The Classification Society Meeting 2018

Book of Abstracts

Stony Brook University
Stony Brook, New York, USA

June 20–23, 2018

Contents

Thursday June 21, 2018

President's Invited Address

Charles Bouveryon, Université Côte d'Azur

Modeling and Clustering of Networks with Textual Edges

Due to the significant increase of communications between individuals via social media (Facebook, Twitter, Linkedin) or electronic formats (email, web, e-publication) in the past two decades, network analysis has become an unavoidable discipline. Many random graph models have been proposed to extract information from networks based on person-to-person links only, without taking into account information on the contents. This paper introduces the stochastic topic block model, a probabilistic model for networks with textual edges. We address here the problem of discovering meaningful clusters of vertices that are coherent from both the network interactions and the text contents. A classification variational expectation-maximization algorithm is proposed to perform inference. Simulated datasets are considered in order to assess the proposed approach and to highlight its main features. Finally, we demonstrate the effectiveness of our methodology on real-word datasets.

Contributed Session 1

Abby Flynt, Bucknell University

***sARI*: A soft agreement measure for class partitions incorporating assignment probabilities**

Agreement indices are commonly used to summarize the performance of both classification and clustering methods. The easy interpretation/intuition that results from the Rand and Adjusted Rand indices, with other desirable properties, has led to their popularity over other available indices. While more algorithmic clustering approaches like k-means and hierarchical clustering produce hard partition assignments (assigning each observation to a single cluster), many modern techniques like model-based clustering, include information about the certainty of allocation of objects through class membership probabilities (soft partitions). In order to assess performance using traditional indices such as the Adjusted Rand Index, the soft partition is mapped to a hard set of assignments, which commonly overstates the certainty of correct assignments. This work proposes an extension of the Adjusted Rand Index (ARI), named the soft Adjusted

Rand Index (sARI). The new index benefits from the natural intuition underlying the ARI while properly incorporating the information from one or two soft partitions. It can be used in conjunction with the ARI to compare the similarities of hard to soft, or soft to soft partitions to the similarities of the transformed hard partitions. Simulation studies results appear to support the intuition that in general, transforming to hard partitions tends to increase the measure of similarity between partitions. In applications, the sARI more accurately reflects the cluster boundary overlap commonly seen in real data.

Kayla Frisoli, Carnegie Mellon University

Incorporating Sociodemographic Transitions and Family Structure into Historical Record Linkage

Record linkage is the process of identifying records corresponding to unique entities (e.g. individuals, companies) across datasets that do not have a unique identifier. While this process is commonly used with administrative data, linking historical databases allows researchers to better characterize topics like population mobility, the impact of events on a local or national scale, generational changes, and familial makeup. Most record linkage algorithms rely on text string similarities (e.g. similarity of last name or address) to determine whether two records originate from the same entity; however sometimes we expect to see particular changes that would not be properly captured by standard similarity metrics (e.g. if someone gets married between census dates, their last name and address may differ). In addition, due to computational constraints, linkage methods often only consider information from pairs of records without incorporating potential relationship information across records (e.g. parents, siblings). Our application of interest is early twentieth century Ireland census records from 1901 and 1911. These datasets have limited, non-standardized fields with errors due to varying education levels, changes in format/style over time, and the digitization of scanned, hand-written original records. These issues, coupled with high frequencies of common names and addresses, are more evidence for modeling additional sociodemographic and family network information to correctly link individuals across censuses. We crowdsource training data that captures the unique true matches that exist in our data through an R Shiny application. Once we acquire training data, we show that traditional record linkage approaches find difficulty uncovering matches that a human was able to identify by hand. We conclude with a discussion of possible framework extensions to incorporate this type of structure.

Forrest Paton, McMaster University

Clustering Gaussian Processes

The covariance kernel of a gaussian process is pivotal in predicting future process realizations and understanding the underlying functions behaviour. This presentation will cover methods to cluster processes based on kernel hyper-parameters and latent variables associated with the cluster.

