A Bayesian Network model for predicting the outcome of in vitro fertilization

Giorgio Corani
IDSIA - Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (Manno, Switzerland)
giorgio@idsia.ch

Cristina Magli
IIRM - International Institute for Reproductive Medicine (Lugano, Switzerland)

Alessandro Giusti
IDSIA - Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (Manno, Switzerland)

Luca Gianaroli
IIRM - International Institute for Reproductive Medicine (Lugano, Switzerland)

Luca Gambardella
IDSIA - Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (Manno, Switzerland)

Abstract

We present a Bayesian network model for predicting the outcome of in-vitro fertilization (IVF). The problem is characterized by a peculiar missingness process, and we propose a simple but effective averaging approach which improves parameter estimates compared to the traditional MAP estimation. The model can provide relevant insights to IVF experts.

1 Introduction

According to the World Health Organization, infertility affects more than 80 million people worldwide; in vitro fertilization (IVF) helps addressing the problem. In IVF, a semen specimen is merged with a female egg in laboratory to yield, after some days of culture, an *embryo*. Several embryos are cultured for each woman: after some days, some or all of them are transferred to the woman. A clinical pregnancy occurs when at least one of the transferred embryos implants. Predicting the outcome of an IVF transfer is a challenging problem, in which models generally achieve only limited accuracy (Saith et al., 1998; Morales et al., 2008).

A pioneering approach for estimating the probability of single pregnancy and multiple pregnancy after an IVF transfer is the EU model (Speirs et al., 1983), which assumes that, for pregnancy to happen, both a *receptive* uterus and a *viable* embryo are necessary. We represent *uterine receptivity* as the binary variable U, with states $\{u, \neg u\}$ (u denoting receptivity, $\neg u$ non-receptivity); we represent *embryo viability* as the binary variable E, with states $\{e, \neg e\}$ (e denoting viability, $\neg e$ non-viability). We denote by θ_e and θ_u respectively the probabilities of the embryo to be viable, namely $\theta_e = P(E = e)$, and $\theta_u = P(U = u)$. The EU model estimates the probability of pregnancy after the transfer of a *single* embryo as $\theta_e \cdot \theta_u$, thus assuming the independence of viability and receptivity. When dealing with the transfer of *multiple* embryos, each embryo is assumed to implant independently from the others. For instance, if *two* embryos are transferred, the probability of a single pregnancy is $2\theta_u\theta_e\ (1 - \theta_e)$, accounting for the fact that two embryos can give rise to pregnancy; the probability of double pregnancy is instead $\theta_e^2\theta_u$. The *EU assumption* is thus that the number of babies born after an IVF transfer is $(U = u) \cdot \sum(E_i = e)$, where E_i

is the viability of the i-th transferred embryo. The main limitation of the original EU model is the unrealistic assumption of θ_e and θ_u being identical for respectively all women and all embryos. Therefore, in (Zhou and Weinberg, 1998) the model has been reworked (adopting a generalized linear model framework) by letting vary both θ_u and θ_e on external covariates; in particular, by letting θ_u depend on the age of the woman and θ_e on the number of cells which the embryo contains (the number of cells contained by an embryo is considered as a marker of its implantation capability). More recently it has been investigated (Roberts et al., 2010) how to select the number and the types of covariates on which θ_u and θ_e should depend. In fact, quantifying how θ_e and θ_u vary with the external covariates such as the age of the woman or the embryo score is an important by-product of the models based on the EU assumption, which allows for instance verifying the effectiveness of the adopted embryo scoring systems, a very important problem for embryologists.

Analyzing the IVF data under the EU assumption implies having to deal with a *partial observability* problem. For instance, if pregnancy does *not* occur, it cannot be ascertained whether a) the uterus was *non-receptive*, b) *all* the transferred embryos were *non-viable* or c) both. If pregnancy occurs, the embryo is known to be receptive, but it is still unknown which of the embryos gave rise to the pregnancy, unless the number of babies equals the number of transferred embryos. The partial observability problem is addressed by (Zhou and Weinberg, 1998) adopting a latent variable formulation, and then estimating the parameters of the model via Expectation-Maximization (EM).

