

ULRR

Variable selection using a smooth information criterion for distributional regression models

Item Type	Article
Authors	O'Neill, Meadhbh;Burke, Kevin
Citation	Statistics and Computing 33, 71
Publisher	Springer
Download date	2026-03-14 23:09:17
Item License	https://creativecommons.org/licenses/by-nc-sa/4.0/
Link to Item	https://doi.org/10.34961/researchrepository-ul.22778522



Variable selection using a smooth information criterion for distributional regression models

Meadhbh O'Neill¹ · Kevin Burke¹

Received: 7 March 2022 / Accepted: 3 January 2023
© The Author(s) 2023

Abstract

Modern variable selection procedures make use of penalization methods to execute simultaneous model selection and estimation. A popular method is the least absolute shrinkage and selection operator, the use of which requires selecting the value of a tuning parameter. This parameter is typically tuned by minimizing the cross-validation error or Bayesian information criterion, but this can be computationally intensive as it involves fitting an array of different models and selecting the best one. In contrast with this standard approach, we have developed a procedure based on the so-called “smooth IC” (SIC) in which the tuning parameter is automatically selected in one step. We also extend this model selection procedure to the distributional regression framework, which is more flexible than classical regression modelling. Distributional regression, also known as multiparameter regression, introduces flexibility by taking account of the effect of covariates through multiple distributional parameters simultaneously, e.g., mean and variance. These models are useful in the context of normal linear regression when the process under study exhibits heteroscedastic behaviour. Reformulating the distributional regression estimation problem in terms of penalized likelihood enables us to take advantage of the close relationship between model selection criteria and penalization. Utilizing the SIC is computationally advantageous, as it obviates the issue of having to choose multiple tuning parameters.

Keywords Variable selection · Information criteria · Penalized maximum likelihood · Heteroscedasticity · Distributional regression · Multiparameter regression

1 Introduction

Enhancements in data collection technologies have highlighted the importance of modern variable selection techniques. Traditional methods, such as best subset selection, are suboptimal and are computationally expensive when the number of variables is high (Fan and Lv 2010). Modern approaches make use of penalization methods to execute simultaneous model selection and estimation. A popular method is the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), which comprises of an L_1 penalty but leads to biased estimates. In contrast, the adaptive LASSO (Zou 2006) has adaptive weights, which reduce

the bias present in the LASSO estimates. These methods have been developed primarily in the context of normal linear regression and have been extended to generalized linear models (GLMs) (Tibshirani 1996; Friedman et al. 2010) and Cox’s proportional hazard models for survival data (Tibshirani 1997). In these classical models, covariates enter through the location parameter (or the hazard scale in the Cox model case). A more modern and flexible approach is to include covariates in multiple distributional parameters, such as the location and dispersion, simultaneously; this approach is known as “distributional regression” (Stasinopoulos et al. 2018) and “multiparameter regression” (MPR) (Burke et al. 2020). The goal of this paper is to expand penalized regression to the flexible MPR setting using a novel differentiable L_0 penalty that does not require tuning parameter selection, which is especially appealing in the MPR setting where one would typically require multiple tuning parameters (one for each distributional parameter).

Originally, methods such as quadratic programming were used to solve these non-differentiable LASSO-type prob-

✉ Meadhbh O’Neill
meadhbh.oneill@ul.ie

Kevin Burke
kevin.burke@ul.ie

¹ Department of Mathematics and Statistics, University of Limerick, Limerick, Republic of Ireland

lems, but Efron et al. (2004) and Friedman et al. (2007), respectively, proposed the least angle regression (LARS) and co-ordinate descent algorithms—with the latter proving to be particularly fast for problems of this type. These are somewhat “non-standard” estimation procedures in the context of classical statistical estimation, where non-differentiable objective functions are relatively less common. As an alternative non-gradient based optimization, perturbing the penalty function slightly to render it differentiable (Hunter and Li 2005; Lloyd-Jones et al. 2018) enables standard optimization methods to be used. Oelker and Tutz (2017) outline a series of approximations of different penalties, which allows for penalized smooth functions. These differentiable penalties can be easily implemented and solved using standard gradient based optimization procedures, i.e., Newton-Raphson. The tuning parameter that controls the strength of the penalty is typically obtained by minimizing the cross-validation error or an information criterion (IC), such as the Akaike IC (AIC) (Akaike 1974) or Bayesian IC (BIC) (Schwarz 1978). This is a two-step estimation process, which tends to be computationally intensive as it involves fitting an array of different models and selecting the best one.

Su (2015) and Su et al. (2018) present an estimation procedure that is not based on the L_1 norm, titled “MIC” (minimum approximated information criterion). They exploit the close connection between model selection criteria and penalization (Fan and Lv 2010) and introduce an approximated information criteria in order to avoid the classic two-step estimation process. At its core, the MIC utilizes an approximation of the “ L_0 norm” with a continuous unit dent function. The L_0 norm is discrete in nature and it is preferable to have a penalty function with a level of smoothness for optimization purposes. Su (2015) describes a “subtle uprooting” method for variable selection, which involves using a smooth surrogate function for approximating cardinality. This is followed by a second technical step for enhancing sparsity, where the final problem becomes non-differentiable. This approach is extended to GLMs in Su et al. (2018). Fixing the tuning parameter at two for the AIC or $\log(n)$ for the BIC is computationally advantageous, as it avoids the tuning parameter selection problem. It is not required to compute the whole regularization path of solutions, nor is it necessary to choose the best tuning parameter using cross-validation, as is typically done.

We propose a more straightforward method of approximating the IC function using a smooth approximation of the L_0 norm, which can be optimized directly. Instead of performing the reparameterization step as outlined in Su (2015), which renders the problem non-differentiable, we achieve sparsity in a different way. Our approach squeezes the coefficient values to zero by optimizing a sequence of objective functions that get successively closer to the non-differentiable one. Consequently, our proposed “smooth IC”

(SIC) function can be optimized directly using standard gradient based optimization techniques. Additionally, we extend this new SIC variable selection procedure for use in the developing area of distributional regression (Stasinopoulos et al. 2018). Our proposed methods are implemented in our publicly available R package “smoothic” (O’Neill and Burke 2021). To date, penalized estimation has been primarily applied in the context of classical regression models, where the covariates are allowed to enter the model through a single parameter (e.g., a location parameter). Other distributional parameters, such as a dispersion parameter, are typically constant. This “single parameter regression” (SPR) does not take into account the possible impact of covariates on the other distributional parameters. Distributional regression, which is also referred to as “multiparameter regression” (MPR), is a more flexible approach where multiple parameters are regressed simultaneously on covariates. For example, covariates can enter the model through the location and dispersion parameters, or scale and shape parameters of the hazard function in the survival context (see Burke and MacKenzie 2017; Burke et al. 2020 and references therein), or indeed in various different distributional parameters as in generalized additive models for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos 2005). Mayr et al. (2012) address the problem of variable selection by utilizing classical gradient boosting techniques to fit GAMLSS models. More recently, Groll et al. (2019) suggest implementing a LASSO-type penalization in the GAMLSS framework. This regularization approach to GAMLSS is highly flexible, but it has the added complexity of separate tuning parameters for each regression component. Groll et al. (2019) state that carrying out the computationally demanding grid search for the optimal tuning parameters is a drawback of their method. In our proposed multiparameter regression with smooth IC (MPR-SIC) procedure, this issue is circumvented as the values of both tuning parameters are known in advance.

The model formulation, including the introduction of the “smooth L_0 norm”, the estimation procedure and the optimization algorithm are outlined in Sect. 2. In Sect. 3, the performance of our proposed methods is evaluated in both variable selection and parameter estimation through extensive simulation studies. We consider three real data analyses to demonstrate our proposed methods in Sect. 4. Finally, we close with some concluding remarks in Sect. 5.

2 Model formulation

2.1 Preliminaries

The classic normal linear regression is a single parameter problem that assumes there is constant variance in the errors.

The model is

$$y_i = x_i^T \beta + \varepsilon_i \tag{1}$$

for $i = 1, \dots, n$, where y_i is the response value and $x_i = (1, x_{1i}, \dots, x_{pi})^T$ is a vector of covariates for the i th individual over the predictor variables $j = 0, 1, \dots, p$. Here, $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ is the vector of regression coefficients for the location parameter and $\varepsilon_i \sim N(0, \sigma^2)$ under the homogeneity assumption. For the multiparameter regression (MPR) approach, where covariates appear in multiple distributional parameters simultaneously, the single parameter model in (1) is extended to include heterogeneity of error variance:

$$\text{Var}(\varepsilon_i) = \sigma_i^2 = e^{x_i^T \alpha}, \tag{2}$$

where the log-linear form ensures that σ_i^2 remains positive. The vector of regression coefficients for the dispersion parameter is $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_p)^T$. There may be different (possibly overlapping) sets of covariates impacting the location and dispersion, and although we use x_i for both, a given β or α coefficient may be set to zero, which removes the covariate from that model component. Because we apply penalized variable selection, the regression coefficients need to be on a similar scale, and therefore we assume that the predictors are scaled to have unit variance.

The log-likelihood function for the MPR normal model is

$$\begin{aligned} \ell(\theta) = & -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n x_i^T \alpha \\ & - \frac{1}{2} \sum_{i=1}^n e^{-x_i^T \alpha} (y_i - x_i^T \beta)^2, \end{aligned} \tag{3}$$

where $\theta = (\beta^T, \alpha^T)^T = (\beta_0, \dots, \beta_p, \alpha_0, \dots, \alpha_p)^T$. Our focus is on variable selection in the location and dispersion components, and we note that model selection criteria, such as the AIC and BIC, have a penalized functional form similar to regularization. In the distributional regression framework, an information criterion (IC) can be formulated as

$$\text{IC} = -2\ell(\theta) + \lambda [\|\tilde{\beta}\|_0 + \|\tilde{\alpha}\|_0 + 2], \tag{4}$$

where λ is fixed at $\lambda = 2$ or $\lambda = \log(n)$ for the AIC and BIC respectively, and $\tilde{\beta} = (\beta_1, \dots, \beta_p)^T$ and $\tilde{\alpha} = (\alpha_1, \dots, \alpha_p)^T$, i.e., the coefficient vectors with the intercepts omitted; there is an addition of two in the penalty to take into account the estimation of the intercept terms β_0 and α_0 . The L_0 norm, $\|\theta\|_0 = \text{card}(\theta) = \sum_{j=1}^p I(\theta_j \neq 0)$, indicates the cardinality or the number of non-zero elements in θ . This is not truly a norm since $\|c\theta\|_0 \neq c\|\theta\|_0$ when $c \neq 0, 1$.

