Empirical means to validate skills models and assess the fit of a student model

Behzad Beheshti
Supervisor : Michel C. Desmarais

Génie Informatique Et Génie Logiciel
École Polytechnique de Montréal

26 février 2016

Problem Specification

- Student skills assessment models

NMF Additive

POKS

NMF Conjunctive

Expected

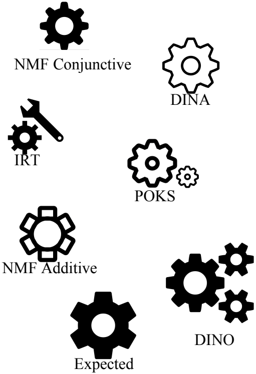
DINO

IRT

DINA

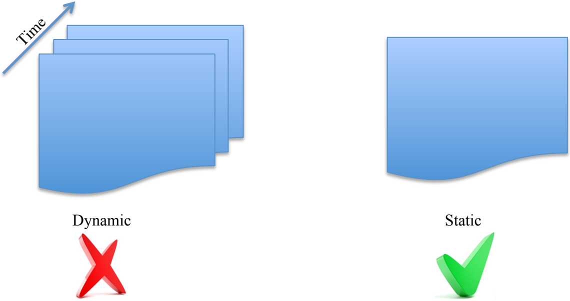
Problem Specification

- Student skills assessment models
- How to decide which are the most representative of the underlying ground truth ?



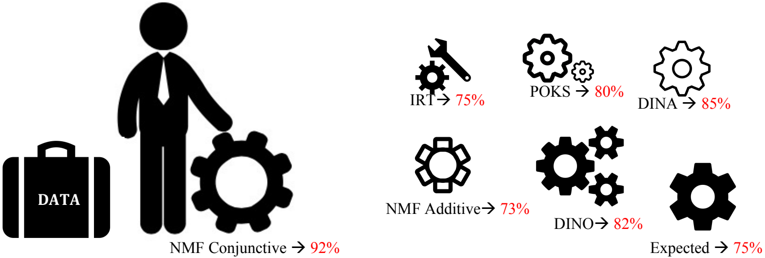
Problem Specification

- Student skills assessment models
- How to decide which are the most representative of the underlying ground truth ?
- Static Vs. Dynamic



Problem Specification

- Student skills assessment models
- How to decide which are the most representative of the underlying ground truth ?
- Static Vs. Dynamic
- Model selection and goodness of fit
- A general answer : best performer



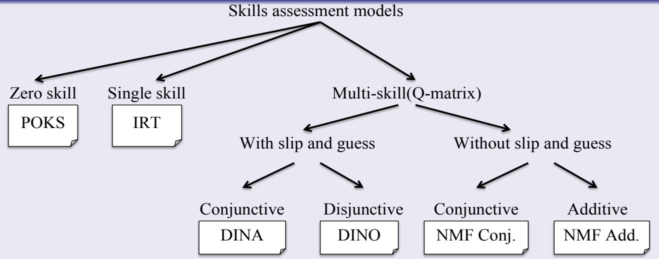
Problem Specification

- Student skills assessment models
- How to decide which are the most representative of the underlying ground truth ?
- Static Vs. Dynamic
- Model selection and goodness of fit
- A general answer : best performer
- Our contribution
    - To make a comprehensive comparison of educational data model performances
    - To propose a new approach to assessing model fit

Problem Specification

- Student skills assessment models
- How to decide which are the most representative of the underlying ground truth ?
- Static Vs. Dynamic
- Model selection and goodness of fit
- A general answer : best performer
- Our contribution
  - To make a comprehensive comparison of educational data model performances
  - To propose a new approach to assessing model fit
- The proposed approach :
  - Assessing the fit of the model to the underlying ground truth using a methodology based on **synthetic data**

Student skills assessment models


Skills assessment models

Zero skill

POKS

Single skill

IRT

Multi-skill(Q-matrix)

With slip and guess

Without slip and guess

Conjunctive

DINA

Disjunctive

DINO

Conjunctive

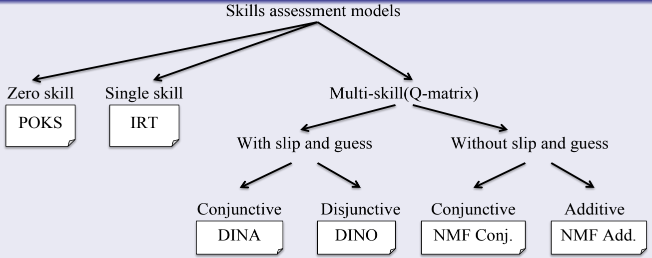
NMF Conj.

Additive

NMF Add.

- Number of Skills

Student skills assessment models



Skills assessment models

- Zero skill
  - POKS
- Single skill
  - IRT
- Multi-skill(Q-matrix)
  - With slip and guess
    - Conjunctive
      - DINA
    - Disjunctive
      - DINO
  - Without slip and guess
    - Conjunctive
      - NMF Conj.
    - Additive
      - NMF Add.

