Aggregation and Population Growth: The Relational Logistic Regression and Markov Logic Cases

David Poole, David Buchman, Sriraam Natarajan, and Kristian Kersting
cs.ubc.ca/~poole/ cs.ubc.ca/~davidbuc/
tsi.wfubmc.edu/fac/snataraj/ www-kd.iai.uni-bonn.de/people.php?kristian.kersting

Abstract

This paper considers how relational probabilistic models adapt to population size. First we show that what are arbitrary choices for non-relational domains become a commitment to how a relational model adapts to population change. We show how this manifests in a directed model where the conditional probabilities are represented using the logistic function, and show why it needs to be extended to a relational logistic function. Second we prove that directed aggregation models cannot be represented by Markov Logic without clauses that involve multiple individuals. Third we show how these models change as a function of population size.

1 Introduction

Relational probabilistic models are characterized by having models that are specified independently of the actual individuals, and where the individuals are exchangeable; before we know anything about the individuals they are treated identically. One of the features of relational probabilistic models is that the predictions of the model depends on the number of individuals (the **population size**). Sometimes, this dependence is desirable. In other cases, the numbers may need to change [Jain et al., 2007, 2010]. In either case, it is important to understand how the predictions change with population size.

Varying population sizes are actually quite common. They can appear in a number of ways including:

- The actual population may be arbitrary. For example, in considering the probability of someone committing a crime (which depends on how many other people could have committed the crime) [Poole, 2003] we could consider the population to be the population of the neighbourhood, the population of the city, the population of the country, or the population of the whole world. It would be good to have a model that does not depend on this arbitrary decision. We would like to be able to compare models where the modelers have made different choices.
- The population can change. For example, the number of people in a neighbourhood or in a school class may change. We would like a model to make reasonable predictions as the population changes. We would also like to be able to apply a model learned at one population size to a different population size. For example, models from drug studies are acquired from very limited populations but are applied much more generally.
- The relevant populations can change from one individual to another. For example, the happiness of a person may depend on how many of her friends are kind (and how many are not kind). We would like a model that makes reasonable predictions for a diverse number of friends.

In the following, we start with a simple model, namely a directed model where the conditional probabilities represent a logistic regression model and show how the population growth of this well-studied model is actually not well understood. This suggests a parametrization for such models that takes the population growth into account. We compare the resulting model to Markov logic parametrizations of the same model, and show that the Markov logic models cannot compactly represent the simple directed model. Finally we show how both models adapt to changing population size.

2 Some Basic Definitions

A **population** is a set of **individuals**. A population corresponds to a domain in logic. The **population size** is the cardinality of the population which can be any non-negative integer. For the examples below, where there is a single population, we write the population as $A_1 \ldots A_n$, where n is the population size.

A **parameter**, which corresponds to a logical variable, is

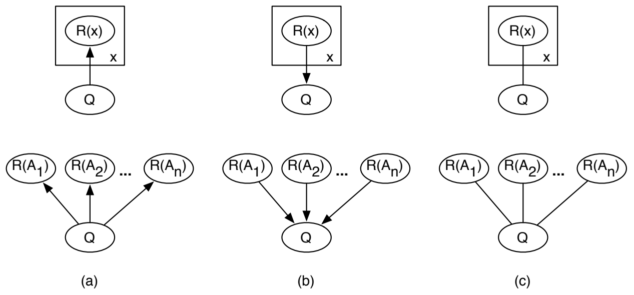

Figure 1: Running example as (a) naïve Bayes (b) logistic regression (with independent priors for each $R(x)$) and (c) Markov network. On the top are the parametrized networks, and on the bottom are the the groundings for the population $\{A_1, A_2, \ldots, A_n\}$

written in lower case. Parameters are typed with a population; if x is a parameter of type τ, $pop(x)$ is the population associated with x and $|x| = |\tau| = |pop(x)|$. We assume that the populations are disjoint (and so the types are mutually exclusive). Constants are written starting with an upper case letter.

A **parametrized random variable (PRV)** is of the form $F(t_1, \ldots, t_k)$ where F is a k-ary functor (a function symbol or a predicate) and each t_i is a parameter or a constant. Each functor has a range, which is $\{True, False\}$ for predicate symbols. A parametrized random variable represents a set of random variables, one for each assignment of an individual to a parameter. The range of the functor becomes the range of each random variable.

A **grounding** of a model with respect to a population for each parameter is a model created by replicating each PRV for each individual in the domain of each parameter, and preserving the structure.

3 Representing Conditional Probabilities

Suppose Boolean parametrized random variable Q is connected to Boolean parametrized random variable $R(x)$, which contains an extra logical variable, x. In the grounding, Q is connected to n instances of $R(x)$, where n is the population size. Ideally, we define the model before we know n; it should be applicable for all values of n.

