AIC for many-regressor heteroskedastic regressions

Stanislav Anatolyev*

CERGE-EI and New Economic School

May 2026

Abstract

The original and corrected Akaike information criteria (AIC) have been routinely used for model selection for ages. The penalty terms in these criteria are tied to the classical normal linear regression, characterized by conditional homoskedasticity and a small number of regressors relative to the sample size. We derive, from the same principles, a general version that takes account of conditional heteroskedasticity and regressor numerosity. The new AICm penalty takes a form of a ratio of certain weighted average error variances. The classical results on asymptotic prediction optimality of AIC are extended to AICm in heteroskedastic possibly misspecified models. Further, the infeasible AICm penalty can be operationalized via unbiased estimation of individual variances. Under suitable conditions, the feasible AICm criterion differs from the infeasible counterpart up to an asymptotically negligible term, and also possesses the property of asymptotic optimality. In simulations, the feasible AICm tends to select models that deliver systematically better out-of-sample predictions than the classical criteria. In the empirical application to an impact of legalized abortion on crime reduction, AICm tends to select a more parsimonious model than non-heteroskedasticity-robust versions of AIC.

Keywords: linear regression, model selection, Akaike information criterion, many regressors, conditional heteroskedasticity, asymptotic optimality.

*Address: Stanislav Anatolyev, Politickych vězňů 7, 11121 Prague 1, Czech Republic; e-mail stanislav.anatolyev@cerge-ei.cz. This research was supported by the grant 24-12720S from the Czech Science Foundation.

...use AICc unless a better form is known.

Burnham and Anderson (2002, p.327)

1 Introduction

The celebrated Akaike information criterion (AIC) (Akaike 1973, 1974) has been routinely used for model selection in applied work for ages. While AIC has been originally derived for a classical normal linear regression, it is used in many settings beyond the classical one as an omnibus model selection tool. AIC has even entered the toolboxes of model averaging (Buckland, Burnham, and Augustin 1997) and smoothing parameter selection in nonparametric estimation (Hurvich, Simonoff, and Tsai 1998). The AIC penalty $2m$ or normalized penalty $2m/n$, where m is a number of regressors and n is a sample size, is very attractive in its simplicity and implementability (in particular, cleanness from unknowns). These properties are natural consequences of the tight assumptions used in derivation of the penalty as a difference of expected out-of-sample and estimated in-sample loglikelihoods that are ingredients of the Kullback-Leibler divergence (KLIC).

In a linear regression context, there have been subsequent improvements of AIC, most renowned being the small-sample corrected AIC (AICc) (Sugiura 1978, Hurvich and Tsai 1989) that modifies the AIC penalty (making it heavier) to reduce the bias that gets stronger with parameter numerosity. AICc shows remarkable improvement in simulations and applied problems relative to the original AIC. Burnham and Anderson (2002) strongly advocate the use of AICc in situations when the sample size is small or the parameters are many, supporting this advice by numerical evidence from simulation and real data examples. The AICc penalty formula is a bit more involved than AIC's but still depends only on m and n. Burnham and Anderson (2002) also show that its derivation in different circumstances, for example, for grouped data or non-normal distributions, will result in different versions of the AICc penalty.

We present a general version of the formula for AIC that explicitly acknowledges possible numerosity of regressors. Our derivation does not rely on homoskedastic normality of the conditional distribution; in particular, conditional heteroskedasticity in the true DGP is allowed. The presumption is that the researcher utilizes the Gaussian quasi-maximum likelihood estimation for a linear mean regression model, but the model may (or may not) contain a sizable number of regressors, m, compared to the number of observations, n, and may be conditionally heteroskedastic. Of course, as any AIC, our generalization

is derived under the assumption that the regression model under consideration[1] is correctly specified. We abbreviate the proposed criterion with 'AICm,' where 'm' stands for 'many [regressors/covariates/explanatory variables]'. We derive the AICm penalty based on the same paradigm the original AIC is derived – as (twice) the gap between the expected estimated normal quasi-likelihood fit to out-of-sample data and the expected normal quasi-likelihood fit to in-sample data. The AICm penalty encompasses the celebrated classical ones: it is approximately equal to the AIC penalty when the regression is conditionally homoskedastic and regressors are few, and to the AICc penalty when the regression is conditionally homoskedastic but the number of regressors is not negligible.

Because the AICm penalty is derived under an implausible presumption of a correctly specified model, we also investigate the properties of AICm in realistic setups, where linear models are misspecified, being approximations to or underspecified versions of the true regression. In particular, we extend the results on asymptotic prediction optimality of AIC (Shao 1997) to AICm in heteroskedastic models. We also analyze the heteroskedasticity non-robust AIC versions, and discover that they may also exhibit asymptotic optimality in favorable, though restrictive, circumstances, but otherwise violate this property. The results reveal that the main driver of optimality is the "correct" asymptotic order of an AIC penalty, this order being the same for AICm and AICc but different and often completely off for the original AIC. On the other hand, compatibility with the heteroskedasticity pattern is important in general, but may not play a role with ideal, though implausible, regressor designs. As a by-product of the analysis of asymptotic optimality, we discover some tendencies how AICm-selection works for models of different types of parameter numerosity – those that we call moderately large (when m, though high, is asymptotically negligible compared to n), and those that we call really large (when m is comparable to n), with the amount of misspecification also playing a role.

Next, because the AICm penalty is unknown, AICm is not immediately implementable, in contrast to original and corrected AIC. It is not a function of m and n only, but in addition depends on the pattern of conditional heteroskedasticity in the sample. More precisely, it is equal to a ratio of certain weighted average error variances. Therefore, we also propose a feasible version, abbreviated as AICm^{F}, that replaces unknown individual error variances by their unbiased estimates based on the leave-one-out idea (Kline, Saggio and Sølvsten 2020, Jochmans 2022). Under suitable conditions, the AICm^{F} penalty differs

[1]That is, concerning only the *form* of the regression function and *composition* of regressors, not the whole conditional *distribution.*

from the ideal AICm penalty by an asymptotically negligible term, and hence, feasible AICm is asymptotically equivalent to ideal AICm. Moreover, under suitable conditions, feasible AICm retains the property of asymptotic prediction optimality. Implementation of AICm^{F} is straightforward and simple, and computation is fast.

Simulations in a stylized heteroskedastic setup, employing correctly specified and underspecified and misspecified models, show that the use of AICm tends to select better predicting models, both when the true regression is contained in the model set and when it is not, both when the regression errors are truly Gaussian and when the Gaussian QML is misspecified. This numerical evidence reflects the differences in asymptotic optimality or non-optimality of different criteria together with finite sample deviations from the asymptotic properties. At the same time, experiments with a homoskedastic model indicate that the gap between ideal AICm and feasible AICm is negligible, unless the sample is rather small. This evidence endorses the use of AICm in applied problems based on linear models with possibly numerous explanatory variables.

Finally, we provide an empirical illustration, where various versions of AIC are applied to the existing study of an impact of legalized abortion on crime reduction (Donohue III and Levitt 2001, Belloni, Chernozhukov and Hansen 2014). Depending on the type of crime, AICm selects the same, slightly more parsimonious, or significantly more parsimonious model than AICc (which, in turn, tends to select a bit smaller model than the original AIC). The penalty differences turn out to be sufficient to secure such changes in selection, which are driven by a positive, though small, association between heteroskedasticity and regressor leverages.

The article is structured as follows. Section 2 lays out the setup based on a correct regression specification, introduces the many-regressor AIC, and discusses the links to AIC and AICc as well as to other celebrated model selection criteria. Section 3 discusses asymptotic properties of AICm in a setup with possible misspecification, such as its asymptotic optimality and some following mechanics of model selection, also in comparison with the classical versions of AIC. In Section 4, we introduce the feasible version of AICm and find out under what circumstances it is asymptotically equivalent to ideal AICm. Section 5 presents simulation evidence and its discussion. Section 6 contains an illustrative empirical application. We conclude in Section 7 by discussion of possible extensions and challenges. Appendices contain mathematical proofs.

2 Many-regressor AIC

2.1 Setup with correct specification

We consider the linear equation

$$y_i = x_i'\beta + e_i, \tag{2.1}$$

where y_i is a scalar, and x_i and β are $m \times 1$. Letting $\mathcal{X} = \{x_i\}_{i=1}^n$, we assume that the errors $\{e_i\}_{i=1}^n$ are independent conditional on $\mathcal{X}$, and for the time being, that the equation (2.1) is a correctly specified mean regression, so that $E[e_i|\mathcal{X}] = 0$ for all $i = 1, ..., n$. In contrast to the classical linear regression for which the AIC and AICc were originally derived, we do not impose homoskedasticity. Define individual variances

$$\sigma_i^2 = E\big[e_i^2|\mathcal{X}\big],$$

assuming that they do exist. In the matrix form, the model then can be written as

$$y = X\beta + e, \quad E[e|\mathcal{X}] = 0, \quad E[ee'|\mathcal{X}] = \operatorname{dg}\left\{\sigma_i^2\right\}_{i=1}^n, \tag{2.2}$$

where $y = (y_1, \cdots, y_n)'$, $X = (x_1, \cdots, x_n)'$, and $e = (e_1, \cdots, e_n)'$. The analysis will be conditional on $\mathcal{X}$, and all expectations and variances, unless otherwise noted, are conditional.

Let $\hat{\beta}$ be the OLS estimator of β:

$$\hat{\beta} = (X'X)^{-1} X'y, \tag{2.3}$$

and introduce the (degree-of-freedom adjusted) residual variance

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n-m}, \tag{2.4}$$

where $\hat{e} = (\hat{e}_1, \cdots, \hat{e}_n)'$, and $\hat{e}_i = y_i - x_i'\hat{\beta}$ for $i = 1, \cdots, n$. Let us define the $n \times n$ projection matrices $P = X(X'X)^{-1}X'$ and $M = I_n - P$, with $\left(i^{th}, j^{th}\right)$ elements P_{ij} and M_{ij}. Note for future use that, in terms of the 'residual projection' matrix,

$$\hat{\sigma}^2 = \frac{y'My}{n-m} = \frac{e'Me}{n-m}.$$

All formulas for different AIC throughout the article have the form

$$AIC_f = \log(\hat{e}'\hat{e}) + pen_{AIC_f} + \zeta_f. \tag{2.5}$$

Here, the index $f \in \{o, c, m\}$ refers to the AIC variation, with o standing for 'original,' c standing for 'corrected,' and m for 'many.' The leading term equals, apart from a constant, (minus twice) the estimated average[2] Gaussian quasi-loglikelihood. The second component, pen_{AIC_f}, is a penalty term that is derived, under assumptions at hand, as (twice) the gap between the expected estimated quasi-loglikelihood fit to out-of-sample data and the expected quasi-loglikelihood fit to in-sample data, possibly less the remainder ζ_f that absorbs asymptotically negligible terms, i.e. $\zeta_f = o_P\big(pen_{AIC_f}\big)$ as $n \to \infty$. The gap arises because the candidate models contain parameters that are estimated from the same sample the likelihood is evaluated from. A presumption used in the derivation of AIC is that the regression model under consideration is not underspecified in terms of its regressor set, i.e., the model (2.1) is deemed true. Eventually though, these criteria are most likely to be applied to misspecified models; see Section 3 for a corresponding analysis. The sum of the leading term and the penalty component in (2.5) is minimized to select a model from a given set of candidate models.

2.2 AICo and AICc penalties

The penalty of the original AICo for a linear regression is derived under the assumption that the Gaussian quasi-density is in fact the true conditional distribution, which includes, in particular, conditional homoskedasticity. The celebrated AICo penalty reads

$$pen_{AIC_o} = 2\,\frac{m}{n},$$

whereas $\zeta_o = O\big(m^2/n^2\big) = o\big(pen_{AIC_o}\big)$. The remainder term is negligible in the few-regressor setup. However, when the regressors are even moderately many, such a remainder may result in a bias and make a strong impact on how big a model is selected, not to mention the case when m is comparable to n.

The influence of numerosity of regressors is explicitly acknowledged by the 'corrected' AIC criterion AICc derived under exact error normality in Sugiura (1978):

$$pen_{AIC_c} = 2\,\frac{m+1}{n-m-2}\,,$$

with $\zeta_c = 0$. See also a derivation and discussion in Cavanaugh (1997) and Burnham and Anderson (2002, p. 378). This correction can be also derived as a second order bias

[2]Note that we have accepted the convention when the total likelihood is normalized per one observation, i.e., without the n multiplier.

adjustment (Hurvich and Tsai 1989) in nonlinear models. Note that $pen_{AICc} > pen_{AICo}$, and so an AICc-selected model is always no larger than an AICo-selected model.

Burnham and Anderson (2002) strongly advocate the use of AICc in situations when the sample size is small or the parameters are many. The comparison between AICo and AICc is abundant in the literature; see, for example, Cavanaugh (1997), Burnham and Anderson (2002) and Cesar, Vivanco and Menezes (2014).

Note a common property of the AICo and AICc penalties: they are immediately implementable, being a function of known m and n only. This pleasant feature is a consequence of a tight distributional specifications of the candidate model and data generation, conditional homoskedasticity in particular.

2.3 Assumptions

We impose the following conditions on the number of regressors relative to the sample size, on the skedastic pattern, and on existence of moments. All the bounding constants are absolute in the sense that they do not depend on n.

Assumption 1 *As $n \to \infty$, $m \to \infty$, we have, almost surely,*
(i) $\mathrm{rk}(P) = m$ *and* $\underline{C}_M \leq \min_{1 \leq i \leq n} M_{ii}$, *where* $\underline{C}_M > 0$;
(ii) $\underline{C}_\sigma < \min_{1 \leq i \leq n} \sigma_i^2 \leq \max_{1 \leq i \leq n} \sigma_i^2 \leq \overline{C}_\sigma$, *where* $\underline{C}_\sigma, \overline{C}_\sigma > 0$;
(iii) $\max_{1 \leq i \leq n} E\left[e_i^8\right] \leq \overline{C}_8$, *where* $\overline{C}_8 > 0$;
(iv) $\left(n^{-1} e' M e\right)^{-6}$ *is uniformly integrable.*

Assumption 1 allows for a spectrum of regressor numerosity – from growing arbitrarily slowly to growing moderately fast to growing as fast as proportionally to n. It is convenient to maintain asymptotically growing m even in case regressors are not that many, to associate model selection with the dimensionality parameter that affects the asymptotic behavior of AIC ingredients. Note that, importantly, we do not make assumptions about asymptotically vanishing diagonal elements of the projection matrix P (often referred to as 'leverage values'), which preclude the numerosity of regressors to be comparable to the sample size. For the conventional regression, such assumptions are typically variations of the condition $\max_{1 \leq i \leq n} P_{ii} = o_P(1)$ unless the number of regressors is allowed to grow proportionately with the sample size.[3]

[3]See Anatolyev and Yaskov (2017) for more on the asymptotic behavior of diagonal elements of projection matrices in the case of proportionality of m to n.

Assumption 1(i) combines a rank condition with a restriction on maximal leverages; both are conventional in the many-regressors literature. Assumption 1(i) also implies that $m \leq C_m n$, where $0 < C_m = 1 - \underline{C}_M < 1$, and in the case of proportionality of regressor numerosity to sample size, makes sure that the gap between regressor numerosity and sample size is still large. Assumption 1(ii) conventionally restricts the pattern of heteroskedasticity. Assumption 1(iii) restricts conditional higher moments of the error term; the conditional moments of order $r < 8$ are then automatically bounded by, say, $\overline{C}_r > 0$. The technical condition in Assumption 1(iv) is imposed in order for expectations involving negative powers of the variance estimator in the quasi-loglikelihood to exist.

2.4 AICm penalty

Introduce the following weighted average individual variances:

$$\bar{\sigma}_P^2 = \frac{1}{m}\sum_{i=1}^{n} P_{ii}\sigma_i^2, \quad \bar{\sigma}_M^2 = \frac{1}{n-m}\sum_{i=1}^{n} M_{ii}\sigma_i^2.$$

The following result emerges when the AIC penalty is derived from the same principle, as twice the (asymptotic) gap between the expected estimated normal quasi-likelihood fit to out-of-sample data and the expected estimated normal quasi-likelihood fit to in-sample data, under the presumption that the model (2.2) is true. However, unlike AICo and/or AICc, it does not assume conditional homoskedasticity or that the true conditional distribution is normal.

