RC25586 (WAT1601-034) January 13, 2016
Computer Science

IBM Research Report

Supervised Item Response Models for Informative Prediction

Tsuyoshi Idé, Amit Dhurandhar
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

Research Division
Almaden – Austin – Beijing – Brazil – Cambridge – Dublin – Haifa – India – Kenya – Melbourne – T.J. Watson – Tokyo – Zurich

Supervised Item Response Models for Informative Prediction

Tsuyoshi Idé · Amit Dhurandhar

Received: date / Accepted: date

Abstract Supporting human decision making is a major goal of data mining. The more decision making is critical, the more interpretability is required in the predictive model. This paper proposes a new framework to build a fully interpretable predictive model for questionnaire data, while maintaining high prediction accuracy with regards to the final outcome. Such a model has applications in project risk assessment, in health care, in social studies and presumably in any real world application that relies on questionnaire data for informative and accurate prediction.

Our framework is inspired by models in Item Response Theory (IRT), which were originally developed in psychometrics with applications to standardized tests such as SAT. We extend these models, which are essentially unsupervised, to the supervised setting. For model estimation, we introduce a new iterative algorithm by combining Gauss-Hermite quadrature with an expectation-maximization algorithm. The learned probabilistic model is linked to the metric learning framework for informative and accurate prediction. The model is validated by three real-world data sets: Two are from information technology project failure prediction and the other is an international social study about people's happiness.

To the best of our knowledge, this is the first work that leverages the IRT framework to provide informative and accurate prediction on ordinal questionnaire data.

Keywords psychometrics · questionnaire data · item response theory · metric learning

IBM Research, T. J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10592, USA
E-mail: {tide,adhuran}@us.ibm.com

1 Introduction

Supporting human decision-making is one of the most important goals of data mining. In recommender systems for example, certain actions are recommended. Depending on the domain these actions could vary from being buying decisions [17] for shoppers to being important business decisions that are recommended to executives or managers based on historical data. Irrespective of the domain the final recommended action presented by itself is rarely sufficient to convince the decision maker of its "plausibility". Ordinarily, additional supporting evidence needs to be provided in support of the recommendation. Hence, a lower likelihood recommendation from a learning model may be a better choice if it can be clearly justified.

The plausibility, or more precisely *interpretability*, is in fact a critical success factor in many business applications. For example, imagine that you are a manager of a company and you are making decisions of lay-offs based on a scorecard for individual employees, which includes a number of qualitative questions such as "Has he/she made good enough contributions to teamwork?" You have a database of the historical records of best practices, which contains a collection of pairs $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is a filled scorecard as represented by a binary or graded vector (see Fig. 1) and y is the binary indicator to represent termination ($y = +1$) or not ($y = -1$). Although the problem can be viewed formally as simple binary classification to predict y given $\boldsymbol{x}$, the nature of the problem is glaringly different in at least two aspects.

First, the input data are typically ordinal. In general it is not valid to naively use standard probabilistic assumptions such as the Gaussian-distributed noise for ordinal variables. Second, the model must have a high degree of interpretability. For the year-end assessment meeting, you as a manager will want to be very clear on the rationale of the suggested outcome from at least three perspectives:

1. Comparison to other employees: What is the difference between lay-off and no lay-off groups?
2. Comparison between different questions in the scorecard: What kind of weighting is used for individual questions? How can we justify the weighting?
3. Comparison between different question choices: Some questions may be easily achieved and others may not. How can we quantify the heterogeneity?

In other words, we need to ensure at least three different interpretabilities: instance-wise, dimension-wise, and ordinal-grade-wise interpretabilities. As long as a predictive model is used to support critical decision-making, the model must be fully interpretable in this sense. This is especially true in applications such as healthcare, project audit, and company reputation analysis, as illustrated in Fig. 1.

In our preliminary work [13], we introduced a new framework to build a fully interpretable predictive model for questionnaire data. The method is inspired by the item response theory (IRT) [30], which was originally developed

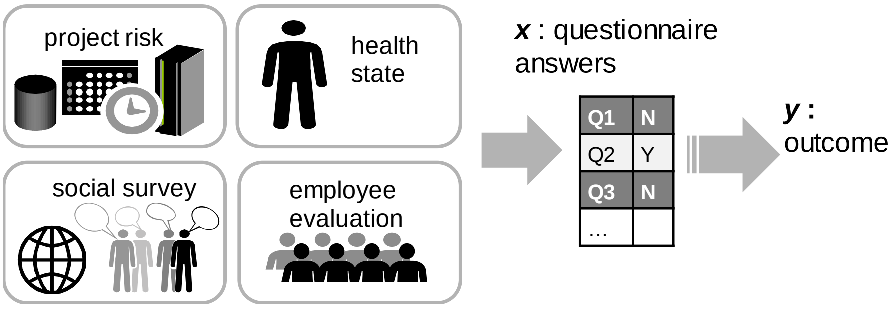


Fig. 1 Questionnaire-based diagnosis is ubiquitous. In many real applications, black-box predictive models are not practical. Full interpretability is often required at all instance-, dimension-, and ordinal grade-levels (see the text).

in psychometrics with applications to standardized academic tests such as SAT [25]. IRT provides a natural way to ensure dimension-wise (*i.e.* between individual question items) and grade-wise (*i.e.* between individual question choices of each question item) interpretabilities. However, the original IRT is unsupervised and cannot incorporate the outcome variable y. Also, there is no direct method to evaluate the *informativeness of each question items* in terms of predictability of the outcome variable.

To address these limitations, in [13], we proposed a supervised version of IRT combined with a framework of distance metric learning approach [32]. Although it was a fully interpretable prediction model, one issue was that it has to solve a high dimensional optimization problem that originates from a non-conjugate Bayesian formulation. In this paper, we introduce an efficient iterative algorithm which we call Gauss-Hermite EM (expectation-maximization) algorithm to address this issue. The algorithm is applicable any models whenever the prior distribution is Gaussian and the observe model satisfies a conditional independence.

We also expand the experimental study in [13] by adding analysis on publicly available social study data, World Values Survey. To the best of our knowledge, our framework is the first approach to fully interpretable informative prediction for questionnaire data.

2 Related Work

There are four categories of previous work that is relevant to the present study.

The *first* category is obviously standard classification methods. As mentioned in Introduction, our task in Fig. 1 mirrors the task of binary classification. However, standard binary classifiers are not very useful in terms of analyzing the quality of the questions. For instance, support vector machines (SVMs) [10] or regularized logistic regression (LR) [10] may be accurate in predicting the outcome variable y, but the information they provide about the questions is in the form of unbounded signed weights, which can be difficult to interpret. On the other hand, in decision trees [10] it can be challenging to evaluate the importance of a variable as it might occur at different levels in different parts of the tree. Ensemble methods [10] may help compute variable importance, but they end up with losing instance-level information. We thus want a systematic way of evaluating the quality of the questions that is more informative and easier to digest, while maintaining predictability.

The *second* category is about modeling of human cognition. As Fig. 1 illustrates, we are interested in modeling the generative process of questionnaire answers. It amounts to modeling the decision-making process of humans, which is one of the typical examples of dynamics of complex systems. To model complex systems, deep learning has become a more and more practical tool in recent years, and dramatic successes in image and speech recognition [20, 12] are well-known. Also, if a fair amount of text data is given, sentiment analysis [24] for text documents provides a powerful method to understand the human cognition. Although we share a part of research motivation of modeling complex dynamics of human decision making, we pursue a completely opposite direction from those approaches that are mostly black-box: we request that our model should achieve interpretability at all different levels of instance, question item, and answer choices within each question. While some recent work addresses personal cognitive process in decision making [4, 23], which may be relevant to questionnaire analysis, our work differs in that we are interested in handling questionnaire data as the primary data source.

In psychometrics, on the other hand, quantitatively modeling human cognition bias has been an important topic for years. For a useful review, the reader may refer to Baker and Kim [1]. In the machine learning community, Lan *et al.* [19] recently extended the original IRT to incorporate factor analysis in an unsupervised setting. As explained in a later section, the original motivation of IRT was to quantitatively estimate the ability of examinees and the difficulty of individual question items in academic tests. If we are allowed to rephrase the ability as, *e.g.*, the medical risk in the case of diagnosis questionnaire, this is exactly what we want. Unlike the traditional setting of academic tests, however, we assume additional data of the final outcome such as occurrence of serious side effects, project failure, or termination of employment. In the context of the SAT test, in addition to the SAT scores themselves, we were as if given information that the individual examinees had succeeded in their life later on. Using the information on the final outcome, we should be able to evaluate the true informativeness of the test items. To the best of the authors' knowledge, little attention has been paid to such a problem setting in psychometrics and data mining.

The *third* category is the study on ordinal data. Modeling ordinal data has been one of the major research topics in statistics and statistical data mining. Well-known examples include ordinal regression [21] and learning to rank [5]. Rank-constrained nonlinear discriminant analysis [27] is another recent instance. These assume that the response variable is ordinal. In our case, however, we are interested in handling ordinal predictor variables instead. From this perspective, the most relevant work will be Koren and Sill [16, 15], which addresses the ordinal nature of human rating in collaborative filtering, although their problem clearly differs from ours.