The contributions of this paper are as follows: a) a novel probabilistic model of IVF transfers, based on a Bayesian network; b) a simple but effective averaging approach to improve the estimate of the parameters from the incomplete samples which characterize the problem; c) a thorough experimental analysis showing good predictive performance of the proposed model on both artificial and real data sets.

2 The model

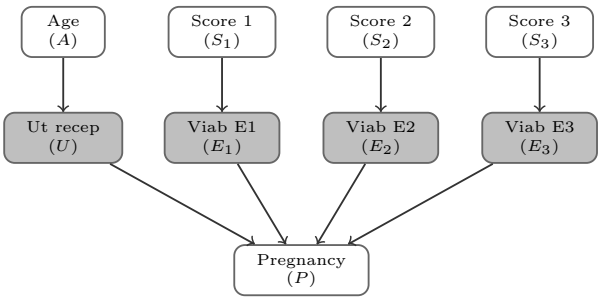

Figure 1: The IVF model: nodes affected by the missingness process are shown with a gray background.

Given a generic variable X, we denote by θ_X the probability mass function which associates a marginal probability to each different value of X; we denote by $\theta_{X|Pa(X)}$ the probability mass function which associates a conditional probability to each different value of X, given each possible configuration of the parents of X, denoted as $Pa(X)$. We moreover denote by $\boldsymbol{\theta}$ the set of all the parameters of the BN model.

We represent the IVF transfer by the BN model shown in Fig. 1. The model manages IVF cycles with up to three embryos, as this is the maximum allowed under the Swiss law; however, it can be straightforwardly extended to manage a higher number of transferred embryos. The woman age is discretized as {<34, 34-40, 40+}.

We denote by $\mathcal{S}$ the set of nodes $\{S_1, S_2, S_3\}$ which represent the score of the embryos and which are referred to as the $\mathcal{S}$-nodes. The score can be *top* of *non-top* depending on the morphology of the embryo, according to criteria which are out of our scope. Moreover, embryos graded as *top* consistently on all the days of the culture are scored as *top-history*. The score of the embryos is also allowed to be *no-transfer*, thus allowing to model cycles with less than 3 transferred embryo: since multiple pregnancies are dangerous for the health of both mother and babies, often only 1 or 2 embryos are transferred. Summing up, the $\mathcal{S}$-nodes take values in {*no-transfer, non-top, top, top-history*}. The embryos occupy the different positions (1,2,3)

in a purely random fashion.

The $\mathcal{S}$-nodes are *tied*: they share the same mass function θ_S instead of having separate mass functions θ_{S_1}, θ_{S_2} and θ_{S_3}. This prevents the same embryo score (e.g., top) being given a different marginal probability depending on whether one refers to node S_1, S_2 or S_3. Moreover, sharing the parameters allows to obtain more accurate estimates than using a different mass function for each $\mathcal{S}$-node. Node U represents uterine receptivity; it is therefore binary, with states $(u, \neg u)$. We denote by $\mathcal{E}$ the set of nodes $\{E_1, E_3, E_3\}$, which are referred to in the following as $\mathcal{E}$-nodes. Each $\mathcal{E}$-node represents the viability of a different embryo; it is thus binary with states $(e, \neg e)$. The $\mathcal{E}$-nodes share the parameter set of the conditional mass function $\theta_{E|S}$, rather than having independent mass functions $\theta_{E|S_1}$, $\theta_{E|S_2}$ and $\theta_{E|S_3}$. Again, this prevents two embryos with the same score being given different probability of being viable just because they occupy a (random) different position; moreover, it enables more accurate estimate than using a different set of parameters for each $\mathcal{E}$-node.

