Sparse Structured Probabilistic Projections for Factorized Latent Spaces

Xinquan Qu, Xinlei Chen
State Key Lab of CAD&CG
College of Computer Science, Zhejiang University
Hangzhou, China
{04dzjsqxq,endernewton}@gmail.com

ABSTRACT

Building a common representation for several related data sets is an important problem in multi-view learning. CCA and its extensions have shown that they are effective in finding the shared variation among all data sets. However, these models generally fail to exploit the common structure of the data when the views are with private information. Recently, methods explicitly modeling the information into shared part and private parts have been proposed, but they presume to know the prior knowledge about the latent space, which is usually impossible to obtain. In this paper, we propose a probabilistic model, which could simultaneously learn the structure of the latent space whilst factorize the information correctly, therefore the prior knowledge of the latent space is unnecessary. Furthermore, as a probabilistic model, our method is able to deal with missing data problem in a natural way. We show that our approach attains the performance of state-of-art methods on the task of human pose estimation when the motion capture view is completely missing, and significantly improves the inference accuracy with only a few observed data.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Algorithms, Experimentation, Performance

Keywords

Multi-view Learning, Factorized Latent Space, Sparse Structured Projections

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

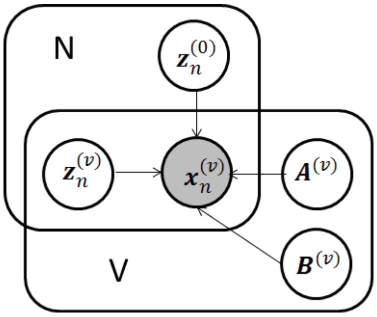
1. INTRODUCTION

In machine learning, we usually meet cases where there are two or more disparate but related data sets. For example, the human pose data is composed of video information and motion capture information; in image retrieval applications, the data is consisted of various image features and surrounding texts. Effective consolidation of these data sets has been proved to be beneficial for many computer vision tasks. Particularly in this paper, we define data consolidation as to find a common latent representation for all data sets. Furthermore, this problem appears naturally in the multi-view learning problem, where the multiple data sets are refereed as different views of one object.

Canonical Correlation Analysis (CCA) aims to find linear transformations which maximize two views' correlation. And there has been work to put CCA into the more flexible probabilistic framework [2]. Other extensions include learning nonlinear transformations [9] in the RKHS, and learning a sparse projection matrix for easy interpretation [6]. Beside models based on CCA, there have been several nonlinear methods proposed [15, 4, 13, 17], which may be more robust than Kernel CCA [9] in noisy cases. However, they do not capture the view-dependent information, and therefore either totally fail to represent them, or mix them with the information shared by all views.

Some methods explicitly accounting the dependencies and independencies have been proposed. They factorize the latent space into shared part across all views and private part of each view. Among these approaches, some could be characterized as deterministic type [14, 7], and are proposed for predicting missing view based on observed views. One limitation of these methods is that they presume to know at least one complete view. Furthermore, these approaches all fail to utilize the partial observed information in views.

On the other hand, methods on factorizing the latent space into shared part and private parts in a probabilistic framework have been proposed [3, 1, 8]. Particularly, [8] firstly extracts the shared information across views by CCA, then learns the private information from the residuals. And the work of [3] can be regarded as its kennel extension. Since these methods use CCA or NCCA to capture the shared information, they inevitably inherit the defats of CCA. Another approach [1] imposes a sparse structure on the projection matrices for better interpretation, but this assumption may not work for missing view prediction task. What's more, these methods generally assume to know the



The corresponding graphical model for the latent variable learning of data $\mathbf{x}_n$. The data in each view $\mathbf{x}_n^{(v)}$ is a mix of two independent continuous components, the shared one $\mathbf{z}_n^{(0)}$ and the private one $\mathbf{z}_n^{(v)}$.

Figure 1: Graphical Model

dimension of the latent space, which is impractical in real-world applications.

In this paper, we build our model in the framework of the probabilistic interpretation of CCA, which is outlined by [2] and further extended by [8, 1]. Compared with previous approaches towards shared/private factorization, our model has three main contributions: (1). while previous extensions of CCA are generally sensitive to highly correlated noise, our model is more robust; (2). our model could learn the dimension of the latent space automatically without prior knowledge; (3). as a probabilistic model, our approach has an innate mechanism to deal with missing data problem. Detailed model will discussed in the later sections.

The remainder of this paper is organized as follows. In the next section, our model will be introduced. A parameter estimation and inference procedure will be provided in section 3 . In Section 4, we will give a through review of related methods. Section 5 is our experimental part. Finally, we will give the conclusion and discuss the future work in Section 6.

