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

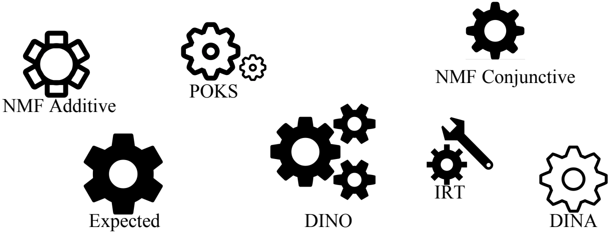
Behzad Beheshti
Supervisor : Michel C. Desmarais

Génie Informatique Et Génie Logiciel
École Polytechnique de Montréal

7 avril 2016

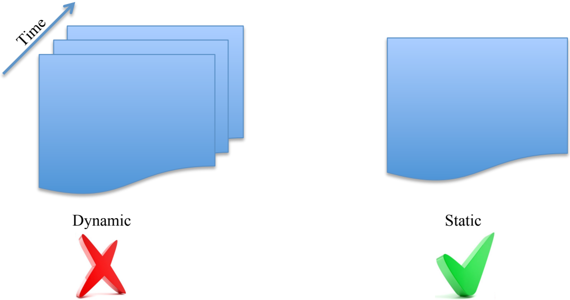
Problem Specification

- Student skills assessment models



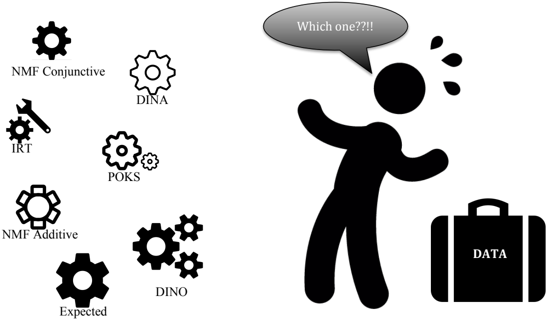
Problem Specification

- Student skills assessment models
- Static Vs. Dynamic



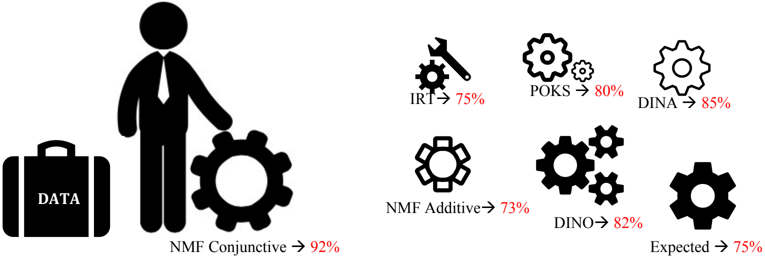
Problem Specification

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



Problem Specification

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



Problem Specification

Our contribution

- To make a comprehensive comparison of educational data model performances
- To propose a new approach to assessing model fit

Problem Specification

Our contribution

- To make a comprehensive comparison of educational data model performances
- To propose a new approach to assessing model fit

General hypothesis :

- Recognizing the ground truth based on the uniqueness of this comprehensive comparison

Problem Specification

Our contribution

- To make a comprehensive comparison of educational data model performances
- To propose a new approach to assessing model fit

General hypothesis :

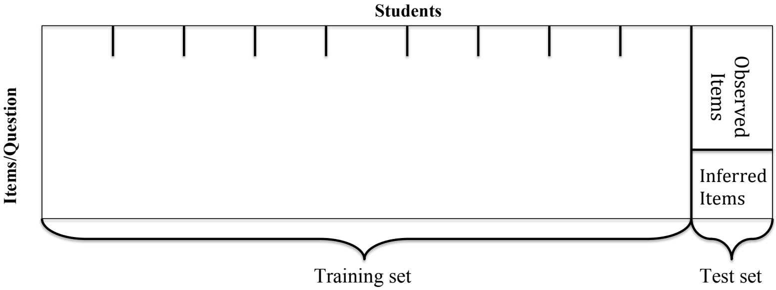
- Recognizing the ground truth based on the uniqueness of this comprehensive comparison

The proposed approach :

- Assessing the fit of the model to the underlying ground truth using a methodology based on **synthetic data**

Presentation terms and concepts

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



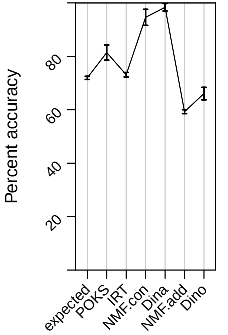
Presentation terms and concepts

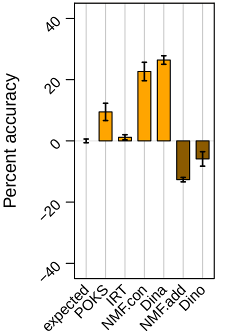
- **Performance of a model** over a data set
- Model parameters

