Model Selection in High-Dimensional Block-Sparse General Linear Regression

Prakash B. Gohain*, Magnus Jansson†
*†Division of Information Science and Engineering, KTH Royal Institute of Technology
*Ericsson
Stockholm, Sweden
prakash.b.gohain@gmail.com, janssonm@kth.se

***Abstract*—Model selection is an indispensable part of data analysis dealing very frequently with fitting and prediction purposes. In this paper, we tackle the problem of model selection in a general linear regression where the parameter matrix possesses a block-sparse structure, i.e., the non-zero entries occur in clusters or blocks and the number of such non-zero blocks is very small compared to the parameter dimension. Furthermore, a high-dimensional setting is considered where the parameter dimension is quite large compared to the number of available measurements. To perform model selection in this setting, we present an information criterion that is a generalization of the Extended Bayesian Information Criterion-Robust (EBIC-R) and it takes into account both the block structure and the high-dimensionality scenario. We name it Generalized EBIC-R (GEBIC-R). The analytical steps for deriving the GEBIC-R are provided. Simulation results show that the proposed method performs considerably better than the existing state-of-the-art methods and achieves empirical consistency at large sample sizes and/or at high-SNR.**

***Index Terms*—Model selection, block-sparsity, compressed sensing, information criterion, orthogonal matching pursuit.**

I. Introduction

Selecting the best model/subset in the high-dimensional (HD) linear regression has been an active research topic for a long time now. In this context, methods based on Information Criterion (IC) have played a pivotal role ever since Akaike proposed the famous Akaike IC [1]. In the present era, popular IC-based methods for model selection in the HD setting include extended Bayesian IC (EBIC) [2], extended Fisher IC (EFIC) [3], and extended BIC-Robust (EBIC_R) [4], [5]. Apart from the IC-based methods, there are other non-IC-based model selection approaches in the HD regime. They include methods based on hypothesis testing framework such as the Residual-Ratio-Thresholding (RRT) [6] and the Multi-Beta-Test (MBT) [7]. Some recent methods also include the significance test of the LASSO [8] and knock-off-filters [9]. Another popular method is cross-validation (CV) [10], [11]. However, CV-based procedures can be computationally intensive, and their performance in the HD setting is not satisfactory [12]–[14].

In this paper, we specifically consider model selection in a general linear regression where the nonzero coefficients in the parameter matrix occur in clusters (or groups). Such signals are referred to as block-sparse [15]–[17]. Block-sparsity inherently arises in a variety of scenarios. For example in multi-band signals [18], [19], in the recovery of signals from compressed microarray measurements [20], and in the multiple measurement vector (MMV) problem [21]–[24]. Furthermore, as shown in [15] and [16], the block-sparsity model can be used to handle the issue of sampling signals that lie in a union of subspaces [25], [26].

A recent method for model selection in block-sparse HD linear regression is the Generalized RRT (GRRT) [27]. GRRT is an extension of RRT [6] developed to treat the block-sparse structure in linear regression. The authors also present a new approach that allows GRRT to perform model selection in non-monotonic predictor sequences generated by LASSO [28]. However, to the best of our knowledge, there are no existing IC-based methods designed to take into account the block structure during model selection. Hence, in their current form, they cannot be applied directly without tailoring them to incorporate the block nature of the underlying linear model into the criterion. In this paper, the main goal is to develop an IC-based model selection method for the general linear regression model assuming a block structure and HD setting.

Notations: matrices and vectors are denoted by boldface letters. The notation $(\cdot)^T$ stands for transpose. $\mathbf{I}_N$ is an $N \times N$ identity matrix. $\mathbf{\Pi}(\mathbf{A}) = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ represents the orthogonal projection matrix on the column space of $\mathbf{A}$ and $\mathbf{\Pi}^\perp(\mathbf{A}) = \mathbf{I}_N - \mathbf{\Pi}(\mathbf{A})$ the orthogonal projection matrix on the null space of $\mathbf{A}^T$. The notation $\left|\mathbf{X}\right|$ denotes the determinant of the matrix $\mathbf{X}$, $\|\cdot\|_2$ denotes the Euclidean norm and $\|\cdot\|_F$ the Frobenius norm. $X \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ signifies a Gaussian distributed random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$. The symbol $\otimes$ represents the Kronecker product and $\text{vec}(\mathbf{A})$ the vectorization of the matrix $\mathbf{A}$. Further, $card(\mathcal{S})$ denotes the cardinality of the set $\mathcal{S}$ and $\mathcal{O}(\cdot)$ denotes the standard Big-O notation.

II. Problem Statement

Technically there can be four different linear regression structures depending on the configuration of the parameter matrix (or vector). They are (a) single measurement vector (SMV), (b) block single measurement vector (BSMV), (c)

This research was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 742648.

