14 Multinomial processing trees

with Dora Matzke

14.1 Multinomial processing model of pair-clustering

Consider a free recall task in which people study a list of words—table, dog, brick, pencil, cat, news, doctor, keys, nurse, soccer—and, after a short delay, are asked to recall the words in any order. An interesting property of this list is that it contains some pairs of semantically related words, like dog and cat, or doctor and nurse, as well as some words, called "singletons," like table or soccer, that are not semantically related to another on the list.

A standard finding is that semantically related words are often recalled consecutively, even when they are not adjacent in the study list. For example, a person may recall soccer, cat, dog, table, doctor, and nurse. The finding that semantically related items are often recalled consecutively can be taken as evidence for the idea that they were stored and retrieved as a cluster.

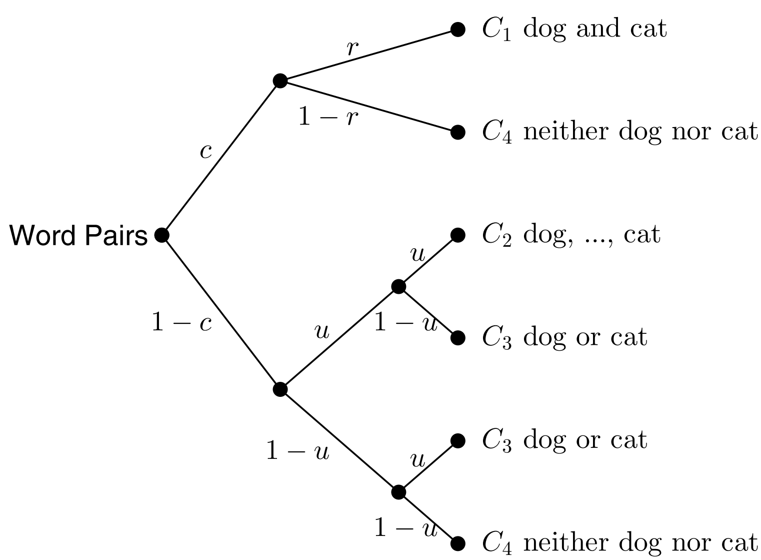


Fig. 14.1 Multinomial processing tree model for pair-clustering effects in recall from memory, with cluster-storage c, cluster-retrieval r, and unique storage-retrieval u parameters.

Multinomial processing trees (MPTs: Batchelder & Riefer, 1980, 1986; Chechile, 1973; Chechile & Meyer, 1976) provide one approach to modeling these memory effects. An MPT model assumes that observed behavior arises from a sequence of cognitive events, able to be represented by a rooted tree architecture such as the one shown in Figure 14.1. In this MPT model, the focus is on the recall of word pairs, but more general models can be developed that also account for singletons.

The model in Figure 14.1 considers four categories of response behavior for word pairs. In the first category, C_1, both words in a word pair are recalled consecutively. In the second category, C_2, both words in a word pair are recalled, but not consecutively. In the third category, C_3, only one word of a word pair is recalled. In the fourth category, C_4, neither word in a word pair is recalled.

The pair-clustering MPT model describes a simple sequence of cognitive processes that can produce these four behavioral outcomes, controlled by three parameters. The cluster-storage parameter c is the probability that a word pair is clustered and stored in memory. The cluster-retrieval parameter r is the conditional probability that a word pair is retrieved from memory, given that is was clustered. The unique storage-retrieval parameter u is the conditional probability that a member of a word pair is stored and retrieved from memory, given that the word pair was not stored as a cluster.

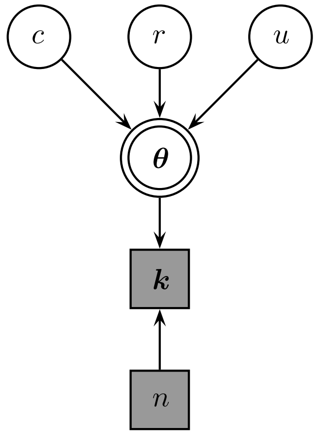
Under the MPT model for the pair-clustering paradigm, the probabilities of the four response categories are as follows:

$$
\begin{aligned}
\Pr\left(C_{11} \mid c, r, u\right) &= cr \\
\Pr\left(C_{12} \mid c, r, u\right) &= \left(1-c\right)u^2 \\
\Pr\left(C_{13} \mid c, r, u\right) &= 2u\left(1-c\right)\left(1-u\right) \\
\Pr\left(C_{14} \mid c, r, u\right) &= c\left(1-r\right) + \left(1-c\right)\left(1-u\right)^2 .
\end{aligned}
$$

