
Sparse Probabilistic Principal Component Analysis

Yue Guan

Electrical and Computer Engineering Dept.
Northeastern University
Boston, MA 02115, USA

Jennifer G. Dy

Electrical and Computer Engineering Dept.
Northeastern University
Boston, MA 02115, USA

Abstract

Principal component analysis (PCA) is a popular dimensionality reduction algorithm. However, it is not easy to interpret which of the original features are important based on the principal components. Recent methods improve interpretability by sparsifying PCA through adding an L_1 regularizer. In this paper, we introduce a probabilistic formulation for sparse PCA. By presenting sparse PCA as a probabilistic Bayesian formulation, we gain the benefit of automatic model selection. We examine three different priors for achieving sparsification: (1) a two-level hierarchical prior equivalent to a Laplacian distribution and consequently to an L_1 regularization, (2) an inverse-Gaussian prior, and (3) a Jeffrey's prior. We learn these models by applying variational inference. Our experiments verify that indeed our sparse probabilistic model results in a sparse PCA solution.

1 Introduction

Principal component analysis (PCA) (Jolliffe, 1986) is a popular tool for data analysis and dimensionality reduction. In PCA, the derived principal components (PCs) are orthogonal to each other and represent the directions of largest variance. PCA captures the largest information in the first few principal components, guarantees minimal information loss and minimal reconstruction error in a least squares sense. However, in PCA, the principal components are a linear combination of all the original features, which makes

the PCs difficult to interpret and explain. Rotation methods are typically used to improve interpretability of the PCs (Jolliffe, 1995). Another approach is simple thresholding of PC loadings to identify important features (Cadima & Jolliffe, 1995). A recent work in linear regression called least absolute shrinkage and selection operator (LASSO) selects relevant variables by adding an L_1 norm regularizer on the predictor weights. The L_1 term serves as a tractable estimate to L_0 and leads to sparse solutions. Elastic net (Zou & Hastie, 2003) is an extension of LASSO that essentially adds both an L_1 and an L_2 (ridge regression (Hoerl, 1962)) regularizers to the least squares objective in a regression framework. The L_2 term allows for the regression algorithm to deal with data when the number of dimensions is much higher than the number of samples as is typical in gene applications. LASSO, ridge regression and elastic net are called coefficient shrinkage methods (Zou & Hastie, 2005) because they select features in linear models by shrinking weights to zero. Inspired by these sparsification techniques, SCoTLASS (Jolliffe, 2003) combines the maximum variance PCA objective with an L_1 penalty (similar to LASSO), and sparse PCA (SPCA) (Zou et al., 2006) combines a PCA least squares error objective with both L_1 and L_2 regularization terms similar to elastic net.

Although there are studies in probabilistic models for sparse regression (Cawley et al., 2007) and sparse classification (Tipping, 2001), there is none to our knowledge on probabilistic sparse PCA¹. A probabilistic model provides several benefits, including extensions to mixture models, dealing with missing data, and

Appearing in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

¹There was none at the time this paper was submitted for review; however, by the time of this publication, there is now a recent work that presents a general sparse probabilistic projection model which includes PCA as a special case (Archambeau & Bach, 2008). The work on (Archambeau & Bach, 2008) focuses on the general model; whereas, this paper focuses on sparse probabilistic PCA and studies three alternative sparsity priors in detail.

Bayesian methods for model selection. A limitation with SPCA is that the level of sparsity is not automatically determined. In this paper, we introduce a probabilistic formulation of sparse PCA and show the benefit of having the probabilistic formulation for model selection. We provide a Bayesian solution to our probabilistic formulation and apply variational inference to make our algorithm tractable. To achieve sparsification, we investigate three alternative priors: a two-level hierarchical prior equivalent to a Laplacian distribution and consequently to an L_1 regularization, an inverse-Gaussian prior, and a Jeffrey's prior.

This paper is organized as follows. We review the probabilistic PCA (PPCA) model introduced by (Tipping & Bishop, 1999) and independently by (Roweis, 1997) in Section 2. Then, we present our sparse probabilistic PCA model in Section 3. Section 4 provides details of our variational approach. Section 5 studies different prior models to enforce sparsity. We discuss and report our empirical investigation on synthetic and benchmark data verifying that our probabilistic model indeed results in a sparse PCA solution in Section 6. Finally, we summarize the paper in Section 7.

2 Review of Probabilistic PCA

Conventional PCA seeks a q -dimensional ($q < d$) linear projection that best represents the data in a least-squares sense. Let D be a data set of observed d -dimensional vector $D = \{\mathbf{t}_n\}$, where $n \in 1, \dots, N$. The sample covariance matrix is $\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{m})(\mathbf{t}_n - \mathbf{m})'$ where $\mathbf{m} = \frac{1}{N} \sum_{n=1}^N \mathbf{t}_n$ and A' stands for the transpose of matrix A . Then, the q principal axes are given by the q dominant eigenvectors (i.e. those with the q largest eigenvalues). The projected value of data \mathbf{t}_n is given by $\mathbf{x}_n = \mathbf{U}'_q (\mathbf{t}_n - \mathbf{m})$, where $\mathbf{U}_q = (\mathbf{u}_1, \dots, \mathbf{u}_q)$. It can be shown that PCA finds the linear projection which maximizes the variance in the projected space.

