From Diversity-based Prediction to Better Ontology & Schema Matching

Dr. Haggai Roitman
Lead Researcher, Master Inventor – Information Retrieval – Cognitive Analytics & Solutions – IBM Research - Haifa

Joint research with Prof. Avigdor Gal & Dr. Tomer Sagi
Technion – Israel Institute of Technology

Schema Matching

- Given two schemata $S=\{a_1,a_2,\ldots,a_n\}$ and $S'=\{b_1,b_2,\ldots,b_m\}$, identify corresponding $\sigma_{i,j} = (a_i,b_j)$ attribute pairs
- Schema Matching is usually a two-stepped process
  - First line matching: determines the similarity $M_{i,j}$ between any pair (a_i,b_j)
  - Second line matching: selects pairs to be included in a match σ

$S_1 \longrightarrow$ / $\downarrow S_2$	cardNum	city	arrival Day	checkIn Time
clientNum	0.84	0.32	0.32	0.30
city	0.29	1.00	0.33	0.30
checkInDate	0.34	0.33	0.35	0.64

$S_1 \longrightarrow$ / $\downarrow S_2$	cardNum	city	arrival Day	checkIn Time
clientNum	1	0	0	0
city	0	1	0	0
checkInDate	0	0	0	1

σ = { (clientNum, cardNum), (city, city), (checkInDate, checkIn Time) }

Schema Matching Performance Prediction

- Prediction Task: Given a pair (M,σ) of 1LM similarity matrix and a 2LM match determine how good the attribute correspondences $\sigma_{i,j}$ are?
- A good match is one with both high Precision and high Recall
- Prediction is made in two main levels (Sagi and Gal)
  - Matrix-level prediction
    - Given (M,σ), a good predictor should provide a prediction that correlates as much as possible with the actual match quality.
    - E.g.,: MAX/STDEV predictors have high correlation to Recall, while AVG/Dominants predictors have high correlation to Precision
  - Entry-level prediction
    - Given entry $\sigma_{i,j}$, a good predictor should assign higher confidence to $\sigma_{i,j}$ whenever this is a true match (and low otherwise).
  - This work propose a new diversity-based schema matching predictor

Match Diversity: Motivation

- Using MWBM as 2LM:

$$\mathcal{Q}_{\text{MWBM}}(\sigma, M) = \sum_{(i,j)\in\sigma} M_{i,j} \text{ =1.9}$$

$$M = \begin{pmatrix} 0.9 & 0.1 & 0.9 \\ 0.1 & 0.1 & 0.1 \\ 0.9 & 0.1 & 0.9 \end{pmatrix}$$

- Yet, pair (2,2) may be a risky selection!
  – It has a relatively low confidence
  – Its "competitor" pairs also have low confidence
- Pair (2,2) should be considered as a false-positive
  – Higher chance for improving Precision than hurting Recall
- Hypothesis: a pair whose confidence deviates more from the confidence of its competitors is a better pair for match selection

Match Diversity: Motivation

- Using MWBM as 2LM:

$$\mathcal{Q}_{\mathsf{MWBM}}(\sigma, M) = \sum_{(i,j)\in\sigma} M_{i,j} = 1.9$$

$$M = \begin{pmatrix} 0.9 & 0.9 & 0.9 \\ 0.9 & 0.1 & 0.9 \\ 0.9 & 0.9 & 0.9 \end{pmatrix}$$

- Yet, pair (2,2) may be a risky selection!
  - It has a relatively low confidence
  - Its "competitor" pairs also have low confidence
- Pair (2,2) should be considered as a false-positive
  - Higher chance for improving Precision than hurting Recall
- Hypothesis: a pair whose confidence deviates more from the confidence of its competitors is a better pair for match selection

Match Competitor Deviation (MCD)

- Single entry deviation:

$$\Delta_{i,j} = (M_{i,j} - \mu_{i,j})^2$$

$$\mu_{i,j} = \frac{1}{n+m-1}\left(\sum_{l=1}^{n} M_{l,j} + \sum_{l=1}^{m} M_{i,l} - M_{i,j}\right)$$

