The More Data There Is, the More a Penalised Criterion Prefers the More Complicated Model

— the epicycles begin to win at $n = 53.777041$ under AIC, agreeing to six digits with the closed form $2\Delta k$ over the logarithm of the residual ratio, and at $n = 131.109261$ under BIC — the separator is how the penalty grows with n — [Paper 701]

Yuuki YAMAGISHI (山岸 勇輝)
Yamagishi Complexity Torus Lab, Japan
September 2026

— From thinking to syncing. —

Abstract

A complicated model is supposed to be penalised. Yet inside a penalised criterion, the more data there is the more the complicated model wins.How penalty and fit grow with n is shown here through the crossing point.No new theorem or law is claimed. First, chosen on fit alone, the complicated model wins——the residuals are 0.004 against 0.005, and the side allowed 11 parameters wins (Section 2). **Second, with penalties, the two criteria disagree at $n = 100$——AIC gives -990.663110 against -980.348755 and chooses the epicycles; BIC gives -962.006238 against -967.322904 and chooses the ellipse** (Section 2). **Third, and this is the core. More data makes the complicated model win——the epicycles begin to win at $n = 53.777041$ under AIC and $n = 131.109261$ under BIC, and above that they are chosen for ever** (Section 3). **Fourth, the crossing point has a closed form——for AIC it is $2\Delta k$ over the logarithm of the residual ratio, agreeing with the bisected 53.777041 to six digits** (Section 3). **Fifth, the reason is only the rate of growth——the difference in fit grows in proportion to n while the difference in penalty grows as a constant (AIC) or a logarithm (BIC); at $n = 5000$ the AIC gap is $+1103.717757$** (Section 3). **Sixth, the separator is how the penalty grows with n——not whether there is a penalty: unless it grows in proportion to n, it is overtaken sooner or later** (Section 4). **Keywords**: model selection, AIC, BIC, penalty term, overfitting, effect of sample size.

Scope of this paper (scope note): **No new theorem or law is claimed——the definitions of AIC and BIC, the growth of their penalty terms and overfitting are all standard (Akaike 1974, Schwarz 1978).Nothing is fitted to astronomical data——the parameter counts and residuals are representative, and no claim is made about how well the historical models actually fitted.Methods of selection are not compared——cross-validation and selection by predictive error are mentioned and not compared.Paradigms are not discussed——theories of how scientific theories are replaced are not treated; only quantities.The treatment of the residual is stated——the residual sum of squares is held fixed, as a property of the model, independent of n. In reality it grows with n, so this is an idealisation.Relation to earlier papers**: **Paper 536 showed that whether a formula was borrowed or derived shows outside the working range——there the dividing line lay outside the range. Here, so long as one stays inside it, a penalty does not divide the two, and more data leans towards the wrong side. This is the other face of Paper 536.**What is added is **showing AIC and BIC to disagree at $n = 100$, giving the crossing points $\mathbf{53.777041}$ and $\mathbf{131.109261}$, showing the AIC crossing to agree with a closed form to six digits, showing the gap to open in proportion to n, reaching $+\mathbf{1103.717757}$ at $n = 5000$, and placing the separator on how the penalty grows with n.

1. Notation and the Question

Setting: **fit two models to the same data. Write k for the number of parameters and S for the residual sum of squares, and take two criteria.**

$$\mathrm{AIC} = n \ln \frac{S}{n} + 2k, \qquad \mathrm{BIC} = n \ln \frac{S}{n} + k \ln n$$

Take $k = 11$ and $S = 0.004$ for five stages of epicycles, and $k = 5$ and $S = 0.005$ for an ellipse plus perturbation. S is held fixed, independent of n, as a property of the model.

Question: **does a penalty put the complicated model away?**

2. At $n = 100$ the Two Criteria Disagree

Computation 2.1: **the values at $n = 100$ were derived.**

criterion	epicycles	ellipse	chosen
residual	$\mathbf{0.004}$	$\mathbf{0.005}$	epicycles
AIC	$\mathbf{-990.663110}$	$\mathbf{-980.348755}$	epicycles
BIC	$\mathbf{-962.006238}$	$\mathbf{-967.322904}$	ellipse

Observation 2.2

the same data and the same two models give different answers under different criteria.AIC takes the epicycles, BIC the ellipse——"chosen by a penalised criterion" says nothing until the criterion is named.

