Factorized Orthogonal Latent Spaces

Mathieu Salzmann EECS & ICSI, UC Berkeley
Carl Henrik Ek EECS & ICSI, UC Berkeley
Raquel Urtasun TTI Chicago
Trevor Darrell EECS & ICSI, UC Berkeley

Abstract

Existing approaches to multi-view learning are particularly effective when the views are either independent (i.e, multi-kernel approaches) or fully dependent (i.e., shared latent spaces). However, in real scenarios, these assumptions are almost never truly satisfied. Recently, two methods have attempted to tackle this problem by factorizing the information and learn separate latent spaces for modeling the shared (i.e., correlated) and private (i.e., independent) parts of the data. However, these approaches are very sensitive to parameters setting or initialization. In this paper we propose a robust approach to factorizing the latent space into shared and private spaces by introducing orthogonality constraints, which penalize redundant latent representations. Furthermore, unlike previous approaches, we simultaneously learn the structure and dimensionality of the latent spaces by relying on a regularizer that encourages the latent space of each data stream to be low dimensional. To demonstrate the benefits of our approach, we apply it to two existing shared latent space models that assume full dependence of the views, the sGPLVM and the sKIE, and show that our constraints improve the performance of these models on the task of pose estimation from monocular images.

1 Introduction

Many machine learning problems inherently involve multiple views, where a view is broadly defined as any sensor stream of a scene or event. The different views can arise either from the same sensor type or from different modalities. Kernel combination approaches to multi-view learning (Bach et al., 2004; Sonnenburg et al., 2006) have recently become very popular since they provide a convenient way of combining information from multiple data streams. They are particularly effective when the views are independent, since the errors occurring in a view can then be corrected by the other views.

In contrast to multi-kernel approaches, methods have been introduced to take advantage of the dependencies in the data. These techniques typically rely on learning latent spaces that capture the relevant information shared by all the views, and are most effective when the streams have significant dependencies. The best-known example is Canonical Correlation Analysis (CCA) (Kuss and Graepel, 2003), which learns latent representations of the views whose correlation is maximal. While this, in essence, is a good idea, it can result in trivial solutions in the presence of highly correlated noise, as shown in Section 3.1. Recently, non-linear shared latent variable models that do not suffer from this problem have been proposed: the shared Gaussian process latent variable model (sGPLVM) (Shon et al., 2006; Ek et al., 2007; Navaratnam et al., 2007), which minimizes the reconstruction error between the data and the model's prediction, and the shared kernel information embedding (sKIE) (Sigal et al., 2009), which maximizes the mutual information between the latent representation and each input stream.

However, in real scenarios, information in the views is typically neither fully independent nor fully correlated, and thus multi-kernel (Bach et al., 2004; Sonnenburg et al., 2006) and shared latent space (Kuss and Graepel, 2003; Ek et al., 2007; Navaratnam et al., 2007; Sigal et al., 2009) techniques are not optimal. Typically, in the latter case, information relevant to only a single stream will be mixed with the shared information, making inference complicated. Only few approaches have tried to factorize the information and learn separate latent spaces for modeling the shared (i.e., correlated across the views) and private (i.e., independent between the views) components of the input signals (Archambeau and Bach, 2008; Klami and Kaski, 2008; Ek et al., 2008; Leen, 2008). However, (Archam-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

