BIOMETRIC METHODOLOGY

Concave likelihood-based regression with finite-support response variables

K.O. Ekvall¹,² M. Bottai²

¹Department of Statistics, University of Florida, Gainesville, Florida, USA
²Division of Biostatistics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

Correspondence
K.O. Ekvall, Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, Gainesville, FL 32611, USA.
Email: k.ekvall@ufl.edu

Abstract
We propose a unified framework for likelihood-based regression modeling when the response variable has finite support. Our work is motivated by the fact that, in practice, observed data are discrete and bounded. The proposed methods assume a model which includes models previously considered for interval-censored variables with log-concave distributions as special cases. The resulting log-likelihood is concave, which we use to establish asymptotic normality of its maximizer as the number of observations $n$ tends to infinity with the number of parameters $d$ fixed, and rates of convergence of $L_1$-regularized estimators when the true parameter vector is sparse and $d$ and $n$ both tend to infinity with $\log(d)/n \to 0$.

We consider an inexact proximal Newton algorithm for computing estimates and give theoretical guarantees for its convergence. The range of possible applications is wide, including but not limited to survival analysis in discrete time, the modeling of outcomes on scored surveys and questionnaires, and, more generally, interval-censored regression. The applicability and usefulness of the proposed methods are illustrated in simulations and data examples.

KEYWORDS
concave, finite support, interval-censoring, maximum likelihood, penalized, regression, survival analysis

1 | INTRODUCTION

In practice observed data are discrete and bounded, be it by design, because of limited measurement precision, or because the data are stored in finite precision, for example as floating point numbers in a computer. However, it is common to ignore this and use models assuming continuous distributions, or continuous models for short. In general, this practice leads to misspecification and biased estimators. While the bias can be small in some settings, it can be substantial in others, and the practice nevertheless persists. In our experience, this is in part due to a lack of reliable methods for the correctly specified likelihood and an unawareness of the potential pitfalls. To address these issues, we propose methods with theoretical and computational guarantees for a flexible class of regression models for finite-support (i.e., discrete and bounded) response variables. In addition, we illustrate the bias that can result from incorrectly applying a continuous model using simulations.

We consider four data examples, two of which are provided as Supporting Information. Each example is in a different setting with different challenges. Given the ubiquitous use of continuous models with data with finite support, the examples can illustrate but a small fraction of the many potential applications for the proposed methods.
The first data example (Section 6.1) focuses on the effects of clinical predictors on plasma lipoprotein(a) \( \text{Lp}(a) \) levels measured in the clinical care. \( \text{Lp}(a) \) is measured with finite precision, has a lower limit of detection, and a natural upper bound. Thus, in practice \( \text{Lp}(a) \) has finite support.

In the second example, cancer patients are observed repeatedly over the course of a study. For each patient, the time to death or distant metastases is recorded. Interest can be in univariable modeling of the time-to-event or, as is the focus in Section 6.2, the effects of clinical predictors and prediction using gene expressions. Either way, time-to-event has finite support: patients do not live forever, and time is measured with finite precision. Additionally, in many studies patients can be observed only at a few specific time points, leading to the observable time-to-event being far from continuous.

The third example (Web Appendix A.1) has an ordinal response, the total score on a depression screening questionnaire, taking values in \( \{0, 1, \ldots, 27\} \), and illustrates how the proposed methods can be used in settings where there need not exist a latent continuous variable of interest. The fourth example (Web Appendix A.2) focuses on discovering genes that predict, or are associated with, glucose intolerance. The response is ordinal with three levels, and the number of predictors is three orders of magnitude larger than the number of observations.

Now, regardless of application, any response variable \( Y \) with finite support \( \mathcal{Y} \) can be modeled using the categorical distribution parameterized by the category probabilities. When the number of categories, that is, the cardinality of \( \mathcal{Y} \), is small relative to the number of observations, it may be possible to estimate those probabilities with acceptable precision using the corresponding sample proportions. However, when the number of categories is large, their probabilities depend on predictors, or there is a known relation between the probabilities, then further modeling is often needed. We consider a model which handles many practically relevant settings and which leads to estimators with theoretical support. Specifically, we assume the probability mass function for \( Y \) given a non-stochastic predictor vector \( x \in \mathbb{R}^p \) can be expressed, for functions \( a \) and \( b \) to be specified, as

\[
\begin{equation}
 f(y \mid x) = \int_{a(y,x,\theta)}^{b(y,x,\theta)} r(w) \, dw = R[b(y,x,\theta)] - R[a(y,x,\theta)],
\end{equation}
\]

where \( r \) is a log-concave Lebesgue-density on \( \mathbb{R} \), \( R \) is the corresponding cumulative distribution function, and \( \theta \) is a parameter vector. We will assume that \( a \) and \( b \) are affine in \( \theta \) for every \( (y,x) \) and give further details on the specification in Section 2.

Intuitively, Equation (1) can be understood as the mass function for an interval-censored latent, continuous random variable \( W \) with density \( r \). In some settings, \( W \) has a practical interpretation. For example, it is typically related to the unobservable continuous time-to-event in settings such as the cancer study discussed above. On the other hand, Equation (1) is also useful in many settings where there is no latent variable of practical interest. In fact, Example 1 establishes that, when there are no predictors, any mass function for a categorical random variable can be obtained as a special case of Equation (1).

Authors considering models like Equation (1) include Burridge (1981, 1982) who note that, in some cases of interest, the log-likelihood is concave. These and some related results are discussed in the review of methods for grouped data by Heitjan (1989). At the time, much of the literature was concerned with adjusting methods for continuous data to address bias introduced by grouping. By contrast, the focus here is the development of methods based on the correct likelihood. Likelihood-based methods for settings related to ours include that by Finkelstein (1986), who proposed a model for interval-censored failure time data. Gentleman and Geyer (1994) gave statistical and computational guarantees for maximum likelihood estimates under interval-censoring of a non-parametric model for survival times, and Huang (1996) provided convergence rates for maximum likelihood estimators in interval-censored proportional hazards models. More recently, Taraldsen (2011) studied the special case of rounded exponential data in detail, Zeng et al. (2016) proposed methods for interval-censored survival times, Couso et al. (2017) discussed different coarsening processes, and Guillaume et al. (2017) proposed robust optimization methods for coarse data in an essentially non-parametric setting. Kowal and Canale (2020) also proposed a non-parametric method, for integer-valued data, mentioning rounded data as a relevant special case. McGough et al. (2021) studied penalized regression for censored and truncated, but not interval-censored, data. Notably, many of the applications are in survival analysis, which is natural given that time is generally measured in discrete units. A thorough treatment of survival analysis in discrete time is given by Tutz and Schmid (2016). Here, we consider a unified framework including some models for survival analysis as special cases.

While some special cases of Equation (1), for example, logistic regression (see Section 2) and cumulative probability models (Example 1), have been studied extensively, the general setting has not. We give intuitive conditions on the density \( r \) and endpoints \( a \) and \( b \) which guarantee asymptotic normality of the maximum likelihood estimator when the number of observations \( n \) grows with the number of parameters \( d \) fixed. Essentially, an asymptotic rank condition on a model matrix and \( r \) being continuously differentiable suffices. We also consider settings where \( d \) tends to infinity with \( \log(d)/n \to 0 \), and give convergence...
rates for an $L_1$-regularized maximum likelihood estimator under a restricted eigenvalue condition on a model matrix and $r$ continuously differentiable. Finally, we establish the numerical convergence of an inexact proximal Newton algorithm under conditions similar to those ensuring statistical convergence.

