Predicting the outcome of in vitro fertilization

G. Corani[1] C. Magli[2] A. Giusti[1] L. Gambardella[1] L. Gianaroli[2]

1: IDSIA (Ist. Dalle Molle Artificial Intelligence, Switzerland)

2: IIRM (Int. Ist. Reproductive Medicine, Switzerland)

giorgio@idsia.ch

In vitro fertilization (IVF)

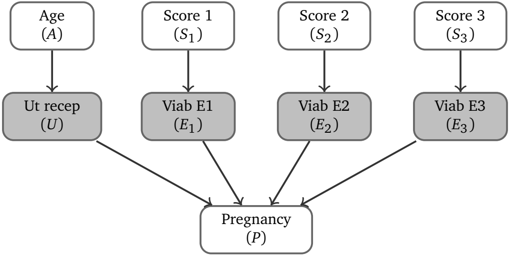
- Infertility affects more than 80 million people worldwide.
- During IVF 1-3 embryos are cultured in vitro; they are then transferred to the woman.
- A *pregnancy* occurs when at least one of the transferred embryos implants.
- Embryos are scored by biologists according to their morphology: {non-top, top, top+}.

The EU assumption (Speirs, 1983)

- Pregnancy requires a *receptive* uterus and a *viable* embryo.
- The probability of the uterus being *receptive* and of the embryo being *viable* are respectively θ_u and θ_e.
- It is assumed the independence of viability and receptivity.
- In case of a single embryo transferred: $P(pregnancy) = \theta_u \cdot \theta_e$.
- Usually, θ_u depends on the age of the woman and θ_e on the score of the embryo.

The BN-EU model

- Models the IVF outcome under the EU assumption
- The *goal*: estimating the probability of pregnancy, given the score of the transferred embryos and the age of the woman.



Nodes with a gray background are affected by a missingness process.

How to estimate θ_u and θ_e?

- There is a partial observability problem (Zhou and Weinberg, 1998; Robert 2007, 2009).
- If pregnancy does *not* occur, we cannot ascertain whether:
  - the uterus is *non-receptive*;
  - *each* embryo is *non-viable*;
  - the uterus is non-receptive **and** each embryo is non-viable.

Training instance, in case of no pregnancy.

A	U	S_1	S_2	S_3	E_1	E_2	E_3	P
40+	*?*	*top*	*ntop*	*toph*	*?*	*?*	*?*	*0*

Training instances

If pregnancy *does* occur:

- the embryo is known to be receptive but ...
- it is unknown which embryo is viable, unless the number of babies matches the number of embryos.

Training instance (single pregnancy).

A	U	S_1	S_2	S_3	E_1	E_2	E_3	P
40+	*u*	*top*	*ntop*	*toph*	*?*	*?*	*?*	*1*

Training instance (triple pregnancy).

A	U	S_1	S_2	S_3	E_1	E_2	E_3	P
40+	*u*	*top*	*ntop*	*toph*	*e*	*e*	*e*	*1*

Estimation procedure

- The missingness process is MAR (*missing at random*); parameters can be estimated via EM (Expectation Maximization).

- *MAP estimation*: among m EM runs, the estimate with the highest posterior probability $P(\boldsymbol{\theta}|D)$ ($\boldsymbol{\theta}$ denotes the parameters of the model) is selected.

- MAP estimation is a good approximation of Bayesian estimation if the posterior is peaked around the maximum; this is *not* the case when learning from incomplete samples.

- Different EM runs achieve close values of $P(\boldsymbol{\theta}|D)$, returning however *very* different parameter estimates.

The averaging approach

Given a parameter θ_X^x, we weighted-average its estimates across the m EM runs:

$$\hat{\theta}_X^x = \frac{\sum_{i=1}^{i=m} \hat{\theta}_X^{x-i} P(\hat{\boldsymbol{\theta}}^i|D)}{\sum_{i=1}^{i=m} P(\hat{\boldsymbol{\theta}}^i|D)}$$

where $\hat{\theta}_X^{x-i}$ and $P(\hat{\boldsymbol{\theta}}^i|D)$ denote the estimate of θ_X^x and the MAP score obtained in the i-th EM run.

Rationale

Consider the query $P(\mathscr{Z}|y,D)$, where $\mathscr{Z}$ is the set of variables being queried, and y is the available evidence.

