# Data Cloning

## Lele et al. (2007, 2010)

Laxman Ghimire and Jennifer La Rosa

McMaster University

November 12, 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 length $n$, where $n$ is the sample size. 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:

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

- Hierarchy 2:

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

- The corresponding likelihood is

$$L(\boldsymbol{\theta}; \mathbf{y}) = \int h(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta_1}) g(\mathbf{x}; \boldsymbol{\theta_2}) \mathbf{dx}$$

# 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.
- Note that the corresponding likelihood

$$L(\boldsymbol{\theta}; \mathbf{y}^{(k)}) = [L(\boldsymbol{\theta}; \mathbf{y})]^k$$

# The Method: The Steps

**Step 2:**

- Generate random variates, $\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_B}$, from the posterior distribution, $\pi_k(\boldsymbol{\theta}|\mathbf{y})$, which is based on the prior, $\pi(\boldsymbol{\theta})$, the hierarchical structure and the $k$-cloned data vector, $\mathbf{y}^{(k)}$, via an MCMC algorithm.

$$\pi_k(\boldsymbol{\theta}|\mathbf{y}) = \frac{[L(\boldsymbol{\theta}; \mathbf{y})]^k \pi(\boldsymbol{\theta})}{\int [L(\boldsymbol{\theta}; \mathbf{y})]^k \pi(\boldsymbol{\theta}) d\theta}$$

- The Metropolis-Hastings algorithm could be used for example.

# The Method: The Steps

**Step 3:**

- Calculate the sample means and sample variances of $\theta_j$ 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.
- This is because we note that via data cloning

$$\pi_k(\boldsymbol{\theta}|\mathbf{y}) \sim MVN(\hat{\boldsymbol{\theta}}, \frac{1}{k} I^{-1}(\hat{\boldsymbol{\theta}}))$$

# Determining the Number of Clones

- The statistical accuracy of the MLE is based on the data, $\mathbf{y}$, and its sample size, $n$. 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.
- To determine if the posterior distribution has become degenerate, we can plot the largest eigenvalue of the posterior as a function of the number of clones, $k$. Then we compare this with the expected value plot of $\frac{1}{k}$ since the largest eigenvalue of the posterior distribution converges to zero at the same rate as $\frac{1}{k}$.

# 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.
- Data cloning allows us to check for identifiability of the parameters.

# Disadvantages of Data Cloning

- Although the method simplifies the computation, it does not mean that the method is necessarily computationally efficient. MCMC runs are longer via data cloning requiring more computing time.
- 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.
- 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.
- Data cloning can be used to obtain point prediction and prediction intervals for the random effects.

# Identifiability

- A challenge of many hierarchical models is non-identifiability of the parameters.
- That is, two parameters may produce the same likelihood function and thus we can't identify the true parameter.
- Intrinsic non-identifiability occurs due to the structure of the model. For instance, a parameter may be confounded with one or more other parameters in the model.
- Extrinsic non-identifiability occurs when the data are inadequate or the parameters are poorly estimated near the boundaries.
- However, the inferences must be based on identifiable parameters to be valid.

# Estimability Diagnostics

- 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 $\theta$ 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.