Theorem Am. *Let Assumption 1 hold. The many-regressor AIC penalty is*

$$pen_{AIC_m} = \frac{2m}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2} = 2\,\frac{\sum_{i=1}^{n} P_{ii}\sigma_i^2}{\sum_{i=1}^{n} M_{ii}\sigma_i^2}\,, \tag{2.6}$$

with the remainder term

$$\zeta_m = O_P\left(\frac{1}{n}\right). \tag{2.7}$$

Notice that by Assumption 1(ii),

$$2\,\frac{\underline{C}_\sigma}{\overline{C}_\sigma}\frac{m}{n-m} \leq pen_{AIC_m} \leq 2\,\frac{\overline{C}_\sigma}{\underline{C}_\sigma}\frac{m}{n-m}\,.$$

Thus, the AICm penalty (2.6) grows in regressor dimensionality, linearly at slowest, while the remainder term (2.7) does not relate to this dimensionality. It also follows that the AICm and AICc penalties have the same asymptotic rates.

We know that an AICc-selected model is always no larger than an AICo-selected model. Whether an AICm-selected model is larger or smaller than an AICc-selected model depends on the leverage and heteroskedasticity patterns. Namely, an AICm-selected model is no larger than an AICc-selected model if

$$\widehat{cov}\left(P_{ii}, \sigma_i^2\right) > 0, \tag{2.8}$$

where $\widehat{cov}\left(\cdot,\cdot\right)$ denotes covariance with respect to the empirical measure. Plausibly, the association between the regressor leverage and conditional variance is positive (e.g., it is in our empirical illustration, and in virtually all Monte Carlo sections in the many regressor literature), so the models selected by AICm are likely to be more parsimonious, though not necessarily.

2.5 Special cases

We can trace the connection of the AICm penalty to the penalties of the original AICo and the corrected AICc in various special cases.

First, let us relate the AICm penalty to the penalty of the original AICo, which equals $2m/n$, in the classical setup. Suppose conditional homoskedasticity holds, but regressors may be many or few. Then, all individual variances are equal to σ^2 so that $\bar{\sigma}_P^2 = \bar{\sigma}_M^2 = \sigma^2$, and

$$pen_{AIC_m} = 2\,\frac{m}{n-m}\,.$$

The same formula emerges as an asymptotic approximation if the errors are heteroskedastic but the regressors have an asymptotically balanced design, in which case all P_{ii}'s are asymptotically equal;[4] the higher order difference can be attributed to the remainder term ζ_m. These computations suggest a 'homoskedastic' many-regressor AIC penalty.

Corollary AmH. *Suppose that there may be many regressors, but the regression is homoskedastic. Then*

$$pen_{AIC_m} = 2\,\frac{m}{n-m}\,.$$

The same relation holds if the regression is heteroskedastic but the regressors have an asymptotically balanced design.

[4]See Anatolyev and Yaskov (2017) for the theory and examples. A leading example is a Gaussian regressor design used as a leading case in the many-regressor statistical literature, but many other designs also satisfy this condition. However, many important designs are ruled out, including the case of many dummy variables.

Now, along with conditional homoskedasticity, assume that the regressors are few, $m \ll n$. Then,

$$2\frac{m}{n-m} \approx 2\,\frac{m}{n}\,.$$

The difference between these is of order $O\big(m^2/n^2\big) = o\big(pen_{AIC_m}\big)$. Up to this asymptotically negligible difference, which can be attributed to the remainder term ζ_m, the AICm and original AICo in the classical setup are equal.

Corollary AmF. *Suppose that there are few regressors, and the regression is homoskedastic. Then, as $n \to \infty$ and $m \to \infty$,*

$$pen_{AIC_m} - pen_{AIC_o} = O\left(\frac{m^2}{n^2}\right).$$

Finally, we can examine a relation of this penalty to that of the finite-sample corrected AIC criterion AICc. It is also straightforward to compare pen_{AIC_c} and pen_{AIC_m} under the conditions AICc is derived, homoskedasticity being among them. We have the following result.

Corollary AmN. *Suppose that there may be many regressors, but the regression is homoskedastic normal. Then, as $n \to \infty$ and $m \to \infty$,*

$$pen_{AIC_m} - pen_{AIC_c} = O\left(\frac{1}{n}\right).$$

Thus, in the special case of conditional homoskedastic normality, not only when m grows slowly but also when m is proportional to n, the AICm and AICc penalties are essentially the same, and the difference can be attributed to the remainder term ζ_m.

2.6 Relation to non-Akaike criteria

In this subsection, we contrast and compare various versions of the Akaike criterion to a couple of other celebrated criteria – Mallows' and Takeuchi's – that also target prediction performance.

The *Mallows criterion* (Mallows, 1973), originally developed for a conditionally homoskedastic regression,

$$C_p = \hat{e}'\hat{e} + 2m\sigma^2, \tag{2.9}$$

is another popular tool of model selection proven to have favorable properties in suitable setups. Let us set σ^2 in the penalty to the degrees-of-freedom adjusted OLS residual

variance (2.4), motivated by the relation $E[\hat{\sigma}^2] = \sigma^2$.[5] Then, consider the following monotonic transformation of C_p:

$$\log\left(\frac{C_p}{n}\right) = \log\left(\frac{\hat{e}'\hat{e}}{n} + 2\frac{m}{n}\frac{\hat{e}'\hat{e}}{n-m}\right) = \log\left(\hat{e}'\hat{e}\right) + \log\left(1 + 2\frac{m}{n-m}\right) - \log n.$$

The last term is independent of m, so it can be ignored. When the number of regressors is small ($m \ll n$), the second term is approximately equal to pen_{AIC_c}, and hence, such a monotonic transformation of C_p is nearly equivalent to AIC_c. That is, in a conditionally homoskedastic linear regression with few regressors, Mallows model selection does almost the same job as model selection using the corrected AIC. Likewise, one can see a similar relation of a heteroskedastic version of the Mallows criterion to AIC_m. Andrews (1991) extended (2.9) to conditionally heteroskedastic models, in which case the *generalized Mallows criterion* reads

$$GC_p = \hat{e}'\hat{e} + 2\sum_{i=1}^{n} P_{ii}\sigma_i^2. \tag{2.10}$$

Similarly to above, exploiting the relation $E\left[\hat{e}'\hat{e}\right] = \sum_{i=1}^{n} M_{ii}\sigma_i^2$, one can find that

$$\log\left(\frac{GC_p}{n}\right) \approx \log\left(\hat{e}'\hat{e}\right) + \log\left(1 + 2\frac{\sum_{i=1}^{n} P_{ii}\sigma_i^2}{\sum_{i=1}^{n} M_{ii}\sigma_i^2}\right) - \log n.$$

Hence, in a conditionally heteroskedastic linear regression with few regressors, Mallows model selection does almost the same job as model selection using AICm.

However, the near-equivalence evaporates when the regressors are many in the sense of proportionality of m to n,[6] in which case the penalty terms of the monotonically transformed Mallows criteria,

$$pen_{C_p} = \log\left(1 + 2\frac{m}{n-m}\right), \quad pen_{GC_p} = \log\left(1 + 2\frac{\sum_{i=1}^{n} P_{ii}\sigma_i^2}{\sum_{i=1}^{n} M_{ii}\sigma_i^2}\right),$$

may seriously deviate from pen_{AIC_c} and pen_{AIC_m}, respectively. Because the $\log(\cdot)$ function is strictly concave, the modified AIC criteria penalize for model complexity more heavily than the corresponding Mallows criteria, and, because $m/(n-m)$ gets unbounded as m gets close to n, the discrepancy is potentially unbounded. This discrepancy is due to the difference in the curvature of the quality-of-fit terms relative to the penalty terms.

[5] The common strategy when using the Mallows criterion in practice is to estimate σ^2 instead by the OLS residual variance from the model that nests all the models under consideration.

[6] Anatolyev (2021) proposes how to operationalize the infeasible generalized Mallows criterion (2.10) when regressors are many, along the lines of Section 4 of the present article.

Now we contrapose the AICm penalty to that of the Takeuchi (1976) information criterion (TIC) that explicitly acknowledges possible misspecification of the form of the conditional density (but not of the form of the regression). The TIC penalty reads

$$pen_{TIC} = 2\,\frac{\mathrm{tr}\big(Q_{\partial s}^{-1}V_s\big)}{n},$$

where $Q_{\partial s}$ is minus a derivative of the quasi-score, and V_s is a variance thereof. Underlying the construction of TIC are a Taylor expansion of the quasi-logdensity and asymptotic normality of parameter estimates, and severe numerosity of regressors (e.g. proportionality of m to n) is likely to invalidate these asymptotic expansions. In contrast, (2.6) is derived without resorting to conventional asymptotic properties of parameter estimates. On the other hand, in the few- or moderately-many regressor frameworks when the TIC criterion is valid, the asymptotic approximations underlying the construction of TIC are too crude for our purposes; for example, under the correct specification of the quasi-density when $Q_{\partial s} = V_s$, the TIC penalty reduces to $2(m+1)/n$, which is essentially the AICo penalty. In contrast, as we saw in the previous subsection, the AICm penalty contains the many-regressor adjustment akin to that of the AICc criterion.

3 Asymptotic Properties

3.1 Setup with misspecification

All the AIC criteria are derived under the presumption that the regression model (2.2) is correctly specified. Henceforth, we part with this presumption and consider a more realistic situation, when no linear combination of regressors may be exactly equal to the true regression function. These misspecified models typically arise as linear regressions employing subsets of variables in X, or as linear approximations of the true regression using nonlinear functions of elements of X. That is, a linear model $X\beta$ represents an approximation to the true mean regression $g = E\,[y|\mathcal{X}]$, and the misspecification error $g - X\beta$ can often be eliminated only asymptotically. The usual measure of misspecification is

$$\Delta_g = \frac{g'Mg}{n}.$$

If $g = X'\beta + v$, where v is the error in a linear projection of g on X, then $Mg = Mv$. We say that the linear model $X\beta$ is *correctly specified* if $v = 0$; then $\Delta_g = 0$. Generally,

$\Delta_g = n^{-1}v'Mv = O_P(1)$, and we say that the linear model $X\beta$ is *asymptotically correctly specified* if $\Delta_g \neq 0$ but $\Delta_g \to^P 0$, and *asymptotically misspecified* if $\Delta_g \neq 0$ and $\Delta_g \not\to^P 0$.

Define

$$\mu_n = \frac{\sum_{i=1}^n P_{ii}\sigma_i^2}{\sum_{i=1}^n \sigma_i^2},$$

which is a monotonic transformation of pen_{AIC_m}.[7] Note that in the homoskedastic case, μ_n is simply equal to m/n, while in the heteroskedastic case, the asymptotic order of μ_n is equal to that of m/n thanks to Assumption 1(ii); see discussion under Theorem Am. We will call a model *moderately large* with *moderately many regressors* if $\mu_n \to^P 0$, so that m asymptotically grows with a rate smaller than n. In contrast, we will call a model *really large* with *really many regressors* if $\mu_n \not\to^P 0$, so that m may asymptotically grow with rate n.

3.2 Selection procedure

Suppose a researcher is interested in using to select a model from a collection of linear models using the AICm criterion. Denote the candidate model set by Q. Let us attach an argument q to the projection matrices corresponding to model $q \in Q$ to obtain $P(q)$ and $M(q)$, and similarly to all the other objects that depend on q, such as $\Delta_g(q)$, $m(q)$, $\mu_n(q)$, etc. Further, when imposing Assumption 1, we assume that its statements that depend on q – in particular, conditions (i) and (iv) – hold uniformly for all $q \in Q$. We also presume that each candidate model can be characterized by, on the one hand, whether it is asymptotically correctly specified or misspecified, i.e. whether its Δ_g is zero in the limit or not, and on the other hand, whether it is moderately large or really large, i.e. whether its μ_n is zero in the limit or not. This implies that the candidate model set Q can be partitioned into four corresponding subsets (some of which may be empty).

The model selected by AICm is

$$\hat{q}_m = \arg\min_{q \in Q} \; AIC_m(q),$$

where $AIC_m(q)$ is the value of AICm for model q. We assume that the minimum is attained by only one model; however, multiple models in the candidate set may well have asymptotically equivalent behavior.

[7] Namely, $\mu_n = 1 - (pen_{AIC_m}/2 + 1)^{-1}$.

3.3 Asymptotic optimality

Shibata (1981) considered a problem of regression variable selection in such a situation under homoskedasticity, when the number of regressors diverges more slowly than the sample size. Shao (1997) studied the properties of various selection criteria in the homoskedastic setup, including the possibility of really many regressors. In particular, Shao (1997) showed that under certain conditions put on model complexity and quality of approximation, AICc is *asymptotically loss efficient*, i.e. asymptotically the selected model delivers a minimal value of the quadratic loss function that measures the quality of fit to the true regression function. A more frequently used term for this property is *asymptotic (prediction) optimality* (e.g., Li, 1987, Andrews, 1991). In this subsection, we study the asymptotic optimality of AICm, and, for completeness, of AICc and AICo, in the heteroskedastic many-regressor setup.

For model q, define the traditional loss function, the average squared error,

$$L(q) = \frac{(g - P(q)\, y)'(g - P(q)\, y)}{n} = \Delta_g(q) + \frac{e' P(q)\, e}{n},$$

where the second equality follows by regression algebra. The model $\hat{q}_m$ selected by AICm will be asymptotically optimal over Q if

$$\frac{L(\hat{q}_m)}{\inf_{q \in Q} L(q)} \to^P 1,$$

i.e., simply speaking, when the minimal loss value is asymptotically indistinguishable from its value attained from the AICm-selected model.[8] For models selected by AICo and AICc, asymptotic optimality is defined analogously.

The following conditions represent alternative (though overlapping) circumstances when the asymptotic optimality of AICm holds.

Condition M: $\sup_{q \in Q} \mu_n(q) \to^P 0$.

Condition G: $\sup_{q \in Q} \Delta_g(q) \to^P 0$.

Condition M implies that all models in the candidate set must be moderately large. Condition G allows models in the candidate set to be really large, but requires all models to asymptotically eliminate misspecification. The following theorem extends the findings

[8]Note that this condition is weaker than a requirement that both objectives would prefer the same model; see Shao (1997).

of Shao (1997) to the heteroskedasticity-robust AICm criterion in heteroskedastic setups.

Theorem Om. *Let Assumption 1 hold. If Condition M or Condition G is satisfied, then model $\hat{q}_m$ is asymptotically optimal over Q.*

Note that if the candidate model set includes really large models (i.e., Condition M fails) and the best fitting model does not asymptotically eliminate the misspecification error (i.e., Condition G fails), the AICm-selected model may not be asymptotically optimal. Arguably, if a model uses a really big set of regressors and still falls short of recovering the true regression even asymptotically, it cannot be regarded as a reasonable one.

It may also be interesting to know how far the heteroskedasticity non-robust AICc falls from AICm in terms of asymptotic optimality. It turns out that if a certain stringent condition on the heteroskedasticity pattern and regressor design is satisfied, AICc may also be optimal. Recall that σ_i^2's are variances of the true regression errors, and $\widehat{cov}(\cdot,\cdot)$ denotes covariance with respect to the empirical measure.

Condition P: $\widehat{cov}(P_{ii}(q), \sigma_i^2) = o_P(m(q)/n + \Delta_g(q))$ *uniformly in* $q \in Q$.

Condition P states that conditional heteroskedasticity is weak, in the sense of being dominated by model complexity and/or asymptotic misspecification. Under this circumstance, it makes intuitive sense that the homoskedasticity-based AICc also exhibits optimality properties. The quantity $\widehat{cov}(P_{ii}, \sigma_i^2)$ is proportional to a difference between heteroskedastic quantity μ_n in the heart of AICm penalty and its homoskedastic analog m/n, and so Condition P ensures asymptotic equivalence of the AICm and AICc penalties, while its violation breaks this equivalence.