2. STRUCTURED PROJECTIONS FOR FACTORIZED LATENT SPACES

2.1 Proposed Model

The graphical model of our model is shown in Fig. 1. It is capable of dealing with the multi-view case, and explicitly considers the effects of private information in each view. Each data $\mathbf{x}_n$ has V representations as $\mathbf{x}_n^{(1)}, \mathbf{x}_n^{(2)}, \ldots, \mathbf{x}_n^{(V)}$. Furthermore, the data $\mathbf{x}_n$ in each view $\mathbf{x}_n^{(v)} \in \mathbb{R}^{P_v}$ is generated as a mix of a shared latent continuous vector $\mathbf{z}_n^{(0)} \in \mathbb{R}^{D_0}$, and a view-dependent (private) continuous latent vector $\mathbf{z}_n^{(v)} \in \mathbb{R}^{D_v}, v \in \{1, 2, \ldots, V\}$, such that:

$$\begin{aligned}
&\mathbf{x}_n^{(v)} = \mathbf{A}^{(v)}\mathbf{z}_n^{(0)} + \mathbf{B}^{(v)}\mathbf{z}_n^{(v)} + \mu^{(v)} + \varepsilon^{(v)} \in \mathbb{R}^{P_v} \\
&\mathbf{A}^{(v)} \in \mathbb{R}^{P_v \times D_0}, \mathbf{B}^{(v)} \in \mathbb{R}^{P_v \times D_v} \\
&\mathbf{z}_n^{(0)} \sim \mathcal{N}(\mathbf{0}_{D_0}, \Lambda^{(0)}), \mathbf{z}_n^{(v)} \sim \mathcal{N}(\mathbf{0}_{D_v}, \Lambda^{(v)}),
\end{aligned} \tag{1}$$

where $\mathbf{A}^{(v)}$ and $\mathbf{B}^{(v)}$ are projection matrices, $\mu^{(v)}$ is the view dependent mean, and $\varepsilon^{(v)} \sim \mathcal{N}(\mathbf{0}_{P_v}, \sigma^{(v)}\mathbf{I}_{P_v})$ is the white noise in view v. Here $\mathbf{0}_{p\times q}/\mathbf{1}_{p\times q}$ is the all-zero/all-one matrix in $\mathbb{R}^{p\times q}$, and $\mathbf{I}_p$ is the identity matrix in $\mathbb{R}^{p\times p}$, respectively.

To obtain a low dimensional latent variable, we choose a simple yet effective way as to impose a column sparsity structure on the projection matrices. In this way, with sufficient number of columns, the model can automatically determine the number of nonzero columns of $\mathbf{A}^{(v)}$,$\mathbf{B}^{(v)}$ and then adapt them to the data. To achieve the column sparsity, we impose an Automatic Relevance Determination (ARD) [18] prior on the columns of $\mathbf{A}^{(v)}$,$\mathbf{B}^{(v)}$ as follows:

$$\begin{aligned}
\mathbf{A}_j^{(v)} &\sim \mathcal{N}(\mathbf{0}_{P_v}, \rho_j^{(v)}\mathbf{I}_{P_v}), \qquad j \in \{1, 2, \ldots, D_0\} \\
\mathbf{B}_j^{(v)} &\sim \mathcal{N}(\mathbf{0}_{P_v}, \gamma_j^{(v)}\mathbf{I}_{P_v}), \qquad j \in \{1, 2, \ldots, D_v\}
\end{aligned}$$

On the other hand, we directly impose the ARD prior on the latent variables $\mathbf{z}^{(v)}$ to make them low dimensional. Therefore, the covariance matrices $\mathbf{\Lambda}^{(0)}, \mathbf{\Lambda}^{(1)}, \ldots, \mathbf{\Lambda}^{(V)}$ are diagonals. Note that if we restrict them to be nonzero diagonal, then the model reduces to Factor analysis. Furthermore, if we set $\mathbf{\Lambda}^{(V)}$ as identity, then the model reduces to Probabilistic PCA[19].

If the initial dimension of latent variables is set to be large, the model tends to over-fit since the residual variance $\sigma^{(v)}$ will approach zero. As a remedy to this problem, we impose a conjugate gamma prior on the inverse residual variance as $\left(\sigma^{(v)}\right)^{-1} \sim \mathcal{G}(a, b)$. Here (a, b) could be regarded as regularization parameters.

