

Statistical Inference:
600 Exercises, 100 Stories, and Solutions

Nils Lid Hjort and Emil Aas Stoltenberg

Department of Mathematics, University of Oslo

Here: Preface, Chapters 11, 12, 13, 14

– This version last touched by Nils and Emil, 16 August 2026 –

©Nils Lid Hjort and Emil Aas Stoltenberg, 2026

Some technical stuff

ISBN - Numbers numbers

The Kioskvelter Project

This is a draft of our book-to-be and it may not be reproduced
or transmitted, in any form or by any means, without permission.

1234-5678

To my somebody
– N.L.H.

To my somebody
– E.Aa.S.

Preface

There are various books on various shelves concerning the broader themes of *statistical inference*, both generally and with its many specialised themes. These include Bayesian analysis (parametric and nonparametric), testing theory, optimal estimation, confidence distributions, large-sample theory with the necessary probability on board, survival analysis and event history analysis, Markov chains and processes, optimal decision theory, time series, machine learning, processing of big data, image analysis, multivariate analysis, model selection and averaging, goodness of fit methods, stochastic processes, change-point identification, prediction, etc. This book is decidedly different, in several key respects.

600 Exercises

First of all, we change gears from the traditional narrative of ‘theory development first, then a few exercises and illustrations’, to ‘crisp introduction text to a theme, then jump into exercises mode’. So when we come to Bayesian analysis in Ch. 7, for example, readers & doers are pushed rather quickly into working with and understanding from our generous lists of exercises. Thus theory is developed as we go, through problems and their solutions, from simpler statements and illustrations to the needed lemmas and theorems, as opposed to the more traditional ‘theory first’ approach. This is in line with what Aristotle wrote, in *Nicomachean Ethics*; “for the things we have to learn before we can do them, we learn by doing them”; people learn to cook by cooking. Somewhat more pointedly, the psychologist and progressive educator John Dewey held, “learn to know by doing, and to do by knowing”. Our approach is hence correlated with educational principles around Experimental Learning, and also the so-called Flipped Classroom.

The introductory Ch. 1 works through a generous list of parametric models, their construction and properties, from the more famous to various extensions and connections to yet other models. A healthy segment of our readers would know the basics of these models already, but may use the chapter for reference, as do we for the rest of the book. Ch. 2 then builds the apparatus for large-sample approximations, the uses of which truly abound in probability theory and both theoretical and applied statistics, with the Law of Large Numbers, the Central Limit Theorem with extensions, the delta method, moment-generating functions, etc. Chs. 3 and 4 then develop the fundamental concepts and methods for statistical estimation, confidence, and testing, with the required Neyman–Pearson theory and other excursions.

The important Ch. 5, on likelihoods and their ramifications, can of course be worked through and appreciated separately, but is also used in various guises in several later

exercises,
exercises

335 b.C.

in 1914

chapters, specifically for the Bayesian methods of Ch. 7, the survival and event history Ch. 11, for model selection in Ch. 12, and for Markov chain models in Ch. 13. Ch. 6 goes into minimum divergence inference, generalising and in various situations robustifying likelihood inference. In Ch. 8 we build confidence, with the frequentist no-prior methodology for summarising inference in terms of distributions for focus parameters, with consequences also for meta-analyses, ways of combining information across diverse sources. Themes related to loss, risk, optimal estimation, minimaxity, admissibility are worked with in Ch. 9.

Often enough statisticians need for modelling and inference purposes to go beyond the finite-dimensional world of data points and vectors. Ch. 10 develops the appropriate tools and results for classes of stochastic processes, with Donsker theorems, Brownian motion and bridges, along with statistically relevant constructions of running t-tests, monitoring processes for models, log-likelihood maxima bridges, and more. Some of these are also used in the survival chapter Ch. 11.

Model selection themes have been part of statistical practice since the early models were invented, but have become more important and ubiquitous in the modern world of computing, access to an increasing volume of datasets, and available software to run through thousands of model candidates. Ch. 12 develops these themes, with emphasis on the established acronyms AIC, BIC, FIC (the Akaike, the Bayes, the Focused Information Criterion). Markov chains and processes are the themes of Ch. 13, where also the important Markov Chain Monte Carlo methods are tended to; the MCMC algorithms play an important role both in the finite-dimensional Bayes world of Ch. 7 but also in our final Ch. 14, on Bayesian nonparametrics, with infinite-dimensional models.

Notably, each chapter is rounded off with a ‘Notes and Pointers’ section, commenting on how key ideas have evolved historically, and with connections to other chapters and yet further themes.

Notes &
Pointers

100 Statistical Stories

Secondly, we, with the active participation of you, the readers & doers, illustrate and showcase concepts, modelling, theory development via a generous list of Statistical Stories. These range from the perhaps simple illustrations and applications to the rather more challenging terrain, where theory needs further development or new twists, in order to solve real-life real-data problems. Our stories are sorted into the broader categories of

stories, stories

- (i) Demography, Epidemiology, Medicine;
- (ii) Biology, Climate, Ecology;
- (iii) Art, History, Literature, Music;
- (iv) Economics, Political Science, Sociology;
- (v) Sports;
- (vi) Simulated stories;
- (vii) Miscellaneous.

Significantly, also the stories are told in the spirit of exercises, from start observations concerning the context of and the problems at hand, to developing the right modelling and then the tools required for solving these. Students and readers are then meant to get involved in the problem solving process from the start, as opposed to listening to a polished final lecture by someone who with struggle and efforts might finally having solved the problem. Ploughing through the pieces of these stories will also provide good training in the subtle tasks of statistical modelling, checking, and selection.

It is to be noted that most stories engage tools from not merely one chapter at a time, so to speak, but often via several themes from several chapters. So as opposed to illustrating a certain Bayesian method, for example, there might be alternative and equally relevant approaches, depending on the context, drawn from other chapters in our book. Also, for some stories, theory needs further development as we go.

solutions,
solutions

Importantly, we also provide *solutions* to all 600 Exercises and 100 Statistical Stories. There is not space in the physical book to include these, however, so they are placed online in the GitHub site pointed to below. That site also has *all datasets* used through the book, with **R** and **Python** code when required.

Our methodological horizon is wide, so also in that sense our book is on the ambitious side of the spectrum. We have chapters on a fair range of methodological themes, from basic models to estimation and testing, on large-sample theory with central limit theorems and so on, empirical processes, parametric and nonparametric Bayes, confidence distributions, survival analysis, Markov chains and processes, and modern methods for statistical learning. Though we do manage to develop several of these themes rather broadly, to the state of strong applicability also for challenging uses, the book is not intended to be a full-semester full-course book for each of its subthemes. Several types of courses can however easily be built from the book, as we offer advice on below.

Appendixes A, B, C, D, E

At the end of the book are four appendixes with exercises that takes the reader through some of the required mathematical and probability background. Our list of exercises are selective, and we have strived to only include material that is strictly necessary for solving and understanding the exercises in other chapters, and that hopefully also sheds some light on the short & crisp exercises as well as on the statistical stories.

Appendix **A** gives a concise and selective introduction to metric spaces. and includes results about such spaces that are used several places in the book, but in particular in Ch. **10**. Statistics builds on probability, and probability is a branch of measure theory. Appendix **B** gives an introduction to measure theory. This includes exercises on σ -algebras, Carathéodory's extension theorem, Dynkin's lemma, the Lebesgue integral, the convergence theorems, the Radon–Nikodym theorem, and so on. Crucially, the theory developed in this appendix underlies all of the statistical theory in the book, even though for most of the exercises and stories this theory is not actively present, so to speak. The measure theory appendix provides the background for Appendix **C** on probability theory. This appendix gives a measure theoretic introduction to probability theory, and develops such familiar concepts as the expectation, independence, and conditional expectation, to mention some, in a rigorous fashion. Appendix **D** contains exercises on martingale theory in discrete and continuous time. Some martingale theory is present in Ch. **10**, and is used very actively in Ch. **11** on survival analysis. Martingale theory is a subject that deserves book length treatises (of which there are many, see Notes & Pointers), in our martingale appendix, however, we have attempted to keep the theory to the bare minimum required to work with survival analysis from the counting process point of view, without having to take too many mathematical results for granted.

Finally, and crucially, Appendix **E** goes through the datasets used for our Statistical Stories, with references and brief backgrounds and descriptions, along with pointers to where in our book the datasets are used.

This book, its students and readers, and its GitHub site

As made clear above the book is broad-horizoned, and will have different strata of readers. The required start level for this book to be readable and useful is meant to be at end-of-bachelor and above, as in Master's and PhD levels (and above!). Prerequisites, for the core of readers, should be basic mathematics, including calculus, linear algebra with matrices and so on; familiarity and experience with programming in or via suitable statistical or mathematical software systems, such as **R** and **Python**, and specifically with the ability to code up things on one's own, outside the terrain of the application of already fine-tuned algorithms of packages; and having operating knowledge of classic statistical models and their parameters. Our start chapter 1 goes through such models, though, along with combinations, variations, extensions, so it is perfectly possible to start almost from scratch, with a certain uphill parameter. Otherwise the already flagged intention and spirit is 'come along', learning away as one proceeds, chapter by chapter, story by story.

Our intended strata of readers are roughly as follows. These boxes are of course far from accurately defined, and our students are invited to cross over to other categories as they might be inclined to. who are you?

(i) The *Linear Readers*, who will benefit from having the stamina to work through chapter by chapter (ideally also exercise for exercise) and appropriate subsets of our stories. These readers will be at a Master's or PhD level. We even hope and trust that experienced scholars above these levels will find things to like, to be surprised and inspired by; some of the material, also in the collection of Stories, is advanced.

(ii) The *Chapter Samplers*, who need to freshen up or learn more about certain themes, like confidence distributions or survival analysis, without necessarily having the time to or the interest for all the other chapters. Such readers will also be guided to and benefit from relevant subsets of our 100 Statistical Stories.

(iii) Long live our lovely linear readers, but the book is also fine for the *Statistical Stories Readers and Samplers*, for those who already know the basics on statistical models, parameter estimation and testing, some Bayes, etc. As noted, our book has Part One, the theory development in chapters, then Part Two, the stories, but could by a segment of our readers be worked through in the opposite order: read stories, and check what you need to understand and do all details. The [McCullagh \(2022\)](#) book, with aspirations partly parallelling ours, is a case in point.

(iv) Our book is also for the *Specialists Inside Themes*, who wish to learn even more and to see how these subfields are applied in new fashions, possibly in combination with methods and themes from other areas.

The GitHub site **Hjort-Stoltenberg-Book** (xx to be fixed), already pointed t above, is seen as a vital companion to the physical book. It has all datasets used; code in **R** and **Python**, i.e. scripts used to generate our figures, tables, and analyses; and, importantly, solutions to all our 600 Exercises and all our 100 Statistical Stories. The site will be actively maintained, and will with time also include follow-up exercises, snippets of course notes, etc.

the Hjort-
Stoltenberg
book website

We have spent efforts making the book largely self-contained, and desert island readers, with few other books on their shelves, can in principle 'learn nearly everything' from this one (given the prerequisites alluded to in the previous paragraph). Just a few exceptions are as follows. (i) The modern Bayesian statistician needs Markov Chain Monte Carlo, as explained and exemplified in Ch. 7 (and used in various stories), on this

occasion pointing to the later-in-the-book Ch. 13 for basic theory. But lifting that basic theory, from discrete settings to continuous ones, would have taken seven extra pages, so there we instead point to relevant sources. (ii) In Ch. 10 on random processes we scrupulously go through the details of the $C[0, 1]$ space of continuous functions on the unit interval, and also need for several further developments, and applications in certain stories, similar but more complicated mathematical details for the bigger function space $D[0, 1]$. Again, all of this would need eleven more pages, so we point to relevant literature, but do sum up the required details and criteria such that we can proceed to further vital developments and applications, like for Poisson, Gamma, and Beta processes, etc. (iii) [xx is there more? completeness theorem for exponential family. Emil Doob-Meyer. xx]

Declaration

Since the work with this endeavour was finished in 2026, with humankind arguably at one of its cross-roads, a small *declaration of independence* might be uttered: the authors have absolutely not used so-called AI or versions thereof when it comes to the actual writing of our texts, sentences, paragraphs, nor for the many cases, on different scales, where ideas have been hatched and developed, modified, finessed, or abandoned, etc. Occasionally we have been helped by modest-level chat-gpt to find certain datasets and references from older literature, however.

We would similarly hope that our readers and users apply themselves, their cleverness, imagination, enthusiasm, perhaps with flecks of fruitful frustration, when working through exercises and stories – i.e. without being too quick to ask machines to solve problems for them. We hold these truths to be self-evident, that thoughtful thinking (fast and slow) & productive struggle remain the best algorithms for deep learning, genuine creativity, and satisfaction.

Road maps for courses to give based on our book

as many as ten
courses can be
taught from our
book

Several types of courses can be taught from this book. Depending on the course parameters of ambition, time allotted, and background level, lecturers could supplement material with course notes and pointers to specialised literature. This is a list of such, reflecting both the book's general ambitions and our own teaching experiences at the Master and PhD level for statistics and data science at the Department of Mathematics, University of Oslo. Below we provide the skeletons of relevant Road Maps, for the different courses. For each of these, a reasonable going-through of the long list of statistical models of Ch. 1 would be required and useful, without necessarily working through all of them, so below we omit further mention of that chapter. The other chapters point back to that initial cavalcade of models when needed.

- A A basic course in probability models and statistical modelling: more of Ch. 1 than for the other courses, and most of Ch. 3. A suitable selection of these Stories (xx more, edit xx): #1, #2, #17, #24, #27, #44, #61, #71, #76, #74, #75, #77, #79, #85, #86, #87, #90. The lecturer and the students could then for these stories focus on the modelling and interpretation aspects, and perhaps leave the inference aspects for other rainy days.
- B Hard-core classical statistical inference, with parametric models, estimation, testing, risk functions, etc.: half of Ch. 2, then most of Chs. 3, 4, half of 5, 7, 9. A

- suitable selection of these Stories: #1, #2, #6, #17, #22, #26, #27, #28, #31, #39, #41, #44, #47, #53, #62, #76, #77, #78, #83, #88, #99.
- C Modern Bayesian statistics, prior to posterior in classical setups and with MCMC, excursions into nonparametric terrain: the essence of Chs. 3, 5, then good portions of Chs. 7, 14; this could also be coupled with addition emphasis on estimation and decision theory, risk functions, with Ch. 9. A suitable selection of these Stories: #1, #4, #18, #19, #23, #31, #41, #52, #56, #60, #81, #86. #90, #101.
- D Confidence distributions, perhaps alongside classical Bayes, with meta-analysis and combination across information sources: the essence of Ch.s. 3, 5, then healthy portions of 7, 8, 14. A suitable selection of these Stories: #1, #3, #4, #6, #8, #12, #13, #30, #33, #45, #46, #52, #59, #60, #77, #90.
- E Large-sample theory, the careful probability theory leading to CLT and more, ideally coupled with parts of empirical processes, convergence, approximations, applications in statistics: most of Chs. 2, 5, 6, 10, perhaps even with parts of Chs. 11, 13. A suitable selection of these Stories (xx more, edit xx): #15, #21, #22, #24, #27, #34, #44, #46, #50, #51, #60, #77, #83, #85, #87, #93, #96, #100.
- F Survival and event history analysis: models and methods for survival data, counting processes, some martingales, Kaplan–Meier and Nelson–Aalen, Cox and Aalen regression: the essence of Chs. 2, 3, 4, 5, then the full 11. A suitable selection of these Stories: #4, #5, #9, #10, #15, #19, #21, #48, #51.
- G Model selection and model averaging, AIC, BIC, FIC, tolerance radii, cross-validation, applications: the essence of Chs. 2, 3, 4, 5, then the full Ch. 12. A suitable selection of these Stories: #5, #8, #16, #17, #22, #24, #26, #27, #28, #37, #38, #45, #51, #50, #53, #55, #58, #61, #76, #87, #98.
- H Markov chains and processes, recurrence times, stationary distributions, MCMC, with applications: half of Ch. 5, then Chs. 10, 13. A suitable selection of these Stories: #3, #14, #15, #25, #46, #61, #86.
- I Finally we point to the possibility of building a different type of course, on statistical modelling and applications, with emphasis on the stories, and less concern for all the technical details. This is also a way to learn about and appreciate classical and modern statistics, concepts, modelling, computation, scientific context. The selection of stories, in volume and in detail, would of course depend on the time given and the interests of the lecturer and the audience.

Other books

There is a veritable ecosystem of books in good libraries about statistical inference, from the basics to more advanced texts and to specialist treatments. The recurrent core themes remain model building with the required portions of probability theory, parametric estimation, confidence intervals, testing of hypotheses, approximate normality. There are also visible Poisson rates for publishing in new categories, in the directions of machine learning and AI, along with still new takes on classic methodology, related to their active uses with emerging types of new data, and the general ‘combination of many types of accessible information’. With the various parameters one may attempt to measure, related to ambition level, attention to details, the scale from theory to applications (or is it the other way around), horizon width, and the pedagogical lines taken, we trust the Mahalanobis distance from our book to most others is significant and positive.

Classic texts on the basis of the theory include [Lehmann \(1959, 1983\)](#), [Casella and](#)

Berger (2013), Bickel and Doksum (2015), Davison (2003), (xx point to 1-2 more, Brazzale et al. (2007), xx).

Books in this landscape differ when it comes to emphasis on applications, where there naturally is a range from ‘simple illustration to check how things work’ to diving into the scientific context and associated key questions. Since it takes ample time to develop theory, concepts, and methodology, many very good texts are perhaps modest when it comes to application details. Some are more ambitious in that regard, however, giving readers Statistical Stories akin to ours, as e.g. Gelman et al. (2020) with its well-told regression stories and Brazzale et al. (2007) where statistical fine-tuning methods are delved into via a list of such stories. In this regard, our own book is admittedly not ten times more ambitious than McCullagh (2022)’s excellent book, even though we have a hundred stories and that book ten; he goes more deeply into several of these. (xx more from the theory-with-applications books. xx)

After having developed the solid groundwork of statistical inference, in chapters 1–5, our book dares to go decently far also into more specialised topics, expanding the horizon of modern statistical modelling and inference, in chapters 6–14. For these themes there are naturally other more specialised texts; here we may point to Brown (1986), McCullagh and Nelder (1989), and Efron (2023) for exponential families and generalised linear models; Basu et al. (2011) for robustness and minimum divergence; Robert (2007) xx and Gelman xx for Bayes; Schweder and Hjort (2016) for confidence distributions and meta-analysis; Billingsley (1968) and Pollard (1984) for processes and convergence; Andersen et al. (1993); Aalen et al. (2008) with emphasis on survival and event history analysis; Claeskens and Hjort (2008b) for model selection; Ghosh and Ramamoorthi (2003); Hjort et al. (2010); Ghosal and van der Vaart (2017) for Bayesian nonparametrics (xx also mueller xx). We place further comments in the Notes and Pointers sections at the end of our chapters.

(xx we point to *some* other books in similar terrain, though they are quite some distance away from our book. xx) Gelman et al. (2020) (yes, story telling), Ferguson (1967, 1996), Li et al. (2027), Lehmann (1959, 1983, 1999); Lehmann and Casella (1998), Keener (2011), ...

Go forth!

поехали!, as
Gagarin said

After having worked your way through significant chunks of this book, both chapters and well-sampled stories, you’ll be fit to do wonders and be sought after on the job market, the segments associated with the words statistics, data science and analysis, modelling, machine learning, prediction (with uncertainty), and even AI.

Thanks

The authors owe special thanks to Céline Cunen, Gudmund Hermansen, Tore Schweder, for having contributed significantly to several of our Statistical Stories, and also for always pleasant and inspiring long-term collaborations. Deep thanks are also due to a long list of colleagues and friends, who have taken part in discussions and rounds of clarification of relevance to various exercises and stories in our book: Marthe Aastveit, Patrick Ball, Bear Braumoeller, Gerda Claeskens, Aaron Clauset, Dennis Christensen, Riccardo de Bin, Ingrid Dæhlen, Arnaldo Frigessi, Abhik Ghosh, Ingrid Glad, Håvard Hegre, Aliaksandr Hubin, Ingrid Hobæk Haff, Kristoffer Hellton, Anders Dahl Hjort,

Martin Jullum, Vinnie Ko, George Lowther, Ian McKeague, Per Mykland, Per August Moen, Jonas Moss, Håvard Møkleiv Nygård, Lars Olsen, Sonia Petrone, Oskar Høgberg Simensen, Catharina Stoltenberg, Gunnar Taraldsen, Ingrid Van Keilegom, Lars Walløe, Jonathan Williams, Lan Zhang, Anti C. Pated.

anticipated

We have also benefitted, directly and indirectly, through the collective efforts of several grander wide-horizoned funded projects: the *FocuStat: Focus Driven Statistical Inference with Complex Data* 2014-2019 project (led by Hjort) at the Department of Mathematics, University of Oslo, funded by the Norwegian Research Council; the *Stability and Change* 2022-2023 project (led by Hjort and Hegre), funded by and hosted at the Centre for Advanced Study (CAS), Academy of Science and Letters, Oslo; and *Integreat: The Norwegian Centre for Knowledge-Driven Machine Learning* 2023-2033 Centre of Excellence (led by Frigessi and Glad), Oslo, funded by the Norwegian Research Council. We finally acknowledge with gratitude a partial support stipend from the Norwegian Non-Fiction Writers and Translators Association (Norsk faglitterær forfatter- og oversetterforening).

Nils Lid Hjort and Emil Aas Stoltenberg
Blindern, some day in 2026

Contents

Preface	iii
Contents	xi
I 600 Exercises	1
1 Statistical models	3
2 Large-sample theory	43
3 Parameters, estimators, precision, confidence	83
4 Testing, sufficiency, power	113
5 Maximum likelihood	145
6 Minimum divergence	195
7 Bayesian inference and computation	213
8 CDs, confidence curves, combining information	245
9 Loss, risk, performance, optimality	281
10 Converging processes, Brownian motion, and bridges	313
11 Survival and event history analysis	355
12 Model selection and model averaging	385
13 Markov chains and Markov processes	413
14 Bayesian nonparametrics	433
II 100 Statistical Stories	459
15 Demography, Epidemiology, Medicine	461
16 Biology, Climate, Ecology	509

17 Art, History, Literature, Music	543
18 Economics, Political Science, Sociology	589
19 Sports	625
20 Simulated stories	659
21 Miscellaneous stories	683
III Appendix	711
A Sets, functions, and metric spaces	713
B Measure and integration	717
C Probability	733
D Martingales	755
E Overview of stories and data	767
References	787
Subject index	801

Part I

600 Exercises

Survival and event history analysis

[xx chapter abstract here xx]

Key words: central limit theorems, counting processes, Cox regression, large-sample approximations, Kaplan–Meier survival curves, Lindeberg conditions, martingales, Nelson–Aalen cumulative hazards

Survival and event history analysis is the statistical study of the occurrence of events in the lives of individuals, animals, nations, firms, or whatever your units of interest may be. Studying the occurrence of events means trying to understand of how the probability of the event (or the probabilities of the events) of interest evolves over time, how the probability of the event of interest differs between individuals, and what factors may contribute to the probability of the event occurring. [xx mention that outcome is typically (always?) the time to an event, time between events xx]. In the *survival analysis* setting one is interested in the occurrence of a single event in the life of an individual, and it owes its name to this event being death. *Event histories*, on the other hand, concerns several different events occurring over time for an individual, or the individual experiencing the same event several times. A recurring theme in survival and event history analysis is *censoring*. In general, censoring means that the exact time of the event is not know, rather we know that it must have happened after some time (right-censoring), before some time (left-censoring), or somewhere in an interval of time (interval-censoring). The most intensely studied form of censoring is right-censoring, and it is to right-censoring we devote most attention in this chapter. Counting processes and the theory of martingales provide a powerful tool for studying estimators and testing procedures arising from censored survival and event history data. Both counting processes and martingales are instances of *stochastic processes*. See Appendix D. [xx write ‘multiplicative intensity model’ somewhere xx]

Discrete time survival analysis

Ex. 11.1 *A discrete time framework for survival analysis.* Consider an event occurring at one and only one of the time points $\{0, 1, 2, \dots\}$. We model this by a random variable X whose probability mass function is $f_j = \Pr(X = j) \geq 0$ for $j = 0, 1, 2, \dots$ and zero otherwise. Let $F_x = \sum_{j \leq x} f_j$ for $x \in \mathbb{R}$ be its c.d.f., and $S_x = 1 - F_{[x]-1}$ its *survival function*. It is fruitful to work with the *hazards*

$$\alpha_j = \Pr(X = j \mid X \geq j) = f_j / (f_j + f_{j+1} + f_{j+1} + \dots), \quad j = 0, 1, 2, \dots$$

along with the *cumulative hazards* $A_j = \alpha_0 + \alpha_1 + \cdots + \alpha_j$.

(a) We see that $S_0 = 1$, and $S_1 = 1 - f_0 = 1 - \alpha_0 S_0$. Proceed by induction to show that $S_{k+1} = \prod_{j=0}^k (1 - \alpha_j)$, and that, consequently, the c.d.f. can be expressed as $F_x = 1 - \prod_{j=0}^{\lfloor x \rfloor} (1 - \alpha_j)$. Deduce from this that the probability mass function can be written

$$f_j = (1 - \alpha_0)(1 - \alpha_1) \cdots (1 - \alpha_{j-1})\alpha_j, \quad \text{for } j = 0, 1, 2, \dots,$$

and give an interpretation of this identity.

(b) Suppose that X is subject to *right-censoring*, meaning that we only observe X if it occurs before some (random) *censoring time* C , that is, we observe $T = \min(X, C)$. In the canonical right-censoring setup it is assumed that C is independent of X ; that its distribution does not depend on the parameter on which we seek to make inference, in which case the censoring distribution is said to be *uninformative*; and that we know whether we observe a censored outcome or the real thing. Thus, our data are a realisation of (T, Δ) , with $\Delta = I(X \leq C)$ an indicator of noncensoring. Define the *counting process* and the *at-risk process* of observed lifetimes

$$N_j = \Delta I(T \leq j), \quad \text{and} \quad Y_j = I(T \geq j),$$

counting
process and
at-risk process

and denote $dN_j = N_j - N_{j-1}$. Show that $E(dN_j | Y_j) = Y_j \alpha_j$ almost surely.

(c) Let $(\mathcal{F}_j)_{j \geq 0}$ be the filtration generated by (N, Y) , i.e., $\mathcal{F}_j = \sigma((N_k, Y_k)_{0 \leq k \leq j})$ for $j = 0, 1, \dots$. Verify that Y_j is $\mathcal{F}_{j-1}$ -measurable for each j , in other words, that the information contained in $\mathcal{F}_{j-1}$ is sufficient to determine whether a subject is still at risk or not. A process possessing this peeking ahead property is said to be *previsible* or *predictable*.

previsible
process

(d) Define the processes M by $M_j = N_j - \sum_{k=1}^j Y_k \alpha_k$ for $j = 1, 2, \dots$. Show that M is a square integrable martingale. That is, M_j is $\mathcal{F}_j$ -measurable and square integrable for each j , and $E(M_j | \mathcal{F}_{j-1}) = M_{j-1}$ almost surely for each j .

counting
process
martingale

(e) To any square-integrable martingale M there is an associated *predictable quadratic variation* process denoted $\langle M, M \rangle$, defined by (see Ex. D.4)

$$\langle M, M \rangle_j = \sum_{k=1}^j \text{Var}(dM_k | \mathcal{F}_{k-1}), \quad \text{for } j = 1, 2, \dots$$

Show that the martingale from (d) has predictable quadratic variation $\langle M, M \rangle_j = \sum_{k=1}^j Y_k \alpha_k (1 - \alpha_k)$, and deduce that $M_j^2 - \langle M, M \rangle_j$ is a martingale.

(f) Consider now a full data set, assuming that $(T_1, \delta_1), \dots, (T_n, \delta_n)$ are independent replicates of (T, Δ) . Form the combined counting process N and the combined at-risk process Y , as

$$N_j = \sum_{i=1}^n \Delta_i I(T_i \leq j) \quad \text{and} \quad Y_j = \sum_{i=1}^n I(T_i \geq j) \quad \text{for } j = 0, 1, \dots$$

Formulate in words what these counts N and Y amount to; and verify that $M_j = N_j - \sum_{k=1}^j Y_k \alpha_k$ is a square integrable martingale, with respect to the filtration generated by the processes N and Y .

(g) Show, with the necessary efforts of book-keeping, that the likelihood information from such a dataset can be expressed as

$$L = \prod_{i:\Delta_i=1} f_{T_i} \prod_{i:\Delta_i=0} (1 - F_{T_i}) = \prod_{j=0}^{\infty} \alpha_j^{\mathrm{d}N_j} (1 - \alpha_j)^{Y_j - \mathrm{d}N_j}.$$

The product above is really a finite one, every factor after the largest T_i being 1.

(h) Discuss how or to what extent the likelihood in (f) can be interpreted as a succession of binomial trials, with $N_j \mid Y_j \sim \text{binom}(Y_j, \alpha_j)$.

Ex. 11.2 *Nelson–Aalen in discrete time.* Let $(T_1, \delta_1), \dots, (T_n, \delta_n)$ be independent replicates of (T, Δ) , where these are as defined in Ex. 11.1. This means that in the background here there are i.i.d. lifetimes $X_1, \dots, X_n$ and i.i.d. censoring times $C_1, \dots, C_n$ with distributions F and G , respectively.

(a) Show that the ML estimator for α_j is $\hat{\alpha}_j = \mathrm{d}N_j / Y_j$ provided $Y_j > 0$. We set $\hat{\alpha}_j = 0$ when $Y_j = 0$, thus from now on $\hat{\alpha}_k = I(Y_k > 0) \mathrm{d}N_k / Y_k$ where we take ‘0/0’ to be 0. The discrete time Nelson–Aalen estimator of the cumulative hazard is then $\hat{A}_j = \hat{\alpha}_0 + \dots + \hat{\alpha}_j = \sum_{k=0}^j I(Y_k > 0) \mathrm{d}N_k / Y_k$.

(b) Introduce $\tilde{A}_j = \sum_{k=0}^j \alpha_k I(Y_k > 0)$. Add and subtract $\tilde{A}_j$ to show that

$$\hat{A}_k - A_k = \sum_{j=0}^k \frac{I(Y_j > 0)}{Y_j} \mathrm{d}M_j + o_{\text{pr}}(1/\sqrt{n}),$$

in terms of the martingale differences $\mathrm{d}M_j = M_j - M_{j-1}$.

(c) Show that $\hat{A}_k - \tilde{A}_k$ is a square integrable martingale; and deduce from Lenglart’s inequality (see Ex. D.7) that this implies that for any $\varepsilon > 0$ and $\delta > 0$

$$\Pr(\max_{j \leq k} |\hat{A}_j - \tilde{A}_j| \geq \varepsilon) \leq \frac{\delta}{\varepsilon^2} + \Pr\left\{\sum_{j=0}^k \frac{I(Y_j > 0)}{Y_j} \alpha_j (1 - \alpha_j) \geq n\delta\right\}.$$

Derive conditions on the distributions F and G under which we are justified in concluding that $\max_{j \leq k} |\hat{A}_j - A_j| \rightarrow_{\text{pr}} 0$ as n tends to infinity.

(d) Consult the martingale central limit theorem in Ex. D.8, and deduce from the above efforts that

$$\sqrt{n}(\hat{A}_j - A_j) \rightarrow_d N(0, \sigma_j^2), \quad \text{with} \quad \sigma_j^2 = \sum_{k=0}^j \frac{\alpha_k(1 - \alpha_k)}{(1 - F_{k-1})(1 - G_{k-1})},$$

as n tends to infinity, for all j such that F_{j-1} and G_{j-1} are strictly smaller than 1.

Ex. 11.3 *Inference on the hazard.* Since $\alpha_j = A_j - A_{j-1}$, large sample results pertaining to the ML estimator $\hat{\alpha}_j$ can be deduced from those for the Nelson–Aalen estimator.

(a) Show that $\max_{j \leq k} |\hat{\alpha}_j - \alpha_j| \rightarrow_{\text{pr}} 0$ under appropriate conditions on F and G .

(b) Derive the limiting distribution of $\sqrt{n}(\hat{A}_j - A_j, \hat{A}_{j-1} - A_{j-1})$, and show that

$$\sqrt{n}(\hat{\alpha}_j - \alpha_j) \rightarrow_d N\left(0, \frac{\alpha_j(1 - \alpha_j)}{(1 - F_{j-1})(1 - G_{j-1})}\right),$$

via an application of the Cramér–Wold device (see Ex. 2.45(b)).

Ex. 11.4 Kaplan–Meier in discrete time. Suppose we have i.i.d. survival data as described in the preceding exercises. Estimating the c.d.f. or the survival function of X is complicated by the fact that we only observe the censored survival time $T = \min(X, C)$ along with an indicator $\Delta = I(X \leq C)$ of noncensoring.

(a) Convince yourself that $\tilde{F}_j = n^{-1} \sum_{i=1}^n I(T_i \leq j)$ is, in the presence of right-censoring, *not* a good estimator for F_j . The failure of this estimator is part of the rationale for constructing estimators based on $\hat{\alpha}_j$, as this latter is consistent for the hazard α_j .

(b) A natural estimator, indeed the ML estimator, for the survival function $S_x = \prod_{j=0}^{\lfloor x \rfloor - 1} (1 - \alpha_j)$ is the Kaplan–Meier estimator $\hat{S}_j = \prod_{k=0}^{j-1} (1 - \hat{\alpha}_k)$. Show that $\hat{S}_j \rightarrow_{\text{pr}} S_j$ for any j such that F_{j-1} and G_{j-1} are both strictly smaller than 1.

(c) Deduce from the above exercises that

$$\log \frac{\hat{S}_j}{S_j} = - \sum_{k=0}^{j-1} \frac{\hat{\alpha}_k - \alpha_k}{1 - \alpha_k} + O_{\text{pr}}(1/n),$$

and use this to find the limiting distribution of $\sqrt{n} \log(\hat{S}_j/S_j)$.

(d) Derive an approximate 95 percent confidence interval for S_k .

Ex. 11.5 Parametric likelihood inference in discrete time. Retain the setup from Ex. 11.1, and suppose that the probability mass function f_j generating the data belongs to a parametric family $\{f(\theta): \theta \in \Theta\}$ of probability mass functions on $\{0, 1, 2, \dots\}$, with $\Theta \subset \mathbb{R}$. Thus $\Pr(X = j) = f_j(\theta_0)$, where θ_0 is the true value of the parameter. Assume also that the processes $(N_1, Y_1), \dots, (N_n, Y_n)$ are only observed up to some finite time τ , for which F_τ and G_τ are both strictly smaller than 1; and that the hazard $\alpha_j(\theta) = f_j(\theta)/\{f_j(\theta) + f_{j+1}(\theta) + \dots\}$ is, for all $j = 1, \dots, \tau$, twice continuously differentiable in an open interval around θ_0 .

