Maria Iannario*, Anna Clara Monti and Pietro Scalera

The number of response categories in ordered response models

https://doi.org/10.1515/ijb-2021-0013
Received August 5, 2020; accepted August 23, 2021; published online September 21, 2021

Abstract: The choice of the number m of response categories is a crucial issue in categorization of a continuous response. The paper exploits the Proportional Odds Models' property which allows to generate ordinal responses with a different number of categories from the same underlying variable. It investigates the asymptotic efficiency of the estimators of the regression coefficients and the accuracy of the derived inferential procedures when m varies. The analysis is based on models with closed-form information matrices so that the asymptotic efficiency can be analytically evaluated without need of simulations. The paper proves that a finer categorization augments the information content of the data and consequently shows that the asymptotic efficiency and the power of the tests on the regression coefficients increase with m. The impact of the loss of information produced by merging categories on the efficiency of the estimators is also considered, highlighting its risks especially when performed in its extreme form of dichotomization. Furthermore, the appropriate value of m for various sample sizes is explored, pointing out that a large number of categories can offset the limited amount of information of a small sample by a better quality of the data. Finally, two case studies on the quality of life of chemotherapy patients and on the perception of pain, based on discretized continuous scales, illustrate the main findings of the paper.

Keywords: collapsibility property; efficiency; hypothesis testing; merging categories; proportional odds models; sample size.

1 Introduction

A critical point in surveys with rating questions is the choice of the number m of response categories to use in the discretization of a measurement obtained on a continuous scale (in which the only marks are those related to the minimum and the maximum level). Although categorization of continuous measurements implies a loss of information, it is a widespread practice in various fields, such as medicine and epidemiology for instance, where researchers often split a continuous scale into ordered categories to make interpretation of the results easy. Examples are in [1–5], among others, where categorization is performed without prearranged meaningful categories. Furthermore [6] underline the benefits of transforming continuous responses in ordinal categories when the measurement variable is skewed because ordinal response models handle floor and ceiling effects better than linear models.

Within the statistical literature the choice of m is discussed by [7] who studies the beneficial impact of an increasing number of categories on standard errors. [8] instead points out that a large m allows a more

***Corresponding author**: **Maria Iannario**, Department of Political Sciences, University of Naples Federico II, Napoli, Italy, E-mail: maria.iannario@unina.it. https://orcid.org/0000-0002-2646-9937
Anna Clara Monti, Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Benevento, Italy, E-mail: acmonti@unisannio.it
Pietro Scalera, Department of Political Sciences, University of Naples Federico II, Napoli, Italy, E-mail: pietroscalera@hotmail.it

powerful detection of associations between variables; a result confirmed by [9] with reference to tests on differential item functioning. Furthermore [10] show that a larger value of m reduces the impact of response errors on the local robustness properties of the estimators in the modeling framework for ordinal data denoted as CUB models [11].

The current paper investigates the impact that the choice of m has on the efficiency of the estimators in case of discretization of a continuous response variable in data analysis (in Section 4) performed through a proportional odds model (POM) [12, 13]. The latter naturally arises when the rating is supposed to be driven by an underlying continuous variable. Each rating corresponds to an interval on the support of this variable. Choosing m is equivalent to deciding in how many classes the support is to be partitioned. With respect to alternative modeling frameworks, the POM is extremely parsimonious and is, by far, the most widely applied model in the biomedical context (see [14–16], among others).

Closely related to the choice of the number of categories, combining values/scores by collapsing adjacent categories of an ordinal response represents another relevant issue and also a widespread practice which arises in processing sample information after data collection. It may be pursued for overcoming sparseness problems, for simplifying interpretation or dealing with *extreme response styles* [17]. For a given dataset, changing the number of categories can affect the inferential results obtained from the data [18, 19]. Other studies focus on merging categories to reduce the size of contingency tables by using the *homogeneity* of the corresponding rows (or columns) or the *structure* criterion (see [20] and reference therein). Section 6 of the current paper shows that collapsing categories – in case of a univariate ordinal response – reduces the information content of the sample generating a loss of efficiency which becomes extremely high in case of dichotomization.

Another critical point in data analysis concerns the appropriate number of categories with respect to a given sample size n [21]. Section 7 shows that increasing the number of categories enhances the efficiency of the estimators even if n is small. The relationship between m and n represents a crucial point discussed also in close field related to the association among categorical variables (see [20, 22, 23], among others).

In summary the paper handles three topics regarding data analysis of discretized continuous variables and ordinal variables: the choice of the number of response categories, the consequences of collapsing categories, and the relationship between m and the sample size n and it is organized as follows. The next Section provides a brief overview of the POM, whereas Section 3 describes the models used for the analysis. The information matrices of these models are analytically derivable so that the evaluation of the asymptotic efficiency of the estimators of the regression coefficients can be carried out without need of simulations. The impact of the choice of the number of response categories on efficiency and on hypotheses testing is investigated in Sections 4 and 5. The effect of various forms of merging categories is examined in Section 6, and the relationship between the number of categories and the sample size is analyzed in Section 7. In Section 8 two case studies related to the medical context illustrate the main findings of the paper. The first is about the perceived health-related quality of life of chemotherapy patients, whereas the second one deals with perceived pain by women during labor. Final remarks end the paper.

2 Theoretical background

In the POM framework, the ordinal response Y depends on an underlying continuous variable Y^* through the relationship

$$Y = j \iff \alpha_{j-1} < Y^* \leq \alpha_j, \quad j = 1, 2, \dots, m, \tag{2.1}$$

where $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_m = +\infty$ are the thresholds of the support of Y^*. The choice of m determines in how many intervals the support of Y^* is divided.

The variable Y^*, in turn, depends on $p \geq 1$ covariates, so that for the ith statistical unit we have the regression model

$$Y_i^* = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p + \epsilon_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \dots, n, \tag{2.2}$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ and ϵ_i is a random variable whose distribution function is denoted by $G(\epsilon)$.

Formula (2.1) implies that from the same underlying variable Y^* a countable set of ordinal variables $\left\{Y^{(m)}\right\}$ can be generated by allowing m to vary in $\{3, 4, \dots\}$. These variables differ from each other for the number of categories. Nevertheless all of them refer to the same regression model (2.2) and therefore the different estimators of the regression coefficients, which are obtained by varying m, estimate always the same $\boldsymbol{\beta}$. This property – known as *invariance to choice of response categories* ([8]; p. 56) or *collapsibility* ([17]; p. 255) - of the POM is exploited to analyze how the choice of m affects the efficiency of the estimators and the accuracy of the derived inferential procedures.

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ be the parameter vector, where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{m-1})'$ is the vector of the thresholds. Given an observed random sample $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \dots, n$, the log-likelihood function is $\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$ with individual term

$$\ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) = \sum_{j=1}^{m} I[y_i = j] \log P(Y_i = j|\boldsymbol{x}_i),$$

where $I[\omega]$ is an indicator function which takes value 1 when ω holds and 0 otherwise, and $P(Y_i = j|\boldsymbol{x}_i) = P\left(\alpha_{j-1} < Y_i^* \le \alpha_j|\boldsymbol{x}_i\right) = G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})$, for $j = 1, 2, \dots, m$. The score function is $S_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} S(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$, where

$$S(\boldsymbol{\theta}, y_i, \boldsymbol{x}_i) = \frac{\partial l(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{m} I[y_i = j] \frac{1}{P(Y_i = j|\boldsymbol{x}_i)} \frac{\partial P(Y_i = j|\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}},$$

and $\partial P(Y_i = j|\boldsymbol{x}_i)/\partial\boldsymbol{\theta} = \{\partial P(Y_i = j|\boldsymbol{x}_i)/\partial\boldsymbol{\alpha}, \partial P(Y_i = j|\boldsymbol{x}_i)/\partial\boldsymbol{\beta}\}'$ (see [24] for the analytic expression of $S(\boldsymbol{\theta}, y_i, \boldsymbol{x}_i)$). The maximum likelihood estimator (MLE) is the solution $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ of $S_n(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$.

