Theoretical Computer Science 752 (2018) 41–55

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Data fusion using Hilbert space multi-dimensional models

Jerome Busemeyer [a,*], Zheng Wang [b]

[a] Indiana University, United States
[b] The Ohio State University, United States

ARTICLE INFO

Article history:
Received 21 May 2017
Received in revised form 24 November 2017
Accepted 4 December 2017
Available online 12 January 2018

Keywords:
Quantum probability
Hilbert space
Multidimensional models
Contingency table analysis
Data fusion

ABSTRACT

General procedures for constructing, estimating, and testing Hilbert space multi-dimensional (HSM) models, built from quantum probability theory, are presented. HSM models can be applied to collections of K different contingency tables obtained from a set of p variables that are measured under different contexts. A context is defined by the measurement of a subset of the p variables that are used to form a table. HSM models provide a representation of the collection of K tables in a low dimensional vector space, even when no single joint probability distribution across the p variables exists. HSM models produce parameter estimates that provide a simple and informative interpretation of the complex collection of tables.

1. Introduction

When large data sets are collected from different contexts, often they can be summarized by collections of contingency tables or cross-tabulation tables. Suppose there are p different variables $(Y_1, \ldots, Y_p)$ that can be used to measure objects, or events, or people. It may not be possible to measure all p variables at once, and instead, only a subset of variables $(Y_{k_1}, \ldots, Y_{k_s})$, $s \leq p$, can be measured at once. Each subset forms a context k of measurement [8] [16]. More than one context can be collected, which forms a collection of K data tables $(T_1, \ldots, T_k, \ldots T_K)$, each collected under a different context k. Each table T_k is a joint relative frequency, or contingency, table based on a subset of variables.

For example, a research problem could involve three variables (Y_1, Y_2, Y_3), but some tables might include only some subset of the three variables. One context might involve the measurement of a single variable Y_1 that has 5 values to form a 1-way frequency table T_1 composed of 5 frequencies. Another context could be used to form another 5×3 table T_2, composed of joint frequencies for two variables (Y_1, Y_2). A third context could form another 3×2 table T_3 containing variables (Y_2, Y_3), and fourth could form a 5×2 table containing variables (Y_1, Y_3).

A critical data fusion problem arises: How to integrate and synthesize these K different tables into a compressed, coherent, and interpretable representation? This question arises in relational data base theory [1], where the problem is to find a *universal* relation capable of reproducing a set of component relations defined on all the data. In statistics, the problem is to find a *single* latent p-way joint distribution of the observed variables that can reproduce the frequencies in the K different tables by marginalizing across variables in the p-way table [5]. Often Bayesian causal networks are used to reduce the number of latent probability parameters by imposing conditional independence assumptions [13]. Unfortunately, however, in many cases, no *universal* relation exists and no p-way joint distribution can reproduce the observed tables (see, e.g., [7] [2])! This occurs when the data tables violate consistency constraints required by classical (Kolmogorov) probability theory.

* Corresponding author.
E-mail address: jbusemey@indiana.edu (J. Busemeyer).

https://doi.org/10.1016/j.tcs.2017.12.007
0304-3975/

In this case, no Bayesian network representation composed of the p-variables can even be formed. In the following sections, we give concrete examples of the various types of possible joint probability violations.

The data fusion problem presented above is not new. Discussions concerning the conditions for the existence of a single joint distribution to reproduce a collection of K different tables has a long history that goes all the way back to George Boole [26]. Vorob'ev [29] was one of the first to begin a rigorous program to identify all of these conditions. For example, the famous Bell inequality describes a condition required for 4 two-way tables constructed according to the Bell experimental design to be described by 4 binary random variables [17]. The relevance of this data fusion problem for human judgments was first pointed out by Aerts and Aerts [4]. Recent work has identified very general conditions required for arbitrary collections of tables [12] [15].

Hilbert space multi-dimensional (hereafter, denoted HSM) modeling is based on quantum probability theory [19] [25]. It provides a promising new solution to the data fusion problem by constructing a model that has (a) a single finite state vector that lies within a low dimensional vector space, and (b) by forming a set of measurement operators that represent the p measurements. In this way, we can achieve a compressed, coherent, and interpretable representation of the p variables that generate the complex collection of K tables, even when no standard p-way joint distribution exists. In a Hilbert space model, the state vector represents respondents' initial tendencies to select responses to each of the p measurements; the measurement operators describe the inter-relations between the p measurements (independent of the initial state of the respondents).[1]

HSM models are similar to traditional multi-dimensional scaling (MDS) models [32], but also different from them in important aspects. Like traditional MDS models, HSM models are based on similarity relations between entities located within a vector space. However, traditional MDS models define the similarity relations by inner products between vectors in a real vector space, whereas HSM models define similarity relations by projections onto subspaces of a complex vector space. Also, MDS models are designed to account for a single 2-way symmetric similarity matrix, whereas HSM models can be applied to multiple similarity matrices (e.g., when the similarity relation is asymmetric, see [27]).

The article is organized as follows. First, we provide an artificial data example that illustrates how consistency requirements of a single p-way joint distribution can be violated. Second, we describe the general procedures for building HSM models. Third, we illustrate the general method by applying the principles to the artificial data set. Finally, we apply the theory to construct a HSM model for a real data set.

2. An artificial example

In this section we present an artificial example that serves to illustrate several ways that a joint probability model can fail. (Later we present a real data example, but the artificial example provides a clearer illustration of all of the problems). Suppose that the relations among three attributes or variables are investigated, labeled A, B, C. For example, these variables could represent ratings about the Adeptness (yes, no), Brilliance (low, medium, high), and Confidence (1, 2, 3, 4) of a political candidate reported on various large social media sources. Suppose that six contingency tables, shown together in Table 1, are collected from various sources. The table labeled $p(C = c_k)$ is a 1-way table containing the relative frequency of ratings for 4 increasing levels of Confidence obtained from one source. For example, the relative frequency of the second level of confidence equals $p(C = c_2) = .2788$. Table $p(A = a_i, B = b_j)$ is a 2×3 contingency table containing the relative frequencies of responses to Adeptness (yes, no) and then Brilliance (low, medium, high) obtained from another source. Table $p(A = a_i, C = c_k)$ is a 2×4 contingency table containing the relative frequencies of responses to Adeptness and then Confidence presented in the AC order. For example, the relative frequency of saying yes to Adeptness and then choosing Confidence level 2 equals $p(A = a_1, C = c_2) = .0312$. Table $p(C = c_k, A = a_i)$ is a table produced when the attributes (Confidence, Adeptness) were asked in the CA order. For example, the relative frequency of choosing Confidence level 2 and then saying yes to Adeptness equals $p(C = c_2, A = a_1) = .0297$. (The CA table is arranged in the same format as the AC table so that they can be directly compared.) The table $p(B = b_j, C = c_k)$ is a 3×4 contingency table containing the relative frequencies of responses to Brilliance and then Confidence in the BC order; and the table $p(C = c_k, B = b_j)$ is a table produced by the opposite CB order. (Again it is displayed in the same format to facilitate comparison.) Each of the six tables forms a context for judgments.

2.1. Does a joint distribution exist?

The following question can be asked about Table 1: Does a *single* 3-way joint probability distribution of the observed variables exist that can reproduce Table 1? The 3-way joint probability distribution is defined by 3 discrete random variables, A with 2 values, B with 3 values, and C with 4 values, that generate $2 \cdot 3 \cdot 4 = 24$ latent joint probabilities that sum to one: $\pi(A = a_i \cap B = b_j \cap C = c_k)$, where, for example, A is a random variable with values a_1 for yes and a_2 for no, and similar definitions hold for the other three random variables. For example, the relative frequency of $(A = a_2, C = c_4)$ in the table $p(A = a_i, C = c_k)$ is predicted by the marginal $\pi(A = a_2, C = c_4) = \sum_j \pi(A = a_2 \cap B = b_j \cap C = c_4)$, and the relative

[1] Technically, a Hilbert space is a complete inner product vector space defined on a complex field (although it can be restricted to the real field). The norm for the vector space is determined by the inner product. Our vector spaces are of finite dimension, and so they are always complete.

Table 1
Six different contingency tables produced by answers to attributes A, B, C.