For example, the MPT model states that the probability of category C_1—that is, consecutively recalling semantically related study words—requires that the words are first stored in memory as a pair, with probability c, and then retrieved as a pair, with probability r. The probabilities for other categories can also be decomposed as products and sums of the MPT parameters. In this way, the MPT model provides an account of the number of times words are recalled according to the different categories of behavior, controlled by the cognitive processes represented by the tree and its parameters.

Table 14.1 Category counts for Trials 1, 2, and 6 from Riefer et al. (2002).

	C_1	C_2	C_3	C_4
Trial 1	45	24	97	254
Trial 2	106	41	107	166
Trial 6	243	64	65	48



$$
\begin{aligned}
c &\sim \text{Beta}(1,1)\\
r &\sim \text{Beta}(1,1)\\
u &\sim \text{Beta}(1,1)\\
\theta_1 &\leftarrow cr\\
\theta_2 &\leftarrow (1-c)\,u^2\\
\theta_3 &\leftarrow 2u\,(1-c)\,(1-u)\\
\theta_4 &\leftarrow c\,(1-r)+(1-c)\,(1-u)^2\\
\boldsymbol{k} &\sim \text{Multinomial}(\boldsymbol{\theta}, n)
\end{aligned}
$$

Fig. 14.2 Graphical model for the pair-clustering MPT model of aggregated data.

Most analyses of MPT models rely on category response data that are aggregated over subjects and items (e.g., Hu & Batchelder, 1994). Our applications use a subset of the free recall data reported in Riefer et al. (2002). We analyze the free recall performance of 21 subjects responding to 20 categorically related word pairs in a series of six study-test trials. Each trial featured exactly the same word materials, and it therefore seems reasonable to expect that all three model parameters will increase over trials. Hence, we focus on performance in the first, second, and sixth session. Table 14.1 shows the aggregated data for each of the four categories of behavior, in all three of these trials.

A graphical model for the MPT account of these data is shown in Figure 14.2. The c, r, and u parameters are given uniform priors, and generate probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)$ for each of the four categories. The aggregated count data k for any trial thus follow $k \sim \text{Multinomial}\big(\boldsymbol{\theta}, n\big)$, where n is the total number of behaviors over all subjects and word pairs.

The script `MPT_1.txt` implements the graphical model in WinBUGS:

```
# Multinomial Processing Tree
model{
  # MPT Category Probabilities for Word Pairs
  theta[1] <- c * r
  theta[2] <- (1-c) * pow(u,2)
  theta[3] <- (1-c) * 2 * u * (1-u)
  theta[4] <- c * (1-r) + (1-c) * pow(1-u,2)
  # Data
  k[1:4] ~ dmulti(theta[1:4],n)
  # Priors
  c ~ dbeta(1,1)
  r ~ dbeta(1,1)
  u ~ dbeta(1,1)
}
```

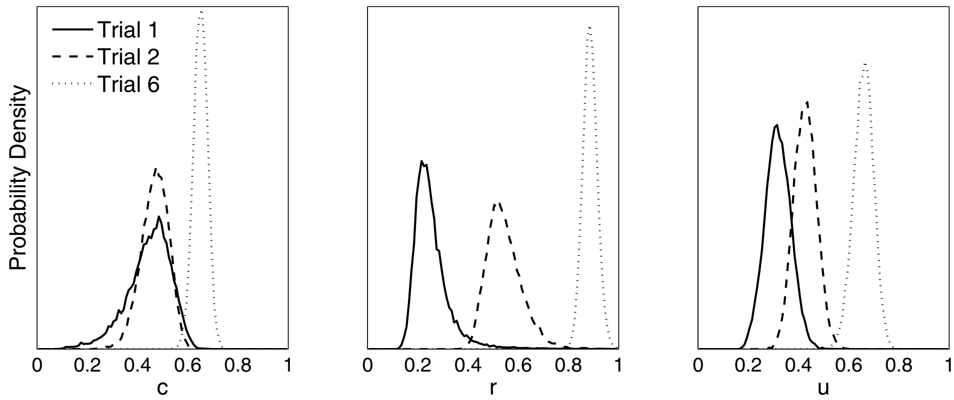


Fig. 14.3 Posterior distributions for the c, r, and u parameters for the Riefer et al. (2002) data set.