The generic term of the information matrix $\mathcal{I}(\boldsymbol{\theta}, \boldsymbol{X})$ for a single statistical unit, conditionally on $\boldsymbol{X} = \boldsymbol{x}$, is given by

$$\begin{aligned}\mathcal{I}_{rs}(\boldsymbol{\theta}, \boldsymbol{x}) &= E_Y \left\{ \frac{\partial \ell(\boldsymbol{\theta}, Y, \boldsymbol{X})}{\partial \theta_r} \frac{\partial \ell(\boldsymbol{\theta}, Y, \boldsymbol{X})}{\partial \theta_s} \middle| \boldsymbol{X} = \boldsymbol{x} \right\} \\ &= \sum_{y=1}^{m} S_r(\boldsymbol{\theta}; y, \boldsymbol{x}) S_s(\boldsymbol{\theta}; y, \boldsymbol{x}) P(Y = y|\boldsymbol{x}), \quad (r, s) = 1, 2, \dots, m + p - 1,\end{aligned}$$

where $S_r(\boldsymbol{\theta}; y, \boldsymbol{x})$ is the element of the score function related to the r-th element θ_r of $\boldsymbol{\theta}$. The elements of the unconditional information matrix $\mathcal{I}(\boldsymbol{\theta})$ are given by

$$\mathcal{I}_{rs}(\boldsymbol{\theta}) = E_{\boldsymbol{X}} \{\mathcal{I}_{rs}(\boldsymbol{\theta}, \boldsymbol{X})\}, \quad (r, s) = 1, 2, \dots, m + p - 1. \tag{2.3}$$

The asymptotic variance-covariance matrix of the MLEs is $\mathcal{I}(\boldsymbol{\theta})^{-1}$.

Asymptotically we have $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N\left(\boldsymbol{0}, \mathcal{I}(\boldsymbol{\theta})^{-1}\right)$. In particular for the estimator $\hat{\beta}_k$ of the single regression coefficient β_k we have

$$\sqrt{n}(\hat{\beta}_k - \beta_k) \to N\left(0, \mathcal{I}(\boldsymbol{\theta})^{\beta_k\beta_k}\right), \tag{2.4}$$

where $\mathcal{I}(\boldsymbol{\theta})^{\beta_k\beta_k}$ is the element on the diagonal of $\mathcal{I}(\boldsymbol{\theta})^{-1}$ corresponding to β_k.

3 The models

To investigate the asymptotic efficiency of the estimators of the regression coefficients when m varies, we focus on models whose (unconditional) information matrix can be analytically derived through (2.3). In particular we consider the following three underlying regression models.

– **Model 1** *(with a continuous covariate)*. The variable Y^* depends on a continuous covariate $Y^* = X\beta + \epsilon$, where $X \sim N(0,1)$ and $\beta = 1.5$. The information matrix is given by

$$\mathcal{I}(\theta) = \int_{\mathcal{R}} \mathcal{I}(\theta; x)\phi(x)\mathrm{d}x,$$

where $\phi(\cdot)$ is the standard normal density function.

– **Model 2** *(with dichotomous covariates)*. The variable Y^* depends on two dichotomous covariates $Y^* = X_1\beta_1 + X_2\beta_2 + \epsilon$, where $X_1 \sim Ber(0.5)$, $X_2 \sim Ber(0.25)$ and X_1 and X_2 are mutually independent. The regression coefficients are $\beta_1 = 1.5$ and $\beta_2 = 0.7$. Denote the conditional information matrix, given $X_1 = x_1$ and $X_2 = x_2$, by $\mathcal{I}(\theta; x_1, x_2)$; the information matrix is given by

$$\mathcal{I}(\theta) = \sum_{x_1=0}^{1}\sum_{x_2=0}^{1} \mathcal{I}(\theta; x_1, x_2)P(X_1 = x_1)\,P(X_2 = x_2)\,.$$

– **Model 3** *(with mixed covariates)*. The variable Y^* depends on a continuous covariate, a dichotomous one and their interaction. The regression model is $Y^* = X_1\beta_1 + X_2\beta_2 + X_1X_2\beta_3 + \epsilon$, where $X_1 \sim N(0,1)$, $X_2 \sim Ber(0.5)$ and X_1 and X_2 are mutually independent. The regression coefficients are $\beta_1 = 2.7$, $\beta_2 = 1.5$ and $\beta_3 = 0.7$. In obvious notation, the information matrix is given by

$$\mathcal{I}(\theta) = \frac{1}{2}\int_{\mathcal{R}} \mathcal{I}(\theta; x_1, 0)\phi(x_1)\mathrm{d}x_1 + \frac{1}{2}\int_{\mathcal{R}} \mathcal{I}(\theta; x_1, 1)\phi(x_1)\mathrm{d}x_1.$$

Notice that Model 3 is generally used in the analysis of differential item functioning in grading scales (see [9] and references therein).

We consider the probit, the logit and the complementary log-log link function which assume the Gaussian, the logistic and the extreme value distribution for ϵ_i in (2.2), respectively.

Furthermore, in the analysis of the asymptotic efficiency, four different sets of thresholds are considered. By following the literature [7, 25, 26], a first set is given by equidistant thresholds α_j^{ED}, which satisfy the constraint $\alpha_j^{\mathrm{ED}} - \alpha_{j-1}^{\mathrm{ED}} = h$. Alternative thresholds α_j^{ID} are obtained by considering smaller classes around the median of Y^*, and by progressively increasing the length of the classes by a factor of $(1 + 1/m)$ when moving towards the tails. A further set of thresholds α_j^{EP} splits the support of Y^* in classes of equal probability, so that $P(Y = j) = 1/m$ for $j = 1, \ldots, m$. A final set of thresholds α_j^{DP} corresponds to classes with larger probability at the center and decreasing probability, by a factor of $(1 - 1/m)$, when moving from the center towards the extremes. Details on the construction of the four sets of thresholds are given in the supplementary material. Each set of thresholds generates its specific distribution of the ordinal response from the same underlying variable. The distributions produced by the four sets of thresholds, in Model 1 with the logit link and $m = 7$, are displayed in Figure 1 and appear markedly different.

Since the analytical expression of the (unconditional) information matrix is available for the three models, most of the analyses carried out in the following sections, and in particular the evaluation of the asymptotic efficiency of the estimators, the approximation of the power of the test and the assessment of the loss of efficiency produced by collapsing categories and dichotomization, do not require simulations. Numerical experiments are carried out only for Figure 3 and more extensively in Section 7 where the appropriate choice of m for a given sample size n is investigated. The purpose is to take into account numerical issues which may arise for specific combinations of m and n especially when n is small. Estimation is implemented in the `R` package `MASS` [27] and the simulation is always performed on 10 000 samples. Both analytical results and simulation experiments consider a number of parameters increasing with m, since estimation involves $m - 1$ thresholds in addition to the p regression coefficients.

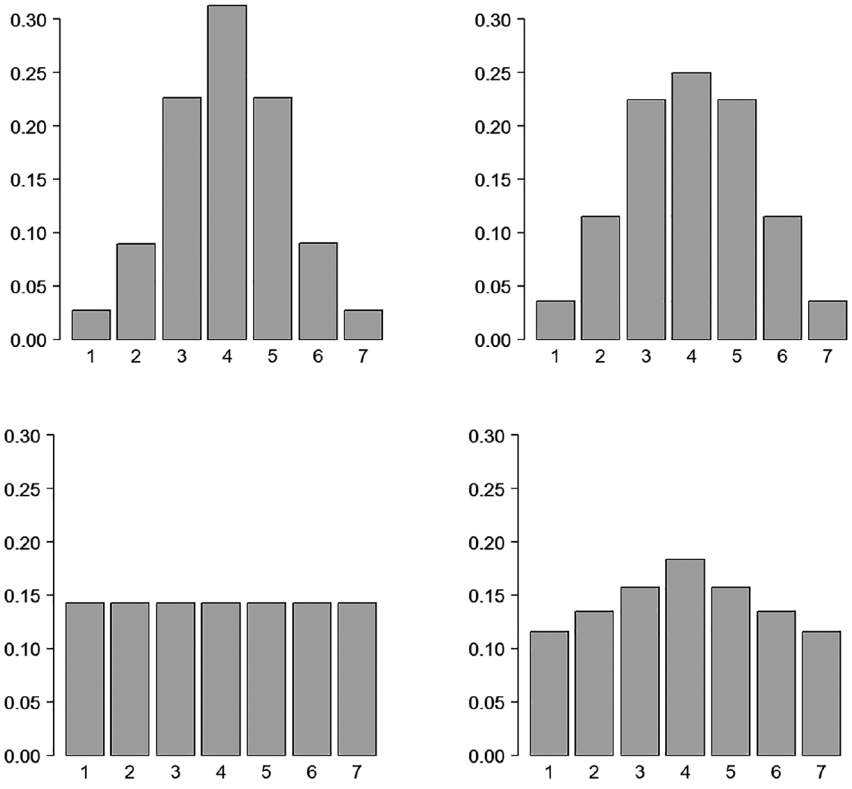

Figure 1: Distribution of the ordinal variable for the α^{ED} thresholds (upper left panel), for the α^{ID} thresholds (upper right panel), for the α^{EP} thresholds (lower left panel), and for the α^{DP} thresholds (lower right panel) in Model 1 with logit link and $m = 7$.

