Algorithm for MCMC-based Pseudo-optimization of an Individual Based Model

The individual-based model (IBM) for dairy cattle has been previously developed (Al-mamun et al., 2016). Briefly, individual animals are tracked from birth, through a heifer-raising process, to adulthood. Adult animals maintain a history of milk production and reproductive status; reproduction decisions are made at the herd level by means of a standard policy. All animals are subject to involuntary culling due to disease or other factors. Herd size is maintained by means of voluntary culling, in which animals are removed for reasons of failing to reproduce or age. The goal of this project is to provide an algorithm for assessing cow value, incorporating both reproductive status and milk production, to improve the voluntary culling process. The goal is to increase the net present value (NPV) of the herd over a set time frame.

**Changes to the IBM**

Currently, voluntary culling in the IBM is based primarily on age and reproduction. We will reprogram the IBM such that voluntary culling is based on the value of the cow, cowval (see below). That will require the following changes:

1. Add a check for herd size – if the adult herd contains more animals than the preset maximum, voluntary culling will occur to bring the number of adults down to the maximum
2. Calculate cowval for each animal in the herd – using the prespecified cowval algorithm (see below), a single value will be assigned to each adult animal
3. Cull the animals with the lowest cowval

We will also have to add a calculation of the herd’s NPV. This will be based on the amount of milk produced, the number of animals inseminated, the number of animals culled, the number of animals fed, and economic values attached to each of these numbers. For any given point in time, the herd value will be:

where pmilk is the price received for milk, fci is the feed costs for group i, Nins is the number inseminated, crepro is the cost of reproduction programs per insemination, Nculls is the number of animals culled, and salvage is the value received for a culled animal. The NPV is a function of the herdval over our time horizon,

where i is the interest (discount) rate.

**Cow Value**

One of the most common herd management software programs contains a function known as cowval, which compares the value of a particular cow to the average value of a replacement heifer (Sorge et al., 2007). While this function is proprietary, it is known to be based on milk production and reproductive status. However, the best combination of these parameters is not clear, as both are time-dependent, variable, and complex. Therefore, we will consider multiple potential cowval algorithms.

Milk production varies over the life of the cow and over the course of a lactation. Therefore, a method exists for standardizing the value of milk production based on age and days in milk (DIM): ME305. Use of ME305 would allow direct comparison of the milk production potential of cows at different life and lactation stages. Therefore, we will consider ME305 as a linear factor in cowva formulation; for simplification, we will also consider the quintile of ME305 as a categorical factor. However, it is possible that ME305 alone could mask some variation in cow value related to age or genetic potential. Therefore, we will consider cowval formulations with an interaction between ME305 and either parity or genetic potential (measured as a categorical factor).

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Format |
| ME | ME305 | Linear |
| qME | Quintile of ME305 | Categorical {1-5} |
| pME | ME305|parity | Set of linear {1-5+} |
| pqME | Quintile of ME305|parity | Set of categorical |
| qqME | Quintile of ME305|quintile of genetic potential | Set of categorical |

Reproductive status is a time-dependent factor. For the first 60 DIM, cows are not inseminated; this is known as the voluntary waiting period (VWP). After the VWP, cows are inseminated on a monthly schedule and pregnancy detection occurs one month following. Therefore, cows can be sorted into pregnancy categories: VWP, inseminated (Ins), pregnant (preg), or non-pregnant (open). Pregnant animals can be further defined by the length of the pregnancy, commonly referred to as days carried calf (DCC) or month in pregnancy (MIP). Further, as the calf will be expected to inherit the mother’s genetic potential (with some improvement factor, negligible for this analysis), the pregnant animal can also be defined by the genetic potential of her calf. Therefore, our cowval formulations may include the pregnancy category alone or the pregnancy category with additional factors for pregnant animals. As the cost of being non-pregnant is likely to increase as the lactation progresses, we may also add an additional variable for month in milk (MIM) in animals classed as Ins or open.

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Format |
| Rep | Reproductive status | Categorical {VWP, Ins, Preg, Open} |
| MIP(Preg) | Months in pregnancy | Categorical {1-9} |
| MIP | Months in pregnancy | Categorical {0-9} |
| MIP|MIM | Months in pregnancy | Set of Categorical {0-9}|{3-12+} |
| qGen(Preg) | Quintile of genetic potential | Categorical {0-5} |
| MIM(Rep) | Months in milk | Set of Categorical {1-12+}|Rep |
| MIM(Open) | Months in milk | Categorical {1-12+} |

The overall algorithm for cowval will be calculated as a linear combination of the milk factors and the reproductive factors, as in a linear regression model:

where **βM** is a vector of the coefficients for all milk variables and **βR** is a vector of the coefficients for all reproduction variables. For each formulation of cowval, the coefficient (β) values will be fit using the MCMC process (below).

