Part I: Model Selection & Model Averaging

focustat

FOCUS DRIVEN STATISTICAL
INFERENCE WITH COMPLEX DATA

Nils Lid Hjort

Department of Mathematics, University of Oslo

Geilo Winter School, January 2017

[Note: This is the pdf version of the 2 x 45 minutes Nils Talk I I gave at the Geilo Winter School, January 2017. In my actual presentation I of course did both of (a) saying quite a bit more than is on the page and (b) skidding semi-quickly over chunks of the material, including parts of the mathematics, complete with the usual mixture of hand-waving, glossing over technicalities, and swiping of details under imaginary carpetry. The pdf notes themselves are meant to be decently coherent, though, and may be suitable for study.]

Statistical selection of statistical lectures

Scenario A: I talk for 6 hours on Single Topic X.

Then 33% might be decently happy, perhaps even with Cumulative Satisfaction Coefficient monotone over $[0, 6]$. But 67% might be lost at sea after 1 hour.

Scenario B: I talk for 2 hours on Topic X, 2 hours on Topic Y, 2 hours on Topic Z.

Then 33% might have wished for more depth & more developed applications. But 67% will at least be given 3 chances to re-focus and get something out of it (being exposed to cute contemporary statistical ideas, even without getting the details).

[Footnote: '1 hour' is in Academic Units, i.e. 45 minutes.]

I go for B – maximising $\mathrm{E}\,(\mathit{utility} \mid \mathit{information}, \mathit{data})$.

- Model selection and model averaging
- Confidence distributions and data fusion
- Bayesian nonparametrics

Model fitting; inference; but which method, and which model?

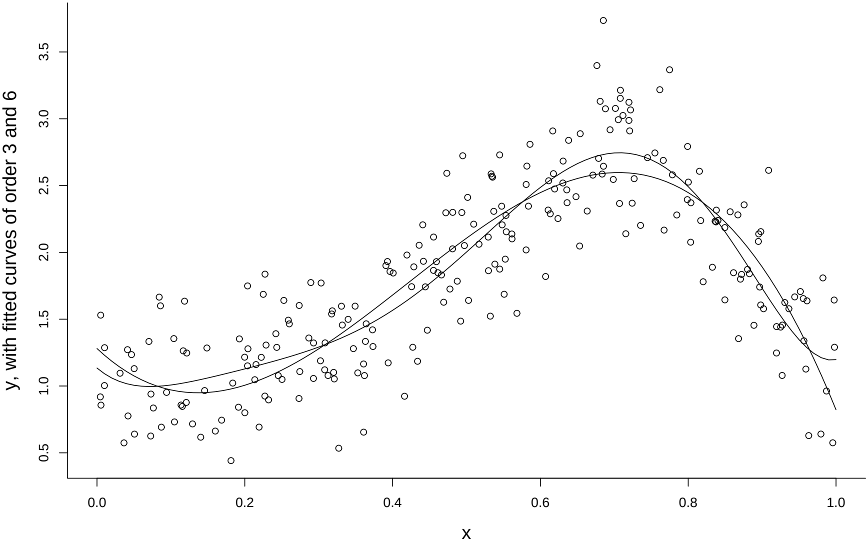
Typical problem setup: data y (i.i.d. data, or regression data, or something more complex); wish to reach inference statements for one or more focus parameters μ.

Typical method: fit a model, say $f(y,\theta)$, via estimate $\widehat\theta$; this leads to $\widehat\mu=\mu(\widehat\theta)$, along with standard error (estimated standard deviation). If things go well, one concludes with $\widehat\mu\pm 1.96\,\mathrm{se}$. (But which estimation method; which methods for reaching se; which fine-tuning tools?)

But which model? With candidate models $f_j(y,\theta_j)$, fitting these leads to $\widehat\mu_j=\mu_j(\widehat\theta_j)$. Which model is best, which focus parameter estimate is best?

All models are wrong, but some are useful. We need to understand not merely variation, but biases.

Example: When to stop, which polynomial order? High order: less bias, more variance. Small order: more bias, less variance. More data: more sophistication.



Plan

Model selection, model averaging, post-selection and post-averaging inference, bagging ...

A Maximum likelihood: basic theory

B AIC, the Akaike Information Criterion

C BIC, the Bayesian Information Criterion

D The Biggest Plagiarism Scandal in the history of literature (well, I'll check it out)

E FIC, the Focused Information Criteria

... then short & quick excursions:

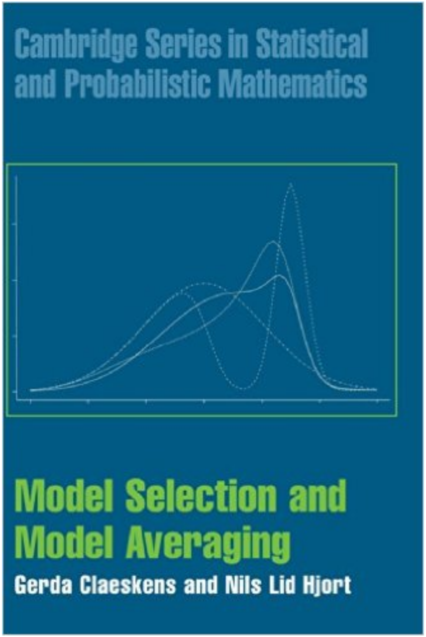
F The Quiet Scandal of Statistics

G Model averaging

H Models with increasing or very big dimension

I Concluding remarks

Check also course website for STK 4160, UiO (starting next week):



Cambridge Series in Statistical and Probabilistic Mathematics

Model Selection and Model Averaging

Gerda Claeskens and Nils Lid Hjort

A: Maximum likelihood: theory, and how to apply it

Fitting data y to a parametric model $f_{\rm joint}(y,\theta)$: we keep data fixed and maximise over the parameter,

$$\widehat\theta = \mathrm{argmax}\{\ell(\theta)\}, \quad \ell(\theta) = \log f_{\rm joint}(y,\theta).$$

(Some) Q:

(i) Is $\widehat\theta$ close to the best [as opposed to ‘the true’] parameter value θ_0?

(ii) What is its distribution?

(iii) What about $\widehat\mu = \mu(\widehat\theta)$, the estimated focus parameter?

(Some) A:

(i) Yes [modulo a sorting-out of what ‘best value’ should mean];

(ii) close to a multinormal distribution; and

(iii) close to a normal

– provided data information content is at least moderately good compared to the parameter dimension.

The i.i.d. situation first

To understand basic issues, methods, results, implications, we start in the i.i.d. situation. Suppose $y_1, \ldots, y_n$ are i.i.d. from density g, and that we try to fit the model $f(y, \theta)$, with θ of dimension p. The ML $\widehat{\theta}$ maximises log-likelihood function

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(y_i, \theta).$$

We have

$$H_n(\theta) = n^{-1}\ell_n(\theta) \to_{\rm pr} H(\theta) = \int g \log f_\theta \,{\rm d}y.$$

(a) Under weak conditions,

$$\widehat{\theta} = \mathrm{argmax}(H_n) \to_{\rm pr} \theta_0 = \mathrm{argmax}(H).$$

So the ML is aiming at $\theta_0 = \mathrm{argmin}\,\mathrm{KL}(g, f_\theta)$, the least false parameter, minimising the Kullback–Leibler distance

$$\mathrm{KL}(g, f_\theta) = \int g \log \frac{g}{f_\theta} \,{\rm d}y.$$

If model is perfect, then least false = true parameter, and $\widehat\theta$ tends to true value. But ML does a sensible job, even outside model conditions. How close is $\widehat\theta$ to θ_0?