- Number of Skills
- Q-matrix

s_1 : fraction multiplication

s_2 : fraction addition

s_3 : fraction reduction

i_1 $\frac{4}{\frac{12}{3}} + \frac{3}{5} = \frac{8}{5}$

i_2 $\frac{4}{\frac{12}{3}} = \frac{4\times3}{12} = \frac{12}{12} = 1$

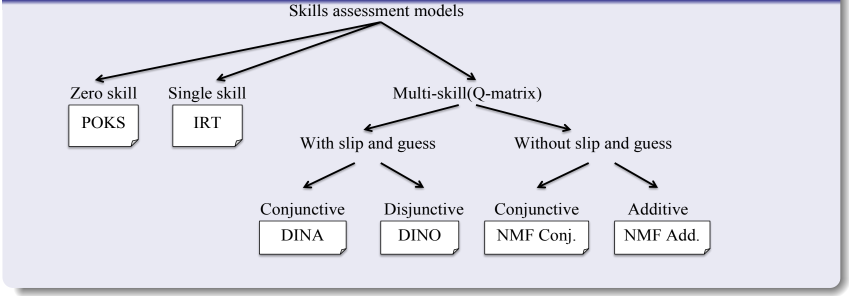
i_3 $1 + \frac{3}{5} = \frac{8}{5}$

i_4 $2\times\frac{1}{2} = 1$

Skills

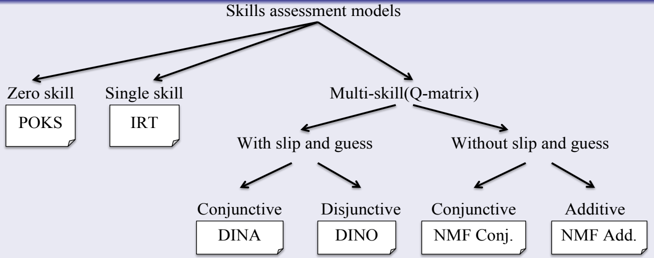
$$\begin{array}{cc|ccc} & & s_1 & s_2 & s_3 \\ \hline \text{Items} & i_1 & 1 & 1 & 1 \\ & i_2 & 1 & 0 & 1 \\ & i_3 & 0 & 1 & 1 \\ & i_4 & 1 & 0 & 1 \end{array}$$

Student skills assessment models



- Number of Skills
- Q-matrix
- Slip and Guess

Student skills assessment models



Skills assessment models

Zero skill
POKS

Single skill
IRT

Multi-skill(Q-matrix)

With slip and guess

Without slip and guess

Conjunctive
DINA

Disjunctive
DINO

Conjunctive
NMF Conj.

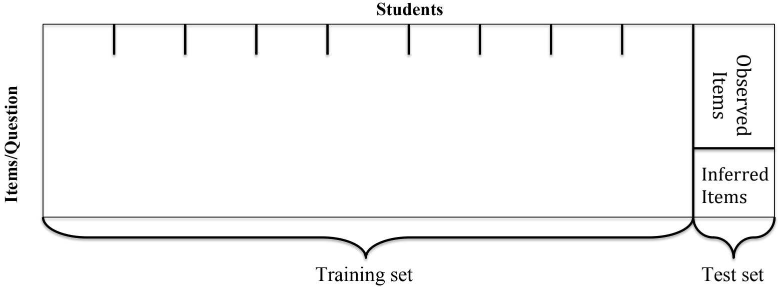
Additive
NMF Add.

- Number of Skills
- **Q-matrices (types)**
- Slip and Guess

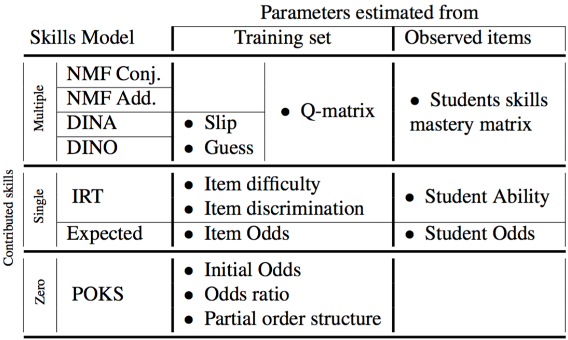
1. Conjunctive
2. Additive
3. Disjunctive

$$\begin{array}{cc} & \text{Skills} \\ \text{Items} & \begin{array}{c} \\ i_1 \\ i_2 \\ i_3 \\ i_4 \end{array} \begin{array}{c} \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \left[\begin{array}{ccc} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{array}\right] \end{array} \end{array}$$

- **Performance of a model** over a data set

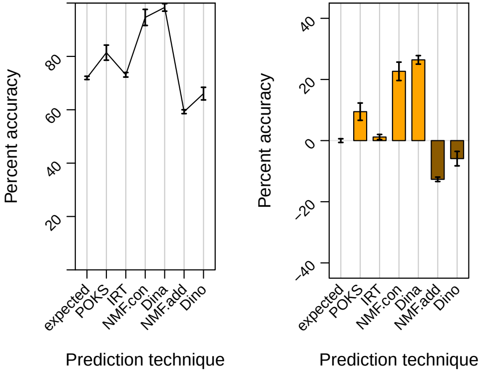


- **Performance of a model** over a data set
- Model parameters



		Skills Model	Parameters estimated from		
			Training set		Observed items
Contributed skills	Multiple	NMF Conj.		• Q-matrix	• Students skills mastery matrix
		NMF Add.			
		DINA	• Slip • Guess		
		DINO			
	Single	IRT	• Item difficulty • Item discrimination		• Student Ability
		Expected	• Item Odds		• Student Odds
	Zero	POKS	• Initial Odds • Odds ratio • Partial order structure		

- **Performance of a model** over a data set
- Model parameters
- Performance vector
- Performance signature



Model	*Performace*
Expected	0.72%
POKS	0.80%
IRT	0.74%
NMF.Conj	0.94%
Dina	0.99%
NMF.Add	0.60%
Dino	0.65%

- **Performance of a model** over a data set
- Model parameters
- Performance vector
- Performance signature
- Performance prototype

The *performance vector* associated with the synthetic data of a model class.

- **Performance of a model** over a data set
- Model parameters
- Performance vector
- Performance signature
- Performance prototype
- Target performance vector

The *performance vector* of the real data set to classify.