**MCMC**

The Markov chain-Monte Carlo (MCMC) algorithm been well established for statistical fitting of models to data (Brooks, 1998). It has also recently been applied to stochastic optimization processes for machine learning algorithms using stochastic gradient MCMC (SG-MCMC) methods (Mandt et al., 2016; Li et al., 2016; Tobias et al., 2016). We propose using an MCMC approach with a Metropolis-Hastings algorithm (MH) and a Gibbs sampler for stochastic optimization of the IBM for cowval. This will follow these steps:

1. sample all coefficients (β values) from their prior distributions (set initially at U(-∞, ∞)), βs
2. simulate herd over time frame and calculate NPVs
3. draw new values of coefficients from a random walk algorithm, βnew
4. simulate herd over time frame and calculate NPVnew
5. calculate your decision value
6. if α>U(0,1), NPVs+1=NPVnew and βs+1=βnew; else, NPVs+1=NPVs and βs+1=βs
7. repeat steps 3-6 until stationarity is achieved for all β (by means of the Gelman-Rubin diagnostic test)

We will have to fine-tune the random walk algorithm to provide efficient estimation, and we will need to set the burn-in empirically unless you know an automatic algorithm.

**Comparing Cowval Formulations**

Our MCMC process will produce a distribution of NPV for each cowval formulation. We can compare cowval formulations simply by comparing the NPV distributions, but that is unlikely to provide a simple answer. We will need to define a rule for preferring one distribution to another. I can provide the first-order and second-order rules for stochastic dominance, at least, once we get to that point.

The other option is a variation on the reversible-jump MCMC (RJMCMC) (Andrieu et al., 1999; Hooten and Hobbs, 2015), in which we sample formulations as well as coefficients within those formulations. In this case (which I would prefer, I think), I will suggest that we fit each formulation separately. We will then perform a variation on the MH algorithm above:

1. select a formulation from the set of possible formulations, fs
   1. sample all coefficients (β values) from their posterior distributions for that model, βf,s
2. simulate herd over time frame and calculate NPVs
3. select a different formulation from the set of possible formulations, fnew
   1. draw new values of coefficients from the posterior distributions for the new model, βnew
4. simulate herd over time frame and calculate NPVnew
5. calculate your decision value
6. if α>U(0,1), NPVs+1=NPVnew and fs+1=fnew; else, NPVs+1=NPVs and fs+1=fs
7. repeat steps 3-6 for a set amount of time and identify which formulations were selected most frequently

Traditionally, the RJMCMC would combine these two approaches, so that the coefficients drawn within a particular formulation would be based on a random walk or similar process. We can debate the better approach.

**Expanding to Include Disease**

The above process will need to be completed on a disease-free herd before considering disease costs. Once the disease-free cowval formulation has been chosen, we will consider adding testing components.

*Changes to the IBM*

The current IBM should have the mechanism in place for adding different testing programs and defining cows by their test status. IMPORTANT: the cowval should always be based on testing history, never on true infection status!

We will need to subtract the cost of testing from the herdvalue calculation. Test cost is simply Ntested\*costtest. We would also need to subtract the cost of any additional hygiene strategy; we will need to define that cost function later.

*Changes to Cowval*

We will have to consider different cowval formulations for test results. Tests can give results that are binary (positive/negative), ordered (negative/low/high), or numeric. With binary results, we also may wish to include a count variable for the number of positive test results. For each potential test, therefore, we may wish to try different formulations of adding the test to the cowval calculation. We will also want to optimize the no-control cowval (the current formulation) with disease present, with and without additional hygiene controls.

For efficiency, I propose that the optimization of the cowval in the diseased herd begin with the existing cowval parameters drawn from their disease-free posterior. We can debate whether these parameters should still be subjected to a random walk process, or if they can be drawn entirely from the disease-free posterior. We will assume that the coefficient values for the test results are negative (bounded on the right by 0).

For each test, we would want to optimize the cowval formulation at a number of different testing frequencies: annual, biannual, quarterly, and at calving (cows are tested at the time of calving). We will compare the NPV distributions for each test/frequency combination, without testing, and with additional hygiene controls. We may also try the RJMCMC approach for all variations on control. In this case, we would select at each “model selection” point the test to be used and the frequency, as well as the amount of hygiene to apply.