EXTENDING THE USE OF MDL FOR HIGH-DIMENSIONAL PROBLEMS: VARIABLE SELECTION, ROBUST FITTING, AND ADDITIVE MODELING

Zhenyu Wei[⋆] *Raymond K. W. Wong*[†] *Thomas C. M. Lee*[⋆]

[⋆]University of California, Davis [†]Texas A&M University

ABSTRACT

In the signal processing and statistics literature, the minimum description length (MDL) principle is a popular tool for choosing model complexity. Successful examples include signal denoising and variable selection in linear regression, for which the corresponding MDL solutions often enjoy consistent properties and produce very promising empirical results. This paper demonstrates that MDL can be extended naturally to the high-dimensional setting, where the number of predictors p is larger than the number of observations n. It first considers the case of linear regression, then allows for outliers in the data, and lastly extends to the robust fitting of nonparametric additive models. Results from numerical experiments are presented to demonstrate the efficiency and effectiveness of the MDL approach.

Index Terms— denoising, heavy-tailed errors, outliers, spline fitting, variable screening

1. INTRODUCTION

The minimum description length (MDL) principle [1, 2] has long been successfully applied to perform signal denoising [3, 4, 5] and model selection in regression and time series problems [6, 7, 8, 9, 10, 11, 12]. This paper extends MDL to solve some high-dimensional problems, including linear regression, nonparametric additive models, and robust fitting. Notice that this paper does **not** claim that MDL is the only approach for doing so. Instead, it shows that MDL can be extended to solve these problems in a conceptually clean and natural manner, and yet produce excellent results.

A typical description of the high-dimensional linear regression problem is as follows. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a vector of n responses and $\boldsymbol{x}_i$ be a p-variate predictor variable for y_i. Write $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ as the design matrix of size $n \times p$. The observed responses and the predictors are related by the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown parameters and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is a vector of i.i.d. random Gaussian errors with zero mean and unknown variance σ^2. It is assumed that $p \gg n$, making this high-dimensional regression problem different from the classical multiple regression regression problem for which $p < n$.

When $p \gg n$, one needs to assume that the number of significant predictors in the true model is small; i.e., the true model is sparse. The problem is then to identify which β_j's are non-zero, which is known as the variable selection problem. Although there are existing methods for addressing this problem [13], this paper seems to be one of the earliest attempts that MDL is being applied and carefully studied in high-dimensional settings.

This paper also considers robust fitting; i.e., it allows the possible presence of heavy-tailed errors or outliers in the response. It achieves this goal by modeling the error component with the Laplace distribution. Lastly this paper extends the variable selection problem for (1) to high-dimensional nonparametric additive models, defined as

$$y_i = \mu + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where μ is an intercept term, the f_j's are unknown nonparametric functions, and x_{ij} is the jth covariate of $\boldsymbol{x}_i$. Again, we consider $p \gg n$ and impose the sparsity assumption. We select the significant predictors, as well as allowing the possibility of outliers.

Due to space limitation, some theoretical derivations and empirical results are delayed to the full version of this paper [14], which will be uploaded to arXiv.org in due course.

2. A BRIEF DESCRIPTION OF MDL

In model selection problems the MDL principle defines the best fitting model as the one that produces the shortest code length of the data [1, 2]. In this context the code length of an object can be treated as the amount of memory space that is required to store the object. Of course comparing code lengths is neither the only nor the best approach for defining a best fitting model, but it is still a sensible one. It is because a common feature of a good encoding (or compression) scheme and a good statistical model is the ability to capture the regularities, or patterns, hidden in the data.

There are different versions of MDL, and this paper focuses on the so-called two-part codes. When applying this, it

is common to split the code length for a set of data into two parts: (i) a fitted model plus (ii) the data "conditioned on" the fitted model; i.e., the residuals. If we denote the data as $\boldsymbol{y}$, any fitted model as $\hat{\boldsymbol{\theta}}$, and the residuals as $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, where $\hat{\boldsymbol{y}}$ is the fitted value of $\boldsymbol{y}$, we split $\boldsymbol{y}$ into $\hat{\boldsymbol{\theta}}$ plus $\hat{\boldsymbol{e}}$. Notice that knowing $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{e}}$ can completely retrieve $\boldsymbol{y}$.