(a) Consider the log-likelihood function for θ based on the data observed up to and including point $k \leq \tau$. Verify that it is given by (disregarding terms constant in θ)

$$\ell_n(\theta)_k = \sum_{j=1}^k \{\log(\alpha_j(\theta)) dN_j + \log(1 - \alpha_j(\theta))(Y_j - dN_j)\},$$

and show that the score function process $s_n(\theta) = \{s_n(\theta)_j: j = 0, 1, \dots, \tau\}$, normalised by $n^{-1/2}$, is $s_n(\theta)_k = n^{-1/2} \partial \ell_n(\theta)_k / \partial \theta = n^{-1/2} \sum_{j=1}^k \psi_j(\theta) \{dN_j - Y_j \alpha_j(\theta)\}$, with $\psi_j(\theta) = (\partial \alpha_j(\theta) / \partial \theta) / \{\alpha_j(\theta)(1 - \alpha_j(\theta))\}$. In particular, define $s_n = s_n(\theta_0)$ and show that $(s_n)_{0 \leq j \leq \tau}$ is a square integrable martingale with respect to the filtration generated by the counting and at-risk processes.

(b) Denote $J_n(\theta) = -\partial s_n(\theta)_\tau / \partial \theta$, and let $J_n = J_n(\theta_0)$. Establish that

$$E \langle s_n, s_n \rangle_\tau = E J_n,$$

This equality indicates that the predictable quadratic variation $\langle s_n, s_n \rangle_k$ plays a role akin to the observed information that we met in Ch. 5 on likelihood theory.

(c) Let $\dot{\psi}_\theta$ be the derivative of ψ_θ with respect to θ . Note that $\dot{\psi}_\theta$ is continuous as $\alpha_j(\theta)$ is twice continuously differentiable for all j . Show that $J_n - \langle s_n, s_n \rangle_\tau \rightarrow_{\text{pr}} 0$. Moreover, show that $J_n(\hat{\theta}_n) - J_n \rightarrow_{\text{pr}} 0$ for any deterministic sequence $\hat{\theta}_n$ tending to θ_0 .

(d) Let $J = \sum_{j=1}^\tau \psi_j(\theta_0)^2 y_j \alpha_j(\theta_0) \{1 - \alpha_j(\theta_0)\}$, where y_j is the probability limit of Y_j/n . Give yourself the assumptions you need (take a look at the relevant exercises in Ch. 5), and show that, with $\hat{\theta}_n$ the ML estimator, $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, J^{-1})$.

Continuous time survival analysis

Ex. 11.6 *Survival functions and hazard rates.* Consider a lifetime variable T with density f and c.d.f. on the half-line; in particular, the distribution is continuous. Define the hazard rate function α as

$$\alpha(t) = \lim_{h \rightarrow 0} (1/h) \Pr\{T \in [t, t+h) \mid T \geq t\}.$$

So, informally, $\alpha(t) dt = \Pr(T \in [t, t+dt) \mid T \geq t)$ for a small time window $[t, t+dt)$, i.e., the conditional probability that T occurs in $[t, t+dt)$ given that it has not occurred prior to time t .

(a) First define the *survival function* as

$$S(t) = \Pr(T \geq t) = 1 - F(t).$$

What are its basic properties?

(b) Show that in fact $\alpha(t) = f(t)/S(t)$. So from knowledge of f we can always find the hazard rate from $\alpha = f/(1 - F)$.

(c) Define also the cumulative hazard rate function as $A(t) = \int_0^t \alpha(s) ds$. Show that

$$F(t) = 1 - \exp\{-A(t)\} \quad \text{and} \quad f(t) = \alpha(t) \exp\{-A(t)\}.$$

(d) Let T have the exponential distribution with density $f(t, \theta) = \theta \exp(-\theta t)$. Find its survival function and hazard rate.

(e) For the Weibull distribution, with $F(t) = 1 - \exp\{-(at)^b\}$, with the hazard rate function, and display it in a plot, for $a = 3.33$ and b equal to 0.9, 1.0, 1.1.

(f) Consider the Gamma distribution with parameters (a, b) , which has the density

$$f(t, a, b) = \frac{b^a}{\Gamma(a)} t^{a-1} \exp(-bt) \quad \text{for } t > 0.$$

Show that the mean and variance are a/b and a/b^2 . Take $b = 2.22$, compute the hazard rates for a equal to 0.8, 1.0, 1.2, and display these in a diagram. Give explicit formulae for the survival function and hazard rate for the case of $a = 2$.

(g) Consider a lifetime distribution with hazard rate $\alpha(t) = 1/(1+t)$. Find its survival function and density.

Ex. 11.7 *Maximum likelihood estimation with censored data.* If we have observed independent lifetime data $X_1, \dots, X_n$, from a suitable parametric density $f(x, \theta)$, the ML estimator is found by maximising the log-likelihood function $\sum_{i=1}^n \log f(X_i, \theta)$. This exercise looks into the required amendments in the case of censored data. In particular, let $C_1, \dots, C_n$ be independent censoring times (i.e., nonnegative random variables) from some common distribution with c.d.f. G , and assume that censoring is *uninformative* in the sense that G does not depend on θ . Assume also that the we have independent censoring, meaning that the X_i and C_i are independent. We observe the survival data $(T_1, \delta_1), \dots, (T_n, \delta_n)$ where $T_i = \min(X_i, C_i)$, thus there is *right-censoring*, and

$$\Delta_i = I(T_i \text{ is the observed lifetime}) = I(X_i \leq C_i),$$

So, in the case of $\Delta_i = 0$, this means that the real lifetime X_i , is at least as large as C_i , but we do not know more than that.

right-censoring

(a) Assume that the parametric model is given and perhaps primarily thought about via its hazard rate function, say $\alpha(t, \theta)$. Assume first that all T_i correspond to genuinely observed lifetimes, i.e. that there is no censoring. Show that the log-likelihood function above can be expressed as

$$\ell_n(\theta) = \sum_{i=1}^n \{\log \alpha(T_i, \theta) - A(T_i, \theta)\},$$

with $A(t, \theta)$ the cumulative hazard function for the model.

(b) For the very simple exponential model, with $\alpha(t, \theta) = \theta$, write up the log-likelihood function from the expression under (a), and show that the ML estimator is $\hat{\theta} = n / \sum_{i=1}^n T_i = 1/\bar{T}_n$.

(c) Then consider the general case with censoring as described above, i.e. some of the Δ_i are equal to zero. Show that the log-likelihood function can be written

$$\ell_n(\theta) = \sum_{\Delta_i=1} \log f(T_i, \theta) + \sum_{\Delta_i=0} \log S(T_i, \theta) = \sum_{i=1}^n \{\Delta_i \log \alpha(T_i, \theta) - A(T_i, \theta)\}.$$

Sometimes the first expression is more practical to work with, sometimes the second; also, as will be seen later, the second expression lends itself more easily to general counting process models.

(d) For the simple exponential model, again, but now with censoring, show that the ML estimator is $\hat{\theta} = \sum_{i=1}^n \Delta_i / \sum_{i=1}^n T_i$, generalising the non-censored case above.

(e) Generalise the previous situation to the model where the hazard rate is $\alpha(t) = \alpha_0(t)\theta$, with a known basis hazard function α_0 but with unknown multiplicative parameter θ .

Ex. 11.8 *Counting process, at-risk process, intensity process, martingales.* The eternal golden braid, for modelling and analysing survival and event history data, is the quadruple (N, Y, λ, M) . The ingredients are the counting process N , the at-risk process Y , the intensity process λ , and the martingale M . This particular exercise gives a separate, brief, and partial introduction to these four items, in the context of survival data. These data are of the form $(T_1, \delta_1), \dots, (T_n, \delta_n)$, as described in Ex. 11.7. The counting process N counts the number of observed events, over time:

$$N(t) = \#\{\text{observed events over } [0, t]\} = \sum_{i=1}^n \Delta_i I(T_i \leq t).$$

It starts out at zero, at time zero, and then increases with jump size 1 each time a new observation is recorded. The at-risk process counts those individuals who are still at risk, for each given time point:

$$Y(t) = \#\{\text{individuals at risk just before time } t\} = \sum_{i=1}^n I\{T_i \geq t\}.$$

The ‘just before’ can be formalised, e.g., via left continuity. Note that $Y(t)$ counts both those T_i with $\Delta_i = 1$ and those with $\Delta_i = 0$ (since we do not know yet when events occur, or when censoring might occur). The intensity process $\lambda(s)$ can be defined in several ways, and with somewhat different, but related, motivations and interpretations [xx point to Ex. in appendix xx]. The simplest way, in this framework, might be

$$\lambda(s) ds = \Pr\{N[s, s+ds] = 1 \mid \mathcal{F}_{s-}\}.$$

First,

$$dN(s) = N[s, s + ds] = N(s + ds) - N(s-)$$

is the number of observed events inside the small time window $[s, s + ds]$. Second, $\mathcal{F}_{s-}$ is the full history of everything that has been observed up to just before time s , i.e. over $[0, s)$. In the present setup of survival data (i.e. without complications of more complex event history constructions), the relevant information in all of $\mathcal{F}_{s-}$ is simply ‘how many are still at risk’, i.e. $Y(s)$.

(a) In this setup, show that

$$dN(s) | \mathcal{F}_{s-} \sim \text{Bin}(Y(s), \alpha(s) ds),$$

a simple binomial situation with $Y(s)$ at risk and with a small probability $\alpha(s) ds$. Show that

$$\begin{aligned} \Pr\{dN(s) = 0 | \mathcal{F}_{s-}\} &= 1 - Y(s)\alpha(s) ds + O((ds)^2), \\ \Pr\{dN(s) = 1 | \mathcal{F}_{s-}\} &= Y(s)\alpha(s) ds + O((ds)^2), \\ \Pr\{dN(s) \geq 2 | \mathcal{F}_{s-}\} &= O((ds)^2), \end{aligned}$$

This means that all the action here is in 0 (high chance) and 1 (slim chance, but important, and sooner or later it will kick in). In particular, show using the above that

$$\lambda(s) = Y(s)\alpha(s)$$

(b) Then consider the random process

$$M(t) = N(t) - \int_0^t \lambda(s) ds = N(t) - \int_0^t Y(s)\alpha(s) ds. \quad (11.1)$$

Demonstrate that it has the magical martingale property,

$$\mathbb{E}\{dM(s) | \mathcal{F}_{s-}\} = 0,$$

with $dM(s) = M(s + ds) - M(s)$ the martingale increment.

(c) Show that the process

$$K(t) = M(t)^2 - \int_0^t Y(s)\alpha(s) ds$$

also is a martingale. We shall see later, in the course and in exercises, that various central properties flow from these martingales, including results on limiting normality for classes of estimators.

(d) Consider again the situation of Ex. 11.7, with log-likelihood functions for censored data. With the golden quadruple on board, show that the log-likelihood function also can be expressed as

$$\ell_n(\theta) = \sum_{i=1}^n \{\delta_i \log \alpha(t_i, \theta) - A(t_i, \theta)\} = \int_0^\tau \{\log \alpha(s, \theta) dN(s) - Y(s)\alpha(s, \theta) ds\}.$$

Here the integral of a function with respect to a counting process is defined, simply, as

$$\int_0^\tau g(s) dN(s) = \sum_{i=1}^n g(t_i) \delta_i,$$

a sum of the function evaluated precisely at the observed timelengths.

Ex. 11.9 *More on counting process martingales.* In this exercise we formalise some of the matters from Ex. 11.8. Let N and Y be as defined in Ex. 11.8. We start out considering $n = 1$, so that the counting process N makes no more than one jump. The filtration $(\mathcal{F}_t)_{t \geq 0}$ generated by (N, Y) , that is, $\mathcal{F}_t = \sigma(\{N(s), Y(s)\}_{s \leq t})$ records the history of these two processes. The aim of this exercise is to show that the process M defined in (11.1) is a square integrable martingale with respect to $(\mathcal{F}_t)_{t \geq 0}$. For more on square integrable martingales, see Appendix D.

(a) Verify that M is adapted, i.e., that $M(t)$ is measurable with respect to $\mathcal{F}_t$ for each t . You might consult the result from Ex. B.19(b) in doing so. Show also that $E M(t)^2 \leq (1 - \log(1 - F(t)))^2$, establishing that M is square integrable.

(b) To conclude that M is indeed a square integrable *martingale* we must show the martingale property

$$E(M(t) | \mathcal{F}_s) = M(s), \quad \text{a.s., for each } s \leq t.$$

One way of proving this (feel free to attempt others) is to start by establishing that the martingale property is the same as $E\{N(t) - N(s) | \mathcal{F}_s\} = E\{\int_s^t Y(u)\alpha(u) du | \mathcal{F}_s\}$ a.s., for all $s \leq t$. Now, suppose that we manage to show that the left- and right hand side of this equation are respectively

$$E\{N(t) - N(s) | \mathcal{F}_s\} = aI(T > s), \quad \text{and} \quad E\{\int_s^t Y(u)\alpha(u) du | \mathcal{F}_s\} = bI(T > s),$$

almost surely, for non-random $a = a(s, t)$ and $b = b(s, t)$. Show that if this is so, then $a = b$, and the martingale property is proven.

(c) To prove the two equalities in the second display in (b), it suffices to show that for any event $A \in \mathcal{F}_s$ such that $\Pr(A) > 0$, we have $\Pr(T > s) = \Pr(A \cap \{T > s\})$. Verify that this is indeed sufficient, and show that the equalities hold.

(d) Suppose that the survival time T and indicator for noncensoring Δ come with a covariate Z . We regard this covariate as ‘realised at time zero’, and define the filtration $\mathcal{G}_t = \mathcal{F}_t \vee \sigma(Z)$ for $t \geq 0$. Generalise the above, and give yourself the proper technical assumptions, to show that in this setup $M(t) = N(t) - \int_0^t Y(s)\alpha(s, Z) ds$ is a square integrable martingale with respect $(\mathcal{G}_t)_{t \geq 0}$.

(e) Going one step further, suppose that we have a covariate process $Z(t)$ that of the previsible sort. In particular, with $\mathcal{Z}_t = \sigma(Z_s; s \leq t)$, the random variable Z_t is $\mathcal{Z}_{t-}$ -measurable (cf. the discussion about the at-risk process). Define $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{Z}_t$, and attempt to show that, under suitable conditions, $M(t) = N(t) - \int_0^t Y(s)\alpha(s, Z(s)) ds$ is a square integrable martingale with respect $(\mathcal{G}_t)_{t \geq 0}$.

(f) Show that a sum of martingales is a martingale. In particular, if $M_1, \dots, M_n$ are i.i.d. replicates of the martingale M defined in (11.1), or, indeed, the ones described in (d) or (e), then $\sum_{i=1}^n M_i(t)$ is a square integrable martingale.

Ex. 11.10 *The previsible variation process.* The previsible quadratic variation process of a square integrable martingale M , is an increasing previsible process, denoted $\langle M, M \rangle$, such that $M^2 - \langle M, M \rangle$ is a martingale. More details are given in Ex. D.16. In this exercise we study the previsible quadratic variation of the counting process martingales introduced in Ex. 11.8 and Ex. 11.9.

(a) Let M be a square integrable martingale. Establish that $E\{M(t)^2 | \mathcal{F}_{t-}\} = \langle M, M \rangle_t$ a.s.. As we will see in Ex. abc, this identity is a special case of what is called the Itô isometry.

(b) Let $M_1, \dots, M_n$ be independent square integrable martingales, and let $a_1, \dots, a_n$ be constants. Demonstrate that

$$\left(\sum_{i=1}^n a_i M_i(t)\right)^2 - \sum_{i=1}^n a_i^2 \langle M_i, M_i \rangle_t, \quad t \geq 0,$$

is a martingale (in t), implying that $\langle \sum_{i=1}^n a_i M_i, \sum_{i=1}^n a_i M_i \rangle_t = \sum_{i=1}^n a_i^2 \langle M_i, M_i \rangle_t$. Thus, in this regard, the previsible quadratic variation behaves like the variance of a sum of independent random variables.

(c) Let M be the martingale defined in (11.1). Show that as k tends to infinity.

$$\sum_{j=1}^k \mathbb{E} \left\{ \left(M(jt/k) - M((j-1)t/k) \right)^2 \mid \mathcal{F}_{(j-1)t/k} \right\} = \int_0^t Y(s) \alpha(s) ds + o_{\text{pr}}(1),$$

almost surely, for each t . Deduce from this that

$$\langle M, M \rangle_t = \int_0^t Y(s) \alpha(s) ds.$$

and, consequently, that with $M_1, \dots, M_n$ i.i.d. replicates of the M in (11.1), we have

$$\left\langle \sum_{i=1}^n M_i, \sum_{i=1}^n M_i \right\rangle_t = \int_0^t \sum_{i=1}^n I(T_i \geq s) \alpha(s) ds = \int_0^t Y(s) \alpha(s) ds,$$

with $Y(s) = \sum_{i=1}^n I(T_i \geq s)$ as before. Thus, for this setup, the previsible quadratic variation is hence identical to the so-called compensator $\int_0^t Y \alpha ds$ of the counting process N .

(d) Generalise the findings from (c) to the martingales introduced in Ex. 11.9(d)-(e).

Ex. 11.11 *A parametric step-function for the hazard rate.* Consider independent lifetime data of the form $(T_1, \delta_1), \dots, (T_n, \delta_n)$, as met with in Ex. 11.7 and subsequent exercises, and assume they stem from a common distribution with hazard rate $\alpha(s)$. We wish to estimate the cumulative hazard rate $A(t) = \int_0^t \alpha(s) ds$. This can famously be done using the (nonparametric) Nelson–Aalen estimator, see Ex. 11.12 below, but we start out working through a parametric version, via a step-function. When the windows become small, this parametric estimator will actually converge to the Nelson–Aalen estimator.

(a) Consider the parametric model where the hazard rate is a step-function, i.e. constant over windows. Suppose $[0, \tau]$ is the full time window of relevance (i.e. with a sufficiently big end-point τ), with windows $W_j = [s_{j-1}, s_j]$ for $j = 1, \dots, k$, and $0 = s_0 < s_1 < \dots < s_k = \tau$. The model is then of the form

$$\alpha(s) = \alpha_j \text{ on time window } W_j \quad \text{for } j = 1, \dots, k.$$

Using the log-likelihood expression of Ex. 11.7, show that the log-likelihood function can be written as

$$\ell_n(\alpha_1, \dots, \alpha_k) = \sum_{j=1}^k \int_{W_j} \{ \log \alpha(s) dN(s) - Y(s) \alpha(s) ds \} = \sum_{j=1}^k (\Delta N_j \log \alpha_j - R_j \alpha_j),$$

in which

$$\Delta N_j = N(W_j) = N(s_j) - N(s_{j-1}) \quad \text{and} \quad R_j = \int_{W_j} Y(s) ds.$$

So ΔN_j is the number of observed events, and R_j is the ‘total time at risk’, over window W_j .

(b) Show the ML estimators for the local constants become

$$\hat{\alpha}_j = \frac{\Delta N_j}{R_j} = \frac{\Delta N_j}{\int_{W_j} Y(s) ds}, \quad \text{for } j = 1, \dots, k,$$

whenever $R_j > 0$. If $R_j = 0$ then we set $\hat{\alpha}_k = 0$ for all $k \geq j$. The ML estimator of the full cumulative hazard rate is hence the integral of the estimated step-function, which becomes the piecewise linear

$$\hat{A}(t) = \begin{cases} \hat{\alpha}_1 t & \text{for } t \in W_1, \\ \hat{\alpha}_1 s_1 + \hat{\alpha}_2(t - s_1) & \text{for } t \in W_2, \\ \hat{\alpha}_1 s_1 + \hat{\alpha}_2(s_2 - s_1) + \hat{\alpha}_3(t - s_2) & \text{for } t \in W_3, \\ \text{etc.} \end{cases}$$

(c) Find the Hessian matrix

$$J(\alpha) = -\partial^2 \ell_n(\alpha_1, \dots, \alpha_k) / \partial \alpha \partial \alpha^\dagger,$$

and show that it is diagonal. Let’s write α for the full parameter vector $(\alpha_1, \dots, \alpha_k)$ where convenient. Find also the observed information, which is this minus the second order derivative matrix computed at the ML estimate. Show indeed that

$$\hat{J} = J(\hat{\alpha}) = \text{diag}(R_1^2/\Delta N_1, \dots, R_k^2/\Delta N_k).$$

(d) Attempt, using the martingale central limit theorem of Ex. D.19, to show that

$$\sqrt{n}(\hat{\alpha} - \alpha) \rightarrow_d N_k(0, J^{-1}),$$

as $n \rightarrow \infty$, and so, approximately, $\hat{\alpha} \approx_d N_k(\alpha, \hat{J}^{-1})$ for large enough sample sizes. Show that this translates to the $\hat{\alpha}_j$ being approximately unbiased, normal, and independent, with $\text{Var } \hat{\alpha}_j \doteq \Delta N_j / R_j^2$.

(e) Argue, perhaps heuristically, that when the time windows become carefully small, then the above $\hat{A}(t)$ is in effect a nonparametric estimator of the cumulative hazard function, approximately unbiased and normal, and with variance estimable by

$$\hat{\kappa}(t)^2 = \sum_{\text{windows left of } t} \frac{\Delta N_j}{(R_j/d_j)^2},$$

with $d_j = s_j - s_{j-1}$ the width of window W_j . Since $R_j/d_j = (1/d_j) \int_{W_j} Y(s) ds$, this is close to $Y(s_j^*)$, with s_j^* the mid-point of W_j . A further approximation, which becomes correct in a fine-tuned large-sample setup with cells becoming small at the right rate, is then

$$\hat{\kappa}(t)^2 = \int_0^t \frac{dN(s)}{Y(s)^2}.$$

Via the step-function model, and some extra analysis, we have essentially reinvented the Nelson–Aalen estimator, along with its properties.

Ex. 11.12 *The Nelson–Aalen estimator.* Consider again independent lifetime data of the form $(T_1, \delta_1), \dots, (T_n, \delta_n)$, as met with in Ex. 11.7 and several subsequent exercises, and assume they stem from a common distribution with hazard rate $\alpha(s)$. In Ex. 11.11 we set up a step-function model for $\alpha(s)$, which led to an almost nonparametric estimator for the cumulative $A(t) = \int_0^t \alpha(s) ds$. The canonical nonparametric estimator is indeed this fine-tuned limit, namely the Nelson–Aalen estimator.

(a) We start with the definition, using the counting process and at-risk process notation of Exercise 5 (and used in the book). The Nelson–Aalen estimator for $A(t)$ is

$$\hat{A}(t) = \int_0^t \frac{dN(s)}{Y(s)} = \sum_{t_i \leq t} \frac{\delta_i}{Y(t_i)}.$$

Again, the integral with respect to a counting process is simply the finite sum over the appropriate integrand, over the observed event times. It is easy to make a programme computing $\hat{A}(t)$. Do this, for a dataset of your choice, perhaps simulated.

(b) Then we ought to spend a few minutes thinking about why the Nelson–Aalen $\hat{A}(t)$ is a natural estimator of $A(t)$. Using the martingale M of Exercise 5, we may write

$$dN(s) = Y(s)\alpha(s) ds + dM(s) = \text{structure} + \text{random fluctuations},$$

which implies

$$dN(s)/Y(s) = \alpha(s) ds + \text{noise}.$$

Argue that this points to the Nelson–Aalen.

(c) With a bit of heuristics, we have

$$\hat{A}(t) - A(t) = \int_0^t \frac{dN(s)}{Y(s)} - A(t) \doteq \int_0^t \frac{dM(s)}{Y(s)} = \int_0^t \frac{1}{\hat{y}(s)} \frac{dM(s)}{n},$$

where $\hat{y}(s) = Y(s)/n$ is a steadily more precise estimate of its limit in probability, say $y(s) = \Pr\{X \geq s, C \geq s\}$, with C the censoring mechanism and X the true uncensored lifetime. It follows that

$$Z_n(t) = \sqrt{n}\{\hat{A}(t) - A(t)\} \doteq \int_0^t \frac{1}{\hat{y}(s)} \frac{dM(s)}{\sqrt{n}}$$

But $M(t)/\sqrt{n} \rightarrow_d V(t)$, say, a zero-mean Gaussian martingale with incremental variances $\text{Var } dV(s) = y(s)\alpha(s) ds$, by the results in Appendix D. This, at least heuristically, is seen to imply

$$Z_n(t) \rightarrow_d Z(t) = \int_0^t \frac{1}{y(s)} dV(s),$$

which is another zero-mean Gaussian martingale with incremental variances

$$\text{Var } dZ(s) = \text{Var } \frac{dV(s)}{y(s)} = \frac{\alpha(s) ds}{y(s)}.$$

(d) So the Nelson–Aalen is for large samples approximately unbiased, approximately normal, and with variance

$$\sigma(t)^2 = \text{Var } \hat{A}(t) \doteq \frac{1}{n} \int_0^t \frac{\alpha(s) ds}{y(s)}.$$

Give arguments supporting the estimator

$$\hat{\sigma}(t)^2 = \int_0^t \frac{dN(s)}{Y(s)^2}.$$

So a programme for the Nelson–Aalen just needs a few lines more to produce also $\hat{\sigma}(t)$. In particular, confidence bands are now easy to construct, say

$$\hat{A}(t) \pm 1.645 \hat{\sigma}(t) \quad \text{for } t \in [0, \tau],$$

where $[0, \tau]$ is a relevant time window for the data. Try to show that this band contains the true $A(t)$ with probability converging to 0.90, for each fixed t .

Ex. 11.13 *Integration with respect to counting process martingales.* We’ve met the full eternal golden quadruple (N, Y, λ, M) in Exercise 11.8. Here we go through more details regarding the martingale machinery, in particular integrals with respect to counting process martingales (for even more details, see Appendix D). The standard setup so far is for the survival data framework of data $(T_1, \delta_1), \dots, (T_n, \delta_n)$, as with Exercises 11.7, for which the counting process martingale is $M(t) = N(t) - \int_0^t Y(s)\alpha(s) ds$. Throughout, $(\mathcal{F}_t)_{t \geq 0}$ is the filtration generated by the counting process N and the at-risk process Y . (a) Now consider any function $H = \{H(s) : s \geq 0\}$, and form its integral with respect to the martingale M ,

$$K(t) = \int_0^t H(s) dM(s) \quad \text{for } t \geq 0.$$

This integral may be defined as the fine limit of Riemann type sums $\sum_j H(s_j)\{M(s_{j+1}) - M(s_j)\}$, when the cells $[s_j, s_{j+1})$ become smaller, but for the present purposes of the survival setup it is sufficient to agree on

$$\int_0^t H dM = \int_0^t H(s)\{dN(s) - Y(s)\alpha(s) ds\} = \sum_{t_i \leq t} H(t_i)\delta_i - \int_0^t H(s)Y(s)\alpha(s) ds.$$

Now, show that $K = \int_0^\cdot H dM$ is also a martingale, provided H is *previsible*, in the sense that the value of $H(s)$ is known when $\mathcal{F}_{s-}$ is known, i.e., $H(s)$ is measurable with respect to $\mathcal{F}_{s-}$.

(b) When H is previsible (with respect to the same filtration of growing history), such that $K = \int_0^\cdot H dM$ is another martingale, show that

$$\langle \int_0^\cdot H dM, \int_0^\cdot H dM \rangle(t) = \int_0^t H(s)^2 d\langle M, M \rangle(s).$$

(c) For the survival data setup, consider the random function

$$K(t) = \int_0^t \frac{J(s)}{Y(s)^{1/2}} dM(s),$$

where $J(s) = I\{Y(s) \geq 1\}$ is equal to 1 with very high probability (unless s becomes large). The point is that $1/Y(s)$ isn’t defined when $Y(s) = 0$, and we take $J(s)/Y(s)$ to be zero in case of $Y(s) = 0$, which also means $J(s) = 0$. Show that K is a martingale, with variance process

$$\langle K, K \rangle(t) = A^*(t) = \int_0^t J(s)\alpha(s) ds,$$

which with very high probability is equal to $A(t)$ itself.

(d) Consider now two martingales, say M_1 and M_2 , with respect to the same filtration. We define their *previsible covariance process* $\langle M_1, M_2 \rangle$ via

$$d\langle M_1, M_2 \rangle(s) = \text{cov}\{dM_1(s), dM_2(s) | \mathcal{F}_{s-}\},$$

again with notation $dM_j(s) = M_j(s + ds) - M_j(s)$. If

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \alpha(s) ds$$

is the little martingale associated with individual i only, so that $N_i(t) = \delta_i I(t_i \leq t)$ and $Y_i(t) = I(T_i \geq t)$, taking on values 0 and 1 only, show that

$$\langle M_i, M_j \rangle(t) = 0 \quad \text{for } i \neq j,$$

whereas $\langle M_i, M_i \rangle(t) = \int_0^t Y_i(s) \alpha(s) ds$.

(e) Finally, at this stage, consider two martingales M_1 and M_2 , along with two previsible processes H_1 and H_2 , such that $K_1 = \int_0^\cdot H_1 dM_1$ and $K_2 = \int_0^\cdot H_2 dM_2$ become martingales. Show that

$$\langle K_1, K_2 \rangle(t) = \langle \int_0^\cdot H_1 dM_1, \int_0^\cdot H_2 dM_2 \rangle(t) = \int_0^t H_1 H_2 d\langle M_1, M_2 \rangle(s).$$

In particular, if M_1 and M_2 are orthogonal, meaning that their covariance process is zero, then also $\int_0^\cdot H_1 dM_1$ and $\int_0^\cdot H_2 dM_2$ are orthogonal, even if H_1 and H_2 might be dependent in complicated ways. It suffices that M_1 and M_2 have uncorrelated increments.

Ex. 11.14 *Martingales have Gaussian process limits.* Here I go through a few things having to do with the remarkable and powerful machinery of *martingale limit theorems*, but without the finer details. Such finer details are included in Appendix D. The overall message is that yes, lo & behold, a martingale $M_n(t)$, indexed by sample size n , is approximately normal, for each t , when n grows. *Even more*, the full random process $M_n = \{M_n(t) : t \geq 0\}$ tends in distribution, when properly scaled, to a full Gaussian martingale process, with independent and normally distributed increments (which is saying much more than merely ‘for each t , the distribution of $M_n(t)$ is close to a normal’, see Ch. 10). All these results are important and very useful generalisations of the results for independent summands that we met in Ch. 2. For survival analysis models and methods we very much need results with even complicated dependencies on the past.

(a) Let $X_1, X_2, \dots$ be a sequence of variables for which

$$E(X_n | \mathcal{F}_{n-1}) = 0 \quad \text{for } n = 2, 3, \dots,$$

where $\mathcal{F}_{n-1}$ means the previous history $(X_1, \dots, X_{n-1})$, that is, $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ for each $n \geq 1$. Show that

$$M_n(t) = \sum_{i \leq [nt]} X_i \quad \text{for } t \geq 0$$

is a martingale with respect to $\mathcal{F}_n$. Note that X_n can depend on the past in even complicated ways, as long as its conditional mean is zero.

(b) Show that

$$\langle M_n, M_n \rangle(t) = \sum_{i \leq [nt]} V_i, \quad \text{where } V_i = \text{Var}(X_i | \mathcal{F}_{i-1}).$$

In the easier special case of independence, as with the previous exercise, the $\langle M_n, M_n \rangle(t)$ is just the sum of the variances of the X_i ; here it is rather the sum of the *conditional variances*, and these, being conditional, are random.

(c) There are now various theorems (for one such, see Ex. D.8) which say that if (i) $\langle M_n, M_n \rangle(t) \rightarrow_{\text{pr}} v(t)$ for each t and (ii) some Lindeberg type condition holds, then there is full process convergence $M_n(t) \rightarrow_d V(t)$, a Gaussian martingale with variance $v(t)$. Show that $V(t)$ has the same distribution as $W(v(t))$, with W standard Brownian motion.

(d) Now consider a more complicated process, namely

$$K_n(t) = \sum_{i \leq [nt]} H_i X_i \quad \text{for } t \geq 0,$$

where the sequence of $H_1, H_2, \dots$ are *previsible*, meaning that H_i is known once $\mathcal{F}_{i-1}$ is known. Show that K_n is a martingale, with

$$\langle K_n, K_n \rangle(t) = \sum_{i \leq [nt]} H_i^2 \Delta \langle M_n, M_n \rangle_i = \sum_{i \leq [nt]} H_i^2 \text{Var}(X_i | \mathcal{F}_{i-1}).$$

These are parallels to what we've seen and worked with in Exercise 9.

(e) Since a martingale limit theorem is a martingale limit theorem, deduce that as long as (i) $\langle K_n, K_n \rangle(t) \rightarrow_{\text{pr}} q(t)$ for each t and (ii) some Lindeberg type condition holds for K_n , then there is full process convergence $K_n(t) \rightarrow_d Q(t) = W(q(t))$, a Gaussian martingale with variance $q(t)$. Note that such K_n processes can be much more complicated than simpler sums of independent components processes.

Ex. 11.15 Integrating sequence converging in probability. When working with convergence (of various sorts) of counting processes, the following situation arise frequently: A sequence of stochastic processes $X_n = \{X_n(s) : s \in [0, \tau]\}$ is such that

$$X_n(s) \rightarrow_{\text{pr}} x(s), \quad (11.2)$$

as $n \rightarrow \infty$ for each (or almost all) $s \in [0, \tau]$, where $\int_0^\tau |x(s)| ds$ is finite. We would like to conclude that

$$\int_0^t X_n(s) ds \rightarrow_{\text{pr}} \int_0^t x(s) ds, \quad \text{for each } t \in [0, \tau], \quad (11.3)$$

as well. This is somewhat similar to the convergence theorems of Ex. B.14, and also the uniform integrability results in Ex. 2.5, but differs from these because we are dealing with sequences of stochastic processes.

(a) To start off, show that if $\sup_{s \in [0, \tau]} |X_n(s) - x(s)| \rightarrow_{\text{pr}} 0$, then (11.3) holds.

(b) Suppose (11.2) holds, that $E X_n(s) \rightarrow x(s)$ for each $s \in [0, \tau]$ and that $E |X_n(s)| \leq k(s)$ for an integrable $k(s)$. Show that this is sufficient for (11.3) to hold. In fact, show that the convergence is uniform over $[0, \tau]$.

(c) We can weaken the result from (b), replacing convergence in mean by uniform integrability. Assume (11.2) holds and that $(X_n(s))_{n \geq 1}$ is uniformly integrable for each $s \in [0, \tau]$; and that $E |X_n(s)| \leq k(s)$ for an integrable function $k(s)$. Show that $\int_0^t X_n(s) ds$ converges in probability to $\int_0^t x(s) ds$, uniformly over $[0, \tau]$.

