model. The term "structural equation model" came about from his use, which Wold borrowed from the matrix terminology of systems of equation regression approaches developed at the Cowles Commission. Social scientists were ultimately not content to let PCA dictate their