Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Sparse Probabilistic Relational Projection

Wu-Jun Li[†] and **Dit-Yan Yeung**[‡]
[†] Shanghai Key Laboratory of Scalable Computing and Systems
Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[‡] Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong, China
liwujun@cs.sjtu.edu.cn, dyyeung@cse.ust.hk

Abstract

Probabilistic relational PCA (PRPCA) can learn a projection matrix to perform dimensionality reduction for relational data. However, the results learned by PRPCA lack interpretability because each principal component is a linear combination of all the original variables. In this paper, we propose a novel model, called *sparse probabilistic relational projection* (SPRP), to learn a sparse projection matrix for relational dimensionality reduction. The sparsity in SPRP is achieved by imposing on the projection matrix a sparsity-inducing prior such as the Laplace prior or Jeffreys prior. We propose an expectation-maximization (EM) algorithm to learn the parameters of SPRP. Compared with PRPCA, the sparsity in SPRP not only makes the results more interpretable but also makes the projection operation much more efficient without compromising its accuracy. All these are verified by experiments conducted on several real applications.

Introduction

Principal component analysis (PCA) (Jolliffe 2002) and probabilistic PCA (PPCA) (Tipping and Bishop 1999) are very popular dimensionality reduction methods which have been widely used to explore the structure of a high-dimensional data set by mapping the data set into a low-dimensional space via a projection (or called transformation) matrix. However, it is difficult to interpret the results of PCA and PPCA because each principal component is a linear combination of all the original variables. To achieve interpretability, some sparse versions of PCA or PPCA have been proposed by enforcing many entries of the projection matrix to go to zero. Sparse PCA (SPCA) (Zou, Hastie, and Tibshirani 2006) first reformulates PCA as a regression-type optimization problem and then applies the *elastic net* (Zou and Hastie 2005) constraint on the regression coefficients to get a sparse projection matrix. In (Sigg and Buhmann 2008), sparsity is achieved by putting a 1-norm (L_1) constraint on the projection matrix during the expectation-maximization (EM) (Dempster, Laird, and Rubin 1977) learning procedure of PPCA. In (Archambeau and Bach 2008) and (Guan and Dy 2009), sparse versions of PPCA are proposed by putting some sparsity-inducing priors such as the Jeffreys prior on the projection matrix.

All the variants of PCA and sparse PCA mentioned above assume that the instances are independent and identically distributed (i.i.d.). Hence, they are not suitable for modeling *relational data* (Getoor and Taskar 2007; Li and Yeung 2009; Li, Zhang, and Yeung 2009; Li 2010; Li and Yeung 2011; Li, Yeung, and Zhang 2011) in which the instances are not i.i.d. In relational data, besides the content information,[1] there also exist links (i.e., relations) between the instances in the data. The attributes of the linked instances are often *correlated* rather than i.i.d. (Li, Yeung, and Zhang 2009). One typical example of relational data is a collection of research papers which contain both paper content and citations between the papers. The existence of a citation relation between two papers often implies that they are about the same topic. In (Li, Yeung, and Zhang 2009), probabilistic relational PCA (PRPCA), which extends PPCA by eliminating the i.i.d. assumption, is proposed to perform dimensionality reduction for relational data. By explicitly modeling the covariance between instances, PRPCA dramatically outperforms PCA and PPCA. However, as in PCA and PPCA, the results learned by PRPCA also lack interpretability.

In this paper, we propose a novel model, called *sparse probabilistic relational projection* (SPRP), to learn a sparse projection matrix for relational dimensionality reduction. Compared with PRPCA, the sparsity in SPRP not only makes the results more interpretable but also makes the projection operation much more efficient without compromising its accuracy.

Notation

For the convenience of presentation and comparison, we adopt the same notation as that in (Li, Yeung, and Zhang 2009). More specifically, we use boldface lowercase letters, such as $\mathbf{v}$, to denote vectors and v_i to denote the ith element of $\mathbf{v}$. Boldface uppercase letters, such as $\mathbf{F}$, are used to denote matrices, with the ith row and the jth column of $\mathbf{F}$ denoted by $\mathbf{F}_{i*}$ and $\mathbf{F}_{*j}$, respectively. F_{ij} is the element at the ith row and jth column of $\mathbf{F}$. We use $|\mathbf{F}|$ to denote the determinant of a matrix $\mathbf{F}$, $\mathrm{tr}(\mathbf{F})$ to denote its trace, $\mathbf{F}^T$

[1] As in (Li, Yeung, and Zhang 2009), we use the term 'content information' to refer to the feature vectors describing the instances.

for its transpose and $\mathbf{F}^{-1}$ for its inverse. $\mathbf{F} \succeq 0$ means that $\mathbf{F}$ is positive semi-definite (psd) and $\mathbf{F} \succ 0$ means that $\mathbf{F}$ is positive definite (pd). $\mathbf{P} \otimes \mathbf{Q}$ denotes the Kronecker product (Gupta and Nagar 2000) of $\mathbf{P}$ and $\mathbf{Q}$. $\mathbf{I}_n$ is the identity matrix of size $n \times n$ and $\mathbf{e}$ is a vector of 1's whose dimensionality depends on the context. $\mathcal{N}(\cdot)$ is overloaded for both multivariate normal distributions and matrix variate normal distributions (Gupta and Nagar 2000). We use $\mathrm{cov}(\cdot)$ to denote the covariance operation and $\langle\cdot\rangle$ to denote the expectation operation. The operation $\mathrm{diag}(\mathbf{v})$ converts the vector $\mathbf{v}$ into a diagonal matrix in which the ith diagonal entry is v_i.