The AIC is reported to be asymptotically “selection inconsistent” and “loss-efficient” as a variable selection criterion (Shao 1997; Yang 2005; Wang et al. 2009). As a result of its consistency property and superior empirical performance in variable selection, we employ a BIC-type criterion (Wang and Leng 2007) where $\lambda = \log(n)$.

Using the likelihood in (3) and arranging (4) as an IC-based penalized likelihood results in

$$\ell^{\text{IC}}(\theta) = \ell(\theta) - \frac{\log(n)}{2} [\|\tilde{\beta}\|_0 + \|\tilde{\alpha}\|_0 + 2]. \tag{5}$$

To enable gradient-based optimization, we define $\|\theta\|_{0,\epsilon} = \sum_{j=1}^p \phi_\epsilon(\theta_j)$ as the “smooth L_0 norm”, (See Sect. 2.2) and substitute the L_0 norm in (5) with $\|\tilde{\beta}\|_{0,\epsilon}$ and $\|\tilde{\alpha}\|_{0,\epsilon}$. This results in our proposed approach of MPR with smooth IC (MPR-SIC), which is the maximization of

$$\ell^{\text{SIC}}(\theta) = \ell(\theta) - \frac{\log(n)}{2} [\|\tilde{\beta}\|_{0,\epsilon} + \|\tilde{\alpha}\|_{0,\epsilon} + 2]. \tag{6}$$

Therefore, since BIC minimization is intrinsic to this formulation, it obviates the usual need for estimating the model at a range of tuning parameter grid points and then evaluating each of these using an external BIC in a second step. Avoiding this grid search is especially useful in the context of distributional regression. For the more commonly used L_1 norm, there is no direct link to the BIC, in which case one must search for the optimal tuning parameter. Moreover, one would typically have a separate tuning parameter for each distributional parameter to account for differing scales in these parameters, and this multidimensional grid search optimization is quite computationally intensive. In contrast, the BIC penalizes all parameters equally: it is $\log(n)$ for all non-zero parameters, irrespective of their size or distributional type (e.g., location or dispersion), and it is zero for zero parameters.

2.2 Smooth L_0 norm

Due to the non-differentiability of the L_0 norm, we propose a smooth function to approximate it:

$$\phi_\epsilon(x) = \frac{x^2}{x^2 + \epsilon^2}. \tag{7}$$

This is differentiable for $\epsilon > 0$ and $\lim_{\epsilon \rightarrow 0} \phi_\epsilon(x) = \|x\|_0$. Figure 1 demonstrates how $\phi_\epsilon(x)$ gets closer to $\|x\|_0$ as ϵ decreases. The smallest value shown ($\epsilon = 10^{-5}$) approximates the L_0 norm very closely, but it is also near the discontinuity at $x = 0$ making it unstable. Ultimately, (7) requires a small ϵ value to produce shrinkage, but we have found that simply fixing it to a small value from the offset

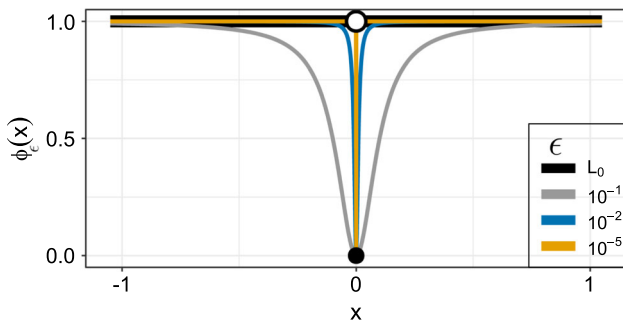


Fig. 1 Smooth L_0 norm

yields poor results due to its closeness to the discontinuity (see the Supplementary Material). Therefore, to create a more stable problem, we recommend the use of a decreasing sequence of ϵ values (described in Sect. 2.4). Interestingly, with a fixed “large” value of $\epsilon = 1$, this penalty has been referred to as a “weight elimination penalty” in the context of neural networks (Rumelhart et al. 1991). It is noteworthy that Oelker and Tutz (2017) develop a general penalized estimation procedure based on smooth approximations to penalties. Within this framework, they consider an L_0 -norm approximation that is slightly less straightforward than ours, since it is based on a logistic function with two smoothing parameters. Devriendt et al. (2021) provide an alternative estimation procedure to Oelker and Tutz (2017), which is exact rather than approximate, but which does not include the L_0 penalty (albeit they suggest adapting to a stochastic algorithm could potentially handle this). Crucially, however, both of these approaches require a grid search to find the optimal tuning parameter, but this is avoided in our work due to the connection to an information criterion established in Sect. 2.1.

Note that the first and second derivatives have a simple analytic form and therefore can be used within the gradient based optimization procedure of Sect. 2.3:

$$\phi'_\epsilon(x) = \frac{2x\epsilon^2}{(x^2 + \epsilon^2)^2}, \quad \phi''_\epsilon(x) = \frac{2\epsilon^2(\epsilon^2 - 3x^2)}{(x^2 + \epsilon^2)^3}. \quad (8)$$

2.3 Estimation procedure

We define the penalized estimates as

$$\hat{\theta} = \arg \max(\ell^{\text{SIC}}(\theta)),$$

where $\ell^{\text{SIC}}(\theta)$ is given by (6). The first and second derivatives with respect to the parameters are

$$\frac{\partial \ell^{\text{SIC}}}{\partial \beta} = \frac{\partial \ell}{\partial \beta} - \frac{\log(n)}{2} v_\beta = X^T z_\beta - \frac{\log(n)}{2} v_\beta,$$

$$\frac{\partial \ell^{\text{SIC}}}{\partial \alpha} = \frac{\partial \ell}{\partial \alpha} - \frac{\log(n)}{2} v_\alpha = X^T z_\alpha - \frac{\log(n)}{2} v_\alpha, \quad (9)$$

where X is an $n \times (p + 1)$ matrix, whose i th row is x_i , z_β and z_α are vectors of length n such that $z_{\beta,i} = e^{-x_i^T \alpha} (y_i - x_i^T \beta)$ and $z_{\alpha,i} = (e^{-x_i^T \alpha} (y_i - x_i^T \beta)^2 - 1)/2$, and v_β and v_α are vectors whose $(j + 1)$ th elements are given by $\phi'_\epsilon(\beta_j)$ and $\phi'_\epsilon(\alpha_j)$ respectively, except whose first elements are zero due to the fact that the intercepts are not penalized.

The matrix of negative second derivatives of $\ell^{\text{SIC}}(\theta)$, i.e., $-\nabla_\theta \nabla_\theta^T \ell^{\text{SIC}}(\theta)$ is given by

$$\begin{aligned} I(\theta) &= I_0(\theta) + \begin{pmatrix} \log(n) \Sigma_\beta / 2 & 0 \\ 0 & \log(n) \Sigma_\alpha / 2 \end{pmatrix} \\ &= \begin{pmatrix} X^T W_\beta X + \log(n) \Sigma_\beta / 2 & X^T W_{\alpha\beta} X \\ X^T W_{\alpha\beta} X & X^T W_\alpha X + \log(n) \Sigma_\alpha / 2 \end{pmatrix} \end{aligned}$$

where $I_0(\theta) = -\nabla_\theta \nabla_\theta^T \ell(\theta)$ is the observed information matrix of the unpenalized likelihood; Σ_β and Σ_α are diagonal matrices that appear due to the penalties and whose $(j + 1)$ th diagonal elements are given by $\phi''_\epsilon(\beta_j)$ and $\phi''_\epsilon(\alpha_j)$ respectively, except whose first diagonal elements are zero due to the fact that the intercepts are not penalized; W_β , W_α and $W_{\beta,\alpha}$ are $n \times n$ diagonal weight matrices whose i th diagonal elements are given by $e^{-x_i^T \alpha}$, $e^{-x_i^T \alpha} (y_i - x_i^T \beta)^2 / 2$ and $e^{-x_i^T \alpha} (y_i - x_i^T \beta)$ respectively. We employ the “RS” algorithm (Rigby and Stasinopoulos 2005), which does not use cross derivatives. This algorithm is motivated by the fact that in many classical models, including location and scale models, the parameters are information orthogonal as discussed in Cox and Reid (1987). However, Stasinopoulos and Rigby (2007) report that the RS algorithm works well even when the parameters are not information orthogonal.

The resulting system of Newton–Raphson equations can be expressed compactly as

$$\begin{pmatrix} X^T W_\beta^{(m)} X + \log(n) \Sigma_\beta^{(m)} / 2 & 0 \\ 0 & X^T W_\alpha^{(m)} X + \log(n) \Sigma_\alpha^{(m)} / 2 \end{pmatrix} \begin{pmatrix} \beta^{(m+1)} - \beta^{(m)} \\ \alpha^{(m+1)} - \alpha^{(m)} \end{pmatrix} = \begin{pmatrix} X^T z_\beta^{(m)} - \log(n) v_\beta^{(m)} / 2 \\ X^T z_\alpha^{(m)} - \log(n) v_\alpha^{(m)} / 2 \end{pmatrix}. \quad (10)$$

They are iteratively solved for $\theta^{(m+1)} = (\beta^{(m+1)T}, \alpha^{(m+1)T})^T$, where the elements super-scripted by (m) depend on θ^m , but this is excluded for notational convenience. Note that, since the RS algorithm sets the off-diagonal blocks to zero, it is possible to optimize the problem by alternating between the estimation of the mean and variance models; however, this is not considered here. We use the classical ordinary least squares estimates as initial values for the location parameter, i.e., $\beta^{(0)} = (X^T X)^{-1} X^T Y$,

where $Y = (y_1, \dots, y_n)^T$ is the response vector. We fix the starting value for the intercept of the dispersion term at $\log(s^2)$, where the classical residual variance estimator $s^2 = \sum_{i=1}^n (y_i - x_i^T \beta^{(0)})^2 / (n - p)$ is used. The remaining elements of the $\alpha^{(0)}$ parameter vector are set to zero (Rutemiller and Bowers 1968; Harvey 1976), which gives $\alpha^{(0)} = (\log(s^2), 0, \dots, 0)^T$. The standard errors of the estimates can be directly acquired by estimating the covariance of the penalized estimates for the true non-zero parameters using the sandwich formula,

$$\text{cov}(\hat{\theta}) = \{I(\hat{\theta})\}^{-1} I_0(\hat{\theta}) \{I(\hat{\theta})\}^{-1}, \tag{11}$$

which has been shown to be accurate for moderate sample sizes (Fan and Li 2001, 2002).

