Posterior predictive checks with missing data

Posterior predictive checks allow us to determine if our model of an ecological process is capable of giving rise to the data, the fundamental assumption of model-based inference. These checks compare statistics computed from the observed data with statistics computed from data simulated from the model. A valid comparison depends on the assumption that the data are complete or, if there are missing observations, that they are missing completely at random. Data missing at random (notice, not completely at random) will support inference on model parameters and posterior prediction if proper covariates have been included in the model. Data missing not at random will support inference on model parameters and posterior prediction if the missing data mechanism has been modeled. However, a strict interpretation of the math means that the Bayesian P values produced by posterior predictive checks are not exact if data are missing at random. An exact value is probably not critical because evidence for lack of fit comes only from reasonably extreme values (e.g. $.10 \lessapprox P \gtrapprox 90$). Missing values are not likely to influence the decision on whether to reject a model for lack of fit, unless they are a large proportion of the data set[1]. However, if you need an exact Bayesian P value, it can be obtained by *multiple imputation*.

Multiple imputation is a like imputing a single missing value except we impute *all* of the missing data in a data set multiple times. Here is how this works.We first create the data vector $\mathbf{o}$ containing the indices of all of the observation in $\mathbf{y}$ that are missing. At each MCMC inteation (indexed by (k)), we simulate a complete, new data set $\mathbf{y}^{new(k)}$ as described in section ——. We we then assign the values in the data we used for model fitting to a new data set, $\mathbf{y}^{imp(k)} = \mathbf{y}$ and assign imputed values from the simulated data set to the missing values in $\mathbf{y}^{imp(k)}$

$$\mathbf{y}^{imp(k)}_{\forall i \in \mathbf{o}} = \mathbf{y}^{new(k)}_{\forall i \in \mathbf{o}}, \tag{1}$$

where the notation $\forall i \in \mathbf{o}$ read "for all elements in $\mathbf{o}$". We now have K imputed and new

[1]In which case, Bayesian$P$ values are the least of your problems.

data sets where K is the number of retained MCMC iterations.

Computing the Bayesian P properly adjusted for missing data proceeds as follows. We randomly select and index from the imputed data $k^{imp} \sim \text{uniform}(1, K)$ and a different index from the simulated data $k^{new} \sim \text{uniform}(1, K)$. Next, we compute a test statistic T from from the simulated data $T(\mathbf{y}^{new(k^{new})})$and from the imputed data $T(\mathbf{y}^{imp(k^{imp})})$. We set an indicator variable I

$$I = \begin{cases} 1 & T(\mathbf{y}^{new(k^{new})}) \geq T(\mathbf{y}^{imp(k^{imp})}) \\ 0 & T(\mathbf{y}^{new(k^{new})} < T(\mathbf{y}^{imp(k^{imp})}). \end{cases} \tag{2}$$

The mean of I is the Bayesian P value for the test statistic T corrected for the missing data.