4 The asymptotic efficiency of the estimators

When m increases, each category of Y corresponds to a smaller class on the support of Y^*, and the resulting finer categorization yields more information on the underlying variable. In this context it is useful to remind that, given an event $\mathcal{E}$, its information content is given by $-\log\{P(\mathcal{E}\}$: the smaller $P(\mathcal{E})$ the larger the information ([28]; p. 4), since occurrence of rare events is more informative than occurrence of likely ones. Now suppose we have a finer discretization of the support of Y^* in m classes $C_1, \dots, C_m$, and a coarser categorization of Y^* in m' classes $\tilde{C}_1, \dots, \tilde{C}_{m'}$, with $m' < m$. Let Y and $\tilde{Y}$ be the responses obtained from the discretization in m and m' classes. When $Y = j$, $\tilde{Y}$ takes a value k such that $C_j \cap \tilde{C}_k \neq \emptyset$. Since $\tilde{Y}$ is based on a coarser categorization, it is reasonable to assume that $P(Y = j) = P(Y^* \in C_j) \leq P(\tilde{Y} = k) = P(Y^* \in \tilde{C}_k)$. Under this condition, the following result holds.

Proposition 1. *Let $Y^* \sim F$. Given two alternative discretizations of its support in the classes $(C_1, \dots, C_m)$ and $(\tilde{C}_1, \dots, \tilde{C}_{m'})$, with $m' < m$, let $\pi_j = P(Y = j) = \int_{C_j} dF(y)$, for $j = 1, \dots, m$, and $\tilde{\pi}_k = P(\tilde{Y} = k) = \int_{\tilde{C}_k} dF(y)$, for $k = 1, \dots, m'$. We have*

$$\log P(Y) = \sum_{j=1}^{m} I[Y = j] \log(\pi_j), \quad \log P(\tilde{Y}) = \sum_{k=1}^{m'} I[\tilde{Y} = k] \log(\tilde{\pi}_k),$$

If $\pi_j < \tilde{\pi}_k$ for all k such that $C_j \cap \tilde{C}_k \neq \emptyset$, for $j = 1, \dots, m$, then

$$-\log P(Y) \geq -\log P(\tilde{Y}). \tag{4.1}$$

Equation (4.1) implies that the information content of Y is larger than that of $\tilde{Y}$. Hence a finer categorization provides more information content to the data, which – in turn – yields more efficient estimators.

Tables 1–4 illustrate the asymptotic efficiency of the estimators of the regression coefficients in Models 1, 2 and 3 for the four sets of thresholds. In Model 1 there is a single regression coefficient, so that the efficiency is measured by its asymptotic variance $\mathcal{I}(\boldsymbol{\theta})^{\beta\beta}$. In case of Models 2 and 3, where there is a vector of regression coefficients, the efficiency is measured by the trace of the portion of the asymptotic variance-covariance matrix related to the regression coefficients, i.e. by the sum of the asymptotic variances of the elements of $\hat{\boldsymbol{\beta}}$, $\sum_{k=1}^{p} \mathcal{I}(\boldsymbol{\theta})^{\beta_k\beta_k}$.

The outcomes clearly point out that, for all the sets of thresholds, the efficiency of $\hat{\boldsymbol{\beta}}$ increases with m accordingly with the larger amount of information associated with a finer categorization. The decrease of the asymptotic variances is especially marked for low values of m while it tapers off when m increases, and this behavior is shared by the three models and by the three links. Increasing m up to 7 usually yields considerable gains in efficiency, while marginal benefits are obtained by increasing m beyond 10.

As concerns the impact of the thresholds on the asymptotic efficiency, Tables 1–4 show that no set of thresholds outperforms the others, and the optimal set depends on the model as well as on the value of m and on the link (see also the Figures S.1, S.2, and S.3 in the supplementary material). In Model 1, with the probit and the logit link, the α_j^{DP} thresholds produce more efficient estimators for small m, while when $m \geq 6$ the α_j^{ED} and α_j^{ID} thresholds are to be preferred. In the same model, the best results for the complementary log-log link are generally obtained with the α_j^{DP} thresholds. In Model 2, with the probit and the complementary log-log

Table 1: Asymptotic efficiency of the estimators with the α^{ED} thresholds when m varies.

m	Model 1			Model 2			Model 3		
	Probit	Logit	C.log-log	Probit	Logit	C.log-log	Probit	Logit	C.log-log
3	5.07	8.00	6.08	16.00	48.11	22.25	65.97	94.20	67.65
4	3.79	6.50	4.78	13.49	37.58	18.50	48.93	72.75	51.40
5	3.18	5.82	4.10	12.51	35.46	17.30	38.40	60.63	41.46
6	2.86	5.46	3.69	11.97	33.76	15.63	31.89	53.94	35.36
7	2.66	5.24	3.42	11.65	32.86	14.50	27.71	49.87	31.33
8	2.54	5.10	3.24	11.44	32.24	13.89	24.93	47.22	28.52
9	2.45	5.00	3.11	11.30	31.82	13.49	23.01	45.40	26.48
10	2.39	4.93	3.01	11.20	31.52	13.18	21.64	44.10	24.95
12	2.31	4.84	2.87	11.07	31.13	12.75	20.95	42.40	22.83
15	2.25	4.77	2.76	10.96	30.81	12.40	19.12	41.01	20.96
20	2.20	4.71	2.66	10.87	30.56	12.11	16.95	39.92	20.30

Table 2: Asymptotic efficiency of the estimators with the α^{ID} thresholds when m varies.

m	Model 1			Model 2			Model 3		
	Probit	Logit	C.log-log	Probit	Logit	C.log-log	Probit	Logit	C.log-log
3	4.54	7.35	5.48	14.73	41.06	19.41	64.01	92.66	67.04
4	3.57	6.24	4.49	13.09	36.30	17.53	45.82	69.71	48.83
5	3.04	5.64	3.86	12.22	33.92	14.95	36.01	58.98	39.62
6	2.77	5.34	3.51	11.79	32.89	14.21	30.19	52.80	33.98
7	2.60	5.15	3.28	11.51	32.14	13.77	26.46	49.09	30.24
8	2.49	5.03	3.13	11.34	31.72	13.33	24.00	46.62	27.65
9	2.42	4.94	3.01	11.22	31.39	13.04	22.30	44.95	25.77
10	2.36	4.89	2.93	11.13	31.18	12.80	21.07	43.73	24.35
12	2.29	4.81	2.81	11.02	30.89	12.48	19.48	42.15	22.40
15	2.24	4.75	2.72	10.93	30.65	12.21	18.17	40.85	20.67
20	2.19	4.70	2.63	10.86	30.47	11.99	17.14	39.83	19.22

Table 3: Asymptotic efficiency of the estimators with the α^{EP} thresholds when m varies.

m	Model 1			Model 2			Model 3		
	Probit	Logit	C.log-log	Probit	Logit	C.log-log	Probit	Logit	C.log-log
3	4.28	7.28	5.01	14.28	35.57	18.15	75.52	110.85	82.48
4	3.41	6.23	4.06	12.81	33.07	15.94	47.85	79.43	53.51
5	3.02	5.74	3.60	12.14	32.01	14.84	36.96	66.28	41.76
6	2.80	5.46	3.33	11.77	31.45	14.17	31.31	59.21	35.48
7	2.67	5.29	3.17	11.54	31.12	13.71	27.91	54.84	31.61
8	2.57	5.16	3.05	11.39	30.91	13.38	25.66	51.90	29.00
9	2.50	5.08	2.97	11.28	30.77	13.14	24.08	49.79	27.14
10	2.45	5.01	2.90	11.20	30.66	12.96	22.90	48.20	25.74
12	2.38	4.92	2.81	11.09	30.53	12.69	21.29	46.01	23.80
15	2.32	4.84	2.73	10.99	30.42	12.44	19.85	44.01	22.03
20	2.26	4.77	2.66	10.91	30.34	12.21	18.56	42.21	20.43

Table 4: Asymptotic efficiency of the estimators with the α^{DP} thresholds when m varies.