$p(C=c_k)$			
1	2	3	4
0.2186	0.2788	0.2551	0.2475

$p(A=a_i, B=b_j)$		
0.0721	0.5777	0.0078
0.1235	0.0374	0.1815

$p(A=a_i, C=c_k)$			
0.0388	0.0312	0.2675	0.3201
0.1554	0.1506	0.0182	0.0183

$p(C=c_k, A=a_i)$			
0.0233	0.0297	0.2279	0.2212
0.1953	0.2491	0.0272	0.0264

$p(B=b_j, C=c_k)$			
0.1266	0.0476	0.0049	0.0165
0.0915	0.0911	0.2158	0.2167
0.1086	0.0320	0.0133	0.0354

$p(C=c_k, B=b_j)$			
0.0680	0.0762	0.0581	0.0713
0.1089	0.1391	0.1273	0.0924
0.0416	0.0635	0.0697	0.0838

Table 2
Three 2×2 tables produced from Table 1. Left table is formed from $p(A=a_i, B=b_j)$, middle is formed from $p(B=b_j, C=c_k)$, right is formed from $p(A=a_i, C=c_k)$.

$p(X=x_i, Y=y_j)$	
0.6498	0.0078
0.1609	0.1815

$p(Y=y_j, Z=z_k)$	
0.3568	0.4539
0.1406	0.0487

$p(X=x_i, Z=z_k)$	
0.0700	0.5876
0.3060	0.0365

frequency of $(C=c_4)$ is predicted by the marginal $p(C=c_4)=\sum_{i,j}\pi(A=a_i \cap B=b_j \cap C=c_4)$. Note that this 3-way joint distribution is completely general without any type of independence restrictions.

The answer to our question is that no single 3-way joint distribution of the three observed variables can reproduce Table 1. First of all, the 3-way distribution requires the marginal distribution of a single random variable to be invariant across contexts. Marginal invariance is based on the law of total probability. This requirement fails (e.g., $p(C=c_2)=.2788$ from the 1-way table, which is smaller than that from the 2-way table $p(A=a_1, C=c_2)+p(A=a_2, C=c_2)=.0312+1506=.1818$.) A second problem is that the order that questions are asked changes the 2-way distributions for some pairs (e.g., the distribution for the context BC is not the same as the distribution for the context CB.) Order effects violate the commutative property required by the joint probability model.

The tables also violate another consistency requirement, called the Leggett–Garg [21] inequality, which concerns the correlations between pairs of variables required by a single 3-way joint distribution. To illustrate this in a simple manner, consider the three tables shown in Table 2. These tables were formed by defining new variables X, Y, Z as follows: $X=A$; $(Y=y_1)=(B=b_1)\cup(B=b_2)$ and $(Y=y_2)=(B=b_3)$; $(Z=z_1)=(C=c_1)\cup(C=c_2)$ and $(Z=z_2)=(C=c_3)\cup(C=c_4)$. The Leggett–Garg inequality implies the following restriction on the 2×2 joint probabilities required by the 3-way joint probability model (see Appendix for a simple proof for this inequality):

$$\pi(X \neq Y)+\pi(Y \neq Z)-\pi(X \neq Z) \geq 0, \tag{1}$$

where for example, $\pi(X \neq Y)=\pi(X=x_1 \cap Y=y_2)+\pi(X=x_2 \cap Y=y_1)$, which is simply the sum of the probabilities that the attributes produce different yes–no answers. Using the data in Table 2, the Leggett–Garg value equals $-.1303$, which is below the zero bound required by the 3-way joint probability model.

Other consistency requirements can be shown to apply for larger sets of variables. For example, four 2×2 tables (AC, AD, BC, BD) generated from 4 binary variables (A, B, C, D) must satisfy the famous Bell inequalities in order to be reproduced by a 4 way joint distribution of the four observed variables. More general formulations for linear inequalities required for a p-way joint distribution to reproduce a collection of K tables produced by subsets of p variables is provided in [15] [16] [12].

2.2. A general test of the joint distribution model

More generally, we can test whether or not a 3-way joint probability distribution can fit the data in Table 1 by estimating all of its parameters from the data to produce a closest fit. The 3-way joint probability model has 24 joint probability parameters, $\pi(A = a_i \cap B = b_j \cap C = c_k)$. Using the law of total probability, these 24 model parameters can be used to predict the 50 cell probabilities in Table 1. Parameter estimation programs can be used to search for parameters that produce the closest fit according to the Kullback–Leibler (KL) divergence. Denote π_n as the predicted cell probability and define p_n, as the observed cell probability. Then the (KL) divergence is defined as

$$D = \sum_n p_n \cdot ln\left(\frac{\pi_n}{p_n}\right).$$

For sample data, with N observations per table, the KL divergence can be converted into a chi square statistic $G^2 = -2 \cdot N \cdot D$. The null hypothesis states that the data in Table 1 was generated by a 3-way joint probability model. If the null hypothesis is correct, then G^2 has a chi-square distribution with degrees of freedom $df = (50 - 6) - (24 - 1)$ (the probabilities in each of the 6 tables sum to one, and so 6 of the 50 probabilities are linearly constrained; the 24 joint probabilities sum to one and so one probability is linearly constrained). Using G^2, we can compute a p-value, which equals the probability of obtaining the observed G^2 or greater under the null hypothesis. If it is below the significance level $\alpha = .05$, then we reject the null hypothesis.

Following these procedures, we fit the 3-way joint probability model to the data in Table 1, and the KL divergence equals $D = -0.5431$. If we assume that each table is based on 100 observations, then $G^2 = 54.31$ and $p = .0001$, which is a statistically significant difference.

The above test of the joint distribution based on the KL divergence, is not limited to this example. The same principles can be extended to p variables, that produce K tables of various sizes collected under different contexts.[2]

The proposed test of a p-way joint distribution to account for a collection of K contingency tables formed by subsets of the p-variables does not rule out *all* Kolmogorov models. A more general $p+q$ joint distribution can be chosen to reproduce the data tables by using an additional q random variables (see, e.g., [16]). The proposed non-parametric method only tests a p-way joint distribution based on the observed p-variables.

3. Empirical evidence from research on document retrieval

Violations of a single joint probability model for the observed variables have observed in several empirical studies investigating document retrieval.

Question order effects were observed by Bruza and Chang [9]. They collected a large sample of participants from the Amazon Mechanical Turk platform. Five query terms about topics were used to examine question order effects. Participants were told to read a brief description of each topic (e.g., research on emerging branding trends), and then they were asked to rate an article (presumably returned from a web search engine) on two different dimensions (e.g., how relevant is this article for the topic, how interesting is the article for this topic). One group of participants rated the pair of topics in a one order (e.g., one group rated relevance first) and another group in the opposite order (e.g., another rated interest first). The results revealed strong an significant order effects from three of the five topics.

Similar findings of order effects on relevance judgments for document retrieval were reported by Wang et al. [31] using Chinese student participants. A total 15 different pairs of topics were taken from articles appearing in Wikipedia and the Chinese Daily. In this study, a pair of articles were retrieved on a topic, and the participants had to rate each article in the pair for its relevance. The order that the articles were rated in a pair was manipulated across groups. Out of 15 topics, 4 produced statistically significant order effects.

Violations of marginal invariance were observed by DiNunzio, Bruza, and Sitbon [14]. Once again, a large sample of participants were collected from the Amazon Mechanical Turk platform. In this study, participants were asked to categorize 82 different articles from Reuters about shipping and crude oil. For each article, one group was asked about a single category (e.g., was this article about shipping), and another group was asked about the conjunction (e.g., crude oil and shipping). The main comparison was between the proportion of answers to a single categorization (e.g., proportion of shipping categories when asked alone) and the total proportion to summed across the two mutually exclusive and exhaustive conjunctions (e.g., proportion of oil and shipping plus proportion of not oil and shipping). Large and significant differences were found between the two conditions: for experienced workers, the total probability from the conjunction task matched the categorization probabilities reported by Reuters, but their proportions for the single categories over or underestimated those reported by Reuters.

[2] see https://arxiv.org/abs/1704.04623 for a different example using 4 variables.

4. Why apply quantum probability to data from information retrieval?

The idea of applying quantum probability to the field of information retrieval was proposed several years ago by Keith van Rijsbergen [28]. Dominic Widdows also promoted the use of quantum theory to information retrieval [30]. For the most recent developments concerning the application of quantum theory to information retrieval, see [22].

