The LASSO on Latent Indices for Regression Modeling with Ordinal Categorical Predictors

Francis K. C. Hui[a,*], Samuel Müller[b], A. H. Welsh[a]

[a]*Research School of Finance, Actuarial Studies & Statistics, Australian National University, Acton, ACT 2601, Australia*
[b]*School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia*

Abstract

Many applications of regression models involve ordinal categorical predictors. Two common approaches for handling ordinal predictors are to form a set of dummy variables, or employ a two stage approach where dimension reduction is first applied and then the response is regressed against the predicted latent indices. Both approaches have drawbacks, with the former running into a high-dimensional problem especially if interactions are considered, while the latter separates the prediction of the latent indices from the construction of the regression model. To overcome these challenges, a new approach called the LASSO on Latent Indices (LoLI) for handling ordinal predictors in regression is proposed, which involves jointly constructing latent indices for each or for groups of ordinal predictors and modeling the response directly as a function of these. LoLI borrows strength from the response to more accurately predict the latent indices, leading to better estimation of the corresponding effects. Furthermore, LoLI incorporates a LASSO type penalty to perform hierarchical selection, with interaction terms selected only if both parent main effects are included. Simulations show that LoLI can outperform the dummy variable and two stage approaches in selection and prediction performance. Applying LoLI to an Australian household-based panel identified three dimensions of psychosocial workplace quality (job demands, stress, and security) which affect an individual's mental health in an additive and

*Corresponding author
Email addresses: francis.hui@anu.edu.au (Francis K. C. Hui), samuel.mueller@sydney.edu.au (Samuel Müller), alan.welsh@anu.edu.au (A. H. Welsh)

pairwise interactive manner.

Keywords: dimension reduction, factor analysis, factor variables, interaction, latent variables, model selection

1. Introduction

Many applications of regression models involve ordinal categorical predictors. For instance, this article is motivated by the Household Income and Labour Dynamics in Australia (HILDA) survey, a nationally representative panel study that has collected data annually in Australia since 2001 (Watson and Wooden, 2012). Among other data collected, individuals are asked about their overall mental health and to respond to a series of statements concerning their current workplace situation e.g., "I have a lot of choice in deciding what I do at work". For each statement, the individual provides an ordinal rating or score from 1 ("strongly disagree") to 7 ("strongly agree"). One of the aims of the HILDA survey is to improve understanding of how various aspects of an individual's workplace quality contribute their overall mental well-being. For instance, having both a lack of job security and increased job stress/strain may compound and lead to a stronger detrimental effect on mental health than just having either aspect on its own (e.g., Butterworth et al., 2013; Milner et al., 2015, 2016).

1.1. Main Modelling Challenges For Ordinal Predictors

How to handle (a potentially large number of) ordinal predictors is a common challenge in regression modeling. If the number of levels (7 in the case of the HILDA survey) is large for each ordinal predictor, and there is *a-priori* knowledge regarding the distances between levels, then it may be possible to use the raw ratings from the ordinal data (or some simple monotone transformation of it) as actual scores and model them as values from a *continuous* predictor (see for instance, Agresti, 2013). However, in many cases such a direct score-based approach may not be appropriate e.g., in the HILDA survey, treating the score as a continuous predictor would mean that the distance between any two consecutive scores is the same, but there is no underlying reason why this should be the case. Instead, the two most popular approaches for handling ordinal predictors are as follows: 1) treat each ordinal predictor as a factor variable using (for example) a set of dummy variables; or 2) use a two stage approach where dimension reduction is first applied on the ordinal predictors (e.g., factor analysis Bartholomew et al., 2011), and then include the predicted indices as continuous covariates in a regression model in the second stage.

The first approach can often result in a high-dimensional problem, especially if we include interactions in the model. The problem of high-dimensionality is frequently encountered in regression modeling, and has spurred considerable research into penalized likelihood methods (among other approaches) for variable selection, including penalties which respect the hierarchical structure of the predictors in various modeling contexts; see for example (Zhao et al., 2009) for generalized linear models, (Hui et al., 2017) for selection in generalized linear mixed models, and Tutz and Gertheiss (2016); Pauger et al. (2019) for categorical data. More recently, prompted by interest in uncovering epistatic effects in genome wide association studies, there has been a further surge in interest on penalties which obey some form of marginality principle (e.g., Bien et al., 2013; Haris et al., 2016; She et al., 2016; Yan and Bien, 2017). While these approaches are capable of selecting from a large number of categorical variables and their interactions, they are perhaps not the most appropriate methods for handling the ordinal predictors in our setting. This is because the statements regarding workplace conditions in the HILDA survey are thought of as manifestations of latent indices related to various aspects of job quality (Leach et al., 2010). In turn, it is more sensible and appealing to explicitly construct these indices and enter these, instead of the ordinal variables, as covariates into a regression model.

This leads to the second commonly used approach for handling ordinal predictors, which first involves fitting latent variable models to the ordinal predictors (e.g.,typically the ordinal ratings in the HILDA survey are treated as continuous and factor analysis is applied, Leach et al., 2010; Butterworth et al., 2011), and then regressing the responses against the predicted latent indices; other approaches such as optimal scaling (Linting et al., 2007) could also be used in the first stage. This two stage approach though does have potential drawbacks. Notably, it fails to utilize the information from the response to better predict the latent indices for each individual. Indeed, by definition latent variable models can only be fitted to more than one manifest (ordinal) predictor, and yet it is common to have cases where we wish to construct a continuous latent index from just a single ordinal predictor e.g., in the HILDA survey there is one particular statement on workplace conditions which has been argued to constitute its own latent dimension on job quality

(Butterworth et al., 2011; Milner et al., 2016).

1.2. A New Approach and Main Contributions

We propose a new method for the analysis of ordinal predictors in regression models called the LASSO on Latent Indices (LoLI), which is motivated by the challenges of the dummy variable and two stage approaches discussed above. The key innovation of our method is to jointly construct a continuous latent index for each or for groups of ordinal predictors, and model the response directly as a function of these (and other predictors if appropriate) including potential pairwise interactions. This joint approach means LoLI can borrow strength from the response to more accurately predict the latent indices i.e., the scores for each individual, which in turn produces better estimation and inference on the corresponding regression coefficients. To perform selection on main and interaction effects between the latent indices, a LASSO type penalty is employed which accounts for the hierarchical nature of the coefficients. That is, the penalty ensures that whenever an interaction term is selected, both its parent main effects must also be included in the model.

Due to the construction of latent indices, LoLI does not require complicated group sparsity penalties to handle dummy variables. Put another way, compared to treating the ordinal predictors as factors, the dimensionality of the problem is already markedly reduced *before* any variable selection is performed. Alternatively, LoLI can be viewed as type of a penalized regression model with unknown latent scores assigned to the levels of ordinal predictors, except that the scores are observation-specific (in contrast to, say, Row-by-Column association models where the scores are the same across observations; see Section 6.3, Agresti, 2010).

We emphasize that LoLI is an *alternative* approach to the construction of dummy variables for handling ordinal predictors, and is ideally suited to settings where there is some scientific belief that the ordinal predictors are manifest variables of some underlying continuous index e.g., in our motivating HILDA survey. There are many other contexts where such a belief may not apply e.g., highest level of education attained with levels "no completion of high school", "high school", "vocational certificate", and "undergraduate degree or above", where indeed it may be better to analyze the ordinal predictor via the dummy variable approach.

We propose an efficient two-step estimation approach for calculating the LoLI estimates, which first involves estimating cutoff parameters (which relate the observed ordinal predictors to the latent indices) by fitting marginal ordinal regression models to the ordinal predictors. Conditional on these estimates, we apply a Monte-Carlo Expectation Maximization (MCEM) algorithm (Wei and Tanner, 1990) to predict the latent indices and estimate and perform selection on all other parameters. We show that this two-step approach produces consistent estimates of the cutoffs. Regarding the choice of the tuning parameter, we adapt the Extended Regularized Information Criterion (Hui et al., 2015; Fu et al., 2017) for use with LoLI. This criterion uses a dynamic model complexity penalty that depends on the tuning parameter itself, resulting in more aggressive shrinkage and often to better finite sample selection performance than other commonly used criteria such as AIC or BIC.

Simulation studies show that LoLI can outperform dummy variable and two stage approaches for handling ordinal predictors, in terms of estimation and selection performance as well as predicting the latent indices. Applying LoLI to the motivating HILDA survey, and adjusting for potential confounders such as age and gender, we identify three dimensions of workplace quality which affect an individual's mental health in an additive manner: job demands/complexity/interest, job stress/strain, and job security. Furthermore, we found evidence that having both increased job interest and increased job security had an effect on mental well-being that was greater than each aspect of job quality on its own i.e., a positive interaction between these two latent indices.