m	Model 1			Model 2			Model 3		
	Probit	Logit	C.log-log	Probit	Logit	C.log-log	Probit	Logit	C.log-log
3	4.29	7.10	5.07	14.37	36.15	17.88	67.22	98.59	71.88
4	3.35	6.08	4.04	12.73	33.18	15.80	44.98	73.61	49.76
5	2.96	5.60	3.57	12.07	32.12	14.61	34.79	61.71	39.13
6	2.74	5.35	3.30	11.69	31.50	13.93	29.64	55.82	33.62
7	2.61	5.18	3.13	11.47	31.16	13.47	26.44	51.94	30.04
8	2.52	5.08	3.01	11.32	30.94	13.17	24.40	49.49	27.69
9	2.45	5.00	2.93	11.22	30.79	12.94	22.93	47.65	25.96
10	2.41	4.94	2.87	11.15	30.68	12.77	21.88	46.35	24.70
12	2.34	4.86	2.78	11.04	30.54	12.53	20.43	44.50	22.92
15	2.28	4.79	2.71	10.96	30.43	12.31	19.15	42.83	21.30
20	2.23	4.74	2.64	10.88	30.35	12.11	18.03	41.36	19.88

link, larger efficiency is usually achieved by using the α_j^{DP} thresholds for small m and the α_j^{ID} thresholds for larger m, while with the logit link the best thresholds are the α_j^{EP}. In Model 3 with the probit link the α_j^{DP} thresholds appear preferable for small m and the α_j^{ID} and the α_j^{ED} ones for $m > 7$, while the best thresholds for the logit link and the complementary log-log link are the α_j^{ID} and the α^{DP}.

Although no set of thresholds dominates the others, appreciable differences in efficiency are observed only for small m, while the efficiency of the estimators obtained with the four sets of thresholds gets progressively closer when m increases. Hence a sufficiently large m can reduce the loss of efficiency produced by an inappropriate choice of the thresholds. In the following sections of the paper, the α_j^{ED} thresholds will be considered. This choice is motivated by their simplicity and frequency of use in applications (whenever no other indications arise from the concrete problem at hand), without implying any preference in this direction. Analogous analyses for the other sets of thresholds are reported in the supplementary material (Section S.1).

The efficiency of $\hat{\boldsymbol{\beta}}$ affects also the efficiency of the derived measures used to investigate the impact of the covariates and, in particular, the efficiency of the estimators of the odds-ratio (OR). As an example consider Model 1 with the logit link. We have $\mathrm{OR}(X_1) = \exp(-\beta)$ which is estimated by $\widehat{\mathrm{OR}}(X_1) = \exp(-\hat{\beta})$. The standard error of the estimator is $SE\left\{\widehat{\mathrm{OR}}(X_1)\right\} = \exp(-\beta)SE(\hat{\beta})$, whose asymptotic version is $\exp(-\beta)\left\{\mathcal{I}(\boldsymbol{\theta})^{\beta\beta}\right\}^{1/2}$. Figure 2 shows the asymptotic standard error of $\widehat{\mathrm{OR}}(X_1)$ when m varies. It can be appreciated how the

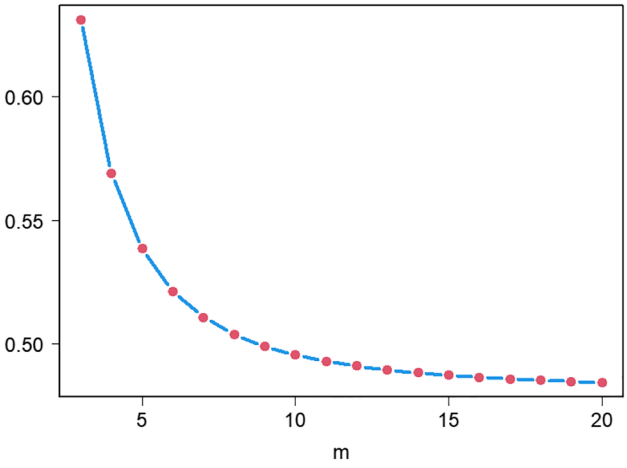

Figure 2: Asymptotic standard error of $\widehat{\mathrm{OR}}(X_1)$ in Model 1 with the logit link and the α^{ED} thresholds when m varies.

efficiency increases with m. Consistently with the results for $\hat{\beta}$, the gain in efficiency is considerably large for small values of m and becomes smaller for $m > 10$.

5 Hypothesis testing

The choice of m affects also the power of the test. Consider the hypotheses on a single regression coefficient $H_0: \beta_k = \beta_k^0$, $H_1: \beta_k \neq \beta_k^0$. They can be tested through a t-type statistic $t_k = (\hat{\beta}_k - \beta_k^0)/SE(\hat{\beta}_k)$ where the standard error of $\hat{\beta}_k$ is given by $SE(\hat{\beta}_k) = \sqrt{\mathcal{I}(\hat{\boldsymbol{\theta}})^{\beta_k\beta_k}/n}$. By (2.4), under H_0, this statistic is asymptotically $N(0,1)$ distributed. The null hypothesis is rejected when $|t_k| > z_{1-\alpha/2}$ where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and $\Phi(\cdot)$ is the standard normal distribution function. Hence the power of the test can be approximated by

$$\gamma(\boldsymbol{\beta}) = \Phi\left(z_{\alpha/2} - \frac{\beta_k - \beta_k^0}{\left\{\mathcal{I}(\boldsymbol{\theta})^{\beta_k\beta_k}/n\right\}^{1/2}}\right) + 1 - \Phi\left(z_{1-\alpha/2} - \frac{\beta_k - \beta_k^0}{\left\{\mathcal{I}(\boldsymbol{\theta})^{\beta_k\beta_k}/n\right\}^{1/2}}\right). \tag{5.1}$$

To investigate the impact of the choice of m on $\gamma(\boldsymbol{\beta})$ we consider the null hypothesis $H_0: \beta_3 = 0$ in Model 3. It implies that the interaction between X_1 and X_2 is omitted from the latent model, which becomes $Y^* = X_1\beta_1 + X_2\beta_2 + \epsilon$. Table 5 shows the power of the test, computed analytically through (5.1), at the 5% significance level, for the sample sizes $n = 250, 500$ and the three links (see also the analogous Tables S.1.1, S.1.2, and S.1.3 in the supplementary material). The power clearly increases with m. Intuitively the gain in the efficiency of $\hat{\boldsymbol{\beta}}$, obtained when m increases, induces a decrease of $SE(\hat{\beta}_k)$ so that high absolute values of the t_k statistic are more likely. A large m is especially recommended when ϵ has a large variance. This is the case of the cumulative logit model: the power of the test can be very low for small m, consequently large values of m are required to offset the variability of the error term.

The results of Table 5 are based on the asymptotic efficiency analytically evaluated. To take into account also the numerical issues which may arise when the estimation is actually implemented, Figure 3 shows the power of the test assessed through a simulation when the logit link is adopted, for sample sizes between 100 and 500. The magnitude of $\gamma(\boldsymbol{\beta})$ is mainly determined by the sample size. Nevertheless it can be appreciated the gain in power which can be achieved by increasing the number of categories especially when the initial m is small, say $m \leq 6$, though the marginal benefits are decreasing with m.

Table 5: Power of the test on $\beta_3 = 0$ in Model 3 with the α^{ED} thresholds, at the 5% significance level, when m varies and $n = 250, 500$.

m	$n = 250$			$n = 500$		
	Probit	Logit	C. log-log	Probit	Logit	C. log-log
3	0.72	0.49	0.70	0.95	0.78	0.94
4	0.84	0.58	0.81	0.99	0.87	0.98
5	0.92	0.67	0.89	1.00	0.92	0.99
6	0.95	0.72	0.93	1.00	0.95	1.00
7	0.97	0.75	0.96	1.00	0.96	1.00
8	0.98	0.78	0.97	1.00	0.97	1.00
9	0.99	0.79	0.98	1.00	0.98	1.00
10	0.99	0.81	0.98	1.00	0.98	1.00
12	0.99	0.82	0.99	1.00	0.98	1.00
15	1.00	0.83	0.99	1.00	0.99	1.00
20	1.00	0.84	1.00	1.00	0.99	1.00

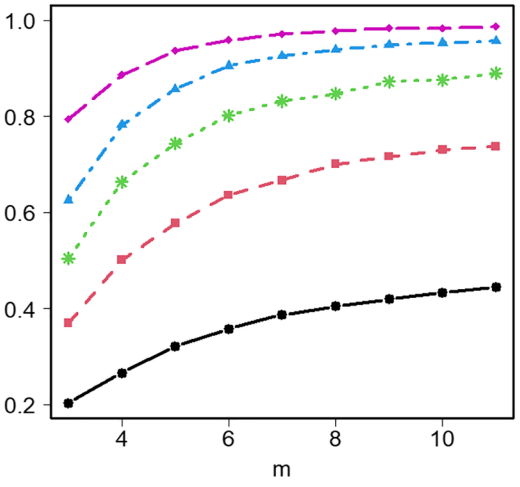

Figure 3: Simulated power of the test on the null hypothesis $\beta_3 = 0$ in Model 3 with the logit link, at the 5% significance level, when m varies and $n = 100$ (continuous line with circles), $n = 200$ (short-dashed line with squares), $n = 300$ (dotted line with asterisks), $n = 400$ (dot-dashed line with triangles) and $n = 500$ (long-dashed line with diamonds).