2.4 ϵ -telescoping

Although smaller values of ϵ lead to a better approximation of the L_0 norm (see Fig. 1), and hence IC optimization, we have found that the procedure becomes less numerically stable (see Supplementary Material). On the other hand, larger values of ϵ lead to a more stable optimization procedure, but one that does not yield coefficients close to zero. Therefore, we propose a method that “telescopes” through a decreasing sequence of ϵ values and makes use of “warm starts”, whereby the solution to the previous optimization problem is used as the initial point for the current nearby problem. The method can produce final estimates of the true zero coefficients that are extremely close to zero, and, so, can be treated as being equal to zero for practical purposes.

In this paper, we treat values below 10^{-8} as being zero. We have found that using a sequence of $T = 100$ steps from $\epsilon_1 = 10$ to $\epsilon_T = 10^{-5}$ performs well. Of course, applying fewer steps in the sequence from ϵ_1 to ϵ_T is an option in practice. However, larger values of T help to ensure that the repeated fitting procedure brings the parameters close to zero while avoiding estimation instability. If T is too small (e.g., $T = 10$), then the variable selection performance declines; simulation results for $T = 50$ and $T = 10$ are provided in the Supplementary Material. Once an adequate number of steps are used, the performance of the method is not highly influenced by the choice of the sequence. A large enough ϵ_1 must be chosen in order to introduce smoothness and give stable estimates, while a smaller ϵ_T more closely approximates the L_0 norm to induce pseudo-sparsity (where we say “pseudo” since the algorithm produces coefficients that can be made arbitrarily close to zero while not being exactly zero). In addition to this, we propose implementing an exponentially decaying sequence of the form of $\epsilon_1 r^{t-1}$, where ϵ_1 is the starting value, $r \in (0, 1)$ is the rate of decay and t is the step number. For our suggested sequence with $T = 100$ steps from $\epsilon_1 = 10$ to $\epsilon_T = 10^{-5}$, the decay param-

eter is $r = 0.87$. This is advantageous as the optimization begins with large increments from $\epsilon_1 = 10$, which provides rapid improvements and estimates that are initially close to the unpenalized values. The smaller increments leading to $\epsilon_T = 10^{-5}$ allow for smaller refinements, especially with regard to squeezing some coefficients to be close to zero.

Although we avoid a grid search over penalty tuning parameters (typically denoted by λ in penalized estimation), we instead have a sequence of ϵ values. However, there is a key distinction between the objectives of these two approaches. In tuning parameter selection, the grid search over λ is an optimization procedure, which, as previously discussed, is computationally demanding in the context of distributional regression due to it being a multidimensional grid. Moreover, the position of the optimal solution is unknown and could potentially be missed—especially if one reduces the number of grid points to combat the aforementioned computational expense. In contrast, our ϵ -telescoping approach is unidimensional and is not itself an optimization procedure since we know in advance, from a mathematical perspective, that ϵ should effectively be zero. Thus, the role of the ϵ -telescope is to move the problem to an arbitrarily small value of ϵ_T in a stable way. It should be noted that, although we use $\epsilon_T = 10^{-5}$, it may be that a relationship between ϵ_T and the sample size could be established using asymptotic analysis, e.g., a larger ϵ_T value might be acceptable at smaller sample sizes; however, this is beyond the scope of the current article.

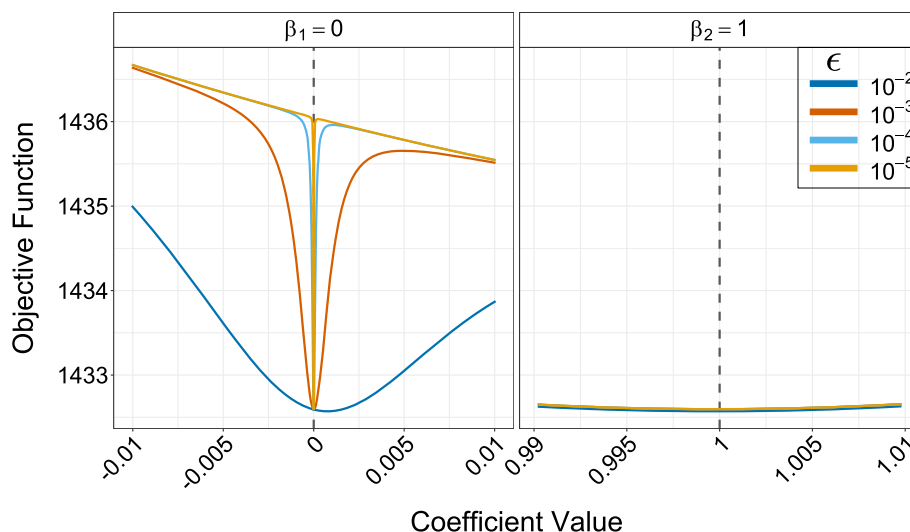
Table 1 presents an example of the coefficient estimates for a true zero and true non-zero coefficient for some simulated data. The shrinkage effect due to decreasing ϵ values is apparent. The value of the true zero coefficient β_1 drastically reduces in magnitude through each step. It is obvious that the final estimate at $\epsilon_T = 10^{-5}$ is extremely close to zero. As a result, it can be treated as a zero coefficient without any issues—and, indeed, could be shrunk further if desired by reducing ϵ_T . The estimate for the true non-zero coefficient β_2 does not vary greatly over the telescoping steps.

Figure 2 provides a visualization of the telescoping effect in terms of the objective function. This is a slice through the objective function, which is plotted as a function of the coefficient value. Different curves are plotted corresponding

Table 1 Coefficient values of β_1 and β_2 as the method telescopes through ϵ

ϵ	$\beta_1 = 0$	$\beta_2 = 1$
10^{-2}	0.0007631062	0.9998
10^{-3}	0.0000078397	0.9998
10^{-4}	0.0000000815	0.9998
10^{-5}	0.0000000008	0.9998

Fig. 2 Slice through objective function for different values of ϵ . Dashed vertical lines mark true value



to the ϵ values in the telescope sequence. In the case of the true zero coefficient β_1 , it is clear that as ϵ decreases, the width of the curves become narrower and therefore there is less uncertainty around the estimate. Additionally, it is evident that the minimum of the curve shifts towards zero as ϵ decreases. For the true non-zero coefficient β_2 , the curves for the different ϵ values are almost identical, i.e., the telescoping has little impact on the shape of the objective function in this case.

2.5 Algorithm

The proposed MPR-SIC variable selection method is summarized in Algorithm 1.

3 Simulation studies

3.1 Setup

We have undertaken a simulation study to investigate the performance of our proposed MPR-SIC method. We simulate data from the normal MPR model from Sect. 2, where X is a matrix of 12 covariates. To achieve a realistic setup, we make use of a range of different regression parameter values and covariate distributions. The regression coefficients, provided in Table 2, take values of 0, 0.5 or 1, respectively, corresponding to covariates having no effect, a weak effect or a strong effect. Moreover, there are several combinations where covariates enter through: both distributional parameters (X_1 to X_4); the location only (X_5 and X_6); the dispersion only (X_7 and X_8); and in neither the location nor the scale (X_9 to X_{12}), this latter group being pure noise covariates, i.e., they have no impact on the response. As for the covariate distributions, we include: two skewed covari-

Algorithm 1 Implementation of the MPR-SIC ϵ -telescope Method

1. **Initialization:** Set $\theta^{(0)} = (\beta^{(0)T}, \alpha^{(0)T})^T$, where $\beta^{(0)}$ and $\alpha^{(0)}$ are the initial values for the location and dispersion parameters respectively (see Sect. 2.3).
2. **Telescoping:** Go through the exponentially decaying sequence of telescope values of length T from ϵ_1 to ϵ_T , where $\epsilon_t = \epsilon_1 r^{t-1}$ for step $t = 1, \dots, T$ and $r \in (0, 1)$ is the rate of decay (see Sect. 2.4).
 - **For $t = 1, \dots, T$:**
 - Optimization:** Maximize $\ell^{\text{SIC}}(\theta)$ in (6) by iteratively resolving the system of equations in (10) with initial values $\theta_{\epsilon_t}^{(0)}$, to obtain $\hat{\theta}_{\epsilon_t}$. Convergence is achieved when $|\theta_{\epsilon_t}^{(m+1)} - \theta_{\epsilon_t}^{(m)}| \leq \omega$ for some small tolerance, e.g., $\omega = 10^{-8}$. For warm starts, set $\theta_{\epsilon_{t+1}}^{(0)} = \hat{\theta}_{\epsilon_t}$ so that the obtained estimates are used as initial values for the next step in the telescope. Note that we set $\hat{\theta}_{\epsilon_0} = \theta^{(0)}$.
3. **Output:** At $t = T$, the final estimates $\hat{\theta}_{\epsilon_T}$ are obtained and any estimates that are very close to zero (below 10^{-8} for example) can be treated as being zero. The corresponding standard errors are computed by evaluating (11) at $\hat{\theta}_{\epsilon_T}$. Note that because we are applying penalized variable selection, the predictors are scaled to have unit variance. However, the final estimates are converted back to their original scale.

ates, $(X_1, X_{11}) \sim \text{Exponential}(1)$; two unbalanced binary covariates $(X_3, X_{10}) \sim \text{Bernoulli}(0.75)$; four independent normal covariates $(X_4, X_5, X_7, X_8) \sim N(0, 1)$; and four correlated multivariate normal covariates $(X_2, X_6, X_9, X_{12}) = (Z_1, Z_2, Z_3, Z_4) \sim \text{MVN}$ wherein $\text{corr}(Z_j, Z_k) = 0.8^{|j-k|}$. Lastly, three different sample sizes ($n = 100, 500$ and 1000) are considered, where each scenario is replicated 1000 times. The Supplementary Material contains some additional simulation studies (not discussed here): scenarios where no covariate enters the dispersion (i.e., classical linear regression) and scenarios with only independent normally distributed covariates.