The remainder of the manuscript is structured as follows. In Section 2, we establish the latent indices models and subsequently define the Lasso on Latent Indices (LoLI). In Section 3, we detail our two-step estimation approach for LoLI and discuss how to choose the tuning parameter using a new information criterion. Section 4 presents a numerical study which shows that, by jointly constructing the latent indices and building the regression model, LoLI can outperform dummy variable and other two stage approaches in selecting and/or predicting the latent indices. In Section 5, we illustrate the application of LoLI on the motivating HILDA survey, including its ability to straightforwardly investigate interaction effects between different dimensions of job quality. We conclude with a discussion of areas of future research in Section 6. We provide `R` code for implementing LoLI as part of the Supporting Information.

2. The LASSO on Latent Indices

Consider a set of $i = 1, \ldots, n$ independent observations, consisting of a univariate continuous response y_i, a q-vector of predictors $\boldsymbol{z}_i$ that will not be dimension reduced, and a p-vector of ordinal predictors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, such that x_{ij} can take values $1, \ldots, L_j$. In this article, we focus on the case where both p and q are less than n, given this is the setting most relevant to the motivating HILDA survey (see Section 5). We acknowledge that future research may be required to handle situations where p and/or q exceed n.

Conditional on the predictors, we assume y_i (or some suitable transformation of it) is normally distributed with mean μ_i as specified below in equation (1) and variance σ^2. We focus on estimation and inference of the main effects and possible pairwise interactions between the ordinal predictors. For ease of presentation, we assume there are no interactions between $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$, although the developments below can be extended to handle such interactions.

As reviewed in Section 1.1, one possible approach is to set up a $(L_j - 1)$-vector of dummy variables for each ordinal predictor and fit a linear model to these. However, this can lead to a high-dimensional regression model: if $\boldsymbol{z}_i$ involves only an intercept, then there are $d_{\text{LM}} = 1 + \sum_{j=1}^{p}(L_j - 1) + \sum_{1 \le j < k \le p}(L_j - 1)(L_k - 1)$ coefficients present. Even if $n > p$, it could be that $d_{\text{LM}} > n$ and thus the coefficients cannot be estimated by standard regression techniques. To overcome this, the principle behind LoLI is to jointly construct a latent index for each or for groups of ordinal predictors and build a regression model directly from these indices. We first discuss the limiting case of LoLI with a separate latent index for each ordinal predictor, and then discuss the case for groups of ordinal predictors in Section 2.1.

For $j = 1, \ldots, p$, define a vector of cutoffs $\xi_{j,0} = -\infty < \xi_{j,1} = 0 < \ldots < \xi_{j,L_j-1} < \xi_{j,L_j} = \infty$, and a continuous latent index u_{ij} where $\xi_{j,l-1} < u_{ij} < \xi_{j,l}$ if and only if $x_{ij} = l$ for $l = 1, \ldots, L_j$. Analogously to cumulative link models for ordinal responses, it is common to set $\xi_{j,1} = 0$ for all $j = 1, \ldots, p$ to ensure the parameters are identifiable (Agresti, 2010). Letting $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^\top$ denote the p-vector of latent indices for observation i, the conditional mean of the response is regressed against these latent indices as

$$\text{E}(y_i|\boldsymbol{z}_i, \boldsymbol{u}_i) = \mu_i = \boldsymbol{z}_i^\top \boldsymbol{\alpha} + \boldsymbol{u}_i^\top \boldsymbol{\beta} + \sum_{1 \le j < k \le p} u_{ij} u_{ik} \gamma_{jk}, \tag{1}$$

where the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the regression coefficients corresponding to

$\boldsymbol{z}_i$ and the main effects for the latent indices, respectively, and γ_{jk} is the interaction coefficient between latent indices j and k. Compared to using dummy variables, we see that by modeling the conditonal expectation in terms of latent variables, the number of coefficients to estimate and select from is substantially reduced: if $\boldsymbol{z}_i$ involves only an intercept, then $d_{\text{LoLI}} = 1 + p + 2^{-1}p(p-1) < d_{\text{LM}}$ with the difference depending on the L_j's. Also, with continuous latent indices we are not limited to just linear terms, and may wish to include polynomial or smoothing terms for u_{ij} depending on the question of interest. For simplicity though, in this article we focus on the model as defined in equation (1). Also, with the inclusion of $\sum_{j=1}^{p}(L_j - 2)$ free cutoff parameters, the total number of parameters to estimate may still be quite high. But the key point is that the number of parameters involved in the *regression* component of the model is markedly reduced.

For observation i and ordinal predictor j, define $\boldsymbol{x}_{ij}^* = (x_{ij1}^*, \ldots, x_{ijL_j}^*)^\top$, where $x_{ijl}^* = 1$ if $x_{ij} = l_j$ for $l_j = 1, \ldots, L_j$ and zero otherwise. Let $\boldsymbol{\Psi} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \sigma^2, \boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_p^\top)^\top$ denote the full parameter vector, where $\boldsymbol{\gamma} = (\gamma_{12}, \ldots, \gamma_{1p}, \gamma_{23}, \ldots, \gamma_{(p-1)p})^\top$ and $\boldsymbol{\xi}_j = (\xi_{j,2}, \ldots, \xi_{j,L_j-1})^\top$. The marginal log-likelihood for the latent indices model, with mean structure given by equation (1), is defined as

$$
\begin{aligned}
\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\Psi}) &= \sum_{i=1}^{n} \log \left\{ \int f(y_i | \boldsymbol{u}_i, \boldsymbol{z}_i, \boldsymbol{\Psi}) \prod_{j=1}^{p} \left(\prod_{l=1}^{L_j} f(x_{ijl}^* | u_{ij}, \boldsymbol{\Psi}) f(u_{ij}) \, du_{ij} \right) \right\} \\
&= \sum_{i=1}^{n} \log \left\{ \int f(y_i | \boldsymbol{u}_i, \boldsymbol{z}_i, \boldsymbol{\Psi}) \prod_{j=1}^{p} \left(\prod_{l=1}^{L_j} \mathbb{I}(\xi_{j,l-1} < u_{ij} < \xi_{j,l})^{x_{ijl}^*} f(u_{ij}) \, du_{ij} \right) \right\},
\end{aligned} \tag{2}
$$

where $f(y_i | \boldsymbol{u}_i, \boldsymbol{z}_i, \boldsymbol{\Psi}) = \mathcal{N}(\mu_i, \sigma^2)$ is a normal density with μ_i given by equation (1) and variance σ^2, $f(u_{ij})$ is the $\mathcal{N}(0,1)$ density function, and we choose $f(x_{ijl}^* | u_{ij}, \boldsymbol{\Psi})$ to be the indicator function $\mathbb{I}(\xi_{j,l-1} < u_{ij} < \xi_{j,l})^{x_{ijl}^*}$. Using the standard normal density for u_{ij} along with the suggested indicator function is analogous to the latent variable parameterization for cumulative probit regression (Agresti, 2010), and is a standard choice in item response and latent variable models (Skrondal and Rabe-Hesketh, 2004). We can also replace the indicator function with probabilistic choices; this is discussed in Section 2.1. The assumption of zero mean and unit variance for $f(u_{ij})$ ensures that the parameters in the latent indices model are identifiable i.e., avoiding location and scale invariance. The assumption of independence between the u_{ij}'s could be relaxed to allow for correlated latent indices, although previous research has shown that this assumption is not overly restrictive in practice, and similarly that the normality assumption can be robust to misspecification of the shape of the latent index distribution (Wedel and Kamakura, 2001); see also the relevant discussion in Section 6.

Equation (2) embodies the joint nature of LoLI in that the latent indices are simultaneously constructed from the $\boldsymbol{x}_{ij}^*$'s and used as covariates in the regression model for the mean of y_i. In doing so, we can borrow strength from the latter in order to better predict the latent indices u_{ij} i.e., the scores for each observation, which in turn should lead to better estimation and inference of coefficients $\boldsymbol{\beta}$ and γ_{jk}'s. Indeed, this "limiting" case where each ordinal predictor has its own latent index demonstrates the clearest advantage of LoLI over two stage approaches: if we were to construct the latent indices based solely on the $\boldsymbol{x}_{ij}^*$, then the predictions would still show the same degree of discretization as the ordinal predictors. By borrowing information from the (continuous) response, LoLI produces improved predictions of the u_{ij}'s.

To perform variable selection on main and interaction effects associated with the latent indices, we propose combining equation (2) with a LASSO type penalty as follows.

