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, modeling 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.

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 along 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

[1] pewglobal.org

[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 a new problem of l*earning 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 unseen data. These versatile capacities are demonstrated on one moderate US-based and two large-scale world-wide surveys by the PewResearch Centre[3].

II. Embedded Restricted Boltzmann Machines for Fusing Mixed Responses

Denote by $\boldsymbol{v} = (v_1, v_2, ..., v_N)$ the set of observed mixed responses, each of which is represented by a type-specific *submodel.* In this paper, we consider *seven* data types: *binary*, *categorical*, *multicategorical, ordinal*, *continuous, count* and *rank* responses. Our modelling goal is to fully specify the joint distribution $P(\boldsymbol{v})$. Obviously, the key is to correctly represent each type (and modality), and at the same time, capture the correlation among types in an efficient way.

A. Model Structure

We assume that each observed response v_i is *independently* generated from a subset of underlying latent *utilities* $u_i \in \mathbb{R}^{D_i}$. The utilities reflect how the user perceives the value of the response choices. Here D_i is the dimensionality of the response, and its value depends on the assumption we make about the response type. For example, in the case of continuous responses, the utility is typically one dimensional, i.e., $D_i = 1$. In the categorical responses, the user needs to consider a set of choices, each of which may has a different utility to the user, and thus D_i equals the number of categories. The generation of response given its utility $P_i(v_i \mid u_i)$ is response-specific. For example, in the case of categorical responses, the user will make a choice whose utility is maximum among those of all possible choices. Finally, we wish to emphasize that the utilities are not observed, and we can only make inference about them conditioned the observed responses, e.g. through the posterior $P(u_i \mid \boldsymbol{v})$.

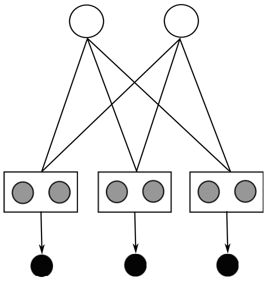

Figure 1. Model architecture. Filled nodes represent observed responses, shaded nodes are latent utilities capturing the type-specific nature of data, and empty nodes represent binary hidden factors. The top two layers form a bipartite Markov Random Field known as Restricted Boltzmann Machine [17], [5], [8].

The correlation among responses $\boldsymbol{v} = (v_1, v_2, ..., v_N)$ is captured through the correlation among corresponding utilities $\boldsymbol{u} = (u_1, u_2, ..., u_N)$, i.e., by using $P(\boldsymbol{v}) = \sum_{\boldsymbol{u}} P(\boldsymbol{v} \mid \boldsymbol{u})P(\boldsymbol{u})$. However, specifying and estimating $P(\boldsymbol{u})$ from data can be highly challenging for large N, which can be several hundreds. For example, the typical multivariate Gaussian approach to model $P(\boldsymbol{u})$ is not scalable with N. Second, if we can infer $P(\boldsymbol{u} \mid \boldsymbol{v})$, this is of limited practical use since the dimensionality of $\boldsymbol{u}$ 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 $\boldsymbol{h} \in \{0, 1\}^K$ on top of the utility layer $\boldsymbol{u}$. 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 $\Psi(\boldsymbol{u}, \boldsymbol{h}) \geq 0$ be the model potential function cap-

[3] pewresearch.org

turing the correlation structure among variables. The RBM architecture specifies that we can factorise the potential function as

$$\Psi(\boldsymbol{u},\boldsymbol{h}) = \left[\prod_i \phi_i(u_i)\right]\left[\prod_{ik}\psi_{ik}(u_i,h_k)\right]\left[\prod_k \phi_k(h_k)\right] \tag{1}$$

where ϕ_i, ψ_{ik} and ϕ_k are local potential functions. Finally, the model distribution is defined as

$$P(\boldsymbol{v},\boldsymbol{u},\boldsymbol{h}) = \frac{1}{Z}\Psi(\boldsymbol{u},\boldsymbol{h})P(\boldsymbol{v}\mid\boldsymbol{u}) \tag{2}$$

where $P(\boldsymbol{v} \mid \boldsymbol{u}) = \prod_i P_i(v_i \mid u_i)$ and $Z = \int_{\boldsymbol{u}}\sum_{\boldsymbol{h}}\Psi(\boldsymbol{u},\boldsymbol{h})d\boldsymbol{u}$ is the *finite* normalising constant[4].

