10.24425/acs.2026.1540

Archives of Control Sciences
Volume 36(LXXII), 2026
No. 2, *pages* 317–343

The best linear sub-model selection with entropy based regularization

Frantisek VOLDRICH and Jaromir KUKAL

This paper presents a unified framework for linear sub-model selection based on the principle of entropy maximization. By maximizing entropy under suitable constraints, we derive generalized normal distributions that define both the error model and the prior on regression weights, linking distributional assumptions directly to regularization penalties. The resulting quasi-likelihood formulation integrates likelihood maximization with regularization and allows flexible control of robustness and sparsity through two shape parameters. We develop a convex optimization approach for parameter estimation and propose a heuristic binary optimization algorithm for efficient sub-model selection guided by Bayesian Information Criteria (BIC). The framework is experimentally evaluated on simulated regression tasks with both binary and continuous weights under varying noise conditions. The experiments systematically examine the influence of the distributional parameters and on the accuracy and stability of sub-model identification. The results show that for low and moderate noise levels, the proposed method reliably recovers the true sub-model and achieves information criterion values comparable to those of established regularization techniques such as ridge and lasso regression.

Key words: entropy maximization, regularization, information criteria, sub-model selection

1. Introduction

The selection of the most appropriate linear sub-model in statistical modeling is challenging, especially when dealing with high-dimensional data where naive least-squares leads to high variance and unstable estimates. A statistical interpretation of these ideas emerged with the introduction of ridge regression by Hoerl and Kennard [1], which brought the concept of penalty term into the

F. Voldrich (corresponding author, e-mail: voldrichfrantisek@gmail.com) and J. Kukal are with Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, Czech Republic.

This work has been supported by the grant SGS26/170/OHK4/3T/14 of Czech Technical University in Prague.

Received 06 November 2025.

context of statistical modeling. Ridge regression addressed multicollinearity in linear regression by adding an L_2 penalty term, shrinking coefficients toward zero to improve stability and reduce overfitting.

Subsequent progress focused on methods for choosing optimal model complexity. Mallows [2] proposed one of the first criteria to balance variance from overfitting against bias from underfitting. Information-theoretic approaches followed, with Akaike [3] introducing the Akaike Information Criterion (AIC) and Schwarz [4] developing the Bayesian Information Criterion (BIC), which established the principle of penalizing model complexity to maintain predictive accuracy. Generalized Cross Validation (GCV) proposed by Golub et al. [5] extended cross-validation to a computationally efficient, rotation-invariant estimator of prediction error, making it particularly suitable for smoothing and regularization problems.

A major milestone came with the introduction of the lasso, short for Least Absolute Shrinkage and Selection Operator, by Tibshirani [6], which employed an L_1 penalty to achieve both coefficient shrinkage and automatic variable selection by setting some coefficients exactly to zero. This innovation addressed a key limitation of ridge regression by introducing model sparsity. The elastic net proposed by Zou and Hastie [7] further combined the strengths of lasso and ridge regression by integrating L_1 and L_2 penalties, improving performance in settings with correlated predictors.

Later developments extended regularization to more complex structures, introducing group-based methods such as group lasso, fused lasso, and sparse group approaches, which exploit prior knowledge of feature relationships. These have proven valuable in high-dimensional domains such as neuroimaging and genomics [8–10].

Two complementary strands have characterized the most recent advances. The first is the Bayesian regularization framework, exemplified by the horseshoe prior of Carvalho et al. [11], which provides adaptive, probabilistic shrinkage robust to unknown sparsity patterns. Extensions of this approach, including the regularized horseshoe and related priors [12–14], achieve near-optimal shrinkage for irrelevant variables while preserving large effects through heavy-tailed behavior.

The second strand focuses on selection reliability and valid post-selection inference, addressing reproducibility and uncertainty after regularized estimation. Stability selection introduced by Meinshausen and Bühlmann [15] provides a framework for reproducible feature identification via subsampling-based selection frequencies. Related approaches, such as debiased estimators [16] and selective inference [17], enable valid statistical inference after model selection, thereby mitigating selection-induced bias that conventional procedures fail to correct [18, 19].

The goal of this paper is to develop a unified, entropy-based framework for linear sub-model selection. The proposed method integrates likelihood and log-likelihood with entropy maximization to identify appropriate distributional assumptions and penalty structures. By maximizing entropy under targeted constraints, the framework yields models that are simultaneously informative, sparse, and robust. Its effectiveness is demonstrated through a series of experiments.

This paper is organized as follows. Section 2 develops the theoretical foundation by deriving optimal probability distributions through entropy maximization under various constraints, establishing the connection between different constraint types and their corresponding regularization penalties. Section 3 presents the quasi-likelihood framework for linear regression, incorporating both data fitting and regularization through generalized normal distributions, and derives the optimization objective that balances model fit with parameter sparsity. Section 4 describes the application of information criteria (AIC and BIC) for sub-model comparison and selection. Section 5 outlines the optimization strategies, covering both convex optimization techniques for smooth cases and binary heuristic methods for discrete feature selection problems. Section 6 details the experimental part of the paper, describing the implementation of the code that demonstrates the framework's effectiveness across different noise levels and dimensional settings. It also presents two testing scenarios that illustrate the practical application and evaluation of the proposed methods.. Finally, Section 7 concludes with a summary of findings and contributions of the entropy-based regularization approach.

2. Entropy maximization for symmetric distributions

Entropy maximization is a fundamental concept in information theory, often used to derive probability distributions [20] that best represent the underlying uncertainty given a set of constraints. In this section, we explore entropy maximization for symmetric distributions under various conditions. By imposing different constraints – such as the mean absolute value, the mean quadratic value, and the maximal absolute value – we derive distributions that maximize entropy, thereby providing the most appropriate model consistent with the given data.

Let $\mathrm{f}\colon \mathbb{R} \to \mathbb{R}_0^+$ be a symmetric probability density function (PDF) of a random variable $X \in \mathbb{R}$ that is both normalized and exhibits even symmetry as

$$\begin{gathered} \int_{-\infty}^{\infty} \mathrm{f}(x)\,\mathrm{d}x = 1, \\ \forall\, x \in \mathbb{R}\colon\ \mathrm{f}(x) = \mathrm{f}(-x). \end{gathered} \tag{1}$$

The differential entropy [21], which is subject of maximization with respect to f(x), is defined as

$$\mathrm{H}(X) = -\int_{-\infty}^{\infty} \mathrm{f}(x)\ \ln \mathrm{f}(x)\,\mathrm{d}x.$$

We will find standardized and optimal solutions under various constraints.