Type	Specifications	$dim(\mathbf{Y})$, $dim(\mathbf{X})$
SMV	$L=1$, $L_B=1$, $p_B=p$	$N\times 1$, $p\times 1$
MMV	$L>1$, $L_B=1$, $p_B=p$	$N\times L$, $p\times L$
BSMV	$L=1$, $L_B>1$, $p_B=p/L_B$	$N\times 1$, $p\times 1$
BMMV	$L>1$, $L_B>1$, $p_B=p/L_B$	$N\times L$, $p\times L$

TABLE I
TYPES OF LINEAR REGRESSION STRUCTURES.

multiple measurement vector (MMV) and (d) block multiple measurement vector (BMMV). For example, as mentioned in [27], SMV models are used in wireless signal detection [29], MMV models in Electroencephalogram (EEG) [30], BSMV models in multi pitch estimation [31] and BMMV models in face recognition [32]. Here, we consider the BMMV model, since it is the general setting and the rest of the models are special cases of BMMV. The BMMV model is as follows:

$$\mathbf{Y}=\mathbf{A}\mathbf{X}+\mathbf{W}, \tag{1}$$

where, $\mathbf{Y}\in\mathbb{R}^{N\times L}$ is the observed response matrix, $\mathbf{A}\in\mathbb{R}^{N\times p}$ is the design matrix, $\mathbf{X}\in\mathbb{R}^{p\times L}$ is the unknown parameter matrix and $\mathbf{W}\in\mathbb{R}^{N\times L}$ is the noise/error matrix, whose elements are assumed to be i.i.d. and $\mathbf{W}[i,j]\sim\mathcal{N}(0,\sigma^2)$. The p rows of $\mathbf{X}$ are divided into $p_B=p/L_B$ unique blocks of equal size L_B. Each of these p_B blocks of size $L_B\times L$ is non-zero or zero at once. The block size L_B is assumed to be known *a-priori*. The jth block consists of the rows of $\mathbf{X}$ indexed by $\mathcal{I}_j=\{(j-1)L_B+1,(j-1)L_B+2,\dots,jL_B\}$. We denote the true block support of $\mathbf{X}$ as $\mathcal{S}_B=\{j:\mathbf{X}[\mathcal{I}_j,:]\neq\mathbf{0}_{L_BL}\}$ where $j\in\{1,2,\dots p_B\}$. Also, $\mathbf{X}$ is assumed to be block-sparse such that $K_B=card(\mathcal{S}_B)\ll p_B$. Table I shows different linear regression structures. The goal of model selection herein is estimating $\mathcal{S}_B$ given $\mathbf{Y}$ and $\mathbf{A}$.

The model selection procedure can be divided into two stages: (i) In the first stage, we pick a competent set of candidate models using an appropriate predictor/subset selection algorithm up to maximum cardinality K under the assumption that $K_B\leq K\ll N$. (ii) In the second stage, we estimate the true model using a suitable model selection criterion. Let us denote $\mathcal{I}_B$ as the block support of a candidate model such that $card(\mathcal{I}_B)=k_B$, where $k_B\in\{1,2,\dots,p_B\}$. Then we can reformulate the linear model in (1) as

$$\mathcal{H}_{\mathcal{I}_B}:\mathbf{Y}=\mathbf{A}_{\mathcal{I}_B}\mathbf{X}_{\mathcal{I}_B}+\mathbf{W}_{\mathcal{I}_B}, \tag{2}$$

where $\mathcal{H}_{\mathcal{I}_B}$ signifies the hypothesis that the data $\mathbf{Y}$ is actually produces in accordance with (2), $\mathbf{A}_{\mathcal{I}_B}\in\mathbb{R}^{N\times(k_BL_B)}$ is the sub-matrix consisting of columns from the known matrix $\mathbf{A}$ with block support $\mathcal{I}_B\subseteq\{1,2,\dots,p_B\}$, $\mathbf{X}_{\mathcal{I}_B}\in\mathbb{R}^{(k_BL_B)\times L}$ is the corresponding unknown parameter coefficient matrix, and $\mathbf{W}_{\mathcal{I}_B}\in\mathbb{R}^{N\times L}$ is the associated noise matrix.