6 Merging categories

A widespread practice in data analysis is collapsing adjacent categories into one larger category (see [8, 18, 19], among others). This is typically done with extreme categories when there is a concern about their frequencies being very low (for instance in case of extreme response styles which cause the observations to be concentrated only on one side of the scale). An alternative reason for merging arises when a limited sample size yields unobserved categories or categories with very low frequencies. Finally the reduction of the number of categories may be finalized to simplify the interpretation, and in the extreme case it reaches its limit when the response is dichotomized.

In the main literature on categorical data, merging categories is a common practice for reducing the dimension of contingency tables (see [29], among others), and avoiding sparseness or small cell entries especially at the edges of the classification scale. However, an easier interpretation of the model is also a recurrent motivation (see [20], among others), and guidelines criteria for merging are *homogeneity* and *structure* (see [30–32], for further details).

In this paper merging categories is considered for a single variable, i.e. the ordinal response Y. Because of the collapsibility property of the POM, the regression parameters remain unchanged when the categories are

merged. Nevertheless, it is important to point out that collapsing categories reduces the information content of the sample outcome, as shown by the following proposition.

Proposition 2. *Let Y be a response with m categories and Y^M be a response obtained by merging two or more categories of Y, then*

$$\log\{P(Y^M)\} - \log\{P(Y)\} > 0.$$

See the Appendix for the proof.

Clearly the loss of information produced by collapsing is likely to turn into a loss of efficiency.

Here we investigate the impact of various forms of merging categories. For the case extreme categories are involved, we consider merging performed symmetrically on both sides of the scale, and merging implemented only on one side. A selection of cases is illustrated in the current section, while a more extensive investigation is carried out in the supplementary material (Section S.2). Furthermore the impact of halving the number of categories is analyzed, and finally the effect of dichotomizing the response is examined (see also Section S.1.4 of the supplementary material for similar analyses with the α^{ID}, α^{EP} and α^{DP} thresholds).

For the first form of merging Model 1 with the logit link is considered. In this model the distribution of the underlying variable is symmetric, so that (to avoid low extreme frequencies) it is reasonable to join both the first two categories and the last two. Consequently the first and the last thresholds α_1 and α_{m-1} are neglected, and the categories are based on the remaining thresholds $\alpha_2, \ldots, \alpha_{m-2}$. This procedure reduces the number of categories from m to $m-2$. Table 6 shows the asymptotic efficiency ratio between the "before merging" estimator $\hat{\beta}_m$ and the "after merging" estimator $\hat{\beta}_{m-2}$. The efficiency loss is considerable when m is small ($m = 5, 6, 7$). When the number of categories is reduced from 5 to 3 the loss of efficiency can be as high as almost 22%. The loss of efficiency is restrained – does not exceed 5% – when the number of categories is $m > 7$ and the probability of the extreme categories (which disappear) is below 0.025. Similar results for the value of m (which should be fairly large, say $m \geq 10$) and the probability of the vanishing categories (which should be sufficiently low, say around 0.025) are observed also for the other models and for the probit link (see Tables S.2.1–S.2.5 of the supplementary material).

To investigate the consequences of merging when it occurs only on one side, reference is still made to Model 1 but the link is now the complementary log-log one, so that the underlying variable has a skewed distribution with low probability on the first categories. When the first two categories are merged the number of scale points is reduced from m to $m-1$, since the threshold α_1 is neglected. Consequently the "before merging" estimator is $\hat{\beta}_m$ and the "after merging" estimator is $\hat{\beta}_{m-1}$. Table 7 shows the efficiency ratio between $\hat{\beta}_m$ and $\hat{\beta}_{m-1}$. When merging occurs only on one side the loss of efficiency is less dramatic than when both tails are involved. The loss of efficiency is below 5% when $m \geq 6$ and the probability of the first category

Table 6: Efficiency loss produced by merging of the extreme categories on both sides in Model 1 with the logit link.

m	Probability of the categories				$\mathrm{Var}(\hat{\beta}_m)$	$\mathrm{Var}(\hat{\beta}_{m-2})$	Efficiency ratio
	$j=1$	$j=2$	$j=m-1$	$j=m$			
5	0.051	0.236	0.236	0.051	5.823	7.100	1.219
6	0.035	0.140	0.140	0.035	5.457	6.009	1.101
7	0.027	0.090	0.090	0.027	5.238	5.526	1.055
8	0.022	0.062	0.062	0.022	5.096	5.266	1.033
9	0.019	0.045	0.045	0.019	4.999	5.109	1.022
10	0.017	0.034	0.034	0.017	4.930	5.005	1.015
12	0.014	0.022	0.022	0.014	4.840	4.882	1.008
15	0.011	0.013	0.013	0.011	4.767	4.788	1.004
20	0.009	0.007	0.007	0.009	4.710	4.720	1.002

Table 7: Efficiency loss produced by merging the lowest categories in Model 1 with the complementary log-log link.

m	Probability of the categories		$\text{Var}(\hat{\beta}_m)$	$\text{Var}(\hat{\beta}_{m-1})$	Efficiency ratio
	$j=1$	$j=2$			
4	0.058	0.310	4.784	5.590	1.169
5	0.037	0.159	4.099	4.353	1.062
6	0.027	0.092	3.689	3.792	1.028
7	0.021	0.058	3.423	3.473	1.014
10	0.014	0.023	3.009	3.020	1.004
15	0.010	0.009	2.757	2.760	1.001
20	0.008	0.005	2.660	2.661	1.000

(which disappears) is around 0.025 or below. Similar results for the probability of the vanishing category hold also for the other models (see Tables S.2.6 and S.2.7 of the supplementary material), although m should be increased with the complexity of the model (for instance in Model 3 we find again $m \geq 10$).

As anticipated, another merging option consists in halving the number of categories by joining adjacent ones. Table 8 shows the efficiency ratio between the estimators obtained from m and $m/2$ categories, which is computed as follows $\sum_{k=1}^{p} \text{Var}\left(\hat{\beta}_{\frac{m}{2},k}\right) / \sum_{k=1}^{p} \text{Var}\left(\hat{\beta}_{m,k}\right)$, where $\text{Var}(\hat{\beta}_{m,k})$ is the asymptotic variance of the k-th element of the estimator $\hat{\boldsymbol{\beta}}_m$ obtained from m categories. Halving the number of scale points can have a remarkably high price in terms of efficiency especially when the number of covariates increases or the link is the probit or the complementary log-log one (the neglected information turns out to be especially valuable in these cases). Furthermore, consistently with previous results, the negative effect of merging is larger when the initial value of m is small, since collapsing produces a much coarser categorization.

Finally a common practice in applications is to reduce an ordinal response into a dichotomous one to easy interpretation (see [33, 34], among others).

Table 9 shows the cost in terms of efficiency to be paid for dichotomization, which is measured by $\sum_{k=1}^{p} \text{Var}\left(\hat{\beta}_{2,k}\right) / \sum_{k=1}^{p} \text{Var}\left(\hat{\beta}_{m,k}\right)$. The loss of efficiency due to dichotomization can be extremely severe (see also [35] and [36] for similar results). If a response with 4 categories is dichotomized, the efficiency ratio varies between roughly 1.2 and more than 5 (see Model 3). The loss of efficiency constantly increases with m. In the worst case, Model 3 with the probit or the complementary log-log link, if a 10-point response is dichotomized the efficiency ratio largely exceeds 10, and it gets even worse for larger m.

Table 9 shows also a different pattern for the three link functions: although for all of them the dichotomization has a considerable impact, the estimators obtained with the logit link appear to exploit better the reduced amount of information limiting the loss of efficiency.