If $\mathrm{CL}(z)$ denotes the code length of an object z, we have $\mathrm{CL}(\boldsymbol{y}) = \mathrm{CL}(\hat{\boldsymbol{\theta}}) + \mathrm{CL}(\hat{\boldsymbol{e}}|\hat{\boldsymbol{\theta}})$. Note that in this expression it is stressed that $\hat{\boldsymbol{e}}$ is conditional on $\hat{\boldsymbol{\theta}}$; i.e., different $\hat{\boldsymbol{\theta}}$'s would give different $\hat{\boldsymbol{e}}$'s. Now the task is to find an expression for $\mathrm{CL}(\boldsymbol{y})$ so that the best MDL $\hat{\boldsymbol{\theta}}$ can be defined and obtained as its minimizer.

3. HIGH-DIMENSIONAL LINEAR REGRESSION

We first consider variable selection for model (1). Let S be a subset of $\{1, \ldots, p\}$. If $j \in S$, it means β_j is significant. Hence S can be used to represent any candidate model. Denote the corresponding design matrix as $\boldsymbol{X}_S$, and the maximum likelihood estimate of the corresponding coefficients $\boldsymbol{\beta}_S$ as $\hat{\boldsymbol{\beta}}_S$. Also, let $|S|$ be the number of elements in S, i.e., the number of significant β_j's. It is shown in Section 3.1 that a MDL criterion for the model specified by S is

$$\mathrm{MDL}(S) = \frac{n}{2}\log\left(\frac{\mathrm{RSS}}{n}\right) + \frac{|S|}{2}\log(n) + |S|\log(p), \quad (3)$$

where $\mathrm{RSS} = (\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\beta}}_S)^T(\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\beta}}_S)$ is the residual sum of squares. When comparing to the classical MDL criterion for $p < n$, this new $\mathrm{MDL}(S)$ has an additional penalty term $|S|\log(p)$, which coincidentally shares the same asymptotic order as the corresponding penalty term in EBIC of [15].

3.1. Derivation of MDL

This section outlines the derivation of (3). Here the parameter vector estimate $\hat{\boldsymbol{\theta}}$ is $\hat{\boldsymbol{\beta}}_S$, so we begin with $\mathrm{MDL}(S) = \mathrm{CL}(\boldsymbol{y}) = \mathrm{CL}(\hat{\boldsymbol{\beta}}_S) + \mathrm{CL}(\hat{\boldsymbol{e}}|\hat{\boldsymbol{\beta}}_S)$. According to [1], the code length for encoding an integer N is approximately $\log_2 N$ bits. To encode $\hat{\boldsymbol{\beta}}_S$, one first needs to identify which of the $|S|$ predictors are selected. Since each of the $|S|$ predictors can be uniquely identified by an index in $\{1, \ldots, p\}$, it takes a total of $|S|\log_2 p$ bits to encode this information. Next, the corresponding parameter estimates need to be encoded. In [1] it is demonstrated that if a maximum likelihood estimate of a real-valued parameter is computed from N data points, then it can be effectively encoded with $\frac{1}{2}\log_2 N$ bits. This gives the total code length for the $|S|$ parameter estimates as $\frac{|S|}{2}\log_2 n$, and hence $\mathrm{CL}(\hat{\boldsymbol{\beta}}_S) = |S|\log_2 p + \frac{|S|}{2}\log_2 n$. Notice that in classical applications of MDL for problems with $p \ll n$, the term $|S|\log_2 p$ is often omitted as it is relatively small compared with $\frac{|S|}{2}\log_2 n$. However, when p is comparable to n or even $p \gg n$, this term cannot be omitted as otherwise it will give erratic results.