<table>
<tr><th colspan="3"></th><th colspan="3">Parameters estimated from</th></tr>
<tr><th colspan="3">Skills Model</th><th colspan="2">Training set</th><th>Observed items</th></tr>
<tr><td rowspan="7">Contributed skills</td><td rowspan="4">Multiple</td><td>NMF Conj.</td><td rowspan="2"></td><td rowspan="4">• Q-matrix</td><td rowspan="4">• Students skills mastery matrix</td></tr>
<tr><td>NMF Add.</td></tr>
<tr><td>DINA</td><td rowspan="2">• Slip
• Guess</td></tr>
<tr><td>DINO</td></tr>
<tr><td rowspan="2">Single</td><td>IRT</td><td colspan="2">• Item difficulty
• Item discrimination</td><td>• Student Ability</td></tr>
<tr><td>Expected</td><td colspan="2">• Item Odds</td><td>• Student Odds</td></tr>
<tr><td>Zero</td><td>POKS</td><td colspan="2">• Initial Odds
• Odds ratio
• Partial order structure</td><td></td></tr>
</table>

Presentation terms and concepts

- **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%

Presentation terms and concepts

- **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.

Presentation terms and concepts

- **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.

Vector space of accuracy performances

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

Model	**Synthetic data set**						
	Random	POKS	IRT	DINA	DINO	L.Conj.	L.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

Vector space of accuracy performances

- Given a target vector, we want to define which columns are the closest to the target column

Model	Synthetic data set							Target
	Random	POKS	IRT	DINA	DINO	L.Conj.	L.Comp.	
Expected	0.75	0.91	0.90	0.72	0.72	0.78	0.93	0.43
POKS	0.75	0.94	0.94	0.81	0.81	0.90	0.94	0.75
IRT	0.75	0.91	0.95	0.73	0.73	0.79	0.89	0.68
DINA	0.75	0.77	0.81	1.00	0.65	0.98	0.89	0.93
DINO	0.75	0.63	0.56	0.66	1.00	0.68	0.91	0.60
NMF.Conj	0.75	0.59	0.53	0.95	0.65	0.97	0.58	0.80
NMF.Comp	0.75	0.76	0.79	0.59	0.93	0.70	0.98	0.70

- Our hypothesis : the similarity of vectors can indicate the nature of data
- The nearest neighbour in this space can be a candidate for the ground truth

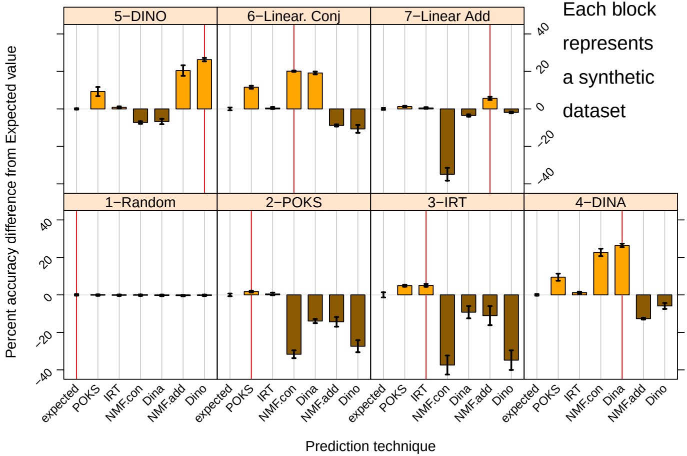
Research questions

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 vector* of models over real and synthetic data sets

Datasets

Data set	Number of Skills	Number of Items	Number of Students	Mean Score	Q-matrix
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	$\mathbf{Q}_{13}$

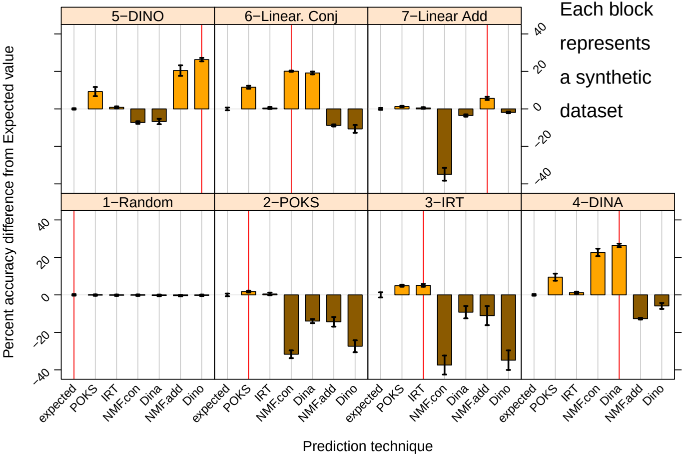
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