The factorisation in Eq. (1) indeed offers nice decomposition of conditional distributions

$$\begin{aligned} P(\boldsymbol{u}\mid\boldsymbol{h}) &= \prod_i P(u_i\mid\boldsymbol{h}) & (3)\\ P(\boldsymbol{u}\mid\boldsymbol{v},\boldsymbol{h}) &= \prod_i P(u_i|v_i,\boldsymbol{h}) & (4)\\ P(\boldsymbol{h}\mid\boldsymbol{u},\boldsymbol{v}) &= P(\boldsymbol{h}\mid\boldsymbol{u}) = \prod_k P(h_k\mid\boldsymbol{u}) & (5)\end{aligned}$$

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 $(\boldsymbol{h},\boldsymbol{u})$. For simplicity, we assume that the latent layer $\boldsymbol{u}$ follows the multivariate Gaussian distribution[6]. The choice of Gaussian is for convenient 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

$$\begin{aligned} \phi_i(u_i) &= \exp\left\{\sum_{d=1}^{D_i}\left(-\frac{u_{id}^2}{2\sigma_i^2} + \alpha_{id}u_{id}\right)\right\} & (6)\\ \psi_{ik}(u_i,h_k) &= \exp\left\{\sum_{d=1}^{D_i} w_{idk}u_{id}h_k\right\} & (7)\\ \phi_k(h_k) &= \exp\{\beta_k h_k\} & (8)\end{aligned}$$

where σ_i is the standard deviation of the i-th variable, $\{\alpha_{id},\beta_k,w_{idk}\}$ 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

$$\begin{aligned} P(u_i\mid\boldsymbol{h}) &\propto \phi_i(u_i)\prod_k \psi_{ik}(u_i,h_k)\\ &= \exp\left\{\sum_{d=1}^{D_i}\left(-\frac{u_{id}^2}{2\sigma_i^2} + \alpha_{id}u_{id} + \sum_k w_{idk}u_{id}h_k\right)\right\}\\ &\propto \prod_{d=1}^{D_i}\mathcal{N}\left(u_{id};\mu_{id}(\boldsymbol{h}),\sigma_i\right) & (9)\end{aligned}$$

i.e., $P(u_{id}\mid\boldsymbol{h})$ is a Gaussian with mean $\mu_{id}(\boldsymbol{h})$ and standard deviation σ_i, where

$$\mu_{id}(\boldsymbol{h}) = \sigma_i^2\left(\alpha_{id} + \sum_{k=1}^{K} w_{idk}h_k\right) \tag{10}$$

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

$$\begin{aligned} P(h_k\mid\boldsymbol{u}) &\propto \phi_k(h_k)\prod_i \psi_{ik}(u_i,h_k)\\ &= \exp\left(\beta_k + \sum_{id} w_{ikd}u_{id}\right)^{h_k} & (11)\end{aligned}$$

which gives rise to the sigmoid form of $P(h_k = 1\mid\boldsymbol{u}) = 1/\left(1+e^{-s(\boldsymbol{u})}\right)$ where $s(\boldsymbol{u}) = \beta_k + \sum_{id} w_{ikd}u_{id}$.