2 | MODEL

Let $\Theta \subseteq \mathbb{R}^d$ be a convex parameter set and suppose $Y_1, \ldots, Y_n$ are independent, each having a mass function consistent with Model 1:

$$f_\theta(y_i | x_i) = \int_{r(u(y, \theta), \theta)} r(u(y, x_i, \theta)) = R[a(y, x_i, \theta)],$$

where $x_i \in X \subseteq \mathbb{R}^p$. We will often write $f_\theta(y_i | x_i)$ instead of $f_\theta(y_1 | x_1)$ for brevity. The support $\mathcal{Y}_i = \{y_i : f_\theta(y_i | x_i) > 0\}$ need not be the same for every $i$, but we will assume $\mathcal{Y} = \cup_{i=1}^n \mathcal{Y}_i$ is finite. Define $r(\infty) = r(-\infty) = R(\infty) = 0$ and $R(\infty) = 1$.

We assume, for $Z_i = Z_i(y_i, x_i) \in \mathbb{R}^{2 \times d}$ and $m_i = m_i(y_i, x_i) \in [-\infty, \infty]^d$ to be defined shortly, $[a_i(y_i, x_i, \theta), b_i(y_i, x_i, \theta)] = Z_i \theta + m_i$. When writing $Z_i$ and $m_i$ for brevity, dependence on $(y_i, x_i)$ is implicit. Denote the first and second element of $m_i$ by, respectively, $m_i^a$ and $m_i^b$. Accordingly, denote the first and second row of $Z_i$ by, respectively, $z_i^a$ and $z_i^b$. We assume that if $m_i^a(y_i, x_i) = -\infty$ for some $y_i \in \mathcal{Y}_i$, then it holds for every $x_i \in X$; and, for those $y_i$, we let $z_i^b(y_i, x_i) = 0$ for every $x_i \in X$. Thus, for every $x_i \in X$; and $z_i^b(y_i, x_i) = 0$.

The following three examples illustrate definitions and connections to some common models. Example 1 shows that, when there are no predictors, any model for a categorical response is a special case of (1), while Examples 2 and 3 include predictors.

Example 1 (Cumulative probability models). Consider a response $Y$ with $m$ possible values, without loss of generality $\mathcal{Y} = \{1, \ldots, m\}$. A possible version of (1) assumes $f_\theta(y)$ is defined by

$$f_\theta(1) = R(\theta_1), \quad f_\theta(2) = R(\theta_2) - R(\theta_1), \ldots, \quad f_\theta(m-1) = R(\theta_{m-1}) - R(\theta_{m-2}), \quad f_\theta(m) = 1 - R(\theta_{m-1}),$$

with parameter set $\Theta = \{\theta \in \mathbb{R}^{m-1} : \theta_j \geq \theta_{j-1}, j \in \{2, \ldots, m-1\}\}$. In the notation of (1), without predictors, $a(y, \theta) = -\infty$ if $y = 1$, and $a(y, \theta) = \theta_{y-1}$ otherwise. Similarly, $b(y, \theta) = \infty$ if $y = m$ and $b(y, \theta) = \theta_y$ otherwise. One may also write (3) as $P_\theta(Y \leq j) = R(\theta_j), j \in \{1, \ldots, m-1\}$, which shows cumulative probability models are a special case of (1); see, for example, Agresti (2019, Section 6.2), who uses a different but equivalent parameterization.

Example 2 (Interval-censored regression). Suppose for some $\sigma > 0$, $\beta \in \mathbb{R}^p$, and $W_i$ with log-concave Lebesgue-distribution $r$ on $\mathbb{R}$, independently for $i \in \{1, \ldots, n\}$.

$$Y_i^* = x_i^T \beta + \sigma W_i.$$ (4)

Suppose also, for some $k_i \geq 1$ and known cut points $-\infty = t_0^i < t_1^i < \cdots < t_{k_i}^i < t_{k_i+1}^i = \infty$, the observed response is

$$Y_i = \begin{cases} y_i^{(0)}, & Y_i^* \in (t_0^i, t_1^i) \\ \vdots & \vdots \\ y_i^{(k_i)}, & Y_i^* \in [t_{k_i}^i, t_{k_i+1}^i) \end{cases},$$

where the interval labels $y_i^{(0)}, \ldots, y_i^{(k_i)}$ are arbitrary. Common binary regression models such as probit and logistic regression are special cases with, for every $i$, $Y_i = \{0, 1\}$, known $\sigma = 1$, and $W_i$ having standard normal or logistic distribution, respectively. More generally, in the parameterization $\theta = [\sigma^{-1}, \sigma^{-1} \beta^T] \in \mathbb{R}^{p+1}$, for $j \in \{0, \ldots, k_i\}$,

$$f_\theta(y_i^{(j)} | x_i) = R([t_{j+1}^i, -x_i^T \theta]) - R([t_j^i, -x_i^T \theta]),$$ (5)

which is consistent with (1). In particular, $a_i(y_i, x_i, \theta) = -\infty$ if $y_i = y_i^{(0)}$ and, otherwise, $a_i(y_i, x_i, \theta) = \theta_I^T z_i^a = \theta_I^T [t_{j+1}^i, -x_i^T \theta] + m_i^a = 0$. Similarly, $b_i(y_i, x_i, \theta) = \infty$ if $y_i = y_i^{(k_i)}$ and, otherwise, $b_i(y_i, x_i, \theta) = \theta_I^T z_i^b = \theta_I^T [t_{j+1}^i, -x_i^T \theta]$ and $m_i^b = 0$.

Without predictors, (5) is similar to (3). However, in (5) without predictors the arguments to $R$ are determined by the known cut points and one parameter, $\theta_1$, while in (3) the cut points are parameters.

Example 3 (Interval-censored flexible parametric survival models). Royston and Parmar (2002) introduced a class of flexible parametric models for survival analysis. One model assumes that a survival time $T$ has
cumulative distribution function $F(t; x, \beta, \gamma) = 1 - \exp \{- \exp[\exp(\log t; \gamma) - \beta^T x]\}$, where $\exp(\log t; \gamma)$ is a spline of log(t) with coefficients $\gamma$. Any $\exp(\cdot; \gamma)$ which is monotone increasing for every $\gamma$ in the parameter set and tends to $\pm \infty$ when its argument does, gives a valid cumulative distribution function. The exponential distribution is a special case with $\exp(\log t; \gamma) = \log t$. In practice, what is observed is often an interval containing $T$, say

$$Y = \begin{cases} y(0), & T \in [0, t_1) \\ \vdots \\ y(k), & T \in [t_k, \infty) \end{cases},$$

where $t_0 = 0 < t_1 < \cdots < t_k < \infty$ the observation subscript $i$ is suppressed for simplicity. Thus, for example, $f_\theta(y^{(1)} \mid x) = F(t_2; x, \beta, \gamma) - F(t_1; x, \beta, \gamma)$. Using this it is straightforward to show the mass function for $Y$ satisfies (1) with $R(u) = 1 - \exp\{-\exp(u)\}$, $\theta = [\gamma^T, \beta^T]^T$, and $z^\theta$ and $z'^\theta$ defined accordingly.