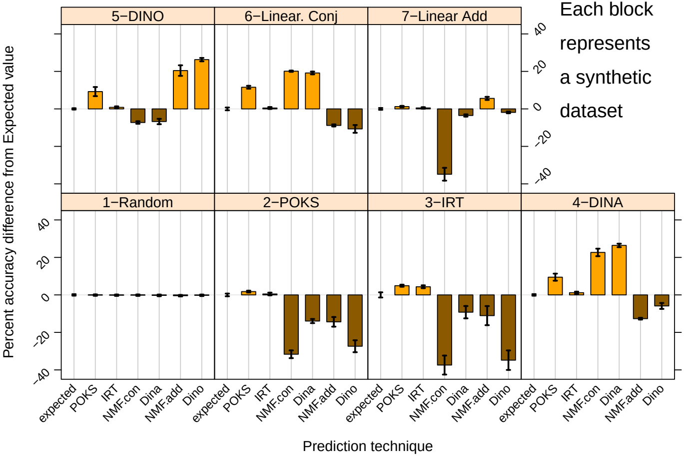
Research questions

1. ✓What is the *performance vector* of student skills assessment models over real and over synthetic data created using the same models ?
    - Experiment 1 : Predictive performance of models over real and synthetic data sets

Datasets

Data set	Number of			Mean Score	Q-matrix
	Skills	Items	Students		
Synthetic					
1.Random	7	30	700	0.75	$\mathbf{Q}_{01}$
2.POKS	7	20	500	0.50	$\mathbf{Q}_{02}$
3.IRT-2PL	5	20	600	0.50	$\mathbf{Q}_{03}$
4.DINA	7	28	500	0.31	$\mathbf{Q}_{5}$
5.DINO	7	28	500	0.69	$\mathbf{Q}_{6}$
Linear (Matrix factorization)					
6. Conj.	8	20	500	0.24	$\mathbf{Q}_{1}$
7. Comp.	8	20	500	0.57	$\mathbf{Q}_{1}$
Real					
8.Fraction	8	20	536	0.53	$\mathbf{Q}_{1}$
9.Vomlel	6	20	149	0.61	$\mathbf{Q}_{4}$
10.ECPE	3	28	2922	0.71	$\mathbf{Q}_{3}$
Fraction subsets and variants of $\mathbf{Q}_1$					
11. 1	5	15	536	0.53	$\mathbf{Q}_{10}$
12. 2/1	3	11	536	0.51	$\mathbf{Q}_{11}$
13. 2/2	5	11	536	0.51	$\mathbf{Q}_{12}$
14. 2/3	3	11	536	0.51	$\mathbf{Q}_{13}$

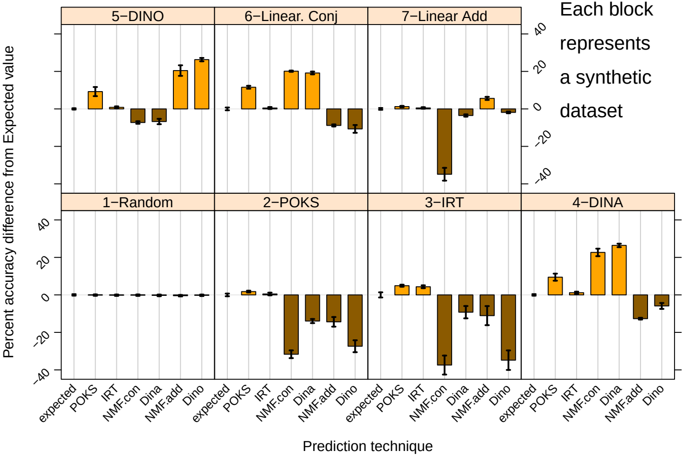
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

The highest performance is for the generative model behind the dataset

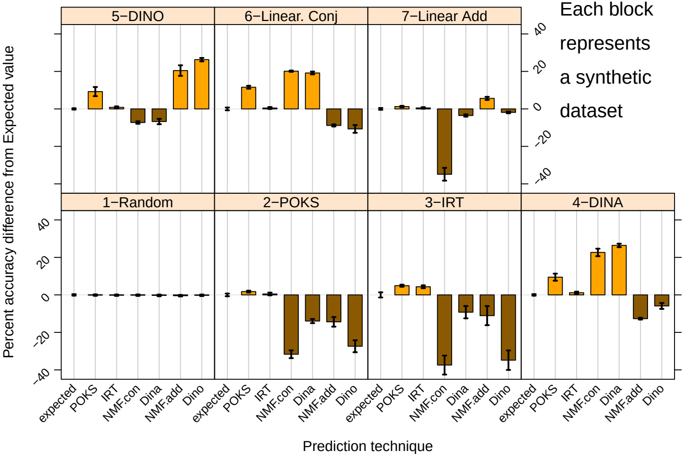
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

Data sets have unique pattern of performance vector across models

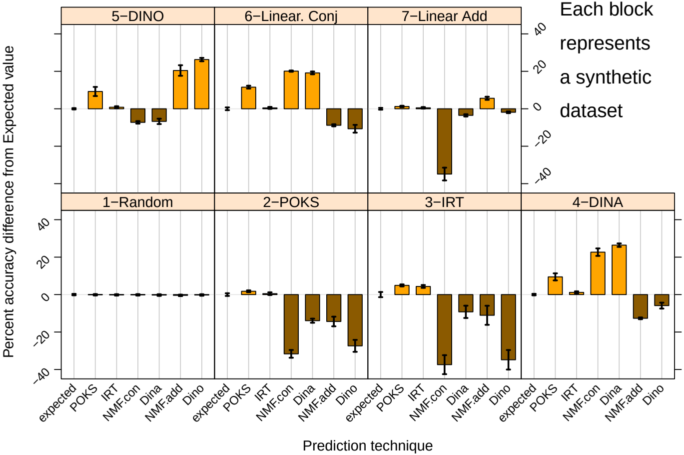
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