The code `MPT_1.m` or `MPT_1.R` applies the model to make inferences for the three trials. Figure 14.3 shows the posterior distributions for each of the c, r, and u parameters for each trial.

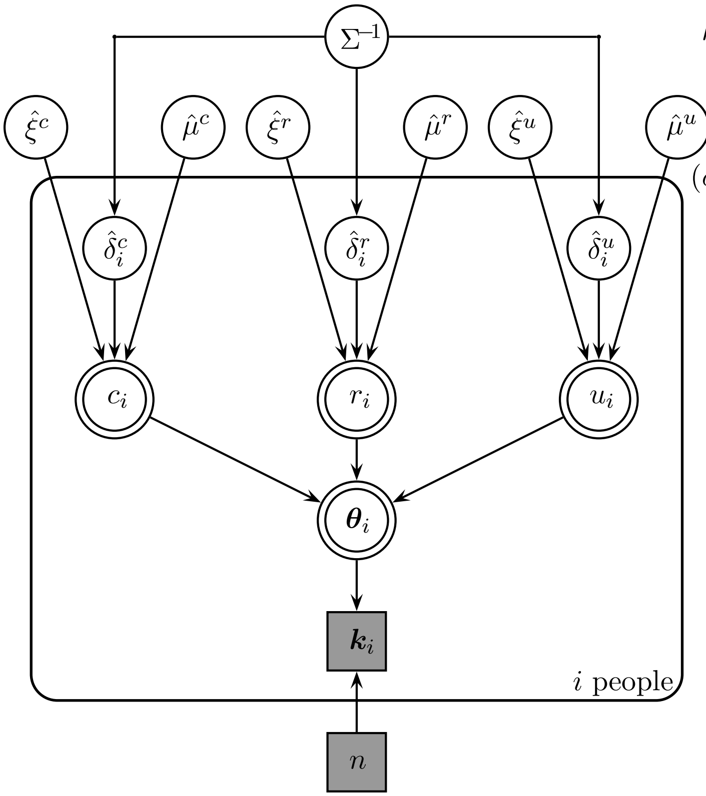
Exercises

Exercise 14.1.1 What do you conclude from the posterior distributions in Figure 14.3 about learning over the course of the trials?

Exercise 14.1.2 Because the u parameter corresponds to both the storage and retrieval of unclustered words, it is typically regarded as a nuisance parameter. In an approach to inference that is not fully Bayesian, the lack of interest in the posterior distribution of u might lead to the shortcut of a reasonable value being substituted, rather than assigning a prior distribution. Modify the graphical model so that u is set to a constant for each trial, given by the expected value of the posterior from the fully Bayesian analysis. How does this change affect the posterior distributions of c and r, the parameters of interest?

14.2 Latent-trait MPT model

The use of aggregated data relies on the assumption that subjects do not differ in terms of the psychological process used by the MPT model. Modeling individual subject data, and allowing the parameters for each subject to vary in some structured way, provides one powerful way to addresses this issue. Here we focus on a latent-trait approach developed by Klauer (2010), which not only allows for variation in parameters between individuals, but provides an explicit model of that variation.



$$
\begin{aligned}
\hat{\mu}^c, \hat{\mu}^r, \hat{\mu}^u &\sim \text{Gaussian}(0, 1) \\
\hat{\xi}^c, \hat{\xi}^r, \hat{\xi}^u &\sim \text{Uniform}(0, 100) \\
\Sigma^{-1} &\sim \text{Wishart}(I, 4) \\
(\hat{\delta}_i^c, \hat{\delta}_i^r, \hat{\delta}_i^u) &\sim \text{MvGaussian}(\mathbf{0}, \Sigma^{-1}) \\
c_i &\leftarrow \Phi(\hat{\mu}^c + \hat{\xi}^c \hat{\delta}_i^c) \\
r_i &\leftarrow \Phi(\hat{\mu}^r + \hat{\xi}^r \hat{\delta}_i^r) \\
u_i &\leftarrow \Phi(\hat{\mu}^u + \hat{\xi}^u \hat{\delta}_i^u) \\
\theta_{i1} &\leftarrow c_i r_i \\
\theta_{i2} &\leftarrow (1 - c_i)\, u_i^2 \\
\theta_{i3} &\leftarrow 2u_i\,(1 - c_i)\,(1 - u_i) \\
\theta_{i4} &\leftarrow c_i\,(1 - r_i) \\
&\quad + (1 - c_i)\,(1 - u_i)^2 \\
\boldsymbol{k}_i &\sim \text{Multinomial}(\boldsymbol{\theta}_i, n)
\end{aligned}
$$