Next, we establish concavity of the log-likelihood. The log-likelihood for one observation is $\ell_i(\theta; y_i, x_i) = \log(f_\theta(y_i \mid x_i))$, and $\ell_n(\theta; Y, X) = \sum_{i=1}^n \ell_i(\theta; y_i, x_i)$, where $Y = [Y_1, \ldots, Y_n]^T \in \mathbb{R}^n$ and $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times p}$.

**Theorem 1.** The log-likelihood $\ell_n(\cdot; Y, X)$ given by model (2) is concave on $\Theta$. Moreover, if $\sum_{i=1}^n Z_i^T Z_i$ is positive definite and $r$ is strictly positive, strictly log-concave, and continuously differentiable; then $\ell_n(\cdot; Y, X)$ is strictly concave on every open, convex subset of $\Theta$.

The proof of Theorem 1 uses classical results on log-concave functions due to Prékopa (1973) and is in the Supporting Information along with proofs of other formally stated results. A special case of the non-strict concavity given by Theorem 1 is discussed without proof by Burridge (1982, p. 150). We have not seen the strict part, which requires substantially more work, stated or proved before. The essential component in its proof is Lemma B.1 (Supporting Information) which establishes strict log-concavity of the map $(t_1, t_2) \mapsto [R(t_2) - R(t_1)]$. The strict log-concavity is critical for our results with diverging number of parameters.

## 3 | ASYMPTOTIC PROPERTIES

### 3.1 | Fixed number of parameters

We consider maximum likelihood estimators

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} G_n(\theta; Y, X),$$

where $G_n(\theta; Y, X) = -n^{-1} \ell_n(\theta; Y, X)$. Because $\ell_n(\cdot; Y, X)$ is concave on the convex $\Theta$ (Theorem 1), $\hat{\theta}_n$ is a solution to a stochastic convex optimization problem, which is used in the proofs of our main asymptotic results.

In results and their proofs $c_j \in (0, \infty)$, $j \in \{1, 2, \ldots\}$, denote generic constants which can change between statements but, in each statement, depend on neither of $i$, $n$, $d$, $Y$, $X$, or $\theta$. We use $\| \cdot \|$ for the spectral norm for matrices and Euclidean norm for vectors, $\| \cdot \|_\infty$ for the max-norm (maximum absolute element), and $\| \cdot \|_1$ for the one-norm (sum of absolute elements). The true parameter is denoted as $\theta^\ast$.

The following assumption will be used in both the low- and high-dimensional settings.

**Assumption 1.** For all small enough $\rho > 0$, there is a compact $E \subseteq \{ t \in \mathbb{R}^2 : t_1 < t_2 \}$ such that, for every $i \in \mathbb{N} = \{1, 2, \ldots\}$, $y_i \in Y$, $x_i \in X$, and $\theta \in \Theta$ with $\|\theta - \theta^\ast\|_1 \leq \rho$, it holds that either $Z_i \theta + m_i \in E$ or an element of $m_i$ is infinite. Moreover, for some $c_1 < \infty$, $\|Z_i\|_\infty \leq c_1$ and, when the left-hand sides are finite, $|m_i^0| \leq c_1$ and $|m_i^0| \leq c_1$.

The particular choice of norms in Assumption 1 is unimportant when $d$ is fixed but will matter in later sections when $d \to \infty$. To get some intuition for the first part of the assumption, consider, for example, the interval-censored regression in Example 2. As noted following (5), when both $t_{ij}$ and $t_{ij+1}$ are finite, $m_i = 0$ and $\delta_i^0 \delta_i^1 = \sigma^{-1} t_{ij}^1 - x_i^1 \beta / \sigma < \sigma^{-1} t_{ij+1}^1 - x_i^1 \beta / \sigma$. Using this and that $\sigma > 0$, it is straightforward to show Assumption 1 holds (see Proof of Corollary 1, Supporting Information, for an example). More generally, Assumption 1 ensures among other things that the support does not depend on $\theta$ near $\theta^\ast$.

To state the first result, let $\lambda_{\min}(\cdot)$ denote the smallest eigenvalue of its matrix argument.

**Theorem 2.** If (a) $Y$ is finite, (b) $r$ is strictly log-concave, strictly positive, and continuously differentiable on $\mathbb{R}$; (c) $\theta^\ast$ is an interior point of $\Theta$; (d) Assumption 1 holds; and (e)

$$\liminf_{n \to \infty} \lambda_{\min}\left\{ \frac{1}{n} \sum_{i=1}^n E(Z_i^T Z_i) \right\} > 0; \quad (6)$$

then as $n \to \infty$ with $d$ fixed, $I_n(\theta^\ast; X)^{1/2} (\hat{\theta}_n - \theta^\ast) \leadsto \mathcal{N}(0, I_d)$, where $I_n(\theta^\ast; X) = \text{cov}\{\nabla \ell_n(\theta^\ast; Y, X)\}$ is the Fisher information.

The proof of Theorem 2 uses a result by Hjort and Pollard (2011) on minimizers of convex processes. The expectation and covariance in the theorem statement are with respect to the distribution of $Y \mid X$ under the true
In the proof, it is argued that \( I_n(\Theta; X) \) has eigenvalues bounded below by \( c \) for some \( c > 0 \). With this, the theorem implies \( \| \hat{\Theta}_n - \Theta_s \| = O_p(1/\sqrt{n}) \). If \( r \) is assumed to be twice continuously differentiable, then the conclusion of the theorem continues to hold if \( I_n(\Theta; X) \) is replaced by the observed information \( -V^2 \ell_n(\hat{\Theta}_n, Y, X) \) (Theorem B.5, Supporting Information). In the case of interval-censored linear regression, Equation (6) reduces to a familiar condition on the design matrix \( X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{nxp} \).

**Corollary 1.** Suppose \( Y_1, \ldots, Y_n \) satisfy the interval-censored regression model in Example 2 with known \( \sigma = 1 \), \( \mathcal{Y} \) is finite, the density of \( W_i \), satisfies requirement (b) of Theorem 2, and \( \| x_i \| \leq c_1 \); then the conclusion of Theorem 2 holds if \( \liminf_{n \to \infty} \lambda_{\min}(X^T X/n) > 0 \) as \( n \to \infty \) with \( d = p \) fixed.

When \( k \geq 2 \) in the setting of Corollary 1, we expect the conclusion can be shown to hold also when \( \sigma \) is unknown. Intuitively, when the support of the response variables has cardinality greater than two, the variance need not be a function of the mean, and it may then be possible to estimate an additional parameter. By contrast, it is well-known \( \sigma \) is unidentifiable in general in logistic and probit regression, which are special cases.

### 3.2 Diverging number of parameters

Our second main result gives convergence rates for maximum \( L_1 \) regularized likelihood estimators when \( d \) tends to infinity with \( n \), \( \Theta_s \) is sparse. Since \( d \) varies \( \Theta_s \) generally depends on \( d \), but we suppress this in notation.