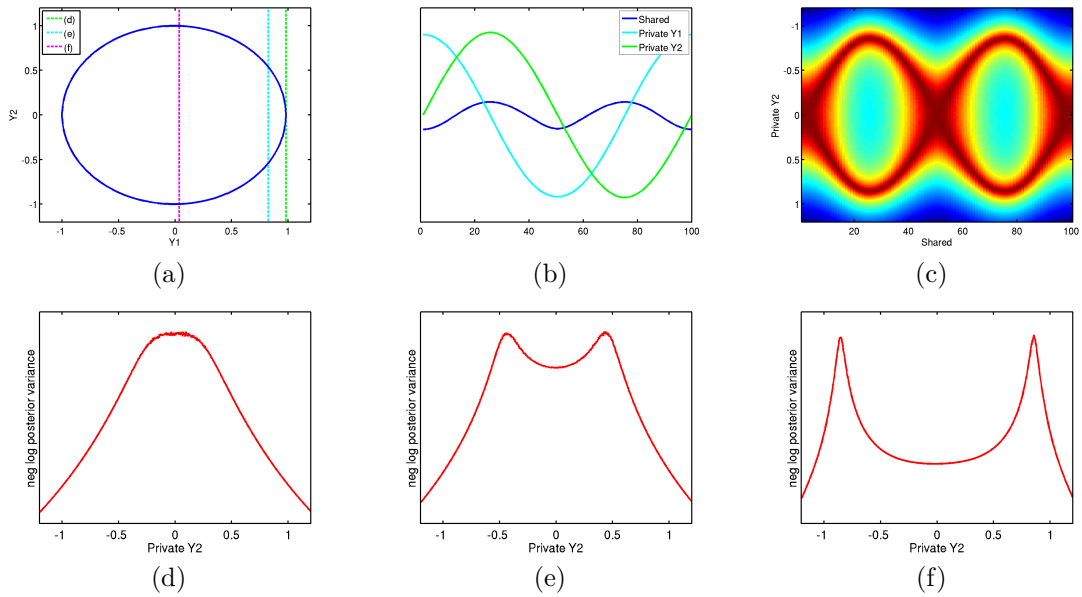

Figure 1: **Latent Factorization**: (a) Circle used to generate the observed data. The observed data was generated by projecting the circle onto the $x-$axis for $\mathbf{Y^{(1)}}$ and onto the $y-$axis for $\mathbf{Y^{(2)}}$. Knowing the location on one axis gives information about the location on the second one, which indicates the presence of shared information. Green dashed lines indicate slices of the circle used for further experiments. (b) Recovered factorized latent spaces. (c) Negative log variance over one private space given the shared location. Given the shared location of each training point, we regularly sampled the private latent space for $\mathbf{Y^{(2)}}$. Given a pair of shared and private coordinates, we computed the negative log. variance of the model's prediction, which indicates where the model expects the true location to be. (d)-(f) 3 slices of the plot in (c) taken at fixed shared locations. These slices correspond to the dashed lines in plot (a). Note that the maxima along each slice correspond to the data points on the circle.

beau and Bach, 2008; Klami and Kaski, 2008) build on probabilistic CCA and therefore do not generalize to non-linear mappings. Furthermore, while (Ek et al., 2008; Leen, 2008) overcome this issue, they are typically initialized with CCA, and thus suffer from its inherent weaknesses, which we illustrate below. Finally, these methods require choosing a priori the dimensionality of the latent space, or use cross-validation to find it, which is computationally expensive.

In this paper, we propose a method to learn shared and private latent spaces that are inherently disjoint by introducing orthogonality constraints. Furthermore, following (Geiger et al., 2009), we discover the structure and dimensionality of the latent representation of each data stream by encouraging it to be low dimensional, while still allowing to generate the data. Combined together, these constraints encourage finding factorized latent spaces that are non-redundant, and that can capture the shared-private separation of the data. We demonstrate the effectiveness of our approach by applying it to two existing models, the sGPLVM (Shon et al., 2006) and the sKIE (Sigal et al., 2009), and show significant performance improvement over the original models, as well as over the existing shared-private factorizations (Ek et al., 2008; Leen, 2008) in the context of pose estimation.

2 Factorized Latent Spaces

We are interested in learning low dimensional representations of multi-view data. To this end, we would like to factorize the latent space into the information shared across all data streams and the independent or private information of each stream. We propose to learn this factorization and infer the dimensionality of each space by introducing orthogonality constraints between the latent spaces, as well as rank constraints that encourage lower dimensional representations of each data stream.

2.1 Factorized Orthogonal Latent Spaces (FOLS)

To have a minimal factorization, we would like the shared and private latent spaces to be non-redundant. Similarly, we would like to penalize the redundancy of different private spaces, and thus encourage representing information common to them in the shared space. Here, we propose to enforce this by using orthogonality constraints.

In addition to factorizing the shared and private information, we would like to estimate the latent spaces dimensionalities at the same time as we learn their

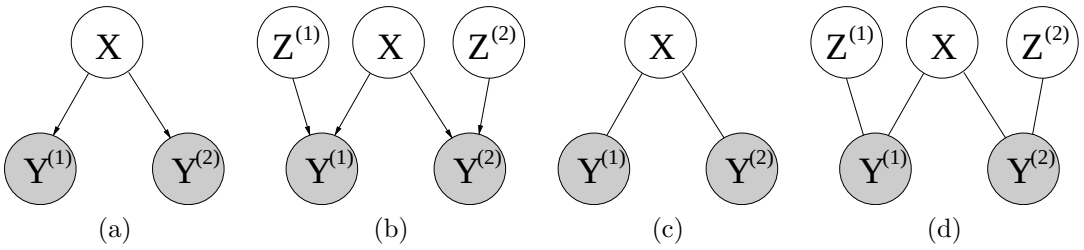

Figure 2: **Graphical models** (a) sGPLVM. (b) FOLS-GPLVM. (c) sKIE. (d) FOLS-KIE

structure. This avoids having to either choose the latent dimensionalities a priori or estimate them by cross-validation. To this end, we initialize the private spaces to their corresponding data stream and the shared space to their concatenation, and introduce a regularizer that encourages each joint shared-private latent space to be low dimensional.

Finally, since both the above terms tend to bring the latent coordinates close to zero, we incorporate a term in the optimization that encourages conservation of the energy of the spectrum of the data.