III. PROPOSED METHOD

In this section, we provide the necessary steps to derive the GEBIC$_\text{R}$. Note that while EBIC$_\text{R}$ [4], [33] is for model selection in SMV scenarios in the absence of any block structure, GEBIC$_\text{R}$ is a generalization (or extension) of EBIC$_\text{R}$ to perform model selection in block general linear regression (e.g. BMMV scenarios which is the most general setting). The generalization involves some crucial steps that are not trivial and affect the performance drastically. The analysis assumes the following property of the design matrix $\mathbf{A}$ [34]–[36]

$$\lim_{N\to\infty}\left\{N^{-1}(\mathbf{A}_{\mathcal{I}_B}^T\mathbf{A}_{\mathcal{I}_B})\right\}=\mathbf{M}_{\mathcal{I}_B}, \tag{3}$$

where $\mathbf{M}_{\mathcal{I}_B}$ is a $(k_BL_B\times k_BL_B)$ positive definite matrix and bounded as $N\to\infty$. The assumption in (3) holds true in many cases but not all (see [34], [37] for more details).

To arrive at the GEBIC$_\text{R}$ for the BMMV model, we first reformulate the linear model in (2) into vector form as

$$\text{vec}(\mathbf{Y})=\mathbf{I}_L\otimes\mathbf{A}_{\mathcal{I}_B}\text{vec}(\mathbf{X}_{\mathcal{I}_B})+\text{vec}(\mathbf{W}_{\mathcal{I}_B}). \tag{4}$$

This step allows us to utilize the same derivation steps as in EBIC$_\text{R}$ [4] without the need to carry out the analysis from scratch. Also, (4) is technically equivalent to (2), hence we do not alter the underlying original linear model but just restructure it for our convenience. Now, let $\mathbf{y}=\text{vec}(\mathbf{Y})\in\mathbb{R}^{NL\times 1}$, $\breve{\mathbf{A}}_{\mathcal{I}}=\mathbf{I}_L\otimes\mathbf{A}_{\mathcal{I}_B}\in\mathbb{R}^{NL\times k_BL_BL}$, $\mathbf{x}_{\mathcal{I}}=\text{vec}(\mathbf{X}_{\mathcal{I}_B})\in\mathbb{R}^{k_BL_BL\times 1}$ and $\mathbf{e}_{\mathcal{I}}=\text{vec}(\mathbf{W}_{\mathcal{I}_B})\in\mathbb{R}^{NL\times 1}$. The elements of $\mathbf{e}_{\mathcal{I}}$ are i.i.d. and $\mathbf{e}_{\mathcal{I}}\sim\mathcal{N}(\mathbf{0},\sigma_{\mathcal{I}}^2\mathbf{I}_{NL})$. Then, we can rewrite (4) as

$$\mathcal{H}_{\mathcal{I}}:\mathbf{y}=\breve{\mathbf{A}}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}}+\mathbf{e}_{\mathcal{I}}, \tag{5}$$

where $\mathcal{I}\subseteq\{1,2,\dots,pL\}$. Then the pdf of $\mathbf{y}$ under $\mathcal{H}_{\mathcal{I}}$ is

$$p\left(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{I}},\mathcal{H}_{\mathcal{I}}\right)=\frac{\exp\{-\|\mathbf{y}-\breve{\mathbf{A}}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}}\|_2^2/2\sigma_{\mathcal{I}}^2\}}{(2\pi\sigma_{\mathcal{I}}^2)^{NL/2}}, \tag{6}$$

where $\boldsymbol{\theta}_{\mathcal{I}}=[\mathbf{x}_{\mathcal{I}}^T,\sigma_{\mathcal{I}}^2]^T$ comprises of all the unknown parameters of the model under $\mathcal{H}_{\mathcal{I}}$. The maximum likelihood estimates (MLE) $\hat{\boldsymbol{\theta}}_{\mathcal{I}}=[\hat{\mathbf{x}}_{\mathcal{I}}^T,\hat{\sigma}_{\mathcal{I}}^2]^T$ are obtained as [38]

$$\hat{\mathbf{x}}_{\mathcal{I}}=\left(\breve{\mathbf{A}}_{\mathcal{I}}^T\breve{\mathbf{A}}_{\mathcal{I}}\right)^{-1}\breve{\mathbf{A}}_{\mathcal{I}}^T\mathbf{y}\quad\&\quad\hat{\sigma}_{\mathcal{I}}^2=\frac{\mathbf{y}^T\boldsymbol{\Pi}^{\perp}(\breve{\mathbf{A}}_{\mathcal{I}})\mathbf{y}}{NL}. \tag{7}$$

GEBIC$_\text{R}$ is derived under the Bayesian framework of model selection. We follow similar steps as in EBIC$_\text{R}$ [4], [33], but incorporate the block structure into it. Let us denote the prior pdf of the parameter vector $\boldsymbol{\theta}_{\mathcal{I}}$ as $p(\boldsymbol{\theta}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$, the marginal of $\mathbf{y}$ as $p(\mathbf{y}|\mathcal{H}_{\mathcal{I}})$ and the prior probability of the model with support $\mathcal{I}$ as $\Pr(\mathcal{H}_{\mathcal{I}})$. Then the MAP estimate of the true support $\mathcal{S}\subseteq\{1,2,\dots pL\}$ is equivalently given by [34], [36]

$$\hat{\mathcal{S}}_{\text{MAP}}=\arg\max_{\mathcal{I}}\left\{\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}})+\ln\Pr\left(\mathcal{H}_{\mathcal{I}}\right)\right\}. \tag{8}$$