Let

$$u(y,\theta) = \partial \log f(y,\theta)/\partial\theta, \quad i(y,\theta) = \partial^2 \log f(y,\theta)/\partial\theta\partial\theta^{\rm t}.$$

Consider the random function

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = U_n^{\rm t} s - \tfrac{1}{2} s^{\rm t} J_n s + o_{\rm pr}(1),$$

with

$$U_n = \ell_n'(\theta_0)/\sqrt{n} = n^{-1/2} \sum_{i=1}^n u(y_i,\theta_0) \to_d U \sim {\rm N}_p(0,K),$$

where $K = {\rm Var}_g\, u(Y,\theta_0)$, and

$$J_n = -n^{-1} \sum_{i=1}^n i(y_i,\theta_0) \to_{\rm pr} J = -{\rm E}_g i(Y,\theta_0).$$

So $\ell_n(\theta)$ is close to a quadratic, close to θ_0.

(b) Now we understand the behaviour of $\ell_n(\theta)$ close to θ_0; it's quadratic in the limit, with

$$A_n(s) \to_d A(s) = U^{\rm t} s - \tfrac{1}{2} s^{\rm t} J s.$$

Corollary 1: Approximate distribution of ML estimator: $\mathrm{argmax}(A_n) \to_d \mathrm{argmax}(A)$,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d J^{-1} U \sim \mathrm{N}_p(0, J^{-1} K J^{-1}).$$

This leads to confidence regions and tests, etc.

Corollary 2: Approximate distribution of deviance: $2 \max A_n \to_d 2 \max A$,

$$D_n(\theta_0) = 2\{\ell_{n,\max} - \ell_n(\theta_0)\} \to_d W = U^{\rm t} J^{-1} U, \ \text{with}\ U \sim \mathrm{N}_p(0, K).$$

Here $W \sim \chi^2_p$, if model is correct. May e.g. form

$$\{\theta_0 \colon D_n(\theta_0) \le z_{0.95}\}.$$

With μ a focus parameter, expressible as $\mu = \mu(\theta)$ under the given model: may read off approximate distribution of $\widehat{\mu} = \mu(\widehat{\theta})$.

Corollary 3: Delta method:

$$\sqrt{n}(\widehat{\mu} - \mu_0) \to_d c^{\rm t} J^{-1} U \sim {\rm N}(0, \kappa^2),$$

where

$$\kappa^2 = c^{\rm t} J^{-1} K J^{-1} c \quad \text{and} \quad c = \partial\mu(\theta_0)/\partial\theta.$$

Good News: These results are very general (valid for all smooth models, even for those you might invent yourself), and increasingly easy to use. Data y, model $f(y, \theta)$: you programme the log-likelihood function $\ell(\theta)$ and ask an R-routine to give you

(i) the ML $\widehat{\theta}$;

(ii) the information matrix $\widehat{J} = -\partial^2\ell(\widehat{\theta})/\partial\theta\partial\theta^{\rm t}$;

(iii) an estimate $\widehat{K}$ of K;

(iv) the derivative $\widehat{c} = \mu(\widehat{\theta})/\partial\theta$;

and you're in business.

Illustration: $n = 141$ lifetimes from Roman era Egypt, c. 100 years BC (with $\bar{y} = 30.7$, max $= 96$, and life more dangerous for women than for men). I fit the model

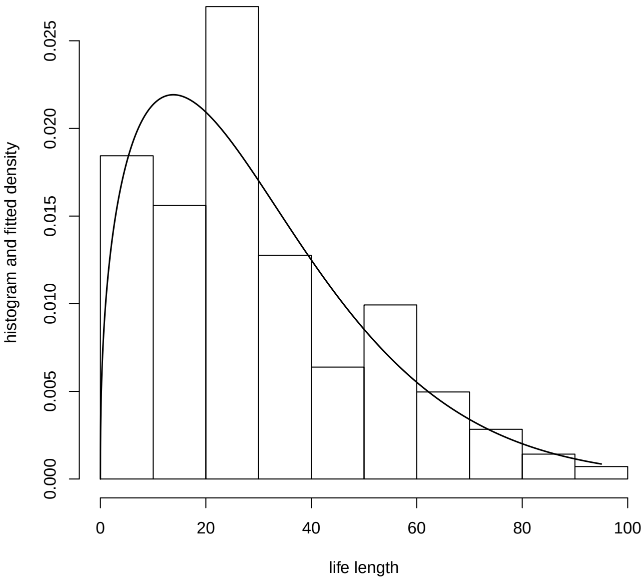
$$F(y, a, b) = 1 - \exp\{-(y/a)^b\} \quad \text{for } y > 0.$$

```
logL = function(ab)
{
a = ab[1]
b = ab[2]
well = -(yy/a)^b + log(b)+(b-1)*log(yy)-b*log(a)
sum(well)
}
```

I then use `nlm` to minimise `minuslogL` (or anything similar):

```
nils = nlm(minuslogL,c(25,1),hessian=T)
ML = nils$estimate
Jhat = nils$hessian
se = sqrt(diag(solve(Jhat)))
```

Histogram and fitted $f(y, \widehat{a}, \widehat{b})$ for 141 ancient Egypt life-lengths:



Parameter estimates, with standard errors, read off from the simple R programme:

```
     ML      se
a  33.563  2.113
b   1.404  0.096
```

With focus parameter $\mu = F^{-1}(0.80) = a(-\log 0.2)^{1/b}$: Estimate is $\widehat{\mu} = \mu(\widehat{a}, \widehat{b}) = 47.104$, and standard error 2.832 is read off from a simple programme, using numerical derivatives for $\widehat{c} = \partial\mu(\widehat{\theta})/\partial\theta$, and $\theta = (a, b)$. A 90% confidence interval for the 0.80 quantile is $[42.445, 51.762]$.

Similar steps can be carried out for other models – see discussion of model selection issues below.

From i.i.d. to regression models

Data are often of the form (x_i, y_i), for $i = 1, \ldots, n$ objects, with outcome y to be explained and interpreted as a consequence of x.

Regression model for y given x: $f(y \,|\, x, \theta)$. Chief examples:

(i) linear regression, $y_i = x_i^{\rm t}\beta + \varepsilon_i$, where $\varepsilon_i \sim {\rm N}(0, \sigma^2)$;

(ii) logistic regression, for 0-1 data,

$$\Pr(y_i = 1 \,|\, x_i) = \frac{\exp(x_i^{\rm t}\beta)}{1 + \exp(x_i^{\rm t}\beta)};$$

(iii) Poisson regression, with $y_i \sim {\rm Pois}(\mu_i)$, and $\mu_i = \exp(x_i^{\rm t}\beta)$.

For each of these (and of lots of others), may start from log-likelihood function:

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(y_i \,|\, x_i, \theta).$$

Good News: concepts, methods, results, algorithms generalise to the regression context – with some efforts, experience, patience; and without too many or overly hard mathematical obstacles.