Conventional PCA does not define a probability model. PCA can be reformulated as a maximum likelihood solution to a latent variable model (Tipping & Bishop, 1999). Let \mathbf{x} be a q -dimensional latent variable. The observed variable \mathbf{t} is then defined as a linear transformation of \mathbf{x} with additional noise ϵ : $\mathbf{t} = \mathbf{W}\mathbf{x} + \mathbf{m} + \epsilon$. Here \mathbf{W} is a $d \times q$ linear transformation matrix, \mathbf{m} is a d -dimensional vector that allows \mathbf{t} to have a non-zero mean. Both the latent variable \mathbf{x} and noise ϵ are assumed to be isotropic Gaussian: $p(\mathbf{x}) \sim \mathcal{N}(0, \mathbf{I}_q)$ and $p(\epsilon) \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$. Therefore, the conditional distribution of \mathbf{t} given \mathbf{x} is: $p(\mathbf{t}|\mathbf{x}) \sim \mathcal{N}(\mathbf{W}\mathbf{x} + \mathbf{m}, \sigma^2 \mathbf{I}_d)$. Then, the marginal distribution of \mathbf{t} is also Gaussian, $p(\mathbf{t}) \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$, where the covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}' + \sigma^2 \mathbf{I}_d$. One can

compute the maximum-likelihood estimator for the parameters \mathbf{m}, \mathbf{W} and σ^2 from a data set D .

The log-likelihood under this model is $\mathcal{L} = \sum_{n=1}^N \ln[p(\mathbf{t}_n)]$. The maximum-likelihood estimate for these parameters are: $\mathbf{m}_{ML} = \frac{1}{N} \sum_{n=1}^N \mathbf{t}_n$, $\sigma_{ML}^2 = \frac{1}{d-q} \sum_{i=q+1}^d \lambda_i$, and $\mathbf{W}_{ML} = \mathbf{U}_q (\Lambda_q - \sigma_{ML}^2 \mathbf{I})^{1/2}$, where the q columns in the $d \times q$ orthogonal matrix \mathbf{U}_q are the q dominant eigenvectors of the sample covariance matrix \mathbf{S} , and diagonal matrix Λ_q contains the corresponding q largest eigenvalues.

3 Sparse Probabilistic PCA

Sparsity is achieved in SPCA by adding an L_1 regularizer (Zou et al., 2006). Similarly, we add an L_1 regularizer by introducing a Laplacian prior to each element, \mathbf{W}_{ij} , of the transformation matrix \mathbf{W} , since Laplacian priors are equivalent to L_1 regularization (Chen et al., 1998; Williams, 1995; Figueiredo, 2001). The Laplacian density has the following form:

$$p(\mathbf{W}_{ij}|\lambda) = \frac{1}{2} \sqrt{\frac{2}{\lambda}} \exp(-\sqrt{\frac{2}{\lambda}} |\mathbf{W}_{ij}|) \quad (1)$$

where $|\cdot|$ is the absolute value operator. Assuming that the elements \mathbf{W}_{ij} are independent, the prior probability for \mathbf{W} is $p(\mathbf{W}|\lambda) = \prod_{i=1}^d \prod_{j=1}^q p(\mathbf{W}_{ij}|\lambda)$. And, the log joint distribution is shown below:

$$\begin{aligned} & \log(p(\mathbf{t}, \mathbf{W}, \mathbf{x}, \mathbf{m}, \sigma^2)) \quad (2) \\ &= \sum_{n=1}^N \left(-\frac{1}{2\sigma^2} (\mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \mathbf{m})' (\mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \mathbf{m}) \right. \\ & \quad \left. - \frac{1}{2} \mathbf{x}'_n \mathbf{x}_n - \frac{1}{2} \log \sigma^2 \right) \\ & \quad + \sum_{i=1}^d \sum_{j=1}^q \left(-\frac{1}{2} \log \lambda - \sqrt{\frac{2}{\lambda}} |\mathbf{W}_{i,j}| \right) \\ & \quad - \frac{1}{2} \mathbf{m}' \beta \mathbf{m} - (a-1) \log \sigma^2 - b(\sigma^{-2}) + const \end{aligned}$$

Here, we assume the prior for the data mean \mathbf{m} follows the Gaussian distribution centered at zero and with variance β . The precision σ^{-2} follows a Gamma distribution with parameters a and b , and $const$ is the additional constant term with respect to \mathbf{W} that is needed to normalize this into a valid probability density. Note that the first part is the log-likelihood for probabilistic PCA and the second part, the Laplacian prior, resulted in an L_1 regularization on \mathbf{W} (the $\sum_{i=1}^d \sum_{j=1}^q |\mathbf{W}_{ij}|$ term). Figure 1 displays the distribution for a Laplacian. Observe that values close to zero have high probabilities leading to sparse solutions.

Bayesian Formulation In this paper, we provide a Bayesian solution to sparse probabilistic PCA. As we introduce a prior for the transformation matrix \mathbf{W} , we

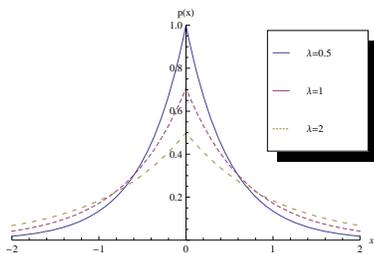


Figure 1: The Laplacian prior distribution.