Applying a second order Taylor series expansion an approximation of $\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}})$ is obtained under the presumption that N is large or/and SNR is high (see [34], [36] for details)

$$\begin{aligned}\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}})\approx{}&\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathcal{I}},\mathcal{H}_{\mathcal{I}})+\ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})\\&+\frac{k_BL_BL+1}{2}\ln(2\pi)-\frac{1}{2}\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right|.\end{aligned} \tag{9}$$

Here, $\hat{\mathbf{F}}_{\mathcal{I}}$ is the sample Fisher information matrix [38] under $\mathcal{H}_{\mathcal{I}}$ evaluated at the MLE which gives us (see [34], [36])

$$\hat{\mathbf{F}}_{\mathcal{I}}=\begin{bmatrix}\frac{1}{\hat{\sigma}_{\mathcal{I}}^2}\breve{\mathbf{A}}_{\mathcal{I}}^T\breve{\mathbf{A}}_{\mathcal{I}} & \mathbf{0}\\ \mathbf{0} & \frac{NL}{2\hat{\sigma}_{\mathcal{I}}^4}\end{bmatrix}. \tag{10}$$

From the linear model in 5 we have

$$-2\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathcal{I}},\mathcal{H}_{\mathcal{I}}) = NL\ln\hat{\sigma}^2_{\mathcal{I}} + \text{const.} \tag{11}$$

Now, using (11), it is possible to rewrite (9) as

$$-2\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}}) \approx NL\ln\hat{\sigma}^2_{\mathcal{I}} + \ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| - 2\ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) \\ -k_B L_B L\ln 2\pi + \text{const.} \tag{12}$$

Furthermore, the prior term in (9), i.e., $\ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$, is ignored under the pretext that it is flat and uninformative. Thus, discarding the constants and the terms not dependent on the block model dimension k_B, we can equivalently reformulate the MAP-based model estimate as

$$\hat{\mathcal{S}}_{\text{MAP}} = \arg\min_{\mathcal{I}}\Big\{NL\ln\hat{\sigma}^2_{\mathcal{I}} + \ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| - k_B L_B L\ln 2\pi \\ - 2\ln\Pr\left(\mathcal{H}_{\mathcal{I}}\right)\Big\}. \tag{13}$$

GEBIC$_{\text{R}}$ is derived from (13) with some further modifications and approximations. The two key terms that require further analysis are $\ln|\hat{\mathbf{F}}_{\mathcal{I}}|$ and the prior term $\Pr(\mathcal{H}_{\mathcal{I}})$. First, we perform normalization of $\hat{\mathbf{F}}_{\mathcal{I}}$ under both large-N and high-SNR assumption. For this we factorize the $\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right|$term in a similar manner as performed in [4], [33], [36]

$$\begin{aligned}\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| &= \ln\left[\left|\mathbf{Q}\right|\left|\mathbf{Q}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{Q}^{-1/2}\right|\right] \\ &= \ln|\mathbf{Q}| + \ln\left|\mathbf{Q}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{Q}^{-1/2}\right|.\end{aligned} \tag{14}$$

The objective here is to choose a suitable $\mathbf{Q}$ matrix that normalizes $\hat{\mathbf{F}}_{\mathcal{I}}$ such that the second term in (14) is $\mathcal{O}(1)$, i.e., it should be bounded as $N\to\infty$ and/or $\sigma^2\to 0$. To achieve this purpose, we choose the following $\mathbf{Q}^{-1/2}$ matrix [4]

$$\mathbf{Q}^{-1/2} = \begin{bmatrix}\sqrt{\frac{L_B}{N}}\sqrt{\frac{\hat{\sigma}^2_{\mathcal{I}}}{\hat{\sigma}^2_0}}\mathbf{I}_{k_B L_B L} & \mathbf{0} \\ \mathbf{0} & \sqrt{\frac{L_B}{N}}\left(\frac{\hat{\sigma}^2_{\mathcal{I}}}{\hat{\sigma}^2_0}\right)\end{bmatrix}, \tag{15}$$