Table 8: Efficiency ratio between $\hat{\boldsymbol{\beta}}_{m/2}$ and $\hat{\boldsymbol{\beta}}_m$ produced by halving the number of categories.

m	Model 1			Model 2			Model 3		
	Probit	Logit	C.log-log	Probit	Logit	C.log-log	Probit	Logit	C.log-log
4	1.91	1.65	1.71	1.43	1.18	3.20	5.16	3.55	4.97
6	1.78	1.47	1.65	1.34	1.42	1.42	2.07	1.75	1.91
8	1.50	1.28	1.48	1.18	1.17	1.33	1.96	1.54	1.80
10	1.33	1.18	1.36	1.12	1.12	1.31	1.77	1.38	1.66
12	1.24	1.13	1.28	1.08	1.08	1.23	1.52	1.27	1.55
14	1.17	1.09	1.23	1.06	1.06	1.16	1.41	1.21	1.46
20	1.09	1.05	1.13	1.03	1.03	1.09	1.28	1.10	1.23

Table 9: Efficiency ratio between $\hat{\beta}_2$ and $\hat{\beta}_m$ produced by dichotomization.

m	Model 1			Model 2			Model 3		
	Probit	Logit	C.log-log	Probit	Logit	C.log-log	Probit	Logit	C.log-log
4	1.91	1.65	1.71	1.43	1.18	3.20	5.16	3.55	4.97
6	2.53	1.97	2.21	1.61	1.31	3.79	7.91	4.79	7.23
8	2.85	2.11	2.52	1.68	1.38	4.27	10.12	5.47	8.96
10	3.02	2.18	2.71	1.72	1.41	4.50	11.66	5.86	10.24
12	3.13	2.22	2.84	1.74	1.43	4.65	12.05	6.09	11.19
16	3.24	2.26	2.99	1.76	1.44	4.81	14.25	6.34	12.45
20	3.29	2.28	3.07	1.77	1.45	4.89	14.88	6.47	12.59

These outcomes call for a recommendation against the use of dichotomization. Similar suggestions can be also found in [8, 12, 37, 38]. In particular [36] define dicothomization an arbitrary researchers' choice and show that the loss of efficiency can be exacerbated by the selection of an inappropriate cut-point.

Overall, the above results point out that a reduction of the number of categories, in any of the forms considered here, by decreasing the amount of information, can produce a remarkable loss of efficiency especially when the merging involves the central categories (with higher frequencies) or reaches the limiting case of dichotomization.

7 Choice of m for given n

A question which frequently arises, when setting the number of response options, concerns the appropriate number of categories for a given sample size n. On one side the positive relationship between efficiency and m would suggest a large number of categories. On the other side, if the sample size n is small, when m increases one or more categories may not produce observations.

In this regard, it is to be pointed out that, although in different statistical contexts categories with null frequencies give rise to the well known problems of sparse data, in the POM context a missing category in the sample produces no computational problems. Indeed, when one or more categories are unobserved, the estimation of the model can be still carried out by considering only the sampled categories.

The relationship between m and n is investigated through a simulation (with the details given in Section 3) to take into account the numerical issues which may arise for specific combinations of n and m, especially when the sample size is small. The analysis is carried out in the context of Model 3 with the logit link, though similar results are obtained for the other models and the other links as reported in Section S.3 of the supplementary material (see also Section S.1.5 for thresholds different from the α^{ED}).

Let m^{obs} be the observed number of categories, Table 10 shows the percentage of samples such that $m^{\mathrm{obs}} < m$. This percentage is extremely large for $n = 50$ or $n = 100$, though it rapidly reduces when n increases: it is below 1.5% for $n = 300$ and it becomes negligible for $n = 400$. These results indicate that, in order to avoid unobserved categories, the samples size should be $n \geq 300$ (see also Tables S.3.1–S.3.8 of the supplementary material for analogous results on the other models and the other links).

Table 11 shows the sum of the mean square errors of $\hat{\boldsymbol{\beta}}$ computed on the same samples of Table 10. The estimation is performed on the observed number of categories, whether $m^{\mathrm{obs}} = m$ or $m^{\mathrm{obs}} < m$. Consequently the mean square errors are computed always on 10 000 samples although they may have a different number of observed scale points. Notice that the efficiency can be very poor for extremely small sample sizes, say $n = 50$ or $n = 100$, although it quickly increases with n. A sample size $n \geq 300$ seems adequate also with respect to the efficiency (consistent results are obtained also for the other models and the other links as shown in Tables S.3.9–S.3.16 of the supplementary material).

Table 10: Percentage of simulated samples with a number of categories smaller than m in Model 3 with the logit link, when n and m vary.

n	m								
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
50	0.02	1.80	11.43	25.10	39.42	52.62	64.41	74.51	82.48
100	0.00	0.04	0.79	3.48	8.71	15.91	23.69	32.41	40.92
200	0.00	0.00	0.00	0.03	0.37	1.35	2.94	5.10	7.71
300	0.00	0.00	0.00	0.00	0.03	0.17	0.43	0.82	1.44
400	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.12	0.35
500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06

Table 11: Simulated efficiency $\left(100 \times \sum_k MSE(\hat{\beta}_k)\right)$ in Model 3 with the logit link, when n and m vary.

n	m								
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
50	1835.07	408.86	280.22	215.09	189.58	175.70	165.87	161.66	155.05
100	142.38	104.39	82.35	72.30	66.27	62.23	59.89	57.46	56.72
200	56.84	43.46	35.15	31.07	28.62	27.38	26.12	25.26	24.89
300	34.97	26.74	22.26	19.63	17.92	17.12	16.48	16.12	15.76
400	25.15	19.54	16.03	14.10	13.20	12.44	11.89	11.59	11.37
500	20.16	15.73	12.87	11.38	10.52	10.01	9.57	9.33	9.09

It is to be pointed out that, regardless whether the observed number of categories corresponds to m or not, the efficiency of $\hat{\boldsymbol{\beta}}$ increases with m for any sample size in agreement with the results of Section 4. Although [39] notice that when n is small and m is large maximum likelihood can yield biased estimators of the regression coefficients, the larger amount of information provided by a finer categorization produces a reduction in variance sufficiently large to offset the bias, yielding decreasing mean square errors. Hence the circumstance that some categories may be missing in the sample, does not alter the positive relationship between efficiency of $\hat{\boldsymbol{\beta}}$ and m, which holds for any sample size.

Notice that comparable efficiency is obtained by the couples $(n = 50, m = 11)$ and $(n = 100, m = 3)$, $(n = 100, m = 11)$ and $(n = 200, m = 3)$, $(n = 200, m = 11)$ and $(n = 400, m = 3)$, $(n = 300, m = 11)$ and $(n = 500, m = 4)$, etc. On the basis of Table 11, Figure 4 sketches couples (n, m) which yield the same efficiency.

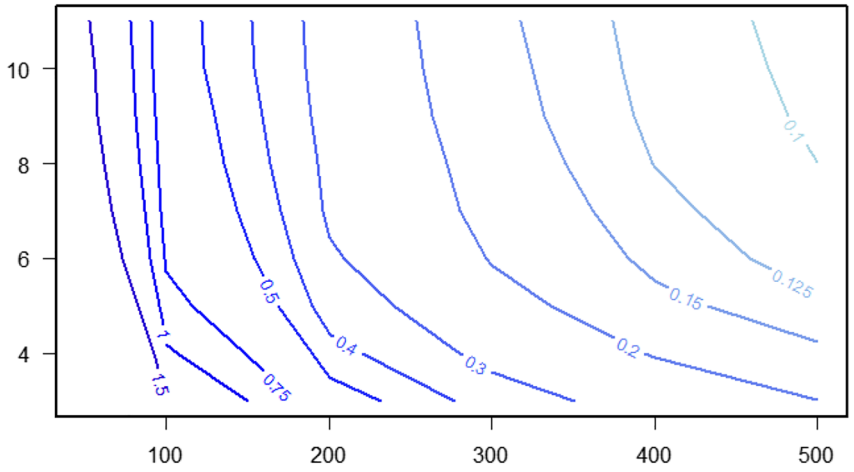

Figure 4: Simulated efficiency $\left(\sum_k MSE(\hat{\beta}_k)\right)$ in Model 3 with the logit link, with increasing n (horizontal axis) and m (vertical axis). Levels indicate the efficiency of the couple (n, m).

These results indicate that a small n requires a larger number of categories to compensate the limited availability of data with more information on the underlying variable, i.e. with a better quality of the data. The choice of m becomes less crucial when n increases, because the waste of information produced by a coarser categorization is balanced by a larger amount of data. Hence, m needs to be large especially if n is small.

