Zero-Inflated Exponential Family Embeddings

Li-Ping Liu [1] [2] **David M. Blei** [1]

Abstract

Word embeddings are a widely-used tool to analyze language, and exponential family embeddings (Rudolph et al., 2016) generalize the technique to other types of data. One challenge to fitting embedding methods is sparse data, such as a document/term matrix that contains many zeros. To address this issue, practitioners typically downweight or subsample the zeros, thus focusing learning on the non-zero entries. In this paper, we develop zero-inflated embeddings, a new embedding method that is designed to learn from sparse observations. In a zero-inflated embedding (ZIE), a zero in the data can come from an interaction to other data (i.e., an embedding) or from a separate process by which many observations are equal to zero (i.e. a probability mass at zero). Fitting a ZIE naturally downweights the zeros and dampens their influence on the model. Across many types of data—language, movie ratings, shopping histories, and bird watching logs—we found that zero-inflated embeddings provide improved predictive performance over standard approaches and find better vector representation of items.

1. Introduction

Word embeddings use distributed representations to capture usage patterns in language data (Harris, 1954; Rumelhart et al., 1988; Bengio et al., 2003; Mikolov et al., 2013a;b; Pennington et al., 2014). The main idea is to fit the conditional distribution of words by using vector representations, called embeddings. The learned parameters—the embedding vectors—are useful as features about the meanings of words. Word embeddings have become a widely used method for unsupervised analysis of text.

In a recent paper, Rudolph et al. (2016) developed *exponential family embeddings*. Their work casts embeddings in a probabilistic framework and generalizes them to model various types of high-dimensional data.

Exponential family embeddings give a recipe for creating new types of embeddings. There are three ingredients. First is a notion of a context, e.g., a neighborhood of surrounding words around a word in a document. Second is a conditional distribution of data given its context, e.g., a categorical distribution for a word. Third is an "embedding structure" that captures parameter sharing, e.g., that the embedding vector for PHILOSOPHY is the same wherever it appears in the data. Rudolph et al. (2016) show that exponential family embeddings embody many existing methods for word embeddings. Further, they easily extend to other scenarios, such as movie ratings, shopping basket purchases, and neuroscience data.

Many applications of embeddings—both the classical application to language and the others types of data—involve *sparse observations*, data that contain many zero entries. As examples, shoppers do not purchase most of the items in the store, authors do not use most of the words in a vocabulary, and movie-watchers do not view most of the movies in an online collection. Sparse observations are challenging for many machine learning methods because the zeros dominate the data. Most methods will focus on capturing and predicting them, and embeddings are no exception.

Folk wisdom says that zeros in sparse data contain less information than the non-zeros. Consequently, practitioners use various methods to downweight them (Hu et al., 2008) or downsample them, as is often the case in word embeddings (Mikolov et al., 2013b; Rudolph et al., 2016). Empirically, methods that downweight and downsample the zeros far outperform their counterparts.

What this folk wisdom suggests is that zeros often occur for one of two reasons—either they are part of the underlying process that we are trying to model, such as capturing the meanings of words, or they part of a different process, such as that only a particular part of speech belongs in a particular location in a sentence. As other examples, a film enthusiast might not watch a movie either because she does not think she will like it or because she has never heard of it; a shopper might not buy a brand of cookies either be-

[1] Columbia University, 500 W 120th St., New York, NY 10027 [2] Tufts University, 161 College Ave., Medford, MA 02155. Correspondence to: Li-Ping Liu <ll3105@columbia.edu>, David M. Blei <david.blei@columbia.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

cause he doesn't like them or because he didn't see them. Motivated by this intuition, we develop *zero-inflated embeddings*, a probabilistic embedding model that captures the special status of the zeros in the data matrix.

The main idea is as follows. Exponential family embeddings model the conditional distribution of each data point given its context, where the parameter to that distribution relates to the embedding vectors. In a zero-inflated embedding, the conditional distribution places extra probability mass on zero, capturing conditions other than the embedding under which an item might not appear. If an observed zero is explained by this extra probability, then the corresponding item is not *exposed* to its relation with other items.

The probability of seeing a zero that is *not* from the embedding model can be a fixed quantity or can depend on other properties, such as the popularity of an item, demographics about the shopper, or the parts of speech of the context words. While zero-inflated embeddings sometimes fall into the class of an exponential family embedding (as in the Bernoulli case), they sometimes reflect a more complex distribution.

The practical effect is that zero-inflated embeddings intelligently downweight the zeros of the data—the embeddings no longer need to explain all of the zeros—and empirically improve the learned representations of the words or other type of data. We will demonstrate zero-inflated embeddings on language, movie ratings, and shopping baskets, as described above. We also study zero-inflated embeddings on bird watching logs (Munson et al., 2015), where features like time and location can influence which birds are possible to see.

Below, we develop zero-inflated embeddings and show how we can flexibly define the "exposure model" alongside the exponential family embedding model. We derive two algorithms to fit them and study them on a variety of data sets. Zero-inflated embeddings improve performance in language, shopping histories, recommendation system data, and bird watching logs.

Related work. The main thread of work on downweighting zeros comes from recommendation systems, where absent user-item interaction can be misconstrued as a user disliking an item. Hu et al. (2008) and Rendle et al. (2009) proposed to manually down-weight zeros and showed excellent empirical performance. Liang et al. (2016) builds on this work and introduces an exposure model to explain zero entries. The exposure model captures whether a user does not see an item or intentionally chooses not to consume it. In a sense, zero-inflated embeddings build on Liang et al. (2016), using a type of "exposure model" in the context of embeddings and capturing item-item interactions. They also share similarities with the spike and slab model (Mitchell & Beauchamp, 1988) and zero-inflated regression models (Lambert, 1992).

2. Zero-Inflated Embeddings

We first review exponential family embeddings. We then develop zero-inflated embeddings.