Now it remains to calculate $\mathrm{CL}(\hat{\boldsymbol{e}}|\hat{\boldsymbol{\beta}}_S)$. It is shown in [1] that this term equals to the negative of the log of the likelihood of $\hat{\boldsymbol{e}}$ conditioned on $\hat{\boldsymbol{\beta}}_S$. For the present problem, it simplifies to $\mathrm{CL}(\hat{\boldsymbol{e}}|\hat{\boldsymbol{\beta}}_S) = \frac{n}{2}\log_2\left(\frac{\mathrm{RSS}}{n}\right)$. Changing $\log_2$ to $\log$ and adding $\mathrm{CL}(\hat{\boldsymbol{\beta}}_S)$ and $\mathrm{CL}(\hat{\boldsymbol{e}}|\hat{\boldsymbol{\beta}}_S)$, one obtains $\mathrm{MDL}(S)$ in (3).

3.2. Practical Minimization of (3)

In practice minimizing (3) is not a trivial task, as $p \gg n$. We propose using a three-stage procedure to locate a good approximated minimizer of (3). The first stage is to apply a screening procedure to remove a large number of insignificant predictors, so that we will only have to consider the remaining m predictors, where $m < n$. Then in the second stage the lasso [16] method is applied to obtain a nested sequence of m candidate models. Lastly, the $\mathrm{MDL}(S)$ values for these m candidate models are calculated and the one with the smallest value is taken as the final, best fitting model. Due to space limitation, details are skipped but can be found in [14].

4. THEORETICAL PROPERTIES

Let S_0 be the index set of the true model, and $\boldsymbol{\mu} = E(\boldsymbol{y}) = \boldsymbol{X}_{S_0}\boldsymbol{\beta}_{S_0}$. Define the projection matrix for any $S \subset \{1, \ldots, p\}$ as $\boldsymbol{P}_S = \boldsymbol{X}_S(\boldsymbol{X}_S^T\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^T$, and write $\delta(S) = \|\boldsymbol{\mu} - \boldsymbol{P}_S\boldsymbol{\mu}\|^2$, with $\|\cdot\|$ being the Euclidean norm. Clearly, if $S_0 \subset S$, we have $\delta(S) = 0$. To proceed, we need the following identifiability condition.

Condition 1. *The true model S_0 is asymptotically identifiable if* $\lim_{n\to\infty} \min\left\{\frac{\delta(S)}{\log n} : S \neq S_0, |S| \leq k|S_0|\right\} = \infty$ *for some fixed* $k > 1$.

Now we can state our main theorem.

Theorem 1. *Consider a data set $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ from model (1). Suppose Condition 1 holds and $p = O(n^\gamma)$ for some fixed γ. Also assume $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Then we have $P\left(\min_{S\neq S_0, |S|\leq k|S_0|} \mathit{MDL}(S) > \mathit{MDL}(S_0)\right) \to 1$ as $n \to \infty$.*

Theorem 1 ensures that, as $n \to \infty$, the probability that $\mathrm{MDL}(S)$ wrongly selects a model of a similar size other than the true model goes to zero. Its proof can be found in [14].

5. ROBUST FITTING

This section considers robust estimation for high-dimensional linear regression. Some non-MDL work include [17, 18, 19].

Robust fitting can be natually embedded into the MDL framework by adopting a heavy tail distribution for the errors ε_i's to allow for outliers. For this we suggest using the zero mean Laplace distribution; i.e., $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{Laplace}(0, b)$, where b is a scale parameter.

Similar to Section 3.1, it can be shown that the MDL criterion for robust fitting with a model specified by S is

$$\text{MDL}_{\text{robust}}(S) = n \log\left(\frac{\text{SAE}}{n}\right) + \frac{|S|}{2}\log(n) + |S|\log(p). \tag{4}$$

In the above $\text{SAE} = \sum_{i=1}^{n} |y_i - \boldsymbol{x}_{S,i}\hat{\boldsymbol{\beta}}_S|$ is the sum of absolute errors, with $\boldsymbol{x}_{S,i}$ denoting the i-th row of $\boldsymbol{X}_S$.

