What is Optimal Data Analysis?

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Preview begins by describing the ODA algorithm, requisite special-purpose software, and applied investigations using ODA to conduct statistical analysis. Discussion next addresses the development and application of multivariable linear and non-linear optimal models. Preview concludes by discussing current research and development foci, including causal inference methodology, system automation, new application arenas, and evolving methods and resources.

Also known as univariable optimal discriminant analysis, the ODA algorithm was derived using operations research optimization methods over a quarter-century ago.\(^1\) In any given application ODA finds the cutpoint for an ordered attribute (i.e., independent variable)—or assignment rule for a categorical attribute—that most accurately (i.e., optimally) discriminates the two or more categories of the class (i.e., dependent) variable.

For example, imagine a hypothetical study involving a two-category class variable (whether or not a patient has cardiovascular disease); an ordered attribute (systolic blood pressure, in mm Hg); and a categorical attribute (gender, male or female). Further imagine that the most accurate possible ODA model for predicting presence versus absence of disease for systolic blood pressure (SBP) is: if SBP≤145 mm Hg, predict the patient has no disease; if SBP>145 mm Hg, predict the patient has disease. Finally imagine the most accurate ODA gender model for predicting presence versus absence of disease is: if gender=female, predict no disease; if gender=male, predict disease.

How is the most accurate possible model for a given attribute and sample identified?

For any given application, identifying the most accurate cutpoint (or assignment rule) requires iterating through every possible value for an ordered attribute, or rule for a categorical attribute, and computing the effect strength for sensitivity (ESS) obtained using each cutpoint (rule) to classify all the sample observations.

An index of model accuracy that adjusts for the effect of chance, ESS is simply the mean sensitivity obtained across class categories by using the cutpoint (or rule) to classify observations, standardized using a 0%-100% scale on which 0% represents the level of discriminatory accuracy expected by chance; 100% represents perfect accuracy; and negative ESS values indicate accuracy that is worse than is expected by chance. In unweighted designs the mean sensitivity that is expected by chance is 50% for a two-category class variable, and is 100%/C for a C-category class variable.

By definition, for any given application, the maximum accuracy, optimal model uses the optimal cutpoint for ordered attributes—or the optimal assignment rule for categorical attributes—that yields the greatest ESS value.\(^4\)
The convention used to evaluate the strength-of-effect of ODA models derives from simulation research: ESS < 25% is a relatively weak effect; < 50% is a moderate effect; < 75% is a relatively strong effect; < 90% is a strong effect; and ESS > 90% is a very strong effect. The optimal model is subjected to a non-parametric permutation test to assess statistical significance of obtained ESS. Model reproducibility and cross-generalizability are assessed via cross-validation methods such as jackknife, K-unfolding, bootstrap, or hold-out analysis.

Statistical software implementing ODA offers many options to control modeling and validation processes. Statistical analysis using the ODA algorithm has been conducted in a wide variety of medical applications, for example identifying radiation-related predictors of hematologic toxicity for cervical cancer; clinical features predictive of gemcitabine-associated lung injury; and risk factors and outcomes associated with non-Enterococcus faecalis, non-Enterococcus faecium enterococcal bacteremia; predicting serious adverse drug reactions; primary care dentist referral of potentially malignant disorders; deep vein thrombosis in stroke; postsurgical outcomes in patients with chronic sinusitis; and degree of independence one year after experiencing a severe brain injury; modeling metronidazole therapy outcomes for Clostridium difficile disease; comparing complication rates of corticosteroid-treated asthmatics undergoing different surgical procedures; and assessing the impact of duration of untreated symptoms in children with juvenile dermatomyositis on clinical and laboratory status at diagnosis; comparing patients with nasal polyps or allergic rhinitis; and assessing measurement reliability in a study of computerized prescriber order entry and prescribing errors.

ODA has also served as the analytic engine in programmatic medical research, for example in occupational medicine studies predicting anhydride-induced immunologic respiratory disease using antibody level, assessing the role of hexahydrophthalic anhydride exposure and other risk factors in predicting immunologically-mediated respiratory disease, evaluating the effect of removal from exposure on anhydride-induced respiratory disease, and determining the effectiveness of respiratory protective devices on the development of antibody and occupational asthma. Or, in emergency medicine studies evaluating accuracy of patient waiting time estimates, and importance of the difference between patient waiting time perceptions and expectations, and of ratings of staff information delivery and expressive quality, in predicting patient satisfaction with care received in the Emergency Department.