2.1. Exponential Family Embedding

An Exponential Family Embedding (EFE) generalizes word embedding to other types of data. It uses vector representations to capture conditional probabilities of data given a context. Specifically, an EFE aims to learn vector representation of items, that is, to represent each item $j \in \{1, \ldots, J\}$ with an embedding vector $\boldsymbol{\rho}_j \in \mathbb{R}^K$ and a context vector $\boldsymbol{\alpha}_j \in \mathbb{R}^K$. Vectors $\boldsymbol{\rho}_j$ and $\boldsymbol{\alpha}_j$ for all j are denoted as $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$, respectively.

An EFE model learns from observation-context pairs. In each pair $i \in \{1, \ldots, N\}$, the observation $\mathbf{x}_i = \{x_{ij} : j \in s_i\}$ contains values observed from one or multiple items in s_i, and the context $\mathbf{y}_i = \{y_{ij} : j \in c_i\}$ contains the values of related items in a context set c_i.

An EFE is defined by three elements: *the context*, *the conditional distribution*, and *the embedding structure*. We have defined the context. The *conditional distribution* of $\mathbf{x}_i$ given its context $\mathbf{y}_i$ is from the exponential family,

$$\mathbf{x}_i \sim \mathrm{ExpFam}\big(\eta(\mathbf{y}_i, s_i), T(\mathbf{x}_i)\big), \tag{1}$$

where the context $\mathbf{y}_i$ provides the natural parameter through $\eta(\mathbf{y}_i, s_i)$, and $T(\mathbf{x}_i)$ is the sufficient statistics.

The embeddings come into play in the natural parameter,

$$\eta(\mathbf{y}_i, s_i) = f\Big(\boldsymbol{\rho}_{s_i}^\top \sum_{j \in c_i} y_{ij} \boldsymbol{\alpha}_j\Big), \tag{2}$$

where the columns of $\boldsymbol{\rho}_{s_i}$ are embedding vectors of items in s_i and $f(\cdot)$ is a link function. The EFE conditional probability $p(\mathbf{x}_i | \mathbf{y}_i; \boldsymbol{\rho}_{s_i}, \boldsymbol{\alpha}_{c_i})$ specifies the distribution of the values of items in s_i in their context. When there is no ambiguity, we write the probability as $p(\mathbf{x}_i | \mathbf{y}_i)$. By fitting the conditional probability, the embedding model captures the interaction between items in s_i and items in c_i.

The *embedding structure* of EFE decides how the vectors of items are shared by different observation-context pairs. It is essentially the definition of c_i and s_i, which indicate how to attach item indices j-s to pair indices i-s.

Finally, an EFE puts Gaussian prior over $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$. For all j,

$$\boldsymbol{\alpha}_j \sim \mathrm{Gaussian}(\mathbf{0}, \sigma_\alpha^2 I) \tag{3}$$

$$\boldsymbol{\rho}_j \sim \mathrm{Gaussian}(\mathbf{0}, \sigma_\rho^2 I) \tag{4}$$

where I is a K-identity matrix and σ_α^2 and σ_ρ^2 are hyperparameters controlling the variance of the context and embedding vectors.

An EFE infers $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ by maximizing the conditional log likelihood of observations given their contexts. The learned vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ are able to capture the correlation between items and their contexts.

We give two examples. In movie ratings, we can extract observation-context pairs from a person's ratings: the observation $\mathbf{x}_i$ is the rating of a single movie in $s_i, |s_i| = 1$, and the context is the ratings $\mathbf{y}_i$ of all other movies rated by the same person. In language, the observation at a text position is a one-hot vector $\mathbf{x}_i$ indicating which word is there and the context $\mathbf{y}_i$ is the vector representation of words in the context. In this case s_i and c_i are both the entire vocabulary, $\{1, \ldots, J\}$. The original word embedding model (Bengio et al., 2003; Mnih & Hinton, 2007) assumes the conditional distribution to be the multinomial distribution with 1 trial. The word2vec model optimized by negative sampling (NEG) (Mikolov et al., 2013b) uses a product of Bernoulli distributions as the conditional distribution.

2.2. Exposure modeling with zero-inflation

An EFE explains every observation $\mathbf{x}_i$ by an item's interaction with its context. However, as we described in Section 1, we may not want the embeddings to explain every observation, especially when they are dominated by zeros.

A Zero-Inflated Embedding (ZIE) places extra probability mass at zero in the embedding distribution. This mass can be thought of as the probability that the corresponding item is not "exposed" to its interaction with other items. In a ZIE, the embeddings vectors need not capture the (zero) data that are explained by the extra mass.

In more detail, for each observed value x_{ij} we explicitly define an exposure indicator b_{ij} to indicate whether the corresponding item j is exposed to the interaction with context items ($b_{ij} = 1$) or not ($b_{ij} = 0$). Each b_{ij} is a random variable from Bernoulli distribution with probability u_{ij},

$$b_{ij} \sim \text{Bernoulli}(u_{ij}). \tag{5}$$

The exposure indicators and exposure probabilities of the items in the observation $\mathbf{x}_i$ are collectively denoted as $\mathbf{b}_i = \{b_{ij} : j \in s_i\}$ and $\mathbf{u}_i = \{u_{ij} : j \in s_i\}$ respectively.

In many applications, we have the information about the exposure probability u_{ij}. Suppose we have a set of covariates $\mathbf{v}_i \in \mathbb{R}^d$ for each i related to the exposure probability. We fit u_{ij} with a logistic regression,

$$u_{ij} = \text{logistic}(\mathbf{w}_j^\top \mathbf{v}_i + w_j^0), \tag{6}$$

where $\mathbf{w}_j$ are the coefficients and w_j^0 is the intercept. (If there is no such covariates, then only the intercept term is used, which means that the exposure probabilities of items are shared by observation-context pairs.)