2.1. Given mean absolute value

Let the expected absolute value of X be finite. The standardization is guaranteed by constraint

$$\mathrm{E}|X| = -\int_{-\infty}^{\infty} |x|\,\mathrm{f}(x)dx = 1. \tag{2}$$

Applying the method of Lagrange multipliers [22], we formulate a variational problem in which we maximize the differential entropy with respect to the unknown PDF and minimize with respect to the multipliers λ_0 and λ_1 to satisfy the constraints given by Eqs. (1) and (2):

$$\mathcal{L} = \frac{1}{2}\int_{-\infty}^{\infty} (-\mathrm{f}(x)\ \ln \mathrm{f}(x) - \lambda_0\,\mathrm{f}(x) - \lambda_1\,|x|\,\mathrm{f}(x))\ \mathrm{d}x = \max_{\mathrm{f}}\ \min_{\lambda_0,\lambda_1},$$

$$\text{and therefore} \int_{0}^{\infty} (-\ln \mathrm{f}(x) - \lambda_0 - \lambda_1\,|x|)\,\mathrm{f}(x)\,\mathrm{d}x = \max_{\mathrm{f}}\ \min_{\lambda_0,\lambda_1}.$$

The occurence of the absolute value is not necessary but reflects the symmetry of the final PDF. Denoting the integrand as F, we apply Euler condition [23]

$$\frac{\partial \mathrm{F}}{\partial \mathrm{f}(x)} = -\ln \mathrm{f}(x) - 1 - \lambda_0 - \lambda_1\,|x| = 0,$$
$$\ln \mathrm{f}(x) = -\lambda_0 - 1 - \lambda_1\,|x|.$$

Having $C = \exp(-\lambda_0 - 1)$ we obtain

$$\mathrm{f}(x) = C\ \exp(-\lambda_1\,|x|)$$

and applying conditions (1) and (2), we arrive at $\lambda_1 = 1$ and $C = 1/2$. Therefore

$$\mathrm{f}(x) = \frac{1}{2}\exp(-|x|),$$

which is two sided exponential distribution, also called Laplace distribution [24]. Corresponding negative log-likelihood [25] contribution of an individual point is

$$\psi(x) = -\ln \mathrm{f}(x) = \ln 2 + |x|.$$

This is a convex but non-smooth function.

2.2. Given mean quadratic value

Let the variance of X as mean quadratic value be constrained as

$$\mathrm{var}X = \mathrm{E}X^2 = \int_{-\infty}^{\infty} x^2\,\mathrm{f}(x)\,\mathrm{d}x = 1. \tag{3}$$

Using the same reasoning, we conclude that

$$\mathcal{L} = \int_0^{\infty} \left(-\ln \mathrm{f}(x) - \lambda_0 - \lambda_1 x^2\right) \mathrm{f}(x)\,\mathrm{d}x = \max_{\mathrm{f}} \min_{\lambda_0,\lambda_1},$$

and hence

$$\ln \mathrm{f}(x) = -\lambda_0 - 1 - \lambda_1 x^2,$$
$$\mathrm{f}(x) = C\,\exp\left(-\lambda_1 x^2\right),$$

and finally

$$\mathrm{f}(x) = \frac{1}{\sqrt{2\,\pi}} \exp\left(-\frac{x^2}{2}\right),$$

which is the standard Gaussian normal distribution [26]. Similarly, we calculate

$$\psi(x) = -\ln \mathrm{f}(x) = \frac{1}{2}\ln(2\,\pi) + \frac{x^2}{2},$$

which is both convex and smooth function.

2.3. Given maximal absolute value

Let the maximum value be constant and equal to one, i.e.,

$$\max |X| = 1.$$

Similarly to previous sections, we formulate the variational problem

$$\mathcal{L} = \int_0^1 (-\ln \mathrm{f}(x) - \lambda_0)\,\mathrm{f}(x)\,\mathrm{d}x = \max_{\mathrm{f}} \min_{\lambda_0}$$

and solve

$$\begin{aligned} \ln \mathrm{f}(x) &= -\lambda_0 - 1, \\ \mathrm{f}(x) &= C \cdot \mathrm{I}\left(|x| \leqslant 1\right), \\ \mathrm{f}(x) &= \frac{1}{2} \cdot \mathrm{I}\left(|x| \leqslant 1\right), \end{aligned}$$

where I denotes the indicator function and PDF corresponds to the uniform distribution [27]. The negative log-likelihood contribution is then given by

$$\psi(x) = \begin{cases} \ln 2 & \text{for } |x| \leqslant 1, \\ +\infty & \text{otherwise}, \end{cases}$$

which is a piecewise constant, discontinuous, unbounded function, but convex for $|x| < 1$.

2.4. General absolute moment

Let us now generalize previous findings. Let $p > 0$ and the constrain is

$$\mathrm{E}|X|^p = \int_{-\infty}^{\infty} |x|^p \, \mathrm{f}(x) \, \mathrm{d}x = 1.$$

Using the same reasoning as in previous sections, we can formulate the variational problem

$$\mathcal{L} = \int_{0}^{\infty} \left(-\ln \mathrm{f}(x) - \lambda_0 - \lambda_1 |x|^p\right) \mathrm{f}(x) \, \mathrm{d}x = \max_{\mathrm{f}} \min_{\lambda_0, \lambda_1},$$

and solve it

$$\begin{aligned} \ln \mathrm{f}(x) &= -\lambda_0 - 1, \\ \mathrm{f}(x) &= C \cdot \exp\left(-\lambda_1 |x|^p\right), \end{aligned}$$

and finally

$$\mathrm{f}(x) = \frac{p}{2\, p^{1/p} \cdot \Gamma\left(1/p\right)} \exp\left(-\frac{|x|^p}{p}\right),$$

which is the Generalized Normal (GN) distribution [28] with

$$\psi(x) = \ln\left(\frac{2\, p^{1/p} \cdot \Gamma\left(1/p\right)}{p}\right) + \frac{|x|^p}{p}. \tag{4}$$

Resulting function $\psi(x)$ is:

- Non-convex if $p < 1$,
- Convex, non-smooth if $p = 1$,
- Convex, smooth if $1 < p < +\infty$,
- Discontinuous $p \to +\infty$

and helps to design convex programming tasks for $p \geqslant 1$. Several cases are depicted in Figure 1.

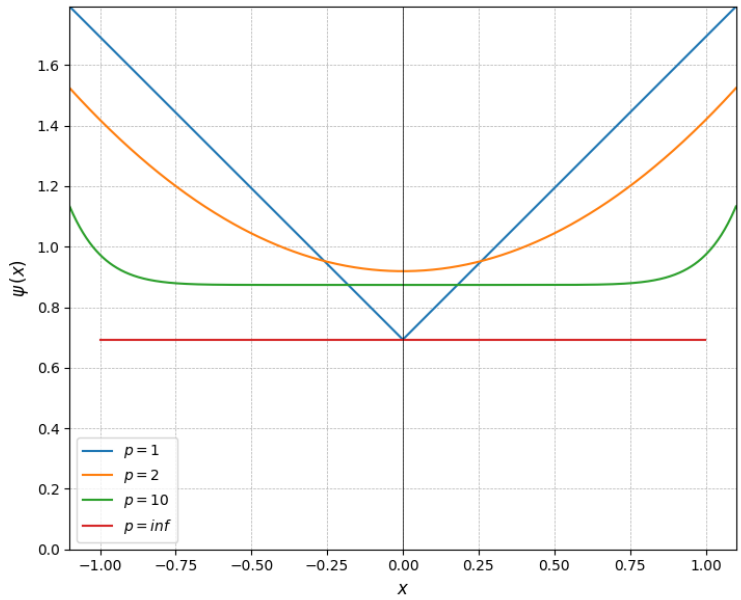

Figure 1: Behavior of $\psi(x)$ for various $p > 0$

3. Quasi-likelihood approach to linear regression model

3.1. Linear model with additive error

Let's consider a standard linear model [29] with observed variable

$$Y = b + \sum_{j=1}^{n} w_j x_j + s E,$$

where $n \in \mathbb{N}_0$ is the number of predictors, $s > 0$ is the noise scale, $E \sim \mathrm{f}$ is the error term following a distribution with f as PDF, $b \in \mathbb{R}$ is the bias, $\mathbf{w}, \mathbf{x} \in \mathbb{R}^n$ are the vector of weights and the vector of features, respectively. The measured data of size $m \in \mathbb{N}$ are collected in matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ and vector $\mathbf{y}^* \in \mathbb{R}^m$. Using

the method of likelihood maximization [30], we estimate the model parameters (weights, bias and noise scale) of likelihood [25]

$$\mathrm{L}\,(p, \mathbf{w}, b, s) = \prod_{i=1}^{m} \frac{\mathrm{f}\,(e_i/s)}{s} = \max_{\mathbf{w},b,s},$$

where

$$e_i = b + \sum_{j=1}^{n} w_j\, x_{ij} - y_i^*$$

and f is PDF of GN distribution with parameter $p \geqslant 1$. This task is equivalent to solving

$$\begin{aligned}\Psi\,(p, \mathbf{w}, b, s) &= -\ln \mathrm{L}\,(\mathbf{w}, b, s)\\ &= m \ln s - \sum_{i=1}^{m} \ln \mathrm{f}\left(\frac{e_i}{s}\right)\\ &= m \ln s + \sum_{i=1}^{m} \psi\left(\frac{e_i}{s}\right) = \min_{\mathbf{w},b,s}.\end{aligned}$$