The pregnancy node P has four states $\{0, 1, 2, 3\}$, corresponding to the number of babies which might be born after having transferred up to three embryos. The CPT of P encodes the deterministic EU assumption, namely it assigns probability 1 to the state whose value equals $(U = u) * \sum_{i=1}^{i=3}(E_i = e)$.

In the following, we discuss the incompleteness which characterizes the instances. Node P represents the class; it is thus always observed in training and always missing in test. The missingness process affecting the U and the $\mathcal{E}$-nodes is instead more complicated. Let us consider an IVF cycle in which all the 3 embryos are transferred. Let us start by the *training stage*; recall that at training stage P is always observed. Given $P = 0$, the observation of node U is missing; given $P > 0$, U is observed $(U = u)$. Given $P < 3$, the observation of the $\mathcal{E}$-nodes is missing; given $P = 3$, the $\mathcal{E}$-nodes are observed $(E_i = e \; \forall \; i)$. Thus, the observation of U and the $\mathcal{E}$-nodes is missing or not depending on the value of the observed variable P; the missingness process is MAR (missing at random). Since most IVF cycles result in no-pregnancy, in most instances both U and the $\mathcal{E}$-nodes are *not* observed.

At *test* stage, the parameters have been already learned and the goal is to assess the predictive ability of the model. The U and the $\mathcal{S}$-nodes are *never* observed at test stage (if they were observed at test stage, the outcome would be known with certainty); they are therefore affected by a MCAR (missing completely at random) missingness process. For more details about MAR and MCAR missing data, see (Koller and Friedman, 2009, Chap.19).

In some cycles less than three embryos are transferred, to reduce the danger of multiple pregnancy. For these cycles, the missingness process affects the $\mathcal{E}_t$-nodes rather than the $\mathcal{E}$-nodes, the $\mathcal{E}_t$-nodes representing the viability of the *transferred* embryos. The $\mathcal{E}_t$-nodes are affected by a MAR and a MCAR missingness process at respectively train and test stage. The viability of non-transferred embryos is alway observed (as $\neg e$), since a non-transferred embryo is by definition non-viable.

3 Estimation procedure

Given a generic variable X, we denote by θ_X^x the probability $P(X = x)$ and by $\theta_{X|Y}^{x|y}$ the probability $P(X = x|Y = y)$; this additional notation allows accessing singletons of the mass functions. We denote as $\mathcal{X}$ the set of all variables which constitute the BN model and by $\boldsymbol{x}$ an *instance*, namely a single row of data containing either an observation or a missing value for each variable.

Because of the incomplete samples, the likelihood contains the summation over all the possible data completions. As an example, consider the following instance $\boldsymbol{x}$ of the training set:

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

in which a single pregnancy has occurred; thus, the uterus is known to be receptive (U=u) but the observations of the $\mathcal{E}$-nodes is missing,

since it is unknown which of the three embryos has implanted. The possible data completions are those in which exactly one out of three embryos is viable; the likelihood of the instance is thus:

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \theta_A^{40+} \cdot \theta_U^u \cdot \theta_S^{top} \cdot \theta_S^{ntop} \cdot \theta_S^{toph} \cdot$$
$$\cdot[\theta_{P|U,\mathcal{E}}^{1|u,e_1,\neg e_2,\neg e_3} \cdot \theta_{E|S}^{e|top} \cdot \theta_{E|S}^{\neg e|ntop} \cdot \theta_{E|S}^{\neg e|toph}$$
$$+\theta_{P|U,\mathcal{E}}^{1|u,\neg e_1,e_2,\neg e_3} \cdot \theta_{E|S}^{\neg e|top} \cdot \theta_{E|S}^{e|ntop} \cdot \theta_{E|S}^{\neg e|toph}$$
$$+\theta_{P|U,\mathcal{E}}^{1|u,\neg e_1,\neg e_2,e_3} \cdot \theta_{E|S}^{\neg e|top} \cdot \theta_{E|S}^{\neg e|ntop} \cdot \theta_{E|S}^{e|toph}]$$

Notice that the likelihood contains terms θ_S^{top}, θ_S^{ntop} and θ_S^{toph} rather than $\theta_{S_1}^{top}$, $\theta_{S_2}^{ntop}$ and $\theta_{S_3}^{toph}$, as a consequence of the $\mathcal{S}$-nodes sharing the same mass function. The same consideration applies to the $\mathcal{E}$-nodes; in the likelihood it appears e.g. $\theta_{E|S}^{\neg e|top}$ rather than $\theta_{E_1|S}^{\neg e|top}$.