The *fourth* category is metric learning. Since the advent of the seminal paper of Xing [32], metric learning has been one of the most active areas in the data mining communities [29, 9, 18, 8]. For a recent review, the reader may refer to Bellet et al. [2]. By definition of the task, metric learning (often implicitly) assumes that the samples distribute in a *metric space*, just like dots placed on a piece of paper, whose coordinates and the distance are well-defined and ready to be calculated using *e.g.* the Euclidean distance. However, it is clear that a special attention is required when handling ordinal variables since the ordinal scale distinguishes only relative goodness or badness. For example, an ordinal variable may ask about the goodness of personal relationship with your boss, and another ordinal variable may be the level of satisfaction to your family life. It is clear that relative comparison between two different ordinal variables is not trivial at all [26]. In spite of the popularity of metric learning research, only limited attention has been paid to metric learning for ordinal variables.

Recently, Ouyang and Gray proposed a rank-constrained approach to kernel learning. Also, Terada and Luxburg [28] proposed a method for order-preserving embedding. These works are somewhat relevant to ours, but their setting differs from ours in that they assume that the ordinal relationship between the instances is given; in our case, what is given is the final rating of projects, which is by no means sufficient to define the total order. Another relevant piece of work is ground metric learning [6], which handles the ordinal nature of the variables by considering histograms. However, their problem setting differs from us since we need to make a prediction for a single project, rather than for a histogram as a collection of projects.

Our framework for informative prediction attempts to achieve practical interpretability and predictability by combining a psychometric model with metric learning. Instead of solving semi-definite programming as large-margin nearest neighbors [29], we take the path of Kostinger et al. [18], which first proposed an "optimization-free" method to metric learning and achieved a state-of-the art performance in the image classification task. As explained in a later section, we introduce an information-theoretic view to their approach.

3 Problem Definition and Motivation

We first formally describe our problem setting. We then provide real world examples of where we encounter this setting thus showcasing its wide presence.

3.1 Problem Statement

Imagine we have a questionnaire containing M question items and N subjects (patients, projects, employees, etc.) take the questionnaire to answer the questions. Our training data set can be formally represented as

$$\mathcal{D} = \{(\boldsymbol{x}^{(n)}, y^{(n)}) \mid n = 1, 2, \ldots, N\}, \tag{1}$$

where $\boldsymbol{x}^{(n)}$ is an M-dimensional vector representing the questionnaire answers of the n-th subject, and $y^{(n)}$ is the class label for the n-th subject. Our goal is to build a fully interpretable model to predict y given a new $\boldsymbol{x}$, and to evaluate the informativeness of the individual question items, through which a qualitative feel to the user in terms of the predictability of the final outcome is provided.

Here we say that a model is *fully interpretable* if a predictive model allows

- quantitative comparison between subjects in terms of their importance,
- quantitative comparison between question items in terms of their importance,
- quantitative comparison between answer choices in terms of probability of choosing each option,

while maintaining a comparable accuracy to other less interpretable methods.

3.2 Application Domains

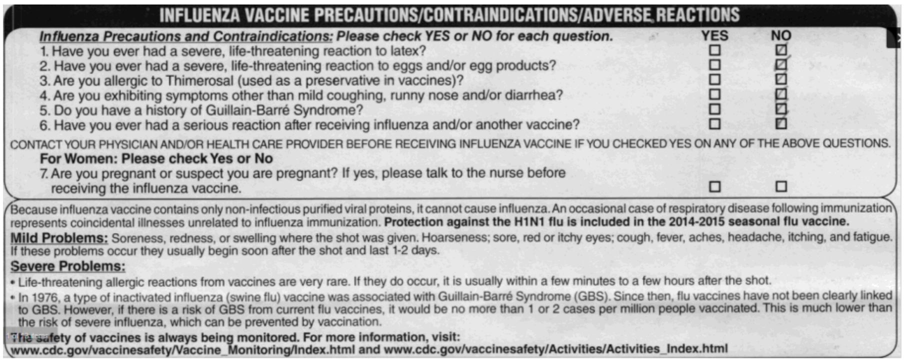
We now instantiate the above problem definition to different domains.

Project Risk Assessment: In our motivating example, $\boldsymbol{x}^{(n)}$ is an M-dimensional vector representing the questionnaire answers, and $y^{(n)}$ is the health rating indicating the troubled or healthy status. Each of the dimensions of $\boldsymbol{x}^{(n)}$ takes an integer value in the predefined risk levels, while $y^{(n)}$ takes either of $+1$ or -1 (troubled or non-troubled). It is known that $\{\boldsymbol{x}^{(n)}\}$ has human bias which modulates the true risk levels of projects in some nonlinear fashion.

Health Care: Before administering any treatment doctors require patients to fill up (yes/no) questionnaires indicative of their condition. An example, flu shot questionnaire from last year is depicted in figure 2. Here the $\boldsymbol{x}^{(n)}$ are binary yes/no questions and $y^{(n)}$ indicates if the treatment was successful or not, *i.e.*, in our example if the person got flu or not.

Employee Evaluations: Many companies during yearly appraisals fill out questionnaires based on various criteria for each employee. For instance the $\boldsymbol{x}^{(n)}$ would be, did the person have at least an δ amount of business impact, how many projects did he see to completion, did he lead any projects, etc. Based on the response they would decide the $y^{(n)}$ which would be to either layoff or keep the employee.

Social Surveys: Policy makers often use social surveys to better understand the motivations behind social events and to improve governmental policies. World Values Survey (WVS) [31] is such a social survey. The main topic of it is the level of satisfaction with the society and the communities. It is evident



INFLUENZA VACCINE PRECAUTIONS/CONTRAINDICATIONS/ADVERSE REACTIONS

Influenza Precautions and Contraindications: ***Please check YES or NO for each question.***	**YES**	**NO**
1. Have you ever had a severe, life-threatening reaction to latex?	□	☑
2. Have you ever had a severe, life-threatening reaction to eggs and/or egg products?	□	☑
3. Are you allergic to Thimerosal (used as a preservative in vaccines)?	□	☑
4. Are you exhibiting symptoms other than mild coughing, runny nose and/or diarrhea?	□	☑
5. Do you have a history of Guillain-Barré Syndrome?	□	☑
6. Have you ever had a serious reaction after receiving influenza and/or another vaccine?	□	☑

CONTACT YOUR PHYSICIAN AND/OR HEALTH CARE PROVIDER BEFORE RECEIVING INFLUENZA VACCINE IF YOU CHECKED YES ON ANY OF THE ABOVE QUESTIONS.

For Women: Please check Yes or No

7. Are you pregnant or suspect you are pregnant? If yes, please talk to the nurse before receiving the influenza vaccine.	□	□

Because influenza vaccine contains only non-infectious purified viral proteins, it cannot cause influenza. An occasional case of respiratory disease following immunization represents coincidental illnesses unrelated to influenza immunization. **Protection against the H1N1 flu is included in the 2014-2015 seasonal flu vaccine.**

Mild Problems: Soreness, redness, or swelling where the shot was given. Hoarseness; sore, red or itchy eyes; cough, fever, aches, headache, itching, and fatigue. If these problems occur they usually begin soon after the shot and last 1-2 days.

Severe Problems:

- Life-threatening allergic reactions from vaccines are very rare. If they do occur, it is usually within a few minutes to a few hours after the shot.
- In 1976, a type of inactivated influenza (swine flu) vaccine was associated with Guillain-Barré Syndrome (GBS). Since then, flu vaccines have not been clearly linked to GBS. However, if there is a risk of GBS from current flu vaccines, it would be no more than 1 or 2 cases per million people vaccinated. This is much lower than the risk of severe influenza, which can be prevented by vaccination.

The safety of vaccines is always being monitored. For more information, visit: www.cdc.gov/vaccinesafety/Vaccine_Monitoring/Index.html and www.cdc.gov/vaccinesafety/Activities/Activities_Index.html

Fig. 2 Above is a brief snapshot of a flu shot vaccination questionnaire.

that interpretability and accountability are extremely important in this case. For more details, see Sec. 7.

4 Supervised Item Response Model

This section introduces a probabilistic framework towards informative prediction to meet the requirements explained in Sec. 3.1.

4.1 Probabilistic Model for Answer Choice

Imagine that we are given a questionnaire having M questions. To be specific, let us use project failure prediction as a running example, and assume that each question has only two options of 1 (at-risk) or 0 (no risk) about a particular aspect of project risks, such as the tightness of development schedule (see Sec. 7 for details).