For this situation, a directed model where $R(x)$ is a child of Q is shown in Fig. 1 (a). It produces a naïve Bayesian model in the grounding with separate factors for Q and for each individual. An undirected model with a potential for Q and a pairwise potential for each factor is shown in Fig. 1 (c). In both these models the joint probability is the product of factors.

For a directed model with $R(x)$ as a parent of Q (Fig. 1 (b)), the variable Q has a unbounded number of parents in the grounding, so we need some way to aggregate the parents. Common ways to aggregate in relational domains, e.g. [Horsch and Poole, 1990; Friedman et al., 1999; Neville et al., 2005; Perlich and Provost, 2006; Natarajan et al., 2010], include logical operators such a *noisy-or*, *noisy-and*, as well as ways to combine probabilities. This requirement for aggregation occurs in a directed model whenever a parent contains an extra logical variable.

3.1 Relational Logistic Regression

Consider a situation in which all the variables $R(x)$ are observed, and we only wish to model the conditional probability $P(Q|R1, \ldots, R_n)$. These conditional probabilities for both the naïve Bayesian model (Fig. 1 (a)) and the Markov model (Fig. 1 (c)) have the logistic regression form. To see this, notice that both can be expressed by a product of non-negative factors:

$$P(Q, R(A_1), \ldots, R(A_n)) \propto \prod_i f(Q, R(A_i)) \times g(Q)$$

where R_i is $R(A_i)$ for some enumeration $A_1, \ldots, A_n$ of the population. We can now choose a particular value q for Q and write $\neg q$ as the negation of the assignment q:

$$\begin{aligned} P(q|R_1, \ldots, R_n) &= \frac{P(q, R_1, \ldots, R_n)}{P(q, R_1, \ldots, R_n) + P(\neg q, R_1, \ldots, R_n)} \\ &= \frac{1}{1 + \frac{P(\neg q, R_1, \ldots, R_n)}{P(q, R_1, \ldots, R_n)}} \\ &= \frac{1}{1 + 1/\left(\prod_i f(q, R_i)/f(\neg q, R_i) \times g(q)/g(\neg q)\right)} \\ &= \frac{1}{1 + e^{-\sum_i \hat{f}(R_i) + \hat{g}}} \end{aligned}$$

where $\hat{f}(R_i) = \log(f(q, R_i)/f(\neg q, R_i))$ and $\hat{g} = \log(g(q)/g(\neg q))$. This last step is only valid if all potentials are positive (contain no zeros). Assume some numerical representation for the two values of each R_i was chosen (e.g., $\{0, 1\}$ or $\{-1, 1\}$). For fixed n, we can always find values for w_0 and w_i, which depend on the numerical representation chosen, such that $\sum_i \hat{f}(R_i) + \hat{g} = w_0 + \sum_i w_i R_i$. Thus:

$$P(q|R_1, \ldots, R_n) = \text{sigmoid}(w_0 + \sum_i w_i R_i)$$

where

$$\text{sigmoid}(x) = 1/(1 + e^{-x}) .$$

$P(q|R_1, \ldots, R_n) > 0.5$ iff $w_0 + \sum_i w_i R_i > 0$. The space of assignments to the w_i so that $w_0 + \sum_i w_i R_i = 0$ is called the **decision threshold**, as it is the boundary of where $P(q \mid R_1, \ldots, R_n)$ changes between being closer to 0 and being closer to 1.

For a relational model where the individuals are exchangeable, w_i must be identical for all variables R_i, so:

$$P(q|R_1,\ldots,R_n) = \text{sigmoid}(w_0 + w_1 \sum_i R_i). \tag{1}$$

Consider what happens with a relational model, in which n can vary.

Example 1. Suppose we want to represent "Q is true if and only if R is true for 5 or more individuals" (i.e., $q \equiv |\{i \mid R_i = true\}| \geq 5$) using a logistic regression model $(P(q) > 0.5) \equiv (w_0 + w_1 \sum_i R_i \geq 0)$, which we fit for a population of 10. Consider what this model represents when the population size is 20.

If the values of R are represented with $false \rightarrow 0$ and $true \rightarrow 1$, this model will have Q true if R is true for 5 or more individuals out of a population of 20. It is easy to see this, as $\sum_i R_i$ only depends on the number of positive individuals.

However, if the values of R are represented with $false \rightarrow -1$ and $true \rightarrow 1$, this model will have Q true if R is true for 10 or more individuals out of a population of 20. The sum $\sum_i R_i$ depends on how many more individuals have R true than have R false.

If the values of R are represented with $false \rightarrow 1$ and $true \rightarrow 0$, then this model will have Q true if R is true for 15 or more individuals out of a population of 20. The sum $\sum_i R_i$ depends on how many individuals have R false.