Invited Lecture

Tanzy Love, University of Rochester

Outlier Detection in Model-Based Cluster Analysis

In model-based clustering based on normal-mixture models, a few outlying observations can influence the cluster structure and number. This paper develops a method to identify these, however it does not attempt to identify clusters amidst a large field of noisy observations. We identify outliers as those observations in a cluster with minimal membership proportion or for which the cluster-specific variance with and without the observation is very different. Results from a simulation study demonstrate the ability of our method to detect true outliers without falsely identifying many non-outliers and improved performance over other approaches, under most scenarios. We use the R package MCLUST for model-based clustering, but propose a modified prior for the cluster-specific variance which avoids degeneracies in estimation procedures.

Distinguished Dissertation Award Session

Yang Tang, McMaster University

Dimension Reduction with non-Gaussian Mixtures

Three novel non-Gaussian mixture models are developed for high-dimensional complex data. A mixture of joint generalized hyperbolic distributions (MJGHD) is introduced for asymmetric clustering for high-dimensional data. The MJGHD approach takes into account the cluster-specific subspace, thereby limiting the number of parameters to estimate while also facilitating visualization of results. We explore the possibility of discovering extreme voting patterns in the U.S. Congressional voting records by drawing ideas from the mixture of contaminated normal distributions. A mixture of latent trait models via contaminated normal distributions is proposed. We assume that the low dimensional continuous latent variable comes from a contaminated normal distribution and, therefore, picks up extreme patterns in the observed binary data while clustering. Finally, a clustering approach is developed for Boston Airbnb reviews, in the English language, collected since 2008. A penalized mixture of latent traits approach is developed to reduce the number of parameters and identify variables that are not informative for clustering. The introduction of component-specific rate parameters avoids the over-penalization that can occur when inferring a shared rate parameter on clustered data. All three models are motivated by, and therefore, applied to real data sets.

Michael Fop, University College Dublin

Advances in Model-based Clustering and Classification of Complex Data

Model-based clustering and classification methods have application in a broad range of contexts, spanning from astrophysics to food science, from medical diagnosis to the social sciences. Current research on the

topic is devoted to the development of novel approaches for the analysis of high-dimensional and complex data. This talk discusses some of the challenges faced and presents recent advancements in variable selection and parsimonious modeling for model-based clustering and classification.

Poster Session

Jeffrey Andrews, University of British Columbia-Okanagan

Addressing Overfitting in Mixtures of Factor Analyzers

The expectation-maximization (EM) algorithm is a common approach for parameter estimation in the context of cluster analysis using finite mixture models. This approach suffers from the well-known issue of convergence to local maxima, but also the less obvious problem of overfitting. Mixtures of factor analyzers assume an underlying factor analysis structure and thereby perform dimensionality reduction simultaneously during the alternating expectation conditional maximization (AECM) model-fitting process. Importantly though, both convergence to local maxima and overfitting remain a concern. We address these concerns by introducing an algorithm that augments the traditional AECM with the nonparametric bootstrap. Further simulations and applications to real data lend support for the usage of this bootstrap augmented AECM-style algorithm.

Katherine Clark, McMaster University

Choosing the Best Clustering Method: A Review

When clustering, it is often the case that the number of clusters is not known a priori. In this sense, model selection is a very important aspect of cluster analysis. Furthermore, there are often other considerations that play into model selection. For example, the number of latent variables may need to be selected, or the covariance structure. A review of model selection criteria for clustering is presented, covering of 50 years of work. Some examples are presented for illustration.

Brian Franczak, MacEwan University

Cluster Analysis using Mixtures of Asymmetric Distributions

Cluster analysis can be lucidly defined as the process of sorting similar objects into groups. When a finite mixture model is used for cluster analysis, we call the process model-based clustering. This talk will discuss the development of a mixture of contaminated shifted asymmetric Laplace factor analyzers (MCSALFA). This model will be well suited for the analysis of high-dimensional data; specifically, where the number of variables exceeds the number of observations. In addition to providing a classification of similar observations, the MCSALFA will also provide a classification of an observation as being either 'good' or 'bad', unifying the fields of model-based clustering and outlier detection. From a methodological

standpoint, the MCSALFA will unify the factor analysis model and the contaminated mixture model. The classification performance of the MCSALFAs will be demonstrated using a real data set.

Regina Kampo, McMaster University

Clustering Incomplete Data using Evolutionary Algorithms

Generally, parameter estimation in mixture model-based clustering is performed using the expectation-maximization (EM) algorithm. However, many clustering problems are plagued with missing data. Computational strategies are presented for handling mixtures of multivariate Gaussian distribution when data are missing at random. Specifically, an evolutionary algorithm that focuses on searching the parameter space, hence providing an approach that is more robust with respect to local maxima, is developed. The proposed methodology is illustrated through both simulated and real data sets with varying proportions of missing values.