For asymptotically misspecified moderately large models, Condition P is automatically satisfied, as $\widehat{cov}(P_{ii}, \sigma_i^2) \to^P 0$, while $\Delta_g \not\to^P 0$. Consider correctly specified models, having shut down misspecification for a clearer picture. Then, a slightly (barring conditional homoskedasticity or other special heteroskedasticity patterns) stronger statement than Condition P is that $\Delta_P \equiv n^{-1}\sum_{i=1}^n |P_{ii} - m/n| = o_P(m/n)$ for all candidate models. For really large models, this condition requires, as a minimum, the regressor design to be asymptotically balanced, i.e. $|P_{ii} - m/n| = o_P(1)$ for almost all i. Anatolyev and Yaskov (2017) give examples of regressor designs when this condition holds and when it fails. While for most plausible random regressor designs the condition $\Delta_P = o_P(1)$ fails, in a leading, though non-practical, example – under rotated IID regressor designs, – it

does hold. For moderately large models, Δ_P is of asymptotic order m/n, so Condition P does restrict Δ_P further. Even for a balanced regressor design, considering that the difference $|P_{ii} - m/n|$ has then the asymptotic order $m^{-1/2}$, it further requires m to grow faster than $n^{2/3}$ (or, if there is misspecification, $m\Delta_g^2$ to grow).

Denote

$$\hat{q}_c = \arg\min_{q \in Q} \; AIC_c\,(q)\,.$$

Theorem Oc. *Let Assumption 1 hold. If Condition M or Condition G is satisfied, and if in addition Condition P holds, then model $\hat{q}_c$ is asymptotically optimal over Q.*

Because of the stringency of Condition P for asymptotically correctly specified models, in empirically relevant situations AICc is likely to not possess the property of asymptotic optimality, unless the regressor design is artificially favorable, or unless the models are not precisely specified so that misspecification dominates mismatches of regressor design.

Finally, it may be interesting to know the problems of the crude AICo related to optimality. It turns out that in situations when AICc is optimal, AICo possesses the optimality property too, but only if the candidate set does not contain too large models. Denote

$$\hat{q}_o = \arg\min_{q \in Q} \; AIC_o\,(q)\,.$$

Theorem Oo. *Let Assumption 1 hold, and let in addition Condition P hold.*
(OoM) *If Condition M holds, then model $\hat{q}_o$ is asymptotically optimal over Q.*
(OoG) *If Condition G holds, and if additionally* $\operatorname{plim}_{n\to\infty} \sup_{q\in Q} \mu_n\,(q) < \frac{1}{2}$, *then model $\hat{q}_o$ is optimal over Q.*

That is, for moderately large and some really large models, the AICo penalty is asymptotically as strong as that of AICc, but at a certain point becomes insufficient. The intriguing additional condition in part (OoG) sets an upper threshold for model complexity, beyond which the feeble AICo penalty cannot keep up with extreme overfitting.

To summarize, AICm exhibits the asymptotic optimality property under weaker conditions than AICc does, which in turn shows optimality more often than AICo. In situations when two or three of them are asymptotically optimal at the same time and deliver the same asymptotic loss efficiency, the ability of delivering a better quality of fit depends on their higher order asymptotic properties.

3.4 Selection mechanics

The derivation of Theorem Om provides, as a by-product, some insights on how AICm ranks different types of models in large samples. In particular, several observations follow.

1. If all the models in the candidate set are really large and asymptotically eliminate the misspecification error, the selection is driven by model complexity, and the smallest one will be selected.

2. A really large model has a chance to be selected only if the candidate set does not contain moderately large models that asymptotically eliminate the misspecification error; otherwise any really large model will lose to one of those.

3. If the candidate set contains only moderately large models, and...

   (a) if some of these models do not asymptotically eliminate the misspecification error, they are unselected in favor of those which do, provided that there are such in the candidate model set.

   (b) if all of these models do not asymptotically eliminate the misspecification error, the selection is driven exclusively by approximation quality, and the selected model is the most successful in elimination of the misspecification error.

   (c) if all of these models do asymptotically eliminate the misspecification error, the selection is driven by the trade-off between model complexity and approximation quality. Among the correctly specified models the minimal model will be selected, but if there are also misspecified ones that are asymptotically correct, some of these may be preferred if misspecification is offset by model parsimony.[9]

The case not covered by Theorem Om is when there are models in the candidate set that are really large and there are models that do not asymptotically eliminate the misspecification error. If really large models have the latter property, they are unselected from the candidate set if it also contains moderately large models that do asymptotically eliminate the misspecification error. However, really large models may compete with

[9]Suppose such models are characterized by $n\Delta_g \asymp n^{\zeta_g}$ and $m \asymp n^{\zeta_m}$ for $0 < \zeta_g, \zeta_m < 1$. Then, the one with the smallest $\max\{\zeta_g, \zeta_m\}$ will be selected on the asymptotic terms.

moderately large models that do not asymptotically eliminate the misspecification error, on the basis of the trade-off between model complexity and approximation quality.[10]

The AICc and AICo criteria in their domains of optimality – Condition P for AICc and Condition P together with the additional restriction for AICo – exhibit similar patterns of model selection. In particular, model unselection often occurs because of an interplay of asymptotic orders of misspecification and model complexity, and the exact form of the AIC penalty term does not matter, as they all have the same asymptotic order. When Condition P is violated (but the additional restriction for AICo holds[11]), AICc and AICo still exhibit these selection patterns, but their selected models cease to be optimal in the sense of asymptotic loss efficiency, while AICm ensures asymptotic "fine-tuning" of the trade-off between model complexity and approximation quality by accounting for the heteroskedasticity shape, and selects an asymptotically loss efficient model.

4 Feasible AICm

4.1 Construction

The traditional AICo and AICc penalties depend only on known m and n, hence these criteria are directly implementable. The many-regressor AIC, in contrast, is infeasible as it depends on unknown quantities $\bar{\sigma}_P^2$ and $\bar{\sigma}_M^2$. An operational version of the AICm criterion and its penalty, call them AIC_m^F and $pen_{AIC_m}^F$, respectively, with F standing for 'feasible,' exploits unbiased estimates of individual variances robust to regressor numerosity. Such estimates are recently proposed in Kline, Saggio and Sølvsten (2020) for unbiased estimation of quadratic forms in parameters of interest, and in Jochmans (2022)

[10]Suppose a moderately large model is characterized by $\delta \equiv \operatorname{plim}_{n\to\infty}(n\Delta_g/e'e)$ and a really large model is characterized by $\mu \equiv \lim_{n\to\infty} m/n$. Which one of these models will be selected depends on whether $\log(1+\delta)$ is smaller (then the former is selected) or larger (then the latter is selected) than $\upsilon(\mu)$, where the function $\upsilon(\cdot)$ can be found in the Appendix. Inspection of the shapes of $\log(1+\cdot)$ and $\upsilon(\cdot)$ on $(0,1)$ reveals that both outcomes are possible.

[11]A particularly pathological selection behavior is exhibited by AICo when the additional condition fails. If $\operatorname{plim}_{n\to\infty}\mu_n(q_{VL}) > \frac{1}{2}$, the "very large" model q_{VL} may beat smaller really large models, which would be preferred by AICm and AICc. Furthermore, when $\operatorname{plim}_{n\to\infty}\mu_n(q_{EL}) > \mu^*$, where $\mu^* \approx 0.80$ is a solution to the equation $\log(1-\mu^*) = -2\mu^*$, the "extremely large" model q_{EL} will beat all the moderately large models, while being definitely unselected by AICm and AICc.

for estimating asymptotic variances when covariates may be many:

$$\hat{\sigma}_i^2 = \frac{y_i \hat{e}_i}{M_{ii}}. \tag{4.11}$$

Such estimates originate from leave-one-out OLS estimation and the well-known equality between OLS residuals $\hat{e}_i$ corrected for leverage and leave-one-out OLS residuals. The estimates (4.11) are indeed conditionally unbiased: $E\left[\hat{\sigma}_i^2\right] = M_{ii}^{-1} E\big[\left(x_i'\beta + e_i\right) \sum_{j=1}^n M_{ij} e_j\big] = M_{ii}^{-1} \sum_{j=1}^n M_{ij} E\left[e_i e_j\right] = \sigma_i^2$. It is worth noting that the estimates (4.11), though unbiased, are not individually consistent for the respective estimands; it is their weighted linear combinations that are typically consistent for the corresponding combinations of the estimands (Kline, Saggio and Sølvsten 2020; Anatolyev and Sølvsten 2023; Boot 2023), like those in the normalized numerator and denominator of the AICm penalty (2.6).

Based on (4.11), we can form an unbiased and consistent estimate of $\bar{\sigma}_P^2$:

$$\hat{\sigma}_P^2 = \frac{1}{m} \sum_{i=1}^n P_{ii} \hat{\sigma}_i^2.$$

Regarding estimation of $\bar{\sigma}_M^2$, we can simply use $\hat{\sigma}^2$ as its unbiased and consistent estimator (cf. Lemma B(i) in the Appendix). This estimator is especially elegant in view of the following sequence of identities, motivated by the same plug-in principle as used for $\hat{\sigma}_P^2$ above:

$$\begin{aligned} \sum_{i=1}^n M_{ii} \hat{\sigma}_i^2 &= \sum_{i=1}^n M_{ii} \frac{y_i \hat{e}_i}{M_{ii}} \\ &= \sum_{i=1}^n y_i \left(My\right)_i \\ &= y'My \\ &= \left(n - m\right) \hat{\sigma}^2. \end{aligned}$$

That is, the plug-in principle leads to the same estimate for $\bar{\sigma}_M^2$. To recapitulate, the feasible AICm penalty is constructed as

$$pen_{AIC_m}^F = \frac{2m}{n-m} \frac{\hat{\sigma}_P^2}{\hat{\sigma}^2} = 2 \, \frac{\sum_{i=1}^n P_{ii} \hat{\sigma}_i^2}{\sum_{i=1}^n M_{ii} \hat{\sigma}_i^2}. \tag{4.12}$$

The recent literature also suggests another construction of individual variance estimates that are robust to many regressors. Cattaneo, Jansson and Newey (2018a) proposed, in the course of constructing valid inference when there are many nuisance covariates along with few regressors of interests, an alternative estimator of individual variances

as $\hat{\sigma}_i^2 = \left((M \odot M)^{-1} \hat{e} \odot \hat{e}\right)_i$, see also Rao (1970). The theory underlying this estimator contains a condition, although sufficient, that the number of regressors be at most half of the sample size. Also, these estimates are more cumbersome to compute, and do not follow the elegant identity mentioned above. In our simulations exercises of Section 6, we exploit these estimates for comparison purposes as well.

Yet another approach is to modify the estimates (4.11). These estimates are not scale-invariant, hence it makes sense to 'demean' them, by subtracting the sample mean of y_i from each y_i in (4.11). We also examine this variation in our simulations experiments of Section 6.

4.2 Asymptotic equivalence

Now we investigate the impact on model selection of the operationalization of the AICm penalty. To ensure that the feasible penalty works properly, some technical conditions need to be imposed. Let us introduce the quantity

$$\pi_4^4 = \frac{1}{n} \sum_{i=1}^{n} P_{ii}^4.$$

Assumption 2 *As $n \to \infty$ and $m \to \infty$, (i) $\sum_{i=1}^{n} g_i^4 = O_P(n)$, (ii) $\pi_4 = O_P(m/n)$.*

Assumption 2(i) precludes trends in the regression function. Kline, Saggio and Sølvsten (2020), when using the same individual variance estimates, impose a much stronger condition $\max_{1 \leq i \leq n} |g_i| = O_P(1)$, but note that it can be relaxed to allow for a slow increase; in particular, Cattaneo, Jansson and Newey (2018) and Jochmans (2022) assume $\max_{1 \leq i \leq n} |g_i| = o_P(\sqrt{n})$. Assumption 2(i) restricts the behavior of a power average instead of a maximum; under the IID regressor design, it is naturally satisfied if the fourth moments of g_i are finite. Assumption 2(ii) is a weak requirement placed on the asymptotic behavior of average powered leverages. If regressors are really many, $m/n \asymp 1$, so $\pi_4 = O_P(1)$ automatically. When regressors are moderately many, in an IID environment each P_{ii} is concentrated around m/n, and π_4 is expected to be $O_P(m/n)$ under existence of sufficiently high moments of elements of x_i. Cattaneo, Jansson and Ma (2019) assume that $\max_{1 \leq i \leq n} P_{ii} = O_P(m/n)$ when $m/n \to 0$, and list examples, in which even this much stronger assumption holds. In non-IID environments, Assumption 2(ii) rules out trends in regressor design imbalances, such as dummy covariates for groups of expanding size.

Now we determine the effect of estimation of individual variances on the penalty term pen_{AICm}. Let us define

$$\Delta_{pen} = pen^F_{AICm} - pen_{AICm},$$

the deviation of the feasible penalty from the ideal penalty. This deviation contains bias and variance components related to both model misspecification and estimation of individual variances.

We first consider asymptotically correctly specified models, i.e. those with $\Delta_g \to^P 0$. It turns out that when misspecification asymptotically vanishes, both components of Δ_{pen} become negligible.

Theorem Fm. *Suppose Assumptions 1 and 2 hold, and let* $\Delta_g \to^P 0$. *Then,*

$$\Delta_{pen} = o_P\left(pen_{AICm}\right).$$

Thus, using the feasible AICm penalty has no impact on model selection in asymptotic terms, and the asymptotic equivalence of AICm^F and AICm holds. Such a conclusion is more straightforward for asymptotically correctly specified really large models, saying that penalty estimation is consistent, as then $\Delta_{pen} = o_P(1)$, while the order of the penalty itself is $O_P(1)$ but not $o_P(1)$. For asymptotically correctly specified moderately large models, Theorem Fm is more informative saying that the penalty estimation error is of a smaller asymptotic order than the penalty itself, although both are asymptotically vanishing. To summarize, the asymptotic equivalence holds for both types of large models if they are sufficiently well specified.

What happens if the model under consideration does not asymptotically eliminate the misspecification error, i.e. when $\Delta_g \not\to^P 0$? Examining the derivations of Theorem Fm reveals that in such a case, Δ_{pen} is of the same asymptotic order as pen_{AICm}, and thus, using the feasible AICm penalty with asymptotically misspecified models is expected to distort AICm-based model selection.

4.3 Asymptotic optimality revisited

Finally, it may be interesting to know whether the results on asymptotic optimality of Section 4 for the ideal AICm extend to the feasible AICm. The selected model guided by the feasible AICm is

$$\hat{q}^F = \arg\min_{q \in Q} \; AIC^F_m(q),$$

where $AIC_m^F(q)$ is the value of feasible AICm for model $q \in Q$, and Q is the model candidate set. We restore the notation that model-specific objects are marked with the model as an argument. Further, when appealing to conditions that depends on q, – in particular, Assumption 1(i,iv), Assumption 2(ii) and the condition $\Delta_g \to^P 0$, – we require that their statements hold uniformly in $q \in Q$.

With the requirement that all the candidate models are at least asymptotically correctly specified, we have $\sup_{q\in Q} \Delta_g(q) \to^P 0$, which is Condition G, and thus one of the prerequisites for Theorem Om holds automatically, irrespective of whether the candidate models are moderately or really large. When Q contains only really large models, asymptotic optimality over Q holds for a simple reason that $\Delta_{pen}(q) = o_P(L(q))$ for all $q \in Q$, because $\Delta_{pen}(q) = o_P(1)$, while the loss $L(q)$ is of order $m(q)/n \succ 1$ for really large models. If, however, Q also contains moderately large models, their $L(q)$ is also $o_P(1)$, as both of its components asymptotically vanish. Note however that the order of $L(q)$ is equal or higher than the order $m(q)/n$ of the penalty. Hence, $\Delta_{pen}(q) = o_P(pen_{AIC_m}(q)) \leq o_P(L(q))$, uniformly in $q \in Q$. Thus, asymptotic prediction optimality of the feasible AICm holds if the asymptotic equivalence of the feasible AICm and ideal AICm holds for all models in the candidate set. If, on the contrary, the latter phenomenon does not hold – which occurs when Condition G is violated – one cannot expect asymptotic optimality to hold either.

4.4 Computational note

Even though the AICm^{F} penalty is more involved than those in AICo and AICc, typically it can be implemented in statistical languages via a one-line statement. Suppose that y and $\hat{e}$ are sitting in arrays y and eh, respectively, and the matrix M is in M. The table below shows relevant statements for computing the AICm^{F} penalty in a number of languages. Note that the denominator is already computed in the first term of any AIC.