Next, we incorporate the exposure indicator into the embedding model. When $\mathbf{x}_i = \{x_{ij}\}$ is an observation with only one item j, the indicator b_{ij} decides whether x_{ij} is zero or from the embedding distribution,

$$x_{ij} \sim \begin{cases} \delta_0 & \text{if } b_{ij} = 0 \\ \text{ExpFam}\big(\eta(\mathbf{y}_i, s_i), T(x_{ij})\big) & \text{if } b_{ij} = 1 \end{cases}. \tag{7}$$

The distribution δ_0 has probability mass 1 at 0.

When $\mathbf{x}_i$ has multiple entries, the indicator vector $\mathbf{b}_i$ decides the exposure of each item separately. The items not exposed have zero values,

$$x_{ij} \sim \delta_0, \quad \forall j \in s_i, b_{ij} = 0. \tag{8}$$

Let the items exposed be $s_i^+ = \{j : j \in s_i, b_{ij} = 1\}$, and their values are $\mathbf{x}_i^+ = \{x_{ij} : j \in s_i, b_{ij} = 1\}$. Then $\mathbf{x}_i^+$ is from a smaller embedding model restricted to s_i^+.

$$\mathbf{x}_i^+ \sim \text{ExpFam}\big(\eta(\mathbf{y}_i, s_i^+), T(\mathbf{x}_i^+)\big). \tag{9}$$

We have $p(\mathbf{x}_i|\mathbf{y}_i, \mathbf{b}_i) = p(\mathbf{x}_i^+|\mathbf{y}_i, \mathbf{b}_i)$ when s_i has either single or multiple items.

Finally, if the exposure probabilities are fit by covariates then each weight vector $\mathbf{w}_j$ is also given a Gaussian prior,

$$\mathbf{w}_j \sim \text{Gaussian}(\mathbf{0}, \sigma_w^2 I), \;\; j = \{1, \ldots, J\}. \tag{10}$$

The identity matrix I has size d here, and σ_w^2 is the hyperparameter.

2.3. Inference

In this subsection, we derive the method of inferring $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ from ZIE. The approach is to maximize the log-likelihood $\sum_i \log p(\mathbf{x}_i|\mathbf{y}_i, \mathbf{u}_i)$ with all hidden variables $\mathbf{b}_i$ marginalized.

In this subsection, we derive the solution with $\mathbf{u}_i$-s instead of $(\mathbf{w}_j, w_j^0)$-s for notational simplicity. Once we can calculate the gradient of each $\mathbf{u}_i$, the gradient calculation of each $(\mathbf{w}_j, w_j^0)$ is straightforward. The probability $p(\mathbf{x}_i^+|\mathbf{y}_i, \mathbf{b}_i)$ is indeed from the basic embedding model, and we denote it as $\hat{p}(\mathbf{x}_i^+|\mathbf{y}_i)$ to explicitly show how the basic embedding model is positioned in the solution.

The inference is easy when the observation $\mathbf{x}_i$ is about one item. We can either use EM to maximize its exact variational lower bound or directly marginalize out the hidden variable. We briefly give both solutions here, as they have different indications of zero inflation.

We first give the EM solution. In E step, we calculate the posterior distribution $p(b_{ij}|x_{ij}, \mathbf{y}_i, u_{ij})$ of each exposure indicator b_{ij}. This distribution is a Bernoulli distribution, and its parameter is denoted as μ_{ij}. Applying the Bayesian rule, we have

$$\mu_{ij} = \begin{cases} \frac{u_{ij}\hat{p}(x_{ij}=0|\mathbf{y}_i)}{1-u_{ij}+u_{ij}\hat{p}(x_{ij}=0|\mathbf{y}_i)} & \text{if } x_{ij} = 0 \\ 1 & \text{if } x_{ij} \neq 0 \end{cases} . \quad (11)$$

The lower bound of the log-likelihood of pair i is

$$E\big[\log p(x_{ij}|\mathbf{y}_i, b_{ij})\big] - \mathrm{KL}(\mu_{ij}, u_{ij}) =$$

$$\begin{cases} \mu_{ij} \log \hat{p}(x_{ij}|\mathbf{y}_i) - \mathrm{KL}(\mu_{ij}, u_{ij}) & \text{if } x_{ij} = 0 \\ \log \hat{p}(x_{ij}|\mathbf{y}_i) + \log u_{ij} & \text{if } x_{ij} \neq 0 \end{cases} . \quad (12)$$

The expectation is taken with respect to $p(b_{ij}|x_{ij}, \mathbf{y}_i, u_{ij})$, and $\mathrm{KL}(\mu_{ij}, u_{ij})$ is the KL-divergence from the prior to posterior of b_{ij}. In M step, the lower bound is maximized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$. Note that, there is no need to take derivative with respect to μ_{ij} when taking gradient steps even though it is a function of $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$, because μ_{ij} already maximizes the lower bound (Hoffman et al., 2013).

Eq. (12) shows that zero-inflation downweights zero entries by μ_{ij} when learning $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$. This method of downweighting zeros is derived in a systematic way instead of a hack.

We can also marginalize b_{ij} directly.

$$p(x_{ij}|\mathbf{y}_i, u_{ij}) =$$

$$\begin{cases} u_{ij}\hat{p}(x_{ij} = 0|\mathbf{y}_i) + (1 - u_{ij}) & \text{if } x_{ij} = 0 \\ u_{ij}\hat{p}(x_{ij}|\mathbf{y}_i) & \text{if } x_{ij} \neq 0 \end{cases} . \quad (13)$$

This equation makes the zero-inflation clearer.

