Relational learning with many relations

Guillaume Obozinski

Laboratoire d'Informatique Gaspard Monge

École des Ponts - ParisTech

École des Ponts ParisTech

Joint work with Rodolphe Jenatton, Nicolas Le Roux and Antoine Bordes.

Labex Bézout - Huawei Seminar - April 3rd, 2015

Modelling relations between pairs of entities

Triplets:

Term 1 - Relation - Term 2

Modelling relations between pairs of entities

Triplets:

Term 1 - Relation - Term 2

Single relation

- Collaborative filtering
- Link prediction
- Modeling of social networks

Modelling relations between pairs of entities

Triplets:

Term 1 - Relation - Term 2

Single relation

- Collaborative filtering
- Link prediction
- Modeling of social networks

Multiple relations

- Collective classification
- Modelling in relational knowledge databases
- Proteins-protein and protein-ligand interactions
- Natural language semantics (and semantic role labelling)

Our motivation : Learning the semantic value of verbs

Model triplets:

Subject	Verb	Object
$\mathcal{S}_i$	$\mathcal{R}_j$	$\mathcal{O}_k$

Our motivation : Learning the semantic value of verbs

Model triplets:

Subject	Verb	Object
$\mathcal{S}_i$	$\mathcal{R}_j$	$\mathcal{O}_k$

View this as the relation:

$$\mathcal{R}_j(\mathcal{S}_i, \mathcal{O}_k) = 1$$

Different kinds of relational learning

Learn to predict relations from object attributes:

- Binary classification from pairs of feature vectors

Different kinds of relational learning

Learn to predict relations from object attributes:

- Binary classification from pairs of feature vectors

Exploit logical properties of relations: transitivity, implication, mutual exclusion, etc

- Markov Logic Networks (Kok and Domingos, 2007)

Different kinds of relational learning

Learn to predict relations from object attributes:

- Binary classification from pairs of feature vectors

Exploit logical properties of relations: transitivity, implication, mutual exclusion, etc

- Markov Logic Networks (Kok and Domingos, 2007)

Predict relations from some observed relations

Different kinds of relational learning

Learn to predict relations from object attributes:

- Binary classification from pairs of feature vectors

Exploit logical properties of relations: transitivity, implication, mutual exclusion, etc

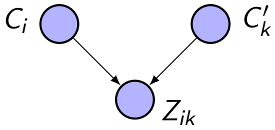
- Markov Logic Networks (Kok and Domingos, 2007)

Predict relations from some observed relations

- Idea: relations derive from unobserved latent attributes.
- Relational learning from *intrinsic latent attributes*

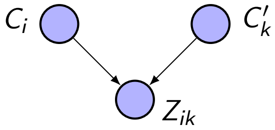
Stochastic Block Model

Wang and Wong (1987); Nowicki and Snijders (2001)



Stochastic Block Model

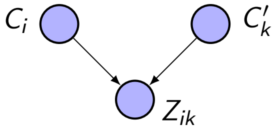
Wang and Wong (1987); Nowicki and Snijders (2001)



$$\mathbb{P}(Z_{ik} = 1) = \sum_{c,c'} \mathbb{P}(Z_{ik} = 1 \mid C_i = c,\ C'_k = c')\, \mathbb{P}(C_i = c)\, \mathbb{P}(C'_k = c')$$

Stochastic Block Model

Wang and Wong (1987); Nowicki and Snijders (2001)

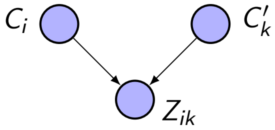


$$\mathbb{P}(Z_{ik} = 1) = \sum_{c,c'} \mathbb{P}(Z_{ik} = 1 \mid C_i = c,\ C'_k = c')\,\mathbb{P}(C_i = c)\,\mathbb{P}(C'_k = c')$$

$$\mathbf{P}_{ik} = \sum_{c,c'} \mathbf{R}_{cc'}\,\mathbf{S}_{ci}\,\mathbf{O}_{c'k} = (\mathbf{s}^i)^\top \mathbf{R}\mathbf{o}^k$$