Other parametrizations can result in different decision thresholds.

The following table gives some possible parameter settings as a function of the numerical representation of $false$ and $true$, which represent the same conditional distribution for $n = 10$, and the corresponding prediction for a population size of 20:

false	**true**	$\mathbf{w_0}$	$\mathbf{w_1}$	**Prediction for n = 20**
0	1	-4.5	1	$Q \equiv \lvert\{R_i = true\}\rvert \geq 5$
-1	1	0.5	1	$Q \equiv \lvert\{R_i = true\}\rvert \geq 10$
-1	0	5.5	1	$Q \equiv \lvert\{R_i = true\}\rvert \geq 15$
-1	2	-4.5	1	$Q \equiv \lvert\{R_i = true\}\rvert \geq 8$
1	2	-14.5	1	$Q \equiv \lvert\{R_i = true\}\rvert \geq 0$

All of these are linear functions of population size for fixed numbers of R true, and linear functions of the number of Rs true for a fixed population. We can prove the following proposition:

Proposition 1. *If false is represented by the number α and true is represented by β, for a fixed w_0/w_1 learned for one population size, the decision threshold for a population of size n is*

$$\frac{w_0}{w_1(\alpha - \beta)} + \frac{\alpha}{\alpha - \beta} n$$

What is important about this proposition is that the way the decision threshold grows with the population size n does not depend on data (which would provide the weights), but on the prior assumptions, which are implicitly encoded into the numerical representation of R.

Thus, (1) with any specific numeric representation of $true$ and $false$ is only able to model one of the dependencies of how predictions depend on population size, and so cannot properly fit data that does not adhere to that dependence.

We need an extra degree of freedom to get a relational model that can model any of the above dependencies on n, regardless of the numerical representation chosen.

Definition 1. Let Q be a Boolean parametrized random variable with a single parent $R(x)$, where x is the set of logical variables in R that are not in Q (so we need to aggregate over the values of x). A (single parent) **relational logistic function (RLF)** for Q with parents $R(x)$ is of the form:

$$P(q|R(A_1),\ldots,R(A_n)) = \\ \text{sigmoid}(w_0 + w_1 \sum_i R_i + w_2 \sum_i (1 - R_i)) \; , \tag{2}$$

where R_i is 1 if $R(A_i)$ is true, and 0 otherwise. So, $\sum_i R_i$ is the number of individuals for which R is $true$ and $\sum_i (1 - R_i)$ is the number of individuals for which R is $false$.

Changing the parametrization, (e.g., representing $false$ as -1) would result in the weights changing, but not the decision thresholds.

A relational logistic function can represent not only the threshold for any particular n, but also a (linear) function of how the threshold depends on the population.

An alternative but equivalent parametrization can be formulated as

$$P(q|R(A_1),\ldots,R(A_n)) = \\ \text{sigmoid}(w_0 + w_2 \sum_i 1 + w_3 \sum_i R_i)$$

where 1 is a function that has value 1 for every individual, and so $\sum_i 1 = n$. The mapping between these parametrizations is $w_3 = w_1 - w_2$; w_0 and w_2 are the same.

While the dependence on the population may be arbitrary when a single population is observed, it affects the ability of a model to simultaneously predict when multiple populations or subpopulations are observed.

Example 2. Suppose we want to model whether someone being happy depends on the number of their friends that are kind to them. Consider the following three hypotheses:

(a) A person is happy as long as they have 5 or more friends who are kind to them.

$$happy(x) \equiv |\{y : friend(y,x) \wedge kind(y)\}| \geq 5$$

(b) A person is happy if half or more of their friends are kind to them.

$$\begin{aligned} happy(x) \equiv & |\{y : friend(y,x) \wedge kind(y)\}| \\ & \geq |\{y : friend(y,x) \wedge \neg kind(y)\}| \end{aligned}$$

(c) A person is happy as long as fewer than 5 of their friends are not kind to them.

$$happy(x) \equiv |\{y : friend(y,x) \wedge \neg kind(y)\}| < 5$$

These three models coincide for people with 10 friends, but make different predictions for people with 20 friends. Only one of these can be represented using standard logistic regression (which one depends on the representation of true and false).

We can use the following relational logistic function to model these cases:

$$\begin{aligned} & P(happy(x)|par) \\ & \quad = \text{sigmoid}(w_0 + w_1 \sum_y friend(y,x)kind(y) \\ & \qquad\qquad + w_2 \sum_y friend(y,x)(1 - kind(y))) \end{aligned} \tag{3}$$

where par is a complete assignment of $friend$ and $kind$ to the individuals.