- Match deviation:

$$\mathcal{Q}_{\text{MCD}}(\sigma, M) = \sqrt{\frac{1}{|\sigma|}\sum_{(i,j)\in\sigma} \Delta_{i,j}}$$

- An optimal MCD match is suggested

- Main idea: find a match with both high confidence (MWBM) and high selection diversity (MCD)

MWBM vs. MCD Optimality Tradeoff

- Bad news: the optimality of MWBM may violate the optimality of MCD (and via versa)
- Moreover, we show that:
  - For any possible match σ: $\mathcal{Q}_{\text{MWBM}}(\sigma, M) \geq \mathcal{Q}_{\text{MCD}}(\sigma, M)$
  - If σ' is MWBM optimal match and σ'' is MCD optimal match, then:

$$\begin{aligned} \mathcal{Q}_{\text{MWBM}}(\sigma', M) &\geq \mathcal{Q}_{\text{MWBM}}(\sigma'', M) \quad \text{(MWBM optimality)} \\ &\geq \mathcal{Q}_{\text{MCD}}(\sigma'', M) \quad \text{(Proposition 1)} \\ &\geq \mathcal{Q}_{\text{MCD}}(\sigma', M) \quad \text{(MCD optimality)} \end{aligned}$$

  - MWBM optimality ratio: $\alpha = \frac{\mathcal{Q}_{\text{MWBM}}(\sigma'', M)}{\mathcal{Q}_{\text{MWBM}}(\sigma', M)}$
    - Worst ratio: $\underline{\alpha} \leq \frac{1}{min(n, m)}$

MCD-based Match Regularization

- Main idea: find a match with both high confidence (MWBM) and high selection diversity (MCD)
  - Essentially it is a bi-objective optimization problem:

$$\max_{\sigma \in \Sigma}\{\mathcal{Q}_{\mathsf{MWBM}}(\sigma, M), \mathcal{Q}_{\mathsf{MCD}}(\sigma, M)\}$$

  - For any given β in [0,1] this is equivalent to maximizing the weighted product mean:

$$\mathcal{Q}(\sigma, M) = \mathcal{Q}_{\mathsf{MCD}}(\sigma, M)^{\beta}\mathcal{Q}_{\mathsf{MWBM}}(\sigma, M)^{1-\beta}$$

  - Therefore, the effect of MCD on the optimization (and as a result, on the decisions made by MWBM) can be controlled.
    - Higher β will result in a more diverse match (with an expected increase in Precision)
  - Unfortunately, the maximization problem is NP-Hard

Match Quality Optimization as a Rare-Event Estimation Problem

- Original deterministic optimization problem:

$$\gamma^* = \mathcal{Q}(\sigma^*, M) = \max_{\sigma \in \Sigma} \mathcal{Q}(\sigma, M)$$

- Associated stochastic problem:

$$l(\gamma) = \mathbb{P}_v(\mathcal{Q}(\mathbf{\Sigma}, M) \geq \gamma) = \mathbb{E}_v(\delta_{[\mathcal{Q}(\mathbf{\Sigma}, M) \geq \gamma]})$$

- Yet, since the problem is NP-Hard, the estimation given γ^* becomes a rare event estimation problem

- Solution: Cross Entropy (CE) method

Cross Entropy Matcher (CEM)