We consider the penalized average negative log-likelihood defined for \( \lambda_n > 0 \) by \( G_n(\Theta; Y, X) = G_0(\Theta; Y, X) + \lambda_n \| \Theta \|_1 \), and \( \hat{\Theta}_n \in \arg \min_{\Theta \in \mathbb{R}^d} G_n(\Theta; Y, X) \). For any \( \Theta \in \mathbb{R}^d \) and \( S \subseteq \{1, \ldots, d\} \), define \( \Theta_S \in \mathbb{R}^d \) to equal \( \Theta \) with the \( j \)th element set to zero if \( j \not\in S \), \( j \in \{1, \ldots, d\} \):

\[
(\Theta_S)_{j} = \begin{cases} 
\Theta_{j} & j \in S \\
0 & j \not\in S 
\end{cases}
\]

We say \( \Theta \) is \( s \)-sparse if \( \Theta = \Theta_S \) for some \( S \subseteq \{1, \ldots, d\} \) with cardinality \( s = |S| \).

To state results, define the cone \( \mathcal{C}(S) = \{ \Theta \in \mathbb{R}^d : \| \Theta_S \|_1 \leq 3 \| \Theta_S \|_1 \} \), where \( S^c = \{1, \ldots, d\} \setminus S \) and, hence, \( \Theta_S^c = -\Theta_S \). Intuitively, \( C(S) \) is a set of nearly-sparse \( \Theta \) in the sense that the elements \( \Theta_j \), \( j \not\in S \) are not too large compared with the \( \Theta_j \), \( j \in S \). Define also for any \( k > 0 \), \( n \), and \( d \) the set

\[
\mathcal{C}_{x,n,d} = \{ (Y, X) : \inf_{\Theta \in \mathcal{C}(S) : \Theta \neq 0} \left\{ \Theta^T \left( \frac{1}{n} \sum_{i=1}^{n} Z_i^T Z_i \right) \Theta \right\} \geq k \}.
\]

We are ready to state the next result.

**Theorem 3.** If (a) \( \Theta \) is open, (b) \( r \) is strictly log-concave, strictly positive, and continuously differentiable on \( \mathbb{R} \); (c) \( \Theta_s \) is \( s \)-sparse and \( \| \Theta_s \|_1 \leq c_1 \); (d) Assumption 1 holds; and (e) \( \lambda_n = c_2 \log(d)/n \to 0 \); then there are \( c_1, c_2, c_3 \) such that, for large enough \( n \) and \( d \), with probability at least \( P(C_{x,n,d}) - d^{-c_3} \),

\[
\| \hat{\Theta}_n^\lambda - \Theta_s \|^2 \leq c_4 \frac{\log(d)}{n}, \quad \| \hat{\Theta}_n^\lambda - \Theta_s \|_1 \leq c_5 \sqrt{\frac{\log(d)}{n}}.
\]

Assumptions (a) and (b) ensure the gradient and Hessian of \( G_n \) exist. In some settings of interest, for example, interval-censored regressions with known error variance, the matrices \( Z_1, \ldots, Z_n \) do not depend on \( Y \). Then, the event \( C_{x,n,d} \) either contains all outcomes or none and is hence better thought of as a restricted eigenvalue condition on the deterministic \( \sum_{i=1}^{n} Z_i^T Z_i/n \). Specifically, if the inequality in the definition of \( C_{x,n,d} \) holds for some \( k > 0 \) and all \( n \) and \( d \), and the other conditions of the theorem hold, then the conclusion of the theorem holds with probability at least \( 1 - d^{-c_3} \). Moreover, in interval-censored regression with known error variance, a restricted eigenvalue condition on \( \sum_{i=1}^{n} Z_i^T Z_i/n \) is equivalent to one on \( X^T X/n \) since, in those cases, \( Z_i = -[x_i, x_i]^T \), with one of the rows replaced by zeros if \( a_i(y_i, x_i, \Theta) = -\infty \) or \( b_i(y_i, x_i, \Theta) = \infty \).

It is common in the literature for the bounds on norms of \( \hat{\Theta}_n^\lambda - \Theta_s \) to depend linearly on \( s \) (e.g., Negahban et al., 2012, Corollary 2). Here, \( s \) is fixed and absorbed in the constants \( c_4 \) and \( c_5 \). This is because our proofs require \( Z_i \Theta_s \) to be contained in a compact subset of \( \mathbb{R}^2 \). We expect the linear dependence on \( s \) can be recovered in many special cases, though it may require substantial work; see, for example, Negahban et al. (2009) for the special case of logistic regression.

### 4 COMPUTING

#### 4.1 Inexact proximal Newton

We propose using an inexact proximal Newton algorithm for computing \( \hat{\Theta}_n^\lambda \) in practice. That is, a proximal Newton algorithm where the sub-problems are solved inexactly. Similar algorithms have proven useful in, for example, the fitting of penalized generalized linear models (Byrd et al., 2016; Friedman et al., 2010; Lee et al., 2006; Yuan et al., 2012). The R package **fnsnet** (2022) implements the algorithm, and an accelerated proximal gradient descent algorithm similar to the Fast Iterative Shrinkage-Thresholding (FISTA) algorithm (Beck & Teboulle, 2009). We focus on the proximal Newton algorithm here because we found it tends to perform well. It is often useful in practice to include a ridge penalty and hence we solve the
convex elastic-net optimization problem
\[
\min_{\theta \in \Theta} \left\{ G_n(\theta; Y, X) + \lambda_1 \|\theta\|_1 + \frac{\lambda_2}{2} \|\theta\|^2 \right\},
\]
where \(\lambda_1 \geq 0\) and \(\lambda_2 \geq 0\) are user-specified penalty parameters. The setting in Section 3.2 is a special case with \(\lambda_2 = 0\). If \(\lambda_2 > 0\) the objective function is strongly convex and has a unique global minimizer. To simplify notation, let us suppress dependence on the data \((Y, X)\) for the remainder of the section and re-define \(G_n^1\) to include the ridge penalty. That is, \(G_n^1(\theta)\) is the objective function in Equation (7).

Proximal Newton solves Equation (7) by iteratively updating and minimizing an \(L_1\)-penalized quadratic approximation of \(G_n^1\). To be more specific, let \(Q(\cdot; \theta^k)\) denote a quadratic approximation of the differentiable part of \(G_n^1\) at the \(k\)th iterate \(\theta^k\), given by
\[
Q(\theta; \theta^k) = \{\nabla G_n(\theta^k) + \lambda_2 \theta^k\}^T \theta + \frac{1}{2}(\theta - \theta^k)^T \{\nabla^2 G_n(\theta^k) + \lambda_2 I_d\}(\theta - \theta^k).
\]
Then, the \((k + 1)\)th iterate in the proximal Newton algorithm is
\[
\theta^{k+1} \approx \arg \min_{\theta \in \Theta} \{ Q(\theta; \theta^k) + \lambda_1 \|\theta\|_1 \},
\]
where \(\approx\) indicates it is not necessary to solve the optimization problem exactly (see Section 4.3). The update (8) does not in general admit a closed form solution but can be solved efficiently to desired tolerance using coordinate descent.

4.2 Coordinate descent

To discuss the coordinate descent algorithm for (8), we assume \(\Theta = \mathbb{R}^d\) for simplicity. Settings where some parameters need to be positive (e.g., to ensure monotonic splines in an interval-censored flexible parametric model) or not penalized (e.g., the error scale parameter in an interval-censored regression), could be treated by minor modifications and are supported in our software.