Observation 2.3

the disagreement comes only from the penalty term.The fit term has the same form in both, so the difference comes from $2k$ against $k \ln n$——at $n = 100$, $\ln n = 4.6$, which exceeds 2, so BIC charges more for complexity.

3. More Data Makes the Complicated Model Win

Computation 3.1: which model is chosen was derived over a range of n; a positive difference means the epicycles.

n	AIC difference	BIC difference
10	-9.768564	-11.584075
50	-0.842822	-12.314960
100	$+10.314355$	-5.316666
500	$+99.571776$	$+74.284127$
5000	$+1103.717757$	$+1064.614597$

Computation 3.2: the epicycles begin to win at $n = 53.777041$ under AIC and $n = 131.109261$ under BIC. The AIC crossing agrees to six digits with the closed form $2\Delta k$ over the logarithm of the residual ratio.

Observation 3.3

this is the core: the direction is the opposite of the intuition.While data are few the simpler model is chosen, and the more data there are the more the complicated one is——"more data will settle on the right model" does not hold between two models whose residuals differ.

Observation 3.4

the reason is only the rate of growth.The difference in fit grows in proportion to n, gaining the logarithm of the residual ratio 1.250000 at every step, while the difference in penalty grows as a constant for AIC and a logarithm for BIC——neither a constant nor a logarithm beats a proportion, so a crossing point must exist at some finite n.

Observation 3.5

so a penalty does not “punish complexity”.What it charges for is the number of parameters, not for overfitting as such——a model that genuinely fits better wins however much penalty it pays. The trouble is that whether that better fit is genuine cannot be settled within the same data.

4. The Separator

Observation 4.1

the separator is how the penalty grows with n.Not whether there is a penalty: unless it grows in proportion to n, the difference in fit overtakes it——the difference between AIC and BIC is itself only constant against logarithm, and neither reaches a proportion.

Observation 4.2

beside Paper 536.Paper 536 showed that whether a formula was borrowed or derived shows outside the working range——there the dividing line lay outside the range; here, inside it, a penalty does not divide the two and more data leans the wrong way. Together they say that only the outside of the range divides models.

Observation 4.3

what is not said.Nothing is fitted to astronomical data or to history, and holding the residual independent of n is an idealisation——what is shown is the single point that penalties do not grow in proportion.

5. Roll Call and Closing

1. **residuals 0.004 against 0.005** (Computation 2.1): **on fit, the epicycles.**
2. **AIC -990.663110 against -980.348755** (Computation 2.1): **epicycles.**
3. **BIC -962.006238 against -967.322904** (Computation 2.1): **ellipse.**
4. **the answer changes with the criterion** (Observation 2.2): **name the criterion.**
5. **AIC crosses at 53.777041** (Computation 3.2): **the core.**
6. **BIC crosses at 131.109261** (Computation 3.2): **later, but it comes.**
7. **agreement to six digits** (Computation 3.2): **bisection against closed form.**
8. **$+1103.717757$ at $n = 5000$** (Computation 3.1): **the gap keeps opening.**
9. **residual ratio 1.250000** (Observation 3.4): **its logarithm counts each time.**
10. **how the penalty grows** (Observation 4.1): **the separator; Paper 536.**

Closing: **The more data there is, the more a penalised criterion prefers the more complicated model.On the same 100 points AIC gives -990.663110 against -980.348755 and takes the epicycles, while BIC gives -962.006238 against -967.322904 and takes the ellipse——"chosen by a penalised criterion" says nothing until the criterion is named.The epicycles begin to win at $n = 53.777041$ under AIC and $n = 131.109261$ under BIC, and above that they are chosen for ever.The AIC crossing agrees with a closed form to six digits, and by $n = 5000$ the gap is $+1103.717757$——the difference in fit grows in proportion to n while the penalty grows as a constant or a logarithm.Neither a constant nor a logarithm beats a proportion, so a crossing point must exist at some finite n.The separator is how the penalty grows with n.Placed among the earlier papers——Paper 536 showed that it shows outside the working range; here it is shown that inside it nothing divides them and more data leans the wrong way.To be honest——holding the residual independent of n is an idealisation, and the parameters and residuals are representative.**

References

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*.
2. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*.
3. Yamagishi, Y. (2026). *Whether a Formula Was Borrowed or Derived Shows Outside the Working Range* [Paper 536]. Zenodo. 10.5281/zenodo.23014123

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)