Fig. 14.4 Graphical model for the latent-trait pair-clustering model.

The latent-trait approach we adopt assumes that there are group means—μ_c, μ_r, and μ_u—for the three parameters. The values of these parameters for the ith subject—c_i, r_i, and u_i—are then modeled in terms of their displacement around these group means. The parameters that control the displacement come from a multivariate distribution, so that the variability and correlation between the displacements themselves are both modeled. The motivation for taking this additional modeling step is that the cognitive processes represented by the model parameters—the retrieval of clustered and unclustered word pairs, for example—may often be highly correlated, and part of the goal of modeling is to make inferences about these relationships from data.

Figure 14.4 presents the graphical model for the latent-trait pair-clustering model. The data $\boldsymbol{k}_i$ for the ith subject consist of counts of responses falling into the different response categories shown in Figure 14.1. There are now cluster-storage c_i, cluster-retrieval r_i, and unique-retrieval u_i parameters for the ith subject, which generate response probabilities $\boldsymbol{\theta}_i\,(\theta_{i1}, \ldots, \theta_{i4})$, and $k_i \sim \text{Multinomial}\big(\boldsymbol{\theta}_i, n\big)$ as before, with n now being just the number of word pairs for a single subject.

Box 14.1 Fundamental flaws

"One argument for Bayesian inference is that it is better to wrestle with the practicalities of a method that is fundamentally sound, than to work with one having fundamental flaws." (O'Hagan & Forster, 2004, p. 17)

The individual subject parameters c_i, r_i, and u_i are all probabilities, as in the original model, but their variability is modeled in a probit-transformed space. The group means in the probit space are $\hat{\mu}^c$, $\hat{\mu}^r$ and $\hat{\mu}^u$, with, for example, $\mu^c = \Phi(\hat{\mu}^c)$. The priors on the group mean in the probit space are $\hat{\mu}^c, \hat{\mu}^r, \hat{\mu}^u \sim \text{Gaussian}(0, 1)$ which corresponds in the probability space to $\mu^c, \mu^r, \mu^u \sim \text{Uniform}(0, 1)$.

Individual differences come from displacement parameters $\hat{\delta}_i^c$, $\hat{\delta}_i^r$, and $\hat{\delta}_i^u$ for the ith subject. These are draws from a multivariate Gaussian distribution, with $(\delta_i^c, \delta_i^r, \delta_i^u) \sim \text{MvGaussian}(\mathbf{0}, \Sigma^{-1})$. The multivariate Gaussian[1] is zero-centered, with an unscaled covariance matrix Σ. The prior for the inverse covariance is $\Sigma^{-1} \sim \text{Wishart}(I, 4)$, where I is the 3×3 identity matrix, and there are 4 degrees of freedom. This is a standard prior that corresponds to a uniform distribution on the correlation between the model parameters (Gelman & Hill, 2007, pp. 284–287 and pp. 376–378; Klauer, 2010, pp. 77–78).

To improve the rate of convergence in MCMC sampling, a parameter expansion method is also used in the graphical model shown in Figure 14.4. This involves redundant multiplicative scale parameters $\hat{\xi}^c, \hat{\xi}^r, \hat{\xi}^u \sim \text{Uniform}(0, 100)$, that combine with the $\hat{\delta}_i^c$, $\hat{\delta}_i^r$, and $\hat{\delta}_i^u$ values to determine the offset in probit space each subject has from the group mean.

Putting all of this together, the cluster-storage parameter, for example, for the ith subject is given by $c_i \leftarrow \Phi(\hat{\mu}^c + \hat{\xi}^c \hat{\delta}_i^c)$. Exactly the same approach is used for the cluster-retrieval and unique-retrieval parameters.