The \((l + 1)\)th iterate for the \(j\)th component in a coordinate descent algorithm for (8) is
\[
\theta_{j}^{k+1} = \arg \min_{\theta_j \in \mathbb{R}} \{ Q(\theta_1^{k+1}, \ldots, \theta_j^{k+1}, \ldots, \theta_d^{k+1}, \theta_j; \theta^k) + \lambda_1 \|\theta_j\|_1 \}.
\]
This is a univariate \(L_1\)-penalized quadratic optimization problem which can be solved in the closed form using the soft-thresholding operator. To be more specific, define \(g: \mathbb{R}^2 \to \mathbb{R}^2\) and \(H: \mathbb{R}^2 \to \mathbb{R}^{2 \times 2}\) as, respectively, the gradient and Hessian of the map \((t_1, t_2) \mapsto \log[R(t_2) - R(t_1)], -\infty < t_1 < t_2 < \infty\). Extend also \(g\) and \(H\) to include points where \(t_1 = -\infty\) by setting the first element of \(g\) and first row and column of \(H\) to zero at such points. Then
\[
\nabla G_n(\theta; Y, X) = -\frac{1}{n} \sum_{i=1}^{n} Z_i^T g(Z_i; \theta + m_i);
\]
\[
\nabla^2 G_n(\theta; Y, X) = -\frac{1}{n} \sum_{i=1}^{n} H(Z_i; \theta + m_i) Z_i.
\]
Let \(\eta_k = Z_i \theta_k + m_i, H^k_{ij} = H(\eta_k), g^k_i = g(\eta_k)\), and
\[
\eta_{k+1}^{i, j} = \sum_{m<j} Z_{i}^{m} \theta_{m}^{k,l+1} + \sum_{m>j} Z_{i}^{m} \theta_{m}^{k,l} + m_i,
\]
where \(Z_{i}^{m}\) is the \(m\)th column of \(Z_i\). Up to terms not depending on \(\theta_j\), the objective function in (9) is
\[
\frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \{-(Z_i^j)^T g_i^k + \lambda_2 \theta_j^k + (\eta_k^i - \eta_{k+1}^{i, j})^T H_i^k Z_i^j \} + \frac{\partial^2}{2 \partial \theta_j^2} \sum_{i=1}^{n} \{-(Z_i^j)^T H_i^k Z_i^j + \lambda_2 \} + \lambda_1 \|\theta_j\|_1.
\]
Using this, a routine calculation shows the minimizer in (9) is
\[
\theta_{j}^{k+1} = \text{soft}\left[\frac{-\frac{1}{n} \sum_{i=1}^{n} (Z_i^j)^T g_i^k + \lambda_2 \theta_j^k + (\eta_k^i - \eta_{k+1}^{i, j})^T H_i^k Z_i^j + \lambda_2}{\frac{1}{n} \sum_{i=1}^{n} (Z_i^j)^T H_i^k Z_i^j + \lambda_2} \right],
\]
where \(\text{soft}(x, \lambda) = \text{sign}(x) \max\{|x| - \lambda, 0\}\). Notably, \(\eta_{k+1}^{i, j}\) is the only term that needs updating in the coordinate descent, making each step fast to compute.

The resulting algorithm is stated in Algorithm 1.

**Algorithm 1** Proximal Newton with coordinate descent

1. Input \(\lambda_1 \geq 0, \lambda_2 \geq 0, \theta^0 \in \Theta^d\)
2. For \(k = 1, 2, \ldots\) until convergence:
   a. Let \(\theta_{k+1} = \theta^k\) and for \(l = 1, 2, \ldots\) until convergence, update \(\theta_{j}^{k+1}\) iteratively for \(j = 1, \ldots, d\) according to Equation (10).
   b. Let \(\theta_{j}^{k}\) be the vector of final iterates from (a) and set \(\theta^{k+1} = (1 - s) \theta^k + s \theta_{j}^{k}\) with \(s \in [0, 1]\) selected by backtracking line-search.
3. Return final iterate \(\theta^{k+1}\) from (2).
4.3 Convergence

Convergence of Algorithm 1 can be guaranteed by selecting appropriate termination criteria for the inner coordinate descent algorithm (step 2(a)) and the backtracking line-search (step 2(b)). It will be convenient to characterize solutions to Equation (7) using the function $J : \Theta \rightarrow \mathbb{R}^d$ defined for $c_1 > 0$ by $J(\theta; c_1) = \nabla G_n(\theta) + \lambda_2 \theta - P_{\lambda_1} \{\nabla G_n(\theta) + \lambda_2 \theta - \theta/c_1\}$, where $P_{\lambda_1}$ is the elementwise projection onto $[-\lambda_1, \lambda_1]$. Routine calculations show, for any $c_1 > 0$, $J(\theta; c_1) = 0$ if and only if $0$ is a sub-gradient of $G_n^\lambda$ at $\theta$ (Byrd et al., 2016; Mizerek & Ulbrich, 2014); that is, $J(\theta; c_1) = 0$ if and only if $\theta$ is a solution to Equation (7). Similarly, $\theta$ is a solution to Equation (8) if and only if $J_Q(\theta; c_1, \theta^k) = 0$, where

$$J_Q(\theta; c_1, \theta^k) = \nabla G_n(\theta^k) + \lambda_2 \theta^k + \{\nabla^2 G_n(\theta^k)$$

$$+ \lambda_2 I_d \theta - \theta/c_1\} + \lambda_2 \theta^k + \{\nabla^2 G_n(\theta^k) + \lambda_2 I_d \theta - \theta/c_1\} - \theta/c_1.$$

Following Byrd et al. (2016), the coordinate descent algorithm for Equation (9) may be terminated when the $i$th coordinate descent iterate $\theta^{k,i} = [\theta_1^{k,1}, ..., \theta_p^{k,1}]^T$ satisfies, for a user-specified $c_2 \in [0,1]$,

$$\|J_Q(\theta^{k,i}; \theta^k, c_1)\| \leq c_2 \|J_Q(\theta^k; \theta^k, c_1)\|. \quad (11)$$

To specify a termination criterion for the line-search in step 2(b), define a first-order approximation of $G_n^\lambda$ at $\theta^k$ by

$$L_n^\lambda(\theta; \theta^k) = G_n(\theta^k) + 0.5 \lambda_2 \|\theta^k\|^2 + \{\nabla G_n(\theta^k)$$

$$+ \lambda_2 \theta^k\}^T(\theta - \theta_k) + \lambda_1 \|\theta\|_1.$$

Given $\theta^{k,i}$ satisfying Equation (11), backtracking line-search starts with step-size $s = 1$ and decreases until, for a user-specified $c_3 \in (0, 1/2)$,

$$G_n^\lambda(\theta^k) - G_n^\lambda((1-s)\theta_k + s\theta^{k,i}) \geq c_3 [L_n^\lambda(\theta^k; \theta^k)$$

$$- L_n^\lambda((1-s)\theta_k + s\theta^{k,i}; \theta^k)]. \quad (12)$$

We are ready to state a convergence result for Algorithm 1.