3.2. Prior parameter distribution as regularization tool

Let's assume a prior symmetric independent distribution of weights W_j with some scaling $s_0 > 0$, i.e.,

$$\frac{W_j}{s_0} \sim g,$$

We define the quasi-likelihood [31] in a Bayesian sense [32] as the product of the likelihood function and the prior density contribution as

$$\mathrm{Q}\,(p, q, \mathbf{w}, b, s, s_0) = \mathrm{L}\,(p, \mathbf{w}, b, s) \cdot \prod_{j=1}^{n} \frac{\mathrm{g}\,(w_j/s_0)}{s_0} = \max_{\mathbf{w},b,s,s_0},$$

where g is PDF of GN distribution with parameter $q \geqslant 1$. Similarly, we define the negative log-quasi-likelihood as

$$\begin{aligned}\Psi^*\,(p, q, \mathbf{w}, b, s, s_0) &= -\ln \mathrm{Q}\,(p, q, \mathbf{w}, b, s, s_0)\\ &= m \ln s + \sum_{i=1}^{m} \psi\left(\frac{e_i}{s}\right) + n \ln s_0 + \sum_{j=1}^{n} \eta\left(\frac{w_j}{s_0}\right),\end{aligned}$$

where $\eta(x) = -\ln \mathrm{g}(x)$ is from the same family as $\psi(x)$. We will refer to the constant term as

$$\begin{aligned}\mathrm{A}(p) &= \ln \frac{2p^{1/p} \cdot \Gamma(1/p)}{p} \\ &= \ln 2 + \left(\frac{1}{p} - 1\right) \ln p + \ln \Gamma\left(\frac{1}{p}\right).\end{aligned}$$

We assume that the residuals follow the GN distribution with parameter $p \geqslant 1$. The prior distribution of the weights is also GN but with a potentially different parameter $q \geqslant 1$.

$$\begin{aligned}\Psi^*(p, q, \mathbf{w}, b, s, s_0) = m\,\mathrm{A}(p) + n\,\mathrm{A}(q) + m \ln s + n \ln s_0 \\ + \frac{1}{p s^p} \mathrm{S_E}(p, \mathbf{w}, b) + \frac{1}{q s_0^q} \mathrm{S_W}(q, \mathbf{w}),\end{aligned}$$

where

$$\mathrm{S_E}(p, \mathbf{w}, b) = \sum_{i=1}^{m} |e_i|^p, \tag{5}$$

$$\mathrm{S_W}(q, \mathbf{w}) = \sum_{j=1}^{n} |w_j|^q. \tag{6}$$

After the substitution

$$s_0 = \left(\frac{s}{\lambda}\right)^{p/q}$$

with $\lambda > 0$ and denoting

$$\mathrm{P}(p, q, \mathbf{w}, b, \lambda) = \mathrm{S_E}(p, \mathbf{w}, b) + \frac{p\,\lambda^p}{q} \mathrm{S_W}(q, \mathbf{w})$$

we obtain

$$\Psi^+(p, q, \mathbf{w}, b, s, \lambda) = m\,\mathrm{A}(p) + n\,\mathrm{A}(q) + \left(m + \frac{np}{q}\right) \ln s - \frac{np}{q} \ln \lambda + \frac{\mathrm{P}(p, q, \mathbf{w}, b, \lambda)}{p s^p}.$$

The optimal value of scaling s_{opt} is obtained from the condition $\dfrac{\partial \Psi^+}{\partial s} = 0$, yielding

$$s_{\mathrm{opt}} = \left(\frac{\mathrm{P}(p, q, \mathbf{w}, b, \lambda)}{m + np/q}\right)^{1/p}.$$

Therefore

$$\Psi^{+}_{\mathrm{opt}}(p,q,\mathbf{w},b,\lambda) = m\,\mathrm{A}(p) + n\,\mathrm{A}(q) - \frac{np}{q}\ln\lambda + \left(\frac{m}{p}+\frac{n}{q}\right)\left(1+\ln\left(\frac{\mathrm{P}(p,q,\mathbf{w},b,\lambda)}{m+np/q}\right)\right).$$

From this form, it is apparent that finding the minimum of Ψ^{+}_{opt} for fixed p, q, λ is equivalent to the minimization of convex function $P(p,q,\mathbf{w},b,\lambda)$ with respect to $\mathbf{w}, b$, as it is the only term which depends on the weights and bias. The optimal penalty is given as

$$\mathrm{P}_{\mathrm{opt}}(p,q,\lambda) = \min_{\mathbf{w},b} \mathrm{P}(p,q,\mathbf{w},b,\lambda)$$

and

$$\Psi_{\mathrm{opt}}(p,q,\lambda) = m\,\mathrm{A}(p) + n\,\mathrm{A}(q) - \frac{np}{q}\ln\lambda + \left(\frac{m}{p}+\frac{n}{q}\right)\left(1+\ln\left(\frac{\mathrm{P}_{\mathrm{opt}}(p,q,\lambda)}{m+np/q}\right)\right)$$

The objective function $\mathrm{P}(p,q,\mathbf{w},b,\lambda)$ and its parts can reach extremely high positive values, which is inacceptable for numeric optimization. Therefore, we prefer the minimization of

$$\mathrm{T}(p,q,\mathbf{w},b,\lambda) = \ln \mathrm{P}(p,q,\mathbf{w},b,\lambda)$$

to avoid the numeric overflow in particular cases. Having $u \geqslant v > 0$, we use identity

$$\ln(u+v) = \ln u + \ln\left(1+\frac{v}{u}\right) = \ln u + \ln\left(1+\exp(\ln v - \ln u)\right),$$

where $\ln v \leqslant \ln u$ and also $0 \leqslant \ln(1+\exp(\ln v - \ln u)) < \ln 2$. In our case

$$\ln u = \max(\ln r, \ln s),$$
$$\ln v = \min(\ln r, \ln s),$$

where

$$\ln r = \ln \mathrm{S_E}(p,\mathbf{w},b) = p\ln M_{\mathrm{E}} + \ln\left(\sum_{i=1}^{m}\left|\frac{e_i}{M_{\mathrm{E}}}\right|^{p}\right), \quad M_{\mathrm{E}} = \max_{i=1,\ldots,m}|e_i|$$

and

$$\ln s = \ln\left(\tfrac{p}{q}\,\lambda^{p}\cdot \mathrm{S_W}(q,\mathbf{w})\right) = \ln p - \ln q + p\,\ln\lambda + q\ln\mu + \ln\left(\sum_{j=1}^{m}\left|\frac{w_j}{M_{\mathrm{W}}}\right|^{q}\right),$$

$$M_{\mathrm{W}} = \max_{j=1,\ldots,m}\left|w_j\right|.$$

3.3. Quasi-likelihood estimation as convex optimization task

The minimized function $\mathrm{P}(p, q, \mathbf{w}, b, \lambda)$ is convex for $p, q \geqslant 1$, but smooth only for $p, q > 1$. Therefore, the regression comes to a free convex minimization for $p, q > 1$.