MATLAB	$2 * \text{sum}(\text{y}. * \text{eh}. * (1 - \text{diag}(\text{M}))./\text{diag}(\text{M}))/(\text{eh}' * \text{eh})$
GAUSS	$2 * \text{sumc}(\text{y}. * \text{eh}. * (1 - \text{diag}(\text{M}))./\text{diag}(\text{M}))/(\text{eh}'\text{eh})$
R	$2 * \text{sum}(\text{y} * \text{eh} * (1 - \text{diag}(\text{M}))/\text{diag}(\text{M}))/(\text{t}(\text{eh})\% * \%\text{eh})$
Python	$2 * \text{np.sum}(\text{y} * \text{eh} * (1 - \text{np.diag}(\text{M}))/\text{np.diag}(\text{M}))/(\text{eh.T@eh})$
Julia	$2 * \text{sum}(\text{y}. * \text{eh}./\text{diag}(\text{M}). * (1. - \text{diag}(\text{M})))/(\text{eh}'\text{eh})$

5 Simulation evidence

It would be interesting to see how the theoretical results of Sections 3 and 4 are compatible with the actual evidence in finite samples, and how they translate to differences in predictive performance of models selected by AICm and feasible AICm and models selected by the classical AICo and AICc. Theoretical asymptotic optimality of AICm and AICm$^{\mathrm{F}}$ in finite samples may be distorted by errors from multiple sources: asymptotically negligible components in the construction of the expected quasi-loglikelihood, approximation errors in linearization of the logarithmic function in the main component of AIC for moderately large models, errors arising from estimation of individual variances in the feasible AICm penalty, and so on. Another factor is the failure of asymptotic optimality when misspecification is severe, or the regressor leverage structure deviates from favorable designs, or model complexity is overwhelming. The net effect of all these aspects may reveal itself in realistic finite samples, and Monte Carlo experimentation is able to show which effects prevail in which circumstances.

5.1 Data generation

The data generating process is inspired by a simulation study in Cattaneo, Jansson, and Newey (2018b). Introduce the basic d-dimensional regressor vector z_i, where d is small. In our experiments, we vary d between 2 and 5 and generate z_i as a vector of independent standard uniform. The regressors x_i are generated as various powers of the basic regressor z_i as shown in Table 1. Beyond the two minimal models, new sets of regressors first become next-order powers of the basic regressor, and then additionally all interactions of the same order, $^{\odot}$ standing for a Hadamard power and $^{\otimes}$ standing for a Kronecker power. For example, x_i for model 3 contains 1, z_i and $z_i \odot z_i$, while x_i for model 4 contains 1, z_i and 15 distinct elements of $z_i z_i'$. In total, there are 10 models corresponding to the composition of regressors x_i of different dimensionality. Table 1 also shows the dimensionality of regressor sets when $d = 2, \cdots, 5$.

The true regression function takes one of two possibilities:

$$g_i = \begin{cases} \exp\left(\|z_i\|^2\right), \\ 1 + \|z_i\|^2 + \frac{1}{2}\|z_i\|^4. \end{cases}$$

In the former case, no model covers the true regression, so every model is an approximation of the true model. In the latter case, model 7 and all larger models are true.

Table 1: Composition of regressors in the simulation experiment

model q	regressors† $x_i(q)$	$m(q)$			
		$d=2$	$d=3$	$d=4$	$d=5$
1	1	1	1	1	1
2	$1, z_i$	3	4	5	6
3	$x_i(2)\, z_i^{\odot 2}$	5	7	9	11
4	$x_i(2), z_i^{\otimes 2}$	6	10	15	21
5	$x_i(4), z_i^{\odot 3}$	8	13	19	26
6	$x_i(4), z_i^{\otimes 3}$	10	20	35	56
7	$x_i(6), z_i^{\odot 4}$	12	23	39	61
8	$x_i(6), z_i^{\otimes 4}$	15	35	70	126
9	$x_i(8), z_i^{\odot 5}$	17	38	74	131
10	$x_i(8), z_i^{\otimes 5}$	21	56	126	252

† Regressor list expands as model label q grows; $\odot$ signifies Hadamard power, $\otimes$ signifies Kronecker power. For example, model 7 combines regressors from model 6 and fourth powers of basic regressors $z_{i,1}^4, \cdots, z_{i,d}^4$; model 4 combines regressors from model 2 and second powers of basic regressors together with their interactions $z_{i,1}^2, z_{i,1}z_{i,2}, \cdots, z_{i,d-1}z_{i,d}, z_{i,d}^2$.

We generate the regression errors as $e_i = \sigma_i \eta_i$, where η_i are independent random variables normalized to have mean zero and variance unity, from one of three distributions:

$$\begin{cases} \text{Gaussian} & N(0,1), \\ \text{Skewed} & \left(\chi^2(1)-1\right)/\sqrt{2}, \\ \text{Bimodal} & \left(N(-1,1)+N(1,1)\right)/\sqrt{2}. \end{cases}$$

Thus, the normal QML estimator is the true ML in the first case but uses Gaussian quasi-likelihood for non-normal errors in the other two cases. To generate heteroskedasticity, we insert the diagonal elements of the projection matrix $P_{ii}^z = z_i' \left(\sum_{j=1}^n z_j z_j'\right)^{-1} z_i$ associated with the basic regressor z_i directly into the skedastic function: $\sigma_i = nP_{ii}^z$. We also examine the homoskedastic case $\sigma_i = d$. This value is comparable with average σ_i in the heteroskedastic case because $\sum_{i=1}^n P_{ii}^z = d$. In our main set of experiments, the sample

size is $n = 800$.

It may be useful to look at some numerical characteristics of the data generating process related to the expected performance of AIC. Table 2 shows such characteristics for the case $d = 5$, when the candidate set contains moderately large and really large models, leaving out the intercept-only model 1. The fourth and fifth column show the amount of misspecification Δ_g, for both exponential and polynomial specifications. In the case of exponential regression, misspecification is severe with smaller models, but the power regressors provide a good approximation starting from model 8, which employs interactions of fourth powers. For the polynomial regression, the true model is 7, but the approximation quality is good already for models 4, 5 and 6, without employing third powers and even interactions of second powers. The low-numbered models provide a poor approximation to the true regression function in either case, and the prerequisites for AICm optimality are not satisfied; however, they will be satisfied if the low-numbered models are excluded from the candidate set, or when, for example, $d = 2$, because the maximal candidate model then would be moderately large. The sixth column in Table 2 shows the relative deviation of μ_n due to heteroskedasticity from the homoskedastic analog m/n. While for balanced regressor designs this measure[12] is of order of 10^{-4}, the figures point at heavily unbalanced regressor designs in all models, invalidating Condition P and breaking down possible optimality of AICc and AICo. Finally, the last column in Table 2 verifies the asymptotic behavior of π_4, whose values turn out comparable to m/n, which is conformable with Assumption 2(ii) for all the models, moderately large and really large. Assumption 2(i) holds because basic regressors are IID and have bounded support.

5.2 Results and discussion

The number of simulation runs is 5,000. We compute averages, across simulations, of the mean squared prediction error computed as $\mathrm{MSPE}(q) = n^{*-1} \sum_{i=1}^{n^*} \left(g_i^* - x_i^*(q)' \hat{\beta}(q)\right)^2$ for model q, where g_i^* is generated according to the same data generating process, $x_i^*(q)$ is the set of regressors employed in model q, $\hat{\beta}(q)$ is the least squares estimate in model q, and $n^* = 10,000$ is the number of pseudo-out-of-sample observations.

Table 3 contains average values of out-of-sample MSPE in the heteroskedastic case. Beneath each MSPE figure, labelled by $\div$ is sitting the ratio of this MSPE to MSPE from

[12] Obtained via simulations using IID fictitious regressor vectors containing a unity and $m-1$ independent standard uniforms.

Table 2: Some properties of data generation in the simulation experiment, $d = 5$

model q	$m(q)$	$\frac{m(q)}{n}$	$\Delta_g(q)$ in exponential	$\Delta_g(q)$ in polynomial	$\frac{\mu_n(q) - m(q)/n}{m(q)/n}$	$\frac{\pi_4(q)}{m(q)/n}$
2	6	0.008	8.53	0.361	0.209	1.15
3	11	0.014	6.37	0.275	0.322	1.48
4	21	0.026	1.81	0.0109	0.352	1.43
5	26	0.033	1.56	0.00874	0.358	1.52
6	56	0.070	0.287	0.000291	0.428	1.69
7	61	0.076	0.276	0.000283	0.419	1.67
8	126	0.158	0.0306	$< 10^{-8}$	0.431	1.70
9	131	0.164	0.0300	$< 10^{-8}$	0.423	1.68
10	252	0.315	0.00201	$< 10^{-8}$	0.361	1.48

AICc, which serves as a most relevant benchmark. One can easily see the failure of the original AICo in case the regressors are many, but we are interested more in comparisons with the benchmark, AICc. Even with $d = 2$ implying the largest model of merely 21 regressors, a moderate number, one can observe big improvements from AICm, around 30% and sometimes even more, in model selection in terms of out-of-sample fit. These percentages increase further to about 50% and even more when we switch to $d = 3$ implying the largest model of 56 regressors. However, these percentages stabilize or revert the direction when higher d are set, and sit in the range 15-55%, depending on the other DGP features. The MSPE figures, naturally, get smaller when the true regression is contained in the model set, more appreciably the higher d is, and so does the ratio of interest. There is not much variation, however, across the error distribution specification – the true one (Gaussian) or a misspecified one (skewed and bimodal), which has an important implication for the use of Gaussian QMLE.

Now we switch our attention to differences between the figures for the ideal and feasible AICm. The gap is pretty small, in all situations, especially when compared to the ratio figures. Often and naturally, operationalizing AICm tends to decrease the ratio because of estimation noise; interestingly though, when the regression error has a skewed distribution, the ratio in fact increases by additional 2-3%, so that AICm^{F} works even

Table 3: Average values of out-of-sample MSPE, heteroskedastic case, $n = 800$, from simulations

Error distribution	AICo	AICc	AICm	AICm$^{\mathrm{F}}$	AICo	AICc	AICm	AICm$^{\mathrm{F}}$
	$d = 2$				$d = 3$			
	true regression is not contained in model set							
Gaussian	0.181 ÷1.02	0.176 ÷1.00	0.122 ÷0.69	0.123 ÷0.70	0.823 ÷1.19	0.690 ÷1.00	0.387 ÷0.56	0.394 ÷0.57
Skewed	0.180 ÷1.02	0.176 ÷1.00	0.125 ÷0.71	0.120 ÷0.68	0.859 ÷1.13	0.758 ÷1.00	0.422 ÷0.56	0.388 ÷0.51
Bimodal	0.183 ÷1.02	0.179 ÷1.00	0.124 ÷0.69	0.127 ÷0.71	0.813 ÷1.17	0.697 ÷1.00	0.388 ÷0.56	0.394 ÷0.56
	true regression is contained in model set							
Gaussian	0.171 ÷1.03	0.166 ÷1.00	0.111 ÷0.67	0.112 ÷0.68	0.680 ÷1.23	0.553 ÷1.00	0.266 ÷0.48	0.270 ÷0.49
Skewed	0.169 ÷1.03	0.164 ÷1.00	0.116 ÷0.70	0.108 ÷0.66	0.730 ÷1.16	0.628 ÷1.00	0.300 ÷0.48	0.272 ÷0.43
Bimodal	0.171 ÷1.03	0.166 ÷1.00	0.112 ÷0.67	0.114 ÷0.68	0.662 ÷1.22	0.544 ÷1.00	0.266 ÷0.49	0.270 ÷0.50
	$d = 4$				$d = 5$			
	true regression is not contained in model set							
Gaussian	2.76 ÷1.65	1.67 ÷1.00	1.12 ÷0.67	1.12 ÷0.67	10.59 ÷2.80	3.78 ÷1.00	3.22 ÷0.85	3.18 ÷0.84
Skewed	3.02 ÷1.55	1.95 ÷1.00	1.15 ÷0.60	1.12 ÷0.57	11.99 ÷2.96	4.05 ÷1.00	3.26 ÷0.80	3.18 ÷0.78
Bimodal	2.78 ÷1.65	1.68 ÷1.00	1.12 ÷0.67	1.12 ÷0.67	10.41 ÷2.76	3.78 ÷1.00	3.22 ÷0.85	3.20 ÷0.85
	true regression is contained in model set							
Gaussian	1.76 ÷1.97	0.897 ÷1.00	0.495 ÷0.55	0.498 ÷0.56	4.71 ÷3.94	1.20 ÷1.00	0.783 ÷0.65	0.781 ÷0.65
Skewed	2.14 ÷1.87	1.14 ÷1.00	0.524 ÷0.46	0.493 ÷0.43	7.40 ÷5.16	1.43 ÷1.00	0.810 ÷0.56	0.773 ÷0.54
Bimodal	1.72 ÷1.96	0.876 ÷1.00	0.497 ÷0.57	0.499 ÷0.57	4.63 ÷3.95	1.17 ÷1.00	0.779 ÷0.66	0.782 ÷0.67

better than the theoretical AICm.[13]

Figure 1 presents histograms of which model is selected using the four criteria and distributions of the four penalties across the models (with the leading log-likelihood AIC

[13]We conjecture that this occurs due to a favorable sign of the correlation between Δ_{pen} and the ζ_m remainder in (2.7).

term, common to all four criteria, superimposed), in the case of Gaussian errors when the true regression is not contained in model set (i.e. the first lines in Table 3). One can immediately see that the classical AICo and AICc, especially AICo, tend to select significantly bigger models. For higher dimensions, one can see a pile-up of the AICo-selected largest model 10, a footprint of the pathological phenomenon exhibited by AICo when "very large" models are present (cf. footnote 11).

Note that the modal model selected by AICm and AICm$^{\mathrm{F}}$ is always model 4 that uses regressors that are only up to quadratics in the basic regressor; in contrast, the modal models selected by AICo and AICc depend on the dimensionality; for higher dimensional models it becomes model 6 that in addition uses cubics. Indeed, models 4 and 6 are exactly those when misspecification becomes much less severe (cf. Table 2 for the case $d = 5$). One can also see not only a rapidly increasing penalty with the model dimension for all criteria, but also a quickly expanding gap between penalties of AICc, on the one hand, and those of AICm and AICm$^{\mathrm{F}}$, on the other. The latter phenomenon is due to strong heteroskedasticity embedded in the data generation.

Table 4: Standard deviations of selected regressor dimensionality, heteroskedastic case, $n = 800$, from simulations

Error	AICm	AICm$^{\mathrm{F}}$	AICm	AICm$^{\mathrm{F}}$	AICm	AICm$^{\mathrm{F}}$	AICm	AICm$^{\mathrm{F}}$
	$d = 2$		$d = 3$		$d = 4$		$d = 5$	
	true regression is not contained in model set							
Gaussian	2.64	2.85	4.76	5.27	9.25	10.12	16.75	18.76
Skewed	2.69	2.72	5.36	5.50	9.45	10.40	16.86	19.06
Bimodal	2.62	2.80	4.74	5.31	9.19	9.83	16.73	18.41
	true regression is contained in model set							
Gaussian	2.24	2.42	3.28	3.45	5.03	5.13	6.47	6.62
Skewed	2.42	2.24	3.90	3.50	5.34	5.11	6.68	6.61
Bimodal	2.22	2.33	3.44	3.61	5.08	5.22	6.47	6.67

Next, we investigate and quantify finite sample losses of precision from AICm penalty estimation. Table 4 presents such evidence by documenting standard deviations of the

number of regressors in AICm- and AICm$^{\mathrm{F}}$-selected models. The loss of precision in identifying the optimal model is not big – the maximal relative loss throughout the entire table is about 13%. It is larger when the true regression is not contained in the model set ($1\% \div 13\%$) than when it is not ($-10\% \div 8\%$), and has a slight tendency to increase with d in the former case but does not in the latter. Again, in the case of a skewed error distribution and the true regression being contained in the model set, the variability of selected models in fact goes *down* as a result of penalty estimation.

Recall that the heteroskedasticity in our our data generation is positively correlated with regressor leverages, and a consequence, AICm selects no larger models than AICc does (cf. subsection 2.4) which, in turn, selects no larger models than AICo does. If the sign of this association is reversed to negative, the opposite selection tendency will result. To this end, we temporarily change the skedastic function from $\sigma_i = nP^z_{ii}$ to alternative $\sigma_i = n(\max_{1\leq j\leq n} P^z_{jj} - P^z_{ii})$, so that the covariance becomes negative but of a comparable magnitude. Table 5 shows the resulting average values of out-of-sample MSPE. Selection of larger models by AICm has a harmful effect on the MSPE in finite samples, although much smaller (up to 5% for the exponential regression and up to 13% for the polynomial regression) than under positive association (up to 44% and 52%, respectively). As all the discrepancies are net effects from the failure of asymptotic optimality of AICc on the one hand, and finite sample distortions of asymptotic properties of AICm on the other, in the case of negative association and resulting larger model selection the second effect seems to prevail.