So ML estimators $\widehat\theta$ can be found, via numerical optimisation. Under model conditions:

$$\widehat\theta \approx_d \mathrm{N}_p(\theta_0, \widehat J^{-1}),$$

with $\widehat J = -\partial^2 \ell_n(\widehat\theta)/\partial\theta\partial\theta^{\rm t}$, the observed information matrix.

Outside model conditions: there's a more elaborate definition of least false parameter value $\theta_{0,n}$, but things pan out as for the i.i.d. case, with sandwich matrix $\widehat J^{-1}\widehat K\widehat J^{-1}$ replacing $\widehat J^{-1}$.

B: Model selection: AIC (and relatives)

For simplicity now: working in the i.i.d. setup, but all concepts, arguments, techniques, results generalise to regressions and more general frameworks.

For the same data $y_1, \ldots, y_n$, consider competing parametric models $f_j(y, \theta_j)$. Which of the fitted models $f_j(y, \widehat{\theta}_j)$ should we choose? We wish

$$\mathrm{KL}(g, f_j(\cdot, \widehat{\theta}_j)) = \int g \log g \, \mathrm{d}y - \int g(y) \log f_j(y, \widehat{\theta}_j) \, \mathrm{d}y$$

to be small. We shall see that

$$n^{-1}\ell_{n,\max,j} = n^{-1}\ell_{n,j}(\widehat{\theta}_j) \text{ overestimates } \int g \log f(\cdot, \widehat{\theta}_j) \, \mathrm{d}y.$$

Hence we need something like

$$\mathrm{IC}_j = \ell_{n,max,j} - \mathrm{pen}_j,$$

with pen_j a measure of complexity for model j.

For a given model, need to understand how much $n^{-1}\ell_{n,\max}$ overshoots $\int g \log f(\cdot,\widehat\theta)\,\mathrm{d}y$. We work with

$$H_n(\theta) = n^{-1}\ell_n(\theta) \quad \text{and its limit} \quad H(\theta) = \int g \log f_\theta \,\mathrm{d}y.$$

Two 2nd order Taylor approximations, using $V_n = \sqrt{n}(\widehat\theta - \theta_0) \to_d V = J^{-1}U$:

$$\begin{aligned} H(\widehat\theta) &= H(\theta_0) - \tfrac12 n^{-1} V_n^{\mathrm{t}} J V_n + o_{\mathrm{pr}}(1/n), \\ H_n(\theta_0) &= H_n(\widehat\theta) - \tfrac12 n^{-1} V_n^{\mathrm{t}} J^{-1} V_n + o_{\mathrm{pr}}(1/n). \end{aligned}$$

Subtraction & book-keeping yields

$$\Delta_n = H_n(\widehat\theta) - H(\widehat\theta) = H_n(\theta_0) - H(\theta_0) + n^{-1}W_n + o_{\mathrm{pr}}(1/n),$$

with $W_n = V_n^{\mathrm{t}} J V_n \to_d W = U^{\mathrm{t}} J^{-1} U$. So

$$\mathrm{E}\,\Delta_n = n^{-1}p^* + o(1/n), \quad \text{with} \quad p^* = \mathrm{E}\,W = \mathrm{Tr}(J^{-1}K).$$

Conclusion:

$$\mathrm{AIC}^* = 2\ell_{n,\max} - 2\widehat{p}^*$$

is doing the minimise expected KL distance job, with $\widehat{p}^* = \mathrm{Tr}(\widehat{J}^{-1}\widehat{K})$. If model is trusted, then $p^* = p$, length of θ.

The simplified version

$$\mathrm{AIC} = 2\ell_{n,\max} - 2p$$

is most typically used – it indirectly takes p as an estimate of or approximation to $p^* = \mathrm{Tr}(J^{-1}K)$.

$$\mathrm{AIC}^* = 2\ell_{n,\max} - 2\widehat{p}^*$$

is the model-robust version.

For candidate models $M_1, \ldots, M_k$, compute $\mathrm{AIC}_j^* = 2\ell_{n,j,\max} - 2\widehat{p}_j^*$ for each, and use the model with the biggest score.

Cross validation methods abound; some of these are related to AIC.

Example: With (x_i, y_i) data, and $x_i = (x_{i,1}, \ldots, x_{i,q})$ of length q: can try different subsets of the q covariates. AIC says: for each subset of the covariates (there are 2^q such), compute the score

$$A = n \log \widehat{\sigma} + p,$$

where p is the length of the covariate and $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (y_i - x_i^{\mathrm{t}} \widehat{\beta})^2$, and select model with smallest A.

Example: It's easy to try out different models for the $n = 141$ life-lengths of Roman era Egypt (c. 2100 years ago). A partial list is:

```
 model dim  logLmax  aic
  1     1 -623.7764 -1249.553  4  Expo
  2     2 -615.3861 -1234.772  3  Gamma
  3     2 -611.3530 -1226.706  1  Gompertz
  4     2 -613.1144 -1230.229  2  Weibull
```

Better models may be found. If a three-parameter model should beat the Gompertz, it needs to have a logLmax of -610.353 or higher.

C: BIC: $\mathrm{Pr}(M_j \,|\, \mathrm{data})$

With candidate models $M_1, \ldots, M_k$ for the data y, why not select the most likely model, given the data? This needs a Bayesian machinery: prior $p(M_j)$ for the models, and a prior $\pi_j(\theta_j)$ for the parameters of model M_j. Bayes' theorem says

$$\mathrm{Pr}(M_j \,|\, \mathrm{data}) = \frac{p(M_j)\lambda_j}{p(M_1)\lambda_1 + \cdots + p(M_k)\lambda_k},$$

where

$$\lambda_j = L_{j,\mathrm{marg}}(y) = \int L_{n,j}(\theta_j)\pi_j(\theta_j)\,\mathrm{d}\theta_j$$

are the marginal likelihoods.

Trouble: Rather hard to set up all priors, conceptually and practically – both for models, and for the parameter vector in each model; also, rather hard to do the integrations.

Good News: There's an easy to use approximation, which also bypasses the need to be explicit about the priors.

The approximation is of the Laplace type. For a given model, with n increasing, and using $\theta = \widehat{\theta} + s/\sqrt{n}$:

$$\begin{aligned}
\lambda_n &= \int L(\theta)\pi(\theta)\,\mathrm{d}\theta \\
&= \int \exp\{\ell_{n,\max} + \ell_n(\theta) - \ell_n(\widehat{\theta})\}\pi(\theta)\,\mathrm{d}\theta \\
&= \exp(\ell_{n,\max})n^{-p/2} \int \exp\{\ell_n(\widehat{\theta} + s/\sqrt{n}) - \ell_n(\widehat{\theta})\}\pi(\widehat{\theta} + s/\sqrt{n})\,\mathrm{d}s \\
&\doteq \exp(\ell_{n,\max})n^{-p/2} \int \exp(-\tfrac{1}{2}s^{\mathrm{t}}J_n s)\,\mathrm{d}s\,\widehat{\pi}(\widehat{\theta}).
\end{aligned}$$

Hence

$$\log \lambda_n = \ell_{n,\max} - \tfrac{1}{2}p\log n + O_{\mathrm{pr}}(1)$$

and

$$\mathrm{Pr}(M_j \,|\, \mathrm{data}) \doteq \frac{\exp(\frac{1}{2}\mathrm{BIC}_j)}{\sum \exp(\frac{1}{2}\mathrm{BIC}_{j'})} \text{ with } \mathrm{BIC}_j = 2\ell_{n,j,\max} - p_j \log n.$$

To explain or to predict: the Two Cultures

Different uses of statistics (modelling, fitting, fine-tuning, model selection, inference) in different schools and applications.