Table 2 True parameter values

	X_0	E X_1	M X_2	B X_3	N X_4	N X_5	M X_6	N X_7	N X_8	M X_9	B X_{10}	E X_{11}	M X_{12}
β	0	1	0.5	0.5	1	0.5	1	0	0	0	0	0	0
α	0	0.5	1	0.5	1	0	0	0.5	1	0	0	0	0

E = Exponential, B = Bernoulli, N = independent normal, M = multivariate normal (correlated)

For each scenario, we perform our proposed procedures, MPR-SIC and SPR-SIC; the latter is the SIC implemented for a single-parameter regression model, i.e., penalized linear regression. We compare the performance of our method to the “`bamlss`” package (Umlauf et al. 2018), a package which implements penalized distributional regression (among other things). More specifically, we assign LASSO penalties to the location and dispersion regression parameters of a normal distribution, where the associated tuning parameters are selected by minimizing the BIC using a two-dimensional grid search with 50×50 grid points; hereafter, we refer to this as BAMLSS. We note that the BAMLSS method does not always necessarily bring parameters very close to zero, and, therefore, our interpretation of a zero effect in BAMLSS is based on the associated 95% credible interval containing zero.

We also apply the adaptive LASSO (ALASSO) method from the “`glmnet`” package (Friedman et al. 2010), which corresponds to penalized linear regression, where we select the value of the tuning parameter by minimizing the BIC (ALASSO-IC). Note that only cross-validation-minimization is available in the `glmnet` package, and, therefore, we compute the BIC by evaluating the normal likelihood at the ALASSO estimates, $\hat{\beta}$, and with the variance estimator

$$\hat{\sigma}^2 = \sum_{i=1}^n \frac{(y_i - x_i^T \hat{\beta})^2}{n - k} \tag{12}$$

as suggested in Reid et al. (2016), where k is the number of non-zero elements in $\hat{\beta}$. Since `glmnet` does not provide parameter inference, we compute standard errors (and hence confidence intervals) using the general sandwich formula provided in (11). (The form of the matrix $I(\hat{\theta})$ is slightly different due to the presence of the L_1 penalty.) The ALASSO method is misspecified in the scenarios we consider here, since it does not cover dispersion effects. However, we include it as a very commonly used method in practice, and it is useful to see how variable selection in the location is impacted when one fails to model the dispersion. The SPR-SIC method is similarly misspecified and is included as an SIC-based alternative to the ALASSO. Note that the Supplementary Material includes scenarios (not discussed here) where these are not misspecified, i.e., the true model only contains location effects.

Table 3 Simulation results: average computation time per simulation replicate (in s)

n	MPR-SIC	BAMLSS	SPR-SIC	ALASSO-IC
100	6.8	384.8	0.8	0.1
500	4.1	165.4	1.0	0.1
1000	4.8	186.5	1.3	0.1

Intel(R) Core(TM) i7-10610U CPU @ 1.80GHz 2.30 GHz

3.2 Simulation results

Before we consider performance in terms of variable selection and parameter inference, we first briefly review the computational expense. To this end, average computation times for each of the methods are given in Table 3. We can see that our MPR-SIC procedure is 40–50 times faster than BAMLSS, in large part due to the two-dimensional grid required by the latter. Even though the SPR-SIC approach is misspecified here, it is still useful to note that it is 4–8 times faster than the MPR-SIC approach. Thus, as expected, the distributional MPR approach is slightly slower than the SPR approach due to the fact that the former specifies (correctly) a dispersion model, and, hence, has twice the number of parameters to estimate (ignoring intercepts). However, the difference is not as dramatic as MPR-SIC versus BAMLSS since the SIC approaches both have the same penalty with unidimensional ϵ -telescoping. The ALASSO-IC approach is the fastest overall, but it should be noted that the core of its implementation is compiled C code. Even so, the SPR-SIC is still relatively competitive computationally at approximately 10 times slower using only R code.

Turning now to variable selection performance, metrics including the average number of true zero coefficients correctly set to zero (C) and the average number of true non-zero coefficients incorrectly set to zero (IC) are investigated. The probability of choosing the true model (PT) is examined by looking at the proportion of times the true model is selected. The mean squared error (MSE) is computed for each simulation replicate in order to assess in-sample prediction accuracy, and is calculated by $MSE(\hat{\theta}) = (\hat{\theta} - \theta)^T X^T X (\hat{\theta} - \theta) / n$ (Tibshirani 1997). These metrics, averaged over simulation replicates, are presented in Table 4.

For all of the methods, the C values are close to six (true value) and improve as the sample size increases. For the MPR-SIC method, the IC values are zero in most cases,

Table 4 Simulation results: model selection metrics

	n	MPR-SIC				BAMLSS			
		C(6)	IC(0)	PT	MSE	C(6)	IC(0)	PT	MSE
β	100	5.25	0.15	0.44	0.14	5.55	0.24	0.52	0.17
	500	5.88	0.00	0.88	0.01	5.67	0.00	0.73	0.02
	1000	5.95	0.00	0.95	0.00	5.70	0.00	0.74	0.01
α	100	5.52	0.80	0.30	0.62	5.60	1.08	0.20	0.46
	500	5.92	0.00	0.93	0.04	5.35	0.00	0.67	0.08
	1000	5.95	0.00	0.95	0.02	5.08	0.00	0.63	0.06
	n	SPR-SIC				ALASSO-IC			
		C(6)	IC(0)	PT	MSE	C(6)	IC(0)	PT	MSE
β	100	5.59	3.20	0.00	2.37	5.43	2.99	0.01	2.02
	500	5.84	1.57	0.11	0.64	5.58	1.13	0.22	0.61
	1000	5.88	0.70	0.37	0.27	5.70	0.45	0.45	0.27
α	100	6.00	6.00	0.00	6.43	6.00	6.00	0.00	6.62
	500	6.00	6.00	0.00	7.12	6.00	6.00	0.00	7.16
	1000	6.00	6.00	0.00	7.15	6.00	6.00	0.00	7.17

C, average correct zeros; IC, average incorrect zeros; PT, the probability of choosing the true model; MSE, the average mean squared error

apart from $n = 100$. This is due to the three smaller-valued weak effects being set to zero incorrectly in some simulation replicates ($\beta_2, \beta_3, \beta_5$ in the location and $\alpha_1, \alpha_3, \alpha_7$ in the dispersion). This behaviour is also conveyed by the probability of choosing the true model (PT). The PT values are low for $n = 100$, which is due to both β and α sometimes having a zero coefficient not set to zero, and, sometimes having a non-zero coefficient incorrectly set to zero. The sample size of $n = 100$ is a challenging scenario as the MPR-SIC method is fitting a total of $2(p + 1)$ parameters, which in this case is 26 parameters for a relatively small sample size. Taking this into account, we suggest that the performance of the method in this setting is reasonable. The PT values are also high for $n = 500$ and 1000, and appear to be converging to one. The BAMLSS procedure has higher IC values than the MPR-SIC method for $n = 100$ and lower C values for the larger sample sizes. The net effect of this is that the PT values are generally lower than for the MPR-SIC method (except for the location parameter at $n = 100$). For the SPR-SIC and ALASSO-IC methods, it only makes sense to consider their performance in the location, since there is no dispersion model. We can see that the IC values for these approaches are quite large, which means that they are setting some of the non-zero parameters to zero (albeit this is reducing with the sample size). The corresponding PT values are also quite low compared to the MPR-SIC approach. Ultimately, this indicates that erroneously ignoring the dispersion has an impact on the estimation of the location, even though the location and dispersion parameters are orthogonal for the normal distribution.

The estimation and inferential performance of our proposed MPR-SIC method is investigated in Table 5. The average estimate over simulation replicates is shown along with the true standard error (SE), which is the standard deviation of the estimates over simulation replicates, and the average estimated standard error (SEE) over simulation replicates, where the SEE in a given replicate is computed using (11); also shown is the empirical coverage probability (CP) for a nominal 95% confidence interval. We can see that, in all cases, the estimated parameter is close to the true value, albeit there is some bias in the larger α coefficients at $n = 100$. The standard errors for both the β and α parameters are underestimated at $n = 100$, leading to CPs below 0.95. However, at $n \geq 500$ the standard errors are well estimated and the coverage is very close to the desired 0.95 level. The equivalent results for the BAMLSS, SPR-SIC and ALASSO-IC methods are deferred to the Supplementary Material, but we briefly outline them here: although BAMLSS appears to be better at the smallest sample size, the dispersion parameter results are not as good as MPR-SIC for the larger sample sizes (with higher SE values and CP values generally below 0.9); both the SPR-IC and ALASSO-IC methods perform poorly in the location parameter in all respects (biased estimates, large SEs that are underestimated by the SEEs, and quite low CP values).

Out-of-sample prediction coverage probabilities (PCPs) for the methods are calculated for a sample 20% the size of the original data and are shown in Table 6. This is calculated as the proportion of times the true response value lies in a nominal 95% prediction interval (PI) in each replicate. The average is then taken over the 1000 replicates. The 95% PIs are calculated as $x_i^T \hat{\beta} \pm 1.96 \sqrt{\exp(x_i^T \hat{\alpha})}$. Note that for the SPR-SIC and ALASSO-IC, the dispersion parameter is held constant across observations and so α is a vector of zeros except for its first element, which is the intercept. The methods appear to perform similarly when examining the overall PCP, although both the MPR-SIC and BAMLSS methods are somewhat poorer than the others in the $n = 100$ case. However, note that the observations can be categorized by their variability σ_i and, therefore, we split them into groups with low, medium and high variability using the thresholds $\sigma_i \leq 1$, $\sigma_i \in (1, 2.2]$ and $\sigma_i > 2.2$, respectively, where these thresholds are the tertiles computed numerically from the true underlying distribution of σ_i . Doing so reveals that none of the methods perform particularly well at $n = 100$ when viewed in terms of these three levels of variability. For $n = 500$ and 1000, the coverage for the MPR-SIC and BAMLSS procedures remain at approximately 95%, which is the desired nominal value. In contrast, for the SPR-SIC and ALASSO-IC methods, the PIs are too wide for the low and medium variability cases (leading to 100% coverage) and too narrow for the large variability cases (leading to

