Embedded Restricted Boltzmann Machines for Fusion of Mixed Data Types and Applications in Social Measurements Analysis

Truyen Tran  
Department of Computing  
Curtin University  
Bentley, Western Australia, Australia  
t.tran2@curtin.edu.au

Dinh Q. Phung and Svetha Venkatesh  
School of Information Technology  
Deakin University  
Geelong, Victoria, Australia  
{dinh.phung,svetha.venkatesh}@deakin.edu.au

Abstract—Analysis and fusion of social measurements is important to understand what shapes the public’s opinion and the sustainability of the global development. However, modelling data collected from social responses is challenging as the data is typically complex and heterogeneous, which might take the form of stated facts, subjective assessment, choices, preferences or any combination thereof. Model-wise, these responses are a mixture of data types including binary, categorical, multicategorical, continuous, ordinal, count and rank data. The challenge is therefore to effectively handle mixed data in the a unified fusion framework in order to perform inference and analysis. To that end, this paper introduces eRBM (Embedded Restricted Boltzmann Machine) — a probabilistic latent variable model that can represent mixed data using a layer of hidden variables transparent across different types of data. The proposed model can comfortably support large-scale data analysis tasks, including distribution modelling, data completion, prediction and visualisation. We demonstrate these versatile features on several moderate and large-scale publicly available social survey datasets.

Keywords—Information fusion; mixed data types; embedded restricted Boltzmann machines; social measurements analysis.

I. INTRODUCTION

Understanding the public opinion, perception and attitude towards various economic, social and cultural issues contributes to sustainable social development and policy making. An important key step towards such an understanding is to design effective measurement and survey tools, collecting data from a diverse population at a large-scale and analyse the data. For example, the Pew Global Attitudes Project\(^1\) has set out the objectives of surveying hundreds of thousands of people across continents to capture timely snapshots of world views over cultures, values, local conditions and state of lives. On the one hand, getting the right global survey data is difficult and expensive: a process that requires expertise from local knowledge and experience, culture and religion, economy, sociology, and political science. On the other hand, making sense of the collected data is equally complicated due to the complex nature of responses.

\(^1\)pewglobal.org

The challenges arise in analyzing such collections of data are manifolds. First, responses are heterogeneous in type: they typically contain a mixture of stated facts, personal assessment, choices and preferences. Put in statistical terms, responses can be either binary, categorical, multicategorical, ordinal, continuous, count or category-ranked. There are no simple scaling and coding techniques that can make these responses more homogeneous without distorting the data. Modelling the covariance among these data types is inherently difficult, let alone inferring meaningful statistics. Second, the subsampled data is prone to bias, noise and missing entries.

The research questions are therefore how can we represent and fuse complex responses in a unified manner? More importantly, how might one infer the implied semantics which are hidden in the raw responses? What is the correlations among any group of responses? Given previous answers, can we predict the an unseen outcome? For example, can we predict a person’s financial success based on his attitudes towards things in life? If a respondent refuses to answer a particular sensitive question, can we make an educated guess? What are the similarity and gulfs in public views between countries and cultures? What is the variance among people of the same ethnic group, same country or the same region? How can we effectively visualise the data? etc.

Researchers have attempted to address these questions using statistical tools (e.g., \([7]\)); however, most existing methods are limited to understanding univariate data or bivariate correlations. There has been limited focus on joint modelling of such complex data types [1].

This paper\(^2\) presents embedded RBM (eRBM), a probabilistic tool that provides answers to these questions. As its name implies, it is based on a probabilistic graphical architecture known as the Restricted Boltzmann Machine (RBM) [17], [5], [8]. A RBM is a bipartite Markov random field [9] wherein the input layer is associated with observed responses, and the output layer typically consists of hidden binary factors of variation. The model is ‘restricted’ in the

\(^2\)This paper is a condensed version of the technical report, which will be made available at http://truyen.vietlabs.com.
sense that connectivity is limited to nodes between layers only. The idea is that an input can be encoded into a set of binary factors, where a factor contributes to the model density only when it is activated.