C. Conditional Bernoulli-Gaussian RBMs

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

$$\begin{aligned} P(u_i|v_i,\boldsymbol{h}) &\propto P_i(v_i|u_i)P(u_i|\boldsymbol{h})\\ &= P_i(v_i|u_i)\prod_{d=1}^{D_i}\mathcal{N}\left(u_{id};\mu_{id}(\boldsymbol{h}),\sigma_i\right) & (12)\end{aligned}$$

Clearly the exact form of $P(u_i|v_i,\boldsymbol{h})$ 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 $\hat{\boldsymbol{h}} = (P(h_1|\boldsymbol{v}), P(h_2|\boldsymbol{v}), ..., P(h_K|\boldsymbol{v}))$. This can be estimated by collecting samples of $\{h_k\}$ from a chain where $\boldsymbol{v}$ 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 $P_j(v_j|\boldsymbol{v})$, where $v_j \notin \boldsymbol{v}$. 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 $\hat{\boldsymbol{h}}$ from $P(\boldsymbol{h}\mid\boldsymbol{v})$, then

[4] That is, we assume the integration over $\boldsymbol{u}$ exists.

[5] A Markov blanket of a node in a MRF is a set of neighbour nodes.

[6] In standard multivariate Gaussian models, we need to specify both the mean and covariance structure. This is not a trivial task both in term of modelling and computation. In our treatment, the covariance is factored into simpler components involving hidden units $\{h_k\}_{k=1}^{K}$.

integrating over the utilities (if the integration has closed form solution) as follows

$$P(v_j|\boldsymbol{v}) \approx P(v_j|\hat{\boldsymbol{h}}) = \int_{u_j} P(v_j, u_j|\hat{\boldsymbol{h}})du_j \tag{13}$$

Data likelihood can be estimated in the same way, i.e., $P(v_i) \approx P(v_i|\hat{\boldsymbol{h}})$, where v_i 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 $P_i(v_i \mid u_i)$. 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 *multicategorical* 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 $\mathbb{S}_i = (c_{i1}, c_{i2}, ..., c_{iM_i})$ 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

$$P_i(v_i = 1 \mid u_i) = \mathbb{I}\left[u_{i1} > \theta_i\right]$$

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

This leads to

$$P(u_i|v_i = 1, \boldsymbol{h}) = \mathcal{N}\left(\mu_{i1}(\boldsymbol{h}), \sigma_i\right)\mathbb{I}\left[u_{i1} > \theta_i\right]$$

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$\hat{\boldsymbol{h}}$, we can estimate the probability of the binary output as

$$P(v_i = 1|\hat{\boldsymbol{h}}) = \int_{\theta_{i1}}^{\infty} \mathcal{N}\left(\mu_{i1}(\hat{\boldsymbol{h}}), \sigma_i\right) = 1 - \Phi(\theta_i^*)$$

where $\theta_i^* = \frac{\theta_i - \mu_{i1}(\hat{\boldsymbol{h}})}{\sigma_i}$, 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., $u_i = (u_{i1}, u_{i2}, ..., u_{iM_i})$ where $D_i = M_i$. A categorical choice is assumed to be made by selecting the one with the maximum latent utility, that is, $v_i = c_{d_{max}}$ where $d_{max} = \arg\max_d u_{id}$. This choice model can be captured by:

$$P_i(v_i = c_d \mid u_i) = \mathbb{I}\left[u_{id} > \max_{m \neq d}\{u_{im}\}\right]$$

The maximisation suggests that the utility variables for the ith observable are correlated, and thus the joint distribution $P(u_{i1}, u_{i2}, ..., u_{iM_i} \mid v_i, \boldsymbol{h})$ is not factorisable. Denote by $u_{i\neg m} = u_i \backslash u_{im}$, the conditional distribution $P(u_{im}|v_i, \boldsymbol{h}, u_{i\neg m})$ can be expressed explicitly as follows

$$P(u_{im}|v_i = c_d, \boldsymbol{h}, u_{i\neg m}) \propto \mathcal{N}\left(\mu_{im}(\boldsymbol{h}), \sigma_i\right)\tau_i(u_i, m, d)$$

where the function $\tau_i(u_i, m, d)$ truncates the domain of the normal distribution $\mathcal{N}\left(\mu_{im}(\boldsymbol{h}), \sigma_i\right)$, i.e.,

$$\tau_i(u_i, m, d) = \begin{cases} \mathbb{I}\left[u_{im} \geq \max_{j \neq m}\{u_{ij}\}\right] & m = d \\ \mathbb{I}\left[u_{im} < u_{id}\right] & m \neq d \end{cases}$$