One natural tendency of human reviewers is that they are cautious about choosing the at-risk option until they observe evident indications of future project failures, possibly worrying about making a mistake. Once they find something convincing enough from their perspective, they start being critical about the project. If we denote the true project failure tendency by $\theta \in \mathbb{R}$, human reviewers would have the following biases:

- They underestimate risks when θ is lower.
- They overestimate risks when θ is higher.
- They may even use random guess when they are not familiar with details of the project.

To capture these natural psychological traits, we use the following model for choosing 1 of the l-th question:

$$P(\theta, a_l, b_l, c_l) \equiv c_l + \frac{1 - c_l}{1 + \mathrm{e}^{-a_l(\theta - b_l)}}. \tag{2}$$

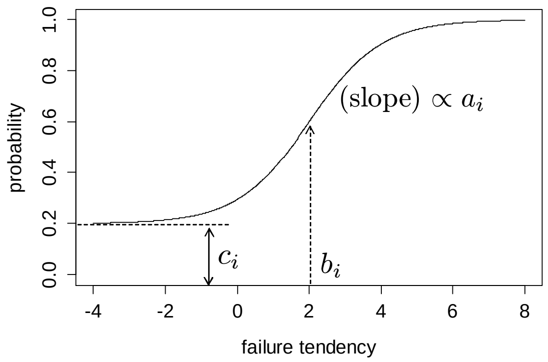


Fig. 3 Item Characteristic Curve for the example of project risk management.

In psychometrics, this model is called three-parameter item response model [1], and Fig. 3 is called the item characteristic curve (ICC). The parameters a_l, b_l, c_l are called the discrimination, difficulty, and guessing parameters, respectively.

The *discrimination* parameter controls the slope of the ICC. Since we are interested in rating projects using θ, we should design the question so $a_l \sim 1$ so the ICC looks like a linear function w.r.t. θ as much as possible. If $|a_l|$ is large and the ICC looks like a step-function, an infinitesimal change in θ may end up with an infinite change in P, which is unintuitive. The *difficulty* parameter plays the role of threshold. Roughly speaking, when θ exceeds b_l, the reviewer becomes pessimistic about the future of projects, while he/she remains optimistic below b_l. Finally, the *guessing* parameter literally represents the probability of choosing the at-risk option even when $\theta \to -\infty$, and thus corresponds to selection just by guess.

By stacking M ICCs, we have the probability of an answer pattern $\boldsymbol{x}$, given θ and model parameters as

$$p(\boldsymbol{x}|\theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \prod_{l=1}^{M} P(\theta, a_l, b_l, c_l)^{\delta(x_l,1)}[1 - P(\theta, a_l, b_l, c_l)]^{\delta(x_l,0)} \tag{3}$$

where δ represents Kronecker's delta and $\boldsymbol{a} \equiv (a_1, \ldots, a_M)^\top$. The other parameter vectors $\boldsymbol{b}, \boldsymbol{c}$ are also defined similarly.

4.2 Prior Distribution to Latent State Variable θ

Although the original IRT is fully unsupervised, we extend the model to include the outcome variable y. In the case of project risk analysis for example, we assume that troubled projects ($y = +1$) have generally higher failure risk tendencies. Thus it is quite natural to assume a prior distribution conditioned

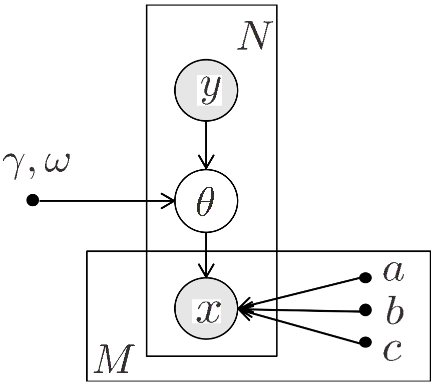


Fig. 4 Graphical model of supervised IRT model.

on y:

$$f(\theta|y) = \begin{cases} \frac{\gamma}{\sqrt{2\pi}} \exp\left(-\frac{\gamma}{2}\theta^2\right) & \text{for} \quad y = -1, \\ \frac{\gamma}{\sqrt{2\pi}} \exp\left(-\frac{\gamma}{2}(\theta - \omega)^2\right) & \text{for} \quad y = +1, \end{cases} \tag{4}$$

where γ and ω are hyper-parameters to be learned from the training data.

By marginalizing out the latent variable θ, the log marginalized likelihood function of the model is written as follows:

$$L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) = \sum_{n=1}^{N} \ln \left[\pi(y^{(n)}) p(\boldsymbol{x}^{(n)}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, y^{(n)})\right] \tag{5}$$

$$p(\boldsymbol{x}^{(n)}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, y^{(n)}) \equiv \int_{-\infty}^{\infty} \mathrm{d}\theta^{(n)}\ p(\boldsymbol{x}^{(n)}|\theta^{(n)}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})\ f(\theta^{(n)}|y^{(n)}) \tag{6}$$

where $\mathcal{D}$ symbolically represents the dependency on the training data, and $y^{(n)}$ is the variable representing the n-th project health indicator. The distribution $\pi(y^{(n)})$ is the prior distribution for $y^{(n)}$, which is assumed to be the same as the ratio of each of the labels to N. In Fig. 4, we summarize the probabilistic model using the plate notation of probabilistic graphical models. We call this model the *supervised IRT* (sIRT) model.

4.3 Maximum likelihood equation and issues in optimization

The model parameters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ can be found by maximizing the likelihood:

$$(\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*) = \arg\max_{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}} L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) \tag{7}$$

$$\text{subject to} \quad 0 \leq c_l \leq 1 \quad (l = 1, \ldots, M).$$

The box constraints on the guessing parameter $\{c_l\}$ can be handled by the method of barrier function [22].

Although Eq. (7) is a formally well-defined optimization problem, direct numerical maximization of $\tilde{L}$ w.r.t. $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ is not straightforward. For a real example, in World Values Survey (see Sec. 7.4), the total number of questions is 430, and the problem (7) is 1 290-dimensional non-convex optimization problem, which is hard to solve with general purpose numerical solvers.

For maximum likelihood (ML) problems with latent variables, one popular approach is to derive an EM (expectation-maximization) algorithm. If the model belongs to the exponential family, it is well-known that ML estimation based on the EM approach is reduced to an iterative estimation of moments, as is in the Gaussian mixture case [3]. Unfortunately, the IRT model $P(\theta, a_l, b_l, c_l)$ does not have a conjugate prior, and the standard EM approach does not lead to a simple iterative algorithm. In particular, it is evident that

$$\ln P(\theta, a_l, b_l, c_l) = \ln \left\{ c_l + \frac{1 - c_l}{1 + \mathrm{e}^{-a_l(\theta - b_l)}} \right\}$$

is not represented by elementary polynomials, and simple application of Jensen's inequality to Eq. (6) does not lead to tractable iterative formula.

To address this issue, in the next section, we derive a new efficient iterative algorithm that reduces the intractable high-dimensional optimization problem to a collection of simple 3-dimensional optimization problems.

5 Gauss-Hermite EM algorithm for efficient optimization

This section derives an efficient algorithm to solve the high-dimensional optimization problem Eq. (7). Although we will focus on the particular model of IRT, our framework is applicable whenever (1) the prior is Gaussian and (2) the parameters are conditionally independent given the latent parameter.

5.1 Reducing integral to summation via Gauss-Hermite quadrature

As mentioned in the previous section, the sIRT model does not allow to derive a simple moment matching solution as is the case of exponential family. The first step towards a tractable EM algorithm is to represent the integration w.r.t. the latent variable θ in Eq. (6) as

$$\int_{-\infty}^{\infty} \mathrm{d}\theta \; f(\theta | y^{(n)}) \; p(\boldsymbol{x}^{(n)} | \theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \approx \sum_{i=1}^{W} w_i \; p\left(\boldsymbol{x}^{(n)} \middle| \phi_i^{(n)} \; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \right), \qquad (8)$$

where W is an input parameter typically chosen as ~ 20. Since $f(\theta|y^{(n)})$ is Gaussian in our case, the weights $\{w_i\}$ and the nodes $\{\phi_i^{(n)}\}$ can be optimally determined via Gauss-Hermite quadrature *independently of* $\mathcal{D}$:

$$\phi_i^{(n)} = \sqrt{\frac{2}{\gamma}}\varphi_i + \omega\delta(y^{(n)}, 1) \tag{9}$$

$$w_i = \frac{2^{W-1}W!}{[WH_{W-1}(\varphi_i)]^2}, \tag{10}$$

where $\{\varphi_i\}$ is the zeros of the W-th order Hermite polynomial $H_W(\varphi)$. These are determined so that the expansion is exact for polynomials up to the order of $2W-1$. For derivation, see Chap. 8 of [11]. We treat W and thus $\{w_i, \phi_i\}$ as given constants hereafter.