- to explain – to understand, to get close to the truth, to interpret the mechanisms;
- to predict – to classify, to get the black box to work, to win prediction contests.

Both are solid, long-standing, valuable statistical pursuits (and sometimes correlated). We wish to predict the average temperature and the polar bear population in a.D. 2067 and to understand underlying causes.

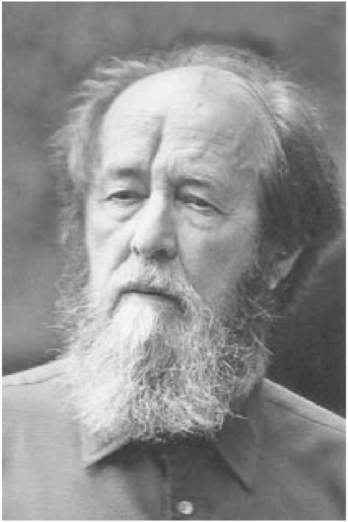
(Die Statistiker haben die Welt nur verschieden interpretiert; es kömmt aber darauf an, sie zu verändern.)

AIC (and cousins): prediction, estimation.

BIC (and cousins): aiming at best explanation.

D: Quiet Does Not Flow the Don:
Statistical analysis of a quarrel between Nobel prize winners

Sholokhov.



Solzhenitsyn.

The Nobel Prize of Literature 1965: Mikhail Sholokhov (1905–1984), for Тихий Дон. He was called 'the greatest of our writers', his works have been published in more than a thousand editions, printed in more than sixty million copies ...

But in 1974 the article *Стремя 'Тихого Дона' (Загадки романа)*, was published in Paris, by the author and critic Д*. He claimed that Тихий Дон was instead written by Fiodor Kriukov (who fought against the Bolsheviks and died in 1920). Д* was backed by a.o. Aleksandr Solzhenitsyn (1918–2008, Nobel prize 1970).

Шолохов (Nobel 1965) vs. **Солженицын** (Nobel 1970) mystery:

An inter-Nordic research team formed in 1975, led by Geir Kjetsaa (Department of Literature, Regional Studies and European Languages at the University of Oslo):

Linguistic analyses, detective work, quantitative data, ...

And Sholokhov and Solzhenitsyn had quarrelled before.

Sholokhov, to the Secretariat of the Union of Soviet Writers, 1967:

"I have read Solzhenitsyn's *Feast of the Conquerors* and *In the First Circle*. What is striking [...] is the sickly shamelessness of the author. Solzhenitsyn not only makes no attempt to hide or somehow veil his anti-Soviet views [...].

As regards the form of the play, it is feeble and foolish [...]. Why are Vlasovites – traitors to the Motherland on whose conscience lie thousands of our dead and tormented soldiers – praised as those who express the hopes of the Russian people? The novel *In the First Circle* stands on this same political and artistic plane.

At one time I formed an impression of Solzhenitsyn [...] that he is an insane person, suffering from megalomania. [...] If Solzhenitsyn is psychologically normal, then he is, in essence, an open and malicious anti-Soviet person. In either case, Solzhenitsyn has no place in the ranks of the Union of Soviet Writers. I am unconditionally in favour of the exclusion of Solzhenitsyn from the СП СССР."

Doris Lessing (in *Walking in the Shade*, 1997), about Sholokhov:

"The only word for this man is *macho*, positively a comic-opera he-man. Vibrations of dislike instantly flowed between us."

She says all Sholokhov's later work is of lesser quality than Quiet Don – precisely one of Solzhenitsyn's arguments.

This is the potentially biggest plagiarism scandal in the history of literature. How should one approach the problem?

- Literary style and themes ...
- Pure detective work ...
- Statistical analysis ...

Data have been extracted from three corpora:

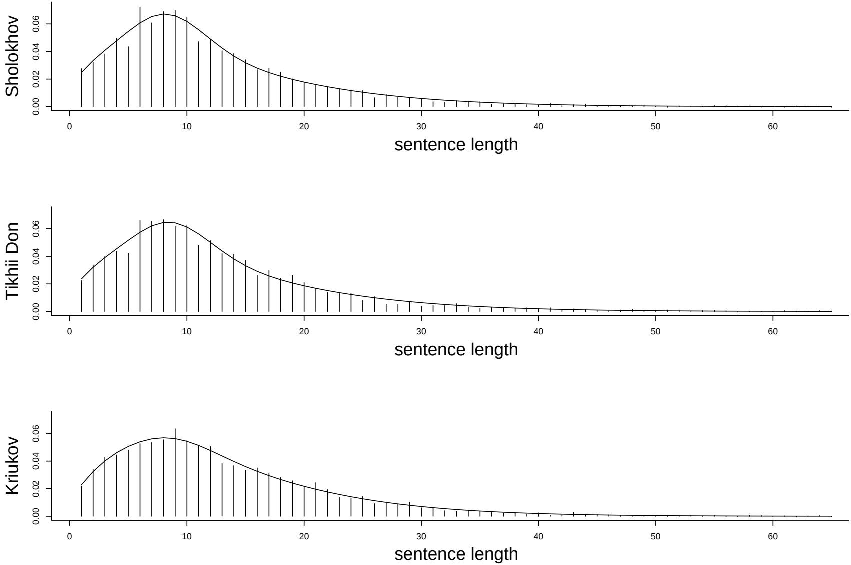
- Ш, or Sh, from published work guaranteed to be by Sholokhov, 4183 sentences;
- Кр, or Kr, from the hand of the alternative hypothesis Kriukov, 3736 sentences; and
- ТД, or TD, the Nobel winning text, 3760 sentences.

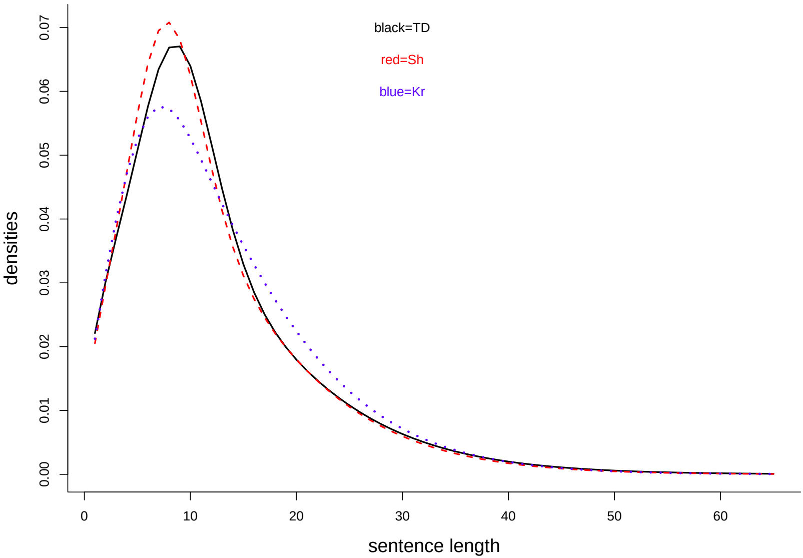
Each of the corpora has about 50,000 words.
Here: focus on sentence lengths.

