Structure Learning in Sequential Data

Liam Stewart
liam@cs.toronto.edu

Richard Zemel
zemel@cs.toronto.edu

2005.09.19

Motivation

A. Cau, R. Kuiper, and W.-P. de Roever. Formalising Dijkstra's development strategy within Stark's formalism. In C. B. Jones, R. C. Shaw, and T. Denvir, editors, Proc. 5th. BCS-FACS Refinement Workshop, 1992.

A. Cau, R. Kuiper, and W.-P. de Roever. Formalising Dijkstra's development strategy within Stark's formalism. In C. B. Jones, R. C. Shaw, and T. Denvir, editors, Proc. 5th. BCS-FACS Refinement Workshop, 1992.

Structure Learning

Structure: patterns and relationships ...

- between labels and observations,
- between labels.

Basic components:

- Observations $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T | \mathbf{x}_i \in \mathbb{R}^d\}$
- Labels $Y = \{y_1, \ldots, y_T | y_i \in \mathcal{Y}\}$

Goal: learn a mapping from observations to labels.

Why is it difficult?

If sequences are considered as a whole:

- Observations have indeterminate dimensionality,
- Number of joint classifications of labels is exponential in T.

If the items are considered separately:

- Information about label structures is lost.

How do we learn a mapping while maintaining tractability and exploiting structure?

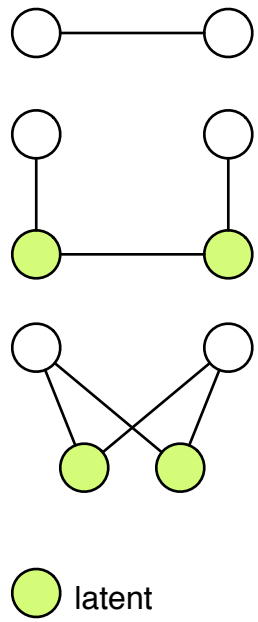
Outline

1. Probabilistic models for structure learning
2. Experimental results
3. Conclusions and future work

Probabilistic Models

Choices:

- Generative vs. conditional.
- Directed vs. undirected model.
- Use latent states?
- Connectivity?



Generative models express the conditional distribution P(Y|X) in terms of the joint distribution P(Y,X) and require a model of the observations P(X).

We focus on conditional models.

Conditional Models

These model the conditional distribution p(Y|X) directly and make extensive use of features: functions of the observations.

Features can be overlapping and non-independent.

Examples:

- Regular Expressions: [A-Z]
- Category: "is a name"
- Exact match: Morgan

Assumption: raw observations have been pre-processed by a set of feature functions.

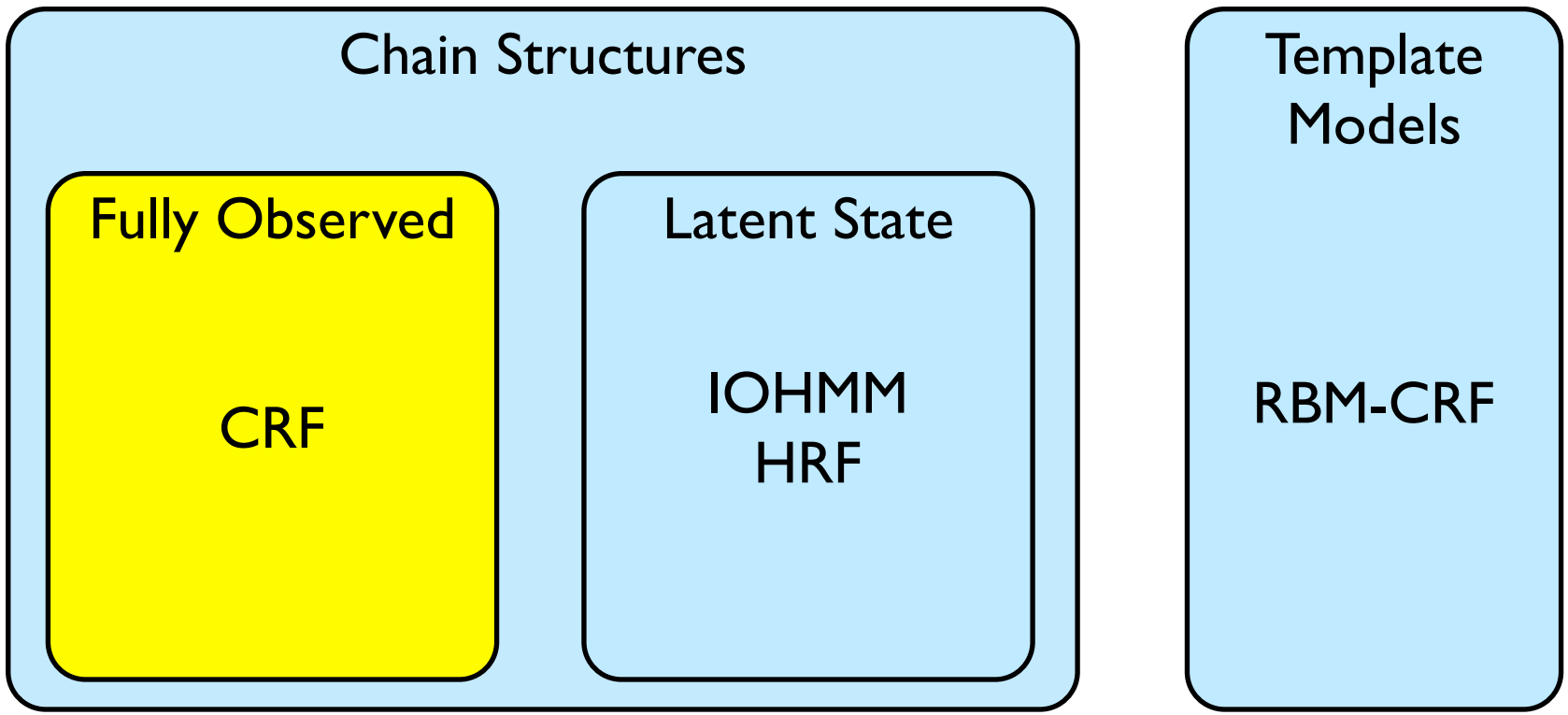
Models

We implemented

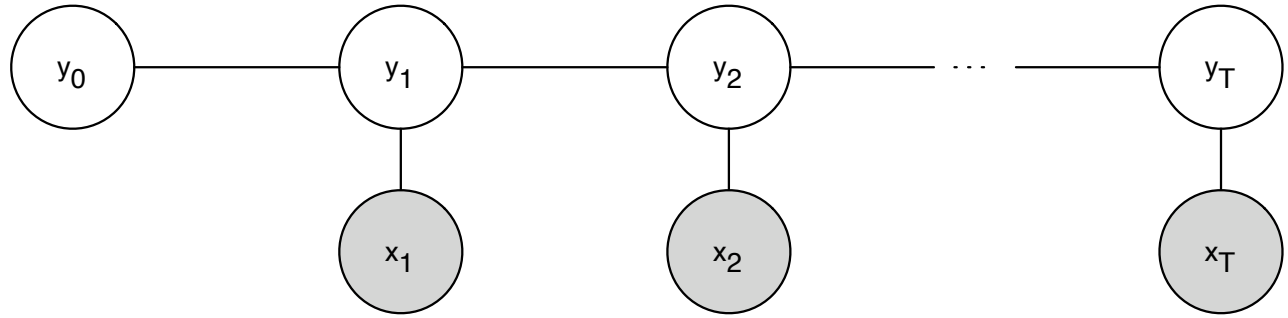
- logistic regression (LR)
- the Maximum Entropy Markov Model (MEMM) and a version with observation independent transitions (the IMEMM),
- the Conditional Random Field (CRF) and a version with observation independent edge potentials (the ICRF).
- the Input Output Hidden Markov Model (IOHMM)
- the Hidden Random Field (HRF)
- the RBM-CRF template model.

We report results for LR, the CRF, the IOHMM, the HRF, and RBM-CRF models.

Conditional Models



CRF

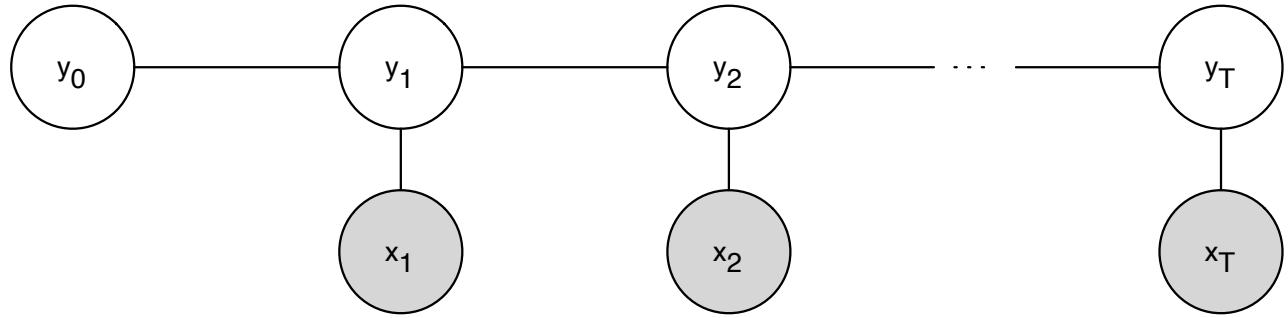


Conditional random field [Lafferty, McCallum, Pereira 2002]

It is the undirected version of the maximum entropy Markov model (MEMM).