To model each of the above three cases, we can set w_0, w_1, and w_2 in (3) as follows:

(a) Let $w_0 = -4.5$, $w_1 = 1$, $w_2 = 0$
(b) Let $w_0 = 0.5$, $w_1 = 1$, $w_2 = -1$
(c) Let $w_0 = 5.5$, $w_1 = 0$, $w_2 = -1$

4 Markov Logic Networks

Markov logic networks (MLNs) [Richardson and Domingos, 2006; Domingos et al., 2008] provide alternative parametrizations for the above example.

Example 3. Consider an MLN representation of the ongoing example, with the following α_i weights and formulae:

$$\begin{array}{ll} \alpha_0 & \neg q \\ \alpha_1 & q \\ \alpha_2 & \neg q \vee \neg r(x) \\ \alpha_3 & \neg q \vee r(x) \\ \alpha_4 & q \vee \neg r(x) \\ \alpha_5 & q \vee r(x) \\ \alpha_6 & \neg r(x) \\ \alpha_7 & r(x) \end{array}$$

where the probability of any world is $\prod_{f_i} e^{\alpha_i} = e^{\sum_{f_i} \alpha_i}$ for all ground formulae f_i true in the world. This MLN can be used to represent the same model as Equation (2) when $R(x)$ is observed for all x.

In particular, if obs is an observation for which R is true for k individuals and false for the remaining $n-k$ individuals:

$$\begin{aligned} & P(q|obs) \\ & \quad = \text{sigmoid}(\alpha_1 - \alpha_0 + (\alpha_4 - \alpha_2)k + (\alpha_5 - \alpha_3)(n-k)) \ . \end{aligned}$$

Thus, we can create the same conditional distribution as (2) by setting $\alpha_0 = 0$, $\alpha_1 = w_0$, $\alpha_2 = 0$, $\alpha_3 = 0$, $\alpha_4 = w_1$, $\alpha_5 = w_2$. Note that α_6 and α_7 are not required for representing the conditional probability (they cancel out), but can be used to affect $P(r(A_i))$.

5 Representing Distributions

It is well known that the Naïve Bayesian model and the Markov model are equivalent to the logistic function when all of the individuals have R observed. When not all of the R_i are observed, a Naïve Bayesian model ignores the unobserved variables. In the (relational) logistic regression and the Markov logic network representations the unobserved variables are marginalized over. However, the independence assumptions are different between the models [Pearl, 1988]:

- In the logistic regression model (Fig. 1 (b)), the $R(A_i)$ are independent of each other (when Q is not observed), and the $R(A_i)$ are dependent given Q. Thus, for logistic regression $P(R(A_1)|R(A_2)) = P(R(A_1))$.
- In a Markov network and in Naive Bayes (Fig. 1 (a) and (c)), the $R(A_i)$ depend on each other when Q is not observed, and independent given Q. Thus, for Markov network and Naive Bayes $P(R(A_1)|R(A_2), Q) = P(R(A_1)|Q)$.

This does not mean that the models cannot represent each other. In particular, if we marry the parents in a Bayesian network, we can treat the resulting graph as a Markov network. Essentially we can enforce the independence by the parametrization. Given a Markov model we can create a Bayesian network by, for every factor f_i on variables $X_1, \ldots, X_k$ in the Markov network, introducing a variable T_i, and creating $P(T_i|X_1, \ldots, X_k)$, where the probability for $T_i = true$ corresponds to the values for the Markov network (suitably scaled) and then conditioning on $T_i = true$.

Unfortunately, it is not obvious that this can be done for models where n varies, as marrying the parents creates a factor of unbounded size. In the next section we show that Markov logic indeed cannot represent the same distribution as the directed model without introducing factors among the individuals.

5.1 Markov Logic Networks

While MLNs can represent any (ground) directed model (without aggregation) by representing the conditional probability tables as factors [Domingos et al., 2008] and they can represent various aggregation models [Natarajan et al., 2010], they cannot compactly represent even the directed model of Fig. 1 (b). The reason is that the MLNs make the variables $R(A_i)$ interdependent. This can be fixed only by introducing factors between the $R(A_i)$ variables. We will prove this for any aggregation function, not just variants of logistic regression.

Consider directed aggregation models, characterized by:

$$P(Q,R_1,\dots,R_n) = P(Q \mid R_1,\dots,R_n)\prod_i P(R_i) \qquad (4)$$

i.e. models where the variables $\{R_i\}$ are independent when Q is not observed. These are directed models depicted by Fig. 1 (b), but with arbitrary conditionals $P(Q \mid R_1,\dots,R_n)$, of which the logistic regression model is a special case. We show that such models, where the R_i actually affect Q, cannot be represented using undirected models such as MLNs, using just formulae over Q, over $R(x)$, and over both, without using formulas which combine different R variables. In other words, by using models with no more than a single logical variable in each formula.