Practical minimization of (4) can be achieved in a similar fashion as the 3-stage procedure described in Section 3.2. To be more specific, Stage 1 remains the same, while in Stage 2 the robust LARS method of [18] is used in place of the original lasso, and in Stage 3 the MDL criterion (4) is used instead of (3) when calculating the MDL values.

6. HIGH-DIMENSIONAL NONPARAMETRIC ADDITIVE MODELS

This section extends our work to the high-dimensional nonparametric additive models (2). The goal is to select those significant ones from the functions $f_1, \ldots, f_p$, as well as to estimate them nonparametrically. We first discuss the use of splines for modeling the f_j's.

6.1. Spline Modeling for Additive Functions

Briefly, a spline function is a piecewise polynomial function. The locations at which two adjacent pieces join are called knots. First we state their standard conditions and definition.

Suppose that $x \in [a, b]$ for some finite numbers $a < b$ and that $E(y^2) < \infty$. To ensure identifiability, it is assumed $E\{f_j(x)\} = 0$ for $j = 1, \ldots, p$. Let K be the number of knots for a partition of $[a, b]$ that satisfy specific conditions stated for example in [20]. Let $\mathcal{S}_n$ be the collection of functions s with domain $[a, b]$ satisfying the following two conditions: (i) s is a polynomial of degree l (or less) on each subinterval, and (ii) for any two integers l and l' satisfying $l \geq 2$ and $0 \leq l' < l - 1$, s is l'-times continuously differentiable on $[a, b]$. Then there exists a normalized B-spline basis $\{\varphi_k(\cdot), k = 1, \ldots, d_n\}$ such that for any $s \in \mathcal{S}_n$, we have

$$s(x) = \sum_{k=1}^{d_n} \alpha_k \varphi_k(x), \tag{5}$$

where α_k is the coefficient of the basis function $\varphi_k(x)$ for $k = 1, \ldots, d_n$ with $d_n = K + l$. Since $\mathcal{S}_n$ is a relatively rich class of smooth functions, in this paper, for the reason of speeding up technical calculations, we shall assume that the spline representation (5) is exact for the functions f_j's. In other words, for $j = 1, \ldots, p$, it is assumed that

$$f_j(x) = \sum_{k=1}^{d_n} \alpha_{jk} \varphi_k(x), \tag{6}$$

where α_{jk}'s are the corresponding coefficients of the bases $\varphi_k(x)$'s.

6.2. MDL Criteria

Recall that for the fitting of the high-dimensional nonparametric additive models (2), we aim to select those significant functions from $f_1, \ldots, f_p$, as well as to estimate them nonparametrically. For any candidate model, denote the number of significant f_j's as q, and the number of basis functions used for each f_j as d_n. Using similar steps as in Section 3.1, it can be shown that an MDL criterion for fitting (2) is:

$$\text{MDL}_{\text{additive}}(S) = \frac{n}{2}\log\left(\frac{\text{RSS}}{n}\right) + \frac{qd_n}{2}\log(n) + q\log(p). \tag{7}$$

One can also perform robust fitting as in Section 5 above, and the resulting MDL criterion is

$$\text{MDL}^{\text{robust}}_{\text{additive}}(S) = n\log\left(\frac{\text{SAE}}{n}\right) + \frac{qd_n}{2}\log(n) + q\log(p). \tag{8}$$

MDL criteria (7) and (8) can be minimized in a similar manner as in Section 3.2. See [14] for details.