where $\hat{\sigma}^2_0 = \|\mathbf{y}\|_2^2/NL$. Also for the considered generating model (5), $\hat{\sigma}^2_0 \to$ const. as $N\to\infty$ and/or $\sigma^2\to 0$ [35], [36]. Two important points to note here regarding the choice of the $\mathbf{Q}^{-1/2}$ matrix are: (i) The ratio $\left(\frac{\hat{\sigma}^2_{\mathcal{I}}}{\hat{\sigma}^2_0}\right)$ is introduced to normalize the $\hat{\mathbf{F}}_{\mathcal{I}}$ w.r.t. σ^2 where the factor $\hat{\sigma}^2_0$ is especially utilized to counteract the data scaling problem (as discussed elaborately in [4], [36]). (ii) The $\frac{1}{N}$ portion of the factor $\frac{L_B}{N}$ is used to normalize the FIM w.r.t. N. However, note that L_B (which is absent in EBIC$_{\text{R}}$ [4], [33]) is also included as part of the normalizing term because for the mean-squared-error of $\hat{\sigma}^2$ to approach the Cramér-Rao bound, we require that $NL \gg K_B L_B L$ or in other words $N/L_B \gg K_B$. Hence, we use the normalization factor L_B/N instead of just $1/N$ in (15). In this way, the penalty will be a function of N/L_B instead of N alone (as will be seen in the subsequent steps). This novel modification helps to counteract the effects of changing L_B on the performance of GEBIC$_{\text{R}}$.

Now, using (3), (10), and (15) we can show that

$$\begin{aligned}\left|\mathbf{Q}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{Q}^{-1/2}\right| &= \left|\begin{matrix}\frac{L_B}{\hat{\sigma}^2_0}\frac{\breve{\mathbf{A}}^T_{\mathcal{I}}\breve{\mathbf{A}}_{\mathcal{I}}}{N} & \mathbf{0} \\ \mathbf{0} & \frac{L_B L}{2\hat{\sigma}^4_0}\end{matrix}\right| \\ &= \frac{L_B^{k_B L_B L+1} L}{2(\hat{\sigma}^2_0)^{k_B L_B L+2}}\left|\mathbf{I}_L \otimes \frac{\mathbf{A}^T_{\mathcal{I}_B}\mathbf{A}_{\mathcal{I}_B}}{N}\right| \\ &= \text{const.}\times\left|\mathbf{I}_L\right|^{k_B\times L_B}\left|\frac{\mathbf{A}^T_{\mathcal{I}_B}\mathbf{A}_{\mathcal{I}_B}}{N}\right|^L = \mathcal{O}(1)\end{aligned} \tag{16}$$

as N grows large and/or $\sigma^2\to 0$. Hence, this term can be removed without significantly affecting the criterion. Next, observe that the $\ln\left|\mathbf{Q}\right|$ term in 14 can be expanded as follows

$$\begin{aligned}\ln|\mathbf{Q}| &= \ln\left|\begin{matrix}\left(\frac{N}{L_B}\right)\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{I}}}\right)\mathbf{I}_{k_B L_B L} & \mathbf{0} \\ \mathbf{0} & \left(\frac{N}{L_B}\right)\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{I}}}\right)^2\end{matrix}\right| \\ &= (k_B L_B L+1)\ln\left(\frac{N}{L_B}\right) + (k_B L_B L+2)\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{I}}}\right).\end{aligned} \tag{17}$$

Therefore, using (16) and (17) we can rewrite (14) as

$$\begin{aligned}\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| =& k_B L_B L\ln\left(\frac{N}{L_B}\right) + (k_B L_B L+2)\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{I}}}\right) + \\ & \mathcal{O}(1) + \ln\left(N/L_B\right).\end{aligned} \tag{18}$$

Next, for the model prior probability term $-2\ln\Pr(\mathcal{H}_{\mathcal{I}})$ in (13), a similar strategy is adopted as in EBIC [2] such that $\Pr(\mathcal{H}_{\mathcal{I}}) \propto \binom{p_B}{k_B}^{-\zeta}$, where $\zeta\geq 0$ is a tuning parameter. If p_B is sufficiently large, the following approximation can be assumed $\ln\binom{p_B}{k_B} \approx k_B\ln p_B$ [3]. This gives

$$-2\ln\Pr(\mathcal{H}_{\mathcal{I}}) = 2\zeta k_B\ln p_B + \text{const.} \tag{19}$$

Now, substituting (18), (19) in (13) and dropping the $\mathcal{O}(1)$, the $\ln\left(N/L_B\right)$ term (since independent of k_B), the constant and the $p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ term we arrive at the GEBIC$_{\text{R}}$

$$\begin{aligned}\text{GEBIC}_{\text{R}}(\mathcal{I}) = & NL\ln\hat{\sigma}^2_{\mathcal{I}} + k_B L_B L\ln\left(\frac{N}{2\pi L_B}\right) \\ & + (k_B L_B L+2)\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{I}}}\right) + 2k_B\zeta\ln p_B.\end{aligned} \tag{20}$$