In words, given the observed category c_d, the conditional distribution of the dth element is truncated from below, where the truncation value is the maximum over all other utilities. On the other hand, the conditional distributions of other categories are truncated from above at u_{id}.

3) Ordinal Responses.: An ordinal variable receives individual values from an ordered set $S_i = \{c_{i1} \prec c_{i2} \prec ..., \prec c_{iM_i}\}$ where $\prec$ denotes the order in some sense. For instance, one can describe their present day as one of $\{$*particularly bad* $\prec$ *typical* $\prec$ *particularly good*$\}$. We assume that there exists an underlying latent utility whose *value intervals* determine the ordinal categories [11]. Translated into the generative distribution $P_i(v_i \mid u_{i1})$, we have $P_i(v_i = c_d \mid u_{i1}) =$

$$\begin{cases} \mathbb{I}\left[u_{i1} < \theta_{i1}\right] & d = 1 \\ \mathbb{I}\left[\theta_{i(d-1)} < u_{i1} \leq \theta_{id}\right] & d \leq M_i - 1 \\ \mathbb{I}\left[\theta_{i(M_i-1)} < u_{i1}\right] & d = M_i \end{cases}$$

where $\theta_{i1} < \theta_{i2} < ... < \theta_{i(M_i-1)}$ are threshold parameters. Substituting this into Eq. (12) yields

$$P(u_{i1}|v_i = c_d, \boldsymbol{h}) \propto \mathcal{N}\left(\mu_{i1}(\boldsymbol{h}), \sigma_i\right)P_i(v_i = c_d \mid u_{i1})$$

In other words, the utility distribution is the normal distribution truncated from above if the first category is chosen, from below if the last category is chosen, and from both sides otherwise.

Given an estimate of hidden factors $\hat{\boldsymbol{h}}$, we can efficiently compute the probability that a particular ordinal level is selected. This is equivalent to computing the probability that the utility belongs to the corresponding interval.

$$P(v_i = c_d|\hat{\boldsymbol{h}}) = \begin{cases} \Phi\left(\theta_{i1}^*\right) & d = 1 \\ \Phi\left(\theta_{id}^*\right) - \Phi\left(\theta_{i(d-1)}^*\right) & d \leq M_i - 1 \\ 1 - \Phi(\theta_{i(M_i-1)}^*) & d = M_i \end{cases}$$

where $\theta_{id}^* = \frac{\theta_{id} - \mu_{i1}(\hat{\boldsymbol{h}})}{\sigma_i}$, and $\Phi(\cdot)$ is the CDF.

4) Count Responses.: We assume counts are distributed as Poisson variables, i.e., they obey the following form

$$P(v_i = r \mid \lambda_i) = \frac{1}{r!}\exp\left(r\log\lambda_i - \lambda_i\right)$$

where $\lambda_i > 0$ is the rate. We maintain one latent utility per Poisson variable, i.e., $u_i = (u_{i1})$, and the rate is assumed to be $\lambda_i = e^{cu_i}$ for some positive constant $c > 0$.