As in (Tipping and Bishop 1999) and (Li, Yeung, and Zhang 2009), $\{\mathbf{t}_n\}_{n=1}^N$ denotes a set of observed d-dimensional data (content) vectors, $\boldsymbol{\mu}$ denotes the data sample mean, the $d \times q$ matrix $\mathbf{W}$ denotes the q principal axes (often called factor loadings or projection matrix), and $\mathbf{x}_n = \mathbf{W}^T(\mathbf{t}_n - \boldsymbol{\mu})$ denotes the corresponding q principal components (or called latent variables) of $\mathbf{t}_n$. We further use the $d \times N$ matrix $\mathbf{T}$ to denote the content matrix with $\mathbf{T}_{*n} = \mathbf{t}_n$ and the $q \times N$ matrix $\mathbf{X}$ to denote the latent variables of $\mathbf{T}$ with $\mathbf{X}_{*n} = \mathbf{W}^T(\mathbf{t}_n - \boldsymbol{\mu})$. As in (Li, Yeung, and Zhang 2009), we assume that the links are undirected. For data with directed links, we first convert the directed links into undirected links which can keep the original physical meaning of the links (Li, Yeung, and Zhang 2009). The adjacency (link) matrix of the N instances is denoted by $\mathbf{A}$. $A_{ij} = 1$ if there exists a link between instances i and j, and otherwise $A_{ij} = 0$. Moreover, $A_{ii} = 0$, which means that there exist no self-links.

Probabilistic Relational PCA

With matrix variate normal distributions (Gupta and Nagar 2000), the generative model of PRPCA (Li, Yeung, and Zhang 2009) is defined as:

$$\begin{aligned}\boldsymbol{\Upsilon} \mid \Theta &\sim \mathcal{N}_{d,N}(\mathbf{0}, \sigma^2\mathbf{I}_d \otimes \boldsymbol{\Phi}),\\ \mathbf{X} \mid \Theta &\sim \mathcal{N}_{q,N}(\mathbf{0}, \mathbf{I}_q \otimes \boldsymbol{\Phi}),\\ \mathbf{T} &= \mathbf{W}\mathbf{X} + \boldsymbol{\mu}\mathbf{e}^T + \boldsymbol{\Upsilon},\end{aligned}$$

where $\Theta = \{\boldsymbol{\mu}, \mathbf{W}, \sigma^2\}$ denotes the set of parameters, $\boldsymbol{\Phi} = \boldsymbol{\Delta}^{-1}$ and $\boldsymbol{\Delta} \triangleq \gamma\mathbf{I}_N + (\mathbf{I}_N + \mathbf{A})^T(\mathbf{I}_N + \mathbf{A})$ with γ being typically a very small positive number to make $\boldsymbol{\Delta} \succ 0$. In (Li, Yeung, and Zhang 2009), $\boldsymbol{\Phi}$ is called *relational covariance* which reflects the covariance between the instances.

Then, we can get the following results:

$$\begin{aligned}\mathbf{T} \mid \mathbf{X}, \Theta &\sim \mathcal{N}_{d,N}(\mathbf{W}\mathbf{X} + \boldsymbol{\mu}\mathbf{e}^T, \sigma^2\mathbf{I}_d \otimes \boldsymbol{\Phi}),\\ \mathbf{T} \mid \Theta &\sim \mathcal{N}_{d,N}\left(\boldsymbol{\mu}\mathbf{e}^T, (\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}_d) \otimes \boldsymbol{\Phi}\right).\end{aligned}$$

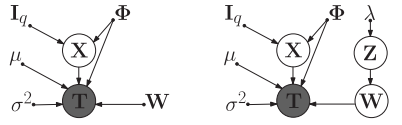
Based on the generative process, Figure 1(a) shows the graphical model of PRPCA.

If we set $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}_d$, the log-likelihood of the observation matrix $\mathbf{T}$ in PRPCA is

$$\mathcal{L} = \ln p(\mathbf{T} \mid \Theta) = -\frac{N}{2}\Big[\ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{H})\Big] + c,$$

where c is a constant independent of the parameters Θ and $\mathbf{H}$ is defined as follows:

$$\mathbf{H} = \frac{(\mathbf{T} - \boldsymbol{\mu}\mathbf{e}^T)\boldsymbol{\Delta}(\mathbf{T} - \boldsymbol{\mu}\mathbf{e}^T)^T}{N}. \tag{1}$$



(a) PRPCA (b) SPRP

Figure 1: Graphical models of PRPCA and SPRP, in which $\mathbf{T}$ is the observation matrix, $\mathbf{X}$ and $\mathbf{Z}$ are the latent variable matrices, $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2 are the parameters to learn, λ is the hyperparameter, and the other quantities are kept constant.