Now let's work on the inference problem when $\mathbf{x}_i$ has values of multiple items. In this case, we need to consider an embedding model $\hat{p}(\mathbf{x}_i^+|\mathbf{y}_i)$ for each configuration of $\mathbf{b}_i$, so there are potentially exponential number of embedding models to consider. We have to exploit the structure of the embedding model to give a tractable MLE problem. In this paper, we consider two special cases, that the embedding models are independent for items in s_i, and that the observation $\mathbf{x}_i$ is from a multinomial distribution.

In the first case, the embedding model is the product of the embedding models with the same context. Word2vec with NEG training is a special case with single models as Bernoulli embedding (Mikolov et al., 2013b).

$$p(\mathbf{x}_i|\mathbf{y}_i, \mathbf{b}_i; \boldsymbol{\rho}_{s_i}, \boldsymbol{\alpha}_{c_i}) = \prod_{j \in s_i} p(x_{ij}|\mathbf{y}_i, b_{ij}; \boldsymbol{\rho}_j, \boldsymbol{\alpha}_{c_i}). \quad (14)$$

The exposure indicators are independent of each other, so the entire model can be decomposed over items in s_i and solved as single models. We omit the detail here.

Now we consider the case when the embedding distribution is multinomial with 1 trial, which is the model assumption of word embedding prior to the proposal of NEG training (Bengio et al., 2003; Mikolov et al., 2013a). The link function $f(\cdot)$ is the identity function, so the vector products in $\eta(\cdot)$ directly give logits of the multinomial distribution. Denote the natural parameter as $\boldsymbol{\eta}_i = \boldsymbol{\rho}_{s_i}^\top \sum_{j \in c_i} y_{ij}\boldsymbol{\alpha}_j$. The probability vector of the multinomial distribution is $\boldsymbol{\pi}_i = \mathrm{softmax}(\boldsymbol{\eta}_i)$.

The logarithm of the joint probability of the model for one pair is

$$\log p(\mathbf{x}_i, \mathbf{b_i}|\mathbf{y}_i) = \log \hat{p}(\mathbf{x}_i^+|\mathbf{y}_i) + \log p(\mathbf{b_i}). \quad (15)$$

Note that the probability $\hat{p}(\mathbf{x}_i^+|\mathbf{y}_i)$ is from the embedding model decided by $\mathbf{b}_i$. To learn the model, we need to maximize the log-likelihood of the data with the hidden variable $\mathbf{b}_i$ marginalized.

To avoid considering exponentially large number of models, we use the following relation between the full embedding model and the one with items exposed only in s_i^+.

$$\log \hat{p}(\mathbf{x}_i^+|\mathbf{y}_i) = \log \hat{p}(\mathbf{x}_i|\mathbf{y}_i) - \log(\mathbf{b}_i^\top \boldsymbol{\pi}_i). \quad (16)$$

Here the last term re-normalizes the probability of $\hat{p}(\mathbf{x}_i|\mathbf{y}_i)$ to get the probability of picking one from these items that are exposed.

Expand $\hat{p}(\mathbf{x}_i|\mathbf{y}_i)$ and cancel the normalizer, then we have

$$\log \hat{p}(\mathbf{x}_i^+|\mathbf{y}_i) = \eta_{ij^*} - \log(\mathbf{b}_i^\top \exp(\boldsymbol{\eta}_i)). \quad (17)$$

Here j^* is the index such that $\mathbf{x}_{ij*} = 1$.

Now we consider the problem of marginalizing $\mathbf{b}_i$ via variational inference. Let $q(\mathbf{b}_i)$ be the variational distribution. Combine Eq. (15) and (17), then the variational lower bound is,

$$\begin{aligned} L_q &= E_{q(\mathbf{b}_i)}\big[\log \hat{p}(\mathbf{x}_i^+|\mathbf{y}_i)\big] - \mathrm{KL}(q(\mathbf{b}_i), p(\mathbf{b}_i)) \\ &= \eta_{ij^*} - E_q\big[\log(\mathbf{b}_i^\top \exp(\boldsymbol{\eta}_i))\big] - \mathrm{KL}(q(\mathbf{b}_i), p(\mathbf{b}_i)). \end{aligned} \quad (18)$$

$\mathrm{KL}(q(\mathbf{b}_i), p(\mathbf{b}_i))$ is the KL-divergence of $p(\mathbf{b}_i)$ from the posterior $q(\mathbf{b}_i)$.

This lower bound often needs to be maximized in the online manner due the large quantity of data, but the expectation of the logarithm is challenging to estimate even with moderate size of s_i. In this work, we find a data-related lower bound of the expectation term.

Let γ be any subset of s_i such that $j^* \in \gamma$ and $|\gamma| = r$, and $\mathbf{b}_{i\gamma}$ be the sub-vector of $\mathbf{b}_i$ indexed by γ. If

$$\frac{\exp(\eta_{ij^*})}{\sum_{j \in s_i} \exp(\eta_{ij})} \geq \frac{r}{|s_i|}, \quad (19)$$

then $\mathbf{b}_i^\top \exp(\boldsymbol{\eta}_i) \le \frac{|s_i|}{r} \mathbf{b}_{i\gamma}^\top \exp(\boldsymbol{\eta}_{i\gamma})$ for any $\mathbf{b}_i$, and then we have another lower bound

$$\max_q L_q \ge \max_{q(\mathbf{b}_{i\gamma})} \; \eta_{ij^*} - E_{q(\mathbf{b}_{i\gamma})}\left[\log(\mathbf{b}_{i\gamma}^\top \exp(\boldsymbol{\eta}_{i\gamma}))\right] + \log\frac{r}{|s_i|} - \mathrm{KL}(q(\mathbf{b}_{i\gamma}), p(\mathbf{b}_{i\gamma})). \quad (20)$$

Here $q(\mathbf{b}_{i\gamma})$ is the marginal of some $q(\mathbf{b}_i)$.