**Theorem 4.** If in Algorithm 1 convergence in step 2(a) is determined using Equation (11), the backtracking linesearch in step 2(b) satisfies (12), $r$ is continuously differentiable, and either

(a) $r$ is strictly log-concave and strictly positive, $\sum_{i=1}^n Z_i^T Z_i$ is positive definite, and $\inf_{\theta \in \Theta} G_n^\lambda(\theta) = G_n^\lambda(\hat{\theta})$ for some $\hat{\theta} \in \Theta$; or

(b) $\lambda_2 > 0$;

then the sequence $\{\theta^k\}$ of iterates satisfies

$$\lim_{k \to \infty} J(\theta^k; c_1) = 0.$$

Conditions (a) and (b) are used to show, among other things, the iterates $\{\theta^k\}$ stay in a compact set. If this can be guaranteed by other means, some conditions can be weakened. For example, it is typically possible to relax the first two requirements in (a) if the gradient is Lipschitz-continuous and the Hessian in the quadratic approximation $Q(\theta; \theta^k)$ is regularized to have eigenvalue bounded away from zero. Notably, we have had no convergence issues in simulations even when $\lambda_2 = 0$ and $\sum_{i=1}^n Z_i^T Z_i$ is indefinite because $d > n$, as long as $\lambda_1 > 0$.

5 | NUMERICAL EXPERIMENTS

We illustrate the proposed methods in two interval-censored regression models (see Example 2). In the first, $W$ has the extreme-value distribution with cumulative distribution function $R(w) = 1 - \exp(-\exp(w))$, and the number of predictors $p = 3$ is smaller than the number of observations $n = 100$. When $W$ has the extreme-value distribution in Equation (4) and $\sigma = 1$, $\exp(Y^*)$ has the exponential distribution with mean $\exp(x^T \theta)$. This model is a special case of that in Example 3. It is also a special case of a gamma generalized linear model with logarithm link function, which we therefore include in comparisons. The observed response indicates whether $\exp(Y^*)$ is in $[0, d), [d, 2d),..., [kd, 5),$ or $[5, \infty)$, where $d$ (the interval size) varies in the simulations and $k$ is the largest integer such that $kd < 5$. Thus, a larger $d$ corresponds to more severe censoring. Because $Y^*$ is not observed, when fitting the generalized linear model we take the upper endpoints of the observed intervals, or $5 + d$ if the interval is $[5, \infty)$, as responses.

In the second setting, $W$ is normally distributed and $p = 200 > 100 = n$. The observed intervals are for $Y^*$ are $(-\infty, 5), [5, -kd),..., [d, 0), [0, d),..., [kd, 5), [5, \infty)$. We compare the estimates from Algorithm 1 with $\lambda_2 = 0$ to those from lasso regression using glmnet (Friedman et al., 2010). For both methods, the regularization parameter $\lambda_1$ is selected by 5-fold cross-validation. For our method, we select the $\lambda_1$ which minimizes the average out-of-sample misclassification rate. Here, the misclassification rate for one fold is the proportion of observations $(Y_i, x_i)$
in that fold for which the predicted mean $x_i^T \hat{\theta}^*_i$ of the $i$th unobservable response $Y^*_i$ is outside the observed interval.

The predictors are generated as centered and scaled realizations from a multivariate normal distribution with mean zero and a covariance matrix with $(i, j)$th element $0.5|j-i|$. When $p = 3$ we include an intercept so there are two jointly normal predictors in addition to the intercept. The true coefficient vector is $\theta^* = \begin{bmatrix} 1, 1/2, -1/2 \end{bmatrix}^T$ when $p = 3$ and $\theta^* = \begin{bmatrix} 1, 1/2, -1/2, 0, \ldots, 0 \end{bmatrix}^T$ when $p = 200$.

Figure 1 shows how sum of squared estimation errors for the non-zero components of $\theta^*$, and mean misclassification rates vary with the interval size $d$. The sum of squared estimation errors is defined as $\sum_{j=1}^{3} \sum_{i=1}^{m} (\hat{\theta}^*_j - \theta^*_j)^2$, where $m = 500$ is the number of replications in the simulations and $\hat{\theta}^*_j$ is an estimate of the $j$th element of $\theta^*$ in the $i$th replication.

The first row of Figure 1 shows, as expected, using the correct likelihood is beneficial, and the benefits are greater the more severe the interval-censoring. We note the mean misclassification rate for the generalized linear model decreases as the interval-censoring gets more severe, which is an effect of it being easier to predict the correct interval when the intervals are larger.

The second row in Figure 1 indicates the proposed method can, when intervals are small enough, perform similarly to that based on the incorrect normal likelihood; that is, to lasso regression. Some intuition for this can be gained by considering the bias-variance trade-off in estimating $\theta^*$: bias is introduced by using the incorrect likelihood, but if the intervals are small enough that bias is small in comparison to the variance. Indeed, the large variance in high-dimensional settings is a key reason regularization, which introduces bias but decreases variance, is often useful. As the censoring becomes more severe, however, the bias is again substantial.

6 | DATA EXAMPLES

6.1 | Lipoprotein data

Lipoprotein(a) [Lp(a)] is a risk factor for cardiovascular complications (see, for example, Littmann et al. (2019, 2022)). Hence, it is of interest to model the distribution of Lp(a) in different populations and to investigate the effects of covariates. One challenge is that Lp(a) has a lower limit of detection of 10 nanomoles per liter (nmol/L), leading to censoring from below. Additionally, in practice Lp(a) is often categorized, into classes, such as those defined by deciles. We consider a regression model for Lp(a) in nmol/L in intervals $[0, 10], (10, 20], \ldots, (110, 120], (120, \infty)$. The data are a subset of those used by Littmann et al. (2019), except they use different classes, $[0, 10], (10, 30], (30, 120], (120, \infty)$. There are $n = 1837$ observations and four covariates: sex, age, smoking status (never-smoker, ex-smoker, or smoker), and hemoglobin A1c (HbA1c) measurements categorized into three levels (low, average, high) corresponding to good, average, and poor metabolic control.

We first fit a model for Lp(a) without predictors. One possibility is to fit the cumulative probability model in...
Example 1. As argued there, this is equivalent to fitting a general categorical model with 12 parameters, the number of categories minus one. For concreteness, take $R$ to be the standard normal cumulative distribution function and denote the maximum likelihood estimate by $\hat{\beta}^c$. This estimate ensures $R(\hat{\beta}^c_j) = n^{-1} \sum_{i=1}^{n} Y_i I(Y_i \leq 10j), j \in \{1, \ldots, 12\}$, where $I(\cdot)$ is an indicator function. That is, the estimated category probabilities equal the sample proportions.

Another possibility is to assume the Lp(a) measurements come from a censored regression model such as in Example 2 with an intercept only. This can be particularly useful when interest is in inference on the unobservable continuous Lp(a). Because the model for the unobservable continuous Lp(a) is the same regardless of the censoring, this model facilitates pooling data from studies with different censoring. Since Lp(a) must be positive, we consider the model which assumes the continuous Lp(a), $Y^*_i$, satisfies

$$\log(Y^*_i) = \beta + \sigma W_i,$$

where $W_i \sim \mathcal{N}(0, 1)$. Our response $Y_i$ is the interval containing $Y^*_i$. Fitting this model we get the maximum likelihood estimates $\hat{\beta} = 3.02$ and $\hat{\sigma} = 1.76$. These can be interpreted as usual in the latent regression, or one can focus on the estimated mass function for $Y_i$ given in Figure 2. Notably, the estimated probabilities are reasonably close to the sample proportions, or equivalently, the estimates from the cumulative probability model.

The maximized likelihood for the latent regression model will always be lower than that of the cumulative probability model since the latter is equivalent to a general categorical model (Example 1). However, the former has fewer parameters and so may still be preferable. For example, computing the BIC for both models shows the smaller is preferable with a BIC of 7481.