The advantage of this latent-trait approach to individual differences—as compared, say, to simply drawing the individual subject parameters from independent group distributions for each parameter—is that the relationship between parameters is being modeled. The covariance matrix Σ contains the information needed to infer both the variance of the cluster-storage, cluster-retrieval, and unique-retrieval parameters, and the correlation between each pair of these parameters. The standard deviation for the cluster-storage parameter, for example, is given by $\sigma^c = |\hat{\xi}^c|\sqrt{\Sigma_{cc}}$, where Σ_{cc} is the diagonal element of the 3×3 covariance matrix Σ corresponding to the c parameter. The correlation between the cluster-storage and cluster-retrieval parameters, for example, is given by $\rho^{cr} = \hat{\xi}^c \hat{\xi}^r \Sigma_{cr} / \left(|\hat{\xi}^c|\sqrt{\Sigma_{cc}}|\hat{\xi}^r|\sqrt{\Sigma_{rr}}\right)$. With priors on the $\hat{\xi}$ parameters that allow only positive values, this simplifies to $\rho^{cr} = \Sigma_{cr} / \left(\sqrt{\Sigma_{cc}}\sqrt{\Sigma_{rr}}\right)$.

[1] Just as WinBUGS uses precisions instead of variances to parameterize the Gaussian distribution, it uses inverse covariance matrices instead of covariance matrices to parameterize the multivariate Gaussian distribution.

The script MPT_2.txt implements the graphical model in WinBUGS:

```
# Multinomial Processing Tree with Latent Traits
model{
  for (i in 1:nsubjs){
    # MPT Category Probabilities for Word Pairs
    theta[i,1] <- c[i] * r[i]
    theta[i,2] <- (1-c[i])*pow(u[i],2)
    theta[i,3] <- (1-c[i])*2*u[i]*(1-u[i])
    theta[i,4] <- c[i]*(1-r[i])+(1-c[i])*pow(1-u[i],2)
    # Data
    k[i,1:4] ~ dmulti(theta[i,1:4],n[i])
    # Probitize Parameters c, r, and u
    c[i] <- phi(muchat + xichat*deltachat[i])
    r[i] <- phi(murhat + xirhat*deltarhat[i])
    u[i] <- phi(muuhat + xiuhat*deltauhat[i])
    # Individual Effects
    deltahat[i,1:nparams] ~
       dmnorm(mudeltahat[1:nparams],SigmaInv[1:nparams,1:nparams])
    deltachat[i] <- deltahat[i,1]
    deltarhat[i] <- deltahat[i,2]
    deltauhat[i] <- deltahat[i,3]
  }
  # Priors
  mudeltahat[1] <- 0
  mudeltahat[2] <- 0
  mudeltahat[3] <- 0
  muchat ~ dnorm(0,1)
  murhat ~ dnorm(0,1)
  muuhat ~ dnorm(0,1)
  xichat ~ dunif(0,100)
  xirhat ~ dunif(0,100)
  xiuhat ~ dunif(0,100)
  df <- nparams+1
  SigmaInv[1:nparams,1:nparams] ~ dwish(I[1:nparams,1:nparams],df)
  # Post-Processing Means, Standard Deviations, Correlations
  muc <- phi(muchat)
  mur <- phi(murhat)
  muu <- phi(muuhat)
  Sigma[1:nparams,1:nparams] <- inverse(SigmaInv[1:nparams,1:nparams])
  sigmac <- xichat*sqrt(Sigma[1,1])
  sigmar <- xirhat*sqrt(Sigma[2,2])
  sigmau <- xiuhat*sqrt(Sigma[3,3])
  for (i1 in 1:nparams){
    for (i2 in 1:nparams){
      rho[i1,i2] <- Sigma[i1,i2]/sqrt(Sigma[i1,i1]*Sigma[i2,i2])
    }
  }
}
```

Note that the script, besides implementing the generative model for the data shown in Figure 14.4, also generates standard deviations and correlations in the sigma and rho variables. These could, in principle, be found by post-processing the posterior samples, but are implemented in the script for convenience.