DEFINITION 2.1. *For a given tuning parameter $\lambda > 0$, the LoLI (LASSO on Latent Indices) method is defined by the penalized likelihood*

$$\ell_{pen}(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) - \lambda \sum_{j=1}^{p} \left(w_j \beta_j^2 + \sum_{k=1}^{j-1} w_{kj} |\gamma_{kj}| + \sum_{k=j+1}^{p} w_{jk} |\gamma_{jk}| \right)^{1/2},$$

where $\{w_j > 0; j = 1, \ldots, p\}$ and $\{w_{jk} > 0; j = 1, \ldots, p; k = 2, \ldots, p\}$ are adaptive weights constructed a-priori *to guide feature selection, and $\ell(\boldsymbol{\Psi})$ as defined in equation (2).*

If the vector $\boldsymbol{z}_i$ contains covariates that we wish to select on, then the above penalized log-likelihood can be augmented with further penalties to select on the elements of $\boldsymbol{\alpha}$. However, we do not consider this extension here given that in the motivating HILDA survey the covariates $\boldsymbol{z}_i$ are included to adjust for potential confounding. Also, if we only consider a subset rather than all possible pairwise interactions in equation (1), then the penalty in Definition 2.1 can be modified to accommodate this setting.

LoLI formally accounts for the hierarchical structure of the coefficients by enforcing two types of sparsity. For latent index $j = 1, \ldots, p$, we first impose individual coefficient sparsity in the form of an adaptive LASSO penalty (Zou, 2006) on all associated interaction effects. This means interaction terms between two latent indices can be removed from the model without affecting selection of the parent main effects. Second, we impose group coefficient sparsity in the form of the group LASSO penalty (Yuan and Lin, 2006), which encourages the entire quantity $w_j\beta_j^2 + \sum_{k=1}^{j-1} w_{kj}|\gamma_{kj}| + \sum_{k=j+1}^{p} w_{jk}|\gamma_{jk}|$ to be shrunk to zero. This implies that if either one of the parent main effects for a latent index is shrunk to zero, then any child interaction term must also be shrunk to zero. The proposed penalty in Definition 2.1 is by no means the only way of constructing penalties that respect this hierarchical nature of the coefficients. For example, we could have implemented various flavors of the family of composite absolute penalties (CAP, Zhao et al., 2009), and indeed the proposed penalty can be regarded as a specific case from the CAP family. Importantly, the innovation of LoLI lies in the construction of the latent indices *and* the regularization of the corresponding coefficients, rather than in the penalty itself.

As an aside, it is possible to use other approaches to perform the model selection instead e.g., using information criteria for comparing candidate latent indices models. We prefer a regularization approach as it is both more computationally efficient (it simplifies the choice of model selection from a discrete space to a one-dimensional search along a continuous solution path dictated by λ, and allows us to make use of warm starts for both the parameter estimates and the latent indices), and tends to be more stable (prediction of the latent indices occurs in a "smooth" manner as the tuning parameter varies, in contrast to approaches such as information criteria where the latent indices are re-predicted for every candidate model).

We construct the adaptive weights in Definition 2.1 from a fit of the saturated model. Specifically, let $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ denote the vectors of main and interaction effect coefficients, respectively, obtained based on maximum likelihood estimation of the unpenalized model. Then we set $w_j = \tilde{\beta}_j^{-2}$ and $w_{jk} = |\tilde{\gamma}_{jk}|^{-1}$ as the adaptive weights. We remark that the construction of the adaptive weights for LoLI is relatively stable precisely because, as pointed out in Section 1, we have substantially reduced the dimensionality of the problem before any variable selection is performed. At the same time, in practice it is possible for potential instability to still arise particularly if the number of latent indices is large relative to the number of observations, and may motivate other methods of constructing the adaptive weights (e.g., Garcia and Mueller, 2016).

2.1. Groups of Ordinal Predictors

Suppose now we want to construct a single latent index for groups of ordinal predictors, but with different cutoffs for each predictor. Such cases commonly arise in item response theory, where a group of ordinal predictors are believed to all correspond to the same latent quantity. Let the p predictors be *a-priori* divided into $G < p$ non-overlapping groups, such that $\mathcal{A}_g$ denotes the set of predictors in group $g = 1, \ldots, G$ with dimension $1 \leq p_g < p$, and $\sum_{g=1}^{G} p_g = p$. Then the latent indices model involves G latent indices and their interactions such that equation (1) is modified to $\mu_i = \boldsymbol{z}_i^\top \boldsymbol{\alpha} + \boldsymbol{u}_i^\top \boldsymbol{\beta} + \sum_{1 \leq g < h \leq G} u_{ig} u_{ih} \gamma_{gh}$ where $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iG})^\top$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_G)^\top$. The penalized likelihood for this particular model is then given by $\ell_{\text{pen}}(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) - \lambda \sum_{g=1}^{G} \left(w_g \beta_g^2 + \sum_{h=1}^{g-1} w_{hg} |\gamma_{hg}| + \sum_{h=g+1}^{G} w_{gh} |\gamma_{gh}| \right)^{1/2}$, where $\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left\{ \int f(y_i | \boldsymbol{u}_i, \boldsymbol{z}_i, \boldsymbol{\Psi}) \prod_{g=1}^{G} \left(\prod_{j \in \mathcal{A}_g} \prod_{l=1}^{L_j} f(x_{ijl}^* | u_{ig}, \boldsymbol{\Psi}) f(u_{ig}) \, du_{ig} \right) \right\}$. We consider two possible choices for the conditional distribution $f(x_{ijl}^* | u_{ig}, \boldsymbol{\Psi})$: if $p_g = 1$, then we set $f(x_{ijl}^* | u_{ig}, \boldsymbol{\Psi}) = \mathbb{I}(\xi_{j,l-1} < u_{ig} < \xi_{j,l})^{x_{ijl}^*}$. This is consistent with the limiting case in equation (2). If $p_g > 1$, then we propose to use $f(x_{ijl}^* | u_{ig}, \boldsymbol{\Psi}) = \{\Phi(\xi_{j,l} - a_j u_{ig}) - \Phi(\xi_{j,l-1} - a_j u_{ig})\}^{x_{ijl}^*}$, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, and a_j is an additional covariate specific slope parameter controlling the "discrimination" between the various levels of the ordinal predictor (Samejima, 1969). The use of soft probabilistic differences, as opposed to hard indicator functions when $p_g = 1$, is motivated from graded response models (Samejima, 1969) which model the conditional distribution of the ordinal variables using differences in cumulative probabilities when groups of the ordinal variables are reduced to the same latent index. It is also possible to use alternative link functions such as the logit, but given the latent index is assumed to be normally distributed, then it is more natural to use the probit link. More importantly, note the joint construction and regression of latent indices mean LoLI continues to have the advantage of being able to borrow strength from y_i to better predict the u_{ig}, which in turn to lead to better inference on β_g and γ_{gh}.

For the remainder of this article, unless stated otherwise, we will focus on the general formulation of LoLI given by Definition 2.1 i.e., where each ordinal predictor has its own continuous latent index.

3. Estimation

We propose a two-step estimation approach to calculate estimates for LoLI. First, we fit a series of marginal regression models using the ordinal predictors as the response to obtain estimates of the cutoffs $(\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_p^\top)^\top$. Conditional on these cutoff estimates, we then estimate the remaining parameters and perform variable selection on the coefficients using an MCEM algorithm. A similar estimation method can be formulated for the case of groups of ordinal predictors in Section 2.1, with the additional complication that we also estimate the slopes $(a_1, \ldots, a_p)^\top$, and we provide details for this in Appendix A.3. Our proposed two-step estimation method bears similarities to other two-step estimation procedures commonly used in factor analytic models (e.g., Lee et al., 1995) as well as copula-based models (e.g., Joe, 2005).

3.1. Marginal Cumulative Probit Regression Models

To estimate of the cutoff parameters, which we denote as $\overline{\boldsymbol{\xi}}_j$ for $j = 1, \ldots, p$, we fit a marginal cumulative probit regression model to each ordinal predictor.

$$
\begin{aligned}
\overline{\boldsymbol{\xi}}_j &= \arg\max_{\boldsymbol{\xi}_j} \sum_{i=1}^{n} \log \left(\int \prod_{l=1}^{L_j} \mathbb{I}(\xi_{j,l-1} < u_{ij} < \xi_{j,l})^{x^*_{ijl}} \, f(u_{ij}) \, du_{ij} \right) \\
&= \arg\max_{\boldsymbol{\xi}_j} \sum_{i=1}^{n} \sum_{l=1}^{L_j} x^*_{ijl} \log \left\{ \Phi(\xi_{j,l}) - \Phi(\xi_{j,l-1}) \right\}.
\end{aligned}
$$

Such cumulative probit models are straightforwardly fitted via maximum likelihood estimation, and in fact analytical solutions can be derived based on the cumulative frequencies of the levels of each ordinal predictor; see Appendix A.1 for details of these solutions. On the other hand, formulating the problem in terms of multinomial log-likelihood estimation as above directly facilities theoretical investigation. Specifically, in Appendix A.1 we show that the cutoffs based on fitting the marginal regression models above are consistent for the true cutoff values. Intuitively, this is because the conditional distribution of the response depends on the latent indices only and not the cutoffs. Therefore, y_i does not provide any direct information regarding the $\boldsymbol{\xi}_j$'s and we can achieve reasonable estimates using only the ordinal predictors. Such a result can be used to further prove, under mild regularity conditions on the likelihood function in equation (2) and the tuning parameter, that all parameters in $\boldsymbol{\Psi}$ are consistently estimated by the two-step procedure.