Stochastic Block Model

Wang and Wong (1987); Nowicki and Snijders (2001)

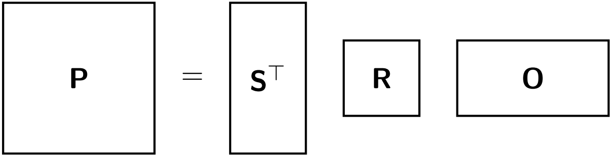


$$\mathbb{P}(Z_{ik} = 1) = \sum_{c,c'} \mathbb{P}(Z_{ik} = 1 \mid C_i = c,\ C'_k = c')\, \mathbb{P}(C_i = c)\, \mathbb{P}(C'_k = c')$$

$$\mathbf{P}_{ik} = \sum_{c,c'} \mathbf{R}_{cc'}\, \mathbf{S}_{ci}\, \mathbf{O}_{c'k} = (\mathbf{s}^i)^\top \mathbf{R} \mathbf{o}^k$$

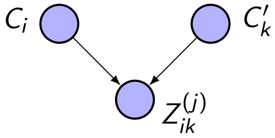
$$\mathbf{P} = \mathbf{S}^\top \mathbf{R}\, \mathbf{O}$$

A matrix factorization problem

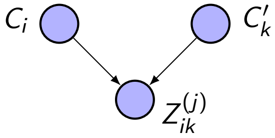


- $0 \leq \mathbf{R}_{ik} \leq 1$
- $\mathbf{o}^k \in \triangle$, $\mathbf{s}^i \in \triangle$ with $\triangle = \{\mathbf{x} \in \mathbb{R}_+^p \mid \|\mathbf{x}\|_1 = 1\}$

Stochastic Block Model for several relation types

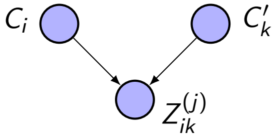


Stochastic Block Model for several relation types



$$\mathbb{P}(Z^{(j)}_{ik} = 1) = \sum_{c,c'} \mathbb{P}(Z^{(j)}_{ik} = 1 \mid C_i = c,\ C'_k = c')\,\mathbb{P}(C_i = c)\,\mathbb{P}(C'_k = c')$$

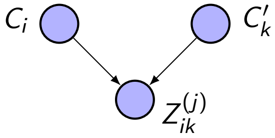
Stochastic Block Model for several relation types



$$\mathbb{P}(Z^{(j)}_{ik} = 1) = \sum_{c,c'} \mathbb{P}(Z^{(j)}_{ik} = 1 \mid C_i = c,\ C'_k = c')\,\mathbb{P}(C_i = c)\,\mathbb{P}(C'_k = c')$$

$$\mathbf{P}^{(j)}_{ik} = \sum_{c,c'} [\mathbf{R}_j]_{cc'}\,\mathbf{S}_{ci}\,\mathbf{O}_{c'k} = (\mathbf{s}^i)^\top \mathbf{R}_j\,\mathbf{o}^k$$

Stochastic Block Model for several relation types

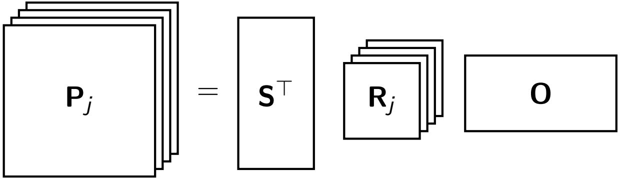


$$\mathbb{P}(Z_{ik}^{(j)} = 1) = \sum_{c,c'} \mathbb{P}(Z_{ik}^{(j)} = 1 \mid C_i = c,\ C'_k = c')\,\mathbb{P}(C_i = c)\,\mathbb{P}(C'_k = c')$$

$$\mathbf{P}_{ik}^{(j)} = \sum_{c,c'} [\mathbf{R}_j]_{cc'}\,\mathbf{S}_{ci}\,\mathbf{O}_{c'k} = (\mathbf{s}^i)^\top \mathbf{R}_j\,\mathbf{o}^k$$