One of the reasons that van Rijsbergen gives for applying quantum theory to information retrieval is that it provides a sufficiently general yet rigorous formulation for integration of all three kinds of approaches – logical, vector space, and probabilistic – used in the past for information retrieval. As he points out, important concepts in quantum probability – state vectors, observables, uncertainty, complementarity, superposition, and compatibility – all readily translate to analogous concepts in information retrieval.

Another important reason for considering quantum theory is that much of the data of interest is generated by human judgments, which frequently violate various rules of classical (Kolmogorov) probability. There is now a large literature that applies quantum theory to human judgment and decision making [10] [20]. For example, human judgments have been found to violate the rule that conjunctive probabilities cannot exceed the probability of a constituent event, which can be interpreted as a violation of total probability. Quantum probability provides a formulation for explaining these and other phenomena that appear puzzling from a classical probability point of view (see, e.g., [11] [2]).

5. Basics of quantum probability theory

HSM models are based on quantum probability theory and so we need to briefly review some of the basic principles used from this theory.[3]

Suppose we have p variables $(Y_i, i = 1, \cdots, p)$ and each variable, such as Y_i, produces one of a finite set of n_i values when measured. In quantum theory, Y_i is called an observable. The measurement outcome generated by measuring one of the p variables produces an event. For example, if variable Y_1 is measured and it produces the value y_i, then we observe the event $A = (Y_1 = y_i)$.

Quantum theory represents events within a Hilbert space H. Quantum theory defines an event A as a *subspace* of the Hilbert space. Each subspace, such as A, corresponds to a projector, denoted P_A for subspace A, which projects vectors into the subspace.

In quantum theory, a sequence of events, such as A and then B, denoted AB, is represented by the sequence of projectors $P_B P_A$. If the projectors commute, $P_A P_B = P_B P_A$, then the product of the two projectors is a projector corresponding to the subspace $A \cap B$, that is, $P_B P_A = P(A \cap B)$; and the events A and B are said to be *compatible*. However, if the two projectors do not commute, $P_B P_A \neq P_A P_B$, then neither their product is a projector, and the events are *incompatible*.

Quantum theory uses a unit length state vector, denoted $|\psi\rangle$, to assign probabilities to events as follows[4]:

$$p(A) = \|P_A\,|\psi\rangle\|^2 . \tag{2}$$

Quantum probabilities satisfy an additive measure: $p(A) \geq 0$, $p(H) = 1$, and if $P_A P_B = 0$, then $p(A \vee B) = p(A) + p(B)$. In fact, Equation (2) is the unique way to assign probabilities to subspaces that form an additive measure for dimensions greater than 2 [18].

In quantum theory, the definition of a conditional probability is

$$p(B|A) = \frac{\|P_B P_A\,|\psi\rangle\|^2}{p(A)},$$

and so the probability of the sequence AB equals $p(AB) = p(A) \cdot p(B|A) = \|P_B P_A\,|\psi\rangle\|^2$. Extensions to sequences with more than two events follows the same principles: The probability of the sequence $(AB)C$ equals $\|P_C\,(P_B P_A)\,|\psi\rangle\|^2$ for quantum theory.

5.1. Building projectors

This section describes a general way to construct the projectors for events in the Hilbert space, and to formally describe the conditions that produce incompatibility. In the following, $|V\rangle$ denotes a vector in the Hilbert space, $\langle V|W\rangle$ denotes an inner product, $|V\rangle\langle V|$ denotes an outer product, and $P^\dagger$ denotes a Hermitian transpose.

In general, a projector, denoted P, operating in an N-dimensional Hilbert space is defined by the two properties $P = P^\dagger = P^2$. By the first property, P is Hermitian, and so it can be decomposed into N orthonormal eigenvectors; by the second property, P has only two eigenvalues, which are simply $(0, 1)$. Define $\left|V_j\right\rangle$, $j = 1, \cdots, N$ as the set of N orthonormal eigenvectors of P. The projector P can be expressed in terms of the eigenvectors as follows

[3] See [10], [20], [28] for tutorials.

[4] A more general approach uses what is called a density operator rather than a pure state vector, but to keep ideas simple, we use the latter.

$$P = \sum_j \lambda_j \left|V_j\right\rangle\left\langle V_j\right|, \tag{3}$$

where the outer product, $\left|V_j\right\rangle\left\langle V_j\right|$, is the projector that projects into the ray spanned by eigenvector $\left|V_j\right\rangle$, and $\lambda_j = 1$ if $\left|V_j\right\rangle$ corresponds to an eigenvalue of 1, and $\lambda_j = 0$ if $\left|V_j\right\rangle$ corresponds to an eigenvalue of 0. These N eigenvectors form an orthonormal basis that spans the Hilbert space. Every vector, such as $|\psi\rangle \in H$ can be expressed as a linear combination of these basis (eigen) vectors

$$|\phi\rangle = \sum_j^N \phi_j \cdot \left|V_j\right\rangle \tag{4}$$

If two projectors, P_A, P_B share all of the same eigenvectors, then they commute. In other words, two events A, B are compatible if they are described in terms of the same basis. If the two projectors do not share all of the same eigenvectors, then they do not commute, and the events A, B are described by two different bases. They are incompatible, and must be evaluated sequentially, because one needs to change from one basis to evaluate the first event, to another basis to evaluate the second event, making them incompatible.

Define $\left|V_j\right\rangle, j = 1, \cdots, N$ as the basis used to describe event A, and define $\left|W_j\right\rangle, j = 1, \cdots, N$ as the basis used to describe event B. We can change from one basis to another by a unitary transformation (a "rotation" in Hilbert space)

$$\left|W_j\right\rangle = U\left|V_j\right\rangle, j = 1, \cdots, N, \tag{5}$$

where U is defined by $U^\dagger U = I$, that is, U is an isometric transformation that preserves inner products. Therefore, the projector for event B can be re-expressed in terms of the event A basis $\left|V_j\right\rangle, j = 1, \cdots, N$ as follows

$$\begin{aligned} P_B &= \sum_j^N \lambda_j \left|W_j\right\rangle\left\langle W_j\right| \\ &= U\left(\sum \lambda_j \left|V_j\right\rangle\left\langle V_j\right|\right)U^\dagger. \end{aligned} \tag{6}$$

According to Equation (5), the unitary transformation U represents the transitions from state $|W_i\rangle$ to state $\left|V_j\right\rangle$ by the inner product $\left\langle V_j|W_i\right\rangle$.

So far, we have presented a general method for building the projectors by defining a basis for the vector space and by transforming from one basis to another using unitary transformation. Then the next question is how to build the unitary transformation? In general, any unitary transformation can be built from a Hermitian operator H as follows:

$$U = exp(-i \cdot H). \tag{7}$$

The right hand side is the exponential function of the Hermitian operator H.

In summary, the HSM program selects a Hermitian operator H for Equation (7), and then uses the Hermitian operator to build the unitary operator U which provides the relation between projectors P_A and P_B for incompatible events. The beauty of using a vector space is that it provides an infinite number of ways to generate incompatible variables by unitary "rotation," and yet remain within the same N-dimensional space. This is how an HSM model maintains a low dimensional representation even when there are a large number of variables.

5.2. Building the Hilbert space

This section describes how we construct a Hilbert space to represent the p variables. This construction depends on the compatibility relations between the variables. For this section, we need to use the Kronecker (tensor) product between two matrices, denoted as $P \otimes Q$.

To begin building the Hilbert space, suppose we measure a single variable, say Y_1, that can produce n_1 values corresponding to the mutually exclusive and exhaustive set of events $(Y_1 = y_i), i = 1, \cdots n_1$. To represent these events in a Hilbert space, we partition the space into n_1 orthogonal subspaces. Each subspace, such as $(Y_1 = y_i)$, corresponds to a projector $P(Y_1 = y_i)$. The projectors for all of the events are pairwise orthogonal, $P(Y_1 = y_i)P(Y_1 = y_j) = \mathbf{0}$, and complete, $\sum_i P(Y_1 = y_i) = \mathbf{I}$ (where $\mathbf{I}$ is the identity that projects onto the entire Hilbert space). These n_1 events are all compatible, and the projectors are all commutative, because they are all orthogonal to each other. Each projector generates $N_1 \geq n_1$ eigenvectors, and the projectors all share the same eigenvectors, but with different eigenvalues. These N_1 eigenvectors provide the basis for spanning a N_1-dimensional Hilbert space, H_{N_1}.