Algorithm 2 Cross Entropy Matcher

input: similarity matrix M, N, ρ, λ
initialize:
for $i = 1, \ldots, m; j = 1, \ldots, n$ **do**
  $v^0_{i,j} = \frac{1}{2}$
end for
$t = 1$
loop
  Randomly draw N matches $\sigma \in \Sigma$ using v^{t-1}
  $\Sigma_l = \mathtt{sort}_{l=1,\ldots,N}(\mathcal{Q}(\sigma_l, M))$
  $\gamma_t = \mathtt{quantile}_{1-\rho}(\overrightarrow{\Sigma_l})$
  for $i = 1, \ldots, n; j = 1, \ldots, m$ **do**
    $v^t_{i,j} := \frac{\sum_{l=1}^{N} \delta[\mathcal{Q}(\sigma_l, M) \geq \gamma_t]\delta[(i,j) \in \sigma_l]}{\sum_{l=1}^{N} \delta[\mathcal{Q}(\sigma_l, M) \geq \gamma_t]}$
    $v^t_{i,j} := \lambda v^{t-1}_{i,j} + (1-\lambda) v^t_{i,j}$
  end for
  if γ_t converged **then**
    stop and return random match σ^* sampled from $f(v^t)$
  else
    $t := t + 1$
  end if
end loop

$$l(\gamma) = \mathbb{P}_{v_\gamma}(\mathcal{Q}(\boldsymbol{\Sigma}, M) \geq \gamma) \geq \rho$$

Using importance sampling, on each iteration, we learn the next reference parameter that is based on an estimation of a less-rare event which advances our target towards the optimal match

Cross Entropy Matcher (CEM)

Algorithm 2 Cross Entropy Matcher

input: similarity matrix M, N, ρ, λ
initialize:
for $i = 1, \ldots, m; j = 1, \ldots, n$ **do**
  $v^0_{i,j} = \frac{1}{2}$
end for
$t = 1$
loop
  Randomly draw N matches $\sigma \in \Sigma$ using v^{t-1}
  $\overrightarrow{\Sigma_l} = \texttt{sort}_{l=1,\ldots,N}\left(\mathcal{Q}\left(\sigma_l, M\right)\right)$
  $\gamma_t = \texttt{quantile}_{1-\rho}(\overrightarrow{\Sigma_l})$
  for $i = 1, \ldots, n; j = 1, \ldots, m$ **do**
    $v^t_{i,j} := \frac{\sum_{l=1}^{N} \delta[\mathcal{Q}(\sigma_l, M) \geq \gamma_t] \delta[(i,j) \in \sigma_l]}{\sum_{l=1}^{N} \delta[\mathcal{Q}(\sigma_l, M) \geq \gamma_t]}$
    $v^t_{i,j} := \lambda v^{t-1}_{i,j} + (1 - \lambda)$
  end
  if γ
    st
  else
    t
  end
end lo

Reinforce those pairs that belong to "elite" samples that guarantee at least some minimum required level of match quality

$$v^t_{i,j} := \frac{\sum_{l=1}^{N} \delta[\mathcal{Q}(\sigma_l, M) \geq \gamma_t] \delta[(i,j) \in \sigma_l]}{\sum_{l=1}^{N} \delta[\mathcal{Q}(\sigma_l, M) \geq \gamma_t]}$$

Datasets & Setup

- **Datasets**:

Dataset	#Schemas	#Attr	#Pairs
Web-forms	147	10-30	247
Thalia	44	6-17	18
OAEI	101	80-100	100
Purchase Order	10	50-400	44
University Applications	16	50-150	182

- **1LMs**:

Matcher	System	Type
Term	Ontobuilder [24]	Syntactic
Token Path	AMC [25]	Syntactic
WordNet [29, 16, 28]	ORE	Semantic

- **2LMs**:
  MWBM, Stable Marriage (SM), Dominants (Harmony), Threshsold(v), Max-Delta(δ)

- **CEM**
  - N=10,000, ρ=0.01, λ=0.3 (default)
  - β={0.1,0.2,….,0.9}

https://bitbucket.org/tomers77/ontobuilder-research-environment/wiki/Home

MCD as Matrix-level Predictor

- Prediction over 960 matrices generated by running all 1LMs on 90 different schema pairs sampled from the three datasets.
- 2LMs: Max-Delta(0.1), Threshold(0.5), MWBM and SM.
- Quality of prediction measured by Pearson's r correlation to actual match quality measures
- MCD has the best correlation to Precision
- The MWBM/MCD bi-objective is expected to yield good quality results (while MAX is highly correlated with Recall, MCD is highly correlated with Precision)