We maximize the objective on the r.h.s. of Eq. (20) with respect to $q(\mathbf{b}_{i\gamma})$ and model parameters. In each iteration of calculation, we randomly sample a subset γ, maximize the lower bound with respect to $q(\mathbf{b}_{i\gamma})$, and calculate the gradient of model parameters. The maximization with respect to $q(\mathbf{b}_{i\gamma})$ is tractable for small r. For larger r, we restrict the form of $q(\mathbf{b}_{i\gamma})$. In our experiment, we set $r = 5$, and let $q(\mathbf{b}_{i\gamma})$ assign zero probability to any $\mathbf{b}_i$ that has more than one zero entry. Then we only need to consider r configurations of $\mathbf{b}_i$: the case that all items in γ are exposed and the cases that only one item, $j \in \gamma, j \neq j^*$, is not exposed, so the maximization with respect to $q(\mathbf{b}_{i\gamma})$ can be calculated efficiently.

The lower bound in Eq. (20) essentially uses $r - 1$ "negative" items in the random set γ to contrast the item j^* in a smaller multinomial distribution. Since the set γ is randomly selected, every $j \neq j^*$ has the chance to be used as a negative sample. The maximization procedure encourages the model to get larger value of η_{ij^*}, and empirically the condition (19) often holds.

2.4. Computation with Subgradient

The data for embedding is often in large amount, so optimization with stochastic gradients is critically important. In problems where s_i has only one item, a pair with nonzero observation often provides more informative gradients than a pair with zero observation. In our optimization, we keep all pairs with non-zero observations and sub-sample zero observations to estimate an unbiased stochastic gradient. The resultant optimization is much faster than that with full gradient.

3. Empirical Study

In this section, we empirically evaluate the Zero-Inflated Embeddings. We compare four models, two baselines and two variants of our model, in the following subsections: 1) EFE is the basic exponential family embedding model; 2) EFE-dz assigns weight 0.1 to zero entries in the training data (same as (Rudolph et al., 2016)); 3) ZIE-0 is the zero-inflated embedding model and fits the exposure probabilities with the intercept term only; and 4) ZIE-cov fits exposure probabilities with covariates.

All models are evluated with four datasets, eBird-PA, MovieLens-100K, Market, and Wiki-S, which will be introduced in detail in the following subsections. Their general information is tabulated in Table 1. The last column lists the number of exposure covariates, which is used by ZIE-cov to fit the exposure probability.

Table 1. Information about datasets.

dataset	# item	# nonzero	sparsity	range	# covar
eBird-PA	213	410k	0.08	$\mathbb{N}$	13
MovieLens-100K	811	78.5k	0.1	0-3	6
Market	7903	737k	10^{-3}	$\mathbb{N}$	20
Wiki-S	10000	365m	10^{-4}	0/1	11

All four models are optimized by AdaGrad (Duchi et al., 2011) implemented in TensorFlow[1], and the AdaGrad parameter η for step length is set to 0.1. One tenth of the training set is separated out as the validation set, whose log-likelihood is used to check whether the optimization procedure converges. The variance parameters of $\boldsymbol{\alpha}$, $\boldsymbol{\rho}$, and $\mathbf{w}$ are set to 1 for all experiments.

We report two types of predictive log-likelihood on the test set, the log-likelihood of all observations (denoted as "all") and that of non-zero entries only (denoted as "pos"). For non-zero entries, the predictive log-likelihood is calculated as $\log p(\mathbf{x}_i | \mathbf{y}_i, \mathbf{x}_i > 0) = \log p(\mathbf{x}_i|\mathbf{y}_i) - \log(1 - p(\mathbf{x}_i = 0|\mathbf{y}_i))$. The predictive log-likelihood is also estimated through sub-sampling in the same way as in training. We use $\boldsymbol{\alpha}$ vectors as embeddings of items.

3.1. Bird embedding from bird observations

In this experiment, we embed bird species into the vector space by studying their co-occurance pattern in bird observations(Munson et al., 2015). The data subset eBird-PA consists of bird observations from a rectangular area that mostly overlaps Pennsylvania and the period from day 180 to day 210 of years from 2002 to 2014. Each datum in the subset is a checklist of counts of 213 bird species reported from one observation event. The values of these counts range from zero to hundreds. Some extraordinarily large counts are treated as outliers and set to the mean of positive counts of that species. Associated with each checklist there are 13 observation covariates, such as effort time, effort distance, and observation time of the day. The dataset is randomly split into two thirds as the training set and one third as the test set.

The embedding model use Poisson distribution to fit the count of each species j given the counts of all other species for each checklist, so $s_i = \{j\}$ and $c_i = \{1, \dots J\} \backslash j$. The link function is $\log \mathrm{softplus}(\cdot)$, which means the Poisson parameter λ is the softplus function of the linear product. The embedding dimension K iterates over the set

[1] https://www.tensorflow.org/

Table 2. ZIE models improve predictive log-likelihood on bird data.

K		EFE	EFE-dz	ZIE-0	ZIE-cov
32	all	$-0.416(.002)$	$-0.555(.001)$	$-0.324(.001)$	$\mathbf{-0.314(.001)}$
	pos	$-2.407(.017)$	$\mathbf{-1.855(.012)}$	$\mathbf{-1.844(.011)}$	$\mathbf{-1.840(.011)}$
64	all	$-0.374(.002)$	$-0.490(.001)$	$-0.308(.001)$	$\mathbf{-0.298(.001)}$
	pos	$-2.140(.016)$	$\mathbf{-1.727(.013)}$	$\mathbf{-1.736(.012)}$	$\mathbf{-1.739(.012)}$
128	all	$-0.348(.002)$	$-0.459(.001)$	$-0.300(.001)$	$\mathbf{-0.291(.001)}$
	pos	$-1.992(.019)$	$\mathbf{-1.681(.015)}$	$\mathbf{-1.705(.015)}$	$\mathbf{-1.708(.015)}$