(d) Again, suppose that (11.2) holds and that $x(s)$ is integrable; and that for each $\delta > 0$ there is an integrable function $k_\delta(s)$ such that $\liminf_n \Pr(|X_n(s)| \leq k_\delta(s)) \geq 1 - \delta$. Define

$$Y_n = \begin{cases} X_n, & X_n \leq k_\delta, \\ k_\delta, & \text{otherwise,} \end{cases} \quad \text{and} \quad y = \begin{cases} x, & x \leq k_\delta, \\ k_\delta, & \text{otherwise.} \end{cases}$$

Explain why $\sup_{t \in [0, \tau]} |\int_0^t Y_n(s) ds - \int_0^t y(s) ds| \rightarrow_{\text{pr}} 0$ as $n \rightarrow \infty$, and deduce that the same holds when Y_n and y are replaced by X_n and x , respectively.

(e) Let Y_n be the at-risk process as described in Ex. 11.8, in particular, the possibly right-censored survival data $(T_1, \delta_1), \dots, (T_n, \delta_n)$ are i.i.d. Show that $Y_n(t)/n \rightarrow_{\text{pr}} y(t)$ for each $t \in [0, \tau]$, where $y(t) = \{1 - F(t)\}\{1 - G(t)\}$. Explain that, in fact, this convergence is uniform in t .

Ex. 11.16 *More on the Nelson–Aalen estimator.* Here are a few more technical details and supplementing remarks regarding the Nelson–Aalen estimators, compared to statements reached in Ex. 11.12. So we work, again, with

$$\hat{A}(t) = \int_0^t \frac{dN(s)}{Y(s)} = \sum_{t_i \leq t} \frac{\delta_i}{Y(t_i)}.$$

We define ‘0/0’ as 0 when time has run long enough to have $Y(s) = 0$, i.e. nobody left.

(a) Note that this construction makes sense also in event history analysis setups, for each intensity direction, so to speak. There would then be Nelson–Aalen estimators of the form

$$\hat{A}_{i,j}(t) = \int_0^t \frac{dN_{i,j}(s)}{Y_i(s)},$$

from box i to box j on the map of possible stations to occupy, with $Y_i(s)$ the number at risk in box i at time just before s , and $dN_{i,j}(s)$ the number of those for which an event occurs inside the short time interval $[s, s + ds)$. For the rest of this exercise, we’re in the simpler survival setup, though.

(b) With $M(t) = N(t) - \int_0^t Y(s)\alpha(s) ds$ the martingale, show that

$$\hat{A}(t) = \int_0^t \frac{J(s)}{Y(s)} dN(s) = A^*(t) + \int_0^t \frac{J(s)}{Y(s)} dM(s),$$

where $J(s) = I\{Y(s) \geq 1\}$ and $A^*(t) = \int_0^t J(s)\alpha(s) ds$. Find a formula for $\Pr\{J(s) = 1\}$, and show that it is very close to 1, unless s is growing big.

(c) Show that $\hat{A} - A^*$ is a martingale, with variance process

$$\langle \hat{A} - A^*, \hat{A} - A^* \rangle(t) = \int_0^t \frac{J(s)}{Y(s)^2} d\langle M, M \rangle(s) = \int_0^t \frac{J(s)}{Y(s)} \alpha(s) ds.$$

(d) Show then that the variance of $\hat{A}(t) - A^*(t)$, which is very nearly the same as the variance of $\hat{A}(t)$, can be expressed as

$$\sigma(t)^2 = \mathbb{E} \int_0^t \frac{J(s)}{Y(s)} \alpha(s) ds.$$

Argue that a natural estimator of $\sigma(t)^2$ is

$$\hat{\sigma}(t)^2 = \int_0^t \frac{dN(s)}{Y(s)^2} = \sum_{T_i \leq t} \frac{\delta_i}{Y(T_i)^2}.$$

This is about the same as covered inside Ex. 11.12, but now with a bit more detail.

(e) A somewhat more elaborate version of the variance estimator above is

$$\hat{\sigma}(t)^2 = \sum_{T_i \leq t} \frac{1}{Y(T_i)} \hat{p}_i (1 - \hat{p}_i) = \sum_{T_i \leq t} \frac{1}{Y(T_i)} \frac{\Delta N(T_i)}{Y(T_i)} \left\{ 1 - \frac{\Delta N(T_i)}{Y(T_i)} \right\}.$$

First, we use $\Delta N(T_i)$, the number of events observed inside $[T_i, T_i + \varepsilon)$, for a small ε , in the case of ties (the theory says that we should not have two events at the very same time, but in practice data are not always given with very fine time precision). Second, going through arguments above one learns that the variance of $dN(s)$ given the past enters the variance process arguments, and this conditional variance is $Y(s)p(1-p)$, with the small $p = \alpha(s) ds$. In the fine limit, where ds becomes infinitesimal, $p(1-p) = p$, so to speak, but it is possible that the above variance estimator is just slightly better when $Y(s)$ is relatively small; then the estimate $\hat{p} = \Delta N(t)/Y(t)$ is not so small, etc.

(f) But we can't quite live on with just a sensible estimator of the sensible variance of a sensible estimator; we need approximate normality, in order to construct confidence intervals and bands, tests, comparisons, etc. So let us prove the $\sqrt{n}\{\hat{A}(t) - A(t)\}$ has a normal limit process. First, show that

$$\sqrt{n}\{A^*(t) - A(t)\} \rightarrow_{\text{pr}} 0, \quad \text{for each } t,$$

which means that it is enough to find the limit distribution in question for the simpler

$$Z_n(t) = \sqrt{n}\{\hat{A}(t) - A^*(t)\}.$$

But this is a martingale. Show that

$$\hat{y}(s) = Y(s)/n \rightarrow_{\text{pr}} y(s), \quad \text{for each } s,$$

under mild conditions on the censoring distribution. Show that

$$\langle Z_n, Z_n \rangle(t) = \int_0^t \frac{nJ(s)}{Y(s)} \alpha(s) ds \rightarrow_{\text{pr}} v(t) = \int_0^t \frac{1}{y(s)} \alpha(s) ds.$$

This secures, with a small extra technical argument having to do with Lindeberg conditions, see Ex. D.19, that

$$Z_n(\cdot) \rightarrow_d V(\cdot) = W(v(\cdot)),$$

a time-transformed Brownian motion, with variance $\text{Var } V(t) = v(t)$.

(g) The final statement we need under our belts is that $\hat{\sigma}(t)$ is consistent for $\sigma(t)$, or more properly that $n\hat{\sigma}(t)^2$ converges in probability to the limit of $n\sigma(t)^2$, which is the $v(t)$ above. Try to prove this.

(h) From these statements prove that

$$\frac{\hat{A}(t) - A(t)}{\hat{\sigma}(t)} = \frac{\sqrt{n}\{\hat{A}(t) - A(t)\}}{\sqrt{n}\sigma(t)} \frac{\sigma(t)}{\hat{\sigma}(t)} \rightarrow_d N(0, 1),$$

for each t . In particular, show from this that

$$\Pr\{A(t) \in \hat{A}(t) \pm z\hat{\sigma}(t)\} \rightarrow \Pr\{-z \leq N(0, 1) \leq z\},$$

yielding pointwise confidence bands, tests for hypotheses of the type $A = A_0$, etc.

Ex. 11.17 *The Kaplan–Meier estimator.* The cumulative hazards of the world are more versatile and useful tools, particularly in more complex event history setups, than ‘only’ the task of estimating the survival curve $S(t) = \Pr(T \geq t)$ for survival data $(T_1, \delta_1), \dots, (T_n, \delta_n)$. The canonical nonparametric estimator for $S(t)$ remains however an important quantity, and this is the Kaplan–Meier estimator. Its definition is

$$\hat{S}(t) = \prod_{[0,t]} \left\{ 1 - \frac{dN(s)}{Y(s)} \right\} = \prod_{t_i \leq t} \left\{ 1 - \frac{\delta_i}{Y(t_i)} \right\}.$$

(a) For non-censored data, say $t_1 < \dots < t_n$, show that $\hat{S}(t)$ becomes the simpler empirical survival function

$$\hat{S}_{\text{emp}}(t) = (1/n) \sum_{i=1}^n I\{t_i \geq t\} = 1 - F_{\text{emp}}(t).$$

(b) From Ex. 11.6, we do, under certain conditions, have the easy formula $S = \exp(-A)$, binding together the survival curve and the cumulative hazard, so it is not at all forbidden to start with the Nelson–Aalen and then exp its minus to arrive at

$$\tilde{S}(t) = \exp\{-\hat{A}(t)\}.$$

The $A = -\log(1 - F)$ formula is valid only for continuous distributions, however, so that particular connection is not as straightforward for non-continuous step-function type estimators as $\hat{A}$ and $\hat{S}$. You may however attempt to prove that the Kaplan–Meier $\hat{S}$ and Aalen-related estimator are quite close; in particular, when n increases,

$$\sqrt{n}[\hat{S}(t) - \exp\{-\hat{A}(t)\}] \rightarrow_{\text{pr}} 0.$$

This means that these two related estimators have the same large-sample properties. Proving the above is easier when working on the log-scale; attempt to show that

$$\sqrt{n}[\hat{A}(t) + \log \hat{S}(t)] \rightarrow_{\text{pr}} 0.$$

(c) (xx briefly on how to assess and estimate the variance of $\hat{S}(t)$. the Greenwood formula (from 1926!). limiting normality and confidence bands. xx)

Ex. 11.18 *Smoothing the Nelson–Aalen estimator.* In this exercises we look into ways of smoothing the Nelson–Aalen estimator to obtain estimates of the hazard rate, akin to kernel density estimates for density functions, see Ch. abc. The naive and perhaps most straightforward way of estimating the hazard rate is by ‘differentiating’ the Nelson–Aalen estimator, that is, using $\hat{\alpha}_n(t) = h^{-1}\{\hat{A}(t+h) - \hat{A}(t)\} = h^{-1} \int_t^{t+h} d\hat{A}(s)$ for some small $h > 0$. A slightly more general version of this takes

$$\tilde{\alpha}_n(t) = \frac{1}{h} \int_0^\tau K\left(\frac{t-s}{h}\right) d\hat{A}(s),$$

where K is a (bounded) kernel density and h is some bandwidth. In this exercise we retain the assumptions from Ex. ?? (e.g., the processes are observed over the time window $[0, \tau]$), and introduce some new ones. We will assume that the kernel has support in $[-1, 1]$ (to ease the notation), and the estimator above is only defined for $t \in [h, \tau - h]$ (to ease the notation), and so we need $0 < h \leq \tau/2$ for this to make sense. For the theoretical results below, the bandwidth h will depend on the sample size n , and tend to zero as $n \rightarrow \infty$, as we are used to from kernel density estimation.

(a) Let $\tilde{A}(t) = \int_0^t J(s)\alpha(s) ds$, and introduce $\tilde{\alpha}_n(t) = h^{-1} \int_t^{t+h} K((t-s)/h) d\tilde{A}(s)$. Show that $\hat{\alpha}_n(t)$ is unbiased for $E \tilde{\alpha}_n(t)$, and that $E \tilde{\alpha}_n(t) = h^{-1} \int_0^t K((t-s)/h) E \{Y(s)\} \alpha(s) ds$. Show also that if α is continuous at t and $\sup_s |Y(s) - 1| \rightarrow_{\text{pr}} 0$ uniformly, then $E \tilde{\alpha}_n(t) \rightarrow \alpha(t)$ as $n \rightarrow \infty$ and $h \rightarrow 0$ with it, and also $\tilde{\alpha}_n(t) \rightarrow_{\text{pr}} \alpha(t)$.

(b) Use martingale theory to show that

$$E (\hat{\alpha}_n(t) - \tilde{\alpha}_n(t))^2 = \frac{1}{h} \int_{-1}^1 K(u)^2 E \frac{J(uh+t)}{Y(uh+t)} \alpha(uh+t) du$$

(c) Assume that $nE J(s)/Y(s) \rightarrow 1/y(s)$ where $y(s) > 0$ on $[0, \tau]$ is the function from Ex. 11.18; and that both α and y are continuous at t . Show that $E (\hat{\alpha}_n(t) - \tilde{\alpha}_n(t))^2 = (nh)^{-1} \alpha(t)/y(t) \int K(u)^2 du + o((nh)^{-1})$ as $n \rightarrow \infty$ and $h \rightarrow 0$. Deduce that provided $h = h_n$ is so that $nh_n \rightarrow \infty$ as $n \rightarrow \infty$, the estimator $\hat{\alpha}_n(t)$ is consistent for $\alpha(t)$.

(d) Suppose that $nJ(s)/Y(s) \rightarrow 1/y(s)$ uniformly in the proximity of $t \in [h, \tau - h]$, and suppose that both α and y are continuous at t . Show that then $(nh)^{1/2}(\hat{\alpha}_n(t) - \tilde{\alpha}_n(t))$ converges in distribution to a mean zero normal with variance $\{\alpha(t)/y(t)\} \int_{-1}^1 K(u)^2 du$ as $n \rightarrow \infty$, $h \rightarrow 0$ and $nh \rightarrow \infty$.

Ex. 11.19 *Confidence bands for the cumulative hazard and survival functions.* Consider again survival data of the familiar form $(T_1, \delta_1), \dots, (T_n, \delta_n)$, leading to the Nelson–Aalen estimator $\hat{A}$ and Kaplan–Meier estimator $\hat{S}$. These are nonparametric estimators of the cumulative hazard rate $A(t) = \int_0^t \alpha(s) ds$ and survival curve $S(t)$, respectively. Here we look into how to construct *confidence bands* for these two crucial quantities.

(a) In Ex. 11.16 We have seen there that

$$\hat{A}(t) \approx_d N(A(t), \hat{\sigma}(t))$$

in the precise sense that $Z_n(t) = (\hat{A}(t) - A(t))/\hat{\sigma}(t) \rightarrow_d Z(t) \sim N(0, 1)$. Show that this implies that the band

$$B_n(t) = \hat{A}(t) \pm 1.96 \hat{\sigma}(t) = [\hat{A}(t) - 1.96 \hat{\sigma}(t), \hat{A}(t) + 1.96 \hat{\sigma}(t)]$$

is an approximate 95% pointwise confidence band, in the sense of

$$\Pr\{A(t) \in B_n(t)\} \rightarrow 0.95 \quad \text{for each } t.$$

(b) Explain how this can be used to test the null hypothesis $H_0(t)$ that $A(t)$ is equal to some given $A_0(t)$.

(c) Sometimes we need something more, however, namely a somewhat bigger band, say $B_n^*(t)$, which covers the full $A(t)$ across some time interval $[a, b]$, as opposed to merely at a given value t . Such bands may be constructed in different ways. Let's work with

$$B_n^*(t) = \hat{A}(t) \pm c \hat{\sigma}(t) = [\hat{A}(t) - c \hat{\sigma}(t), \hat{A}(t) + c \hat{\sigma}(t)],$$

and attempt to scale c such that it succeeds in being a 95% *simultaneous* confidence band over the time window $[a, b]$. What is needed is then

$$\Pr\{A(t) \in B_n^*(t) \text{ for all } t \in [a, b]\} \rightarrow 0.95.$$

Show that this corresponds to

$$\Pr(M_m \leq c) \rightarrow 0.95,$$

where

$$M_n = \max_{t \in [a, b]} \left| \frac{\hat{A}(t) - A(t)}{\hat{\sigma}(t)} \right|.$$

(d) From earlier results, of Ex. 11.16 and elsewhere, show that

$$Z_n(t) = \frac{\hat{A}(t) - A(t)}{\hat{\sigma}(t)} = \frac{\hat{A}(t) - A(t)}{\sigma_n(t)} \frac{\sigma_n(t)}{\hat{\sigma}(t)} \rightarrow_d Z(t) = \frac{W(v(t))}{\sqrt{v(t)}},$$

where $W(v(t))$ is a Gaussian martingale with independent increments and variance

$$v(t) = \int_0^t \frac{\alpha(s) \, ds}{y(s)}.$$

Argue that this implies

$$M_n = \max_{t \in [a, b]} |Z_n(t)| \rightarrow_d M = \max_{t \in [a, b]} |Z(t)|.$$

The threshold level c can then, in principle, be read off from the limit distribution, i.e. that of M .

(e) Here we're helped by first showing that

$$M = \max_{t \in [a, b]} \left| \frac{W(v(t))}{\sqrt{v(t)}} \right| = \max_{s \in [v(a), v(b)]} \left| \frac{W(s)}{\sqrt{s}} \right|.$$

Show that this lead to the following recipe: first, obtain estimates $\hat{v}(a)$ and $\hat{v}(b)$, via

$$\hat{v}(t) = n \int_0^t \frac{dN(s)}{Y(s)^2} = n\hat{\sigma}(t)^2.$$

Then simulate say 10^5 paths of Brownian motion $W(s)$ over the interval

$$[\hat{v}(a), \hat{v}(b)] = [n\hat{\sigma}(a)^2, n\hat{\sigma}(b)^2],$$

and record for each run the value of M , the maximal value of $W(s)/\sqrt{s}$ over that interval. You've now simulated the M distribution, and can read off c as the 0.95 quantile.

(f) For a Brownian motion $W = \{W(s) : s \geq 0\}$, consider the rescaled and time-transformed process W^* , with $W^*(s) = W(cs)/\sqrt{c}$, where c is a positive constant. Take a cup of tea (as the preeminent botanist Robert Brown did in 1827, indirectly pointing to a problem which was not mathematically solved until Einstein wrote a paper on this in 1905). Show that W^* is another Brownian motion. Show also that the variable

$$M = M_{a,b} = \max_{a \leq s \leq b} |W(s)/\sqrt{s}|$$

above has the property that $M_{a,b}$ has the same distribution as $M_{ca,cb}$, for any positive scaling factor c . So $M_{1.13,2.23}$ has the same distribution as $M_{113,223}$, etc.

(g) Explain how you can use such a plot to test whether $A(t)$ is equal to a given $A_0(t)$ over a given time interval.

(h) Explain finally how you can translate pointwise and simultaneous confidence bands for $A(t)$ to such bands for the survival curve $S(t) = \exp\{-A(t)\}$.

Ex. 11.20 *The likelihood theory works also for censored data.* Ch. 5 deals, among other things, with ML estimation in the classic terrain of independent and fully observed data points. It was not so easy to lift these machines to the world of survival analysis models, however, since the censoring business makes the models, the estimators, and their behaviour, more complicated. For early contributions to the theory, see the Notes & Pointers.

Consider survival data of the usual form (T_i, δ_i) for individuals $i = 1, \dots, n$, and assume that the hazard rate is of a suitable parametric form $\alpha(s, \theta)$. Examples include the exponential, the Weibull, the gamma, the Gompertz, the log-normal, etc. Below, let θ_0 denote the true parameter value.

(a) Check that you are able to deduce several (equivalent) expressions for the log-likelihood function, met also in several earlier exercises:

$$\begin{aligned}\ell_n(\theta) &= \sum_{\delta_i=1} \log f(T_i, \theta) + \sum_{\delta_i=0} \log S(T_i, \theta) = \sum_{i=1}^n \{\delta_i \log \alpha(T_i, \theta) - A(T_i, \theta)\} \\ &= \int_0^\tau \{\log \alpha(s, \theta) dN(s) - Y(s)\alpha(s, \theta) ds\}.\end{aligned}$$

Here τ is an upper bound for the time window under consideration.

(b) We attempt to follow the path of the ‘standard’ theory from Ch. 5, and start with the first derivative. Consider therefore the random function

$$U_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} = \int_0^\tau \psi(s, \theta) \{dN(s) - Y(s)\alpha(s, \theta) ds\},$$

with $\psi(s, \theta) = \partial \log \alpha(s, \theta) / \partial \theta$. At the true parameter value, show that $U_n(\theta_0) = \int_0^\tau \psi(s, \theta_0) dM(s)$, involving the martingale $M(t) = N(t) - \int_0^t Y(s)\alpha(s, \theta_0) ds$. Hence $U_n(\theta_0)$ is very conveniently a martingale, evaluated at the end-point τ .

(c) Show that the previsible quadratic variation of the score is

$$\begin{aligned}\langle (1/\sqrt{n})U_n(\theta_0), (1/\sqrt{n})U_n(\theta_0) \rangle &= (1/n) \int_0^\tau \psi(s, \theta_0) \psi(s, \theta_0)^t d\langle M, M \rangle(s) \\ &= (1/n) \int_0^\tau \psi(s, \theta_0) \psi(s, \theta_0)^t Y(s) \alpha(s, \theta_0) ds,\end{aligned}$$

and that, provided $Y(s)/n$ tends uniformly in probability to a limit function $y(s)$, this converges in probability to the matrix

$$J_0 = \int_0^\tau \psi(s, \theta_0) \psi(s, \theta_0)^t y(s) \alpha(s, \theta_0) ds.$$

Conclude from martingale limit theorems, that

$$(1/\sqrt{n})U_n(\theta_0) \rightarrow_d U \sim N_p(0, J_0)$$

So we’re in good shape, with the required result for the first derivative random function, though it took more work and more advanced limit theory than for the i.i.d. case.

(d) Next up is the second derivative $p \times p$ matrix function. Show that

$$\begin{aligned}I_n(\theta) &= \frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta^t} \\ &= \int_0^\tau [\psi^*(s, \theta) \{dN(s) - Y(s)\alpha(s, \theta) ds\} - \psi(s, \theta) \psi(s, \theta)^t Y(s) \alpha(s, \theta) ds],\end{aligned}$$

where $\psi^*(s, \theta) = \partial^2 \log \alpha(s, \theta) / \partial \theta \partial \theta^t$.

(e) Try to show, whether this means ‘prove’ or ‘make very plausible’, that

$$-(1/n)I_n(\theta_0) \rightarrow_{\text{pr}} J_0.$$

So, still, we’re in good shape, and have essentially been able to modify and extend the arguments of the i.i.d. case for the present more complicated world of survival models, partly thanks to the martingale theory. Conclude that for the ML estimator $\hat{\theta}$, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d J_0^{-1}U \sim N_p(0, J_0^{-1}).$$

This is the clear extension of the classical ML result for i.i.d. data to the world of survival models, involving now a new definition of the J_0 matrix.

(f) For the special case of no censoring, and with time window the full half-line, corresponding to $\tau = \infty$, show indeed that J_0 becomes equal to the Fisher information matrix for i.i.d. data.

(g) We need a good estimator of J_0 . There are several options, actually. The perhaps simplest to work with, since it comes out of computer optimisation programmes, as the Hessian matrix, is

$$\hat{J}_0 = -\frac{1}{n} \frac{\partial^2(\hat{\theta})}{\partial \theta \partial \theta^t} = \int_0^\tau \hat{y}(s) \psi(s, \hat{\theta}) \psi(s, \hat{\theta})^t \alpha(s, \hat{\theta}) ds - (1/n) \int_0^\tau \psi^*(s, \hat{\theta}) d\hat{M}(s),$$

in which $\hat{y}(s) = Y(s)/n$ and $\hat{M}(t) = N(t) - \int_0^t Y(s) \alpha(s, \hat{\theta}) ds$ the estimated martingale, at the model. Try to prove that $\hat{J}_0 \rightarrow_{\text{pr}} J_0$, i.e. that it is consistent. Part of this involves showing that the second term here goes to zero in probability. Another estimator, which also works, based on the form of J_0 , is

$$\tilde{J}_0 = \int_0^\tau \hat{y}(s) \psi(s, \hat{\theta}) \psi(s, \hat{\theta})^t \alpha(s, \hat{\theta}) ds,$$

corresponding to ignoring the second term above.

(h) For a simple illustration, consider the exponential model with a constant hazard rate, $\alpha(s, \theta) = \theta$. Find an expression for the ML estimator $\hat{\theta}$, and check each step of the arguments above. Put up the limit distribution for $\sqrt{n}(\hat{\theta} - \theta_0)$. How much is lost in precision, if the censoring distribution is another exponential, with rate γ ?

Ex. 11.21 *Parametric hazard rate regression models: theory.* The ML machinery of Ex. 11.20 extends nicely to the case of regression models. The setting is that of independent survival data $(T_1, \delta_1, x_1), \dots, (T_n, \delta_n, x_n)$, with covariate vectors x_i of length p . We start from

$$\alpha_i(s) = \alpha(s, \theta) \exp(x_i^t \beta), \quad \text{for } i = 1, \dots, n.$$

The theory briefly sketched below will work, mutatis mutandis. for the case of $\alpha(s, \theta) r(x_i, \beta)$, i.e. not merely for everyone’s favourite choice $r(x_i, \beta) = \exp(x_i^t \beta)$.

(a) Show that the log-likelihood function may be expressed as

$$\ell_n(\theta, \beta) = \sum_{i=1}^n \int_0^\tau [\{\log \alpha(s, \theta) + x_i^t \beta\} dN_i(s) - Y_i(s) \alpha(s, \theta) \exp(x_i^t \beta) ds].$$

The ML estimator $(\hat{\theta}, \hat{\beta})$ are the maximisers of $\ell_n(\theta, \beta)$.

(b) With θ and β of dimensions q and p , we start from the first derivatives

$$U_n(\theta, \beta) = \begin{pmatrix} U_{n,1}(\theta, \beta) \\ U_{n,2}(\theta, \beta) \end{pmatrix} = \begin{pmatrix} \partial \ell_n(\theta, \beta) / \partial \theta \\ \partial \ell_n(\theta, \beta) / \partial \beta \end{pmatrix}$$

With $\psi(s, \theta) = \partial \alpha(s, \theta) / \partial \theta$, show that

$$U_{n,1}(\theta, \beta) = \sum_{i=1}^n \int_0^\tau \psi(s, \theta) \{dN_i(s) - Y_i(s) \alpha(s, \theta) \exp(x_i^t \beta) ds\}$$

and that

$$U_{n,2}(\theta, \beta) = \sum_{i=1}^n \int_0^\tau x_i \{dN_i(s) - Y_i(s) \alpha(s, \theta) \exp(x_i^t \beta) ds\}.$$

(c) At the true parameter value (θ_0, β_0) , show that

$$U_n(\theta_0, \beta_0) = \sum_{i=1}^n \int_0^\tau \begin{pmatrix} \psi(s, \theta_0) \\ x_i \end{pmatrix} dM_i(s),$$

with

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \alpha(s, \theta_0) \exp(x_i^t \beta_0) ds$$

the martingale associated with counting process N_i . Hence the score function process $U_n(\theta_0, \beta_0)$ is a martingale.

(d) Show that the variance process of U_n at the true parameter position can be expressed as

$$\langle U_n(\theta_0, \beta_0), U_n(\theta_0, \beta_0) \rangle = \sum_{i=1}^n \int_0^\tau \begin{pmatrix} \psi(s, \theta_0) \\ x_i \end{pmatrix} \begin{pmatrix} \psi(s, \theta_0) \\ x_i \end{pmatrix}^t Y_i(s) \alpha(s, \theta_0) \exp(x_i^t \beta_0) ds.$$

Argue that under reasonable conditions,

$$U_n(\theta_0, \beta_0) \rightarrow_d U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \sim N_{q+p}(0, J_0),$$

with an appropriate $(q+p) \times (q+p)$ information matrix J_0 .

(e) Then find expressions for the second derivative process, say

$$I_n(\theta_0, \beta_0) = \frac{\partial^2 \ell_n(\eta_0)}{\partial \eta \partial \eta^t},$$

writing for brevity η for (θ, β) with true value $\eta_0 = (\theta_0, \beta_0)$. Show that

$$-(1/n)I_n(\theta_0, \beta_0) \rightarrow_{\text{pr}} J_0,$$

the same matrix as met above.

(f) Use this to establish that

$$\begin{pmatrix} \sqrt{n}(\hat{\theta} - \theta_0) \\ \sqrt{n}(\hat{\beta} - \beta_0) \end{pmatrix} \rightarrow_d J_0^{-1} U \sim N_{q+p}(0, J_0^{-1}).$$

The practical translation of all of this, for direct use when ML fitting parametric regression models for survival data, is that

$$\begin{pmatrix} \hat{\theta} \\ \hat{\beta} \end{pmatrix} \approx_d N_{q+p} \left(\begin{pmatrix} \theta_0 \\ \beta_0 \end{pmatrix}, \hat{J}_{\text{obs}}^{-1} \right),$$

with the $(q+p) \times (q+p)$ observed information matrix

$$\hat{J}_{\text{obs}} = -\frac{\partial^2 \ell_n(\hat{\eta}_{\text{ml}})}{\partial \eta \partial \eta^t}.$$

For further elaborations, extensions, and applications, see the references in the Notes & Pointers section.

Ex. 11.22 *Cox's partial likelihood.* In Ex. 11.21 we studied parametric forms of Cox's regression model, with hazard rates of the generic form $\alpha(s, \theta) r(x_i, \beta)$, for datasets of the classic survival type (T_i, δ_i, x_i) . The classical case is that of $r(x_i, \beta) = \exp(x_i^t \beta)$, but theory goes nicely through also for the more general case. We saw that the basic ML machinery works well, after all, with approximate joint normality for $(\hat{\theta}, \hat{\beta})$, an explicit estimator for its variance matrix, etc.. In the present exercise we rather consider the classic semiparametric setup of Cox's proportional hazards, where hazard rates are modelled as

$$\alpha_i(s) = \alpha_0(s) \exp(x_i^t \beta) \quad \text{for } i = 1, \dots, n,$$

but without modelling the baseline hazard rate $\alpha_0(\cdot)$ at all. Seen from the perspective of Ex. 11.21, there's simply no θ or attempt to model the $\alpha_0(\cdot)$. As earlier we use (N_i, Y_i, M_i) , for the counting process, the at-risk-process, and the martingale

$$dM_i(s) = dN_i(s) - Y_i(s) \alpha_0(s) \exp(x_i^t \beta_0) ds,$$

associated with individual i ; also, we posit that the model holds, with the true parameter denoted β_0 .

(a) For a generic observations (T, δ, x) , show that the density is

$$\begin{aligned} & [\alpha_0(t) \exp(x^t \beta_0) \exp\{-\exp(x^t \beta_0) \int_0^t \alpha_0(s) ds\} \{1 - G(c|x)\}]^\delta \\ & \times [\exp\{-\exp(x^t \beta_0) \int_0^t \alpha_0(s) ds\} g(c|x)]^{1-\delta}, \end{aligned}$$

where $G(c|x)$ is the conditional c.d.f. of the censoring times, and $g(c|x)$ its density. Explain that, under the assumption of non-informative censoring, this density is, viewed as a function of β_0 and α_0 , is proportional to the likelihood function

$$(\beta, \alpha_0) \mapsto \{\alpha_0(t) \exp(x^t \beta)\}^\delta \exp\{-\exp(x^t \beta) \int_0^t \alpha_0(s) ds\}.$$

(b) Suppose we have n independent realisations $(T_1, \delta_1, x_1), \dots, (T_n, \delta_n, x_n)$ of the vector specified above, and let $L_n(\beta, \alpha_0) = \prod_{i=1}^n \{\alpha_0(t_i) \exp(x_i^t \beta)\}^{\delta_i} \exp\{-\exp(x_i^t \beta) \int_0^{t_i} \alpha_0(s) ds\}$ be the likelihood function. Explain why $\sup_{\beta, \alpha_0} L_n(\beta, \alpha_0)$ is infinite, and consequently no ML estimator exists for this likelihood.

(c) Cox proposed estimating β parameters by maximising the function

$$L_n(\beta) = \prod_{i=1}^n \left(\frac{\exp(x_i^t \beta)}{\sum_{j:t_j \geq t_i} \exp(x_j^t \beta)} \right)^{\delta_i} = \prod_{t \geq 0} \prod_{i=1}^n \left(\frac{\exp(x_i^t \beta)}{\sum_{j=1}^n Y_j(t) \exp(x_j^t \beta)} \right)^{dN_i(t)}.$$

This is typically called Cox's partial likelihood. It can also be derived as a profile likelihood, where the baseline hazard is profiled out. We will now understand how. Note that the cumulative baseline hazard $A_0(t) = \int_0^t \alpha_0(s) ds$ is an absolutely continuous function, meaning that when we parametrise the Cox model in terms of α_0 , the implied parameter space is

$$\{(\beta, A_0) : \beta \in \mathbb{R}^p, A_0 \in \{\text{absolutely continuous functions on } [0, \infty)\}\}.$$

And with this parameter space, the likelihood $L_n(\beta, \alpha_0)$ has no maximiser. The trick is to instead parametrise the Cox model in terms of β and the cumulative A_0 , and allow A_0 to live in the much larger space of all nonnegative functions on $[0, \infty)$ such that $A_0(0) = 0$, thus admitting functions with jumps. Consider the likelihood function

$$L_n(\beta, A_0) = \prod_{i=1}^n \{\Delta A_0(t_i) \exp(x_i^t \beta)\}^{\delta_i} \exp\{-\exp(x_i^t \beta) A_0(t_i)\}, \quad (11.4)$$

where $\Delta A_0(t) = A_0(t) - A_0(t-)$ is the jump of A_0 at t . Treat β as known, and show that the maximiser, $A_0(t, \beta)$ say, of $L_n(\beta, A_0)$ is a step function whose jump at time t_i is $\Delta A_0(t_i, \beta) = \delta_i / \{\sum_{j:t_j \geq t_i} \exp(x_j^t \beta)\}$, that is

$$A_0(t, \beta) = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{j \geq i} \exp(x_j^t \beta)} = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{j=1}^n Y_j(t_i) \exp(x_j^t \beta)}.$$

Profile out A_0 , that is, insert the expression for $A_0(t, \beta)$ in $L_n(\beta, A_0)$, and show that

$$L_n(\beta, A_0(\cdot, \beta)) = L_n(\beta) \exp\left(-\sum_{i=1}^n \delta_i\right).$$

This entails that the maximiser $\hat{\beta}$ of $L_n(\beta)$ and $A_0(t, \hat{\beta})$, known as the Breslow-estimator, are ML estimators relative to the likelihood in (11.4).

(d) Compare partial to full likelihood inference in a case where that is possible. Simulate one thousand datasets $(t_i, \delta_i, x_i)_{1 \leq i \leq n}$, where $n = 47$, the x_i are i.i.d. standard normal, the t_i stem from the Cox model whose hazard is $\alpha_0 \exp(x\beta_0)$ for a constant baseline hazard. Estimate β_0 using the partial likelihood and the full likelihood. Compute the mean, the variance, and mean squared error of your one thousand estimates. Comment on and give an heuristic explanation of what you observe.