The joint distribution between the Poisson response and its underlying utility is then $P(v_i = r, u_i \mid \boldsymbol{h}) \propto$

$$\frac{1}{r!}\exp\left(-\frac{1}{2\sigma_i^2}u_i^2 + \left(w_i + \sum_k w_{ik}h_k + cr\right)u_i - e^{cu_i}\right)$$

[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*. 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 μ_i $f(u_i) \approx f(\mu_i) + f''(\mu_i)\frac{(u_i-\mu_i)^2}{2}$, leading to the distribution approximation:

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

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 $P(v_i \mid \boldsymbol{h}) = \int P(v_i, u_i \mid \boldsymbol{h}) du_i$ can be estimated as

$$\begin{aligned} P(v_i = r \mid \boldsymbol{h}) \quad & \propto \quad \frac{1}{r!} e^{f(\mu_i)} \int \exp\left(-\frac{(u_i-\mu_i)^2}{2\kappa_i^2}\right) du_i \\ & = \quad \frac{1}{r!} e^{f(\mu_i)} \sqrt{2\pi\kappa_i^2} \qquad (15) \end{aligned}$$

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

$$P(u_i \mid v_i = r, \boldsymbol{h}) \quad = \quad \frac{P(v_i, u_i \mid \boldsymbol{h})}{P(v_i = r \mid \boldsymbol{h})} \approx \mathcal{N}\left(\mu_i, \kappa_i^2\right)$$

F. Learning with Persistent Markov Chains

We are interested in two learning problems. One is the estimation of the model distribution $P(\boldsymbol{v})$ given the sub-sampled empirical distribution $\tilde{P}(\boldsymbol{v})$. 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 $P(\boldsymbol{v}_c|\boldsymbol{v}_{\neg c})$ for the output responses $\boldsymbol{v}_c$. The latter is somewhat easier than the former since we already know $\boldsymbol{v}_{\neg c}$.

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 $P(\boldsymbol{v})$, the general strategy is to iteratively maximising the data likelihood $P(\tilde{\boldsymbol{v}}) = \sum_{\boldsymbol{h}} \int_{\boldsymbol{u}} P(\tilde{\boldsymbol{v}}, \boldsymbol{u}, \boldsymbol{h})$. 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 $(\boldsymbol{u}, \boldsymbol{h})$ constrained by observed $\tilde{\boldsymbol{v}}$ 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.

Data	*Bin*	*Cat*	*Mcat*	*Cont*	*Ord*	*Count*	*Rank*
GA02	43	12	3	3	125	0	2
GA08	52	124	0	3	165	0	0
HEH11	53	32	6	6	4	3	0

Table I
Number of responses per type for the three datasets. *Mcat*: multicategorical.

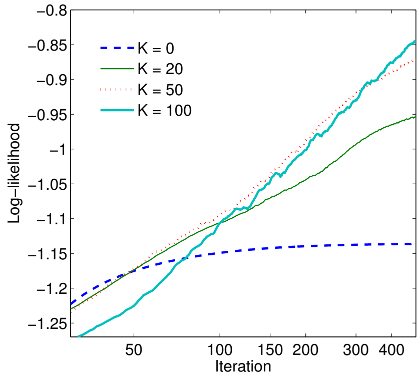

Figure 2. Learning curve on *GA02*.

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:

[8] The datasets are publicly available from http://pewresearch.org/

	Baseline	$K = 20$	$K = 50$	$K = 100$
Bin.	0.285	0.174	0.149	0.128
Cat.	0.499	0.306	0.232	0.195
Multicat.	0.363	0.245	0.182	0.146
Cont.(*)	1.002	0.749	0.606	0.585
Ord.	0.264	0.170	0.139	0.133
Count	0.382	0.338	0.317	0.305

Table II
ERROR RATES WHEN RECONSTRUCTING *HEH11* FROM POSTERIORS. THE BASELINE IS ESSENTIALLY THE eRBM WITHOUT HIDDEN LAYER (I.E., ASSUMING RESPONSES ARE INDEPENDENT). (*) THE ERROR IS RELATIVE TO THE EMPIRICAL STANDARD DEVIATION.

