Predictive Accuracy of Linear Models with Ordinal Regressors

Bachelor Thesis

Fall 2015

Jim Modin Larsson
Supervisor: Shaobo Jin

UPPSALA UNIVERSITY
DEPARTMENT OF STATISTICS
Abstract

This paper considers four approaches to ordinal predictors in linear regression to evaluate how these contrast with respect to predictive accuracy. The two most typical treatments, namely, dummy coding and classic linear regression on assigned level scores are compared with two improved methods; penalized smoothed coefficients and a generalized additive model with cubic splines. A simulation study is conducted to assess all on the basis of predictive performance. Our results show that the dummy based methods surpass the numeric at low sample sizes. Although, as sample size increases, differences between the methods diminish. Tendencies of overfitting are identified among the dummy methods. We conclude by stating that the choice of method not only ought to be context driven, but done in the light of all characteristics.

Keywords: variable classification; predictive performance; model specification; discretized continuous data.
# Table of Contents

1. Introduction ................................................................................................. 4
   1.1 Research Question and Disposition....................................................... 5

2. Methodology ............................................................................................... 6
   2.1 Traditional Methods ............................................................................. 6
      2.1.1 Dummy Coding ............................................................................. 6
      2.1.2 Standard Metric Method .............................................................. 7
   2.2 Improved Methods .................................................................................. 7
      2.2.1 Cubic Splines ............................................................................... 8
      2.2.2 Penalized Coefficient Smoothing ................................................. 9
   2.3 Alternative Methods .............................................................................. 9

3. Simulation Studies ..................................................................................... 10
   3.1 Simulation Settings ............................................................................... 10
   3.2 Results .................................................................................................. 11
      3.2.1 Model Error ................................................................................ 14

4. Discussion .................................................................................................. 15

5. Conclusions ............................................................................................... 18

Acknowledgements ....................................................................................... 19
References ...................................................................................................... 19
Appendix A .................................................................................................... 21
1. Introduction

Practitioners of statistics are constantly faced with various choices of analysis, and in these situations classification of variables proves to be notably useful. Such classifications offer guidance in selection of methods and tests, thought at the same time, restricts the amount of justifiable methods. This serves as a vital foundation in the motivation of various approaches, and is also provided when new methods are proposed to argue where said methods are found appropriate. Consequently, to have a generally accepted system unites theorizing and makes statistical methods accessible. A brief description of the most common ways of classifying data will be provided but complications are also brought forth.

To determine the preciseness of a variable, a level of measurement is often assessed following the classification scheme first introduced by Stevens (1946), that is; nominal, ordinal, interval and ratio, ranging from the weakest level of measurement to the highest. Demands on certain mathematical properties are placed when moving up successive levels, ordinal scale requires possible ordering of categories, interval scale needs a valid measurement of distance between the categories/values and ratio-scale possess a true or absolute zero. These levels are cumulative, i.e. a variable on interval scale maintain the properties as those of nominal and ordinal, which in turn supports a researcher to analyse data from the perspective of a lower level (Kleinbaum et al., 2007). Variables on nominal and ordinal scale are referred to as qualitative or categorical, while variables on interval and ratio scale are numerical.

Problems arise when one encounters situations where variables are felt to be in-between typologies, the difference between ordinal and interval in particular have received some controversy. The debate of how to analyse these have gone so far as to create two camps of statisticians, with Stevens himself giving small acknowledgment to the side questioning his initial guidelines of not treating ordinal data as numerical, “for this ‘illegal’ statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results.” (Stevens, 1946, p.679). Knapp (1990) provides a valuable commentary on the recent views of the matter. This dispute is the effect of classification schemes with no allowance for grey areas between types.
Taking this overlap into account, there has been a following debate regarding the appropriate way of including data on ordinal scale in regression analysis. However, much of the attention has been directed towards ordinal response variables, see McCullagh (1980) for the starting point of many developments, and Liu & Agresti (2005) for a review of various approaches. Whereas, the case of ordinal predictors is left somewhat in the shadows as of up to recent years, with upspring in testing and establishment of new techniques. Greene & Hensher (2010) provide a collection of contributions from a number of disciplines, while Pedhazur (1997) argues that many general techniques for categorical variables can be utilized with ordinal data. Standard approaches are still, ever so often, being applied without deliberation. Be that as it may, it is important that we have an understanding for the conceivable consequences of all techniques, despite them being regarded as standard or not.

