Chapter 2

Daniel C. Furr

November 12, 2015

Contents

1 Introduction

2 Simulation and analysis methods

2.1 Data generation

Data are simulated for varying numbers of persons (P) and items (I) using the model described in Chapter 1. Specifically, the composite item difficulties are generated as

$$\delta_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{2i}x_{3i}\beta_5 + \epsilon_i \tag{1}$$

$$\epsilon_i \sim \mathrm{N}(0, \tau^2), \tag{2}$$

where $x_{1i} = 1$ is an intercept and x_{2i}, x_{3i}, and x_{4i} are indicator variables which each equal 1 for half of the items and 0 for the remainder. Each possible combination of the indicators occurs an equal number of times, and the generating model includes one interaction, $x_{2i}x_{3i}$. Table 1 provides the design matrix for item covariates. The rows of the design matrix are repeated to accommodate $I > 8$ items.

Table 1: Items design matrix

x_1	x_2	x_3	x_4
1	0	0	0
1	0	0	1
1	0	1	0
1	0	1	1
1	1	0	0
1	1	0	1
1	1	1	0
1	1	1	1

The composite abilities are generated as

$$\theta_p = w_{1p}\gamma_1 + w_{2p}\gamma_2 + \zeta_p \tag{3}$$

$$\zeta_p \sim \mathrm{N}(0, \sigma^2), \tag{4}$$

where w_{1p} and w_{2p} are also crossed indicator variables. Table 2 presents the design matrix for person covariates, the rows of which are repeated to accommodate the P persons.

A key feature of the generated datasets (and data of this type more generally) is the extent to which the item covariates account for the composite item difficulties. To this end, let $\upsilon^2 = \mathrm{var}(x'\beta)$ represent the variance of the structural part of item difficulty. Because of the item design, υ^2 does not vary between simulated datasets, even if they have differing numbers of items (so long as I is a multiple of 8.). The total item variance is $\upsilon^2 + \tau^2$. Then

$$R^2 = \frac{\upsilon^2}{\upsilon^2 + \tau^2} \tag{5}$$

Table 2: Persons design matrix

w_1	w_2
0	0
0	1
1	0
1	1

represents the proportion of item variance accounted for by the item predictors.

The generating values for the structural part of item difficulties are $\beta = \{-.5, .5, .5, .5, -.5\}$ in all simulation conditions, and so $\upsilon^2 = 0.11$ in all conditions. Figure 1 displays R^2 as a function of τ with υ^2 fixed to this value. The points marked indicate the generating values of τ, which are $\tau \in \{0.00, 0.10, 0.30, 0.50\}$ (or equivalently, $\tau^2 \in \{0.00, 0.01, 0.09, 0.25\}$). This choice yields $R^2 \in \{1.00, 0.92, 0.55, 0.30\}$. On the person side, the generating values are fixed across conditions, with $\gamma = \{.5, .5\}$ and $\sigma = 1$. The numbers of persons and items vary: $I \in \{32, 128\}$ and $P \in \{300, 1000\}$.

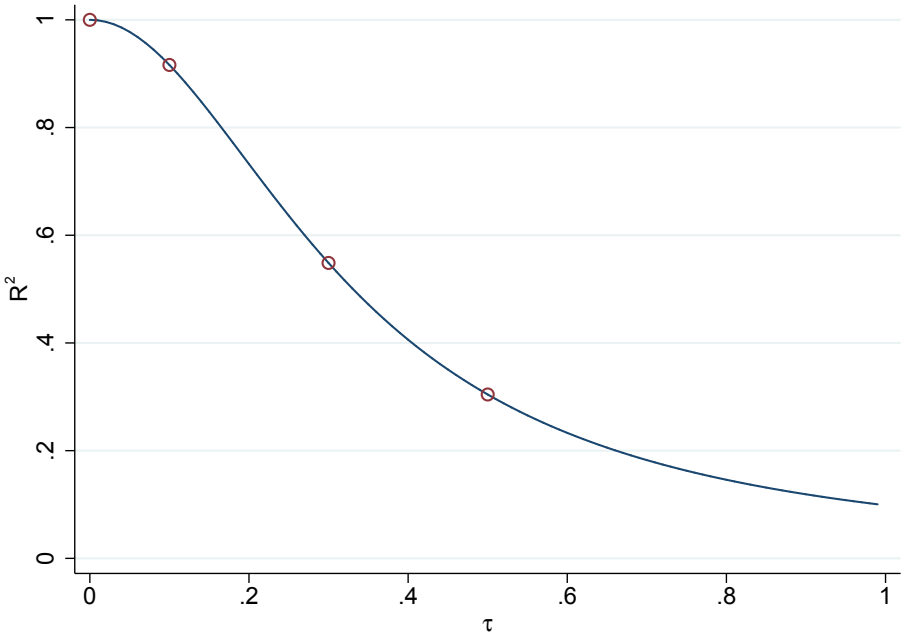

Figure 1: R^2 versus τ for the simulations. The points indicate generating values of τ.