In (Li, Yeung, and Zhang 2009), two *maximum likelihood estimation* (MLE) methods, one based on a closed-form solution and another based on EM, are proposed to learn the parameters of PRPCA.

Sparse Probabilistic Relational Projection

In this paper, we propose to put a sparsity-inducing prior on $\mathbf{W}$ to encourage many of its entries to go to zero. The resulting model is called *sparse probabilistic relational projection* (SPRP) due to its sparsity property. Although there exist many sparsity-inducing priors in the literature, e.g., (Figueiredo 2003; Caron and Doucet 2008; Archambeau and Bach 2008; Guan and Dy 2009), here we consider only two of them, the Laplace prior and Jeffreys prior. Using other sparsity-inducing priors in SPRP is expected to follow the same principle and will be left to our future pursuit.

SPRP with Laplace Prior

In SPCA (Zou, Hastie, and Tibshirani 2006), sparsity is achieved by putting an L_1 regularization term on the projection matrix. Here we learn a sparse $\mathbf{W}$ for SPRP in a similar way. However, unlike SPCA which is formulated from a non-probabilistic view, SRPR is based on a probabilistic formulation which can automatically learn the hyperparameters while the non-probabilistic formulation cannot.

We adopt the Laplace (i.e., double-exponential) prior (Park and Casella 2008; Guan and Dy 2009) for $\mathbf{W}$:

$$\begin{aligned}p(W_{ij} \mid \lambda) &= \frac{\sqrt{\lambda}}{2}\exp\left\{-\sqrt{\lambda}\|W_{ij}\|_1\right\},\\ p(\mathbf{W} \mid \lambda) &= \prod_{i=1}^{d}\prod_{j=1}^{q} p(W_{ij} \mid \lambda)\\ &= \left(\frac{\sqrt{\lambda}}{2}\right)^{dq}\exp\left\{-\sqrt{\lambda}\|\mathbf{W}\|_1\right\},\end{aligned}$$

where $\|\cdot\|_1$ denotes the absolute value for a scalar and the 1-norm for a matrix.

Using Bayes' rule, the log-posterior of Θ can be computed as follows:

$$\begin{aligned}\ln p(\Theta \mid \mathbf{T}) &= \ln p(\mathbf{T} \mid \Theta) + \ln p(\Theta) + c_0\\ &= -\frac{N}{2}\Big[\ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{H})\Big] - \sqrt{\lambda}\|\mathbf{W}\|_1\\ &\quad + \ln p(\boldsymbol{\mu}) + \ln p(\sigma^2) + c_1,\end{aligned} \tag{2}$$

where c_0 and c_1 are constants independent of Θ.

For simplicity, here we adopt the *maximum a posteriori* (MAP) strategy to estimate the parameters. Due to the L_1 constraint on $\mathbf{W}$, the MAP estimation of $\mathbf{W}$ will naturally induce sparsity, which means that many entries of $\mathbf{W}$ will be automatically driven to zero during the learning procedure. In this paper, we assume that $p(\boldsymbol{\mu})$ and $p(\sigma^2)$ are uniform.

Remark 1 *Of course, we may also put non-uniform priors, such as conjugate priors, on $\boldsymbol{\mu}$ and σ^2. Here, uniform priors for $\boldsymbol{\mu}$ and σ^2 are adopted mainly for fair comparison because they are also adopted in PRPCA. Alternatively, we may also treat all the parameters as random variables and resort to fully Bayesian methods, such as variational methods (Jordan et al. 1999), for learning and inference. However, since the focus of this paper is on demonstrating the promising advantages of sparsity under the PRPCA framework, all these possible variants are left to future extensions.*

It is not easy to directly optimize the objective function in (2) though. As in (Park and Casella 2008; Guan and Dy 2009), we adopt a hierarchical interpretation of the Laplace prior:

$$p(Z_{ij} \mid \lambda) = \frac{\lambda}{2} \exp\left\{-\frac{\lambda}{2} Z_{ij}\right\}, \text{ for } Z_{ij} \geq 0, \tag{3}$$

$$W_{ij} \mid Z_{ij} \sim \mathcal{N}(0, Z_{ij}). \tag{4}$$

It is easy to show that this hierarchical reformulation is equivalent to the original Laplace prior, because

$$\begin{aligned} p(W_{ij} \mid \lambda) &= \int p(W_{ij} \mid Z_{ij}) p(Z_{ij} \mid \lambda) dZ_{ij} \\ &= \frac{\sqrt{\lambda}}{2} \exp\left\{-\sqrt{\lambda}\|W_{ij}\|_1\right\}. \end{aligned} \tag{5}$$

Figure 1(b) depicts the graphical model of SPRP as compared with that of PRPCA in Figure 1(a).