need a closed form for the posterior distribution of \mathbf{W} . However, since the Laplace distribution is not a conjugate prior for the Gaussian distribution, we utilize an alternative prior – a two-level hierarchical decomposition of the Laplacian distribution. We apply a two-level hierarchy for the Laplacian similar to (Lange & Sinsheimer, 1993). The first level assumes a Gaussian prior, $p(\mathbf{W}_{ij}|z_{ij}) \sim \mathcal{N}(0, z_{ij})$. We set the mean of the Gaussian to zero because our goal is to sparsify \mathbf{W}_{ij} , and a small z_{ij} pushes \mathbf{W}_{ij} to zero. And, the second level is an exponential distribution hyperprior to the variance z_{ij} , $p(z_{ij}) = \frac{1}{\lambda} \exp(-\frac{z_{ij}}{\lambda})$, with $z_{ij} \geq 0$. In automatic relevance determination (MacKay, 1995), a Gaussian prior is introduced to automatically determine the importance, whereas here we introduce a Laplace prior to induce sparsity. Note that when the intermediate random variable z_{ij} is marginalized out, we obtain a Laplace distribution.

$$\begin{aligned} p(\mathbf{W}_{ij}|\lambda) &= \int p(\mathbf{W}_{ij}|z_{ij})p(z_{ij})dz_{ij} \\ &= \frac{1}{2} \sqrt{\frac{2}{\lambda}} \exp(-\sqrt{\frac{2}{\lambda}}|\mathbf{W}_{ij}|) \end{aligned} \quad (3)$$

To form a full Bayesian model, we also introduce prior distributions for the other parameters in the PPCA model: the prior for the mean \mathbf{m} follows a Gaussian distribution and the prior for the precision of the isotropic noise ϵ follows a gamma distribution.

$$p(\mathbf{m}) \sim \mathcal{N}(0, \beta^{-1}\mathbf{I}) \quad (4)$$

$$p(\sigma^{-2}) \sim \Gamma(\sigma^{-2}|a, b) \quad (5)$$

In summary, Figure 2 is a graphical model for our Bayesian formulation of sparse PPCA. The joint distribution of the data D and parameters θ for our sparse probabilistic PCA model is

$$\begin{aligned} p(D, \theta) &= \prod_{n=1}^N p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{W}, \mathbf{m}, \sigma^{-2}, \mathbf{z})p(\mathbf{x}_n) \\ &\quad p(\mathbf{W}|\mathbf{z})p(\mathbf{z})p(\mathbf{m})p(\sigma^{-2}). \end{aligned} \quad (6)$$

4 Variational Inference

It is computationally intractable to evaluate the marginal likelihood, $p(D) = \int p(D, \theta)d\theta$, where $\theta =$

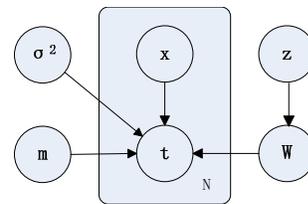


Figure 2: Graphical model for sparse PPCA.

$\{\theta_i\}$ represents the set of all parameters and latent variables. Variational methods allow us to approximate the marginal likelihood by maximizing a lower bound, $\mathcal{L}(Q)$, on the true log marginal likelihood (Bishop, 1999). $\ln p(D) = \ln \int p(D, \theta)d\theta = \ln \int Q(\theta) \frac{p(D, \theta)}{Q(\theta)} d\theta \geq \int Q(\theta) \ln \frac{p(D, \theta)}{Q(\theta)} d\theta = \mathcal{L}(Q(\theta))$, using Jensen's inequality. The difference between the log marginal $p(D)$ and the lower bound $\mathcal{L}(Q)$ is the Kullback-Leibler divergence between the approximating distribution $Q(\theta)$ and the true posterior $p(\theta|D)$. The idea is to choose a $Q(\theta)$ distribution that is simple enough that the lower bound can be tractably evaluated and flexible enough to get a tight bound. Here, we assume a distribution for $Q(\theta)$ that factorizes over all the parameters $Q(\theta) = \prod_i Q_i(\theta)$. For our model, this is

$$Q(\mathbf{x}, \mathbf{W}, z, \mathbf{m}, \sigma^{-2}) = Q(\mathbf{x})Q(\mathbf{W})Q(z)Q(\mathbf{m})Q(\sigma^{-2}) \quad (7)$$

The $Q_i(\theta_i)$ that minimizes the KL divergence over all factorial distributions is

$$Q_i(\theta_i) = \frac{\exp \langle \ln P(D, \theta) \rangle_{k \neq i}}{\int \exp \langle \ln P(D, \theta) \rangle_{k \neq i} d\theta_j} \quad (8)$$

Applying Equation 8 and the explicit form for $p(D, \theta)$ provided in Equation 6, we obtain

$$Q(\mathbf{x}) = \prod_{n=1}^N N(\mathbf{x}_n|\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}) \quad (9)$$

$$Q(\mathbf{m}) = N(\mu_{\mathbf{m}}, \Sigma_{\mathbf{m}}) \quad (10)$$

$$Q(\mathbf{W}) = \prod_{i=1}^d N(\mathbf{W}_{i,\cdot}|\mu_{\mathbf{W}_{i,\cdot}}, \Sigma_{\mathbf{W}_{i,\cdot}}) \quad (11)$$

$$Q(\sigma^{-2}) = \Gamma(\sigma^{-2}|c, d) \quad (12)$$

$$\begin{aligned} Q(z_{i,j}) &= \frac{1}{\sqrt{\pi z_{i,j} \lambda}} \exp\left(-\frac{1}{2z_{i,j}}(\mathbf{W}_{i,j})'(\mathbf{W}_{i,j})\right. \\ &\quad \left. - \frac{z_{i,j}}{\lambda} + \sqrt{\frac{2}{\lambda}}|\mathbf{W}_{i,j}|\right) \end{aligned} \quad (13)$$