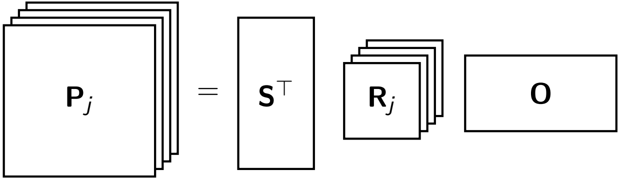
$$\mathbf{P}_j = \mathbf{S}^\top \mathbf{R}_j\,\mathbf{O}.$$

Collective matrix factorization



- $0 \leq [\mathbf{R}_j]_{ik} \leq 1$
- $\mathbf{o}^k \in \triangle$, $\mathbf{s}^i \in \triangle$ with $\triangle = \{\mathbf{x} \in \mathbb{R}_+^p \mid \|\mathbf{x}\|_1 = 1\}$

Collective matrix factorization

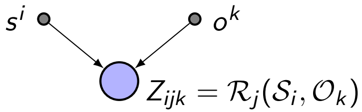


- $0 \leq [\mathbf{R}_j]_{ik} \leq 1$
- $\mathbf{o}^k \in \triangle$, $\mathbf{s}^i \in \triangle$ with $\triangle = \{\mathbf{x} \in \mathbb{R}_+^p \mid \|\mathbf{x}\|_1 = 1\}$

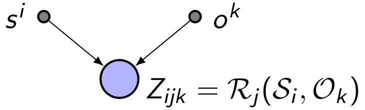
Corresponds to the approach used in RESCAL (Nickel et al., 2012)

$$\min_{\mathbf{S}=\mathbf{O},\mathbf{R}_j} \|\mathbf{Z}_j - \mathbf{P}_j\|_F^2$$

A bilinear logistic model



A bilinear logistic model

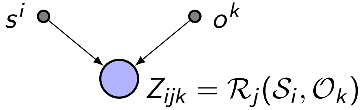


$$\mathbb{P}(\mathcal{R}_j(\mathcal{S}_i, \mathcal{O}_k) = 1) = \mathbf{P}_{ik}^{(j)} = \left(1 + \exp -\eta_{ik}^{(j)}\right)^{-1}$$

with an "energy"

$$\mathcal{E}(\mathbf{s}^i, \mathbf{R}_j, \mathbf{o}^k) = \eta_{ik}^{(j)} = \langle \mathbf{s}^i, \mathbf{R}_j \, \mathbf{o}^k \rangle$$

A bilinear logistic model



$$\mathbb{P}(\mathcal{R}_j(\mathcal{S}_i, \mathcal{O}_k) = 1) = \mathbf{P}^{(j)}_{ik} = \left(1 + \exp -\eta^{(j)}_{ik}\right)^{-1}$$

with an "energy"

$$\mathcal{E}(\mathbf{s}^i, \mathbf{R}_j, \mathbf{o}^k) = \eta^{(j)}_{ik} = \langle \mathbf{s}^i, \mathbf{R}_j \, \mathbf{o}^k \rangle$$

So that with

$$\mathbf{H}^{(j)} = (\eta^{(j)}_{ik})_{1 \le i,k \le n}$$

we have

$$\mathbf{H}^{(j)} = \mathbf{S}^\top \mathbf{R}_j \mathbf{O}$$

Dealing with the number of parameters? : related work

Dealing with the number of parameters? : related work

Clustering of Entities and Relations

- Miller et al. (2009); Zhu (2012)
- Bayesian Non-parametric clustering: Kemp et al. (2006); Sutskever et al. (2009)
- Clustering in the context of Markov Logic Network: Kok and Domingos (2007)

Dealing with the number of parameters? : related work

Clustering of Entities and Relations

- Miller et al. (2009); Zhu (2012)
- Bayesian Non-parametric clustering: Kemp et al. (2006); Sutskever et al. (2009)
- Clustering in the context of Markov Logic Network: Kok and Domingos (2007)

Embeddings

- Collective Matrix Factorization by (Nickel et al., 2012) (RESCAL)
- Semantic Matching Energy (SME) model of Bordes et al. (2012): encodes relations as vectors for scalability.