7. NUMERICAL EXPERIMENTS

7.1. Linear Regression

Following the settings in [21], the data were generated with the model $y_i = b(x_{i1} + \ldots + x_{id}) + \varepsilon_i$ for $i = 1, \ldots, n$, where the coefficient b controls the signal-to-noise ratio. The $\boldsymbol{x}_i$'s are standard normal variables with the correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ set to be $\rho^{|i-j|}$ with $\rho = 0.5$. The number of observations was n, and the number of predictors was p, where only the first d are significant. Five combinations of (n, p, d) were used: $(100, 1000, 3)$, $(200, 3000, 5)$, $(300, 10000, 8)$, $(200, 100000, 5)$ and $(300, 200000, 8)$. For each of these 5 combinations, 3 values of b were used: $b = 2/\sqrt{d}, 3/\sqrt{d}$ and $5/\sqrt{d}$. For the error term $\boldsymbol{\varepsilon}$, 4 distributions were used: $N(0, 1)$, Laplace$(0, 1)$, t_3 and a Gaussian mixture with two components: $95\%N(0, 1)$ and $5\%N(0, 7^2)$. The last one represents the situation where roughly 5% of the observations are outliers. Therefore, a total of $5 \times 3 \times 4 = 60$ experimental configurations were considered. The number of repetitions for each experimental configuration was 500.

For each generated data set, six methods were applied to select a best fitting model: (i) MDL: the MDL method proposed in Section 3, (ii) RobustMDL: the robust version proposed in Section 5, (iii) RLARS: the robust LARS method of [18], (iv) LAD-LASSO: the least absolute deviation lasso of [17] (v) SparseLTS: the sparse least trimmed squares method of [19], and (vi) Welsh: the adaptive welsh estimators of [22]. Since this method is quite computationally expensive, we only applied it to the cases with $n = 100$ and $p = 1000$.

To evaluate the performances of different methods on variable selection, we calculated the false negative error of selection (FN) and the false positive error of selection (FP), defined respectively as FN $= \#$ of $\{i : \beta_i \neq 0 \ \& \ \hat{\beta}_i = 0\}$

and FP $= \#$ of $\{i : \beta_i = 0 \ \& \ \hat{\beta}_i \neq 0\}$. We also calculated the F1 score and mean squared error (MSE) between the estimated and true signal $\boldsymbol{X}\boldsymbol{\beta}$. The FN, FP, F1 score and MSE values for all the 60 different experimental configurations are given in [14]. Here we summarize a subset of the results in Tables 1 and 2, which is representative for the full results. When considering computational speeds and performances, it seems that RobustMDL is the preferred method.

Table 1. The FN, FP, F1 score and MSE values for the methods compared in Section 7.1 for those experimental settings with (n, p) = (100,1000) and $\varepsilon_i \sim$ Laplace$(0, 1)$.

(d, b)	method	FN	FP	F1	MSE	time(s)
$(3, \frac{2}{\sqrt{3}})$	MDL	0.03	0.05	0.99	0.12	0.05
	RobustMDL	0.00	0.05	0.99	0.09	0.04
	RLARS	0.00	1.99	0.81	0.24	0.14
	LAD-lasso	0.00	0.47	0.94	0.45	1.10
	SparseLTS	0.00	15.24	0.35	0.45	3.28
	Welsh	0.63	0.13	0.78	1.56	1045.80
$(3, \frac{3}{\sqrt{3}})$	MDL	0.00	0.02	1.00	0.09	0.05
	RobustMDL	0.00	0.02	1.00	0.09	0.04
	RLARS	0.00	1.64	0.84	0.20	0.14
	LAD-lasso	0.00	0.47	0.94	0.45	1.15
	SparseLTS	0.00	7.79	0.53	0.45	2.74
	Welsh	0.39	0.07	0.87	2.12	1034.46
$(3, \frac{5}{\sqrt{3}})$	MDL	0.00	0.01	1.00	0.09	0.05
	RobustMDL	0.00	0.01	1.00	0.09	0.04
	RLARS	0.00	1.22	0.87	0.16	0.14
	LAD-lasso	0.00	0.48	0.94	0.43	1.26
	SparseLTS	0.00	4.11	0.64	0.49	2.40
	Welsh	0.37	0.18	0.86	5.26	1031.58

Table 2. Similar to Table 1 but for settings with (n, p) = $(200, 3000)$ and $\varepsilon_i \sim t_3$.