$$
\begin{aligned}
p(Y|X) &= \frac{1}{Z(X)} \prod_{t=1}^{T} \phi(y_t, y_{t-1}, \mathbf{x}_t) \\
&= \frac{1}{Z(X)} \prod_{t=1}^{T} \exp\{\boldsymbol{\theta}_{y_t, y_{t-1}} \cdot \mathbf{x}_t\}
\end{aligned}
$$

CRF



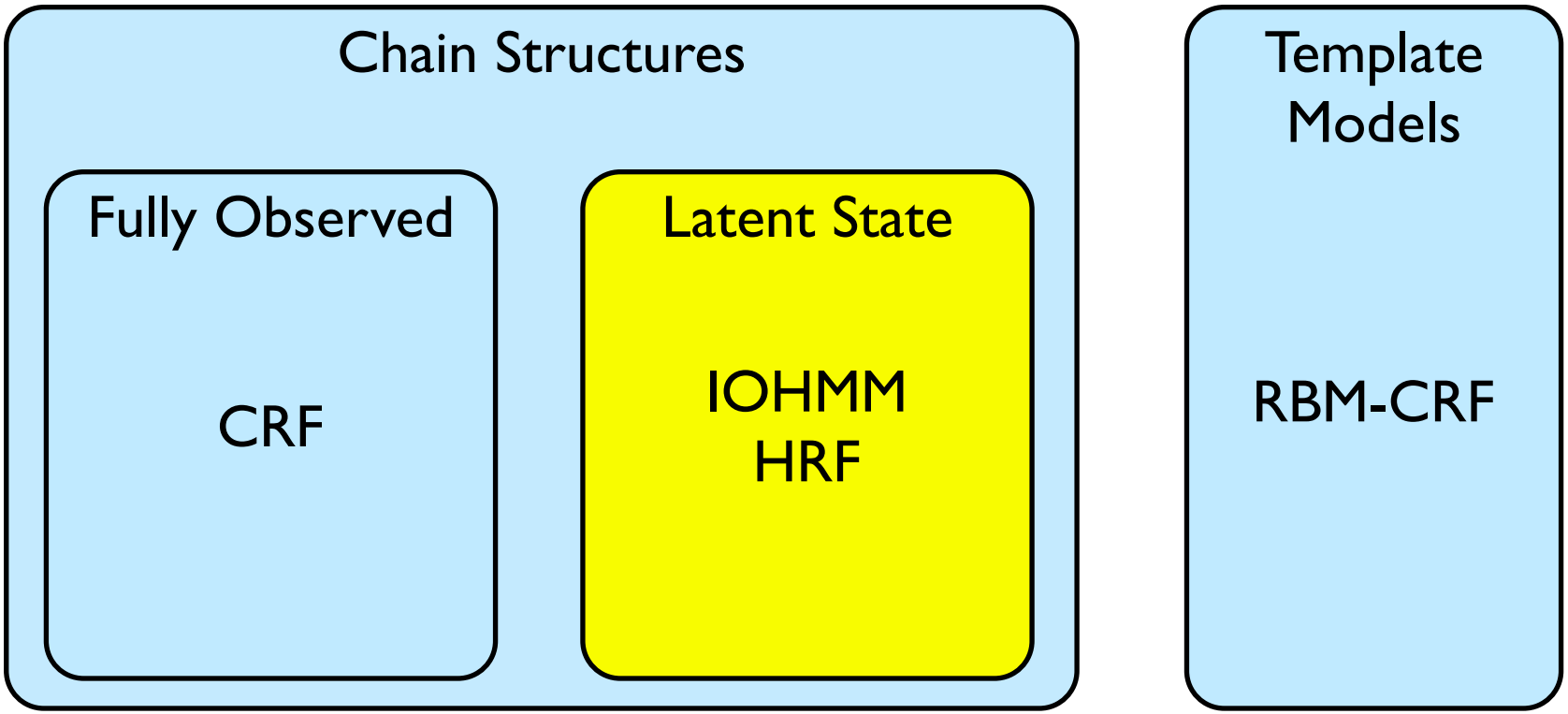
Most implementations use a second-order quasi-Newton optimizer (e.g. L-BFGS) to optimize the weights.

Let $m_{i,t,j}$ be one when $y_t^{(i)} = j$.

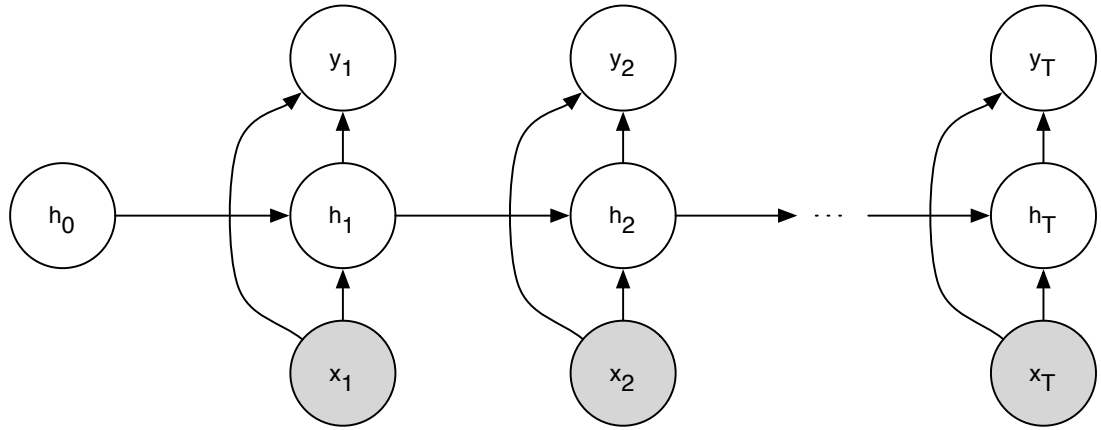
$$\begin{aligned}\ell(\Theta; \mathcal{D}) &= \sum_{i=1}^{N} \log p(Y_i|X_i) \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T_i} \boldsymbol{\theta}_{y_t^{(i)}, y_{t-1}^{(i)}} \cdot \mathbf{x}_t^{(i)} - \sum_{i=1}^{N} \log Z(X_i)\end{aligned}$$

$$\nabla_{\boldsymbol{\theta}_{jk}} \ell(\Theta; \mathcal{D}) = \sum_{i=1}^{N}\sum_{t=1}^{T_i} [m_{i,t,j} m_{i,t-1,k} - p(y_t = j, y_{t-1} = k|X_i)]\mathbf{x}_t^{(i)}$$

Conditional Models



IOHMM



Input output HMM [Bengio, Frasconi 1995]

It is an HMM where the transition and emission distributions are conditional on the observations.

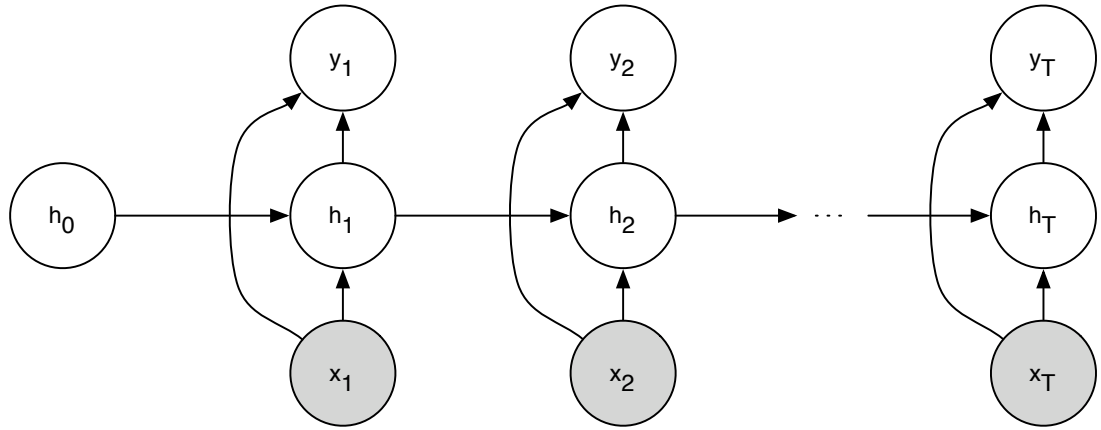
A set of latent random variables H that form a chain structure are used to maintain state information.

$$p(Y|X) = \sum_{H} \prod_{t=1}^{T} p(h_t|h_{t-1}, \mathbf{x}_t) p(y_t|h_t, \mathbf{x}_t).$$

$$p(h_t = j|h_{t-1} = k, \mathbf{x}_t) \propto \exp\{\boldsymbol{\lambda}_{jk} \cdot \mathbf{x}_t\}$$

$$p(y_t = j|h_t = k, \mathbf{x}_t) \propto \exp\{\boldsymbol{\theta}_{jk} \cdot \mathbf{x}_t\}$$

IOHMM



The EM algorithm is used to train the IOHMM.

