Mallows criterion for heteroskedastic linear regressions with many regressors

Stanislav Anatolyev*

CERGE-EI (Czechia) and New Economic School (Russia)

April 2021

Abstract

We present a feasible generalized Mallows criterion for model selection for a linear regression setup with conditional heteroskedasticity and possibly numerous explanatory variables. The feasible version exploits unbiased individual variance estimates from recent literature. The property of asymptotic optimality of the feasible criterion is shown. A simulation experiment shows large discrepancies between model selection outcomes and those yielded by the classical Mallows criterion or other available alternatives.

Keywords: model selection, Mallows criterion, conditional heteroskedasticity, many regressors.

*Address: Stanislav Anatolyev, CERGE-EI, Politickych vězňů 7, 11121 Prague 1, Czech Republic; e-mail stanislav.anatolyev@cerge-ei.cz. I thank an anonymous expert referee for helpful suggestions. This research was supported by the grant 20-28055S from the Czech Science Foundation.

1 Introduction and setup

The Mallows criterion (Mallows, 1973) is a powerful tool of model selection and averaging for linear regressions. Originally developed for homoskedastic regressions, it is proven to have favorable properties in suitable setups, such as asymptotic optimality (Li, 1987). It is often applied bluntly in more general settings; however, theoretically correct adaptation to more modern regression setups, such as those containing conditional heteroskedasticity and at the same time allowing use of extended regressor sets, is highly desirable.

Adaptation of the Mallows criterion to conditionally heteroskedastic regressions was made by Andrews (1991) who also showed asymptotic optimality of the infeasible version. Liu and Okui (2013) operationalize the generalized Mallows criterion in a model averaging context via a clever use of the Eicker-White asymptotic variance formula and the weighted average structure of the criterion. However, if one is allowed to utilize extended sets of regressors, whose number may be comparable to the number of observations, the Eicker-White formula may unfortunately fail to correctly estimate asymptotic variances, as was recently shown in Cattaneo, Jansson, and Newey (2018a).

We propose alternative implementation of the generalized Mallows criterion by using estimates of individual error variances in linear models from Kline, Saggio and Sølvsten (2020) and Jochmans (2021). These allow exploiting regressor sets whose numerosity is comparable to sample sizes. We show that the resulting feasible generalized Mallows criterion keeps the asymptotic optimality property. We also verify how the feasible criterion performs in prediction terms in model selection experiments when many regressors are allowed, and compare outcomes across alternative variance estimates.

Consider a mean regression in an IID environment:

$$y_i = g_i + e_i,$$

where, for the i^{th} unit, $g_i = E\left[y_i|x_i\right]$ is conditional mean given the vector x_i of basic regressors, e_i is regression error with conditional variance $\sigma_i^2 \equiv E\left[y_i|x_i\right]$, $i = 1, ..., n$, and the sample $\{(x_i, y_i)\}_{i=1}^n$ is random. The regression function is approximated by a sequence of linear regression models

$$y_i = z_i\left(q\right)'\beta\left(q\right) + u_i$$

estimated by ordinary least squares (OLS)

$$\hat{\beta}(q) = \left(\sum_{i=1}^{n} z_i(q) z_i(q)' \right)^{-1} \sum_{i=1}^{n} z_i(q) y_i$$

so that the fitted values are $\hat{g}_i(q) = z_i(q)' \hat{\beta}(q) = (P(q) y)_i$. Here, $q \in \{1, ..., \bar{q}\} \equiv Q$ is the model index, model $\bar{q}$ being the largest, $z_i(q)$ is a respective $\dim(z_i(q)) \times 1$ vector of regressors that are functions of the basic regressors x_i, y is an $n \times 1$ vector of all response variables, and $P(q)$ is an associated projection matrix. We assume for simplicity that, as $n \to \infty$, the number of models under consideration stays finite, but the precision of the largest model increases. The dimensionality of regressor sets is restricted only by the sample size $\dim(z_i(q)) < n$ and by full rank of the corresponding matrices $\sum_{i=1}^{n} z_i(q) z_i(q)'$ for all $q \in Q$. Our analysis is conditional on all basic regressors $\{x_i\}_{i=1}^{n}$.

