Kernelized Bayesian Matrix Factorization

Mehmet Gönen[1,2] **Suleiman A. Khan**[1,2] **Samuel Kaski**[1,2,3]

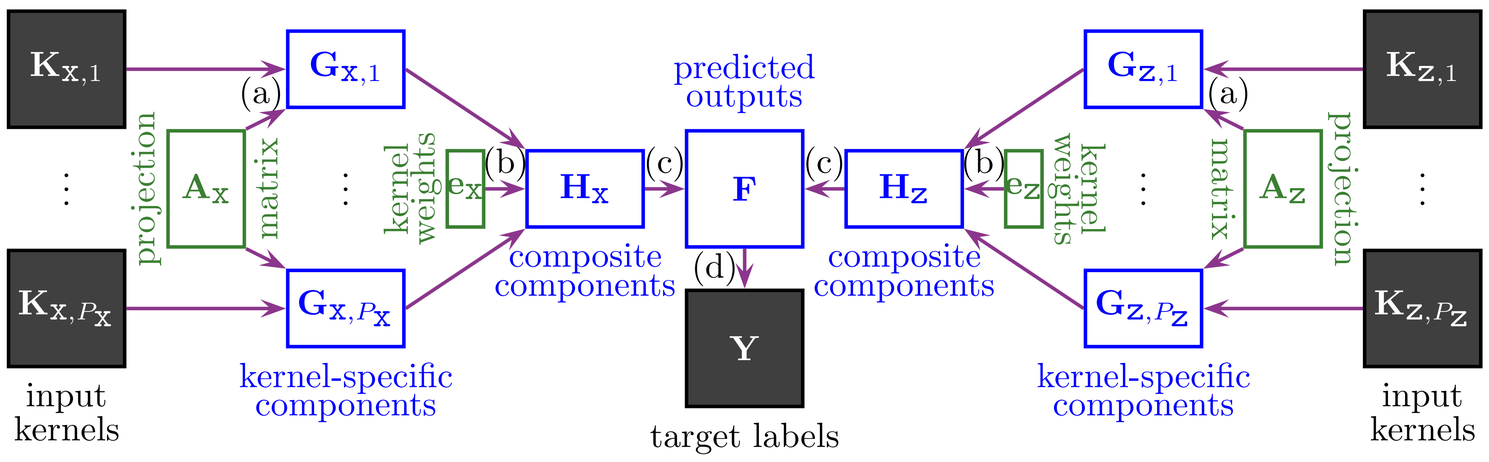
[1]Helsinki Institute for Information Technology HIIT
[2]Department of Information and Computer Science, Aalto University
[3]Department of Computer Science, University of Helsinki

http://research.ics.aalto.fi/mi/software/kbmf

Summary

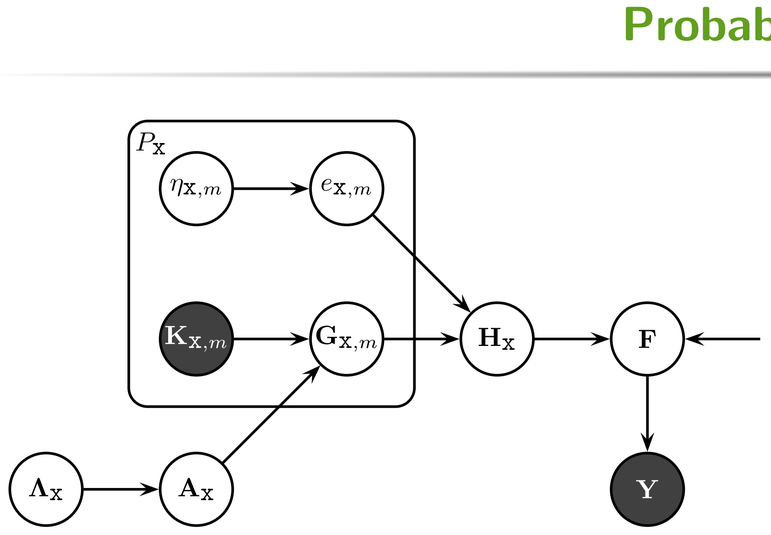
- We extend *kernelized matrix factorization*
  - with a fully Bayesian treatment,
  - with an ability to work with multiple side information sources.
- Side information is necessary for making out-of-matrix predictions.
- We mainly discuss *bipartite graph inference*, where the output matrix is binary.
- 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