In the E step, we need to calculate the posterior distributions:

$$p(h_t^{(i)} = j | Y_i, X_i)$$

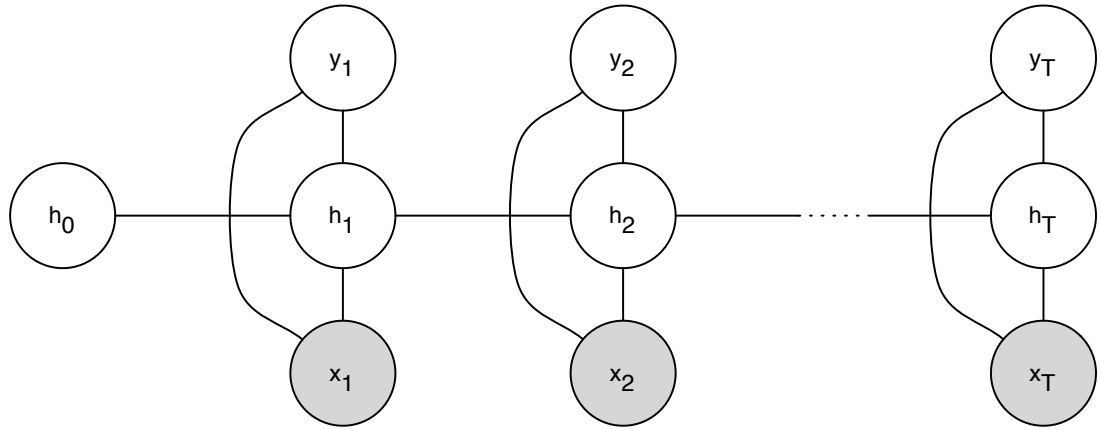
$$p(h_t^{(i)} = j, h_{t-1}^{(i)} = k | Y_i, X_i)$$

The M step updates:

1. the transition weights,
2. the emission weights.

Both of these updates are weighted logistic regression problems.

HRF



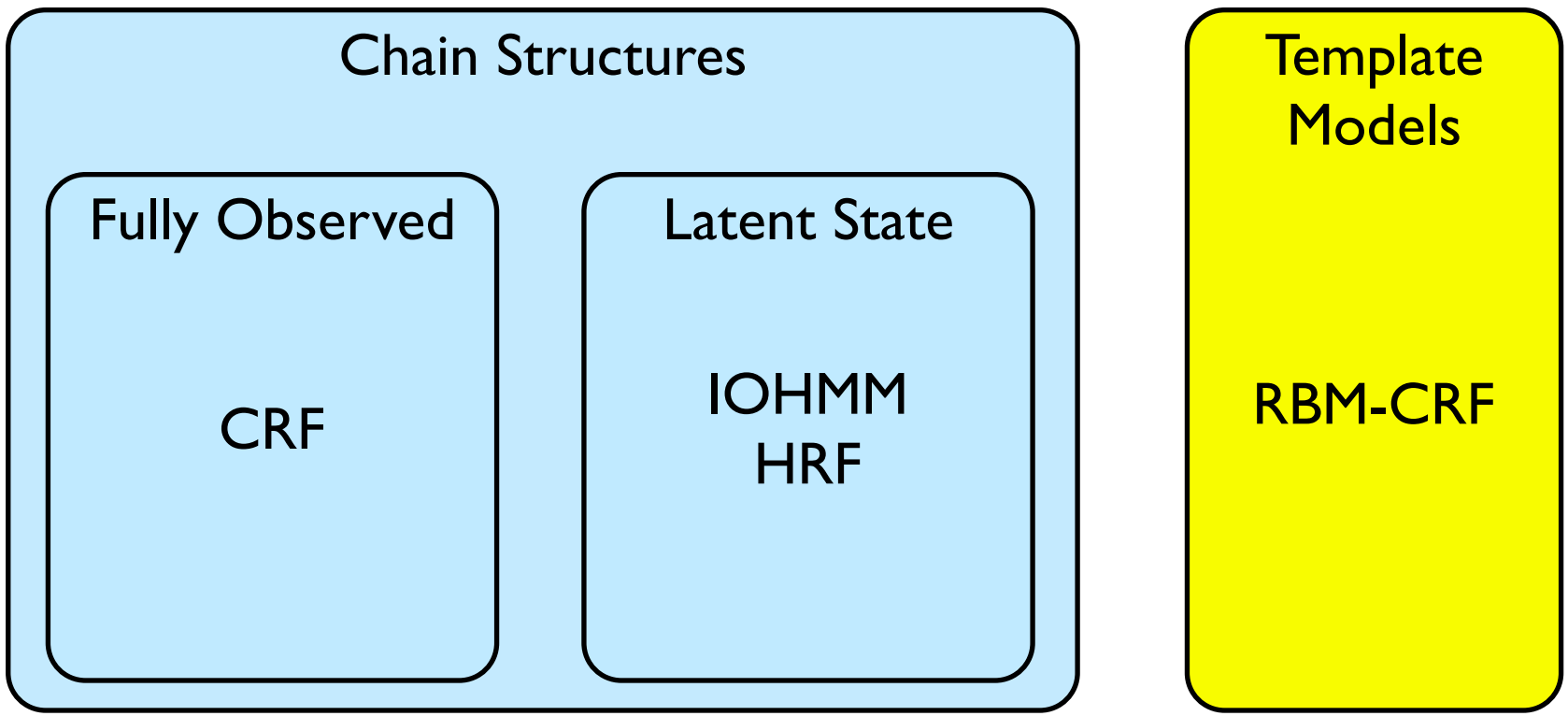
Hidden random field [Kakade, Teh, Roweis 2002]

It is the undirected equivalent of the IOHMM.

$$
\begin{aligned}
p(Y|X) &= \sum_H p(Y, H|X) \\
&= \frac{1}{Z(X)} \sum_H \prod_{t=1}^{T} \phi(h_t, h_{t-1}, \mathbf{x}_t)\phi(y_t, h_t, \mathbf{x}_t) \\
&= \frac{1}{Z(X)} \sum_H \prod_{t=1}^{T} \exp\{\boldsymbol{\lambda}_{h_t,h_{t-1}} \cdot \mathbf{x}_t + \boldsymbol{\theta}_{y_t,h_t} \cdot \mathbf{x}_t\}
\end{aligned}
$$

Like the IOHMM, training uses the EM algorithm. The M step involves optimizing a fully-observed CRF.

Conditional Models



Label Features

CRF models are fully specified with respect to label configurations.

Adding higher-order features results in an exponential increase in model complexity.

Idea: learn higher-order structures from the data using label features.

Label Feature: parameterized feature on a group of label variables. The weights can be adjusted to change what the label feature looks for.

RBM-CRF

Restricted Boltzmann machine CRF [He,Zemel,Carreira-Perpinan 2004]

A label feature is a binary random variable connected to J labels.

Define label features through groups:

$$g = \left\langle n_g, \mathbf{o}_g, \{W_{g,n}\}_{n=1}^{n_g}, \mathbf{b}_g \right\rangle$$

n_g : number of label features in the group.

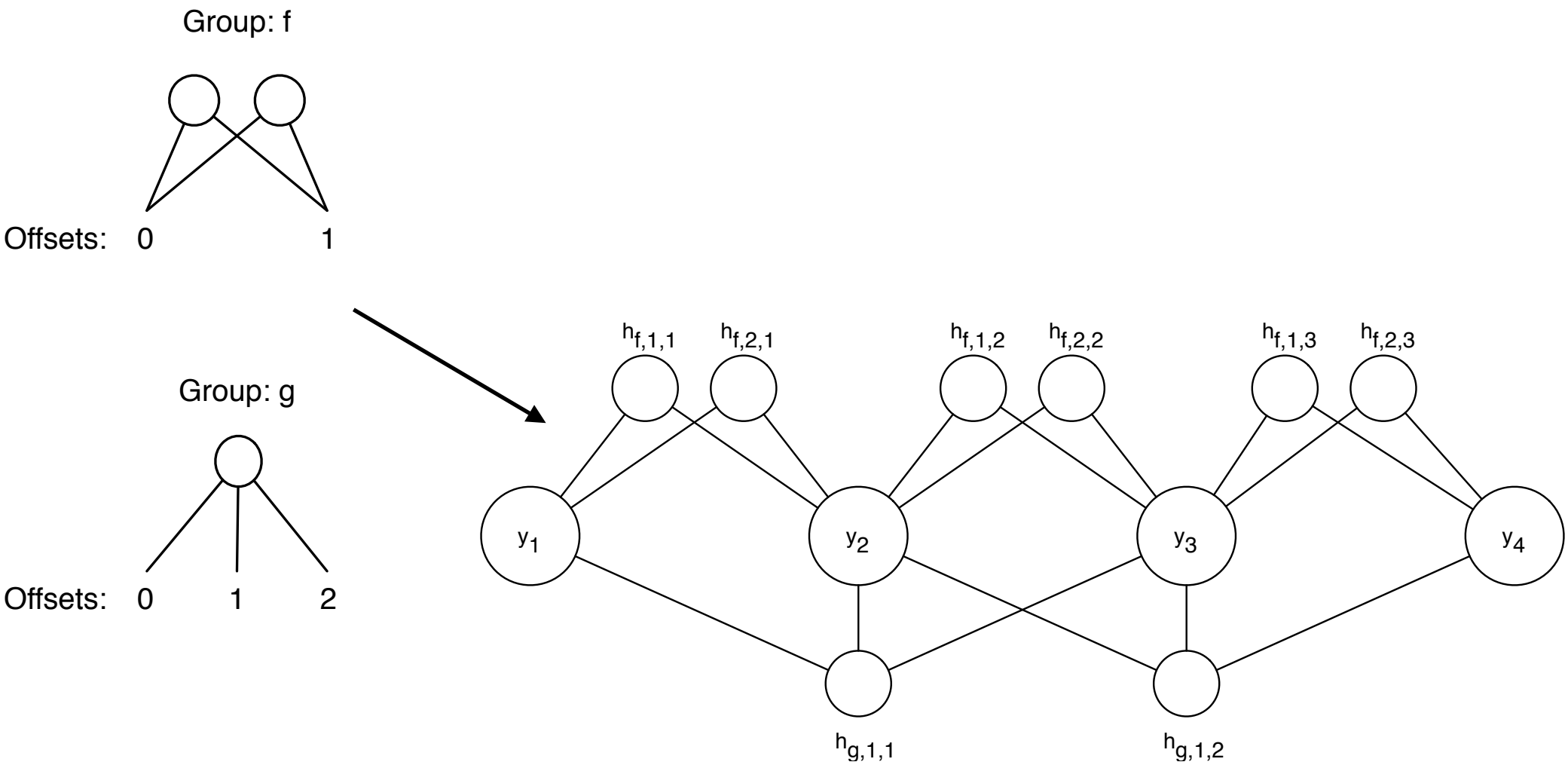
$\mathbf{o}_g$: vector of offsets that specifies the connectivity.