2.2 Compact Reformulation of the Model

Before we give a detailed parameter estimation and inference procedure in section 3, we would like to first rewrite the above model in a more compact form. We denote the data in each view, the projection matrices and the latent variables as follows:

$$\begin{aligned}
\mathbf{X}^{(v)} &= \left(\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \ldots, \mathbf{x}_N^{(v)}\right), \\
\mathbf{C}^{(v)} &= \left(\mathbf{A}^{(v)}, \mathbf{0}_{P_1\times D_1}, \ldots, \mathbf{B}^{(v)}, \ldots, \mathbf{0}_{P_V\times D_V}\right), \\
\mathbf{z}_n &= \left(\mathbf{z}_n^{(0)T}, \mathbf{z}_n^{(1)T}, \ldots, \mathbf{z}_n^{(V)T}\right)^T, \\
\mathbf{Z} &= (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N),
\end{aligned}$$

where N is the number of observations. The generative model can then be reformulated as follows:

$$\begin{aligned}
&\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}_D, \mathbf{\Lambda}), \\
&\mathbf{x}_n^{(v)}|\mathbf{z}_n \sim \mathcal{N}\left(\mathbf{C}^{(v)}\mathbf{z}_n, \sigma^{(v)}\mathbf{I}_P\right), \\
&\mathbf{\Lambda} = diag\left(\Lambda^{(0)}, \Lambda^{(1)}, \ldots, \Lambda^{(V)}\right) \\
&\mathbf{C}_j^{(v)} \sim \mathcal{N}\left(\mathbf{0}, \lambda_j^{(v)}\mathbf{I}\right), \quad j \in \{1, 2, \ldots, D\},
\end{aligned} \tag{2}$$

where $D = \sum_{v=0}^{V} D_v$ and $P = \sum_{v=1}^{V} P_v$. Note that we incorporate the column variance ρ and γ into τ. Additionally, we abuse the notation that if $\mathbf{C}_j^{(v)} = 0$, then $\lambda_j^{(v)} = 0$. The reason for this notation is that, $\mathbf{C}^{(v)}$ is initially dense, and we wish to learn a sparse column projection by setting columns not fitting to data to zero.

3. PARAMETER ESTIMATION AND INFERENCE

In this section we first derive an algorithm to estimate the parameters and latent variables. After that we give a inference procedure in case that some data of the measurements are missing.

3.1 Parameter Estimation

A traditional approach for parameter estimation is the renowned Expectation-Maximization (EM) algorithm, in which we treat $\mathbf{Z}$ as the missing variables, and other random variables as parameters to be estimated. We however, choose to directly take the Maximum Posteriori (MAP) of both $\mathbf{Z}$, the projection matrices $\mathbf{C}^{(v)}$ and the residual variance $\sigma^{(v)}$[1].

Now we write the log-likelihood function as follows:

$$\begin{aligned} L &= \log P(\mathbf{X}, \mathbf{Z}, \mathbf{C}, \Lambda, \tau, \sigma) \\ &= \sum_{v=1}^{V} \sum_{n=1}^{N} \log P\left(\mathbf{x}_n^{(v)} | \mathbf{z}_n, \mathbf{C}^{(v)}, \sigma^{(v)}\right) \\ &+ \sum_{i=1}^{D} \log P(\mathbf{Z}_{i,:}) + \sum_{v=1}^{V}\sum_{j=1}^{D} \log P(\mathbf{C}_j^{(v)}) + \sum_{v=1}^{V} \log P(\sigma^{(v)}), \end{aligned} \tag{3}$$

where $\mathbf{Z}_{i,:}$ denotes the i-th row of $\mathbf{Z}$. And each item expand as:

$$\begin{aligned} &\log P(\mathbf{x}_n^{(v)} | \mathbf{z}_n, \mathbf{C}^{(v)}, \sigma^{(v)}) = \\ &-\frac{1}{2}\log(2\pi) - \frac{1}{2}P_v \log(\sigma^{(v)}) - \frac{1}{2\sigma^{(v)}}\left\|\mathbf{x}_n^{(v)} - \mu^{(v)} + \mathbf{C}^{(v)}\mathbf{z}_n\right\|^2 \\ &\log P(\mathbf{Z}_{i,:}) = -\frac{1}{2}N\log(2\pi) - \frac{1}{2}N\log(\lambda_i) - \frac{1}{2\lambda_i}\|\mathbf{Z}_{i,:}\|^2 \\ &\log P(\mathbf{C}_j^{(v)}) = -\frac{1}{2}P_v \log(2\pi) - \frac{1}{2}P_v\log(\tau_j^{(v)}) - \frac{1}{2\tau_j^{(v)}}\left\|\mathbf{C}_j^{(v)}\right\|^2 \\ &\log P(\sigma^{(v)}) = -(a-1)\log\left(\sigma^{(v)}\right) + a\log(b) - \frac{b}{\sigma^{(v)}} - \Gamma(a). \end{aligned} \tag{4}$$