5.2 Lower bound of log marginalized likelihood

Using Eq. (8), the log marginalized likelihood is written as

$$L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) = \sum_{n=1}^{N} \ln \pi(y^{(n)}) + L_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) \tag{11}$$

$$L_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) \equiv \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{W} w_i p(\boldsymbol{x}^{(n)}|\phi_i^{(n)}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \right\} \tag{12}$$

Notice that L_1 looks like a mixture model. Encouraged by this observation, we derive the lower bound (LB) by applying Jensen's inequality [3]

$$L_1 \geq L_1^{\mathrm{LB}} \equiv \sum_{n=1}^{N}\sum_{i=1}^{W} Q_{i,n} \ln \left\{ \frac{w_i}{Q_{i,n}} p(\boldsymbol{x}^{(n)}|\phi_i^{(n)}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \right\}, \tag{13}$$

where $\{Q_{i,n}\}$ is determined so L_1^{LB} is maximized under the constraint of $\sum_{i=1}^{W} Q_{i,n} = 1$. By solving

$$0 = \frac{\partial}{\partial Q_{i,n}} \left\{ L_1^{\mathrm{LB}} + \sum_{n=1}^{N} \lambda_n \sum_{i=1}^{W} Q_{i,n} \right\}, \tag{14}$$

we easily get the solution for $Q_{i,n}$ as

$$Q_{i,n} = \frac{w_i \exp\left(\sum_{l=1}^{M} J_{i,n}^{l}\right)}{\sum_{j=1}^{M} w_j \exp\left(\sum_{l=1}^{M} J_{j,n}^{l}\right)}, \tag{15}$$

where $J_{j,n}^{l}$ is defined as

$$J_{i,n}^{l} \equiv \begin{cases} \ln P(\phi_i^{(n)}, a_l, b_l, c_l) & \text{if} \quad x_l^{(n)} = 1, \\ \ln[1 - P(\phi_i^{(n)}, a_l, b_l, c_l)] & \text{if} \quad x_l^{(n)} = 0. \end{cases} \tag{16}$$

Using this notation, the lower bound L_1^{LB} is written as

$$L_1^{\mathrm{LB}} = \sum_{l=1}^{M} \mathrm{Tr}\left(\mathsf{Q}^\top \mathsf{J}^l\right) + \sum_{n=1}^{N}\sum_{i=1}^{W} Q_{i,n} \ln \frac{w_i}{Q_{i,n}}, \tag{17}$$

where Q and J^l are $W \times N$ matrices whose (i,n) elements are $Q_{i,n}$ and $J^l_{i,n}$, respectively.

5.3 Gauss-Hermite EM algorithm

Given Q, L_1^{LB} in Eq. (17) separates the parameters into individual l's. To optimize w.r.t. (a_l, b_l, c_l), we care only about the J^l part:

$$(a_l^*, b_l^*, c_l^*) = \arg\max_{a_l, b_l, c_l} \mathrm{Tr}\left(\mathsf{Q}^\top \mathsf{J}^l\right) \quad \text{subject to} \quad 0 \leq c_l \leq 1. \tag{18}$$

This is just a 3-parameters constrained optimization problem. Although we cannot obtain an analytic solution for this, we can use any numerical solver such as `constrOptim` in R.

To get the final solution, Eqs. (15) and (18) are solved alternatingly until convergence. Algorithm 1 summarizes the procedure to fit the sIRT model, which we call the *Gauss-Hermite EM (GHEM) algorithm*.

Algorithm 1 Gauss-Hermite EM algorithm for the supervised IRT model.

```
Input: Training data $\mathcal{D}$. Hyper-parameters $\omega, \gamma$. Initial values of the model
parameters $\boldsymbol{a}^0, \boldsymbol{b}^0, \boldsymbol{c}^0$.
Output: The maximizer $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$.
repeat
  Given the current $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$, compute the $\mathsf{Q}$ matrix using Eq. (15).
  Given the current $\mathsf{Q}$, solve Eq. (18) independently for $l = 1, \ldots, M$.
until Convergence
Return $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$.
```

In addition to the IRT parameters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$, we may output the Q matrix for descriptive analysis. The n-th column of this matrix gives the posterior distribution of the latent parameter θ:

$$p(\theta = \phi_i^{(n)} \mid \boldsymbol{x}^{(n)}, y^{(n)}, \mathcal{D}) = Q_{i,n}. \tag{19}$$

In problem project failure prediction, this tells what the latent failure tendency looks like for the n-th project.

In Algorithm 1, the hyper-parameters γ, ω are determined so the cross-validated performance is maximized. For details, see Sec. 6.4.

6 Making prediction

We have discussed how to estimate the model parameters so far. We now consider how to make a prediction of y given a new questionnaire answer $\boldsymbol{x}$. We first derive a classifier based only on the probabilistic model. We then discuss a metric learning method for k-NN classification to further enhance the interpretability. One practically important outcome of the metric learning algorithm is the informativeness score for each of the questions, which gives valuable insights for questionnaire design.

6.1 Neyman-Pearson decision rule

Once we obtain the optimized parameters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ that maximizes the marginal likelihood, the predictive distribution of $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x}|\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*, y) = \int_{-\infty}^{\infty} \mathrm{d}\theta \, p(\boldsymbol{x}|\theta, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*) \, f(\theta|y). \tag{20}$$

This is a conditional distribution given y. To make a prediction for y, we define a classification score by

$$s(\boldsymbol{x}) = \ln \frac{p(\boldsymbol{x}|\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*, y = +1)}{p(\boldsymbol{x}|\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*, y = -1)} = \ln \frac{\sum_{i=1}^{W} w_i p(\boldsymbol{x}|\phi_i^{+1}, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}{\sum_{i=1}^{W} w_i p(\boldsymbol{x}|\phi_i^{-1}, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}, \tag{21}$$

where

$$\phi_i^{+1} \equiv \sqrt{\frac{2}{\gamma}} \varphi_i + \omega, \quad \phi_i^{-1} \equiv \sqrt{\frac{2}{\gamma}} \varphi_i. \tag{22}$$

If $s(\boldsymbol{x})$ is greater than a threshold, s_{th}, a newly observed $\boldsymbol{x}$ is classified into the class of $y = +1$.

The score $s(\boldsymbol{x})$ is based on the distribution of $\boldsymbol{x}$, not y, unlike the standard Bayes' decision rule [10]. Proposition 1 tells the optimality of this rule, which we call the *Neyman-Pearson (NP) decision rule.*

Proposition 1 *(Neyman-Pearson decision rule) Consider a decision rule of classification for an instance* $\boldsymbol{x}$

$$\begin{aligned} y &= +1, \quad \text{if} \quad s(\boldsymbol{x}) \geq s_{\mathrm{th}} \\ y &= -1, \quad \text{if} \quad s(\boldsymbol{x}) < s_{\mathrm{th}}. \end{aligned}$$

This is optimal in the sense that it maximizes the positive sample accuracy while keeping the negative sample accuracy constant.

Proof By the definition of the major and positive sample accuracy, the optimal decision criterion s^* can be formally written as

$$s^* = \arg\max_{s} \int \mathrm{d}\boldsymbol{x} \, I\left[s(\boldsymbol{x}) \geq s_{\mathrm{th}}\right] p(\boldsymbol{x}|\boldsymbol{\eta}, y = +1), \tag{23}$$

where $\boldsymbol{\eta}$ represents the set of model parameter, which is $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ in our case. Also, $I[\cdot]$ is the indicator function, which is 1 if the condition $[\cdot]$ holds, and 0 otherwise. The threshold s_{th} is related with the negative sample accuracy α by the equation

$$\int \mathrm{d}\boldsymbol{x}\ I\left[s(\boldsymbol{x}) \geq s_{\mathrm{th}}\right]\ p(\boldsymbol{x}|\boldsymbol{\eta}, y=-1) = 1-\alpha \tag{24}$$

Using a Lagrange multiplier λ, this problem can be rephrased as the maximization of $\Psi[s|\lambda]$ w.r.t. s:

$$\Psi[s|\lambda] = \int \mathrm{d}\boldsymbol{x}\ I\left[s(\boldsymbol{x}) \geq s_{\mathrm{th}}\right] \{p(\boldsymbol{x}|\boldsymbol{\eta}, y=1) - \lambda p(\boldsymbol{x}|\boldsymbol{\eta}, y=-1)\} \tag{25}$$

To maximize the integral, the indicator function $I[\cdot]$ must be 1 wherever $\{\cdot\} > 0$. The condition is readily given as

$$s(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\eta}, y=+1)}{p(\boldsymbol{x}|\boldsymbol{\eta}, y=-1)},\quad \lambda = s_{\mathrm{th}} \tag{26}$$

If we re-define a new criterion by transforming it using the logarithm function as $s(\boldsymbol{x})$, this coincides with Eq. (21). □

The NP decision rule is reduced to the Bayes decision rule when the data is balanced. For imbalanced data, however, they give different criteria.