The update equations we obtain for the variational sparse probabilistic PCA model are:

$$\mu_{\mathbf{x}_n} = \langle \sigma^{-2} \rangle \Sigma_{\mathbf{x}} \langle \mathbf{W}' \rangle (\mathbf{t}_n - \langle \mathbf{m} \rangle) \quad (14)$$

$$\Sigma_{\mathbf{x}} = (I + \langle \sigma^{-2} \rangle \langle \mathbf{W}' \mathbf{W} \rangle)^{-1} \quad (15)$$

$$\mu_{\mathbf{m}} = \langle \sigma^{-2} \rangle \Sigma_{\mathbf{m}} \sum_{n=1}^N (\mathbf{t}_n - \langle \mathbf{W} \rangle \langle \mathbf{x}_n \rangle) \quad (16)$$

$$\Sigma_{\mathbf{m}} = (\beta + N \langle \sigma^{-2} \rangle)^{-1} I \quad (17)$$

$$\mu_{\mathbf{w}_{i,\cdot}} = \langle \sigma^{-2} \rangle \Sigma_{\mathbf{w}_{i,\cdot}} \sum_{n=1}^N \langle \mathbf{x}_n \rangle (\mathbf{t}_n - \langle \mathbf{m} \rangle)_{i,1} \quad (18)$$

$$\Sigma_{\mathbf{w}_{i,\cdot}} = \left[\langle \sigma^{-2} \rangle \sum_{n=1}^N \langle \mathbf{x}_n \mathbf{x}_n' \rangle \text{dg}(\langle z_{i,\cdot} \rangle) + \mathbf{I} \right]^{-1} \text{dg}(\langle z_{i,\cdot} \rangle) \quad (19)$$

$$\mu_{z_{i,j}} = \frac{1}{2} \left(\lambda + \sqrt{2\lambda} |\mathbf{W}_{i,j}| \right) \quad (20)$$

$$c_{\sigma^{-2}} = \frac{Nd}{2} + a \quad (21)$$

$$d_{\sigma^{-2}} = \frac{1}{2} \sum_{n=1}^N \{ \|\mathbf{m}\|^2 + \text{Tr}(\langle \mathbf{W}' \mathbf{W} \rangle \langle \mathbf{x}_n \mathbf{x}_n' \rangle) + 2 \langle \mathbf{m} \rangle' \langle \mathbf{W} \rangle \langle \mathbf{x}_n \rangle - 2 \mathbf{t}_n' \langle \mathbf{W} \rangle \langle \mathbf{x}_n \rangle - 2 \mathbf{t}_n' \langle \mathbf{m} \rangle + \|\mathbf{t}_n\|^2 \} + b \quad (22)$$

Here, $\mathbf{w}_{i,\cdot}$ denotes the i th row of the transformation matrix \mathbf{W} , $\text{dg}(\cdot)$ converts a vector into a diagonal matrix, $z_{i,\cdot}$ is the vector formed from the z s at the i th row, and $\langle \cdot \rangle$ is the expectation of a random variable. The moments needed in the update equations are: $\langle \mathbf{x} \rangle = \mu$, $\langle \mathbf{x} \mathbf{x}' \rangle = \Sigma + \mu \mu'$ for a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, which applies to the variables \mathbf{W} , \mathbf{x} , \mathbf{m} , and $\langle \sigma^{-2} \rangle = a/b$ for a Gamma distribution $\Gamma(\sigma^{-2} | a, b)$.

In summary, variational sparse probabilistic PCA proceeds as follows. We first initialize the parameters, then update those parameters according to equations (15) to (23) until convergence (i.e., the change in the lower bound $\mathcal{L}(Q(\theta))$ is small). In our experiments, we initialize W with conventional PCA, set $a = 10^{-3}$, $b = 10^{-3}$, and $\beta^{-1} = 10^{-3}$ obtaining broad priors. We set our condition for convergence tolerance to be 10^{-3} . The λ parameter controls the sparsity of SPPCA, similar to the sparsity parameter in SPCA. Figure 1 displays the distribution for a Laplacian with varying values of λ . Note that in our model, the smaller λ is, the sparser the solution. Unlike in SPCA where the parameter λ is provided by the user, in sparse probabilistic PCA, we can automatically learn the hyperparameter λ through a type II maximum likelihood (Bishop, 2006). Applying type II maximum likelihood provide us with an update equation for λ as

$$\langle \lambda \rangle = \frac{1}{dq} \sum_{i=1}^d \sum_{j=1}^q z_{i,j}. \quad (23)$$

5 Alternative Prior Models for Enforcing Sparsity

We introduced the Laplacian prior earlier because it is the probabilistic counterpart for L1-norm regularization utilized in SPCA. However, there are several other ways to induce sparsity in W . In this paper, we investigate two other prior models that result in sparse solutions: an inverse-Gaussian prior, and a non-

informative Jeffrey's prior.