Nothing requires the number of eigenvectors of a projector to be equal to the number of observed values of the variable used by the researcher. For example, it would be an arbitrary decision by a researcher to use only 2 values (yes, no) for attribute A, but 3 values for attribute B, and 4 values for attribute C. The researcher could have used 5 rating values for all three variables. So it would not make sense to assume that variable A is represented in a 2 dimensional space, when the

choice of a binary measure was an arbitrary choice by the researcher. Given that the number of values of the variable is an arbitrary choice of the researcher, there is no reason to think that this measurement will be a complete measurement, i.e., a measurement that is represented as a set of mutually exclusive and exhaustive one dimensional projections [24]. Of course, the simplest assumption is that a projector is one dimensional, but this assumption often fails, and we shall see that it does in this example.

Continuing with the case of a single variable represented by the Hilbert space H_{N_1}, we can express each vector $|\phi\rangle$ in terms of its coordinates with respect to the eigenvectors of $P(Y_1 = y_i)$ by using Equation (4). Using this basis, the coordinate representation of each projector, say $P(Y_1 = y_i)$ is simply an $N_1 \times N_1$ diagonal matrix, $M_1(i)$ with ones located in the rows corresponding to basis vectors that have an eigenvalue of one associated with the projector $P(Y_1 = y_i)$, and zeros otherwise. The coordinate representation of $|\psi\rangle$ with respect to this basis is a $N_1 \times 1$ column matrix ψ with coordinate ψ_i in row i, which satisfies $\psi^\dagger \psi = 1$. Then the probability distribution over the values of Y_1 for $i = 1, \cdots, n_1$ is given by

$$\|P(Y_1 = y_i) \cdot |\psi\rangle\|^2 = \|M_1(i) \cdot \psi\|^2 = |\psi_i|^2 . \tag{8}$$

There is little difference between classical and quantum probability at this point.

Next suppose we measure two variables, Y_1 with n_1 values and Y_2 with n_2 values, with $n_1 \geq n_2$. If these two variables are compatible, then the joint event $(Y_1 = y_i \cap Y_2 = y_j)$ is well defined for all pairs of values. Therefore the Hilbert space is partitioned into $n_1 \cdot n_2$ orthogonal subspaces. Each subspace corresponds to a projector $P(Y_2 = y_j)P(Y_1 = y_i) = P(Y_1 = y_i)P(Y_2 = y_j) = P(Y_1 = y_i \cap Y_2 = y_j)$. These projectors are pairwise orthogonal and complete, and every pair of projectors is commutative. Each projector shares $(N_1 \cdot N_2) \geq (n_1 \cdot n_2)$ eigenvectors, but with different eigenvalues, to span a Hilbert space $H_{N_1 \cdot N_2}$. Using this basis, the projector $P(Y_1 = y_i)$ is represented by the Kronecker product $M_1(i) \otimes I_{N_2}$, where I_{N_2} is an $N_2 \times N_2$ identity matrix. The projector $P(Y_2 = y_j)$ is represented by the matrix Kronecker product $I_{N_1} \otimes M_2(j)$. Then $P(Y_2 = y_j)P(Y_1 = y_i) = P(Y_1 = y_i \cap Y_2 = y_j)$ is represented by the product $\left(M_1(i) \otimes I_{N_2}\right) \cdot \left(I_{N_1} \otimes M_2(j)\right) = M_1(i) \otimes M_2(j)$, which is simply a diagonal matrix with ones located in the rows corresponding to $(Y_1 = y_i \cap Y_2 = y_j)$ and zeros otherwise. The coordinate representation of $|\psi\rangle$ with respect to this basis is a $(N_1 \cdot N_2) \times 1$ column matrix, $\left(\psi, \psi^\dagger \psi = 1\right)$, with coordinate ψ_{ij} in row $n_2 \cdot (i-1) + j$. Then the joint probability for a pair of values equals

$$\left\| P(Y_2 = y_j)P(Y_1 = y_i)\,|\psi\rangle \right\|^2 = \|M_1(i) \otimes M_2(j) \cdot \psi\|^2 = \left|\psi_{ij}\right|^2 . \tag{9}$$

There is still little difference between the classical and quantum theories at this point. Adding variables increases the dimensionality of the space, just like it does with a Bayesian model.

Now suppose that variables Y_1 (with n_1 values) and Y_2 (with $n_2 \leq n_1$ values) are incompatible. In this case, we cannot define the joint occurrence of two events $(Y_1 = y_i \cap Y_2 = y_j)$, and we can only represent a sequence of two single events, e.g., $(Y_1 = y_i)$ and then $\left(Y_2 = y_j\right)$ by the sequence of projectors $P(Y_2 = y_j)P(Y_1 = y_i)$. As before, we define $P(Y_1 = y_i)$ as the projector for the event $(Y_1 = y_i)$, and likewise, we define $P(Y_2 = y_j)$ as projector for the event $\left(Y_2 = y_j\right)$. Both projectors are represented with a Hilbert space, H_{N_1}, of dimension $N_1 \geq n_1$. We can choose to express each vector $|\phi\rangle$ in terms of the coordinates with respect to the eigenvectors of $P(Y_1 = y_i)$ by using Equation (4). Using this basis, the coordinate representation of projector $P(Y_1 = y_i)$ is simply an $N_1 \times N_1$ diagonal matrix, $M_1(i)$ with ones located in the rows corresponding to basis vectors that have an eigenvalue of one associated with this projector, and zeros otherwise. Using Equation (5), the projector $P(Y_2 = y_j)$ can be expressed in terms of the $P(Y_1 = y_i)$ basis by a unitary matrix, U. Then the matrix representation of $P(Y_2 = y_j)$ is $\left(U \cdot M_1(j) \cdot U^\dagger\right)$. Finally, the coordinate representation of the state vector $|\psi\rangle$ with respect to the Y_1 basis is a $N_1 \times 1$ column matrix ψ. The probability of the sequence of events $(Y_1 = y_i)$ and then $\left(Y_2 = y_j\right)$ equals

$$\left\| P(Y_2 = y_j)P(Y_1 = y_i)\,|\psi\rangle \right\|^2 = \left\| \left(U \cdot M_1(j) \cdot U^\dagger\right) \cdot M_1(i) \cdot \psi \right\|^2 . \tag{10}$$

This is where a key difference between the classical and quantum theories occurs. Note that, unlike a Bayesian model, adding variable Y_2 does not increase the dimensionality of the space.

Finally suppose that we measure three variables, Y_1 with n_1 values, Y_2 with n_2 values, and Y_3 with n_3 values. Suppose Y_1 is compatible with Y_2 and Y_2 is compatible with Y_3 but Y_1 is incompatible with Y_3. In this case, we can partition the Hilbert space using projectors $P(Y_1 = y_i \cap Y_2 = y_j), i = 1, \cdots n_1, j = 1, \cdots n_2$, which are pairwise orthogonal and complete, and every pair of these projectors is commutative. Using the eigenvectors of these projectors as the basis, the projector $P(Y_1 = y_i)$ is represented by the Kronecker product $M_1(i) \otimes I_{N_2}$, and the projector $P(Y_2 = y_j)$ is represented by the Kronecker product $I_{N_1} \otimes M_2(j)$. Using a unitary transformation, U, the matrix representation of the projector $P(Y_3 = y_k)$ is given $\left(U \cdot M_1(k) \cdot U^\dagger\right) \otimes I_{N_2}$. Then, the probability of the two compatible events $(Y_1 = y_i)$ and $\left(Y_2 = y_j\right)$ equals

$$\left\| P(Y_2 = y_j)P(Y_1 = y_i)\,|\psi\rangle \right\|^2 = \|M_1(i) \otimes M_2(j) \cdot \psi\|^2 . \tag{11}$$

The probability of the two compatible events $(Y_2 = y_i)$ and $\left(Y_3 = y_j\right)$ equals

$$\| P(Y_3 = y_k) P(Y_2 = y_i) \, |\psi\rangle \|^2 = \left\| \left(U \cdot M_1(k) \cdot U^{\dagger} \right) \otimes M_2(i) \cdot \psi \right\|^2, \tag{12}$$

and the probability of the sequence of two incompatible events $(Y_1 = y_i)$ and then $(Y_3 = y_k)$ equals

$$\| P(Y_3 = y_k) P(Y_1 = y_i) \, |\psi\rangle \|^2 = \left\| \left(U \cdot M_1(k) \cdot U^{\dagger} \cdot M_1(i) \right) \otimes I_{N_2} \cdot \psi \right\|^2. \tag{13}$$

5.3. Constraints on HSM models

The methods described above generalize in a fairly straightforward manner for more variables. Note that when variables are compatible, quantum probability theory works like classical probability theory, and the Hilbert space dimensionality increases exponentially as the number of compatible variables increases. However, when variables are incompatible, it is unlike classical probability theory, and the Hilbert space dimensionality remains constant as the number of incompatible variables increases.