We directly use maximum likelihood to estimate λ and τ. By setting the derivative to be zero, we have:

$$\begin{aligned} \frac{dL}{d\lambda_i} = 0 \Rightarrow \qquad \lambda_i &= \frac{1}{N}\|\mathbf{Z}_{i,:}\|^2 \\ \frac{dL}{d\tau_j^{(v)}} = 0 \Rightarrow \quad \tau_j^{(v)} &= \frac{1}{N}\left\|\mathbf{C}_j^{(v)}\right\|^2 \end{aligned} \tag{5}$$

Note that the above variance is directly related to the magnitude norm of rows of $\mathbf{Z}$ or columns of $\mathbf{C}$ respectively, and the ones not fitting to the data well will be driven to zero. This is the internal mechanics why our model could do automatic dimension determination.

By substituting the λ and τ in Eq.(3) with Eq.(4) and (5), and further discarding the constants, we then formulate it as a minimization problem:

$$\begin{aligned} \min_{\{\mathbf{C}^{(v)}, \sigma^{(v)}\}_{v=1}^{V}, \mathbf{Z}} & \sum_{v=1}^{V} (\frac{1}{2}P_v N + a - 1)\log\left(\sigma^{(v)}\right) \\ &+ \sum_{v=1}^{V} \frac{1}{\sigma^{(v)}}\left(\frac{1}{2}\left\|\mathbf{X}^{(v)} - \mu^{(v)}\mathbf{1}_{N\times P_v}^T - \mathbf{C}^{(v)}\mathbf{Z}\right\|_{Fro}^2 + b\right) \\ &+ N\sum_{i=1}^{D}\log\|\mathbf{Z}_{i,:}\| + \sum_{v=1}^{V}\sum_{j=1}^{D} P_v \log\left\|\mathbf{C}_j^{(v)}\right\|. \end{aligned} \tag{6}$$

We can directly solve problem Eq. (6) by the Majorization-Minimization (MM) algorithm [10]. For numerical stability, we can slightly modify the objective function of Eq. (6) by replacing the last two terms:

$$\begin{aligned} N\sum_{i=1}^{D}\log\|\mathbf{Z}_{i,:}\| &\to N\sum_{i=1}^{D}\log(\|\mathbf{Z}_{i,:}\| + \alpha), \\ \sum_{v=1}^{V}\sum_{j=1}^{D} P_v \log\left\|\mathbf{C}_j^{(v)}\right\| &\to \sum_{v=1}^{V} P_v \sum_{j=1}^{D}\log\left(\left\|\mathbf{C}_j^{(v)}\right\| + \alpha\right). \end{aligned} \tag{7}$$

where α is a regularization parameter. We denote the solution obtained in the t-th iteration as $\mathbf{Z}^{(t)}$, $\mathbf{C}^{(v,t)}$. In the $(t+1)$-th iteration, due to the concavity property, we can bound the last two terms in Eq. (eq:minreg) by the following inequation:

$$\begin{aligned} &\sum_{i=1}^{D}\log\left(\|\mathbf{Z}_{i,:}\| + \alpha\right) \le \\ &\qquad \sum_{i=1}^{D}\left[\log\left(\left\|\mathbf{Z}_{i,:}^{(t)}\right\| + \alpha\right) - \frac{\|\mathbf{Z}_{i,:}\| - \left\|\mathbf{Z}_{i,:}^{(t)}\right\|}{\left\|\mathbf{Z}_{i,:}^{(t)}\right\| + \alpha}\right] \\ &\sum_{i=1}^{D} P_v \mathrm{log}\left(\left\|\mathbf{C}_j^{(v)}\right\| + \alpha\right) \le \\ &\qquad \sum_{i=1}^{D} P_v\left[\log\left(\left\|\mathbf{C}_j^{(v,t)}\right\| + \alpha\right) - \frac{\left\|\mathbf{C}_j^{(v)}\right\| - \left\|\mathbf{C}_j^{(v,t)}\right\|}{\left\|\mathbf{C}_j^{(v,t)}\right\| + \alpha}\right]. \end{aligned} \tag{8}$$