The log-likelihood for the whole training set is obtained by summing the logs of the likelihood of all instances; it has a complex expression, which would be very difficult to analytically optimize. However, recalling that the missingness process at the training stage is MAR, the Expectation-Maximization algorithm (EM) can be used to learn the parameters. We estimate the parameters by maximizing the posterior probability of the data $P(\boldsymbol{\theta}|D)$ rather than the likelihood; this approach improves the estimates and reduces the danger of overfitting. In particular, we adopt a Dirichlet prior for for the parameters, setting to 1 the equivalent sample size. In the following, we refer to $P(\boldsymbol{\theta}|D)$ as the *MAP score*.

Given the multi-modality of $P(\boldsymbol{\theta}|D)$, EM converges only to a *local* maximum of the MAP score. It is thus common to initialize EM from m different starting points and to execute m different *EM runs*, to eventually select the estimate yielding the highest MAP score. In the following, we refer this procedure as *MAP estimation*. By definition, MAP estimation selects the parameterization which is the most probable a posteriori, rather than integrating over the full posterior: for this reason, it "*does not offer the same benefits as a full Bayesian estimation. It does not attempt to represent the shape of the posterior and thus does not differentiate between a flat posterior and a sharply peaked one.*" (Koller and Friedman, 2009, Sec.17.4.4). In particular, MAP estimation is a good approximation of Bayesian estimation when the posterior is sharply peaked around the maximum; this is however not the case when learning from incomplete samples. Typically, different EM runs achieve close values of the MAP score, returning however very different parameter estimates. Therefore, MAP estimation in this context is hardly robust. This consideration remain valid even if informative starting point are adopted for EM, which allows improving the estimates.

As an alternative to MAP estimation, we propose the following averaging approach. With reference to a generic parameter θ_X^x, we average its estimates obtained in the m EM runs as follows:

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

where $\hat{\theta}_X^{x-i}$ and $P(\hat{\boldsymbol{\theta}}^{\boldsymbol{i}}|D)$ denote respectively the estimate of θ_X^x and the MAP score obtained in the i-th EM run, once it has converged. We average all parameters of the model according to Equation (1).

To illustrate the rationale of our approach, consider the general query $P(\mathcal{Z}|\boldsymbol{y}, D)$, where $\mathcal{Z}$ is the set of variables being queried, and $\boldsymbol{y}$ is the evidence available on the subset of variables $\mathcal{Y} \in \mathcal{X}$. A fully Bayesian inference would be:

$$P(\mathcal{Z}|\boldsymbol{y}, D) = \int P(\mathcal{Z}|\boldsymbol{y}, D, \boldsymbol{\theta}) P(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \quad (2)$$

while, under MAP estimation, the above integral is roughly approximated as:

$$P(\mathcal{Z}|\boldsymbol{y}, D) \approx P(\mathcal{Z}|\boldsymbol{y}, D, \hat{\boldsymbol{\theta}}) \quad (3)$$

where $\hat{\boldsymbol{\theta}}$ represents the most probable parameters estimate a posteriori.

The following *pseudo-Bayesian* approach moves towards the Bayesian inference, by sam-

pling the posterior in correspondence of the local maxima identified by the m EM runs:

$$P(\mathcal{Z}|\boldsymbol{y}, D) \simeq \sum_{i=1}^{i=m} P(\mathcal{Z}|\boldsymbol{y}, D, \hat{\boldsymbol{\theta}}^i) P(\hat{\boldsymbol{\theta}}^i|D) \quad (4)$$

The pseudo-Bayesian approach should generate more accurate inferences than the MAP approach, because it partially reconstructs the shape of the posterior. A similar idea has been for instance used with good results in clustering with naive Bayes (Santafé et al., 2006). Yet, it requires keeping a collection of e.g. m=20 networks, each characterized by the same structure but different parameters; this compromises the possibility for IVF experts of easily interpreting the model.

