

---

# Learning Features with Structure-Adapting Multi-view Exponential Family Harmoniums

---

Yoonseop Kang<sup>1</sup>    Seungjin Choi<sup>1,2,3</sup>

Department of Computer Science and Engineering<sup>1</sup>,

Division of IT Convergence Engineering<sup>2</sup>,

Department of Creative Excellence Engineering<sup>3</sup>,

Pohang University of Science and Technology (POSTECH)

Pohang, South Korea, 790-784.

{e0en, seungjin}@postech.ac.kr

## Abstract

We propose a graphical model for multi-view feature extraction that automatically adapts its structure to achieve better representation of data distribution. The proposed model, *structure-adapting multi-view harmonium* (SA-MVH) has *switch parameters* that control the connection between hidden nodes and input views, and learn the switch parameter while training. Numerical experiments on synthetic and a real-world dataset demonstrate the useful behavior of the SA-MVH, compared to existing multi-view feature extraction methods.

## 1 Introduction

Earlier multi-view feature extraction methods including canonical correlation analysis [1] and dual-wing harmonium (DWH) [2] assume that all views can be described using a single set of shared hidden nodes. However, these methods fail when real-world data with partially correlated views are given. More recent methods like factorized orthogonal latent space [3] or multi-view harmonium (MVH) [4] assume that views are generated from two sets of hidden nodes: view-specific hidden nodes and shared ones. Still, these models rely on the pre-defined connection structure, and deciding the number of shared and view-specific hidden nodes requires a great human effort.

In this paper, we propose structure-adapting multi-view harmonium (SA-MVH) which avoids all of the problems mentioned above. Instead of explicitly defining view-specific and hidden nodes in prior to the training, we only use one set of hidden nodes and let each one of them to decide the existence of connection to views using *switch parameters* during the training. In this manner, SA-MVH automatically decides the number of view-specific latent variables and also captures partial correlation among views.

## 2 The Proposed Model

The definition of SA-MVH begins with choosing marginal distributions of visible node sets  $\mathbf{v}^{(k)}$  and a set of hidden nodes  $\mathbf{h}$  from exponential family distributions:

$$\begin{aligned} p(v_i^{(k)}) &\propto \exp\left(\sum_a \xi_{ia}^{(k)} f_{ia}^{(k)}(v_i^{(k)}) - A_i^{(k)}(\{\xi_{ia}^{(k)}\})\right), \\ p(h_j) &\propto \exp\left(\sum_b \lambda_{jb} g_{jb}(h_j) - B_j(\{\lambda_{jb}\})\right), \end{aligned} \quad (1)$$

$f(\cdot), g(\cdot)$  are sufficient statistics,  $\xi, \lambda$  are natural parameters, and  $A, B$  are log-partition functions.

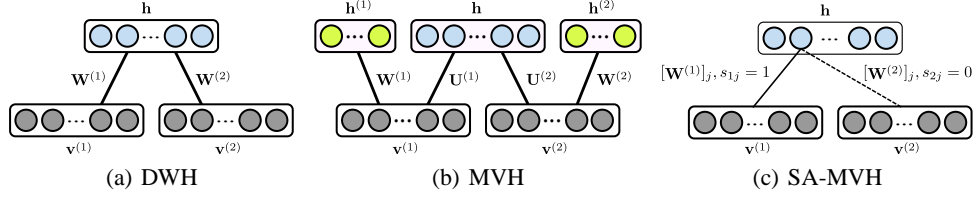


Figure 1: Graphical models of (a) dual-wing harmonium, (b) multi-view harmonium, and (c) structure-adapting multi-view harmonium.