In summary, three factors are varied between simulation conditions in a crossed design: τ (and by extention, R^2), I, and P. All other elements are fixed across conditions. Because cross-validation features prominently in this chapter, multiple datasets are simulated within each replication. A "training" dataset is created as described above, and along with it three "test" datasets are formed: one representing a sample with new items (corresponding to new draws of ϵ_i), one representing a sample with new persons (new draws of ζ_p), and the last representing a sample with both new items and new persons.

2.2 Models

Three models, differing only in specification of δ_i, are fit. Model 1 includes only the "main effects" for the item covariates:

$$\delta_i^{(1)} = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4. \tag{6}$$

Model 2 adds an interaction:

$$\delta_i^{(2)} = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{2i}x_{3i}\beta_5. \tag{7}$$

Model 3 adds an additional (spurious [*find a better word*]) interaction:

$$\delta_i^{(3)} = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{2i}x_{3i}\beta_5 + x_{3i}x_{4i}\beta_6. \tag{8}$$

None of the analysis models includes the residual ϵ_i. Each analysis model models ability as in Equations 3 and 4.

3 Naive cross-validation methods

One naive approach to model selection is the use of significance testing for parameters. In order to select among the three analysis models, a researcher may fit Model 2 and make a judgment based on the p-value for β_5, the parameter associated with the interaction. If non-significant, the researcher may select Model 1. Otherwise, the researcher may fit Model 3. If the additional interaction (β_6) is significant, Model 3 would be selected. Otherwise, Model 2 would be selected. This is a forward stepwise procedure.

Figure 2 presents the proportion of times each model was selected across conditions (combination of I, P, and τ) when selection is performed via p-values. Within each condition, 200 replications are performed. This method works well when τ is small but poorly otherwise, and this trend is similar for all combinations of P and I. When $\tau = 0$, Model 2 matches the data generating model exactly, and Models 1 and 3 are close. The result is that the correlated nature of responses within an item cluster are appropriately accounted for, yielding correct standard errors for β. The preceding is approximately true for small values of τ like $\tau = .1$. However, with greater values of τ, the analysis models fail to account for the within-cluster dependency, resulting in standard errors that are too low. This shortcoming leads to Model 3 being selected the majority of the time when τ takes medium to large values.

Adding item residuals ϵ_i to the analysis models would provide correct standard errors and p-values. Such a model is prohibitively difficult to fit without resorting to Monte Carlo methods, though. Further, in practical application there may be many more than three models under consideration, which brings up complexities around multiple hypothesis testing. For this reason an appealing alternative is AIC, defined as

$$\text{AIC} = \text{deviance} + 2k, \tag{9}$$

where k is the number of model parameters. The model with the lowest value of AIC is selected. The results of using AIC with the simulated datasets are presented in Figure 3. These results follow similar patterns as for selection by p-values, though a bit worse.

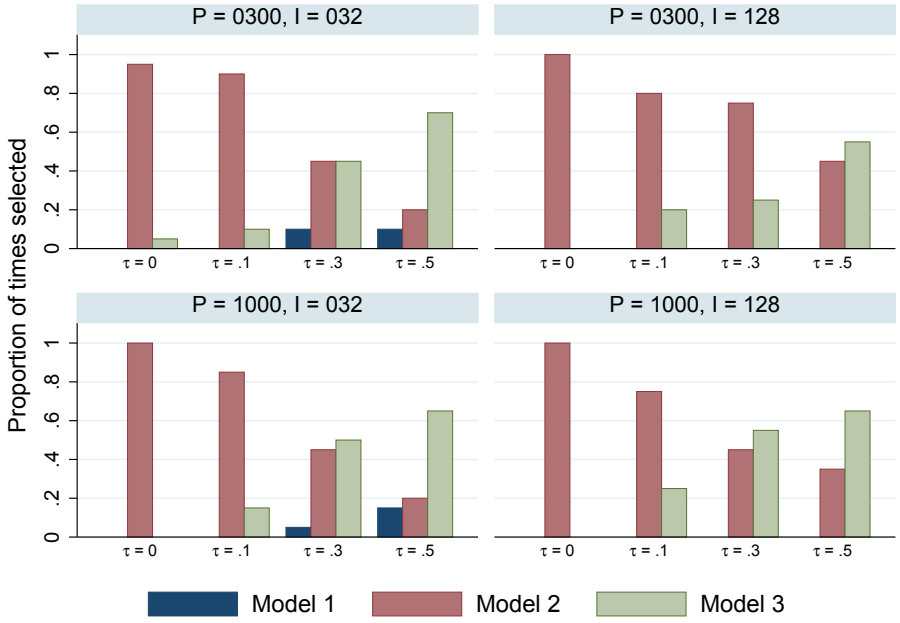

Figure 2: Proportion of times each model was selected using significance tests.