The random data set has a flat performance across techniques

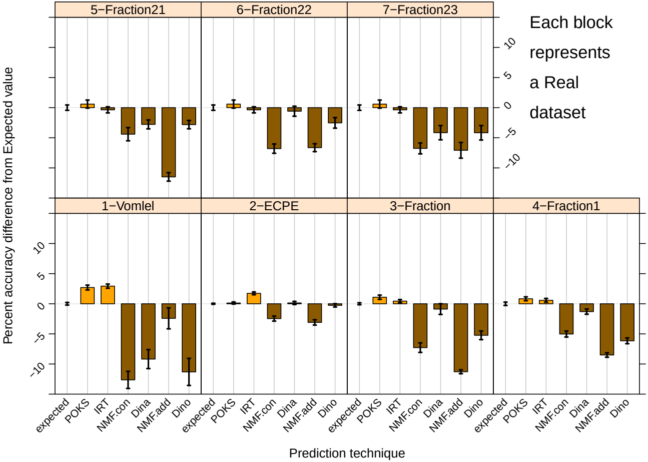
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

The capacity of recognizing a data set's true model relies on this uniqueness characteristic

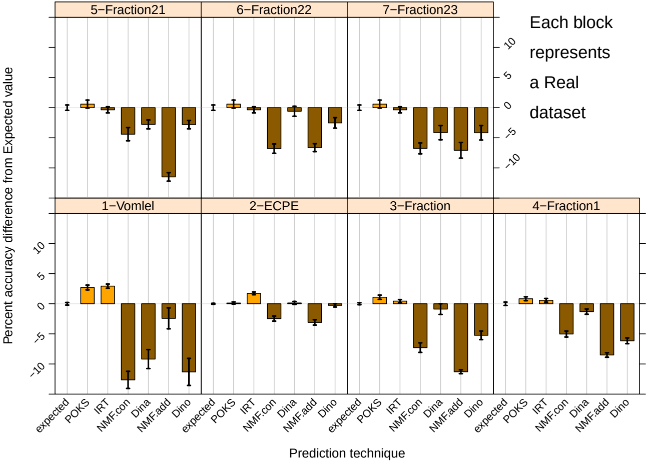
Predictive performance of models over real datasets



Each block represents a Real dataset

In most cases, the best performer is close to the baseline

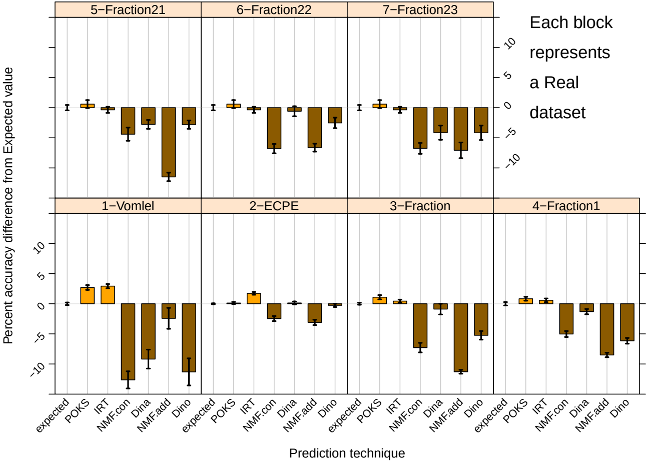
Predictive performance of models over real datasets



Each block represents a Real dataset

The pattern of the Fraction performance data set repeats over its subsets

Predictive performance of models over real datasets



Each block represents a Real dataset

None of the real data sets show the large the amplitude and the differences found in the synthetic data sets models

Vector space of accuracy performances

- Performance vectors of datasets in columns(Data points in the performance space)

Model	**Synthetic data set**						
	Random	POKS	IRT	DINA	DINO	Linear .Conj	Linear .Comp
Expected	**0.75**	0.91	0.90	0.72	0.72	0.78	0.93
POKS	0.75	**0.94**	0.94	0.81	0.81	0.90	0.94
IRT	0.75	0.91	**0.95**	0.73	0.73	0.79	0.89
DINA	0.75	0.77	0.81	**1.00**	0.65	**0.98**	0.89
DINO	0.75	0.63	0.56	0.66	**1.00**	0.68	0.91
NMF.Conj	0.75	0.59	0.53	0.95	0.65	0.97	0.58
NMF.Comp	0.75	0.76	0.79	0.59	0.93	0.70	**0.98**

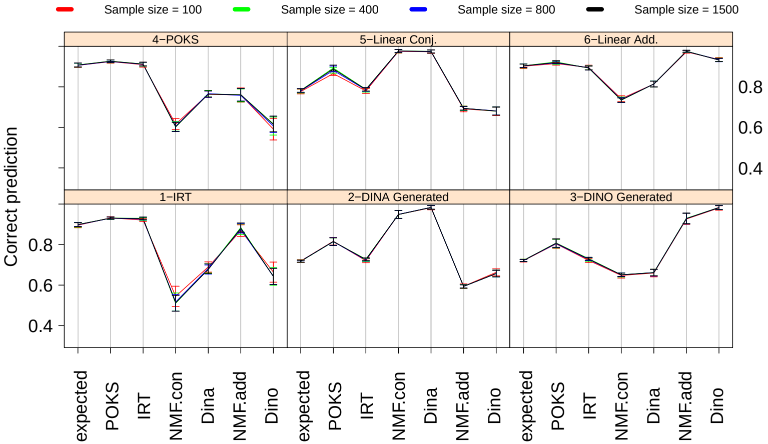
The diagonal generally displays the best performance