The random data set has a flat performance across techniques

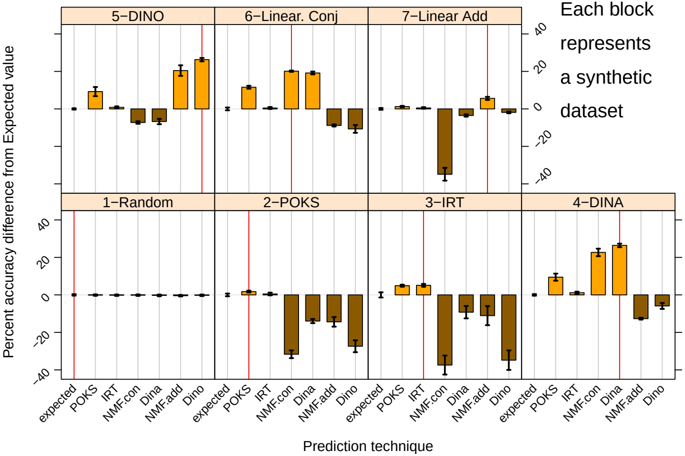
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

The highest performance is for the generative model behind the dataset

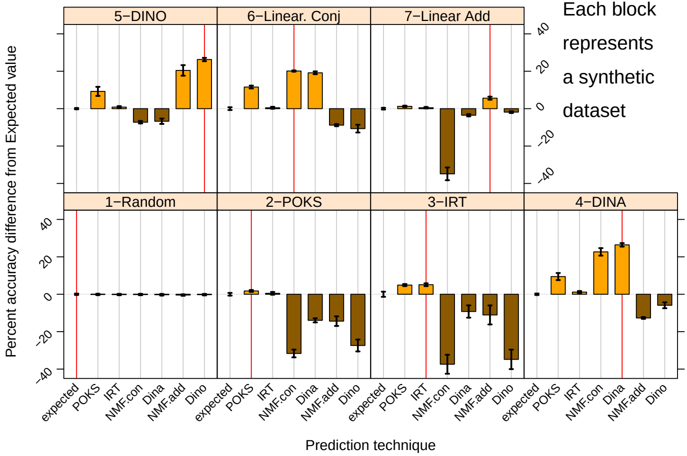
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

Motivation : Most of the models preform worse than simple expected prediction model

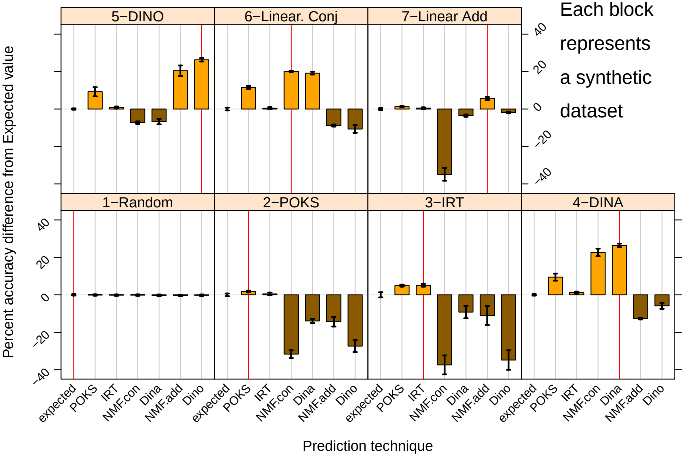
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

Data sets have discriminant pattern of performance vector across models

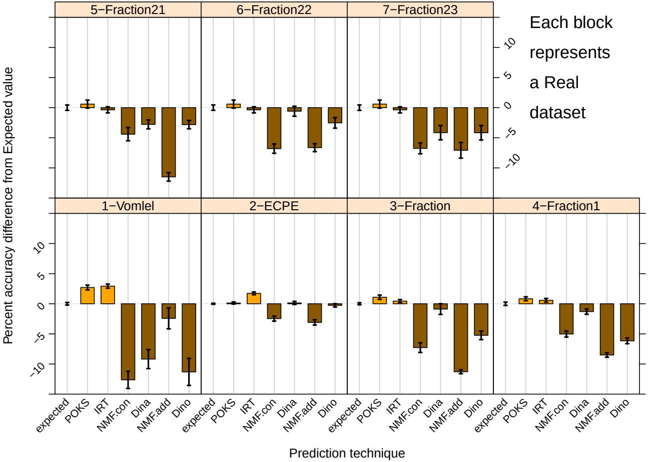
Predictive performance of models over synthetic datasets