The score function $s(\boldsymbol{x})$ uses the parameters $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ computed via the GHEM algorithm. In the same spirit, noting that $\sum_{i=1}^{W} w_i = 1$ holds, we approximate $s(\boldsymbol{x})$ as

$$s(\boldsymbol{x}) \approx \sum_{i=1}^{W} w_i \ln \frac{p(\boldsymbol{x}|\phi_i^{+1}, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}{p(\boldsymbol{x}|\phi_i^{-1}, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)} \tag{27}$$

$$= {\boldsymbol{z}^1}^\top \boldsymbol{d}^1 + {\boldsymbol{z}^0}^\top \boldsymbol{d}^0, \tag{28}$$

where $\boldsymbol{z}^1$ and $\boldsymbol{z}^0$ are M-dimensional vectors whose l-th components are $\delta(x_l, 1)$ and $\delta(x_l, 0)$, respectively. Also, $\boldsymbol{d}^1$ and $\boldsymbol{d}^0$ are M-dimensional vectors whose l-th components are defined by

$$d_l^1 \equiv \sum_{i=1}^{W} w_i h_{i,l}^1,\quad h_{i,l}^1 \equiv \ln \frac{P_{i,l}^{+1}}{P_{i,l}^{-1}},\quad d_l^0 \equiv \sum_{i=1}^{W} w_i h_{i,l}^0,\quad h_{i,l}^0 \equiv \ln \frac{1-P_{i,l}^{+1}}{1-P_{i,l}^{-1}}, \tag{29}$$

with $P_{i,l}^{+1}$ and $P_{i,l}^{-1}$ being a shorthand notation of $P(\phi_i^{+1}, a_l^*, b_l^*, c_l^*)$ and $P(\phi_i^{-1}, a_l^*, b_l^*, c_l^*)$, respectively.

The expression Eq. (28) suggests that the NP decision rule is essentially a linear classifier w.r.t. the 1-of-K coded version of $\boldsymbol{x}$. The linearity comes from the conditional independence expressed in Eq. (3), where the distribution of $\boldsymbol{x}$ is represented as the product over individual dimensions. Although Proposition 1 guarantees the optimality, the resulting classifier can be suboptimal if the model class is too limited. In the next subsection, we introduce one approach to overcome the limitation.

6.2 Deriving Distance Metric from sIRT

To overcome the limitation due to the conditional independence, we combine the sIRT model with the k-NN classification framework, which naturally handles nonlinear decision boundaries. This is indeed preferable in terms of interpretability since k-NN naturally achieves the comparability between different instances. For example, in healthcare questionnaire analysis, finding similar subjects or patients is a part of doctors' daily routines. In project risk management, lessons and learned from historical records is also an important part of the quality assurance process.

The k-NN algorithm first finds k nearest neighbors from the training data. To capture complex heterogeneity in the space defined by the ordinal variables, we use the Riemannian distance parametrized by a Riemannian metric A:

$$d^2_{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{x}') \equiv (\boldsymbol{x} - \boldsymbol{x}')^\top \mathsf{A} (\boldsymbol{x} - \boldsymbol{x}'). \tag{30}$$

For a newly observed $\boldsymbol{x}$, we compute the classification score

$$s^{k\mathrm{NN}}(\boldsymbol{x}) \equiv \ln \frac{\pi(-1)N^{+1}}{\pi(+1)N^{-1}}, \tag{31}$$

where N^{+1} and N^{-1} are the number of positive and negative instances in the neighborhood. The symbols $\pi(-1)$ and $\pi(-1)$ are defined in Eq. (5). If this score is greater than a certain threshold, s_{th}, we classify $\boldsymbol{x}$ into the positive class. The threshold as well as the number of nearest neighbors k is determined by leave-one-out (LOO) cross validation (CV).

To learn A, we use the following criterion:

If $p(\cdot \mid \boldsymbol{\eta}, y = +1)$ differs from $p(\cdot \mid \boldsymbol{\eta}, y = -1)$, the difference should be explained by the difference between $\mathcal{N}(\cdot \mid \boldsymbol{x}', \nu \mathsf{I}_M)$ and $\mathcal{N}(\cdot \mid \boldsymbol{x}', \mathsf{A})$.

Here $\boldsymbol{\eta} = (\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)$, and $\mathcal{N}(\cdot \mid \boldsymbol{x}', \nu \mathsf{I}_M)$ denote the Gaussian distribution of the mean $\boldsymbol{x}'$ and the covariance matrix $\nu \mathsf{I}_M$ with I_M being the M-dimensional identity matrix and ν being a scale parameter. Also, $\boldsymbol{x}'$ is an arbitrary location to learn A, which will be averaged out in the final result.

Let us use the Kullback-Leibler (KL) divergence to quantify the difference between the probability distributions. First, for the Gaussian part, we have

$$\int \mathrm{d}\boldsymbol{x}\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{x}', \nu \mathsf{I}_M) \ln \frac{\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{x}', \nu \mathsf{I}_M)}{\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{x}', \mathsf{A})} = \frac{1}{2} \mathrm{Tr} \left((\mathsf{A} - \nu \mathsf{I}_M) \Sigma(\boldsymbol{x}') \right), \tag{32}$$

where we put a constraint of $|\mathsf{A}| = 1$ and defined

$$\Sigma(\boldsymbol{x}') \equiv \int \mathrm{d}\boldsymbol{x}\, \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{x}', \mathsf{I}_M)(\boldsymbol{x} - \boldsymbol{x}')(\boldsymbol{x} - \boldsymbol{x}')^\top. \tag{33}$$

Second, for the sIRT model part, based on Eq. (27), we have

$$\begin{aligned} &\sum_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{\eta}, y = -1) \ln \frac{p(\boldsymbol{x}|\boldsymbol{\eta}, y = -1)}{p(\boldsymbol{x}|\boldsymbol{\eta}, y = +1)} \\ &\qquad = -\sum_{l=1}^{M} \sum_{i=1}^{W} w_i \left\{ P^{-1}_{i,l} h^1_{i,l} + (1 - P^{-1}_{i,l}) h^0_{i,l} \right\} \end{aligned} \tag{34}$$

By equating Eqs. (32) with (34), we have a solution to A as

$$\mathsf{A}_{l,m} = -\frac{\delta(l,m)}{\sigma_l^2}\sum_{i=1}^{W} w_i \left\{ P_{i,l}^{-1} h_{i,l}^{1} + (1 - P_{i,l}^{-1}) h_{i,l}^{0} \right\}, \tag{35}$$

where we have adjusted the proportionality factor between the two KL divergences to absorb the unimportant 1/2 factor and replaced the diagonal element of the local scatter matrix $\mathbf{\Sigma}(\boldsymbol{x}')$ with the variance σ_l^2 over all of the samples. We also dropped the ν parameter by regarding as a small constant. This is the equation that bridges the sIRT and the Riemanian metric. Note that the r.h.s. of Eq. (35) is positive in spite of the minus sign because it is originally defined as the KL divergence.

6.3 Informativeness Scores

Equation (35) essentially says that the KL divergence between the positive and negative distributions is proportional to the diagonal elements of the distance metric. Since the diagonal elements work as a weighting factor in the distance (see Eq. (30)), in the context of k-NN classification, they also can be thought of as the informativeness in the classification. We explicitly define the *informativeness* of the l-th variable V_l as

$$V_l^{\mathrm{KL}} = -\frac{1}{\sigma_l^2}\sum_{i=1}^{W} w_i \left\{ P_{i,l}^{-1} h_{i,l}^{1} + (1 - P_{i,l}^{-1}) h_{i,l}^{0} \right\}, \tag{36}$$

where the superscript represents the KL divergence.

Along the same line, we may define another informativeness score by replacing the KL divergence with another divergence measure. One practically useful measure is the Kolmogorov-Smirnov (KS) goodness-of-fit statistic:

$$\mathrm{KS}_l = \left| p(x_l = 1 \mid \boldsymbol{\eta}, y = +1) - p(x_l = 1 \mid \boldsymbol{\eta}, y = -1) \right|, \tag{37}$$

where $p(x_l \mid \boldsymbol{\eta}, y)$ is the marginal distribution w.r.t. the l-th variable. One major practical advantage of the KS statistic is that it is bounded within [0,1]. If it is 0, the positive and negative distributions have no difference and thus not important when making prediction. If it is 1, the variable should play a major role in prediction.