Instantiate a group at time t (replication r):

$$g(t) = \left\langle n_g, \mathbf{o}_g(t), \{W_{g,n}\}_{n=1}^{n_g}, \mathbf{b}_g \right\rangle$$

$$\mathbf{o}_g(t) = \mathbf{o}_g + t$$

RBM-CRF: Example



RBM-CRF

An RBM-CRF is just a collection of groups.

When rolled-out across a sequence, the model has the form of a restricted Boltzmann machine (RBM).

$$
\begin{aligned}
p_g(Y, H) &\propto \exp\{\sum_r \sum_n [b_{g,n} + \sum_k \mathbf{w}_{g,n,k} \cdot \mathbf{l}_k] h_{g,n,r}\} \\
&= \prod_r \prod_n \exp\{[b_{g,n} + \sum_k \mathbf{w}_{g,n,k} \cdot \mathbf{l}_k] h_{g,n,r}\} \\
&= \prod_r \prod_n \tilde{p}_{g,n,r}(h_{g,n,r}, Y)
\end{aligned}
$$

RBM-CRF: Inference and Training

Exact inference is hard because of the loopy structure.

However, we can exploit the bipartite nature of the graph to implement a block Gibbs sampler.

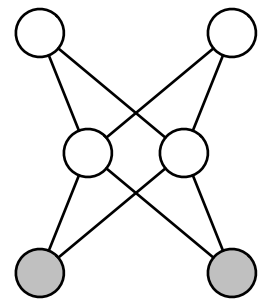
RBM-CRF models are products of experts so we can use contrastive divergence to train them.

RBM-CRF: Observations

There are several ways to incorporate observations into RBM-CRF models.

1. Use a pre-trained classifier (e.g.: logistic regression, CRF).
2. Train a classifier at the same time as the RBM-CRF.
3. Incorporate observations directly into the parameterization of the label feature.

$$p_g(Y, H|X) \propto \exp\{\sum_r \sum_n [b_{g,n} + \sum_j \boldsymbol{\theta}_{g,n,j} \cdot \mathbf{x}_j + \sum_k \mathbf{w}_{g,n,k} \cdot \mathbf{l}_k] h_{g,n,r}\}$$



Outline

1. Probabilistic models for structure learning
2. Experimental results
3. Conclusions and future work

Evaluation Metrics

With respect to label l, let

- A be the number of true positives, B be the number of false negatives
- C be the number of false positives, D be the number of true negatives

F1 score: $F1 = (2A)/(2A + B + C)$

Accuracy: number of items labeled correctly divided by the total number of items in a sequence.

Instance accuracy: percentage of sequences labeled entirely correctly.

We use Viterbi decoding for the CRF, IOHMM, and HRF and maximum marginals for logistic regression and the RBM-CRF.

Toy Problem 1

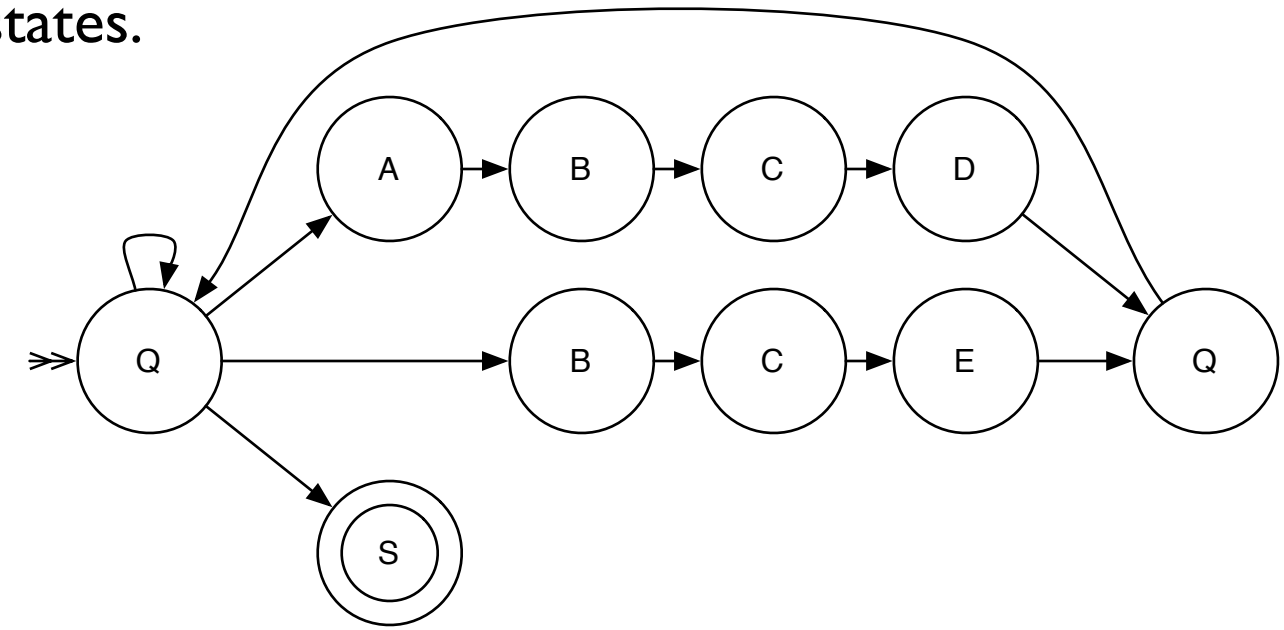
100 training instances &100 test instances, each of length 50.

Models require memory / knowledge of structures to classify observation 5.

RBM-CRF models look at groups of six label variables.

IOHMM/HRF have 3 states.

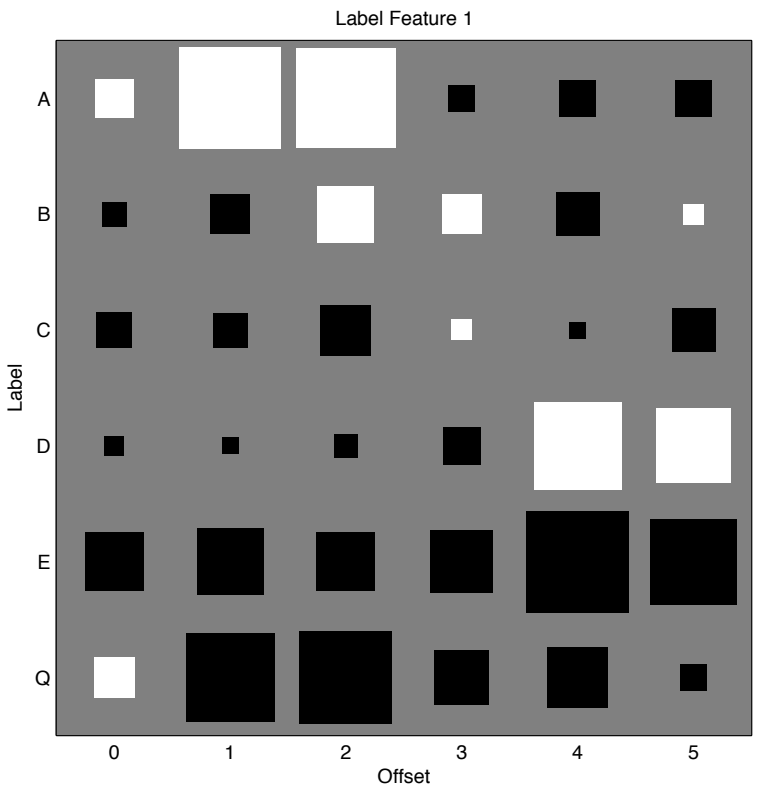
Observation	Labels
1	A
2	B
3	C
4	Q
5	D,E