Predictor	P Correlation	R Correlation
BMM	.379**	.206**
LMM	.246**	.338**
Max	.180**	.506**
STDEV	.124**	**.630****
Avg	.565**	.077**
Dominants	.429**	.039
LC	.425**	.048
MCD	**.568****	-.002

$$\text{Max}(M) = \frac{1}{n}\sum_{i=1}^{n} \max_i$$

MCD as Entry-level Predictor

- Based on sample obtained from two randomly selected schema pairs form the Web-forms dataset, matched using all 1LMs.
- Overall 5869 entries were obtained.
- Quality of prediction is measured by Goodman-Kruskal Gamma correlation
- MCD exhibits significantly better correlation
- Using MCD would allow to reduce the number of false-positively matched pairs

	MCD		CRV		CNV		Val	
	Γ	sig.	Γ	sig.	Γ	sig.	Γ	sig.
Term	**0.98**	0.018	0.91	0	0.95	0	0.96	0
Token Path	**0.93**	0.002	0.67	0	0.67	0	0.34	0.042
WordNet	**0.93**	0	0.51	0.01	0.59	0	0.67	0

CEM Match Quality

- CEM compared with all other 2LMs (Threshold and Max-Delta parameters were further tuned so as to maximize F1)

	Threshold			Max-Delta			Dominants			SM			MWBM			CEM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Token Path	.02	.03	.02	.20	**.67**	.30	**.48**	**.45**	.45	.27	.62	.36	.32	.58	.41	.29	.60	.38
Term	.51	.43	.41	.27	**.78**	.38	.09	.67	.15	.28	.64	.37	.41	.63	.48	**.53****	.60	**.55****
WordNet	.36	.52	.38	.15	**.67**	.24	.20	.62	.29	.20	.45	.27	.26	.46	.32	**.40****	.47	**.42****

(a) Web-forms

	Threshold			Max-Delta			Dominants			SM			MWBM			CEM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Token Path	.00	.00	.00	.25	.53	.33	**.46**	.46	**.45**	.31	**.56**	.40	.33	.54	.41	.42	.54	**.45**
Term	.53	.48	.48	.25	.55	.33	.44	.53	.47	.32	**.58**	.40	.30	.52	.37	**.59****	.46	**.50****
WordNet	.57	.51	.51	.34	**.72**	.45	.50	.63	.53	.39	.71	.50	.43	.66	.51	**.67****	.52	**.56****

(b) Thalia

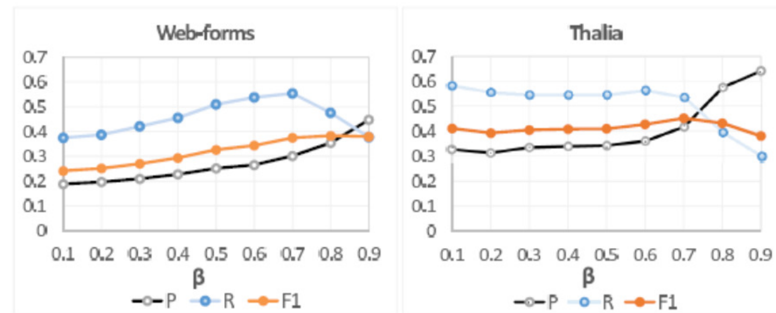
	Threshold			Max-Delta			Dominants			SM			MWBM			CEM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Token Path	.43	.28	.22	.10	**.69**	.17	**.61**	.52	**.55**	.39	.47	.43	.50	.50	.50	.29	.27	.27
Term	.10	.61	.17	.07	**.66**	.13	.31	.61	.38	.37	.45	.40	.46	.45	.45	**.48**	.44	**.46**
WordNet	.13	.27	.16	.15	**.54**	.23	.22	.46	.29	.28	.34	.31	.39	.35	.36	**.53****	.41	**.45****