More formally, let $\mathbf{Y}^{(i)} = [\mathbf{y}_1^{(i)}, \cdots, \mathbf{y}_N^{(i)}]^T$ be the set of observations from a single view i, with $1 \leq i \leq V$. Additionally, let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T$ be the latent space shared across different views, $\mathbf{Z}^{(i)} = [\mathbf{z}_1^{(i)}, \cdots, \mathbf{z}_N^{(i)}]^T$ be the private space for i-th view, and $\mathbf{M}^{(i)} = [\mathbf{m}_1^{(i)}, \cdots, \mathbf{m}_N^{(i)}]^T$ be the joint shared-private latent space for each view, with $\mathbf{m}_j^{(i)} = [\mathbf{x}_j, \mathbf{z}_j^{(i)}]$. By imposing the above-mentioned constraints as a soft penalty, a FOLS model can be learned by minimizing

$$\mathcal{L} = L + \alpha \underbrace{\sum_i \left(||\mathbf{X}^T \cdot \mathbf{Z}^{(i)}||_F^2 + \sum_{j>i} ||(\mathbf{Z}^{(i)})^T \cdot \mathbf{Z}^{(j)}||_F^2 \right)}_{\text{Orthogonality}} + \gamma \underbrace{\sum_i \phi(s_i)}_{\text{Low dimensionality}} + \eta \underbrace{\sum_i (E_0^{(i)} - \sum_j s_{i,j}^2)^2}_{\text{Energy conservation}}, \quad (1)$$

where $\|\cdot\|_F$ represents the Frobenius norm of a matrix, s_i are the singular values of $\mathbf{M}^{(i)}$, and $E_0^{(i)}$ is the energy of stream i. α, γ and η are scalars that set the relative influence of the different terms. L is the loss function of the particular model into which we introduce our factorization constraints. As described later for the sGPLVM and the sKIE models, L can represent the square loss, or the negative mutual information between each joint latent space and its corresponding data stream.

As can be observed from Eq. 1, our orthogonality constraints are encoded as minimizing the Frobenius norm of the inner product between latent spaces, which results in minimizing the scalar products between each pair of latent dimensions. Note that, since both shared and private latent spaces are initialized from the observed data, they can potentially have twice as many dimensions as the streams, which would prevent the spaces from being orthogonal. However, in addition to being continuous and differentiable, the Frobenius norm has the advantage of also being minimized when latent dimensions shrink to zero. Therefore, the penalty associated with our orthogonality constraints is also minimized when some of the latent dimensions become negligible.

Furthermore, we would like to learn the intrinsic dimensionalities of the individual joint shared-private latent spaces, $\mathbf{m}^{(i)}$. This can be done by introducing a regularizer that encourages $\mathbf{M}^{(i)}$ to be low rank. Penalizing the rank of a matrix results in a difficult non-smooth optimization problem. Different relaxations of the rank minimization problem have been proposed. The most widely used relaxation is the nuclear norm, also known as the trace norm, which is a particular instance of the Schatten p-norm with $p = 1$. The Schatten p-norm is defined as

$$||\mathbf{M}^{(i)}||_p = \left(\sum_i s_{i,j}^p \right)^{1/p}.$$

When $p < 2$ then the Schatten p-norm can be used for the rank minimization problem since it will encourage sparsity of the singular values.

The nuclear norm is typically used since it is a convex function, and thus when the loss function is convex the problem has a unique solution. In our case however, the loss functions of the sKIE and sGPLVM models are non-convex, and more sophisticated regularizers than the nuclear norm can be used. In particular we want to drive the smaller singular values faster to zero, since they represent the noise of the data. Following (Geiger et al., 2009), we use a logarithmic penalty

$$\phi(s_{i,j}) = \sum_j \log(1 + \beta s_{i,j}^2), \quad (2)$$

where β is a constant.

Since both previous terms are minimized when all latent dimensions go to zero, we encourage the energy of the data spectrum to remain constant. This

can be achieved by minimizing for each data stream the squared difference between the energies of the latent and observation spaces, $(E_o^{(i)} - \sum_j s_{i,j}^2)^2$. The energy of the observation space can be computed as $E_o^{(i)} = \sum_j p_{i,j}^2$, with $p_{i,j}$ the singular values of $\mathbf{Y}^{(i)}$.

Fig. 1 shows a data-set exemplifying where two 1D observation spaces $\mathbf{Y^{(1)}}$ and $\mathbf{Y^{(2)}}$ were generated by projecting a circle onto the $x-$ and $y-$axis. In general, knowing the location along one axis of the circle does not completely disambiguate the location along the other one. However, it reduces its possible values to at most two. In Figure 1 (b), we show the latent representation recovered with our approach. The model learned a 3D latent space factorized into a 1D shared space and two 1D private spaces corresponding to each data stream. The shared space encodes the information that is common for both axes, while the private spaces reflect the remaining ambiguities of the observed data given a shared location. To demonstrate this, we took the shared latent location corresponding to each training data point of $\mathbf{Y^{(1)}}$ and uniformly sampled locations in the private space of $\mathbf{Y^{(2)}}$. Fig. 1 (c) depicts the negative log of the variance for these locations. As can be observed from the plot, there are clear modes in the variance. These modes depict the fact that the location in $\mathbf{Y^{(1)}}$ either completely disambiguates the location in $\mathbf{Y^{(2)}}$, as is the case at the largest $x-$axis value on the circle and shown in plot (d), or leaves an ambiguity between two locations, as is the case in the middle of the circle and shown in plots (e) and (f).