Dealing with the number of parameters? : related work

Clustering of Entities and Relations

- Miller et al. (2009); Zhu (2012)
- Bayesian Non-parametric clustering: Kemp et al. (2006); Sutskever et al. (2009)
- Clustering in the context of Markov Logic Network: Kok and Domingos (2007)

Embeddings

- Collective Matrix Factorization by (Nickel et al., 2012) (RESCAL)
- Semantic Matching Energy (SME) model of Bordes et al. (2012): encodes relations as vectors for scalability.

Tensor factorization

- CANDECOMP/PARAFAC Tucker (1966); Harshman and Lundy (1994)
- Probabilistic formulation of Chu and Ghahramani (2009)

Our solution: *Latent relational factors*

Idea: Modelling the relations between the relations...

Our solution: *Latent relational factors*

Idea: Modelling the relations between the relations...

$$\mathbf{R}_j = \sum_{r=1}^{d} \alpha_r^j \, \boldsymbol{\Theta}_r, \qquad \text{with} \quad \boldsymbol{\Theta}_r = \mathbf{u}_r \mathbf{v}_r^\top$$

for some sparse vector $\alpha^j \in \mathbb{R}^d$.

Our solution: *Latent relational factors*

Idea: Modelling the relations between the relations...

$$\mathbf{R}_j = \sum_{r=1}^{d} \alpha_r^j \, \mathbf{\Theta}_r, \qquad \text{with} \quad \mathbf{\Theta}_r = \mathbf{u}_r \mathbf{v}_r^\top$$

for some sparse vector $\alpha^j \in \mathbb{R}^d$.

Given

- n_r number of relations
- p embedding dimension: $\mathbf{R}_j \in \mathbb{R}^{p \times p}$
- d number of latent relational factors
- $\bar{s} \leq \lambda \, d$ average number of non-zero α coefficients

Our solution: *Latent relational factors*

Idea: Modelling the relations between the relations...

$$\mathbf{R}_j = \sum_{r=1}^{d} \alpha_r^j \, \mathbf{\Theta}_r, \qquad \text{with} \quad \mathbf{\Theta}_r = \mathbf{u}_r \mathbf{v}_r^\top$$

for some sparse vector $\alpha^j \in \mathbb{R}^d$.

Given

- n_r number of relations
- p embedding dimension: $\mathbf{R}_j \in \mathbb{R}^{p \times p}$
- d number of latent relational factors
- $\bar{s} \leq \lambda\, d$ average number of non-zero α coefficients

$\Rightarrow$ we reduce the # of parameters from $n_r\, p^2$ to $2pd + \bar{s} n_r$

Algorithmic approach

- Large scale $|\mathcal{P}| = 10^6$

Algorithmic approach

- Large scale $|\mathcal{P}| = 10^6$
- Stochastic projected block-coordinate gradient descent algorithm

Algorithmic approach

- Large scale $|\mathcal{P}| = 10^6$
- Stochastic projected block-coordinate gradient descent algorithm
- Mini-batches of 100 triplets

Algorithmic approach

- Large scale $|\mathcal{P}| = 10^6$
- Stochastic projected block-coordinate gradient descent algorithm
- Mini-batches of 100 triplets
- For each positive triplet (i, j, k), sampling negative triplets (i, j', k).