Proposition 1. *Let a distribution be characterized by Eq.* (4)*, i.e. in which the variables R_i are independent of each other (when Q is not observed). If $n \geq 2$, and if the distribution is representable by an undirected model which does not contain factors between multiple R_i's, then Q is independent of some parent variable R_i.*

Proof. If the undirected model does not contain factors over multiple R_i variables, then the distribution can be characterized by:

$$P(Q,R_1,\dots,R_n) = \frac{1}{Z} f_1(Q) \prod_i f_{2,i}(R_i) f_{3,i}(Q,R_i)$$

where Z is the normalization constant, which may depend on the population size n.

Define: $g_i(Q,R_i) = f_{2,i}(R_i) f_{3,i}(Q,R_i)$

$$P(Q,R_1,\dots,R_n) = \frac{1}{Z} f_1(Q) \prod_i g_i(Q,R_i) \qquad (5)$$

Define: $\quad f_5(Q) = \frac{1}{Z} f_1(Q) \prod_{i=3}^{n} \sum_{R_i} g_i(Q,R_i)$

$$\begin{aligned}
P(Q,R_1,R_2) &= \sum_{R_3} \cdots \sum_{R_n} P(Q,R_1,\dots,R_n) \\
&= \frac{f_1(Q)}{Z} g_1(Q,R_1) g_2(Q,R_2) \prod_{i=3}^{n} \sum_{R_i} g_i(Q,R_i) \\
&= f_5(Q) g_1(Q,R_1) g_2(Q,R_2) \\
P(R_1,R_2) &= \sum_{Q} f_5(Q) g_1(Q,R_1) g_2(Q,R_2)
\end{aligned}$$

We will use $P_{12}(\cdot,\cdot)$ as a shorthand notation, e.g. $P_{12}(T,F) = P(R_1 = T, R_2 = F)$. We will mark $f_5 = \begin{bmatrix} a \\ b \end{bmatrix}, g_1 = \begin{bmatrix} c & d \\ e & f \end{bmatrix}$ and $g_2 = \begin{bmatrix} c' & d' \\ e' & f' \end{bmatrix}$, i.e. $f_5(F) = a, f_5(T) = b, g_1(F,F) = c, g_1(F,T) = d, g_1(T,F) = e, g_1(T,T) = f$, and similarly for g_2. If the distribution can also be represented by (4), then R_1 and R_2 are independent (when Q is not observed), and therefore:

$$\begin{aligned}
P(R_1,R_2) &= P(R_1)P(R_2) \\
P_{12}(F,F)P_{12}(T,T) &= P_{12}(F,T)P_{12}(T,F) \\
(acc'+bee')(add'+bff') &= (acd'+bef')(adc'+bfe') \\
a^2cc'dd' + abcc'ff' + abdd'ee' + b^2ee'ff' & \\
= a^2cc'dd' + abcd'e'f + abc'def' + b^2ee'ff' & \\
abcc'ff' + abdd'ee' &= abcd'e'f + abc'def' \\
ab(cf-de)(c'f'-d'e') &= 0
\end{aligned}$$

Thus $a = 0,\ b = 0,\ cf = de$ or $c'f' = d'e'$.

In the case that $a = 0$ or $b = 0$, Q is deterministic, and therefore does not depend on any R_i. Consider the case $cf = de$. If $c = 0$, then $d = 0$ or $e = 0$. In this case, either Q is deterministic, and therefore does not depend on any R_i, or R_1 is deterministic, in which case it carries no information, and Q does not depend on it. If $c \neq 0$, then $g_1 = \begin{bmatrix} c & d \\ e & \frac{de}{c} \end{bmatrix}$. It can therefore be decomposed as: $g_1(Q,R_1) = f_6(Q) f_7(R_1)$, with $f_6 = \begin{bmatrix} c \\ e \end{bmatrix}$ and $f_7 = \begin{bmatrix} 1 & \frac{d}{c} \end{bmatrix}$, i.e. $f_7(F) = 1$ and $f_7(T) = \frac{d}{c}$. Substituting into (5) yields:

$$P(Q,R_1,\dots,R_n) = \frac{1}{Z} f_1(Q) f_6(Q) f_7(R_1) \prod_{i \neq 1} g_i(Q,R_i) \qquad (6)$$

Therefore, Q does not depend on R_1. Similarly, the case of $c'f' = d'e'$ leads to Q not depending on R_2. □

Proposition 2. *Let a distribution characterized by (4) be representable using an undirected model which contains no factors between multiple R_i's. Then at most a single variable R_j affects Q, i.e. the distribution can be characterized by:*

$$P(Q,R_1,\dots,R_n) = P(Q \mid R_j) \prod_i P(R_i).$$

Proof. By Proposition 1, Q does not depend on some parent variable R_i. We can disconnect R_i from Q and repeat the argument, until Q has a single parent variable. □