Ex. 11.23 Basic theory for the Cox regression model. Cox's proposal for estimating β is to maximise what he called the partial likelihood,

$$L_n(\beta) = \prod_{i: \delta_i=1} \frac{\exp(x_i^t \beta)}{S_n(t_i, \beta)} = \prod_{i=1}^n \left\{ \frac{\exp(x_i^t \beta)}{S_n(t_i, \beta)} \right\}^{\delta_i},$$

with

$$S_n(s, \beta) = \sum_{\text{risk set}} \exp(x_j^t \beta) = \sum_{j=1}^n Y_j(s) \exp(x_j^t \beta).$$

We call this $\hat{\beta}$ the Cox estimator. Its use hinges on basic large-sample theory for the Cox estimators $\hat{\beta}$, the maximiser of $\ell_n(\beta)$. The key points are that yes, indeed, the theory paralleling that of Ex. 11.21 still goes through, even though this is not ML theory, and, it requires new proofs.

(a) Show that the log-partial-likelihood can be expressed as

$$\ell_n(\beta) = \sum_{i=1}^n \{x_i^t \beta - \log S_n(t_i, \beta)\} \delta_i = \sum_{i=1}^n \int_0^\tau \{x_i^t \beta - \log S_n(s, \beta)\} dN_i(s).$$

(b) Show that its first derivative, with respect to the regression parameters β , can be expressed as

$$U_n(\beta) = \sum_{i=1}^n \{x_i - E_n(t_i, \beta)\} \delta_i = \sum_{i=1}^n \int_0^\tau \{x_i - E_n(s, \beta)\} dN_i(s),$$

with

$$E_n(s, \beta) = \frac{\partial S_n(s, \beta)}{\partial \beta} = \frac{S_n^{(1)}(s, \beta)}{S_n(s, \beta)} = \frac{\sum_{j=1}^n Y_j(s) x_j \exp(x_j^t \beta)}{\sum_{j=1}^n Y_j(s) \exp(x_j^t \beta)}.$$

Introduce also the second order derivative function

$$S_n^{(2)}(s, \beta) = \frac{\partial S_n(s, \beta)}{\partial \beta \partial \beta^t} = \sum_{j=1}^n Y_j(s) x_j x_j^t \exp(x_j^t \beta).$$

(c) At the true parameter β_0 , show that, crucially and half magically,

$$U_n(\beta_0) = \sum_{i=1}^n \int_0^\tau \{x_i - E_n(s, \beta_0)\} dM_i(s).$$

Hence it is a martingale (here evaluated at the right end-point τ of the time observation window).

(d) Show that the variance process associated with this martingale is

$$\begin{aligned} \langle U_n(\beta_0), U_n(\beta_0) \rangle &= \sum_{i=1}^n \int_0^\tau \{x_i - E_n(s, \beta_0)\} \{x_i - E_n(s, \beta_0)\}^t Y_i(s) \alpha_0(s) \exp(x_i^t \beta_0) ds \\ &= \int_0^\tau \{S_n^{(2)}(s, \beta_0) - E_n(s, \beta_0) E_n(s, \beta_0)^t S_n(s, \beta_0)\} \alpha_0(s) ds \\ &= \int_0^\tau V_n(s, \beta_0) S_n(s, \beta_0) \alpha_0(s) ds, \end{aligned}$$

in which

$$V_n(s, \beta) = \frac{S_n^{(2)}(s, \beta)}{S_n(s, \beta)} - E_n(s, \beta) E_n(s, \beta)^t.$$

(e) Consider a random covariate vector x^* , sampled from the risk set at time s , with probabilities proportional to the $\exp(x_j^t \beta)$. Show that the mean of such an x^* is $E_n(s, \beta)$ and that its variance matrix is $V_n(s, \beta)$. This provides some insight and intuition into the E_n and V_n , and certain bootstrapping schemes might be invented flowing from this, but it is not actively used in the mathematics below.

(f) Under reasonable regularity conditions, it follows from martingale limit theorems that

$$(1/\sqrt{n})U_n(\beta_0) \rightarrow_d U \sim N_p(0, J_0),$$

with J_0 the well-defined limit in probability of $(1/n)\langle U_n(\beta_0), U_n(\beta_0) \rangle$. Indeed such regularity conditions include or imply convergence in probability of the sample averages of $Y_j(s) \exp(x_j^t \beta)$, $Y_j(s)x_j \exp(x_j^t \beta)$, $Y_j(s)x_j x_j^t \exp(x_j^t \beta)$, to well-defined limits, say

$$n^{-1}S_n(s, \beta) \rightarrow_{\text{pr}} S(s, \beta), \quad n^{-1}S_n^{(1)}(s, \beta) \rightarrow_{\text{pr}} S^{(1)}(s, \beta), \quad n^{-1}S_n^{(2)}(s, \beta) \rightarrow_{\text{pr}} S^{(2)}(s, \beta),$$

uniformly in $s \in [0, \tau]$ and in a neighbourhood of β values around β_0 . Show that this implies

$$E_n(s, \beta) \rightarrow_{\text{pr}} E(s, \beta) = \frac{S^{(1)}(s, \beta)}{S(s, \beta)}$$

and also

$$V_n(s, \beta) \rightarrow_{\text{pr}} v(s, \beta) = \frac{S^{(2)}(s, \beta)}{S(s, \beta)} - E(s, \beta)E(s, \beta)^t.$$

Hence the crucial J_0 matrix above can be expressed as

$$J_0 = \int_0^\tau v(s, \beta_0) S(s, \beta_0) ds.$$

(g) So now things are fine for the score function part, that the first derivative process becomes normal. The second crucial ingredient, as per the chain of arguments used in Exercises 30, 31, 33, is to secure that the second derivative process, when normalised by sample size, converges to a matrix in probability. Show that

$$\begin{aligned} I_n(\beta) &= \frac{\partial^2 \ell_n(\beta)}{\partial \beta \partial \beta^t} \\ &= - \sum_{i=1}^n \int_0^\tau \frac{S_n^{(2)}(s, \beta) S_n(s, \beta) - S_n^{(1)}(s, \beta) S_n^{(1)}(s, \beta)^t}{S_n(s, \beta)^2} dN_i(s) \\ &= - \sum_{i=1}^n \int_0^\tau V_n(s, \beta) dN_i(s). \end{aligned}$$

Show, using $dN_i(s) = Y_i(s)\alpha_0(s) \exp(x_i^t \beta_0) ds + dM_i(s)$, that

$$-(1/n)I_n(\beta_0) \rightarrow_{\text{pr}} \int_0^\tau v(s, \beta_0) S(s, \beta_0) \alpha_0(s) ds,$$

which, lo & behold, is precisely the J_0 matrix above.

(h) Try to prove, in some more generality, that if $H(s)$ is a given function, then

$$\begin{aligned} (1/n) \sum_{i=1}^n \int_0^\tau H(s) dN_i(s) &= (1/n) \sum_{i=1}^n \int_0^\tau H(s) \{Y_i(s)\alpha_0(s) \exp(x_i^t \beta_0) ds + dM_i(s)\} \\ &\rightarrow_{\text{pr}} \int_0^\tau H(s) S(s, \beta_0) \alpha_0(s) ds. \end{aligned}$$

Also, attempt to generalise this to

$$(1/n) \sum_{i=1}^n \int_0^\tau H_n(s) dN_i(s) \rightarrow_{\text{pr}} \int_0^\tau H(s) S(s, \beta_0) \alpha_0(s) ds,$$

if $H_n(s)$ is such that it converges to $H(s)$, uniformly in probability.

(i) The essential part of the work is now done. Argue, pointing to earlier and similar efforts above, and also in Ch. 5, that

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightarrow_d J_0^{-1}U \sim N_p(0, J_0^{-1}).$$

(j) The crucial result above underlies most software packages etc. for survival analysis via Cox regression. The ‘practical translation’ is that

$$\hat{\beta} \approx_d N_p(\beta_0, \hat{J}_{\text{obs}}^{-1}),$$

where

$$\hat{J}_{\text{obs}} = -\frac{\partial^2 \ell_n(\hat{\beta})}{\partial \beta \partial \beta^t} = \int_0^\tau V_n(s, \hat{\beta}) \sum_{i=1}^n dN_i(s) = \sum_{t_i \leq \tau} V_n(t_i, \hat{\beta}) \delta_i.$$

Ex. 11.24 *Aalen’s linear hazard rate risk regression model: theory.* With survival data of the familiar form (T_i, δ_i, x_i) , [Aalen \(1980, 1989\)](#) introduced an additive risk regression model, rather different from Cox’s multiplicative one. The hazard rates are now taken as

$$\alpha_i(s) = \beta_0(s) + x_{i,1}\beta_1(s) + \cdots + x_{i,q}\beta_q(s) \quad \text{for } i = 1, \dots, n,$$

in terms of a baseline hazard function $\beta_0(\cdot)$ in addition to ‘regressor functions’ $\beta_j(\cdot)$ for $j = 1, \dots, q$. The setup is quite general, but it suits perhaps particularly well situations where the covariates indicate ‘risky values’, with zero the value for ‘normal and healthy’. In such a setup, $\beta_0(s)$ is the hazard rate for all ‘normal and healthy’ individuals, those having all covariates equal to zero, and any increase in any of these covariates contributes to a higher hazard rate.

For the mathematical development it is convenient to ‘move notation around’ a little bit to land at

$$\alpha_i(s) = x_{i,1}\beta_1(s) + \cdots + x_{i,p}\beta_p(s) = x_i^t \beta(s),$$

where the first of the columns in the $n \times p$ covariate matrix X typically (but not necessarily) consists of $(1, \dots, 1)$. With the usual counting process N_i , at-risk process Y_i , and martingale

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \alpha_i(s) ds = N_i(t) - \int_0^t Y_i(s) x_i^t \beta(s) ds,$$

we may write

$$dN_i(s) = Y_i(s) x_i^t \beta(s) ds + dM_i(s),$$

to capture the main structure. We are interested in estimators and large-sample approximations for the cumulative regressor functions

$$B(t) = (B_1(t), \dots, B_p(t))^t, \quad \text{with} \quad B_j(t) = \int_0^t \beta_j(s) ds.$$

(a) Show that

$$\sum_{i=1}^n x_i dN_i(s) = \sum_{i=1}^n Y_i(s) x_i x_i^t \beta(s) ds + \sum_{i=1}^n x_i dM_i(s),$$

which is convenient to rewrite as

$$(1/n) \sum_{i=1}^n x_i dN_i(s) = G_n(s) dB(s) + (1/n) \sum_{i=1}^n x_i dM_i(s),$$

featuring

$$G_n(s) = (1/n) \sum_{i=1}^n Y_i(s) x_i x_i^t.$$

This matrix has full rank $p \times p$ as long as there are at least p linearly independent x_i vectors in the risk set at time s .

(b) Show that this leads to

$$G_n(s)^{-1} (1/n) \sum_{i=1}^n x_i dN_i(s) = dB^*(s) + G_n(s)^{-1} (1/n) \sum_{i=1}^n x_i dM_i(s),$$

featuring $B^*(t)$, with cumulatives $B_j^*(t) = \int_0^t J(s) \beta_j(s) ds$, where $J(s) = I\{\text{full rank at time } s\}$. In practice, this means $J(s) = 1$, and $B_j^*(t) = B_j(t)$, for all t in the ‘good terrain’ where there are enough individuals at risk. This leads to the Aalen estimator

$$\hat{B}(t) = \int_0^t G_n(s)^{-1} (1/n) \sum_{i=1}^n x_i dN_i(s) = \sum_{i \leq t} G_n(t_i)^{-1} (1/n) x_i \delta_i,$$

with all the cumulative $B_j(t)$ estimated jointly.

(c) So we have

$$d\hat{B}(s) = J(s) dB(s) + J(s) G_n(s)^{-1} (1/n) \sum_{i=1}^n x_i dM_i(s),$$

with the second term equal to a martingale increment. Show that the variance of that increment is the mean of

$$\begin{aligned} & \left\langle J(s) G_n(s)^{-1} (1/n) \sum_{i=1}^n x_i dM_i(s), J(s) G_n(s)^{-1} (1/n) \sum_{i=1}^n x_i dM_i(s) \right\rangle \\ &= (1/n) J(s) G_n(s)^{-1} (1/n) \sum_{i=1}^n Y_i(s) x_i x_i^t x_i^t dB(s) G_n(s)^{-1} \\ &= (1/n) J(s) G_n(s)^{-1} dH_n(s) G_n(s)^{-1}. \end{aligned}$$

(d) Argue that this leads to the natural variance matrix estimator for the $\hat{B}(t)$:

$$(1/n) \int_0^t J(s) G_n(s)^{-1} d\hat{H}_n(s) G_n(s)^{-1} = (1/n) \sum_{t_i \leq t} J(t_j) G_n(t_i)^{-1} d\hat{H}_n(t_i) G_n(t_i)^{-1}.$$

Here

$$d\hat{H}_n(s) = (1/n) \sum_{i=1}^n Y_i(s) x_i x_i^t x_i^t d\hat{B}(s),$$

which also means

$$d\hat{H}_n(t_i) = (1/n) x_i x_i^t x_i^t d\hat{B}(t_i) \delta_i.$$

(e) For n large, argue that

$$\sqrt{n} \{ \hat{B}(t) - B(t) \} \approx_d W_n(t),$$

a p -dimensional Gaussian zero-mean martingale process, with variance matrix

$$\Sigma(t) = \int_0^t G(s)^{-1} dH(s) G(s)^{-1},$$

where matrix functions $G(s)$ and $H(s)$ are the limits in probability of $G_n(s)$ and $H_n(s)$.

(f) (xx a bit more. xx)

Notes and pointers

[xx for notes and pointers xx] Parts of the material presented in Ex. 11.1 was also included in Hjort (1990b), along with further results, extensions, and discussion. The framework and methods for the discrete time setup has separate interest, and it inspired the invention of the Beta process, as a fine limit of the time-discrete grid.

(xx mention that Kaplan-Meier used the ‘product limit estimator’ term to allude to the mesh to zero limit. this matches some of the above. point to Beta process of Ch14 in that regard. xx)

[xx some notes and pointers here xx] The result you have proven in Ex. D.3 has an analogue in the continuous time world, known as the Doob–Meyer theorem, see e.g. Jacod and Shiryaev (2013, Theorem I.3.15, p. 32) and the references therein.

for Lengart in continuous time see e.g. Andersen et al. (1993, p. 86) or Jacod and Shiryaev (2013, p. 35).

for Ex. 11.18, point to Ramlau-Hansen (1983).

for Ex. 11.11 we point to Hermansen and Hjort (2015).

confidence bands of Ex. 11.19(c) were proposed by Hjort (1985b), inside a long discussion contribution to the 1984 SJS Lecture (Andersen and Borgan, 1985), along with various other points.

commenting on theory in Ex. 11.23. This is what it took the profession a few years to sort out properly, from when Sir David Cox (Cox, 1972) introduced the model and the basic estimation and inference strategy [as of October 2018, when I write this, that paper has more than 48,000 citations on Google Scholar], all the way, via other good but still partial results, to Andersen and Gill (1982).

Model selection and model averaging

Statistical model selection has in principle been around from the inventions of the more classic models, but has now become much more vital and nearly omnipresent due to the ease with which statisticians can sift through thousands of candidate models. There are classes of relevant tools, from graphical checks to a rich cavalcade of goodness-of-fit tests, cf. Ch. 10, but this chapter focuses on direct selection criteria. Chief amongst these are the acronyms AIC, BIC, FIC, i.e. the Akaike, the Bayes, and the Focused Information Criterion. The latter is set up to work for a given focus parameter, with the FIC aiming for the candidate model that delivers the best inference for this focus; so the same data and the same list of candidate models might have different winners for different foci. Connections to cross-validation are also tended to. When taking weighted averages of the best results, or need to understand the properties of the finally produced estimators, we enter the terrain of model averaging. The theory is motivated and worked through first for i.i.d. setting, but is naturally extended to cover also regression frameworks.

Key words: AIC, BIC, choosing covariates, FIC, model averaging, model selection, narrow model, tolerance radius, wide model

Suppose a dataset y is studied, with competing parametric models, say $M_1, \dots, M_k$, with perhaps different parameter dimensions $p_1, \dots, p_k$. A simple, classical, powerful model selection scheme is then *the AIC*, the Akaike Information Criterion, working as follows. For each model, fit its parameters maximising the log-likelihood function, say $\ell_j(\theta_j)$, involving in particular the maximum likelihood estimate $\hat{\theta}_{j,\text{ml}}$. Then compute the AIC score

$$\text{AIC}_j = 2\ell_{j,\text{max}} - 2p_j = 2\ell_j(\hat{\theta}_j) - 2p_j, \quad (12.1)$$

and declare the model with the highest AIC score the winner.

In exercises below we illustrate model selection using this criterion and also go through the chain of arguments behind it, with variations. We also go into *the BIC*, the Bayesian Information Criterion, which from quite different start arguments and certain approximations land at

$$\text{BIC}_j = 2\ell_{j,\text{max}} - p_j \log n = 2\ell_j(\hat{\theta}_j) - p_j \log n, \quad (12.2)$$

with n the number of observations. The third model selection criterion given emphasis in this chapter is *the FIC*, the Focused Information Criterion. Whereas the AIC and the

BIC work in overall modus, comparing a model's complexity with how well it fits the data, the FIC starts from the notion of a given focus parameter, say μ , a quantity of interest with clear interpretation across models. Model j leads to the parameter estimate $\hat{\mu}_j = \mu_j(\hat{\theta}_j)$. These estimators are associated with certain risks, the mean squared errors the FIC

$$\text{mse}_j = E_{\text{true}}(\hat{\mu}_j - \mu_{\text{true}})^2 = v_j + \text{bsq}_j$$

say, for $j = 1, \dots, k$, with v_j the variance and bsq_j the squared bias of $\hat{\mu}_j$ as an estimator of μ_{true} . The FIC aims at estimating these, with

$$\text{FIC}_j = \widehat{\text{mse}}_j = \hat{v}_j + \widehat{\text{bsq}}_j, \quad (12.3)$$

and with different formulae for these depending on the complexities of the models and the focus parameter. Models with lowest FIC scores are then judged the best. The FIC might choose different optimal models for different focus parameters.

It should be noted that model selection methods are related to but also different in attitude from the many types of goodness-of-fit tests. The latter aim at giving yes-no answers to the question of whether the data adequately fit a given model, whereas AIC, BIC, FIC leave that question partly to the side; we aim at having the best model, from a list of candidate models, without necessarily claiming that the winning model is correct.

Model selection criteria, like AIC, BIC, FIC, can be used to decide on which covariates to exclude, and which to include, and in which ways, in regression models. We also include material on cross validation and discuss its relation to the AIC.

These methods, along with deliberations to find the appropriate tools depending on the context, are amply illustrated and discussed in Stories [#5](#), [#8](#), [#16](#), [#17](#), [#22](#), [#24](#), [#26](#), [#27](#), [#28](#), [#37](#), [#38](#), [#51](#), [#45](#), [#50](#), [#53](#), [#55](#), [#58](#), [#61](#), [#76](#), [#87](#), [#98](#).

AIC and BIC in action

Ex. 12.1 *Fitting models and computing AIC and BIC.* To get used to the operations and calculations involved in fitting different parametric candidate models and then ranking these with AIC or BIC, simulate say $n = 200$ datapoints from the Beta distribution with parameters $(a, b) = (1.10, 0.90)$.

(a) Fit the following four candidate models to your data: (i) the Beta($a, 1$); (ii) the Beta($1, b$); (iii) the symmetric Beta(a, a); (iv) the two-parameter Beta(a, b). For each, programme the log-likelihood function and maximise it numerically. In that process, table also the implied estimates of the focus parameter $\mu = EY$, along with standard errors (estimated standard deviation).

(b) Then as a consequence of your model fitting operations, compute AIC and BIC scores, and find the winning models. How does this compare with results from testing hypotheses about the two Beta parameters?

(c) Play with your code a bit, to see how your conclusions about the best models change with increasing sample size.

(d) Include one more candidate model, the three-parameter model with c.d.f. $G(y, a, b, c) = \text{Be}(y, a, b)^c$, where $\text{Be}(y, a, b)$ is the c.d.f. for the Beta(a, b). Do the data tell you that it is worth it, to include the extra c parameter?

Ex. 12.2 *Which probabilities are equal?* Suppose you have two coins, A and B, with probabilities p_a and p_b for 'heads'. You flip each $n = 40$ times, and get $x_a = 17$ and $x_b = 24$ heads. Are the probabilities equal, or are they not?

(a) When $X \sim \text{binom}(n, p)$, we have seen in Ex. 5.23 that the ML estimator is $\hat{p} = X/n$. Show that the log-likelihood maximum is

$$\ell_{\max} = nH(\hat{p}) + \log \binom{n}{x}, \quad \text{with } H(p) = p \log p + (1-p) \log(1-p).$$

(b) Consider model 1, that $p_a = p_b$, and model 2, where $p_a \neq p_b$. Compute AIC scores for the two models, and decide which is judged the best. If you first get $x_a = 17$ for the first coin, how far away must x_b be, for the second coin, in order for AIC to determine that the two-probabilities model is better than the common-probability model?

(c) Suppose now that you actually have three coins, with probabilities p_a, p_b, p_c for ‘heads’, with results 17, 24, 15 in $n = 40$ flips. Consider as many as five models: model 1 takes $p_a = p_b = p_c$; model 2 takes $p_a = p_b$ but different from p_c ; model 3 takes $p_a = p_c$ but different from p_b ; model 4 takes $p_b = p_c$ but different from p_a ; and model 5 says the three are different. Compute AIC scores and list the five models in ranked order. For an application of this machinery, along such lines, do the corresponding AIC analysis for the five models for the Star Trek casualties data of Story #52. Should your uniform be blue, to better your survival chances?

Ex. 12.3 *Are some multinomial probabilities equal?* Let (Y_1, Y_2, Y_3, Y_4) be multinomial (n, p_1, p_2, p_3, p_4) , with the p_j summing to one and the Y_j summing to n . Suppose there is a possibility that $p_1 = p_2$ and that $p_3 = p_4$, in which case the two first and the two last boxes could be lumped together. That’s a question that AIC can answer.

(a) The wide model has 3 unknown parameters. Check with Ex. 5.23 to see that its log-likelihood function can be written $y_1 \log p_1 + y_2 \log p_2 + y_3 \log p_3 + y_4 \log p_4$, that it is maximised at $\hat{p}_j = y_j/n$, leading also to

$$\ell_{\max, \text{wide}} = n(\hat{p}_1 \log \hat{p}_1 + \cdots + \hat{p}_4 \log \hat{p}_4).$$

(b) The narrow model, which lumps together the two first and then the two last probabilities, takes $p_1 = p_2 = \frac{1}{2}\theta_1$ and $p_3 = p_4 = \frac{1}{2}\theta_2$, say, with $\theta_1 + \theta_2 = 1$. Show that the log-likelihood for this submodel is

$$\ell_{\text{narr}}(\theta_1) = (y_1 + y_2) \log(\tfrac{1}{2}\theta_1) + (y_3 + y_4) \log(\tfrac{1}{2}\theta_2),$$

that is is maximised at $\hat{\theta}_1 = \hat{p}_1 + \hat{p}_2$ and $\hat{\theta}_2 = \hat{p}_3 + \hat{p}_4$, with

$$\ell_{\max, \text{narr}} = n[(\hat{p}_1 + \hat{p}_2) \log(\hat{p}_1 + \hat{p}_2) + (\hat{p}_3 + \hat{p}_4) \log(\hat{p}_3 + \hat{p}_4) - \log 2].$$

(c) Show that $D = \ell_{\max, \text{wide}} - \ell_{\max, \text{narr}}$ is always nonnegative, and that AIC says the wide model is best if $D > 2$.

(d) Generate 1000 outcomes of the Beta(5, 3) distribution, and display a standard equal-bin-sized histogram for the data with 20 cells. Explain that this corresponds to estimating a multinomial with probabilities $p_1, \dots, p_{20}$; see Ex. 1.6. Then display a different histogram for the same data, but with 10 cells. Use the AIC to determine which histogram is best.

Ex. 12.4 *AIC and BIC for linear regressions.* For regression data (y_i, x_i) , with available covariate vectors $x_i = (x_{i,1}, \dots, x_{i,p})^t$ of length p , there is often a question of which covariates to include, as including too many leads to higher variability. Inside given model frameworks, AIC and BIC offer general answers. Here we study the general normal linear regression model, with independent $y_i \sim N(x_i^t \beta, \sigma^2)$ for $i = 1, \dots, n$; see Ex. 3.33.

- (a) For this model we know that the ML estimator $\hat{\beta}$ is the same as the minimum sum of squares estimator, leading to the minimum $Q_0 = Q(\hat{\beta})$, where $Q(\beta) = \sum_{i=1}^n (y_i - x_i^t \beta)^2$. Show that the log-likelihood function is $\ell_n(\beta, \sigma) = -n \log \sigma - \frac{1}{2} Q(\beta) / \sigma^2 - \frac{1}{2} n \log(2\pi)$, and that $\ell_{n, \max} = -n \log \hat{\sigma} - \frac{1}{2} n - \frac{1}{2} n \log(2\pi)$, in terms of $\hat{\sigma}^2 = Q_0/n$. Deduce that $\text{AIC} = -2n \log \hat{\sigma} - 2(p+1) - n - n \log(2\pi)$ and that $\text{BIC} = -2n \log \hat{\sigma} - (p+1) \log n - n - n \log(2\pi)$.
- (b) Suppose now that a list of submodels is considered, say $M_1, \dots, M_k$, with model M_j having a subset C_j of the covariates $x_{i,1}, \dots, x_{i,p}$ on board. Show that doing well, according to the AIC or the BIC, is equivalent to having small value of the scores

$$a_j = n \log \hat{\sigma}_j + p_j \quad \text{and} \quad b_j = n \log \hat{\sigma}_j + \frac{1}{2} p_j \log n,$$

respectively, with $\hat{\sigma}_j^2 = Q_{0,j}/n$ involving the minimum sum of squares for subset C_j , with its p_j covariates. Explain here that $\hat{\sigma}_j$ for a submodel estimates a different parameter than the σ in the full model. Show also that a given submodel, with $p_0 < p$ covariates, is judged by the AIC to be better than the full model provided $n \log(\hat{\sigma}_0/\hat{\sigma}_{\text{full}}) < p - p_0$. Turn the narrative around: explain that if one starts with a perhaps acceptable model M_0 , with p_0 covariates, then it's worth including additional covariates only if $\hat{\sigma}_{\text{full}}$ becomes small enough to achieve $n \log(\hat{\sigma}_0/\hat{\sigma}_{\text{full}}) > p - p_0$.

(c) Use linear regression methods and results from Ch. 3 to give the variance of the estimator $\hat{\gamma} = x_0^t \hat{\beta}$ for $\gamma = E(Y|x_0) = x_0^t \beta$. Give also the confidence interval for γ , using Ex. 3.36. This assumes the model is correctly specified, however, and ignores potential modelling bias. The FIC methods developed below address this problem.

Ex. 12.5 *AIC for polynomial order.* Assume regression data (x_i, y_i) are recorded, with the covariate x_i taking values in $[0, 1]$. Consider the polynomial regression models

$$y_i = m_j(x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \dots + \beta_j x_i^j + \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

for polynomials of order $j = 1, \dots, p$, and for each model taking the ε_i as i.i.d. $N(0, \sigma^2)$. Thus candidate model j has $j + 2$ parameters.

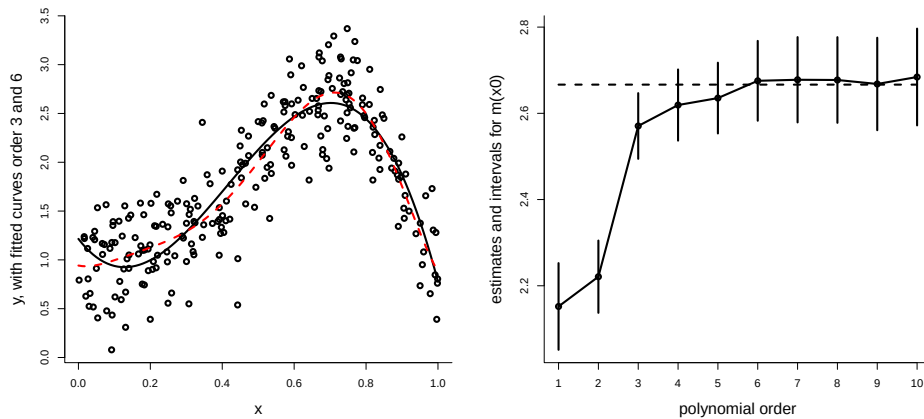


Figure 12.1: Left panel: a total of $n = 250$ data points generated as described, along with fitted polynomial regression curves of order three and six. Right panel: the different estimates of $m(x_0)$, at $x_0 = 0.75$, along with 95 percent confidence intervals, not taking modelling bias into account; the true value is 2.666.

(a) Explain that the σ parameter depends on the order j of the polynomial here (so a more pedantic notation would use σ_j here, for $j = 1, \dots, p$). Use results from Ex. 12.4 to show that the ML estimator inside model j is $\hat{\sigma}_j^2 = Q_{j,\min}/n$, with the minimum $\sum_{i=1}^n \{y_i - \hat{m}_j(x_i)\}^2$, involving the estimated $m_j(x) = \beta_0 + \beta_1 x + \dots + \beta_j x^j$. Find a formula for the AIC score, and show that the AIC chooses as winner the model with the smallest $n \log \hat{\sigma}_j + j$.

(b) Try this out, in a simulation setting, generating for a given n the x_i from the uniform distribution on the unit interval and then $y_i = m_{\text{true}}(x_j) + \varepsilon_{i,0}$, with $m(x) = \exp(\sin(\pi x^2))$, and $\varepsilon_{i,0}$ taken i.i.d. $N(0, \sigma_0^2)$, with say $\sigma_0 = 0.333$. Generate $n = 250$ data points following this recipe, fit the first ten polynomial regression models, compute AIC scores, and find the winning model. Construct a version of Fig. 12.1, left panel.

(c) Play with your simulation code, checking perhaps $n = 50, 100, 150, 200, \dots, 1000$, and demonstrate that with more data come increasingly sophisticated winning models. Also, for a fixed $n = 250$, repeat the experiment several times, to see how often the different models become winners.

(d) Assume one wishes to estimate $\gamma = m(x_0)$, for some given x_0 . For this illustration, use the generated data to compute the ten different estimates of γ , along with 95 percent confidence intervals, with these computed assuming that the model used is correct. Construct a version of Figure 12.1, right panel. The first few of these err since the models are not adequate. We address modelling bias when developing the FIC theory below. Explain that the coverage of the AIC winning interval is not 95 percent, due to (i) modelling bias and (ii) the extra randomness coming from model selection.

Ex. 12.6 AIC and BIC for logistic regression order. For 0-1 data Y , associated with a covariate $x \in [0, 1]$, assume the associated true probability curve is $p(x) = 0.5 + 0.5 \sin(2\pi x^2)$. We work here with the logistic models $p_k(x) = H(\beta_0 + \beta_1 x + \dots + \beta_k x^k)$, with $H(u) = \exp(u)/\{1 + \exp(u)\}$ the logistic transform and k the order of the polynomial. How well will the estimated $\hat{p}_k(x)$ perform, when the order is selected using AIC or BIC?

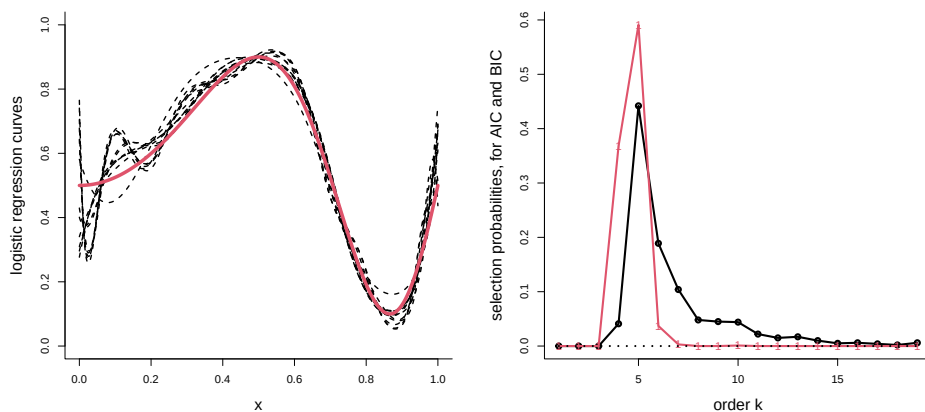


Figure 12.2: Left panel: for a simulated dataset of size $n = 500$, fitted logistic curves $\hat{p}_k(x)$ are given for $k = 4, 5, \dots, 15$ (dashed lines), along with the underlying true $p(x)$ (full curve). Right panel: the AIC and BIC order selection probabilities, for $n = 1000$, for 1000 simulated datasets, with the BIC on average preferring lower order.

(a) The true $p(x)$ is outside the logistic regression models. Explain however via the Weierstrass theorem, see Ex. 2.17, that for given $\varepsilon > 0$, there is a perhaps high k such that $\max |p_k(x) - p(x)| \leq \varepsilon$, with coefficients properly selected.

(b) With $n = 500$ and $x_i = i/n$ for $i = 1, \dots, n$, simulate a dataset, then fit this to the logistic models $p_k(x)$, for order say $k = 1, 2, \dots, k_{\text{up}} = 15$. Make a version of Figure 12.2, left panel, indicating that the fitted curves manage to mimic the true $p(x)$ function with increasing order.

(c) Fitting a given dataset to the order k model, via ML, show that the maximised log-likelihood function is $\ell_{n,\max} = \sum_{i=1}^n \{y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)\}$, with $\hat{p}_i$ the estimated $p(x_i)$. Having fitted a given dataset to these logistic models, explain that the AIC and BIC scores are $\text{AIC}_k = 2\ell_{n,\max} - 2(k + 1)$ and $\text{BIC}_k = 2\ell_{n,\max} - (k + 1) \log n$. With $n = 1000$, simulate 1000 datasets, finding and keeping track of $\hat{k}_{\text{aic}}$ and $\hat{k}_{\text{bic}}$, those giving largest AIC and BIC scores. Construct a version of Figure 12.2, right panel. One learns that the BIC typically prefers smaller models than does the AIC.

(d) Let $\hat{p}_A(x)$ and $\hat{p}_B(x)$ be the finally selected estimators of $p(x)$, with orders $\hat{k}_a$ and $\hat{k}_b$. For the 1000 simulated datasets, compute and keep track of the values of $D_A = \max_x |\hat{p}_A(x) - p(x)|$ and $D_B = \max_x |\hat{p}_B(x) - p(x)|$. Plot these (D_A, D_B) and comment.