The averaging idea of Equation (1) aims at keeping as much as possible the benefits of the pseudo-Bayesian approach, but instantiating only a single network. In particular, averaging the parameters according to Eq. (1) produces the same inferences of the pseudo-Bayesian approach of Eq. (4), in case of a query of type $P(X = x|pa(X))$, where $pa(X)$ denotes an instantiation of all the parents of X. In these cases, the returned inference correspond in fact to the parameter $\theta^{x|pa(X)}_{X|Pa(X)}$ of the network; averaging the parameters according to Eq. (1) is equivalent to averaging the inferences according to Eq. (4). This is especially important for our application, in which IVF experts are interested in analyzing the estimates of the conditional mass functions $\theta_{E|S}$ and $\theta_{U|A}$.

However, a single network with parameters averaged according to Eq.(1) does not yield the same inferences than the pseudo-Bayesian approach of Eq.(4) in more general queries. In general, it is not possible replicating by a single network the inference produced by a set of networks. Moreover, the property of parameter decomposability, which allows estimating independently the different conditional probability mass function, does not hold if the training set is incomplete (Koller and Friedman, 2009, Chap. 19.1.3). This prevents in principle averaging the parameters referring to the same conditional mass functions across the different EM runs.

Yet the experiments of the next section show that the averaging approach consistently outperforms MAP estimation, both as for the accuracy of the parameters estimates and of the predictive inferences: the benefits of going towards Bayesian estimation outweigh the effects of the introduced approximations.

4 Experiments with generated data

We perform experiments with generated data, in order to compare the estimates generated by MAP estimation and by the averaging approach. We consider the network structure of Figure 1 and the sample sizes $n \in \{50, 150, 300, 450, 600\}$. For each sample size n, we perform 100 *experiments* constituted by the following steps: a) random drawing of the parameters of the structure, thus instantiating the *true network*; b) sampling of n complete instances (*training set*) from the true network; c) application of the MAR missingness process of the training stage, described in Section 2; d) execution of EM from m=20 different initializations and estimation of the parameters adopting the MAP and the averaging approach; e) evaluation of the estimated parameters; f) generation of a test set of 1000 instances, making missing the U and the $\mathcal{E}$-nodes; g) classification of the test instances. On average, in the simulations U is missing in 75% of the instances and the $\mathcal{E}$-nodes in 80% of the instances.

Since the problem has 4 classes, we measure the classification performance by computing 4 AUCs: one for each of no-pregnancy, single, double and triple pregnancy; they are denoted as $AUC_0, AUC_1, AUC_2, AUC_3$. For this problem, AUC is a more appropriate measure than accuracy: typically, some 70-80% of the cycles ends with non-pregnancy and thus a *trivial* predictor which always returns no-pregnancy would achieve an *apparently* high accuracy of 70-80%. Previous studies in this area (Saith et al., 1998; Morales et al., 2008) show that even sophisticated classifiers only achieve a small improvement of accuracy over the trivial predictor, whose performance is inflated by the skew in the

class distribution. The AUC is insensitive on the class distribution and thus allows assessing more clearly the performance of the classifiers.