Thus in the $(t+1)$-th iteration, what we need is to solve a weighted version of the Eq. (7):

$$\begin{aligned} \min_{\{\mathbf{C}^{(v)}, \sigma^{(v)}\}_{v=1}^{V}, \mathbf{Z}} & \sum_{v=1}^{V} (\frac{1}{2}P_v N + a - 1)\log\left(\sigma^{(v)}\right) \\ &+ \sum_{v=1}^{V} \frac{1}{\sigma^{(v)}}\left(\frac{1}{2}\left\|\mathbf{X}^{(v)} - \mu^{(v)}\mathbf{1}_{N\times P_v}^T - \mathbf{C}^{(v)}\mathbf{Z}\right\|_{Fro}^2 + b\right) \\ &+ N\sum_{i=1}^{D}\frac{\|\mathbf{Z}_{i,:}\|}{\left\|\mathbf{Z}_{i,:}^{(t)}\right\| + \alpha} + \sum_{v=1}^{V} P_v \sum_{j=1}^{D}\frac{\left\|\mathbf{C}_j^{(v)}\right\|}{\left\|\mathbf{C}_j^{(v,t)}\right\| + \alpha} \end{aligned} \tag{9}$$

According to [10], the MM algorithm is guaranteed to converge to a local optimum. In each iteration, the object function is not convex. Here we simply optimize one variable while fixing the other variables and the sub-optimization problem turns out to be convex. Moreover, note that although the optimization is not convex w.r.t. $\sigma^{(v)}$, it is convex w.r.t. its inverse.

[1] Since $\max_{\mathbf{Z}} \log P(\mathbf{X}, \mathbf{Z}) \ge E(\log P(\mathbf{X}, \mathbf{Z})|\mathbf{X})$, therefore taking the MAP of $\mathbf{Z}$ is a reasonable alternate to EM

3.2 Inference

An important application of multi-view learning is to infer unobserved views based on observed ones [7, 14]. And very frequently we may have partial information about the missing views. Unfortunately, many deterministic methods, such as [5, 14, 7] fail to utilize the partial observed data since their models can only deal with sound data. Here we treat it as a missing data problem and wish to utilize the incomplete information in inference.

We treat the latent variables $\mathbf{Z}$ and $\mu^{(v)}$ as parameters, and the missing data $\mathbf{X}_{mis}$ as latent variables. The EM algorithm is then applied. Denoting $\mathbf{X} = (\mathbf{X}_{ob}, \mathbf{X}_{mis})$, the corresponding object function is:

$$\min_{(\{\mu^{(v)}\}_{v=1}^{V}, \mathbf{Z})} \quad \sum_{i=1}^{D} \frac{\|\mathbf{Z}_{i,:}\|^2}{\lambda_i} + \mathbb{E}_{X_{mis}} \left[\sum_{v=1}^{V} \frac{1}{2\sigma^{(v)}} \left\| \mathbf{X}^{(v)} - \mu^{(v)} \mathbf{1}_{N \times P_v}^{T} - \mathbf{C}^{(v)} \mathbf{Z} \right\|^2 \middle| \mathbf{Z}, \mathbf{X}_{ob} \right] \tag{10}$$

Note that $\mathbb{E}_{\mathbf{X}_{mis}}(\mathbf{X}_{mis}^{(v)} | \mathbf{Z}, \mathbf{X}_{ob})$ can be calculated according to $\mathbf{C}^{(v)}\mathbf{Z}$, and $\mathbb{E}_{\mathbf{X}_{mis}}(\mathbf{X}_{mis}^{(v)} \mathbf{X}_{mis}^{(v)T} | \mathbf{Z}, \mathbf{X}_{ob})$ can be deduce from $\mathbf{C}^{(v)}\mathbf{Z}\mathbf{Z}^T\mathbf{C}^{(v)}$. We omit the detailed computations here.

4. RELATED WORK

There has been some work taking emphasis on the structure of the latent space [7, 14, 3, 1, 8, 12]. That is, they all explicitly model the shared information across all views and private information for each view. Those models can all be formulated in a general generative model as

$$\mathbf{x}_n^{(v)} = f(\mathbf{z}_n^{(0)}) + g^{(v)}(\mathbf{z}_n^{(v)}) + \varepsilon^{(v)},$$

where f,$g^{(v)}$ could either be linear [7, 1, 8], or nonlinear functions, such as the functions in RKHS [3, 14, 12]. Existing work in the literature can also be characterized in two categories, namely probabilistic framework [1, 8, 12], or deterministic framework [7, 14, 3].

Both linear and kernel extensions on CCA to model the private information of views have been proposed in [3]. The key idea is that as CCA incorporates private information into the residuals, it explicitly extracts the view-dependent effects from the residuals. To be specific, it first applies CCA or kernel CCA to the data to calculate f and $\mathbf{z}_n^{(0)}$, and then optimize $g^{(v)}$ and $\mathbf{z}_n^{(v)}$ from the residuals. While it seems to be a good idea, this model inevitably inherits the drawbacks of CCA, i.e., sensitive to highly correlated noise.