In the next set of experiments, we decrease n to investigate finite sample distortions in AICm$^{\mathrm{F}}$ compared to AICm arising in smaller samples because of estimation of $\bar{\sigma}^2_P$ and $\bar{\sigma}^2_M$. It is most clear to see this in the homoskedastic case, where these quantities are constant and the estimation noise is isolated; in this case the ideal criterion is AICc which we take as a baseline for comparison. We set n to 800 and smaller multiples of 100, with a finer grid in the region where the distortions become sizable. Instead of reporting average MSPEs that may be contaminated by outliers, in Table 6 we report median MSPEs together with percentages of outliers exhibited by AICm$^{\mathrm{F}}$, defined as all those MSPEs produced by AICm$^{\mathrm{F}}$, which exceed maximal MSPEs produced by AICm. In addition, we show ratios of dimensionality of maximal models to sample size.

One can see that as long as the sample size is not very small, the gap between median MSPE for AICm and AICm$^{\mathrm{F}}$ is not large in relative terms. Of course, the critical sample size is larger the larger d is, but one cannot say that the gap in those critical situations

Table 5: Average values of out-of-sample MSPE, alternative heteroskedastic case, Gaussian errors, $n = 800$, from simulations

AICo	AICc	AICm	AICm^{F}	AICo	AICc	AICm	AICm^{F}
$d = 2$				$d = 3$			
true regression is not contained in model set							
0.238 ÷1.01	0.236 ÷1.00	0.248 ÷1.05	0.248 ÷1.05	0.587 ÷1.00	0.587 ÷1.00	0.590 ÷1.01	0.590 ÷1.02
true regression is contained in model set							
0.177 ÷1.01	0.175 ÷1.00	0.192 ÷1.10	0.193 ÷1.10	0.334 ÷1.01	0.330 ÷1.00	0.360 ÷1.09	0.364 ÷1.10
$d = 4$				$d = 5$			
true regression is not contained in model set							
1.36 ÷1.00	1.36 ÷1.00	1.39 ÷1.03	1.41 ÷1.04	3.31 ÷1.01	3.27 ÷1.00	3.42 ÷1.05	3.47 ÷1.06
true regression is contained in model set							
0.816 ÷1.01	0.553 ÷1.00	0.611 ÷1.11	0.616 ÷1.12	0.816 ÷1.01	0.806 ÷1.00	0.910 ÷1.13	0.915 ÷1.14

increases with d and maximal relative dimensionality: for example, with $d = 2$ and $n = 100$ corresponding to relative dimensionality of 0.21, the relative gap is in fact larger than that with $d = 5$ and $n = 300$, corresponding to relative dimensionality of 0.84 (as $0.324/0.299 > 4.24/4.08$). The outliers do happen when the relative dimensionality gets large together with the sample getting small, but still their share does not exceed around 15% even for critical combinations.

Finally, Table 7 contains average values of out-of-sample MSPE in the baseline heteroskedastic case under Gaussian errors, when alternative estimators of individual variances are used in the feasible AICm. In addition to the first three columns from Table 3, including the results from the use of Kline, Saggio and Sølvsten (2020) estimates in the column $\text{AICm}^{\text{F}}_{\text{KKS}}$, we add the results from the use of Cattaneo, Jansson, and Newey (2018a) estimates in the column $\text{AICm}^{\text{F}}_{\text{CJN}}$, and the results from the use of 'demeaned' Kline, Saggio and Sølvsten (2020) estimates in the column $\text{AICm}^{\text{F}}_{\text{KKS}-}$. One can see that the variations in MSPE are little, with a uniform tendency to be slightly worse for the

Table 6: Median values of out-of-sample MSPE and percentages of outliers, homoskedastic case, Gaussian errors, true regression is not contained in model set, from simulations

n	$\frac{\dim}{n}$	AICc	AICm	AICm$^{\mathrm{F}}$	%out	$\frac{\dim}{n}$	AICc	AICm	AICm$^{\mathrm{F}}$	%out
			$d=2$					$d=3$		
800	0.03	0.0519	0.0520	0.0521	0%	0.07	0.234	0.234	0.235	0%
400	0.05	0.0940	0.0942	0.0946	0%	0.14	0.374	0.374	0.380	0%
200	0.11	0.189	0.189	0.193	0%	0.28	0.746	0.744	0.777	3%
100	0.21	0.297	0.299	0.324	0%	0.56	1.17	1.17	1.30	16%
			$d=4$					$d=5$		
800	0.16	0.829	0.829	0.828	0%	0.32	2.53	2.53	2.56	0%
400	0.32	1.19	1.20	1.20	1%	0.63	3.51	3.52	3.53	2%
300	0.42	1.44	1.44	1.48	2%	0.84	4.08	4.08	4.24	14%
200	0.63	2.16	2.14	2.36	8%	>1		—		

CJN-estimates. 'Demeaning' the KKS-estimates, in contrast, may impact MSPE in the opposite directions: it slightly improves figures when the true regression is contained in the model set, but slightly worsens them when it is not; these tendencies are, however, barely noticeable.

6 Empirical illustration

We demonstrate how the AIC model selection tools work with the famous empirical model of Donohue III and Levitt (2001) that relates crime rate to abortion rate. In the first differences form advocated in Belloni, Chernozhukov, and Hansen (2014), the regression reads

$$c_{s,t} - c_{s,t-1} = \alpha \left(a_{s,t} - a_{s,t-1} \right) + z_{s,t}' \gamma + \tau_t + \varepsilon_{s,t},$$

where $c_{s,t}$ is crime rate in state s at time t, $a_{s,t}$ is abortion rate in state s at time t, $z_{s,t}$ is a vector of covariates, τ_t is a fixed time effect, and $\varepsilon_{s,t}$ is an idiosyncratic component. We utilize the same cleaned database containing 576 observations in total, for 46 states and 13 years. The model is estimated for three types of crimes: violent crime, property

Table 7: Average values of out-of-sample MSPE, heteroskedastic case, $n = 800$, Gaussian errors, from simulations

d	AICc	AICm	$\text{AICm}^{\text{F}}_{\text{KKS}}$	$\text{AICm}^{\text{F}}_{\text{CJN}}$	$\text{AICm}^{\text{F}}_{\text{KKS}-}$
	true regression is not contained in model set				
2	0.176 ÷1.00	0.122 ÷0.69	0.123 ÷0.70	0.125 ÷0.71	0.124 ÷0.70
3	0.690 ÷1.00	0.387 ÷0.56	0.394 ÷0.57	0.402 ÷0.58	0.395 ÷0.57
4	1.67 ÷1.00	1.12 ÷0.67	1.12 ÷0.67	1.14 ÷0.68	1.13 ÷0.67
5	3.78 ÷1.00	3.22 ÷0.85	3.18 ÷0.84	3.28 ÷0.87	3.19 ÷0.85
	true regression is contained in model set				
2	0.166 ÷1.00	0.111 ÷0.67	0.113 ÷0.68	0.114 ÷0.69	0.113 ÷0.68
3	0.537 ÷1.00	0.266 ÷0.50	0.272 ÷0.51	0.278 ÷0.52	0.271 ÷0.51
4	0.882 ÷1.00	0.496 ÷0.56	0.500 ÷0.57	0.504 ÷0.57	0.499 ÷0.57
5	1.204 ÷1.00	0.780 ÷0.65	0.783 ÷0.65	0.789 ÷0.66	0.781 ÷0.65

crime, and murder.

Of ultimate interest is the treatment coefficient α, so the abortion variable is present in all specifications. Specifications otherwise differ in which covariates, if any, are included, and in whether the time effects are included (if not, then a constant is included). The "full" model contains all of these; the other specifications have compositions of covariates and/or time effects selected by one of the three AIC criteria – AICo, AICc, AICm. The selection set spans all combinations of the basic covariates – state-level controls `prison`, `police`, `unemp`, `income`, `pover`, `afdc15`, `gunlaw`, `beer`, as well as their squares $\texttt{prison}^2$, $\texttt{police}^2$, $\texttt{unemp}^2$, $\texttt{income}^2$, $\texttt{pover}^2$, $\texttt{afdc15}^2$, $\texttt{beer}^2$.[14] Additionally, we allow combinations of various blocks of variables – `intcov` containing interactions among the basic

[14]These covariates stand for: `prison` – the first difference of log(prisoners per capita), `police` – the first difference of log(police per capita), `unemp` – the first difference of state unemployment rate, `income` – the first difference of log(state income per capita), `pover` – the first difference of poverty rate, `xxafdc15` – AFDC generosity 15 years ago, `gunlaw` – the first difference of shall-issue concealed weapons law, `beer` – first difference of beer consumption per capita. The list does not include squared `gunlaw` as its underlying variable is a dummy.

covariates, `inicond` containing initial conditions, `intrend` containing interactions of all the previously mentioned variables with a quadratic trend, and, finally, `tdum` containing 12 dummies for year fixed effects. A choice of such blocks of controls, in addition to the basic covariates used by Donohue III and Levitt (2001), is inspired by very-many-regressor experiments in Belloni, Chernozhukov, and Hansen (2014). In the full specification, there are $m = 220$ regressors, and with the sample size $n = 576$, it is viewed as a really large model.

Table 8 shows specifications selected by the three AIC crireria (rows "AICo," "AICc," "AICm"), in addition to the full specification (rows "no selection") – the total number of regressors, its ratio to the sample size, the composition of covariates (in addition to the treatment variable), and the OLS coefficient estimate for the treatment variable. The standard errors given in parentheses are robust to many covariates and heteroskedasticity, see Jochmans (2022).

Consistent with our discussion in Section 3, all the AIC criteria unselect all possible combinations of covariate blocks that would make a model really large, and instead prefer much smaller models, which could be characterized as moderately large. AICc tends to select a bit smaller specification than AICo does, discarding two, one or zero covariates. AICm may or may not shrink the model further. In one of the non-homicidal crime equations, AICm rejects one more covariate; in the other it agrees with AICc, while in the murder equation AICm unselects a whole lot more of covariates, in particular, the time effect dummies. The fact that the model complexity selected by AICm is (weakly) lower that selected by AICc is consistent with a positive, though small, sample correlation coefficient (around $0.02 \div 0.05$) between the leverage values and robust estimates of individual variances computed from the full model.

Finally, one can see in the last column how model selection affects the treatment coefficient estimates and their standard errors. The removal of many initially employed covariates changes the estimates enormously, nearly halving them, while the standard errors shrink by ten to twenty times. Further unselection of one or two covariates for the non-homicidal crime affects estimates in the third digit after the decimal point, while unselection of even one, or removal of a dozen, in the murder equation is able to change the second digit after the decimal point. In the last case, such unselection makes the treatment estimate much sharper, almost halving its standard error.

Table 8: Model selection and estimation results in the illustrative empirical application

selection criteria	m	m/n	selected covariates	treatment
			violent crime	
no selection	220	0.382	all	−0.270 (0.461)
AICo	15	0.026	unemp, unemp^2, tdum	−0.151 (0.042)
AICc	13	0.023	tdum	−0.155 (0.041)
AICm	13	0.023	tdum	−0.155 (0.041)
			property crime	
no selection	220	0.382	all	−0.219 (0.461)
AICo	15	0.026	unemp, prison, tdum	−0.107 (0.022)
AICc	15	0.026	unemp, prison, tdum	−0.107 (0.022)
AICm	14	0.024	unemp, tdum	−0.109 (0.022)
			murder	
no selection	220	0.382	all	−0.639 (1.952)
AICo	17	0.030	police, income, income^2, afdc15, tdum	−0.341 (0.182)
AICc	16	0.028	police, income, income^2, tdum	−0.331 (0.182)
AICm	4	0.007	income, afdc15	−0.365 (0.098)

7 Concluding remarks

We have extended the Akaike information criterion originally derived for the classical normal linear regression, characterized by conditional homoskedasticity and a negligible number of regressors compared to sample size, to the general case that takes account of conditional heteroskedasticity and regressor numerosity, up to proportionality to sample size. Under suitable conditions, the heteroskedasticity-consistent AICm criterion possesses the asymptotic prediction optimality property in a heteroskedastic environment

with possible misspecification, while the classical AIC criteria share such a property only in restrictive circumstances. The really large models, those with parameter dimension proportional to a number of observations, tend to be unselected by AICm in favor of moderately large models. The general AICm penalty depends on all the individual error variances, and can be operationalized via their unbiased estimation, following the recent developments in the literature on linear regressions with many regressors/covariates. In simulations, the feasible AICm tends to select models that deliver better out-of-sample predictions than the original and corrected AIC.

Our setup presumes linear regression modeling even when the true regression is non-linear; the difference is attributed to misspecification error, which is possible to keep under control. An obvious venue of prospective research is a generalization to truly non-linear high-dimensional models. Even abstracting from computational issues when there are many parameters, they pose additional challenges, at the levels of both AICm construction and parameter estimation. First, it appears that derivation of the AICm penalty needs to lean on the Taylor expansion tools, which ties one's hands and forces consideration of only moderately large models; even then, it is not a fact that the AICm penalty will depend solely on individual variances and not on the shape of the regression function in addition. Second, estimation of high-dimensional nonlinear regressions (e.g., Sur and Candès 2019) and even indexes that are linear in moderately many parameters and enter the regression function (or more generally, moment conditions) non-linearly (e.g., Cattaneo, Jansson and Ma 2019) creates asymptotic biases that complicate the theory and call for measures of their elimination (e.g., jackknifing), which are proved to be working in restricted circumstances, in particular, when parameters are only moderately many. Thus, it seems that extensions of AICm may be possible only for moderately large models, excluding really large models. On the other hand, the fact that under linear regression modeling, AICm tends to unselect really large models in favor of moderately large ones, undermines constructiveness of setups with really large nonlinear models in the context of model selection.

Another challenge is adaptation to models with conditionally dependent data. We conjecture that the methods of this article can be extended to clustered dependence with possibly expanding clusters, as many methods used here have cluster analogues (e.g., leave-cluster-out machinery). However, for stationary time series data, the absence of strict exogeneity combined with really many regressors creates parameter estimation biases, which are tricky to remove (Mikusheva and Sølvsten 2025). We remain pessimistic

about extensions to situations of time series endogeneity even if estimation problems can be fixed, as even AICm penalty derivation remains a challenge because of lack of conditioning tools because of endogeneity, which are available under independence.

Acknowledgements

This article is dedicated to the memory of Grigory Kosenok, my friend and co-author. I am grateful to the Editor and Associate Editor for helpful guidance, and two anonymous referees for their insightful suggestions, which greatly improved the presentation. I also thank audiences at numerous events for useful comments and discussions, especially at the 2024 Aarhus Workshop in Econometrics, 2025 World Congress of the Econometric Society, and Econometrics Seminar at Erasmus University Rotterdam. Filip Staněk and Elizaveta Volgina provided excellent research assistance.

A Appendix

We introduce some notation to be used throughout. Define the quantities

$$\Delta_{P\sigma,n} = \sum_{i=1}^{n} P_{ii}\left(\hat{\sigma}_i^2 - \sigma_i^2\right), \quad \Delta_{M\sigma,n} = \sum_{i=1}^{n} M_{ii}\left(\hat{\sigma}_i^2 - \sigma_i^2\right) = (n-m)\left(\hat{\sigma}^2 - \bar{\sigma}_M^2\right).$$

A.1 Auxiliary Lemmas

Lemma B. (i) Under Assumption 1, $\hat{\sigma}^2 = \bar{\sigma}_M^2 + O_P\left(\Delta_g + 1/\sqrt{n}\right)$; (ii) Under Assumptions 1 and 2(i), $\Delta_{P\sigma,n} = O_P\left(\pi_4 n\sqrt{\Delta_g} + \pi_4\sqrt{n}\right)$ and $\Delta_{M\sigma,n} = O_P\left(n\Delta_g + \sqrt{n}\right)$.