Research questions

1. ✓What is the *performance vector* of student skills assessment models over real and over synthetic data created using the same models ?
    - Experiment 1 : Predictive performance of models over real and synthetic data sets
2. **Is the performance vector unique to each synthetic data type (data from the same ground truth model) ?**
    - Experiment 2 : Sensitivity of the Model performance over different data generation parameters

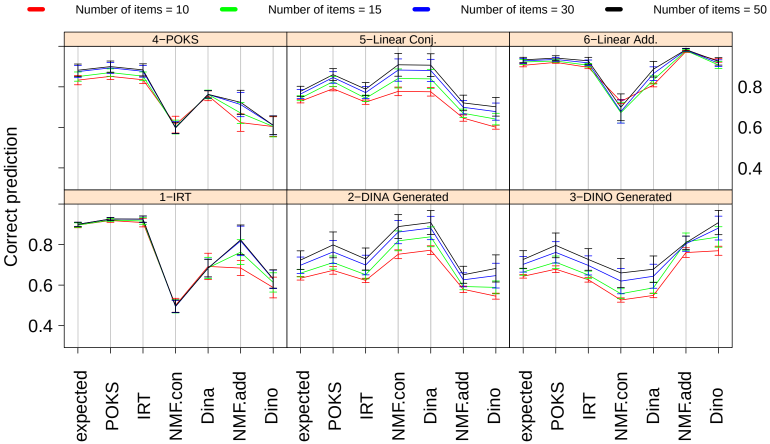
Are they stable in addition to be unique.

Variation of sample size over synthetic data sets



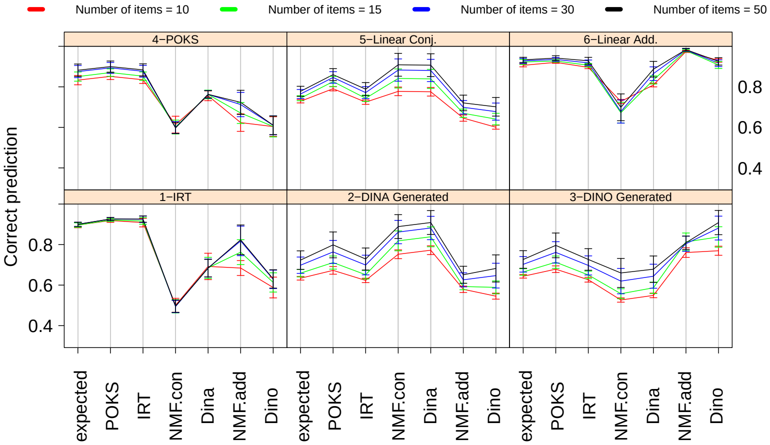
Obviously, the signature pattern did not change significantly for some parameters such as **Sample size**.

Variation of number of items over synthetic data sets



Even for synthetic data the ground truth should not necessarily be close to 100

Variation of number of items over synthetic data sets



The Performance signature shifts down once the number of items degrades.

Data specific parameters

1. Sample size (Number of students)
2. Number of items
3. Number of latent skills
4. Item score variance
5. Student score variance
6. Average success rate

Data specific parameters

1. Sample size (Number of students)
2. Number of items
3. Number of latent skills
4. Item score variance
5. Student score variance
6. Average success rate

Conclusion :

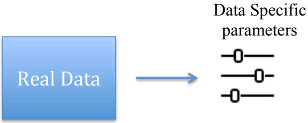
- Contextual factors can potentially influence the performance of a model
- For better comparison of the results, we can also consider **data specific parameters** of the real data in the generation process

Research questions

1. ✓What is the *performance vector* of student skills assessment models over real and over synthetic data created using the same models ?
    - Experiment 1 : Predictive performance of models over real and synthetic data sets
2. ✓Is the *performance vector* unique to each synthetic data type (data from the same ground truth model) ?
    - Experiment 2 : Sensitivity of the Model performance over different data generation parameters
3. **Can the performance vector be used to define a method to reliably identify the ground truth behind the synthetic data ?**
    - Experiment 3 : Model selection based on performance vector classification

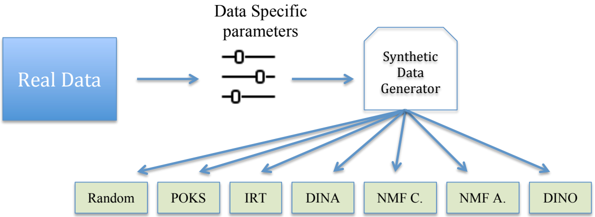
Signature framework

This approach relies on generating synthetic datasets



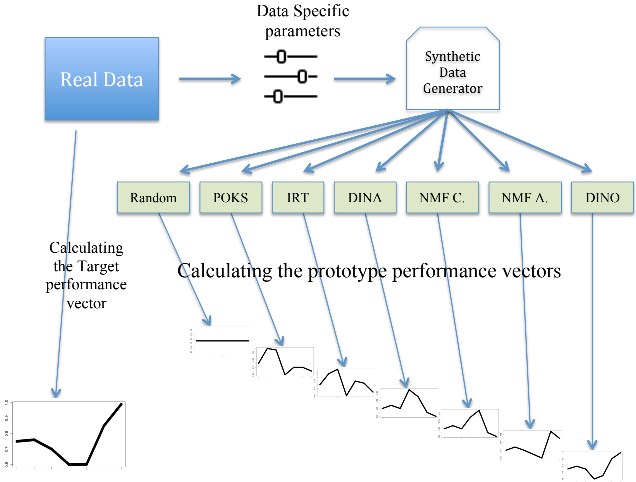
Signature framework

This approach relies on generating synthetic datasets



Signature framework

This approach relies on generating synthetic datasets



Data Specific parameters

Real Data

Synthetic Data Generator

Random

POKS

IRT

DINA

NMF C.