Our factorization scheme is general and can be applied to a variety of existing shared latent space models. To demonstrate this, we apply our constraints to two recently developed models, the sGPLVM (Shon et al., 2006; Ek et al., 2007; Navaratnam et al., 2007) and the sKIE (Sigal et al., 2009). In the remaining of the section we briefly describe the FOLS versions of these models.

2.2 FOLS-GPLVM

The shared Gaussian process latent variable model (sGPLVM) (Shon et al., 2006; Ek et al., 2007; Navaratnam et al., 2007) is an extension of the GPLVM (Lawrence, 2005) to learn a latent space that is shared across feature streams. It is a directed model, where the mapping from the latent space to the data streams is modeled as a product of independent Gaussian processes. Its graphical model is depicted in Fig. 2 (a).

To learn a FOLS version of the GPLVM, we introduce private spaces into the sGPLVM as shown in Fig. 2 (b). For each data stream, we model the mapping from the joint latent space to the observation stream as a product of Gaussian processes

$$p(\mathbf{Y}^{(i)}|\mathbf{Z}^{(i)}, \mathbf{X}) = \prod_{d=1}^{D_i} \mathcal{N}(\mathbf{Y}_{:,d}^{(i)}|0, \mathbf{K}^{(i)}) \ , \tag{3}$$

where $\mathbf{Y}_{:,d}^{(i)}$ is the d-th column in $\mathbf{Y}^{(i)}$, $\mathbf{y}_j^{(i)} \in \Re^{D_i}$, and $\mathbf{K}^{(i)}$ is the covariance matrix which is typically defined in terms of a kernel function. $\mathbf{K}^{(i)}$ is restricted to be positive definite, and thus all Mercer kernels are valid. Here, we use the sum of an RBF, a bias and a white noise kernel, such that

$$k(\mathbf{m}_i, \mathbf{m}_j) = \theta_1^{(i)} \exp\left(-\frac{\|\mathbf{m}_i - \mathbf{m}_j\|_2^2}{2(\theta_2^{(i)})^2}\right) + \theta_3^{(i)} + \theta_4^{(i)}\delta_{ij} \ , \tag{4}$$

with hyper-parameters $\Theta^{(i)} = [\theta_1^{(i)}, \theta_2^{(i)}, \theta_3^{(i)}, \theta_4^{(i)}]$.

Assuming conditional independence between the latent spaces and the data streams, the FOLS-GPLVM can be learned by minimizing Eq. 1 with respect to $\{\mathbf{m}_j^{(i)}\}$ and $\{\theta_1^{(i)}, \theta_3^{(i)}, \theta_4^{(i)}\}$, where the negative log likelihood L is defined up to a constant as

$$L = \sum_{i=1}^{V} \left(\frac{D_i}{2} \ln |\mathbf{K}^{(i)}| + \frac{D_i}{2} tr\left[(\mathbf{K}^{(i)})^{-1}\mathbf{Y}^{(i)}(\mathbf{Y}^{(i)})^T\right]\right) \ . \tag{5}$$

The kernel width $\theta_2^{(i)}$ is determined by cross-validation, and fixed during the optimization.

For inference, the mean prediction of the mapping from a joint shared-private latent space to its corresponding view is given by

$$\bar{\mathbf{y}}_*^{(i)} = (\mathbf{k}_*^{(i)})^T(\mathbf{K}^{(i)})^{-1}\mathbf{Y}^{(i)} \ , \tag{6}$$

where $\mathbf{k}_*^{(i)}$ contains the evaluation of the kernel function between the training and test data for the i-th view.

2.3 FOLS-KIE

The shared kernel information embedding (sKIE) (Sigal et al., 2009) is an extension of the KIE (Memisevic, 2006) model to learn a latent space shared across multiple data streams. As illustrated by Fig. 2 (c), the sKIE is an undirected model. The model is trained by maximizing the mutual information between the shared latent space and the data streams, which is approximated by kernel density estimation.

We create a FOLS version of KIE by introducing private latent spaces into the sKIE model, as shown in Fig. 2 (d). The FOLS-KIE model can be learned by maximizing the mutual information between each joint shared-private space and its corresponding data stream. Assuming independence of the views given the