Learning By setting the gradient of $\ln p(\Theta \mid \mathbf{T})$ with respect to $\boldsymbol{\mu}$ to zero, we get the (closed-form) MAP estimate for $\boldsymbol{\mu}$ as follows:

$$\widehat{\boldsymbol{\mu}} = \frac{\mathbf{T}\boldsymbol{\Delta}\mathbf{e}}{\mathbf{e}^T\boldsymbol{\Delta}\mathbf{e}}. \tag{6}$$

For the other parameters ($\mathbf{W}$ and σ^2) of SPRP, we derive an (iterative) EM (Dempster, Laird, and Rubin 1977) algorithm to learn them. For the rest of this paper, we still use Θ to refer to the parameters but they only contain $\mathbf{W}$ and σ^2 because $\boldsymbol{\mu}$ can be directly computed by (6). During the EM learning procedure, we treat $\{\mathbf{Z}, \mathbf{X}\}$ as missing data and $\{\mathbf{T}, \mathbf{Z}, \mathbf{X}\}$ as complete data. The EM algorithm for MAP estimation operates by alternating between the following two steps:

- **E-step:** The expectation of the complete-data log-posterior with respect to the distribution of the missing variables $\{\mathbf{Z}, \mathbf{X}\}$ is computed. This expected value is often called the Q-function which is defined as follows:

$$Q\left(\Theta \mid \Theta(t)\right) = \int d\mathbf{Z}\, d\mathbf{X}\, p(\mathbf{Z}, \mathbf{X} \mid \Theta(t), \mathbf{T}) \ln p(\Theta \mid \mathbf{T}, \mathbf{Z}, \mathbf{X}).$$

- **M-step:** The Q-function is maximized to update the parameters:

$$\Theta(t+1) = \operatorname*{argmax}_{\Theta} Q\left(\Theta \mid \Theta(t)\right).$$

The whole EM learning procedure is summarized in Algorithm 1 and the detailed derivation can be found in (Li 2010). Note that as in (Li, Yeung, and Zhang 2009), we use $\mathbf{W}$ and σ^2 to denote the old values and $\widetilde{\mathbf{W}}$ and $\widetilde{\sigma}^2$ for the updated ones.

Algorithm 1 EM algorithm for SPRP

Initialize $\mathbf{W}$ and σ^2.
for $t = 1$ to T
  E-step: Compute the *sufficient statistics*
    $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}_q$,
    $\langle \mathbf{X} \rangle = \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{T} - \boldsymbol{\mu}\mathbf{e}^T)$,
    $\langle \mathbf{X}\boldsymbol{\Delta}\mathbf{X}^T \rangle = N\sigma^2\mathbf{M}^{-1} + \langle \mathbf{X} \rangle \boldsymbol{\Delta} \langle \mathbf{X} \rangle^T$,
    $\langle \frac{1}{Z_{ij}} \rangle = \frac{\sqrt{\lambda}}{\|W_{ij}\|_1}$.
  M-step: Update the parameters
    for $i = 1$ to d
      $\boldsymbol{\Sigma}_i = \operatorname{diag}\left(\frac{\|W_{i1}\|_1}{\sqrt{\lambda}}, \cdots, \frac{\|W_{iq}\|_1}{\sqrt{\lambda}}\right)$,
      $\widetilde{\mathbf{W}}_{i*} = \mathbf{H}_{i*}\mathbf{W}\mathbf{M}^{-1}\boldsymbol{\Sigma}_i \big[(\sigma^2\mathbf{I}_q + \mathbf{M}^{-1}\mathbf{W}^T\mathbf{H}\mathbf{W})\mathbf{M}^{-1}\boldsymbol{\Sigma}_i + \frac{\sigma^2}{N}\mathbf{I}_q\big]^{-1}$.
    end for
    $\widetilde{\sigma}^2 = \frac{\operatorname{tr}[\mathbf{H} - \mathbf{H}\mathbf{W}\mathbf{M}^{-1}\widetilde{\mathbf{W}}^T]}{d}$.
end for

To learn the hyperparameter λ, we may resort to the cross-validation strategy. Alternatively, we may treat λ as one of the parameters, just like $\mathbf{W}$, to get the following EM updating equation: $\lambda = \frac{2dq}{\sum_{i=1}^{d}\sum_{j=1}^{q}\langle Z_{ij} \rangle}$.

SPRP with Jeffreys Prior

The Jeffreys prior (Figueiredo 2003; Guan and Dy 2009) is a noninformative prior. To use the Jeffreys prior, we only need to change the density function of Z_{ij} in (3) to: $p(Z_{ij}) \propto \frac{1}{Z_{ij}}$. We can see that there exist no hyperparameters in the Jeffreys prior but the Laplace prior does have the hyperparameter λ which needs to be learned.