3.4. Role of regularization parameter

We will study only the case $\lambda > 0$, $p, q \in (1, \infty)$, when the function $\mathrm{P}(\lambda, p, q, \mathbf{w}, b)$ is smooth, strictly convex, and has a unique minimum for $\mathbf{w} \in \mathbb{R}^n$, $b \in \mathbb{R}$. Therefore, $\mathrm{P}_{\mathrm{opt}}(p, q, \lambda)$ is continuous function of $\lambda > 0$. When $\lambda \to 0^+$, the prior scale $s_0 \to +\infty$ and the regularization is omitted. We will study the limit

$$\begin{aligned}\lim_{\lambda\to 0^+}\left(\Psi_{\mathrm{opt}}(p,q,\lambda)+\frac{np\ln\lambda}{q}\right) &= m\,\mathrm{A}(p)+n\,\mathrm{A}(q)\\ &\quad+\left(\frac{m}{p}+\frac{n}{q}\right)\left(1+\lim_{\lambda\to 0^+}\ln\frac{\mathrm{P}_{\mathrm{opt}}(p,q,\lambda)}{m+np/q}\right)\\ &= m\mathrm{A}(p)+n\mathrm{A}(q)+\left(\frac{m}{p}+\frac{n}{q}\right)\left(1+\ln\frac{\mathrm{S}_{\mathrm{E}}^{\mathrm{opt}}(p,\mathbf{w},b)}{m+np/q}\right)\\ &= A_1\end{aligned}$$

and recognize the left asymptote

$$\psi_{\mathrm{left}} = A_1 - \frac{np\ln\lambda}{q}$$

in the $\ln\lambda$ vs. Ψ_{opt} diagram with slope $-n/q$ as seen in Figure 2.

When $\lambda \to +\infty$, the prior scale $s_0 \to 0^+$, and a constant model with $\mathbf{w} = \mathbf{0}$ is approached. In this case:

$$\lim_{\lambda\to+\infty}\left(\Psi_{\mathrm{opt}}(p,q,\lambda)+\frac{np\ln\lambda}{q}\right) = m\mathrm{A}(p)+n\mathrm{A}(q)+\left(\frac{m}{p}+\frac{m}{q}\right)\left(1+\ln\frac{\mathrm{R}(p,q)}{m+np/q}\right),$$

where

$$\begin{aligned}\mathrm{R}(p,q) &= \lim_{\lambda\to+\infty}\mathrm{P}_{\mathrm{opt}}(p,q,\lambda) = \lim_{\lambda\to+\infty}\min_{\mathbf{w},b}\left(\mathrm{S}_E(p,\mathbf{w},b)+\frac{p\,\lambda^p}{q}\,\mathrm{S}_W(q,\mathbf{w})\right)\\ &\leqslant \lim_{\lambda\to+\infty}\left(\mathrm{S}_E(p,\mathbf{0},b)+\frac{p\,\lambda^p}{q}\,\mathrm{S}_W(q,\mathbf{0})\right)\\ &= \min_b \mathrm{S}_E\,(p,\mathbf{0},b) = \mathrm{S}_E^{\mathrm{const}}(p),\end{aligned}$$

and also

$$\mathrm{R}(p,q) \geqslant \min_{\mathbf{w},b}\mathrm{S}_E\,(p,\mathbf{w},b) = \mathrm{S}_E^{\mathrm{opt}}(p) \leqslant \mathrm{S}_E^{\mathrm{const}}(p).$$

Threrefore, if the limit

$$A_2 = \lim_{\lambda \to +\infty} \left(\Psi_{\text{opt}} + \frac{np \ln \lambda}{q} \right)$$

exists, then

$$m\mathrm{A}(p) + n\mathrm{A}(q) + \left(\frac{m}{p} + \frac{n}{q} \right) \left(1 + \ln \frac{\mathrm{S}_E^{\text{opt}}(p)}{m + np/q} \right) \leqslant A_2 \leqslant m\mathrm{A}(p) + n\mathrm{A}(q) + \left(\frac{m}{p} + \frac{n}{q} \right) \left(1 + \ln \frac{\mathrm{S}_E^{\text{const}}(p)}{m + np/q} \right).$$

Corresponding right asymptote is

$$\psi_{\text{right}} = A_2 - \frac{np \ln \lambda}{q}$$

in the $\ln \lambda$ vs. Ψ_{opt} Figure 2. with identical slope as in the left case. The asymptotes are parallel and $A_2 \geqslant A_1$ because $\mathrm{S}_E^{\text{const}}(p) \geqslant \mathrm{S}_E^{\text{opt}}(p)$.

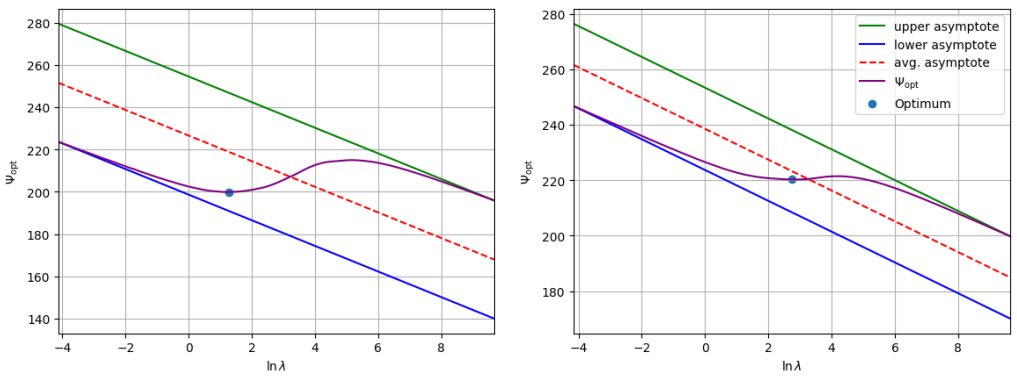

Figure 2: ψ_{opt} as a function of $\ln \lambda$, with asymptotes of slope $-n/q$ and the estimated local minimum, shown for $p = 1.01$, $q = 2$, and $n = 12$ in the wide-range plot (left), and for $n = 11$ in the narrow-range plot (right)

Thanks to the asymptotic behavior, the quasi-log-likelihood Ψ_{opt} reaches its supremum for $\lambda \to 0^+$ and its infimum for $\lambda \to +\infty$. Avoiding the fall into the infimum, we will find the optimal value $\lambda_{\text{opt}}(p, q)$ heuristically as the minimum of $\Psi_{\text{opt}}(p, q, \lambda)$ in the domain

$$\mathcal{D} = \left\{ \lambda \in \mathbb{R}^+ \colon \ \Psi_{\text{opt}}(p, q, \lambda) \leqslant \frac{A_1 + A_2}{2} - \frac{np \ln \lambda}{q} \right\}.$$

i.e. under the middle line between the asymptotes. Therefore, resulting

$$\lambda_{\text{opt}}(p,q) \in \arg\min_{\lambda\in\mathcal{D}} \Psi_{\text{opt}}(p,q,\lambda),$$

brings the final negative quasi-log-likelihood

$$\Psi^*_{\text{opt}}(p,q) = \Psi_{\text{opt}}\left(p,\ q, \lambda_{\text{opt}}(p,q)\right)$$

recommended in this study.