Each block represents a synthetic dataset

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

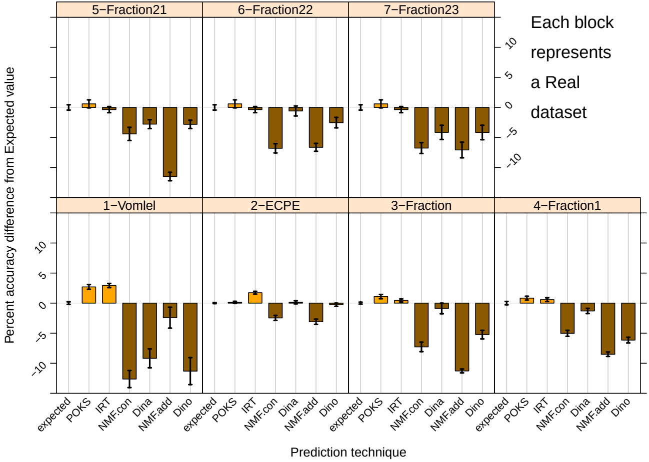
Predictive performance of models over real datasets



Each block represents a Real dataset

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

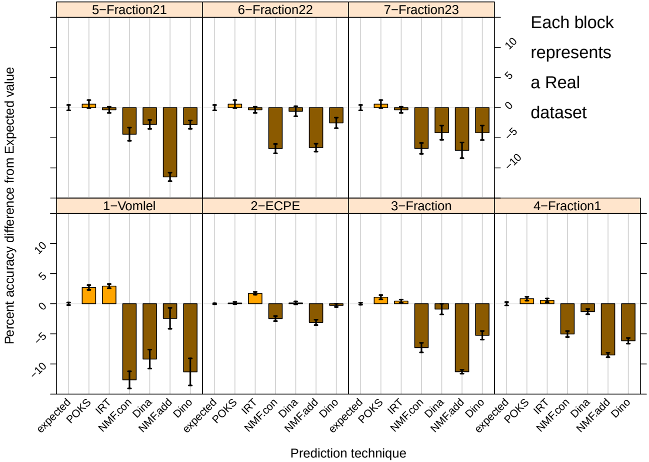
Predictive performance of models over real datasets



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

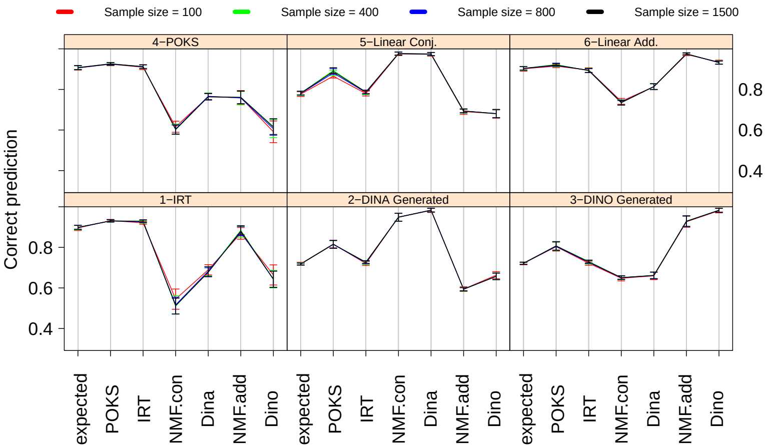
None of the real data sets show the large variance and the differences found in the synthetic data sets models

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

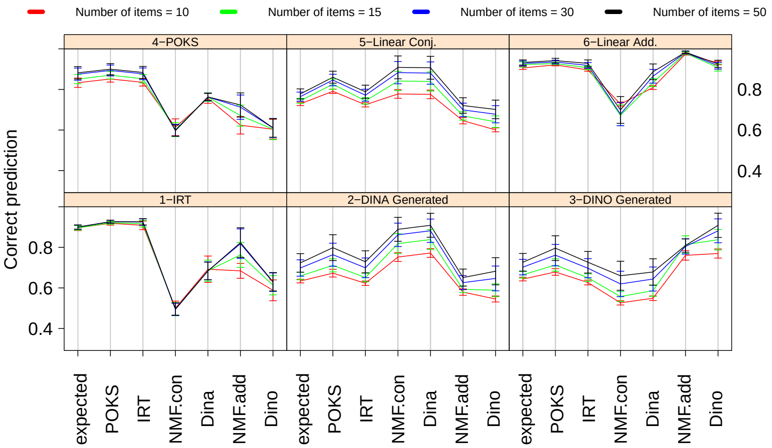
   Are they stable in addition to be discriminant ?

