



Kernelized Bayesian Matrix Factorization

30th International Conference on Machine Learning (ICML 2013)
Atlanta, Georgia, USA

Mehmet Gönen^{1,2} Suleiman A. Khan^{1,2} Samuel Kaski^{1,2,3}

¹ *Helsinki Institute for Information Technology HIIT*

² *Department of Information and Computer Science, Aalto University*

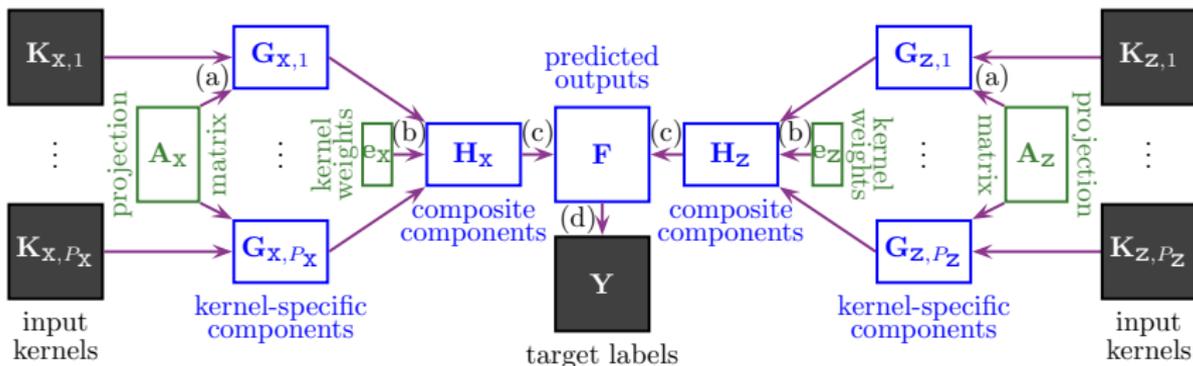
³ *Department of Computer Science, University of Helsinki*

June 18, 2013

Summary

- We extend kernelized matrix factorization
 - with a fully Bayesian treatment,
 - with an ability to work with multiple side information sources.
- Side information is necessary for making out-of-matrix predictions.
- We mainly discuss bipartite graph inference, where the output matrix is binary.
- We show the performance of our method
 - by predicting drug–protein interactions on two data sets,
 - by performing multilabel classification on 14 benchmark data sets.

Proposed Method



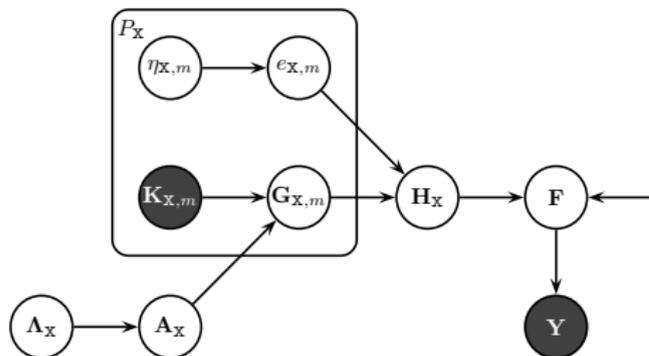
(a) *kernel-based nonlinear dimensionality reduction*

(b) *multiple kernel learning*

(c) *matrix factorization*

(d) *binary classification*

Probabilistic Model



dimensionality reduction

$$\lambda_{x,s}^i \sim \mathcal{G}(\lambda_{x,s}^i; \alpha_\lambda, \beta_\lambda)$$

$$a_{x,s}^i | \lambda_{x,s}^i \sim \mathcal{N}(a_{x,s}^i; 0, (\lambda_{x,s}^i)^{-1})$$

$$g_{x,m,i}^s | \mathbf{a}_{x,s}, \mathbf{k}_{x,m,i} \sim \mathcal{N}(g_{x,m,i}^s; \mathbf{a}_{x,s}^\top \mathbf{k}_{x,m,i}, \sigma_g^2)$$

multiple kernel learning

$$\eta_{x,m} \sim \mathcal{G}(\eta_{x,m}; \alpha_\eta, \beta_\eta)$$

$$e_{x,m} | \eta_{x,m} \sim \mathcal{N}(e_{x,m}; 0, \eta_{x,m}^{-1})$$

$$h_{x,i}^s | \{e_{x,m}, g_{x,m,i}^s\}_{m=1}^{P_x} \sim \mathcal{N}\left(h_{x,i}^s; \sum_{m=1}^{P_x} e_{x,m} g_{x,m,i}^s, \sigma_h^2\right)$$

matrix factorization

$$f_j^i | \mathbf{h}_{x,i}, \mathbf{h}_{z,j} \sim \mathcal{N}(f_j^i; \mathbf{h}_{x,i}^\top \mathbf{h}_{z,j}, 1)$$