3.2. Monte-Carlo Expectation Maximization Algorithm

To calculate the remaining LoLI estimates, we employ a MCEM algorithm with an importance sampling algorithm to perform the E-step. The unpenalized complete log-likelihood for the latent indices model is given by

$$\begin{aligned}\ell_c(\boldsymbol{\Psi}, \boldsymbol{u}) &= \sum_{i=1}^{n} \ell_{ci}(\boldsymbol{\Psi}, \boldsymbol{u}_i) \\ &= -\frac{1}{2} n \log(\sigma^2) - (2\sigma^2)^{-1} \sum_{i=1}^{n} \left(y_i - \boldsymbol{z}_i^\top \boldsymbol{\alpha} - \boldsymbol{u}_i^\top \boldsymbol{\beta} - \sum_{1 \le j < k \le p} u_{ij} u_{ik} \gamma_{jk} \right)^2 \\ &\quad + \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{l=1}^{L_j} x_{ijl}^* \log \left\{ \mathbb{I}(\overline{\xi}_{j,l-1} < u_{ij} < \overline{\xi}_{j,l}) \right\} - \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{u}_i^\top \boldsymbol{u}_i,\end{aligned}$$

where $f(\boldsymbol{u}_i) = \prod_{j=1}^{p} f(u_{ij})$ is the $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$ density, and terms constant with respect to $\boldsymbol{\Psi}$ are dropped. In practice, one could add a small amount $\epsilon > 0$ to the third term e.g., $\log\left\{\mathbb{I}(\overline{\xi}_{j,l-1} < u_{ij} < \overline{\xi}_{j,l}) + \epsilon\right\}$ so that $\ell_c(\boldsymbol{\Psi}, \boldsymbol{u})$ remains finite for all $\boldsymbol{u}$. However, as we shall see below our proposed importance sampling algorithm for the E-step ensures that $\mathbb{I}(\overline{\xi}_{j,l-1} < u_{ij} < \overline{\xi}_{j,l}) = 1$ is always satisfied for every j. For fixed λ and a set of adaptive weights, the MCEM algorithm involves iterating between the following two steps until convergence. At iteration t, suppose we have estimates $\hat{\boldsymbol{\Psi}}^{(t)}$. In the E-step, we calculate the expectation of the complete log-likelihood with respect to the conditional distribution of the latent indices, also known as the Q function, $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(t)}) = \int \ell_c(\boldsymbol{\Psi}, \boldsymbol{u}) f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}^*, \hat{\boldsymbol{\Psi}}^{(t)}) d\boldsymbol{u}$. In the M-step, we obtain an updated estimate $\hat{\boldsymbol{\Psi}}^{(t+1)}$ that maximizes (or at least leads to an increase in) the function

$$Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(t)}) - \lambda \sum_{j=1}^{p} \left(w_j \beta_j^2 + \sum_{k=1}^{j-1} w_{kj} |\gamma_{kj}| + \sum_{k=j+1}^{p} w_{jk} |\gamma_{jk}| \right)^{1/2}.$$

To perform the E-step, we propose using importance sampling. Specifically, for each $i = 1, \ldots, n$, suppose we obtain M samples $\{\boldsymbol{u}_i^m = (u_{i1}^m, \ldots, u_{ip}^m); m = 1, \ldots, M\}$ from a proposal distribution $h(\boldsymbol{u}_i)$. For all the simulations and applications later on, we used $M = 1000$. Then we approximate the Q-function as

$$Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(t)}) \approx \sum_{i=1}^{n} \sum_{m=1}^{M} v_i^m \ell_{ci}(\boldsymbol{\Psi}, \boldsymbol{u}_i^m), \tag{3}$$

where

$$v_i^m = \frac{f(y_i|\boldsymbol{u}_i^m, \boldsymbol{z}_i, \hat{\boldsymbol{\Psi}}^{(t)}) \prod_{j=1}^{p} f(\boldsymbol{x}_{ij}^*|u_{ij}^m, \hat{\boldsymbol{\Psi}}^{(t)}) f(\boldsymbol{u}_i^m) h(\boldsymbol{u}_i^m)^{-1}}{\left(\sum_{m=1}^{M} f(y_i|\boldsymbol{u}_i^m, \boldsymbol{z}_i, \hat{\boldsymbol{\Psi}}^{(t)}) \prod_{j=1}^{p} f(\boldsymbol{x}_{ij}^*|u_{ij}^m, \hat{\boldsymbol{\Psi}}^{(t)}) f(\boldsymbol{u}_i^m) h(\boldsymbol{u}_i^m)^{-1}\right)}.$$

We propose sampling from a truncated multivariate normal distribution as follows. Let $\mathcal{TN}_p(\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{a}, \boldsymbol{b})$ generically denote the truncated p-dimensional multivariate normal distribution with location vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{A}$, and $\boldsymbol{a}$ and $\boldsymbol{b}$ are the vectors of the lower and upper truncation points respectively. Then we use

$$h(\boldsymbol{u}_i) = \mathcal{TN}_p\left(\hat{\boldsymbol{\Sigma}}^{(t)} \hat{\boldsymbol{\beta}}^{(t)} (y_i - \boldsymbol{z}_i^\top \hat{\boldsymbol{\alpha}}^{(t)}), \hat{\boldsymbol{\Sigma}}^{(t)}, \overline{\boldsymbol{\zeta}}_-, \overline{\boldsymbol{\zeta}}_+\right), \tag{4}$$

where $\hat{\boldsymbol{\Sigma}}^{(t)} = (\boldsymbol{I}_p + (\hat{\sigma}^{(t)})^{-2} \hat{\boldsymbol{\beta}}^{(t)} (\hat{\boldsymbol{\beta}}^{(t)})^\top)^{-1}$, $\boldsymbol{I}_p$ is the identity matrix of dimension p, $\overline{\boldsymbol{\zeta}}_-^{(t)} = (\sum_{l=1}^{L_1} x_{i1l}^* \overline{\xi}_{1,l-1}, \ldots, \sum_{l=1}^{L_p} x_{ipl}^* \overline{\xi}_{p,l-1})$, and $\overline{\boldsymbol{\zeta}}_+ = (\sum_{l=1}^{L_1} x_{i1l}^* \overline{\xi}_{1,l}, \ldots, \sum_{l=1}^{L_p} x_{ipl}^* \overline{\xi}_{p,l})$. There are three connected advantages for using the above as the proposal distribution: 1) suppose all the interaction terms between the latent indices are zero for all j and k. Then applying straightforward algebra to the complete log-likelihood $\ell_c(\boldsymbol{\Psi}, \boldsymbol{u})$, we can show that $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}^*, \hat{\boldsymbol{\Psi}}^{(t)})$ is exactly equal to equation (4) and the E-step collapses to directly sampling from the conditional distribution. This result, namely that an exact conditional distribution to sample from can be obtained, relies on the assumption of normality for the latent indices, and indeed is an additional advantage of assuming the $u_{ij}'s$ are normally distributed.; 2) in many applications of LoLI, we expect the true interactions to be sparse i.e., most elements of $\boldsymbol{\gamma}$ are equal to zero. In such cases, even though equation (4) is not exactly equal to $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}^*, \hat{\boldsymbol{\Psi}}^{(t)})$, it should still be a relatively good approximation; 3) it is clear from the complete log-likelihood $\ell_c(\boldsymbol{\Psi}, \boldsymbol{u})$ that the conditional distribution of the latent indices, $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}^*, \hat{\boldsymbol{\Psi}}^{(t)})$, is bounded above and below by $\overline{\boldsymbol{\zeta}}_+$ and $\overline{\boldsymbol{\zeta}}_-$ respectively. Therefore, it is sensible to choose a proposal distribution whose support coincides with that of the conditional distribution, rather than a proposal distribution defined on $\mathbb{R}^p$ (say). Indeed, using equation (4) simplifies calculation of the importance weights to $v_i^m = f(y_i|\boldsymbol{u}_i^m, \boldsymbol{z}_i, \hat{\boldsymbol{\Psi}}^{(t)}) f(\boldsymbol{u}_i^m) h(\boldsymbol{u}_i^m)^{-1} \left(\sum_{m=1}^M f(y_i|\boldsymbol{u}_i^m, \boldsymbol{z}_i, \hat{\boldsymbol{\Psi}}^{(t)}) f(\boldsymbol{u}_i^m) h(\boldsymbol{u}_i^m)^{-1} \right)^{-1}$ since $\prod_{j=1}^p f(\boldsymbol{x}_{ij}^*|u_{ij}^m, \hat{\boldsymbol{\Psi}}^{(t)}) = \prod_{j=1}^p \prod_{l=1}^{L_j} \mathbb{I}\left(\overline{\xi}_{j,l-1} < u_{ij} < \overline{\xi}_{j,l}\right)^{x_{ijl}^*} = 1$ by definition of the proposal distribution.