Table 5 Simulation results: estimation and inference metrics

	θ	MPR-SIC				$n = 500$				$n = 1000$			
		$\hat{\theta}$	SE	SEE	CP	$\hat{\theta}$	SE	SEE	CP	$\hat{\theta}$	SE	SEE	CP
β_0	0.0	-0.01	0.22	0.13	0.76	-0.00	0.06	0.06	0.93	-0.00	0.04	0.04	0.94
β_1	1.0	1.00	0.15	0.09	0.78	1.00	0.04	0.04	0.93	1.00	0.03	0.03	0.94
β_2	0.5	0.46	0.23	0.10	0.73	0.50	0.05	0.05	0.94	0.50	0.03	0.03	0.96
β_3	0.5	0.50	0.11	0.07	0.82	0.50	0.03	0.03	0.93	0.50	0.02	0.02	0.95
β_4	1.0	1.00	0.12	0.07	0.78	1.00	0.03	0.03	0.94	1.00	0.02	0.02	0.95
β_5	0.5	0.49	0.11	0.07	0.80	0.50	0.03	0.03	0.93	0.50	0.02	0.02	0.94
β_6	1.0	1.03	0.23	0.11	0.70	1.00	0.05	0.05	0.94	1.00	0.03	0.03	0.95
α_0	0.0	-0.13	0.45	0.23	0.66	-0.04	0.10	0.10	0.92	-0.02	0.07	0.07	0.96
α_1	0.5	0.45	0.29	0.12	0.70	0.50	0.07	0.07	0.95	0.50	0.05	0.05	0.96
α_2	1.0	1.10	0.38	0.17	0.73	1.01	0.07	0.07	0.93	1.01	0.05	0.05	0.92
α_3	0.5	0.49	0.37	0.13	0.58	0.51	0.08	0.08	0.95	0.50	0.05	0.05	0.94
α_4	1.0	1.12	0.24	0.17	0.82	1.01	0.07	0.07	0.95	1.01	0.05	0.05	0.94
α_7	0.5	0.51	0.30	0.13	0.69	0.51	0.07	0.07	0.93	0.50	0.04	0.05	0.96
α_8	1.0	1.11	0.24	0.17	0.81	1.02	0.07	0.07	0.93	1.01	0.05	0.05	0.94

SE, standard deviation of estimates over 1000 replications; SEE, average of estimated standard errors over 1000 replications; CP, the empirical coverage probability of a nominal 95% confidence interval

Table 6 Simulation results: out-of-sample prediction coverage probabilities

n	MPR-SIC			BAMLSS			SPR-SIC			ALASSO-IC		
	100	500	1000	100	500	1000	100	500	1000	100	500	1000
Low	0.77	0.93	0.94	0.82	0.93	0.95	1.00	1.00	1.00	1.00	1.00	1.00
Medium	0.90	0.94	0.95	0.91	0.95	0.95	0.98	1.00	1.00	0.99	1.00	1.00
High	0.95	0.95	0.95	0.94	0.94	0.94	0.79	0.85	0.86	0.81	0.85	0.86
Overall	0.86	0.94	0.95	0.88	0.94	0.95	0.93	0.95	0.95	0.94	0.95	0.95

Variability categorized as low ($\sigma_i \leq 1$), medium ($\sigma_i \in (1, 2.2]$) and high ($\sigma_i > 2.2$). Out-of-sample coverage is calculated for a sample 20% the size of the original data

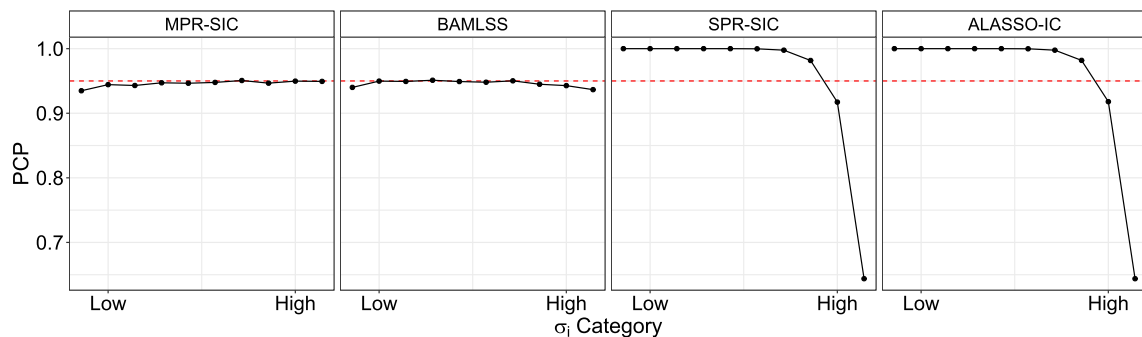


Fig. 3 Prediction coverage probabilities (PCPs) of observations for different dispersion levels, σ_i . Solid black line indicates the coverage and the red dashed line is a reference line at 0.95

approximately 85% coverage). This is unsurprising since these methods assume a constant σ , and, therefore, cannot adapt to heterogeneity in the data. This can also be visualized in Fig. 3, which shows the coverage for ten σ_i categories in the case where $n = 1000$; again we see that the MPR-SIC and BAMLSS methods lie close to 95%, whereas the other methods are either too high or too low.

In order to examine the generalizability of our simulation study, we have also considered several additional simulation scenarios, whose results can be found in the Supplementary Material. In particular, we have changed the effect sizes and cardinality of the active sets (i.e., the sets of covariates with non-zero effects) so they differ across the location and dispersion parameters; we also consider settings where we have doubled the number of covariates (to 24). In general,

the performance is comparable to the results presented here, but two noteworthy differences are as follows: (1) when the dispersion effects are much larger than the location effects, the selection performance in the location reduces considerably (albeit inferential performance remains good); and (2) when the number of covariates is increased to 24, the problem becomes unstable for $n = 100$ and the larger-sample PT values are reduced (by about 10–15 percentage points).

4 Real data analyses

4.1 Overview

We consider three real data analyses to illustrate our proposed MPR-SIC method, which is implemented using the ϵ -telescope (Algorithm 1). For each dataset, the resulting MPR-SIC, BAMLSS, SPR-SIC and ALASSO-IC estimates ($\hat{\beta}$, $\hat{\alpha}$) are presented, and note that, for the SPR-SIC and ALASSO-IC methods, $\hat{\alpha}$ is a vector of zeros except for its first element (the intercept). We also compare these methods in terms of out-of-sample PCP values. Additionally, for the proposed MPR-SIC, we provide the associated standard errors for each non-zero coefficient and the change in BIC, denoted ΔBIC , that arises upon setting that coefficient to zero. The ΔBIC value provides a measure of the impact of dropping a variable from the location (β coefficient) or the dispersion (α coefficient), and, therefore, indicates its importance in these model components. For the other methods, these additional metrics are deferred to Supplementary Material, but we indicate statistical significance by emboldening coefficient values for all methods in the main text. (Note that, for BAMLSS, “statistical significance” is based on the credible intervals excluding zero.)

4.2 Prostate cancer data

We examine the prostate cancer data, which come from a study by Stamey et al. (1989) and which appear in Tibshirani (1996) and Zou and Hastie (2005). The correlation between the level of prostate-specific antigen (PSA) and various clinical measures in 97 men who were about to receive a radical prostatectomy is examined. The predictors consist of eight clinical measures: log(cancer volume (cm^3)) (**lcavol**), log(prostate weight (g)) (**lweight**), presence of seminal vesicle invasion (SVI) (**svi**), age of the patient (**age**), log(amount of benign prostatic hyperplasia (cm^2)) (**lbph**), log(capsular penetration (cm)) (**lcp**), Gleason score (**gleason**) and percentage of Gleason scores four of five (**pgg45**). The logarithm of PSA (ng/mL) is the response variable. The presence of SVI (**svi**) is a binary variable (1 = yes, 0 = no) and **gleason** is a discrete numerical variable with four values. The Gleason score relates to

prostate cancer grades and the **pgg45** predictor provides information on the history of the patient. This is the percentage of Gleason scores they received before their final Gleason score in **gleason**. PSA is a protein that is produced by normal and malignant prostate cells, and is useful as a preoperative marker, as prostate cancer causes PSA to be discharged into the blood.

Figure 4 plots the standardized coefficient values with respect to the MPR-SIC ϵ -telescope, which shows how the method works as ϵ moves towards zero. We note that the coefficients are essentially unpenalized at the largest $\epsilon = 10^1$ value where there is no variable selection; this is because the penalty in (7) is close to zero for large ϵ values. Then, decreasing ϵ moves the problem towards L_0 penalization such that variable selection occurs. In particular, **lcavol** is selected only in the location component while **lweight** and **svi** are selected in both the location and dispersion components. Interestingly, although we decrease to $\epsilon = 10^{-5}$, the results here do not change appreciably beyond $\epsilon = 10^{-2}$.

From Table 7, we can see that, like the MPR-SIC method, the other three methods also select **lcavol**, **lweight**, and **svi** in the location component; in all cases, their location coefficients are positive. Thus, increased values of log(cancer volume) and log(prostate weight), and the presence of SVI are associated with increased log(PSA) values, and therefore may be indicative of prostate cancer. As for the dispersion component, while the MPR-SIC method selects both **lweight** and **svi**, BAMLSS only identifies **lweight** as being important. Both of these methods have similar BIC values (224 and 222 units respectively), which are lower than the models with only location regression components (227 and 228 units, respectively, for SPR-SIC and ALASSO-IC). Distributional regression approaches improve on classical single parameter regression approaches since they can capture more complex covariate effects, e.g., **lweight** and **svi** appear in both the location and dispersion within the MPR-SIC model. Given this additional complexity, it can be helpful to visualize the effects. To this end, inspired by Stadlmann and Kneib (2021), we provide a series of model-based (MPR-SIC) conditional density curves for different covariate combinations in Fig. 5.