Fully Bayesian inference

$$P(\mathscr{Z}|y,D) = \int P(\mathscr{Z}|y,D,\boldsymbol{\theta})P(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$

MAP inference

$$P(\mathscr{Z}|y,D) \approx P(\mathscr{Z}|y,D,\hat{\boldsymbol{\theta}})$$

Pseudo-Bayesian inference

$$P(\mathscr{Z}|y,D) \simeq \sum_{i=1}^{i=m} P(\mathscr{Z}|y,D,\hat{\boldsymbol{\theta}}^i)P(\hat{\boldsymbol{\theta}}^i|D)$$

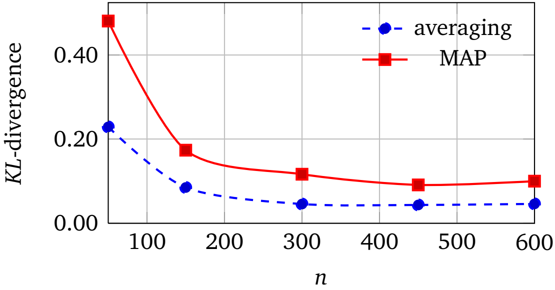
Rationale

- The *pseudo-Bayesian* approach partially reconstructs the shape of the posterior but keeps a collection of m networks, preventing model interpretability.
- The goal of the averaging approach is to retain the advantages of the pseudo-Bayesian approach, but instantiating only a single model.

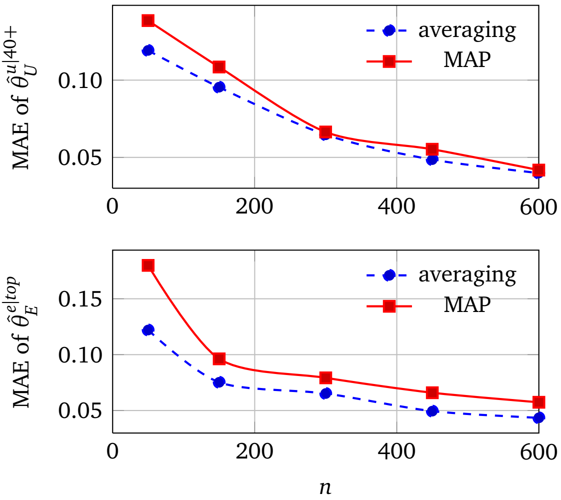
Experiments with generated data

For each sample size, 100 repetitions of the following:

- random drawing of the parameters;
- generation of incomplete instances;
- learning of the parameters by the MAP and the averaging approach;
- classification of the test instances.

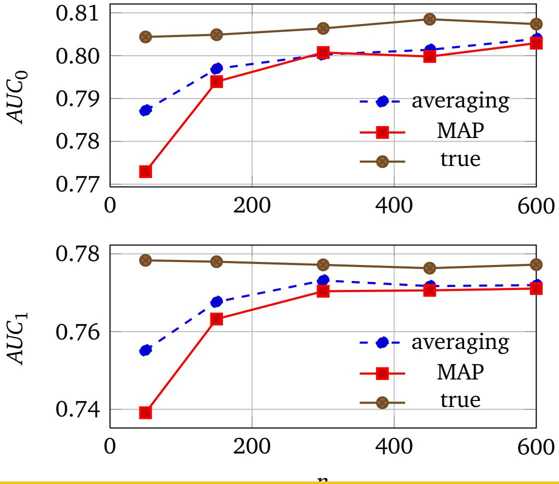


Averaging decreases the estimation error of both receptivities and viabilities.



Averaging increases AUC.

- 4 AUCs: one for each of no-pregnancy (AUC_0), single, double and triple pregnancy.



Comments

- Averaging *significantly* increases AUC and decrease KL-divergences.
- The AUC of the true model is only slightly better than that of the estimated models.
- At test stage receptivity and viabilities are always unknown: this is where lies a major difficulty of predicting IVF!

Test instance

A	U	S_1	S_2	S_3	E_1	E_2	E_3	P
40+	*?*	*top*	*ntop*	*toph*	*?*	*?*	*?*	*?*

The IIRM data set (Lugano, 388 cycles)

- We test BN-EU vs. the high-performance AODE classifier (Webb et al., 2005).
- To learn AODE we build a *complete* data set, with features: the age of the woman and the number of embryos of each type transferred to the woman.
- Despite being learned on a complete data set, AODE does not outperform BN-EU.

	BN-EU	AODE
AUC_0	74.1	**74.8**
AUC_1	67.0	**68.0**
AUC_2	**83.4**	81.6

Conclusions

BN-EU is more interpretable than AODE:

- uterine receptivity drops from 78% to 58% and 26% for woman aged respectively {<34, 34-40, 40+};
- embryo viability increases from 7% to 21% to 39% for embryos scored respectively as non-top, top and top+.

- The BN-EU model can be used to cross-check the effectiveness of the embryo scoring system.

- Future direction of research: studying more covariates on which letting depend both receptivity and viability.

- The averaging approach can be easily added to any EM implementation.