With the Q-function approximated using equation (3) and equation (4), a series of conditional M-steps can then be performed to obtain updates $\hat{\boldsymbol{\Psi}}^{(t+1)}$. The details of these updates are provided in Appendix A.2. For both the interaction γ_{jk} and main effect β_j terms, we approximate the penalty in Definition 2.1 using the local linear approximation, thereby facilitating the use of soft threshold operators to efficiently perform coordinate wise optimization. Note predictions of the latent indices can be straightforwardly obtained as part of the MCEM algorithm e.g., for the i-th observation, the prediction $\mathrm{E}(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}^*, \hat{\boldsymbol{\Psi}})$ can be approximated by $M^{-1} \sum_{m=1}^M v_i^m \boldsymbol{u}_i^m$ where v_i^m is discussed above.

It is important to discuss the challenges that would be involved, if we were to also estimate the cutoffs as part of the M-step above, in contrast to our proposed computationally efficient method of estimating them separately. Since the proposal distribution in equation (4) is non-zero in precisely the region defined by the cutoff estimates at iteration t of the MCEM algorithm, it follows that these estimates maximize the Q-function equation (3) and therefore $\hat{\boldsymbol{\xi}}_j^{(t+1)} = \hat{\boldsymbol{\xi}}_j^{(t)}$ i.e., no update can be achieved directly using the EM algorithm. This problem is a special case of a more general issue first formalized by (Ruud, 1991), who showed that the EM algorithm does not work if the support of the conditional distribution of the missing data depends on parameters to be estimated. There are a number of possible ways to overcome this issue. For example, we can reparameterize the model such that cutoff parameters appear in other parts of the complete log-likelihood instead of in the log indicator functions. Even for simple ordinal probit models however, this approach is computationally burdensome as it involves having to construct a vector of latent indices for each u_{ij} itself. Another approach to estimating the cutoffs is to sample from $f(u_{ij})$ directly in the E-step, or at least a distribution with a support not defined by the cutoffs.

However, this is extremely inefficient since the cutoffs themselves will result in a large proportion of Monte-Carlo samples of u_{ij} contributing no weight to the integration. In summary, estimating the cutoffs within the MCEM algorithm presents a major bottleneck in the estimation procedure, and motivates us to propose the above two-step estimation approach.

3.3. Tuning Parameter Selection

We choose the single tuning parameter in Definition 2.1 using the Extended Regularized Information Criterion (ERIC, Hui et al., 2015) developed originally for penalized regression modeling. With our specific data and model structure, ERIC is defined as $\text{ERIC}(\lambda) = -2\ell(\hat{\boldsymbol{\Psi}}) + \log(n\lambda^{-1})\left\{\sum_{j=1}^{p}\mathbb{I}(\hat{\beta}_j \neq 0) + \sum_{1\leq j<k\leq p}\mathbb{I}(\hat{\gamma}_{jk} \neq 0)\right\}$, where $\ell(\hat{\boldsymbol{\Psi}})$ is the unpenalized marginal log-likelihood evaluated at the LoLI estimates, and the model complexity is based on counting the number of estimated non-zero main and interaction coefficients. Note the original definition of ERIC included an additional parameter for tuning the severity of model complexity penalization, but we choose to omit that here for simplicity.

ERIC features a *dynamic* model complexity penalty which depends on the tuning parameter itself. This means the degree of penalization induced by ERIC differs depending on how complex the model is already, as captured by λ. Smaller values of λ lead to more aggressive shrinkage, and result in less overfitting and sparser models. This contrasts with many other information criteria that employ *static* complexity penalties and thus penalize a fixed amount for every coefficient entered into the model (e.g., the AIC and BIC, Zhang et al., 2010). The use of a more aggressive approach to shrinkage, as promoted by ERIC, is particularly appropriate here given both the number of interaction coefficients in LoLI can still be quite large, and *a-priori* we believe that the underlying model is sparse; see the discussion below equation (4). Based on extensive simulations (not shown), we found that this aggressive shrinkage enforced by ERIC leads to better overall selection performance (as assessed based on the mean number of false positives and false negative for the main and interaction effects, similar to the simulation study below) compared to using, say, BIC to choose the tuning parameter for LoLI.

4. Simulation Study

We conducted two simulations to assess the relative performance of LoLI (in conjunction with ERIC) in terms of estimation, variable selection, and prediction of the latent indices. Note that, while estimation consistency of the two-step procedure can be established (meaning the estimates from LoLI are asymptotically unbiased), it is also important to investigate the finite sample bias and variability of these estimates. In the first setting each ordinal predictor is associated with its own latent index, while in the second setting each latent index is associated with a group of ordinal predictors.

For both simulation settings, we considered datasets of size $n = 50, 100, 200$, and for each sample size simulated 500 datasets. We also performed simulations at $n = 400$ and 800, and found similar trends to those discussed below and present these in Appendix B. We assessed estimation performance based on the mean squared errors (MSE), averaged across simulated datasets, of the quantities $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|^2$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$, and $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|^2$, where $(\hat{\boldsymbol{\alpha}}^\top, \hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\gamma}}^\top)^\top$ denotes the estimates from a particular method and $(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ denotes the true parameter values. We assessed selection performance based on the mean number of false positives (FP) and false negatives (FN) separately for main $\boldsymbol{\beta}$ and interaction $\boldsymbol{\gamma}$ effects. We also recorded the mean computation time (in seconds) for each method, along with results for the MSE of the estimates of the cutoffs. The latter are of secondary interest compared to the parameters in the latent indices regression model, but nevertheless still present similar trends to the MSE of these other parameters; we provide the results for these in Appendix B.

4.1. Setting 1

We considered a true model with $q = 3$ predictors which are not dimension reduced and $p = 6$ ordinal predictors. For the former, we generated the covariate vector $\boldsymbol{z}_i$ by setting the first element equal to one representing the intercept, and simulating the remaining two elements independently from a standard normal distribution. We set the corresponding true coefficient vector as $\boldsymbol{\alpha} = (2, 1, -1)^\top$. Next, we generated a vector of $p = 6$ latent indices $\boldsymbol{u}_i$ from a multivariate standard normal distribution, set $\boldsymbol{\beta} = (1, -1, 0.5, 0, 0, 0)^\top$, and set $\gamma_{12} = -0.5$ and $\gamma_{23} = 0.4$ while all remaining interaction terms were set to zero. Hence only the first three latent indices are truly informative and there are two pairwise interactions between these. The mean μ_i was then constructed based on equation (1), and given this the response was generated as $y_i \sim \mathcal{N}(\mu_i, 1)$ i.e., $\sigma^2 = 1$. Finally, we constructed the ordinal six predictors based on the following set of cutoffs: $L_1 = L_2 = 3$ with $\check{\boldsymbol{\xi}}_1 = \check{\boldsymbol{\xi}}_2 = (-1, 1)^\top$, $L_3 = L_4 = 4$ with $\check{\boldsymbol{\xi}}_3 = \check{\boldsymbol{\xi}}_4 = (-1, 0, 1.25)^\top$, and $L_5 = L_6 = 5$ and $\check{\boldsymbol{\xi}}_5 = \check{\boldsymbol{\xi}}_6 = (-1.5, -1, 0.5, 1.5)^\top$. Afterward, we generated the ordinal predictors x_{ij} as per equation (2). That is, we simulated the vector $(x^*_{ij1}, \ldots, x^*_{ijL_j})^\top$ from a multinomial distribution with trial size 1 and probabilities $\mathbb{I}(\check{\xi}_{j,l-1} < u_{ij} < \check{\xi}_{j,l}); l = 1, \ldots, L_j$, and set $x_{ij} = l$ if $x^*_{ijl} = 1$. Finally, to obtain the "true" vector of cutoff parameters associated with LoLI, we set $\boldsymbol{\xi}_j = \check{\boldsymbol{\xi}}_j - \check{\xi}_{j1}$ for $j = 1, \ldots, 6$ such that the first element of $\boldsymbol{\xi}_j$ is always equal to zero (which is required for parameter identifiability in LoLI). To clarify, in analyzing the data we only have access to y_i, $\boldsymbol{z}_i$, and $\boldsymbol{x}_i$. We also conducted simulations with $\sigma^2 = 4$ and 16, reflecting a weaker signal-to-noise ratio; results for these are presented in Appendix B, and exhibit similar conclusions to those below (except all methods performed worse compared with when $\sigma^2 = 1$, as anticipated).