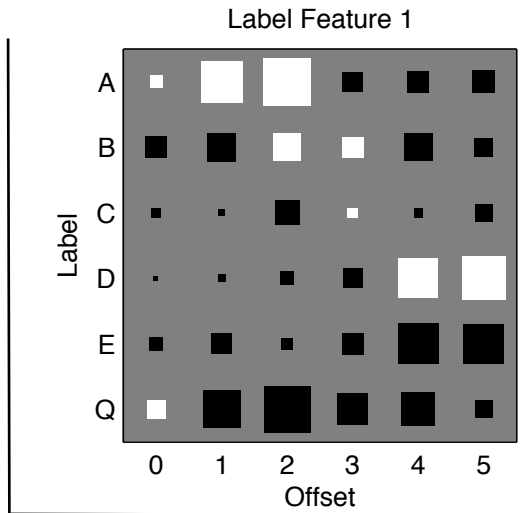


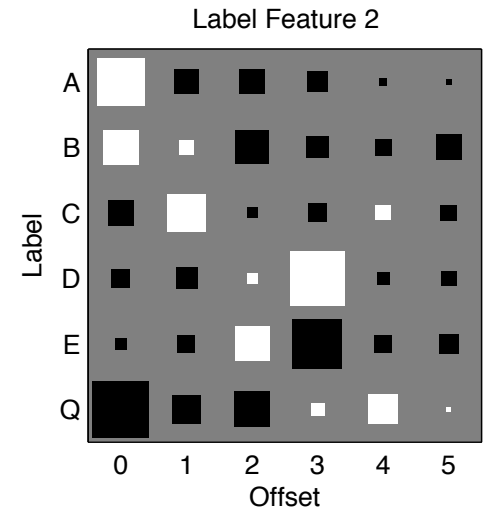
Toy Problem 1

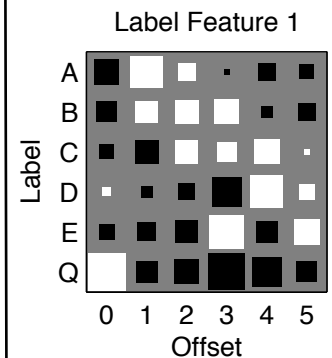
	LR/CRF	IOHMM	HRF	RBM(1)	RBM(2)	RBM(3)
D	0.0	100	79.5	99.2	99.6	97.2
E	66.5	100	80.8	99.2	99.6	96.7
Avg. F1	77.8	100	93.4	99.7	99.9	98.8
Avg. Acc.	97.5	100	99.0	99.9	99.9	99.8
Inst. Acc.	26.0	100	52.0	98.0	99.0	92.0

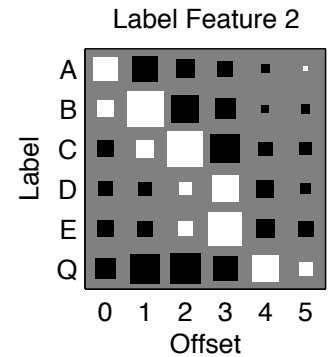
Toy Problem 1

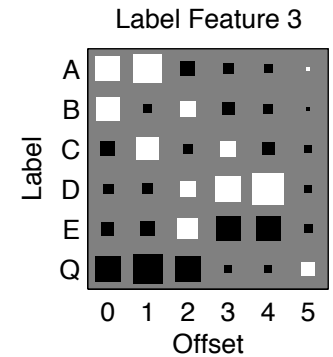


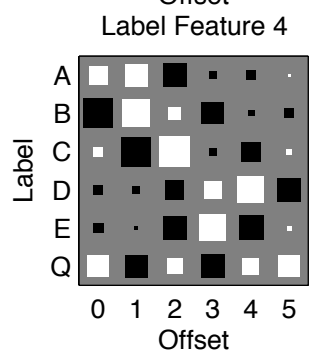


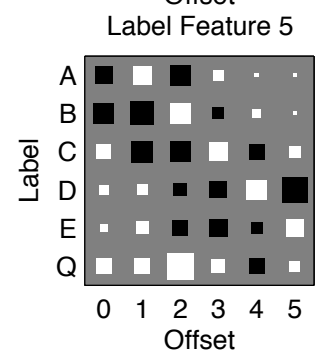


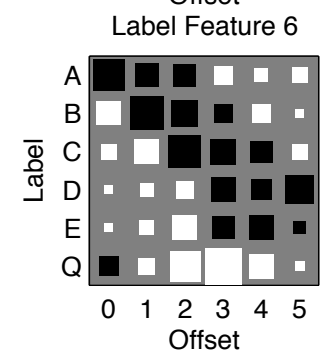












Toy Problem 2

Three labels: A,B,O

Label structures: AAAAAA, BBBBBB, BBB, AA, BA. Separated by O

Each label has a distribution over observations, but the interior of the larger structures is a mixture distribution.

100 training instances, 1000 test instances.

RBM(1) has 9 pair-wise label features; RBM(2) has 9 pair-wise features and 3 features that look at groups of 8 label variables.

Five-fold cross-validation used to choose number of states in the IOHMM and the HRF (8).

Toy Problem 2

	LR	CRF	IOHMM	HRF	RBM(1)	RBM(2)
A	73.9	82.2	71.7	71.28	80.37	83.0
B	65.9	84.4	73.4	73.9	82.9	85.0
Avg. F1	83.0	88.61	81.5	81.4	87.6	89.2
Avg. Acc.	90.8	93.7	90.0	89.8	93.4	94.2
Inst. Acc.	21.2	52.3	17.9	18.4	52.7	63.9

Toy Problem 3

From [Kakade, Teh, Roweis 2002].

Investigates variable term memory.

The CRF, IOHMM, and HRF can model well, but achieve sub-optimal performance.

RBM-CRF models with label features conditional on the input can improve improve upon the performance of the CRF but are not able to model the data perfectly because they maintain no state.

Cora References

Consists of 500 bibliography entries from research papers; 350 are used for training and 150 are used for testing.

Labels: author, book title, date, editor, institution, journal, location, note, pages, publisher, tech, title, and volume.

Entries tokenized by whitespace. Each token processed by 4191 features:

3*(19 regular expressions + 4 categorical + 1374 vocabulary)

The IOHMM and HRF took too long to train.

Cora References

Name	Groups	Observations
RBM-LR(2)	(10,0:4), (10,0:9), (10,0:19)	Pre-trained LR
RBM-CRF(2)	(10,0:4), (10,0:9), (10,0:19)	Pre-trained CRF
RBM(3)	(15,0:1), (20,0:2), (5,0:9), (5,0:19)	LR

Notation: (number of nodes, offsets)

Cora References

	LR	RBM-LR (2)	CRF*	CRF	RBM-CRF (2)	RBM(3)
Avg. F1	76.9	81.4	91.5	87.7	89.4	68.4
Avg. Acc.	85.1	90.1	95.4	94.5	95.1	76.8
Inst. Acc.	16.0	42.0	77.3	66.0	65.3	29.3

CRF*: CRF trained by [Peng, McCallum 2005].

Outline

1. Probabilistic models for structure learning
2. Experimental results
3. Conclusions and future work

Conclusions

We evaluated several methods for doing sequence labeling including latent state models and parameterized templated models.

IOHMMs and HRFs can be more expressive than CRFs, but it can be difficult to pick a good number of latent states and training take a long time and is subject to local optima.

Template models may improve performance, but it easy to pick suboptimal architectures. They do not maintain state so they may not be appropriate for all tasks.

Future Work

RBM-CRF:

- More experiments on real data are required.
- Feature induction to learn the architecture of the model.
- Extension to a hierarchical model.
- Integration of the RBM-CRF with one or more IOHMM/HRF models in a product model.

IOHMM/HRF:

- Investigation of training issues and methods.

All:

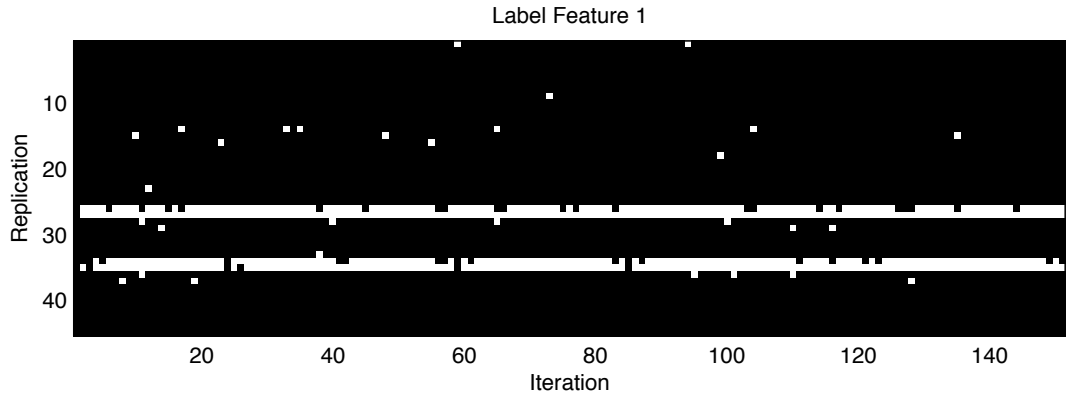
- Put them in a the Bayesian framework.

end.