As mentioned, within the MPR-SIC model (and all models considered), increased **lcavol** values are associated with increased log(PSA). Moreover, the large ΔBIC value of 40.39 identifies the **lcavol** location effect as being the most important (across all location and dispersion effects)—and there is a clear location shift in the associated conditional density plots in Fig. 5. Similarly, increased **lweight** values are also associated with increased log(PSA), but to a lesser extent than with **lcavol**. In line with this, there is more overlap between the conditional densities for high and low **lweight** values and the ΔBIC value is smaller (19.79). In other words, the cancer volume (**lcavol**) is

Fig. 4 Prostate Cancer Data: standardized coefficient values through the ϵ -telescope for the location and dispersion components. Lines are coloured corresponding to the selected variables, where grey indicates that the variable is set to zero and not selected in the final model. (Color figure online)

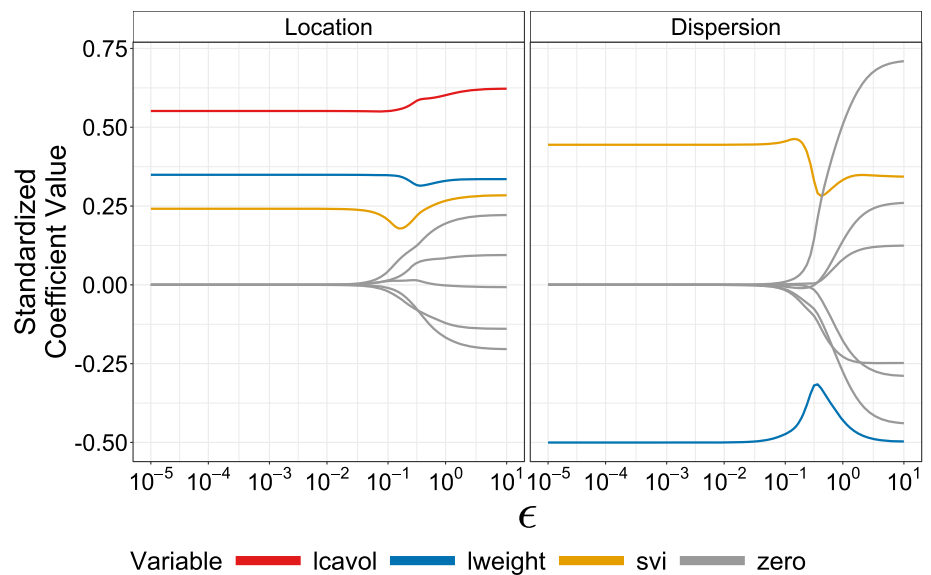


Table 7 Prostate Cancer Data: estimation metrics

BIC	MPR-SIC 224		BAMLSS 222		SPR-SIC 227		ALASSO-IC 228			
	$\hat{\beta}_j$	Δ BIC	$\hat{\alpha}_j$	Δ BIC	$\hat{\beta}_j$	$\hat{\alpha}_j$	$\hat{\beta}_j$	$\hat{\alpha}_j$		
inter	-1.26 (0.53)		3.15 (1.36)		-0.79	2.66	-0.78	-0.73	-0.27	-0.68
lcavol	0.47 (0.06)	40.39			0.52	-0.19	0.53		0.54	
lweight	0.82 (0.14)	19.79	-1.17 (0.38)	4.78	0.81	-0.93	0.66		0.52	
svi	0.58 (0.22)	1.64	1.07 (0.38)	4.09	0.73	0.77	0.67		0.53	
age					-0.01	0.02				
lbph					0.06	0.05				
lcp					-0.16	0.45				
gleason					0.02	-0.12				
pgg45					0.01	-0.01				

Significant effects indicated in bold

more strongly associated with PSA values than the prostate weight (*lweight*). However, the nature of the *lweight* effect differs from that of *lcavol* since it impacts the dispersion: increased *lweight* values are associated with reduced dispersion. As for *svi*, its presence primarily increases the dispersion. This can be seen both visually from the density plots and confirmed by the fact that removing *svi* from the dispersion increases the BIC 4.09, whereas, its removal from the location increases the BIC by just 1.64 units.

4.3 Sniffer data

When gasoline is pumped into a tank, hydrocarbon vapours are forced out and emitted into the atmosphere. This is a source of air pollution and in order to reduce this, devices that capture the vapour are set up. Testing these vapour recovery systems involves a “sniffer” device to measure the amount of vapour that is recovered. A method of estimating the total

amount released is required to estimate the efficiency of the system. A laboratory experiment was carried out to discover factors that impact the amount of hydrocarbon vapour released when gasoline is pumped into a tank. Four factors are varied—vapour pressure (psi) of the dispensed gasoline (*gaspres*), temperature (°F) of the dispensed gasoline (*gastemp*), initial tank temperature (°F) (*tanktemp*) and initial vapour pressure (psi) in the tank (*tankpres*). The quantity of emitted hydrocarbon (g) is the response variable. There are 125 runs in the data. These data have previously been considered by Weisberg (2013) who noted that the dispersion may depend on the predictors but did not apply a heteroscedastic model, and Bedrick (2000) who used a model with all four predictors in the location along with *gastemp* and *gaspres* in the dispersion.

The MPR-SIC, BAMLSS and SPR-SIC methods each select a different combination of variables for the location parameter (see Table 8). In terms of the selected statisti-

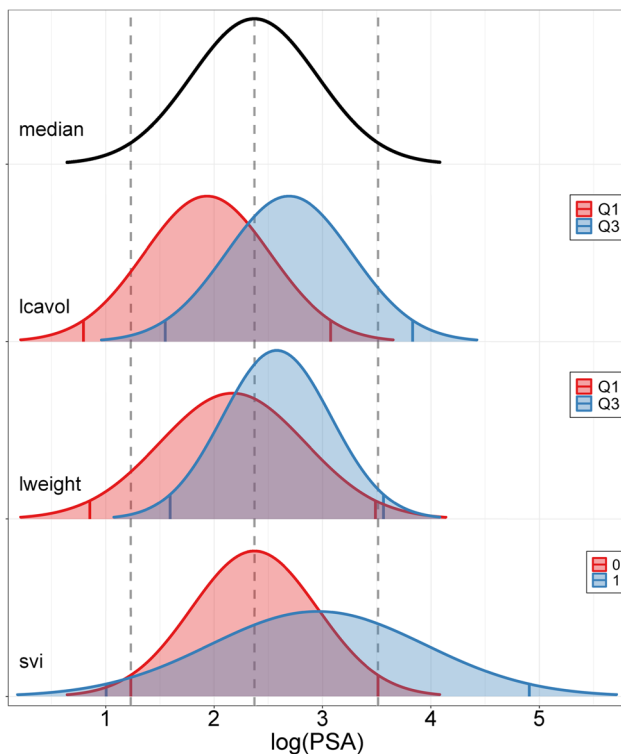


Fig. 5 Prostate Cancer Data: MPR-SIC model-based conditional density curves. The black curve corresponds to an individual whose covariates (*lcavol*, *lweight*, *svi*) are all equal to median values (serving as a “baseline” or “average” individual); dashed grey lines mark the 2.5th, 50th and 97.5th quantiles of this density. Keeping two of the covariates fixed at the median values, the red and blue densities correspond to the modification of the third covariate as: “low” (Q1, the first quartile) and “high” (Q3, the third quartile) for the continuous covariates, *lcavol* and *lweight*; and “absence” (= 0) and “presence” (= 1) for the binary covariate, *svi*; red and blue vertical lines mark the 2.5th and 97.5th quantile values of each density. (Color figure online)

cally significant effects, the two location regression models (SPR-SIC and ALASSO-IC) select *gaspres*, *gastemp*, and *tankpres*. However, these models have higher BIC values than the distributional regression models (MPR-SIC and BAMLSS), with the latter models choosing *tanktemp* rather than *tankpres* as being important in the location. In any case, the location effects of *gaspres* and *gastemp* are positive across all models (albeit *gaspres* is not statistically significant in BAMLSS), indicating that increased gasoline pressure and temperature values are related to increased amounts of emitted hydrocarbon; moreover, the MPR-SIC model identifies these as the most important effects with ΔBIC values of 68.41 and 56.85, respectively. The initial tank temperature (*tanktemp*) appears to be less important ($\Delta\text{BIC} = 5.60$), but its negative location coefficient in the MPR-SIC and BAMLSS models indicates that higher temperatures reduce the emitted hydrocarbon. In addition to the location effect of *gastemp*, the MPR-SIC model

also finds this variable to increase the dispersion. With a ΔBIC value of 17.62, the *gastemp* dispersion effect is far greater than the *tanktemp* location effect; this demonstrates the fact that modelling only the location—as is most often done in practice—can miss important features of the process under study. We note that the BAMLSS model is somewhat more complex than the MPR-SIC model, in that there are more coefficients that are far from zero. Overall, the MPR-SIC model achieves the lowest BIC of 616 units.

Conditional density plots display the various effects in Fig. 6a (and these are analogous to those shown in Fig. 5 for the prostate cancer data). Here we can clearly see: the large impact of *gaspres* on the location; the lesser impact of *gastemp* on the location and its impact on the dispersion; and the weak impact of *tanktemp* on the location. Moreover, since there are only three selected variables, and this is an industrial setup where each one can be altered in practice, Fig. 6b also displays all of the eight combinations of conditional densities that arise from varying each covariates at either “low” or “high” values. There is a clear optimal configuration here, which yields both the lowest and least variable hydrocarbon emissions: reduce the gasoline pressure (*gaspres*) and temperature (*gastemp*) and increase the initial tank temperature (*tanktemp*). This setting will yield emissions approximately between 20 and 30, with a mean of 25. In contrast, for the worst setting (high *gaspres* and *gastemp* and low *tanktemp*), emissions will generally be between 27 and 42, with a mean of 35. Thus, air pollution can be reduced considerably by optimizing the setup.

4.4 Boston house price data

Data from a cross-sectional study of 506 communities in the Boston area carried out in 1970 (Harrison and Rubinfeld 1978) are available in Wooldridge (2015). The association between median house prices in a particular community with various community characteristics is examined. There are eight explanatory variables: average number of rooms per house (*rooms*), percentage of the population that are “lower status” (*lowstat*), average student-teacher ratio of schools in the community (*stratio*), log(property tax per \$1000) (*lproptax*), log(weighted distances to five employment centres in the Boston region) (*ldist*), crimes committed per capita (*crime*), log(annual average nitrogen oxide concentration (*pphm*)) (*lnox*) and index of accessibility to radial highways (*radial*). The log(median house price (\$)) is the dependent variable. DiCiccio et al. (2019) applied a weighted least squares approach to these data (which accounts for heterogeneity but does not model the dispersion), where they only considered the *rooms*, *stratio*, *ldist* and *lnox* variables.