Variation of sample size over synthetic data sets



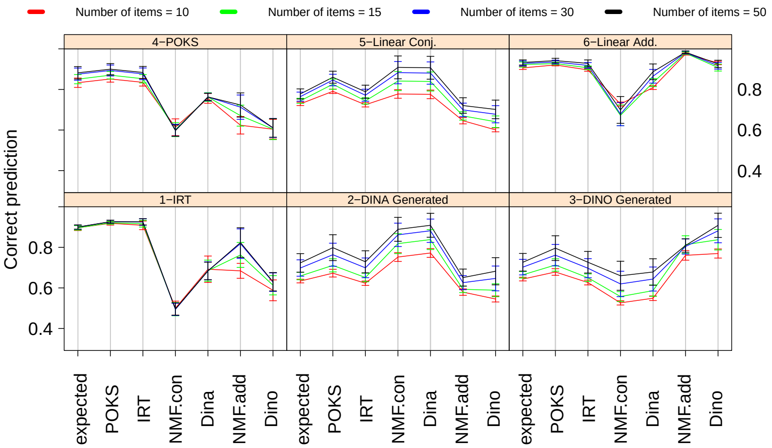
Obviously, the signature pattern did not change significantly for some parameters such as **Sample size**.

Variation of number of items over synthetic data sets



Even for synthetic data the performance of the ground truth model 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 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 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 :

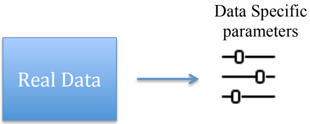
- Data parameters can potentially influence the performance of a model
- For better comparison of the vectors, we also consider **data characteristic parameters** of the real data in the generation process

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

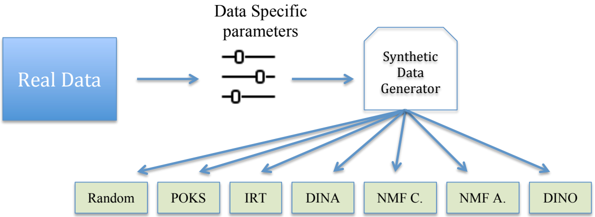
Signature framework

This approach relies on generating synthetic datasets



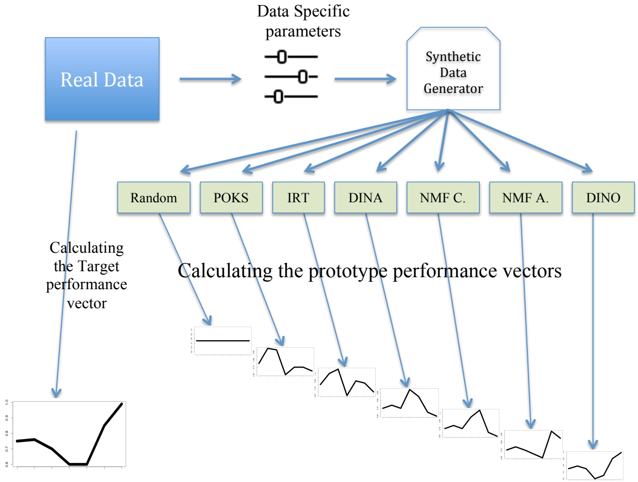
Signature framework

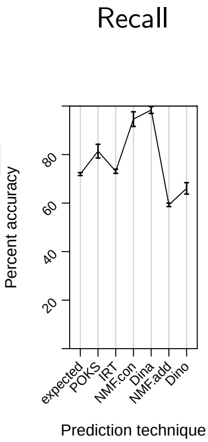
This approach relies on generating synthetic datasets



Signature framework

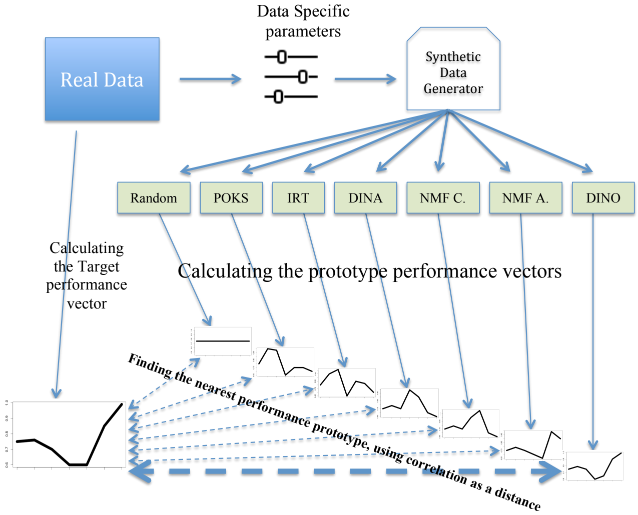
This approach relies on generating synthetic datasets