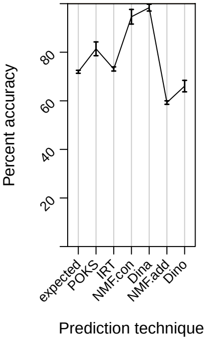
NMF A.

DINO

Calculating the Target performance vector

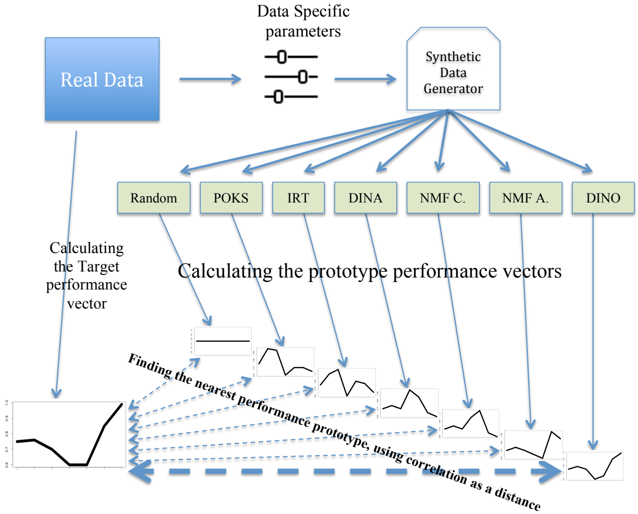
Calculating the prototype performance vectors

Recall



Signature framework

This approach relies on generating synthetic datasets



Data Specific parameters

Real Data

Synthetic Data Generator

Random

POKS

IRT

DINA

NMF C.

NMF A.

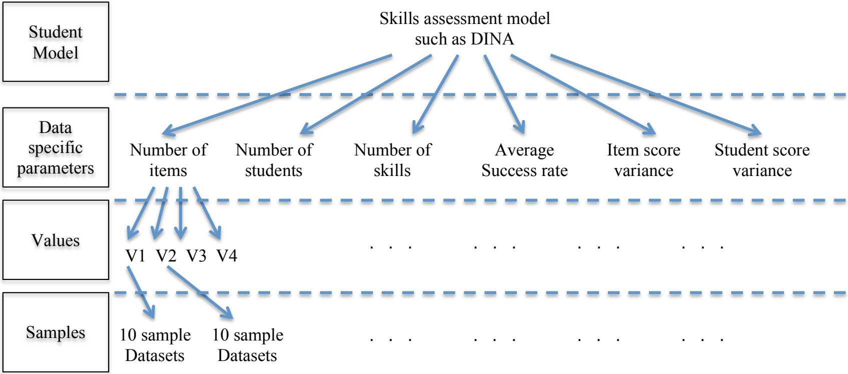
DINO

Calculating the Target performance vector

Calculating the prototype performance vectors

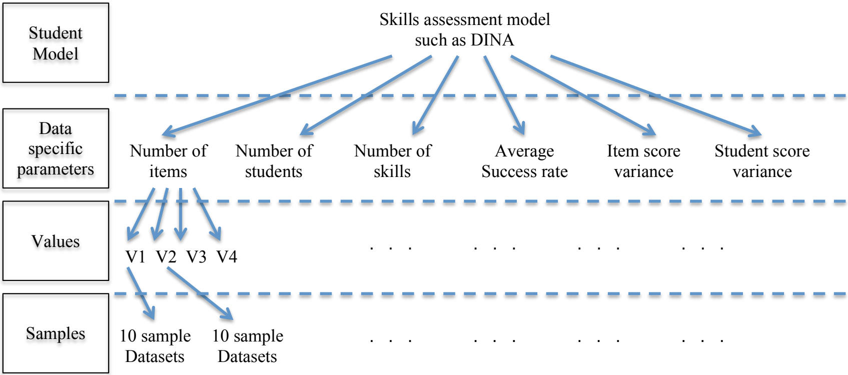
Finding the nearest performance prototype, using correlation as a distance

Pool of synthetic datasets



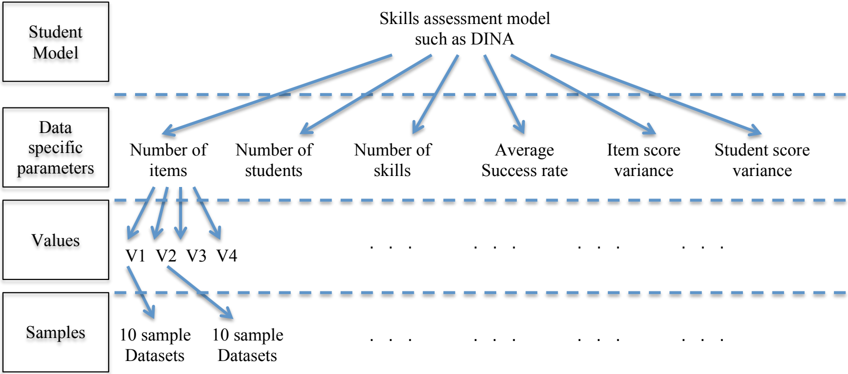
There exists 6 skills assessment models

Pool of synthetic datasets



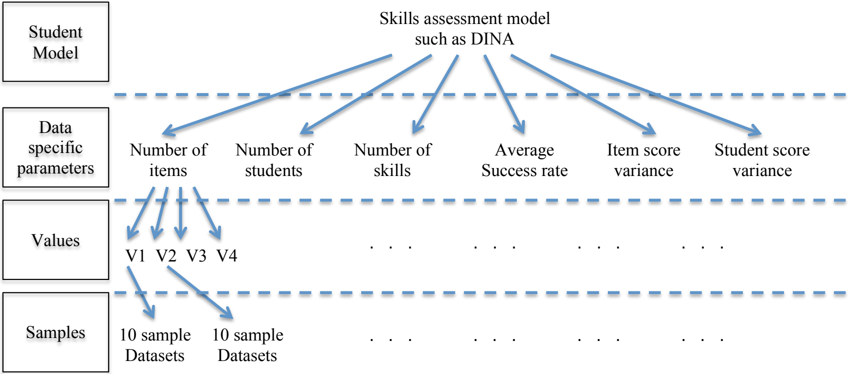
There exists 6 skills assessment models X 6 data specific parameters