In practice, we compute the GEBIC$_{\text{R}}$ score block-wise, i.e., GEBIC$_{\text{R}}(\mathcal{I}_B)$ where $\mathcal{I}_B\subseteq\{1,\ldots,p_B\}$. Then the $\hat{\sigma}^2_{\mathcal{I}}$ can be replaced by $\hat{\sigma}^2_{\mathcal{I}_B} = \|\boldsymbol{\Pi}^{\perp}(\mathbf{A}_{\mathcal{I}_B})\mathbf{Y}\|^2_F/NL$. Finally, the true block support is estimated as

$$\hat{\mathcal{S}}_B = \arg\min_{\mathcal{I}_B}\{\text{GEBIC}_{\text{R}}(\mathcal{I}_B)\}. \tag{21}$$

Observe that for $L=L_B=1$, GEBIC$_{\text{R}}$ boils down to EBIC$_{\text{R}}$ [4]. Thus GEBIC$_{\text{R}}$ can cater to all forms of linear regression scenarios.

IV. Simulation Results

In this section, we provide numerical simulations to highlight the performance of GEBIC$_{\text{R}}$ for model selection in BMMV models. We consider the linear model $\mathbf{Y}=\mathbf{AX}+\mathbf{W}$, where the design matrix $\mathbf{A}$ is generated with independent entries following normal distribution $\mathcal{N}(0,1)$. The cardinality of the true block-support $\mathcal{S}_B$ is chosen to be $K_B=4$. Also,

Algorithm 1 BMMV-OMP with K iterations

Inputs: Design matrix $\mathbf{A}$, measurement $\mathbf{Y}$.
Initialization: $\|\mathbf{a}_j\|_2 = 1\ \forall j$, $\mathbf{R}^0 = \mathbf{Y}$, $\mathcal{S}^0_{\text{B-OMP}} = \emptyset$
for $i = 1$ to K **do**
  Next block index: $d^i = \underset{j=1,\ldots,p_B}{\arg\max} \left\|\mathbf{A}[:, \mathcal{I}_j]^T \mathbf{R}^{i-1}\right\|_F$
  Add current index: $\mathcal{S}^i_{\text{B-OMP}} = \mathcal{S}^{i-1}_{\text{B-OMP}} \cup \{d^i\}$
  Update residual: $\mathbf{R}^i = \mathbf{\Pi}^{\perp}(\mathbf{A}_{\mathcal{S}^i_{\text{B-OMP}}})\mathbf{Y}$
end for
Output: B-OMP generated block index sequence $\mathcal{S}^K_{\text{B-OMP}}$

Algorithm 2 Model selection GEBIC$_{\text{R}}$ with B-OMP

Run B-OMP for K iterations to obtain $\mathcal{S}^K_{\text{B-OMP}}$
for $k_B = 1$ to K **do**
  $\mathcal{I}_B = \mathcal{S}^{k_B}_{\text{B-OMP}}$
  Compute GEBIC$_{\text{R}}(\mathcal{I}_B)$
end for
Block support estimate: $\hat{\mathcal{S}}_B = \underset{\mathcal{I}_B}{\arg\min}\{\text{GEBIC}_{\text{R}}(\mathcal{I}_B)\}$

without loss of generality, we assume $\mathcal{S}_B = [1, 2, 3, 4]$. The non-zero entries in $\mathbf{X}$ are randomly assigned ± 1. The SNR in dB = $10\log_{10}(\sigma_s^2/\sigma^2)$, where σ_s^2 and σ^2 denote signal and true noise power, respectively. The signal power is computed as $\sigma_s^2 = \|\mathbf{AX}\|_F^2/NL$. The chosen SNR (dB) and σ_s^2 are then used to determine the noise power as $\sigma^2 = \sigma_s^2/10^{\text{SNR (dB)}/10}$. Using this σ^2, the elements of the noise matrix $\mathbf{W}$ are generated following $\mathbf{W}[i,j] \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. The probability of correct model selection (PCMS), i.e., $\Pr(\hat{\mathcal{S}}_B = \mathcal{S}_B)$ is evaluated over 1000 Monte Carlo trials. At each Monte Carlo trial, a new design matrix $\mathbf{A}$ is generated in order to preserve the randomness in the data. For predictor/subset selection, BMMV-OMP (B-OMP) [27], [39] (Algorithm 1) is utilized because of its ease of use and broad application. BMMV-OMP (B-OMP) [27], [39] (Algorithm 1) is used for predictor/subset selection for its simplicity and wide range of applicability. The performance of GEBIC$_{\text{R}}$ is compared with GRRT and the oracle, which is B-OMP with *a-priori* knowledge of the block sparsity K_B. Hence, the oracle provides the upper bound on the maximum achievable PCMS for any given setting. The tuning parameters chosen are $\alpha = 0.01$ for GRRT (as mentioned in [27]) and $\zeta = 1$ (EBIC$_{\text{R}}$) [4].