Tensor factorization interpretation of our model

$$\eta_{ik}^{(j)} = \langle \mathbf{s}^i, \mathbf{R}_j \mathbf{o}^k \rangle \quad =$$

Tensor factorization interpretation of our model

$$\eta_{ik}^{(j)} = \langle \mathbf{s}^i, \mathbf{R}_j \mathbf{o}^k \rangle \quad = \quad (\mathbf{s}^i)^\top \Big[\sum_{r=1}^{d} \alpha_r^j \mathbf{u}_r \mathbf{v}_r^\top \Big] \mathbf{o}^k$$

Tensor factorization interpretation of our model

$$
\begin{aligned}
\eta_{ik}^{(j)} = \langle \mathbf{s}^i, \mathbf{R}_j \mathbf{o}^k \rangle &= (\mathbf{s}^i)^\top \Big[\sum_{r=1}^{d} \alpha_r^j \mathbf{u}_r \mathbf{v}_r^\top \Big] \mathbf{o}^k \\
&= \sum_{r=1}^{d} \alpha_r^j \, ((\mathbf{s}^i)^\top \mathbf{u}_r)(\mathbf{v}_r^\top \mathbf{o}^k)
\end{aligned}
$$

Tensor factorization interpretation of our model

$$\eta_{ik}^{(j)} = \langle \mathbf{s}^i, \mathbf{R}_j \mathbf{o}^k \rangle = (\mathbf{s}^i)^\top \Big[\sum_{r=1}^{d} \alpha_r^j \mathbf{u}_r \mathbf{v}_r^\top \Big] \mathbf{o}^k$$

$$= \sum_{r=1}^{d} \alpha_r^j \, ((\mathbf{s}^i)^\top \mathbf{u}_r)(\mathbf{v}_r^\top \mathbf{o}^k)$$

$$= \sum_{r=1}^{d} \alpha_r^j \beta_r^i \gamma_r^k \quad \text{with} \quad \boldsymbol{\beta}_r = \mathbf{S}^\top \mathbf{u}_r, \quad \boldsymbol{\gamma}_r = \mathbf{O}^\top \mathbf{v}_r$$

Tensor factorization interpretation of our model

$$\eta_{ik}^{(j)} = \langle \mathbf{s}^i, \mathbf{R}_j \mathbf{o}^k \rangle = (\mathbf{s}^i)^\top \Big[\sum_{r=1}^{d} \alpha_r^j \mathbf{u}_r \mathbf{v}_r^\top \Big] \mathbf{o}^k$$

$$= \sum_{r=1}^{d} \alpha_r^j \, ((\mathbf{s}^i)^\top \mathbf{u}_r)(\mathbf{v}_r^\top \mathbf{o}^k)$$

$$= \sum_{r=1}^{d} \alpha_r^j \beta_r^i \gamma_r^k \quad \text{with} \quad \boldsymbol{\beta}_r = \mathbf{S}^\top \mathbf{u}_r, \quad \boldsymbol{\gamma}_r = \mathbf{O}^\top \mathbf{v}_r$$

So, $\mathbf{H}$ is related to $\mathbf{R}$ via

$$\mathbf{H} = (\mathbf{I} \otimes \mathbf{S}^\top \otimes \mathbf{O}^\top)\, \mathbf{R} = \sum_{r=1}^{d} (\mathbf{I}\, \boldsymbol{\alpha}_r) \otimes (\mathbf{S}^\top \mathbf{u}_r) \otimes (\mathbf{O}^\top \mathbf{v}_r)$$

i.e. $\mathbf{H}$ is constrained to be the image of the lower dimensional tensor $\mathbf{R}$.

Experiments

Learning semantic representation for verbs

Data

- 2,000,000 Wikipedia articles
- POS-tagging + chunking+ lemmatization+ semantic role labelling using SENNA (Collobert et al., 2011)
- keeping sentences with syntax subject - verb - direct object
- with each term = a single word from the WordNet lexicon

Learning semantic representation for verbs

Data

- 2,000,000 Wikipedia articles
- POS-tagging + chunking+ lemmatization+ semantic role labelling using SENNA (Collobert et al., 2011)
- keeping sentences with syntax subject - verb - direct object
- with each term = a single word from the WordNet lexicon