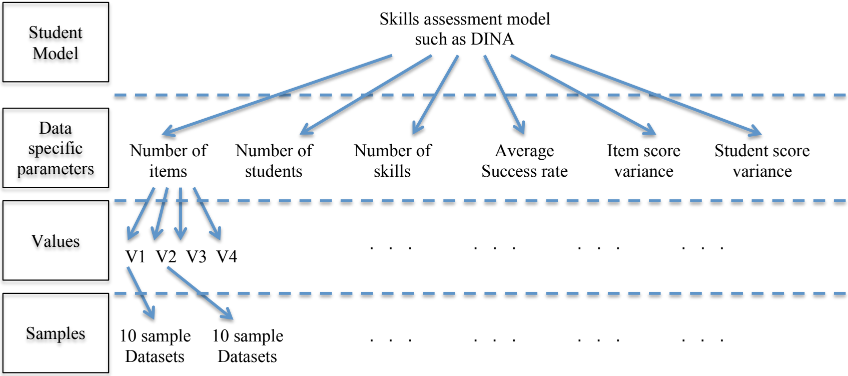


Signature framework

This approach relies on generating synthetic datasets

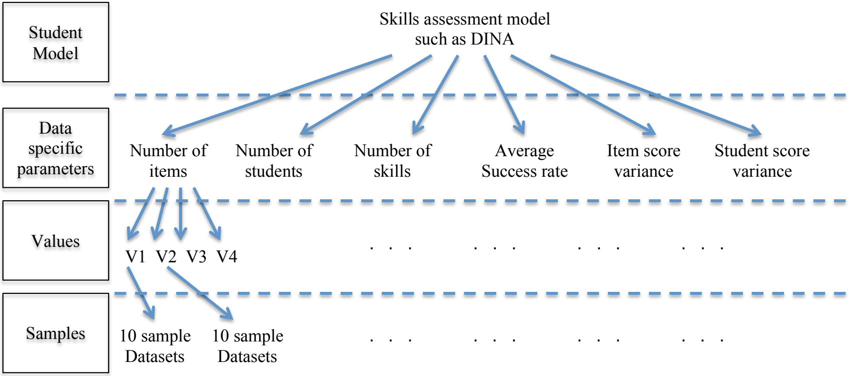


Database of synthetic datasets



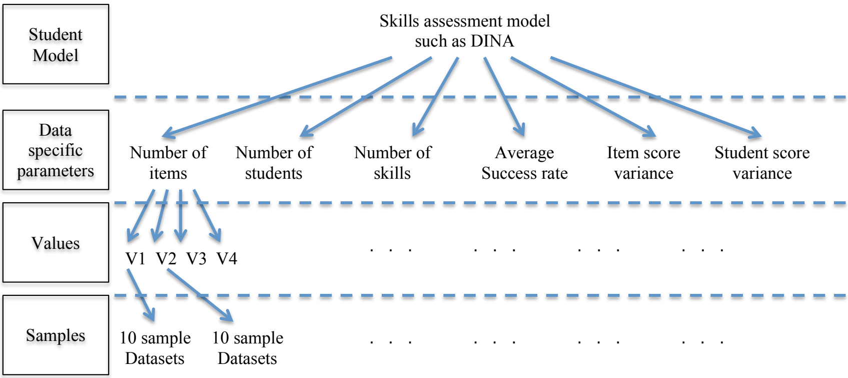
There exists 6 skills assessment models

Database of synthetic datasets



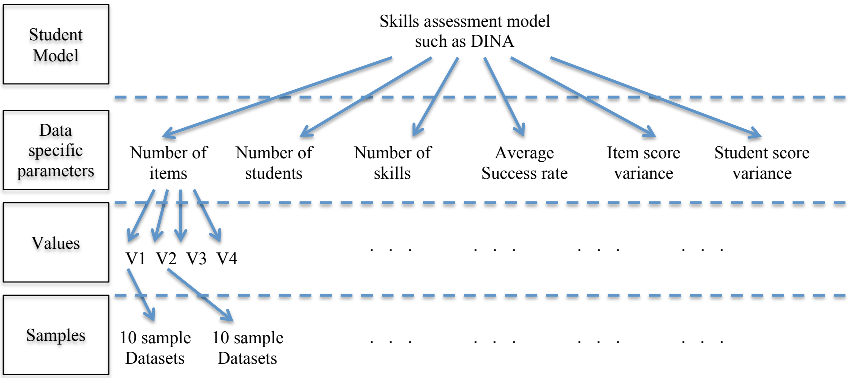
There exists 6 skills assessment models X 6 data characteristic parameters