The two most typical treatments of ordinal predictors are dichotomous dummy coding (Walter et al., 1987) and simple straight-line regression, originally designed for numeric variables. While both are practically sound and easily taught, applications like these have shown to regularly overfit data, resulting in poor predictive performance (Babyak, 2004). In regard to this, along with other aspects, Gertheiss & Tutz (2009) proposed a penalized dummy based method of which they found to surpass the traditional techniques on delivering good predictions. Advancements to the metric framework when handling variables comparable to categorical ones have also been made with assorted smoothing suggestions (Green & Silverman, 1994; Wahba, 1980; Härdle, 1992, among others). In particular, models of additive nature have shown promising results (Hastie & Tibshirani, 1986; 1990). Gertheiss et al. (2014) extended the case to explicitly categorical covariates.

1.1 Research Question and Disposition

The objective of this paper is to evaluate predictive performance of the traditional methods regarding ordinal predictors, together with two improved methods, seen as advancements of respective traditional approach. They will all be accounted for solely on their predictive accuracy under identical conditions, taking the possibility of overfitting into consideration. Capability is extended to the case where all linear models are approximations of the truth, with a varying degree of information concerning unidentified variables.
The paper is organized as follows; Section 2 outlines the theoretical framework of the four methods of handling ordinal predictors that are to be compared. Section 3 conducts a simulation study to investigate predictive accuracy of the four methods. This is followed by a discussion of how the techniques contrast in section 4. Lastly, section 5 gives a summarization of the work and results. Limitations and suggestions for future research are included. Supplementary results are presented in Appendix A.

2. Methodology

In this section, four methods are briefly reviewed as approaches to the inclusion of ordinal predictors in linear models. Two traditional methods and two improved methods, one of which is thoroughly accounted for in Wood (2006) and the other recently proposed in Gertheiss & Tutz (2009).

2.1 Traditional Methods

2.1.1 Dummy Coding

To represent a multi-categorical ordinal predictor in regression analysis, the coding scheme introduced by Walter et al. (1987) has been widely adopted. Originally designed for biostatistics, it is now prominent across all disciplines of statistics. The scheme operates generally by including \( k - 1 \) dummy variables accompanied by an intercept to automatically index all \( k \) categories of an ordinal (or nominal) variable, and is thus able to identify contrasts in the response variable between successive levels of the independent variable. Given a metric response, the following linear model for a single categorical variable is,

\[
y = \alpha + \sum_{k=1}^{k-1} \beta_k x_k + \varepsilon, \quad iid. \varepsilon \sim N(0, \sigma^2)
\]

where \( x_k \) only takes the values of 1 or 0 to indicate category association. This type of model does not take into account the ordinal structure of the data, resulting in a loss of information, and \( k - 1 \) variables are now used to represent a single ordinal variable which in turn consumes degrees of freedom. In cases where categories cannot be properly differentiated, observations can be further grouped by merging categories. This approach is commonly introduced in
textbooks on regression analysis, on both basic and advanced level, *eg.* Kleinbaum *et al.* (2007) and Gelman & Hill (2006).

### 2.1.2 Standard Metric Method

In contrast to the use of dummy variables, there is the case of regarding a categorical variable as numerical in the sense of legitimizing methods designed for continuous variables. The clear-cut way of doing this is to perform classic linear regression, ordinary least-squares, on assigned level scores for ordered categories. Creating the plain linear model formation,

\[ y = \alpha + \beta x + \varepsilon, \quad iid. \ v \sim N(0, \sigma^2) \]

where \( x \) takes on the values of the assigned level scores. If the levels correspond to interval subgroups, midpoints of these could be analysed as well (Diamantopoulos & Schlegelmilch, 1997). However, uncertainty of the exact values will not be addressed directly, nor will open endpoint limits be. Compared to the dummy variable model, only one predictor is used to express each respective input variable.

Situations where data on ordinal scale often is treated as numeric and referred to as originally continuous or discrete is in survey studies, with classic examples such as age or income (Mare & Winship, 1984). This is a less accepted treatment but frequently used in practice, especially for categorical variables with high number of levels or when the scale can be considered linear (Pasta, 2009).