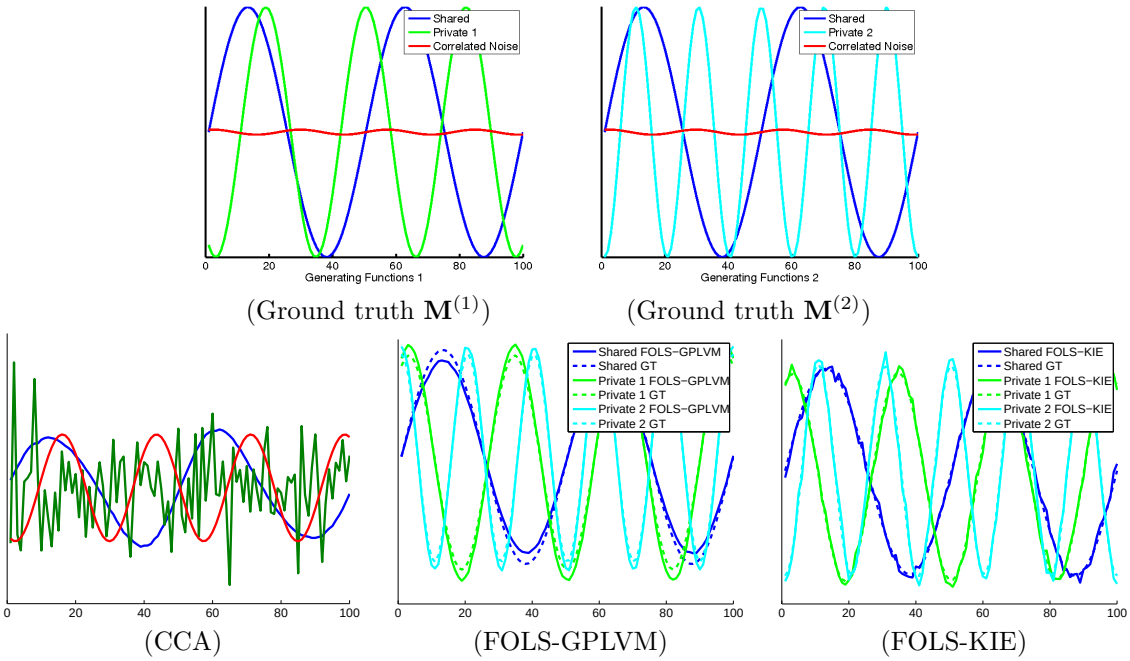

Figure 3: **Recovering shared and private information**: Top: Generative signals for two data streams. The shared information is shown in blue, and correlated noise in red. We projected these signals to a 20D space and added Gaussian noise to them. Bottom left: CCA recovered the true shared, but also the correlated noise, as well as another noise signal. Bottom right: The FOLS-GPLVM and FOLS-KIE models both accurately recovered the generative signals.

joint shared-private spaces, such mutual information can be written as

$$I(\mathbf{y},\mathbf{x},\mathbf{z}) = \sum_{i=1}^{V} I\left(\mathbf{y}^{(i)},(\mathbf{x},\mathbf{z}^{(i)})\right)$$
$$= \sum_{i=1}^{V} H\left(\mathbf{y}^{(i)}\right) + H\left((\mathbf{x},\mathbf{z}^{(i)})\right) - H\left(\mathbf{y}^{(i)},(\mathbf{x},\mathbf{z}^{(i)})\right) ,$$

where $H(\cdot)$ is the Shannon entropy. As in the sKIE model we approximate the mutual information using kernel density estimation, which for each individual view yields

$$\hat{I}\left(\mathbf{y}^{(i)},(\mathbf{x},\mathbf{z}^{(i)})\right) = -\frac{1}{N}\sum_j \log\sum_t k_m(\mathbf{m}_j^{(i)},\mathbf{m}_t^{(i)}) - \frac{1}{N}\sum_j \log\sum_t k_y(\mathbf{y}_j^{(i)},\mathbf{y}_t^{(i)}) + \frac{1}{N}\sum_j \log\sum_t k_m(\mathbf{m}_j^{(i)},\mathbf{m}_t^{(i)})k_y(\mathbf{y}_j^{(i)},\mathbf{y}_t^{(i)}) . \quad (7)$$

Here, we use an RBF kernel to model k_m and k_y. In practice, we determined the kernel widths by cross-validation. The FOLS-KIE model can then be learned by minimizing Eq. 1 with respect to $\{\mathbf{m}_j^{(i)}\}$, with

$$L = -\sum_{i=1}^{V} \hat{I}\left(\mathbf{y}^{(i)},(\mathbf{x},\mathbf{z}^{(i)})\right) . \quad (8)$$

For inference, the FOLS-KIE model provides both a forward and an inverse mappings, whose mean predictions can be expressed as

$$\bar{\mathbf{y}}_*^{(i)} = \sum_{j=1}^{N} \frac{k_m(\mathbf{m}_*{}^{(i)},\mathbf{m}_j^{(i)})}{\sum_{t=1}^{N} k_m(\mathbf{m}_*{}^{(i)},\mathbf{m}_t^{(i)})}\mathbf{y}_j^{(i)} , \quad (9)$$

$$\bar{\mathbf{m}}_*^{(i)} = \sum_{j=1}^{N} \frac{k_y(\mathbf{y}_*{}^{(i)},\mathbf{y}_j^{(i)})}{\sum_{t=1}^{N} k_y(\mathbf{y}_*{}^{(i)},\mathbf{y}_t^{(i)})}\mathbf{m}_j^{(i)} , \quad (10)$$

where $\mathbf{m}_*{}^{(i)}$ and $\mathbf{y}_*{}^{(i)}$ represent the latent coordinates and i-th view of the test example.