Table of sentence lengths, observed and fitted from model:

from	to	Sh	Kr	TD	Sh	Kr	TD
1	5	799	714	684	803.4	717.6	690.1
6	10	1408	1046	1212	1397.0	1038.9	1188.5
11	15	875	787	826	884.8	793.3	854.4
16	20	492	528	480	461.3	504.5	418.7
21	25	285	317	244	275.9	305.2	248.1
26	30	144	165	121	161.5	174.8	151.1
31	35	78	78	75	91.3	96.1	89.7
36	40	37	44	48	50.3	51.3	52.1
41	45	32	28	31	27.2	26.8	29.8
46	50	13	11	16	14.5	13.7	16.8
51	55	8	8	12	7.6	6.9	9.4
56	60	8	5	3	4.0	3.5	5.2
61	65	4	5	8	2.1	1.7	2.9

This is not a statistical walk in the park (and not Solstad vs. Hemingway):




black=TD
red=Sh
blue=Kr
densities
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
10
20
30
40
50
60
sentence length

If $Y \mid \lambda$ is Poisson(λ), but $\lambda \sim \mathrm{Gamma}(a, b)$, then Y has distribution

$$f^*(y, a, b) = \frac{b^a}{\Gamma(a)} \frac{1}{y!} \frac{\Gamma(a + y)}{(b + 1)^{a+y}} \quad \text{for } y = 0, 1, 2, \ldots,$$

which is the negative binomial, with variance $>$ mean.

Fitting this two-parameter model to the data is also found to be too simplistic; clearly the muses had inspired the novelists to transform their passions into patterns more variegated than those dictated by a mere negative binomial, their artistic outpourings also appearing to display the presence of two types of sentences, the rather long ones and the rather short ones, spurring in turn the present author on to the following mixture of one Poisson, that is to say, a degenerate negative binomial, and another negative binomial, with a modification stemming from the fact that sentences containing zero words do not really count among Nobel literature laureates (with the notable exception of a 1958 story by Heinrich Böll):

$$f(y, p, \xi, a, b) = p \frac{\exp(-\xi)\xi^y / y!}{1 - \exp(-\xi)} + (1 - p) \frac{f^*(y, a, b)}{1 - f^*(0, a, b)}$$

I fit the parametric model

$$f(y,p,\xi,a,b) = p\frac{\exp(-\xi)\xi^y/y!}{1-\exp(-\xi)} + (1-p)\frac{f^*(y,a,b)}{1-f^*(0,a,b)}$$

to each of the three corpora, via maximum likelihood:

	Sh	se	Kr	se	TD	se
p	0.184	0.021	0.057	0.023	0.173	0.023
ξ	9.099	0.299	9.844	0.918	9.454	0.387
a	2.093	0.085	2.338	0.092	2.114	0.095
b	0.163	0.007	0.178	0.008	0.161	0.008

Goodness of fit analysis: model is fine. Standard errors by

$$\mathrm{Var}\,\widehat{\theta} = \mathrm{Var}\begin{pmatrix}\widehat{p}\\ \widehat{\xi}\\ \widehat{a}\\ \widehat{b}\end{pmatrix} \doteq n^{-1}J(\widehat{\theta}).$$

Is TD closer to Sh or to Kr? Can we see the difference[s]?

Can conduct an hypothesis test. Note: Crucial difference between

H_0: Sholokhov is the author vs. H_1: he is not the author

and

H_0: Sholokhov is the author vs. H_1: Kriukov is the author.

I choose instead: Model selection problem:

- M_1: Sholokhov is the author: Sh and TD come from the same distribution, while Kr represents another;
- M_2: Д* and Solzhenitsyn were correct: Kr and TD come from the same distribution, while Sh is different;
- M_0: Sh, Kr, TD represent three statistically disparate corpora.

Tell me what you think, by giving me $p(M_1), p(M_2), p(M_0)$ – and I'll tell you what you ought to think, giving you

$$p(M_1 \mid \mathrm{data}),\ p(M_2 \mid \mathrm{data}),\ p(M_0 \mid \mathrm{data}).$$

Write θ_1, θ_2, θ_3 for the three parameter vectors (p, ξ, a, b), for respectively Sh, Kr, TD.

- M_1 says $\theta_1 = \theta_3$ while θ_2 is different;
- M_2 says $\theta_2 = \theta_3$ while θ_1 is different;
- M_0 makes no assumption re the three parameter vectors.

Let $p(M_1)$, $p(M_2)$, $p(M_0)$ be prior probabilities for the three possibilities and let $L_1(\theta_1)$, $L_2(\theta_2)$, $L_3(\theta_3)$ be the three likelihoods. Then by Bayes' theorem:

$$
\begin{aligned}
p(M_1 \,|\, \mathrm{data}) &= p(M_1)\lambda_1/\{p(M_1)\lambda_1 + p(M_2)\lambda_2 + p(M_0)\lambda_0\}, \\
p(M_2 \,|\, \mathrm{data}) &= p(M_2)\lambda_2/\{p(M_1)\lambda_1 + p(M_2)\lambda_2 + p(M_0)\lambda_0\}, \\
p(M_0 \,|\, \mathrm{data}) &= p(M_0)\lambda_0/\{p(M_1)\lambda_1 + p(M_2)\lambda_2 + p(M_0)\lambda_0\},
\end{aligned}
$$

in terms of marginal observed likelihoods $\lambda_0, \lambda_1, \lambda_2$.

These marginal likelihoods are

$$\begin{aligned}
\lambda_1 &= \int \{L_1(\theta) L_3(\theta)\} L_2(\theta_2) \pi_{1,3}(\theta) \pi_2(\theta_2)\, \mathrm{d}\theta\, \mathrm{d}\theta_2, \\
\lambda_2 &= \int \{L_2(\theta) L_3(\theta)\} L_1(\theta_1) \pi_{2,3}(\theta) \pi_1(\theta_1)\, \mathrm{d}\theta\, \mathrm{d}\theta_1, \\
\lambda_0 &= \int L_1(\theta_1) L_2(\theta_2) L_3(\theta_3) \pi_1(\theta_1) \pi_2(\theta_2) \pi_3(\theta_3)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2\, \mathrm{d}\theta_3.
\end{aligned}$$

Here π_1, π_2, π_3 are the prior distributions for θ_1, θ_2, θ_3. Under M_1: one prior $\pi_{1,3}$ for $\theta_1 = \theta_3$; under M_2: one prior $\pi_{2,3}$ for $\theta_2 = \theta_3$.

The integrals are 8-, 8-, 12-dimensional.