### 2.2 Improved Methods

Comparisons of coefficient estimates of dummy variables against metric on level scores have received some attention, see Johnson (2009), Gertheiss & Tutz (2009) among others, and these studies show that while dummy estimates frequently make skipping-like patterns, metric estimates are smooth by cause of their linear restrictions. As an effect of this, metric estimates are not always able to capture variation in separate categories while dummy estimates have tendencies to be over reactive. Both are seen to perform alike under the condition of linear relationship between the ordinal predictor and the response variable (Agresti, 2010).
2.2.1 Cubic Splines

If we pursue a more receptive estimation technique compared to classic linear regression, assuming that the categorical variable originally was of metric structure, we have to ease the linear restrictions imposed. This can be done by utilizing the assigned level scores in a generalized additive model (GAM) but applying a smoother to diverge from the “skipping” noticed in dummy coefficient estimates (Eubank, 1999). It can be shown that cubic splines is the smoothest possible interpolant over all potential datasets (Wood, 2006; Gu, 2002), and thus serves as a good choice of smoothing function.

A model containing one smooth function of one covariate takes the form:

\[ y = \alpha + f(x) + \epsilon, \quad iid. \ \epsilon \sim N(0, \sigma^2) \]

With the basis function to sustain linearity of the above model formation,

\[ f(x) = \sum_{i=1}^{q} \gamma_i(x) \beta_i \]

for values on the unknown parameter \( \beta_i \). Let \( \gamma_i(x) \) be defined by cubic polynomials such that it is continuous in values, first and second derivatives to produce a curve smooth in sections, known as a cubic spline. A control of smoothness is imposed by penalizing the function \( f \) by the second derivatives over the complete interval in order to avoid overfitting, and a smoothing parameter, \( \lambda \), is set so that the spline estimate \( \hat{f} \) is as close as possible to the hypothesized true function \( f \) (see Wood, 2006, for further details). Additionally, a natural boundary condition is established by setting the endpoints second derivatives to zero.

Criticism against the use of cubic regression splines generally rests on the fact that this technique inevitably assumes an underlying continuous feature for variables on ordinal scale, which is further discussed in section 4. This is why it is also rejected as an alternative method in Gertheiss & Tutz (2009), instead suggesting a spline technique penalized to be made discrete (Eilers & Marx, 1996) for categorical predictors.
2.2.2 Penalized Coefficient Smoothing

Returning to the instance of dummy coding, Gertheiss & Tutz (2009) developed a method including penalization functions to adhere differences between coefficients of adjacent categories for an ordinal predictor. The preference is a smoother coefficient vector. This method also recognizes ordering of categories, and therefore, differentiates factors from those of nominal scale. For easier description, the linear regression is given in matrix notation,

\[ y = \alpha + X\beta + \epsilon, \quad iid. \, \epsilon \sim N(0, \sigma^2) \]

where \( X \) denotes the \( N \times (k - 1) \) design matrix with full rank and the response, intercept, and error component are given in respective vectors. A generalized ridge estimator

\[ \hat{\beta}^{**} = (X^T X + \lambda^* \Omega)^{-1} X^T y \]

is obtained, with the penalty matrix \( \Omega = U^T U \) \( [(U)_{ii} = 1, (U)_{i+1,i} = -1] \) and tuning parameter \( \lambda^* = \psi \sigma^2 \) (see Gertheiss & Tutz, 2009, p.348 for further details).

The tuning parameter is chosen by minimizing a corrected version of the Akaike Information Criterion (Hurvich et al., 1998) within the framework of possible \( \lambda^* \)-values. Gertheiss & Tutz (2009) further demonstrate that this base for selection of tuning parameters performs very well comparing to theoretically optimal values of \( \lambda^* \).