Database of synthetic datasets



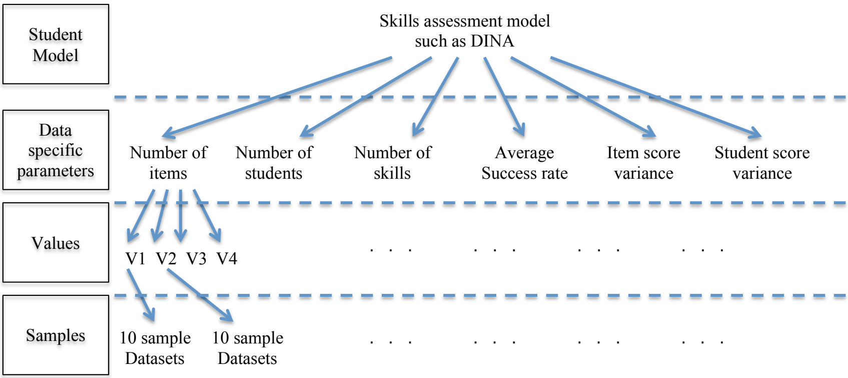
There exists 6 skills assessment models X 6 data characteristic parameters X 4 values

Database of synthetic datasets



There exists 6 skills assessment models X 6 data characteristic parameters X 4 values X 10 samples = 1440 samples in the Database (*DB*)

Database of synthetic datasets



$|DB| = 1440$ and $|DC| = 24$

Prototype definition

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$
where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

Prototype definition

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$ where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

- Given $c \in DC$ and $m \in \mathcal{M}$: $D = \{\forall D_i^{(j)} \in DB | i = c, j = m\}$
- $\mathcal{N}^m$ is a centroid performance prototype defined for model m

Prototype definition

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$ where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

- Given $c \in DC$ and $m \in \mathcal{M}$: $D = \{\forall D_i^{(j)} \in DB | i = c, j = m\}$
- $\mathcal{N}^m$ is a centroid performance prototype defined for model m
- $\mathcal{N}^m = AVG(Performance.vector(D))$

Prototype definition

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$
where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

- Given $c \in DC$ and $m \in \mathcal{M}$: $D = \{\forall D_i^{(j)} \in DB | i = c, j = m\}$
- $\mathcal{N}^m$ is a centroid performance prototype defined for model m
- $\mathcal{N}^m = AVG(Performance.vector(D))$
- Pearson correlation coefficient is used as a measure of similarity between $\mathcal{T}^{(tm)}$ and $\mathcal{N}^m$ (for data condition c)

Prototype definition

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$ where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

- Given $c \in DC$ and $m \in \mathcal{M}$: $D = \{\forall D_i^{(j)} \in DB | i = c, j = m\}$
- $\mathcal{N}^m$ is a centroid performance prototype defined for model m
- $\mathcal{N}^m = AVG(Performance.vector(D))$
- Pearson correlation coefficient is used as a measure of similarity between $\mathcal{T}^{(tm)}$ and $\mathcal{N}^m$ (for data condition c)
- Next slid shows the average of correlation between $\mathcal{T}^{(tm)}$ and $\mathcal{N}^m$ for all data parameter conditions in DC

Degree of similarity between six synthetic performance vector based on the correlation

Target performance vectors

Centroid synthetic performance vectors

	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**

- The diagonal shows high correlations because it compares performance vectors of the same model generated datasets.
- Performance vectors with similar ground truth also show a high correlation.

Degree of similarity between six synthetic performance vector based on the correlation

Target performance vectors

Centroid synthetic performance vectors	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**

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

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

Target performance vectors of Real Datasets

Performance Prototype of Synthetic Datasets	Vomlel	ECPE	Fraction	Fraction subsets: 1	Fraction subsets: 21	Fraction subsets: 22	Fraction subsets: 23
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

The centroid performance vector of each model is the average of performance vectors of 10 generated datasets

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

Target performance vectors of Real Datasets

Performance Prototype of Synthetic Datasets

	Vomlel	ECPE	Fraction	Fraction subsets: 1	21	22	23
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 real datasets and the ground truth based on the correlation

Target performance vectors of Real Datasets

Performance Prototype of Synthetic Datasets

	Vomlel	ECPE	Fraction	Fraction subsets: 1	21	22	23
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 real datasets and the ground truth based on the correlation

Target performance vectors of Real Datasets

Performance Prototype of Synthetic Datasets

	Vomlel	ECPE	Fraction	Fraction subsets 1	21	22	23
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

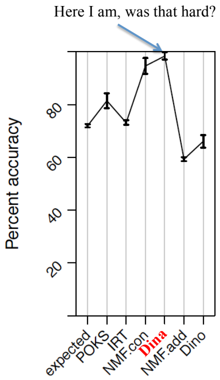
1. ✓What is the *performance vector* of student skills assessment models over real and over synthetic data created using the same models ?
   - Experiment 1 : Predictive performance of models over real and synthetic data sets
2. ✓Is the *performance vector* unique to each synthetic data type (data from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

- What we did ? to evaluate the ability of the Signature approach to identify the ground truth model

Problem specification

- What we did ? to evaluate the ability of the Signature approach to identify the ground truth model
- What we want to do ? to compare the results of classification based on signature approach with the best performer approach.