Ex. 12.7 *A start model and one-parameter extensions.* Assume a start model $f_0(y, \theta)$ is fitted for a given dataset, leading to maximal log-likelihood $\ell_{0,\max}$.

(a) Suppose then that a certain one-parameter extension of the start model is contemplated, say $f_0(y, \theta, \gamma)$, leading to maximal log-likelihood $\ell_{1,\max}$. Show that AIC prefers this wider model if and only if the log-likelihood is increased with at least 1. In other words, one needs $\Delta = 2(\ell_{1,\max} - \ell_{0,\max}) > 2$. Note that this is the log-likelihood-ratio statistic worked with in (xx give pointers to likelihood chapter xx), though not with some customary 0.05 level. Show indeed that the implied significance level is 0.157.

(b) Show then that BIC needs harder evidence of the wider model being better, namely that $\Delta > \log n$.

(c) For an illustration, consider regression data (x_i, y_i) , where the narrow model has $y_i = \beta_0 + \varepsilon_i$ and the wide model takes $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon'_i$, in both cases taking the error terms as i.i.d. $N(0, \sigma^2)$. Derive a clear expression for $\ell_{1,\max} - \ell_{0,\max}$, in the notation above, and state what it takes for AIC to select the wider model.

(d) Invent a similar situation and check out how methods perform.

Ex. 12.8 *Is modelling variance heterogeneity worth the trouble?* Suppose we have regression data (x_i, y_i) , with x_i of length p , for which the standard linear regression model takes $N(x_i^t \beta, \sigma^2)$ with a common variability level σ ; see Ex. 3.33. There are models and methods for factoring in non-constant variances, as we have seen in Ex. 5.49, involving by necessity some extra sophistication and efforts. So when is it worthwhile?

(a) For some one-dimensional covariate w_i , normalised to have mean zero, consider the heteroscedastic extension taking $\sigma_i = \sigma \exp(w_i \gamma)$. Let $\hat{\sigma}(\gamma)^2 = Q_0(\gamma)/n$, with the weighted sum of squares $Q_0(\gamma) = \sum_{i=1}^n \{y_i - x_i^t \hat{\beta}(\gamma)\}^2 / \exp(2\gamma w_i)$; the standard case is that of $\gamma = 0$. Using methods and results from the exercise pointed to, show that AIC prefers such a one-parameter extension of the basic $p + 1$ -dimensional linear regression model provided $n \log\{\hat{\sigma}(0)/\hat{\sigma}(\hat{\gamma})\} > 1$. How much smaller does $\hat{\sigma}(\hat{\gamma})$ need to be, compared to $\hat{\sigma}(0)$, in order for the BIC to judge this worthwhile?

(b) More generally, consider the heteroscedastic setup with $\sigma_i = \sigma \exp(w_i^t \gamma)$, for covari-

ate vectors w_i of length q . It is again fruitful to normalise these to have mean zero, for each component, also to simplify the mathematics; see Ex. 5.49. Show that the AIC score, for such a model, becomes $\text{AIC} = -2n \log \hat{\sigma}(\hat{\gamma}) - 2(p+q)$, ignoring constants. When comparing many such models, perhaps with different subsets of the x_i for the mean structure and the w_i for the variance structure, good AIC models are those with small values of $A = n \log \hat{\sigma}(\hat{\gamma}) + p + q$. In some detail, consider candidate models, indexed as $M_{j,k}$, using subsets C_j of $x_{i,1}, \dots, x_{i,p}$ for the mean and subsets D_k of $w_{i,1}, \dots, w_{i,q}$ for the variance. Fitting such a model involves finding $\hat{\sigma}_{j,k} = \hat{\sigma}_j(\hat{\gamma}_k)$, with the appropriate subsets. Explain that the best AIC models are those with lowest $A_{j,k} = n \log \hat{\sigma}_{j,k} + p_j + q_k$, where p_j and q_k are the sizes of these subsets C_j and D_k . How many such submodels can there be?

Ex. 12.9 Comparing AIC and BIC, in prototype setup I. To learn about how AIC and BIC differ in their selection strategies, consider the following simple setup: i.i.d. $Y_1, \dots, Y_n$ are either from $N(0, 1)$, with no unknown parameters (model 0), or from $N(\theta, 1)$, with θ unknown (model 1).

(a) With $\ell_{0,\max}$ and $\ell_{1,\max}$ for the maximised log-likelihood functions, for models 0 and 1, show that $\ell_1 - \ell_0 = \frac{1}{2}n\bar{Y}^2$, with as usual $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$. Show also that

$$\text{AIC}_1 - \text{AIC}_0 = n\bar{Y}^2 - 2, \quad \text{BIC}_1 - \text{BIC}_0 = n\bar{Y}^2 - \log n.$$

Deduce that AIC selects model 1 if $\sqrt{n}|\bar{Y}| \geq \sqrt{2}$, but that the BIC demands more, choosing model 1 if $\sqrt{n}|\bar{Y}| \geq (\log n)^{1/2}$.

(b) Consider the probabilities that AIC and BIC selects model 1,

$$p_{n,A}(\theta) = \Pr_\theta(\text{AIC selects model 1}), \quad p_{n,B}(\theta) = \Pr_\theta(\text{BIC selects model 1})$$

as a function of the parameter. Compute these, and display them in a graph, for $n = 10, 50, 100$.

(c) Study in particular $p_{n,A}(\theta)$ and $p_{n,B}(\theta)$, and their limits, for cases (i) $\theta = 0$, i.e. model 0 is correct, and (ii) $\theta \neq 0$, i.e. model 1 is correct. Comment on your findings.

(d) Consider

$$\hat{\theta}_{a,\text{final}} = \begin{cases} 0 & \text{if AIC selects model 0} \\ \bar{y} & \text{if AIC selects model 1} \end{cases} = \bar{y} I(|\sqrt{n}\bar{Y}| \geq \sqrt{2}).$$

Find an expression for the risk function $\text{risk}_a(\theta) = E\{\sqrt{n}|\hat{\theta}_{a,\text{final}} - \theta|\}^2$, and similarly for $\text{risk}_b(\theta)$. Again compute these, display them, for $n = 10, 50, 100$, and comment on your findings. (One may simulate to produce these curves, for a grid of θ values, without having to work out mathematical formulae.)

Ex. 12.10 Comparing AIC and BIC, in prototype setup II. Suppose we have sample $x_1, \dots, x_n$ from $N(a, 1)$ and another sample $y_1, \dots, y_n$ from $N(b, 1)$. Consider model 1, that $a = b$, and model 2, that $a \neq b$.

(a) Show that ML estimators are the sample means $\hat{a} = \bar{X}$ and $\hat{b} = \bar{Y}$. A natural parameter to focus on in this setup is the mean difference $\delta = b - a$. Show that $\hat{\delta} = \hat{b} - \hat{a} \sim N(\delta, 2/n)$.

(b) Find expressions for the maximised log-likelihoods, say $\ell_{1,\max}$ and $\ell_{2,\max}$. Give also expressions for $\text{AIC}_2 - \text{AIC}_1$ and $\text{BIC}_2 - \text{BIC}_1$. Letting $Z_n = (n/2)^{1/2}\hat{\delta}$, show indeed that AIC prefers model 2 if $|Z_n| \geq \sqrt{2}$, whereas BIC sets the mark at $|Z_n| \geq \{\log(2n)\}^{1/2}$.

(c) Show that

$$\begin{aligned} p_{n,A}(a, b) &= \Pr_{a,b}(\text{AIC selects model 1}) = \Pr\{\chi_1^2(\tfrac{1}{2}n\delta^2) \geq 2\}, \\ p_{n,B}(a, b) &= \Pr_{a,b}(\text{BIC selects model 1}) = \Pr\{\chi_1^2(\tfrac{1}{2}n\delta^2) \geq \log(2n)\}, \end{aligned}$$

and plot these in a graph, for a few values of n . Characterise what happens when n grows.

(d) Then consider the final estimator for $\delta = b - a$, after the model selection step. For AIC, show that this is $\hat{\delta}_{\text{final}} = (\hat{b} - \hat{a}) I(|Z_n| \geq \sqrt{2})$. Show that the associated risk function for the AIC procedure may be expressed as

$$\text{risk}_A(a, b) = n \mathbb{E}_{a,b} (\hat{\delta}_{\text{final}} - \delta)^2 = n \mathbb{E} \{ (\delta + (2/n)^{1/2} N) I(|\delta + (n/2)^{1/2} N| \geq \sqrt{2}) - \delta \}^2,$$

with N a standard normal. Compute and display both this curve and the corresponding one for the BIC case in a diagram. Compute also the risk ratio $\text{risk}_B(a, b)/\text{risk}_A(a, b)$, and show that the max of this risk ratio tends to infinity with growing n . Translate this to prose.

Motivation and arguments underlying AIC and BIC

Ex. 12.11 *The motivation and arguments behind the AIC.* Suppose $Y_1, \dots, Y_n$ are i.i.d. from a density g , and that using a parametric model $f(y, \theta)$ is contemplated as an approximation, via its ML estimated density $f(y, \hat{\theta})$. Consider the Kullback–Leibler distance $\text{KL}(g, f(\cdot, \hat{\theta}))$ from the true density to the implied approximation,

$$\int g(y) \log \frac{g(y)}{f(y, \hat{\theta})} dy = \int g \log g dy - R_n, \quad \text{with } R_n = \int g(y) \log f(y, \hat{\theta}) dy.$$

The task becomes that of estimating R_n , or its expectation $E_n = \mathbb{E}_g R_n$, from the data. Then different models can be compared and ranked via their E_n values. A natural start estimator is $\hat{R}_n = n^{-1} \ell_n(\hat{\theta})$, the normalised log-likelihood maximum. The details to follow tell us by how much this start estimator overshoots its target.

Below we need a few general results reached in Ch. 5 for the behaviour of the ML estimator $\hat{\theta}$ outside model conditions; see Ex. 5.10–5.11. Specifically, $\hat{\theta} - \theta_0 = J^{-1} \bar{U}_n + o_{\text{pr}}(1/\sqrt{n})$, where θ_0 is the least false parameter value, the minimiser of $\text{KL}(g, f(\cdot, \theta))$, assumed here to be an inner point of its parameter region; J is the agnostic information matrix defined in those exercises; and $\bar{U}_n = n^{-1} \sum_{i=1}^n u_0(Y_i)$, the average of the score functions $u(Y_i, \theta_0)$, computed at this θ_0 . With p the dimension of θ , this led to $\sqrt{n} \bar{U}_n \rightarrow_d U \sim N_p(0, K)$, with K the variance matrix of $u(Y_i, \theta_0)$, and the fundamental limit distribution $J^{-1} U \sim N_p(0, J^{-1} K J^{-1})$ for $\sqrt{n}(\hat{\theta} - \theta_0)$, with the sandwich matrix.

(a) Writing G_n for the empirical distribution of the data, tending to the c.d.f. G , show that

$$\begin{aligned} \hat{R}_n - R_n &= \int \log f(y, \hat{\theta}) \{dG_n(y) - dG(y)\} \\ &= \int \{\log f(y, \theta_0) + u_0(y)^t (\hat{\theta} - \theta_0) + O_{\text{pr}}(\|\hat{\theta} - \theta_0\|^2)\} \{dG_n(y) - dG(y)\} \\ &= a_n - a + n^{-1} W_n + o_{\text{pr}}(n^{-1}), \end{aligned}$$

where $a_n - a = n^{-1} \sum_{i=1}^n \log f(Y_i, \theta_0) - \int g(y) \log f(y, \theta_0) dy$ has mean zero, and where

$$W_n = n \bar{U}_n^t (\hat{\theta} - \theta_0) = n \bar{U}_n^t J^{-1} \bar{U}_n \rightarrow_d W = U^t J^{-1} U.$$

(b) Show that $\mathbb{E} W = p^* = \text{Tr}(J^{-1}K)$. If the parametric model holds, show that $p^* = p$, and we may call p^* the generalised dimension for the model. Show indeed that under model conditions, $W \sim \chi_p^2$.

(c) Deduce that $n^{-1}\ell_{n,\max} - n^{-1}p^*$ is approximately unbiased for $E_n = \mathbb{E}_g R_n$. Argue that this leads to $\ell_{n,\max} - \hat{p}^*$ as the natural correction to $\ell_{n,\max}$, with $\hat{p}^*$ any natural estimator of p^* , like $\hat{p}^* = \text{Tr}(\hat{J}^{-1}\hat{K})$, with these matrix estimators worked with in Ex. 5.29.

(d) Argue that all of this points to using

$$\text{AIC}^* = 2\ell_{n,\max} - 2\hat{p}^* \quad (12.4)$$

AIC*, the
robust AIC

as information criterion. This is sometimes called the robust AIC, in comparison with the simpler (12.1). When can the AIC and AIC* be expected to be close?

(e) For later purposes it is good to know also the following, regarding the attained log-likelihood maximum. Using $J_n = -n^{-1}\partial^2\ell_n(\theta_0)/\partial\theta\partial\theta^t$, show via Taylor expansion that

$$\begin{aligned} n^{-1}\ell_{n,\max} &\doteq n^{-1} \sum_{i=1}^n \{\log f(y_i, \theta_0) + u(y_i, \theta_0)^t(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^t i(y_i, \theta_0)(\hat{\theta} - \theta_0)\} \\ &= n^{-1}\ell_n(\theta_0) + \bar{U}_n^t(\hat{\theta} - \theta_0) - \frac{1}{2}(\hat{\theta} - \theta_0)^t J_n(\hat{\theta} - \theta_0) \\ &= n^{-1}\ell_n(\theta_0) + \frac{1}{2}\bar{U}_n^t J^{-1} \bar{U}_n + o_{\text{pr}}(1/n). \end{aligned} \quad (12.5)$$

Show furthermore that with $R_0 = \int g(y) \log f(y, \theta_0) dy$, we have

$$R_n \doteq R_0 - \frac{1}{2}(\hat{\theta} - \theta_0)^t J(\hat{\theta} - \theta_0) = R_0 - \frac{1}{2}\bar{U}_n^t J^{-1} \bar{U}_n,$$

(f) Work through details for regression setups, via Ex. 5.42 and 5.55, to see that the arguments above carry through, with essentially the same results, modulo small regression modifications to the formulae.

Ex. 12.12 *Leave-one-out statistics.* Consider a parameter functional $T(F)$ with influence function $\text{IF}(F, y)$, as for Ex. 6.17, with estimator $\hat{\theta} = T(F_n)$, based on observations $y_1, \dots, y_n$. Here we learn valuable generic ways of de-biasing, related to a little list of connected key words: leave-one-out, cross validation, jackknifing.

(a) Let $\hat{\theta}_{(i)}$ be the estimator in question computed for the dataset where y_i is pushed out, with $\hat{\theta}_{\text{jack}} = n^{-1} \sum_{i=1}^n \hat{\theta}_{(i)}$ their average again. Assuming sufficient smoothness of the $T(F)$ for the (6.6) representation to hold, deduce $\mathbb{E} \hat{\theta} = \theta + b/n + o(1/n)$ and $\mathbb{E} \hat{\theta}_{(i)} = \theta + b/(n-1) + o(1/n)$. From this construct $\hat{b}_{\text{jack}}/n = (n-1)(\hat{\theta}_{\text{jack}} - \hat{\theta})$, and via this again form the bias corrected estimator $\hat{\theta}^* = n\hat{\theta} - (n-1)\hat{\theta}_{\text{jack}}$. Show that $\mathbb{E} \hat{\theta}^* = \theta + o(1/n)$, i.e. the order $1/n$ bias term has been shaved off. Common terms describing this operation include cross-validation and jackknifing.

(b) For numbers $a_1, \dots, a_n$ with average $\bar{a}$, show that the average of those remaining after having pushed out a_i can be presented as

$$\bar{a}_{(i)} = \frac{1}{n-1} \sum_{j \neq i} a_j = \frac{(n-1)\bar{a} - (a_i - \bar{a})}{n-1} = \bar{a} - \frac{a_i - \bar{a}}{n-1}.$$

(c) Let $\hat{\theta}_{(i)}$ be the estimator in question computed for the dataset where y_i is pushed out. Show that $\hat{\theta}_{(i)} \doteq \hat{\theta} - (n-1)^{-1} \text{IF}(F_n, y_i)$.

(d) Show that

$$\Sigma_n = \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta})(\hat{\theta}_{(i)} - \hat{\theta})^t \doteq (1/n^2) \sum_{i=1}^n \text{IF}(F_n, y_i) \text{IF}(F_n, y_i)^t,$$

and argue that this is a natural estimator of the variance matrix for $\hat{\theta}$. There are other ways of estimating this variance matrix; the point at present is that one via jackknifing can estimate both bias and variance, in a nonparametric fashion.

Ex. 12.13 *log-likelihood cross-validation and the robust AIC.* Consider the setup of Ex. 12.11, where we saw that the direct estimator $n^{-1}\ell_{n,\max}$ tends to overshoot its target $R_n = \int g(y) \log f(y, \hat{\theta}) dy$ with a certain amount, namely $n^{-1}p^*$, with $p^* = \text{Tr}(J^{-1}K)$, involving the two matrices from the sandwich story of ML behaviour outside model conditions; see Ex. 5.10. There is another take on this story, via cross-validation, as follows.

(a) Show that R_n is the same as $E_g \log f(Y_{n+1}, \hat{\theta})$, with Y_{n+1} a new observation from g . Argue that this invites the cross-validation statistic $xv_n = n^{-1} \sum_{i=1}^n \log f(Y_i, \hat{\theta}_{(i)})$, with $\hat{\theta}_{(i)}$ the leave-one-out estimator computed for the dataset where y_i is pushed out. The xv_n can be computed directly from the data, though needing n separate evaluations of the parameter estimate.

(b) Now use the approximation

$$\hat{\theta}_{(i)} \doteq \hat{\theta} - n^{-1} \text{IF}(F_n, y_i) \doteq \hat{\theta} - n^{-1} \hat{J}^{-1} \hat{u}_i,$$

from Ex. 12.12, with $\hat{J}$ the normalised information matrix, as in Ex. 12.11, and $\hat{u}_i = u(Y_i, \hat{\theta})$, the estimated score for datum i . Show that

$$\begin{aligned} xv_n &\doteq n^{-1} \sum_{i=1}^n \log f(y_i, \hat{\theta} - n^{-1} \hat{J}^{-1} \hat{u}_i) \\ &\doteq n^{-1} \sum_{i=1}^n \{ \log f(y_i, \hat{\theta}) - n^{-1} \hat{u}_i^t \hat{J}^{-1} \hat{u}_i \} = n^{-1} \ell_{n,\max} - n^{-1} \text{Tr}(\hat{J}^{-1} \hat{K}). \end{aligned}$$

We have shown that xv_n is to the first order of approximation equal to $\text{AIC}^*/(2n)$, hence equivalent to the robust AIC as a selection criterion.

(c) For the lifelength dataset of Roman Era Egypt, studied for various parametric models in Story #50, consider the Gamma and the Weibull models. For these two models, compute $\hat{J}$ and $\hat{K}$ and hence $\hat{p}^* = \text{Tr}(\hat{J}^{-1} \hat{K})$, and carry out robust AIC analysis. Then compare with cross validation, and check the $\text{AIC}^* = 2n xv_n$ approximation.

Ex. 12.14 *Cross-validation for other criteria.* (xx above: $xv_n = n^{-1} \sum_{i=1}^n \log f(y_i, \hat{\theta}_{(i)})$, seen to be closely related to the robust aic. can cross-validate other criteria than the log-density. e.g. $n^{-1} \sum_{i=1}^n |y_i - \hat{y}_{i,(i)}|$ etc. spell out, give a simple example. xx)

Ex. 12.15 *The motivation and arguments behind the BIC.* The basic ideas underlying the BIC are rather different from those leading to the AIC, see Ex. 12.11. For BIC the starting idea is to put priors on all unknowns and then sort candidate models, say $M_1, \dots, M_k$, by how likely they are, given the data. This is often a tall order, as it involves (a) having prior probabilities $p_1, \dots, p_k$ on the models and (b) placing priors on each of the parameter vectors θ_j . We shall find a simple to use approximation to the posterior probabilities, however.

(a) Given such a setup, with a prior $\pi_j(\theta_j)$ on the parameter vector for model j , show that

$$\Pr(M_j | \text{data}) = \frac{p_j \bar{f}_j}{p_1 \bar{f}_1 + \cdots + p_k \bar{f}_k}, \quad \text{where } \bar{f}_j = \int L_{n,j}(\theta_j) \pi_j(\theta_j) d\theta_j.$$

(b) For an illustration, consider the simple setup with the flipping of two coins, in Ex. 12.2, with probabilities p_a, p_b . For model M_1 , take a uniform for the common $p_a = p_b$, and for model M_2 , take independent uniforms on p_a and p_b ; to complete the Bayesian prior description, assume the two models are equally likely before the coin flipping starts. With outcomes $(x_a, x_b) = (17, 24)$ for the two binomials, find the marginals $\bar{f}_1$ and $\bar{f}_2$, and compute $\Pr(M_1 | \text{data})$ and $\Pr(M_2 | \text{data})$.

(c) Generalise to the follow-up question for Ex. 12.2, with three coins, and outcomes 17, 24, 15 for the three binomials, each with $n = 40$. For the five models, take uniforms for the probabilities involved, and find the marginals $\bar{f}_1, \dots, \bar{f}_5$ for the five models. Complete the analysis by presenting the posterior probabilities for the five models, for the case of the five models being equally likely a priori.

(d) It is clear that a careful and accurate analysis for $\Pr(M_j | \text{data})$ is demanding, in more complex setups, conceptually and computationally. There is a simple approximation, however, using methods and results from Ex. 7.38. With data size n observations, with $\hat{\theta}_j$ the ML estimator for model M_j , of dimension p_j , associated with log-likelihood function $\ell_j(\theta_j)$, show that

$$\log \bar{f}_j \doteq \ell_{j,\max} - \frac{1}{2} p_j \log n + \log \pi_j(\hat{\theta}_j) - \frac{1}{2} \log |J_{n,j}| + \frac{1}{2} p_j \log(2\pi).$$

(e) Show that the two first terms in this approximations dominate the others, with growing n . Show that this leads to the simple approximation

$$\Pr(M_j | \text{data}) \doteq p_j q_j / (p_1 q_1 + \cdots + p_k q_k),$$

where $q_j = \exp(\frac{1}{2} \text{BIC}_j)$, in terms of $\text{BIC}_j = 2 \ell_{j,\max} - p_j \log n$. Show also that the prior probabilities p_j become unimportant here, as data size n grows. In yet other words, all Bayesian aspects of the postulated prior setup disappear, for growing data size. The models with highest posterior probabilities are those with highest BIC scores.

Squared bias, variance, the FIC

Ex. 12.16 *Estimating squared biases.* When we come to FIC building, in exercises below, one of its components is to estimate squared biases inside certain general setups. Here we examine relevant themes, inside simplified prototype setups.

(a) Assume $X \sim N(b, 1)$, and that $\phi = b^2$ is to be estimated. Revisit Ex. 9.37, to see three different estimators for b^2 , along with their risk functions. Attempt to construct an additional estimator, and check its performance.

(b) Assume $D \sim N_q(\delta, Q)$, with δ unknown and Q known, in dimension q . Suppose further that the main parameter of interest is the squared mean for a given linear combination, say $\mu = (c^t \delta)^2$. From $c^t D \sim N(c^t \delta, c^t Q c)$, draw on what you found above, to compare the pluses and minuses of these three estimators: (i) $\hat{\mu}_1 = (c^t D)^2$; (ii) $\hat{\mu}_2 = (c^t D)^2 - c^t Q c$; (iii) $\hat{\mu}_3 = \max((c^t D)^2 - c^t Q c, 0)$.

Ex. 12.17 *Parametric or nonparametric: a FIC view.* Consider a setup with i.i.d. observations $Y_1, \dots, Y_n$ from some distribution G , with a focus parameter $\mu(G)$ to be estimated. The nonparametric choice is $\hat{\mu}_{\text{np}} = \mu(G_n)$, in terms of the empirical c.d.f. G_n . With a parametric model, with density $f_\theta(y) = f(y, \theta)$ and c.d.f. F_θ , we may construct an estimator $\hat{\theta}$, leading to the parametric choice $\hat{\mu}_{\text{para}} = \mu(\hat{\theta})$, where $\mu(\theta) = \mu(F_\theta)$. We suppose $\hat{\theta} \rightarrow_{\text{pr}} \theta_0$, a least false parameter value, without assuming that $F(\cdot, \theta_0)$ is equal to G ; for the ML estimator, the θ_0 is the best parametric approximation in terms of Kullback–Leibler distance; see Ex. 5.3. We address situations where we first can establish that

$$\begin{pmatrix} \sqrt{n}\{\hat{\mu}_{\text{para}} - \mu(\theta_0)\} \\ \sqrt{n}(\hat{\mu}_{\text{np}} - \mu_{\text{true}}) \end{pmatrix} \rightarrow_d \begin{pmatrix} U \\ V \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_{\text{para}} & k \\ k & v_{\text{np}} \end{pmatrix}\right),$$

for suitable variances v_{para} and v_{np} and covariance k . This typically follows from first showing $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d U_0 \sim N_p(0, \Sigma)$, say, with p the dimension of θ and Σ a $p \times p$ matrix, and then applying the delta method to reach $\sqrt{n}\{\hat{\mu}_{\text{para}} - \mu(\theta_0)\} \rightarrow_d U = d^t U_0$, with variance $v_{\text{para}} = d^t \Sigma d$, where $d = \partial \mu(\theta_0) / \partial \theta$.

(a) Assume first that $\hat{\mu}_{\text{np}}$ and $\hat{\mu}_{\text{para}}$ are approximately unbiased estimators, in the sense that their biases are $o(1/n)$; as we have seen in Ex. 12.12 this can be accomplished via jackknifing when the parameter functionals are smooth. Explain that this leads to the approximations

$$\text{mse}_{\text{para}} = b^2 + v_{\text{para}}/n, \quad \text{mse}_{\text{np}} = 0^2 + v_{\text{np}}/n,$$

where we write $b = \mu(\theta_0) - \mu_{\text{true}}$ for the modelling bias. Hence the parametric choice is better than the nonparametric, in cases where $nb^2 \leq v_{\text{np}} - v_{\text{para}}$, or $\sqrt{n}|b| \leq (v_{\text{np}} - v_{\text{para}})^{1/2}$. Assuming $b \neq 0$, i.e. the model is not perfect, show that the nonparametric estimator wins, for sufficiently large n . For moderate or small n , however, the parametric may win, even if the presence of modelling bias.

(b) We shall construct our first FIC, the Focused Information Criterion, by estimating the mse_{para} and mse_{np} risks from data. To estimate the squared bias, consider $\hat{b} = \hat{\mu}_{\text{para}} - \hat{\mu}_{\text{np}}$, and show that $\sqrt{n}(\hat{b} - b) \rightarrow_d U - V \sim N(0, w)$, with $w = v_{\text{np}} + v_{\text{para}} - 2k$. Explain under the unbiasedness assumption that $E\hat{b}^2 = b^2 + w/n + o(1/n)$, i.e. $\hat{b}^2$ overshoots its target b^2 ; see Ex. 12.16. With $\hat{w}$ and $\hat{v}_{\text{para}}$ consistent estimators for w and v_{para} , with means $O(1/n)$ away from these, show that $\text{FIC}_{\text{para}}^u = \hat{b}^2 - \hat{w}/n + \hat{v}_{\text{para}}/n$ is unbiased for mse_{para} up to and including order $1/n$. In case of $\hat{b}^2 - \hat{w}/n$ being negative, we truncate this biased squared estimator to zero, and we are led to

$$\text{FIC}_{\text{para}} = \max(\hat{b}^2 - \hat{w}/n, 0) + \hat{v}_{\text{para}}/n, \quad \text{FIC}_{\text{np}} = 0^2 + \hat{v}_{\text{np}}/n.$$

(c) Naturally estimators $\hat{\mu}_{\text{np}}$ and $\hat{\mu}_{\text{para}}$ are not always unbiased, and we do need to take $O(1/n)$ size biases into account. Assume it can be established that

$$\begin{aligned} \hat{\mu}_{\text{para}} &= \mu(\theta_0) + c_{\text{para}}/n + n^{-1/2}Z_{n,\text{para}} + o(1/n), \\ \hat{\mu}_{\text{np}} &= \mu_{\text{true}} + c_{\text{np}}/n + n^{-1/2}Z_{n,\text{np}} + o(1/n), \end{aligned}$$

with $(Z_{n,\text{para}}, Z_{n,\text{np}})$ having zero means and tending to joint binormality, with variances v_{para} and v_{np} and covariance k , as above. In particular the parametric bias is $b_n = \mu(\theta_0) - \mu_{\text{true}} + c_{\text{para}}/n = b + c_{\text{para}}/n$, up to order $1/n$. Show from this that

$$\text{mse}_{\text{para}} = b_n^2 + v_{\text{para}}/n = b^2 + 2bc_{\text{para}}/n, \quad \text{mse}_{\text{np}} = v_{\text{np}}/n,$$

up to and including order $1/n$. Explain hence that we may keep the nonparametric FIC score as $\widehat{v}_{\text{np}}/n$.

(d) Show that $\widehat{b} = b + (c_{\text{para}} - c_{\text{np}})/n + n^{-1/2}(Z_{n,\text{para}} - Z_{n,\text{np}})$. With $\widehat{c}_{\text{np}}$ consistent and order $1/n$ unbiased for c_{np} , show that this leads to

$$E_{\text{true}} \widehat{b}^2 = b^2 + 2b(c_{\text{para}} - c_{\text{np}})/n + w/n, \quad E_{\text{true}}(\widehat{b}^2 + 2\widehat{b}\widehat{c}_{\text{np}}/n - \widehat{w}/n) = b_n^2,$$

again up to and including order $1/n$ terms. Hence we have a nearly unbiased and a truncated-to-zero modified estimators of the mean squared errors,

$$\begin{aligned} \text{FIC}_{\text{para}}^u &= \widehat{b}^2 + 2\widehat{b}\widehat{c}_{\text{np}}/n - \widehat{w}/n + \widehat{v}_{\text{para}}/n, \\ \text{FIC}_{\text{para}} &= \max(\widehat{b}^2 + 2\widehat{b}\widehat{c}_{\text{np}}/n - \widehat{w}/n, 0) + \widehat{v}_{\text{para}}/n. \end{aligned}$$

These scores can then be computed, compared, and ranked, from the winning parametric model with the smallest FIC to the others; ranking can of course all include the nonparametric model, with $\text{FIC}_{\text{np}} = \widehat{v}_{\text{np}}/n$. It is to be noted that only the c_{np}/n leading bias term needs to be assessed and estimated, as the various c_{para}/n terms cancel out.

(e) A rather unified recipe for estimating the $v_{\text{para}}, v_{\text{np}}, k, w$ quantities can be given via influence function representation of the estimators. Suppose indeed that

$$\begin{aligned} \widehat{\mu}_{\text{para}} &= \mu(\theta_0) + c_{\text{para}}/n + n^{-1} \sum_{i=1}^n \text{IF}_{\text{para}}(G, Y_i) + o_{\text{pr}}(n^{-1/2}), \\ \widehat{\mu}_{\text{np}} &= \mu_{\text{true}} + c_{\text{np}}/n + n^{-1} \sum_{i=1}^n \text{IF}_{\text{np}}(G, Y_i) + o_{\text{pr}}(n^{-1/2}), \end{aligned}$$

i.e. with influence functions $\text{IF}_{\text{para}}(G, y)$ and $\text{IF}_{\text{np}}(G, y)$, having finite variances. Then use

$$\widehat{v}_{\text{para}} = n^{-1} \sum_{i=1}^n \widehat{\text{IF}}_{\text{para}}(Y_i)^2, \quad \widehat{v}_{\text{np}} = n^{-1} \sum_{i=1}^n \widehat{\text{IF}}_{\text{np}}(Y_i)^2, \quad \widehat{k} = n^{-1} \sum_{i=1}^n \widehat{\text{IF}}_{\text{para}}(Y_i) \widehat{\text{IF}}_{\text{np}}(Y_i),$$

followed by $\widehat{w} = \widehat{v}_{\text{para}} + \widehat{v}_{\text{np}} - 2\widehat{k}$ and the FIC scores. Here $\widehat{\text{IF}}_{\text{para}}(y)$ and $\widehat{\text{IF}}_{\text{np}}(y)$ are the empirical and consistent versions of $\text{IF}_{\text{para}}(G, y)$ and $\text{IF}_{\text{np}}(G, y)$. The nonparametric bias ingredient $\widehat{c}_{\text{np}}/n$ needs to be tended to for each application (and is in many cases equal to zero). A generic jackknifing procedure for this step is in Ex. 12.12.

Ex. 12.18 *Parametric or nonparametric for quantiles?* Suppose $Y_1, \dots, Y_n$ are i.i.d. from some smooth c.d.f. G , with positive density g . The task is to estimate the quantile $\mu = Q(p) = G^{-1}(p)$, for one or perhaps many p . Should we do this nonparametrically or parametrically, say using the normal distribution?

(a) Below we need a certain technical detail. With $Q_n(p)$ the sample quantile, computed via linear interpolation as per the default choice of $\mathbf{R}$,

$$E Q_n(p) = Q(p) + c(p)/n \quad \text{with } c(p) = (1 - 2p)/f(\mu) - \frac{1}{2}p(1 - p)f'(\mu)/f(\mu)^3.$$

So for the c_{np}/n ingredient in the general FIC setup of Ex. 12.17, we use this formula, with standard kernel density based estimates f_n and f'_n for f and f' , see Ex. 3.42, evaluated at position $Q_n(p)$. Illustrate and check this, by simulating the exact $n\{E Q_n(p) - Q(p)\}$ for e.g. the normal density, for a little range of p , perhaps with $n = 100$, and comparing to (i) the $c(p)$ formula, and (ii) to the estimated version of that formula, with f_n and f'_n .

check with
Story #22 to
see this in
practice

from Dæhlen
et al. (2024,
Section 2)

(b) For the sample quantile $Q_n(p)$, check from Ex. 6.18 that its influence function $\text{IF}_{\text{np}}(G, y)$ is $-(1-p)/g(Q(p))$ to the left of $Q(p)$ and $p/g(Q(p))$ to the right. Now consider the normal alternative, which is $\hat{\mu}_p = \hat{\xi} + \hat{\sigma} z_p$, with the familiar estimators for mean and standard deviation, and with z_p the normal quantile. Follow the general recipe of Ex. 12.17 to show that

$$\widehat{\text{IF}}_{\text{para}}(y) = \begin{pmatrix} 1 \\ z_p \end{pmatrix}^t \hat{\sigma}^2 \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \frac{1}{\hat{\sigma}} \begin{pmatrix} z(y) \\ z(y)^2 - 1 \end{pmatrix} = \hat{\sigma}[z(y) + \frac{1}{2}z_p\{z(y)^2 - 1\}],$$

where $z(y) = (y - \hat{\xi})/\hat{\sigma}$. The FIC scheme is now to be pursued, involving $\hat{b} = \hat{\xi} + \hat{\sigma} z_p - Q_n(p)$.

