Predicting the outcome of in vitro fertilization

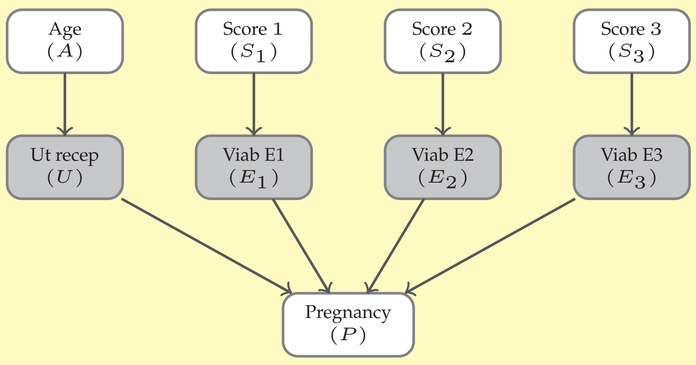
G. Corani, C. Magli, A. Giusti, L. Gambardella, L. Gianaroli

IDSIA and IIRM, Switzerland **giorgio@idsia.ch**

Introduction

- Embryos are cultured *in vitro* and scored on the basis of their morphology.
- A clinical pregnancy occurs when at least one of the transferred embryos implants.
- The EU assumption: pregnancy requires a *receptive* uterus and a *viable* embryo.
- The *goal*: estimating the probability of pregnancy, given the score of the transferred embryos and the age of the woman.

The BN-EU model



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

- A: age of the woman {<34, 34-40, 40+}.
- S_x: score of embryo in position x {non-top, top, top+}.
- U: uterus receptivity (u,$\neg u$).
- E_x: viability of embryo in position x (e,$\neg e$).
- P: pregnancy {0, 1, 2, 3}.

Partial observability

If pregnancy does *not* occur, either:

- the uterus is *non-receptive* (U=$\neg u$);
- *each* embryo is *non-viable* ($E_x = \neg e \ \ \forall x$)
- the uterus is non-receptive **and** each embryo is non-viable.

If pregnancy *does* occur:

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

Training instance (no pregnancy).

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

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*

Test instance (U, E_x and P never observed).

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

Averaging approach

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

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

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

Rationale

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

Fully Bayesian inference

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

MAP inference

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

Pseudo-Bayesian inference

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

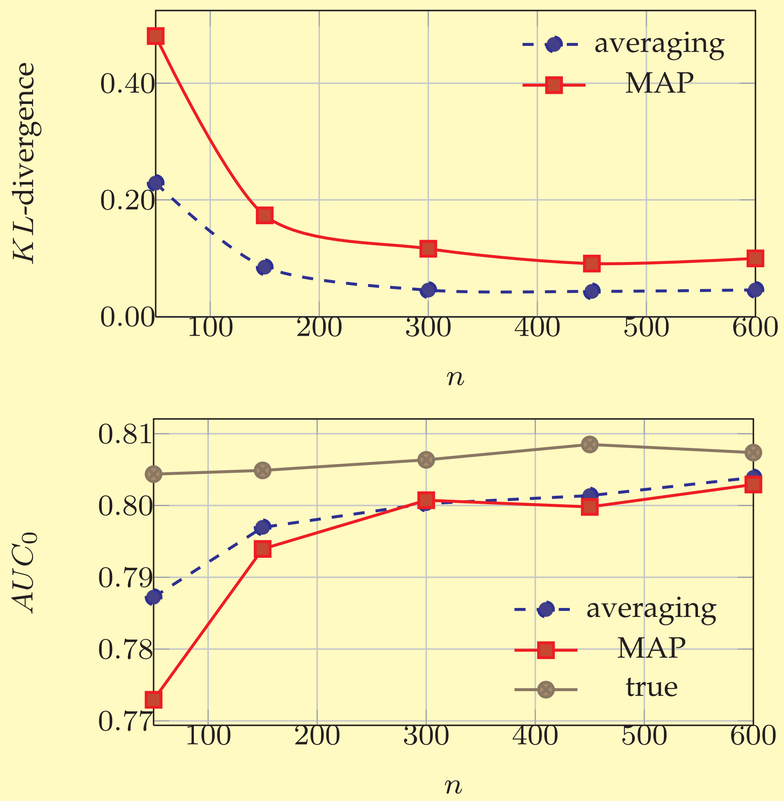
- The *pseudo-Bayesian* approach partially reconstructs the shape of the posterior but keeps a collection of m networks, preventing model interpretability.
- The averaging approach instantiates a single model and produces the same inferences of the pseudo-Bayesian approach, at least in specific queries.

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.

We measure 4 AUCs: one for each of no-pregnancy (AUC_0), single, double and triple pregnancy; we only show AUC_0 in the following



Compared to MAP, averaging decreases the KL-divergence and increases AUC_0.

- The *true* model does not achieve a much higher AUC than the estimated ones, due to the incompleteness of the test set.

The IIRM data set (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

However, BN-EU is more interpretable:

- 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+; these estimates can serve as a cross-check of the embryo scoring system.