Problem specification

- What we did ? to evaluate the ability of the Signature approach to identify the ground truth model
- What we want to do ? to compare the results of classification based on signature approach with the best performer approach.
- Reporting the accuracy of these classification in terms of F1 and accuracy measures

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

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

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

$$F_1 = \frac{Precision.Recall}{Precision + Recall}$$

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

Signature classifier

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$
where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

Signature classifier

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$ where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

- Given $c \in DC$: $D = \{\forall D_i^{(j)} \in DB | i = c, j \in \mathcal{M}\}$

Signature classifier

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$ where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

- Given $c \in DC$: $D = \{\forall D_i^{(j)} \in DB | i = c, j \in \mathcal{M}\}$
- Neighbors of $\mathcal{T}^{(tm)}$ are performance vectors of D

Signature classifier

Target performance vector : $\mathcal{T}^{(tm)} = Performance.vector(D_c^{(tm)} \in DB)$ where : $D_c^{(tm)}$ is characterized with data parameter condition $c \in DC$ and ground truth (tm)

- Given $c \in DC$: $D = \{\forall D_i^{(j)} \in DB | i = c, j \in \mathcal{M}\}$
- Neighbors of $\mathcal{T}^{(tm)}$ are performance vectors of D
- A majority voting among the ground truth of 10 nearest neighbors to $\mathcal{T}^{(tm)}$ defines the estimated ground truth

Results of signature vs. best performer classification

Performance

Models		Best Performer		Signature approach	
		F-Measure	Accuracy	F-Measure	Accuracy
Models	POKS	0.719	0.871	0.998	99.9
	IRT	0.625	0.908	0.998	99.9
	NMF Conj.	0.502	0.887	0.985	99.5
	DINA	0.734	0.891	1	100
	NMF Add.	0.905	0.969	1	100
	DINO	0.963	0.988	0.970	0.990
Total accuracy		75%		99%	

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

Results of signature vs. best performer classification

Performance

		Best Performer		Signature approach	
		F-Measure	Accuracy	F-Measure	Accuracy
Models	POKS	0.719	0.871	0.998	99.9
	IRT	0.625	0.908	0.998	99.9
	NMF Conj.	0.502	0.887	0.985	99.5
	DINA	0.734	0.891	1	100
	NMF Add.	0.905	0.969	1	100
	DINO	0.963	0.988	0.970	0.990
Total accuracy		75%		99%	

- 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		Signature approach	
		F-Measure	Accuracy	F-Measure	Accuracy
Models	POKS	0.719	0.871	0.998	99.9
	IRT	0.625	0.908	0.998	99.9
	NMF Conj.	0.502	0.887	0.985	99.5
	DINA	0.734	0.891	1	100
	NMF Add.	0.905	0.969	1	100
	DINO	0.963	0.988	0.970	0.990
Total accuracy		75%		99%	

- 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
- In terms of signature pattern : it is unique for each generative model
- In terms of data points in the performance vector space : They are data points in this space
- For real data sets, the performances are not better than the expected performance
- For synthetic data, datasets with different ground truths that share some concepts, show a high correlation.

Research questions

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
- Different parameters can create different points in the performance vector space
- There would be a cloud of points for a particular model (Ground truth)
- The cloud is not too dispersed

Research questions

1. ✓ Predictive performance of models over real and synthetic data sets
2. ✓ Sensitivity of the Model performance over different data 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
- A comparison between the target and the prototype signature
- Datasets that share a common source have correlated performance vectors.

Research questions

1. ✓Predictive performance of models over real and synthetic data sets
2. ✓Sensitivity of the Model performance over different data parameters
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
- The ground truth model does not always correspond to the best performer and our approach provides a more reliable means

Future works

- to be extend over different models, different domains and more datasets

Future works

- to be extend over different models, different domains and more datasets
- generalize to dynamic data and skills assessment models

Future works

- to be extend over different models, different domains and more datasets
- generalize to dynamic data and skills assessment models
- Candidate models and their complexity where the data is created with a mixture of models

Future works

- to be extend over different models, different domains and more datasets
- generalize to dynamic data and skills assessment models
- Candidate models and their complexity where the data is created with a mixture of models
- Application in other fields of study

Thank you