Hans-Friedrich Köhn, University of Illinois at Champaign-Urbana

Additive Trees for Fitting Three-Way (Multiple Source) Proximity Data

Additive trees are graph-theoretic network models for proximity data collected on a set of objects. Each object is represented as a node in a connected graph, so that the length of the paths connecting the nodes reflects the proximities observed among objects.

For additive trees, Carroll, Clark, and DeSarbo (1984) developed the INDTREES algorithm to accommodate three-way two-mode data in explicitly modeling individual differences that might underlie the input proximity judgments. The path lengths of the individual trees are estimated using a conjugate gradient routine for minimizing a least-squares loss function that is augmented by a penalty term to account for violations of the constraints imposed by the four-point condition (that determines an additive tree structure).

This study presents an alternative method for fitting additive trees to three-way two-mode proximity data that does not rely on gradient-based optimization nor on penalty terms. Instead, the path lengths of the trees are estimated by an iterative projection algorithm minimizing a constrained least-squares loss function. The constraints are defined by the four-point condition.

Simulations are reported for evaluating the performance of the proposed method in direct comparison with that of the INDTREES algorithm. A real-world data set is used for illustration.

Yajun Liu, Northwestern University

Canadian Charity Vulnerability Prediction Using Two-step Support Vector Machine (SVM)

In the past years, support vector machine (SVM) is widely used in multivariate data classification and one variable functional data analysis (FDA). With the multivariate time-series data set (Canadian charitable

organization development records), we develop a two-step SVM algorithm to classify charities into two groups: success or failure (i.e. vulnerable). The first step of the algorithm compresses data from each year to a vector containing the information from all the variables, then the second step processes the transformed data as a typical functional data classification problem. In real data sets processing, imputation methods (zero-imputation and mean-imputation) and underlying information in unselected variables are used synthetically to impute missing values in the data sets. Rare events classification is also solved. Finally, we can give a prediction of the charities vulnerability (success or failure) which is helpful for the government to make financial decisions and charities to assess their financial situations.

Alan Mishler, Carnegie Mellon University

Clustering Students and Inferring Skill Set Profiles with Skill Hierarchies

Cognitive diagnosis models (CDMs) are a popular tool for assessing students' mastery of sets of skills or abilities (Junker & Sijtsma, 2001). Given a set of K skills that are tested on an assessment, students are classified into one of 2^K latent classes based on whether they have mastered each skill or not. Traditional approaches to estimating these classes are computationally intensive and may be infeasible on large datasets. Instead, proxy skill estimates can be generated from the observed responses and then clustered, and these clusters can be assigned to different skill profiles (Chiu & Douglas, 2009).

Building on the work of Nugent, et al. (2010), we consider how to optimally perform this clustering when not all 2^K classes exist, e.g. because of hierarchical pre-requisite relationships among the skills, and when not all classes are present in the population. We consider how many clusters to extract from a dendrogram generated by hierarchical agglomerative clustering, and we compare this approach to the empty k-means algorithm described in Nugent, et al. (2010), with starting centers derived randomly, derived from the Q-matrix that species the skills that each item tests, or derived from simulated student responses generated under conjunctive and disjunctive models. We also use simulated student responses to investigate semisupervised clustering approaches. We explore methods for assigning clusters to skill profiles when the skill space is constrained due to hierarchical skill pre-requisite relationships. Finally, we consider ways to learn these relationships among skills when these are not assumed to be known.

Mengyuan Ren, San Jose State University

Classification via a family of Parsimonious Generalized Hyperbolic Mixtures

Model-based clustering is an effective tool for identifying homogeneous subpopulations within a heterogeneous population. Traditionally, mixture-model based clustering and classification approaches have focused on the Gaussian mixture model. However, due to the limitations of the Gaussian mixture model, statisticians have proposed using mixtures of non-Gaussian mixtures for classification. As such, there has been an increase in work using mixtures of t-distributions, mixtures of skewed distributions, etc. A drawback of mixture-model based approaches is that they are ill-suited when fitted to data with a large number of variables. This project will focus on the development of a family of parsimonious generalized hyperbolic mixture models that will be well suited for the analysis of high-dimensional data. The parsimonious models are based on the mixtures of factor analyzers model. The mathematical development of our mixture of

generalized hyperbolic distributions model relies on its relationship with the generalized inverse Gaussian distribution. The expectation-maximization (EM) algorithm will be used for estimation. The clustering performance of our mixture models will be demonstrated using simulated data sets.