8 Case studies

Two different case studies concerning the Linear Analogue Self-Assessment (LASA) scale and the visual analog pain scale (VAPS) are considered. Aim of the analysis is to show the impact of the choice of m in the discretization of scales which are endpoint-anchored lines. Researchers are interested in investigating the self-assessment of the quality of life (in the first example) and the perceived pain (in the second example) originally measured on an interval scale. In both examples an increasing number of categories allows to reduce the standard errors of the estimates and improve their significance. The second case study is also related to a small sample size to illustrate the opportunity of a relatively large m when the number of interviewees is limited.

8.1 Quality of life measured on linear analogue self-assessment scale

Data stem from the ANZ0001 trial conducted by the ANZ Breast Cancer Trials Group with the aim of assessing health-related quality of life of patients with advanced breast cancer [40]. Our analysis focuses on the overall quality of life, recorded on an LASA scale, normalized to $(0, 100)$ where 0 represents *'as bad as it can be'* and 100 *'as good as it can be'*. The treatments intermittent capecitabine (IC) and continuous capecitabine (CC) are compared with the standard combination treatment (CMF), each with its own protocol.

The chemotherapy cycle number (cycle num.) and the body surface area (m^2) (body surface) are recorded for each assessment of the quality of life, in addition to the treatment (Treatment). The dataset, which contains 2473 observations, is available in the `R` package `ordinalCont` [41], see also [42]. The regression model corresponding to (2.2) is

$$Y_i^* = \text{cycle num}_i\, \beta_1 + \text{body surface}_i\, \beta_2 + Z_i^{\text{IC}}\, \beta_3 + Z_i^{\text{CC}}\, \beta_4 + \epsilon, \quad i = 1, \dots, n, \tag{8.1}$$

where Z_i^{IC} and Z_i^{CC} are dichotomous variables which identify the modalities IC and CC of the nominal variable Treatment, while CMF is the reference category.

The LASA scale has been discretized into equal-length intervals with m varying between 3 and 15. The fitted models with the logit link are shown in Table 12.

Consistently with the analytical results of the previous sections, the standard errors of the estimators decrease with m. Consequently the estimated coefficient of the variable Z^{IC}, which is not significant for $m = 3$ and $m = 5$, became significant for $m \geq 7$, pointing out that this type of Treatment can negatively affect the patients' quality of life.

Different effects of the two treatments IC and CC can be tested by considering the null hypothesis $\beta_4 - \beta_3 = 0$. The t-statistic, for varying m, is reported in Table 13. As m increases, it becomes evident that the CC treatment leads to a better quality of life with respect to the IC treatment. There is instead no significant difference between CC and CMF.

These outcomes show that an increasing number of categories may enhance model specification.

8.2 Pain measured on visual analog pain scale

Data are about a small sample of 56 women aged between 23 and 44 years interviewed on their perceived pain during labor until childbirth. These women delivered at the Città di Roma Hospital (Rome) or at the Saint

Table 12: Fitted model (8.1).

	Estimate	St. error	*t*-statistic
	(a) $m = 3$		
Cycle num.	−0.048	0.006	−7.808
Body surface	0.372	0.332	1.120
IC	−0.071	0.113	−0.628
CC	−0.093	0.114	−0.824
	(b) $m = 5$		
Cycle num.	−0.039	0.005	−8.420
Body surface	0.615	0.296	2.081
IC	−0.094	0.102	−0.918
CC	−0.014	0.102	−0.134
	(c) $m = 7$		
Cycle num.	−0.037	0.004	−8.848
Body surface	0.629	0.283	2.223
IC	−0.138	0.099	−1.401
CC	−0.011	0.098	−0.113
	(d) $m = 10$		
Cycle num.	−0.035	0.004	−8.624
Body surface	0.757	0.275	2.759
IC	−0.128	0.097	−1.314
CC	0.010	0.097	0.107
	(e) $m = 15$		
Cycle num.	−0.035	0.004	−8.965
Body surface	0.725	0.270	2.681
IC	−0.153	0.096	−1.590
CC	0.015	0.096	0.161

Table 13: Test on the hypothesis $\beta_4 - \beta_3 = 0$ when m varies.

m	3	5	7	10	15
t-statistic	−0.225	0.929	1.538	1.710	2.106

Raffaele Hospital (Milan), and most of them attended hospitals' childbirth preparation classes there. Details on these data are in [43]. The perceived pain has been collected by means of a VAPS. It consists in a slide rule with the patient's side unmarked and the observer's side marked from 0 to 100 mm, where 0 represents *'no pain'* and 100 represents '*worst pain ever*'. We consider the discretization of the VAPS with a number of intervals from $m = 3$ to the maximum rating considered for the analysis of pain $m = 11$ (for comparison with the Numerical Rating Pain Scale, see [44].

The position of the unborn (head), the participation to the pre-birth course (course) and the occurrence of previous events in which women perceived pain (previous pain) are the three covariates. The regression model corresponding to (2.2) is

$$Y_i^* = \text{head}_i\, \beta_1 + \text{course}_i\, \beta_2 + \text{previous pain}_i\, \beta_3 + \epsilon, \quad i = 1, \dots, n. \tag{8.2}$$

The fitted models with the complementary log-log link are in Table 14.

Table 14: Fitted model (8.2).

	Estimate	St. error	*t*-statistic
	(a) $m = 3$		
Head	0.030	0.365	0.083
Course	−0.273	0.540	−0.505
Previous pain	−1.785	0.561	−3.184
	(b) $m = 5$		
Head	0.730	0.279	2.612
Course	−0.722	0.459	−1.574
Previous pain	−1.485	0.415	−3.577
	(c) $m = 11$		
Head	0.541	0.226	2.394
Course	−0.678	0.386	−1.757
Previous pain	−1.111	0.348	−3.197

Table 15: Likelihood ratio test – null hypothesis no Head and Course covariates.

m	3	5	11
LR-statistic	0.260	8.161	7.765
p-value	0.878	0.017	0.021

In accordance with previous results the standard errors decrease with m. The estimated coefficients of Head and Course, which are not significant for $m = 3$, become significant for $m \geq 5$. The hypothesis $\beta_1 = \beta_2 = 0$, which implies that Head and Course do not affect the perceived pain, can be tested trough the likelihood ratio (LR) test. The corresponding statistic and the related p-value are reported in Table 15. A large m is required to detect the relevance of these two covariates as explanatory factors. Despite the small sample size, a large number of categories is necessary to convey power to the tests.

9 Final remarks

The paper exploits the collapsibility property of the cumulative models with proportional odds assumption, which allows to generate ordinal responses with a different number of categories from the same underlying variable, and investigates the impact of the choice of m on the reliability of inferential analyses. It proves that increasing m augments the information content of the data, yielding more efficient estimators and more powerful tests. However the benefits of increasing m are considerable when the initial number of categories is small, and become progressively smaller when m increases. The analyses carried out in the paper suggest values of m between 7 and 10. This range of values for m limits also the impact of inappropriate thresholds used in the categorization of continuous measurements.

Since the variance of the estimators decreases with m, the opportunity of merging categories should be carefully evaluated. Combining extreme categories should be applied only when $m \geq 10$ and the probability of the vanishing category is sufficiently small (say around 0.025). Halving the number of categories appears an inconvenient procedure in terms of efficiency. The dichotomization is the most critical practice because it produces an extremely severe loss of information and, consequently, of efficiency, which can be only partially restrained by choosing the logit link instead of alternative link functions.

Finally numerical simulations show that increasing m enhances the efficiency even if the sample size is small. A high number of scale points is recommended to gather all the information contained in the sample especially if it is of limited size. These experiments indicate also that a sample size $n \geq 300$ allows to avoid unobserved categories to a large extent and produces sufficiently efficient estimators.

These findings are illustrated through two case studies based on discretization of continuous scales. In both cases an increasing number of categories allows to reveal the relevance of the explanatory variables, which may remain undetected if the categorization is too coarse. Hence a large m enhances model specification.

Acknowledgments: We would like to thank Alan Agresti for helpful discussions and constructive comments and Alessio Farcomeni for graciously allowing us to utilize the Pain data in the Case studies Section.
Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: None declared.
Conflict of interest statement: The authors have declared no conflict of interest.

Appendix

Proof of Proposition 2.