Different from optimizing $f, \mathbf{z}^{(0)}$ and $g^{(v)}, \mathbf{z}^{(v)}$ step by step, a model iteratively optimizing the linear projections and latent variables has been proposed in [8]. In each inner iteration, it first models the shared information f and $\mathbf{z}^{(0)}$, and then extract the private information from the residuals. Unfortunately, this model essentially inheritress the drawbacks of CCA just as [3]. Their relationship is similar to the relationship between Maximum Likelihood (ML) and the EM optimization procedure for probabilistic CCA, therefore [8] behaves sensitively to correlated-noise as CCA.

In the model proposed by [1], the linear projection functions f and $g^{(v)}$ are simultaneously optimized. The key contribution of their model is to impose priors on the elements of projection matrices to incur sparsity, i.e. the ARD prior, which results in a model easy to interpret. However, their model only works well when the projection matrices are really sparse in the generative model, which is scarcely the case in practice. Furthermore, this model, together with the above two models all require to know the latent dimension beforehand, which is usually not possible in real applications.

The model introduced in [12] is quite similar to the kernel version proposed in [3], while it actually uses an ARD polynomial kernel function as the covariance function. Although the prior knowledge of the latent dimension is unnecessary here, the model still inherits limitations of CCA as [3, 8] do.

The model proposed by [14] encourages the private-shared factorization to be non-redundant by explicitly adding a penalty term to sKIE [17] and sGPLVM [4]. Although the resulting model has been shown to yield more accurate results in the context of human pose estimation, the optimization is computationally expensive. Furthermore, extension from two views to multiple views is non-trivial since the number of of shared/private latent spaces that need to be explicitly modeled grows exponentially with the number of views.

[7] casts the latent variable learning problem as a matrix decomposition task, and uses recent advances in the sparse coding in helping learning their representation. They introduces ℓ_p norm on the column of dictionary matrix for each view, and the ℓ_p norm also impose structured sparsity on the rows of the coefficient matrix. While this seems a good idea, their method lacks mechanism to deal with views with missing data, and thus implicitly require the training views to be complete. Furthermore, in case there is auxiliary information of target view, i.e. data motion view is partly available in human pose estimation, it is not straightforward to incorporate them into the inference process and thus the result is doomed to degrade.

Compared with the above models, our method avoids the noise sensitiveness of CCA. Furthermore, it can automatically detect the latent dimension without prior information. Finally, our method could deal with missing data in a natural way, which is either critical or beneficial in many multi-view learning applications.

5. EXPERIMENTS

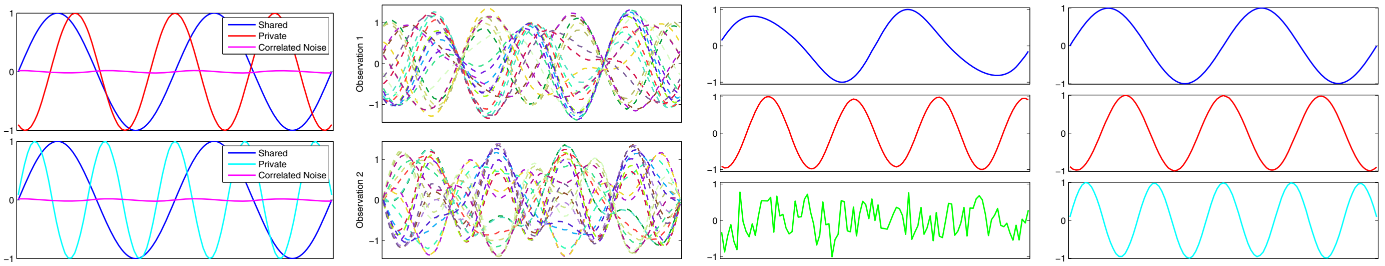
In this section, we conduct experiments to evaluate the effectiveness of our approach. We first give synthetic examples for illustration of the properties of our method, and then apply it to the real application of human pose estimation.

5.1 Latent Space Factorization

To study the property of our method, we conducted the same toy experiment in [14, 7]. We show that our method can also correctly factorize a latent space into shared part and private part, even when the noise is highly correlated. We generated 100 points of two data streams containing one shared and one private information per view. Specifically, we used sinusoidal signals at different frequencies such that:

$$\mathbf{z}^{(0)} = \sin(2\pi t), \quad \mathbf{z}^{(1)} = \cos(\pi^2 t), \quad \mathbf{z}^{(2)} = \cos(2\sqrt{5}\pi t)$$

with t from the same interval $(-1, 1)$. Therefore, the ground-truth latent space is composed of 3 dimensions, 1 shared and 2 private. We then randomly projected the joint shared-private spaces $[\mathbf{z}^{(0)}, \mathbf{z}^{(1)}]$ and $[\mathbf{z}^{(0)}, \mathbf{z}^{(2)}]$ into two 20-dimensional



(a) Ground-truth of latent spaces

(b) Observed views

(c) Dimensions recovered by CCA

(d) 3-dimensional latent space recovered by our method

Figure 2: Synthetic example 2: latent spaces factorization

spaces respectively. Then both correlated noise $0.02\sin(3.6\pi t)$ and zero-mean Gaussian noise of variance 0.01 are added.