Proof of Lemma B. (i) Because

$$\hat{\sigma}^2 = \frac{(g+e)' M (g+e)}{n-m} = \frac{e'Me}{n-m} + \frac{2g'Me}{n-m} + \frac{g'Mg}{n-m}, \tag{A.1}$$

we have $E\left[\hat{\sigma}^2\right] = \bar{\sigma}_M^2 + (1-m/n)^{-1}\Delta_g$ and

$$\begin{aligned}(n-m)^2 var\left(\hat{\sigma}^2\right) &= \sum_{i=1}^{n} M_{ii}^2 E\left[e_i^4\right] + 2\sum_{i=1}^{n}\sum_{j\neq i} M_{ij}^2\sigma_i^2\sigma_j^2 + 4\sum_{i=1}^{n}(Mg)_i^2\sigma_i^2 \\ &\leq \overline{C}_8^{1/2}\sum_{i=1}^{n} M_{ii} + 2\overline{C}_\sigma^2\sum_{i=1}^{n} M_{ii} + 4\overline{C}_\sigma g'Mg \leq O_P\left(n+n+n\Delta_g\right),\end{aligned}$$

using Assumption 1(ii,iii) and properties of M, and the conclusion in (i) follows.

(ii) Note that $\Delta_{P\sigma,n} = a_{P\sigma 1,n} + a_{P\sigma 2,n} + a_{P\sigma 3,n} + a_{P\sigma 4,n} + a_{P\sigma 5,n}$, where

$$\begin{aligned}a_{P\sigma 1,n} &= \sum_{i=1}^{n} P_{ii}\left(e_i^2 - \sigma_i^2\right),\ a_{P\sigma 2,n} = \sum_{i=1}^{n}\frac{P_{ii}}{M_{ii}}e_i\sum_{j\neq i} M_{ij}e_j,\ a_{P\sigma 3,n} = \sum_{i=1}^{n}\frac{P_{ii}}{M_{ii}}\sum_{j=1}^{n} g_i M_{ij}e_j, \\ a_{P\sigma 4,n} &= \sum_{i=1}^{n}\frac{P_{ii}}{M_{ii}}(Mg)_i e_i,\ a_{P\sigma 5,n} = \sum_{i=1}^{n}\frac{P_{ii}}{M_{ii}}g_i(Mg)_i.\end{aligned}$$

Note that $E\left[a_{P\sigma 1,n}\right] = E\left[a_{P\sigma 2,n}\right] = E\left[a_{P\sigma 3,n}\right] = E\left[a_{P\sigma 4,n}\right] = 0$. Compute their variances:

$$var\left(a_{P\sigma 1,n}\right) = \sum_{i=1}^{n} P_{ii}^2 var\left(e_i^2 - \sigma_i^2\right) \leq \sum_{i=1}^{n} P_{ii}^2 E\left[e_i^4\right] \leq \overline{C}_4\sqrt{\sum_{i=1}^{n} 1^2}\sqrt{\sum_{i=1}^{n} P_{ii}^4} = O_P\left(n\pi_4^2\right)$$

and

$$var\left(a_{P\sigma 2,n}\right) = 2\sum_{i=1}^{n}\frac{P_{ii}^2}{M_{ii}^2}\sum_{j\neq i} M_{ij}^2\sigma_i^2\sigma_j^2 \leq 2\underline{C}_M^{-2}\overline{C}_\sigma^2\sum_{i=1}^{n} P_{ii}^2 M_{ii} = O_P\left(n\pi_4^2\right),$$

using the Cauchy-Schwarz inequality, Assumptions 1(i,ii,iii) and 2(i), and properties of P and M. Next,

$$\begin{aligned} var\left(a_{P\sigma 3,n}\right) &= \sum_{j=1}^{n}\left(\sum_{i=1}^{n}\frac{P_{ii}M_{ij}}{M_{ii}}g_i\right)^2\sigma_j^2 \leq \overline{C}_\sigma\sum_{i=1}^{n}\frac{P_{ii}}{M_{ii}}g_i\sum_{k=1}^{n}\frac{P_{kk}}{M_{kk}}g_k\sum_{j=1}^{n}M_{ij}M_{kj} \\ &= \overline{C}_\sigma\sum_{i=1}^{n}\frac{P_{ii}}{M_{ii}}g_i\sum_{k=1}^{n}\frac{P_{kk}}{M_{kk}}g_kM_{ik} = \overline{C}_\sigma g'\mathrm{dg}\left(P\oslash M\right)M\,\mathrm{dg}\left(P\oslash M\right)g \\ &\leq \overline{C}_\sigma\sigma\left(M\right)g'\mathrm{dg}\left(P\oslash M\right)^2 g \leq \frac{\overline{C}_\sigma}{\underline{C}_M^2}\sqrt{\sum_{i=1}^{n}g_i^4}\sqrt{\sum_{i=1}^{n}P_{ii}^4} = O_P\left(n\pi_4^2\right), \end{aligned}$$

using the Cauchy-Schwarz inequality, Assumptions 1(i,ii) and 2(i), and properties of P and M. Here, the spectral radius of M is $\sigma\left(M\right)=1$. Next,

$$\begin{aligned} var\left(a_{P\sigma 4,n}\right) &= \sum_{i=1}^{n}\frac{P_{ii}^2}{M_{ii}^2}\left(Mg\right)_i^2\sigma_i^2 \leq \underline{C}_M^{-2}\overline{C}_\sigma\sum_{i=1}^{n}P_{ii}^2\left(Mg\right)_i^2 \\ &\leq \underline{C}_M^{-2}\overline{C}_\sigma\sqrt{\sum_{i=1}^{n}\left(Mg\right)_i^4}\sqrt{\sum_{i=1}^{n}P_{ii}^4} \leq O_P\left(n^{3/2}\pi_4^2\Delta_g\right), \end{aligned}$$

because $\sum_{i=1}^{n}\left(Mg\right)_i^4 \leq \left(\sum_{i=1}^{n}\left(Mg\right)_i^2\right)^2 = n^2\Delta_g^2$, and using the Cauchy-Schwarz inequality, Assumptions 1(i,ii) and 2(i), and properties of P and M. Finally,

$$a_{P\sigma 5,n} = g'\mathrm{dg}\left(P\oslash M\right)Mg \leq \underline{C}_M^{-1}\sqrt[4]{\sum_{i=1}^{n}g_i^4}\sqrt[4]{\sum_{i=1}^{n}P_{ii}^4}\sqrt{g'Mg} \leq O_P\big(n\pi_4\sqrt{\Delta_g}\big),$$

using the Cauchy-Schwarz inequality, Assumptions 1(i) and 2(i), and properties of P and M. To summarize, $\Delta_{P\sigma,n} = O_P\big(\pi_4 n\sqrt{\Delta_g}+\pi_4\sqrt{n}\big)$. For $\Delta_{M\sigma,n}$, we follow the same lines with M_{ii} replacing P_{ii} to get $\Delta_{M\sigma,n} = a_{M\sigma 1,n}+a_{M\sigma 2,n}+a_{M\sigma 3,n}+a_{M\sigma 4,n}+a_{M\sigma 5,n}$, where

$$\begin{aligned} a_{M\sigma 1,n} &= \sum_{i=1}^{n}M_{ii}\left(e_i^2-\sigma_i^2\right),\ a_{M\sigma 2,n} = \sum_{i=1}^{n}e_i\sum_{j\neq i}M_{ij}e_j,\ a_{M\sigma 3,n} = \sum_{i=1}^{n}\sum_{j=1}^{n}g_iM_{ij}e_j, \\ a_{M\sigma 4,n} &= \sum_{i=1}^{n}\left(Mg\right)_i e_i,\ a_{M\sigma 5,n} = \sum_{i=1}^{n}g_i\left(Mg\right)_i, \end{aligned}$$

with $E\left[a_{M\sigma 1,n}\right] = E\left[a_{M\sigma 2,n}\right] = E\left[a_{M\sigma 3,n}\right] = E\left[a_{M\sigma 4,n}\right] = 0$, $var\left(a_{M\sigma 1,n}\right) = O_P\left(n\right)$, $var\left(a_{M\sigma 2,n}\right) = O_P\left(n\right)$, $var\left(a_{M\sigma 3,n}\right) = O_P\left(n\Delta_g\right)$, $var\left(a_{M\sigma 4,n}\right) = O_P\left(n\Delta_g\right)$, and $a_{P\sigma 5,n} = O_P\left(n\Delta_g\right)$, leading to $\Delta_{M\sigma,n} = O_P\left(n\Delta_g+\sqrt{n}\right)$.

Lemma C. Under Assumption 1 and $\Delta_g = 0$, (i) $E\big[\left|\hat{\sigma}^2-\bar{\sigma}_M^2\right|^3\big] = O_P\left(n^{-3/2}\right)$ and $E\big[\left(\hat{\sigma}^2-\bar{\sigma}_M^2\right)^4\big] = O_P\left(n^{-2}\right)$; (ii) $E\big[\left(e'Pe\right)^3\big] \leq O_P\left(m^3\right)$.

Proof of Lemma C. (i) We have $E[e_i] = 0$ and $E\left[|e_i|^{2r}\right] \leq \overline{C}_{2r}$ for $r \leq 4$ by Assumption 1(iii). Then, by Whittle's inequality, for $2 \leq r \leq 4$, we obtain

$$E\big[\,|e'Me - E[e'Me]|^r\,\big] \leq \overline{C}_{2r}\mathrm{tr}\,(M'M)^{r/2},$$

which implies

$$E\big[\,\big|\hat{\sigma}^2 - \bar{\sigma}_M^2\big|^r\,\big] \leq \overline{C}_{2r}n^{-r/2}.$$

Setting $r = 3$ and $r = 4$, we obtain the statements in (i).

(ii) Useful statements can be derived for the moments $E\big[\,|e'Pe - E[e'Pe]|^r\,\big]$ like for $E\big[\,|\hat{\sigma}^2 - \bar{\sigma}_M^2|^r\,\big]$ above, namely

$$E\big[\,|e'Pe - E[e'Pe]|^r\,\big] \leq \overline{C}_{2r}\mathrm{tr}\,(P'P)^{r/2} = O_P\left(m^{r/2}\right).$$

In particular, $E\big[\,|e'Pe - E[e'Pe]|^3\,\big] = O_P\left(m^{3/2}\right)$, $var\,(e'Pe) = E\big[\,(e'Pe - E[e'Pe])^2\,\big] = O_P(m)$, and $E[|e'Pe - E[e'Pe]|] = O_P\left(m^{1/2}\right)$. Taking also into account that $E[e'Pe] = \sum_{i=1}^n P_{ii}\sigma_i^2 \leq \overline{C}_\sigma \sum_{i=1}^n P_{ii} = O_P(m)$ using also Assumption 1(ii) and properties of P, we obtain the statement in (ii):

$$\begin{aligned} E\big[\,(e'Pe)^2\,\big] &= E[e'Pe]^2 + var\,(e'Pe) = O_P\left(m^2\right) + O_P(m) \\ &= O_P\left(m^2\right) \end{aligned}$$

and

$$\begin{aligned} E\big[\,(e'Pe)^3\,\big] &= E\big[\,(e'Pe - E[e'Pe] + E[e'Pe])^3\,\big] \\ &\leq E\big[\,|e'Pe - E[e'Pe]|^3\,\big] + 3E\big[\,(e'Pe - E[e'Pe])^2\,\big]E[e'Pe] \\ &\quad + 3\,E[|e'Pe - E[e'Pe]|]\,E[e'Pe]^2 + E[e'Pe]^3 \\ &= O_P\left(m^3\right). \end{aligned}$$

□

Lemma S. Under Assumption 1 and $\Delta_g = 0$,

$$E\left[\frac{\bar{\sigma}_M^2}{\hat{\sigma}^2}\right] = 1 + O_P\left(\frac{1}{n}\right)$$

and

$$E\left[\frac{\bar{\sigma}_M^2}{\hat{\sigma}^2}e'Pe\right] = m\bar{\sigma}_P^2 + O_P\left(\left(\frac{m}{n}\right)^{1/2}\right).$$

Proof of Lemma S. Note that thanks to uniform integrability in Assumption 1(iv), $E\left[(e'Me)^{-r}\right] = O_P\big(n^{-r}\big)$ for all $r \leq 6$, and hence $E\left[1/\hat{\sigma}^4\right]$ and $E\left[1/\hat{\sigma}^{12}\right]$ are $O_P\left(1\right)$. Now consider the identity

$$\begin{aligned}\frac{\bar{\sigma}_M^2}{\hat{\sigma}^2} &= 1 - \frac{\hat{\sigma}^2 - \bar{\sigma}_M^2}{\hat{\sigma}^2} = 1 - \frac{\hat{\sigma}^2 - \bar{\sigma}_M^2}{\bar{\sigma}_M^2}\frac{\bar{\sigma}_M^2}{\hat{\sigma}^2} = 1 - \frac{\hat{\sigma}^2 - \bar{\sigma}_M^2}{\bar{\sigma}_M^2}\left(1 - \frac{\hat{\sigma}^2 - \bar{\sigma}_M^2}{\hat{\sigma}^2}\right) \\ &= 1 - \frac{\hat{\sigma}^2 - \bar{\sigma}_M^2}{\bar{\sigma}_M^2} + \frac{\left(\hat{\sigma}^2 - \bar{\sigma}_M^2\right)^2}{\hat{\sigma}^2\bar{\sigma}_M^2}. \qquad \text{(A.2)}\end{aligned}$$

Then, we have

$$E\left[\frac{\bar{\sigma}_M^2}{\hat{\sigma}^2}\right] = 1 - E\left[\frac{\hat{\sigma}^2 - \bar{\sigma}_M^2}{\bar{\sigma}_M^2}\right] + E\left[\frac{\left(\hat{\sigma}^2 - \bar{\sigma}_M^2\right)^2}{\hat{\sigma}^2\bar{\sigma}_M^2}\right] = 1 + O_P\left(\frac{1}{n}\right),$$

because the second term is zero, and the third term, using the Hölder inequality and Lemma C(i), is bounded by

$$\begin{aligned}E\left[\frac{\left(\hat{\sigma}^2 - \bar{\sigma}_M^2\right)^2}{\hat{\sigma}^2\bar{\sigma}_M^2}\right] &\leq \frac{1}{\underline{C}_\sigma}E\left[\frac{1}{\hat{\sigma}^4}\right]^{1/2} E\big[\left(\hat{\sigma}^2 - \bar{\sigma}_M^2\right)^4\big]^{1/2} \\ &= O_P\left(1\right)^{1/2} O_P\left(n^{-2}\right)^{1/2} = O_P\left(n^{-1}\right).\end{aligned}$$

So, the first statement of this Lemma follows. For the second statement, we have, using the identity (A.2),

$$E\left[\frac{\bar{\sigma}_M^2}{\hat{\sigma}^2}e'Pe\right] = E\left[e'Pe\right] - E\left[\frac{\hat{\sigma}^2 - \bar{\sigma}_M^2}{\bar{\sigma}_M^2}e'Pe\right] + E\left[\frac{\left(\hat{\sigma}^2 - \bar{\sigma}_M^2\right)^2}{\hat{\sigma}^2\bar{\sigma}_M^2}e'Pe\right].$$

The first term here equals $m\bar{\sigma}_P^2$. Let us find upper bounds for the second and third terms. For the second term, let us look at

$$\begin{aligned}E\left[\Delta_{M\sigma,n}e'Pe\right] &= E\left[\Big(\sum_{i=1}^n M_{ii}\left(e_i^2 - \sigma_i^2\right) + \sum_{i\neq j} M_{ij}e_ie_j\Big)\sum_{i=1}^n\sum_{j=1}^n P_{ij}e_ie_j\right] \\ &= \sum_{i=1}^n P_{ii}M_{ii}\left(E\left[e_i^4\right] - \sigma_i^4\right) + 2\sum_{i\neq j} M_{ij}P_{ij}\sigma_i^2\sigma_j^2.\end{aligned}$$

By Assumption 1(iii), the first sum is bounded above by $m\overline{C}_4$, while, using Assumption 1(ii), the second sum is bounded above in absolute value by $2\overline{C}_\sigma^2$ times

$$\sum_{i\neq j}|M_{ij}P_{ij}| \leq \Big(\sum_{i\neq j}M_{ij}^2\Big)^{1/2}\Big(\sum_{i\neq j}P_{ij}^2\Big)^{1/2} = O_P\big((mn)^{1/2}\big),$$

as $\sum_{i=1}^n\sum_{j=1,j\neq i}^n M_{ij}^2 = \sum_{i=1}^n\left(M_{ii} - M_{ii}^2\right) < \sum_{i=1}^n M_{ii} = \mathrm{rk}\left(M\right) = n - m$ and similarly $\sum_{i=1}^n\sum_{j=1,j\neq i}^n P_{ij}^2 < \mathrm{rk}\left(P\right) = m$. Hence,

$$E\left[\frac{\hat{\sigma}^2 - \bar{\sigma}_M^2}{\bar{\sigma}_M^2}e'Pe\right] = \frac{1}{\left(n-m\right)\bar{\sigma}_M^2}E\left[\Delta_{M\sigma,n}e'Pe\right] = O_P\big((m/n)^{1/2}\big).$$