With the Jeffreys prior, the log-posterior can be computed as follows:

$$\begin{aligned} \ln p(\Theta \mid \mathbf{T}) &= \ln p(\mathbf{T} \mid \Theta) + \ln p(\Theta) + c_2 \\ &= -\frac{N}{2}\Big[\ln|\mathbf{C}| + \operatorname{tr}(\mathbf{C}^{-1}\mathbf{H})\Big] - \ln\|\mathbf{W}\|_1 \\ &\quad + \ln p(\boldsymbol{\mu}) + \ln p(\sigma^2) + c_3, \end{aligned} \tag{7}$$

where c_2 and c_3 are constants independent of the parameters. We can see that the only difference between (2) and (7) lies in the difference between the regularization terms $\sqrt{\lambda}\|\mathbf{W}\|_1$ and $\ln\|\mathbf{W}\|_1$.

To learn the parameters for SPRP with the Jeffreys prior, we only need to change $\langle \frac{1}{Z_{ij}} \rangle$ and $\boldsymbol{\Sigma}_i$ in Algorithm 1 as follows: $\langle \frac{1}{Z_{ij}} \rangle = \frac{1}{W_{ij}^2}$, $\boldsymbol{\Sigma}_i = \operatorname{diag}\big(W_{i1}^2, \ldots, W_{iq}^2\big)$.

Time Complexity

To train the model, we need $O(dN)$ time to compute $\mathbf{H}$ and $O(Tqd^2 + Tdq^3)$ for the T EM iterations. Hence, the total time complexity is $O(dN + Tqd^2 + Tdq^3)$. If $d > q^2$, Tqd^2 will be larger than Tdq^3 and so the time complexity will become $O(dN + Tqd^2)$ which is equal to that of PRPCA.

If we want to use the learned $\mathbf{W}$ to perform projection, the time complexity will depend on the number of nonzero entries in $\mathbf{W}$. Generally speaking, SPRP has lower projection cost than PRPCA because the $\mathbf{W}$ in SPRP is more sparse than that in PRPCA.

Experimental Evaluation

As in (Li, Yeung, and Zhang 2009), we adopt PCA to initialize $\mathbf{W}$, initialize σ^2 to 10^{-6}, and set γ to 10^{-6}. We set the number of EM iterations T to 30 because 30 iterations are sufficient for both PRPCA and SPRP to achieve good performance. The baseline methods for comparison include PCA, sparse probabilistic projection (SPP) (Archambeau and Bach 2008) and PRPCA. Through the experiments we want to verify the following claims: (1) PCA cannot effectively exploit the relational information in relational data. Furthermore, it cannot learn interpretable results. (2) Due to its i.i.d. assumption, SPP cannot achieve satisfactory performance even though it can learn interpretable results. (3) PRPCA can effectively exploit the relational information, but it cannot learn interpretable results. (4) SPRP not only can effectively exploit the relational information, but it can also learn interpretable results.

Data Sets

Three data sets are used for our experimental evaluation. The first two are the *preprocessed* WebKB (Craven et al. 1998) and Cora (McCallum et al. 2000) data sets used in (Zhu et al. 2007; Li, Yeung, and Zhang 2009). The third data set is called Cora-IR, which contains the information retrieval papers from the original Cora data set (McCallum et al. 2000). All these data sets use the bag-of-words representation for the content information.

The WebKB data set contains 4,017 web pages from the computer science departments of four universities (Cornell, Texas, Washington, and Wisconsin). Each web page is labeled with one of seven categories: *student*, *professor*, *course*, *project*, *staff*, *department*, and "*other*". The original links are directed. We adopt the same strategy as that in (Li, Yeung, and Zhang 2009) to convert the directed links into undirected ones. Some characteristics of the WebKB data set are summarized in Table 1.

Table 1: Characteristics of the WebKB data set.

Data Set	#classes	#instances	#words
Cornell	7	827	4,134
Texas	7	814	4,029
Washington	7	1,166	4,165
Wisconsin	6	1,210	4,189

The Cora data set used in (Li, Yeung, and Zhang 2009) contains 4,343 research papers from the computer science community. The content information refers to the paper abstracts and the links refer to the citations. The task is to classify each paper into one of the subfields of *data structure* (DS), *hardware and architecture* (HA), *machine learning* (ML), and *programming language* (PL). Some characteristics of the Cora data set are summarized in Table 2.

Table 2: Characteristics of the Cora data set.

Data Set	#classes	#instances	#words
DS	9	751	6,234
HA	7	400	3,989
ML	7	1,617	8,329
PL	9	1,575	7,949

Because we do not have the dictionary for generating the bag-of-words representation in the preprocessed WebKB and Cora data sets, we collect another data set, called Cora-IR. The Cora-IR data set contains 350 information retrieval papers from the original Cora set (McCallum et al. 2000). There are four subfields (classes) in Cora-IR: *retrieval*, *filtering*, *extraction*, and *digital_library*. We use the title of each paper for the content information. After preprocessing, we get a dictionary of 787 words. For each word, there is at least one instance (paper) containing it. We will use this dictionary to demonstrate the interpretability of SPRP.