Since each ordinal predictor is associated with its own index, we compared LoLI with three available methods: 1) a penalized likelihood method using a hierarchical LASSO penalty via the `hierNet` package (Bien and Tibshirani, 2014), treating each ordinal predictor in $\boldsymbol{x}_i$ as continuous and using the default ten–fold cross validation to choose the tuning parameter. Note that, in the same way we view LoLI in conjunction with ERIC as a single approach, we also view the hierarchical LASSO penalty in conjunction with cross validation as a single approach, although we acknowledge that future research and comparisons could explore choosing the tuning parameter in LoLI via cross validation, and the general issue of tuning parameter selection for the hierarchical LASSO; 2) backward elimination from a saturated model i.e., all main and pairwise interaction effects between the elements of $\boldsymbol{x}_i$ included, using BIC and treating each ordinal predictor in $\boldsymbol{x}_i$ as continuous; 3) backward elimination from a saturated model using BIC and setting up dummy variables for each ordinal predictor in $\boldsymbol{x}_i$. All three alternative methods respect the hierarchical nature of the covariates i.e., main effects can only be removed from the model if all interaction effects involving it have already been removed. We also included a "gold standard" method where we treated the latent indices as if they were observed and performed backward elimination from a saturated model using BIC. Two stage approaches were not considered in this setting, since we cannot fit latent variable models when each ordinal predictor corresponds to its own latent index.

Not surprisingly, LoLI performs substantially better than the alternative methods (Table 1), with its estimation and selection performance much closer to the "gold standard" method compared to either treating the ordinal predictors as either continuous or constructing dummy variables. LoLI

Table 1: Simulation results for Setting 1 with $\sigma^2 = 1$, comparing LoLI, penalized likelihood using `hierNet`, backward elimination treating the ordinal predictors as continuous (Backward-Cont), backward elimination treating the ordinal predictors as categorical (Backward-Cat), backward elimination assuming the latent indices are assumed known (Backward-True). In the results, FP/FN ($\boldsymbol{\beta}$) refers to the mean number of false positives/mean number of false negatives for the estimates of $\boldsymbol{\beta}$, say. Results are not available for Backward-Cat when $n = 50$, due to the inability to estimate the saturated model with such a small sample size.

n	Criterion	LoLI	hierNet	Backward-Cont	Backward-Cat	Backward-True
	MSE ($\boldsymbol{\alpha}$)	0.160	0.381	129.81	-	0.098
	MSE ($\boldsymbol{\beta}$)	0.366	0.739	50.595	-	0.211
50	MSE ($\boldsymbol{\gamma}$)	0.336	0.385	3.820	-	0.250
	FP/FN ($\boldsymbol{\beta}$)	0.180/0.372	1.382/0.466	1.588/0.140	-	1.554/0.032
	FP/FN ($\boldsymbol{\gamma}$)	0.080/1.460	0.688/1.640	1.862/1.222	-	1.892/0.558
	MSE ($\boldsymbol{\alpha}$)	0.069	0.152	37.35	621.306	0.034
	MSE ($\boldsymbol{\beta}$)	0.140	0.275	17.538	-	0.055
100	MSE ($\boldsymbol{\gamma}$)	0.209	0.307	1.233	-	0.088
	FP/FN ($\boldsymbol{\beta}$)	0.110/0.102	1.362/0.046	0.826/0.050	2.202/0.138	0.758/0.002
	FP/FN ($\boldsymbol{\gamma}$)	0.066/0.922	0.692/1.226	0.806/0.796	6.574/1.000	0.658/0.094
	MSE ($\boldsymbol{\alpha}$)	0.034	0.053	17.217	10.050	0.015
	MSE ($\boldsymbol{\beta}$)	0.051	0.209	12.333	-	0.020
200	MSE ($\boldsymbol{\gamma}$)	0.071	0.187	0.762	-	0.003
	FP/FN ($\boldsymbol{\beta}$)	0.060/0.004	1.564/0	0.508/0	0/0.154	0.424/0
	FP/FN ($\boldsymbol{\gamma}$)	0.052/0.222	0.776/0.518	0.436/0.224	0.002/1.576	0.360/0.002

almost always had the lowest mean number of false positives (indicative of overfitting) without any considerable increase in the mean number of false negatives (indicative of underfitting). The discrete nature of the backward elimination procedure led to poorer estimation performance compared to the two "continuous" penalized likelihood methods (LoLI and hierNet), while hierNet continued to overfit at large sample sizes relative to LoLI. In terms of computation time (see Appendix B), LoLI was the slowest of the methods, which was not surprising since none of the other methods attempt to recover a latent index for each ordinal predictor (and thus leading to worse performance compared to LoLI). Overall, this simulation provides strong evidence of the benefit of LoLI in a scenario where each ordinal predictor results from discretization of a continuous latent index.

4.2. Setting 2

We considered a true model with $q = 4$ predictors which are not to be dimension reduced, and $p = 10$ ordinal predictors divided into $G = 5$ groups and latent indices. For the former, we generated the covariate vector $\boldsymbol{z}_i$ by setting the first element equal to one and simulating the remaining three elements from a multivariate normal distribution with zero mean vector and an AR1 correlation matrix such that $\text{Cov}(z_{ir}, z_{is}) = 0.4^{|r-s|}; r, s = 2, \ldots, q$. We set the corresponding true coefficient vector as $\boldsymbol{\alpha} = (-1, 1, -1, 0)^\top$. Next, we generated a vector of latent indices $\boldsymbol{u}_i$ from a multivariate standard normal distribution, set $\boldsymbol{\beta} = (1, 0.5, 0, 0, 1)^\top$, and set $\gamma_{12} = -0.8$ while the remaining nine interaction terms were set to zero. This implies the first, second, and fifth latent indices are truly informative, and there is only one non-zero pairwise interaction between the first and second indices. The mean μ_i was then constructed as discussed in Section 2.1, and given this the response was generated as $y_i \sim \mathcal{N}(\mu_i, 1)$ i.e., $\sigma^2 = 1$. Again, we conducted simulations with $\sigma^2 = 4$ and 16, and the results for these are presented in Appendix B and exhibit similar trends to those below.

We constructed the ten ordinal predictors based on the following groupings: $\mathcal{A}_1 = \{1, 2, 3\}, \mathcal{A}_2 = \{4, 5\}, \mathcal{A}_3 = \{6, 7\}, \mathcal{A}_4 = \{8, 9\}, \mathcal{A}_5 = \{10\}$. Note the fifth group contains one ordinal predictor. Furthermore, we considered the following set of cutoffs for the ten predictors: $L_1 = \ldots = L_5$ with $\check{\boldsymbol{\xi}}_j = (-1, 0, 2)^\top$ for $j = 1, 2$ and $\check{\boldsymbol{\xi}}_j = (-1, 0, 1.25)^\top$ for $j = 3, 4, 5$, then $L_6 = \ldots = L_{10} = 5$ and $\check{\boldsymbol{\xi}}_j = (-1.5, -1, 0.5, 1.5)^\top$ and $j = 6, \ldots, 10$. For groups 1 to 4 where $p_g > 1$, we set the slope parameter $a_j = 1$. Based on these parameters, we generated the ordinal predictors x_{ij} as in Section 2.1. Specifically, we simulated $(x^*_{ij1}, \ldots, x^*_{ijL_j})$ from a multinomial distribution with trial size 1 and probabilities given by $\{\Phi(\xi_{j,l} - a_j u_{ig}) - \Phi(\xi_{j,l-1} - a_j u_{ig})\}; l = 1, \ldots, L_j$ for $j = 1, \ldots, 9$ and by $\mathbb{I}(\check{\xi}_{j,l-1} < u_{ig} < \check{\xi}_{j,l}); l = 1, \ldots, L_j$ for $j = 10$, and set $x_{ij} = l$ if $x^*_{ijl} = 1$. Finally, to obtain the "true" vector of cutoff parameters associated with LoLI, we set $\boldsymbol{\xi}_j = \check{\boldsymbol{\xi}}_j - \check{\xi}_{j1}$ for $j = 1, \ldots, 6$ such that the first element of $\boldsymbol{\xi}_j$ is always equal to zero.

We compared LoLI with two commonly used two stage approaches: 1) a factor analytic model assuming five factors is fitted to all 10 ordinal predictors in the first stage, and then backward elimination using BIC is applied to a linear model with the five predicted latent indices included at the second stage (FA); 2) a graded response model assuming five factors is fitted in the first stage, and then backward elimination using BIC is applied to a linear model with the four predicted latent indices included at the second stage (GRM). We also included a "gold standard" method where the latent indices are treated as observed and performed backward elimination from a saturated model using BIC. In addition to point estimation and selection performance, because all methods produced predictions of $\boldsymbol{u}_i$, we also assessed predictive performance based on the MSE of the quantity, $n^{-1}\sum_{g=1}^{G}\sum_{i=1}^{n}(\hat{u}_{ig}-u_{ig})^2$, where $\hat{u}_{ig}$ and u_{ig} denotes the predicted and true latent indices respectively.