To investigate the effect of covariates, we continue with a latent regression model: $\log(Y^*_i) = x_i^T \beta + \sigma W_i$, $W_i \sim \mathcal{N}(0, 1)$. Following Littmann et al. (2019), we consider the effect of age on Lp(a), and whether there are interactions between age and the other covariates. Considering the interactions first, we compare two models using a likelihood ratio test, a smaller one where

$$x_i = [1, \text{age}_i, \text{male}_i, \text{never_smoker}_i, \text{smoker}_i, \text{average_hbac}_i, \text{high_hbac}_i]^T,$$

and a larger one where $x_i$ also includes age interacted with all the other predictors. The likelihood ratio test with 5 degrees of freedom gave a $p$-value of 0.19, indicating the interactions are not important.

Coefficient estimates and standard errors based on the observed information for the smaller model are in Table 1. The reported $p$-values are for Wald-type tests for whether a regression coefficient is zero and whether the scale parameter $\sigma = 1$. Any $p$-value less than $10^{-4}$ is reported as 0. In summary, there is evidence that Lp(a) increases with age and is associated with poor metabolic control.

6.2 Breast cancer data

We use data from the Netherlands Cancer Institute on $n = 144$ lymph node positive women (van de Vijver et al., 2002). Following Tutz and Schmid (2016, Examples 7.1 and 7.2), we model the time to development of distant metastases or death, in 3-month intervals up to 15 months. For each patient, the data include a follow-up time and an event indicator. The observable intervals are [0, 3), ..., [12, 15), or one of those intervals with the upper endpoint replaced.
Table 1 shows regression for censored Lp(a)

| Scale | Int. | Age | Male | Nev. Smoker | Smoker | Med. HbA1c | High HbA1c |
|-------|------|-----|------|-------------|--------|------------|------------|
| Est.  | 1.74 | 2.6 | 0.011| -0.13       | -0.20  | -0.41      | 0.16       |
| S.E.  | 0.046| 0.21| 0.0028| 0.088      | 0.12   | 0.16       | 0.11       |
| p-Value | 0   | 0   | 0   | 0.13 | 0.082 | 0.010 | 0.14 | 0.0067 |

Figure 3 shows estimated survival probabilities for breast cancer data.

by ∞ if the event (death or distant metastases) was not observed.

The data also include five clinical predictor variables (diameter of tumor > 2 cm or not, number of affected lymph nodes ≤ 3 or not, estrogen receptor status positive or negative, tumor grade in three levels, and age) and gene expression measurements for 70 genes. We first consider a model using the clinical variables only, and then investigate whether the gene expression data can be used to improve out-of-sample predictions.

Suppose, as in Example 3, the continuous, unobservable time-to-event $T_i$ has cumulative distribution function

$$F(t; x_i, \beta, \gamma) = 1 - \exp[-\exp[sp(\log t; \gamma) - \beta^T x_i]]$$

where $sp$ is a spline function. Specifically, we pick the I-splines discussed by Ramsay (1988) with no knots and 3 degrees of freedom, implemented in the R package splines2 (Wang & Yan, 2021). These splines are monotone if the elements of $\gamma$ are non-negative, which we therefore enforce when fitting. Exponential and Weibull interval-censored models are special cases corresponding to, respectively, $sp(\log t; \gamma) = \log t$ and $sp(\log t; \gamma) = \gamma \log t$. The three models are nested and upon fitting and comparing them using likelihood ratio tests, we got the $p$-value 0.83 when testing the flexible I-splines against Weibull, 0.88 for Weibull against exponential, and 0.53 for the flexible I-splines against the exponential.

Figure 3 shows the two models give similar survival probabilities at the relevant points, consistent with the large $p$-values obtained when comparing the different models. We focus on the exponential model for the remainder of the section.

Table 2 shows results from fitting the exponential model. The reported standard errors are square roots of diagonal entries of the inverse of the observed Fisher information matrix. The $p$-values are Wald-type and are for the null hypotheses that coefficients are zero. The number of affected lymph nodes appears to be an important predictor, and there is some evidence the tumor grade may be important.

We next consider prediction using the gene expression measurements. Let $h_i \in \mathbb{R}^{70}$ be a vector of gene expression measurements, standardized to have sample mean zero and unit sample variance. We are interested
in whether the $h_i$ can be used to improve the predictive performance of our method, and if so, selecting genes useful for that purpose. To investigate we randomly split the data into a test set of $144/3 = 48$ observations and a training set of 96 observations. We consider the exponential model with predictor vector $[x_i^T, h_i^T]^T \in \mathbb{R}^7$ and coefficient vector $\theta = [\beta^T, \alpha^T]^T \in \mathbb{R}^7$, so $\alpha \in \mathbb{R}^7$ is the coefficient vector for the gene expressions. Consider the estimators $\hat{\theta} = (\hat{\beta}, \hat{\alpha}) \in \arg\min_{\theta \in \mathbb{R}^7} \{G_n(\theta) + \lambda \|\alpha\|_1\}$ and $\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^7} G_n(\beta^T, 0^T)$, where $n_i = 96$ is the number of observations in the training set. The former estimator penalizes the coefficients for the gene expression measurements while the latter assumes those coefficients are zero. Thus, $\hat{\beta}$ is the maximum likelihood estimator in the exponential model without gene expressions, using the training set only.

The penalty parameter $\lambda$ was selected from the set ${\exp(-10), \ldots, \exp(0)}$ by five-fold cross-validation on the training set. This gave $\lambda = \exp(-5) \approx 0.0067$, which attained an average classification rate of 0.29 over the five folds. Predictions on the test-set with the selected $\lambda$ gave an out-of-sample classification rate of 0.31. By comparison, using the clinical predictors only, that is, the predictions $\exp(x_i^T \hat{\beta})$, gave a classification rate of 0.44. We conclude the gene expression measurements can improve prediction, agreeing with the findings of Tutz and Schmid (2016).

With $\lambda = \exp(-5)$, 32 of the 70 elements of $\hat{\alpha}$ were zero. The Supporting Information contains a trace plot showing how the number of non-zero coefficients and their sizes vary with $\lambda$.

### 7 CONCLUSION

The fact that observed data have finite support ought to be considered before using models for continuous random variables, which in general leads to misspecification bias. Roughly speaking, the smaller the cardinality of the support and the variance of maximum likelihood estimators are, the more pronounced the misspecification bias is. Even in settings where the bias is small, however, the effects of using a misspecified likelihood can be difficult to assess, leading to unreliable inference. With the methods proposed here practitioners have access to fast and reliable likelihood-based inference, in both low- ($n \leq d$) and high-dimensional ($d > n$) regression problems. There is a wide range of possible applications, including but not limited to survival analysis in discrete time, ordinal regression, and interval-censored linear regression. Moreover, while the presented theory made repeated use of the concavity of the log-likelihood, Equation (1) gives a valid model even if $r$ is not log-concave. Thus, the modeling framework can be extended to many settings not discussed in the present paper.