(c) Work through the FIC details also for the three-parameter t model with $Y_i = \xi + \sigma \varepsilon_i$, with $\varepsilon_i \sim t_\nu$. For implementation, necessary for the Oslo boys and girls of Story #22, we use numerical derivation for finding both $u(y, \hat{\theta}_{\text{ml}})$, the $\hat{d}$, the $\hat{J}$. Explain that the influence function is $\widehat{\text{IF}}_{\text{para}}(y) = \hat{d}^t \hat{J}^{-1} u(y, \hat{\theta}_{\text{ml}})$, with $\hat{d} = \partial \mu(\hat{\theta}_{\text{ml}})/\partial \theta$, for $\mu(\theta) = \xi + \sigma G_\nu^{-1}(p)$, in terms of the t-quantile function.

(d) With positive data $y_1, \dots, y_n$, should one use the nonparametric median, or rather $\log 2/\hat{\theta}$ from having fitted an $\text{Expo}(\theta)$ model? Work out the FIC details.

Tolerance radii around models

Ex. 12.19 *ML behaviour under local alternatives, I.* Consider a parametric density $f(y, \theta)$ for observations $Y_1, \dots, Y_n$. Assume however that these do not precisely follow the model, but rather a density in its vicinity. What can we deduce for the behaviour of the ML estimator $\hat{\theta}$? Results reached here are useful for local power analysis and indeed for the construction of model selection criteria, in particular for the FIC below. We start in the i.i.d. framework, for sorting out matters in a clean fashion, though results reached here will be generalised to regression contexts in Ex. 12.28.

(a) Assume the real state of affairs is a density

$$f_n(y) = f(y, \theta_0)\{1 + \delta h(y)/\sqrt{n}\}, \quad \text{for some } h \text{ with } \int f(y, \theta_0)h(y) dy = 0. \quad (12.6)$$

Write $f_0(y) = f(y, \theta_0)$ and $u_0(y) = u(y, \theta_0)$ for density and score function at θ_0 . With notation and arguments as in Ex. 5.10, show that $U_n = n^{-1/2} \sum_{i=1}^n u_0(Y_i)$ has mean $b\delta = \int f_0(y)h(y) dy \delta$ and variance matrix tending to $J = \int f_0(y)u_0(y)u_0(y)^t dy$. Show via the Lindeberg theorem that indeed $U_n \rightarrow_d U \sim N_p(b\delta, J)$. With $\ell_n(\theta)$ the log-likelihood function and $J_n = -n^{-1} \partial^2 \ell_n(\theta_0)/\partial \theta \partial \theta^t$, show also that $J_n \rightarrow_{\text{pr}} J$, even when the data density f_n is $O(1/\sqrt{n})$ away from f_0 . Proceed to show

$$A_n(s) = \ell_n(\theta_0 + s/\sqrt{n}) - \ell_n(\theta_0) = U_n^t s - \frac{1}{2} s^t J_n s + r_n(s) \rightarrow_d A(s) = U^t s - \frac{1}{2} s^t J s,$$

as with (5.5), but now with the additional $b\delta$ mean for U . Deduce from this that

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d J^{-1} U \sim N_p(J^{-1} b\delta, J^{-1}),$$

and also that $2\{\ell_{n,\text{max}} - \ell_n(\theta_0)\} \rightarrow_d U^t J^{-1} U$, which has a $\chi_p^2(b^t J^{-1} b \delta^2)$ distribution. We learn that when the real data mechanism is $O(1/\sqrt{n})$ away from the model, then bias is affected, but not the variance, for large n .

(b) Consider a focus parameter $\mu(\theta)$, with true value now determined from (12.6), and with $c = \partial\mu(\theta_0)/\partial\theta$. For log-likelihood profiling and the deviance, show using Ex. 5.13–6.14 that

$$D_n(\mu_0) = 2\{\ell_{n,\text{prof,max}} - \ell_{n,\text{prof}}(\mu_0)\} \rightarrow_d \frac{(c^t J^{-1} U)^2}{c^t J^{-1} c} \sim \chi_1^2(\lambda \delta^2),$$

with $\lambda = (c^t J^{-1} b)^2 / c^t J^{-1} c$. (xx invent two simple illustrations. point to local power things already in Ch 4. xx)

Ex. 12.20 *ML behaviour under local alternatives, II.* As an important special case of the machinery of Ex. 12.19, assume that a certain narrow model with density $f_0(y, \theta)$ is used, with dimension p , but that the reality is that data come from a wider model, needing an extra parameter γ of dimension q . Suppose indeed that the real density is $f_{\text{true}}(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n})$, where setting $\gamma = \gamma_0$ in the $p+q$ -dimensional model $f(y, \theta, \gamma)$ gives us back the $f_0(y, \theta)$ model. For this wider model, introduce score vector component $u_0(y) = \partial \log f(y, \theta_0, \gamma_0) / \partial \theta$ and $u_1(y) = \partial \log f(y, \theta_0, \gamma_0) / \partial \gamma$. It is now fruitful to introduce

the important
Q matrix

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \quad \text{with inverse} \quad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & Q \end{pmatrix},$$

the $(p+q) \times (p+q)$ Fisher information matrix for the wider model, with inverse, computed at the position (θ_0, γ_0) . The $q \times q$ matrix $Q = J^{11} = (J_{11} - J_{10} J_{00}^{-1} J_{01})^{-1}$ will play a crucial role, in particular for the FIC to come below. It is a special case of what we worked with in Ex. 5.14.

(a) Let $\bar{U}_0 = n^{-1} \sum_{i=1}^n u_0(Y_i)$ and $\bar{U}_1 = n^{-1} \sum_{i=1}^n u_1(Y_i)$. Show now that

$$\begin{pmatrix} U_{n,0} \\ U_{n,1} \end{pmatrix} = \begin{pmatrix} \sqrt{n} \bar{U}_0 \\ \sqrt{n} \bar{U}_1 \end{pmatrix} \rightarrow_d \begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim N_{p+q} \left(\begin{pmatrix} J_{01} \delta \\ J_{11} \delta \end{pmatrix}, \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \right).$$

With $\hat{\theta}_{\text{narr}}$ the ML estimator in the narrow model, show that

$$\sqrt{n}(\hat{\theta}_{\text{narr}} - \theta_0) = J_{00}^{-1} \sqrt{n} \bar{U}_{n,0} + o_{\text{pr}}(1) \rightarrow_d J_{00}^{-1} U_0 \sim N_p(J_{00}^{-1} J_{01} \delta, J_{00}^{-1}).$$

(b) Next consider ML estimation in this wide model, with $(\hat{\theta}_{\text{wide}}, \hat{\gamma}_{\text{wide}})$. Show that

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_{\text{wide}} - \theta_0) \\ \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0) \end{pmatrix} = J^{-1} \begin{pmatrix} U_{n,0} \\ U_{n,1} \end{pmatrix} + o_{\text{pr}}(1) \rightarrow_d J^{-1} \begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim N_{p+q} \left(\begin{pmatrix} 0 \\ \delta \end{pmatrix}, J^{-1} \right).$$

In particular, note that

$$D_n = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0) \rightarrow_d D \sim N_q(\delta, Q). \quad (12.7)$$

(c) So how do narrow model estimation and wide model estimation pan out, when it comes to some given focus parameter? Consider such a $\mu = \mu(\theta, \gamma)$, with true value $\mu_{\text{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. The two estimators under consideration are $\hat{\mu}_{\text{narr}} = \mu(\hat{\theta}_{\text{narr}}, \gamma_0)$ and $\hat{\mu}_{\text{wide}} = \mu(\hat{\theta}_{\text{wide}}, \hat{\gamma}_{\text{wide}})$. Show that

$$\begin{aligned} \sqrt{n}(\hat{\mu}_{\text{narr}} - \mu_{\text{true}}) &\rightarrow_d N(\omega^t \delta, \tau_0^2), \\ \sqrt{n}(\hat{\mu}_{\text{wide}} - \mu_{\text{true}}) &\rightarrow_d N(0, \tau_0^2 + \omega^t Q \omega), \end{aligned} \quad (12.8)$$

featuring

$$\tau_0^2 = \left(\frac{\partial \mu}{\partial \theta} \right)^t J_{00}^{-1} \frac{\partial \mu}{\partial \theta} \quad \text{and} \quad \omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}, \quad (12.9)$$

with these partial derivatives computed at the null position (θ_0, γ_0) .

(d) So when is narrow model estimation still best, even if that model might be somewhat wrong? Narrow model means bias but smaller variance; wide model means lower bias but higher variance. This bias-variance trade-off is nicely captured in the (12.8) limits. Show that the limit risks, mean squared errors for the limit distributions, are

$$\text{mse}_{\text{narr}} = \tau_0^2 + (\omega^t \delta)^2 \quad \text{and} \quad \text{mse}_{\text{wide}} = \tau_0^2 + \omega^t Q \omega.$$

Conclude that narrow is best, provided $|\omega^t \delta| \leq (\omega^t Q \omega)^{1/2}$. We may translate this to $|\omega^t(\gamma - \gamma_0)| \leq (\omega^t Q \omega)^{1/2}/\sqrt{n}$, where it is to be noted that the tolerance becomes smaller with growing sample size; with more data, models might need to become more sophisticated.

Ex. 12.21 *Computing the tolerance radius.* Consider a model $f_0(y, \theta)$, extended with one extra parameter γ , in a manner where the density $f(y, \theta, \gamma)$ becomes $f_0(y, \theta)$ when γ is equal to its null value γ_0 . We use the same local asymptotic framework as above, with data seen as coming from $f_n(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n})$, with δ unknown.

(a) For a focus parameter $\mu = \mu(\theta, \gamma)$, with estimators $\hat{\mu}_{\text{narr}}$ and $\hat{\mu}_{\text{wide}}$, show from Ex. 12.20 that the associated limiting risks are $\text{mse}_{\text{narr}} = \tau_0^2 + (\omega \delta)^2$ and $\text{mse}_{\text{wide}} = \tau_0^2 + \omega^2 \kappa^2$, with $\kappa^2 = Q$ the lower right element in the $(p+1) \times (p+1)$ matrix J^{-1} . Deduce that narrow estimation is better than wide estimation as long as $|\delta| \leq \kappa$, regardless of the particularities of μ . This translates to the tolerance radius $|\gamma - \gamma_0| \leq \kappa/\sqrt{n}$.

(b) Consider nonnegative data $y_1, \dots, y_n$ from the $\text{Gam}(a, \theta)$, with $a = 1$ corresponding to the simpler $\text{Expo}(\theta)$ model. For estimating the mean, for instance, there is then the choice between the simple $\hat{\mu}_{\text{narr}} = 1/\hat{\theta}_{\text{narr}} = \bar{Y}$ and the somewhat more complicated $\hat{\mu}_{\text{wide}} = \hat{a}_{\text{wide}}/\hat{b}_{\text{wide}}$. Writing V for a unit exponential, show that the Fisher information matrix J , computed at the narrow model, becomes

$$J = \text{Var} \begin{pmatrix} (1/\theta)(1 - V) \\ \log V - \psi(1) \end{pmatrix} = \begin{pmatrix} 1/\theta^2 & a/\theta \\ a/\theta & b \end{pmatrix},$$

with $\psi(1) = -0.5772\dots$, minus the Euler constant. Also, $b = \text{Var } V$ and $a = \text{cov}(-V, \log V)$ are found mathematically to be $\pi^2/6$ and -1 , though can also be found via numerical integration, in more complicated settings. Show from this that $\kappa^2 = 1/(b - a^2)$, leading to the conclusion that as long as $|a - 1| \leq 1.245/\sqrt{n}$, the simpler exponential model will fare better than the two-parameter Gamma.

(c) For another illustration, also with $\text{Expo}(\theta)$ being the narrow model, consider the Weibull extension with c.d.f. $F(y, \theta, \gamma) = 1 - \exp\{-(\theta y)^\gamma\}$, with null value $\gamma = 1$ giving back the exponential model. If the median is to be estimated, the choice is between the simple $\hat{\mu}_{\text{narr}} = (\log 2)/\hat{\theta}_{\text{narr}}$ and the more complicated $\hat{\mu}_{\text{wide}} = (\log 2)^{1/\hat{\gamma}_{\text{wide}}}/\hat{\theta}_{\text{wide}}$. Show that the Fisher information matrix J at the null model is the variance matrix of $((1/\theta)(1 - V), 1 + \log V - V \log V)$, taking a form as above but with different a and b . Find these values and conclude that narrow estimation is best provided $|\gamma - 1| \leq 0.779/\sqrt{n}$.

(d) Perhaps we occasionally ought to stretch the Gaussian tail? Consider the three-parameter density model

$$f(y, \xi, \sigma, \gamma) = \frac{1}{\sigma} \frac{1}{a(\gamma)} \exp\left\{-\frac{1}{2} \left| \frac{y - \xi}{\sigma} \right|^\gamma\right\}.$$

With $\gamma = 2$ we are back to Gauß, with $a(2) = (2\pi)^{1/2}$; also, $\gamma = 1$ corresponds to the double exponential. Tails are fatter than normal for $\gamma < 2$ and slimmer for $\gamma > 2$. Show

that actually $a(\gamma) = 2^{1+1/\gamma}\Gamma(1+1/\gamma)$, and further that the Fisher matrix, computed at normality, becomes

$$J = \text{Var} \begin{pmatrix} \varepsilon/\sigma \\ (\varepsilon^2 - 1)/\sigma \\ -\frac{1}{2}\varepsilon^2 \log |\varepsilon| + b \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 & 0 \\ 0 & 2/\sigma^2 & c/\sigma \\ 0 & c/\sigma & k \end{pmatrix},$$

where ε is a standard normal. Find numerical values for b, c, k here, and use these to find that if $|\gamma - 2| \leq 4.459/\sqrt{n}$, then tailstretching is not worth the extra trouble.

(e) Consider the three-parameter extension of the traditional two-parameter normal, with c.d.f. $F(y, \xi, \sigma, \lambda) = \Phi((y - \xi)/\sigma)^\lambda$. How much can the usual normality based estimation methods tolerate, in terms of λ away from 1, before they lose to the three-parameter based ML methods?

Ex. 12.22 *Gammaweibulling the exponential.* Consider the three-parameter density model

$$f(y, a, b, c) = k(a, b, c)y^{a-1} \exp(-by^c) \quad \text{for } y > 0,$$

with positive parameters a, b, c . Note that ML analysis is fully feasible here, via the log-likelihood function, as per the theory of Ch. 5; see Story #98 to see this in action. Here we examine the level of tolerance around the simple one-parameter model $\text{Expo}(b)$, sitting in the middle, corresponding to $(a, c) = (1, 1)$.

(a) Show that the Gamma and Weibull distributions are indeed special cases, and carry out integration to show that $k(a, b, c) = cb^{a/c}\Gamma(a/c)$. For tolerance analysis we need the Fisher information matrix $J(a, b, c)$, and its inverse, computed at exponentiality, i.e. $(a, c) = (1, 1)$. Show that the score functions, computed there, are

$$U(y) = -(1/b)W_1, \quad V_1(y) = W_2, \quad V_2(y) = (\log b)W_1 - W_3,$$

in terms of $W_1 = y - 1$, $W_2 = \log y - \psi(1)$, $W_3 = y \log y - \{1 + \psi(1)\}$, each with mean zero. Carry out numerical integration, to determine $J(1, b, 1)$ and its inverse, from which you should find the crucial lower-right 2×2 matrix

$$Q = \begin{pmatrix} 4.762 & -5.824 \\ -5.824 & 10.594 \end{pmatrix},$$

not depending on b .

(b) Go through as many as four different foci, $\mu = \mu(a, b, c)$, and check when $\text{mse}_{\text{narr}} \leq \text{mse}_{\text{wide}}$, which needs the partial derivatives and ω : (i) the mean $\xi = EY$; (ii) the standard deviation $\sigma = \text{sd}(Y)$; (iii) the probability $p = \Pr(Y \leq y_0)$ with $y_0 = 0.10$; (iv) the hazard rate $h = f(y_0)/\{1 - F(y_0)\}$ with $y_0 = 0.25$. Draw tolerance bands, which are infinite strips. Show then that $(\omega^t \delta)^2 \leq \omega^t Q \omega$ for all ω is equivalent to $\delta^t Q^{-1} \delta \leq 1$. Draw the tolerance ellipse of (a, c) inside which estimation using the narrow 1-parameter model will do better than estimation using the wide 3-parameter model, for *every* estimand. Show that this tolerance ellipsoid is the intersection of infinitely many tolerance strips.

tolerance
ellipsoid

Master theorem for ML behaviour in submodels

Ex. 12.23 *ML estimation and AIC in narrow and wide models.* Let the setup be as in Ex. 12.20, with a wide model $f(y, \theta, \gamma)$, having $p + q$ parameters, and the narrow

model $f_0(y, \theta) = f(y, \theta, \gamma_0)$, with p parameters, corresponding to setting γ equal to its null value γ_0 . In the exercise mentioned we have sorted out the basic of ML estimation, where a focus parameter $\mu = \mu(\theta, \gamma)$ has narrow and wide model based estimators $\hat{\mu}_{\text{narr}} = \mu(\hat{\theta}_{\text{narr}}, \gamma_0)$ and $\hat{\mu}_{\text{wide}} = \mu(\hat{\theta}_{\text{wide}}, \hat{\gamma}_{\text{wide}})$. The limit risk expressions also had consequences for tolerance radii. Here we work out further consequences, also for the AIC.

(a) In the notation of Ex. 12.20, let $U_{n,0} = \sqrt{n}\bar{U}_0$ and $U_{n,1} = \sqrt{n}\bar{U}_1$, with averages $\bar{U}_0 = n^{-1} \sum_{i=1}^n u_0(Y_i)$ and $\bar{U}_1 = n^{-1} \sum_{i=1}^n u_1(Y_i)$. For ML estimation in the narrow model, we have seen that

$$\sqrt{n}(\hat{\theta}_{\text{narr}} - \theta_0) \doteq J_{00}^{-1}\bar{U}_0 \sim N_p(J_{00}^{-1}J_{01}\delta, J_{00}^{-1}).$$

Then consider $W_n = \sqrt{n}(\bar{U}_1 - J_{10}J_{00}^{-1}\bar{U}_0)$, and show that $(U_{n,0}, W_n) \rightarrow_d (U_0, W)$, with these being independent, and $W \sim N_q(Q^{-1}\delta, Q^{-1})$. Show furthermore for $D_n = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0)$ of (12.7) that $D_n = QW_n + o_{\text{pr}}(1) \rightarrow_d QW = D \sim N_q(\delta, Q)$.

(b) With log-likelihood functions $\ell_{n,\text{narr}}(\theta)$ and $\ell_{n,\text{wide}}(\theta, \gamma)$, show via (12.5) that

$$\begin{aligned} \ell_{n,\text{wide},\text{max}} &= \ell_{n,0} + \frac{1}{2}n \begin{pmatrix} \bar{U}_{n,0} \\ \bar{U}_{n,1} \end{pmatrix}^t J^{-1} \begin{pmatrix} \bar{U}_{n,0} \\ \bar{U}_{n,1} \end{pmatrix} + o_{\text{pr}}(1), \\ \ell_{n,\text{narr},\text{max}} &= \ell_{n,0} + \frac{1}{2}n \bar{U}_{n,0}^t J_{00}^{-1} \bar{U}_{n,0} + o_{\text{pr}}(1), \end{aligned}$$

where $\ell_{n,0} = \sum_{i=1}^n \log f(y_i, \theta_0, \gamma_0)$.

(c) With notation for the blocks of the $(p+q) \times (p+q)$ matrices J and J^{-1} , from 12.20, demonstrate the algebraic fact that

$$\begin{pmatrix} u_0 \\ u_1 \end{pmatrix}^t J^{-1} \begin{pmatrix} u_0 \\ u_1 \end{pmatrix} - u_0^t J_{00}^{-1} u_0 = (u_1 - J_{10}J_{00}^{-1}u_0)^t Q (u_1 - J_{10}J_{00}^{-1}u_0).$$

Explain that $W^t Q W = D^t Q^{-1} D$, and show that

$$\text{AIC}_{\text{wide}} - \text{AIC}_{\text{narr}} = D_n^t Q^{-1} D_n - 2q + o_{\text{pr}}(1) \rightarrow_d D^t Q^{-1} D - 2q \sim \chi_q^2(\delta^t Q^{-1} \delta) - 2q.$$

We have accordingly understood the limit probability that AIC will favour the wider model, in terms of the local parameter δ , in $\gamma = \gamma_0 + \delta/\sqrt{n}$. Show in particular that this probability is $\Pr(\chi_q^2 > 2q)$, if the narrow model is correct. Compute these probabilities, for say $q = 1, \dots, 10$, and comment.

(d) For the case of γ being one-dimensional, write $Q = \kappa^2$ and $\hat{Q} = \hat{\kappa}^2$ from having estimated the limit information matrix J . Explain that $D_n/\hat{\kappa} = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0)/\hat{\kappa} \rightarrow_d D/\kappa \sim N(\delta/\kappa, 1)$. Show further that $\text{AIC}_{\text{wide}} - \text{AIC}_{\text{narr}} \rightarrow_d D^2/\kappa^2 - 2 \sim \chi_1^2(\delta^2/\kappa^2) - 1$. The AIC choosing the wide model is hence for large n equivalent with $|D_n|/\hat{\kappa} \geq \sqrt{2}$, an event having probability 0.157 for large n if the narrow model is correct.

(e) (xx round off this. need to define Λ_0 properly. more in Master Theorem. xx) After having selected the winning model, the final AIC selected estimator becomes

$$\hat{\mu}_{\text{final}} = \hat{\mu}_{\text{narr}} I(\text{AIC}_{\text{wide}} < \text{AIC}_{\text{narr}}) + \hat{\mu}_{\text{wide}} I(\text{AIC}_{\text{wide}} \geq \text{AIC}_{\text{narr}}).$$

Show that

$$\sqrt{n}(\hat{\mu}_{\text{final}} - \mu_{\text{true}}) \rightarrow_d \Lambda_0 + \omega^t \delta I(W^t Q W < 2q) + \omega(\delta - D) I(W^t Q W \geq 2q)$$

(f) (xx do and polish. xx) then estimate $\delta\delta^t$ via $DD^t - Q$, and make a FIC, for this narrow-vs-wide setup. generalised later, to include all $\hat{\mu}_S$.

Ex. 12.24 Master Theorem for ML estimation in submodels: the Limit Experiment. The setting is as with Ex. 12.20 and 12.23, where we so far have worked with the narrow p -parameter model $f_0(y, \theta) = f(y, \theta, \gamma_0)$ inside a wider $(p + q)$ -parameter model $f(y, \theta, \gamma)$. We shall build a framework for analysing estimators from all the intermediate models. Let S be a subset of $\{1, \dots, q\}$, of cardinality $|S|$, with ensuing model $f_S(y) = f(y, \theta, \gamma_S, \gamma_{0,S^c})$ of dimension $p + |S|$; here γ_S is used for the subset of γ_j with $j \in S$. This means allowing γ_j as a free parameter for $j \in S$, but setting $\gamma_j = \gamma_{0,j}$ for $j \in S^c$, the complement set. For any focus parameter $\mu = \mu(\theta, \gamma)$, this leads to submodel estimators of the type $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$, with $(\hat{\theta}_S, \hat{\gamma}_S)$ the ML estimator for submodel S .

Below we shall need the $Q = J^{11}$ matrix from the $(p + q) \times (p + q)$ Fisher information matrix, computed at the narrow model, along with $\tau_0^2 = (\frac{\partial \mu}{\partial \theta})^t J_{00}^{-1} \frac{\partial \mu}{\partial \theta}$ and $\omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$, worked with in Ex. 12.20, again with derivatives computed under the null model where $\gamma = \gamma_0$. Introduce also independent normal variables $\Lambda_0 \sim N(0, \tau_0^2)$ and $D \sim N_q(\delta, Q)$. The Master Theorem reached in (12.11) below is a clear representation of the limit distribution for $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$, in terms of these independent Λ_0 and D , from which limit bias and limit variance can be read off. (xx put in two sentences, here and or for previous, about regularity conditions; μ needs derivatives, J is finite and positive definite; and (θ_0, γ_0) is an inner point. xx)

(a) There will be results similar to those reached for $\hat{\mu}_{\text{narr}}$ and $\hat{\mu}_{\text{wide}}$, but we need a machinery for handling the book-keeping. Let $\pi_S: \mathbb{R}^q \rightarrow \mathbb{R}^{|S|}$ be the projection, taking $x = (x_1, \dots, x_q)^t$ to its subset x_S indexed by S . We may thus view π_S as a $|S| \times q$ matrix of zeros and ones. When handling the S model, there is a Fisher information matrix of size $(p + |S|) \times (p + |S|)$ and its inverse, say

$$J_S = \begin{pmatrix} J_{00} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{pmatrix} \quad \text{with inverse} \quad J_S^{-1} = \begin{pmatrix} J^{00,S} & J^{01,S} \\ J^{10,S} & J^{11,S} \end{pmatrix}.$$

Show that the lower-right matrix of the inverse can be written $Q_S = (\pi_S Q^{-1} \pi_S^t)^{-1}$, of size $|S| \times |S|$. An important book-keeping matrix of size $q \times q$ will be

$$G_S = \pi_S^t Q_S \pi_S Q^{-1} = \pi_S^t (\pi_S Q^{-1} \pi_S^t)^{-1} \pi_S Q^{-1}. \quad (12.10)$$

For the full subset $S = \{1, \dots, q\}$, show that indeed $G_{\text{wide}} = I_q$, the $q \times q$ identity matrix; for $S = \emptyset$ we define $G_{\text{narr}} = 0$. Show that $\text{Tr}(G_S) = |S|$. When Q happens to be diagonal, show that Q_S is just the appropriate $|S| \times |S|$ submatrix of Q .

(b) A useful exercise, to get acquainted with the book-keeping of subset models, is as follows. Generate any covariance matrix J of size 6×6 , thought about as via θ and γ each of length 3. For the $2^3 = 8$ submodels S of $\{1, 2, 3\}$, pushing $\gamma_1, \gamma_2, \gamma_3$ in and out, compute the G_S matrices, and verify $\text{Tr}(G_S) = |S|$.

(c) Use arguments similar to those seen in earlier exercises for narrow and wide model estimators to show that

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_S - \theta_0) \\ \sqrt{n}(\hat{\gamma}_S - \gamma_{0,S}) \end{pmatrix} = J_S^{-1} \begin{pmatrix} \sqrt{n}\bar{U}_{0,n} \\ \sqrt{n}\bar{U}_{1,n,S} \end{pmatrix} + o_{\text{Pr}}(1) \rightarrow_d \begin{pmatrix} C_S \\ D_S \end{pmatrix} = J_S^{-1} \begin{pmatrix} J_{01}\delta + U'_0 \\ \pi_S J_{11}\delta + U'_{1,S} \end{pmatrix},$$

in which (U'_0, U'_1) has the distribution $N_{p+q}(0, J)$. In particular, all these limit distributions are linear functions of the same $(p + q)$ -dimensional multinormal and are hence multivariate normal.

(d) We are now in a position to work with the submodel estimators $\hat{\mu}_S$ for the focus parameter, the true value of which is $\mu_{\text{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. Work with

$$\hat{\mu}_S - \mu_{\text{true}} = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c}) - \mu(\theta_0, \gamma_{0,S}, \gamma_{0,S^c}) + \mu(\theta_0, \gamma_0) - \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$$

to see that the limit distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ can be written

$$\Lambda_S = (\partial\mu/\partial\theta)^t C_S + (\partial\mu/\partial\gamma_S)^t D_S - (\partial\mu/\partial\gamma)^t \delta.$$

Now introduce

$$\Lambda_0 = (\frac{\partial\mu}{\partial\theta})^t J_{00}^{-1} U'_0 \quad \text{and} \quad D = \delta + Q(U'_1 - J_{10} J_{00}^{-1} U'_0).$$

Show that these two are independent, with distributions $N(0, \tau_0^2)$ and $N_q(\delta, Q)$. (xx one more sentence re D , same as limit of $\sqrt{n}(\hat{\gamma} - \gamma_0)$ as seen in (12.7). xx) Then carry out the somewhat laborious algebraic reformulations to verify that

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \rightarrow_d \Lambda_S = \Lambda_0 + \omega^t(\delta - G_S D). \quad (12.11)$$

(e) We have started with a quite general situation, any smooth regular model $f(y, \theta, \gamma)$, and found that there is a simply structured Limit Experiment, characterised by the two independent random variables Λ_0 and D . In this Limit Experiment, essentially all parameters are known (i.e. well estimated, with high precision, for large n), except for the single unknown δ , for which we observe is $D \sim N_q(\delta, Q)$. The object, in this Limit Experiment, is to estimate the linear combination $\psi = \omega^t \delta$. Show that Λ_S is normal, find its mean and variance, and establish that its mse is

$$\text{mse}_S = E \Lambda_S^2 = \tau_0^2 + \omega^t G_S Q G_S^t \omega + \{\omega^t (I_q - G_S) \delta\}^2.$$

For the two extreme cases verify that we recover results of Ex. 12.20, that $\text{mse}_{\text{narr}} = \tau_0^2 + (\omega^t \delta)^2$ and $\text{mse}_{\text{wide}} = \tau_0^2 + \omega^t Q \omega$.

(f) How correlated are different submodel estimators, for the same estimand? For two subsets S and T of $\{1, \dots, q\}$, explain that the Master Theorem result (12.11) holds jointly, so there is a binormal limit for $\hat{\mu}_S, \hat{\mu}_T$. Show that in the Limit Experiment, the correlation is

$$\rho(S, T) = \frac{\tau_0^2 + \omega^t G_S Q G_T^t \omega}{(\text{mse}_S)^{1/2} (\text{mse}_T)^{1/2}}.$$

In particular, the limit correlation between estimators from narrow and wide models is

$$\frac{\tau_0^2}{\{\tau_0^2 + (\omega^t \delta)^2\}^{1/2} (\tau_0^2 + \omega^t Q \omega)^{1/2}} = \frac{1}{\{1 + (\omega^t \delta)^2 / \tau_0^2\}^{1/2} (1 + \omega^t Q \omega / \tau_0^2)^{1/2}}.$$

We learn that the size of $\lambda = \omega^t Q \omega / \tau_0^2$ matters significantly. If this ratio is small, then the extra variability introduced by further unknown parameters is small compared to the variance of the narrow model, most correlations might be high, and model selection might not be fruitful. If the ratio is moderate or big, however, there is more extra room and benefit for cleverness in model selection, as finding a clever S , for the given occasion of the $\mu(\theta, \gamma)$, might be significantly better than competitors.

The FIC, the Focused Information Criterion

Ex. 12.25 *The FIC in the Limit Experiment.* We may derive our first Focused Information Criterion for the Limit Experiment, described and discussed in Ex. 12.24. It consists in estimating the limiting risk mse_S found there, in the situation where J and the partial derivatives of $\mu(\theta, \gamma)$ are known, hence also τ_0, ω, Q, G_S , and with only δ unknown, for which we observe $D \sim N_q(\delta, Q)$. The Real FIC, so to speak, for a given dataset and candidate models, then emerges by estimating all unknowns; see Ex. 12.26.

(a) Via Ex. 12.16, explain that the natural risk function estimator is

$$\text{FIC}_S = \tau_0^2 + \omega^t G_S Q G_S^t \omega + \max(\omega^t (I_q - G_S) (D D^t - Q) (I_q - G_S)^t \omega, 0). \quad (12.12)$$

We are essentially taking the unbiased estimator of the squared bias term and truncating this to zero if that estimator turns out negative. There are 2^q submodels, each with its FIC score, and the winning model is the one with lowest FIC_S . (xx sentence on bias and variance balance. slim S : first small, second big; fatter S : first big, second small. xx) Note that the τ_0^2 term is common to all, and might be ignored, but it useful to retain it in order for $\text{FIC}_S^{1/2}$ to have the interpretation of estimated rmse. For the two extreme choices, show that

$$\begin{aligned} \text{FIC}_{\text{narr}} &= \tau_0^2 + \max((\omega^t D)^2 - \omega^t Q \omega, 0), \\ \text{FIC}_{\text{wide}} &= \tau_0^2 + \omega^t Q \omega. \end{aligned}$$

Explain that FIC prefers the narrow to the wide model when $(\omega^t D)^2 \leq 2 \omega^t Q \omega$, i.e. when $|T| \leq \sqrt{2}$, where $T = \omega^t D / (\omega^t Q \omega)^{1/2}$, which is normal with variance 1. When the narrow model is correct, show that the probability of it being selected by the FIC is 0.843.

(b) Consider the case where the extension parameter γ is one-dimensional, where the choice is between the p -parameter narrow model and the wide $(p+1)$ -parameter model. Assume ω of (12.9) is nonzero, show that it cancels out for the FIC comparison, and that the narrow model wins provided $\max(D^2 - \kappa^2, 0) \leq \kappa^2$, or $\max(D^2/\kappa^2 - 1, 0) \leq 1$ where we as earlier write $Q = \kappa^2$. Explain that this is the same as $D^2/\kappa^2 \leq 2$, and that this is actually precisely the same as with the AIC; see Ex. 12.23. So FIC is only useful on its own for $q \geq 2$, where different foci are associated with different ω , and these do not cancel out.

(c) (xx one or two more things, inside Limit Experiment. xx)

Ex. 12.26 *The Real FIC.* Turning to The Real FIC, assume we have data $Y_1, \dots, Y_n$ from the model $f(y, \theta, \gamma)$, with the wide model envisaged to be in the vicinity of the narrow model where $\gamma = \gamma_0$. Then letting $\hat{J}$ be the observed normalised Fisher informaton matrix, compute from this $\hat{Q}$ and all $\hat{G}_S$ via (12.10), along with $\hat{\tau}_0$ and $\hat{\omega}$ via the estimated partial derivatives of our focus $\mu(\theta, \gamma)$. We also need $D_n = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0)$, discussed earlier, as with (12.7).

(a) Explain that the empirical version of the Limit Experiment version above becomes

$$\text{FIC}_S = \hat{\tau}_0^2 + \hat{\omega}^t \hat{G}_S \hat{Q} \hat{G}_S^t \hat{\omega} + \max(\hat{\omega}^t (I_q - \hat{G}_S) (D_n D_n^t - \hat{Q}) (I_q - \hat{G}_S)^t \hat{\omega}, 0). \quad (12.13)$$

This is the real FIC for real datasets. It is to be computed for all relevant candidate models S , and the FIC score ranks them from the smallest (best) to the biggest (worst).