5.1 Inverse-Gaussian Prior

Another prior model is the inverse-Gaussian prior. In (Caron & Doucet, 2008), the inverse-Gaussian has been shown to produce sparse models for the regression problem. We apply a two-level hierarchy with $p(w|z)$ modeled as a zero-mean Gaussian distribution, followed by an inverse-Gaussian prior. The probability density function for an inverse-Gaussian is:

$$\sqrt{\frac{a}{2\pi x^3}} \exp -\frac{a(x-b)^2}{2x\mu^2}. \quad (24)$$

Using this prior, changes the function $Q(z_{i,j})$ for variational inference in Equation 13 to:

$$Q(z_{i,j}) = \frac{b_z e^{-\frac{\mathbf{w}_{i,j}^2}{2z_{i,j}} - \frac{a_z}{b_z} - \frac{a_z(z-b_z)^2}{2zb_z^2}}}{2z^2 \sqrt{\frac{a_z}{\mathbf{w}_{i,j}^2 + a_z}} K_1 \left(\sqrt{\frac{a_z(\mathbf{w}_{i,j}^2 + a_z)}{b_z^2}} \right)} \quad (25)$$

where $K_a(b)$ is the modified Bessel function of the second kind with order a and evaluated at b . The update equation for $\mu_{z_{i,j}}$ in Equation 20 now becomes:

$$\mu_{z_{i,j}} = e^{a_z/b_z} \sqrt{\frac{2}{\pi}} K_0 \left(\sqrt{\frac{a_z(\mathbf{w}_{i,j}^2 + a_z)}{b_z^2}} \right). \quad (26)$$

As shown in Figure 3, the inverse-Gaussian distribution presents high density in regions near zero. The

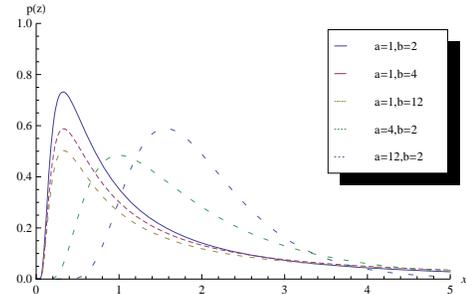


Figure 3: The inverse-Gaussian distribution with different parameter settings.

figure also shows that varying the hyperparameters a_z and b_z changes the level of sparsity. Similar to the Laplacian model, we learn the hyperparameters by type II maximum likelihood giving us the following update equations:

$$\frac{1}{a_z} = \frac{1}{dq} \sum_{i=1}^d \sum_{j=1}^q \left(\frac{1}{z_{i,j}} - \frac{1}{\mu_z} \right) \quad (27)$$

$$b_z = \frac{1}{dq} \sum_{i=1}^d \sum_{j=1}^q z_{i,j} \quad (28)$$

5.2 Jeffrey’s Prior

The Jeffrey’s prior has been shown to enforce sparsity in classification and regression models (Figueiredo, 2001). Similar to the Laplacian and inverse-Gaussian models, we assume a two-level hierarchy with $p(w|z)$ modeled as a Gaussian distribution with mean zero:

$$p(w|z) = \frac{1}{\sqrt{2\pi z}} \exp -\frac{w^2}{2z}, \quad (29)$$

followed by the Jeffrey’s prior. The non-informative Jeffrey’s prior is the square root of the Fisher information, which is:

$$\text{Jeffrey’s}(z) \sim \sqrt{E_{p(w|z)}((\frac{\partial}{\partial z} \lg p(w|z))^2)} = \frac{1}{z}. \quad (30)$$

An example of the Jeffrey’s prior for a Gaussian distribution is shown in Figure 4. Note that this “den-

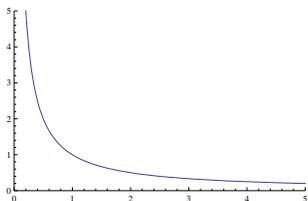


Figure 4: Jeffrey’s prior.

sity”² gives high probabilities to values close to zero leading to sparsity. A computational advantage of the Jeffrey’s prior compared to the Laplacian and inverse-Gaussian is that it has no hyperparameters that needs to be learned or tuned. The update equation for $\mu_{z_{i,j}}$ in Equation 20 now becomes:

$$\mu_{z_{i,j}} = \mathbf{W}_{i,j}^2. \quad (31)$$

5.3 Comparison of Sparsity Characteristics for the Different Models

In this section, we provide a comparison of the different prior models for achieving sparsity. The log joint probability distribution in Equation 6 can in general be expressed as:

$$\begin{aligned} & \log(p(\mathbf{t}, \mathbf{W}, \mathbf{x}, \mathbf{m}, \sigma^2)) \\ &= -\frac{1}{2\sigma^2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}\mathbf{x}_n)' (\mathbf{t}_n - \mathbf{W}\mathbf{x}_n) \\ & \quad - \text{penalty} + \text{constant}. \end{aligned} \quad (32)$$

The first term measures the fitness between our model with the observed data; the penalty term penalizes for model complexity to avoid over-fitting; and constant are terms that do not depend on \mathbf{W} . Table 1 summarizes the penalty terms resulting from the three prior models: Laplace, Jeffrey’s and inverse-Gaussian.

²As opposed to the other two prior models, the Jeffrey’s prior is not a proper density.

Table 1: Penalty Terms

Prior	Penalty Term
Laplace	$\sum_i \sum_j \lambda \mathbf{W}_{ij} $
Jeffrey’s	$\sum_i \sum_j \log(z_{ij})$
Inverse Gaussian	$-\frac{1}{2} \sum_i \sum_j (\log(\mathbf{W}_{ij}^2 + \lambda) + \log(K_1 \left(\frac{\sqrt{\lambda} \sqrt{\mathbf{W}_{ij}^2 + \lambda}}{\mu} \right)))$

Figure 5 displays the contour plots of the penalty functions in two dimensions resulting from each of the different prior models. For Jeffrey’s, we replace z_{ij} with W_{ij}^2 to plot them all with respect to W . Our objective function Equation 32 optimizes for fitness and minimizes for model complexity. Geometrically, the fit term are elliptical contours, and the solution is the first place that this contour touches the penalty contour. The Laplace prior results in sparse solutions if the ellipse contour touches the corner first. In other cases, the ellipse will touch the side of the contour resulting in non-sparse solutions. Note that the penalty contour of Jeffrey’s prior (which has sharp corners and sides close to the origin) has the highest chance of leading to sparse solutions compared to the Laplacian or inverse-Gaussian prior. The plot for the inverse-Gaussian prior shows the contours for varying values of its hyperparameters. This shows that the inverse-Gaussian provides a general model which can adjust to have sparse properties close to that of Jeffrey’s prior, the Laplacian (L1), or L2 (ridge-regression (Hoerl, 1962)).