Table 3. The measures (Δ_p / Δ_λ) of embedded vectors calculated at 4 levels of downsampling (smaller is better). Embedded vectors learned by ZIE models are more resistant to down-sampling.

d.s. ratio	EFE	EFE-dz	ZIE-0	ZIE-cov
0.005	$0.870/0.123$	$0.628/0.172$	$0.463/\mathbf{0.024}$	$\mathbf{0.460}/0.032$
0.01	$1.099/0.947$	$0.649/0.306$	$0.384/\mathbf{0.007}$	$\mathbf{0.351}/0.052$
0.05	$0.373/0.084$	$0.313/0.078$	$\mathbf{0.217}/0.023$	$0.245/\mathbf{0.018}$
0.1	$0.404/0.085$	$0.287/0.067$	$\mathbf{0.212}/0.022$	$0.263/\mathbf{0.017}$

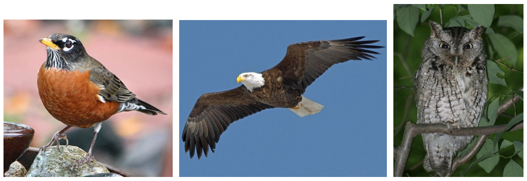

Figure 1. Three bird species in study. Left: American Robin, middle: Bald Eagle, right: Eastern Screech-Owl.

$\{32, 64, 128\}$.

Performance comparison: Table 2 shows the predictive log-likelihoods of the four models with different values of K. The two models with zero-inflation get much better predictive log-likelihood on the entire test set. On positive observations, ZIE models get similar results with the model downweighting zero entries. ZIE-cov performs slightly better than ZIE-0 on all observations and has similar performance with ZIE-0 on positive observations. We have also tried other negative weights (0.05, 0.2) for EFE-dz and found a similar trend: smaller weight gives slightly better predictive log-likelihood on positive observations but worse overall predictive log-likelihood.

Sensitiveness to data sparsity: We down-sample positive observations of one common species, American Robin (Figure 1, left), and test how the embedded vectors changes. Specifically, we randomly set positive counts of American Robin to zero and keep only $r = \{0.005, 0.01, 0.05, 0.1\}$ of positive counts. For each r, we compared the embedded vectors learned from down-sampled data with those learned from the original data.

The vectors are compared by their respectively induced Poisson parameter λ. Let j^* be the index of American Robin, then for each species $j \neq j^*$, the distribution of the count of j given one American Robin has parameter $\lambda_j = \mathrm{softplus}(\boldsymbol{\rho}_j^\top \boldsymbol{\alpha}_{j^*})$.

From the original and down-sampled data, we get two embeded vectors and then have two Poisson parameters $\lambda_j^{\mathrm{orig}}$ and $\lambda_j^{\mathrm{down}}$. The first measure of difference is the symmetric KL divergence (sKL) of the predictive probabilities of presence calculated from the two λ values with Poisson distribution,

$$\Delta_p = \mathrm{sKL}\big(p(x_j > 0; \lambda_j^{orig}), p(x_j > 0; \lambda_j^{down})\big).$$

The second measure is the absolute difference of λ values, $\Delta_\lambda = |\lambda_j^{orig} - \lambda_j^{down}|$. These two measures are averaged over all species.

Table 3 shows these measures calculated from different models with $K = 64$. We can see that the embedding model with exposure explanation is less sensitive to missing observations.

Exposure explanation: We also explore what the exposure probability captures about zero observations. We check the learned coefficients $\mathbf{w}$ of the species Bald Eagle (Figure 1, middle). Its coefficient corresponding to the effort hours of the observation is large, at the percentage of 0.83 of all birds. It agrees with bird watching, since eagles are usually rare and need long time to be spotted. Another example is Eastern Screech-Owl (Figure 1, right), which is only active at night. Its two coefficients corresponding to the observation time at hours 7-12 and hours 12-18 of the day are the smallest among all species. With the learned coefficients, the exposure probabilites are able distinguish the generative process of zeros according to their observation conditions and thus downweight zeros more correctly.

3.2. Movie embedding from movie ratings

In this experiment, we study movie ratings and embed movies as vectors. The MovieLens-100K dataset (Harper & Konstan, 2015) consists of movie ratings from different users. It is preprocessed in the same way as Rudolph et al. (2016): translating ratings above 2 to the range 1-3 and treating absent ratings and ratings no greater than 2 as zeros. We remove all movies with less than 20 ratings. Six covariates, the age, the gender, and four profession categories, are extracted from each user and used as exposure covariates.

For the ratings from the same person, the embedding model fit the rating of one movie given the ratings of all other movies. The embedding distribution is the binomial distribution with 3 trials. The parameter K ranges over $\{8, 16, 32\}$. We run 10 fold cross validation on this dataset.

Performance comparison: In this result, ZIE-0 and ZIE-cov performs much better than the two baselines in predictive performance on all entries. With zero-entries downweighted, EFE-dz is able to predict non-zeros better than

Table 4. ZIE models improves predictive log-likelihood values on movie rating data.

K		EFE	EFE-dz	ZIE-0	ZIE-cov
8	all	$-0.461(.001)$	$-0.740(.001)$	$\mathbf{-0.350(.001)}$	$\mathbf{-0.349(.001)}$
	pos	$-1.870(.007)$	$\mathbf{-1.145(.003)}$	$-1.170(.004)$	$-1.163(.004)$
16	all	$-0.450(.001)$	$-0.706(.001)$	$\mathbf{-0.348(.001)}$	$\mathbf{-0.348(.001)}$
	pos	$-1.795(.007)$	$\mathbf{-1.146(.003)}$	$-1.214(.004)$	$-1.207(.004)$
32	all	$-0.450(.001)$	$-0.669(.001)$	$\mathbf{-0.348(.001)}$	$\mathbf{-0.349(.001)}$
	pos	$-1.758(.007)$	$\mathbf{-1.152(.004)}$	$-1.267(.005)$	$-1.265(.005)$

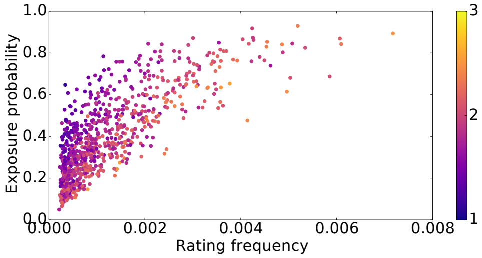

Figure 2. Exposure probability versus rating frequency. Movies with high rating frequency and movies with low ratings tend to get high exposure probability.