Data Characteristics

- Dictionary of $30,605$ words
- $n_r = 4,547$ relations
- Training set: $1,000,000$ unique triplets
- Validation set: $50,000$ unique triplets
- Testing set: $250,000$ unique triplets

Learning semantic representation of verbs

Learning semantic representation of verbs

Hyperparameters

- Embedding dimension $p \in \{25, 50, 100\}$
- Number of latent decompositions matrices $d \in \{50, 100, 200\}$
- Sparsity level as $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1\} \times (n_r \times d)$
- Weighting of negative triplets

Learning semantic representation of verbs

Hyperparameters

- Embedding dimension $p \in \{25, 50, 100\}$
- Number of latent decompositions matrices $d \in \{50, 100, 200\}$
- Sparsity level as $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1\} \times (n_r \times d)$
- Weighting of negative triplets

Actual reduction of the number of parameters

"From $n_r\, p^2$ parameters to $2pd + \bar{s} n_r$"

Learning semantic representation of verbs

Hyperparameters

- Embedding dimension $p \in \{25, 50, 100\}$
- Number of latent decompositions matrices $d \in \{50, 100, 200\}$
- Sparsity level as $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1\} \times (n_r \times d)$
- Weighting of negative triplets

Actual reduction of the number of parameters

"From $n_r\, p^2$ parameters to $2pd + \bar{s} n_r$"

With $n_r = 4,547$, $p = 25$ and $d = 200$,

From 2,841,875 to 19,104.

Verb prediction

Verb prediction

- Rank of the correct verb
- Fraction of examples where the correct verb is in the top $z\%$ (average Recall at precision $(100 - z)\%$)

Verb prediction

- Rank of the correct verb
- Fraction of examples where the correct verb is in the top z% (average Recall at precision $(100 - z)$%)

	synonyms not considered		
	median/mean rank	p@5	p@20
Our approach	50 / **195.0**	**0.78**	**0.95**
SME Bordes et al. (2012)	56 / 199.6	0.77	**0.95**
Bigram	**48** / 517.4	0.72	0.83

Verb prediction

- Rank of the correct verb
- Fraction of examples where the correct verb is in the top $z\%$ (average Recall at precision $(100 - z)\%$)

	synonyms not considered		
	median/mean rank	p@5	p@20
Our approach	50 / **195.0**	**0.78**	**0.95**
SME Bordes et al. (2012)	56 / 199.6	0.77	**0.95**
Bigram	**48** / 517.4	0.72	0.83

	best synonyms considered		
	median/mean rank	p@5	p@20
Our approach	19 / **96.7**	**0.89**	**0.98**
SME Bordes et al. (2012)	19 / 99.2	**0.89**	**0.98**
Bigram	**17** / 157.7	0.87	0.95

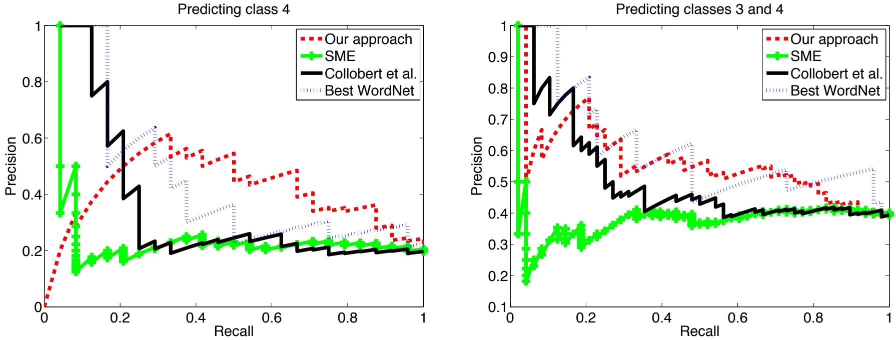
Lexical Similarity Classification

Given two verbs are they similar semantically or not?