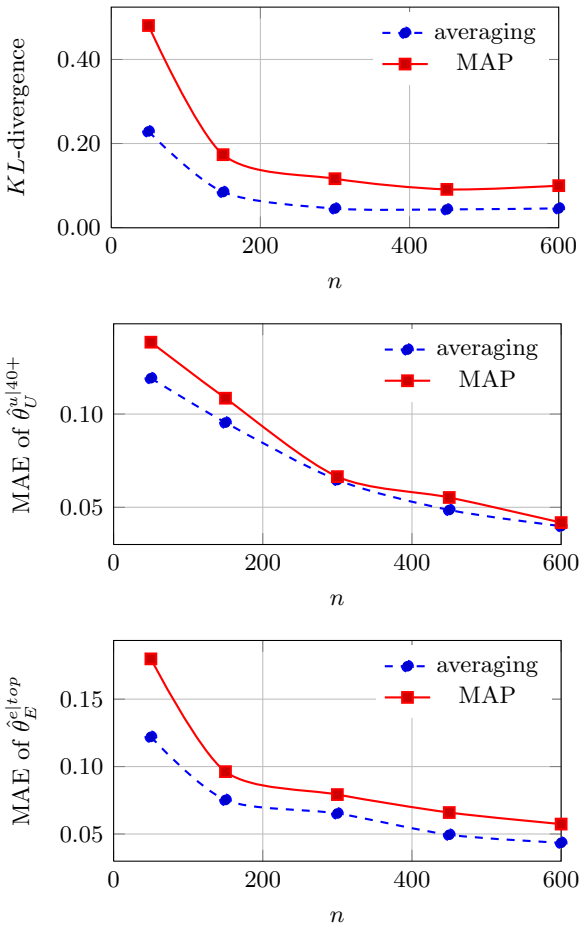

Figure 2: Experimental results on simulated data; the averaging approach *reduces* the KL-divergence from the true model and thus the mean absolute error (MAE) in the estimation of the parameters. Each point represents the average over 100 experiments.

The averaging approach largely reduces, compared to the MAP estimation, both the mean and the standard deviation of the KL-divergence from the true network, as shown in Figure 2. The reduction of the KL-divergence is significant at *each* sample size (t-test, p<0.01). For instance the mean KL-divergences are for n=50: 0.22±0.12 for the averaging approach and 0.48±0.35 for MAP estimation; for n=600: 0.03±0.05 and 0.05±0.09. Thus even for large

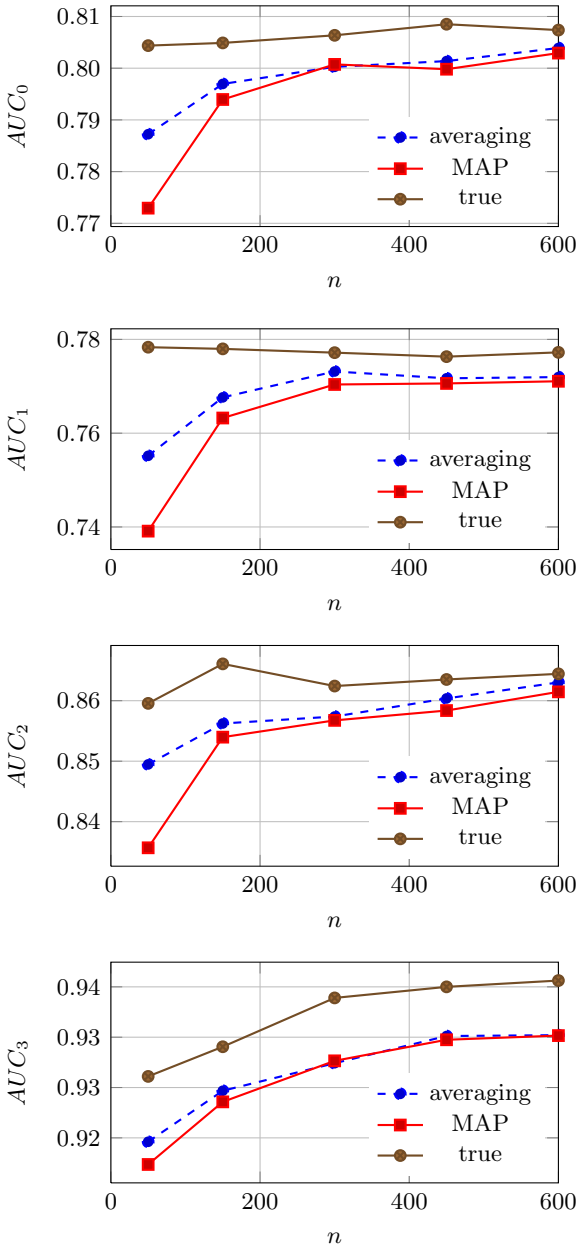

Figure 3: Experimental results on simulated data; the averaging approach *increases* the AUCs compared to MAP estimation. Each point represents the average over 100 experiments.