4. The best sub-model selection via information criteria

4.1. Akaike information criterion

The Akaike Information Criterion (AIC) [33] is based on information theory and aims to minimize the information loss when approximating the true data-generating process. AIC penalizes models with more parameters, thereby discouraging overfitting. Suppose that we have a statistical model of some data. Let Ψ^*_{opt} be the optimal value of negative likelihood function for the model and n the number of weights. AIC in this case is given by

$$\text{AIC} = 2(n+1) + 2\Psi^*_{\text{opt}}\,.$$

4.2. Bayesian information criterion

The Bayesian Information Criterion (BIC) [34] is similar to AIC but imposes a stronger penalty for models with more parameters. BIC is derived from a Bayesian perspective, estimating the likelihood of the model given the data. Suppose that we have a statistical model of some data. Let m be the number of data points (observations). The BIC in this case is given by

$$\text{BIC} = (n+1)\ln m + 2\Psi^*_{\text{opt}}\,.$$

4.3. Sub-model comparison

The main advantage of the information criteria (IC), such as AIC or BIC, is the absence of nested sub-model requirements. Having preferred IC, error and prior weights distributions, we fix $p, q \in (1,\infty)$ and design all 2^N sub-models [35] for comparison, where N is the number of parameters (features) in the full model. The parameter selection can be driven by binary vector $\mathbf{v} \in \{0,1\}^N$, where $v_k = 1$ indicates presence of the original feature x_k in the given sub-model. The number of estimated parameters is therefore $n = \sum_{k=1}^{N} v_k$ with $n = N$ indicating the full model, and $n = 0$ the constant sub-model. We evaluate adequate information

criterion (and find the best) for every sub-model, if 2^N is not too large, of course. Anyway, we can employ any binary optimization heuristic as stochastic tool for the minimum BIC or AIC finding. But we can also compare the sub-models with various values of exponents p, q, when we prefer several distributions of model errors and weight priors.

5. Optimization strategies

First, we will discuss various cases of finding $\mathbf{w} \in \mathbb{R}^n$, $b \in \mathbb{R}$ for fixed $p, q \in [1, +\infty) \cup \{+\infty\}$ and $\lambda > 0$.

5.1. Convex optimization techniques

Convex optimization techniques [36] are particularly well-suited to the problem of model selection because they guarantee that any solution found is globally optimal. We will apply them when $p, q \in (1, \infty)$, and minimize the convex and smooth function $\mathrm{P}(\lambda, p, q, \mathbf{w}, b)$ with respect to $(\mathbf{w}, b) \in \mathbb{R}^{n+1}$. There are a lot of gradient, conjugate gradient [37–40], Newtonian, quasi-Newtonian [41,42], and nondifferentiable methods, including corresponding free minimization solvers [36].

Convex optimization task of size $n \in N$ can be described as

$$\mathrm{f}(\mathbf{x}) = \min_{\mathbf{x} \in \mathbb{R}^n},$$

where $\mathbf{x} \in \mathbb{R}^n$ and $\mathrm{f}\colon \mathbb{R}^n \to \mathbb{R}$ is a convex function. But from the pragmatic point of view, we prefer solution simplicity for $p, q \in (1, +\infty)$ and use $p = q = 1 + \varepsilon$ instead of $p = q = 1$, and $p = q = \dfrac{1}{\varepsilon}$ instead of $p = q \to +\infty$, for reasonably small $\varepsilon > 0$.

5.2. Binary optimization heuristics

Having N features in the full model, the information criterion (IC as AIC or BIC), and binary selection vector $\mathbf{v} \in \{0, 1\}^N$ selects the best sub-model via solving the binary optimization task [43]

$$\mathrm{IC}(\mathbf{v}) = \min_{\mathbf{v} \in \mathcal{D}}, \quad \text{where } \mathcal{D} = \{0, 1\}^N.$$

There are many heuristic approaches [43] suitable for given tasks, but we focused on single point metaheuristics based on mutation operator [44] and its application in the steepest descent (SD) or random descent (RD) search.

Introducing Hamming norm in bipolar space $\{0, \pm 1\}^N$

$$\|\mathbf{v}\|_{\mathrm{H}} = \sum_{j=1}^{N} \mathrm{I}(v_j \neq 0) = \sum_{j=1}^{N} |v_j| = \|\mathbf{v}\|_1 ,$$

we will study the mutation process as generation of $\mathbf{v}_{\mathrm{new}} \in \mathcal{D}$ near $\mathbf{v} \in \mathcal{D}$ in deterministic or stochastic sense.

First two heuristics are frequently used in applied statistics, e.g. as step-wise regression [45].

General step-wise approach is a kind of SD:

- top-down approach begins with $\mathbf{v} = \mathbf{1}_N$ and generates $\mathbf{v}_{\mathrm{new}}$ systematically satisfying $\|\mathbf{v}_{\mathrm{new}} - \mathbf{v}\|_1 = 1$ & $\|\mathbf{v}_{\mathrm{new}}\|_1 < \|\mathbf{v}\|_1$ in every step.
- bottom-up approach begins with $\mathbf{v} = \mathbf{0}_N$, and generates $\mathbf{v}_{\mathrm{new}}$ systematically satisfying $\|\mathbf{v}_{\mathrm{new}} - \mathbf{v}\|_1 = 1$ & $\|\mathbf{v}_{\mathrm{new}}\|_1 > \|\mathbf{v}\|_1$ in every step.

The step-wise approach is deterministic, but does not guarantee reaching of optimal IC value.

It might be preferable to use stochastic approach with random initial point $\mathbf{v} \sim \mathrm{U}(\mathcal{D})$ and RD strategy with $\mathbf{v}_{\mathrm{new}} \sim \mathrm{U}(\mathcal{N})$, where $\mathcal{N} = \{\mathbf{x} \in \mathbb{D}\colon\ 0 < \|\mathbf{x} - \mathbf{v}\|_1 \leqslant k\}$ is the ring Hamming neighborhood of $\mathbf{v}$ and size $k \in \mathbb{N}$.

The size of $\mathcal{N}$ is therefore $\sum_{j=1}^{k} \binom{N}{j} = O(n^k)$ and $k = 3$ is preferred as Lin-3 [46] heuristic.

Empirically, the most effective heuristic built mutation probabilities from weight magnitudes: variables with larger absolute weights were less likely to flip, while smaller ones were more likely. This ensured at least one change per iteration, keeping the search stochastic yet guided by model structure and biased toward retaining influential features.

The random descent procedure iteratively searches the binary feature space by applying mutation operators to a current best mask and accepting only improving solutions with respect to its BIC value Starting from an initialized binary vector, the algorithm repeatedly generates candidate masks through stochastic mutations guided by model weights, fits a regression model under the selected features, and evaluates the result across a grid of regularization parameters λ to identify the optimal penalty.[1] If a candidate mask yields a lower BIC than the current best, it is accepted as the new reference point; otherwise, further mutations are attempted up to a configurable limit. Multiple restarts are supported to reduce dependence on initialization and avoid premature convergence. The process continues until stopping criteria are met.

[1]Note that the "optimal" penalty is obtained heuristically as the local minimum lying under average of upper and lower asymptotesa as depicted in Figure 2.