(c) OAEI

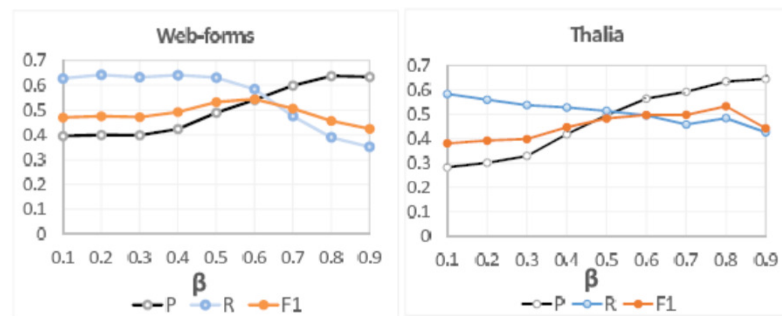
- Up to 25% improvement in F1 (compared to second-best 2LM)
- Specifically, up to 35% and 55% improvement in F1 and Precision compared to MWBM

MCD and the Precision vs. Recall Tradeoff

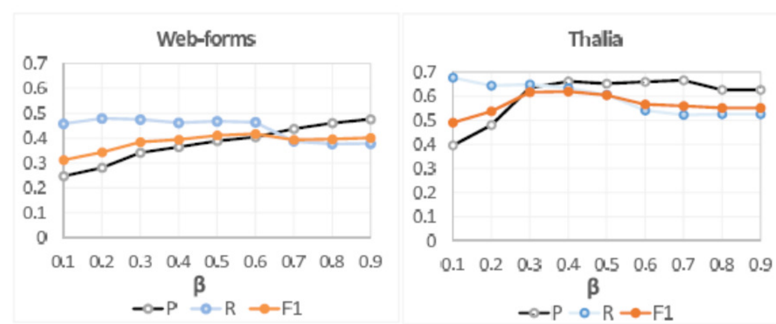
- For all 1LMs, higher β gives more emphasis to the MCD objective yielding increased Precision at the expense of Recall.
- Trend is most notable for the Term 1LM (with $R^2 = 0.93$ and $R^2 = 0.97$ for the Web-forms and Thalia datasets, respectively) compared to the two other 1LM (with an average of $R^2 = 0.92$ and $R^2 = 0.60$).



(a) Token Path



(b) Term



(c) WordNet

Conclusions & Future Work

- We presented a new schema and ontology matching predictor, MCD, discussed its properties, and used it to enhance the performance of an existing state-of-the-art matcher.
- Our empirical evaluation shows MCD to be more predictive than any known matching predictor in the literature so far. We also demonstrated empirically its usefulness for matching.
- Future work:
  - Evaluate the impact of MCD predictor on additional matchers
  - Explore additional match diversification methods
  - Develop new baseline 1LMs whose decisions include diversification considerations

Thanks

haggai@il.ibm.com

Backup slides

MCD Optimization

Algorithm 1 MCD

input: $M(n, m)$
for $(i, j) \in M$ **do**
    $\Delta_{i,j} := (M_{i,j} - \mu_{i,j})^2$
end for
$k := \min(n, m)$
$\sigma^* := \emptyset$
for $p = 1, \ldots, k$ **do**
    $\sigma := \mathsf{MWBM}(\Delta, p)$
    if $\mathcal{Q}_{\mathsf{MCD}}(\sigma, M) > \mathcal{Q}_{\mathsf{MCD}}(\sigma^*, M)$ **then**
        $\sigma^* := \sigma$
    end if
end for
return σ^*