n, the averaging approach reduces of about 40% the mean KL-divergence, although this is hard to grasp from Figure 2, because of its scale. The reduction in the KL-divergence implies a much better estimate of the conditional mass func-

tions $\theta_{E|S}$ and $\theta_{U|A}$. The lower plots of Figure 2 show the reduction of the mean absolute error in the estimation of two (randomly chosen) parameters taken from $\theta_{E|S}$ and $\theta_{U|A}$.

The improved estimates result in better predictive inferences about the probability of pregnancy, as measured on the test sets. The averaging approach improves all AUCs over the MAP estimation, as shown in Figure 3; the improvement is however smaller in percentage than on the parameter estimates. Considering 4 different AUCs and 5 sample sizes, there are 20 possible combinations n-AUC; in 7 out of such 20 cases the AUC improvement yielded by the averaging approach is significant (t-test, $p<0.01$).

Another interesting finding is that the AUC of the *true* model is generally not very far from that of the *estimated* models: the average AUC (averaging $AUC_0, AUC_1, AUC_2, AUC_3$ over all experiments) is 83.4, 83.7 and 84.5 for respectively the MAP, the averaging approach and true model. The point is that at test stage, as already discussed, U and the $\mathcal{E}$-nodes are never observed; this is the main difficulty of predicting the outcome of IVF cycles, under the EU assumption. There is thus a major limit on predictive performance, which holds also if the model parameters are perfectly known.

4.1 Analysis of a a real data set

We analyze 388 cycles performed at the International Institute for Reproductive Medicine (IIRM) of Lugano. The percentage of no pregnancy, single pregnancy and twin pregnancy is respectively 80%, 16% and 4%; no triple pregnancies are present.

It is not possible measuring the KL-divergence of the estimated networks from the true model, which is indeed unknown. However, we exploit cross-validation to provide some results about the quality of the parameter estimates. By adopting a 5-folds cross-validation, 4/5 of the instances (310) are used to learn the model, and the remaining ones to test the predictions. We consider as a *reference model* the BN model learned on the full data set of 388 instances. We repeat 10 times the cross-validation; this yields a sample of 50 (5 folds $\cdot$ 10 repetitions) KL-divergences between the models estimated on the sampled training sets and the reference model. We stratify training and test sets, which contain the same proportion of no-pregnancy, single pregnancies and twin pregnancy of the complete data set. The averaging approach reduces the mean KL-divergence from the reference model of about 5% compared to MAP estimation; this reduction is significant (t-test, $p<0.01$); moreover, it also reduces the standard deviation of about 15%. The advantage is larger if a smaller training set is available; repeating the same procedure with a 2-folds cross-validation (194 instances in the sampled training set), the averaging approach reduces the mean KL-divergence of about 42% and the standard deviation of about 75%. To avoid any bias in favor of the averaging approach, the reference model has been estimated using the MAP approach.

We report in the *first* column of Table 1 the AUCs measured by 5-folds cross-validation, adopting the averaging approach (the model is indicated as BN-EU); they are *not* significantly different from those obtained using MAP estimation. In fact, the training set is quite large, and estimation methods tend to converge as the sample size increases; moreover, as shown in the previous section, the averaging approach yields larger improvements on the parameter estimates than on the AUC.