The use of incompatible variables comes with some costs. HSM theory is not completely general, and it must satisfy several of its own consistency constraints. For example, consider once again a collection of four 2 by 2 tables (AC, AD, BC, BD) generated from 4 binary variables (A, B, C, D). Although quantum probabilities do not have to satisfy the famous Bell inequalities, they do have to satisfy another inequality called the Cirelson inequality. See [12] for general methods for determining constraints on correlations by quantum models. Even more demands are placed on the transition probabilities generated by the unitary transformation used to form incompatible variables. The transition probability matrix is produced by squaring the magnitude of the entries of the unitary matrix, and this transition matrix must be doubly stochastic, and it must satisfy the law of reciprocity (see [24], Ch. 2).

6. Application to the artificial data set

An HSM is constructed from the following 6 steps.

- Step 1 determines the compatibility and incompatibility relations among the variables.
- Step 2 determines the dimension of the Hilbert space based on assumed compatibility relations.
- Step 3 defines the initial state given the dimension of the Hilbert space.
- Step 4 defines the projectors for the variables using unitary transformations to change the basis
- Step 5 computes the choice probabilities given the initial state and the projectors
- Step 6 interprets the model parameters to provide an deeper understanding of the complex tables

We apply the above principles to the artificial data set using six steps[5]:

Step 1. Determine compatibility of variables. Psychologically, this step determines whether two variables can be measured simultaneously (compatible) or they have to be measured sequentially (incompatible). Based on the order effects observed in Table 1, we infer that the pair of variables A, C were incompatible, as well as the pair B, C. The design did not include manipulations of order to test compatibility between variables A, B. In this case, another way to empirically test compatibility is to compare model fits that make compatibility vs. incompatibility assumptions about these variables. Here for the purpose of illustration, we assumed that they were compatible.

Step 2. Define the Hilbert space. Assuming that A, B are compatible means that we can define all of the events obtained from all of the combination of values of these two variables: $(A = a_i \cap B = b_j)$, for $(a_i = a_1, a_2)$ and $(b_j = b_1, b_2, b_3)$. However, we cannot define combinations for more variables because of the incompatibilities with C. The simplest model is a model that assumes that each event $(A = a_i \cap B = b_j)$ is represented by only 1 dimension, which produces a total of 6 dimensions. Therefore, the minimum size of the Hilbert space was set to 6 dimensions, and we started with this minimum.

Step 3. Define the initial state. We chose a basis that provided the most meaningful parameters for the initial state. For this application, we chose to use the basis defined by the combination of variables A and B. Using this basis, the matrix representation of the basis vector $\left| A = a_i \cap B = b_j \right\rangle$ is simply a 6×1 column matrix with zeros everywhere except for a 1 in row $3 \cdot (i-1) + j$. The initial state $|\psi\rangle$ is represented by

$$|\psi\rangle = \sum_{i,j} \psi_{ij} \cdot \left| A = a_i \cap B = b_j \right\rangle. \tag{14}$$

[5] The Matlab code used to perform these computations is available at http://mypage.iu.edu/~jbusemey/quantum/HilbertSpaceModelPrograms.htm.

The six coefficients in Equation (14) form a 6×1 column matrix

$$\psi = \begin{bmatrix} \psi_{11} \\ \psi_{12} \\ \psi_{13} \\ \psi_{21} \\ \psi_{22} \\ \psi_{23} \end{bmatrix}.$$

For example, $|\psi_{13}|^2$ equals the probability of yes to Adeptness and a high rating to Brilliance. The parameters in ψ are estimated from the data under the constraint that $\psi^\dagger \psi = 1$. In general, the 6 coefficients can be complex valued, and so each coefficient contributes a magnitude and a phase. However, the magnitudes must satisfy the unit length constraint that $\psi^\dagger \psi = 1$. Also, one phase for one coefficient can be set to an arbitrary value without changing the final choice probabilities. Therefore, only $6 \times 2 - 2 = 10$ free parameters are required for the initial state. In many applications, the initial state can be restricted to real values, in which case there are only $6 - 1 = 5$ parameters.

The initial state parameters tell us what the initial state of the psychological system (e.g., initial belief or attitude towards attributes A and B) is before any measurement is taken on the system, and can be used to compute the probability of certain response to an attribute when it is measured alone. That is, we can estimate more "context free" responses from the respondents–free from influences from measurement effects from the other attributes–even though we didn't collect such actual empirical data.

Step 4. Define the projectors. We can define the matrix for the projectors for the events $(A = a_i)$ as follows. Define $A_1 = diag\begin{bmatrix} 1 & 0 \end{bmatrix}$, $A_2 = diag\begin{bmatrix} 0 & 1 \end{bmatrix}$, and $I_3 = diag\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$. The 6×6 matrix representation of the projector $P(A = a_i)$ is the Kronecker product $(A_i \otimes I_3)$, which picks out the coordinates in ψ that are associated with the answer a_i to attribute A. We can define the matrix for the projectors for the events $(B = b_j)$ as follows. Define $B_1 = diag\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$, $B_2 = diag\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$, $B_3 = diag\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$, and $I_2 = diag\begin{bmatrix} 1 & 1 \end{bmatrix}$. The 6×6 matrix representation of the projector $P(B = b_j)$ is the Kronecker product $\left(I_2 \otimes B_j\right)$, which picks out the coordinates in ψ that are associated with a rating b_j given to attribute B.

The 6×6 matrix representation of the projector $P(C = c_j)$ is defined as follows. The columns of U provide the matrix representations of the basis vectors $|C = l\rangle$ for the C basis. First we need to construct 6 indicator matrices, denoted C_l, $l = 1, \ldots, 6$, using the joint events from the AB basis as follows: $C_l = P(A = a_i) \cdot P(B = b_j)$, where $l = 3 \cdot (i - 1) + j$, and so each C_l is just a diagonal matrix with zeros in all rows except for row j which contains the value 1. Then we rotate the basis from the AB basis to the C basis using a unitary matrix: $U \cdot C_j \cdot U^\dagger$, which simply picks the j-th column of U.

Although there are 6 basis vectors in the C basis, we need to map these into the observed 4 confidence rating values. So, we need to form 4 projectors from the 6 matrices, C_l, to represent the 4 possible confidence ratings. This choice can affect the results. Here we define the projector for the first confidence level as $P(C = c_1) = U(C_1 + C_2)U^\dagger$; we define the projector for the second confidence level as $P(C = c_2) = UC_3U^\dagger$; we define the projector for the third confidence level as $P(C = c_3) = UC_4U^\dagger$; and finally, we define the projector for the final confidence level as $P(C = c_4) = U(C_5 + C_6)U^\dagger$. In practice, this assignment can be made by fitting different combinations of coarse measurements and choosing the best fit.

Step 4 requires defining the unitary matrix U that rotates to the new C basis. We define a unitary matrix U_A that operates on the two dimensional A subspace, and another unitary matrix U_B that operates on the three dimensional B subspace. The complete unitary matrix is formed by the tensor product: $U = U_A \otimes U_B$ which forms the 6×6 unitary matrix that operates on the full six dimensional Hilbert space.

The 2×2 matrix representation, U_A was determined from Equation (7) by selecting a 2×2 Hermitian matrix, H_A; the 3×3 matrix representation, U_B was determined from Equation (7) by selecting a 3×3 Hermitian matrix. The parameters of each of these Hermitian matrices were estimated from the data. In general, the 2×2 Hermitian matrix has four coefficients, two real diagonal values and one complex off diagonal. However, one diagonal entry can be arbitrarily fixed, and so only $4 - 1 = 3$ parameters are required to produce a total of 6 parameters. The 3×3 Hermitian matrix has 9 coefficients, three real diagonal values and three complex off diagonals. However, one diagonal entry can be arbitrarily fixed, and so only $9 - 1 = 8$ parameters are required to fit the data. In many applications, the Hermitian matrix can be restricted to real values, in which case H_A has only $1 + 1 = 2$ parameters, and H_B has only $2 + 3 = 5$ parameters.