binary classification

$$y_j^i | f_j^i \sim \delta(f_j^i > \nu)$$

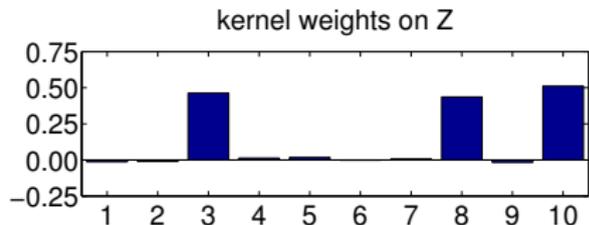
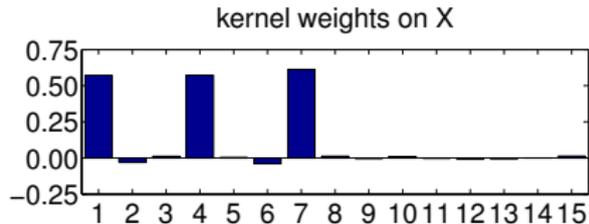
Toy Data Set

$$x_i^m \sim \mathcal{N}(x_i^m; 0, 1)$$

$$z_j^n \sim \mathcal{N}(z_j^n; 0, 1)$$

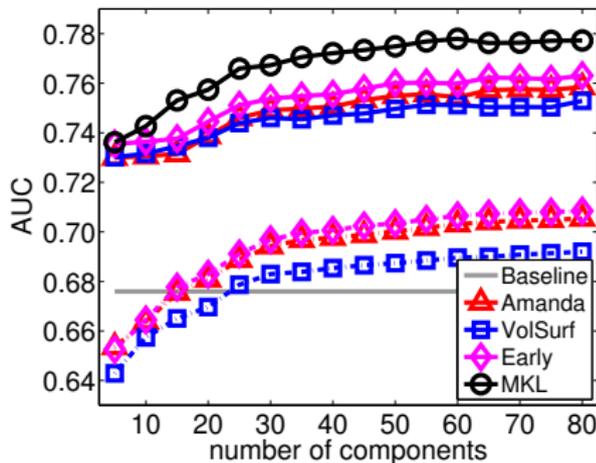
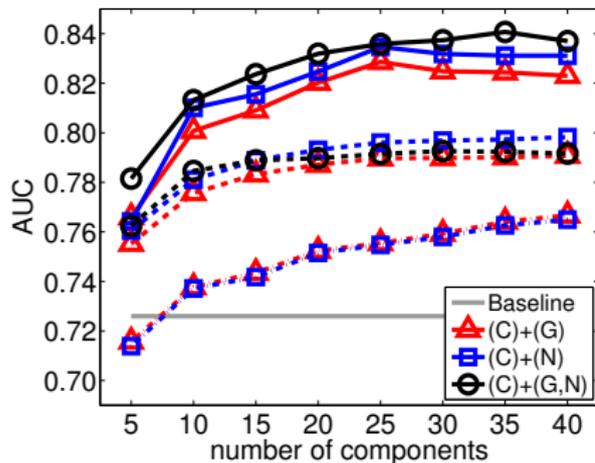
$$y_j^i | \mathbf{x}_i, \mathbf{z}_j \sim \mathcal{N}(y_j^i; x_i^1 z_j^3 + x_i^4 z_j^8 + x_i^7 z_j^{10}, 1)$$

- $(N_x, N_z) = (40, 60)$
- $(P_x, P_z) = (15, 10)$
- A separate linear kernel for each feature



Drug-Protein Interaction Data Sets

- Experiments on two drug-protein interaction data sets
- KBMF is statistically significantly better than KPMF of Zhou et al. (2012) according to paired t -test ($p < 0.01$) on both data sets.



Multilabel Classification Data Sets

- We report classification performances (i.e., Hamming loss values) on 14 multilabel classification benchmark data sets.

Data Set	N_{train}	N_{test}	D	L	KBMF	Zhang's	ML-KNN	RML	Tang's	RankSVM
Emotions	391	202	72	6	0.176	0.195	0.202	0.241	0.240	0.234
Scene	1211	1196	294	6	0.086	0.089	0.099	0.109	0.130	0.127
Yeast	1500	917	103	14	0.189	0.196	0.195	0.204	0.190	0.201
Arts	2000	3000	462	26	0.057	0.057	0.061	0.058	0.094	0.063
Business	2000	3000	681	33	0.025	0.026	0.027	0.032	0.092	0.027
Computers	2000	3000	640	21	0.036	0.036	0.041	0.037	0.097	0.042
Education	2000	3000	606	22	0.039	0.038	0.039	0.050	0.038	0.048
Entertainment	2000	3000	743	40	0.046	0.055	0.063	0.059	0.053	0.062
Health	2000	3000	636	27	0.036	0.037	0.047	0.041	0.222	0.042
Recreation	2000	3000	438	30	0.044	0.057	0.062	0.057	0.057	0.064
Reference	2000	3000	550	33	0.027	0.025	0.032	0.027	0.087	0.034
Science	2000	3000	612	32	0.032	0.031	0.033	0.051	0.057	0.038
Social	2000	3000	793	33	0.022	0.021	0.022	0.101	0.072	0.027
Society	2000	3000	1047	39	0.038	0.052	0.054	0.096	0.056	0.060
Average Rank					1.536	1.964	3.750	4.464	4.607	4.679