Statistical analysis conducted via ODA has also occurred in a broad spectrum of psychometric and psychological (e.g., health, child, clinical, neurological, educational, personality, military, and social) applications, including research evaluating validity characteristics of scores on a survey of affect intensity, of the use of vascular endothelial growth factor as a diagnostic biomarker for major depression, and of androgyny as a measure of physician empathy; investigations of the phenomenal variance of primary progressive aphasia and of the accuracy of an informant-rated scale of estimated cognitive decline; studies predicting neurobehavioral function after severe brain injury as a function of neurostimulants received during rehabilitation, incomplete effort using digit span from the Wechsler Adult Intelligence Scale, and Li+ response and toxicity in Li+-treated bipolar patients using biochemical and psychiatric variables; discriminating high- versus low-quality child therapy sessions; comparing the use of MMPI-2 and MMPI-2-RF validity scales in assessing effort on cognitive tests in a military sample, and identifying gender-based differences in scores on different models of Type A behavior.
weighting of observations, are presented for a variety of research designs and hypotheses from a multitude of different disciplines (e.g., transportation science, finance, biology, chemistry, engineering, insurance, agriculture, accounting, political science, anthropology, advertising, law) in two texts that are dedicated to the ODA algorithm. Examples of ODA analyses are also routinely reported in research using non-linear optimal models (discussed ahead).

**Linear Multiattribute Models**

At the time that the ODA algorithm was discovered, one of two legacy statistical methods was routinely selected to discriminate two class categories (“groups”) using two or more attributes—either logistic regression or Fisher’s linear discriminant analysis. However, issues such as the common failure of empirical data to satisfy underlying assumptions, and mediocre levels of model predictive accuracy, motivated the development of heuristic methods seeking (but not explicitly proving) optimal, maximum-accuracy solutions.

Legacy methods identify a linear model used to obtain a predicted response function score (i.e., Y-hat, or predicted score) for every observation in the sample. Each Y-hat is compared with an *a priori* threshold value (i.e., 0.5) in order to make classification assignments into one of the two groups being discriminated. The ODA algorithm may be used to “fine-tune” the threshold used to make classification decisions, by iterating through all possible threshold values and identifying the optimal threshold that yields maximum predictive accuracy for the sample. So-called “ODA-refinement” has been reported to increase the overall classification accuracy of legacy models by approximately five per cent in hold-out validity analysis. Limited additional study of the efficacy of ODA-based optimization of suboptimal multiattribute linear classification models was promising.

Also occurring during this time, synergy from the combination of more capable optimization software and faster computers enabled experimentation with mathematical programs identifying explicitly optimal linear multiattribute models. Optimal models yield greater overall classification accuracy than suboptimal models, but are feasible only for small samples (N ≤ 100) for formulations involving attributes assessed on ordered scales. Adopting the Warmack-Gonzalez algorithm increased feasible N by an order of magnitude.

Increasing feasible N for optimal linear multiattribute models requires greatly reducing computations, so the “01-1” algorithm was formulated to constrain model attribute coefficients to assume one of three possible values: 1 (add the value of the attribute to the sum); 0 (ignore the attribute); or -1 (subtract the value of the attribute from the sum). Using this algorithm the sum obtained for an observation (for any user-specified number of attributes) is compared with a threshold value (computed by the algorithm) to classify the observation into one of the two classes with explicitly optimal accuracy. When tested for an application predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation, the 01-1 model was more accurate in training and hold-out validity analysis, using half as many attributes, as a logistic regression model using real-number coefficients.

While it was known that optimal linear multiattribute models can identify strong effects in applications for which legacy models find nothing, it was discovered later that optimal nonlinear models can identify linear effects that can’t be identified using linear models.

At this time research investigating optimal linear multiattribute models slowed due to confluence of three events: (a) disruption in systems engineering, quantitative methods and management science programs with which most faculty active in this arena were affiliated; (b) an increasing awareness of the proclivity of linear models for paradoxical confounding; and (c) the discovery of optimal non-linear models that obviate paradoxical confounding.
The HO-CTA Algorithm

Three unique classification tree analysis (CTA) algorithms have been developed, which govern the construction of optimal nonlinear models. Regardless of which algorithm is used to create the CTA model, the least-complex CTA model is an ODA model involving only one attribute. More complex CTA models chain multiple ODA models together.

Every CTA model classifies each sample observation into one of two or more subgroups, each represented as a model endpoint (terminal node). Endpoints are called “strata” because the CTA model stratifies the sample into subgroups homogeneous within, heterogeneous between, the different endpoints defined by the attributes (and the corresponding optimal cutpoints and/or assignment rules) selected by the CTA model.