Working with the marginal likelihoods:

$$\begin{aligned}\lambda_1 &= L_{1,3}(\widehat\theta_{1,3})(2\pi)^{4/2}(n_1+n_3)^{-4/2}|J_{1,3}|^{-1/2}\pi_{1,3}(\widehat\theta_{1,3})C_{1,3}\\ &\qquad L_2(\widehat\theta_2)(2\pi)^{4/2}n_2^{-4/2}|J_2|^{-1/2}\pi_2(\widehat\theta_2)C_2,\\ \lambda_2 &= L_{2,3}(\widehat\theta_{2,3})(2\pi)^{4/2}(n_2+n_3)^{-4/2}|J_{2,3}|^{-1/2}\pi_{2,3}(\widehat\theta_{2,3})C_{2,3}\\ &\qquad L_1(\widehat\theta_1)(2\pi)^{4/2}n_1^{-4/2}|J_1|^{-1/2}\pi_1(\widehat\theta_1)C_1,\\ \lambda_0 &= \prod_{j=1,2,3} L_j(\widehat\theta_j)(2\pi)^{4/2}n_j^{-4/2}|J_j|^{-1/2}\pi_j(\widehat\theta_j)C_j.\end{aligned}$$

These involve likelihoods, ML estimates, Fisher information matrices and certain correction factors. (The above uses mathematics similar to 'Laplace approximations' often used to derive the BIC.)

Two steps to reach a conclusion:
(i) $C_j \doteq 1$, $C_{i,j} \doteq 1$. Can be justified by looking at expansions (sample sizes are big here).
(ii) No real differences between the priors involved: we let the data (and Sholokhov and Kriukov) speak for themselves.

With $\mathrm{BIC}_j^* = 2\log\lambda_j$,

$$\begin{aligned}\mathrm{BIC}_1^* &= 2(\ell_{1,3,\max} + \ell_{2,\max}) - 4\log(n_1+n_3) - 4\log n_2 \\ &\qquad - \log|J_{1,3}| - \log|J_2|, \\ \mathrm{BIC}_2^* &= 2(\ell_{2,3,\max} + \ell_{1,\max}) - 4\log(n_2+n_3) - 4\log n_1 \\ &\qquad - \log|J_{2,3}| - \log|J_1|, \\ \mathrm{BIC}_0^* &= 2(\ell_{1,\max} + \ell_{2,\max} + \ell_{3,\max}) - 4\log(n_1+n_2+n_3) \\ &\qquad - \log|J_1| - \log|J_2| - \log|J_3|.\end{aligned}$$

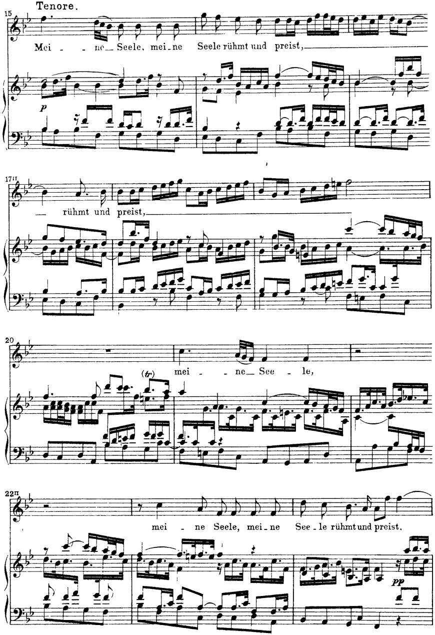
Further calculations, involving ML for the common θ of Sh and TD under M_1, and for the common θ of Kr and TD under M_2:

	M_1	M_2	M_0
$\Delta\mathrm{AIC}$	0.0	−13.7	−11.9
$\Delta\mathrm{BIC}^*$	0.0	−15.1	−28.0

Conclusion:

$$\Pr\{M_1 \mid \mathrm{data}\} \doteq 0.998, \quad \Pr\{M_2 \mid \mathrm{data}\} \doteq 0.002.$$

This agrees with Kjetsaa et al. (1975): No reason to doubt that Sholokhov is the rightful author of Тихий Дон (and the rightful winner of the Stalin Prize 1941).


Tenore.
15
Mei - - ne Seele, mei - ne Seele rühmt und preist,
p
17
rühmt und preist,
20
(tr)
mei - - ne See - le,
22
mei - ne Seele, mei - ne See - le rühmt und preist.
pp

Maxims & Quotations

‘All models are wrong, but some are useful’ (G.E.P. Box).

‘Entia non sunt multiplicanda praeter necessitatem’ (more or less: entities should not be multiplied beyond necessity, called Ockham’s razor, 1323, after the 14th century English logician and Franciscan friar William of Ockham). Slightly vulgarised version: The simplest explanation is the best.

‘How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service’ (C. Darwin).

‘The purpose of models is not to fit the data, but to sharpen the questions’ (S. Karlin).

‘It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience’ (A. Einstein, 1934 – the somewhat vulgarised version of this is ‘everything should be made as simple as possible, but not simpler’).

A famous exchange, after the 1782 premiere of KV 384 in Wien: Emperor Joseph II: “Gewaltig viel Noten, lieber Mozart.” Mozart: “Gerade soviel Noten, Euer Majästät, als nötig sind.”

E: Ranking models by FIC scores

Traditional methods like AIC (with variations), BIC (with variations), DIC (with variations) all work in overall modus – not concerned with the final use of the selected model.

FIC idea: for focus parameter μ, (i) compute all $\widehat{\mu}_j$ for candidate models M_j; (ii) use various tricks to find good approximation formula for

$$\mathrm{mse}_j = (\mathrm{E}_{\mathrm{true}}\,\widehat{\mu}_j - \mu_{\mathrm{true}})^2 + \mathrm{Var}_{\mathrm{true}}\,\widehat{\mu}_j = b_j^2 + v_j\,;$$

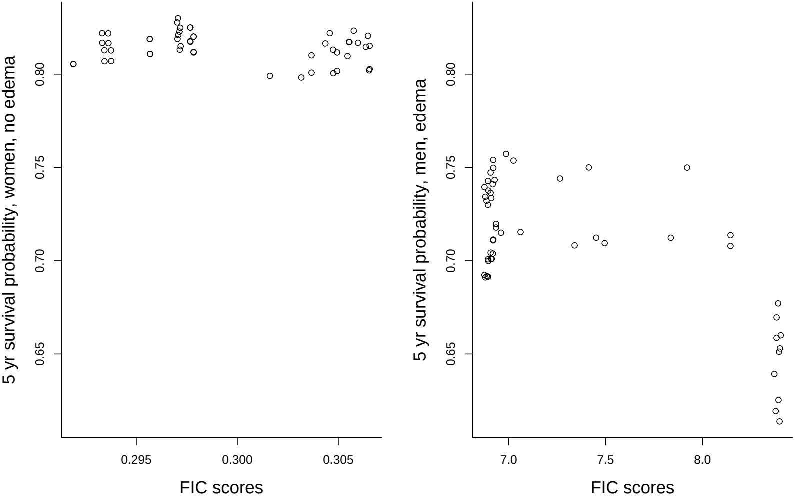
(iii) use other tricks to estimate these from data,

$$\mathrm{fic}_j = \widehat{(b_j^2)} + \widehat{v}_j\,;$$

(iv) plot all $(\sqrt{\mathrm{fic}_j}, \widehat{\mu}_j)$ in a FIC plot.

- wider models yield bigger variances and smaller biases;
- narrower models yield smaller variances but perhaps modelling bias.

PBC data set: FIC plots for the 50 best estimates of five-year survival probability, for two strata: (a) women of age 50, without oedema; (b) men of age 50, with oedema.