–Binary : $\frac{1}{N_{bin}} \sum_y \sum_i \mathbb{I}\left[v_i^{(y)} \neq \hat{v}_i^{(y)}\right]$

–Categorical : $\frac{1}{N_{cat}} \sum_y \sum_i \mathbb{I}\left[v_i^{(y)} \neq \hat{v}_i^{(y)}\right]$

–Multicategorical : $1 - {}^{2\mathrm{RP}}/_{(\mathrm{R+P})}$,

–Continuous : $\sqrt{\frac{1}{D_{cont}} \sum_y \sum_i \left(v_i^{(y)} - \hat{v}_i^{(y)}\right)^2}$

–Ordinal : $\frac{1}{N_{ord}} \sum_y \sum_i \frac{1}{M_i - 1} \left|v_i^{(y)} - \hat{v}_i^{(y)}\right|$

–Category-ranking :

$$\frac{1}{D_{rank}} \sum_y \sum_i \frac{2}{M_i(M_i-1)} \sum_{l,m>l} E_{lm}$$

$$\text{where} \quad E_{lm} = \mathbb{I}\left[(\pi_{il}^{(y)} - \pi_{im}^{(y)})(\hat{\pi}_{il}^{(y)} - \hat{\pi}_{im}^{(y)}) < 0\right]$$

–Count : $\frac{1}{N_{count}} \sum_y \sum_i \frac{1}{\bar{v}_i} \left|v_i^{(y)} - \hat{v}_i^{(y)}\right|$

where $\mathbb{I}[\cdot]$ is the identity function, $\pi_{im} \in \{1, 2, ..., M_i\}$ is the rank of the m-th category of the i-th variable, $\bar{v}_i$ is the mean of the i-th variable, R is the recall rate and P is the precision. The recall and precision are defined as:

$$\mathrm{R} = \frac{\sum_u \sum_i \frac{1}{M_i} \sum_{m=1}^{M_i} \mathbb{I}\left[a_{im}^{(y)} = \hat{a}_{im}^{(y)}\right]}{\sum_u \sum_i \frac{1}{M_i} \sum_{m=1}^{M_i} a_{im}^{(y)}}$$

$$\mathrm{P} = \frac{\sum_u \sum_i \frac{1}{M_i} \sum_{m=1}^{M_i} \mathbb{I}\left[a_{im}^{(y)} = \hat{a}_{im}^{(y)}\right]}{\sum_u \sum_i \frac{1}{M_i} \sum_{m=1}^{M_i} \hat{a}_{im}^{(y)}}$$

where $a_{im} \in \{0, 1\}$ is the m-th component of the i-th multicategorical variable. Note that the summation over i for each type only consists of relevant variables.

B. *Factor Extraction and Visualisation*

Our eRBM transforms a multimodal input $\boldsymbol{v}$ into a real-valued posterior vector $\hat{\boldsymbol{h}} = \left(\hat{h}_1, \hat{h}_2, ..., \hat{h}_K\right)$, where $\hat{h}_k = P(h_k = 1 \mid \boldsymbol{v})$. To quantify the representation faithfulness, we reconstruct the original data using $\hat{v}_i = \arg\max_{v_i} P\left(v_i \mid \hat{\boldsymbol{h}}\right)$ as in Eq. (13). The reconstruction errors for the *HEH11* dataset are reported in Table II, where it is clear that with more hidden units, the model approximate the data better.

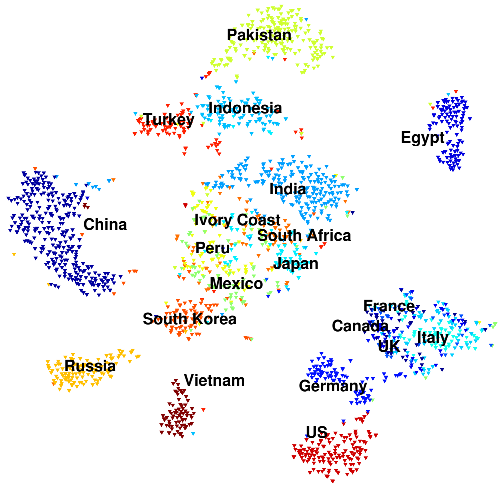

Figure 3. t-SNE projection of *GA02* posteriors ($K = 50$) with country information removed. Each point is a person from one of the 21 selected countries. Each colour represents a country. Best viewed in colour.