Connections between visible nodes and hidden nodes of SA-MVH are defined by weight matrices  $\{\mathbf{W}^{(k)}\}$  and *switch parameters*  $\sigma(s_{kj}) \in [0, 1]$ , where  $\sigma(\cdot)$  is a sigmoid function. A switch  $s_{kj}$  controls the connection between  $k$ -th view and  $j$ -th hidden node by being multiplied to the  $j$ -th column of weight matrix  $\mathbf{W}^{(k)}$  (Figure 1). When  $\sigma(s_{kj})$  is large ( $> 0.5$ ), we consider the view and the hidden node to be connected. With the quadratic term including weights and switch parameters, the joint distribution of SA-MVH is defined as below:

$$p(\{\mathbf{v}^{(k)}\}, \mathbf{h}) \propto \exp\left(\sum_{k,i,j} \sigma(s_{kj}) \mathbf{W}_{ij}^{(k)} f_i^{(k)}(\mathbf{v}_i^{(k)}) g_j(h_j) - \sum_{k,i} \xi_i^{(k)} f_i^{(k)}(\mathbf{v}_i^{(k)}) - \sum_j \lambda_j g_j(h_j)\right). \quad (2)$$

note that indices  $a$  and  $b$  are omitted to keep the notations uncluttered.

We learn the parameters  $\mathbf{W}^{(k)}$ ,  $\xi^{(k)}$ ,  $\lambda$ , and switch parameters  $s_{kj}$  by maximizing the likelihood of model via gradient ascent. The likelihood of SA-MVH is defined as the joint distribution of nodes summed over hidden nodes  $\mathbf{h}$ :

$$\mathcal{L} = \langle \log p(\{\mathbf{v}^{(k)}\}) \rangle_{data} = \langle \log \sum_{\mathbf{h}} p(\{\mathbf{v}^{(k)}\}, \mathbf{h}) \rangle_{data}, \quad (3)$$

where  $\langle \cdot \rangle_{data}$  represents expectation over data distribution. Then the gradient of log-likelihood with respect to the parameters  $\mathbf{W}^{(k)}$ ,  $\xi^{(k)}$ ,  $\lambda$ , and  $s_{kj}$  are derived as follows:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{ij}^{(k)}} \propto \langle \sigma(s_{kj}) f_i(\mathbf{v}_i^{(k)}) B'_j(\hat{\lambda}_j) \rangle_{data} - \langle \sigma(s_{kj}) f_i(\mathbf{v}_i^{(k)}) B'_j(\hat{\lambda}_j) \rangle_{model} \quad (4)$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i^{(k)}} \propto \langle f_i^{(k)}(\mathbf{v}_i^{(k)}) \rangle_{data} - \langle f_i^{(k)}(\mathbf{v}_i^{(k)}) \rangle_{model} \quad (5)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_j} \propto \langle B'_j(\hat{\lambda}_j) \rangle_{data} - \langle B'_j(\hat{\lambda}_j) \rangle_{model}, \quad (6)$$

$$\frac{\partial \mathcal{L}}{\partial s_{kj}} \propto \langle \sigma'(s_{kj}) \mathbf{W}_{ij}^{(k)} f_i(\mathbf{v}_i^{(k)}) B'_j(\hat{\lambda}_j) \rangle_{data} - \langle \sigma'(s_{kj}) \mathbf{W}_{ij}^{(k)} f_i(\mathbf{v}_i^{(k)}) B'_j(\hat{\lambda}_j) \rangle_{model} \quad (7)$$

where  $\langle \cdot \rangle_{model}$  represents expectation over model distribution  $p(\{\mathbf{v}^{(k)}\}, \mathbf{h})$  and  $\hat{\xi}_i^{(k)} = \xi_i^{(k)} + \sum_j \sigma(s_{kj}) \mathbf{W}_{ij}^{(k)} g_j(h_j)$ ,  $\hat{\lambda}_j = \lambda_j + \sum_{k,i} \sigma(s_{kj}) \mathbf{W}_{ij}^{(k)} f_i(\mathbf{v}_i^{(k)})$  are shifted parameters.