The ground-truth latent spaces together with correlated noise are depicted in Fig. 2(a), and the input views are depicted in Fig. 2(b). Fig. 2(c) shows the result obtained by CCA when the latent space is set by a priori to be 3-dimensional. We see that CCA recovers the shared signal, but mixes correlated noise as well. This phenomenon complies with the theoretical claim that approaches based on CCA would tend to be correlated-noise sensitive [8]. Fig. 2(d) depicts the reconstructed latent spaces for both views with our method, which clearly demonstrates the shared-private factorization.

5.2 Pose Estimation

We then apply our method to the problem of the human pose estimation on HumanEva dataset [16], which consists of Videos recorded by 7 cameras and motion capture data describing the $3D$ locations of joints of a human skeleton. The video and the motion capture data have been synchronized and thus can be perceived as different representations for an object. Therefore, the task of pose estimation, i.e, infer the 3D poses from the 2D images, can be naturally formulated as a multi-view learning problem.

In our experiments, several methods that directly perform a regression from the images features to 3D poses were compared, namely linear regression (Lin-Reg), Gaussian Process regression with a linear kernel (GP-Lin). And we also implemented the Gaussian Process regression with an RBF kernel (GP-Rbf) as a competition method. Besides, we also implemented the method proposed in [7], which has been shown to outperform most multi-view learning methods for the task of human pose estimation on HumanEva. For clarity, we name their model FLSSP in short. By directly comparing our model to FLSSP, redundant comparisons with other existing algorithms [14, 11, 3, 4] are avoided. We use cross-validation to determine the regularization parameter for each algorithm.

We consider the walking and jogging video sequences of the first and second subject seen from the BW1 and BW2 camera. Since the objects move in circles, we used the first loop for training, and the remaining for testing. Each image is represented using a 100 dimensional integral HOG descriptoror 84 dimensional PHOG descriptor. In each test case, we have all together three views: two from the video recorded by BW1 and BW2, and one from the motion capture data. The standardized mean squared error (SMSE) is chosen as the metric function, since it's scale-invariant and thus easy for reproduction.

We treat the inference task as a missing data problem, in which we uniformly chose some features of the data points to be missing. In each trial, we varied the missing rate ζ, (the number of missing elements over the total number of elements in the data matrix), from 1 to 0.1. For each value, we sampled 20 different testing sets on the original inference framework. Fig. 3 and 4 show the mean error of the 20 trials.

We see that when the motion data is completely missing ($\zeta = 1$), our method is among the most promising methods. Since there is high ambiguity across the multiple views [3], the outstanding performance demonstrates the power of our model to factorize the intrinsic structure of the latent space.

Furthermore, our method effectively captured the structured information from the observed data, i.e. the error dropped dramatically as the missing rate decreased. For example, the estimation error reduces to about 1/2 with around 10% visible motion capture data in walking people with HOG features.

6. CONCLUSION

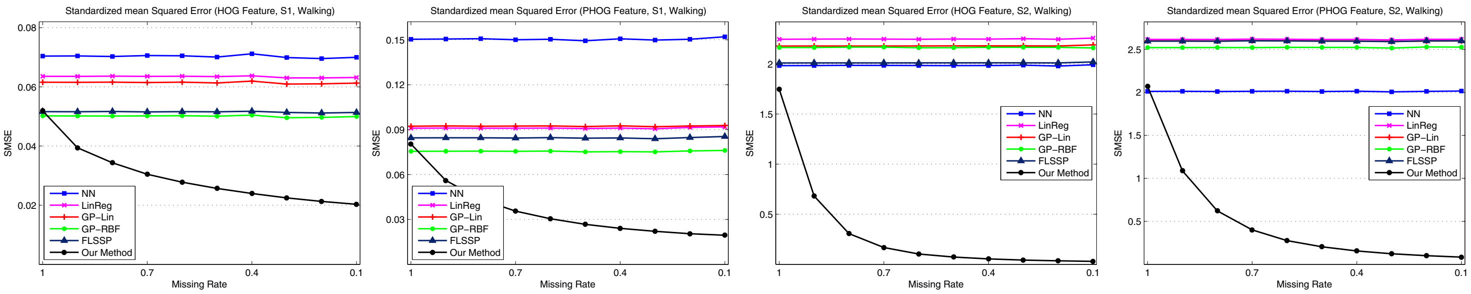
We have proposed a probabilistic model aiming at factorizing the latent space into shared part across views and private part for each view. By imposing ARD priors, we learn column sparse projection matrices. We have demonstrated the effectiveness of our approach on both synthetic data and the human pose estimation task. Currently, we are making two extensions. Since many multi-view learning data sets such as HumanEva are temporal, we are trying to capture the embedded time evolution in analogous to Kalman Filter. On the other hand, we are trying to develop a full Bayesian parameter estimation procedure.