The resulting representation can be used for other tasks such as clustering, prediction and visualisation. Here we first learn the eRBM 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 *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[9]. 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[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.

	Baseline	$K = 10$	$K = 20$	$K = 50$
Bin.	0.327	0.260	0.243	0.240
Cat.	0.565	0.499	0.477	0.459
Cont.(*)	1.024	0.930	0.921	0.910
Ord.	0.393	0.225	0.216	0.211

Table III
IMPUTATION ERROR RATES ON *GA08*. ON AVERAGE, THERE ARE $\rho = 0.2$ ANSWERS MISSING AT RANDOM. (*) SEE TABLE II.

	Baseline	$K = 10$	$K = 20$	$K = 50$
Bin.	0.274	0.238	0.222	0.249
Cat.	0.920	0.670	0.563	0.395
Multicat.	0.484	0.487	0.446	0.448
Cont.(*)	1.060	0.922	0.907	0.873
Ord.	0.178	0.155	0.162	0.162
Rank	0.312	0.272	0.265	0.266

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.

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 [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 of use 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: "[...] rapid pace of change in modern life; global interconnectedness through trade, foreign investment and immigration; [...] democracy and governance, [...] economic globalization, the reach of multinational corporations to terrorism, and the U.S. response". Sample questions and their corresponding data types:

- **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.

- **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?
- **Q.34a** (*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

[1] AR de Leon and K. Carrière Chough. Mixed-outcome data. *Encyclopedia of Biopharmaceutical Statistics*, 2010.

[2] AR de Leon, A. Soo, and T. Williamson. Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistics*, 38(5):1021–1032, 2011.

[3] D.B. Dunson and A.H. Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11, 2005.

[4] G.M. Fitzmaurice and N.M. Laird. Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, pages 110–122, 1997.

[5] Y. Freund and D. Haussler. Unsupervised learning of distributions on binary vectors using two layer networks. *Advances in Neural Information Processing Systems*, pages 912–919, 1993.

[6] P.V. Gehler, A.D. Holub, and M. Welling. The rate adapting poisson model for information retrieval and object recognition. In *Proceedings of the 23rd international conference on Machine learning*, pages 337–344. ACM New York, NY, USA, 2006.

[7] S. Heeringa, B.T. West, and P.A. Berglund. *Applied survey data analysis*. Taylor & Francis, 2010.

[8] G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[9] S.L. Lauritzen. *Graphical Models*. Oxford Science Publications, 1996.

[10] R.J.A. Little and M.D. Schluchter. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72(3):497–512, 1985.

[11] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142, 1980.

[12] J.S. Murray, D.B. Dunson, L. Carin, and J.E. Lucas. Bayesian Gaussian copula factor models for mixed data. *Arxiv preprint arXiv:1111.0317*, 2011.

[13] I. Olkin and RF Tate. Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 32(2):448–465, 1961.

[14] M. Opper and D. Saad. *Advanced mean field methods: Theory and practice*. Massachusetts Institute of Technology Press (MIT Press), 2001.

[15] R. Salakhutdinov, A. Mnih, and G. Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 791–798, 2007.

[16] M.D. Sammel, L.M. Ryan, and J.M. Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678, 1997.

[17] P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1:194–281, 1986.

[18] T. Tran, D.Q. Phung, and S. Venkatesh. Mixed-variate restricted Boltzmann machines. In *Proc. of 3rd Asian Conference on Machine Learning (ACML)*, Taoyuan, Taiwan, 2011.

[19] T.T. Truyen, D.Q. Phung, and S. Venkatesh. Ordinal Boltzmann machines for collaborative filtering. In *Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Canada, June 2009.

[20] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[21] L. Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82(4):625–645, 1989.