Leveraging from the RBM architecture, we make a number of innovations in this paper. First, it presents a novel use of the RBM, borrowed from the machine learning literature, for the social surveying analysis. Second, our RBM differs from standard RBM in that the input layer is also latent – their values are not observed directly. As such, our RBM is embedded into the data. Through the observed responses, we can deterministically infer a set of type-specific constraints to each input variable. This setting gives rise to a new problem of learning a probabilistic graphical model given only variable constraints. Third, through the use of (Gaussian) RBM, we show that all most common response types can be represented naturally – learning and inference given such a joint complex set of responses can be made efficient. This is because the bipartite architecture of the RBM offers a factoring scheme in which decouples the responses representation from their implied semantics, at the same time, still maintain long-range and high-order dependencies.

The eRBM is capable of supporting a variety of survey inference tasks. In particular, the posteriors of binary factors given the responses can be used as a homogeneous vectorial representation, hiding away the complex nature of the responses. This representation is natural for further data analysis needs, including visualisation, dimensionality reduction and clustering. On the other hand, given the binary factors, we can reconstruct the responses in a generative way, making eRBM a candidate for data compression. The eRBM also directly supports predictive analysis – it can learn to perform classification, regression and ranking at the same time, still maintain long-range and high-order dependencies.

The eRBM is capable of supporting a variety of survey inference tasks. In particular, the posteriors of binary factors given the responses can be used as a homogeneous vectorial representation, hiding away the complex nature of the responses. This representation is natural for further data analysis needs, including visualisation, dimensionality reduction and clustering. On the other hand, given the binary factors, we can reconstruct the responses in a generative way, making eRBM a candidate for data compression. The eRBM also directly supports predictive analysis – it can learn to perform classification, regression and ranking at the same time, still maintain long-range and high-order dependencies.

The correlation among responses $v = (v_1, v_2, ..., v_N)$ is captured through the correlation among corresponding utilities $u = (u_1, u_2, ..., u_N)$, i.e., by using $P(v) = \sum_u P(v | u) P(u)$. However, specifying and estimating $P(u)$ from data can be highly challenging for large $N$, which can be several hundreds. For example, the typical multivariate Gaussian approach to model $P(u)$ is not scalable with $N$. Second, if we can infer $P(u | v)$, this is of limited practical use since the dimensionality of $u$ is typically large (hundreds or thousands).

Here we pursue a different approach by factorising the correlation structure among the utilities. More specifically, we introduce an additional binary factor layer $h \in \{0, 1\}^K$ on top of the utility layer $u$. This serves double purposes: one is the data representation – we will have a collection of $2^K$ possible ways to explain the variations between users in the population using only a compact set of $K$ dimensions. The other is computational efficiency – we will be able to make inference faster as described below.

To achieve the second goal, we limit the interaction between latent variables so that direct within-layer influences are not allowed. As depicted in Fig. 1, the top two layers form a bipartite network of Markov random field, also known as restricted Boltzmann Machine (RBM) [17], [5], [8]. Let $\Psi(u, h) \geq 0$ be the model potential function cap-
turing the correlation structure among variables. The RBM architecture specifies that we can factorise the potential function as

\[ \Psi(u, h) = \prod_i \phi_i(u_i) \prod_{ik} \psi_{ik}(u_i, h_k) \prod_k \phi_k(h_k) \]  

(1)

where \( \phi_i, \psi_{ik} \) and \( \phi_k \) are local potential functions. Finally, the model distribution is defined as

\[ P(v, u, h) = \frac{1}{Z} \Psi(u, h) P(v | u) \]  

(2)

where \( P(v | u) = \prod_i P_i(v_i | u_i) \) and \( Z = \int_u \sum_h \Psi(u, h) du \) is the finite normalising constant.4 The factorisation in Eq. (1) indeed offers nice decomposition of conditional distributions

\[ P(u | h) = \prod_i P(u_i | h) \]  

(3)

\[ P(u | v, h) = \prod_i P(u_i | v_i, h) \]  

(4)

\[ P(h | u, v) = P(h | u) = \prod_k P(h_k | u) \]  

(5)

following the standard theory of Markov blankets (MBs)5 in Markov random fields (MRF). The theory says that a variable is conditionally independent of all other variables given its MB. For example, the binary factor layer acts as the MB for each utility, and conversely the utility layer plays the role of the MB for each binary factor.