According to the estimates of the BN-EU model (under the averaging approach), uterine receptivity drops from 78% to 58% and eventually 26% for woman aged respectively {<34, 34-40, 40+}; moreover, embryo viability increases from 7% to 21% to 39% for embryos scored respectively as non-top, top and top-history. These findings show that embryo viability is generally a more critical factor than uterine receptivity, as it is generally accepted in the IVF literature; moreover they show a clear pattern of embryo viability as a function of the score, and of uterine receptivity as a function of the woman age.

As a last finding we present in Table 1 also the AUCs obtained by Bayesian network classifiers such as AODE, TAN and naive Bayes (NB); a

	BN-EU	AODE	TAN	NB
AUC_0	74.1	74.8	73.0	**75.2**
AUC_1	67.0	68.0	65.1	**68.4**
AUC_2	**83.4**	81.6	79.6	80.1

Table 1: Comparison of different classifiers over the IIRM data set (average over 10 runs of 5-folds cross-validation).

presentation of these models can be found in (Webb et al., 2005). All these classifiers are induced on a data set which contains the same information provided to model BN-EU; however, the data set is complete as it does not model the EU assumption. In particular, it contains the following features: age of the woman; total number of transferred embryos; the number of non-top, top, and top-history embryos transferred. According to the terminology of (Roberts, 2007), such a data set is aggregated at the *recipient level*, rather the modeling the interaction between uterus and embryos. Despite such classifiers being learned on a complete data set, theirs AUCs are not significantly better than those of the BN-EU model. However, the BN-EU model is more interpretable and provides more insights to IVF experts than a traditional classifier. We thus share the viewpoint of (Roberts, 2007): "*Most clinical studies of embryo-level factors avoid the partial observability problem by either considering only patients with single embryo transfer, or by using a recipient-level aggregated measure. Modelling approaches which correctly incorporate the structure of the data are to be preferred both for the analysis of studies of embryo-level viability predictors, and in order to derive parameters which can be used in the utilization of such procedures.*'

5 Conclusions

We have proposed a novel BN model for modelling IVF transfer and a simple but effective averaging approach for improving both parameter estimates and predictive inferences. As a further step we plan to study, both from a medical and a statistical viewpoint, further covariates as parents of both U and the $\mathcal{E}$-nodes.

Acknowledgments

Research partially supported by CTI (Commission for Technology and Innovation) grant n. 9707.1 PFSL-LS, Swiss NSF grant no. 200020-132252 and the Hasler foundation grant n. 10030.

References

D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

D.A. Morales, E. Bengoetxea, P. Larrañaga, M. García, Y. Franco, M. Fresnada, and M. Merino. 2008. Bayesian classification for the selection of in vitro human embryos using morphological and clinical data. *Computer Methods and Programs in Biomedicine*, 90(2):104–116.

SA Roberts, WM Hirst, DR Brison, A. Vail, et al. 2010. Embryo and uterine influences on IVF outcomes: an analysis of a UK multi-centre cohort. *Human Reproduction*.

Stephen A. Roberts. 2007. Models for assisted conception data with embryo-specific covariates. *Statistics in Medicine*, 26(1):156–170.

RR Saith, A. Srinivasan, D. Michie, and IL Sargent. 1998. Relationships between the developmental potential of human in-vitro fertilization embryos and features describing the embryo, oocyte and follicle. *Human Reproduction Update*, 4(2):121–134.

G. Santafé, J.A. Lozano, and P. Larrañaga. 2006. Bayesian model averaging of naive bayes for clustering. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(5):1149–1161.

AL Speirs, A. Lopata, MJ Gronow, GN Kellow, and WI Johnston. 1983. Analysis of the benefits and risks of multiple embryo transfer. *Fertility and sterility*, 39(4):468.

G.I. Webb, J.R. Boughton, and Z. Wang. 2005. Not so naive bayes: Aggregating one-dependence estimators. *Machine Learning*, 58(1):5–24.

H. Zhou and C.R. Weinberg. 1998. Evaluating effects of exposures on embryo viability and uterine receptivity in vitro fertilization. *Statistics in medicine*, 17(14):1601–1612.