Fig. 1 shows the PCMS vs SNR (dB) with $N = 150$ and $p = 1000$. Since $L_B = 10$, hence, $p_B = p/L_B = 100$. Additionally, the performance is shown for two different settings of the L parameter, viz. $L = 5$ and 15 to highlight the influence of L on the overall behaviour of the methods. The first clear observation is that for the considered tuning parameter setting, both GEBIC$_{\text{R}}$ and GRRT are empirically consistent in high-SNR, i.e., PCMS $\to 1$ as SNR $\to \infty$ (or inversely $\sigma^2 \to 0$). Second, compared to GRRT, the performance curve of GEBIC$_{\text{R}}$ is much closer to the oracle, especially for low values of SNR. Furthermore, compared to $L = 5$, the oracle plot shifts toward the left when $L = 15$. This indicates that increasing L improves the true support recovery ability of B-OMP, which ultimately improves the model selection performance of the methods.

Fig. 2 presents the PCMS vs N plot. Here, a fixed value of $p = 5000$ is chosen. Additionally, the performance is shown for two different values of the L_B variable, viz. $L_B = 5$ and 20 to highlight the impact of L_B on the overall model selection performance. A similar trend is observed here as well. Both methods achieve empirical consistency (PCMS $\to 1$) as N grows large. However, GEBIC$_{\text{R}}$ provides slightly better performance compared to GRRT for smaller N values, and is much closer to the oracle performance. Furthermore, we also observe that increasing L_B lowers the support recovery performance of B-OMP, which is evident from the shift in the oracle performance towards the right. Thus, requiring more samples to achieve the same PCMS for $L_B = 20$ as compared to $L_B = 5$. This ultimately lowers the overall performance of all model selection methods. Hence, we can say that increasing L_B affects the performance negatively.

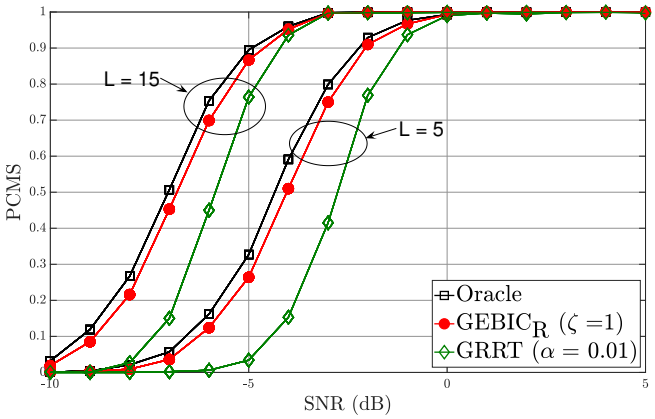

Fig. 1. PCMS vs SNR (dB) for $N = 150$, $p = 1000$, $L = [5, 15]$, $L_B = 10$ and $K_B = 4$.

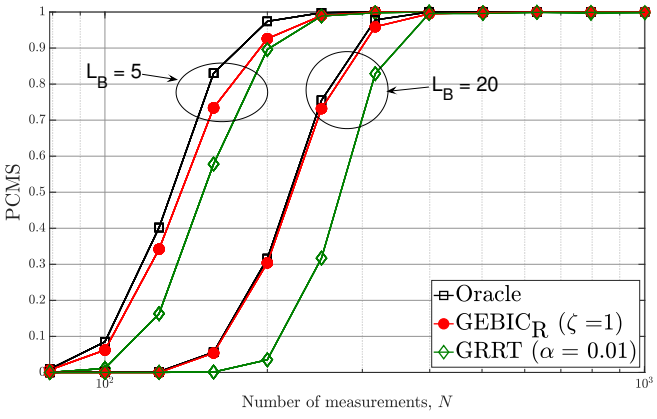

Fig. 2. PCMS vs N for SNR = -4 dB, $p = 5000$, $L = 5$, $L_B = [5, 20]$ and $K_B = 4$.

V. Conclusion

In this paper, we presented GEBIC_R which is a generalized version of EBIC_R to handle model selection in the block-sparse HD general linear regression. GEBIC_R is applicable to all forms of the linear regression structure such as SMV, BSMV, MMV, and BMMV thus making it a versatile IC. The steps to arrive at the criterion are shown in detail. Simulation results show that GEBIC_R is an empirically consistent criterion as $N \to \infty$ and/or SNR $\to \infty$. Also, its performance for lower SNR and N values is close to the oracle behaviour. Furthermore, we also underline the manner in which the parameters L and the block length L_B affect the overall model selection performance.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[2] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[3] A. Owrang and M. Jansson, "A model selection criterion for high-dimensional linear regression," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3436–3446, 2018.