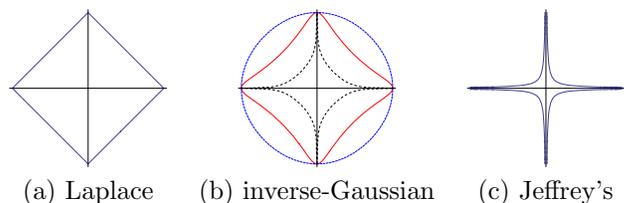


Figure 5: Contour plots for the penalty functions resulting from the different prior models: (a) Laplace, (b) inverse-Gaussian, and (c) Jeffrey’s.

Another plot that shows sparsity properties of regularizers is a plot of the solution of the penalized problem as a soft-threshold estimator (Chen & Donoho, 1994; Fan & Li, 2001). These plots are displayed in Figure 6. The 45° line is the solution without shrinkage. The Laplacian translates the solution by a constant factor and truncates at zero. The Jeffrey’s prior has a smooth penalty with non-zero coefficients asymptotically approaching the solution without shrinkage. The inverse-Gaussian also has a smooth penalty and the characteristics vary with the hyperparameters. For some hyperparameters, it does not truncate coefficients to zero

but instead smoothly pushes them to small values close to zero, unlike the Laplacian and Jeffrey’s prior.

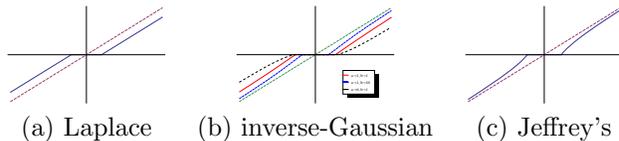


Figure 6: Soft-thresholding plots for: (a) Laplacian, (b) inverse-Gaussian, and (c) Jeffrey’s prior.

6 Experiments

In this section, we verify whether or not our sparse probabilistic PCA (SPPCA) model results in a sparse solution with performance comparable to SPCA, show that our Bayesian formulation is able to automatically determine the sparsity level, and compare the performances of the three different sparse prior models.

We report \mathbf{W} which provides the PC directions derived by our model and the corresponding adjusted variance resulting from each PC. Results presented here reflects \mathbf{W} with each column normalized to one for easier interpretation and arranged according to decreasing variance. Since the principal components obtained from sparse PPCA are no longer orthogonal to each other, we calculate the adjusted variance explained by each PC similar to (Zou et al., 2006). We first rank the PCs according to variance. Then, we apply Gram-Schmidt (Golub & Loan, 1996) to get an orthogonal set of adjusted vectors. The adjusted variance (AVar or AV) explained by each PC is the variance along the adjusted vector divided by the total variance of the data. We also report the cumulative adjusted variance (CVar or CV) which is simply the sum of the current adjusted variance for PC i and all previous PCs $i-1$ to 1. In all our experiments, we initialize the \mathbf{W} matrix for SPPCA and SPCA with PCA. The SPCA results here are obtained using Matlab code from (Sjöstrand, 2005).

6.1 Synthetic Data

To get a good understanding of our approach, we first test it on synthetic data. We generate 200 samples of a ten variable data similar to that in (Zou et al., 2006), where there are two underlying factors. The first two factors V_1 and V_2 are equally important. The first four features are related to V_1 , the next four are related to V_2 . We construct our synthetic data as follows: $D(:, 1 : 4) = V_1 + N(0, 1)$, $D(:, 5 : 8) = V_2 + N(0, 1)$, and $D(:, 9 : 10) = N(0, 1)$, where $V_1 \sim N(0, 100)$ and $V_2 \sim N(0, 98)$. We can see that the first four columns

of the data is controlled by V_1 , while the second four columns of the data is controlled by V_2 and the last two columns of the data is controlled by noise, $N(0, 1)$. In the result from SPPCA, we expect to see that the first two PCs are composed of only those columns from the V_1 and V_2 . And the first PC should be composed of the first four columns of the data, since it has the largest variance.