As the RBM is fully unobserved and embedded into our architecture, we term the proposed model as embedded RBM (eRBM).

B. Bernoulli-Gaussian RBMs

We now specify the distribution of the top layers \((h, u)\). For simplicity, we assume that the latent layer \(u\) follows the multivariate Gaussian distribution6. The choice of Gaussian is for convenience of computation only and in fact we can use any distributions in the exponential family. More precisely, let \(u_i = (u_{i1}, u_{i2}, ..., u_{iD_i})\), the local potentials are specified as

\[ \phi_i(u_i) = \exp \left\{ \sum_{d=1}^{D_i} \left( \frac{w_{id}^2}{2\sigma_i^2} + \alpha_{iid} u_{iid} \right) \right\} \]  

(6)

\[ \psi_{ik}(u_i, h_k) = \exp \left\{ \sum_{d=1}^{D_i} w_{idk} u_{iid} h_k \right\} \]  

(7)

\[ \phi_k(h_k) = \exp \left\{ \beta_k h_k \right\} \]  

(8)

where \( \alpha_i \) is the standard deviation of the \( i \)-th variable, \( \{ \alpha_{iid}, \beta_k, w_{idk} \} \) are free parameters. The two upper layers now form a Bernoulli-Gaussian RBM [8].

We can derive from Eq. (2) that, the distribution of utilities conditioned on all binary factors are indeed Gaussian

\[ P(u_i | h) \propto \phi_i(u_i) \prod_k \psi_{ik}(u_i, h_k) \]

\[ = \exp \left\{ \sum_{d=1}^{D_i} \left( \frac{w_{id}^2}{2\sigma_i^2} + \alpha_{iid} u_{iid} + \sum_k w_{idk} u_{iid} h_k \right) \right\} \]

\[ \propto \prod_k \mathcal{N}(u_{iid}; \mu_{iid}(h), \sigma_i) \]  

(9)

i.e., \( P(u_{iid} | h) \) is a Gaussian with mean \( \mu_{iid}(h) \) and standard deviation \( \sigma_i \), where

\[ \mu_{iid}(h) = \sigma_i^2 \left( \alpha_{iid} + \sum_{k=1}^{K} w_{idk} h_k \right) \]  

(10)

The binary posterior given utilities, on the other hand, assumes the form of a logistic regression model

\[ P(h_k | u) \propto \phi_k(h_k) \prod_i \psi_{ik}(u_i, h_k) \]

\[ = \exp \left\{ \beta_k + \sum_{id} w_{idk} u_{iid} \right\} \]  

(11)

which gives rise to the sigmoid form of \( P(h_k = 1 | u) = 1/(1 + e^{-s(u)}) \) where \( s(u) = \beta_k + \sum_{id} w_{idk} u_{iid} \).

C. Conditional Bernoulli-Gaussian RBMs

Now when \( v \) are observed, we need to take it into account when computing the conditional distributions as in Eqs. (9,11). Eq. (9) now becomes

\[ P(u_i | v_i, h) \propto P_i(v_i | u_i) P(u_i | h) \]

\[ = P_i(v_i | u_i) \prod_{d=1}^{D_i} \mathcal{N}(u_{iid}; \mu_{iid}(h), \sigma_i) \]  

(12)

Clearly the exact form of \( P(u_i | v_i, h) \) depends critically on the type-specific distribution \( P_i(v_i | u_i) \), which we will detail in Section II-E.