Data (Yang and Powers, 2006)

- 130 pairs of verbs
- labeled with score in $\{0, 1, 2, 3, 4\}$
- Ex:
  - `(divide, split)` score 4
  - `(postpone, show)` score 0

Lexical Similarity prediction results: PR curves



Similarity measures between verbs from

- our approach,
- SME Bordes et al. (2012),
- Collobert et al. (2011)
- the best (out of three) WordNet similarity measure (counting the number of nodes along te shortest path in the "is-a" hierarchy).

Conclusions

- Highly multi-relational data is worth modelling
- Relational learning from *intrinsic latent attributes*
- Matrix factorization models arising from variants on the stochastic block model

Conclusions

- Highly multi-relational data is worth modelling
- Relational learning from *intrinsic latent attributes*
- Matrix factorization models arising from variants on the stochastic block model
- Our approach ties or beats existing approaches on benchmark datasets
- Scales to
  - almost 5000 relations
  - more than 30,000 entities
  - 1,000,000 training triplets
- Trigram modeling
  - crucial in benchmark relational learning datasets
  - marginal in the NLP experiment

References I

Bordes, A., Glorot, X., Weston, J., and Bengio, Y. (2012). A semantic matching energy function for learning with multi-relational data. *Machine Learning*. To appear.

Chu, W. and Ghahramani, Z. (2009). Probabilistic models for incomplete multi-dimensional arrays. *Journal of Machine Learning Research - Proceedings Track*, 5:89–96.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *JMLR*, 12:2493–2537.

Harshman, R. A. and Lundy, M. E. (1994). Parafac: parallel factor analysis. *Comput. Stat. Data Anal.*, 18(1):39–72.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proc. of AAAI*, pages 381–388.

Kok, S. and Domingos, P. (2007). Statistical predicate invention. In *Proceedings of the 24th international conference on Machine learning*, pages 433–440.

Miller, K., Griffiths, T., and Jordan, M. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems 22*, pages 1276–1284.

Nickel, M., Tresp, V., and Kriegel, H.-P. (2012). Factorizing YAGO: scalable machine learning for linked data. In *Proc. of the 21st intl conf. on WWW*, pages 271–280.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41.

References II

Sutskever, I., Salakhutdinov, R., and Tenenbaum, J. (2009). Modelling relational data using bayesian clustered tensor factorization. In *Adv. in Neur. Inf. Proc. Syst. 22*.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397).

Yang, D. and Powers, D. M. W. (2006). Verb similarity on the taxonomy of wordnet. *Proceedings of GWC-06*, pages 121–128.

Zhu, J. (2012). Max-margin nonparametric latent feature models for link prediction. In *Proceedings of the 29th Intl Conference on Machine Learning*.

Formulation of the optimization problem

$$\min_{\substack{\mathbf{S},\mathbf{O},\{\boldsymbol{\alpha}^j\},\{\boldsymbol{\Theta}_r\},\\ \mathbf{y},\mathbf{y}',\mathbf{z},\mathbf{z}'}} \quad \sum_{(i,j,k)\in\mathcal{P}} \eta_{ik}^{(j)} - \sum_{(i,j,k)\in\mathcal{P}\cup\mathcal{N}} \log(1+\exp(\eta_{ik}^{(j)})),$$

$$\begin{aligned}
\text{s.t.} \quad & \eta_{ik}^{(j)} = \mathcal{E}(\mathbf{s}^i, \mathbf{R}_j, \mathbf{o}^k), \\
& \mathbf{R}_j = \sum_{r=1}^{d} \alpha_r^j \, \mathbf{u}_r \cdot \mathbf{v}_r^\top, \quad \|\boldsymbol{\alpha}^j\|_1 \leq \lambda, \\
& \mathbf{O} = \mathbf{S}, \quad \mathbf{z} = \mathbf{z}', \\
& \mathbf{s}^j, \mathbf{o}^k, \mathbf{y}, \mathbf{y}', \mathbf{z}, \mathbf{u}_r \text{ and } \mathbf{v}_r \text{ in the ball } \big\{\mathbf{w};\, \|\mathbf{w}\|_2 \leq 1\big\}
\end{aligned}$$