In the hierarchically-optimal CTA (HO-CTA) algorithm the initial (root) node is the attribute achieving the highest ESS value for the entire sample: additional nodes yielding greatest ESS are iteratively added on all model branches. A sequentially-rejective Šidák multiple comparisons procedure is used to ensure statistical significance of model nodes at the experimentwise \( p \leq Šidák 
\) criterion or generalized (per-comparison \( p \leq 0.05 \)) criterion, and finally the model is pruned to ensure maximum ESS.

Statistical analysis using the HO-CTA algorithm has also occurred in a broad spectrum of psychological applications: modeling adaptation in adolescents with spina bifida; visuographic tests of set shifting and inhibitory control; antecedents and consequences of early gains in child psychotherapy; clinically significant sexual concerns in a child welfare population; Oppositional Defiant Disorder symptoms in preschool children and pediatric primary care; the role of race, socioeconomic status, and system-of-care services in placement decision-making; psychiatric hospital admission decisions for children in foster care and reducing depression symptoms for children and adolescents in foster care; diagnosing attention deficit disorders; and evaluating the clinical efficacy of alternative therapies; as well as predicting substance use disorder in hospitalized severely ill psychiatric patients.
amnestic mild cognitive impairment\textsuperscript{119} and change in job status following traumatic brain injury in a military population\textsuperscript{130}; escalation to major depressive disorder of adolescents with subthreshold depressive symptoms\textsuperscript{121}; bipolar spectrum disorders in youth\textsuperscript{112}; cognitive effort using MMPI-2 items\textsuperscript{125}; and early sexual debut among adolescents in psychiatric care.\textsuperscript{124}

And, HO-CTA has been productively used in criminal justice, predicting injuries of women in episodes of intimate partner violence\textsuperscript{125}, determining risk factors for partner violence among a national sample of combat veterans\textsuperscript{126}, identifying subgroups at high risk of dropping out of domestic batterer treatment\textsuperscript{127}, evaluating the accuracy of risk assessments on sexual recidivism\textsuperscript{128}, and identifying three types of violent offenders and predicting violent recidivism while on probation.\textsuperscript{129}

The EO-CTA Algorithm

The enumerated-optimal CTA (EO-CTA) algorithm enumerates all of the statistically valid combinations of attributes that exist for the first three nodes—which dominate the solution—and selects the model yielding highest ESS.\textsuperscript{5}

In an application having a large sample and many attributes that yield a moderate ESS, manually constructing HO-CTA models using ODA software is an arduous task. However, in these circumstances, manually constructing an EO-CTA model using ODA software is essentially infeasible. Accordingly, special-purpose CTA software was created, offering numerous options to control model building and validation processes, and automating repetitive mechanical steps executed by the HO-CTA and EO-CTA algorithms.\textsuperscript{5}

Examples of statistical analyses using CTA software to execute the EO-CTA algorithm include predicting fibromyalgia symptom reduction by automated delivery of individualized guidance\textsuperscript{130}; resolving antisocial behavior among foster care youth\textsuperscript{131,132}, and Oppositional Defiant Disorder in early grammar school\textsuperscript{132}; and comparing initial placement in shelter care versus kinship care, and determining variables predictive of children staying less than 30 days in the shelter versus 30 days or longer.\textsuperscript{133} More examples of EO-CTA models are available in the current textbook on the ODA paradigm.\textsuperscript{5}

The GO-CTA Algorithm

Introduced in 2016, the globally-optimal CTA (GO-CTA) algorithm identifies all the unique statistically valid combinations of attributes (i.e., CTA models) that exist for the sample.\textsuperscript{5} The set of all of statistically valid CTA models for a sample is called the descendant family (DF) of optimal models. Identifying the model in the DF that reflects the “best” combination of accuracy and parsimony is a two-step process.

First, ESS is computed for every model in the DF in order to measure model predictive accuracy after removing the effect of chance—which is represented as 0\% for every model.