(b) This FIC_S relies on consistent estimators for the population quantities $\partial\mu(\theta_0, \gamma_0)/\partial\theta$, $\partial\mu(\theta_0, \gamma_0)/\partial\gamma$, and those coming from the Fisher information matrix $J = J(\theta_0, \gamma_0)$, i.e. τ_0 ,

ω , Q , G_S . Explain that the FIC will work whether one uses ‘narrow’ or ‘wide’ estimation for the focus derivatives and for J . FICologists often prefer using estimators coming from the wide model, as this tends to make inferences more robust. (xx but if formulae are very simple for narrow, etc. xx)

(c) (xx a simple illustration. to be used in one or two stories, perhaps one iid and one regression, and perhaps alongside AIC. xx)

Ex. 12.27 *The AFIC.* (xx spell out. basics, an illustration. and in regression, below. xx) The FIC has been built around a single focus parameter μ_μ , with formulae above being extended in the next subsection for various practical situations. Sometimes there is not such a one-and-only μ of primary interest, however, but perhaps a slim collection of μ_j , with $j = 1, \dots, m$, a finite set, or an index representing a stretch of parameter values. Here we develop the essence of the AFIC, the *Average-weighted Focus Information Criterion*. For simplicity we develop results below in terms of a finite index set.

(a) Explain that the ML estimator for μ_α is $\hat{\mu}_\alpha = \mu_\alpha(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$, and that previous master theorem results imply

$$\sqrt{n}(\hat{\mu}_{S,j} - \mu_{\text{true},j}) \rightarrow_d \Lambda_{S,j} = \Lambda_{0,j} + \omega_j^t(\delta - G_S D),$$

with $\omega_j = J_{10} J_{00}^{-1} \frac{\partial \mu_j}{\partial \theta} - \frac{\partial \mu_j}{\partial \gamma}$, and otherwise the same G_S and D as above; see Ex. 12.9 and 12.24. Also, $\tau_{0,j}^2 = (\frac{\partial \mu_j}{\partial \theta})^t J_{00}^{-1} (\frac{\partial \mu_j}{\partial \theta})$, not depending on submodel S .

(b) For any weight-of-importance scores v_j over the index set, show that

$$n \sum_j v_j (\hat{\mu}_{S,j} - \mu_{\text{true},j})^2 \rightarrow_d \sum_j v_j [\Lambda_{0,j} + \omega_j^t(\delta - G_S D)]^2.$$

with this limit variable having expected value

$$\sum_j v_j \tau_{0,j}^2 + \sum_j v_j [\omega_j^t G_S Q G_S^t \omega_j + \{\omega_j^t (I_q - G_S) \delta\}^2].$$

(c) When searching for a good model S , explain that we may ignore the first term here. Show that an unbiased estimator for the second term, in the limit experiment, is

$$\begin{aligned} \text{risk}_S &= \sum_j v_j [\omega_j^t G_S Q G_S^t \omega_j + \omega_j^t (I_q - G_S) (D D^t - Q) (I_q - G_S)^t \omega_j] \\ &= \text{Tr}(G_S Q G_S^t A) + \text{Tr}\{(I_q - G_S) (D D^t - Q) (I_q - G_S) A\}, \end{aligned}$$

in terms of $A = \sum_j v_j \omega_j \omega_j^t$. The second term here is essentially unbiased for a squared bias, so we truncate it to zero in case of a negative value. The empirical version of the AFIC becomes

$$\text{AFIC}_S = \text{Tr}(\hat{G}_S \hat{Q} \hat{G}_S^t \hat{A}) + \max[0, \text{Tr}\{(I_q - \hat{G}_S) (D_n D_n^t - \hat{Q}) (I_q - \hat{G}_S) \hat{A}\}],$$

with $\hat{A} = \sum_j v_j \hat{\omega}_j \hat{\omega}_j^t$, and $D_n = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0)$, as above.

ML behaviour, tolerance, FIC, model averaging, in regression models

Ex. 12.28 *ML behaviour under local alternatives, in regression models.* We have seen in various exercises in Ch. 5 that the basic theory for ML behaviour, with limiting normality, Wilks theorems, etc., could be lifted from the i.i.d. context to general regression models.

Basically, in smooth regression models, (i) ML estimators still have clear definitions, as maximisers of log-likelihood functions; (ii) behaviour under model conditions is partly dictated by the normalised observed Fisher information matrix $\hat{J}_n = -n^{-1} \partial^2 \ell_n(\hat{\theta}) / \partial \theta \partial \theta^t$, and there is a well-defined limit in probability J ; (ii) there is large-sample normality and a chi-squared for deviances with Wilks type statistics. The present point is that importantly, theory developed above for local alternatives goes through when lifted to regression models. This concerns in particular Ex. 12.20, 12.23, 12.24. Essentially, this involves careful use of Lindeberg theorems, with care for the biases implied by data mechanisms being $O(1/\sqrt{n})$ away from the models used.

To look into the basic mechanisms underlying these ‘methods and results hold up when lifting from i.i.d. to regression’ statements, consider a model $f(y_i | x_i, \theta, \gamma)$ for Y_i given x_i , with θ and γ of dimensions p and q , in generalisation of the setup of Ex. 5.42. As above, in the development that led to both tolerance radii and the FIC, there is a narrow model with p parameters, corresponding to a null value $\gamma = \gamma_0$, and a wide model, where all of $\gamma_1, \dots, \gamma_q$ are free. We need the score functions $u_0(y_i | x_i) = \partial \log f(y_i | x_i, \theta_0, \gamma_0) / \partial \theta$ and $u_1(y_i | x_i) = \partial \log f(y_i | x_i, \theta_0, \gamma_0) / \partial \gamma$, computed at the null position. With a dataset of (x_i, Y_i) for $i = 1, \dots, n$, where the Y_i are independent given the x_i , there is a log-likelihood function $\ell_n(\theta, \gamma) = \sum_{i=1}^n \log f(y_i | x_i, \theta, \gamma)$. Writing $\alpha = (\theta, \gamma)$ for the full parameter, there is a normalised Fisher information matrix

$$J_n = -n^{-1} \frac{\partial \ell_n(\theta_0, \gamma_0)}{\partial \alpha \partial \alpha^t} = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix} \text{ with inverse } J_n^{-1} = \begin{pmatrix} J^{n,00} & J^{n,01} \\ J^{n,10} & Q_n \end{pmatrix},$$

of size $(p+q) \times (p+q)$, computed at the null value, assumed to be an inner point in the parameter region for the model. We postulate that Lindeberg regularity holds, including a clear positive definite limit in probability matrix $J_n \rightarrow_{\text{pr}} J$.

(a) Now examine the local neighbourhood situation where the real mechanism is $f_{i,\text{true}}(y_i | x_i) = f(y_i | x_i, \theta_0, \gamma_0 + \delta/\sqrt{n})$. From the log-likelihood function we form versions for the narrow and wide model, leading to ML estimators $\hat{\theta}_{\text{narr}}$ and $(\hat{\theta}_{\text{wide}}, \hat{\gamma}_{\text{wide}})$. We need

$$\sqrt{n} \bar{U}_0 = n^{-1/2} \sum_{i=1}^n u_0(Y_i | x_i) \quad \text{and} \quad \sqrt{n} \bar{U}_1 = n^{-1/2} \sum_{i=1}^n u_1(Y_i | x_i).$$

Use Taylor expansion techniques, or generalisation of arguments used in Ex. 5.42, to show that

$$\sqrt{n}(\hat{\theta}_{\text{narr}} - \theta_0) = J_{n,00}^{-1} \sqrt{n} \bar{U}_{n,0} + o_{\text{pr}}(1) \rightarrow_d J_{00}^{-1} U_0 \sim N_p(J_{00}^{-1} J_{01} \delta, J_{00}^{-1}).$$

and that

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_{\text{wide}} - \theta_0) \\ \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0) \end{pmatrix} = J_n^{-1} \begin{pmatrix} U_{n,0} \\ U_{n,1} \end{pmatrix} + o_{\text{pr}}(1) \rightarrow_d J^{-1} \begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim N_{p+q} \left(\begin{pmatrix} 0 \\ \delta \end{pmatrix}, J^{-1} \right).$$

Note also that we as in the i.i.d. case have $D_n = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0) \rightarrow_d D \sim N_q(\delta, Q)$

(b) Having established these basic results, for smooth regression models under Lindeberg conditions, verify that also other results reached for i.i.d. setups in Ex. 12.20, 12.23, 12.24 continue to hold in the wider regression context. This in particular holds for the Master Theorem of Ex. 12.24, with a clear limit distribution $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \rightarrow_d \Lambda_0 + \omega^t(\delta - G_S D)$. When using this to form FIC in regression models, as with (12.13), it is sensible to see the ingredients $\hat{Q}_n$ and $\hat{G}_S$ as the natural estimators of quantities Q_n and $G_{n,S}$, associated

with J_m , for the given n of the dataset being analysed. A similar comment applies to $\hat{\tau}_{0,n}$ and $\hat{\omega}_n$, estimates of $\tau_{0,n}$ and $\omega_n = J_{10,n} J_{n,00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$.

Ex. 12.29 *Tolerance radii in regression models.* We have learned in Ex. 12.28 that the basic limiting risk expressions hold for mse_{narr} and mse_{wide} , associated with estimating a focus parameter via the narrow and the wide models. This leads to the same tolerated radius expressions as in the i.i.d. case. In particular, if γ is one-dimensional, representing one extra parameter extension of some regression model, then narrow based estimation is better than wide model estimation provided $|\gamma - \gamma_0| \leq \kappa_n / \sqrt{n}$, with $Q_n = \kappa_n^2$ the lower-right element in the $(p+1) \times (p+1)$ matrix J_n^{-1} .

(a) For a simple linear regression dataset of (x_i, y_i) , with regression line $a + b(x_i - \bar{x})$, as in Ex. 3.32, might it be worth it to include a quadratic term $c(x_i - \bar{x})^2$? Show that this wider regression model, with three parameters for the mean in addition to the normal scale σ , has normalised Fisher information matrix $J_n(\sigma, a, b, c) = (1/\sigma^2) \text{diag}(2, \Sigma_n)$, where

$$\Sigma_n = n^{-1} \sum_{i=1}^n \begin{pmatrix} 1 \\ x_i - \bar{x} \\ (x_i - \bar{x})^2 \end{pmatrix} \begin{pmatrix} 1 \\ x_i - \bar{x} \\ (x_i - \bar{x})^2 \end{pmatrix}^t = \begin{pmatrix} 1 & 0 & M_n \\ 0 & M_n & T_n \\ M_n & T_n & F_n \end{pmatrix},$$

in terms of covariate moments $M_n = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$, $T_n = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^3$, $F_n = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^4$. With E_n the $(3,3)$ element in Σ_n^{-1} , show that simple linear regression is better than quadratic regression provided $|c| \leq E_n^{1/2} \sigma / \sqrt{n}$.

(b) More generally, consider ordinary linear regression, with covariates $x_1, \dots, x_p$. Is it worth it, including also an extra x_{p+1} ? Let Σ_n be the $(p+1) \times (p+1)$ matrix $n^{-1} \sum_{i=1}^n x_{i,\text{plus}} x_{i,\text{plus}}^t$, with $x_{i,\text{plus}}$ being $(x_{i,1}, \dots, x_{i,p}, x_{i,p+1})^t$. Show that the tolerance radius for the extra β_{p+1} is $|\beta_{p+1}| \leq E_n^{1/2} \sigma / \sqrt{n}$, now with E_n being the lower-right element of Σ_n^{-1} . We learn that the tolerance level depends on scale σ , the actual covariates on board, and the sample size.

(c) For the general linear regression model of Ex. 3.33, with $y_i \sim N(x_i^t \beta, \sigma^2)$, consider the modelling option of introducing heterogeneous variance, as worked with in Ex. 5.49. Specifically, suppose there is a covariate w_i , normalised to have average zero and $n^{-1} \sum_{i=1}^n w_i^2 = 1$, for ease of interpretation, where the wider model takes $\sigma_i = \exp(\gamma w_i)$; the w_i could e.g. be one of x_i covariates, normalised. With $(\hat{\beta}, \hat{\sigma}, \hat{\gamma})$ the ML estimators, show via ML theory of Ch. 5 that $\sqrt{n}(\hat{\gamma} - \gamma) \rightarrow_d N(0, 1)$. Show that the tolerance radius becomes $|\gamma| \leq 1/\sqrt{n}$.

Ex. 12.30 *FIC selection of covariates in linear regression.* Consider a linear regression setup with covariates $x_1, \dots, x_p$, deemed important and protected, along with $z_1, \dots, z_q$, seen as potentially fruitful but needing care and checks for possible inclusion. The wide model is $Y_i \sim N(x_i^t \beta + z_i^t \gamma, \sigma^2)$, the narrow model has only the $x_i^t \beta$ term, and there are 2^q candidate submodels corresponding to pushing the $z_1, \dots, z_q$ in and out. Thus submodel S has mean of the form $x_i^t \beta + z_{i,S}^t \gamma_S$.

(a) Show via the general linear regression theory of Ex. 3.33 that ML estimation in the wide model yields least squares estimators

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n \begin{pmatrix} x_i y_i \\ z_i y_i \end{pmatrix}, \quad \text{where } \Sigma_n = n^{-1} \sum_{i=1}^n \begin{pmatrix} x_i \\ z_i \end{pmatrix} \begin{pmatrix} x_i \\ z_i \end{pmatrix}^t,$$

along with $\hat{\sigma}^2 = R_0/n$, say, with $R_0 = \sum_{i=1}^n (y_i - x_i^t \hat{\beta} - z_i^t \hat{\gamma})^2$.

check Story
#16 for similar
FIC work for
logistic
regressions

(b) Show that the Fisher matrix $J_n(\sigma, \beta, \gamma)$, of size $(p + q + 1) \times (p + q + 1)$, is

$$J_n = (1/\sigma^2) \begin{pmatrix} 2 & 0 & 0 \\ 0 & \Sigma_{n,00} & \Sigma_{n,01} \\ 0 & \Sigma_{n,10} & \Sigma_{n,11} \end{pmatrix}, \quad \text{with inverse } J_n^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \Sigma_n^{00} & \Sigma_n^{10} \\ 0 & \Sigma_n^{10} & \Sigma_n^{11} \end{pmatrix},$$

featuring the blocks of Σ_n and Σ_n^{-1} .

(c) Following the general recipe of Ex. 12.28, show that the crucial Q matrix becomes $\hat{Q}_n = \hat{\sigma}^2 \Sigma_n^{11}$. Show next that when finding the $G_{n,S}$ matrices from (12.10), the σ^2 cancels out, leading to $\Sigma_{n,S} = \pi_S \Sigma_n \pi_S^t$ and

$$G_{n,S} = \pi_S^t \hat{Q}_S \pi_S \hat{Q}^{-1} = \pi_S^t (\pi_S \hat{Q}^{-1} \pi_S^t)^{-1} \pi_S \hat{Q}^{-1} = \pi_S^t (\pi_S (\Sigma_n^{11})^{-1} \pi_S^t)^{-1} \pi_S (\Sigma_n^{11})^{-1}.$$

Again, $G_{n,\emptyset} = 0$, $G_{n,\text{wide}} = I_q$, and $\text{Tr}(G_{n,S}) = |S|$.

(d) The task is now to find good models for estimating $\mu = E(Y | x_0, z_0) = x_0^t \beta + z_0^t \gamma$, for given x_0, z_0 . Show that $\omega = \omega(x_0, z_0) = \Sigma_{n,10} \Sigma_{n,00}^{-1} x_0 - z_0$, free of unknown parameters. In the general FIC notation introduced above, show that $\tau_{n,0}^2 = \sigma^2 x_0^t \Sigma_{n,00}^{-1} x_0$, that $D_n = \sqrt{n} \hat{\gamma}$, and that the FIC score for submodel S can be expressed as

$$\begin{aligned} \text{FIC}_S &= \hat{\tau}_{n,0}^2 + \omega^t G_{n,S} \hat{Q}_n G_{n,S}^t \omega + \max\{\omega^t (I_q - G_{n,S}) (D_n D_n^t - \hat{Q}_n) (I_q - G_{n,S})^t \omega, 0\} \\ &= \hat{\sigma}^2 [x_0^t \Sigma_{n,00}^{-1} x_0 + \omega^t G_{n,S} \Sigma_n^{11} G_{n,S}^t \omega \\ &\quad + \max\{\omega^t (I_q - G_{n,S}) (n \hat{\gamma} \hat{\gamma}^t / \hat{\sigma}^2 - \Sigma_n^{11}) (I_q - G_{n,S})^t \omega, 0\}]. \end{aligned}$$

Point to the parts relating to estimating respectively variance and squared bias. For an application of these methods, see Story #17.

Ex. 12.31 *Master Theorem for post-selection and model averaging estimators.* (xx do and polish. xx) Consider again the general setup of Ex. 12.23–12.24, involving a wide model $f(y, \theta, \gamma)$ of dimension $p + q$, inside which is the narrow model $f_0(y, \theta) = f(y, \theta, \gamma_0)$ with a null value $\gamma = \gamma_0$. We have seen the Master Theorem spelling out the large-sample behaviour for all subset model estimators $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$. Indeed, under $\gamma = \gamma_0 + \delta/\sqrt{n}$ circumstances, $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \rightarrow_d \Lambda_0 + \omega^t(\delta - G_S D)$. One crucial component here is the limit variable of $D_n = \sqrt{n}(\hat{\gamma}_{\text{wide}} - \gamma_0)$, namely $D \sim N_q(\delta, Q)$. Note that this convergence also holds jointly, for the collection of all the 2^q subsets S of $\{1, \dots, q\}$. Now we are widening the class of estimators significantly, forming in effect model averaging estimators

$$\mu^* = \sum_S w(S | D_n) \hat{\mu}_S.$$

We require that these potentially data-dependent weights sum to 1 and have at most a finite number of discontinuities in D_n . Thus included are post-selection estimators, for the AIC and the FIC and other regimes, with weight 1 given to the winning model and 0 to the losers.

(a) Demonstrate the Master Theorem for model averaging estimators, under local alternatives, that

$$\sqrt{n}(\hat{\mu}^* - \mu_{\text{true}}) \rightarrow_d \Lambda^* = \Lambda_0 + \omega^t\{\delta - \delta^*(D)\}, \quad (12.14)$$

where $\delta^*(D) = \sum_S w(S | D) G_S D$. Explain that the previous Master Theorem of (12.11) is a special case. Show also that the same limit obtains more generally, if the weights are of the less strict form of $\hat{w}_n(S)$, as long as $\hat{w}_n(S) - w(S | D_n) \rightarrow_{\text{pr}} 0$ for each S . (xx the post-AIC and post-FIC are of this form; do the details. xx)

(b) Note that the limit Λ^* is not normal, but rather a nonlinear mixture of normals (unless the weights are fixed). Using that $\Lambda^* | D$ is normal, explain that

$$\Pr(\Lambda^* \leq u) = \int \Phi\left(\frac{u - \omega^t\{\delta - \delta^*(x)\}}{\tau_0}\right) \phi_q(x - \delta, Q) dx,$$

with $\phi_q(x, Q)$ the multinormal $N_q(0, Q)$ density. The Λ^* distribution can be multimodal and skewed, depending on the nonlinearities of the weights.

(c) Suppose we only care about the narrow and wide models, and interpolate between these, using $\mu^* = (1 - \hat{h})\hat{\mu}_{\text{narr}} + \hat{h}\hat{\mu}_{\text{wide}}$, for some suitable $\hat{h}$, with the property that $\hat{h} - h(D_n) \rightarrow_{\text{pr}} 0$ for a certain associated $h(D) \in [0, 1]$. Demonstrate that

$$\sqrt{n}(\mu^* - \mu_{\text{true}}) \rightarrow_d \Lambda_0 + \omega^t\{\delta - h(D)D\}.$$

Show that the FIC selector corresponds to setting $\hat{h} = 1$ when $(\hat{\omega}^t D_n)^2 \geq 2\hat{\omega}^t \hat{Q} \hat{\omega}$ and 0 otherwise, and that $\hat{h} - h(D_n) \rightarrow_0$, with $h(D)$ indicator for $(\omega^t D)^2 \geq 2\omega^t Q \omega$.

(d) Studying $\mu^* = (1 - h)\hat{\mu}_{\text{narr}} + h\hat{\mu}_{\text{wide}}$, for fixed h , show that the theoretically optimal h is $h_0 = (\omega^t \delta)^2 / \{(\omega^t \delta)^2 + \omega^t Q \omega\}$, in the sense of minimising the limiting risk. This invites using

$$\hat{h} = \frac{(\hat{\omega}^t D_n)^2}{(\hat{\omega}^t D_n)^2 + \hat{\omega}^t \hat{Q} \hat{\omega}}.$$

Find the limit distribution of $\sqrt{n}(\mu^* - \mu)$. (xx one more, smooth function of fic scores, spell out, give risk functions. convey the insight that different ideas may be compared clearly and informatively with risk functions in the Limit Experiment. xx) something like this: let $T_n = \frac{1}{2}(\hat{\omega}^t D_n)^2 / \hat{\omega}^t \hat{Q} \hat{\omega} - 1$. fic says wide if $T_n \geq 0$, narrow if $T_n < 0$. make it smooth:

$$\hat{h} = \exp(\lambda T_n) / \{1 + \exp(\lambda T_n)\},$$

for suitable fine-tuning parameter λ . a large λ means FIC. Again, find the limit distribution of $\sqrt{n}(\mu^* - \mu_{\text{true}})$. Draw risk functions, for the case of $q = 1$.

(e) (xx finding theoretically optimal weights, from the limit experiment, then estimating these, inserting D for δ . xx)

Ex. 12.32 *Hybrid estimators: smoothing between narrow and wide.* (xx nils rant; to be polished. may place a bit of the risk stuff in Ch8. xx) Starting with the estimators $\hat{\mu}_{\text{narr}} = \mu(\hat{\theta}_{\text{narr}}, \gamma_0)$ and $\hat{\mu}_{\text{wide}} = \mu(\hat{\theta}_{\text{wide}}, \hat{\gamma}_{\text{wide}})$, we may form a version linearly balancing these two,

$$\hat{\mu} = (1 - a)\hat{\mu}_{\text{narr}} + a\hat{\mu}_{\text{wide}},$$

with a a balance parameter. With the framework above (xx pointer xx), show that

$$\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \rightarrow_d \Lambda_0 + (1 - a)\omega^t \delta + a\omega^t(\delta - D) = \Lambda_0 + \omega^t \delta - a\omega^t D.$$

For this limit distribution, find the mean and variance, and establish that

$$\text{mse}(a) = (1 - a)^2(\omega^t \delta)^2 + a^2\omega^t Q a.$$

This is minimised at

$$a_0 = \frac{\omega^t \delta)^2}{(\omega^t \delta)^2 + \omega^t Q \omega} = \frac{\rho}{\rho + 1}, \quad \text{with } \rho = (\omega^t \delta)^2 / \omega^t Q \omega,$$

and with risk minimum

$$\text{mse}(a_0) = \frac{(\omega^t \delta)^2 \omega^t Q \omega}{(\omega^t \delta)^2 + \omega^t Q \omega},$$

a number smaller than either of the two risks $\text{mse}(0) = (\omega^t \delta)^2$ and $\text{mse}(1) = \omega^t Q \omega$. note that when the model extension parameter γ is one-dimensional, the ω cancels out, leading to

$$a_0 = \delta^2 / (\delta^2 + Q) = \rho / (\rho + 1) \quad \text{with } \rho = \delta^2 / Q,$$

and risk minimum $\text{mse}(a_0) = \delta^2 Q / (\delta^2 + Q)$. then a separate point regarding the estimation of ρ , and of $\rho / (\rho + 1)$, based on $D \sim N_q(\delta, Q)$. on the scale of $\psi = \omega^t \delta$, this means $\hat{\psi} = \omega^t D \sim N(\psi, \tau^2)$, with known $\tau^2 = \omega^t Q \omega$. prototype version of this: with $\hat{\psi} \sim N(\psi, \tau^2)$, how to estimate not merely ψ^2 well, but more specifically $\rho / (\rho + 1) = 1 - 1 / (\rho + 1)$, with $\rho = \psi^2 / \tau^2$. may take $\tau = 1$ to understand different options and their risk functions: $X \sim N(\theta, 1)$, how to estimate $1 / (1 + \theta^2)$ well under squared loss. what is Bayes solution under flat prior? how does $1 / (1 + X^2)$ do, $\min(1 / X^2, 1)$? since we like $\max(X^2 - 1, 0)$ for θ^2 , may take

$$\frac{1}{1 + \max(X^2 - 1, 0)}$$

after this, with a sensible but not unique $\hat{a}$, we analyse $\hat{\mu}_{\text{final}} = \hat{\mu}(\hat{a})$, for which we have clear theory.

Ex. 12.33 *Selecting regression model with variance heterogeneity.* (xx finding an illustration. $\sigma_i = \sigma \exp(\gamma w_i)$ part of wide model. different foci. xx)

Ex. 12.34 *ML behaviour under local alternatives, III.* (xx to come here, nils hoping to make it not long and not complicated. to be used in Ch7 and some stories. starting point: $y \sim N(\theta, 1)$, where we know $\theta \geq 0$. then ML is $\hat{\theta} = y I(y \geq 0)$, and with a positive probability of being zero. then $\sqrt{n}(\bar{y}_n - \delta / \sqrt{n})$ in the case of $\theta = \delta / \sqrt{n}$. then general ML and profiled, aiming for $D_n = \sqrt{n}(\hat{\gamma} - \gamma_0)$ when $\gamma = \gamma_0 + \delta / \sqrt{n}$ and we have $\gamma \geq \gamma_0$ a priori. consequences also for log-likelihood-ratio. xx)

Notes and pointers

well:

mention [Sala-i Martin \(1997\)](#), ‘I just ran two million regressions’. point to review paper [Claeskens and Hjort \(2027\)](#).

(xx mention the Nils point: a good real BIC should use real coherent priors for the candidate models, so that $\mu(\theta_j)$ has the same prior across models, when $\mu(\theta_j)$ is a focus parameter. xx)

The focused view, for data planning, model selection, and inference, aligns with C. Darwin’s view: “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.” (xx Mozart, Figaro. xx)

do AFIC too. and a bit more on model averaging. include one model averaging thing in Stories.

Schwarz 1978 and others for BIC, Rissanen. [Claeskens and Hjort \(2008b\)](#) for AIC, for BIC, and in particular for FIC, AFIC, master theorems I and II. also [Claeskens and Hjort \(2003\)](#), [Hjort and Claeskens \(2003a,b\)](#).

cross validation. Mention generalisation of model averaging to bagging, in [Hjort \(2014\)](#).

for tolerance discussion: include the limit probability thing with 0.157.

point to [Jullum and Hjort \(2017\)](#); [Claeskens et al. \(2019\)](#); [Cunen et al. \(2020c\)](#); [Dæhlen et al. \(2024\)](#) for FIC Generation Two.

include brief exercise with fic based on two other loss functions, the 0-1 one and the skewed exp'd one.

be more consistent with $\hat{\theta}$ and $\hat{\gamma}$ being used for $\hat{\theta}_{\text{wide}}$ and $\hat{\gamma}_{\text{wide}}$.

Markov chains and Markov processes

For Markov chains, tomorrow's outcome depends on today's. This leads to crucially important dependence models, used broadly in probability theory and statistics. We go through the key issues regarding first visiting times, recurrence, transience, and stationary distributions; these are used, fundamentally, for the rich world of MCMC, Markov Chain Monte Carlo, including posterior sampling for the many Bayesian setups of Ch. 7. We also go into likelihood analyses for such processes, with various applications in our Statistical Stories. When Markov chains are studied in finer time scales, some key processes converge in distribution to some of those worked with in Ch. 10; in particular, the Ornstein–Uhlenbeck process is seen to be a limit of the properly scaled Ehrenfest model, as seen with tools of this chapter in Story #14.

Key words: first visiting time, generating functions, Markov chains, MCMC, ML for Markov chains, random walks, recurrence, stationary distribution

Consider a stochastic process $X_0, X_1, \dots$ on the finite state space $\{1, \dots, k\}$, with the basic property that

$$\Pr(X_{n+1} = j \mid X_0 = i_0, \dots, X_n = i_n) = p_{i_n, j} \quad \text{for } j = 1, \dots, k \quad (13.1)$$

depends on the past only via the current state $X_n = i_n$, and does not change over time. We call such a process a *stationary Markov chain*. This probabilistic aspect, that what happens next depends only on the present, is called the *Markov property*. The transition probabilities $p_{i,j} = \Pr(X_{n+1} = j \mid X_n = i)$ govern the process, and can be modelled, from prior knowledge concerning the process in question, often with parameters to be estimated from data. Instances and applications abound, as we shall see in this chapter and indeed in several stories, including #3, #14, #15, #25, #46, #61, #86.

Material covered here include stationary distributions, first visiting times, generating functions, recurrence, CLT for Markov chains, likelihood theory for estimation and testing. The theory is also extended to Markov chains with continuous outcomes. We have already learned about and seen the MCMC machinery in action in the Bayesian Ch. 7, and Ex. 13.19–13.21 explain and spell out the details.

Markov chains

Ex. 13.1 *Markov chains.* Consider a Markov chain, as in (13.1).

(a) For such a chain, define the transition probability matrix by

$$P = \begin{pmatrix} p_{1,1}, & \dots, & p_{1,k} \\ p_{2,1}, & \dots, & p_{2,k} \\ \vdots & & \\ p_{k,1}, & \dots, & p_{k,k} \end{pmatrix},$$

where $p_{i,j} = \Pr(X_{n+1} = j \mid X_n = i)$. Show that the probability distribution aspects of the chain are fully described by giving such a P and the initial state $X_0 = i_0$, by giving a formula for $\Pr(X_1 = i_1, \dots, X_n = i_n \mid X_0 = i_0)$.

(b) Explain that all rows of P sum to 1, and that 1 is an eigenvalue for P . Prove next, for the two-step transition probabilities, that

$$p_{i,j}^{(2)} = \Pr(X_{n+2} = j \mid X_n = i) = \sum_{r=1}^k p_{i,r} p_{r,j} \quad \text{for all } i, j.$$

Show that this perfectly matches matrix multiplication: the two-step probabilities $p_{i,j}^{(2)}$ are the elements of P^2 .

(c) Generalise to m -step transition probabilities, showing that the $p_{i,j}^{(m)} = \Pr(X_{n+m} = j \mid X_n = i)$ are the elements of P^m ; so indeed $P^{(m)} = P^m$. Also show that Chapman–Kolmogorov

$$p_{i,j}^{(m+n)} = \Pr(X_{n+m} = j \mid X_n = i) = \sum_{r=1}^k p_{i,r}^{(m)} p_{r,j}^{(n)} \quad \text{for all } i, j. \quad (13.2)$$

These are called the Chapman–Kolmogorov equations. Since transition probabilities across several steps match matrix multiplications, the result is merely the unsurprising $P^{m+n} = P^m P^n$.

Ex. 13.2 *The equilibrium distribution.* Consider a Markov chain $X_0, X_1, X_2, \dots$ on a finite state space $\{1, \dots, k\}$, with transition probability matrix $P = (p_{i,j})_{i,j=1,\dots,k}$. For simplicity here we assume all $p_{i,j}$ are positive, e.g. to avoid deterministic periodicity. Assume X_0 has a start distribution $(\pi_1, \dots, \pi_k)$ with the property that

$$\pi_j = \sum_{i=1}^k \pi_i p_{i,j} \quad \text{for all } j.$$

(a) Show that this requirement on $\pi = (\pi_1, \dots, \pi_k)$ can be expressed as $\pi P = \pi$. Explain that this also implies $\pi P^n = \pi$ for any $n \geq 1$, and that this means

$$\Pr(X_n = j) = \sum_{i=1}^k \Pr(X_0 = i, X_n = j) = \sum_{i=1}^k \pi_i p_{i,j}^{(n)} = \pi_j \quad \text{for } j = 1, \dots, k.$$

The $(\pi_1, \dots, \pi_k)$ is called the chain's equilibrium or stationary distribution. Explain that P^n tends to the matrix Q having these $(\pi_1, \dots, \pi_k)$ for each of its rows. A straightforward way of finding these probabilities, for a given Markov transition matrix P , is to ask your computer to compute P^n for $n = 2, 3, \dots$ and then read off any of its rows after numerical convergence. A different way is provided next; for an example to see how this pans out, do Ex. 13.4.

(b) Let $\lambda_1, \dots, \lambda_k$ be the eigenvalues of P , i.e. the solutions to the k -degree polynomial $\det(P - \lambda I) = 0$, with I the identity matrix of appropriate size $k \times k$. Show that 1 is an eigenvalue, and that no other eigenvalue can have $|\lambda_j| > 1$. We arrange the eigenvalues with $\lambda_1 = 1$ and $\lambda_2, \dots, \lambda_k$ of absolute value less than 1. From linear algebra, there is a representation $P = U\Lambda U^{-1}$, for the appropriate eigen-matrix U , where Λ is the diagonal matrix with $(\lambda_1, \dots, \lambda_k)$ on its diagonal, in decreasing order of magnitude. With G_j the matrix having 1 in position (j, j) and zeros elsewhere, show that

$$P^n = U\Lambda^n U^{-1} = \sum_{j=1}^k \lambda_j^n U G_j U^{-1} = U^{(1)} (U^{-1})_{(1)} + \sum_{j=2}^k \lambda_j^n U G_j U^{-1},$$

with $U^{(1)}$ and $(U^{-1})_{(1)}$ respectively column 1 and row 1 for the matrices U and U^{-1} . From this, explain that $U^{(1)}(U^{-1})_{(1)}$ must be equal to the limit matrix Q of P^n , with the stationary probabilities $(\pi_1, \dots, \pi_k)$ for each row. Show also that the convergence $P^n \rightarrow Q$ is exponentially fast, with $\max_{a,b} |p_{a,b}^{(n)} - \pi_b| \leq c\rho^n$ for all n , for $\rho = \max(|\lambda_2|, \dots, |\lambda_k|)$.

Ex. 13.3 *The two-state Markov chain.* Consider a Markov chain on the states $\{0, 1\}$, with transition probabilities $p_{0,1} = a$ and $p_{1,0} = b$.

(a) Show that the full 2×2 transition probability matrix must be

$$P = \begin{pmatrix} p_{0,0} & p_{0,1} \\ p_{1,0} & p_{1,1} \end{pmatrix} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix},$$

and give a formula for $p_{0,0}^{(2)} = \Pr(X_{n+2} = 0 \mid X_n = 0)$.

(b) Show that the equilibrium distribution becomes $(\pi_0, \pi_1) = (b/(a+b), a/(a+b))$. Show further that P has eigenvalues 1 and $\lambda = 1 - a - b$.