Proposition 3. *Let a distribution characterized by (4) be representable using an MLN with at most a single logical variable in each formula. Then, if $n \geq 2$, then all the variables are independent.*

Proof. Let us notice that an MLN is an undirected model, in which the factors are constrained such that they are identical if we exchange the identities of the objects. An MLN with no more than a single logical variable in each formula corresponds to (5), but in which all factors g_i are equal ($g_i = g$). When repeating the proof of Proposition 1 but with $g_i = g$, the substitution of the decomposition of g into (5) becomes the simplified:

$$P(Q, R_1, \ldots, R_n) = \frac{1}{Z} f_1(Q) \prod_i f_6(Q) f_7(R_i)$$

Hence all variables are independent. □

5.2 Dependence on Population

Given the relational logistic parametrization of Definition 1, we can sum out the unobserved variables. If none of the $R(x)$ are observed, and if p_r is the prior probability of $R(x)$:

$$P(q) = \sum_{i=0}^{n} \binom{n}{i} \text{sigmoid}(w_0 + iw_1 + (n-i)w_2) p_r^i (1-p_r)^{n-i}$$

which is an instance of first-order variable elimination [de Salvo Braz et al., 2007].

To compute $P(q)$ given the MLN parametrization of Example 3, it can be noticed that when Q is conditioned on, the graph is disconnected, with each component having the same probability. So we can compute the probability of one of them and raise it to the power of n [Poole, 2003], giving:

$$P(q) = \text{sigmoid}(\alpha_1 - \alpha_0 + n(\alpha_4 - \alpha_2 + \alpha_5 - \alpha_3 + c)) \quad (7)$$

where

$$c = \log(e^{\alpha_4} + e^{\alpha_5 + \alpha_7 - \alpha_6}) - \log(e^{\alpha_2} + e^{\alpha_3 + \alpha_7 - \alpha_6})$$

What can be noticed about this is that this is a sigmoid function of n, α_0 and α_1, but is not a sigmoid of the other parameters.

We also compare these to a simple mean-field approximation:

$$P(q) = \text{sigmoid}(w_0 + np_r w_1 + n(1-p_r)w_2)$$

where np_r is the expected number of R's true and $n(1-p_r)$ is the expected number of R's false.

Example 4. A plot of the probability of q function of the population size n is given in Fig. 2. The parameters settings are fixed to $w_0 = -4.5$, $w_1 = 1$, $w_2 = -1$, and $p_r = 0.7$. On the x-axis is the population size (n). The solid blue line gives $P(q)$ for relational logistic regression. The dashed red line gives the mean-field approximation sigmoid$(-4.5 + 0.4n)$. The dotted line gives the MLN with $\alpha_7 = 2.82$, chosen to give it the same probability as the relational logistic regression for $n = 1$.

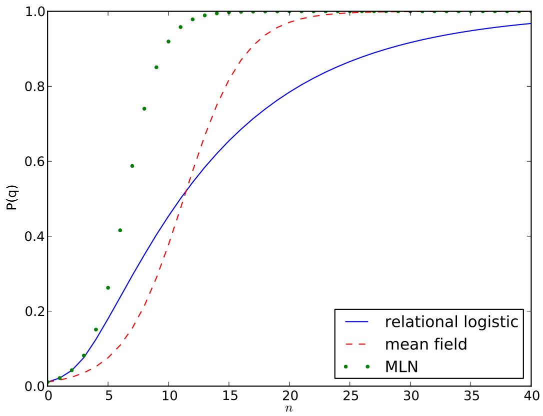

Figure 2: Probability of q as a function of population size for Example 4

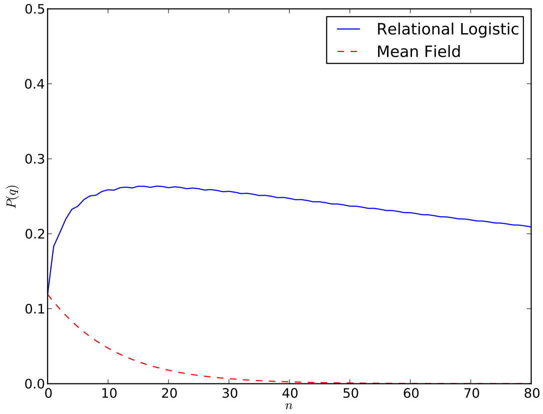

Figure 3: Probability of q as a function of population size in Example 5

It might be conjectured that the MLN representation is qualitatively similar to the relational logistic regressions formulation of the same problem. The probability of q in the MLN as a function of n is a logistic function; the sigmoid of a linear function. The simplest property of the sigmoid is that it is monotonic. Thus, we can make the following conjecture:

Conjecture 1. *The probability of q in the relational logistic function representation of the network in Fig. 1 (b) is monotonic in n.*

It turns out that this conjecture is false.