Let us consider how to find an expression of the marginal distribution $p(x_l \mid \boldsymbol{\eta}, y)$. For this, we exploit the approximation made in Eq. (27). Specifically, we start from an approximated relationship as

$$\ln p(\boldsymbol{x}|\boldsymbol{\eta}, y) \approx \sum_{i=1}^{W} w_i \ln p(\boldsymbol{x}|\phi_i^y, \boldsymbol{\eta}) = \sum_{l=1}^{M}\sum_{i=1}^{W} w_i J_{i,y}^{l}, \tag{38}$$

where $J^l_{i,y}$ is defined by replacing $\phi_i^{(n)}$ with ϕ_i^y of Eq. (22). Using this, we get

$$p(x_l \mid \boldsymbol{\eta}, y) = \frac{1}{Z_{l,y}} \exp\left(\sum_{i=1}^{W} w_i J^l_{i,y}\right), \tag{39}$$

where $Z_{l,y}$ is the normalization constant. By inserting this into Eqs. (37) and (35), we have another definition of the informativeness as

$$V_l^{\mathrm{KS}} = \left| \frac{1}{Z_{l,+1}} \exp\left(\sum_{i=1}^{W} w_i J^l_{i,+1}\right) - \frac{1}{Z_{l,-1}} \exp\left(\sum_{i=1}^{W} w_i J^l_{i,-1}\right)\right|. \tag{40}$$

6.4 Algorithm Summary

Before summarizing our informative prediction framework, consider the predictive distribution of the latent variable θ. In the GHEM framework, the distribution of $\boldsymbol{x}$ is written as a mixture model (see Eq. (12)). This allows us to readily get the posterior distribution for θ as a discrete distribution, given newly observed data $(\boldsymbol{x}, y)$, as

$$p(\theta = \phi_i^y \mid \boldsymbol{x}, y, \mathcal{D}) = \frac{w_i p(\boldsymbol{x}|\phi_i^y, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}{\sum_{j=1}^{W} w_j p(\boldsymbol{x}|\phi_j^y, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}. \tag{41}$$

If y is not available, the posterior predictive distribution for θ is given by

$$p(\theta \mid \boldsymbol{x}, \mathcal{D}) = \frac{\sum_{y=-1,+1} \pi(y) f(\theta|y)\, p(\boldsymbol{x}|\theta, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}{\sum_{y'=-1,+1} \pi(y') \int \mathrm{d}\theta' f(\theta'|y')\, p(\boldsymbol{x}|\theta', \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}. \tag{42}$$

To compute the denominator, we use the GH quadrature in Eqs. (8)-(10).

We have described the questionnaire-based informative prediction framework. In terms of interpretability stated in Sec 3.1, first, the k-NN framework readily allows comparison to the existing instances to explain how similar/different a newly observed instance is. Second, the informativeness scores as well as the IRT parameters allow us to understand which questions are more informative and how. Finally, the slope (discrimination), the threshold (difficulty), and the guessing parameters of each of the ICCs allow quantitatively comparing different answer choices from a probabilistic perspective.

The major steps of the training phase in our questionnaire-based informative prediction approach are given in Algorithm 2. For the performance metric to tune the hyper-parameters, we use the F-score,

$$F \equiv \frac{2\alpha\beta}{\alpha + \beta}, \tag{43}$$

where α is the negative sample accuracy and β is the positive sample accuracy. To compute α, we first create a subset of the data by collecting samples having the true label of $y = -1$, then count the number of correctly/incorrectly

Algorithm 2 Training supervised IRT model.

```
Input: Training data $\mathcal{D}$. Candidate values of hyper-parameters $\{\gamma, \omega, k\}$. Initial values of the model parameters $\boldsymbol{a}^0, \boldsymbol{b}^0, \boldsymbol{c}^0$.
Output: Optimal model parameters $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ and hyper-parameters $\{\gamma^*, \omega^*, s^*_{\mathrm{th}}, k^*\}$. Distance metric $\mathsf{A}$ and the informativeness score $\{V_1, \ldots, V_M\}$.
for Each set of the candidate values of $\{\gamma, \omega, k\}$ do
  for Each partition $\mathcal{D} = \mathcal{D}^{\mathrm{r}} \cup \mathcal{D}^{\mathrm{e}}$ do
    Use Algorithm 1 to determine $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ on $\mathcal{D}^{\mathrm{r}}$.
    Use Eqs. (36) and (40) to compute $\mathsf{A}$ and $\{V_1, \ldots, V_M\}$ on $\mathcal{D}^{\mathrm{r}}$.
    Compute $s^{k\mathrm{NN}}(\boldsymbol{x})$ of Eq. (31) on $\mathcal{D}^{\mathrm{e}}$.
  end for
  Find $s_{\mathrm{th}}$ that gives the highest F-score for the current choice of the hyper parameters.
end for
Find $\{\gamma, \omega, k\}$ and $\{s_{\mathrm{th}}\}$ that give the best F-score.
Return $\{\gamma^*, \omega^*, s^*_{\mathrm{th}}, k^*\}$, $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$, and $\mathsf{A}$.
```

predicted samples. The positive sample accuracy is defined on the subset of truly $y = +1$ samples. The use of α and β for performance evaluation is theoretically supposed by Proposition 1 so the F-score is.

Once training is done, we are ready to make a prediction for newly observed samples. Algorithms 3 summarizes the procedure.

Algorithm 3 Questionnaire-based informative prediction.

```
Input: Training data $\mathcal{D}$. Optimal hyper-parameters $\{s^*_{\mathrm{th}}, k^*\}$. Distance metric $\mathsf{A}$.
Output: Predicted label $y$ for a new entity $\boldsymbol{x}$.
Identify $k^*$-NNs from $\mathcal{D}$ using Eq. (30).
Compute $s^{k\mathrm{NN}}(\boldsymbol{x})$ using Eq. (31).
Predict $y = +1$ if $s^{k\mathrm{NN}}(\boldsymbol{x}) \geq s^*_{\mathrm{th}}$. Otherwise predict $y = -1$.
Return $y$.
```

7 Experiment

This section presents results of experimental evaluation of our metric learning framework based on sIRT. We first use a synthetic data set for illustration. Then we show results based on three real data sets.

7.1 Design of Experiment

In this subsection, we briefly describe data sets we use and alternative methods compared to our approach.

7.1.1 Data sets

1. `Synthetic`. Randomly generated bi-level data of $(N, M) = (100, 2)$ as summarized in Table 1.
2. `CRA`. Real information technology (IT) project assessment data called Contractual Risk Assessment [14] of $(N, M) = (262, 22)$. The questionnaire is designed to record reviewers' impression rather qualitatively about major risk factors of project failures. Major items include service providers relationship with the customer, experience in the planned solution, completeness of the cost case, and feasibility of the schedule. The questionnaire is used in the final project management review to internally approve contract signing. Original data was standardized for each question to take 1 (at-risk) or 0 (no-risk). Each of filled questionnaires are associated with the label of project success $(y = -1)$ or non-success $(y = +1)$. This data is highly imbalanced. Majority of the samples belong to the negative class.
3. `PBA`. Another real IT project assessment data called Project Baseline Assessment of $(N, M) = (1056, 56)$. The questionnaire is designed to collect detailed facts of the status of solution design by asking rather objectively about the approach used to estimate labor costs. Similarly to `CRA`, the data is standardized to be bi-level on both question answers and the outcome. This data is highly imbalanced.
4. `WVS`. Publicly available survey data called World Values Survey [31]. The original survey consists of 430 questions in total, including many rather formal questions on demographics such as gender and language spoken. From the version of "Wave 6" data, we chose 17 questions that would be relevant to the level of happiness in life, as listed in Table 2. We standardized the answers to bi-level using certain threshold values, and made "V10" the target variable. We focus only on instances having a country code of 840, which corresponds to the USA. We removed all of instances having any missing entries, resulting in $(N, M) = (934, 16)$. This is also highly imbalanced data because majority of people tell that they are happy.

Table 1 Summary of `Synthetic` data.

$\boldsymbol{x}$	$(y = +1)$	$(y = -1)$
(0,0)	8	9
(0,1)	6	16
(1,0)	20	20
(1,1)	16	16

7.1.2 Methods compared

We call the approach summarized in Algorithms 1-3 `sIRT-KL` (supervised IRT based on the KL divergence metric). As a variant, we also use `sIRT-KS` (supervised IRT based on the KS statistics) by replacing the r.h.s. of Eq. (35) with Eq. (40) (except for Kronecker's delta). To compare the performance between the NP decision rule and the k-NN approach, we also consider `sIRT-LR` (supervised IRT based on the likelihood ratio), which does not use k-NN (Eq. (31)) but the classification rule of Eq. (21).

As we stated in Sec. 3.1, our goal is to develop a classifier that has full interpretability. For further comparison, we use the following existing classifiers. Since our goal is *not* to achieve the highest ever accuracy but to make informative prediction, we focus only on methods that are thought of as interpretable in practice. Some of the ensemble methods and advanced nonlinear classifiers such as deep learning methods may achieve better accuracies, but they are not comparable to our approach because they do not care about the notion of full interpretability.