Jordyn Walton, McMaster University

Clustering with Matrix Variate Distributions: A Review

Clustering is the search for underlying group structure in data. Although mixture model-based clustering is well-established in the multivariate case, there is a relative paucity of work on clustering with matrix variate distributions. Existing work on clustering using matrix variate mixtures will be reviewed. Although the body of literature spans less than a decade, approaches have already been developed for heavy-tailed and asymmetric clusters as well as some work on clustering in high-dimensions. All of this work is reviewed along with some peripheral work.

Friday June 22, 2018

President's Address

Paul McNicholas, McMaster University

Selected Problems in Classification

Selected problems in classification are considered, either via specific datasets or general problem types. In each case, the problem is introduced before one or more potential solutions are discussed and applied. The problems discussed include data with outliers, longitudinal data, and three-way data. The proposed approaches are predominantly mixture model-based.

Invited Lecture

Katrijn Van Deun, Tilburg University

Finding the hidden link: Statistical methods for multi-source high-dimensional data

Research in many disciplines, including the behavioural and social sciences, has entered the era of big data. Many detailed measurements are taken and multiple sources of information are used to unravel complex multivariate relations. For example, in studying obesity or depression as the outcome of environmental and genetic influences, researchers increasingly collect survey, dietary, biomarker and genetic data from the same individuals. Revealing the variables that are linked throughout these different types of data gives crucial insight in the complex interplay between the multiple factors that determine human behavior, e.g., the concerted action of genes and environment in the emergence of obesity or depression.

Although linked more-variables-than-samples (or, high-dimensional) multi-source data form an extremely rich resource for research, extracting meaningful and integrated information is challenging and not appropriately addressed by current statistical methods. The challenge is to select - in an automated way - those variables that are linked throughout the different blocks and this eludes current available methods for data analysis. A first problem is that relevant information is hidden in a bulk of irrelevant variables with a high risk of finding incidental associations. Second, the sources are often very heterogeneous, which may obscure apparent links between the shared mechanisms.

In this presentation we will discuss the challenges associated to the analysis of large scale multi-source data and present a sparse common and distinctive components approach to address the challenges.

Invited Session on Clustering: Recent Advances and Historical Context

Sanjeena Dang, Binghamton University

Clustering skewed data using mixtures of multivariate normal-inverse Gaussian distributions using a Bayesian framework

Multivariate normal inverse Gaussian (MNIG) distributions possess an appealing property that they can represent symmetric as well as skewed populations with computational simplicity. This makes MNIG attractive for model-based clustering. The MNIG model arises from a mean-variance mixture of a multivariate normal distribution with the inverse Gaussian distribution. Here, a Bayesian approach using Gibbs sampler, an alternate approach to the traditional EM algorithm, is presented for mixtures of MNIG models. A novel approach to simulating from matrix generalized-inverse Gaussian distribution is also discussed. Application on simulated data sets with symmetric and skewed subpopulations as well as a real data set is presented.

Stanley Sclove, University of Chicago at Illinois

From Histograms to Clusters to Predictive Distributions

Consider the problem of modeling a dataset of employee days ill, or accidents in a population of insureds. One can consider a spectrum of levels of granularity in describing the population, from a single distribution, to a bimodal distribution, to a mixture of two or more distributions, to modeling the population at the individual level.

Cluster Analysis, Mixture Model, and Bayesian models will be considered as one moves from one end of this spectrum to the other. In 1920, Greenwood and Yule developed a model with a distribution across the population parameter which would now be called a Bayesian model. These developments are reviewed in relation to the "new" Predictive Analytics. Extensions are made in terms of mixture priors and posterior estimation.

Invited Lecture

Volodymyr Melynkov, The University of Alabama

On Matrix Mixture Modelling and Model-based Clustering

The finite mixture modeling and model-based clustering literature mainly considers the analysis of data observed in the vector form, where each vector element represents a specific variable. In cases when data are presented in the matrix form, the number of available mixture models is rather limited. We discuss the existing literature on the topic and propose an approach relying on the application of transformations to normality to produce mixtures capable of modeling skewed matrix data groups. The proposed methodology is illustrated on a variety of simulated and real-life sets of data.