Pool of synthetic datasets



There exists 6 skills assessment models X 6 data specific parameters X 4 values

Pool of synthetic datasets



There exists 6 skills assessment models X 6 data specific parameters X 4 values X 10 samples = 1440 samples in the pool

Degree of similarity between six synthetic datasets based on the correlation

Synthetic Datasets

Synthetic Datasets

	POKS	IRT	NMF Conj.	DINA	NMF Add.	DINO
POKS	**0.96**					
IRT	0.86	**0.96**				
NMF Conj.	0.22	-0.20	**0.96**			
DINA	0.02	-0.40	0.94	**0.96**		
NMF Add.	0.44	0.75	-0.62	-0.73	**0.93**	
DINO	-0.15	0.20	-0.70	-0.69	0.63	**0.95**

- he diagonal shows high correlations because it compares the same model generated datasets.
- Datasets with similar ground truth also show a high correlation.

Degree of similarity between six synthetic datasets based on the correlation

Synthetic Datasets

		POKS	IRT	NMF Conj.	DINA	NMF Add.	DINO
Synthetic Datasets	POKS	**0.96**					
	IRT	0.86	**0.96**				
	NMF Conj.	0.22	-0.20	**0.96**			
	DINA	0.02	-0.40	0.94	**0.96**		
	NMF Add.	0.44	0.75	-0.62	-0.73	**0.93**	
	DINO	-0.15	0.20	-0.70	-0.69	0.63	**0.95**

In general, correlation similarity provides a very good measure of model fit.

Degree of similarity between six synthetic datasets and the ground truth based on the correlation

Real Datasets

					Fraction subsets			
		Vomlel	ECPE	Fraction	1	21	22	23
Synthetic Datasets	Random	0.58	**0.73**	0.61	0.43	0.24	0.61	0.57
	IRT	**0.90**	0.42	0.72	0.88	0.60	0.77	0.61
	DINA	-0.38	-0.09	0.23	0.30	0.56	0.06	0.38
	DINO	0.34	0.15	-0.18	-0.31	0.10	-0.08	0.38
	POKS	0.75	0.40	**0.83**	**0.95**	**0.70**	**0.83**	**0.80**
	NMF Conj.	-0.05	0.54	0.51	0.55	0.66	0.33	0.57
	NMF Add.	0.39	0.06	-0.04	-0.19	-0.03	0.13	0.28

Vomlel dataset shows a high correlation with IRT model

Degree of similarity between six synthetic datasets and the ground truth based on the correlation

Real Datasets

					Fraction subsets			
		Vomlel	ECPE	Fraction	1	21	22	23
Synthetic Datasets	Random	0.58	**0.73**	0.61	0.43	0.24	0.61	0.57
	IRT	**0.90**	0.42	0.72	0.88	0.60	0.77	0.61
	DINA	-0.38	-0.09	0.23	0.30	0.56	0.06	0.38
	DINO	0.34	0.15	-0.18	-0.31	0.10	-0.08	0.38
	POKS	0.75	0.40	**0.83**	**0.95**	**0.70**	**0.83**	**0.80**
	NMF Conj.	-0.05	0.54	0.51	0.55	0.66	0.33	0.57
	NMF Add.	0.39	0.06	-0.04	-0.19	-0.03	0.13	0.28

Fraction with its subset datasets show similarity with POKS model.

Degree of similarity between six synthetic datasets and the ground truth based on the correlation

Real Datasets

					Fraction subsets			
		Vomlel	ECPE	Fraction	1	21	22	23
Synthetic Datasets	Random	0.58	**0.73**	0.61	0.43	0.24	0.61	0.57
	IRT	**0.90**	0.42	0.72	0.88	0.60	0.77	0.61
	DINA	-0.38	-0.09	0.23	0.30	0.56	0.06	0.38
	DINO	0.34	0.15	-0.18	-0.31	0.10	-0.08	0.38
	POKS	0.75	0.40	**0.83**	**0.95**	**0.70**	**0.83**	**0.80**
	NMF Conj.	-0.05	0.54	0.51	0.55	0.66	0.33	0.57
	NMF Add.	0.39	0.06	-0.04	-0.19	-0.03	0.13	0.28

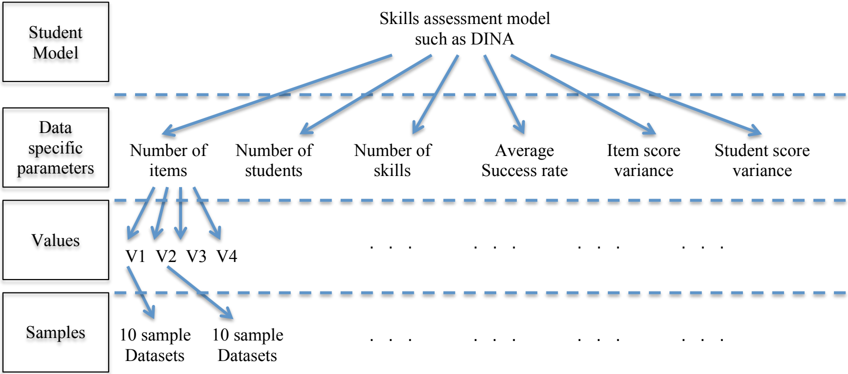
As expected, ECPE has the highest correlation with random generated dataset.