Possible directions for future research include the development of theory for the interplay between the severity of censoring and the properties of maximum likelihood estimators. For example, it may be informative to consider asymptotics where the length of the censoring intervals is allowed to change with the sample size and the number of parameters. Additionally, several special cases of the models considered herein are also of significant interest in their own right, and may hence merit further study. As noted in Section 3, more informative high-dimensional convergence bounds can likely be obtained for special cases. It may also be worthwhile to explore settings with dependent data. In the present setting, some types of dependent responses may be analyzed by joining their supports. For example, two dependent binary responses can be recoded as one response with four possible outcomes. The present setting could also in principle be extended to include random effects in the linear predictors, but the theory and implementation would require substantial work.

### ACKNOWLEDGMENTS

We thank Aaron Molstad for helpful discussions and Jonatan Risberg for contributions to the software implementing the proposed methods. We are grateful for comments from two reviewers and an Associate Editor which led to significant improvements.

### DATA AVAILABILITY STATEMENT

The data that support the findings in Section 6.1 are not openly available as the authors do not have permission to share it. The data that support the findings in Section 6.2 are openly available in the R package penalized (2022). The data that support the findings in Web Appendix A.1 are available in the Supporting Information. The data that support the findings in Web Appendix A.2 are openly available in the R package glmnet (2022).

### ORCID

K.O. Ekvall [https://orcid.org/0000-0001-8085-4353](https://orcid.org/0000-0001-8085-4353)
M. Bottai [https://orcid.org/0000-0002-3531-2739](https://orcid.org/0000-0002-3531-2739)
REFERENCES

Agresti, A. (2019) An introduction to categorical data analysis, 3rd edition, Wiley series in probability and statistics. Hoboken, NJ: Wiley.

Beck, A. & Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1), 183–202.

Burridge, J. (1981) A note on maximum likelihood estimation for regression models using grouped data. Journal of the Royal Statistical Society. Series B (Methodological), 43(1), 41–45.

Burridge, J. (1982) Some unimodality properties of likelihoods derived from grouped data. Biometrika, 69(1), 145–151.

Byrd, R.H., Nocedal, J. & Oztoprak, F. (2016) An inexact successive quadratic approximation method for L-1 regularized optimization. Mathematical Programming, 157(2), 375–396.

Couso, I., Dubois, D. & Hüllermeier, E. (2017) Maximum Likelihood Estimation and Coarse Data. In: Moral, S., Pivert, O., Sánchez, D. & Marin, N. (Eds.), Scalable uncertainty management, Lecture notes in computer science, Springer, pp. 3–16.

Finkelstein, D.M. (1986) A proportional hazards model for interval-censored failure time data. Biometrics, 42(4), 845.

Friedman, J.H., Hastie, T. & Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22.

Gentleman, R. & Geyer, C.J. (1994) Maximum likelihood for interval censored data: consistency and computation. Biometrika, 81(3), 618–623.

Guillaume, R., Couso, I. & Dubois, D. (2017) Maximum likelihood with coarse data based on robust optimisation. In: Proceedings of the tenth international symposium on imprecise probability: theories and applications, Lugano, Switzerland: JMLR.org. pp. 169–180.

Heitjan, D.F. (1989) Inference from grouped continuous data: a review. Statistical Science, 4(2), 164–179.

Hjort, N.L. & Pollard, D. (2011) Asymptotics for minimisers of convex processes. arXiv, https://doi.org/10.48550/arXiv.1107.3806

Huang, J. (1996) Efficient estimation for the proportional hazards model with interval censoring. Annals of Statistics, 24(2), 540–568.

Kowal, D.R. & Canale, A. (2020) Simultaneous transformation and rounding (STAR) models for integer-valued data. Electronic Journal of Statistics, 14(1), 1744–1772.

Lee, S.I., Lee, H., Abbeel, P. & Ng, A.Y. (2006) Efficient L1 regularized logistic regression. Aaaai, 6, 401–408.

Littmann, K., Hagström, E., Häbel, H., Bottai, M., Eriksson, M., Parini, P. & Brinch, J. (2022) Plasma lipoprotein(a) measured in the routine clinical care is associated to atherosclerotic cardiovascular disease during a 14-year follow-up. European Journal of Preventive Cardiology, 28(18), 2038–2047.

Littmann, K., Wodaje, T., Alvarsson, M., Bottai, M., Eriksson, M., Parini, P. & Brinch, J. (2019) The association of Lipoprotein(a) plasma levels with prevalence of cardiovascular disease and metabolic control status in patients with Type 1 diabetes. Diabetes Care, 43(8), 1851–1858.

McGough, S.F., Incerti, D., Lyalina, S., Copping, R., Narasimhan, B. & Tibshirani, R. (2021) Penalized regression for left-truncated and right-censored survival data. Statistics in Medicine, 40(25), 5487–5500.

Milzarek, A. & Ulbrich, M. (2014) A semismooth Newton method with multidimensional filter globalization for L1-optimization. SIAM Journal on Optimization, 24, 298–333.

Negahban, S., Yu, B., Wainwright, M.J. & Ravikumar, P. (2009) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. & Culotta, A. (Eds.) Advances in neural information processing systems (Vol. 22). La Jolla, CA: Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/dc583a306451c9d60c6d37004f48f-Paper.pdf

Negahban, S.N., Ravikumar, P., Wainwright, M.J. & Yu, B. (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statistical Science, 27(4), 538–557.

Prékopa, A. (1973) On logarithmic concave measures and functions. Acta Scientiarum Mathematicarum, 34, 335–343.

Ramsay, J.O. (1988) Monotone regression splines in action. Statistical Science, 3(4), 425–441.

Royston, P. & Parmar, M.K.B. (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine, 21(15), 2175–2197.

Talaldsen, G. (2011) Analysis of rounded exponential data. Journal of Applied Statistics, 38(5), 977–986.

The fsnet package (2022) GitHub. Available at: https://github.com/koekvall/fsnet [Accessed 9th September 2022].

The glmnetcr package (2022) The comprehensive R network. Available at: https://CRAN.R-project.org/package=glmnetcr [Accessed 9th August 2022].

The penalized package (2022) The comprehensive R network. Available at: https://CRAN.R-project.org/package=penalized [Accessed 9th August 2022].

Tutz, G. & Schmid, M. (2016) Modeling discrete time-to-event data, 1st edition). Springer series in statistics, Cham: Springer.

van de Vijver, M.J., He, Y.V., van’t Veer, L.J., Dai, H., Hart, A.A.M., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Martin, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H. & Bernards, R. (2002) A gene-expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine, 347(25), 1999–2009.

Wang, W. & Yan, J. (2021) Shape-restricted regression splines with R package splines2. Journal of Data Science, 19(3), 498–517.

Yuan, G.X., Ho, C.H. & Lin, C.J. (2012) An improved glmnet for l1-regularized logistic regression. Journal of Machine Learning Research, 13(64), 1999–2030.

Zeng, D., Mao, L. & Lin, D. (2016) Maximum likelihood estimation for semiparametric transformation models with interval-censored data. Biometrika, 103(2), 253–271.

SUPPORTING INFORMATION

Web Appendices and Figures referenced in Sections 1, 2, 3, and 6, as well as code for the simulations and data examples, are available with this paper at the Biometrics website on Wiley Online Library. The code is also available at https://github.com/koekvall/finite-suppl/.

How to cite this article: Ekvall, K.O. & Bottai, M. (2023) Concave likelihood-based regression with finite-support response variables. Biometrics, 79, 2286–2297. https://doi.org/10.1111/biom.13760