Contributed Session 2

Angelina Pesevski, McMaster University

Clustering, Outlier Detection, and Visualization for High-Dimensional Data

High-dimensional data sets present one of the most profound analytic challenges in the “big data" era. A common objective is to look for homogeneous subgroups within high-dimensional data (e.g. finding groups of customers with similar behaviour). Model-based clustering is a popular method for finding homogeneous partitions and uses a mixture of distributions where each component density corresponds to a group, or cluster. A well-known method called high-dimensional data clustering (HDDC) has given rise to a computationally efficient family of Gaussian mixture models for clustering high-dimensional data. HDDC is based on the idea that most of high-dimensional data live around lower-dimensional subspaces, hence the clustering procedure can be carried out in these lower-dimensional subspaces. This, in turn, reduces the number of parameters to be estimated significantly, creating a tractable clustering problem. The HDDC family of models has gained vast attention due to its superior performance compared to other families of mixture models. As all Gaussian mixture model-based approaches, these models are sensitive to outlying or spurious points. A variation of these models using the multivariate-t distribution (tHDDC models) has been proposed to account for outliers in data. Through the presence of a concentration parameter (i.e., the degrees of freedom), these models can give very good clustering performance in the presence of outliers. However, they do not provide a way to identify the outlying observations. A mixture of contaminated Gaussian distributions is developed for robust clustering of high-dimensional data. Each component of our contaminated mixture has parameters, which can be estimated from the data, that control the proportion of outlying observations and the degree of contamination, respectively. Hence, the proportion of outlying points does not need to be specified prior to fitting the models. The performance of our 14 parsimonious models is evaluated using both simulated and real data and compared to the HDDC and tHDDC models.

Xu (Sunny) Wang, Wilfred Laurier University

Interdisciplinary Approaches to Automated Obstructive Sleep Apnea Diagnosis through High-dimensional Multiple Scaled Data Analysis

As we collect more complex and high-dimensional data, analysis and interpretation can be challenging and require sophisticated analytic techniques. It may be no longer effective to independently apply methods from a specific scientific discipline such as statistics, mathematics, or computing science. Alternatively, developing new methods by combining techniques from these three sciences is likely more powerful and useful in detecting hidden features in complex data. Obstructive sleep apnea (OSA) is a multifactorial disorder, thus making it necessary to study many different types of data; DNA sequences, multiple time series, metabolites, airflow in airway, and shape analysis of airway and patientsâĂŹ faces, to give some examples. OSA data are an example of complex and multi-dimensional data. Traditionally, each different type of data was analyzed separately using techniques in statistics or in machine learning. In this article, combining the analyses of three data sets from independent OSA studies, we illustrate the effectiveness of combining techniques in statistics, machine learning, and computational topology. A novel geometric OSA severity index (GSI) is developed using methods from computational geometry. This index measures the volume of the airway obstruction in OSA patients. The lower GSI value is, the severer the airway obstruction is. Persistent homology is uniquely employed to extract the importance information from 28 high-dimensional polysomnography (PSG). Random forest and principal component analysis are used and compared to identify important variables, while logistic regression and random forest are used and compared to verify the prediction power of the identified importance variables. The results strongly indicate that persistent homology can accurately extract importance information from PSG, and the identified important variables are very helpful of predicting obstructive apnea-hypopnea index (ahi). Cluster analysis is used to identify the pattern of the survey information, and the importance variables of survey questionnaires are also identified by random forest. The results from all three independent studies are very meaningful in clinical studies and can be used as guidance for clinical practitioners.

Michael Gallaugher, McMaster University

Finite Mixtures of Skewed Matrix Variate Distributions

Clustering is the process of finding underlying group structures in data. Although mixture model-based clustering is firmly established in the multivariate case, there is a relative paucity of work on matrix variate distributions and none for clustering with mixtures of skewed matrix variate distributions. Four finite mixtures of skewed matrix variate distributions are considered, the matrix variate skew-t, generalized hyperbolic, variance gamma and normal inverse Gaussian distributions. Parameter estimation is carried out using an expectation-conditional maximization algorithm, and both simulated and real data are used for illustration.