2 Classical and generalized Mallows criteria

When the regression model is homoskedastic, i.e. when $\sigma_i^2 = \sigma^2$ for all $i = 1, ..., n$, the *classical Mallows criterion* to be minimized with respect to q is defined as

$$C_p(q) = \hat{e}(q)' \hat{e}(q) + 2 \dim(z_i(q)) \hat{\sigma}^2, \tag{1}$$

where $\hat{e}(q)$ is an $n \times 1$ vector of OLS residuals for model q, and $\hat{\sigma}^2$ is the OLS residual variance from the model with largest $\dim(z_i(q))$. The motivation of the criterion (1) is that its expected value minimizes the prediction risk defined as expected squared deviation of fitted values from the true regression values:

$$R(q) = \mathbb{E}\left[(\hat{g}(q) - g)' (\hat{g}(q) - g) \right],$$

where g and $\hat{g}(q)$ are $n \times 1$ vectors of true regression values and fitted values for model q, respectively. Indeed, $\hat{g}(q) - g = -M(q) g + P(q) e$, where $M(q) = I_n - P(q)$ is the orthogonal to $P(q)$ projection matrix. Hence, $R(q) = g' M(q) g + \dim(z_i(q)) \sigma^2$, while for the infeasible version of $C_p(q)$, call it $C_p^0(q)$, that uses σ^2 in place of $\hat{\sigma}^2$, one obtains

$$\mathbb{E}\left[C_p^0(q) \right] = \mathbb{E}\left[(g + e - \hat{g}(q))' (g + e - \hat{g}(q)) \right] + 2 \dim(z_i(q)) \sigma^2 = R(q) + \text{const}.$$

Thus, minimization of the infeasible version of (1) minimized the prediction risk. The unknown σ^2 is replaced by the OLS residual variance from the largest model to minimize the impact of the misspecification error.

Andrews (1991) generalizes the infeasible version of (1) for heteroskedastic models and names it the *generalized Mallows criterion*. This criterion reads

$$GC_p^0(q) = \hat{e}(q)' \hat{e}(q) + 2\sum_{i=1}^{n} P_{ii}(q)\,\sigma_i^2, \tag{2}$$

where $P_{ii}(q)$ is the i^{th} diagonal entry of $P(q)$, also called a leverage for i^{th} observation, in model q. The motivation of (2) is the same: the choice of q that minimizes $GC_p^0(q)$ also minimizes the prediction risk defined analogously; both criteria have the same expressions up to the replacement of $\dim(z_i(q))\,\sigma^2$ by $\sum_{i=1}^{n} P_{ii}(q)\,\sigma_i^2$. This results in prediction risk

$$R(q) = g'M(q)\,g + \sum_{i=1}^{n} P_{ii}(q)\,\sigma_i^2. \tag{3}$$

Andrews (1991) shows that under suitable conditions, when the errors are heteroskedastic, the $GC_p^0(q)$ criterion is asymptotically, as $n \to \infty$, optimal in the sense that the prediction risk of the regression model selected by the GC_p^0 procedure is close to the minimal prediction risk in large samples.

Andrews (1991) notes that the feasible version of $GC_p^0(q)$ in a heteroskedastic context is generally not available because of unknown individual error variances σ_i^2 for all i in the penalty term. However, recent literature suggests availability of estimates of individual error variances that are in addition robust to presence of many regressors. Then, it is possible to construct a feasible version of (2), say,

$$GC_p(q) = \hat{e}(q)' \hat{e}(q) + 2\sum_{i=1}^{n} P_{ii}(q)\,\hat{\sigma}_i^2, \tag{4}$$

where $\hat{\sigma}_i^2$ are (approximately) unbiased estimates of error variances σ_i^2, $i = 1, ..., n$, such that $n^{-1}GC_p(q)$ differs from $n^{-1}GC_p^0(q)$ by no more than $o_p(1)$, and so the feasible generalized Mallows criterion also retains the asymptotic optimality property.

Let M_{ii} denote the i^{th} diagonal entry of $M(q)$ for some model q. We exploit the following estimator of i^{th} error variance:

$$\hat{\sigma}_i^2 = \frac{y_i \hat{e}_i}{M_{ii}}, \tag{5}$$

which originates from leave-one-out OLS estimation and well-known equality between OLS residuals $\hat{e}_i$ corrected for leverage and leave-one-out OLS residuals. The estimates (5) were proposed by Kline, Saggio and Sølvsten (2020) for estimation of quadratic forms in parameters of correctly specified linear regressions, and by Jochmans (2021) for estimation of asymptotic variances in presence of many covariates in possibly misspecified linear regressions. This estimator is unbiased when the regression model is correctly specified:

$$E\left[\hat{\sigma}_i^2\right] = \frac{1}{M_{ii}} E\left[(g_i + e_i)\sum_{j=1}^{n} M_{ij}\,(g_j + e_j)\right] = \frac{g_i\,(Mg)_i}{M_{ii}} + \sigma_i^2,$$

which is exactly σ_i^2 when g is in the span of $z_i\,(q)$'s. When model q does not nest the true regression, there is bias from the first term. To reduce this bias, we employ error variance estimates from the largest model $\bar{q}$ akin to how it is done under homoskedasticity. The use of these error variance estimates allows the approximating models to be large in the sense that $\dim\left(z_i\,(q)\right)$ for some or even all q may be comparable to n.