6. Experimental part

6.1. Implementation

The implementation of our optimization framework is structured into two main components: the regression model and the heuristic optimization. The regression model provides the mathematical foundation for fitting data, while the heuristic optimization performs a guided random descent in binary feature space to search for optimal submodels.

6.2. Regression model

The regression model is implemented in the class `GeneralFreeModel`, which supports generalized priors with user-defined hyperparameters p and q controlling the shape of data and weight penalties. The model uses numerical optimization methods (L-BFGS-B) to minimize the regularized loss.

6.3. Heuristic optimization

The heuristic search is encapsulated in the class `RandomDescentWithMutation`. It explores the space of binary feature masks by iteratively applying stochastic mutations and evaluating the resulting submodels. The procedure is governed by an `OptimizerConfig` object, which specifies the regression model parameters, mutation strategy, stopping criteria, and other runtime parameters.

The optimization starts from a random or user-specified mask. At each iteration, candidate masks are generated by a mutation operator. Although multiple operators were implemented, the experiments rely on a weighted probability mutation strategy, where each feature receives a flip probability inversely proportional to the magnitude of its fitted weight. This ensures that small-weight variables are more likely to be flipped, while large-weight variables are more likely to be retained. The probabilities are applied elementwise, with the guarantee that at least one feature changes per iteration. This guided stochastic search balances exploration with structural bias toward influential features.

Each candidate mask is evaluated by fitting the regression model and computing the BIC under the optimal λ. The search range for λ is scaled by the maximum eigenvalue of $\mathbf{X}_{\text{sub}}^{\top}\mathbf{X}_{\text{sub}}$, where $\mathbf{X}_{\text{sub}}$ represents the feature matrix after applying binary masks. The scaling also uses the coefficients specified in the configuration: `lmbda_search_coeff_min`, `lmbda_search_coeff_max`, and `lmbda_search_delta_step`.

If the candidate improves upon the current best solution, it is accepted as the new reference point; otherwise, new mutations are attempted up to a configurable limit. To avoid premature convergence, the framework supports multiple random restarts, enabling exploration from different regions of the search space.

6.4. Optimization result and logging

The results of the optimization are aggregated in the `OptimizationResult` class. It records both the trajectory of accepted solutions and the full history of all explored masks, including their BIC values and selected λ. The class also stores runtime statistics, iteration counts, and mutation coverage, allowing reproducibility and detailed post-hoc analysis. Logging functions provide fine-grained tracking of the optimization dynamics, such as the evolution of BIC, loss values, and feature selection patterns across iterations and restarts.

6.5. Scenario A: simulated bipolar weights

The feature matrix $\mathbf{X} \in \mathbb{R}^{m\times N}$ is generated by independent generator as

$$x_{i,j} \sim \mathrm{N}(0,1) \quad \text{for } i = 1,\dots,m;\ j = 1,\dots,N.$$

The weight vector $\mathbf{w} \in \{0,\pm 1\}^N$, which consists of $n = \|\mathbf{w}\|_1$ nonzero elements, is generated randomly using uniform distribution $\mathrm{U}(\mathcal{S})$ over any measurable set $\mathcal{S}$.

Let $\mathcal{P}$ be the set of all permutation vectors $\boldsymbol{\pi} \in \{1,\dots,N\}^N$.

We generate a random permutation vector $\boldsymbol{\pi} \sim \mathrm{U}(\mathcal{P})$ and set

$$w_j \sim \mathrm{U}(\{\pm 1\}) \quad \text{for all } j = 1,\dots,N \text{ satisfying } \boldsymbol{\pi}_j \leqslant n.$$

We set $w_j = 0$ for $\pi_j > n$. Bias is generated as $b \sim \mathrm{U}((-1,+1))$. Then, we evaluate observation vector $\mathbf{y}^* \in \mathbb{R}^m$ using formula

$$y_i^* = b + \sum_{j=1}^{N} x_{i,j} w_j + e_i\,,$$

where $e_i \sim \mathrm{N}(0,\sigma^2)$, with noise level $\sigma > 0$.

6.6. Scenario B: simulated positive weights

This approach is used in Python scikit-learn function `make_regression` and also produces $\mathbf{X} \in \mathbb{R}^{m\times N}$, $\boldsymbol{\pi} \in \mathcal{P}$, $\mathbf{w} \in \mathbb{R}^n$, $b \in \mathbb{R}$, and $\mathbf{y}^* \in \mathbb{R}^m$ [47].

Again $x_{i,j} \sim \mathrm{N}(0,1)$, $\boldsymbol{\pi} \sim \mathrm{U}(\mathcal{P})$, but $w_j \sim \mathrm{U}((-10,10))$ for all $j = 1,\dots,N$ satisfying $\pi_j \leqslant n$ and with $w_j = 0$ in the remaining cases. Weights with magnitude smaller than 0.1 were clipped to this value. The random bias $b \sim \mathrm{U}((-10,10))$, and again $y_i^* = b + \sum_{j=1}^{N} x_{i,j} w_j + e_i$, where $e_i \sim \mathrm{N}(0,\sigma^2)$. The main disadvantage of this procedure is in possible generation of false zero weights when $0 < \pi_j^* < \varepsilon$ for ϵ comparable with σ.

6.7. Results

The results are presented for different noise levels, chosen specifically to demonstrate the method's performance across various values of p and q. In all tables, N denotes the total number of features, n the number of informative features, and σ the corresponding noise level.

For bipolar weights, the most suitable values were found to be $p \in [1, 2]$ in combination with any $q \in [1, 100]$. For noise levels $\sigma \leqslant 1.5$, no submodel identification errors were observed, as shown in Table 1. When the noise level

Table 1: Submodel selection with bipolar weights for $N = 20$, $n = 5$, and $\sigma = 1.5$

		Number of false			Investigated
p	q	zeros	non-zeros	BIC	sub-models
1.01	1.01	0	0	413.57	51
1.01	1.50	0	0	412.00	60
1.01	2	0	0	411.01	41
1.01	4	0	0	408.96	57
1.01	100	0	0	407.20	65
1.50	1.01	0	0	407.43	68
1.50	1.50	0	0	405.82	39
1.50	2	0	0	404.82	46
1.50	4	0	0	402.92	36
1.50	100	0	0	400.95	47
2	1.01	0	0	405.17	36
2	1.50	0	0	403.56	36
2	2	0	0	402.57	77
2	4	0	0	400.74	77
2	100	0	0	398.83	54
4	1.01	0	0	409.71	65
4	1.50	0	14	192.02	239
4	2	0	0	407.03	48
4	4	0	0	405.36	48
4	100	0	0	404.24	67
100	1.01	0	4	415.34	470
100	1.50	1	11	297.68	239
100	2	0	15	184.97	260
100	4	0	14	87.98	368
100	100	0	6	409.97	242

increases to $\sigma = 2$, isolated errors appear in the localization of nonzero weights (Table 2). For higher noise levels $\sigma \geqslant 3$, the error rate in submodel determination becomes unacceptable, as demonstrated in Table 3.