2.4 Computational Complexity

The computational overhead of a FOLS model is dominated by the computation of the singular values of $\mathbf{M}^{(i)}$. The complexity of this operation for each view is $\mathcal{O}(\min\{N^3, d_i^3\})$, with N the number of examples, and d_i the initial dimensionality of the shared-private latent space. When both N and d_i are large, estimating the singular values is computationally expensive.

To reduce the complexity, following the thresholding strategy of (Cai et al., 2008), after a small fixed number of iterations we set the smallest singular values to zero. As a consequence, the intrinsic dimensionality of the latent space d_i decreases. In particular, we keep the dimensions accounting for 95% of the vari-

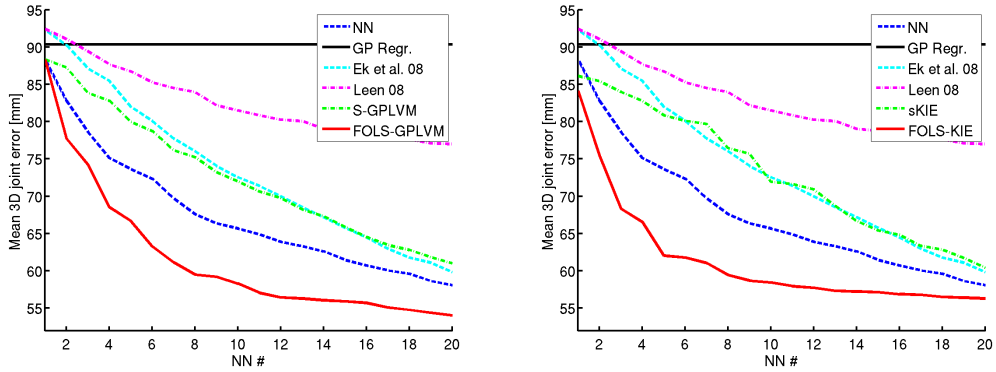

Figure 4: **Humaneva Jog**: For each model, we computed the mean 3D body joint error obtained with the best of k-NN. The NN were computed in the shared latent spaces. We plot this error as a function the number of nearest-neighbors on the left for sGPLVM models, and on the right for sKIE models. We also display the error obtained by computing NN in the feature space, by GP regression from the feature space to the pose space, and by the shared-private factorizations (Ek et al., 2008; Leen, 2008). GP regression does not rely on NN computation. Note that both FOLS models outperform the other techniques.

ance. This makes the computational overhead of our constraints negligible compare to the complexity of the GPLVM which is $\mathcal{O}(N^3)$ and the complexity of the KIE model, which is $\mathcal{O}(N^2)$. Moreover, sparsification techniques can be use to reduce the cost of the original models (Lawrence, 2007). Note that for inference there is an additional benefit with respect to computational complexity since the latent space is factorized and the shared space is lower dimensional than the shared spaces of the sKIE and sGPLVM models.

3 Experimental Evaluation

In this section, we compare our FOLS models to sKIE, sGPLVM, Nearest-Neighbors (NN), GP regression and the factorization techniques of (Ek et al., 2008) and (Leen, 2008) on a synthetic example and on real data for the task of pose estimation from monocular images.

For these experiments, we subtracted the mean of each view, and initialized the shared latent space of the FOLS models with the PCA representation of the concatenated views. Similarly, each private space was initialized with the PCA representation of its corresponding view. For both shared and private spaces we kept 95% of the variance. For the sKIE and the sGPLVM, which assume a known latent dimensionality, we applied PCA to the concatenated views and kept as many components as the global dimensionality found by the FOLS models (i.e., the sum of the dimensionalities of the shared and private spaces).

When learning FOLS models, the relative weights in Eq. 1 were set such that the orthogonality constraints and the energy conservation regularizer initial influences were roughly 10 times that of the other terms. Furthermore, as mentioned in Section 2.4, we sped up the optimization process by removing the latent dimensions accounting for less than 5% of the variance of their corresponding latent space (i.e. shared or private) every 10 iterations of the minimizer. Once a stable dimensionality has been found, we set γ to 0 and optimize until convergence.

3.1 Synthetic Example

To illustrate the weaknesses of CCA, and thus of models initialized from it, we constructed an example where the noise is highly correlated. We generated 100 points of two data streams containing both shared and private information. The ground-truth latent spaces were generated from sinusoidal signals of different frequencies such that

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

where $\mathbf{t}$ is uniformly distributed in the interval $(-1, 1)$. The observations $\mathbf{Y}^{(i)}$ were generated by randomly projecting the joint shared-private spaces $\mathbf{m}^{(1)} = [\mathbf{x}, \mathbf{z}^{(1)}]$ and $\mathbf{m}^{(2)} = [\mathbf{x}, \mathbf{z}^{(2)}]$ into 20D spaces, and adding Gaussian noise with variance 0.01 and correlated noise of the form $\mathbf{x}_{noise} = 0.02\sin(3.6\pi\mathbf{t})$.