Table 9 shows that all eight covariates are included in the location component across the MPR-SIC, BAMLSS,

Table 8 Sniffer Data: estimation metrics

BIC	MPR-SIC 616				BAMLSS 624		SPR-SIC 630		ALASSO-IC 632	
	$\hat{\beta}_j$	ΔBIC	$\hat{\alpha}_j$	ΔBIC	$\hat{\beta}_j$	$\hat{\alpha}_j$	$\hat{\beta}_j$	$\hat{\alpha}_j$	$\hat{\beta}_j$	$\hat{\alpha}_j$
	inter	0.76 (0.85)		-1.35 (0.64)		-1.20	-0.96	0.45	2.01	0.21
gaspres	5.19 (0.51)	68.41			3.34	-3.46	10.84		9.79	
gastemp	0.23 (0.03)	56.85	0.06 (0.01)	17.62	0.26	0.09	0.15		0.19	
tanktemp	-0.09 (0.03)	5.60			-0.15	0.01			-0.07	
tankpres					2.69	2.72	-5.73		-4.08	

Significant effects indicated in bold

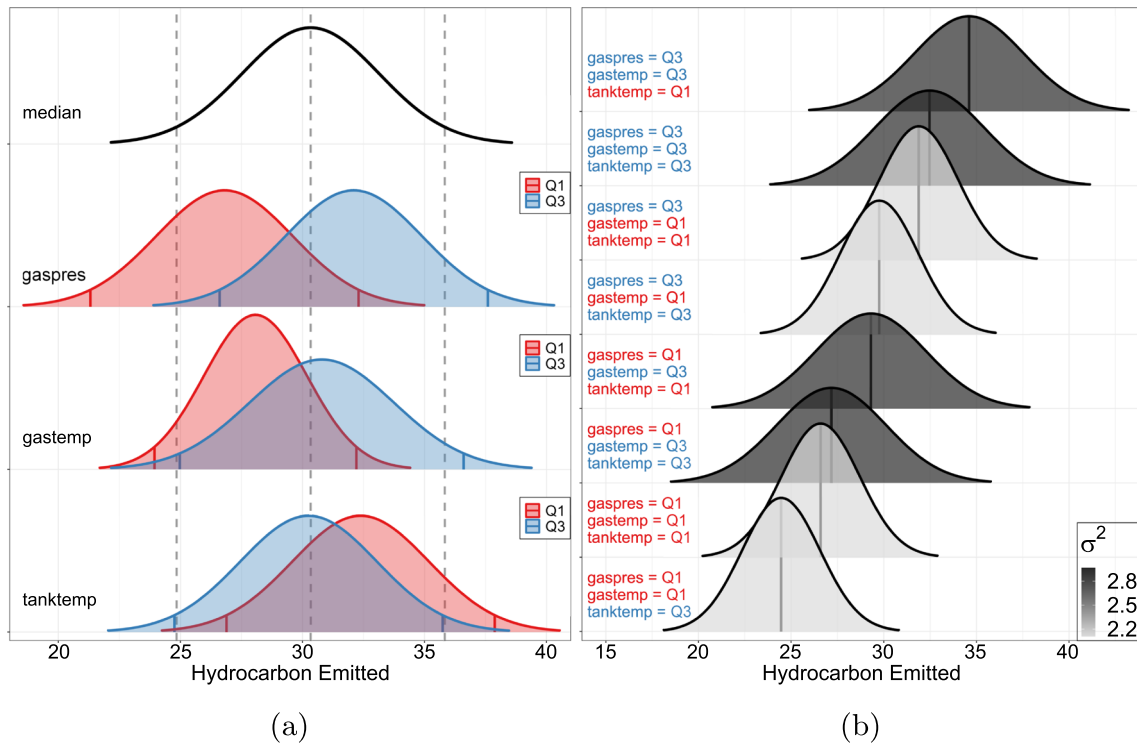


Fig. 6 Sniffer Data: **a** MPR-SIC model-based conditional density curves for each of the selected variables (see Fig. 5 for more details); **b** all eight conditional density curves obtained from the combinations of “low” (Q1, the first quartile) and “high” (Q3, the third quartile) for each of the three covariates *gaspres*, *gastemp*, and *tankpres*; these are ordered based on mean and coloured based on variance. (Color figure online)

SPR-SIC and ALASSO-IC methods. All the coefficients are statistically significant, and the signs of the estimated location coefficients are alike across the methods. Only two covariates (*rooms* and *radial*) have a positive effect on house prices. Houses with a greater number of rooms and access to radial highways are generally desirable, which results in increased house prices. The remaining variables may be considered to be undesirable, thus reducing the median house prices. In particular, the percentage of the population in the community that are “low status” and the student-teacher ratio of schools in the community both have a sizeable impact on the BIC when they are removed from the location parameter. The MPR-SIC method selects three covariates in the dispersion parameter, *lowstat*, *ldist*

and *radial*, of which *ldist* has a negative coefficient while the other two have positive coefficients. We note that *ldist*, when dropped from the dispersion, leads to a greater ΔBIC value than three of the eight variables in the location component, again highlighting that modelling only the location ignores important effects. Interestingly, the BAMLSS approach finds *rooms* rather than *ldist* to be statistically significant, but is otherwise quite comparable to the MPR-SIC approach in the values of the (statistically significant) model coefficients. Moreover, we can see that the BIC values for these two models are much lower than those of the location-regression models (SPR-SIC and ALASSO-IC).

It is useful to estimate the average sales price and the variability for a given community, and, therefore, for each of

Table 9 Boston House Price Data: estimation metrics

BIC	MPR-SIC -360				BAMLSS -337		SPR-SIC -169		ALASSO-IC -169	
	$\hat{\beta}_j$	ΔBIC	$\hat{\alpha}_j$	ΔBIC	$\hat{\beta}_j$	$\hat{\alpha}_j$	$\hat{\beta}_j$	$\hat{\alpha}_j$	$\hat{\beta}_j$	$\hat{\alpha}_j$
	inter	11.16 (0.28)		-3.53 (0.30)		10.51	-2.39	13.26	-3.30	13.18
rooms	0.24 (0.01)	178.27			0.26	-0.20	0.10		0.10	
lowstat	-0.02 (0.00)	95.03	0.03 (0.01)	5.55	-0.02	0.03	-0.03		-0.03	
stratio	-0.03 (0.00)	51.66			-0.02	-0.01	-0.04		-0.04	
lproptax	-0.20 (0.03)	39.81			-0.16	0.45	-0.26		-0.25	
ldist	-0.16 (0.02)	38.50	-0.92 (0.16)	26.80	-0.11	-1.21	-0.28		-0.27	
crime	-0.01 (0.00)	25.87			-0.01	-0.01	-0.01		-0.01	
lnox	-0.39 (0.08)	19.18			-0.28	-1.19	-0.62		-0.60	
radial	0.01 (0.00)	16.74	0.05 (0.01)	20.54	0.00	0.05	0.01		0.01	

Significant effects indicated in bold

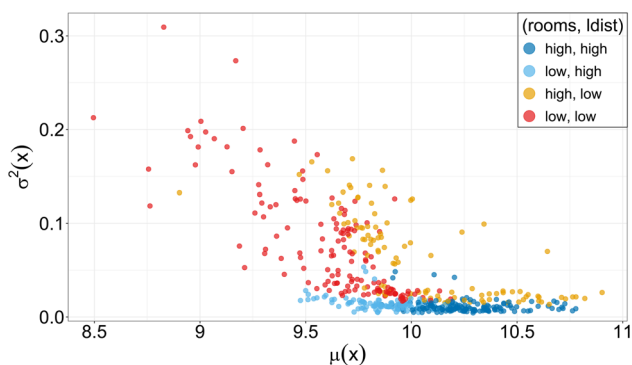


Fig. 7 Boston House Price Data: mean-variance pairs for each of the 506 communities. The colour of the points relate to whether a community has higher or lower values for `rooms` and `ldist` than the medians for each, e.g., the dark blue points correspond to communities where both the `rooms` and `ldist` values are higher than the median values. (Color figure online)

the 506 communities in the dataset, we compute the mean-variance pairs $(\mu(x_i), \sigma^2(x_i))^T$ where $\mu(x) = x^T\beta$ and $\sigma^2(x) = e^{x^T\alpha}$. These are displayed in Fig. 7 and note also that each point is equivalent to an underlying conditional density (as the normal distribution is fully characterized by its mean and variance). Interestingly, we see that communities with higher prices also tend to have lower variability in these prices. The points are coloured according to the `rooms` and `ldist` values, these being the most important location and dispersion variables, respectively. From this, we see that higher values of `rooms` are associated with increased prices, while higher values of `ldist` are associated with reduced variability. Thus, from the perspective of the real estate agent, desirable homes are those with a higher number of rooms located a greater distance away from employment centres. That being said, there are certainly other factors influencing house prices and their variability as previously discussed based on Table 9.

4.5 Prediction coverage probabilities

Table 10 contains the out-of-sample PCPs overall and split by category of variability calculated using 10-fold cross-validation for the prostate cancer, sniffer, and Boston house price data. Considering the PCPs from an overall point of view, the MPR-SIC, BAMLSS, SPR-SIC and ALASSO-IC methods perform similarly across all three data analyses. As expected based on the simulation, the coverage improves with respect to sample size, as we compare results from the smaller prostate cancer and sniffer datasets with the larger Boston house price dataset. The overall pattern is that both the SPR-SIC and ALASSO-IC methods tend to produce wider PIs for the low and medium variability categories, and narrower PIs for the high category—but do okay in terms of coverage for the low variability cases in the two smaller datasets (prostate cancer and sniffer data). The MPR-SIC and BAMLSS approaches are more balanced and tend towards good coverage with increasing sample size—whereas the other two methods continue to produce overly wide intervals for the low and medium variability categories and overly narrow intervals for the high category. This effect can be seen in Fig. 8 for the Boston house price data where the coverage is displayed for six σ_i categories; the coverage for both the MPR-SIC and BAMLSS methods lie close to 95%, while this is not the case for the other methods (that do not model the dispersion).