Table 2: Submodel selection with bipolar weights for $N = 20$, $n = 5$, and $\sigma = 2$

		Number of false			Investigated
p	q	zeros	non-zeros	BIC	sub-models
1.01	1.01	0	0	471.00	32
1.01	1.50	0	0	469.50	72
1.01	2	0	0	468.51	53
1.01	4	0	0	466.67	63
1.01	100	0	0	464.98	63
1.50	1.01	1	0	464.08	273
1.50	1.50	1	0	462.66	245
1.50	2	0	0	462.37	49
1.50	4	0	0	460.67	92
1.50	100	0	0	458.81	56
2	1.01	1	0	461.38	251
2	1.50	1	0	459.96	251
2	2	1	0	459.13	251
2	4	1	0	457.64	251
2	100	0	0	456.77	54
4	1.01	0	0	467.24	43
4	1.50	0	15	198.35	121
4	2	1	0	463.76	253
4	4	1	0	462.36	253
4	100	0	0	462.55	50
100	1.01	1	3	472.21	309
100	1.50	0	10	322.66	448
100	2	0	11	240.84	268
100	4	0	15	98.76	136
100	100	1	4	468.02	238

Moreover, Table 1 shows that for low noise levels, only a relatively small number of submodels need to be explored by the binary optimization heuristic. Both lasso and ridge regression satisfy the imposed constraints on parameters p and q. To reduce computational effort, we recommend the compromise choice

Table 3: Submodel selection with bipolar weights for $N = 20$, $n = 5$, and $\sigma = 3$

		Number of false			Investigated
p	q	zeros	non-zeros	BIC	sub-models
1.01	1.01	3	0	544.75	273
1.01	1.50	3	0	543.94	243
1.01	2	3	0	543.47	294
1.01	4	3	0	542.65	282
1.01	100	1	0	543.73	449
1.50	1.01	3	9	539.95	702
1.50	1.50	3	0	536.51	286
1.50	2	3	0	536.05	332
1.50	4	3	0	535.19	271
1.50	100	1	0	535.56	258
2	1.01	3	0	534.62	271
2	1.50	3	0	533.79	276
2	2	3	0	533.32	276
2	4	3	0	532.42	275
2	100	1	0	533.21	259
4	1.01	3	0	538.94	251
4	1.50	0	13	279.82	231
4	2	2	15	270.00	257
4	4	3	1	536.55	311
4	100	2	0	537.54	256
100	1.01	2	5	550.79	367
100	1.50	1	10	391.49	253
100	2	2	12	309.46	259
100	4	0	15	167.47	112
100	100	2	4	543.55	235

$p = q = 1.5$ for low noise scenarios, where the best bipolar submodel was identified after evaluating only 39 instances of the BIC. With increasing noise levels, a higher value of $q = 100$ appears to yield fewer errors in identifying nonzero weights, while not increasing the number of explored submodels.

For random weights, the values $p, q \in [1, 2]$ proved most effective, as shown in Tables 4, 5 and 6, which illustrates the scenario for low noise levels $\sigma \leqslant 4$, following the same considerations discussed for bipolar weights.

Increasing noise levels further leads to a constant sub-model selection.

Table 4: Submodel selection with random weights with $N = 20$, $n = 5$ and $\sigma = 5$

		Number of false			Investigated
p	q	zeros	non-zeros	BIC	sub-models
1.01	1.01	0	0	674.48	44
1.01	1.50	0	0	674.79	23
1.01	2	0	0	672.32	23
1.01	4	0	0	746.03	23
1.01	100	3	0	856.68	248
1.50	1.01	0	0	672.39	37
1.50	1.50	0	0	670.77	37
1.50	2	0	0	669.82	37
1.50	4	0	0	683.54	36
1.50	100	3	0	831.93	245
2	1.01	0	0	672.52	77
2	1.50	0	0	669.59	77
2	2	0	0	668.66	77
2	4	0	0	667.21	77
2	100	3	0	833.72	249
4	1.01	0	0	677.97	56
4	1.50	0	15	495.75	197
4	2	0	0	676.65	56
4	4	0	0	675.14	54
4	100	3	0	793.23	241
100	1.01	0	3	695.18	543
100	1.50	0	12	580.98	679
100	2	1	0	693.51	464
100	4	0	13	434.23	191
100	100	1	0	689.70	270

7. Conclusions

This paper presents a unified entropy-based framework for linear sub-model selection that integrates likelihood maximization with regularization through generalized normal distributions controlled by two parameters. By maximizing entropy under various constraints, we derived a quasi-likelihood function that provides a probabilistic foundation for the proposed approach and naturally links distributional assumptions with regularization penalties. The framework was applied to linear models with the objective of estimating the parameters of the

Table 5: Submodel selection with random weights with $N = 20$, $n = 5$ and $\sigma = 7$

		Number of false			Investigated
p	q	zeros	non-zeros	BIC	sub-models
1.01	1.01	0	0	741.97	33
1.01	1.50	1	0	733.21	256
1.01	2	1	0	734.58	254
1.01	4	1	0	768.37	234
1.01	100	3	0	861.41	240
1.50	1.01	0	0	740.17	36
1.50	1.50	0	0	738.04	36
1.50	2	0	0	737.09	36
1.50	4	0	0	742.81	36
1.50	100	3	0	843.07	270
2	1.01	0	0	738.93	30
2	1.50	0	0	735.74	30
2	2	0	0	735.13	30
2	4	0	0	734.13	30
2	100	3	0	815.46	236
4	1.01	1	0	744	237
4	1.50	0	15	527.65	141
4	2	1	0	741.12	237
4	4	0	0	742.37	78
4	100	3	0	824.19	263
100	1.01	2	0	753.74	305
100	1.50	2	12	641.70	640
100	4	0	15	387.17	93
100	100	1	2	753.07	295

best-performing sub-model, and it enables flexible modeling of both data noise and parameter priors through the parameters p and q, while maintaining convex optimization guarantees for $p, q > 1$.

A novel optimization strategy for the regularization parameter λ was proposed, differing substantially from traditional selection methods such as ridge regression, lasso or Mallows' Cp statistic, and providing interpretable and consistent results. The selection of the optimal sub-model was guided by the Bayesian Information Criterion (BIC) in combination with a binary optimization heuristic based on weighted probability mutation, which proved effective for exploring the discrete

Table 6: Submodel selection with random weights with $N = 20$, $n = 5$ and $\sigma = 8$

		Number of false			Investigated
p	q	zeros	non-zeros	BIC	sub-models
1.01	1.01	1	0	760.22	251
1.01	1.50	1	0	760.51	251
1.01	2	1	0	759.46	251
1.01	4	1	0	787.65	224
1.01	100	3	0	860.21	251
1.50	1.01	1	0	757.97	258
1.50	1.50	1	0	757.84	253
1.50	2	1	0	756.86	253
1.50	4	1	0	759.96	253
1.50	100	3	0	848.26	271
2	1.01	1	0	759.05	224
2	1.50	1	0	757.97	224
2	2	1	0	757	224
2	4	1	0	755.25	224
2	100	3	0	845.44	253
4	1.01	1	0	771.58	256
4	1.50	0	13	544.51	203
4	2	1	0	767.29	404
4	4	1	0	767	256
4	100	3	0	823.32	263
100	1.01	2	1	778.15	262
100	1.50	1	10	640.29	252
100	2	0	15	500.81	241
100	4	0	15	403.51	167
100	100	2	3	782.01	257

feature space when exhaustive enumeration was computationally infeasible. All components of the proposed framework were implemented in Python.

The experimental results confirm that the choice of parameters p and q has a substantial impact on the sub-model selection performance, and not all values are equally suitable. In the low-noise regime ($\sigma \leqslant 2$), the parameters have little effect on the optimization outcome, while in the high-noise regime the optimization fails to yield meaningful results. Within the intermediate range $\sigma \in [1, 2]$, the approach performs reliably, and the results indicate that setting $p, q \in [1, 2]$,

particularly $p = q = 1.5$, is generally recommended configuration. In this setting, the proposed method achieves performance comparable to established techniques such as ridge and lasso regression.