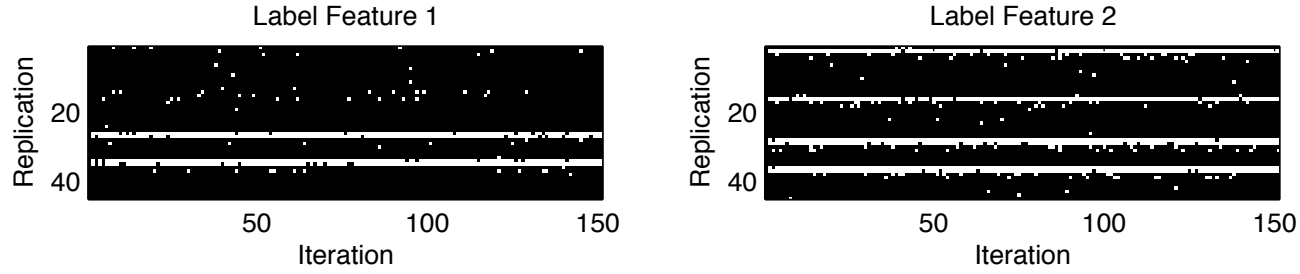
Toy Problem 1

Method	Mean Time (s)	Standard Deviation
LR	29.87	8.61
MEMM	14.76	0.18
IMEMM	246.62	87.74
CRF	387.09	160.36
ICRF	118.07	38.80
IOHMM	2169.60	220.54
HRF	3307.10	2970.70
RBM(1)	424.65	8.03
RBM(2)	945.44	102.97
RBM(6)	27201.00	1582.90

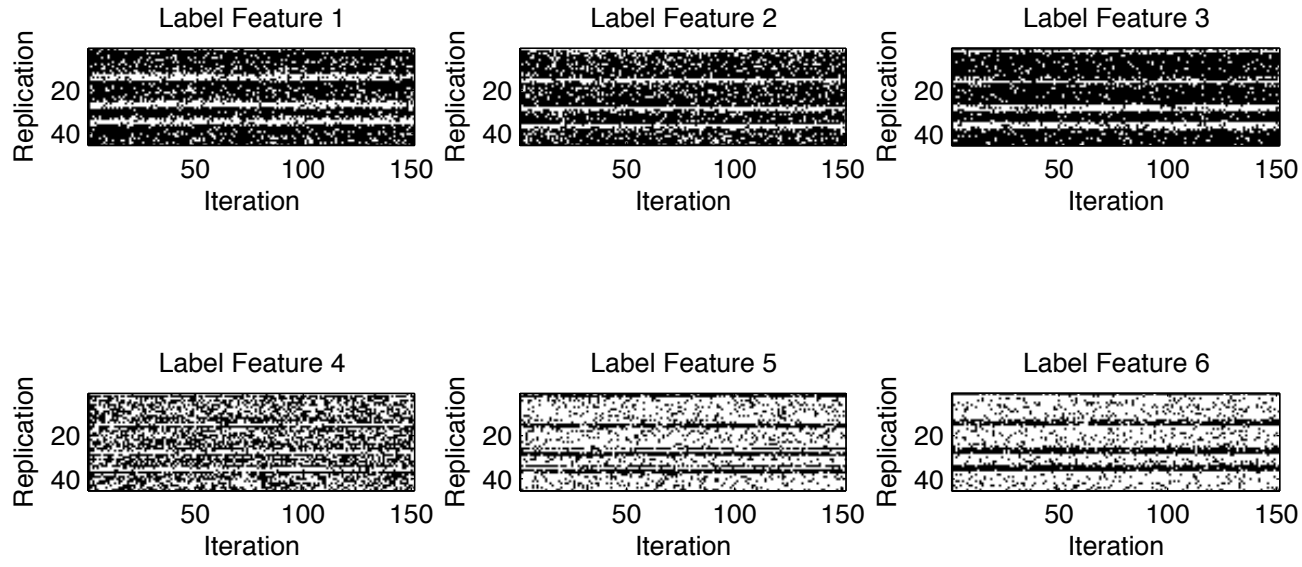
RBM(1)



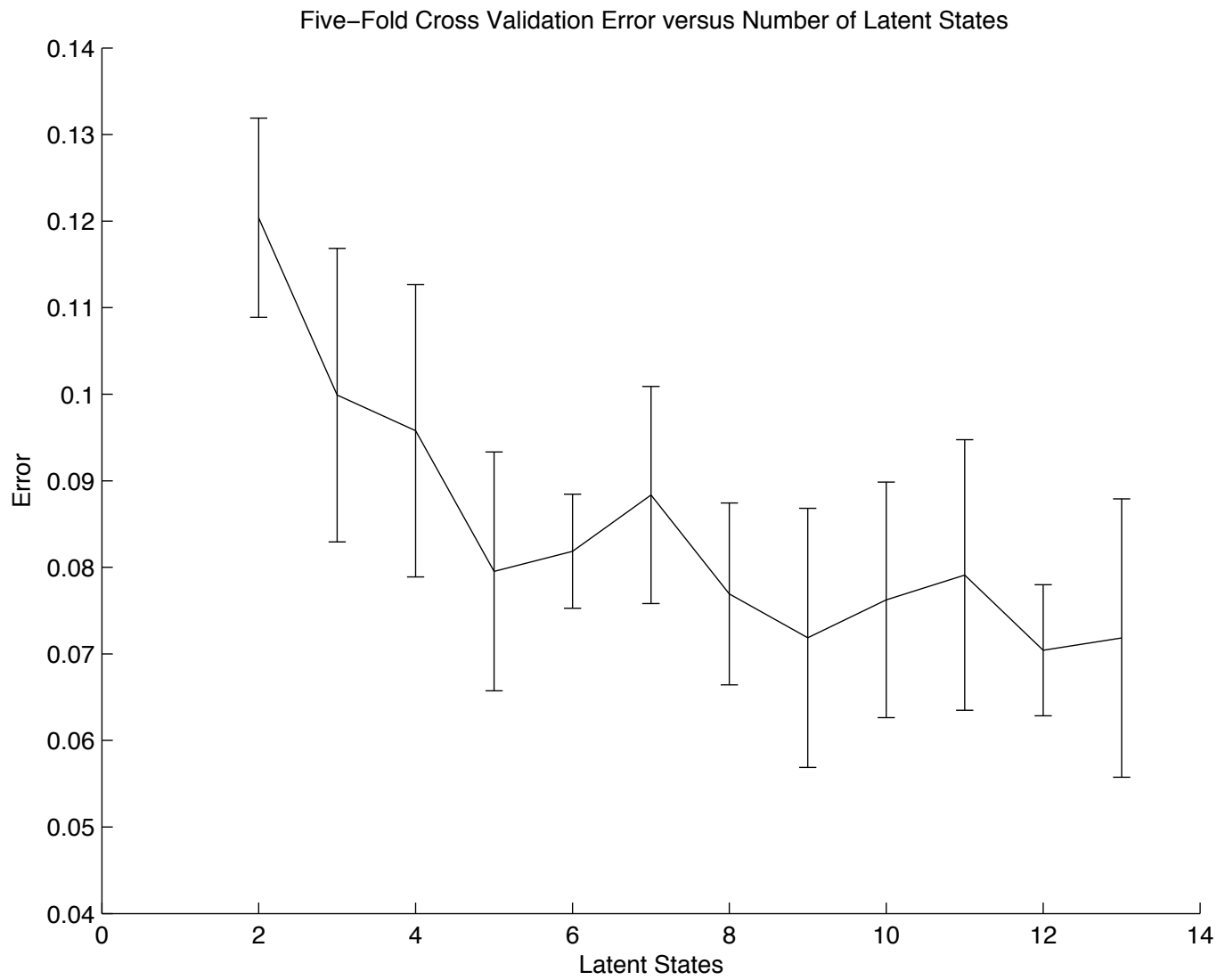
RBM(2)



RBM(3)