2.3 Alternative Methods

Other possibilities to address ordinal predictors in regression analysis are those based of shrinkage methods such as; ridge regression (Hoerl & Kennard, 1970), group lasso (Lin & Yuan, 2006) or sparse modelling (Gertheiss & Tutz, 2010), with the general objective of lowering sample variance (e.g. when a high degree of multicollinearity is present) at the trade-off of unbiasedness. Alternatively, one can fit a curve under monotonic constrains when an order is expected, known as monotonic/isotonic regression (Barlow et al., 1972), with the benefit of making no assumptions regarding the form of the function.
3. Simulation Studies

Analyses will be performed under the model formation of an infinite-order regression, based on the rationale that the assumption of correct model specification is unverifiable. It is solely in simulation studies that we can know for sure that the proposed model accurately depicts suggested relationships (Funk et al., 2010). Thus, the following constructions of models will inevitably be misspecified by intention. We can however measure how different methods of estimation compare under these conditions. The application of an infinite-order also serves a second purpose in that it allows us to control an expansion of an effect to the direct relationship between our predictor of interest and the response variable, which are initially set to have a strict linear dependency.

To illustrate the issue of method selection for inclusion of ordinal data as predictors, discretized continuous variables will be the centre for testing. This is by reason of their inherent metric structure that are straightforward to conceptualize, although commonly labelled as ordinal if classification is to be carried out. Continuous variables will be partitioned into $k=5$ subintervals, without the restrain of equal amount of observations in each factor.

3.1 Simulation Settings

An infinite-order model is used in Hansen (2007) and Hansen & Racine (2012), as a basis for comparison of estimators. Our formation is adapted from their simulation design. The true model is,

$$y_i = \beta x_i^* + \sum_{j=1}^{\infty} \theta_j z_{ji} + \epsilon_i,$$

with the error component $\epsilon_i \sim N(0, 1)$, independent of $x_i^* \sim N(0, 1)$ and $z_{ji} \sim N(0, 1)$. An intercept is included in the model by setting $z_{1i}$ equal to 1. The theta-coefficients are determined by the rule $\theta_{j+1} = c\sqrt{2}a j^{-(a+\frac{1}{2})}$, $j = 1, 2, \ldots$, where a larger value on the parameter $a$ marks an accelerated rate of decline of the coefficients $\theta_j$ with $j$. However, only a subset of $z$ variables are used to fit the linear regression model, which mimics the case of incomplete information. In addition, we let $\beta = \theta_1 = \theta_2$. A pair-wise correlation of 0.4 between all predictors is fixed.
The first predictor $x_i^*$ is separated from the rest, as this variable will be accounted for with the four methods specified earlier. It is therefore latter partitioned into five subintervals and the original continuous covariate, $x_i^*$, is considered not observed. The boundary values to determine factor level or group association of an observation are determined by the percentiles of the standard normal distribution with cumulative probabilities 0, 0.2, 0.4, 0.7, 0.85, 1. Score values of 1 to 5 are assigned to each observation for respective factor level in rising order.

Three sample sizes are used here, \textit{i.e.} $n = 50, 200$ and 400. The number of $z$ predictors included in each examined model construction is determined by $3n^{1/3}$. The parameter $a$ is set to three levels, 0.5, 1 and 1.5. Lastly, the values of the parameter $c$ is set so that the theoretical population coefficient of determination, $R^2$, varies on a grid of 0.1 to 0.9 with 20 values of exact equal spacing. 3000 repetitions over all combinations of conditions are executed.

Examined here is what can be viewed as general applications of the two improved methods. Selection of smoothing parameter for the cubic splines, as well as, the tuning parameter of smoothed coefficients are determined by 5-fold cross validation over a set of plausible values on the criteria given in the methodology section. Programming is built upon the R packages; \textit{mgcv} (Wood, 2006) for GAM splines and \textit{ordPens} (Gertheiss, 2015) for penalized coefficient smoothing.

3.2 Results

To examine how the four methods contrast with respect to predictive performance, a mean squared error (MSE) is calculated and averaged over all iterations of sets of conditions on the formula:

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

These values are solely used for comparison purposes to determine the extent to which each model fits identical sample data points, and thus provide an approximation of accuracy, or if viewed the other way around, a loss function. The resulting MSE computations when $a = 0.5$ are graphically displayed in figure 1. The plots are further divided by sample sizes. $a = 1$ and 1.5 produce a similar pattern to that of $a = 0.5$, therefore, results for $a = 1$ and 1.5 are not displayed here but are presented in Appendix A.
Main similarities over all figures show a relative increase in MSE with $R^2$. This is anticipated as even though the proportion between the model construction and the noise alternate, larger influence is given to the predictors which in turn raise coefficient values and subsequently enables prediction errors of larger magnitude. Moreover, as a higher sample size is employed, the relative difference in MSE of each method becomes smaller, although union is halted by a low value on the parameter $a$.