This FIC idea, assessing and estimating $\mathrm{mse} = b^2 + v$, pans out differently from situation to situation, and depends on how we model 'true':

$$v_j = \mathrm{Var}_{\mathrm{true}}\, \widehat{\mu}_j \quad \text{simpler than} \quad b_j = \mathrm{E}_{\mathrm{true}}\, \widehat{\mu}_j - \mu_{\mathrm{true}}.$$

- Parametric models, inside well-defined range from narrow to wide, and comparing ML estimates: various papers and book by Claeskens and Hjort (2008). Ok for regression models, choosing among 2^q models.
- Nonparametric Aalen model: Hjort (2009).
- Semiparametric Aalen model: Gandy and Hjort (2015), choosing among 3^q models.
- Parametric vs. nonparametric: Jullum and Hjort (2016).
- 'What price Cox regression?' Jullum and Hjort (2016).
- FIC for time series: Hermansen and Hjort (2014).
- Many others, for different sets of models.
- FRIC, with robust M-estimators: Hjort and Walker (2017).

A master theorem for competing estimators: Suppose narrow model has $f(y \,|\, x, \theta)$ with $\dim(\theta) = p$ and wide model has $f(y \,|\, x, \theta, \gamma)$ with $\dim(\gamma) = q$. For a focus parameter $\mu = \mu(\theta, \gamma)$, look at

$$\widehat{\mu}_S = \mu(\widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c}),$$

the ML of μ in submodel S (which has γ_j on board if $j \in S$).

With local neighbourhood model $f_{\rm true}(y) = f(y, \theta, \gamma_0 + \delta/\sqrt{n})$: For each of these 2^q competitors,

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}(\delta - G_S D),$$

with $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$ and $D \sim {\rm N}_q(\delta, Q)$ independent, ω depending on μ, and G_S a certain $q \times q$ matrix.

This provides a clear large-sample picture of everything going on, also with model average estimators $\widehat{\mu}^* = \sum_S \widehat{c}(S)\widehat{\mu}_S$:

$$\sqrt{n}(\widehat{\mu}^* - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}\{\delta - \widehat{\delta}(D)\},$$

with $\widehat{\delta}(D) = \sum_S c(S \,|\, D) G_S D$.

From the Master Theorem above: limiting mean squared error for the S based estimator $\widehat{\mu}_S$ is

$$\begin{aligned} \mathrm{mse} &= \mathrm{E}\,\{\Lambda_0 + \omega^{\mathrm{t}}(\delta - G_S D)\}^2 \\ &= \tau_0^2 + \omega^{\mathrm{t}} G_S Q G_S^{\mathrm{t}} \omega + \{\omega^{\mathrm{t}}(I - G_S)\delta\}^2. \end{aligned}$$

Most of the quantities are the same, across all questions – but

$$\omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$$

depends on the focus parameter μ. Four different focus questions lead to four different optimal models (and to four different leaderboard lists).

X1: FIC for Aalen's linear hazard regression model

Primary biliary cirrhosis dataset (a rare autoimmune liver disease), 312 randomised patients. Eight covariates: intercept, the treatment indicator, edema, sex, age, bilirubin, albumin and prothrombin time:

$$h_i(s) = \alpha_1(s)x_{i,1} + \cdots + \alpha_8(s)x_{i,8},$$

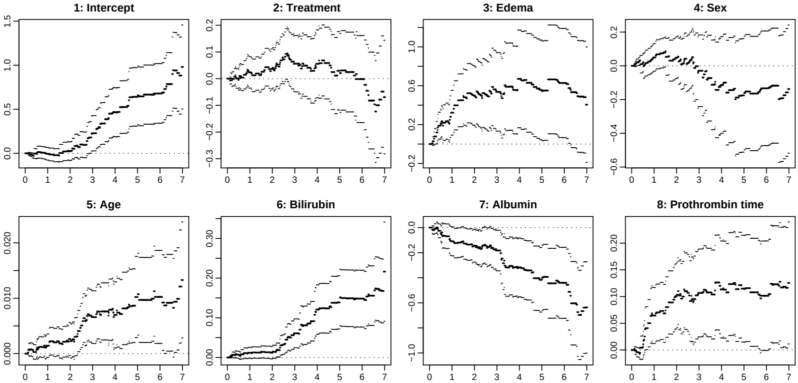
with survival curves

$$S_i(t) = \mathrm{Pr}\{T_i \geq t \,|\, x_i\} = \exp\{-A_1(t)x_{i,1} - \cdots - A_8(s)x_{i,8}\}.$$

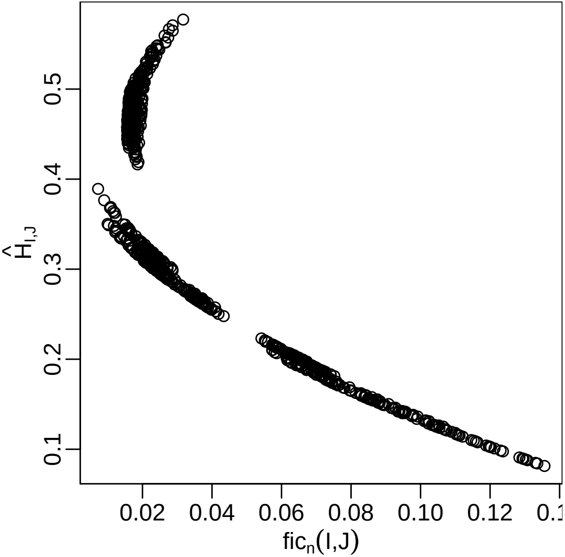
Each α_j: taken as time-varying, or as constant, or set to zero.

Hence (up to) $3^8 = 6561$ models. Have developed $\mathrm{fic}(I, J)$, with $I \cup J \subset \{1, \ldots, 8\}$: I, those with time-varying effects, J, those with time-constant effects, K, those set to zero. Here: no asymptotics, but hard-core mean and variance calculations (via lots o' martingales).

PBC data set: Estimates of cumulative regression functions $A_j(t)$ based on the full additive model with pointwise 95% asymptotic confidence intervals. Time is in years.



PBC data set: Estimating the integrated hazard rate for a 70 year old male at time $t = 1$ with higher-risk values of bilirubin, albumin and prothrombin time.



A GAMIC for extended GAM?

GAM, generalised additive models, work for regression data $(x_1, \ldots, x_p, y)$, where the ground model is

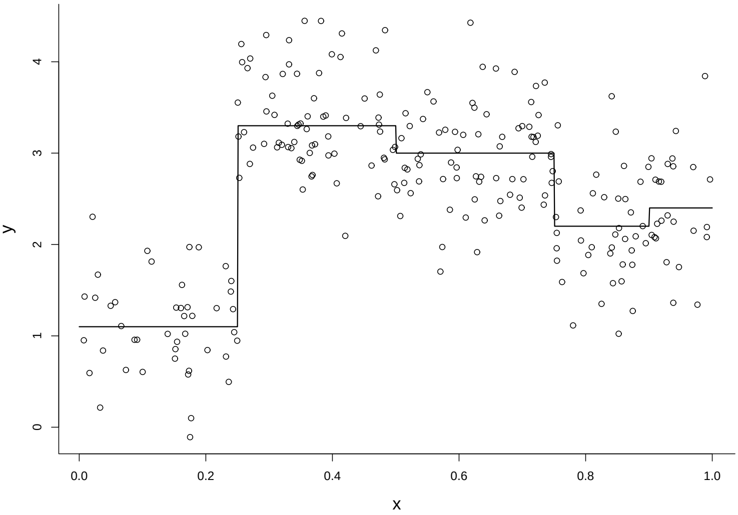
$$m(x_1, \ldots, x_p) = \mathrm{E}(y \mid x_1, \ldots, x_p) = \beta_0 + m_1(x_1) + \cdots + m_p(x_p).$$

The method provides nonparametric smoothing based estimates $\widehat{m}_j(x_j)$, along with confidence bands etc. – and is widely successful.