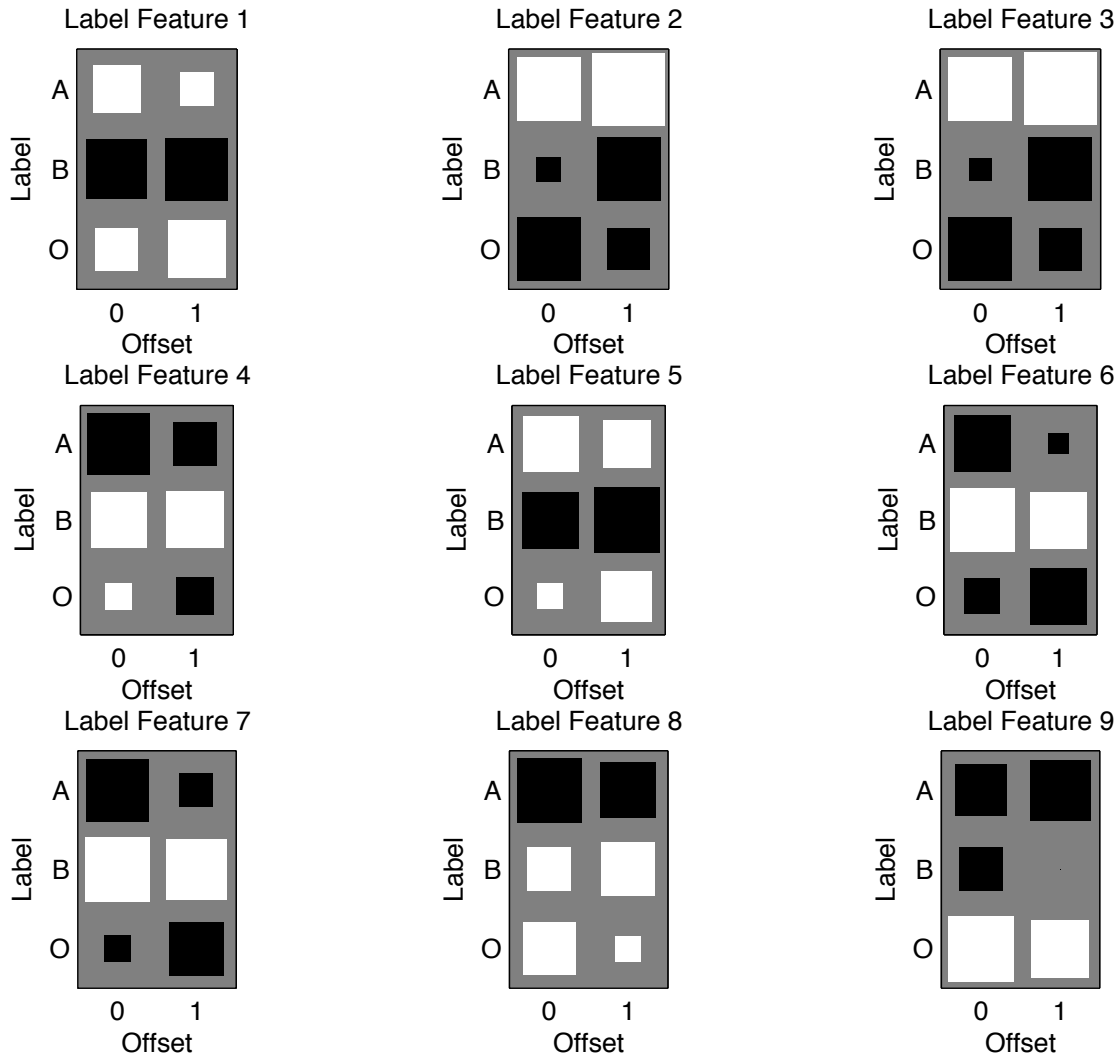
Toy Problem 2



Toy Problem 2

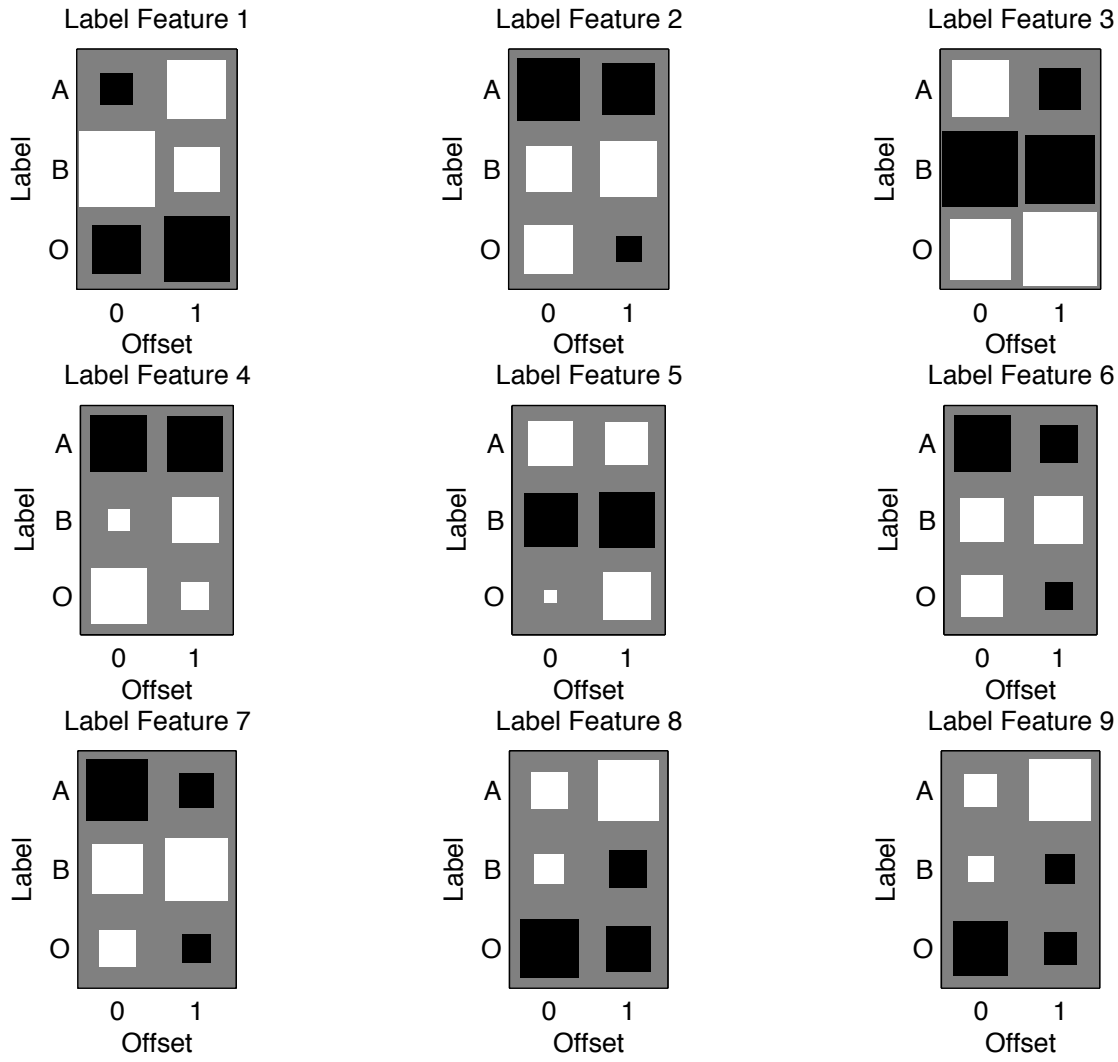
Method	Mean Time (s)	Standard Deviation
LR	158.01	40.45
MEMM	11.08	1.50
IMEMM	430.17	75.19
CRF	229.09	39.58
ICRF	389.81	0.58
IOHMM	7344.30	940.12
HRF	5769.50	1533.00
RBM(1)	13491.00	712.38
RBM(2)	19795.00	3690.50

Toy Problem 2



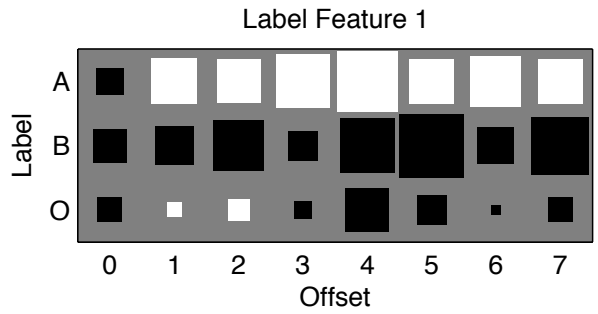
RBM(1)

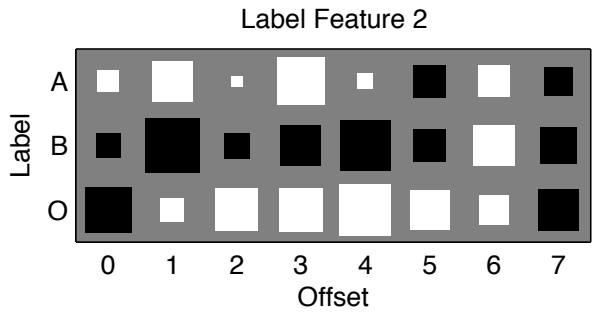
Toy Problem 2

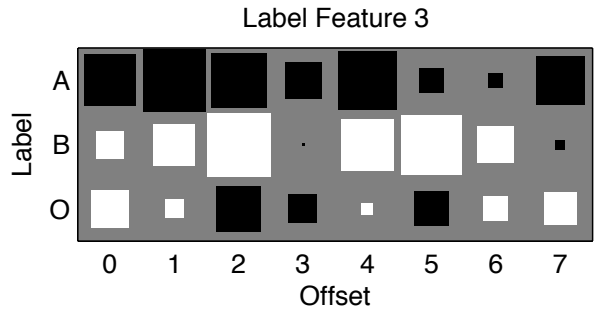


RBM(2)