7. REFERENCES

[1] C. Archambeau and F. Bach. Sparse probabilistic projections. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, volume 21. MIT Press, 2008.

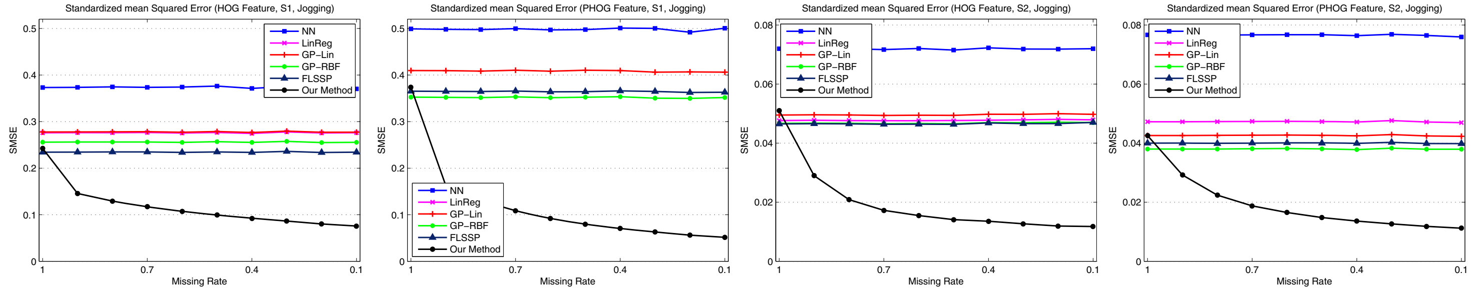
[2] F. Bach and M. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley, 2005.

[3] C. Ek, J. Rihan, P. Torr, G. Rogez, and N. Lawrence. Ambiguity modeling in latent spaces. In *Proceedings of the International Conference on Machine Learning for*



(a) S1_Walking+HOG (b) S1_Walking+PHOG (c) S2_Walking+HOG (d) S2_Walking+PHOG

Figure 3: SMSE of missing features with varying missing rate ζ: Walking



(a) S1_Jogging+HOG (b) S1_Jogging+PHOG (c) S2_Jogging+HOG (d) S2_Jogging+PHOG

Figure 4: SMSE of missing features with varying missing rate ζ: Jogging

Multimodal Interaction, pages 62–73. Springer-Verlag, 2008.

[4] C. Ek, P. Torr, and N. Lawrence. Gaussian process latent variable models for human pose estimation. In *Proceedings of the International Conference on Machine learning for Multimodal Interaction*, pages 132–143. Springer-Verlag, 2007.

[5] J. Fan, W. G. Aref, A. K. Elmagarmid, M.-S. Hacid, M. S. Marzouk, and X. Zhu. Multiview: Multilevel video content representation and retrieval. *J. Electronic Imaging*.

[6] D. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, pages 1–23, 2008.

[7] Y. Jia, M. Salzmann, and T. Darrell. Factorized Latent Spaces with Structured Sparsity. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, volume 23. MIT Press, 2010.

[8] A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72(1-3):39–46, 2008.

[9] M. Kuss and T. Graepel. The geometry of kernel canonical correlation analysis. Technical report, Max Planck Institution, 2003.

[10] K. Lange, D. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000.

[11] N. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, volume 16, pages 329–336. MIT Press, 2004.

[12] G. Leen. Context assisted information extraction. *PhD thesis, Univerisyt of the West of Scotland,*, 2008.

[13] R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007.

[14] M. Salzmann, C. Ek, R. Urtasun, and T. Darrell. Factorized orthogonal latent spaces. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[15] A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning shared latent structure for image synthesis and robotic imitation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, volume 18, pages 1233–1240. MIT Press, 2006.

[16] L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical report, Brown Univertsity, 2006.

[17] L. Sigal, R. Memisevic, and D. Fleet. Shared kernel information embedding for discriminative inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[18] A. Tipping. Analysis of sparse Bayesian learning. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 383–390. MIT Press, 2002.

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