Let $\boldsymbol{Y} = (Y_1, \dots, Y_m)$ be a random vector with a multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_m)$, where $\pi_j \geq 0$ for $j = 1, \dots, m$, and $\sum_{j=1}^{m} \pi_j = 1$. Its probability mass function is

$$P(\boldsymbol{Y}) = \pi_1^{Y_1} \pi_2^{Y_2} \dots \pi_{m-1}^{Y_{m-1}} (1 - \pi_1 - \dots \pi_{m-1})^{1 - \sum_{j=1}^{m-1} Y_j}.$$

Suppose, without loss of generality, that the first two categories are merged. Since $Y_1 = 1$ and $Y_2 = 1$ are incompatible events, the event $(Y_1 = 1) \cup (Y_2 = 1)$ is observed with probability $\pi_1 + \pi_2$. The probability of the merged variable $\boldsymbol{Y}^M = (Y_1 + Y_2, Y_3, \dots, Y_m)$ is

$$P^M(\boldsymbol{Y}^M) = (\pi_1 + \pi_2)^{Y_1 + Y_2} \pi_3^{Y_3} \dots \pi_{m-1}^{Y_{m-1}} (1 - \pi_1 - \dots \pi_{m-1})^{1 - \sum_{j=1}^{m-1} Y_j}.$$

Let $\ell(\boldsymbol{Y}) = \log\{P(\boldsymbol{Y})\}$ and $\ell^M(\boldsymbol{Y}^M) = \log\{P^M(\boldsymbol{Y}^M)\}$. Their difference is always negative

$$\begin{aligned} \ell(\boldsymbol{Y}) - \ell^M(\boldsymbol{Y}^M) &= Y_1 \log(\pi_1) + Y_2 \log(\pi_2) - \{Y_1 + Y_2\} \log(\pi_1 + \pi_2) \\ &= Y_1 \log\left(\frac{\pi_1}{\pi_1 + \pi_2}\right) + Y_2 \log\left(\frac{\pi_2}{\pi_1 + \pi_2}\right) < 0. \end{aligned} \tag{A.1}$$

Inequality (A.1) shows that there is more information in the original distribution, with a larger number of categories, than in the distribution derived from merging, i.e. collapsing categories reduces the amount of sample information. □

References

1. Hosmer DW, Lehemenshow S. Applied logistic regression. New York: John Wiley & Sons; 2000.
2. Gurland J, Lee I, Dahm PA. Polychotomous quantal response in biological assay. Biometrics 1960;16:382–98.
3. Snapinn S, Small R. Tests of significance using regression models for ordered categorical data. Biometrics 1986;42:583–92.
4. Peracchi F, Perotti V. Subjective survival probabilities and life tables: an empirical analysis of cohort effects. Genus 2009;LXV:23–57.
5. O' Brien SM. Cutpoint selection for categorizing a continuous predictor. Biometrics 2004;60:504–9.

6. Winship C, Mare RD. Regression models with ordinal variables. Am Socio Rev 1984;1:512–25.
7. Ramsay JO. The effect of number of categories in rating scales on precision of estimation of scale values. Psychometrika 1973;38:513–32.
8. Agresti A. Analysis of ordinal categorical data, 2nd ed. Hoboken: John Wiley & Sons; 2010.
9. Allahyari E, Jafari P, Bagheri Z. A simulation study to assess the effect of the number of response categories on the power of ordinal logistic regression for differential item functioning analysis in rating scales. Comput Math Methods Med 2016:1–8. https://doi.org/10.1155/2016/5080826.
10. Iannario M, Monti AC, Piccolo D. Robustness issues for CUB models. Test 2016;25:731–50.
11. Piccolo D. On the moments of a mixture of uniform and shifted binomial random variables. Quad Stat 2003;5:85–104.
12. McCullagh P. Regression models for ordinal data. J Roy Stat Soc B 1980;42:109–42.
13. McCullagh P, Nelder JA. Generalized linear models, 2nd ed. London: Chapman & Hall; 1989.
14. Van Meter EM, Garrett-Mayer E, Bandyopadhyay D. Proportional odds model for dose finding clinical trial designs with ordinal toxicity grading. Stat Med 2011;30:2070–80.
15. Everitt BS, Palmer CR, Horton R. Encyclopaedic companion to medical statistics, 2nd ed. New York: Wiley; 2011.
16. Kotz S, Read CB, Balakrishnan N, Vidakovic B, Johnson NL, Liu I, et al. Proportional odds model. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B, Johnson NL, editors. Encyclopedia of statistical sciences. New York: Wiley; 2014.
17. Tutz G. Regression for categorical data. Cambridge: Cambridge University Press; 2012.
18. Strömberg S. Collapsing ordered outcome categories: a note of concern. Am J Epidemiol 1996;144:421–4.
19. Johnson VE, Albert JH. Ordinal data modeling. New York: Springer-Verlag; 1999.
20. Kateri M. Contingency table analysis. Methods and implementation using R. Birkhãuser, Basel: Springer; 2014.
21. Whitehead J. Sample size calculations for ordered categorical data. Stat Med 1993;12:2257–71.
22. Oyeyemi GM, Adewara AA, Adebola FB, Salau SI. On the estimation of power and sample size in test of independence. Asian J Math Stat 2010;3:139–46.
23. Iannario M, Lang JB. Testing conditional independence in sets of I × J tables by means of moment and correlation score tests with application to HPV vaccine. Stat Med 2016;35:4573–87.
24. Iannario M, Monti AC, Piccolo D, Ronchetti E. Robust inference for ordinal response models. Electron J Stat 2017;11:3407–45.
25. Rattray J, Jones MC. Essential elements of questionnaire design and development. J Clin Nurs 2007;16:234–43.
26. Christensen RHB. ordinal - regression models for ordinal data. R package version 2019.4-25; 2019. Available from: http://www.cran.r-project.org/package=ordinal/.
27. Venables WN, Ripley BD. Modern applied statistics with S, 4th ed. New York: Springer; 2002.
28. McMahon DM. Computing explained. Hoboken, NJ: Wiley-Interscience; 2008.
29. Bishop YMM. Effects of collapsing multidimensional contingency tables. Biometrics 1971;27:5453–562.
30. Goodman LA. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. J Am Stat Assoc 1981;76:320–34.
31. Gilula Z. Grouping and association in contingency tables: an exploratory canonical correlation approach. J Am Stat Assoc 1986;81:773–9.
32. Kateri M, Iliopoulos G. On collapsing categories in two-way contingency tables. Statistics. J Theor Appl Stat 2003;37:443–55.
33. Manor O, Matthews S, Power C. Dichotomous or categorical response? Analysing self-rated health and lifetime social class. Int J Epidemiol 2000;29:149–57.
34. Purgato M, Barbui C. Dichotomizing rating scale scores in psychiatry: a bad idea? Epidemiol Psychiatr Sci 2013;22:17–9.
35. Coehn J. The cost of dichotomization. Appl Psychol Meas 1983;7:249–53.
36. Armstrong BG, Sloan M. Ordinal regression models for epidemiologic data. Am J Epidemiol 1989;129:191–204.
37. Archer KJ, Williams AAA. L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. Stat Med 2012;31:1464–74.
38. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and application. Int J Epidemiol 1997;26:1323–33.
39. Lipsitz SR, Fitzmaurice GM, Regenbogen SE, Sinha D, Ibrahim JG, Gawade AA. Bias correction for the proportional odds logistic regression model with application to a study of surgical complications. J Roy Stat Soc C 2012;62:233–50.
40. Stockler M, Sourjina T, Grimison P, Gebski V, Byrne M, Harvey V, et al. A randomized trial of capecitabine (C) given intermittently (IC) rather than continuously (CC) compared to classical CMF as first-line chemotherapy for advanced breast cancer (ABC). J Clin Oncol 2007;25:1031.
41. Manuguerra M, Heller G. ordinalCont - ordinal regression analysis for continuous scales. R package version 2019.2.0.1; 2019. Available from: http://www.cran.r-project.org/package=ordinalCont/.
42. Manuguerra M, Heller G. Ordinal regression models for continuous scales. Int J Biostat 2010;6:14.
43. Capogna G, Camorcia M, Stirparo S, Valentini G, Garassino A, Farcomeni A. Multidimensional evaluation of pain during early and late labor: a comparison of nulliparous and multiparous women. Int J Obstet Anesth 2010;19:167–70.

44. Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Loge JH, et al. Studies comparing numerical rating scales, verbal rating scales, and visual Analogue scales for assessment of pain intensity in adults: a systematic literature review. J Pain Symptom Manag 2011;41:1073–93.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/ijb-2021-0013).