ZIE-0 and ZIE-cov, but it is much less capable to predict which entries are zero. ZIE-cov slightly improves over ZIE-0 in predictive log-likelihood of both types.

Exposure probability versus rating frequency: We investigate the exposure probability learned without covariates. We plot the exposure probability of movie versus its frequency of ratings in Fig.2. Each dot represents a movie, its position indicating its exposure probability and rating frequency, and its color being the average of its positive ratings. The exposure probability is from the vector $\mathbf{u}$.

This figure shows that the exposure probability generally correlates with the popularity of the movie. However, some movies with low rating frequency are given high exposure probabilities. These movies have low ratings, which can and should be explained by the embedding component. For example, the movie "Money Train", whose average rating is at the percentage of 0.08 among all movies, has the largest ratio of exposure probability over rating frequency. The movies with high rating frequency get high exposure probabilities no matter their average rating is high or not.

3.3. Product embedding from market baskets

In this experiment, we embed products in grocery stores into the vector space. The Market dataset (Bronnenberg et al., 2008) consists of purchase records of anonymous households in grocery stores. Each record is for a single shopping trip and contains the costumer ID, respective counts of items purchased, and other information. From the dataset, we take 20 related covariates, such as income range, working hours, age group, marital status, and smoking status, to describe each user.

Table 5. ZIE models improves the predictive log-likelihood on Market data

K		EFE	EFE-dz	ZIE-0	ZIE-cov
32	all	$-0.014(.000)$	$-0.024(.000)$	$-0.009(.000)$	$\mathbf{-0.007(.000)}$
	pos	$-4.719(.024)$	$-2.169(.012)$	$-2.195(.012)$	$\mathbf{-1.060(.004)}$
64	all	$-0.014(.000)$	$-0.022(.000)$	$-0.009(.000)$	$\mathbf{-0.008(.000)}$
	pos	$-4.581(.023)$	$-2.030(.011)$	$-2.190(.012)$	$\mathbf{-1.127(.005)}$
128	all	$-0.015(.000)$	$-0.023(.000)$	$-0.009(.000)$	$\mathbf{-0.008(.000)}$
	pos	$-4.746(.023)$	$-2.041(.011)$	$-2.290(.012)$	$\mathbf{-1.295(.005)}$

The embedding model fits the count of one item given counts of other items in the same shopping trip. All embedding models use Poisson distribution as the conditional distribution and $\log \mathrm{softplus}(\cdot)$ as the link function. The covariates of each item-context pair are the covariates of the customer making the shopping trip. This dataset is randomly split into a training set and a test set by 2:1. The number K of embedding dimensions ranges over $\{16, 32, 64\}$.

Performance comparison: On this dataset, we compare the performances by the predictive log-likelihood on the test set. Table 5 shows the predictive log-likelihood of different models. ZIE-cov gives the largest values of predictive log-likelihood of both type. The exposure covariates of this dataset is informative, so the improvement of ZIE-cov is significant.

3.4. Word embedding

In this subsection, we test word embedding with Wikipedia documents. We take the first million documents of the Wikipedia corpus prepared by Reese et al. (2010). The words in these documents have been tagged by FreeLing with part-of-speech (POS) tags. We keep the top 10,000 frequent words as our vocabulary and remove all words not in the vocabulary. The covariates of each word is from the POS tag of the preceding word no matter the word is removed or not. The original FreeLing POS tags has 64 types, and we combine them into 11 larger types, such as noun, verb, and adverb. The covariates are indicators of these 11 types. The subset is further split into a training set and a test set by 2:1.

On this dataset, we test embedding models with Bernoulli distribution (ZIE) and our relaxed multinomial distribution (ZIE-m). ZIE-cov and ZIE-m-cov use the tag type as the exposure covariates. For Bernoulli distributions, zero entries are downweighted by weight 0.0005 (the target word versus 5 negative words). For multinomial distribution, we randomly sample 5 words as the set r for each word-context pair. We use a context window size of 1, so the context of a word position is the two words before and after the position. Note that word2vec with CBOW and NEG training

Table 6. Predictive log-likelihood per document.

K	EFE	ZIE-0	ZIE-cov
32	$-50.1(0.5)$	$-48.0(0.5)$	$\mathbf{-43.5(0.4)}$
64	$-51.1(0.6)$	$-47.2(0.4)$	$\mathbf{-43.1(0.4)}$
128	$-58.5(2.1)$	$-50.0(0.5)$	$\mathbf{-45.5(0.4)}$

Table 7. Performance of different models on word similarity tasks.