D. MCMC Inference

At this point all our local conditional distributions have been specified in Eqs. (9,11,12). Given the factorisation property in Eqs. (3–5), we can run a layer-wise Gibbs chain updating all elements at each layer in parallel. Given the sampling scheme, homogeneous data representation can be derived from the posterior vector \( \hat{h} = (P(h_1 | v), P(h_2 | v), ..., P(h_K | v)) \). This can be estimated by collecting samples of \( \{h_k\} \) from a chain where \( v \) is kept fixed to the observed responses. The new representation can be used for a number of tasks including visualisation, clustering, prediction and likelihood estimation.

Prediction of an unseen response (e.g., as in data imputation or predictive modelling) can be made using \( P_j(\nu_j | v) \), where \( \nu_j \notin v \). Although this can be achieved by collecting samples as usual, we propose here a more efficient approximation based on mean-field techniques [14]. First, we estimate the data representation \( \hat{h} \) from \( P(h | v) \), then
integrating over the utilities (if the integration has closed form solution) as follows

\[
P(v_j | v) \approx P(v_j | \hat{h}) = \int_{u_j} P(v_j, u_j | \hat{h}) du_j
\]

Data likelihood can be estimated in the same way, i.e.,

\[
P(v_i) \approx P(v_i | \hat{h}), \text{ where } v_i \text{ is now the seen response.}
\]

E. Type-specific Submodels

There is no doubt that the usefulness of the eRBM lies in the correct specification of type-specific submodels \( P_i(v_i | u_i) \). For space limit we present here the cases for binary, categorical, ordinal and count responses. Continuous responses are assumed to follow Gaussian distributions, and thus the utility can be simply fixed to the response value itself. A multivariate response is treated as multiple binary responses. A rank response is an ordering of categories, and thus is a generalisation of the categorical response\(^7\). Before describing the models, let us denote by \( S_i = \{ c_{i1}, c_{i2}, \ldots, c_{iM_i} \} \) the set of categories in the case of discrete variables.

1) Binary Responses.: A binary response outputs one of the two possible options, e.g., \{yes/no\}. In many ways it resembles the decision making process in which we evaluate the utility of the choice against some threshold. The decision can be formalised as

\[
P_i(v_i = 1 | u_i) = \mathbb{I}[u_i > \theta_i]
\]

i.e., we maintain one utility variable per response.

This leads to

\[
P(u_i | v_i = 1, h) = \mathcal{N}(\mu_i(h), \sigma_i) \mathbb{I}[u_i > \theta_i]
\]

i.e., the utility conditional distribution is a Gaussian truncated from below. Likewise, when \( v_i = 0 \), the distribution is truncated from above. Finally, given an estimate of the hidden layer \( h \), we can estimate the probability of the binary output as

\[
P_i(v_i = 1 | \hat{h}) = \int_{\hat{\theta}_i}^{\infty} \mathcal{N}(\mu_i(\hat{h}), \sigma_i) = 1 - \Phi(\theta_i^*)
\]

where \( \theta_i^* = \mu_i(h) / \sigma_i \), and \( \Phi(\cdot) \) is the cumulative distribution function (CDF).

2) Categorical Responses.: This refers to choosing a single element from a categorical set, e.g., the status of a person is one of \{married, living-with-a-partner, widowed, divorced, separated, never-been-married\}. We maintain a latent element per category, i.e., \( u_i = (u_{i1}, u_{i2}, \ldots, u_{iM_i}) \) where \( D_i = M_i \). A categorical choice is assumed to be made by selecting the one with the maximum latent utility, that is, \( u_i = c_{id_{max}} \) where \( d_{max} = \arg \max_{d} u_{id} \). This choice model can be captured by:

\[
P_i(v_i = c_{id} | u_i) = \mathbb{I}[u_{id} > \max_j u_{im}]
\]

The maximisation suggests that the utility variables for the \( i \)th observable are correlated, and thus the joint distribution \( P(u_{i1}, u_{i2}, \ldots, u_{iM_i} | v_i, h) \) is not factorisable.