Finally, for the third term, using the Hölder inequality and Lemma C(i,ii),

$$\begin{aligned} E\left[\frac{\left(\hat{\sigma}^2-\bar{\sigma}_M^2\right)^2}{\hat{\sigma}^2\bar{\sigma}_M^2}e'Pe\right] &\leq \frac{1}{\bar{\sigma}_M^2}E\left[\frac{1}{\hat{\sigma}^{12}}\right]^{1/6}E\big[\left(\hat{\sigma}^2-\bar{\sigma}_M^2\right)^4\big]^{1/2}E\big[\left(e'Pe\right)^3\big]^{1/3} \\ &\leq \underline{C}_\sigma^{-1}O_P\left(1\right)^{1/6}O_P\left(n^{-2}\right)^{1/2}O_P\left(m^3\right)^{1/3}=O_P\left(m/n\right). \end{aligned}$$

After collecting the terms, we get the second statement of the Lemma. □

A.2 Proofs of Theorems

Proof of Theorem Am. Let $\hat{\beta}_{QML}$ and $\hat{\sigma}^2_{QML}$ be Gaussian QML estimates:

$$\hat{\beta}_{QML}=\left(X'X\right)^{-1}X'y$$

and

$$\hat{\sigma}^2_{QML}=\frac{\left(y-X\hat{\beta}_{QML}\right)'\left(y-X\hat{\beta}_{QML}\right)}{n}=\frac{n-m}{n}\hat{\sigma}^2.$$

The (twice) Gaussian quasi-loglikelihood fit to in-sample data is equal to

$$\begin{aligned} 2\log L\big(y,X,\hat{\beta}_{QML},\hat{\sigma}^2_{QML}\big) &= \left[-n\log\left(2\pi\sigma^2\right)-\frac{\left(y-X\beta\right)'\left(y-X\beta\right)}{\sigma^2}\right]_{\hat{\beta}_{QML},\hat{\sigma}^2_{QML}} \\ &= -n\log\left(2\pi\hat{\sigma}^2_{QML}\right)-\frac{n\hat{\sigma}^2_{QML}}{\hat{\sigma}^2_{QML}}, \end{aligned}$$

having expectation

$$E\big[2\log L\big(y,X,\hat{\beta}_{QML},\hat{\sigma}^2_{QML}\big)\big]=-nE\left[\log\left(2\pi\hat{\sigma}^2_{QML}\right)\right]-n.$$

The estimated (twice) Gaussian quasi-loglikelihood $2\log L\big(y^*,X,\hat{\beta}_{QML},\hat{\sigma}^2_{QML}\big)$ fit to out-of-sample data $y^*=X\beta+e^*$, where $e^*|\mathcal{X}\sim\left(0,\mathrm{dg}\left\{\sigma_i^2\right\}_{i=1}^n\right)$ is conditionally independent of y, is equal to

$$\begin{aligned} &\left[-n\log\left(2\pi\sigma^2\right)-\frac{\left(y^*-X\beta\right)'\left(y^*-X\beta\right)}{\sigma^2}\right]_{\hat{\beta}_{QML},\hat{\sigma}^2_{QML}} \\ =\ & -n\log\left(2\pi\hat{\sigma}^2_{QML}\right)-\frac{e^{*\prime}e^*+\left(\hat{\beta}_{QML}-\beta\right)'X'X\left(\hat{\beta}_{QML}-\beta\right)-2e^{*\prime}X\left(\hat{\beta}_{QML}-\beta\right)}{\hat{\sigma}^2_{QML}}. \end{aligned}$$

The conditional expectation of the second term, apart from the minus sign, is

$$\begin{aligned} \mathcal{E}_2 &\equiv E\left[\frac{e^{*\prime}e^*+\left(\hat{\beta}_{QML}-\beta\right)'X'X\left(\hat{\beta}_{QML}-\beta\right)-2e^{*\prime}X\left(\hat{\beta}_{QML}-\beta\right)}{\hat{\sigma}^2_{QML}}\right] \\ &= E\left[\frac{1}{\hat{\sigma}^2_{QML}}\right]E\left[e^{*\prime}e^*\right]+E\left[\frac{\left(\hat{\beta}-\beta\right)'X'X\left(\hat{\beta}-\beta\right)}{\hat{\sigma}^2_{QML}}\right]-2E\left[e^*\right]'E\left[\frac{X\left(\hat{\beta}-\beta\right)}{\hat{\sigma}^2_{QML}}\right] \\ &= \frac{n}{n-m}E\left[\frac{1}{\hat{\sigma}^2}\right]\left(m\bar{\sigma}_P^2+(n-m)\bar{\sigma}_M^2\right)+\frac{n}{n-m}E\left[\frac{e'Pe}{\hat{\sigma}^2}\right], \end{aligned}$$

using conditional independence of e^* of y and thus of $\hat{\sigma}^2_{QML}$ and $X\big(\hat{\beta}-\beta\big)/\hat{\sigma}^2_{QML}$, and the properties $E\left[e^*\right]=0$ and $E\left[e^{*\prime}e^*\right]=\sum_{i=1}^n\sigma_i^2=m\bar{\sigma}_P^2+(n-m)\bar{\sigma}_M^2$, as well as the fact that $\big(\hat{\beta}-\beta\big)'X'X\big(\hat{\beta}-\beta\big)=e'Pe$. Using Lemma S, we compute that

$$\begin{aligned}\mathcal{E}_2 &= \frac{n}{n-m}\frac{m\bar{\sigma}_P^2+(n-m)\bar{\sigma}_M^2}{\bar{\sigma}_M^2}E\left[\frac{\bar{\sigma}_M^2}{\hat{\sigma}^2}\right]+\frac{n}{n-m}\frac{1}{\bar{\sigma}_M^2}E\left[\frac{\bar{\sigma}_M^2}{\hat{\sigma}^2}e'Pe\right]\\ &= \frac{n}{n-m}\frac{m\bar{\sigma}_P^2+(n-m)\bar{\sigma}_M^2}{\bar{\sigma}_M^2}\left(1+O_P\left(n^{-1}\right)\right)+\frac{n}{n-m}\frac{1}{\bar{\sigma}_M^2}\left(m\bar{\sigma}_P^2+O_P\big((m/n)^{1/2}\big)\right)\\ &= \frac{2mn}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2}+n+O_P\left(1\right).\end{aligned}$$

In total,

$$E\big[2\log L\big(y^*,X,\hat{\beta}_{QML},\hat{\sigma}^2_{QML}\big)\big]=-nE\left[\log\left(2\pi\hat{\sigma}^2_{QML}\right)\right]-\frac{2mn}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2}-n+O_P\left(1\right).$$

The n^{-1}-normalized (twice) difference between the expected estimated out-of-sample Gaussian quasi-loglikelihood and the expected in-sample Gaussian quasi-loglikelihood constitutes the AICm penalty together with the remainder:

$$\begin{aligned}pen_{AIC_m}+\zeta_m &= \frac{1}{n}E\big[2\log L\big(y,X,\hat{\beta}_{QML},\hat{\sigma}^2_{QML}\big)\big]-\frac{1}{n}E\big[2\log L\big(y^*,X,\hat{\beta}_{QML},\hat{\sigma}^2_{QML}\big)\big]\\ &= \frac{1}{n}\left(-nE\left[\log\left(2\pi\hat{\sigma}^2_{QML}\right)\right]-n\right)\\ &\qquad -\frac{1}{n}\left(-nE\left[\log\left(2\pi\hat{\sigma}^2_{QML}\right)\right]-\frac{2mn}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2}-n+O_P\left(1\right)\right)\\ &= \frac{2m}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2}+O_P\left(n^{-1}\right),\end{aligned}$$

as in the statement of the Theorem. The resulting value of AICm is

$$\begin{aligned}AIC_m &= -\frac{1}{n}2\log L\big(y,X,\hat{\beta}_{QML},\hat{\sigma}^2_{QML}\big)+pen_{AIC_m}+\zeta_m\\ &= -\left[-\log\left(2\pi\sigma^2\right)-\frac{\left(y-X\beta\right)'\left(y-X\beta\right)}{n\sigma^2}\right]_{\hat{\beta}_{QML},\hat{\sigma}^2_{QML}}+\frac{2m}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2}+O_P\left(n^{-1}\right)\\ &= \text{const}+\log\left(\hat{e}'\hat{e}\right)+\frac{2m}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2}+O_P\left(n^{-1}\right),\end{aligned}$$

with const not depending on the data or m. Omitting const, we have (2.5) for $f=m$. □

Proof of Corollaries AmH and AmF. See the text. □

Proof of Corollary AmN. The difference in question is

$$\begin{aligned} pen_{AIC_c} - pen_{AIC_m} &= 2\frac{m+1}{n-m-2} - 2\frac{m}{n-m} = 2\frac{1}{n-m}\frac{n+m}{n-m-2} \\ &\leq 2\frac{(1+C_m)n}{(1-C_m)n((1-C_m)n-2)} = O\left(\frac{1}{n}\right), \end{aligned}$$

thanks to Assumption 1(i). □

Proof of Theorem Om. Assumptions (A.1), (A.2*) and (A.3') of Andrews (1991) hold by Assumption 1(ii,iii) and because $nE\left[L\left(q\right)\right] = n\Delta_g\left(q\right) + \sum_{i=1}^n P_{ii}\left(q\right)\sigma_i^2 \geq \underline{C}_\sigma m\left(q\right) \to \infty$ uniformly in $q \in Q$. The intermediate results in the proof of Theorem 2.1* of Andrews (1991), which is an adaptation of the proof of Theorem 2.1 of Li (1987) to the heteroskedastic setup, yield, in particular, that $\sup_{q\in Q} n^{-1}\left|g'M\left(q\right)e\right|/L\left(q\right) = o_P\left(1\right)$ and $\sup_{q\in Q} n^{-1}\left|e'P\left(q\right)e - \sum_{i=1}^n P_{ii}\left(q\right)\sigma_i^2\right|/L\left(q\right) = o_P\left(1\right)$. Also, $\left|n^{-1}e'e - n^{-1}\sum_{i=1}^n \sigma_i^2\right| = o_P(1)$ because $E\left[e'e\right] = \sum_{i=1}^n \sigma_i^2$ and $var\left(e'e\right) \leq \sum_{i=1}^n E\left[e_i^4\right] = O\left(n\right)$ by Assumption 1(iii). Note that because

$$\begin{aligned} \left(\mu_n\left(q\right) - \frac{e'P\left(q\right)e}{e'e}\right)\frac{n^{-1}e'e}{L\left(q\right)} &= \left(\frac{e'e}{\sum_{i=1}^n \sigma_i^2} - 1\right)\frac{n^{-1}e'P\left(q\right)e}{L\left(q\right)} \\ &\quad - \frac{e'e}{\sum_{i=1}^n \sigma_i^2}\frac{n^{-1}\left(e'P\left(q\right)e - \sum_{i=1}^n P_{ii}\left(q\right)\sigma_i^2\right)}{L\left(q\right)}, \end{aligned}$$

we have

$$\begin{aligned} \sup_{q\in Q}\left|\mu_n\left(q\right) - \frac{e'P\left(q\right)e}{e'e}\right|\frac{n^{-1}e'e}{L\left(q\right)} &\leq \left|\frac{e'e}{\sum_{i=1}^n \sigma_i^2} - 1\right|\sup_{q\in Q}\frac{n^{-1}e'P\left(q\right)e}{L\left(q\right)} \\ &\quad + \frac{e'e}{\sum_{i=1}^n \sigma_i^2}\sup_{q\in Q}\frac{n^{-1}\left|e'P\left(q\right)e - \sum_{i=1}^n P_{ii}\left(q\right)\sigma_i^2\right|}{L\left(q\right)} \\ &\leq o_P(1)O_P\left(1\right) + O_P\left(1\right)o_P(1) = o_P(1), \end{aligned}$$

and also

$$\begin{aligned} \sup_{q\in Q}\frac{\mu_n\left(q\right)}{L\left(q\right)} &\leq \frac{1}{\sum_{i=1}^n \sigma_i^2}\left(\sup_{q\in Q}\frac{n^{-1}e'P\left(q\right)e}{L\left(q\right)} + \sup_{q\in Q}\frac{n^{-1}\left|\sum_{i=1}^n P_{ii}\left(q\right)\sigma_i^2 - e'P\left(q\right)e\right|}{L\left(q\right)}\right) \\ &\leq O_P\left(1\right)\left(O_P\left(1\right) + o_P\left(1\right)\right) = O_P\left(1\right). \end{aligned}$$

These will be exploited below.

Divide the model candidate set into three subsets: $Q = Q_\mu \cup Q_g \cup Q_L$, where Q_μ contains moderately large models with $\Delta_g\left(q\right) \to^P 0$, Q_g contains moderately large models with $\Delta_g\left(q\right) \nrightarrow^P 0$, and Q_L contains really large models with $\Delta_g\left(q\right) \to^P 0$. Note the following asymptotic properties of $L\left(q\right)$ over the three sets: (i) when $q \in Q_\mu$, both $\Delta_g\left(q\right)$

and $n^{-1}e'P(q)e$ are $o_P(1)$, so $L(q) = n^{-1}e'P(q)e + \Delta_g(q) = o_P(1)$; (ii) when $q \in Q_g$, we have $n^{-1}e'P(q)e = o_P(1)$ and $\Delta_g(q)$ is $O_P(1)$ but not $o_P(1)$, so $L(q) = \Delta_g(q) + o_P(1) = O_P(1)$ but not $o_P(1)$; (iii) when $q \in Q_L$, we have $n^{-1}e'P(q)e$ is $O_P(1)$ but not $o_P(1)$, and $\Delta_g(q)$ is $o_P(1)$, so $L(q) = n^{-1}e'P(q)e + o_P(1) = O_P(1)$ but not $o_P(1)$.

Consider first the case when Condition M, $\sup_{q\in Q}\mu_n(q) \to^P 0$, holds. Then, Q_L is empty, and $Q = Q_\mu \cup Q_g$.

(a) Let Q_g be empty as well so that $Q = Q_\mu$. Then, for any $q_\mu \in Q_\mu$, we have

$$\begin{aligned}\hat{e}(q)'\hat{e}(q) &= (e+g)'M(q)(e+g)\\ &= e'e + (g'M(q)g + e'P(q)e) + 2g'M(q)e - 2e'P(q)e,\end{aligned}$$

therefore

$$\frac{\hat{e}(q)'\hat{e}(q) - e'e}{e'e} = \frac{L(q)}{n^{-1}e'e} + 2\frac{n^{-1}g'M(q)e}{n^{-1}e'e} - 2\frac{n^{-1}e'P(q)e}{n^{-1}e'e} = o_P(1),$$

and hence the following monotonically transformed values of $AIC_m(q)$ is

$$\begin{aligned}\overline{AIC}_m(q) &\equiv (AIC_m(q) - \log(e'e))\frac{e'e}{n}\\ &= \log\left(1 + \frac{\hat{e}(q)'\hat{e}(q) - e'e}{e'e}\right)\frac{e'e}{n} + pen_{AIC_m}(q)\frac{e'e}{n}\\ &= \frac{\hat{e}(q)'\hat{e}(q) - e'e}{e'e}(1 + o_P(1))\frac{e'e}{n} + \frac{2\mu_n(q)}{1-\mu_n(q)}\frac{e'e}{n}\\ &= L(q) + 2\frac{g'M(q)e}{n} + o_P(L(q))\\ &\quad + 2\left(\mu_n(q) - \frac{e'P(q)e}{e'e}\right)\frac{e'e}{n} + o_p(\mu_n(q))\\ &= L(q) + o_P(L(q)),\end{aligned}$$

as $|\mu_n(q) - e'P(q)e/e'e|\, e'e/n \le o_P(L(q))$ and $\mu_n(q) \le O_P(L(q))$ uniformly in q. Hence, minimization of $\overline{AIC}_m(q)$ is asymptotically equivalent to minimization of $L(q)$ uniformly in $q \in Q_\mu = Q$.