(d, b)	method	FN	FP	F1	MSE	time(s)
$(5, \frac{2}{\sqrt{5}})$	MDL	0.17	0.05	0.97	0.19	0.11
	RobustMDL	0.00	0.05	1.00	0.10	0.10
	RLARS	0.00	1.67	0.88	0.15	0.27
	LAD-lasso	0.00	0.42	0.96	0.44	5.30
	SparseLTS	0.00	10.68	0.60	0.36	20.06
$(5, \frac{3}{\sqrt{5}})$	MDL	0.00	0.05	1.00	0.11	0.11
	RobustMDL	0.00	0.03	1.00	0.09	0.11
	RLARS	0.00	1.15	0.91	0.11	0.28
	LAD-lasso	0.00	0.44	0.96	0.42	5.68
	SparseLTS	0.00	8.21	0.63	0.36	18.82
$(5, \frac{5}{\sqrt{5}})$	MDL	0.00	0.03	1.00	0.10	0.11
	RobustMDL	0.00	0.02	1.00	0.09	0.10
	RLARS	0.00	0.77	0.94	0.08	0.27
	LAD-lasso	0.00	0.42	0.96	0.40	6.27
	SparseLTS	0.00	2.54	0.82	0.41	17.83

7.2. Nonparametric Additive Models

For nonparametric additive models defined as (2), we set $n = 400$ and $p = 1000$. Only the first four f_j's are significant: $f_1(x) = 5x$, $f_2(x) = 3(2x-1)^2$, $f_3(x) = 4\sin(2\pi x)/\{2 - \sin(2\pi x)\}$, $f_4(x) = 6\{0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)\}$ and $f_j(x) = 0$ for $5 \leq j \leq p$. The errors ε_i were generated from four distributions: $N(0, 1)$, Laplace$(0, 1)$, t_5 and a Gaussian mixture with $95\%N(0, 1)$ and $5\%N(0, 5^2)$. For each i, the x_{ij}'s were generated from $x_{ij} = (\omega_{ij} + tu_i)/(1+t)$ for $j = 1, \ldots, 4$ and $x_{ij} = (\omega_{ij} + tk_i)/(1 + t)$ for $j = 5, \ldots, p$, where $\omega_{i1}, \ldots, \omega_{ip}, u_i, k_i$ were i.i.d Uniform(0,1). The parameter t controls the correlation among the predictors, and we used $t = (0, 1)$ in our simulation. Also, we used the cubic B-spline with six evenly spaced knots for all the function f_j's; that is, we used $d_n = 9$ basis functions to approximate each f_j. Therefore, in total there were 8 experimental configurations, and the number of replications in each configuration was 500. For each replication, we obtained MDL and RobustMDL estimates by minimizing (7) and (8). The results are summarized in Table 3. We only tested these two methods as we are not aware of any other method that performs robust fitting for high-dimensional additive models.

From the simulation results, RobustMDL gave better performances in terms of FN and MSE, and provided similar results in FP as with MDL.

Table 3. The FN, FP and MSE values for the two methods compared in Section 7.2.

error distribution	t	method	FN	FP	MSE
$N(0, 1)$	0	MDL	0.01	0.00	0.93
		RobustMDL	0.01	0.00	0.93
	1	MDL	0.05	0.02	0.94
		RobustMDL	0.05	0.02	0.94
Laplace(0,1)	0	MDL	0.01	0.00	1.84
		RobustMDL	0.01	0.08	1.83
	1	MDL	0.47	0.02	1.99
		RobustMDL	0.23	0.09	1.89
t_5	0	MDL	0.01	0.00	1.56
		RobustMDL	0.01	0.00	1.56
	1	MDL	0.31	0.04	1.65
		RobustMDL	0.16	0.08	1.59
Outliers	0	MDL	0.14	0.00	3.17
		RobustMDL	0.02	0.00	3.10
	1	MDL	1.49	0.00	3.83
		RobustMDL	0.73	0.17	3.40

8. CONCLUDING REMARKS

The MDL principle has long been adopted by researchers in different fields to perform various estimation tasks. In this paper we extended its use to some "large p small n" problems, including high-dimensional linear regression, nonparametric additive models, as well as their robust counterparts. As can be seen from above, one attractiveness of the MDL principle is that it can be applied to handle such problems in a natural manner; that is, by incorporating the code length of the additional parameters that are needed to specify the models.