	EFE	ZIE-0	ZIE-cov	ZIE-m-0	ZIE-m-cov	GloVe
MEN	0.610	**0.616**	0.592	0.608	0.611	0.737
WS353	0.537	0.533	0.543	0.557	**0.562**	0.522
SIMLEX999	**0.290**	0.281	0.276	0.268	0.264	0.371

Table 8. Nearest words find by three embedding models.

	battle	philosophy	novel	class	english	bible
EFE	combat (0.75)	sociology (0.81)	book (0.85)	division (0.67)	welsh (0.88)	hebrew (0.75)
	defeat (0.73)	theology (0.79)	novels (0.80)	k (0.66)	french (0.87)	study (0.73)
	invasion (0.73)	tradition (0.77)	poem (0.79)	field (0.63)	spanish (0.86)	biblical (0.73)
ZIE-0	fire (0.62)	religion (0.69)	story (0.72)	division (0.56)	french (0.85)	poetry (0.63)
	battles (0.62)	society (0.68)	book (0.71)	family (0.55)	swedish (0.81)	hebrew (0.61)
	assault (0.61)	theology (0.67)	novels (0.70)	classes (0.54)	irish (0.79)	dictionary (0.60)
ZIE-cov	battles (0.61)	principles (0.69)	novels (0.76)	classes (0.56)	french (0.84)	biblical (0.64)
	assault (0.60)	religion (0.68)	story (0.74)	rank (0.54)	spanish (0.83)	hebrew (0.63)
	attack (0.59)	theology (0.67)	fantasy (0.71)	grade (0.53)	swedish (0.75)	texts (0.61)

(Mikolov et al., 2013b) is generally the same as EFE when it takes some settings related to implementation details.

Performance comparison:

Tab. 6 shows the average predictive log-likelihood per document on the test set by Bernoulli embedding models. The embedding model gets the best predictive performance when it uses POS tags to fit the exposure probability.

We also compare all models on three benchmark datasets of word similarities, MEN (Baroni et al., 2014), WS353 (Finkelstein et al., 2002), and SIMLEX999 (Hill et al., 2014), and show the results in Table 7, with each entry being Spearman's correlation of embedding similarities and human-rated similarities. The code is from the repository [2] constructed by Jastrzebski et al. (2017). The last column as a reference shows the performance of GloVe (Pennington et al., 2014) word vectors with dimension 300 trained on 6 billion documents. The results on the three tasks shows that ZIE models perform slightly better than the baseline model.

Word relation learned with exposure explanation: To better understand how the exposure model affects the embedded vectors of words, we check cosine similarities of word vectors $\boldsymbol{\alpha}$ learned by different models. Table 8 shows examples of similar words discovered by Bernoulli embedding models.

We have two interesting observations. First, word embedding with zero-inflation often gives lower similarity scores than the one without. This means the embedded word vectors of ZIE models are more spread out in the space. Second, the distance between a noun word and its plural form often have shorter distance with the vectors learned by ZIE-cov. The difference of a word and its plural form are more in grammar than semantics. The POS tags can partially explain the grammatical difference. For example, numbers (tagged as *number*) often go before a plural instead of a singular word. In such cases, the singular word as a negative example will be downweighted, and thus the embedding component is more likely to treat the singular word and its plural as similar words.

4. Summary

In this work, we have proposed zero-inflated exponential family embedding. With the exposure indicator explaining unrelated zero observations, the real embedding component is able to focus more on observations from item interactions, so the embedded vectors can better represent items in terms of item relations. We have investigated ZIE for two types of embedding models: the observation being from one and multiple items. We have also developed the inference algorithms for different embedding models with zero inflation. Experiment results indicate that ZIE improves the predictive log-likelihood of the data. Qualitative analysis shows that the embedded vectors from the ZIE model have better quality than the vectors learned with the basic embedding model.

Acknowledgements

This work is supported by NSF IIS-1247664, ONR N00014-11-1-0651, DARPA PPAML FA8750-14-2-0009, DARPA SIMPLEX N66001-15-C-4032, and the Alfred P. Sloan Foundation. Thank all reviewers of the paper for their feedback. Thank Francisco J. R. Ruiz, Maja R. Rudolph, Kriste Krstovski, and Aonan Zhang for helpful comments and discussion.

[2] https://github.com/kudkudak/word-embeddings-benchmarks

References

Baroni, M., Dinu, G., and Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 238–247. Association for Computational Linguistics, 2014.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, March 2003.

Bronnenberg, B. J., Kruger, M. W., and Mela, C. F. Database paper—the IRI marketing data set. *Marketing Science*, 27(4):745–748, 2008.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:21212159, July 2011.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.

Harper, F. M. and Konstan, J. A. The MovieLens datasets: history and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19:1–19:19, 2015.

Harris, Z. Distributional structure. *Word*, 10(23):146–162, 1954.

Hill, F., Reichart, R., and Korhonen, A. SimLex-999: evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456 [cs.CL]*, 2014.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Hu, Y., Koren, Y., and Volinsky, C. Collaborative filtering for implicit feedback datasets. In *The 8th IEEE International Conference on Data Mining*, pp. 263–272, 2008.

Jastrzebski, S., Leśniak, D., and Czarnecki, W. M. How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks. *arXiv preprint arXiv:1702.02170 [cs.CL]*, 2017.

Lambert, D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34 (1):1–14, 1992.

Liang, D., Charlin, L., McInerney, J., and Blei, D. M. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 951–961, 2016.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781[cs.CL]*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013b.

Mitchell, T. J. and Beauchamp, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Mnih, A. and Hinton, G. T. Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 641–648, 2007.

Munson, M. A., Webb, K., Sheldon, D., Fink, D., Hochachka, W. M., Iliff, M., Riedewald, M., Sorokina, D., Sullivan, B., Wood, C., and Kelling, S. The eBird reference dataset, version 2014, 2015.

Pennington, J., Socher, R., and Manning, C. D. GloVe: global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Reese, S., Boleda, G., Cuadros, M., Padró, L., and Rigau, G. Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In *Proceedings of the 7th Language Resources and Evaluation Conference*, pp. 1418–1421, 2010.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, 2009.

Rudolph, M., Ruiz, F., Mandt, S., and Blei, D. M. Exponential family embeddings. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 478–486. Curran Associates, Inc., 2016.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. In Anderson, J. A. and Rosenfeld, E. (eds.), *Neurocomputing: Foundations of Research*, pp. 696–699. MIT Press, Cambridge, MA, USA, 1988.