FOLS: Factorized Orthogonal Latent Spaces

Mathieu Salzmann[1], Carl Henrik Ek[1], Raquel Urtasun[2], and Trevor Darrell [1]
[1]EECS & ICSI, UC Berkeley , {salzmann,ek,trevor}@eecs.berkeley.edu
[2]TTI - Chicago, rurtasun@ttic.edu

1. Introduction

Many machine learning problems inherently involve multiple views. Kernel combination approaches to multi-view learning [1] are particularly effective when the views are independent. In contrast, other methods take advantage of the dependencies in the data. The best-known example is Canonical Correlation Analysis (CCA), which learns latent representations of the views whose correlation is maximal. Unfortunately, this can result in trivial solutions in the presence of highly correlated noise. 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) [4], and the shared kernel information embedding (sKIE) [5]. However, in real scenarios, information in the views is typically neither fully independent nor fully correlated. The few approaches that have tried to factorize the information into shared and private components [2, 3] are typically initialized with CCA, and thus suffer from its inherent weaknesses.

In this paper, we propose a method to learn shared and private latent spaces that are inherently disjoint by introducing orthogonality constraints. Furthermore, 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 [4] and the sKIE [5], and show significant performance improvement over the original models, as well as over the existing shared-private factorizations [2, 3] in the context of pose estimation.

2. Factorized Orthogonal Latent Spaces

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. Here, we propose to enforce this by using orthogonality constraints. In addition, we would like to estimate the latent spaces dimensionalities at the same time as we learn their structure. To this end, we introduce a regularizer that encourages each joint shared-private latent space to be low dimensional. Finally, 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

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

where s_i are the singular values of $\mathbf{M}^{(i)}$, and $E_0^{(i)} = \sum_j t_{i,j}^2$ is the energy of stream i, with $t_{i,j}$ the singular values of $\mathbf{Y}^{(i)}$. α, β, 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.

In practice, we applied our constraints to two recently developed models, the sGPLVM [4] and the sKIE [5]. The resulting graphical models are depicted in Fig. 1. For the FOLS-GPLVM, the loss function of Eq. 1 can be written as

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

where $\mathbf{K}$ is a covariance matrix defined by a kernel function that depends on hyper-parameters. For the FOLS-KIE model, the loss function becomes

$$L = -\sum_{i=1}^{V} \hat{I} \left(\mathbf{y}^{(i)}, (\mathbf{x}, \mathbf{z}^{(i)}) \right) , \quad (3)$$

where $\hat{I}$ is an approximation of the mutual information based on kernel density estimation.

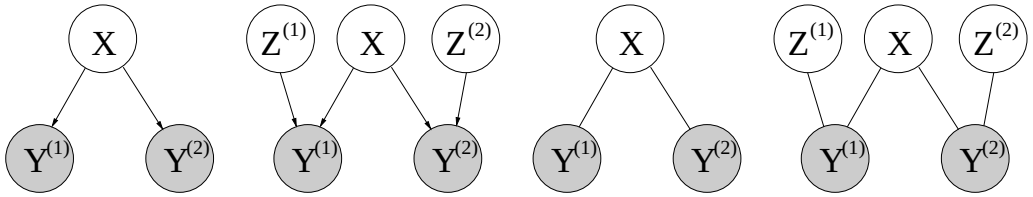
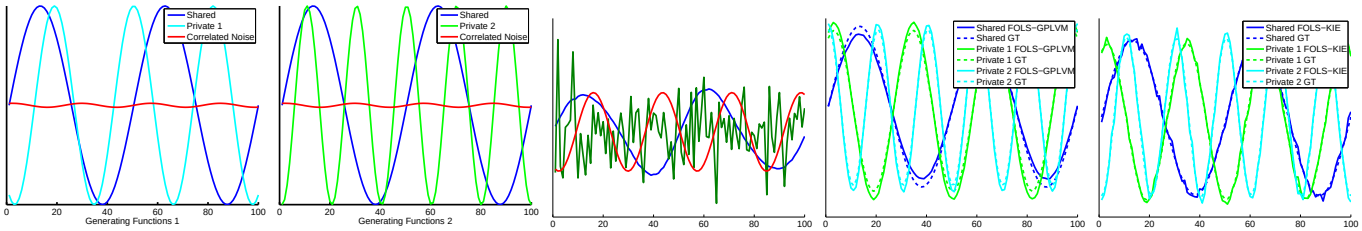


Figure 1. **Graphical models** (a) sGPLVM. (b) FOLS-GPLVM. (c) sKIE. (d) FOLS-KIE



(Ground truth $\mathbf{M}^{(1)}$) (Ground truth $\mathbf{M}^{(2)}$) (CCA) (FOLS-GPLVM) (FOLS-KIE)

Figure 2. **Recovering shared and private information**: From left to right: 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. CCA recovered the true shared, but also the correlated noise, as well as another noise signal. The FOLS-GPLVM and FOLS-KIE models both accurately recovered the generative signals.

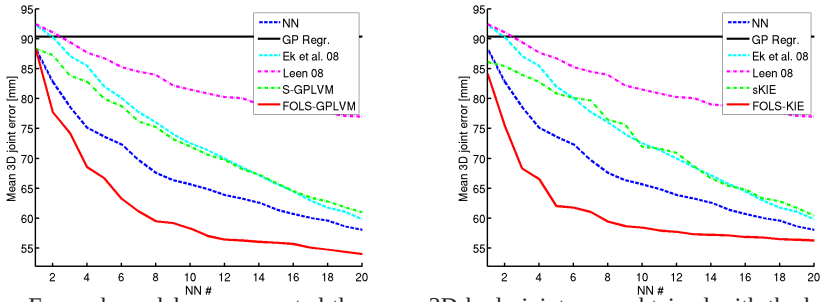


Figure 3. **Humaneva Jog**: For each model, we computed the mean 3D body joint error obtained with the best of k-NN. The NN were computed in the shared latent spaces. We plot this error as a function the number of nearest-neighbors on the left for sGPLVM models, and on the right for sKIE models. We also 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 [2, 3]. GP regression does not rely on NN computation. Note that both FOLS models outperform the other techniques.

3. Experimental Evaluation

Fig. 2 depicts a toy example that illustrates the weaknesses of CCA. The observations were generated by randomly projecting the joint shared-private spaces of the two leftmost plots into two 20D spaces, and adding Gaussian noise with variance 0.01. As expected, CCA retrieved the shared signal, but failed to remove the highly correlated noise and discovered another highly correlated noise signal. In contrast, our FOLS models correctly recover the generative signals.

In Fig. 3, we compare several methods on human pose estimation from hierarchical features using the HumanEva jogging motion. For inference, 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 [2, 3], 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. Fig. 3 depicts the mean squared point-to-point distance as a function of the number of NNs used. Note that the FOLS models outperform the purely shared models, NN, GP regression, and [2, 3].

References

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

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

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

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

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