\(^7\)For full description of all the types please consult our full technical report.
Now we want to simplify this complex distribution by some approximation. Let \( f(u_i) \) be the expression inside the bracket of the \( \exp \) function. The idea is to approximate \( f(u_i) \) by a quadratic function of \( u_i \); first we find the mode \( u_i = \mu_i \) and build a quadratic surrogate around that point. We can verify that \( f''(u_i) < 0 \), i.e., the function \( f \) has a global maximum. Finally, we have a Taylor’s expansion around the mode \( \mu_i \),

\[
f(u_i) \approx f(\mu_i) + f'(\mu_i)(u_i - \mu_i)^2,
\]

leading to the distribution approximation:

\[
P(v_i, u_i | h) \propto \frac{1}{k!} e^{f(\mu_i)} \exp\left(\frac{-(u_i - \mu_i)^2}{2\kappa_i^2}\right) \quad (14)
\]

where \( \kappa_i^2 = -1/f''(\mu_i) \). The method is often referred to as the Laplace’s approximation.

Given the approximation above, the response marginal \( P(v_i | h) = \int P(v_i, u_i | h) du_i \) can be estimated as

\[
P(v_i = r | h) \propto \frac{1}{r!} e^{f(\mu_i)} \int \exp\left(\frac{-(u_i - \mu_i)^2}{2\kappa_i^2}\right) du_i = \frac{1}{r!} e^{f(\mu_i)} \sqrt{2\pi \kappa_i^2} \quad (15)
\]

Finally, using Eqs. (14,15) the utility posterior becomes:

\[
P(u_i | v_i = r, h) = \frac{P(v_i, u_i | h)}{P(v_i = r | h)} \approx \mathcal{N}(\mu_i, \kappa_i^2)
\]

F. Learning with Persistent Markov Chains

We are interested in two learning problems. One is the estimation of the model distribution \( P(v) \) given the subsampled empirical distribution \( \hat{P}(v) \). The other is predictive modelling in that we want to predict unseen responses given a set of seen responses. This is often translated to estimating the conditional distribution \( P(v_c | v_{-c}) \) for the output responses \( v_c \). The latter is somewhat easier than the former since we already know \( v_{-c} \).

Due to space limit, we omit the details here – interested readers may consult the technical report version of this paper.

For the case of learning \( P(v) \), the general strategy is to iteratively maximising the data likelihood \( \hat{P}(\hat{v}) = \sum_h \int_u \hat{P}(\hat{v}, u, h) \). Since computing the likelihood and its gradient is clearly intractable, we pursue MCMC techniques to approximate the likelihood’s gradient – this results in a stochastic gradient ascent method. For each data point (e.g., a respondent), we maintain two Markov chains, one for the pair \((u, h)\) constrained by observed \( \hat{v} \) and another one for same pair without constraints. Parameters are updated after every several MCMC steps. This technique bears some similarity with the Persistent Contrastive Divergence [21], but our case is more complicated due to the constraints.

III. APPLICATION: SOCIAL MEASUREMENTS ANALYSIS

In this section, we apply our proposed eRBM for fusing complex social measurements and perform a number of analysis tasks, including discovering and visualising hidden profile of users, handling missing data and predicting responses.

---

### Table I

<table>
<thead>
<tr>
<th>Data</th>
<th>Bin</th>
<th>Cat</th>
<th>Mcat</th>
<th>Cont</th>
<th>Ord</th>
<th>Count</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA02</td>
<td>43</td>
<td>12</td>
<td>3</td>
<td>3</td>
<td>125</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GA08</td>
<td>52</td>
<td>124</td>
<td>0</td>
<td>3</td>
<td>165</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HEH11</td>
<td>55</td>
<td>52</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>-0.8</td>
</tr>
<tr>
<td>100</td>
<td>-1.2</td>
</tr>
<tr>
<td>150</td>
<td>-1.5</td>
</tr>
<tr>
<td>200</td>
<td>-1.8</td>
</tr>
<tr>
<td>250</td>
<td>-2.0</td>
</tr>
<tr>
<td>300</td>
<td>-2.2</td>
</tr>
<tr>
<td>350</td>
<td>-2.4</td>
</tr>
<tr>
<td>400</td>
<td>-2.6</td>
</tr>
</tbody>
</table>

Figure 2. Learning curve on GA02.