In (Li, Yeung, and Zhang 2009), only information about the words (bag-of-words) is used to represent the content information. We expand the original content features by adding some extra features extracted from the original *directed* links. The ith link feature is referred to as *link-to-instance*$_i$. For example, if instance k links to instance i, the ith link feature of instance k will be 1. Otherwise, it will be 0. In fact, this kind of link features can also be treated as content information. For example, given a paper, the *link-to-instance*$_i$ feature actually reflects whether the reference part of that paper contains paper i. For a web page, the *link-to-instance*$_i$ feature can also be directly extracted from the HTML file (content) of that page. Note that it is somewhat impractical to treat the *linked-by-instance*$_i$ features as content features because they cannot be directly extracted from the content of the instances. For example, the papers citing a specific paper i are not included in the content of paper i. After extracting the link features, we combine the original bag-of-words with the link features to obtain the *expanded content features*. We can see that the way to get the expanded content features also assumes that the instances are i.i.d. We will show that this way of using link information is not good enough to capture the structure information in relational data. On the contrary, PRPCA and SPRP, which are not based on the i.i.d. assumption, can provide more effective ways to model relational data. In what follows, we will refer to the original bag-of-words representation as *original content features*.

Laplace Prior vs. Jeffreys Prior

We define the *degree of sparsity* (DoS) of $\mathbf{W}$ as follows:

$$DoS = \frac{\text{number of zero entries in } \mathbf{W}}{dq} \times 100\%.$$

From (2), we can see that λ in the Laplace prior controls the DoS of the learned $\mathbf{W}$. The larger λ is, the larger will the DoS of $\mathbf{W}$ be. Here, we vary λ to get different DoS and then evaluate the corresponding accuracy. Due to space limitation, we only report here results on the DS data set because other data sets exhibit similar properties. The accuracy of PRPCA on the DS data set is 68.1%. The corresponding results of SPRP with the Laplace prior are shown in Table 3.

Table 3: Accuracy (Acc) against DoS for SPRP with the Laplace prior.

DoS (%)	30	50	60	70	76	80	90	96
Acc (%)	68.8	68.7	68.1	67.5	66.9	66.7	65.9	63.2

From Table 3, we can discover some interesting properties of SPRP:

- In general, the larger the DoS is, the lower will the accuracy be. This is reasonable because less information about the features will be used to construct the principal components with larger DoS.
- Compared with PRPCA, SPRP can achieve a DoS as large as 60% without compromising its accuracy. Even when $DoS = 70\%$, the accuracy of SPRP is still comparable with that of PRPCA. This shows that the sparsity pursuit in SPRP is very meaningful because it can obtain interpretable results without compromising its accuracy.

For the Jeffreys prior, there are no hyperparameters to tune. After learning, we get an accuracy of 68.1% with $DoS = 76\%$. Hence, with similar DoS, the Jeffreys prior can achieve slightly higher accuracy than the Laplace prior. From Table 3, we also find that a relatively good tradeoff between the DoS and accuracy can be obtained if $70\% < DoS < 80\%$. Hence, we can say that the Jeffreys prior can *adaptively* learn a good DoS. Due to this nice property, we only report the results of SPRP with the Jeffreys prior in the rest of this paper. For fair comparison, we also use the Jeffreys prior for SPP (Archambeau and Bach 2008).

Interpretability

For all the projection methods, we set the dimensionality of the latent space to 50. For Cora-IR, we adopt the original content features because we need to use the selected words for illustration. For all the other data sets, we use the expanded content features.

The DoS comparison of PCA, SPP, PRPCA and SPRP is shown in Table 4. We can see that the DoS of both PCA and PRPCA on WebKB and Cora-IR is either 0 or close to 0, which means that all the original variables (i.e., words) will be used to compute the principal components for PCA and PRPCA. For Cora, there exist some features (words) that no instances (papers) contain them. That is to say, all entries in the corresponding rows of the content matrix $\mathbf{T}$ will be zero. We also find that the zeroes in $\mathbf{W}$ are from those rows corresponding to the all-zero rows in $\mathbf{T}$. Hence, we can say that on Cora, PCA and PRPCA cannot learn sparse projection matrices either. Due to this non-sparse property, the results of PCA and PRPCA lack interpretability. On the contrary, both SPP and SPRP can learn sparse projection matrices. Compared with SPP, SPRP achieves lower DoS. However, the discrimination ability of SPRP is much higher than SPP, as to be shown later.

Table 4: DoS (in %) comparison of PCA, SPP, PRPCA and SPRP.