Example 5. Consider the case: $w_0 = -2$, $w_1 = 2$, $w_2 = -1$, $P(R(x)) = 0.3$. We can plot $P(q)$ as a function of n, shown as the solid blue line in Fig. 3. $P(q)$ is at a maximum when $n = 18$. On the x-axis is the population size (n). The dashed red line gives the mean field approximation, sigmoid$(-2 - 0.1n)$. The growth for an MLN representa-

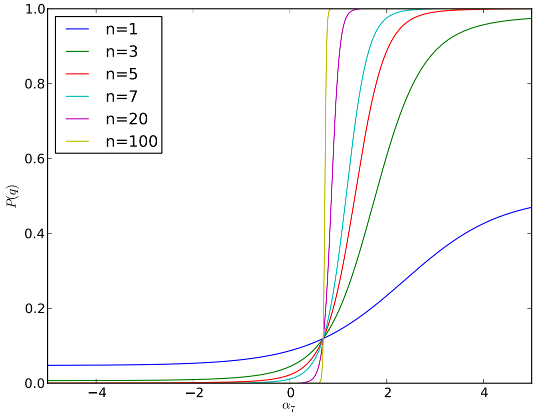

Figure 4: Probability of q as a function of α_7 in an MLN for various population sizes. See Example 6.

tion for this example is given in Example 6.

5.3 Phase Transitions

One of the properties of the relational logistic regression is that $P(R(A_i))$ does not depend on n and can be given as input to the model. Except for the special case of a naive Bayesian model, in MLNs $P(R(A_i))$ is not independent of n. We show that for some MLNs, $P(R(A_i))$ cannot be arbitrarily set in the limit as the population increases.

Example 6. Consider the same parametrization as Example 5, and the mapping to MLNs given in Example 3. Under this mapping, the MLN and the relational logistic regression both represent the same conditional probability of q given an assignment of R to each element of the population. To fully specify the model, the probabilistic relational regression requires p_r, representing $P(r(x))$ for all x. The MLN requires α_6 and α_7. As the model only depends on their difference, we can arbitrarily set $\alpha_6 = 0$.

Fig. 4 shows the probability of q as a function of α_7 for different population sizes. The least steep slope is a population of 1. The other plots are, in order of steepness, for populations of 3,5,7,20,100. All of these slopes are logistic functions; as the population increases the slope becomes steeper.

There is a phase transition at approximately $\alpha_7 = 0.7$. Below this, the probability goes down with population size and above this phase transition point, the probability increases with population size. At the phase transition point, the probability does not depend on n. The phase transition occurs when the coefficient of n in Equation (7) is zero.

Fig. 5 shows the probability of $r(A_1)$ as a function of α_7 for different population sizes. Note that all of the individuals have the same probability that r is true. This is for the same set of populations used in Figure 4; again, the smallest population has the least steep slope.

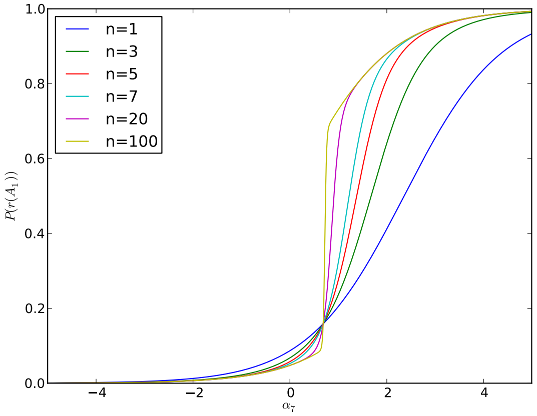

Figure 5: Probability of $r(A_1)$ as a function of α_7 in an MLN for various population sizes. See Example 6.

Notice the way the parameter α_7 affects the probability depends on the population size. We cannot set the parameters so that the MLN represents the logistic regression as the population varies.

At the phase transition, there is an approximately vertical line segment for large populations. The corresponding probabilities for $r(A_1)$ cannot be represented in the limit. We known in the limit that $P(q)$ approaches either 0 or 1 (or is not affected by the population size). For example, if in the limit we have $P(q) \to 1$ and we adjust α_7 to fit $P(r(A_1)) = 0.3$ when $P(q) = 1$, the new value found for α_7 implies that $P(q) \to 0$ in the limit. Similarly, if $P(q) \to 0$ and we adjust α_7 to fit $P(r(A_1)) = 0.3$ when $P(q) = 0$, the new value found for α_7 implies that $P(q) \to 1$. Thus α_7 cannot be set to make $P(r(A_1)) = 0.3$ in the limit.