The unitary parameters determine the rotation from the basis of one variable to the basis of another variable. Psychologically, they tell us the relationship between the variables or attributes being examined, and can reveal the similarity between these variables, independent of the initial state (i.e., step 3) of the person. In addition, based on the unitary matrices, we can compute the probabilities the transition probabilities between basis vectors.

Step 5. Compute choice probabilities for each response sequence. The choice probabilities for each sequence were computed by the product of projectors corresponding to the sequence. For example, the probability of $(C = c_3, A = a_1)$ for the AC table equals

$$\| P(A = a_1) P(C = c_3) |\psi\rangle \|^2 = \left\| (A_1 \otimes I_3) \cdot \left(U \cdot C_3 \cdot U^\dagger \right) \cdot \psi \right\|^2, \tag{15}$$

the probability of $(A = a_1, C = c_3)$ for the CA table equals

Table 3
Transition matrices between basis vector for pairs of incompatible attributes.

	$\|c_1\rangle$	$\|c_2\rangle$	$\|c_3\rangle$	$\|c_4\rangle$	$\|c_5\rangle$	$\|c_6\rangle$
$\|a_1b_1\rangle$	0.0242					
$\|a_1b_2\rangle$	0.0531	0.0002				
$\|a_1b_3\rangle$	0.0291	0.0531	0.0242			
$\|a_2b_1\rangle$	0.2035	0.4458	0.2442	0.0242		
$\|a_2b_2\rangle$	0.4458	0.0019	0.4458	0.0531	0.0002	
$\|a_2b_3\rangle$	0.2442	0.4458	0.2035	0.0291	0.0531	0.0242

$$\|P(C=c_3)P(A=a_1)\,|\psi\rangle\|^2 = \left\|\left(U\cdot C_3\cdot U^{\dagger}\right)\cdot(A_1\otimes I_3)\cdot\psi\right\|^2, \tag{16}$$

Step 6. Estimate parameters by minimizing the KL divergence between model and observed probabilities. We fit the HSM model under the constraint that the initial state was real and the Hermitian matrices for the unitary transformation were real. This model required fitting $5+2+5=12$ parameters. The 5 model parameters representing ψ, along with the 7 parameters representing U_A and U_B, are presented in the Appendix. The HSM model has a total of 12 free parameters, which is far fewer than the 23 required by the 3-way joint probability model. Nevertheless, the HSM model almost perfectly fits all the relative frequencies in Table 1. The KL divergence for the quantum model fit is $D < 10^{-8}$. Recall that the data were artificially generated for illustration, and so the results are not to be taken seriously. However, they help to show the application of a HSM model.

The unitary matrices, U, can be used to describe the transitions between incompatible measurements. The squared magnitudes of the entries in the unitary matrix describes the probability of transiting from a basis vector representing a row attribute (e.g., $|a_ib_j\rangle$) to a basis vector representing a column attribute (e.g., $|c_l\rangle$). Table 3 presents the transition probabilities for the two incompatible pairs, AB, and C. For example, the probability of transiting from $|a_2b_1\rangle$ to $|c_2\rangle$ equals 0.4458, which is one of the highest transition rates.

The transition matrices produced by unitary matrices are always symmetric (see chapter 2 in [10] for a discussion). This is because each entry in the unitary matrix contains the inner product between vectors from different bases, and the squared magnitude is the same in both directions. This can provide a test of the quantum model, but only assuming the most restrictive assumption that the projectors are uni-dimensional (see, e.g., [6]). Unfortunately, what has been often overlooked is the fact that when the projectors are multi-dimensional, the conditional probabilities of events are not necessarily symmetric (see, e.g., [27]). Very often in applications, the projectors must be multi-dimensional, as they have to be in this case, and this is not an arbitrary assumption. If events are represented by multi-dimensional projectors, then the conditionals can be asymmetric. For example, the projector for the Confidence rating ($C=c_1$) is 2-dimensional and the projector for the low judgment of Brilliance ($B=b_1$) is 2 dimensional, and the quantum model quantum model predicts $p(C=c_1|B=b_1)=p(B=b_1,C=c_1)/p(B=b_1)=0.3875$, but $p(B=b_1|C=c_1)=p(C=c_1,B=b_1)/p(C=c_1)=0.2485$, which exhibits asymmetry.

7. Application to real data

7.1. Test of joint probability model

Now we apply HSM theory to a real data set. A total of 184 participants (70% female students from the Ohio State University) observed pictures of female avatars and made binary (Yes, No) judgments about Attractiveness (A), Intelligence (I), Sociability (S), and Honesty (H) of each avatar. For each presentation of an avatar, the participant was asked to judge a pair of attributes (e.g., judge Attractiveness and Intelligence of this avatar) by choosing one of four pairs of answers (YY, YN, NY, NN, where for example YN indicates Yes to the first attribute and No to the second). The choice from the 4 answers was entered into a 2×2 contingency table. A total of 6 tables $\{AI, AS, AH, IS, IH, SH\}$ were formed from all combinations of pairs generated from the 4 attributes. The avatars were sampled from two different types: Sexualized versus non-sexualized avatars. Altogether, each participant judged both types of avatars under all 6 contexts, and each avatar for each contex was replicated on 4 different presentations. The avatars and contexts were randomized organized across presentations with a different random ordering for each participant.

The aggregate results are presented in Table 4. The results are presented separately for each stimulus type. For example, when the non-Sexualized avatar was presented, the relative frequency of Y to Attractive and N to Intelligent was 0.0312, and the corresponding result for the sexualized avatar was 0.4226. Altogether, each 2×2 joint frequency table for each type of avatar (sexualized or not) and for each of the 6 contexts contains a total $(184\cdot 4=736)$ observations.

We conducted a statistical chi-square test of a 4-way joint probability model composed from four binary random variables, A, I, S, H used in the study. The test was based a comparison of the model predictions with the data in Table 4.[6] The

[6] This test should be evaluated with some caution because we are ignoring individual differences for this analysis. In future work, we plan to use a hierarchical Bayesian model that introduces assumptions about the distribution of individual differences and priors on these hyper parameters.

Table 4
Observed results for 6 contexts and 2 types of stimuli. Each row is a 2×2 table containing the relative frequency of answers. Each column indicates a pair of answers. Table on the left is the non-sexualized type of avatar, and table on the right is the sexualized type.

	Non-sexualized				Sexualized			
	YY	YN	NY	NN	YY	YN	NY	NN
AI	0.6644	0.0312	0.2323	0.0720	0.2582	0.4226	0.0394	0.2799
AS	0.6399	0.0639	0.1644	0.1318	0.4905	0.1698	0.0924	0.2473
AH	0.6454	0.0394	0.2527	0.0625	0.2568	0.4049	0.0503	0.2880
IS	0.7622	0.1481	0.0204	0.0693	0.2704	0.0326	0.3668	0.3302
IH	0.8723	0.0394	0.0299	0.0584	0.1753	0.1481	0.1087	0.5679
SH	0.7677	0.0611	0.1155	0.0557	0.2554	0.3546	0.0516	0.3383

joint probability model states that the 6 rows of 2×2 tables for each stimulus type are produced by a joint distribution, $\pi(A = w \cap I = x \cap S = y \cap H = z)$ where $w = 0, 1$, $x = 0, 1$, $y = 0, 1$ and $z = 0, 1$, that has $16 - 1 = 15$ free parameters per stimulus type or 30 parameters altogether. A completely unconstrained saturated model requires 3 parameters for each 2×2 table, producing a total of 18 parameters per stimulus type or 36 parameters altogether. Using maximum likelihood estimation, we computed the G^2_{sat} and G^2_{joint} for each table and computed the difference $G^2_{Diff} = G^2_{sat} - G^2_{joint} = 20.6521$, which produced a statistically significant G^2 difference: based on 6 degrees using the $p = .0021$. This provides some evidence that the 4-way joint probability model systematically deviates from the observed results for the two tables. It is noteworthy that violations of the joint probability model occur in this collection of contexts even though there are no changes in order of measurement. In this case, the violations must arise from either violations of the marginals or correlations between variables.