Conclusion

- Our Matlab implementation is at <http://research.ics.aalto.fi/mi/software/kbmf>.
- More details in the poster session

Kernelized Bayesian Matrix Factorization

Mehmet Gönen^{1,2} Sulaiman A. Khan^{1,2} Samuel Kaski^{1,2,3}

¹Helsinki Institute for Information Technology HIIT
²Department of Information and Computer Science, Aalto University
³Department of Computer Science, University of Helsinki
<http://research.ics.aalto.fi/mi/software/kbmf>

Summary

- We review **kernelized matrix factorization**
 - with a fully Bayesian treatment
 - with an ability to work with multiple side information sources
- Side information is necessary for making **well-posed matrix predictions**
- We modify **kernelized graph inference** where the output matrix is binary
- We show the performance of our method
 - by predicting drug-protein interaction on two data sets
 - by performing multilabel classification on 14 benchmark data sets

Proposed Method

(i) kernel-based nonlinear dimensionality reduction
 (ii) multiple kernel learning
 (iii) matrix factorization
 (iv) binary classification

Probabilistic Model

(i) **Kernel-based learning**

$$K_{ij} = \sum_{k=1}^K \alpha_k \kappa(x_i, x_j; \lambda_k)$$

(ii) **Matrix factorization**

$$Y_{ij} = \sum_{k=1}^K \beta_k \phi_k(x_i) \psi_k(x_j)$$

(iii) **Binary classification**

$$p(y_{ij} = 1) = \frac{1}{1 + \exp(-\theta_{ij})}$$

Toy Data Set

Drug-Protein Interaction Data Sets

A drug-protein network by Varamini et al. (2008)
 485 drugs, 668 proteins, and 3200 validated interactions
 C: chemical similarity for drugs
 P: genetic similarity for proteins
 R: network similarity for proteins

$\kappa(x_i, x_j) = \exp(-\frac{d(x_i, x_j)}{\lambda})$ - 5-fold CV over drugs
 RMR is statistically significantly better than RRF of Zhou et al. (2012) according to paired t-test ($p < 0.05$) on both data sets.

Another drug-protein interaction network by Khan et al. (2012)
 650 drugs, 900 proteins, and 4800 validated interactions
 T: Tm standard ID chemical structure descriptors for drugs
 Amindist (Dixon et al. 2005) and MolSurf (Crippen et al. 2000)
 C: Gaussian kernel robust with $\lambda = 10^7$
 R: 5 replications of 5-fold cross validation over drugs
 A: Active task of finding or recovering drugs with similar functions

Multilabel Classification Data Sets

- Complexes and labels are returned to be four domains: X and Z , respectively
- Class membership matrix corresponds to target label matrix Y in our model
- The similarity between samples are measured with the different Gaussian kernels whose widths are selected as $\sqrt{0.01}$, $\sqrt{0.1}$, $\sqrt{1}$, $\sqrt{10}$ and $\sqrt{100}$
- The similarity between labels is measured with the learned rules over the labels of training samples

We compare with two algorithms: (i) Katushi (Snelard & Houton, 2012), (ii) M-008 (Zhou & Zhou, 2007), (iii) Tang et al. (Zhang et al. 2010), (iv) ML (Hettner & Cramer, 2010), and (v) Zhang et al. (Zhang et al. 2010)

We report classification performance F_1 (Harmonic loss) on the multilabel classification data sets.

Dataset	Kernel	Complexes	Labels	F_1	Std
Complexes	0.01	100	100	0.85	0.02
	0.1	100	100	0.85	0.02
	1	100	100	0.85	0.02
	10	100	100	0.85	0.02
	100	100	100	0.85	0.02
Labels	0.01	100	100	0.85	0.02
	0.1	100	100	0.85	0.02
	1	100	100	0.85	0.02
	10	100	100	0.85	0.02
	100	100	100	0.85	0.02

A! Aalto University
School of Science

Kernelized Bayesian Matrix Factorization
Mehmet Gönen
HIIT & Aalto ICS

7/7
June 18, 2013
ICML 2013