Research questions

1. ✓What is the *performance vector* of student skills assessment models over real and over synthetic data created using the same models ?
    - Experiment 1 : Predictive performance of models over real and synthetic data sets
2. ✓Is the *performance vector* unique to each synthetic data type (data from the same ground truth model) ?
    - Experiment 2 : Sensitivity of the Model performance over different data generation parameters
3. ✓Can the *performance vector* be used to define a method to reliably identify the ground truth behind the synthetic data ?
    - Experiment 3 : Model selection based on performance vector classification
4. **How does the method compare with the standard practice of using the model with the best performance ?**
    - Experiment 4 : Signature vs. best performer classification

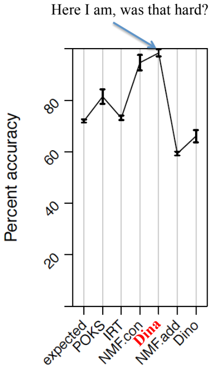
Problem specification

- Evaluating the ability of the Signature approach to identify the ground truth model **over datasets with different data specific parameters**



Problem specification

- Evaluating the ability of the Signature approach to identify the ground truth model **over datasets with different data specific parameters**
- Comparing the results with the best performer approach.



Problem specification

- Evaluating the ability of the Signature approach to identify the ground truth model **over datasets with different data specific parameters**
- Comparing the results with the best performer approach.
- Reporting the accuracy of these classification in terms of precision, recall, F1 measure

		Prediction outcome	
		Positive	Negative
Actual value	Positive	TP	FN
	Negative	FP	TN

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F_\beta = (1 + \beta^2).\frac{Precision.Recall}{\beta^2.Precision + Recall}$$

Results of signature vs. best performer classification

		Performance							
		Best Performer				Nearest Neighbor			
		Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure	Accuracy
Models	POKS	0.564	0.992	0.719	0.871	0.793	0.908	0.847	0.945
	IRT	0.982	0.458	0.625	0.908	0.846	0.867	0.856	0.951
	NMF Conj.	0.943	0.342	0.502	0.887	0.711	0.750	0.730	0.907
	DINA	0.617	0.921	0.739	0.891	0.777	0.696	0.734	0.916
	NMF Add.	0.938	0.875	0.905	0.969	0.946	0.879	0.911	0.971
	DINO	1	0.929	0.963	0.988	0.996	0.946	0.970	0.990
Total accuracy		0.75%				0.84%			

- The F-measure increases when the signature approach is used for classification.

Results of signature vs. best performer classification

		Performance							
		Best Performer				Nearest Neighbor			
		Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure	Accuracy
Models	POKS	0.564	0.992	0.719	0.871	0.793	0.908	0.847	0.945
	IRT	0.982	0.458	0.625	0.908	0.846	0.867	0.856	0.951
	NMF Conj.	0.943	0.342	0.502	0.887	0.711	0.750	0.730	0.907
	DINA	0.617	0.921	0.739	0.891	0.777	0.696	0.734	0.916
	NMF Add.	0.938	0.875	0.905	0.969	0.946	0.879	0.911	0.971
	DINO	1	0.929	0.963	0.988	0.996	0.946	0.970	0.990
Total accuracy		0.75%				0.84%			

- The F-measure increases when the signature approach is used for classification.
- In terms of individual scores per method, the accuracy increases when signature approach is used.

Results of signature vs. best performer classification

		Performance							
		Best Performer				Nearest Neighbor			
		Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure	Accuracy
Models	POKS	0.564	0.992	0.719	0.871	0.793	0.908	0.847	0.945
	IRT	0.982	0.458	0.625	0.908	0.846	0.867	0.856	0.951
	NMF Conj.	0.943	0.342	0.502	0.887	0.711	0.750	0.730	0.907
	DINA	0.617	0.921	0.739	0.891	0.777	0.696	0.734	0.916
	NMF Add.	0.938	0.875	0.905	0.969	0.946	0.879	0.911	0.971
	DINO	1	0.929	0.963	0.988	0.996	0.946	0.970	0.990
Total accuracy		0.75%				0.84%			

- The F-measure increases when the signature approach is used for classification.
- In terms of individual scores per method, the accuracy increases when signature approach is used.
- The total accuracy considers true positive numbers over number of datasets regardless of individual models

Research questions

1. ✓What is the *performance vector* of student skills assessment models over real and over synthetic data created using the same models ?
    - Experiment 1 : Predictive performance of models over real and synthetic data sets
2. ✓Is the *performance vector* unique to each synthetic data type (data from the same ground truth model) ?
    - Experiment 2 : Sensitivity of the Model performance over different data generation parameters
3. ✓Can the *performance vector* be used to define a method to reliably identify the ground truth behind the synthetic data ?
    - Experiment 3 : Model selection based on performance vector classification
4. ✓How does the method compare with the standard practice of using the model with the best performance ?
    - Experiment 4 : Signature vs. best performer classification

Conclusion

- 1

Conclusion

- 1
- 2

Conclusion

- 1
- 2
- 3

Conclusion

- 1
- 2
- 3
- 4

Conclusion

- 1
- 2
- 3
- 4
- 5

Conclusion

- 1
- 2
- 3
- 4
- 5
- 6

Conclusion

- 1
- 2
- 3
- 4
- 5
- 6

Future works

- 1

Future works

- 1
- 2

Future works

- 1
- 2
- 3

Future works

- 1
- 2
- 3
- 4

Future works

- 1
- 2
- 3
- 4
- 5

Future works

- 1
- 2
- 3
- 4
- 5

Thank you