7.2. HSM model of real data

Each pair of attributes forms a measurement context, and so there were 6 different measurement contexts. Each event in the model represents an answer to a question about an attribute within a context. We applied a simple HSM model to the data as follows

First, we used a 4-dimensional space to represent the events and the state vector. After selecting a basis for an event, the state vector can be represented by 4 coordinates, where each coordinate contains the amplitude assigned to one of the pair of answers

$$\psi = \begin{bmatrix} \psi_{YY} \\ \psi_{YN} \\ \psi_{NY} \\ \psi_{NN} \end{bmatrix}.$$

We set the initial state to a uniform distribution $\psi_{ij} = 1/2$.

Second, for a context, (A, B) where A represents the first attribute and B represents the second, we constructed two projectors: one projector, $P_A(Y)$ for answering Yes to the first attribute, and another projector $P_B(Y)$ for answering Yes to the second attribute. These two projectors were defined as follows:

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$M_Y = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

$$P_A(Y) = U_A \cdot (M_Y \otimes I_2) \cdot U_A^{\dagger},$$

$$P_B(Y) = U_B \cdot (I_2 \otimes M_Y) \cdot U_B^{\dagger}.$$

The projector for the No answers were then defined by $P_A(N) = I_4 - P_A(Y)$ and $P_B(N) = I_4 - P_B(Y)$, where $I_4 = I_2 \otimes I_2$.

Third, the unitary matrix for each attribute was computed from a Hermitian matrix using the matrix exponential function

$$U_A = exp(-i \cdot (\pi/2) \cdot H_A),$$

$$U_B = exp(-i \cdot (\pi/2) \cdot H_B).$$

Fourth, the Hamiltonian matrix for the first attribute within a context (A, B) was defined as follows:

$$H_A = V_A \otimes I_2$$

Table 5
Parameters estimated for each type of avatar.

	μ_A	μ_I	μ_S	μ_H
Non sexualized	0.4065	1.1792	0.7627	1.0852
Sexualized	0.3482	−0.4430	0.3339	−0.5470

	γ_{AI}	γ_{AS}	γ_{AH}	γ_{IS}	γ_{IH}	γ_{SH}
Non sexualized	0.3163	0.4465	0.2110	−1.6456	−1.9623	0.3717
Sexualized	0.1924	−1.5400	0.2267	0.2798	−2.0993	0.2769

Table 6
Probabilities predicted by the HSM model. The predictions are organized in the same manner as the observed data table.

	Non-sexualized				Sexualized			
	YY	YN	NY	NN	YY	YN	NY	NN
AI	0.6584	0.0344	0.2373	0.0700	0.2751	0.3923	0.0645	0.2682
AS	0.6224	0.0704	0.1789	0.1284	0.5008	0.1666	0.1003	0.2323
AH	0.6544	0.0384	0.2470	0.0603	0.2593	0.4081	0.0497	0.2829
IS	0.7619	0.1517	0.0274	0.0590	0.2326	0.0594	0.3467	0.3613
IH	0.8261	0.0875	0.0406	0.0458	0.1569	0.1350	0.1112	0.5969
SH	0.7514	0.0699	0.1327	0.0460	0.2776	0.3834	0.0457	0.2933

$$V_A = \frac{1}{\sqrt{1+\mu_A^2}} \begin{bmatrix} \mu_A & 1 \\ 1 & -\mu_A \end{bmatrix}.$$

This "rotates" the amplitudes for the first attribute toward or away from the Yes answer depending on the parameter μ_A. The unitary matrix for the second attribute within the same context (A, B) was initially defined in a similar manner as

$$H_B = (I_2 \otimes V_B),$$
$$V_B = \frac{1}{\sqrt{1+\mu_B^2}} \begin{bmatrix} \mu_B & 1 \\ 1 & -\mu_B \end{bmatrix},$$

which "rotates" the amplitudes for the second attribute toward or away from the Yes answer depending on the parameter μ_B. However, we also added another component, H_C, to the Hamiltonian for the second attribute

$$H_B = (I_2 \otimes V_B) - \gamma_{A,B} \cdot H_C,$$
$$H_C = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

The H_C component in the last equation was a key parameter used to entangle the state. When γ is positive, it creates an entangled state with high amplitudes on ψ_{YY} and ψ_{NN}, and when γ is negative it creates high amplitudes on ψ_{YN} and ψ_{NY}. An entangled state is required to produce dependencies in the predicted joint probability tables. The parameter, $\gamma_{A,B}$, was permitted to vary across each of the 6 contexts to allow different dependencies across the 6 tables.

Finally, we used the quantum rules to compute the choice probabilities. For example, the probability of YN when asked about attributes AI was computed by

$$p(YN|AI) = \|P_I(N) \cdot P_A(Y) \cdot \psi\|^2$$

and the probability of NY when asked about the attributes IS was computed by

$$p(NY|SI) = \|P_I(Y) \cdot P_S(N) \cdot \psi\|^2.$$

This HSM model entails estimating 10 parameters for each type of stimulus: one parameter μ_X associated with each of the 4 variables, and one parameter $\gamma_{A,B}$ associated with each of the 6 contexts. The parameters were estimated from the data separately for each type of stimulus using maximum likelihood methods. The parameter estimates are shown in Table 5. Positive values of μ tend to increase the marginal probabilities of a variable, and negative values tend to decrease the marginals. Positive values of γ tend to entangle answers that agree (YY, NN), and negative values tend to entangle answers that disagree (YN, NY).

The predictions of the HSM model are shown in Table 6. Comparing Tables 4 and 6, one can see that the model makes fairly accurate predictions for the contexts from both stimuli.

7.3. Comparison of HSM and Joint probability models

We compared the fits of the 4-way Joint probability model to the HSM model by using the Bayesian information criterion (BIC). The HSM model is not nested within (not a special case of) the Joint probability model, and so we used the BIC criterion to compare models, which can be used to compare non-nested models. The Joint probability model has $15 \cdot 2 = 30$ free parameters for both types of stimuli, which is 10 more than the $10 \cdot 2 = 20$ parameters used by the HSM model. Usually a model with more parameters fits better only because of the advantage produced by extra parameters. The BIC provides a balance between accuracy and parsimony. It compares the G^2 for each model with a penalty for extra parameters: BIC(HSM vs. Joint) $= \left(G^2_{HSM} - G^2_{Joint}\right) - p \cdot ln(N)$, where $p =$ the difference in number of parameters (10 in this case) and $N =$ total number of observations ($N = 8832$ in this case). If BIC(HSM vs. Joint) is negative, then the simpler HSM model is preferred, and if it is positive, then the more complex Joint model is preferred. Recall that for the Joint probability model, $G^2_{sat} - G^2_{joint} = 20.6521$; the corresponding value for the HSM model is $G^2_{sat} - G^2_{HSM} = 82.79$ and the difference equals $G^2_{HSM} - G^2_{joint} = 62.14$, which is far below the penalty $10 \cdot ln(8832) = 90.86$ for the extra 10 parameters used by the Joint model. Therefore the BIC(HSM vs. Joint) is negative and clearly favors the HSM model over the Joint probability model.

8. Summary and alternative probabilistic models

8.1. Summary

HSM models provide a simple and low dimensional method for data fusion when researchers collect multiple contingency tables formed from measurement of subsets of p variables. The power of HSM models to perform data fusion is produced by the inclusion of incompatible variables. When variables are compatible, quantum probability theory works like classical probability theory, and the Hilbert space dimensionality increases exponentially as the number of compatible variables increases. However, when variables are incompatible, it is unlike classical probability theory, and the Hilbert space dimensionality remains constant as the number of incompatible variables increases. This reduction in dimensionality is achieved by using rotation of the basis vectors to generate new incompatible variables. In this way, the inclusion of additional variables does not increase the dimension of the vector space that is used to represent the data.

This article describes the general methods that we use to build HSM models. We illustrated these principles using an artificial data set in which 3 variables were used to generate 6 different contingency tables. An HSM model was built that perfectly reproduced the 6 tables, even though no 3-way joint probability could fit the data. We also applied these principles to a real data set consisting of 6 different 2×2 tables constructed from pairs of 4 binary variables. The joint probability model based on the 4 observed variables produced statistically significant deviations from the 6 observed data tables. A simpler HSM model produced a better account than the joint probability model of the 6 observed data tables based on a BIC model comparison index.