[4] P. B. Gohain and M. Jansson, "New improved criterion for model selection in sparse high-dimensional linear regression models," in *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5692–5696.

[5] ——, "Robust information criterion for model selection in sparse high-dimensional linear regression models," *IEEE Transactions on Signal Processing*, pp. 1–16, 2023.

[6] S. Kallummil and S. Kalyani, "Signal and noise statistics oblivious orthogonal matching pursuit," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2429–2438.

[7] P. B. Gohain and M. Jansson, "Relative cost based model selection for sparse high-dimensional linear regression models," in *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5515–5519.

[8] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani, "A significance test for the lasso," *Annals of statistics*, vol. 42, no. 2, p. 413, 2014.

[9] R. F. Barber and E. J. Candès, "A knockoff filter for high-dimensional selective inference," *The Annals of Statistics*, vol. 47, no. 5, pp. 2504–2537, 2019.

[10] J. Shao, "Linear model selection by cross-validation," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

[11] R. R. Picard and R. D. Cook, "Cross-validation of regression models," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 575–583, 1984.

[12] L. de Torrenté and T. Hastie, "Does cross-validation work when $p \gg n$?" 2012.

[13] M. Chichignoud, J. Lederer, and M. J. Wainwright, "A practical scheme and fast algorithm to tune the lasso with optimality guarantees," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8162–8181, 2016.

[14] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[15] Y. C. Eldar and M. Mishali, "Block sparsity and sampling over a union of subspaces," in *2009 16th International Conference on Digital Signal Processing*. IEEE, 2009, pp. 1–8.

[16] ——, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[17] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[18] M. Mishali and Y. C. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Transactions on signal processing*, vol. 57, no. 3, pp. 993–1009, 2009.

[19] ——, "From theory to practice: Sub-nyquist sampling of sparse wideband analog signals," *IEEE Journal of selected topics in signal processing*, vol. 4, no. 2, pp. 375–391, 2010.

[20] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 275–285, 2008.

[21] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[22] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Transactions on Signal processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[23] A. Owrang and M. Jansson, "Weighted covariance matching based square root lasso," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 3751–3755.

[24] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2009.

[25] M. Mishali and Y. C. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Transactions on signal processing*, vol. 57, no. 3, pp. 993–1009, 2009.

[26] K. Gedalyahu and Y. C. Eldar, "Time-delay estimation from low-rate samples: A union of subspaces approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3017–3031, 2010.

[27] S. Kallummil and S. Kalyani, "Generalized residual ratio thresholding," *Signal Processing*, vol. 197, p. 108531, 2022.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[29] J. W. Choi and B. Shim, "Detection of large-scale wireless systems via sparse error recovery," *IEEE Transactions on Signal Processing*, vol. 65, no. 22, pp. 6038–6052, 2017.

[30] S. Aviyente, "Compressed sensing framework for eeg compression," in *2007 IEEE/SP 14th workshop on statistical signal processing*. IEEE, 2007, pp. 181–184.

[31] T. Kronvall, S. I. Adalbjörnsson, S. Nadig, and A. Jakobsson, "Group-sparse regression using the covariance fitting criterion," *Signal Processing*, vol. 139, pp. 116–130, 2017.

[32] I. Fedorov, R. Giri, B. D. Rao, and T. Q. Nguyen, "Robust Bayesian method for simultaneous block sparse signal recovery with applications to face recognition," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3872–3876.

[33] P. B. Gohain and M. Jansson, "Robust information criterion for model selection in sparse high-dimensional linear regression models," *arXiv preprint arXiv:2206.08731*, 2022.

[34] P. Stoica and P. Babu, "On the proper forms of BIC for model order selection," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4956–4961, 2012.

[35] D. F. Schmidt and E. Makalic, "The consistency of MDL for linear regression models with increasing signal-to-noise ratio," *IEEE transactions on signal processing*, vol. 60, no. 3, pp. 1508–1510, 2011.

[36] P. B. Gohain and M. Jansson, "Scale-invariant and consistent Bayesian information criterion for order selection in linear regression models," *Signal Processing*, p. 108499, 2022.

[37] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2726–2735, 1998.

[38] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice Hall PTR, 1993.

[39] Y. Shi, L. Wang, and R. Luo, "Sparse recovery with block multiple measurement vectors algorithm," *IEEE Access*, vol. 7, pp. 9470–9475, 2019.