The ground-truth latent spaces together with correlated noise are depicted in the top row of Fig. 3. The bottom left plot of Fig. 3 shows the result obtained by CCA when the latent space is set a priori to be 3D. As expected, CCA retrieved the true shared signal, in blue, but failed to remove the highly correlated noise, in red. Additionally, it discovered another highly correlated noise signal, which frequency does not correspond to any of the frequencies of the signals used to generate the data. In contrast, as depicted by the two bottom right plots of Fig. 3, our FOLS-GPLVM and FOLS-KIE models closely recover the generative signals, and find the correct separation between shared

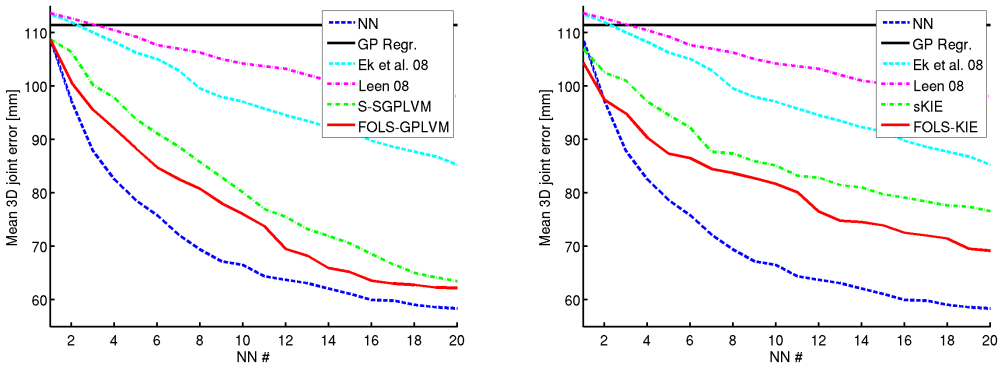

Figure 5: **Humaneva Walk**: As for the jog, we plot the mean 3D joint error as a function of the number of nearest-neighbors. Note in this case that NN in the feature space outperforms other techniques. This suggests that training and test examples are very similar. More importantly, note that the FOLS models outperform the purely shared ones and the shared-private factorizations (Ek et al., 2008; Leen, 2008).

and private spaces, as well as correct latent dimensionalities.

3.2 Pose Estimation

Next, we applied our models to the problem of human pose estimation from monocular images. For this purpose, we used the HumanEva dataset (Sigal and Black, 2006) consisting of pairs of motion captures and corresponding images. We computed hierarchical features (Kanaujia et al., 2007) for the walking and jogging video sequences of the first subject seen from a single camera. The image stream consists of 19D features, and the pose stream consists of 57D observations. Note that estimating the pose directly from the features is known to be multi-modal (Ek et al., 2008) and cannot be solved as a regression task, and therefore, poses contain private information. As the subject moves in circles, we used one loop for training, and tested our models on the remaining one. We compare our FOLS models to GP regression, Nearest-Neighbor in the feature space, the shared-private factorizations (Ek et al., 2008; Leen, 2008), and the sKIE and sGPLVM models. In the latter cases, the dimensionality of the shared space was taken as the global one found by the FOLS models, which in all cases was 5.

For inference, similar to (Sigal and Black, 2006), we relied on the following strategy: We took the latent representation of the first nearest-neighbor (NN) in feature space, computed its k-NN in latent space, and mapped them to the pose space using the forward mapping provided by the different models. In the FOLS case and for (Ek et al., 2008; Leen, 2008), the k-NN were computed in the shared space only. The joint shared-private latent representation was formed by keeping the shared latent variables constant while taking the private ones from the corresponding NNs. Finally the mappings from joint shared-private spaces to the pose were computed using Eq. 6 for the FOLS-GPLVM and Eq. 9 for the FOLS-KIE.

Fig. 4 depicts the error for the jogging dataset as a function of the number of NNs used. The error is computed as the mean squared distance between the model's reconstruction of the 3D joint locations and ground-truth data. Note that GP regression does not rely on NN, and therefore remains constant. The FOLS models significantly outperformed the purely shared models (i.e., sKIE and sGPLVM), NN, GP regression, and the shared-private models (Ek et al., 2008; Leen, 2008). Moreover, note that all models outperform GP regression, which confirms the multimodality of the data (i.e., an image observation can be generated from more than one pose).

Fig. 5 depicts similar plots for the walking dataset. In this case, when using multiple neighbors, NN outperforms our approach. This is due to the fact that there is very little variation between the training and test examples. More importantly, we can again observe that the two FOLS models outperform the purely shared models. This confirms the benefits of factorizing the latent spaces into shared and private parts. Note that, when considering the first nearest-neighbor only, the FOLS-KIE model outperforms the other methods for both jogging and walking.