(b) Let instead Q_μ be empty so that $Q = Q_g$. Then, for any $q \in Q_g$, we have

$$n^{-1}\hat{e}(q)'\hat{e}(q) = n^{-1}e'e + L(q) + o_P(1) = O_P(1),$$

and

$$\begin{aligned}AIC_m(q) - \log(n) &= \log\left(\frac{e'e}{n} + L(q) + o_P(1)\right) + 2\frac{\mu_n(q)}{1-\mu_n(q)}\\ &= \log\left(\frac{e'e}{n} + L(q)\right) + o_P(L(q)).\end{aligned}$$

Hence, thanks to the strict monotonicity of the log function, minimization of $AIC_m(q)$ is asymptotically equivalent to minimization of $L(q)$ uniformly in $q \in Q_g = Q$.

(c) It remains to consider the case when neither Q_μ or Q_g is empty. Then, $L(q) = o_P(1)$ when $q \in Q_\mu$ and $L(q) = O_P(1)$ but not $o_P(1)$ when $q \in Q_g$, hence $\inf_{q \in Q} L(q)$ is asymptotically reached on some model $q_\mu \in Q_\mu$. Then, $\overline{AIC}_m(q_\mu) = o_P(1)$, while for any $q_g \in Q_g$,

$$\begin{aligned} \overline{AIC}_m(q_g) &= (AIC_m(q_g) - \log(e'e))\frac{e'e}{n} \\ &= \log\left(1 + \frac{L(q_g)}{n^{-1}e'e}\right)\frac{e'e}{n} + o_P(1), \end{aligned}$$

which is $O_P(1)$ but not $o_P(1)$. Thus, AICm will prefer the model $q_\mu \in Q_\mu$ to any model in Q_g, and optimality over $Q_\mu \cup Q_g$ reduces to optimality over Q_μ, which has been already established.

Consider now the cases when $\sup_{q \in Q} \mu_n(q) \nrightarrow^P 0$, i.e. really large models are present in the candidate set, and when Condition G, $\sup_{q \in Q} \Delta_g(q) \to^P 0$, holds.

(a) Suppose $Q = Q_L$, i.e. only really large models are compared. Then, for any $q \in Q_L$,

$$\begin{aligned} n^{-1}\hat{e}(q)'\hat{e}(q) &= n^{-1}e'e - \left(\Delta_g(q) + n^{-1}e'P(q)e\right) + 2\left(\Delta_g(q) + n^{-1}g'M(q)e\right) \\ &= n^{-1}e'e - L(q) + o_P(1), \end{aligned}$$

using Condition G and that $n^{-1}g'M(q)e = o_P(L(q)) = o_P(1)$ uniformly in $q \in Q_L = Q$. Also, $n^{-1}e'P(q)e = L(q) + o_P(1)$. Therefore,

$$\begin{aligned} AIC_m(q) - \log(e'e) &= \log\left(\frac{e'e}{n} - L(q) + o_P(1)\right) - \log\left(\frac{e'e}{n}\right) \\ &\quad + 2\frac{L(q) + o_P(1)}{n^{-1}e'e - L(q) + o_P(1)} \\ &= \log\left(1 - \frac{L(q)}{n^{-1}e'e}\right) + 2\frac{L(q)/n^{-1}e'e}{1 - L(q)/n^{-1}e'e} + o_P(L(q)) \\ &= \upsilon\left(\frac{L(q)}{n^{-1}e'e}\right) + o_P(L(q)), \end{aligned}$$

where $\upsilon(\mu) = \log(1-\mu) + 2\mu/(1-\mu)$. Note that the function $\upsilon(\cdot)$ is strictly positive and strictly increasing on $(0,1)$, see Figure 2. Thanks to the strict monotonicity of $\upsilon(\cdot)$, minimization of $AIC_m(q)$ is asymptotically equivalent to minimization of $L(q)$ uniformly in $q \in Q_L$.

(b) Suppose $Q \supset Q_L$, i.e. moderately large models from $Q_\mu \cup Q_g$ are also present. Condition G rules out any models from Q_g, so Q_g is empty. We already know that, if

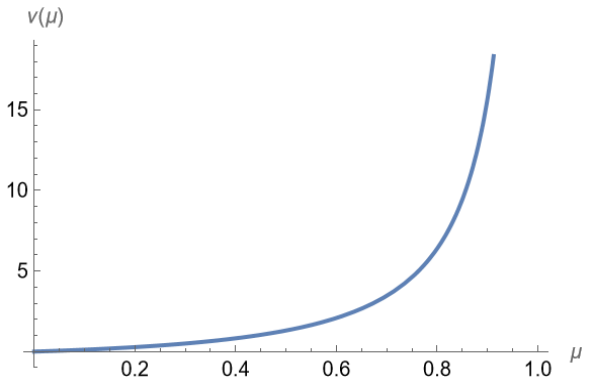

Figure 2: function $\upsilon(\mu)$ on $[0,1)$.

Q_μ is non-empty, the best model $q_\mu \in Q_\mu$ attains $AIC_m(q_\mu) - \log(e'e) = o_P(1)$. At the same time, $L(q_L) = n^{-1}e'P(q_L)e + o_P(1) = (n^{-1}e'e)\mu_n(q_L) + o_P(1)$ and hence $AIC_m(q_L) - \log(e'e) = \upsilon(\mu_n(q_L)) + o_P(1)$, which is strictly positive asymptotically, thanks to the properties of the function $\upsilon(\cdot)$. Thus, AICm will prefer the best model from Q_μ to any model in Q_L, and optimality over Q reduces to optimality over Q_μ, which has been already established.

To summarize, taking into account that certain models are unselected, minimization of $AIC_m(q)$ is asymptotically equivalent to minimization of $L(q)$ uniformly in $q \in Q$, and hence the minimizer $\hat{q}_m$ is asymptotically optimal over Q. □

Proof of Theorem Oc. The absolute difference between $pen_{AIC_m}(q)$ and $pen_{AIC_c}(q)$ is

$$\begin{aligned} 2\left|\frac{\mu_n(q)}{1-\mu_n(q)} - \frac{m(q)}{n-m(q)}\right| &= 2\left|\frac{\mu_n(q) - m(q)/n}{(1-\mu_n(q))(1-m(q)/n)}\right| \\ &\geq 2\left|\mu_n(q) - \frac{m(q)}{n}\right|, \text{ and} \\ &\leq \frac{2}{\left(1 - C_m\overline{C}_\sigma/\underline{C}_\sigma\right)(1-C_m)}\left|\mu_n(q) - \frac{m(q)}{n}\right|, \end{aligned}$$

using Assumption 1(i,ii), so the absolute difference is of the same asymptotic order as that of

$$\mu_n(q) - \frac{m(q)}{n} = \frac{n^{-1}\sum_{i=1}^n (P_{ii}(q) - m(q)/n)\sigma_i^2}{n^{-1}\sum_{i=1}^n \sigma_i^2} = \frac{\widehat{cov}(P_{ii}(q), \sigma_i^2)}{n^{-1}\sum_{i=1}^n \sigma_i^2},$$

and, as $\underline{C}_\sigma \leq n^{-1}\sum_{i=1}^n \sigma_i^2 \leq \overline{C}_\sigma$, by Condition P, has no asymptotic effect of $AIC_m(q)$ uniformly in $q \in Q$. Further,

$$\left|\mu_n(q) - \frac{m(q)}{n}\right| \leq \frac{\sum_{i=1}^n |P_{ii}(q) - m(q)/n|\sigma_i^2}{\sum_{i=1}^n \sigma_i^2} \leq \frac{\overline{C}_\sigma}{\underline{C}_\sigma}\frac{\sum_{i=1}^n |P_{ii}(q) - m(q)/n|}{n},$$

which has the same asymptotic order as $\Delta_P(q)$. □

Proof of Theorem Oo. Suppose first that Condition M holds. The absolute difference between $pen_{AIC_o}(q)$ and $pen_{AIC_c}(q)$ is

$$2\left|\frac{m(q)}{n-m(q)}-\frac{m(q)}{n}\right| \leq \frac{2}{(1-C_m)}\frac{m(q)^2}{n^2} = o_P(m(q)/n)$$

by Condition M, and so has no asymptotic effect of $AIC_m(q)$ uniformly in $q \in Q$.

Suppose now that Condition G holds. Then, Q_g is empty, but the set Q_L may not be empty. If Q_L is empty, then Condition M holds and the best model from Q_μ achieves optimality. If Q_L is non-empty, then for any $q \in Q_L$,

$$\begin{aligned} AIC_m(q_L) - \log(e'e) &= \log\left(1-\frac{L(q_L)}{n^{-1}e'e}\right) + 2\frac{L(q_L)}{n^{-1}e'e} + o_P(L(q_L)) \\ &= \varphi(\mu_n(q_L)) + o_P(1), \end{aligned}$$

where $\varphi(\mu) = \log(1-\mu) + 2\mu$. Note that the function $\varphi(\cdot)$ is ∩-shaped on $[0,1)$, monotonically strictly increasing on $\left[0,\frac{1}{2}\right]$, monotonically strictly decreasing on $\left[\frac{1}{2},1\right)$, reaches zero for μ^* satisfying $\log(1-\mu^*) = -2\mu^*$, and is minimized to $-\infty$ by $\mu \to 1$, see Figure 3. Thus, for $\lim \mu_n(q)$ on $\left(0,\frac{1}{2}\right]$, minimization of $AIC_m(q_L) - \log(e'e)$ coincides with minimization of $L(q)$ uniformly in $q \in Q_L$.

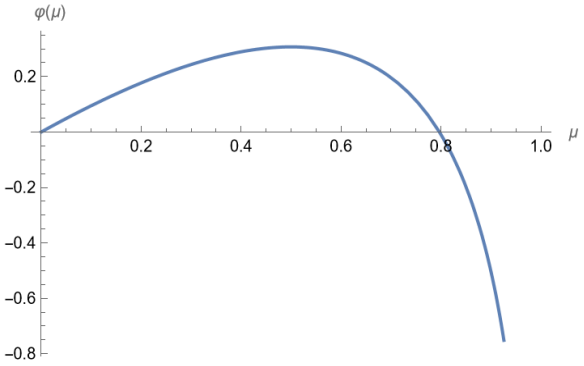

Figure 3: function $\varphi(\mu)$ on $[0,1)$.

Suppose $Q \supset Q_L$, i.e. moderately large models from Q_μ are also present, then the best model $q_\mu \in Q_\mu$ attains $AIC_m(q_\mu) - \log(e'e) = o_P(1)$, and asymptotically will be selected over q_L provided that $\sup_{q\in Q} \mu_n(q) < \frac{1}{2}$. □

Proof of Theorem Fm. Represent the difference in question as

$$\frac{\Delta_{pen}}{2} = \frac{\sum_{i=1}^n P_{ii}\hat{\sigma}_i^2}{\sum_{i=1}^n M_{ii}\hat{\sigma}_i^2} - \frac{\sum_{i=1}^n P_{ii}\sigma_i^2}{\sum_{i=1}^n M_{ii}\sigma_i^2} = \frac{1}{n-m}\frac{1}{\hat{\sigma}^2}\left(\Delta_{P\sigma,n} - \frac{m}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2}\Delta_{M\sigma,n}\right).$$

Assumption 1(i,ii) guarantees $\bar{\sigma}_P^2 \leq \overline{C}_\sigma$, $\bar{\sigma}_M^2 \geq \underline{C}_\sigma$ and $n - m \geq (1 - C_m)\, n$. Using the triangular inequality,

$$\left|\frac{\Delta_{pen}}{2}\right| < \frac{1}{(1-C_m)\, n}\frac{1}{\hat{\sigma}^2}\left(|\Delta_{P\sigma,n}| + \frac{m}{n-m}\frac{\bar{\sigma}_P^2}{\bar{\sigma}_M^2}\,|\Delta_{M\sigma,n}|\right).$$

By Lemma B(i), $E\left[\hat{\sigma}^2\right] = \bar{\sigma}_M^2 + O_P\left(\Delta_g\right)$ and $var\left(\hat{\sigma}^2\right) = O_P\left(n^{-1}\right)$, hence $\hat{\sigma}^2 = \bar{\sigma}_M^2 + o_P\left(1\right)$. By Lemma B(ii), $|\Delta_{P\sigma,n}| = O_P\big(n\pi_4\sqrt{\Delta_g} + \pi_4\sqrt{n}\big)$ and $|\Delta_{M\sigma,n}| = O_P\left(n\Delta_g + \sqrt{n}\right)$. To summarize, also using the Slutsky theorem,

$$\begin{aligned}\left|\frac{\Delta_{pen}}{2}\right| &< \frac{1}{(1-C_m)\, n}\frac{1}{\bar{\sigma}_M^2 + O_P\left(\Delta_g + n^{-1/2}\right)} \\ &\quad \times \left(O_P\big(n\pi_4\sqrt{\Delta_g} + \pi_4\sqrt{n}\big) + \frac{m}{(1-C_m)\, n}\frac{\overline{C}_\sigma}{\underline{C}_\sigma}O_P\left(n\Delta_g + \sqrt{n}\right)\right) \\ &= O_P\left(\pi_4\Big(\sqrt{\Delta_g} + \frac{1}{\sqrt{n}}\Big)\right),\end{aligned}$$

as π_4 dominates m/n because $m = \sum P_{ii} \leq \sqrt{\sum_{i=1}^n 1^2}\sqrt{\sum_{i=1}^n P_{ii}^2} \leq \sqrt{n}\sqrt[4]{n}\sqrt[4]{\sum_{i=1}^n P_{ii}^4} = n\pi_4$, and $\sqrt{\Delta_g}$ dominates Δ_g when $\Delta_g \leq O_P(1)$. Further, with the condition $\Delta_g = o_P\left(1\right)$ and Assumption 2(ii), we have

$$\pi_4\Big(\sqrt{\Delta_g} + \frac{1}{\sqrt{n}}\Big) \leq O_P\left(\frac{m}{n}\right) o_P\left(1\right) = o_P\left(\frac{m}{n}\right).$$

Recalling that pen_{AIC_m} has the asymptotic order of m/n, we conclude. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and F. Csaki (eds.), *Proceedings of the 2nd International Symposium on Information Theory*, Budapest: Akademiai Kiado, pp. 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

Anatolyev, S. (2021). Mallows criterion for heteroskedastic linear regressions with many regressors. *Economics Letters*, 203, 109864.

Anatolyev, S. and Yaskov, P. (2017). Asymptotics of diagonal elements of projection matrices under many instruments/regressors. *Econometric Theory*, 33(3), 717-738.

Anatolyev, S. and Sølvsten, S. (2023). Testing many restrictions under heteroskedasticity. *Journal of Econometrics*, 236(1), 105473.

Anderson, T. W. and Fang, K.-T. (1987). Cochran's theorem for elliptically contoured distributions. *Sankhyā* (Series A), 49(3), 305-315.

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.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608-650.

Boot, T. (2023). Joint inference based on Stein-type averaging estimators in the linear regression model. *Journal of Econometrics*, 235(2), 1542-1563.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53, 603-618.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A practical information-theoretic approach.* 2nd ed., Springer-Verlag.

Cattaneo, M., Jansson, M., and Ma, X. (2019). Two-step estimation and inference with possibly many included covariates. *Review of Economic Studies*, 86(3), 10951122.

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

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

Cavanaugh, J.E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, 33(2), 201-208.

Cesar, P., Vivanco, M. and Menezes, F. (2014). Information criteria: How do they behave in different models? *Computational Statistics & Data Analysis*, 69, 141-153.

Donohue III, J. J. and Levitt, S. D. (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics*, 116, 379-420.

Hurvich, C.M. and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.

Hurvich, C. M., Simonoff, J. S., and Tsai, C-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, 60, 271-293.

Jochmans, K. (2022). Heteroskedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, 117(538), 887-896.

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

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

Mikusheva, A. and Sølvsten, M. (2025). Linear regression with weak exogeneity. *Quantitative Economics*, 16, 367-403.

Rao, C.R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of American Statistical Association*, 65(329), 161-172.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2), 221-242

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, 68(1), 45-54.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics – Theory and Methods*, 7, 13-26.

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

Takeuchi, K. (1976). Distributions of information statistics and criteria for adequacy of models (in Japanese). *Suri-Kagaku (Mathematical Science)*, 153, 1218.

Figure 1: Histograms of models selected and distributions of penalties. Top row: $d = 2$ and $d = 3$, bottom row: $d = 4$ and $d = 5$.