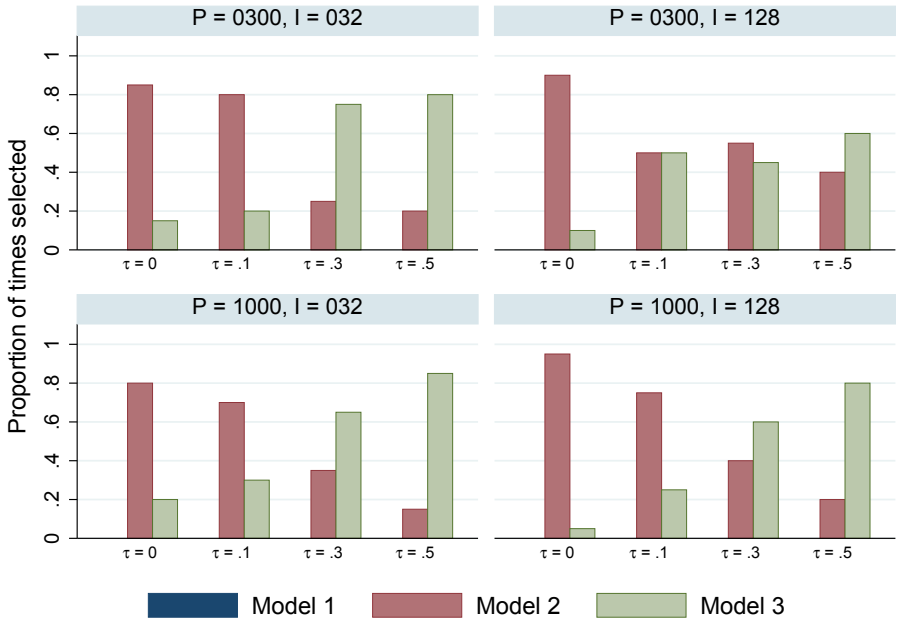

Figure 3: Proportion of times each model was selected using AIC.

AIC is an approximation for holdout cross-validation, in which a model is estimated using a "training" dataset and then evaluated on a "test" dataset. In this instance, AIC approximates the deviance that would result from applying the trained model to a test dataset consisting of new persons and the same items. Figure 4 provides results for the selection procedure using this form of holdout cross-validation. This method performs worse across all conditions than AIC.

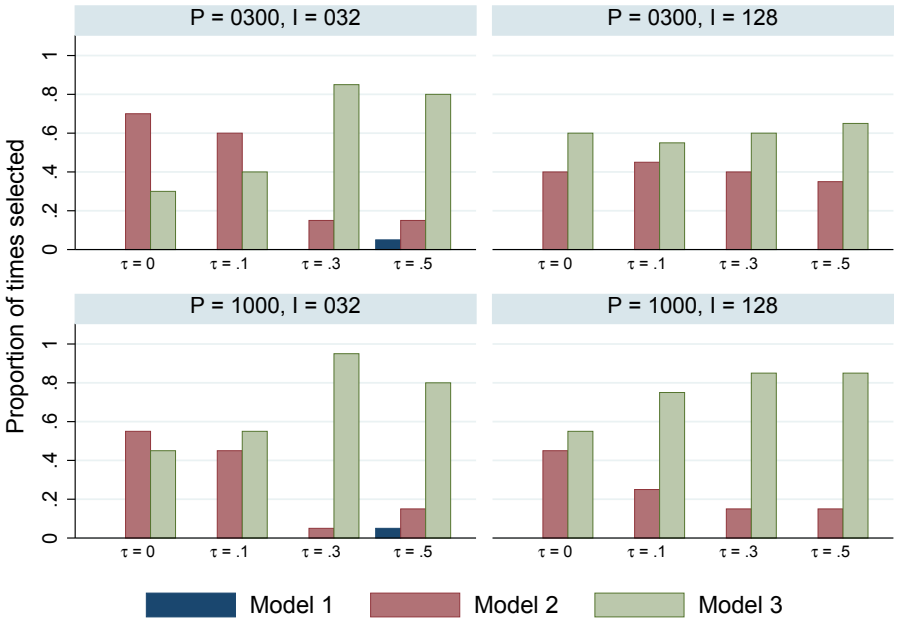

Figure 4: Proportion of times each model was selected using holdout cross-validation with holdout data consisting of new persons and the same items.

4 Holdout cross-validation for item predictors

If the focus of model selection is the choice of item predictors, cross-validation schemes based on test data with the same items are wrongheaded. A useful approach instead is to consider how the item predictors will fare for a new set of items constructed from the same item design. To this end, the analysis models are fit to a training dataset and then evaluated on a holdout dataset representing new persons and new items. Figure 5 provides the proportion of times each model was selected using this scheme. Across all conditions, Model 2 is selected the majority of times. For $I = 32$ items, larger values of τ are associated with a lower selection proportion for Model 2, while this trend is mitigated when $I = 128$.