A. Data and Settings

Three public survey datasets collected by PewResearch Centre\(^8\) are used as input for our model. The first is the Global Attitude (GA02) survey, which interviewed 38,263 people in 44 countries in 2001–2002. The second is the Global Attitude (GA08) survey conducted in 24 countries on March 17 – April 21, 2008. The data contains answers from 24,717 respondents. The third dataset is the Higher Education/Housing (HEH11) survey of 2,142 adults living in the continental United States on March 15-29, 2011. See Appendix for more information on these datasets.

Each questionnaire contains hundreds of questions, whose responses vary greatly in types. Response types are summarised in Table I. For all datasets, continuous responses are pre-normalised across samples.

As specified in Section II-F, the eRBM parameters are estimated from each dataset using our method of persistent Markov chains. Start from small but random parameters, the method gradually improves the data likelihood until a local maximum is reached. Figure 2 shows the learning curve for the GA02, where the data log-likelihood is monitored.

The eRBM is evaluated against the baseline, which is essentially the eRBM without the binary hidden layer, i.e., by assuming that variables are independent. Performance metrics are type-specific normalised error rates. In particular, let \( y \) be the user index, \( \hat{v}_i \) be the predicted value of the \( i \)-th variable, and \( N_t \) is the number of variables of type \( t \) in the test data, we compute the prediction errors as follows:

\( \text{Error rate} = \frac{1}{N_t} \sum (\hat{v}_i - v_i)^2 \)

\( \text{Normalised error rate} = \frac{\text{Error rate}}{\text{Baseline Error rate}} \)

\( \text{Baseline Error rate} = \frac{1}{N_t} \sum (v_i - \mu_i)^2 \)
Table II
ERROR RATES WHEN RECONSTRUCTING HEH11 FROM POSTERIORS.
THE BASELINE IS ESSENTIALLY THE eRBH WITHOUT HIDDEN LAYER
(IE., ASSUMING RESPONSES ARE INDEPENDENT). (*) THE ERROR IS
RELATIVE TO THE EMPIRICAL STANDARD DEVIATION.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>K = 20</th>
<th>K = 50</th>
<th>K = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin.</td>
<td>0.285</td>
<td>0.174</td>
<td>0.149</td>
<td>0.128</td>
</tr>
<tr>
<td>Cat.</td>
<td>0.499</td>
<td>0.306</td>
<td>0.232</td>
<td>0.195</td>
</tr>
<tr>
<td>Multicat.</td>
<td>0.363</td>
<td>0.245</td>
<td>0.182</td>
<td>0.146</td>
</tr>
<tr>
<td>Cont(*)</td>
<td>1.002</td>
<td>0.749</td>
<td>0.606</td>
<td>0.585</td>
</tr>
<tr>
<td>Ord.</td>
<td>0.264</td>
<td>0.170</td>
<td>0.139</td>
<td>0.133</td>
</tr>
<tr>
<td>Count</td>
<td>0.382</td>
<td>0.338</td>
<td>0.317</td>
<td>0.350</td>
</tr>
</tbody>
</table>

The resulting representation can be used for other tasks such as clustering, prediction and visualisation. Here we first learn the eRBH from a subset of the GA02 dataset with \( K = 50 \) hidden units. The obtained representation is further projected onto a 2D plane using a recent visualisation tool known as t-SNE [20]. t-SNE estimates the XY coordinates of each data point so that the \textit{relative distances} between points are probabilistically preserved.

