Data Cloning

Lele et al. (2007, 2010)

Laxman Ghimire and Jennifer La Rosa

McMaster University

November 9, 2015

The Model

- We wish to model our data via hierarchical models with both fixed parameter values and random effects.
- Let $\mathbf{y}$ be our observed data vector of sample size n. Let $\mathbf{x}$ be our unobserved states we wish to predict. Let $\boldsymbol{\theta}=(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ be the parameters we wish to estimate.
- Hierarchy 1:

$$\mathbf{y}|\mathbf{X} = \mathbf{x} \sim \mathbf{h}(\mathbf{y}; \mathbf{X} = \mathbf{x}, \boldsymbol{\theta_1})$$

- Hierarchy 2:

$$\mathbf{X} \sim \mathbf{g}(\mathbf{x}; \boldsymbol{\theta_2})$$

Difficulties Encountered in Analyzing These Models

- Computational approaches to find the mle's are difficult when the likelihood function must be simulated.
- To compute the mle, evaluation of a high-dimensional integral is required.
- Models are complex and analytical results are sparse creating concerns about potential model pitfalls such as non-estimability.

The Method: General

- Construct a Bayesian model and specify proper priors for the unknown parameters.
- Use k clones of the observed data and obtain the corresponding likelihood.
- We assume k is large and that the clones are independent.
- Calculate the posterior distribution via MCMC.
- Set the mle to be the mean of the posterior distribution.
- The asymptotic variance of the mle is equal to k times the variance of the posterior distribution.

The Method: The Steps

Step 1:

- Create the new k-cloned data set

$$\mathbf{y}^{(k)} = (\mathbf{y}, \mathbf{y}, ..., \mathbf{y})$$

  where the observed data vector, $\mathbf{y}$, is repeated k times.
- The k clones are assumed to be independent of each other.

The Method: The Steps

Step 2:

- Generate random numbers from the posterior distribution, which is based on the prior, the hierarchical structure and the k-cloned data vector, via an MCMC algorithm.
- The Metropolis-Hastings algorithm could be used for example.

The Method: The Steps

Step 3:

- Calculate the sample means and sample variances of $(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})_j$ for j=1,2, ...,B, generated from the marginal posterior distribution.
- The mle's correspond to the posterior mean values.
- The approximate variances of the mle's are k times the posterior variances.

Determining the Number of Clones

- The statistical accuracy of the mle is based on the data and its sample size, $\mathbf{y}$. Increasing the number of clones or the length of the MCMC run only improves the numerical accuracy of the approximation to the mle.
- The number of clones is determined by the analyst.
- To determine an adequate number of clones, we must determine when the posterior distribution is nearly degenerate.

Advantages of Data Cloning

- Uses Bayesian framework and MCMC, so it is computationally simple. There is no difficult high-dimensional integration, differentiation or numerical maximization of a noisy likelihood function.
- The Bayesian framework is simply a device to conduct likelihood calculations and the method provides maximum likelihood estimates.
- The inferences do not depend on the prior distribution chosen (as long as the prior is not degenerate and the model satisfies some regularity conditions). So a proper and computationally convenient prior may be used.
- Data cloning now gives ecologists and statisticians the option to use frequentist inference for hierarchical models based on the relevance of the prior for scientific inferences.

Disadvantages of Data Cloning

- Theoretically, as k becomes infinite, the data cloning algorithm arrives at the global maximum. However, since k is finite in practice, we must check that we do not arrive at a local maximum instead. The issue may be solved by rerunning the algorithm with different priors and with increasing values of k since the posterior mean values should converge to the same values for different priors when k is large enough.
- The standard errors are large-sample approximations.
- Data cloning does not make up for a lack of data.
- The likelihood and Bayesian inferences can be ill-behaved for data which does not contain information about the parameters. Data cloning will not remedy over-parameterized or ill-parameterized models. The method assumes the parameters are identifiable.

Other Notes

- The mle's obtained via data cloning result from maximizing the full likelihood function with the random effects integrated out.
- Using informative prior distributions can help to speed the convergence process of the posterior mean values.
- Data cloning can be used to obtain point prediction and prediction intervals for the random effects.

Estimability Diagnostics

- A challenge of many hierarchichal models is non-identifiability of the parameters.
- However, the inferences must be based on identifiable parameters to be valid.
- A benefit of data cloning is that we can check if a parameter is estimable by seeing if the variance of the posterior distribution of the parameter of interest converges to zero.
- To do this, we can plot the posterior variance or the largest eigenvalue of the posterior variance matrix as a function of the number of clones. An estimable function of $\boldsymbol{\theta}$ will have the property that the posterior variance will converge to zero as k increases.

Conclusion

- Data cloning provides a simple way to compute the maximum likelihood estimates using MCMC.
- The results are not dependent on the prior that we choose.
- The method can bring awareness to non-estimability and non-identifiability issues.
- We will now provide some examples and demonstrate how the data cloning method has been put into practice by other ecologists and statisticians.

References

- Lele, S. R., Dennis, B., and Lutscher, F., (2007). "Data cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models Using Bayesian Markov Chain Monte Carlo Methods." *Ecology Letters*, **10**, 551-563.
- Lele, S. R., Nadeem, K., and Schmuland, B., (2010). "Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning." *Journal of the American Statistical Association*, **105**, 1617-1625.