		PCA	SPP	PRPCA	SPRP
Cora	Cora-IR	0	90	0	72
	DS	18	88	20	76
	HA	16	86	18	72
	ML	10	90	12	76
	PL	11	90	13	76
WebKB	Cornell	0	74	0	48
	Texas	0	77	0	42
	Washington	1	74	1	47
	Wisconsin	0	75	0	47

To further compare the results of SPP and SPRP in terms of interpretability, we show some details of the first six columns of $\mathbf{W}$ in Table 5. In the table, the 'Selected Words', arranged in descending order in terms of their $\mathbf{W}$ values, correspond to the top 10 nonzero entries in $\mathbf{W}$. It is easy to see that the learned projection matrix of either SPRP or SPP does show some discrimination ability. More specifically, W_{*1} mainly corresponds to the class *retrieval*, W_{*2} to *filtering*, W_{*3} and W_{*4} to *extraction*, and W_{*5} and W_{*6} to *digital_library*. This is very promising because we can use the magnitude of the corresponding principal components to measure the class proportions of each instance. Detailed comparison between the words selected by SPP and SPRP shows that the words selected by SPRP is more discriminative than those selected by SPP. For example, 'dictionary' is more related to *retrieval* than 'agents', and 'symbolic' and 'wrapper' are more related to *extraction* than 'multi' and 'empirical'.

For SPRP, 116 out of 787 words are not used to construct any principal component, which means that the entire rows of $\mathbf{W}$ corresponding to those words are zero. Hence, SPRP can also be used to perform feature elimination, which will speed up the collection process for new data. For example, some eliminated words include 'wwww', 'aboutness', 'uu', 'erol', 'stylistic', 'encounter', 'classificatin', 'hypercode', 'broswer', 'lacking', 'multispectral', and 'exchanges'. It is interesting to note that most of them are typos or not related to information retrieval at all.

Accuracy

As in (Li, Yeung, and Zhang 2009), we adopt 5-fold cross validation to evaluate the accuracy. The dimensionality of the latent space is set to 50 for all the dimensionality reduction methods. After dimensionality reduction, a linear support vector machine (SVM) is trained for classification based on the low-dimensional representation. The average classification accuracy for 5-fold cross validation, together with the standard deviation, is used as the performance metric.

The results on Cora and WebKB are shown in Figure 2 and Figure 3, respectively. PRPCA based on the original content features is denoted as PRPCA0, which achieves performance comparable with the state-of-the-art methods (Li,

Table 5: Some details of the projection matrices learned by SPRP and SPP.

		Selected Words (arranged in descending order in terms of their $\mathbf{W}$ values)
SPRP	W_{*1}	information; retrieval; cross; language; extraction; system; evaluation; techniques; dictionary; incremental
	W_{*2}	text; categorization; learning; classification; information; feature; retrieval; selection; classifiers; algorithm
	W_{*3}	web; wide; learning; world; information; extract; symbolic; aid; formatting; extraction
	W_{*4}	extraction; information; learning; structured; rules; wrapper; documents; induction; grammatical; machine
	W_{*5}	text; language; digital; wide; world; high; processing; structured; sources; information
	W_{*6}	digital; library; learning; libraries; image; market; services; decoding; stanford; metadata
SPP	W_{*1}	information; retrieval; agents; evaluation; system; cross; language; intelligent; dissemination; distributed
	W_{*2}	text; categorization; learning; classification; information; feature; selection; extraction; study; case
	W_{*3}	web; wide; world; learning; information; search; multi; patterns; server; performance
	W_{*4}	extraction; information; learning; rules; disclosure; automatically; structured; basis; dictionary; empirical
	W_{*5}	text; digital; world; wide; information; system; library; categorization; processing; high
	W_{*6}	digital; library; learning; services; libraries; market; video; access; navigating; agents

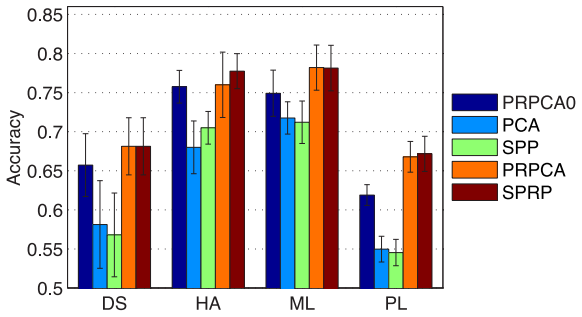


Figure 2: Results in average classification accuracy with standard deviation on the Cora data set.

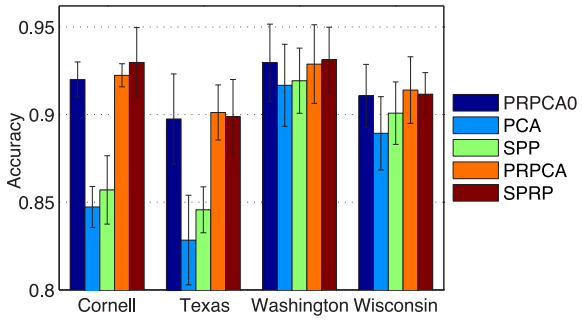


Figure 3: Results in average classification accuracy with standard deviation on the WebKB data set.