The opinions from 21 countries are presented in Figure 3. This clearly shows the nation-wise clustering property, and how cultures may be related in their views in the context of 2002, shortly after the world-shaking September 11th attack in 2011. We can see there are groups of Islam-dominant nations (Pakistan, Turkey and Indonesia), and their distance from the US at the time. Although being an Islam-dominant nation, Egypt stands out, possibly due to its different historical heritages. It is not surprising that the US and its European fellows join a big group but the US slightly stays separated. Although sitting next to the US, Canada is not as close to it as Germany. Instead, it is more close to France, suggesting that the culture link may play the role here. China, Russia and Vietnam claim a group each, but it is interesting to see how they depart from the US.

C. Response Imputation

Missing answers occur frequently in real survey data, and thus it may be beneficial to make a best guess\(^\text{10}\). For evaluation, we randomly remove a portion \( \rho \in (0, 1) \) of

\(^9\)Note that here we refer to how citizens of nations think, not the choice of their political leaders.

\(^10\)We note in passing that this subsumes the standard collaborative filtering problem as a special case.
answers for each person\(^{11}\), then train the eRBM on the remaining answers. Using Eq. (13), the missing answers are then predicted and evaluated against known answers (see Table II).

**D. Learning Predictive Models**

Predictive models attempt to uncover the functional relationship between a set of input responses and output responses. Table IV reports the results for six representative data types on the GA02 dataset: (i) satisfaction with the country (binary), (ii) country of origin (categorical, size of 44), (iii) problems facing the country (multicategorical, size of 11), (iv) age of the person (continuous), (v) ladder of life (ordinal, size of 11), and (vi) rank of dangers of the world (category-ranking, size of 5). The models are trained on answers by 80% randomly selected respondents and tested on the rest.

**IV. RELATED WORK**

The survey analysis literature offers a rich set of tools to model and infer about complex data [7]. However, most existing methods are limited to understanding univariate data, or correlation between a few variables. In statistics, there has been a moderate amount of work addressing mixed data types [2], [3], [4], [10], [12], [13], [16], [18]. There are two general approaches to this problem: one is to specify conditional distribution of one type given another, and another is to assume underlying latent variables for each type. Our work adopts the latter. However, compared to existing work, ours is more extensive as it addresses seven most common types, as opposed to existing combinations of two or three types. Second, the use of RBM, a machinery from the area of AI, for factorising the correlation structure is novel.

The use of RBMs for data processing has been popular in recent years, possibly due to the recent advances in efficient learning and inference. However, most work is limited to single types such as binary [5] and continuous [8], categorical [15], ordinal [19] and count [6]. The only known RBM-based work addressing the mixed data type is [18] but like all previous RBM-based models, it does not model the underlying utilities that generate the observed responses. Thus their models are hardly interpretable in social analysis.

**V. Conclusion**

We have introduced a probabilistic framework called Embedded Restricted Boltzmann Machines (eRBM) for fusing multiple data types, which can be any combination of binary, categorical, multicategorical, ordinal, continuous, count and category-ranking types. The key feature is the uniform use of latent variables to model the utilities which the user perceive certain choices. Thus the correlation structure among utilities reflects that among responses. The correlation structure is further factorised by introducing an additional hidden binary layer on top of the utility layer, creating a bipartite embedded network. The model is efficient to support a variety of large-scale data analysis tasks including distribution estimation, data completion, prediction and data visualisation. The model is highly suitable for analysing social measurements such as those in surveys, most of which are of qualitative and subjective nature. We have demonstrated the effectiveness of our eRBM on large surveys in the US and world-wide.

We have applied our model for multiple responses but from a single source. Future work will include applications from multiple, diverse sources, e.g., information collected from multiple social networks.

**APPENDIX**

**A. Global Attitude 2002 (GA02)**

Topics covered: 

- Q1 (Ordinal): How would you describe your day today—{bad, typical, good}?  
- Q7 (Binary): [...] Are you satisfied or dissatisfied with the way things are going in our country today?  
- Q10,11 (Category-ranking): In your opinion, which one of these poses the greatest/second greatest threat to the world: {a list of threats}?  
- Q74 (Continuous): How old were you at your last birthday?  
- Q91 (Categorical): Are you currently married or living with a partner, widowed, divorced, separated, or have you never been married?