Overall, the entropy-based regularization framework contributes to the statistical modeling literature by providing a theoretical justification of regularization penalties through maximum entropy principles, offering a computationally efficient approach to parameter selection via convex optimization, and demonstrating practical effectiveness across different experimental scenarios. Future work could explore extensions to non-linear models and investigate the theoretical properties of the proposed binary optimization heuristics in high-dimensional settings.

Declarations

Code availability

The code is available from the corresponding author upon reasonable request.

Author contributions

J.K. provided the overall conceptual framework and derived selected parts of the theory. F.V. derived other theoretical components and implemented the software. Both authors reviewed and approved the manuscript.

References

[1] A.E. Hoerl and R.W. Kennard: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), (1970), 55–67.

[2] C.L. Mallows: Some comments on cp. *Technometrics*, **42**(1), (2000), 87–94.

[3] H. Akaike: A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), (1974), 716–723. DOI: 10.1109/TAC.1974.1100705

[4] G. Schwarz: Estimating the dimension of a model. *The annals of statistics*, 461–464, 1978.

[5] G.H. Golub, M. Heath, and G. Wahba: Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**(2), (1979), 215–223.

[6] R. Tibshirani: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**(1), (1996), 267–288.

[7] H. Zou and T. Hastie: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **67**(2), (2005), 301–320.

[8] J.C. Beer, H.J. Aizenstein, S.J. Anderson, et al.: Incorporating prior information with fused sparse group lasso: Application to prediction of clinical measures from neuroimages. *Biometrics*, **75**(4), (2019), 1299–1309. DOI: 10.1111/biom.13075

[9] D. Kong, R. Fujimaki, J. Liu, et al.: Exclusive feature learning on arbitrary structures via $\ell_{1,2}$-norm. *Advances in Neural Information Processing Systems*, **27** (2014).

[10] F. Bach, R. Jenatton, J. Mairal, et al.: Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, **4**(1), (2012), 1–106.

[11] C.M. Carvalho, N.G. Polson, and J.G. Scott: Handling sparsity via the horseshoe. In: *Artificial intelligence and statistics*, PMLR, pp. 73–80, 2009.

[12] C.M. Carvalho, N.G. Polson, and J.G. Scott: The horseshoe estimator for sparse signals. *Biometrika*, (2010), 465–480.

[13] J. Piironen and A. Vehtari: *Sparsity information and regularization in the horseshoe and other shrinkage priors*, 2017.

[14] A. Bhadra, J. Datta, N.G. Polson, et al.: Lasso meets horseshoe. *Statistical Science*, **34**(3), (2019), 405–427.

[15] N. Meinshausen and P. Bühlmann: Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **72**(4), (2010), 417–473.

[16] S. Van de Geer, P. Bühlmann, Y. Ritov, et al.: *On asymptotically optimal confidence regions and tests for high-dimensional models*, 2014.

[17] J.D. Lee, D.L. Sun, Y. Sun, et al.: *Exact post-selection inference, with application to the lasso*, 2016.

[18] R.J. Tibshirani, J. Taylor, R. Lockhart, et al.: Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, **111**(514), (2016), 600–620.

[19] D. Zhang, A. Khalili, and M. Asgharian: Post-model-selection inference in linear regression models: An integrated review. *Statistic Surveys*, **16**, (2022), 86–136.

[20] N. Wu: *The maximum entropy method*, vol 32. Springer Science & Business Media, 2012.

[21] J.V. Michalowicz, J.M. Nichols, and F. Bucholtz: *Handbook of differential entropy*. Crc Press, 2013.

[22] K. Ito and K. Kunisch: *Lagrange multiplier approach to variational problems and applications*. SIAM, 2008.

[23] K. Rektorys: *Variational methods in mathematics, science and engineering*. Springer Science & Business Media, 2012.

[24] S. Kotz, T. Kozubowski, and K. Podgorski: *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2012.

[25] T. Pawitan: *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[26] W. Bryc: *The Normal Distribution: Characterizations with Applications*, Lecture Notes in Statistics, vol 100. Springer New York, New York, NY, 1995. DOI: 10.1007/978-1-4612-2560-7

[27] B. Xu, D. Wang, and H. Qi: Comparative studies between the bayesian estimation and the maximum likelihood estimation of the parameter of the uniform distribution. *International Journal of Modelling, Identification and Control*, **35**(3), (2020), 211–216.

[28] T.K. Pogány and S. Nadarajah: On the characteristic function of the generalized normal distribution. *Comptes Rendus Mathématique*, **348**(3–4), (2010), 203–206.

[29] N. Ravishanker, Z. Chi, and D.K. Dey: *A first course in linear model theory*. Chapman and Hall/CRC, 2021.

[30] J.X. Pan and K.T. Fang: *Maximum likelihood estimation*. In: Growth curve models and statistical diagnostics. Springer, 77–158, 2002.

[31] J. Nelder and Y. Lee: Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **54**(1), (1992), 273–284.

[32] W.M. Bolstad and J.M. Curran: *Introduction to Bayesian statistics*. John Wiley & Sons, 2016.

[33] J.E. Cavanaugh and A.A. Neath: The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, **11**(3), (2019), e1460.

[34] A.A. Neath and J.E. Cavanaugh: The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, **4**(2), (2012), 199–203.

[35] J. Ding, V. Tarokh, and Y. Yang: Model selection techniques: An overview. *IEEE Signal Processing Magazine*, **35**(6), (2018), 16–34.

[36] D. Bertsekas: *Convex optimization algorithms*. Athena Scientific, 2015.

[37] L. Zhang, W. Zhou, and D. Li: Global convergence of a modified fletcher–reeves conjugate gradient method with armijo-type line search. *Numerische Mathematik*, **104**(4), (2006), 561–572.

[38] L. Grippo and S. Lucidi: A globally convergent version of the polak-ribiere conjugate gradient method. *Mathematical Programming*, **78**(3), (1997), 375–391.

[39] M.J.D. Powell: Some convergence properties of the conjugate gradient method. *Mathematical Programming*, **11**(1), (1976), 42–49.

[40] M.J. Colaco and H.R. Orlande: Comparison of different versions of the conjugate gradient method of function estimation. *Numerical Heat Transfer: Part A: Applications*, **36**(2), (1999), 229–249.

[41] Y.H. Dai: Convergence properties of the bfgs algoritm. *SIAM Journal on Optimization*, **13**(3), (2002), 693–701.

[42] A. Mokhtari and A. Ribeiro: Res: Regularized stochastic bfgs algorithm. *IEEE Transactions on Signal Processing*, **62**(23), (2014), 6089–6104.

[43] J.S. Pan, P. Hu, V. Snášel, et al.: A survey on binary metaheuristic algorithms and their engineering applications. *Artificial Intelligence Review*, **56**(7), (2023), 6101-–6167.

[44] T. Bäck, D.B. Fogel, and Z. Michalewicz: Mutation operators. In: *Evolutionary Computation 1*. CRC Press, 2018, 275–293.

[45] P. Ruengvirayudh and G.P. Brooks: Comparing stepwise regression models to the best-subsets models, or, the art of stepwise. *General Linear Model Journal*, **42**(1), (2016), 1–14.

[46] S. Liu and C. Lin: A heuristic method for the combined location routing and inventory problem. *The International Journal of Advanced Manufacturing Technology*, **26**(4), (2005), 372–381.

[47] scikit-learn developers (2025) scikit-learn: sklearn.datasets.make_regression (module documentation). https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_regression.html, accessed: 2025-09-24