### 3 Numerical Experiments

#### 3.1 Feature Extraction on Noisy Arabic-Roman Digit Dataset

To simulate the view-specific and shared properties of multi-view data, we designed a synthetic dataset which contains 11,800 pairs of Arabic digits and the corresponding Roman digits written in various fonts. For each pair, we added random vertical line noises to Arabic digits, and horizontal line noises to Roman digits (Figure 2-(a)). SA-MVH trained with 200 hidden nodes found 95 shared features (with connection to both views), and 47 view-specific features for Roman digits, and 32 for Arabic digits. Remaining 26 were not connected to any views and ignored. Most of the shared features were noise-free and encoded parts of Roman and Arabic numbers (Figure 2-(b)). On the other hand, the view-specific features had components with horizontal or vertical noises, as well as the parts of the numbers (Figure 2-(c)). In this example, SA-MVH automatically separated view-specific and shared information without any prior specification of the graph structure.

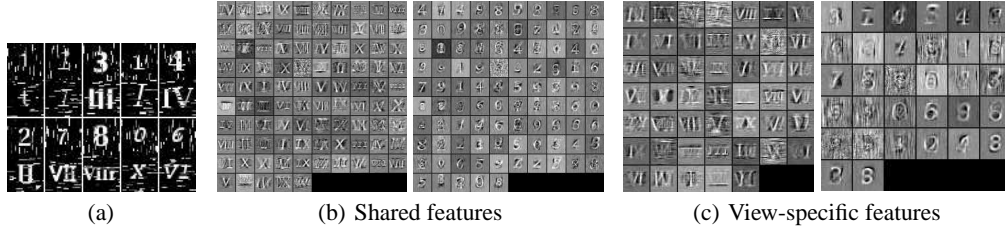


Figure 2: (a) 10 samples from Noisy Arabic-Roman digit dataset, (b) shared features, and (c) view-specific features learned by SA-MVH.

Table 1: Image classification accuracy of k-nn classifier using feature extraction methods trained on Caltech-256 dataset. For each value of  $k$ , the best result is marked as bold text.

Method	# 10-NN	30-NN	50-NN	70-NN	100-NN
Sparse Filtering	0.161	0.165	0.163	0.16	0.155
DWH	0.237	0.231	0.217	0.207	0.194
MVH	0.239	0.225	0.216	0.203	0.191
SA-MVH	<b>0.246</b>	<b>0.232</b>	<b>0.223</b>	<b>0.212</b>	<b>0.198</b>

### 3.2 Image Classification on Caltech-256 Dataset

We extracted 512 dimensions of GIST features and 1,536 dimensions of histogram of gradients (HoG) features from Caltech-256 dataset to simulate multi-view settings. SA-MVH and other multi-view feature extraction methods based on harmonium – DWH and MVH were trained on the dataset for comparison. We also compared our method to Sparse Filtering [5], which is not a harmonium-based method. We trained the feature extraction methods and tested the methods with k-nearest neighbor classifiers (Table 1). SA-MVH resulted better than other feature extraction models in this experiment, regardless of the value of  $k$  for nearest neighbor classifier.

## 4 Conclusion

In this paper, we have proposed the multi-view feature extraction model that automatically decides relations between latent variables and input views. The proposed method, SA-MVH models multi-view data distribution with less restrictive assumption and also reduces the number of parameters to tune by human hand. SA-MVH introduces *switch parameters* that control the connections between hidden nodes and input views, and find the desirable configuration while training. We have demonstrated the effectiveness of our approach by comparing our model to existing models in experiments on synthetic dataset, and image classification with simulated multi-view setting.

## References

- [1] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with applications to learning methods,” *Neural Computation*, vol. 16, pp. 2639–2664, 2004.
- [2] E. P. Xing, R. Yan, and A. G. Hauptmann, “Mining associated text and images with dual-wing harmonium,” in *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Edinburgh, UK, 2005.
- [3] M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell, “Factorized orthogonal latent spaces,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Sardinia, Italy, 2010.
- [4] Y. Kang and S. Choi, “Restricted deep belief networks for multi-view learning,” in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, Athens, Greece, 2011.
- [5] J. Ngiam, P. W. Koh, Z. Chen, S. A. Bhaskar, and A. Y. Ng, “Sparse filtering,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 23. MIT Press, 2011.