Fig. 6 shows how q and $r(A_1)$ vary with population size for two different parametrizations, $\alpha_7 = 0.66$ and 0.73. The lines that approach 1 are for $\alpha_7 = 0.73$ and the lines that approach 0 are for $\alpha_7 = 0.66$.

6 Beyond the Simple Example

The general case for relational logistic regression is to use a linear function of arbitrary relational features on arbitrary populations. It is even possible to have a definition of the conditional probability as a weighted set of clauses in much the same way as an MLN is represented. This is a different class of models than the MLNs.

The MLN

$$q \vee r(x) \vee r(y)$$

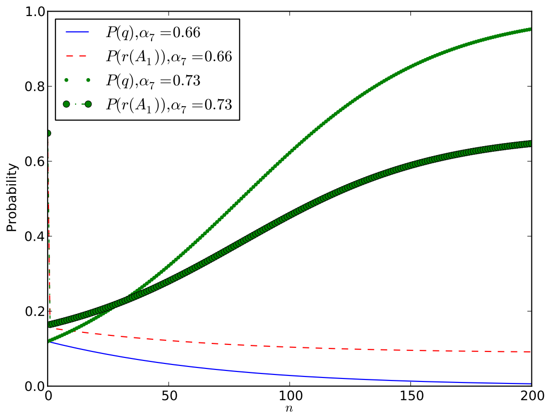

Figure 6: Probability of q and $r(A_1)$ as a function of population sizes for $\alpha_7 = 0.66$ and 0.73. See Example 6.

allows for a squared growth with population. The MLN

$$q \vee r(x) \vee r(y) \vee r(z)$$

allows for a cubic growth with population.

There are many issues that remain open. What parametrizations allow for a k-degree polynomial growth with population? What about a $\sqrt{n}$ growth with population? (E.g., if the individuals are arcs in a dense network, a property of nodes grows with the square root of the population of arcs.)

Noisy-and, noisy-or and averages as aggregation functions can also be represented with clauses [Natarajan et al., 2010]. Proposition 3 is also applicable to these methods; while MLNs without factors among the individuals may be able represent the aggregation, they have side effects that may make them less applicable.

7 Conclusion

We had expected complex dependence on population for complex cases (e.g., when the R_i in the running example are dependent due to common ancestors). We were surprised to find that even the well-understood case of logistic regression has complex dependence on population size.

- If we learn a model for some population sizes, we may want to apply it to other population sizes. We want to make explicit assumptions and know the consequences of these assumptions.
- We want to know the effect of choosing particular parametrizations. What assumptions are we making? Why should we choose one representation over another?
- If one model fits some data, it is important to understand why it fits the data better. This will enable us to know what models to consider for different data.

The other message is that undirected models such as MLNs are different to directed models such as relational logistic regression. It is important to understand these differences if we are to choose an appropriate model for a domain.

References

de Salvo Braz, R., Amir, E., and Roth, D. (2007). Lifted first-order probabilistic inference. In L. Getoor and B. Taskar (Eds.), *Introduction to Statistical Relational Learning*. M.I.T. Press.

Domingos, P., Kok, S., Lowd, D., Poon, H., Richardson, M., and Singla, P. (2008). Markov logic. In L.D. Raedt, P. Frasconi, K. Kersting, and S. Muggleton (Eds.), *Probabilistic Inductive Logic Programming*, pp. 92–117. Springer, New York.

Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (1999). Learning probabilistic relational models. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*, pp. 1300–1309. Morgan Kaufman, Stockholm, Sweden.

Horsch, M. and Poole, D. (1990). A dynamic approach to probabilistic inference using Bayesian networks. In *Proc. Sixth Conference on Uncertainty in AI*, pp. 155–161. Boston.

Jain, D., Kirchlechner, B., and Beetz, M. (2007). Extending markov logic to model probability distributions in relational domains. In *KI*, pp. 129–143.

Jain, D., Barthels, A., and Beetz, M. (2010). Adaptive markov logic networks: Learning statistical relational models with dynamic parameters. In *9th European Conference on Artificial Intelligence (ECAI)*, pp. 937–942.

Natarajan, S., Khot, T., Lowd, D., Kersting, K., Tadepalli, P., and Shavlik, J. (2010). Exploiting causal independence in markov logic networks: Combining undirected and directed models. In *European Conference on Machine Learning (ECML)*.

Neville, J., Simsek, Ö., Jensen, D., Komoroske, J., Palmer, K., and Goldberg, H. (2005). Using relational knowledge discovery to prevent securities fraud. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Perlich, C. and Provost, F. (2006). Distribution-based aggregation for relational learning with identifier attributes. In *Machine Learning*.

Poole, D. (2003). First-order probabilistic inference. In *Proc. Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pp. 985–991. Acapulco, Mexico.

Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62: 107–136.