Back

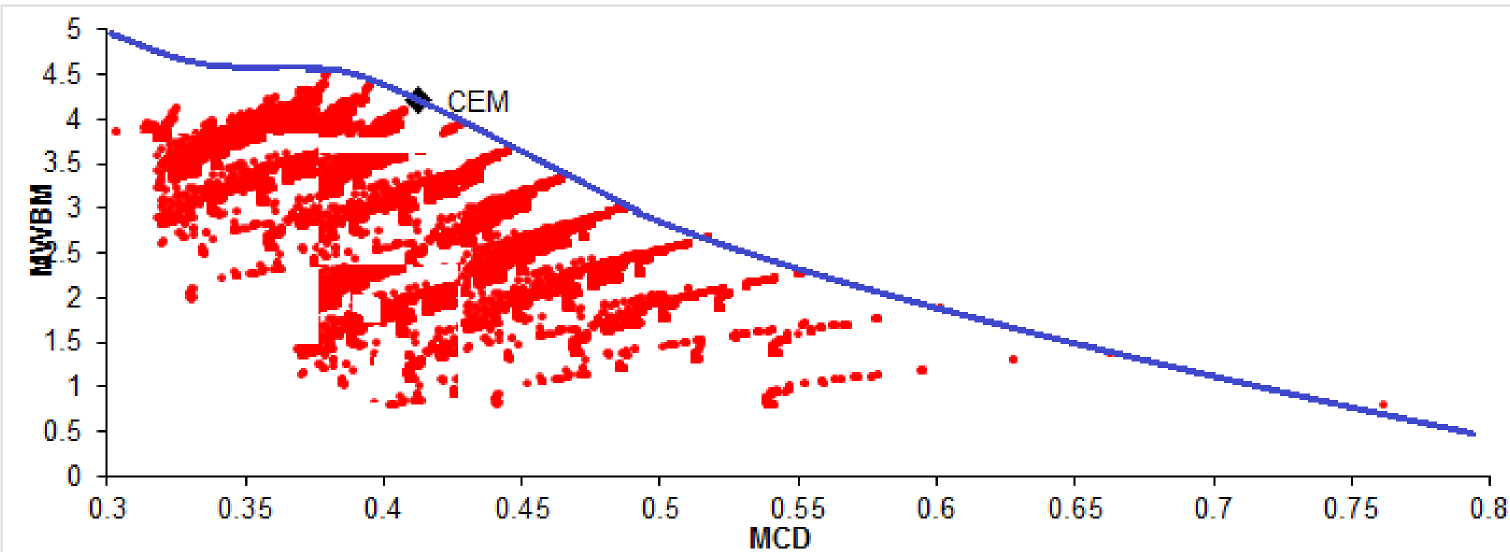
Pareto Optimality

Definition 2 (Pareto Optimal Match). *Given a similarity matrix M, match $\sigma \in \Sigma$ is a Pareto optimal solution to the bi-objective optimization problem (Eq. 5) if for any other match $\sigma' \in \Sigma$ one of the following holds:*

$$\mathcal{Q}_{\text{MWBM}}(\sigma, M) \leq \mathcal{Q}_{\text{MWBM}}(\sigma', M) \Rightarrow \mathcal{Q}_{\text{MCD}}(\sigma, M) > \mathcal{Q}_{\text{MCD}}(\sigma', M),$$

or

$$\mathcal{Q}_{\text{MCD}}(\sigma, M) \leq \mathcal{Q}_{\text{MCD}}(\sigma', M) \Rightarrow \mathcal{Q}_{\text{MWBM}}(\sigma, M) > \mathcal{Q}_{\text{MWBM}}(\sigma', M).$$



Back

Random Match Sampling

- Odds for selecting a single edge:

$$\mathbb{P}_{v_{i,j}}(\delta_{i,j}) = v_{i,j}^{\delta_{i,j}} (1 - v_{i,j})^{1-\delta_{i,j}}$$

- Odds for selecting a sub-set of E:

$$f(E'; v) = \prod_{(i,j) \in E'} v_{i,j}^{\delta_{i,j}} (1 - v_{i,j})^{1-\delta_{i,j}}$$

- Adjustment for correct 1:1 match:

$$Q'(E', M) = \begin{cases} Q(E', M), & E' \in \Sigma \\ -\infty & otherwise \end{cases}$$