Someone ought to invent a suitable GAMIC, the GAM Information Criterion, which should go through 3^p possibilities – for each x_j, the $m_j(x_j)$ could be nonparametric, or a straight line (or more generally, something parametric), or zero. The FIC and AFIC methods pointed to above, for Aalen's linear hazard regression model, will have parallels of this type.

X2: the jump information criterion JIC

Regression .. with how many discontinuities?



Regression (or time series) with discontinuities:

$$y_i = m(x_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where $m(x) = a_j$ on window $[\gamma_{j-1}, \gamma_j)$, for d windows. With d windows on $[0, 1]$, and

$$0 < \gamma_1 < \cdots < \gamma_{d-1} < 1,$$

there are $d - 1 + d = 2d - 1$ unknown parameters (plus d itself!). Need JIC.

Usual BIC: $2\ell_{n,\max} - (2d - 1) \log n$. Grønneberg, Hermansen, Hjort (2014): rather better with

$$\mathrm{BJIC} = 2\ell_{n,\max} - (3d - 1) \log n.$$

Usual AIC: $2\ell_{n,\max} - 2(2d - 1)$. Rather better with

$$\mathrm{AJIC} = 2\ell_{n,\max} - 2\Big(1 + d\frac{\widehat{\sigma}^2}{\sigma_0^2} + \frac{1}{\widehat{\sigma}_0^2}\sum_{j=1}^{d-1}\widehat{\kappa}_j\Big).$$

X3: the copula information criterion CIC

With data (x_i, y_i) fitted to several copulae (Gauß, Clayton, Gumbel, Ali-Mikhail-Haq, Frank, Joe): which is best?

Usual AIC: $2\ell_{n,\max} - 2p$, involving the pseudo-log-likelihood. Grønneberg and Hjort (2014): Rather better with

$$\mathrm{CIC} = 2\ell_{n,\max} - 2(\widehat{p}^* + \widehat{r}^*),$$

for certain (somewhat complex) $\widehat{p}^*$ and $\widehat{r}^*$.

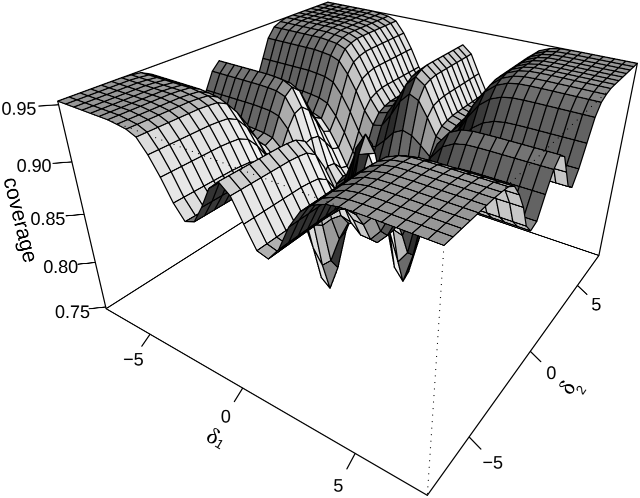
More generally valid alternative, via influence functions to approximate cross-validation:

$$\mathrm{CIC}_{\mathrm{xv}} = 2\ell_{n,\max} - 2\{p + \mathrm{Tr}(\widehat{J}^{-1}\widehat{W})\},$$

for certain (somewhat complex) $\widehat{J}$ and $\widehat{W}$.

F: The Quiet Scandal of Statistics

Your 95% confidence intervals (after model selection) have (much) lower confidence!



G: Model Averaging

With candidate models $M_1, \ldots, M_k$ leading to model-based estimates $\widehat{\mu}_1, \ldots, \widehat{\mu}_k$, for the same focus quantity μ, I can combine them:

$$\widehat{\mu}^* = \sum_j \widehat{w}_j \widehat{\mu}_j.$$

This often reduces variance and/or bias (depending on the weights). Special case:

$$\widehat{w}_j = \exp(-\lambda\,\mathrm{FIC}_j) \Big/ \sum_{j'} \exp(-\lambda\,\mathrm{FIC}_{j'}),$$

perhaps with λ fine-tuned via cross validation.

$\exists$ Master Theorem: Can characterise the limit distribution for all such $\widehat{\mu}^*$, including bagging, and then study properties and performance (Hjort and Claeskens, JASA 2003; Hjort, JASA 2014).

H: Bigger models, bigger data, machine learning

Traditional methods (including basic theory for likelihood, AIC, BIC, FIC): partly rely on

(data information content) / (model complexity) $>$ small,

more or less 'n/p is moderate or large'.

But in lots of modern applications p is large, or even $p \gg n$. So traditional methods need to be extended – modelling; estimation; inference; approximate distributions; model selection; ...

New Guys on the Block (1990ies and expanding): the lasso; ridge estimation; regularisation; (more) Bayes and empirical Bayes; sparsity, 'clever fixes'; partial least squares, ... However, the model selection parts lags a bit behind the other components (so far).

AIC and BIC are in trouble – but

$$\mathrm{FIC} = (\widehat{b}^2 - \mathrm{Var}\,\widehat{b})_+ + \widehat{v}$$

can be set to use (with some work, application by application), along with cross validation.

Example: $y_1, \ldots, y_n$, big model for the density:

$$f(y) = f_0(y, \theta) \exp\Big\{\sum_{j=1}^{100} a_j \psi_j(F_0(y, \widehat{\theta}))\Big\} / c_{100}(a_1, \ldots, a_{100}),$$

with $\psi_1, \psi_2, \ldots$ orthogonal functions on $[0, 1]$. ML: fat chance (difficult operationally and estimates will be bad). Good solution: Maximise

$$\ell_n(\theta, a) - \lambda \sum_{j=1}^{100} a_j^2 j^2.$$

This is regularisation, and/or empirical Bayes. Cf. Hellton and Hjort (2016), the Fridge.

Similar regularisation in lots o' other and bigger models. So far: Various ad hoc fixes, both for modelling and estimation – inference and model selection attempting to catch up.

I: Concluding remarks

- Durable & dependable war-horses: AIC and BIC (with cousins)
- Becoming mainstream: FIC (with variations, AFIC, but demands new efforts for new situations)
- FIC etc. towards personalised solutions
- Special models and special needs $\Longrightarrow$ special tools
- Other estimators, other loss functions $\Longrightarrow$ other variations
- Explain vs. predict
- $\widehat{\mu}^* = \sum_j \widehat{w}_j \widehat{\mu}_j$, e.g. $\widehat{w}_j \propto \exp(-\lambda \mathrm{FIC}_j)$
- Bigger models: regularisation, sparsity; new tools (and additional tools)