In Figure 6 the results are broken down into pairwise comparisons: Model 1 versus Model 1 and Model 3 versus Model 2. This allows for a closer look at how Models 1 and 3 interfere with the selection of Model 2. Model 1 is selected most often over Model 2 when τ is large, and this relationship is mitigated when the number of items is large. Model 3

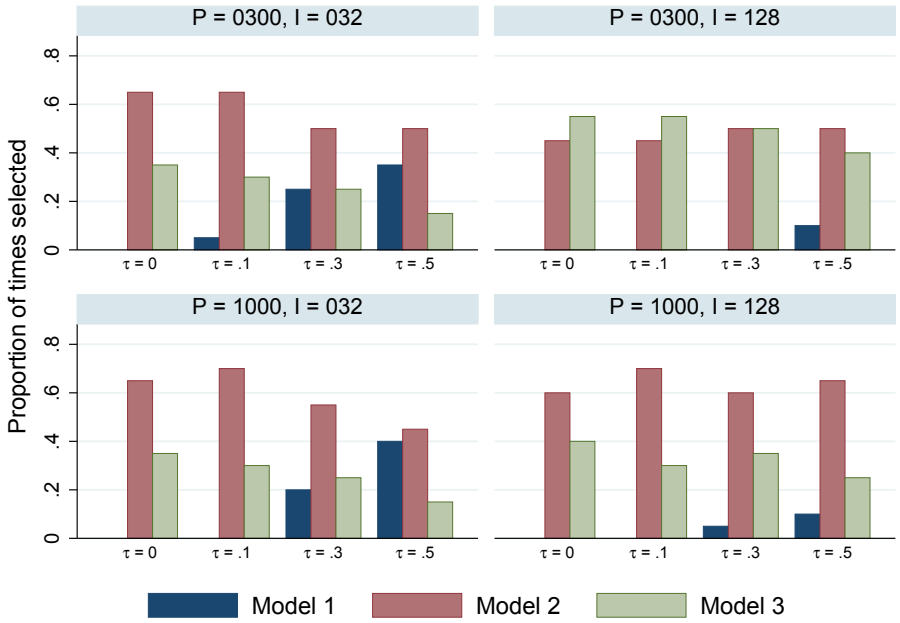

Figure 5: Proportion of times each model was selected using holdout cross-validation with holdout data consisting of new persons and new items.

is selected over Model 2 with some frequency, though there is no clear relationship between selection of Model 3 over Model 2 and the simulation conditions.

5 Approximations for a single dataset

6 Discussion

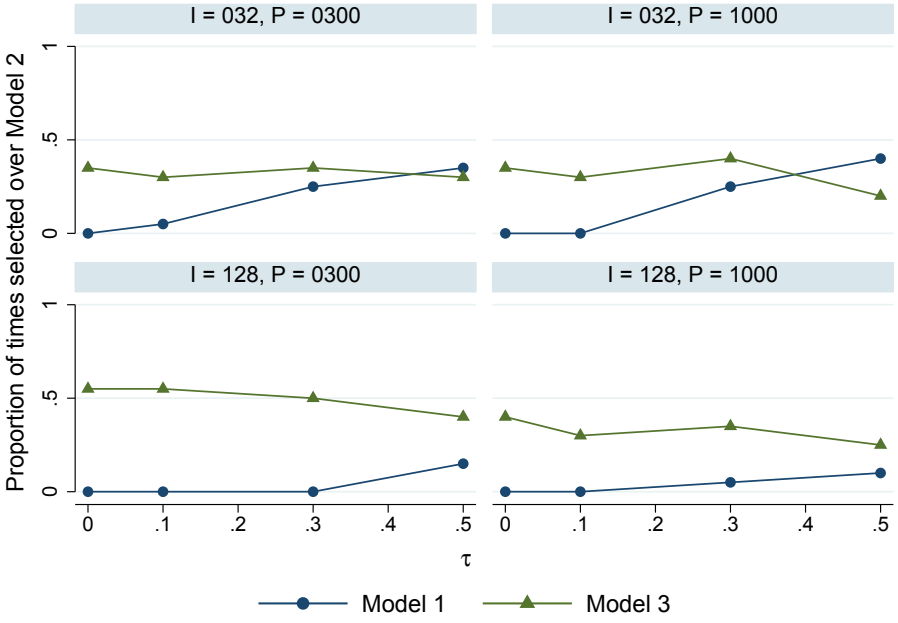

Figure 6: Pairwise comparison of selection proportions using holdout cross-validation with holdout data consisting of new persons and new items.