Algorithm 3 Random Match Sampling

input: M,v
$E := \{(i,j); i = 1, \ldots, n; j = 1, \ldots, m\}$
$\sigma := \emptyset$
while $E \neq \emptyset$ **do**
  select next edge $(i,j) \in E$ to consider at random
  draw $u \sim U[0,1]$
  if $v_{i,j} \geq u$ **then**
    $\sigma := \sigma \cup \{(i,j)\}$
    $E := E \setminus \{(i,j)\}$
  end if
  for $(i', j') \in S$ **do**
    if $i' = i \vee j' = j$ **then**
      $E := E \setminus \{(i', j')\}$
    end if
  end for
end while
return σ

Back

Reference Parameter Derivation

$$l(\gamma_t) = \mathbb{P}_{v^{t-1}}(\delta_{[\mathcal{Q}(\mathbf{\Sigma},M)\geq\gamma_t]}) = \mathbb{E}_{v^{t-1}}(\delta_{[\mathcal{Q}(\mathbf{\Sigma},M)\geq\gamma_t]})$$

$$\mathbb{E}_{v^{t-1}}(\delta_{[\mathcal{Q}(\mathbf{\Sigma},M)\geq\gamma_t]})\frac{f(\mathbf{\Sigma};v^t)}{f(\mathbf{\Sigma};v^t)} = \int_{\Sigma} \delta_{[\mathcal{Q}(\mathbf{\Sigma},M)\geq\gamma_t]} f(\mathbf{\Sigma};v^{t-1})\frac{f(\mathbf{\Sigma};v^t)}{f(\mathbf{\Sigma};v^t)}d\sigma$$

$$l_{LR}(\gamma_t) = \mathbb{E}_{v^t}(\delta_{[\mathcal{Q}(\mathbf{\Sigma},M)\geq\gamma_t]})\frac{f(\mathbf{\Sigma};v^{t-1})}{f(\mathbf{\Sigma};v^t)}$$

$$\hat{l}_{LR}(\gamma_t) = \frac{1}{N}\sum_{k=1}^{N}\delta_{[\mathcal{Q}(\sigma_k,M)\geq\gamma_t]}\frac{f(\sigma_k;v^{t-1})}{f(\sigma_k;v^t)}$$

$$\sigma_k \sim f(\cdot;v^t); k = 1,\ldots,N.$$

$$f^*(\mathbf{\Sigma}) = \frac{\delta_{[\mathcal{Q}(\mathbf{\Sigma},M)\geq\gamma_t]} f(\mathbf{\Sigma},v^{t-1})}{l(\gamma_t)}$$

$$\mathcal{D}_{KL}(f^*(\mathbf{\Sigma}), f(\mathbf{\Sigma};v^t)) = \mathbb{E}_{f^*}\ln\frac{f^*(\mathbf{\Sigma})}{f(\mathbf{\Sigma};v^t)}$$
$$= \int_{\Sigma} f^*(\mathbf{\Sigma})\ln f^*(\mathbf{\Sigma})d\sigma - \int_{\Sigma} f^*(\mathbf{\Sigma})\ln f(\mathbf{\Sigma};v^t)d\sigma$$

Back

Reference Parameter Derivation

$$\max_{v^t} \int_{\Sigma} \frac{\delta_{[\mathcal{Q}(\boldsymbol{\Sigma},M)\geq\gamma_t]} f(\boldsymbol{\Sigma}, v^{t-1})}{l(\gamma_t)} \ln f(\boldsymbol{\Sigma}, v^t) d\sigma$$

$$\max_{v^t} \mathbb{E}_{v^{t-1}} (\delta_{[\mathcal{Q}(\boldsymbol{\Sigma},M)\geq\gamma_t]}) \ln f(\boldsymbol{\Sigma}, v^t)$$

$$\max_{v^t} \frac{1}{N} \sum_{k=1}^{N} \delta_{[\mathcal{Q}(\sigma_k,M)\geq\gamma_t]} \ln f(\sigma_k, v^t)$$

$$\sigma_k \sim f(\cdot; v^{t-1}); k = 1, \ldots, N$$

$$\frac{\partial}{\partial v_{i,j}^t} \ln f(\cdot, v^t) = \frac{\delta_{i,j}}{v_{i,j}^t} - \frac{1-\delta_{i,j}}{1-v_{i,j}^t} = \frac{1}{v_{i,j}^t(1-v_{i,j}^t)}(\delta_{i,j} - v_{i,j}^t)$$