The result of the \mathbf{W} matrix for all three prior models on this synthetic data are similar and is simply reported as SPPCA in Table 2. The results for SPCA and PCA are also shown. We can see that our proba-

Table 2: Result for Synthetic Data by SPPCA

	SPPCA		SPCA		PCA	
	PC1	PC2	PC1	PC2	PC1	PC2
t_1	0.5001	0	-0.5001	0	-0.5000	0.0110
t_2	0.5002	0	-0.5002	0	-0.5001	0.0111
t_3	0.4998	0	-0.4998	0	-0.4996	0.0108
t_4	0.5000	0	-0.4999	0	-0.4998	0.0109
t_5	0	-0.5002	0	0.5001	0.0111	0.5000
t_6	0	-0.4999	0	0.4999	0.0110	0.4998
t_7	0	-0.5002	0	0.5002	0.0109	0.5000
t_8	0	-0.4998	0	0.4998	0.0110	0.4997
t_9	0	0	0	0	-0.0000	0.0001
t_{10}	0	0	0	0	-0.0000	-0.0000
AVar%	0.5971	0.4028	0.5971	0.4028	0.5972	0.4027
CVar%	0.5971	0.9999	0.5971	0.9999	0.5972	0.9999

bilistic sparse PCA model is able to correctly find the variables associated with PC1 and PC2, and remove the noise variables. The \mathbf{W} matrix found is almost the same as that for SPCA with similar performance in terms of adjusted variance. Here, the λ parameter in SPCA was set to 10^{-3} . Note that unlike SPCA, we learn the sparsity parameter λ automatically for our probabilistic models. Moreover, we initially set the number of PCs to 9 for SPPCA and our SPPCA models learned correctly that there are only two PCs (the other PCs have all zero entries).

6.2 Real-World Benchmark Data

We verify our models on three real-world benchmark data sets with different sizes: glass, chart, and face data. The glass data from the UCI repository (Asuncion & Newman, 2007) has nine attributes, six classes, and 214 samples. Chart data, also from the UCI repository, has 600 samples with 60 features. Face data is from the UCI KDD repository (Hettich & Bay, 1999) and consists of 640 face images from 20 people. Each person has 32 images with an image resolution of 32 by 30. We then remove the missing data to form a 960 by 624 data matrix. Note that this data set has a higher dimension than the number of data points. We show that SPPCA can handle such a data set.

The glass data is a small data set which allows us to report the actual \mathbf{W} matrix found by our approach. Table 3 presents the transformation matrix \mathbf{W} obtained

with SPPCA and Jeffrey’s prior, and that of SPCA. In SPPCA, the number of PCs is set to $9 - 1$. For fair comparison, we set the sparsity level of SPCA for each PC based on the sparsity chosen by SPPCA. Note that the number of nonzero coefficients per PC in this table is the same. The results show that SPPCA results in sparse PCA solutions and the cumulative adjusted variance is slightly better than that of SPCA. Note too that even though we set the number of PCs to be 8, SPPCA automatically determines that only seven PCs are needed.

Table 3: The \mathbf{W} matrices obtained by SPPCA-Jeffrey’s and SPCA for the glass data.

SPPCA-Jeffrey’s								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
t_1	-.144	-.047	0	-.014	0	-.008	0	0
t_2	.060	.050	.055	-.049	0	.030	0	0
t_3	0	-.129	0	-.032	0	-.010	0	0
t_4	.097	.067	-.044	0	0	0	.034	0
t_5	.054	-.020	.046	.070	0	0	-.009	0
t_6	.055	-.018	-.092	0	.010	0	-.024	0
t_7	-.140	-.065	0	.028	0	.005	0	0
t_8	.052	.099	0	-.018	-.008	-.037	0	0
t_9	-.036	0	-.027	.011	-.078	0	0	0
AV	.277	.228	.155	.129	.102	.059	.042	0
CV	.277	.505	.660	.789	.891	.950	.992	.992
SPCA								
t_1	-.700	-.149	0	-.093	-.005	0	0	0
t_2	.218	.256	.407	-.514	0	.544	0	0
t_3	-.064	-.748	0	-.490	-.001	-.344	0	0
t_4	.105	.088	-.339	0	0	.000	.902	0
t_5	.143	-.161	.403	.676	0	0	0	0
t_6	.141	0	-.746	.081	0	.306	-.402	0
t_7	-.589	.159	0	.048	0	0	0	0
t_8	.247	.542	-.010	-.146	0	-.701	-.155	0
t_9	0	-.020	0	0	-1.00	0	0	0
AV	.245	.204	.153	.120	.111	.070	.060	0
CV	.245	.449	.602	.722	.833	.903	.963	.963

In Figures 7, 8, and 9, we display the cumulative adjusted variance results versus number of PCs kept for SPPCA with Laplacian prior, SPPCA with inverse-Gaussian prior, and SPPCA with Jeffrey’s prior in bold lines, and the SPCA results in dashed lines with sparsity levels set to be the same as each of the three corresponding prior models for the glass, chart and face data respectively. In these plots, we compare the performance of SPPCA versus SPCA in capturing variance, the PCA objective. We also compare the performance of the different prior models. These plots show that SPPCA is slightly better than SPCA for the glass and chart data and is much better than SPCA for the face data (which has high dimensions). We also observe that among the different prior models, Jeffrey’s has the best performance, Laplacian the worst, and inverse-Gaussian somewhere in between these two consistently for all three data sets. Our Bayesian SPPCA models automatically learn the level of sparsity per PC. To check the performance of the different models in maximizing variance with respect to sparsity, we plot the cumulative variance versus the percentage of the number of nonzero weights. Figure 10 displays such a plot for the chart data. Plots for the

glass and face data have similar characteristics and are not shown here due to space limitations. These plots show that Jeffrey’s is the best, followed by the inverse-Gaussian, and the Laplacian last.

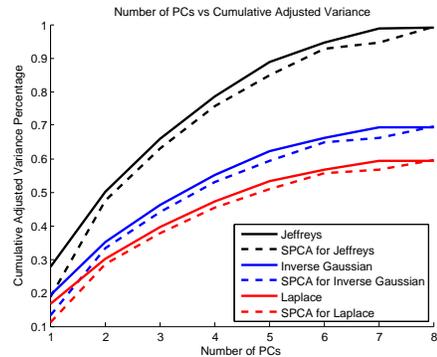


Figure 7: Cumulative adjusted variance versus number of PCs kept for the glass data.