5 Discussion

Our proposed variable selection procedure uses a smooth L_0 norm to facilitate smooth information criterion optimization, and is extended for application in the developing area of distributional regression. This enables straightforward selection of more complex covariate effects afforded by

Table 10 Real data analyses results: out-of-sample prediction coverage probabilities

	MPR-SIC			BAMLSS			SPR-SIC			ALASSO-IC		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
Low	0.88	0.86	0.91	0.95	0.89	0.94	0.95	0.94	0.99	0.95	0.96	0.99
Medium	0.89	0.90	0.94	0.92	0.87	0.96	0.89	0.99	0.98	0.90	0.99	0.98
High	0.96	0.95	0.92	0.96	0.98	0.91	0.79	0.87	0.80	0.79	0.87	0.80
Overall	0.89	0.90	0.93	0.90	0.92	0.94	0.86	0.92	0.93	0.87	0.93	0.93

(a) Prostate cancer data, low: $\sigma_i \leq 0.6$, medium: $\sigma_i \in (0.6, 1.2]$, high: $\sigma_i > 1.2$
 (b) Sniffer data, low: $\sigma_i \leq 2.5$, medium: $\sigma_i \in (2.5, 2.9]$, high: $\sigma_i > 2.9$
 (c) Boston house price data, low: $\sigma_i \leq 0.1$, medium: $\sigma_i \in (0.1, 0.2]$, high: $\sigma_i > 0.2$

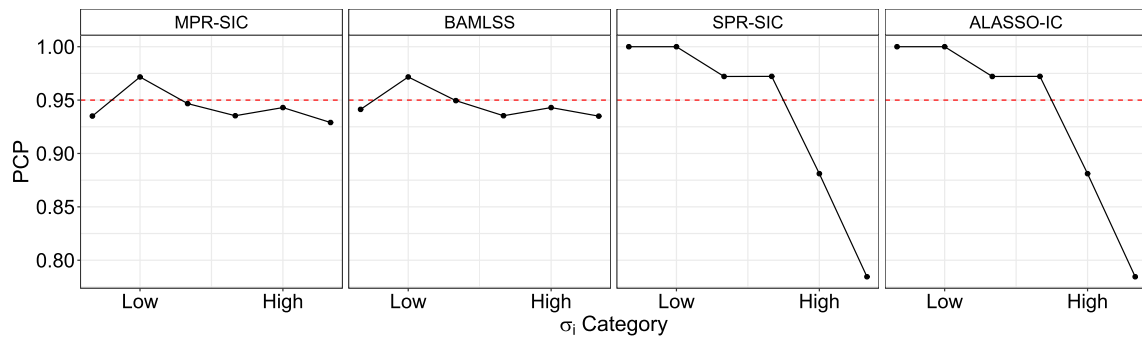


Fig. 8 Boston House Price Data: prediction coverage probabilities (PCPs) of observations for different dispersion levels, σ_i . Solid black line indicates the coverage and the red dashed line is a reference line at 0.95. (Color figure online)

modelling multiple distributional parameters simultaneously. The smooth objective function means that this is achieved using standard gradient based optimization procedures, i.e., Newton–Raphson. Moreover, because the objective function is an information criterion, the approach circumvents the need for penalty tuning parameter optimization, e.g., this is fixed at $\lambda = \log(n)$ in the BIC case. This is something that can be computationally intensive in LASSO-type problems, especially in the context of distributional regression modelling due to the additional parameters to be estimated and the fact that, in theory, there may be a separate tuning parameter for each distributional parameter. We provide a publicly available package `smoothic` for the implementation of our proposed methods (O’Neill and Burke 2021).

Through extensive simulation studies, we have demonstrated that the procedure has very favourable performance in terms of variable selection, parameter inference, and out-of-sample prediction; this is true in both single and multiparameter settings. Results from the real data analyses illustrate the effectiveness of our procedure, and in particular, the advantage of modelling the dispersion is clear from the fact that we have found dispersion effects that are stronger (in BIC terms) than location effects.

The methods proposed in this article are not restricted for use with only the normal distribution. We believe that the techniques presented herein can be extended for use with non-normal models, for example, the generalized linear model family. Moreover, we anticipate that the methods

will also extend to the setting of robust statistical modelling, which is particularly important given the ever-growing presence of complex datasets (Fan et al. 2014; Avella Medina and Ronchetti 2015; Maronna et al. 2019). Such an extension would be capable of dealing with heteroscedasticity and outliers, while also carrying out variable selection and parameter estimation simultaneously. Additionally, an anonymous reviewer has pointed out that, in combination with our proposal, a fused or group penalty would also be useful in practice for nominal or ordinal covariates. Such extensions will be a focus of our future work.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11222-023-10204-8>.

Acknowledgements This work was carried out within the Confirm Smart Manufacturing Research Centre (<https://confirm.ie/>) funded by Science Foundation Ireland (Grant No.: 16/RC/3918). We would also like to thank Prof. James Gleeson for his support.

Funding Open Access funding provided by the IReL Consortium

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your

intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akaike, H.: A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**(6), 716–723 (1974)
- Avella Medina, M., Ronchetti, E.: Robust statistics: a selective overview and new directions. *Wiley Interdiscip. Rev. Comput. Stat.* **7**(6), 372–393 (2015)
- Bedrick, E.J.: Checking for lack of fit in linear models with parametric variance functions. *Technometrics* **42**(3), 226–236 (2000)
- Burke, K., MacKenzie, G.: Multi-parameter regression survival modeling: an alternative to proportional hazards. *Biometrics* **73**(2), 678–686 (2017)
- Burke, K., Jones, M., Noufaily, A.: A flexible parametric modelling framework for survival analysis. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* **69**(2), 429–457 (2020)
- Cox, D.R., Reid, N.: Parameter orthogonality and approximate conditional inference. *J. R. Stat. Soc.: Ser. B (Methodol.)* **49**(1), 1–18 (1987)
- Devriendt, S., Antonio, K., Reynkens, T., et al.: Sparse regression with multi-type regularized feature modeling. *Insur.: Math. Econ.* **96**, 248–261 (2021)
- DiCiccio, C.J., Romano, J.P., Wolf, M.: Improving weighted least squares inference. *Econom. Stat.* **10**, 96–119 (2019)
- Efron, B., Hastie, T., Johnstone, I., et al.: Least angle regression. *Ann. Stat.* **32**(2), 407–499 (2004)
- Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**(456), 1348–1360 (2001)
- Fan, J., Li, R.: Variable selection for Cox’s proportional hazards model and frailty model. *Ann. Stat.* **30**, 74–99 (2002)
- Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **20**(1), 101 (2010)
- Fan, J., Fan, Y., Barut, E.: Adaptive robust variable selection. *Ann. Stat.* **42**(1), 324 (2014)
- Friedman, J., Hastie, T., Höfling, H., et al.: Pathwise coordinate optimization. *Ann. Appl. Stat.* **1**(2), 302–332 (2007)
- Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**(1), 1 (2010)
- Groll, A., Hambuckers, J., Kneib, T., et al.: LASSO-type penalization in the framework of generalized additive models for location, scale and shape. *Comput. Stat. Data Anal.* **140**, 59–73 (2019)
- Harrison, D., Jr., Rubinfeld, D.L.: Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* **5**(1), 81–102 (1978)
- Harvey, A.C.: Estimating regression models with multiplicative heteroscedasticity. *Econom. J. Econom. Soc.* **44**, 461–465 (1976)
- Hunter, D.R., Li, R.: Variable selection using MM algorithms. *Ann. Stat.* **33**(4), 1617 (2005)
- Lloyd-Jones, L.R., Nguyen, H.D., McLachlan, G.J.: A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Comput. Stat. Data Anal.* **119**, 19–38 (2018)
- Maronna, R.A., Martin, R.D., Yohai, V.J., et al.: Robust Statistics: Theory and Methods (with R). Wiley, New York (2019)
- Mayr, A., Fenske, N., Hofner, B., et al.: Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* **61**(3), 403–427 (2012)
- Oelker, M.R., Tutz, G.: A uniform framework for the combination of penalties in generalized structured models. *Adv. Data Anal. Classif.* **11**(1), 97–120 (2017)
- O’Neill, M., Burke, K.: smoothic: variable selection using a smooth information criterion. <https://CRAN.R-project.org/package=smoothic>, R package version 0.1.0 (2021)
- Reid, S., Tibshirani, R., Friedman, J.: A study of error variance estimation in lasso regression. *Stat. Sin.* **26**, 35–67 (2016)
- Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* **54**(3), 507–554 (2005)
- Rumelhart, D., Weigend, A., Huberman, B.: Generalization by weight-elimination with application to forecasting. In: *Advances in Neural Information Processing Systems*, vol. 3 (1991)
- Rutemiller, H.C., Bowers, D.A.: Estimation in a heteroscedastic regression model. *J. Am. Stat. Assoc.* **63**(322), 552–557 (1968)
- Schwarz, G., et al.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
- Shao, J.: An asymptotic theory for linear model selection. *Stat. Sin.* **7**, 221–242 (1997)
- Stadlmann, S., Kneib, T.: Interactively visualizing distributional regression models with distreg.vis. *Stat. Model.* **22**, 1471082X211007308 (2021)
- Stamey, T.A., Kabalin, J.N., McNeal, J.E., et al.: Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *J. Urol.* **141**(5), 1076–1083 (1989)
- Stasinopoulos, D.M., Rigby, R.A., et al.: Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* **23**(7), 1–46 (2007)
- Stasinopoulos, M.D., Rigby, R.A., Bastiani, F.D.: GAMLSS: a distributional regression approach. *Stat. Model.* **18**(3–4), 248–273 (2018)
- Su, X.: Variable selection via subtle uprooting. *J. Comput. Graph. Stat.* **24**(4), 1092–1113 (2015)
- Su, X., Fan, J., Levine, R.A., et al.: Sparse estimation of generalized linear models (GLM) via approximated information criteria. *Stat. Sin.* **28**(3), 1561–1581 (2018)
- Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **58**, 267–288 (1996)
- Tibshirani, R.: The lasso method for variable selection in the Cox model. *Stat. Med.* **16**(4), 385–395 (1997)
- Umlauf, N., Klein, N., Zeileis, A.: BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *J. Comput. Graph. Stat.* **27**(3), 612–627 (2018)
- Wang, H., Leng, C.: Unified LASSO estimation by least squares approximation. *J. Am. Stat. Assoc.* **102**(479), 1039–1048 (2007)
- Wang, H., Li, B., Leng, C.: Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **71**(3), 671–683 (2009)
- Weisberg, S.: *Applied Linear Regression*. Wiley, New York (2013)
- Wooldridge, J.M.: *Introductory econometrics: a modern approach*. Cengage Learning, Boston (2015)
- Yang, Y.: Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**(4), 937–950 (2005)
- Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429 (2006)
- Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **67**(2), 301–320 (2005)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.