Toy Problem 2

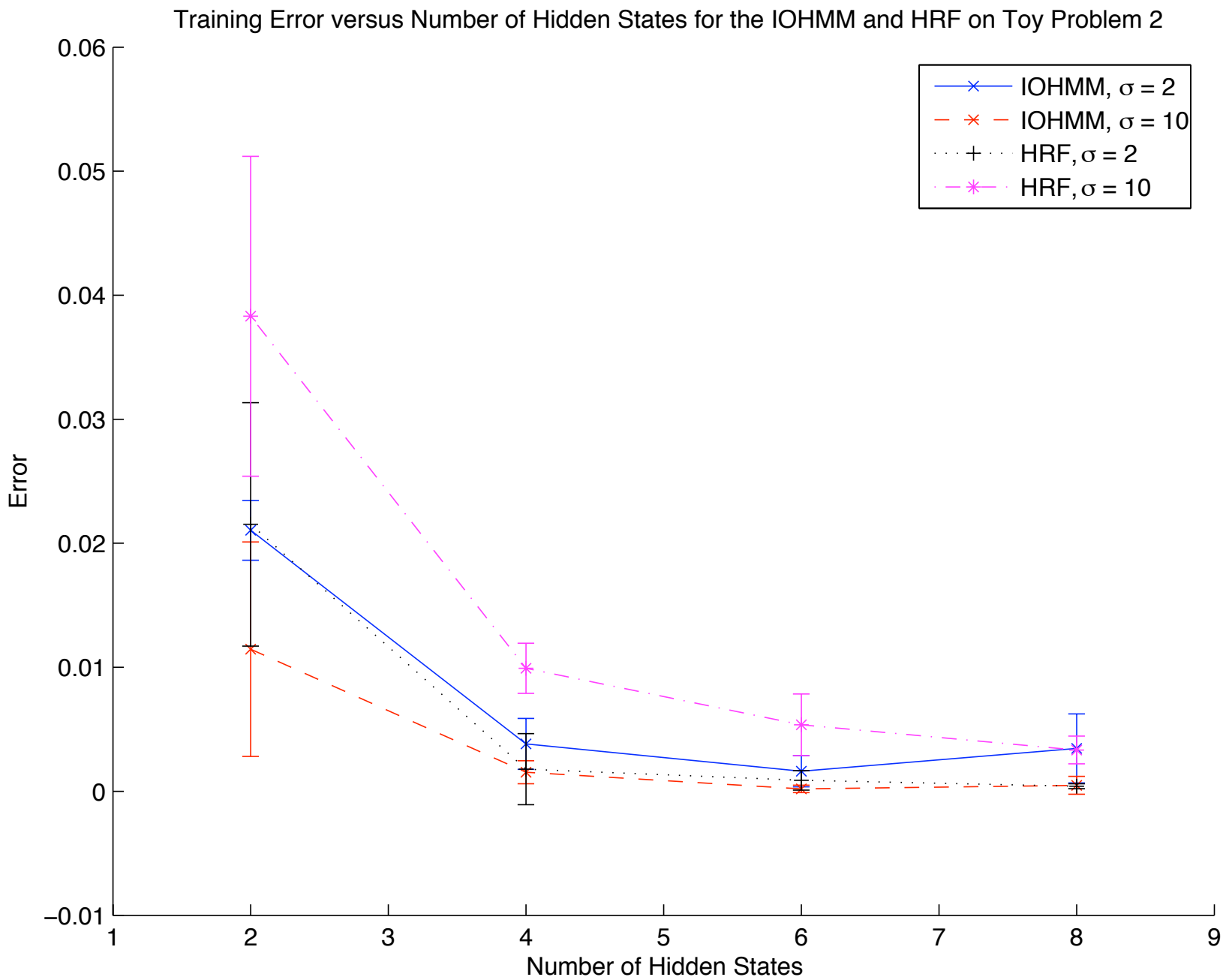




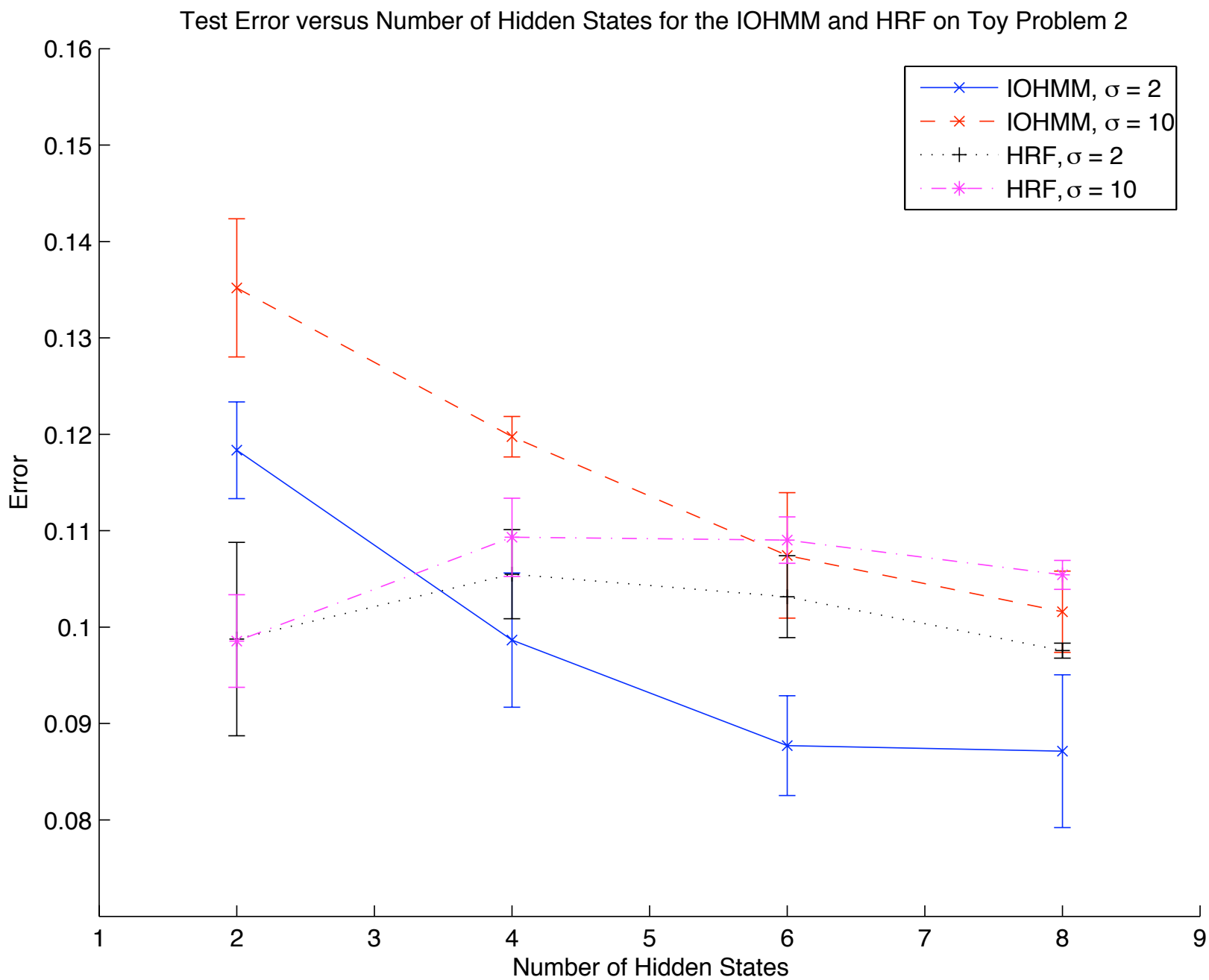


RBM(2)

Toy Problem 2



Toy Problem 2



Toy Problem 3

From [Kakade, Teh, Roweis 2002].

The label of the first I is always 1.

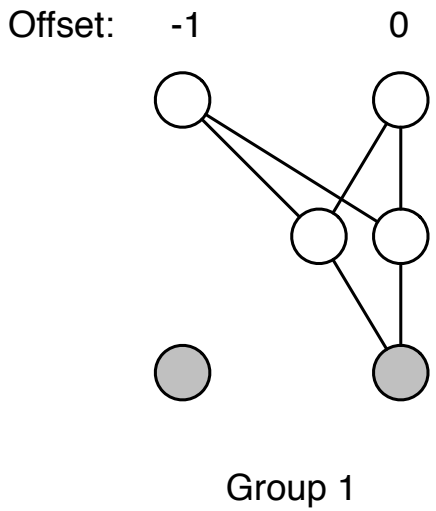
A CRF can only model A/B/R.

No RBM-CRF can model A/B/R/I as no state is maintained.

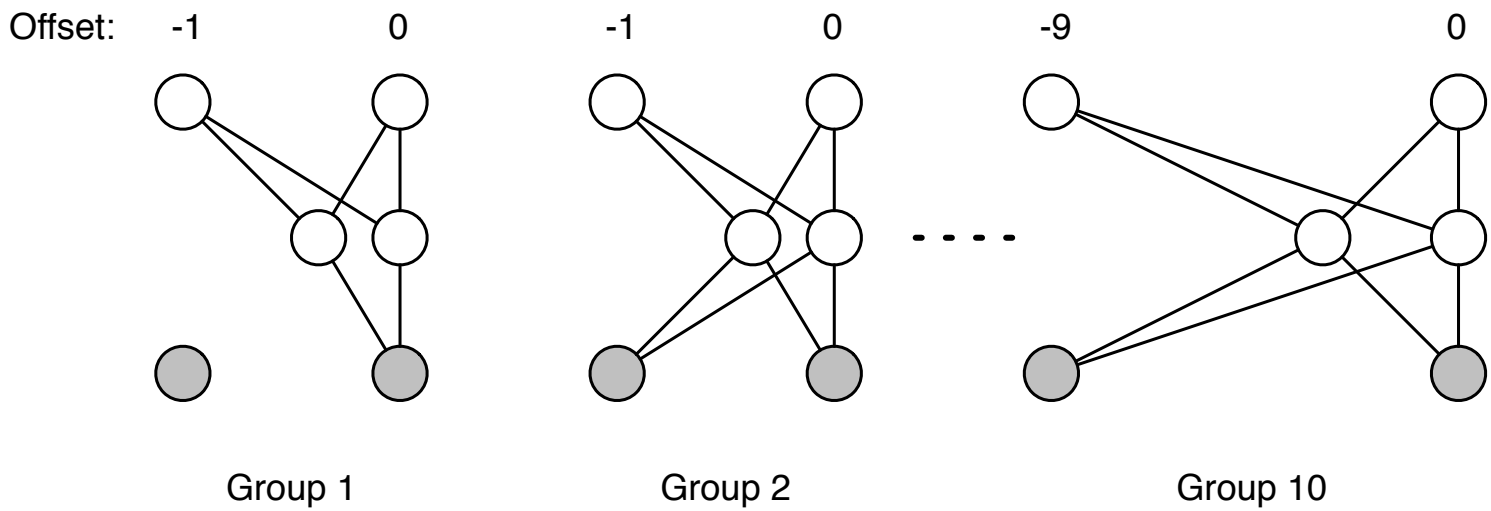
Observation	Label
A	0
B	I
R	Last A/B
I	Inverse of last I

Toy Problem 3

RBM(1)



RBM(2)



Toy Problem 3

	LR	CRF	IOHMM	HRF	RBM(1)	RBM(2)
Avg. F1	75.7	91.3	94.8	96.0	90.4	93.7
Avg. Acc.	76.4	91.3	94.8	96.0	90.4	93.7
Inst. Acc.	0.0	1.0	5.0	7.0	1.0	4.0

Per-observation error rates

	LR	CRF	IOHMM	HRF	RBM(1)	RBM(2)
R	48.2	11.4	1.8	0.1	12.5	4.8
I	44.4	42.5	46.2	41.3	47.2	45.0