Another important observation is that the results of (Leen, 2008), which optimizes the solution of (Ek et al., 2008), are much worse than the initialization itself. This shows that only focusing on data reconstruction does not yield a meaningful latent representation. In contrast, the FOLS objective encourages the factorization into shared and private spaces to be non-redundant. As a result, our FOLS models outperform these methods.

Finally, Fig. 6 depicts the sensitivity of our approach to the image and pose kernel widths for both the FOLS-KIE and FOLS-GPLVM models. For each

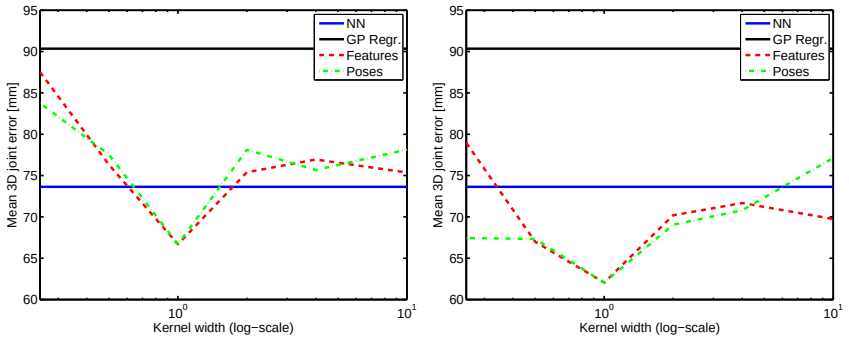

Figure 6: **Kernel width cross-validation**: For both FOLS models, we found the kernel widths for the feature and pose spaces by fixing one and varying the other between 0.25 and 10.0. On the left, we show the errors for 5 NN for the FOLS-GPLVM model, and on the right, for the FOLS-KIE model. Note that the same widths were found for all models and spaces.

model, we fixed one width to 1 (i.e. the value used in Fig. 4 and Fig. 5) and varied the other between 0.25 and 10. Fig. 6 displays the mean errors obtained by using 5 NN for each kernel width. Note that the two models find the same optimal widths for both spaces.

4 Conclusions

In this paper, we have proposed the use of orthogonality and rank constraints to learn the structure and dimensionality of factorized latent spaces that are non-redundant, and can capture the shared-private separation of the data. We have demonstrated the effectiveness of our approach by applying our constraints to two shared models, the sGPLVM and the sKIE, and show clear improvement over the original models in the context of pose estimation from monocular images. Moreover, our approach has shown beneficial over the models of (Ek et al., 2008) and (Leen, 2008), which, to our knowledge, are the only attempts at factorizing latent spaces into private and shared components.

In the future, we plan to investigate the application of the FOLS model to other domains such as multi-agent modeling. We also intend to study the influence of additional constraints such as smoothness or prior knowledge about the task at hand (Urtasun et al., 2008), as well as discriminative constraints for classification tasks (Urtasun and Darrell, 2007). An alternative topic for future research is the extension of our model to semi-supervised learning. Provided that the particular loss function of interest allows for missing data, this could be done by computing the orthogonality constraints between the private spaces only on the samples observed in both views.

References

C. Archambeau, and F. Bach. Sparse Probabilistic Projections. *NIPS*, 2008.

F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*, 2004.

J. F. Cai, E. Candes, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. *Arxiv preprint arXiv:0810.3286*, 2008.

C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In *MLMI*, 2007.

C. H. Ek, P. H. Torr, and N. D. Lawrence. Ambiguity modeling in latent spaces. In *MLMI*, 2008.

A. Geiger, R. Urtasun, and T. Darrell. Rank priors for continuous non-linear dimensionality reduction. In *CVPR*, 2009.

A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised Hierarchical Models for 3D Human Pose Reconstruction. In *CVPR*, 2007.

A. Klami, and S. Kaski. Probabilistic approach to detecting dependencies between data sets. In *Neurocomputing*, 72:39–46, 2008.

M. Kuss and T. Graepel. The geometry of kernel canonical correlation analysis. Technical Report TR-108, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2003.

N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 11 2005.

N. D. Lawrence. Learning for larger datasets with the Gaussian process latent variable model. *AISTATS*, 2007.

G. Leen. *Context assisted information extraction*. PhD thesis, University of the West of Scotland, High Street, Paisley PA1 2BE, Scotland, 2008.

R. Memisevic. Kernel information embeddings. In *ICML*, 2006.

R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The joint manifold model. In *ICCV*, 2007.

A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning shared latent structure for image synthesis and robotic imitation. *NIPS*, 2006.

L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 2006.

L. Sigal, R. Memisevic, and D. J. Fleet. Shared kernel information embedding for discriminative inference. In *CVPR*, 2009.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

R. Urtasun and T. Darrell. Discriminative gaussian process latent variable model for classification. *ICML*, 2007.

R. Urtasun, D. Fleet, A. Geiger, J. Popović, T. Darrell, and N. Lawrence. Topologically-constrained latent variable models. In *ICML*, 2008.