Saturday June 23, 2018

Invited Session on Applications of Classification

Stephen France, Mississippi State University

Overlapping Clustering: A Framework, Software, Empirical Analysis, and Applications

Overlapping clustering is a variant of clustering where each item may be a member of more than one cluster. It has found particular use in marketing segmentation, where products may be members of more than one usage segment. Overlapping clustering methods have been developed from different clustering traditions. Additive decomposition methods, such as ADCLUS and INDCLUS, are discrete variants of continuous mapping methods. Fuzzy clustering methods can generate overlapping clustering solutions by setting thresholds for cluster membership. Partitioning clustering methods, such as k-means clustering, can be extended to overlapping clustering by relaxing the cluster membership constraints.

The R software package and associated framework described in this talk implements overlapping clustering methods from all of these traditions and also implements the generalized omega metric for cluster validation. Attention is paid to the optimization of the models and to the issues of solution initialization and locally optimal solutions. Empirical work on both synthetic and real world datasets is described. Applications are given in customer and product segmentation.

Justin Gross, University of Massachusetts Amherst

Clustering and Classification of Ideological Writings by Salience and Valence of Core Values

Common techniques of ideological measurement sacrifice our ability to analyze the patterns of ideas constituting the conceptual basis of ideologies in exchange for a low-dimensional (typically unidimensional) representation easily included as a variable in an explanatory model. Employing, by contrast, the more natural definition of ideology as a shared configuration of core political beliefs, I identify several challenges to clustering and classification that arise. Using a collection of work by contemporary American political opinion writers, I investigate a number of possible approaches to deal with non-randomness of missing data, asymmetric "similarity", fuzzy class boundaries, and multiple raters (differing reader perceptions).

Contributed Session 3

Ron Yurko, Carnegie Mellon University

Detecting Data Analysis Patterns in Text and Graphs to Characterize Student Learning in an Introductory Statistics & Data Science Course

As part of a revamp of the general education introductory statistics course at Carnegie Mellon, an interactive data explorer platform was built that allows students to fully engage in the entire data analysis workflow without relying on a particular programming language. Its functionality includes tracking actions and storing answers including open-ended questions where students describe graphs and interpret results. Under the assumption that text gives a richer picture of student comprehension (vs a right/wrong multiple choice question), we use clustering procedures to compare the topics, semantics, and complexity structure of student answers in lab sessions over the course of the semester, as well as in their final data analysis reports. Rather than employing topic modeling or natural language processing alone, we are able to identify the relationship between the descriptive text and analysis decisions by students. This allows us to flag students who answered "differently" and use their actions, such as which graph they created, to assist us in understanding "why". We discuss implications of our results on gaining insight into how students from different backgrounds approach introductory statistics and data analysis, potentially establishing a first-step autograder, and improving our overall understanding of the science of data science.

Tyler Roick, McMaster University

Clustering Discrete Valued Time Series

Despite a rapidly growing literature on clustering techniques, and model-based clustering in particular, there is a relative dearth of approaches for clustering discrete data and, in particular, discrete time series data. A review of the application of thinning operators to adapt the ARMA recursion to the integer-valued case is first discussed. A class of integer-valued ARMA (INARMA) models arises from this application. Our focus falls on INteger-valued AutoRegressive (INAR) type models. The INAR type models can be used in conjunction with existing model-based clustering techniques to cluster discrete valued time series data. This approach is then illustrated with the addition of autocorrelations to determine relevant lag times. With the use of a finite mixture model, several existing techniques such as the selection of the number of clusters, estimation using expectation-maximization and model selection are applicable. The proposed model is then demonstrated on various simulated and real data to illustrate its clustering capability.

Youn Seon Lim, Hofstra University

Estimation of The Parameters of The Reduced RUM Model by Simulated Annealing

In this study, a simulation-based method for computing joint maximum likelihood estimates of the reduced reparameterized unified model parameters is proposed. The central theme of the approach is to reduce the complexity of models to focus on their most critical elements. In particular, an approach analogous to

joint maximum likelihood estimation is taken, and the latent attribute vectors are regarded as structural parameters, not parameters to be removed by integration with this approach, the joint distribution of the latent attributes does not have to be specified, which reduces the number of parameters in the model.