1. `k-NN`. The k-NN approach based on the Eucleadian distance, i.e. $\mathsf{A} = \mathsf{I}_M$. For the score, Eq. 31 is used. The threshold and k are determined via LOO CV.
2. `LMNN`. Large-margin nearest neighbors (LMNN) [29], which is the standard baseline method in metric-learning-based k-NN classifier. In LMNN, the *full* Riemannian metric A is determined by minimizing the objective func-

Table 2 Variables in World Values Survey data.

question ID	variable	question summary
V10	y	Overall happiness
V49	x_1	Goal in life is to make parents proud
V50	x_2	When mother works, children suffer
V55	x_3	Have a great deal of choice in life
V56	x_4	Other people are fair
V58	x_5	Have children
V59	x_6	Satisfied with household financial situation
V70	x_7	Liked by those who value creativity
V71	x_8	Liked by those who want to be rich
V77	x_9	Liked by those who aways behave properly
V78	x_{10}	Liked by those who care about natural environment
V79	x_{11}	Liked by those who value tradition
V102	x_{12}	Trust family
V103	x_{13}	Trust neighborhood
V105	x_{14}	Trust people met for the first time
V107	x_{15}	Trust people from another nationality
V170	x_{16}	Neighborhood is secure

tion

$$E(\mathsf{A}) = \frac{1-\mu}{N} \sum_{n=1}^{N} \sum_{j \in \mathcal{N}_n} d^2_{\mathsf{A}}(n,j) \tag{44}$$

$$+ \frac{\mu}{N} \sum_{n=1}^{N} \sum_{j \in \mathcal{N}_n} \sum_{l:\, y^{(l)} \neq y^{(n)}} \left[1 + d^2_{\mathsf{A}}(n,j) - d^2_{\mathsf{A}}(n,l) \right]_+ , \tag{45}$$

where $d^2_{\mathsf{A}}(n,j)$ is a shorthand notation of $d^2_{\mathsf{A}}(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(j)})$, $\mathcal{N}_n$ represents the set of the nearest-neighbors of $\boldsymbol{x}^{(n)}$ chosen from the same label samples, *i.e.*, $y^{(j)} = y^{(n)}$, and $[h]_+ = \max\{0, h\}$ for $\forall h \in \mathcal{R}$. LMNN is believed to be one of the best off-the-shelf metric learning methods [2], thanks mainly to the hinge loss function and its convex formulation [29].

3. `RegLgst`. L_1-regularized logistic regression [7], whose central model is given by
$$\ln \frac{p(y=+1 \mid \boldsymbol{x})}{p(y=-1 \mid \boldsymbol{x})} = \boldsymbol{d}^\top \boldsymbol{x} + d_0.$$
The parameters $\boldsymbol{d} \equiv (d_1, d_2)^\top$ and d_0 are learned via maximum likelihood under the L_1 constraint on $\boldsymbol{d}$. The regularization constant was optimized using LOO CV.

7.2 Illustration using `Synthetic` data

To explain why informativeness matters, we compare `sIRT-KL` and `RegLgst` based on the `Synthetic` data. In this 2-dimensional setting, we have only four choices in $\boldsymbol{x}$ as shown in Table 1.

Figure 5 shows learned coefficients of `RegLgst`. In the figure, we see that both dimensions have significant weights, and it is not clear how each of them affects classification. Almost the only conclusion we can draw would be something like “you cannot ignore either one”.

Figure 6 shows results of `sIRT-KL`. The informativeness score calculated by Eq.(40) clearly shows that x_1 is more important than x_2. This is confirmed by the ICC, where x_2 is less sensitive and even negatively depends on θ. If this is a diagnostic inquiry and a doctor is trying to infer the level of medical risk, the doctor may decide to use only the inquiry x_1 based on the ICCs to distinguish between low risk (small θ) and high risk (large θ) subjects. We see that decision-making becomes much easier with the aid of ICCs. Note that putting a stronger regularizer and thus getting a sparser solution does not improve the situation because the logistic regression coefficients still look like black-box metric that may take negative values.

7.3 Project Risk Assessment: `CRA` and `PBA` data

Following the procedure summarized in Sec. 6.4, we calculated $\{\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*, \}$ based on the `CRA` and `PBA` data. For the hyper-parameters (ω, γ, k), we had

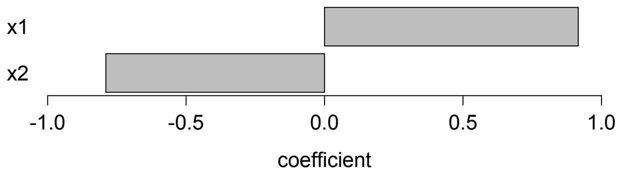


Fig. 5 Learned coefficients of `RegLgst` for `Synthetic`.

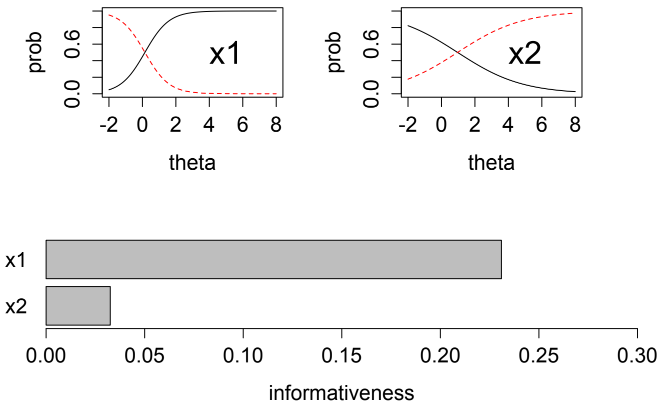


Fig. 6 Item characteristic curves and informativeness score for `Synthetic`.

$(0.1, 0.5, 7)$ for `CRA`, and $(4.0, 5.0, 86)$ for `PBA`. For the initial values, we used $a_l^0 = 1.0, b_l^0 = 0.5, c_l^0 = 10^{-5}$ for all $l = 1, \dots, M$. To handle the imbalanced nature between troubled and healthy samples, in the training phase, we did bootstrap resampling for the non-success instances to obtain the same sample size in either class, although the prediction was made for the original data.

Figure 7 shows the sIRT parameters and the informativeness for `CRA`. We see that the 7th and 9th questions have major informativeness. Interestingly, these ones have negative discrimination parameters. This is due to the nature of risk management process. Since this risk assessment is done after completing all of risk mitigation actions previously suggested by human auditors, readily visible risks cannot exist. The situation is similar to accounting audits of businesses to find out creative accounting. The negative $\{a_l\}$'s suggest that some of the trouble project had questionnaire answers that were "too good to be true" or "unnaturally good."

Figure 8 shows some examples ICCs. We drew $P(\theta, a_l, b_l, c_l)$ with the solid lines as well as $[1 - P(\theta, a_l, b_l, c_l)]$ with the dashed lines. We clearly see that the 10th question is hardly informative, being consistent to the negligible informativeness score in Fig. 7. Interestingly, this question is about future project

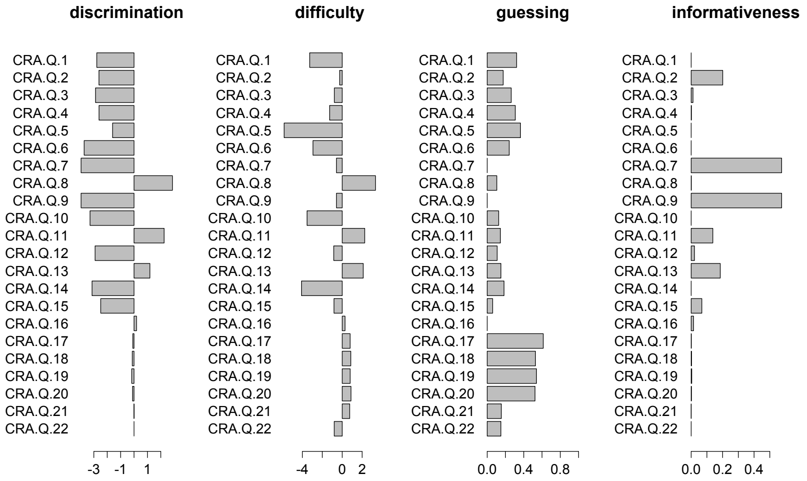


Fig. 7 IRT parameters and informativeness scores learned from the CRA data.

plan after contract signing, which will be conducted by a different team from the one being reviewed. Thus negligible informativeness makes a lot of sense.

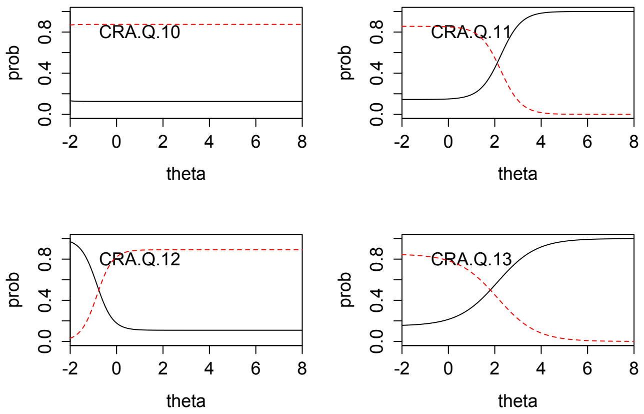


Fig. 8 Examples of ICCs from the CRA data.