Yeung, and Zhang 2009). All the other methods are based on the expanded content features. Compared with PCA, the higher accuracy of PRPCA0 shows that it is not good enough to just extract the extra information from the links and still assume the instances to be i.i.d. Comparison between PRPCA and PRPCA0 shows that slightly better performance can be achieved with the expanded content features, particularly for the Cora data set. Comparison between PRPCA and PCA verifies the claim in (Li, Yeung, and Zhang 2009) that PRPCA dramatically outperforms PCA by eliminating the i.i.d. assumption. Comparison between SPP and PCA shows that the sparsity pursuit does not necessarily deteriorate the accuracy for the case with the i.i.d. assumption. Comparison between SPRP and SPP shows that under the sparsity pursuit case, dramatic accuracy improvement can also be achieved by explicitly modeling the covariance between instances, which once again verifies that the i.i.d. assumption is unreasonable for relational data. Finally, comparison between SPRP and PRPCA shows that under the PRPCA framework, we can also achieve sparsity without compromising accuracy.

Table 6: Projection time (in seconds) comparison.

		PRPCA	SPRP
Cora	DS	2.431	0.749
	HA	0.834	0.284
	ML	8.225	2.272
	PL	7.507	2.123
WebKB	Cornell	2.362	1.241
	Texas	2.273	1.323
	Washington	3.588	1.946
	Wisconsin	3.778	2.029

Projection Cost

When the projection matrix learned is used to perform projection, the sparsity of SPRP will make its projection cost much lower than that of PRPCA. Table reports the projection time needed to perform projection on the Cora and WebKB data sets. The test is performed with MATLAB implementation on a 2.33GHz personal computer. We can see that SPRP is much faster than PRPCA for the projection operation.

Conclusion and Future Work

In this paper, we have proposed a novel model, SPRP, to learn a sparse projection matrix for relational dimensionality reduction. Compared with PRPCA, the sparsity in SPRP not only makes its results interpretable, but it also makes the projection operation much more efficient without compromising its accuracy. Furthermore, SPRP can also be used to perform feature elimination, which will speed up the collection process for new data. Compared with traditional sparse projection methods based on the i.i.d. assumption, SPRP can learn a more discriminative projection by explicitly modeling the covariance between instances. SPRP brings about some theoretical contributions to the area of sparsity pursuing because SPRP is the first model that pursues sparsity

without requiring the i.i.d. assumption. Hence, it can inspire us to relax the i.i.d. assumption in other sparse models as well to further boost model performance.

Acknowledgments

Li is supported by the NSFC (No. 61100125) and the 863 Program of China (No. 2011AA01A202, No. 2012AA011003). Yeung is supported by General Research Fund 621310 from the Research Grants Council of Hong Kong.

References

Archambeau, C., and Bach, F. 2008. Sparse probabilistic projections. In *NIPS 21*.

Caron, F., and Doucet, A. 2008. Sparse Bayesian nonparametric regression. In *ICML*.

Craven, M.; DiPasquo, D.; Freitag, D.; McCallum, A.; Mitchell, T. M.; Nigam, K.; and Slattery, S. 1998. Learning to extract symbolic knowledge from the world wide web. In *AAAI/IAAI*.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38.

Figueiredo, M. A. T. 2003. Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 25(9):1150–1159.

Getoor, L., and Taskar, B. 2007. *Introduction to Statistical Relational Learning*. The MIT Press.

Guan, Y., and Dy, J. G. 2009. Sparse probabilistic principal component analysis. In *AISTATS*.

Gupta, A. K., and Nagar, D. K. 2000. *Matrix Variate Distributions*. Chapman & Hall/CRC.

Jolliffe, I. T. 2002. *Principal Component Analysis*. Springer, second edition.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine Learning* 37(2):183–233.

Li, W.-J., and Yeung, D.-Y. 2009. Relation regularized matrix factorization. In *IJCAI*.

Li, W.-J., and Yeung, D.-Y. 2011. Social relations model for collaborative filtering. In *AAAI*.

Li, W.-J.; Yeung, D.-Y.; and Zhang, Z. 2009. Probabilistic relational PCA. In *NIPS 22*.

Li, W.-J.; Yeung, D.-Y.; and Zhang, Z. 2011. Generalized latent factor models for social network analysis. In *IJCAI*.

Li, W.-J.; Zhang, Z.; and Yeung, D.-Y. 2009. Latent Wishart processes for relational kernel learning. In *AISTATS*.

Li, W.-J. 2010. *Latent Factor Models for Statistical Relational Learning*. Ph.D. Dissertation, Hong Kong University of Science and Technology.

McCallum, A.; Nigam, K.; Rennie, J.; and Seymore, K. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval* 3(2):127–163.

Park, T., and Casella, G. 2008. The Bayesian lasso. *Journal of the American Statistical Association* 103(482):681–686.

Sigg, C. D., and Buhmann, J. M. 2008. Expectation-maximization for sparse and non-negative PCA. In *ICML*.

Tipping, M. E., and Bishop, C. M. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* 61(3):611–622.

Zhu, S.; Yu, K.; Chi, Y.; and Gong, Y. 2007. Combining content and link for classification using matrix factorization. In *SIGIR*.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67(2):301–320.

Zou, H.; Hastie, T.; and Tibshirani, R. 2006. Sparse principal component analysis. *Journal of Computational and Graphical Statistics* 15(2):265–286.