(a) kernel-based nonlinear dimensionality reduction

$$\lambda^i_{\mathbf{x},s} \sim \mathcal{G}(\lambda^i_{\mathbf{x},s}; \alpha_\lambda, \beta_\lambda) \qquad \forall(i,s)$$
$$a^i_{\mathbf{x},s}|\lambda^i_{\mathbf{x},s} \sim \mathcal{N}(a^i_{\mathbf{x},s}; 0, (\lambda^i_{\mathbf{x},s})^{-1}) \qquad \forall(i,s)$$
$$g^s_{\mathbf{x},m,i}|\boldsymbol{a}_{\mathbf{x},s}, \boldsymbol{k}_{\mathbf{x},m,i} \sim \mathcal{N}(g^s_{\mathbf{x},m,i}; \boldsymbol{a}^\top_{\mathbf{x},s}\boldsymbol{k}_{\mathbf{x},m,i}, \sigma^2_g) \qquad \forall(m,s,i)$$

(b) multiple kernel learning

$$\eta_{\mathbf{x},m} \sim \mathcal{G}(\eta_{\mathbf{x},m}; \alpha_\eta, \beta_\eta) \qquad \forall m$$
$$e_{\mathbf{x},m}|\eta_{\mathbf{x},m} \sim \mathcal{N}(e_{\mathbf{x},m}; 0, \eta^{-1}_{\mathbf{x},m}) \qquad \forall m$$
$$h^s_{\mathbf{x},i}|\{e_{\mathbf{x},m}, g^s_{\mathbf{x},m,i}\}_{m=1}^{P_\mathbf{x}} \sim \mathcal{N}\left(h^s_{\mathbf{x},i}; \sum_{m=1}^{P_\mathbf{x}} e_{\mathbf{x},m} g^s_{\mathbf{x},m,i}, \sigma^2_h\right) \qquad \forall(s,i)$$

(c) matrix factorization

$$f^i_j|\boldsymbol{h}_{\mathbf{x},i}, \boldsymbol{h}_{\mathbf{z},j} \sim \mathcal{N}(f^i_j; \boldsymbol{h}^\top_{\mathbf{x},i}\boldsymbol{h}_{\mathbf{z},j}, 1) \qquad \forall(i,j)$$