9. REFERENCES

[1] Jorma Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[2] Jorma Rissanen, *Information and Complexity in Statistical Modeling*, Springer Science & Business Media, 2007.

[3] Israel Cohen, Shalom Raz, and David Malah, "Translation-invariant denoising using the minimum description length criterion," *Signal Processing*, vol. 75, pp. 201–223, 1999.

[4] Jorma Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, pp. 2537–2543, 2000.

[5] Teemu Roos, Petri Myllymaki, and Jorma Rissanen, "MDL denoising revisited," *IEEE Transactions on Signal Processing*, vol. 57, pp. 3347–3360, 2009.

[6] Alexander Aue, Rex C. Y. Cheung, Thomas C. M. Lee, and Ming Zhong, "Segmented model selection in quantile regression using the minimum description length principle," *Journal of the American Statistical Association*, vol. 109, pp. 1241–1256, 2014.

[7] Rex C. Y. Cheung, Alexander Aue, and Thomas C. M. Lee, "Consistent estimation for partition-wise regression and classification models," *IEEE Transactions on Signal Processing*, vol. 65, pp. 3662–3674, 2017.

[8] Qi Gao, Thomas C. M. Lee, and Chun Yip Yau, "Nonparametric modeling and break point detection for time series signal of counts," *Signal Processing*, vol. 138, pp. 307–312, 2017.

[9] Sreejith Kallummil and Sheetal Kalyani, "High SNR consistent linear model order selection and subset selection," *IEEE Transactions on Signal Processing*, vol. 64, pp. 4307–4322, 2016.

[10] Thomas C. M. Lee, "Regression spline smoothing using the minimum description length principle," *Statistics and Probability Letters*, vol. 48, pp. 71–82, 2000.

[11] Daniel F Schmidt and Enes Makalic, "The consistency of MDL for linear regression models with increasing signal-to-noise ratio," *IEEE Transactions on Signal Processing*, vol. 60, pp. 1508–1510, 2011.

[12] Raymond K. W. Wong, Randy C. S. Lai, and Thomas C. M. Lee, "Structural break estimation of noisy sinusoidal signals," *Signal processing*, vol. 90, pp. 303–312, 2010.

[13] Jianqing Fan and Jinchi Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, pp. 101–148, 2010.

[14] Zhenyu Wei, Raymond K. W. Wong, and Thomas C. M. Lee, "Extending the use of MDL for high-dimensional problems: Variable selection, robust fitting, and additive modeling (full version)," https://arxiv.org/abs/2201.11171, 2022.

[15] Jiahua Chen and Zehua Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, pp. 759–771, 2008.

[16] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, pp. 267–288, 1996.

[17] Hansheng Wang, Guodong Li, and Guohua Jiang, "Robust regression shrinkage and consistent variable selection through the lad-lasso," *Journal of Business & Economic Statistics*, vol. 25, pp. 347–355, 2007.

[18] Jafar A Khan, Stefan Van Aelst, and Ruben H Zamar, "Robust linear model selection based on least angle regression," *Journal of the American Statistical Association*, vol. 102, pp. 1289–1299, 2007.

[19] Andreas Alfons, Christophe Croux, and Sarah Gelper, "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *The Annals of Applied Statistics*, pp. 226–248, 2013.

[20] Randy C. S. Lai, Hsin-Cheng Huang, and Thomas C. M. Lee, "Fixed and random effects selection in nonparametric additive mixed models," *Electronic Journal of Statistics*, vol. 6, pp. 810–842, 2012.

[21] Jianqing Fan, Shaojun Guo, and Ning Hao, "Variance estimation using refitted cross-validation in ultrahigh dimensional regression," *Journal of the Royal Statistical Society: Series B*, vol. 74, pp. 37–65, 2012.

[22] Umberto Amato, Anestis Antoniadis, Italia De Feis, and Irene Gijbels, "Penalised robust estimators for sparse and high-dimensional linear models," *Statistical Methods & Applications*, vol. 30, no. 1, pp. 1–48, 2021.