An attractive modification of (5) is[1]

$$\hat{\sigma}_i^2 = \frac{(y_i - \bar{y}_n)\,\hat{e}_i}{M_{ii}}, \tag{6}$$

where by $\bar{a}_n$ we denote the sample mean of a variable a_i. In (6), demeaning of outcome variables corrects (5) for a scale effect, but introduces a bias, which, however, when aggregated in (4), has a negligible effect compared to the leading terms.

3 Asymptotic optimality

We impose the following conditions, which include and expand those imposed in Andrews (1991) needed to show asymptotic optimality of the infeasible criterion $GC_p^0\,(q)$.

Assumption 1 *There are finite positive constants $\underline{C}_\sigma$, $\overline{C}_\sigma$, $\overline{C}_\kappa$, $\underline{C}_M$, $\underline{C}_R$ not depending on n, such that*

(i) $\underline{C}_\sigma < \min_{1\leq i\leq n}\sigma_i^2 < \max_{1\leq i\leq n}\sigma_i^2 \leq \overline{C}_\sigma$, $\max_{1\leq i\leq n} E\left[e_i^4\right] \leq \overline{C}_\kappa$ *and* $E\left[g_i^2\right] < \infty$;

(ii) $\underline{C}_M \leq \min_{1\leq i\leq n} M_{ii}\,(\bar{q})$;

(iii) $g'M\,(\bar{q})\,g \leq O_P(n^{\delta_{Mg}})$ *for* $\delta_{Mg} < 1$, *and* $\min_{q\in Q}\left\{g'M\,(q)\,g + \dim(z_i(q))\right\} > \underline{C}_R n$.

[1] I thank Mikkel Sølvsten for suggesting this descaled version.

Assumption 1(i) restricts heteroskedasticity and kurtosity of the error term, as well as imposes integrability of the conditional mean; Assumption 1(ii) restricts leverages. Assumption 1(iii) makes sure that the precision of the largest model asymptotically increases, and that any model under consideration is either not precise enough, or has many enough regressors. Importantly, we do not make assumptions about asymptotically vanishing diagonal elements of the projection matrices $P(q)$, which preclude the numerosity of regressors to be comparable to the sample size. In the literature, such assumptions are typically variations of the condition $\max_{q\in Q}\max_{1\leq i\leq n} P_{ii}(q) \xrightarrow{P} 0$.

The following theorem proved in online Appendix[2] is an analog of Andrews's (1991) optimality result for the feasible version of the generalized Mallows criterion.

Theorem 1. Let Assumption 1 hold. The feasible generalized Mallows criterion (4) with individual variances estimates in (5) or (6) that use the OLS residuals from the largest model is asymptotically optimal:

$$\frac{R(\hat{q})}{\min_{q\in Q} R(q)} \xrightarrow{P} 1$$

as $n\to\infty$, where $\hat{q}\equiv \arg\min_{q\in Q} GC_p(q)$.

4 Alternative variance estimates

Liu and Okui (2013) operationalize the generalized Mallows criterion in a model averaging context via the identity

$$\sum_{i=1}^{n} P_{ii}\sigma_i^2 = \operatorname{tr}\left(\Big(\sum_{i=1}^{n} z_i z_i'\Big)^{-1} \sum_{i=1}^{n} z_i z_i' \sigma_i^2 \right). \tag{7}$$

The second matrix under the trace can be handled via Eicker-White heteroskedasticity consistent estimation or various "almost unbiased" improvements thereof, referred to in the literature as HCK (see MacKinnon, 2012, for a review):

$$\sum_{i=1}^{n} \mu_i z_i z_i' \hat{e}_i^2, \tag{8}$$

[2]Available at `is.gd/MallowsM`.

where $\mu_i = 1$ for the Eicker-White estimator sometimes referred to as 'HC0', $\mu_i = n/\left(n - \dim(z_i(\bar{q}))\right)$ for the modification known as 'HC1', $\mu_i = 1/M_{ii}$ for the modification known as 'HC2', and $\mu_i = 1/M_{ii}^2$ for the modification known as 'HC3' (MacKinnon, 2012). However, Cattaneo, Jansson, and Newey (2018a) show that all these variants yield inconsistent results when there are many explanatory variables.