Considering the case where $n = 50$ in figure 1, notable MSE differences for the four methods are implied. The order is the same for all levels of $a$, but successively apparent as $a$ increases. Methods using dummy variables outperforms those considering the assigned level scores as numeric, with traditional dummy coding producing a lower MSE on average than
smoothed dummies. The difference between the two is relatively large at low values of $R^2$ but smoothed dummies show to improve with $R^2$ in relation to the other methods. GAM with cubic splines comes ahead of classic linear regression in terms of lower MSE and stays practically parallel in growth.

As a substitute to incorporate confidence bounds in the MSE figures, variation in MSE estimates is presented through a second summary statistic, as deviation from the mean. Smoothed dummies are noticed to have the largest overall variance moving along $R^2$ points as shown in figure 2, and as a consequence of this, give rise to the largest range of values among the methods. This should be taken into account when interpreting MSE averages in that both better and worse results for each combination of conditions are produced.

Figure 2. Standard deviations from the average MSE at the sample size of $n = 50$
3.2.1 Model Error

Given that we know our test models are misspecified, a measurement for estimating how well each of the four model constructions describe the proposed true model for each combination of conditions are calculated. As the error component from the true model has no presence, we can be inclined to think of functions that are possibly too closely fit to the set of data points given, rather than the correct model, and receive an indication of overfit. Which can have fateful consequences in the presence of inaccurate data that would spoil the model. Let $\eta_i = \beta x_i^* + \sum_{j=1}^{\infty} \theta_j z_{ji}$ and define:

$$\text{Model Error} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \eta_i)^2$$

The results are displayed in figure 3 when $a = 0.5$. Similar to the prediction error, $a = 1$ and 1.5 demonstrate matching patterns to that of $a = 0.5$, and thus, are not displayed here but are placed in Appendix A. As the range of model error values vary greatly between sample sizes, individual scaling of the y-axis is employed.

It is evident from figure 3 that the numeric methods match the true model composition closer compared to the dummy based methods, with the exception of smoothed dummies at the low end of $R^2$, where it is shown to depict the true model very closely. The results are displaying how the dummy based methods in general are more receptive to actual data points, and eventual noise, which is in line with theory. This is most likely also an effect of introducing several predictors to indexes all factors of the discretized continuous variables, with a relatively low observation per predictor ratio at a sample size of 50. Observed is a direct reverse case to that of figure 1, indicating how model fit not necessarily goes hand in hand with predictive accuracy, measured by MSE of prediction.
Figure 3. Model error of each respective model when $a = 0.5$

4. Discussion

It shall be taken into consideration that the analyses are performed on discretized continuous variables of five levels and that the count of levels and subsequent factors may have an effect on performance of the methods. There are existing mechanisms for arriving at the best bins being developed when partitioning a variable, see Fayyad et al. (1993) as an example. Furthermore, it is intuitive that analysts favour the use of methods designed for metric variables on an ordinal variable of many levels and vice versa. The approach here, however,
differs in the way that we seek to portray ordinal variables encountered in everyday life rather than finding an optimal representation for the particular data at hand. Commonly encountered in survey studies are variables partitioned into a level of options ranging from three to nine in order to create exhaustive answering possibilities with symmetry, but also for practical purposes. It is in regard to this, pursuing a way to mirror credible cases of real life, $k$ is set to five.

Nonetheless, under the circumstances created here, arguments can be made for the preference of dummy based methods upon inclusion of ordinal predictors or more specifically; grouped intervals, when small samples are considered. The same support for selection is not granted moving up in sample size as the relative difference in MSE between the methods is practically very small. Observe that these statements only are relevant under similar conditions as assumed in this study.

Noteworthy and crucial to discuss is the fact that the results are not consistent with previous research, concerning the two dummy based methods in particular. Penalized smoothed coefficients are partly constructed to reduce the risk of overfitting, which is commonly encountered using traditional dummy schemes, also indicated by figure 3, and shall therefore in theory lead to better predictive performance. This is shown by Gertheiss & Tutz (2009) when introducing the method through several tests in both a Bayesian and frequentist perspective, as well as, an application to real world data including evaluation on MSE of prediction. In their study, attention is however directed towards analysing ordinal variables with a higher number of levels, 11, mostly isolated with close awareness to coefficient values. While applied here, are ordinal variables with a strict count of levels, 5, and under the influence of outside effects. Their data generating process also differs in the way that the response values are built from generated dummy variables, whereas our responses are established from continuous variables alone. At low theoretical $R^2$ where the noise has greater influence on the response variable, it is not surprising that a more receptive method to this performs rather well delivering matching predictions.