Second, the distance statistic (D) is used to adjust the ESS for every model in the DF, removing the effect of complexity—which is operationalized in terms of the number of endpoints. By definition, as the number of model endpoints increases, the model’s complexity and precision increases. And, by definition, as the number of model endpoints decreases (at least two endpoints are required for a model), the model’s complexity and precision decreases—and the model’s parsimony increases. The smallest value that D can attain is zero (which requires that ESS=100): the greater the value of D, the less accurate and the less parsimonious the model. By definition, the GO-CTA model has the smallest D statistic in the DF.\textsuperscript{5,134}

Different GO-CTA models may be reported for increasing levels of complexity. For example, if supported by substantive theory and justified by sufficient statistical power—then, for example, the best (smallest D) of the least-precise (2-strata), intermediate-precision (3-strata), and most-precise (4-strata) GO-CTA models may be reported.\textsuperscript{5,135}


Current Research and Development Directions in the ODA Laboratory

Collaborating with researchers developing new optimal methods, and/or using optimal methods in applied research, is a primary research and development priority for the ODA laboratory. Dr. Ariel Linden’s project\(^{136}\) illuminates use of optimal machine learning algorithms in causal inference, recently publishing numerous significant advances in this area: for example, the use of optimal machine learning algorithms to characterize participation in observational studies\(^{137}\), to assess covariate balance in matching studies\(^{138}\), identify structural breaks in single-group interrupted time-series designs\(^{139}\), and model dose-response relationships\(^{140,141}\); combining machine learning and matching methods to improve causal inference in program evaluation\(^{142}\); combining machine learning and propensity score weights to estimate causal effects in multivalued treatments\(^{143}\), and using the GO-CTA algorithm to create propensity score weights.\(^{135}\) Dr. Paul Yarnold’s collaborative project\(^{144}\) recently focused attention on designs involving ordered class variables.\(^{145-148}\) Two projects using optimal methods, recently underway, expand the cross-cultural aspects and substantive domains of research. Dr. Victor Oguoma’s project\(^{149}\) uses maximum-accuracy methods to study cardiovascular disease risk factors within low-to-middle-income sub-Saharan African populations.\(^{88,93}\) And, Dr. Timothy Ebert’s project\(^{150}\) explores the use of optimal algorithms and software\(^{151}\) in agricultural research: Dr. Ebert notes that a specific, unique set of priorities and risks applies in agricultural investigations.

In addition to promoting collaboration between laboratories, another primary research and development priority of the ODA laboratory is promoting within-laboratory collaboration. In this regard, exposing university faculty and students to optimal methods is a demonstrated effective method of amplifying collaboration using these methods within laboratories. As described by Dr. Fred Bryant\(^{152}\): “An initial set of ODA-based articles by Loyola faculty laid the groundwork for a sustained upsurge in the use of ODA among graduate students which has lasted for more than a decade and a half. These student projects subsequently fueled an increase in ODA-based publications by other Loyola Psychology faculty, who directly supervised the various student projects. Thus, ODA initially trickled down from faculty to students, but later grew up in the opposite direction” (p. 4). During most of the time period covered in the article the resources available to researchers were articles in the literature, and the first ODA textbook and software.\(^{4,153-155}\) Since the article was published a textbook\(^{5,156}\) covering the paradigm from its inception through 2016, an open source, peer-reviewed eJournal\(^{157}\) focusing on theoretical, applied, and pragmatic (“i.e., “how to”) aspects of maximum-accuracy analysis was created, and updated software became available for use in BIG DATA applications.\(^{5,158}\) Current projects in the ODA laboratory in this domain are focused on two additional resources that support within-laboratory collaboration: an instructor’s manual, and a site license for software/documentation.

Finally, an anticipated primary future research arena involves longitudinal functional studies made possible by imminent release of a validated intelligent health diary that identifies patient-tailored symptom-management strategies and provides individualized guidance. This automated system is capable of generating the most thorough real-time data ever obtained for individual cases representing numerous chronic and acute disease and injury diagnoses.\(^{130,159}\)

References


30Thompson DA, Yarnold PR (1995). Relating patient satisfaction to waiting time perceptions and expectations: The Disconfirmation Para-


105 Diesfeldt HFA (2009). Visuographic tests of set shifting and inhibitory control: The contri-
The role of race, socioeconomic status, and System of Care services in placement decision-making. *Children and Youth Services Review*, 50, 3-11.


120. Han SD, Suzuki H, Drake AI, Jak AJ, Houston WS, Bondi MW (2009). Clinical, cog-
nitive, and genetic predictors of change in job status following traumatic brain injury in a military population. *Journal of Head Trauma Rehabilitation, 24*, 57-64.


https://www.researchgate.net/project/Applying-machine-learning-to-improve-causal-inference


https://www.researchgate.net/project/Novometric-Analysis


https://www.researchgate.net/project/Optimal-Data-Analysis


http://ODAJournal.com


Author Notes

This review article is exempt from IRB review. No conflict of interest was reported.