Compared to the two stage approaches, LoLI consistently had the lowest mean squared errors for the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (Table 2). LoLI also performed strongly in terms of estimating the coefficients for covariates that were not dimension reduced, $\boldsymbol{\alpha}$, although the differences between the three methods were small at larger sample sizes. The strong point estimation performance of LoLI is further reflected in its selection performance, where it almost always had a smaller mean number of false positives and false negatives for both the main and interaction effects. Both two stage approaches had a comparably high number of false negatives even at larger sample sizes, and a more detailed analysis suggested that these methods tended to erroneously shrink the fifth element of $\boldsymbol{\beta}$ (i.e., the latent index with only one ordinal predictor in its group) as well as the single non-zero interaction effect to zero. LoLI also performed best with regards to predicting the latent indices across all three sample sizes, reflecting the benefits of being able to borrow strength from the response to better predict the latent indices. Finally, in terms of computation time (see Appendix B) the two stage approach using FA was the fastest, followed by LoLI, while the two stage approach using GRM was by far the slowest.

5. Application to HILDA survey

We applied LoLI to the HILDA survey to understand the association between different aspects of an individual's psychosocial job quality and their mental health. We considered cross-sectional data from Wave 14 (corresponding to observations collected in 2014) of the survey, and focused on a set of $n = 327$ individuals who had a permanent job, no long-term health condition, and a postgraduate degree as their highest education level attained. For the response, we considered a composite mental health score which varies continuously from 0 to 100 with higher scores representing better mental health. The score is derived from the mental component summary of the Short Form 36 (SF-36) questionnaire within the HILDA survey (see Butterworth et al.,

Table 2: Simulation results for Setting 2 with $\sigma^2 = 1$, comparing LoLI, a two stage approach using a factor analytic model (FA), a two stage approach using a graded response model (GRM), and backward elimination assuming the latent indices are assumed known (Backward-True). In the results, FP/FN ($\boldsymbol{\beta}$) refers to the mean number of false positives/mean number of false negatives for the estimates of $\boldsymbol{\beta}$, say.

n	Criterion	LoLI	FA	GRM	Backward-True
50	MSE ($\boldsymbol{\alpha}$)	0.664	0.728	0.849	0.164
	MSE ($\boldsymbol{\beta}$)	0.654	1.237	1.660	0.122
	MSE ($\boldsymbol{\gamma}$)	0.652	1.164	1.537	0.218
	FP/FN ($\boldsymbol{\beta}$)	0.372/0.814	0.900/1.01	0.993/1.272	0.780/0
	FP/FN ($\boldsymbol{\gamma}$)	0.426/0.738	0.984/0.794	1.240/0.830	1.022/0
	MSE ($\boldsymbol{u}_i$)	0.783	0.954	0.991	-
100	MSE ($\boldsymbol{\alpha}$)	0.139	0.154	0.133	0.053
	MSE ($\boldsymbol{\beta}$)	0.253	1.134	1.808	0.039
	MSE ($\boldsymbol{\gamma}$)	0.446	0.932	1.292	0.037
	FP/FN ($\boldsymbol{\beta}$)	0.292/0.176	0.706/0.658	0.732/1.096	0.434/0
	FP/FN ($\boldsymbol{\gamma}$)	0.700/0.356	0.696/0.712	0.808/0.800	0.440/0
	MSE ($\boldsymbol{u}_i$)	0.718	0.913	0.914	-
200	MSE ($\boldsymbol{\alpha}$)	0.041	0.041	0.047	0.025
	MSE ($\boldsymbol{\beta}$)	0.173	0.789	1.289	0.018
	MSE ($\boldsymbol{\gamma}$)	0.264	0.577	0.878	0.014
	FP/FN ($\boldsymbol{\beta}$)	0.109/0.054	0.320/0.678	0.467/1.065	0.250/0
	FP/FN ($\boldsymbol{\gamma}$)	0.091/0.200	0.348/0.674	0.483/0.787	0.243/0
	MSE ($\boldsymbol{u}_i$)	0.716	0.893	0.875	-

2013, and references therein). Of the $n = 327$ individuals, the lowest mental score was 4, while six individuals had the maximum possible mental health score of 100. To remove the boundaries at 0 and 100, we chose to apply a logit transformation, $\log\{(y+4)/(100-y+4)\}$, where the minimum score of 4 was added to ensure all transformed responses were finite (Warton and Hui, 2011). A normal probability plot (not shown) suggested the transformed mental health score was approximately normally distributed.

As covariates which are not dimension reduced i.e., $\boldsymbol{z}_i$, we included age in years (standardized to have zero mean and unit variance) as a linear effect, gender (0 for female; 1 for male), and marital status (0 for married, 1 for otherwise). For the ordinal categorical predictors to be dimension reduced i.e., $\boldsymbol{x}_i$, we considered $p = 12$ statements concerning workplace conditions, to which each individual gives an ordinal score from 1 (strongly disagree) to

7 (strongly agree). A table of the statements can be found in Appendix C. Based on existing literature on the design of the statements (e.g., Butterworth et al., 2011), as well as exploratory analysis involving fitting graded response models with various numbers of latent variables, we grouped the $p = 12$ ordinal predictors into $G = 5$ groups reflecting different underlying aspects of workplace quality: 1) degree of job demands/complexity/interest (3 predictors); 2) degree of job control (3 predictors); 3) degree of job stress and strain (2 predictors); 4) degree of job security (3 predictors); 5) effort-reward unfairness (1 predictor). We refer the reader to Appendix C for these groupings. We then applied LoLI based on these $G = 5$ groupings, allowing for all ten pairwise interactions between the latent indices, and using ERIC to select the tuning parameter. Analogously to simulation Setting 2 in Section 4.2, we compared LoLI to two alternative methods: 1) a two stage method where a factor analytic model with five factors is fitted to all 12 predictors in the first stage, and then backward elimination using BIC is applied to a linear model with the predicted factor scores included (FA), 2) a two stage method where a graded response model with five latent variables is fitted to all 12 predictors, and then backward elimination using BIC is applied to a linear model with the predicted latent indices included (GRM).

Based on point estimates alone, all three approaches suggested that improved mental health was associated with individuals who were older, male, and married (Table 3). All three approaches also indicated that increased job demands/complexity/interest improved mental health, while higher job stress/strain had a strong detrimental impact on mental health. Only LoLI and the two stage approach using GRM provided evidence of a non-zero effect of increased job security on improved mental health, with a similar magnitude of effect to that of increased job demands/complexity/interest. LoLI further indicated a positive interaction between job demands/complexity/interest and job security. That is, the positive effect of both increased job interest and increased job security on an individual's mental health was greater than each aspect acting on its own. We also considered scatterplot pairs of the predicted latent indices, and provide the results and discussion for these in Appendix C.

To assess the variability of these point estimates, we calculated estimates of uncertainty for all three methods. For the two stage approaches using FA and GRM, these were obtained based on standard errors calculated from the linear model in the second model from the `lm` function in `R`. Keep in mind that these do not account for the uncertainty in the prediction of the latent

Table 3: Estimated coefficients and residual variance based on: 1) LoLI, 2) a two stage approach using a factor analytic model (FA) with five factors in the first stage, 3) a two stage approach using a graded response model (GRM) with five factors in the first stage. Coefficients eliminated from the final model are denoted with a ".", while uncertainty estimates are shown for all parameters in parentheses.

Predictor	LoLI	FA	GRM
Intercept	1.088 (0.068)	1.095 (0.068)	1.131 (0.069)
Age	0.065 (0.041)	0.058 (0.041)	0.045 (0.042)
Gender (male)	0.108 (0.089)	0.108 (0.083)	0.101 (0.083)
Martial Status (no)	-0.148 (0.084)	-0.126 (0.086)	-0.163 (0.086)
$\hat{\beta}_1$ (job demands/complexity/interest)	0.089 (0.039)	0.131 (0.046)	0.151 (0.036)
$\hat{\beta}_2$ (job control)	.	.	.
$\hat{\beta}_3$ (job stress/strain)	-0.258 (0.043)	-0.264 (0.048)	-0.247 (0.039)
$\hat{\beta}_4$ (job security)	0.085 (0.040)	0.115 (0.045)	.
$\hat{\beta}_5$ (effort-reward unfairness)	.	.	.
$\hat{\gamma}_{12}$	.	.	.
$\hat{\gamma}_{13}$	.	.	.
$\hat{\gamma}_{14}$	0.084 (0.032)	.	.
$\hat{\gamma}_{15}$	.	.	.
$\hat{\gamma}_{23}$	.	.	.
$\hat{\gamma}_{24}$	.	.	.
$\hat{\gamma}_{25}$	.	.	.
$\hat{\gamma}_{34}$	.	.	.
$\hat{\gamma}_{35}$	.	.	.
$\hat{\gamma}_{45}$	.	.	.
$\hat{\sigma}^2$	0.531	0.546	0.549