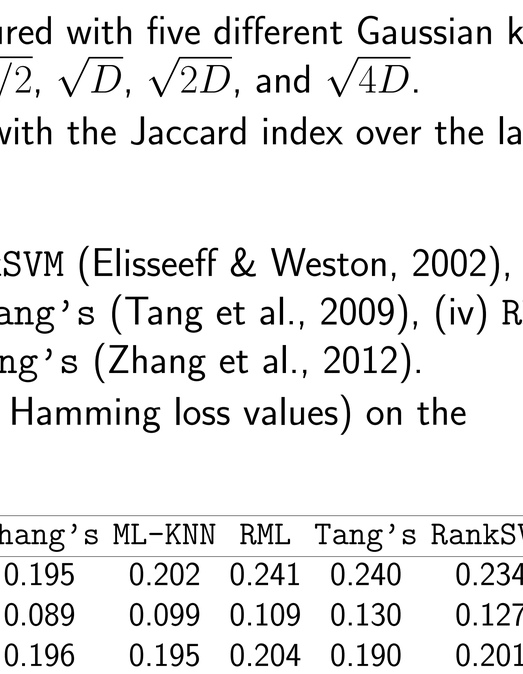
(d) binary classification

$$y^i_j|f^i_j \sim \delta(f^i_j y^i_j > \nu) \qquad \forall(i,j)$$

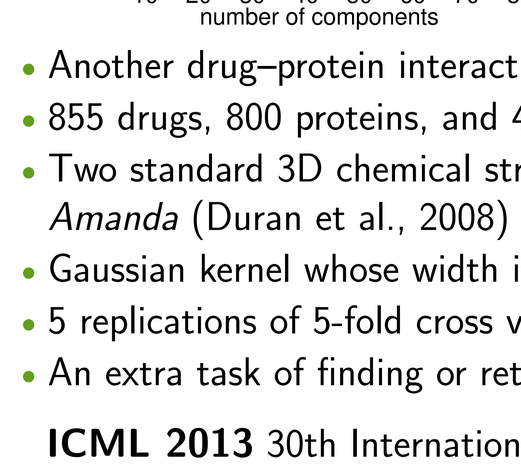
Toy Data Set

$$x^m_i \sim \mathcal{N}(x^m_i; 0, 1) \qquad \forall(m,i)$$
$$z^n_j \sim \mathcal{N}(z^n_j; 0, 1) \qquad \forall(n,j)$$
$$y^i_j|\boldsymbol{x}_i, \boldsymbol{z}_j \sim \mathcal{N}(y^i_j; x^1_i z^3_j + x^4_i z^8_j + x^7_i z^{10}_j, 1) \qquad \forall(i,j)$$
$$(N_\mathbf{x}, N_\mathbf{z}) = (40, 60) \text{ and } (P_\mathbf{x}, P_\mathbf{z}) = (15, 10)$$

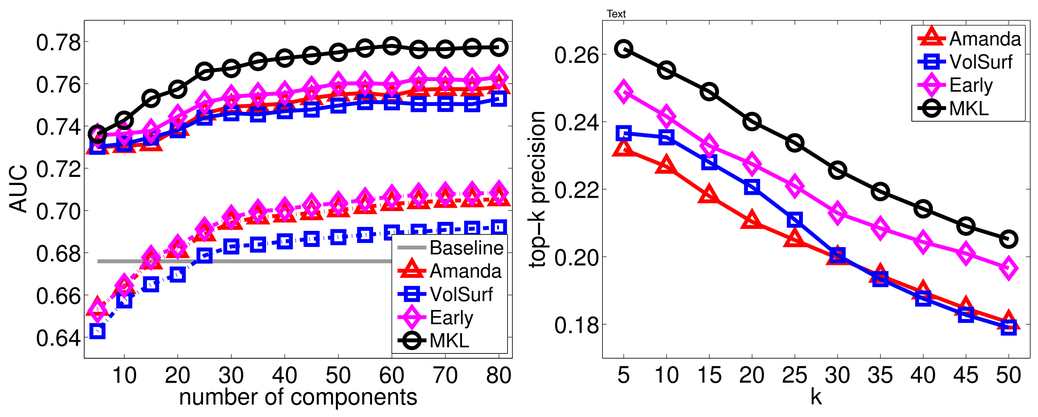
- A separate linear kernel for each feature



Drug–Protein Interaction Data Sets



- A drug–protein network by Yamanishi et al. (2008)
- 445 drugs, 664 proteins, and 2926 validated interactions
- `C`: chemical similarity for drugs
- `G`: genomic similarity for proteins
- `N`: network similarity for proteins
- 5 repl. of 5-fold CV over drugs
- `KBMF` is statistically significantly better than `KPMF` of Zhou et al. (2012) according to paired t-test ($p < 0.01$) on both data sets.



- Another drug–protein interaction network by Khan et al. (2012)
- 855 drugs, 800 proteins, and 4659 validated interactions
- Two standard 3D chemical structure descriptors for drugs: *Amanda* (Duran et al., 2008) and *VolSurf* (Cruciani et al., 2000)
- Gaussian kernel whose width is selected as $\sqrt{D}$
- 5 replications of 5-fold cross validation over drugs
- An extra task of finding or retrieving drugs with similar functions

Multilabel Classification Data Sets

- Samples and labels are assumed to be from domains $\mathcal{X}$ and $\mathcal{Z}$, respectively.
- Class membership matrix corresponds to target label matrix $\mathbf{Y}$ in our model.

- The similarities between samples are measured with five different Gaussian kernels whose widths are selected as $\sqrt{D/4}$, $\sqrt{D/2}$, $\sqrt{D}$, $\sqrt{2D}$, and $\sqrt{4D}$.
- The similarity between labels is measured with the Jaccard index over the labels of training samples.

- We compare with five algorithms: (i) `RankSVM` (Elisseeff & Weston, 2002), (ii) `ML-KNN` (Zhang & Zhou, 2007), (iii) `Tang's` (Tang et al., 2009), (iv) `RML` (Petterson & Caetano, 2010), and (v) `Zhang's` (Zhang et al., 2012).
- We report classification performances (i.e., Hamming loss values) on the multilabel classification 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

ICML 2013 30th International Conference on Machine Learning, 16–21 June 2013, Atlanta, Georgia, USA