Also, Cattaneo, Jansson, and Newey (2018a) propose an alternative estimator of the individual variances in linear models, which is robust to regressor numerosity. This estimator restricts the number of regressors to be at most half of the sample size. In the simulation study of Section 5, where this property does hold, using these estimates yields comparable yet a bit less convincing results. Thus, we stick to the variance estimates (6).

5 Simulation evidence

The data generating process is inspired by simulations in Cattaneo, Jansson, and Newey (2018b). The true regression function is $g_i = \exp\left(\|x_i\|^2\right)$, where the basic regressor vector x_i contains $d = 5$ independent standard uniform. There are $\bar{q} = 10$ models corresponding to the following composition of regressors z_i. shown in Table 1. Beyond the two minimal models, new sets of regressors first become next-order powers of the basic regressors, and then also all interactions of the same order. For example, $z_i(3)$ contains 1, x_i and $x_i \odot x_i$, while $z_i(4)$ contains 1, x_i and 15 distinct elements of $x_i x_i'$. The sample size is $n = 800$.

We insert the diagonal elements of the projection matrix P^x associated with the basic regressors x_i directly into the skedastic function: $\sigma_i = nP_{ii}^x$. This induces correlation between P_{ii}^x and σ_i^2 of about 0.96, and correlations between $P_{ii}(q)$ and $\hat{\sigma}_i^2$ in the range $0.22 \div 0.26$ depending on $q \in Q\backslash\{1\}$. We also examine the homoskedastic case $\sigma_i = \sum_{i=1}^n P_{ii}^x = \dim(x_i) = 5$. The error term is generated as $e_i = \sigma_i u_i$, where u_i, $i = 1, ..., n$, are independent standard normal. The signal-to-noise ratio is $var(g_i)/var(e_i) \approx 1.2$.

We compute averages, across 10,000 simulations, of the following two measures: one is the normalized exact value of expected prediction risk computed from (3) for the selected model $\hat{q}$, and the other is the mean squared prediction error computed as $\text{MSPE}(\hat{q}) = n^{*-1} \sum_{i=1}^{n^*} \left(g_i^* - z_i^*(\hat{q})' \hat{\beta}(\hat{q})\right)^2$, where g_i^* is generated according to the same data generating

Table 1: Composition of regressors

model q	regressors $z_i(q)$	$\dim(z_i(q))$
1	1	1
2	$1, x_i$	6
3	$z_i(2), x_i^{\odot 2}$	11
4	$z_i(2), x_i^{\otimes 2}$	21
5	$z_i(4), x_i^{\odot 3}$	26
6	$z_i(4), x_i^{\otimes 3}$	56
7	$z_i(6), x_i^{\odot 4}$	61
8	$z_i(6), x_i^{\otimes 4}$	126
9	$z_i(8), x_i^{\odot 5}$	131
10	$z_i(8), x_i^{\otimes 5}$	252

Notes: $a^{\odot k}$ denotes a list of length $\dim(a)$ of k^{th} powers of all elements of a, and $a^{\otimes k}$ denotes a list of all k^{th} cross-powers of all elements of a.

process, $z_i^*(q)$ is formed as in model q, and $n^* = 50,000$ is the number of pseudo-out-of-sample observations.

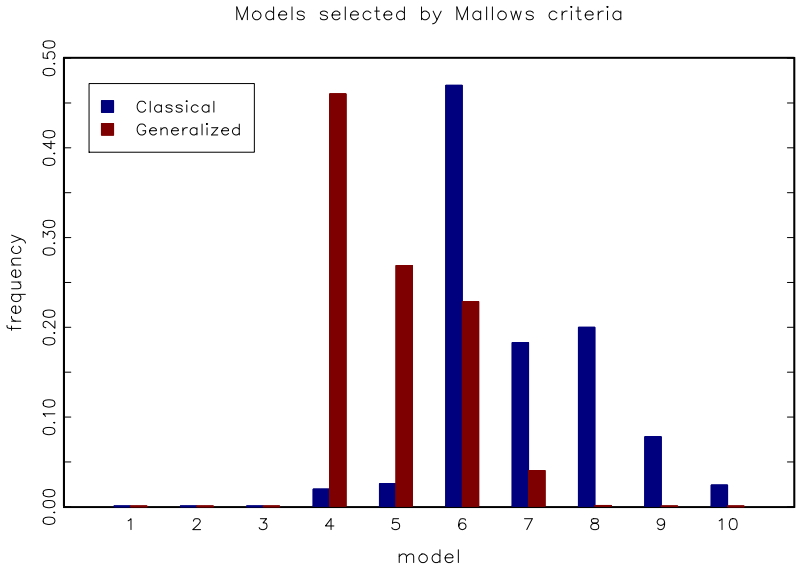
Table 2 contains the values of the C_p and GC_p criteria for four experiments. In the heteroskedastic case, both the predictive risk and mean squared prediction error are larger for the classical criterion by more than 50% and more than 80%, respectively. At the same time, in the homoskedastic setup, the finite-sample distortions because of more complex estimation of individual error variances are meager.