Turning to the numeric methods, it is clear that GAM with cubic splines tends not to perform worse than simple linear regression in respect to predictive accuracy. This is not a surprise as cubic splines natural boundary condition better handles eventual extreme endpoints in that a linear function is enforced even beyond these, which in turn imposes additional smoothing constraints inside the range. The approximate cross validation also
comes in handy creating an adaptive function, where, if needed, the smoothing parameter would tend to infinite resulting in a practically straight line. Although it is seemingly not always inappropriate to implement numeric methods on ordinal data based on the results here, researchers ought to be cautious at small sample sizes as substantial differences compared to the dummy based methods have been shown possible.

In this study, we have not accounted for methods being considered appropriate or not depending on variable classification restrictions discussed earlier. This is with the aim to achieve an open view of how each model construction functions without limiting the work by fixed perspectives of typologies.

So far, the four approaches are only evaluated on MSE and model error, this should by no means be seen as a ruler for deciding if one method is strictly superior compared to another. Other qualities such as interpretation value, unbiasedness and goodness of fit must then be taken into account. Certain characteristics could particularly be ranked high in perspective of various scenarios. Penalized coefficient smoothing has the pleasant feature that it is reasonably easy to interpret coefficient values for categories on ordinal scale as adjacent factors are taken into account, while traditional dummy coding schemes have tendencies to form random-like patterns as each factor stands by itself. Based on the results here however, what is gained in interpretation value for penalized coefficient smoothing is a seemingly trade-off of accuracy at low sample sizes. Simple linear regression can be argued not adequately describing relationships in categorical data as these are expressed in general trends. GAM with cubic splines, on the other hand, delivers independent coefficients open for interpretation for all sections of knots.
5. Conclusions

We have evaluated four linear regression models available for utilizing ordinal data as input on the basis of prediction accuracy. The variable of interest has been treated in different perspectives; strictly categorical with the use of dummy variables on both nominal and ordinal level, or as numeric with the motivation of an underlying continuous structure. This is by reason of no unanimous answer to variable classification, and subsequently, correct method usage. The aim is not to find an overall superior way for the inclusion of ordinal predictors, but rather to accentuate the effects of different applications, here with respect to predictive performance and model specification under the influence of an effect diverging from direct linearity. The results suggest disparity among the models markedly in small samples, for which our recommendation is that researchers are best to estimate several models and weigh desirable qualities for selection. Furthermore, there is no need to be straight-out reluctant to the application of metric methods on data classified as ordinal, as promising results have been shown. Model formation should, however, be in line with the aim of one’s research or investigation.

There are, however, a number of unavoidable limitations imposed for this study. Firstly, the work is concentrated towards analysing discretized continuous variables and not ordinal variables in general which makes generalization to the grand scheme difficult. Secondly, the lack of application to real world data abbreviates the study to a strictly theoretical work. And finally, only two measures are used to evaluate how the methods contrast, clearly not accounting for the entire picture at once, instead providing a compliment to previous research.

Suggestions for future research in light of what has been discussed throughout this paper begins with expansion of the theoretical framework in new context. This could be done with the addition of new constructs to model formation, as varying results have been shown regarding method performance under shifting conditions. Alternatively, by assessing particular aspects of the current theoretical models where focus has not been for this study such as model selection, multicollinearity issues, stability associated with the error components and model complexity/parsimony. Continued testing of features associated to each method is crucial as the field of statistics progresses, new methods must also be thoroughly examined.
Acknowledgements

I would like to take this opportunity to express my gratitude to my supervisor Shaobo Jin for his extraordinary guidance and support throughout the entire process.

References


Appendix A

Figure A1. Average mean squared error of all four methods when $a = 1$
Figure A2. Average mean squared error of all four methods when $a = 1.5$
Figure A3. Model error of each respective model when $a = 1$
Figure A4. Model error of each respective model when $a = 1.5$