\(^{11}\)This simulates the so-called “missing at random” assumption since we do not known the missingness mechanism.

<table>
<thead>
<tr>
<th>Bin.</th>
<th>K = 10</th>
<th>K = 20</th>
<th>K = 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.327</td>
<td>0.260</td>
<td>0.243</td>
<td>0.240</td>
</tr>
<tr>
<td>0.565</td>
<td>0.499</td>
<td>0.477</td>
<td>0.459</td>
</tr>
<tr>
<td>1.024</td>
<td>0.930</td>
<td>0.921</td>
<td>0.910</td>
</tr>
<tr>
<td>0.393</td>
<td>0.225</td>
<td>0.216</td>
<td>0.211</td>
</tr>
</tbody>
</table>

**Table III**

**IMPUTATION ERROR RATES ON GA08.** ON AVERAGE, THERE ARE \(p = 0.2\) ANSWERS MISSING AT RANDOM. \(^{(*)}\) SEE TABLE II.

<table>
<thead>
<tr>
<th>Bin.</th>
<th>K = 10</th>
<th>K = 20</th>
<th>K = 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.274</td>
<td>0.238</td>
<td>0.222</td>
<td>0.249</td>
</tr>
<tr>
<td>0.920</td>
<td>0.670</td>
<td>0.563</td>
<td>0.395</td>
</tr>
<tr>
<td>0.484</td>
<td>0.487</td>
<td>0.446</td>
<td>0.448</td>
</tr>
<tr>
<td>1.090</td>
<td>0.922</td>
<td>0.907</td>
<td>0.873</td>
</tr>
<tr>
<td>0.178</td>
<td>0.155</td>
<td>0.162</td>
<td>0.162</td>
</tr>
<tr>
<td>0.312</td>
<td>0.272</td>
<td>0.265</td>
<td>0.266</td>
</tr>
</tbody>
</table>

**Table IV**

**PREDICTIVE ERROR RATES ON GA02 WITH 80/20 TRAIN/TEST SPLIT.** TYPE-SPECIFIC RESPONSES ARE: SATISFACTION (BIN.), COUNTRY (CAT.), PROBLEMS (MULTICAT.), AGE (CONT.), LIFE LADDER (ORD.) AND WORLD-DANGERS (RANK). \(^{(*)}\) SEE TABLE II.
• Q5 (Multicategorical): What do you think is the most important problem facing you and your family today {a list of problems}?

B. Global Attitude (GA08)

The goal is similar to that of GA02, but carried out 6 year later. Sample questions and their corresponding data types:

• Q4 (Ordinal): [...] how would you describe the current economic situation in (survey country) – {very good, somewhat good, somewhat bad, or very bad}?

• Q11a (Binary): How do you think people in other countries of the world feel about China? – {like, disliked}?

• Q35, 35a (Category-ranking): Which one of the following, if any, is hurting the world’s environment the most/second-most {India, Germany, China, Brazil, Japan, United States, Russia, Other}?

• Q76 (Continuous): How old were you at your last birthday?

• Q85 (Categorical): What is your current employment situation {A list of employment categories}?

C. Higher Education/Housing (HEH11)

Sample questions and their types:

• Q12 (Ordinal): [...] how important you think [a college education] is in helping a young person succeed in the world today {Extremely important, Very important, Somewhat important, Not too important}?

• RSCHL (Binary): Do you ever plan to return to school?

• AGE (Continuous): What is your age?

• PARTY (Categorical): In politics today, do you consider yourself a Republican, Democrat, or Independent?

• Q34a (Multicategorical): What was your major field of study in college? {more than one answers}

• HH1 (Count): How many people, including yourself, live in your household?

REFERENCES