Next, we compare prediction performance of our approach with alternatives. The results are shown in Fig. 9. We see that sIRT-KL and sIRT-KL are consistently better or comparable to the alternative methods. As already discussed in Sec. 6.1, sIRT-LR, which is naively based on the Neyman-Pearson decision rule, gives systematically worse results than sIRT-KL. Although RegLgst and k-NN give good performance for WVS, they perform poorly for CRA and

PBA. The fact that our method achieved comparable or even better prediction performance guarantees that the descriptive analysis with the ICCs and the informativeness scores capture the major features of the data.

It is interesting to see that the LMNN approach, which optimizes the full Riemannian metric via Eq. (44), does not help improve the performance very much. This clearly suggests the importance of the nonlinear transformation by the logistic curve of IRT, and the risk of naively applying metric learning in non-metric spaces. This also suggests that explicitly taking account of human cognition bias is critical in questionnaire data analysis. Our approach successfully captured the latent failure tendency with the aid of the psychoanalytical approach.

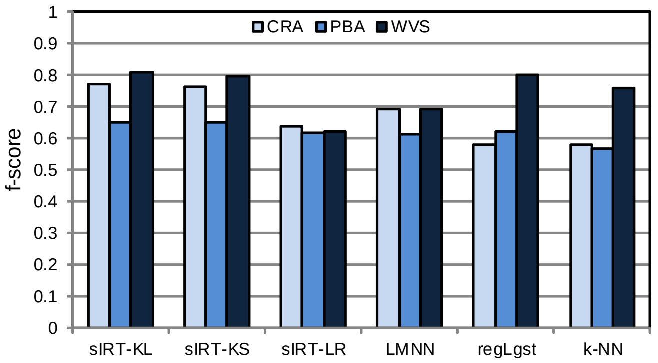


Fig. 9 Comparison of F-values in prediction.

7.4 World Values Survey: Informative Prediction of Happiness

As we summarized in Table 2, our target variable ("V10") is overall happiness, and we are interested in understanding what makes them feel happy with their life. We trained sIRT model using the WVS data. Figure 10 shows the sIRT parameters and the informativeness.

Unlike CRA data, we see that all of the discrimination parameters ($\boldsymbol{a}$) are positive. This means that all of the questions in Table 2 affect the overall happiness positively. We also see that V49, V71, and V102 have very high guessing values, which are almost 1, and their informativeness scores are quite low. These questions are not about respondents themselves but about their parents, other people who seem to want to be rich, and the level of their trust for their family. The result shows that the overall happiness is not significantly affected by these factors.

On the other hand, we also see that V55 and V59 have very high informativeness scores. This is very interesting because V55 asks about how much controllable their life is and V59 asks about the level of financial satisfaction of their household. This result looks consistent to the general beliefs in the USA: anyone can succeed by properly exercising their own capabilities. It would be interesting to compare this analysis for the USA to other countries. However, we leave it to future work.

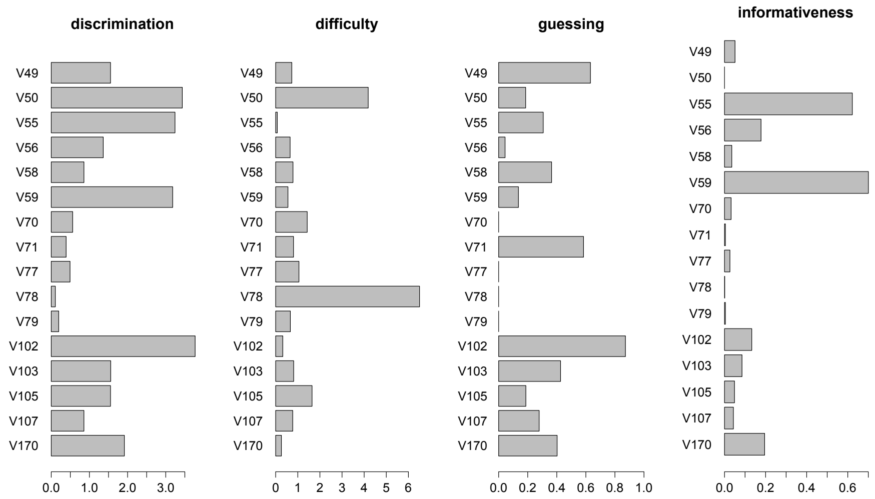


Fig. 10 IRT parameters and informativeness scores learned from the `WVS` data.

8 Concluding remarks

We have addressed the task of informative prediction for questionnaire data. Our primary goal was to establish a method to quantitatively evaluate the informativeness of question items based on the predictability of the final outcome of individual samples.

To tackle the task, we introduced a new framework of supervised item response theory (sIRT) by extending an existing theory in psychometrics. We introduced a prior distribution for the latent variable conditioned on the outcome variable, and developed an efficient iterative algorithm named the Gauss-Hermite EM (GHEM) algorithm for parameter estimation. By combining the sIRT model with metric learning, we successfully developed an informative prediction approach that achieves full interpretability (as defined in Sec. 3.1). Using real-world data sets, we confirmed that our approach gives valuable insights in practice. In particular, in an analysis on World Values Survey, we successfully characterized factors making people feel happy.

For future work, it would be possible to extend the present framework to multi-level questions. One simple approach is to leverage the 1-of-K notation and reduce each of the K-level questions to K bi-level questions. Other extensions such as introduction of multi-variate latent variable would also be possible.

References

1. Baker FB, Kim SH (2004) Item Response Theory: Parameter Estimation Techniques , 2nd edn. CRC Press
2. Bellet A, Habrard A, Sebban M (2013) A survey on metric learning for feature vectors and structured data. ArXiv e-prints 1306.6709
3. Bishop CM (2006) Pattern Recognition and Machine Learning. Springer-Verlag
4. Borji A, Itti L (2013) Bayesian optimization explains human active search. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Advances in Neural Information Processing Systems 26, pp 55–63
5. Chapelle O, Chang Y, Liu T (eds) (2011) Proceedings of the Yahoo! Learning to Rank Challenge, held at ICML 2010, Haifa, Israel, June 25, 2010, JMLR Proceedings, vol 14
6. Cuturi M, Avis D (2014) Ground metric learning. Journal of Machine Learning Research 15(1):533–564
7. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1–22
8. Goldberger J, Roweis S, Hinton G, Salakhutdinov R (2005) Neighbourhood component analysis. In: Advances in Neural Information Processing Systems, 17, pp 513–520
9. Guillaumin M, Verbeek J, Schmid C (2009) Is that you? metric learning approaches for face identification. In: Computer Vision, 2009 IEEE 12th International Conference on, pp 498–505
10. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction , 2nd edn. Springer
11. Hildebrand HB (1974) Introduction to Numerical Analysis, 2nd edn. Dover
12. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainathand T, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29(6):82–97
13. Idé T, Dhurandhar A (2015) Informative prediction based on ordinal questionnaire data. In: Proceedings of 2015 IEEE International Conference on Data Mining (ICDM 15), pp 191–200
14. Idé T, Güven S, Jan EE, Makogon S, Venegas A (2015) Latent trait analysis for risk management of complex information technology projects. In: Proceedings of the 14th IFIP/IEEE International Symposium on Integrated Network Management, IM 15, pp 305–312

15. Koren Y, Sill J (2011) Ordrec: An ordinal model for predicting personalized item rating distributions. In: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, pp 117–124
16. Koren Y, Sill J (2013) Collaborative filtering on ordinal user feedback. In: IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, pp 3022–3026
17. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37
18. Kostinger M, Hirzer M, Wohlhart P, Roth P, Bischof H (2012) Large scale metric learning from equivalence constraints. In: Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2288–2295
19. Lan AS, Waters AE, Studer C, Baraniuk RG (2014) Sparse factor analysis for learning and content analytics. Journal of Machine Learning Research 15(1):1959–2008
20. Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp 609–616
21. McCullagh P (1980) Regression models for ordinal data. Journal of the Royal Statistical Society Series B (Methodological) 42(2):109–142
22. Murray W, Wright MH (1995) Line search procedures for the logarithmic barrier function. SIAM Journal on Optimization 4(2):229–246.
23. Osogami T, Otsuka M (2014) Restricted boltzmann machines modeling human choice. In: Advances in Neural Information Processing Systems 27, pp 73–81
24. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135
25. SAT (2015) Wikipedia; `http://en.wikipedia.org/wiki/SAT`
26. Stevens SS (1946) On the theory of scales of measurement. Science 103(2684):677–680
27. Sun BY, Li J, Wu D, Zhang XM, Li WB (2010) Kernel discriminant learning for ordinal regression. IEEE Transactions on Knowledge and Data Engineering 22(6):906–910
28. Terada Y, Luxburg UV (2014) Local ordinal embedding. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), JMLR Workshop and Conference Proceedings, pp 847–855
29. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244
30. Wilson M (2004) Constructing Measures. Psychology Press
31. World Values Survey Association (2015) World Values Survey. `www.worldvaluessurvey.org`, Wave 6, 2010-2014, Official Aggregate v.20150418
32. Xing EP, Jordan MI, Russell S, Ng AY (2002) Distance metric learning with application to clustering with side-information. In: Advances in neural information processing systems, pp 505–512