The code MPT_2.m or MPT_2.R applies the model to make inferences for the three trials, based on individual subject data. Figure 14.5 shows the posterior distribu-

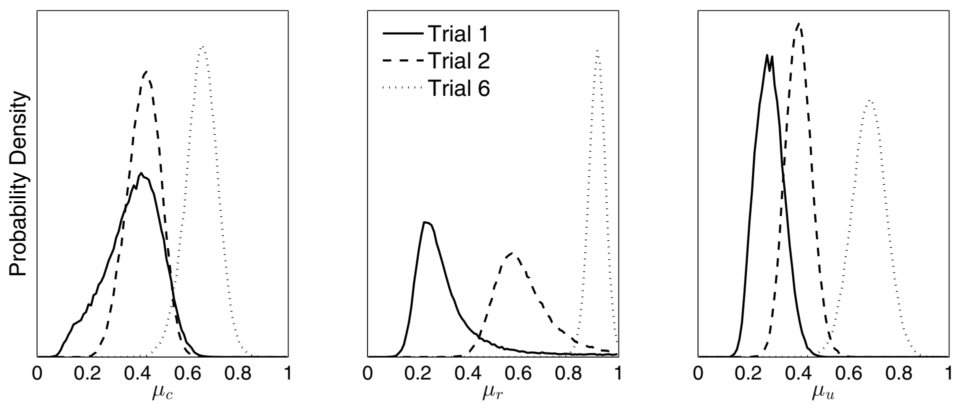


Fig. 14.5 Posterior distributions for group mean μ^c, μ^r, and μ^u parameters for the Riefer et al. (2002) data set, based on a latent-trait MPT model assuming individual differences.

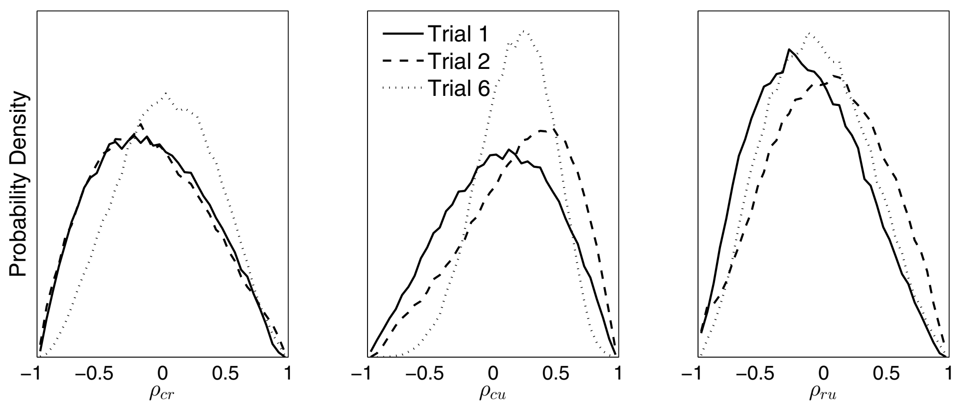


Fig. 14.6 Posterior distributions for the correlations ρ_{cr}, ρ_{cu}, and ρ_{cu} between parameters for the Riefer et al. (2002) data set, based on a latent-trait MPT model.

tions for the group means μ^c, μ^r, and μ^u for each trial. Figure 14.6 shows the posterior distributions for the correlations ρ_{cr}, ρ_{cu}, and ρ_{cu} between model parameters for each trial.

Exercises

Exercise 14.2.1 What do you conclude from the posterior distributions in Figure 14.5 about learning over the course of the trials? Compare your conclusions from the latent-trait model to the conclusions from the original MPT model.

Exercise 14.2.2 Extend the WinBUGS script to collect samples from the prior distributions for the standard deviation and correlation parameters. This will involve including variables `SigmaInvprior`, `Sigmaprior`, `rhoprior`, `sigmacprior`, `sigmarprior`, `sigmauprior`. Examine the prior and posterior distributions for the standard deviations and correlations. What can you conclude about the usefulness of including the correlation parameters in the latent-trait approach?

Exercise 14.2.3 The latent-trait approach deals with parameter heterogeneity as a result of individual differences between subjects, and relies on data that is aggregated over items. In many applications, however, it is reasonable to assume that the model parameters do not only differ between subjects but also between items. For example, it might be easier to cluster some pairs of semantically related words than others. This suggests using MPT models that incorporate both subject and item variability. Develop the graphical model that incorporates this extension. What is preventing the model from being applied to the current data?