HSM models provide new contributions to the current set of probabilistic and statistical tools for contingency table analysis. Loglinear/categorical data models only apply to a single table containing all p variables, whereas the HSM models can be applied to multiple tables containing different subsets of the p variables. Bayesian network models can also be applied to collections of tables; however, they assume the existence of a complete p-way joint distribution, and it is often the case that no complete p-way joint distribution exists. HSM models can be applied to collections of tables even when no p-way joint distribution exists to reproduce the collection.

8.2. Alternative probabilistic models

In this article, we have provided empirical evidence that HSM theory provides a useful way to model collections of contingency tables formed from subsets of p variables that cannot be reproduced by a complete p-way joint distribution. However, this is not the only way, and there are other probabilistic models that could be considered. One way that we mentioned earlier is to postulate a higher dimensional joint distribution with additional random variables [16] and use probabilistic data base programming methods [8] to form the joint distribution. A second way is to relax some of the axioms of Kolmogorov theory [23] and form a generalized probability theory. A third way is to expand the field over the vector space from complex to hyperbolic [20]. A fourth way is to propose a general probabilistic mechanism capable of generating either classical or quantum probabilities as a special case [3]. However, at this point in time, the advantage of HSM over these alternatives is that HSM is based on a coherent set of axioms (supporting Gleason's theorem), HSM provides a specific and well defined algorithm, the models derived from HSM use a reasonably small number of parameters, and HSM models can be rigorously tested with empirical data. At this time, the alternatives mentioned above either lack an axiomatic foundation, or they are too general to specify and apply to real data, or they do not permit rigorous empirical tests.

Table 7
Table of probabilities used to prove the Leggett–Garg inequality.

A	B	C	$p(A \neq B)$	$p(B \neq C)$	$p(A \neq C)$
1	1	1			
1	1	2		X	X
1	2	1	X	X	
1	2	2	X		X
2	1	1	X		X
2	1	2	X	X	
2	2	1		X	X
2	2	2			

Appendix A

A.1. Leggett–Garg inequality

The three columns on the left, labeled A, B, C represent three binary valued random variables of a hypothesized 3-way joint probability model. The eight rows represent the eight possible joint events produced by the eight combinations of values of these three binary random variables. The last three columns on the right are used to compute three probabilities used in the Leggett–Garg inequality. According to the 3-way joint probability model, the probability $p(A \neq B)$ is computed by summing the probabilities of the events in the rows containing an X in the fourth column; the probability $p(B \neq C)$ is computed by summing the probabilities of the events in the rows containing an X in the fifth column; the probability $p(A \neq C)$ is computed by summing the probabilities of the events in the rows containing an X in the sixth column. The Leggett–Garg inequality follows from the fact that columns 4 and 5 contain column 6. (See Table 7.)

A.2. Quantum model parameters

The following parameters were used to almost perfectly fit all the data shown in Table 1.

$$\psi = \begin{bmatrix} 0.2685 \\ -0.7600 \\ 0.0883 \\ 0.3514 \\ -0.1935 \\ -0.4260 \end{bmatrix}.$$

$$H_A = \begin{bmatrix} 0.5236 & 1.5708 \\ 1.5708 & -0.5236 \end{bmatrix}.$$

$$H_B = \begin{bmatrix} -0.1047 & 3.2987 & 0 \\ 3.2987 & 0 & 3.2987 \\ 0 & 3.2987 & 0.1047 \end{bmatrix}$$

References

[1] S. Abramsky, Relational databases and Bells theorem, in: V. Tannen, L. Wong, L. Libkin, W. Fan, T. Wang-Chiew, M. Fourman (Eds.), In Search of Elegance in the Theory and Practice of Computation, 2013, pp. 13–35.
[2] D. Aerts, L. Gabora, S. Sozzo, Concepts and their dynamics: a quantum – theoretic modeling of human thought, Top. Cognit. Sci. 5 (2013) 737–773.
[3] D. Aerts, M. Sassoli de Bianchi, The unreasonable success of quantum probability I: quantum measurements as uniform fluctuations, J. Math. Psych. 67 (2015) 737–773.
[4] D. Aerts, S. Aerts, Applications of quantum statistics in psychological studies of decision processes, Found. Sci. 1 (1) (1995) 85–97.
[5] A. Agresti, M. Katera, Categorical Data Analysis, Springer, 2011.
[6] Thomas Boyer-Kassem, Sébastien Duchêne, Eric Guerci, Testing quantum-like models of judgment for question order effect, Math. Social Sci. 80 (2016) 33–46.
[7] P.D. Bruza, K. Kitto, B.J. Ramm, L. Sitbon, A probabilistic framework for analysing the compositionality of conceptual combinations, J. Math. Psych. 67 (2015) 26–38.
[8] P.D. Bruza, Syntax and operational semantics of a probabilistic programming language with scopes, J. Math. Psych. 74 (2016) 46–57.
[9] P.D. Bruza, V. Chang, Perceptions of document relevance, Front. Psychol. 5 (2014) 612.
[10] J.R. Busemeyer, P.D. Bruza, Quantum Models of Cognition and Decision, Cambridge University Press, 2012.
[11] J.R. Busemeyer, E.M. Pothos, R. Franco, J.S. Trueblood, A quantum theoretical explanation for probability judgment errors, Psychol. Rev. 118 (2) (2011) 193–218.
[12] A. Cabello, S. Severini, A. Winter, A graph theoretic approach to quantum correlations, Phys. Rev. Lett. 112 (2014) 040401, pp. 1–5.
[13] A. Darwiche, Modeling and Reasoning with Bayesian Networks, Cambridge University Press, 2009.
[14] G.M. DiNunzio, P. Bruza, L. Sitbon, Interference in text categorization experiments, in: H. Atmanspacher, E. Haven, K. Kitto, D. Raine (Eds.), Quantum Interaction 2013, in: Lecture Notes in Computer Science, vol. 8369, Springer, Berlin, Heidelberg, 2014.

[15] E. Dzhafarov, J.V. Kujala, Selectivity in probabilistic causality: where psychology runs into quantum physics, J. Math. Psych. 56 (2012) 54–63.
[16] E. Dzhafarov, J.V. Kujala, Context-content systems of random variables: the Contextuality-by-Default theory, J. Math. Psych. 74 (2016) 11–33.
[17] A. Fine, Joint distributions, quantum correlations and commuting observables, J. Math. Phys. 23 (7) (1982) 1306–1310.
[18] A.M. Gleason, Measures on the closed subspaces of a Hilbert space, J. Math. Mech. 6 (1957) 885–893.
[19] S.P. Gudder, Quantum Probability, Academic Press, 1988.
[20] A.Y. Khrennikov, Ubiquitous Quantum Structure: From Psychology to Finance, Springer, 2010.
[21] A.J. Leggett, A. Garg, Quantum mechanics versus macroscopic realism: is the flux there when nobody looks, Phys. Rev. Lett. 54 (1985) 857–860.
[22] M. Melucci, Introduction to Information Retrieval and Quantum Mechanics, Springer, 2015.
[23] L. Narens, Probabilistic Lattices: With Applications to Psychology, World Scientific, London, 2014.
[24] A. Peres, Quantum Theory: Concepts and Methods, Kluwer, 1995.
[25] I. Pitowsky, Quantum Probability, Quantum Logic, Lecture Notes in Physics, vol. 321, 1989.
[26] I. Pitowsky George, Booles conditions of possible experience and the quantum puzzle, British J. Philos. Sci. 45 (1) (1994) 95–125.
[27] E.M. Pothos, J.R. Busemeyer, J.S. Trueblood, A quantum geometric model of similarity, Psychol. Rev. 120 (3) (2013) 679–696.
[28] C.J. van Rijsbergen, The Geometry of Information Retrieval, Cambridge University Press, 2004.
[29] N.N. Vorob'ev, Consistent families of measures and their extensions, Theory Probab. Appl. 7 (2) (1962) 147–163.
[30] D. Widdows, Geometry and Meaning, CSLI Publications, 2004.
[31] B. Wang, P. Zhang, J. Li, D. Song, Y. Hou, Z. Shang, Exploration of quantum interference in document relevance judgement discrepancy, Entropy 18 (4) (2016) 144.
[32] Forrest W. Young, Multidimensional Scaling: History, Theory, and Applications, Psychology Press, 2013.