indices or the model selection uncertainty (to our knowledge, there are no publicly available R packages that implement the two stage methods *and* account for either source of uncertainty). However for LoLI, it is not obvious how to produce estimates of uncertainty for the non-zero estimates with the two-step estimation approach. Therefore, we adopted an *ad-hoc* approach and calculated an empirical information matrix based on the unpenalized log-likelihood $\ell(\mathbf{\Psi})$ in equation (2), where the cutoffs were held fixed at the estimates obtained from the penalized fit i.e., as in Section 3.1, and all coefficients not selected by LoLI were set to zero. Put another way, we can interpret these as uncertainty estimates for a type of "post-LoLI" unpenalized maximum likelihood estimator. In detail, we calculated the information matrix $\hat{\boldsymbol{I}}(\hat{\mathbf{\Psi}}_1) = n^{-1}\sum_{i=1}^{n}(\partial\ell_i(\hat{\mathbf{\Psi}})/\partial\mathbf{\Psi}_1)(\partial\ell_i(\hat{\mathbf{\Psi}})/\partial\mathbf{\Psi}_1)^\top$, where $\mathbf{\Psi}_1$ denotes the coefficients that were selected from LoLI and $\hat{\mathbf{\Psi}}$ denotes the full vector of parameter estimates obtained from LoLI. We then constructed estimates of uncertainty based on the the diagonal elements of $\hat{\boldsymbol{I}}^{-1}(\hat{\boldsymbol{\Psi}}_1)$. We recognize that future research should explore other approaches to calculate information matrices for LoLI, the related issue of developing uncertainty estimates and confidence intervals when using an adaptive LASSO penalty in general (Potscher and Schneider, 2009; Potscher and Leeb, 2009), as well as the more general problem of post model selection inference, (although note this problem would apply to all three methods of selection here; see Lee et al., 2016).

Interestingly, all three methods showed no clear evidence that age, gender, or marital status had substantial effects on mental health. On the other hand, all three methods declared their respective selected main and interaction effects of job quality as having substantive effects. In particular, LoLI confirmed clear evidence of main effects of job demands/complexity/interest, job stress/strain, and job security, as well as an important synergistic effect of job demands/complexity/interest and job security.

6. Discussion

We have proposed a new approach called the LASSO on Latent Indices for handling ordinal predictors in regression modeling, which jointly constructs a latent index for each or for groups of ordinal predictors and models the response directly as a function of these and their interactions. LoLI incorporates a LASSO type penalty to perform selection of the main and interaction effects associated with the latent indices in a hierarchical manner. Simulations show that, compared to dummy variables or two stage approaches, LoLI, in conjunction with a more aggressive approach to choosing the tuning parameter, produced more accurate predictions of the latent indices and better selection of the associated coefficients. Applying LoLI to the HILDA survey revealed the compounding effects of high job demands and job strain on poor mental health, and a positive synergistic effect of high job security and low job strain on improved mental health.

One way to view LoLI is as a special type of (penalized) measurement error type model (Carroll et al., 2006), where instead of an additive error the true latent covariate is discretized into an ordinal predictor. While this connection is not particularly useful in terms of its actual application, it nevertheless offers an interesting insight into how the nature and implications of the measurement error in LoLI is more complicated than that of the standard measurement error model. We explore this idea in detail in Appendix D.

There are a multitude of ways in which LoLI can be extended and explored, with noteworthy ones being to constrain some of the cutoffs across ordinal predictors to be the same if (for example) the same rating scale is used for multiple predictors, how to handle cases of where some levels are not observed at all for one or more ordinal predictors, assessing the robustness of the LoLI approach to different sources of model misspecification, and, along related lines, considering distributions aside from the normal for the latent indices (although with such an extension the attractiveness of the truncated multivariate normal distribution for importance sampling is possibly diminished). A related extension would be to allow the latent indices to be correlated (but perhaps still normally distributed), in which case the two-step estimation procedure would still be possible except the multiple cutoffs in the first step would be estimated simultaneously via a joint cumulative probit regression (say); the penalty in LoLI may have to be altered though to account for the possible collinearity between the latent indices. In addition, how to construct predictions using LoLI e.g., predict the response given a set of covariates and ordinal predictors for a new individual, would be of interest in further explorations. A simple approach may be to construct the prediction based on the marginal log-likelihood function in equation (2), but extended further to account for the uncertainty of the estimated parameters. However, more sophisticated and efficient approaches may also be possible, such as a hot-deck imputation type method based on matching the new set of covariates to those in the existing dataset and then developing some sort of weighted prediction for the latent indices from this.

The issue of high-dimensionality i.e., when the number of ordinal predictors p grows with sample size n, is also worthy of future theoretical and empirical study. Finally, one important extension of LoLI is data driven approaches to choosing both the groupings and the number of groupings (latent indices). As proposed in this article, LoLI requires any groupings of the ordinal predictors to be defined *a-priori*, and for the motivating HILDA survey there was considerable existing literature we could utilize to construct these groups. To relax this, we could draw each ordinal response x_{ij} from a finite mixture of G multinomial distributions, where each component multinomial distribution is associated with a different latent u_{ig}. Alternatively, we may not explicitly form groups at all but instead model all p ordinal predictors against a set of $G < p$ (possibly correlated) latent indices, and then use penalties to select both G and the implicit groupings by shrinking elements and/or entire columns of the relevant loading matrix to zero (Hui et al., 2018).

Acknowledgements

This research was supported by the Australian Research Council discovery project grant. Thanks to the Associated Editor and Reviewers for their insightful comments, and to Peter Straka and Pauline O'Shaughnessy for useful discussions.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. Wiley Series in Probability and Statistics. Wiley.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41:1111–1141.

Bien, J. and Tibshirani, R. (2014). *hierNet: A Lasso for Hierarchical Interactions*. R package version 1.6.

Butterworth, P., Leach, L., McManus, S., and Stansfeld, S. (2013). Common mental disorders, unemployment and psychosocial job quality: is a poor job better than no job at all? *Psychological medicine*, 43:1763–1772.

Butterworth, P., Leach, L. S., Rodgers, B., Broom, D. H., Olesen, S. C., and Strazdins, L. (2011). Psychosocial job adversity and health in Australia: analysis of data from the HILDA Survey. *Australian and New Zealand journal of public health*, 35:564–571.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press, Florida.

Fu, Z., Parikh, C. R., and Zhou, B. (2017). Penalized variable selection in competing risks regression. *Lifetime data analysis*, 23:353–376.

Garcia, T. P. and Mueller, S. (2016). Cox regression with exclusion frequency-based weights to identify neuroimaging markers relevant to Huntingtons disease onset. *The Annals of Applied Statistics*, 10:2130–2156.

Haris, A., Witten, D., and Simon, N. (2016). Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25:981–1004.

Hui, F. K. C., Mueller, S., and Welsh, A. H. (2017). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica*, 27:501–518.

Hui, F. K. C., Tanaka, E., and Warton, D. I. (2018). Order selection and sparsity in latent variable models via the ordered factor LASSO. *Biometrics*, 74, 1311–1319.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2015). Tuning Parameter Selection for the Adaptive Lasso using ERIC. *Journal of the American Statistical Association*, 110:262–269.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

Leach, L., Butterworth, P., Rodgers, B., and Strazdins, L. (2010). Deriving an Evidence-Based Measure of Job Quality from the HILDA Survey. *Australian Social Policy Journal*, 9:67–86.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44:907–927.

Lee, S.-Y., Poon, W.-Y., and Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, 48:339–358.

Linting, M., Meulman, J. J., Groenen, P. J., and van der Koojj, A. J. (2007). Nonlinear principal components analysis: introduction and application. *Psychological methods*, 12:336.

Milner, A., Aitken, Z., Kavanagh, A., LaMontagne, A. D., and Petrie, D. (2016). Persistent and contemporaneous effects of job stressors on mental health: a study testing multiple analytic approaches across 13 waves of annually collected cohort data. *Occupational & Environmental Medicine*, 73:787–793.

Milner, A., Butterworth, P., Bentley, R., Kavanagh, A. M., and LaMontagne, A. D. (2015). Sickness Absence and Psychosocial Job Quality: An Analysis From a Longitudinal Survey of Working Australians, 2005–2012. *American journal of epidemiology*, 181:781–788.

Pauger, D., Wagner, H., et al. (2019). Bayesian effect fusion for categorical predictors. *Bayesian Analysis*, 14:341–369.

Potscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100:2065–2082.

Potscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference*, 139:2775–2790.

Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm. *Journal of Econometrics*, 49:305–341.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34:1–97.

She, Y., Wang, Z., and Jiang, H. (2016). Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association*, Accepted for publication.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* CRC Press.

Tutz, G. and Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, 16:161–200.

Warton, D. I. and Hui, F. K. C. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 92:3–10.

Watson, N. and Wooden, M. P. (2012). The HILDA survey: a case study in the design and development of a successful household panel survey. *Longitudinal and Life Course Studies*, 3:369–381.

Wedel, M. and Kamakura, W. (2001). Factor analysis with (mixed) observed and latent variables. *Psychometrika*, 66:515–530.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704.

Yan, X. and Bien, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, 32:531–560.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67.

Zhang, Y., Li, R., and Tsai, C. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105:312–323.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468–3497.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.