Figure 1 presents a histogram of models selected by the two criteria in the heteroskedastic case. One can clearly see that the GC_p criterion selected a more parsimonious model, with a peak on model 4 with only second powers of the basic regressors included and no bigger than model 7 that allows only own 4^{th} powers. In contrast, the models selected by C_p are much bigger on average ranging between model 4 and model 10 thus sometimes employing even all 5^{th} powers, and almost half of the time prefers model 6 with all 3^{rd} powers of the basic regressors included. As Table 2 indicates, including these higher

Table 2: Values of predictive measures

	heteroskedastic design		homoskedastic design	
measure	classical C_p	generalized GC_p	classical C_p	generalized GC_p
$n^{-1}R(\hat{q})$	4.41	2.85	2.07	2.11
MSPE $(\hat{q})$	6.18	3.37	2.54	2.58

Figure 1: Models selected by the two Mallows criteria in the simulation experiment



powers may harm the predictive power in heteroskedastic environments enormously.

Table 3 with self-explaining labels contains the out-of-sample criteria values for the heteroskedastic experiment, with alternative variance estimates from Section 4.[3] The CJN variance estimates lead to a bit less attractive performance than the modified KSSJ variance estimates (6) do. The modal selected model is no. 4 in both cases, with a bit more right-skewed distribution for CJN. The variance estimates based on the Eicker-White formula, in accordance to theory, lead to worse performance of the generalized criterion in model selection, with HC3 yielding highest improvements. The modal selected model is no. 6 with all HCK variations, the same as what the classical criterion yields.

[3]We omit the results for the non-modified HC0 estimates; these are terrible.

Table 3: Values of predictive measures under different variance estimates (GC_p-KSSJ copied from Table 2), heteroskedastic design

measure	GC_p-KSSJ	GC_p-CJN	GC_p-HC1	GC_p-HC2	GC_p-HC3
$n^{-1}R(\hat{q})$	2.85	3.04	4.32	4.20	3.25
MSPE$(\hat{q})$	3.37	3.65	5.92	5.47	3.80

6 Concluding remarks

In modern, conditionally heteroskedastic, many-regressor linear regression setups, one needs to use correct tools of model selection, those which are compatible with heteroskedasticity and the presence of numerous explanatory variables. The presented feasible form of the generalized Mallows criterion should be used in such setups, for both model selection and model averaging.

We should stress that the method outlined here may be valid only for linear models. In nonlinear setups, conventional estimation in the presence of many regressors leads to severe biases (see, e.g., Cattaneo, Jansson, and Ma, 2018; Sur and Candés, 2019), and so requires additional adjustments. This is a promising avenue for future research.

References

Andrews, D.W.K. (1991). Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 47, 359-377.

Cattaneo, M., M. Jansson, and X. Ma (2018) Two-step estimation and inference with possibly many included covariates. *Review of Economic Studies*, 86(3), 1095-1122.

Cattaneo, M., M. Jansson, and W.K. Newey (2018a) Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523), 1350-1361.

Cattaneo, M., M. Jansson, and W.K. Newey (2018b). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34(2), 277-301.

Jochmans, K. (2021). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, forthcoming.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica*, 88(5), 1859-1898.

Li, K.-C. (1987) Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: Discrete index set. *Annals of Statisttcs*, 15, 958-975.

Liu, Q. and R. Okui (2013). Heteroscedasticity-robust C_p model averaging. *Econometrics Journal*, 16, 463-472.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, 15, 661-675.

MacKinnon, J. G. (2012). "Thirty years of heteroskedasticity-robust inference." In: Chen, X. and N. R. Swanson, eds., *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, 437-461, Springer, New York.

Sur, P. and E.J. Candés (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *PNAS*, 116(29), 14516-14525.