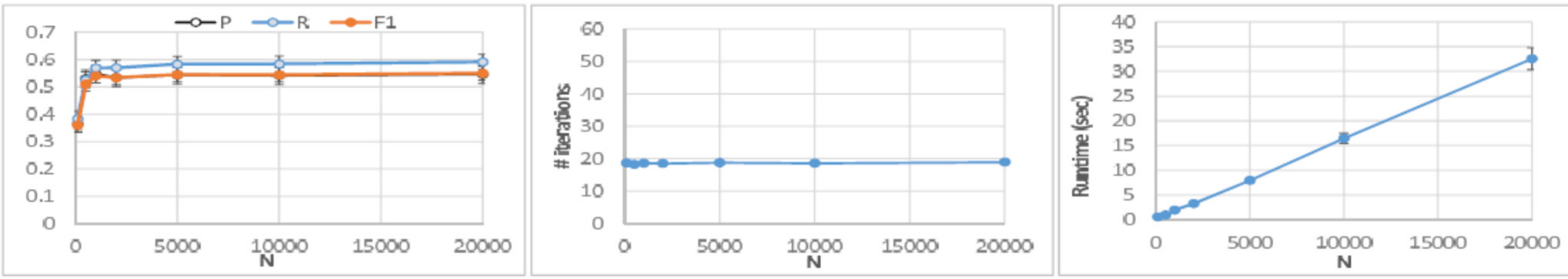
$$\frac{\partial}{\partial v_{i,j}^t} \left(\frac{1}{N} \sum_{k=1}^{N} \delta_{[\mathcal{Q}'(\sigma_k,M)\geq\gamma_t]} \ln f(\sigma_k, v^t) \right) = 0$$

$$\frac{1}{v_{i,j}^t(1-v_{i,j}^t)} \frac{1}{N} \sum_{k=1}^{N} \delta_{[\mathcal{Q}'(\sigma_k,M)\geq\gamma_t]} (\delta_{i,j} - v_{i,j}^t) = 0$$

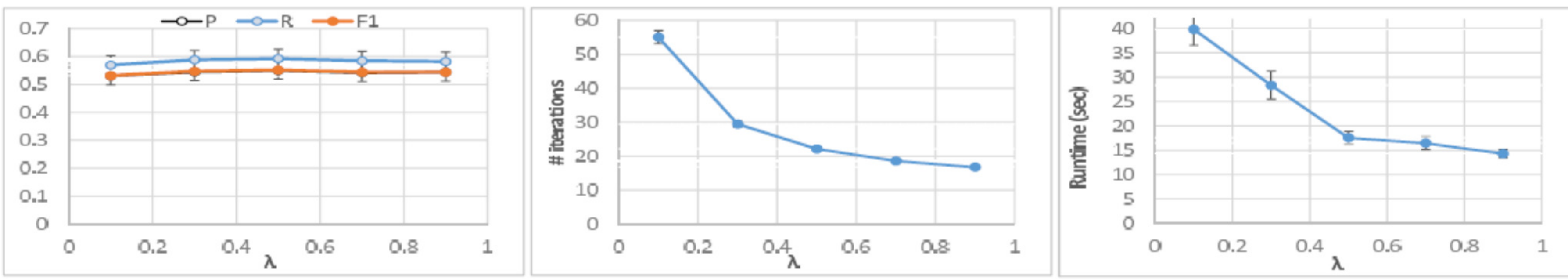
$$v_{i,j}^t = \frac{\sum_{k=1}^{N} \delta_{[\mathcal{Q}'(\sigma_k,M)\geq\gamma_t]} \delta_{i,j}}{\sum_{k=1}^{N} \delta_{[\mathcal{Q}'(\sigma_k,M)\geq\gamma_t]}}$$

Back

CEM Sensitivity Analysis



(a) Sample Size (N)



(b) Model Smoothing (λ)