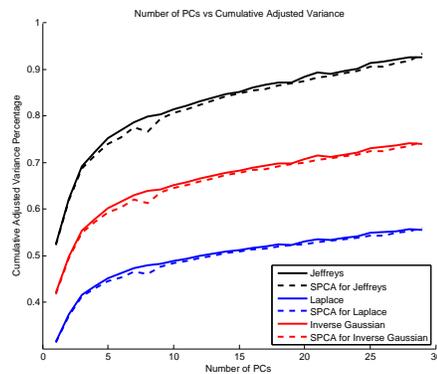


Figure 8: Cumulative adjusted variance versus the number of PCs kept for the chart data.

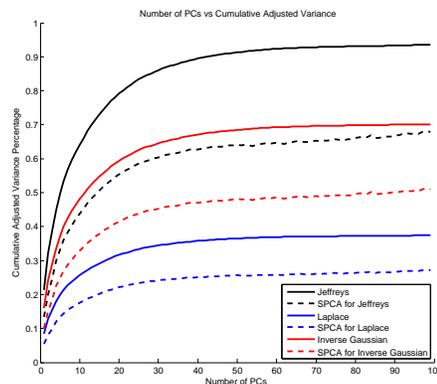


Figure 9: Cumulative adjusted variance versus number of PCs kept for the face data.

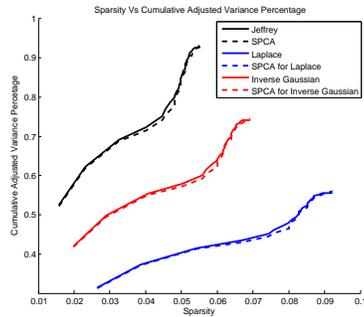


Figure 10: Cumulative adjusted variance vs. percentage of nonzero weights for chart data.

7 Summary

We have developed a probabilistic Bayesian model for sparse PCA, gaining interpretability of the PCs and the benefits of a probabilistic model. We achieve sparsity by adding a Laplacian prior to the transformation matrix \mathbf{W} , which results in an equivalent L_1 constraint. Since the Laplacian distribution is not a conjugate prior for the Gaussian distribution, we utilize an alternative prior, a two-level hierarchical decomposition of the Laplacian distribution: normal with zero-mean and unknown variance in the first level, and an exponential hyperprior on the variance in the second level. We then applied variational inference to learn our model. Our experiments on synthetic and benchmark data confirmed that SPPCA can find sparse PCs similar to SPCA. Moreover, SPPCA gains the benefit of a Bayesian model in being able to automatically determine the level of sparsity, which is a parameter that requires tuning in SPCA. Besides the Laplacian prior, we explored other ways of inducing sparsity using an inverse-Gaussian prior and a non-informative Jeffrey's prior. We provided a comparison of the three models and experiments showed that the Jeffrey's prior resulted in the best performance, Laplacian the worst and inverse-Gaussian somewhere in between.

Acknowledgments We thank NSF IIS-0347532.

References

- Archambeau, C., & Bach, F. (2008). Sparse probabilistic projections. *NIPS*.
- Asuncion, A., & Newman, D. (2007). UCI ML repository.
- Bishop, C. (2006). Pattern recognition and machine learning. *Springer*.
- Bishop, C. M. (1999). Variational principal components. *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, 1, 509–514.
- Cadima, J., & Jolliffe, I. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22, 203–214.
- Caron, F., & Doucet, A. (2008). Sparse bayesian nonparametric regression. *ICML*, 88–95.
- Cawley, G., Talbot, N., & Girolami, M. (2007). Sparse multinomial logistic regression via bayesian L_1 regularization. *NIPS* (pp. 209–216).
- Chen, S., & Donoho, D. (1994). *Basis pursuit* (Technical Report). Statistics Dept., Stanford University, CA.
- Chen, S., Donoho, D., & Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computation*, 20, 33–61.
- Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96, 1348–1360.
- Figueiredo, M. A. T. (2001). Adaptive sparseness using jeffreys prior. *NIPS*, 679–704.
- Golub, G., & Loan, C. V. (1996). *Matrix computations*. Johns Hopkins University Press.
- Hettich, S., & Bay, S. D. (1999). UCI KDD archive.
- Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chem. Eng. Progress*, 58, 54–59.
- Jolliffe, I. (1986). Principal component analysis. *Springer Verlag, New York*.
- Jolliffe, I. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22, 29–35.
- Jolliffe, I. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12, 531–547.
- Lange, K., & Sinsheimer, J. (1993). Normal/independent distributions and their applications in robust regression. *J. of Comp. and Graphical Statistics*, 2, 175–198.
- MacKay, D. J. C. (1995). Probable networks and plausible predictions: a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6, 469–505.
- Roweis, S. (1997). *EM algorithms for PCA and sensible PCA* (Technical Report). California Institute of Technology, Computation and Neural Systems.
- Sjöstrand, K. (2005). Matlab implementation of LASSO, LARS, the elastic net and SPCA. Ver. 2.0.
- Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *JMLR*, 1, 211–244.
- Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, 61, 611–622.
- Williams, P. (1995). Bayesian regularization and pruning using a laplace prior. *Neural Comp.*, 7, 117–143.
- Zou, H., & Hastie, T. (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. *Technical report, Statistical Department, Stanford University*.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society Series B*, 67, Part 2, 301–320.
- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15, 262–286.