(c) Show that

$$P^n = \begin{pmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{pmatrix} + (1-a-b)^n \begin{pmatrix} \pi_1 & -\pi_1 \\ -\pi_0 & \pi_0 \end{pmatrix}.$$

If the chain starts out in its equilibrium distribution, show that $\text{cov}(X_0, X_n) = \pi_1 p_{1,1}^{(n)} - \pi_1^2 = \pi_0 \pi_1 (1-a-b)^n$. Note that the correlation tends to zero exponentially fast.

(d) In particular, for the symmetric Markov chain, with transition matrix

$$P = \begin{pmatrix} 1-a & a \\ a & 1-a \end{pmatrix},$$

show that $p_{1,1}^{(n)} = \frac{1}{2} + \frac{1}{2}(1-2a)^n$, and that $\text{cov}(X_m, X_{m+n}) = \frac{1}{4}(1-2a)^n$.

Ex. 13.4 *Social mobility.* A classic theme inside sociology, demography, economics concerns ‘social mobility’, from one generation to the next. Van Bavel et al. (2011) have counted and tabulated occurrences $N_{i,j}$, the number of father-son pairs in categories (i, j) , in Antwerpen, Belgium, 1860–1920, with the 1–5 categories being unskilled, semiskilled, skilled, middle class, elite.

from father	to son	1	2	3	4	5	total:
1 unskilled		108	39	26	30	2	205
2 semiskilled		36	47	26	32	2	143
3 skilled		14	38	29	36	4	121
4 middle class		32	31	34	81	5	183
5 elite		5	1	1	6	3	15
total:		195	155	116	185	16	668

for an
application,
check Story
#85

there is more in
Story #61

(a) Translate the information above to estimated transition probabilities $p_{i,j} = \Pr(Y = j | X = i)$, forming a Markov matrix P , with X for father and Y for son. If you are middle class, what's the probability that your grandchild also will also be middle class? And your grandchild's children?

(b) Find the equilibrium probabilities $\pi_1, \dots, \pi_5$, for this society, assuming that the transition probabilities remain unchanged over time. Use both methods outlined in Ex. 13.2; the answer should be 0.287, 0.236, 0.174, 0.279, 0.024.

(c) To explore this a little further, find eigenvalues numerically (they are 1.000, 0.351, 0.203, 0.132, 0.039), study the P^n representation of Ex. 13.2 and the speed with which it converges to the the limit, with the stationary distribution $(\pi_1, \dots, \pi_5)$ in each row. Also, simulate a long chain from P , and check that the relative frequencies in states 1, ..., 5 match the equilibrium distribution. A formal proof of the intuitive $N_{n,a}/n \rightarrow_{\text{pr}} \pi_a$, with $N_{n,a}$ the number of visits to a during the first n observations, is given in Ex. 13.13. Why is it not enough to point to the Law of Large Numbers from Ch. 2?

after 4-5
generations
equilibrium is
reached

(d) Let us shift perspectives, reading Turgenev backwards, from sons to fathers. Show that the time-reversed transition probabilities become

time reversal

$$p_{i,j}^* = \Pr(X = j | Y = i) = \pi_j p_{j,i} / \pi_i \quad \text{for } i, j = 1, \dots, 5.$$

Compute the matrix P^* with these, and compare with P . Show that the stationary distribution of the time-reversed chain is identical to that of the original chain.

Ex. 13.5 *Today, given the past and the future.* Consider a Markov chain $X_0, X_1, \dots, X_n$, as with (13.1).

(a) Show that

$$\Pr(X_i = x | x_0, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \Pr(X_i = x | x_{i-1}, x_{i+1}),$$

i.e. today's outcome depends only on yesterday and tomorrow, not other aspects of the past and the future.

(b) For the symmetric simple two-state chain studied in Ex. 13.3, with 0-1 outcomes, find the four probabilities $\Pr(X_i = 1 | (x_{i-1}, x_{i+1}))$, for the neighbour pair being (0, 0), (0, 1), (1, 0), (1, 1). Parametrise the stay-put probability $1 - a = \Pr(X_i = 1 | X_{i-1} = 1)$ as $\exp(\beta) / \{1 + \exp(\beta)\}$. Show that $\Pr(X_i = x | x_{i-1}, x_{i+1}) \propto \exp(H\beta)$, where H is the number of neighbours equal to x . This is significant, leading to fruitful models in dimension two, with $X_{i,j}$ on a chessboard, say. For the famous Ising model, the leading characteristic is $\Pr(X_{i,j} = x | \text{rest}) \propto \exp\{\beta H(x)\}$, with $H(x)$ the number among the four neighbours being equal to x . Give prose descriptions for how such a model will behave, for β clearly positive, clearly negative, or close to zero.

the Ising model,
in dimension 1;
check Story
#90
basically from
1925

First visiting times, generating functions

Ex. 13.6 *First visits, and generating functions.* For a Markov chain $X_0, X_1, X_2, \dots$ with states $1, \dots, k$, suppose $X_0 = i$, and consider $T_{i,j}$, the first time the chain visits j . Let

$$f_{i,j}^{(n)} = \Pr(T_{i,j} = n | X_0 = i) = \Pr(X_1 \neq j, \dots, X_{n-1} \neq j, X_n = j | X_0 = i)$$

for $n = 1, 2, 3, \dots$. For these first visit probabilities, as well as for the transition probabilities $p_{i,j}^{(n)}$, we define their associated generating functions,

$$F_{i,j}(s) = \sum_{n=1}^{\infty} f_{i,j}^{(n)} s^n \quad \text{and} \quad P_{i,j}(s) = \sum_{n=0}^{\infty} p_{i,j}^{(n)} s^n,$$

for s inside the convergence interval for the power series in question. Here $p_{i,i}^{(0)} = 1$, whereas $p_{i,j}^{(0)} = 0$ for $j \neq i$.

(a) Show that $f_{i,i}^* = F_{i,i}(1) = \sum_{n=1}^{\infty} f_{i,i}^{(n)}$ is the probability that $T_{i,i}$ is finite, given that $X_0 = i$; in other words, $1 - f_{i,i}^*$ is the probability that the chain never returns to i . We say that the chain is *recurrent* if $f_{i,i}^* = 1$ and *transient* if $f_{i,i}^* < 1$.

(b) Show that $p_{i,i}^{(n)} = \sum_{k=1}^n f_{i,i}^{(k)} p_{i,i}^{(n-k)}$, and from this that $P_{i,i}(s) = 1/\{1 - F_{i,i}(s)\}$. In situations where $P_{i,i}(s)$ can be found, explain that $F_{i,i}(s) = 1 - 1/P_{i,i}(s)$. When i is recurrent, show that the mean of the first revisit variable $T_{i,i} = \min\{n \geq 1: X_n = i\}$, for a chain starting in $X_0 = i$, is equal to $F'_{i,i}(1)$.

(c) With the chain starting in $X_0 = i$, let $N_i = \sum_{n=1}^{\infty} I(X_n = i)$ be the number of times the chain visits i again. Show that EN_i is finite if and only if $f_{i,i}^* < 1$. If state i is recurrent, show that the chain will visit i not only once, but, with probability one, infinitely many times.

(d) For $j \neq i$, show $p_{i,j}^{(n)} = \sum_{k=1}^n f_{i,j}^{(k)} p_{j,j}^{(n-k)}$, and then that $P_{i,j}(s) = F_{i,j}(s)P_{j,j}(s)$.

Ex. 13.7 *A well, unwell, sick model.* Suppose that a man is either well, or unwell, or sick, with these three states here called 0, 1, 2. Letting X_n be the person's state at time n , say recorded in months, suppose $X_0, X_1, \dots$ forms a Markov chain, with transition probability matrix

$$P = \begin{pmatrix} 1-p_0 & p_0 & 0 \\ 0 & 1-p_1 & p_1 \\ p_2 & 0 & 1-p_2 \end{pmatrix}.$$

(a) Translate the probabilistic assumptions here to a few sentences of prose, describing the person's life.

(b) Suppose $p_0 = 0.02$. If the person is well on January 1st, what is the probability that he remains well, all through January, February, March?

(c) Suppose $X_0 = 0$, and consider the first time T_0 he gets unwell. Find the distribution for T_0 , and show that $ET_0 = 1/p_0$.

(d) Let then T be the full cycle, from being well, to then becoming unwell and sick, and becoming well again. With T represented as $T_0 + T_1 + T_2$, find the mean and variance of T .

(e) Find the equilibrium distribution π_0, π_1, π_2 for this man's health process over time, expressed via p_0, p_1, p_2 . In fact, show that $\pi_0 = p_0^{-1}/k, \pi_1 = p_1^{-1}/k, \pi_2 = p_2^{-1}/k$, with $k = 1/p_0 + 1/p_1 + 1/p_2$.

(f) Let B_n be the number of days our person has been ill, in the course of the first n days, after having started off being well. Find formulae for the mean and variance of B_n/n , and comment.

Ex. 13.8 *From falling ill to having recovered.* Mr Jones has fallen ill. Thanks to a clear diagnosis and correct medicine, along with his own good constitution, he'll soon enough have fully recovered, however. In fact he reports that every day he is about ten percent better than the day before. When asked to make this optimistic statement a bit more precise, he explains that on his scale of wellbeing, normal health corresponds to level 2.0, and this particular illness to level 1.0. His health index is $H_0 = 1.0$ on the day when he fell ill, but then $H_m = (1 + \delta_1) \cdots (1 + \delta_m)$ after m days, where he experiences the δ_j as i.i.d. with mean 0.10 and standard deviation 0.03. So when will he have fully recovered?

so Markov
chains can be
worked with
also with
continuous
outcomes

(a) Clearly precise answers, about the day M where H_m first exceeds 2.0, depends on the particularities of the distribution of the δ_j . One modelling possibility, which captures the essence of the setup here, is to take $1 + \delta_j = \exp(y_j)$, with the y_j i.i.d. from an appropriate $\text{Gam}(a, b)$. Show that

$$\mathbb{E} \delta_j = \left(\frac{b}{b-1} \right)^a - 1, \quad \text{Var} \delta_j = \left(\frac{b}{b-2} \right)^a - \left(\frac{b}{b-1} \right)^{2a}.$$

Find the Gamma distribution parameters to match $\mathbb{E} \delta_j = 0.10$ and $\text{Var} \delta_j = 0.03^2$ (answers, modulo numerical imprecision: (12.151, 127.991)). Produce a plot, displaying densities for his health index after 2, 4, 6 days.

(b) Find and present the full distribution for M , the number of days it takes Mr Jones to be fully recovered. What is the most probable number of days needed for his recovery, based on the ‘about ten percent better each new day’ starting assumption?

(c) Investigate a few other possibilities, regarding the distribution of the δ_j , but still keeping the $\mathbb{E} \delta_j = 0.10$ and $\text{Var} \delta_j = 0.03^2$ starting information. Are the gamma-based results reached fairly robust?

Ex. 13.9 *The number of throws to get six-six.* You roll your k -sided die, yielding $X_1, X_2, \dots$ uniform on $\{1, \dots, k\}$. The time to the first k is a simple geometric, with parameter $p = 1/k$ and mean k , but the time needed to see (k, k) is a more complicated random variable, since (X_{n-1}, X_n) and (X_n, X_{n+1}) are dependent.

(a) Explain that the sequence of $X_n^* = (X_n, X_{n+1})$ follows a Markov chain, on the sample space $\{1, \dots, k\} \times \{1, \dots, k\}$ of pairs (i, j) . For $k = 3$, write down the transition matrix P , of size $k^2 = 9$. Each row has 3 elements equal to p and 6 elements zero.

(b) Show next that the two-step transition matrix P^2 has all elements equal to $1/k^2$, each row having $(1/k^2, \dots, 1/k^2)$. With generating function notation from Ex. 13.6, for this Markov chain on sates (i, j) , show that

$$P_{(i,j),(k,k)}(s) = \delta_{(i,j),(k,k)} + \delta_{j,k}ps + \sum_{n=2}^{\infty} p^2 s^n = \delta_{i,k}\delta_{j,k} + \delta_{j,k}ps + p^2 s^2/(1-s),$$

with Kronecker delta notation, $\delta_{i,k}$ being 1 or 0 if $i = k$ or $i \neq k$. In particular, writing $q = 1 - p$, deduce

$$P_{(k,k),(k,k)}(s) = 1 + ps + \frac{p^2 s^2}{1-s} = \frac{1-s+ps(1-s)+p^2 s^2}{1-s} = \frac{1-qs-pqs^2}{1-s},$$

(c) Consider $T_{i,i}$, the first time one sees $X_n^* = (i, i)$ again, after having started with $X_0^* = (i, i)$. Show from Ex. 13.6 that the first revisiting probabilities $\Pr(T_{i,i} = n)$ have generating function

$$F_{(i,i),(i,i)}(s) = 1 - \frac{1-s}{1-qs-pqs^2}.$$

Show from this that the expected recurrence time, from $(X_0, X_1) = (i, i)$ to seeing $(X_n, X_{n+1}) = (i, i)$ again, is $1/p^2 = k^2$.

(d) For start position $X_0^* = (i, j)$, show that the time T to reach $X_n^* = (k, k)$ has a distribution with generating function

$$F_{(i,j),(k,k)}(s) = \frac{p_{(i,j),(k,k)}s(1-s) + p^2 s}{1-qs-pqs^2}$$

(e) The task is to find the mean of $T = \min\{n \geq 2: (X_{n-1}, X_n) = (k, k)\}$. The distribution and mean of T is affected by the start position (X_1, X_2) from the first two throws. For start position (i, j) , with $j < k$, show that T , with $T \geq 4$, has mean $k^2 + k + 2$; for good start position (i, k) , with $i < k$, with $T \geq 3$, show that the mean is $k^2 + 2$; for perfect start position (k, k) , we have of course $T = 2$. From these ingredients, show that the time to first seeing (k, k) has expected value

$$ET = \frac{k(k-1)}{k^2}(k^2 + k + 2) + \frac{k-1}{k^2}(k^2 + 2) + \frac{1}{k^2}2 = k^2 + k.$$

(f) The above calculations have used general tools from Ex. 13.6 to catalogue the means of T from different start positions (i, j) for the two first throws. Another venue, bypassing the Markov chains, is as follows. Show that T can be represented as $T_0 + 1$, with probability p , and $T_0 + T^* + 1$, with probability $1 - p$, where T_0 is the geometric time needed to see k the first time, and T^* has the same distribution as T . With generating function $G_0(s) = E s^{T_0} = ps/(1 - qs)$, deduce that the generating function of T satisfies $G(s) = pG_0(s)s + qG_0(s)sG(s)$, and from this deduce

$$G(s) = \frac{pG_0(s)s}{1 - qG_0(s)s} = \frac{p(1 - qs)G_0(s)s}{(1 - qs)(1 - qG_0(s)s)} = \frac{p^2 s^2}{1 - qs - pq s^2}.$$

Check that $G'(1) = k^2 + k$, and find also the variance of T .

Ex. 13.10 *The symmetric random walk.* Consider a simple up-and-down process $X_0, X_1, X_2, \dots$ on the integers, as follows. It starts in $X_0 = 0$, and if $X_n = i$, then X_{n+1} is $i + 1$ or $i - 1$, with probabilities $\frac{1}{2}, \frac{1}{2}$. We call this a simple random walk, with symmetric probabilities.

(a) Simulate a few paths, say up to $n = 1000$, to learn aspects of the behaviour of a random walk. Show that $X_n = U_n - D_n$, with $U_n \sim \text{binom}(n, \frac{1}{2})$ and $D_n = n - U_n$. Use this to give the mean and variance of X_n . What is the probability that $|X_n| \leq 2\sqrt{n}$?

(b) Show that $p_{0,0}^{2n} = \binom{2n}{n}(\frac{1}{2})^{2n}$. Using Stirling's formula, see Story #103, show that $p_{0,0}^{(2n)} \doteq 1/(\pi n)^{1/2}$. Deduce that the random walk process is recurrent; if it starts at 0, it will with probability 1 come back, some time, to 0.

(c) It is useful to find the generating function of the $p_{0,0}^{(n)}$, also since that gives us the generating function for the $f_{0,0}^{(n)}$, via results from Ex. 13.6. To this end, work first with $G(x) = \sum_{n=0}^{\infty} \binom{2n}{n} x^n$. Show that $G'(x) = 2G(x) + 4xG'(x)$; use this to find the derivative of $\log G(x)$; and work out from this that

$$G(x) = \sum_{n=0}^{\infty} \binom{2n}{n} x^n = \frac{1}{(1 - 4x)^{1/2}} \quad \text{for } |x| < \frac{1}{4}.$$

Show also that

$$(1 - x)^{1/2} = 1 - \sum_{n=1}^{\infty} \binom{2n}{n} \frac{1}{2n-1} \left(\frac{1}{4}\right)^n \quad \text{for } |x| < 1.$$

(d) Show that $P_{0,0}(s) = 1/(1 - s^2)^{1/2}$, and deduce that state 0 is recurrent; the chain will indeed come back to 0, infinitely many times.

(e) Then show that $F_{0,0}(s) = 1 - (1 - s^2)^{1/2}$, and deduce from the above that

$$f_{0,0}^{(n)} = \binom{2n}{n} \frac{(\frac{1}{4})^n}{2n-1} \quad \text{for } n \geq 1.$$

Show that its mean is big, so the time for the first return is expected to be long.

Ex. 13.11 *The biased random walk.* Consider now a biased random walk, where the chain goes up with probability p and down with probability $q = 1 - p$, and where p might be different from $\frac{1}{2}$.

(a) Write $X_n = U_n - D_n = 2U_n - n$, where $U_n \sim \text{binom}(n, p)$, to find the mean and variance of X_n . In particular, the biased random walk has a drift of $(2p - 1)n$.

(b) Show that $P_{0,0}(s) = 1/(1 - 4pqs^2)^{1/2}$, and that the expected number of visits in 0, given that the walk started there, is

$$E(N_0 | X_0 = 0) = \frac{1}{|1 - 2p|} - 1 = \begin{cases} 2p/(1 - 2p) & \text{if } p \leq \frac{1}{2}, \\ 2(1 - p)/(2p - 1) & \text{if } p \geq \frac{1}{2}. \end{cases}$$

(c) Show that $F_{0,0}(s) = 1 - (1 - 4pqs^2)^{1/2}$, and derive from this that

$$f_{0,0}^{(n)} = \binom{2n}{n} \frac{(pq)^n}{2n - 1} \quad \text{for } n \geq 1.$$

What is the probability of returning to zero, given that the chain started there?

(d) Suppose the chain starts at zero and that p is a bit bigger than $\frac{1}{2}$. It is then a chance that the chain will drift off to infinity, without ever coming back to zero. What is the expected time to return, given that it actually will return (i.e. not drift off to infinity)? And what is the probability that the chain will never return, given that it has not been seen at zero for the first 100 times?

(e) Find $p_{0,1}^{(2n+1)}$ and then a formula for $P_{0,1}(s)$. What is $E(N_1 | X_0 = 0)$, with N_1 the number of visits to 1?

(f) What is the probability $f_{0,100}^*$, that the chain sooner or later will find its way to 100, given that it started in $X_0 = 0$?

Ex. 13.12 *You'll eventually find your car keys (in dimension two).* We saw in Ex. 13.10 that for the symmetric random walk in dimension 1, we have $p_{0,0}^{(2n)} \doteq 1/(\pi n)^{1/2}$, a divergent series, implying recurrence. Now study the random walk in dimension two, with the four movements right, left, up, down, having equal probabilities $\frac{1}{4}$.

(a) If the walk starts at the origo, explain that the position after n steps is $(X_n, Y_n) = (R_n - L_n, U_n - D_n)$, with (R_n, L_n, U_n, D_n) a multinomial $(n, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

(b) Explain next that $p_{2n} = p_{00,00}^{(2n)}$, the probability of having come back to origo after $2n$ steps, is that of having $R_{2n} = L_{2n}$ and $U_{2n} = D_{2n}$. Show from this that

$$p_{2n} = \sum_{(i,j): i+j=n} \frac{(2n)!}{i! i! j! j!} \left(\frac{1}{4}\right)^{2n}.$$

The above can be worked with directly, but the expression is somewhat awkward. An easier route is as follows, via $N_X = R_{2n} + L_{2n}$, $N_Y = U_{2n} + D_{2n}$. Explain that $N_X \sim \text{binom}(2n, \frac{1}{2})$, with $N_Y = 2n - N_X$. Conditioning on $N_X = 2k$, show that

$$p_{2n} = \sum_{k=0}^n \binom{2n}{2k} \binom{2k}{k} \binom{2n-2k}{n-k} \left(\frac{1}{2}\right)^{4n}.$$

(c) This can be simplified further. Argue that the coming-back problem splits into two independent one-dimensional random walks, with

in dimension
three you might
never return

$$p_{2n} = \Pr(R_{2n} = L_{2n}) \Pr(U_{2n} = L_{2n}) = \left\{ \binom{2n}{n} \left(\frac{1}{2}\right)^{2n} \right\}^2 \doteq \frac{1}{\pi n}.$$

Conclude that the chain is recurrent, i.e. with probability 1 coming back to where it started (though slower than in dimension one, of course).

Recurrence times, CLTs, and likelihood for Markov chain models

Ex. 13.13 *Occupation and recurrence times.* Consider a Markov chain $X_0, X_1, \dots, X_n$ on a state space $\{1, \dots, k\}$ with transition probabilities $p_{a,b}$. For simplicity for the details to follow we assume the chain starts in its equilibrium distribution $\pi = (\pi_1, \dots, \pi_k)$, given by $\pi P = \pi$. For various purposes, regarding analysis of both probabilistic and statistical issues, we need a CLT for such chains.

(a) For a recurrent state a , consider $N_{n,a} = \sum_{i=1}^n I(X_i = a)$, the number of visits to a . Show that $E N_{n,a} = n\pi_a$ and

$$\text{Var } N_{n,a} = n\pi_a(1 - \pi_a) + 2 \sum_{i < \ell} \pi_a(p_{a,a}^{(\ell-i)} - \pi_a).$$

With $g_{a,r} = p_{a,a}^{(r)} - \pi_a$, show that $n^{-1} \text{Var } N_{n,a} = \pi_a(1 - \pi_a) + 2\pi_a \sum_{r=1}^n (1 - r/n)g_{a,r}$. Explain from this that $N_{n,a}/n \rightarrow_{\text{pr}} \pi_a$, for each a . Writing $Z_{n,a} = \sqrt{n}(N_{n,a}/n - \pi_a)$, show also that

$$\text{Var } \{\sqrt{n}(N_{n,a}/n - \pi_a)\} \rightarrow \sigma_a^2 = \pi_a(1 - \pi_a) + 2\pi_a \gamma_{a,a},$$

where $\gamma_{a,a} = \sum_{r=1}^{\infty} (p_{a,a}^{(r)} - \pi_a)$. Explain that each $\gamma_{a,b} = \sum_{r=1}^{\infty} (p_{a,b}^{(r)} - \pi_b)$ is finite via exponential convergence established in Ex. 13.2.

(b) We wish to prove that $\sqrt{n}(N_{n,a}/n - \pi_a) \rightarrow_d N(0, \sigma_a^2)$, which is an instance of a CLT for Markov chains. One method is via mixing sequences and bounds and approximations (where Ex. 2.30 might find use). Here we follow another path, roundabout but perhaps with simpler tools. Let $T_{a,1}, T_{a,2}, \dots$ be the times between successive visits to state a , so that $S_{a,m} = T_{a,1} + \dots + T_{a,m}$ is the time for visit m to state a . Show that these $T_{a,1}, T_{a,2}, \dots$ are i.i.d., with $\Pr(T_{a,1} = r) = f_{a,a}^{(r)}$, as in Ex. 13.6. Show also that $m^{1/2}(S_{a,m}/m - \xi_a) \rightarrow_d N(0, \tau_a^2)$, with ξ_a and τ_a^2 being the mean and variance of these waiting times.

mean recurrence
time is inverse
equilibrium

(c) Show that $N_{n,a} = m$ is equivalent to $S_{a,m} \leq n < S_{a,m+1}$. From

$$S_{a,N_n} \leq n < S_{a,N_n+1} \quad \text{deduce that} \quad \frac{S_{a,N_n}}{N_{n,a}} \leq \frac{n}{N_{n,a}} < \frac{S_{a,N_n+1}}{N_{n,a}}.$$

Show that the mean waiting time, also called the mean recurrence time, is $\xi_a = 1/\pi_a$, and verify again from this that indeed $N_{n,a}/n \rightarrow_{\text{pr}} \pi_a$, as we discovered above via somewhat laborious variance calculations.

(d) From

$$\sqrt{n} \left(\frac{S_{a,N_n}}{N_{n,a}} - \xi_a \right) \leq \sqrt{n} \left(\frac{n}{N_{n,a}} - \frac{1}{\pi_a} \right) < \sqrt{n} \left(\frac{S_{a,N_n+1}}{N_{n,a}} - \xi_a \right),$$

deduce that $\sqrt{n}(N_{n,a}/n - \pi_a)$ indeed must have a zero-mean normal limit distribution, with a variance that needs to be equal to the σ_a^2 above.

(e) Show also that $\sigma_a = p_a^{3/2} \tau_a$, connecting the σ_a formula above with the standard deviation of the waiting time distribution. If a chain is such that τ_a is more easily found, this also gives a recipe for finding and perhaps later estimating the σ_a .

(f) Since we shall need the joint limiting distribution for such visit-counting variables in material below, consider $(N_{n,a}, N_{n,b})$ for a pair with $a \neq b$. Work through the details of

$$\begin{aligned} \text{cov}(N_{n,a}, N_{n,b}) &= \sum_{i,j} \text{cov}\{I(X_i = a), I(X_j = b)\} \\ &= -n\pi_a\pi_b + \sum_{i < j} (\pi_a p_{a,b}^{(j-i)} - \pi_a\pi_b) + \sum_{i > j} (\pi_b p_{b,a}^{(i-j)} - \pi_a\pi_b), \end{aligned}$$

and establish, with details similar to those above, that

$$\text{cov}(Z_{n,a}, Z_{n,b}) \rightarrow -\pi_a\pi_b + \pi_a\gamma_{a,b} + \pi_b\gamma_{b,a}.$$

Ex. 13.14 *Ergodic convergence of averages.* Suppose $X_1, \dots$ is a Markov chain on the state space $\{1, \dots, k\}$, with all transition probabilities positive, and for simplicity that it starts in its equilibrium distribution $\pi = (\pi_1, \dots, \pi_k)$, with $\pi P = \pi$. For any function h , consider $\bar{h}_n = n^{-1} \sum_{i=1}^n h(X_i)$.

(a) Show that $\bar{h}_n = \sum_{a=1}^k h(a) N_{n,a}/n$, with $N_{n,a}/n$ the relative frequency of visits at a . Use Ex. 13.13 to prove that $\bar{h}_n \rightarrow_{\text{pr}} \sum_{a=1}^k h(a)\pi_a = E_\pi h(X)$, with X here denoting a variable having the equilibrium distribution.

this is called
ergodic
convergence

(b) Find also an expression for the variance of $\bar{h}_n$.

Ex. 13.15 *CLT for counting in Markov chains.* Consider a Markov chain $X_0, X_1, \dots, X_n$ with transition probabilities $p_{a,b}$ on the states $1, \dots, k$, and assume for simplicity that it starts out with X_0 drawn from the stationary distribution $\pi = (\pi_1, \dots, \pi_k)$, with $\pi = P\pi$. Write $N_{a,b} = \sum_{i=1}^n I((X_{i-1}, X_i) = (a, b))$ for the number of cases with transitions from a to b , and $N_{a,\cdot} = \sum_b N_{a,b}$ for the number of cases we have seen $X_{i-1} = a$. Write also $N_{\cdot,\cdot} = \sum_b N_{a,b}$ for the number of visits seen at state a for $X_0, X_1, \dots, X_{n-1}$.

(a) For the model that does not impose further structure on the $k(k-1)$ unknown parameters $p_{a,b}$, show that the likelihood and log-likelihood functions become

$$L_n = \prod_{a,b} p_{a,b}^{N_{a,bn}} \text{ and } \ell_n = \sum_{a,b} N_{a,b} \log p_{a,b}, \quad \text{with } \sum_b p_{a,b} = 1 \text{ for each } a.$$

Show that this is maximised for $\hat{p}_{a,b} = N_{a,b}/N_{a,\cdot}$, the natural observed transition rates.

(b) For the first and second derivatives, show that

$$\begin{aligned} \frac{\partial \ell_n}{\partial p_{a,b}} &= \frac{N_{a,b}}{p_{a,b}} - \frac{N_{a,k}}{p_{a,k}} \quad \text{for } b = 1, \dots, k-1, \\ \frac{\partial^2 \ell_n}{\partial p_{a,b} \partial p_{a,c}} &= -\frac{N_{a,b}}{p_{a,b}^2} \delta_{b,c} - \frac{N_{a,k}}{p_{a,k}^2} \quad \text{for } b, c = 1, \dots, k-1. \end{aligned}$$

For the information matrix $J = -n^{-1} E \partial^2 \ell_n / \partial \theta \partial \theta^t$, here not depending on n , writing θ for the collection of $k(k-1)$ parameters $p_{a,b}$, show that

$$J = \text{diag}(\pi_1 M_1, \dots, \pi_k M_k),$$

where M_a of size $(k-1) \times (k-1)$ has elements $1/p_{a,b} + 1/p_{a,k}$ on the diagonal and $1/p_{a,k}$ outside. Check with Ex. 5.23 to recognise this as the multinomial structure. If maximum

likelihood theory works as anticipated, therefore, we should have

$$\sqrt{n}(\hat{p}_{a,b} - p_{a,b}) \rightarrow_d (1/\pi_a)^{1/2} V_{a,b} \quad \text{for } b = 1, \dots, k,$$

with covariance structure $\text{cov}(V_{a,b}, V_{a,c}) = p_{a,b}(\delta_{b,c} - p_{a,c})$ for $b, c = 1, \dots, k$; also, the vector of $V_{a,b}$ is independent of the vector of $V_{c,d}$ when $a \neq c$.

(c) To verify the anticipated multinomial limiting structure above we shall use $N_{a,\cdot}/n \rightarrow_{\text{pr}} \pi_a$, for each a , from Ex. 13.13. Argue that given $N_{a,\cdot}$, the collection $(N_{a,1}, \dots, N_{a,k})$ forms a multinomial count, with probabilities $p_{a,1}, \dots, p_{a,k}$. Deduce from this that

$$\sqrt{n}(\hat{p}_{a,b} - p_{a,b}) = (n/N_{a,\cdot})^{1/2} N_{a,\cdot}^{1/2} (\hat{p}_{a,b} - p_{a,b}) \rightarrow_d (1/\pi_a)^{1/2} V'_{a,b},$$

say, where these form a zero-mean normal limit with covariance structure as above; in other words, the $V'_{a,b}$ collection has exactly the same distribution as the anticipated $V_{a,b}$ collection above.

Ex. 13.16 *Maximum likelihood for Markov chain models.* Consider a parametric model $p_{a,b}(\theta)$ for the transition probabilities of a Markov chain, leading also to an equilibrium distribution $p_a(\theta)$. We assume the model is smooth and regular with two continuous derivatives. Write $u_{a,b}(\theta) = \partial \log p_{a,b}(\theta) / \partial \theta$ and $i_{a,b}(\theta) = \partial^2 \log p_{a,b}(\theta) / \partial \theta \partial \theta^t$ for the first and second log-derivatives.

(a) For each fixed a , take two derivatives of $\sum_b p_{a,b}(\theta) = 1$ to learn that $\sum_b p_{a,b}(\theta) u_{a,b}(\theta) = 0$ and

$$J_a(\theta) = - \sum_b p_{a,b}(\theta) i_{a,b}(\theta) = \sum_b p_{a,b}(\theta) u_{a,b}(\theta) u_{a,b}(\theta)^t.$$

Define also $J(\theta) = \sum_a \pi_a(\theta) J_a(\theta)$, where $\pi(\theta)$ is the equilibrium distribution implied by the $p_{a,b}(\theta)$ transition probabilities.

(b) With data $X_0, X_1, \dots, X_n$ for such a chain, show that the log-likelihood function can be written $\ell_n(\theta) = \sum_{a,b} N_{a,b} \log p_{a,b}(\theta)$ (ignoring the start information). Write $U_n(\theta)$ and $I_n(\theta)$ for the first and second order derivatives. At the true value θ_0 , show that

$$-n^{-1} I_n(\theta_0) = - \sum_{a,b} n^{-1} N_{a,b} i_{a,b}(\theta_0) \rightarrow_{\text{pr}} J = J(\theta_0).$$

(c) We next need to work with $U_n(\theta) = \sum_{a,b} N_{a,b} u_{a,b}(\theta)$. Write $\pi_a, p_{a,b}, u_{a,b}$ for these functions evaluated at the true position θ_0 . Let $\hat{p}_a = N_{a,\cdot}/n$, the proportion of time the chain has visited a , tending in probability to π_a . Cleverly writing

$$\hat{\pi}_a \hat{p}_{a,b} - \pi_a p_{a,b} = \hat{\pi}_a (\hat{p}_{a,b} - p_{a,b}) + (\hat{\pi}_a - \pi_a) p_{a,b},$$

deduce

$$n^{-1/2} U_n(\theta_0) = \sum_{a,b} \hat{\pi}_a \sqrt{n} (\hat{p}_{a,b} - p_{a,b}) u_{a,b} + \sum_{a,b} \sqrt{n} (\hat{\pi}_a - \pi_a) p_{a,b} u_{a,b},$$

and verify also that the second term disappears. Use then results and notation of Ex. 13.13–13.15 to establish that

$$n^{-1/2} U_n(\theta_0) \rightarrow_d U = \sum_{a,b} \pi_a^{1/2} V_{a,b} u_{a,b}.$$

Via covariance expressions and algebraic efforts, show that indeed $\text{Var } U = J$.

see e.g. Ex. 5.10 (d) Via these results, along with ML theory arguments resembling those used in Ch. 2, show that

$$\sqrt{n}(\hat{\theta} - \theta_0) = J^{-1}U_n(\theta_0)/\sqrt{n} + o_{\text{pr}}(1) \rightarrow_d J^{-1}U \sim N_p(0, J^{-1}).$$

For smooth focus parameters $\phi = \phi(\theta)$, estimated as $\hat{\phi} = \phi(\hat{\theta})$, explain via the general delta methods of Ch. 2 that $\sqrt{n}(\hat{\phi} - \phi) \rightarrow_d N(0, c^t J^{-1} c)$, with $c = \partial\phi(\theta_0)/\partial\theta$.

for a good illustration, check with Story #25

(e) Show furthermore that $W_n = 2\{\ell_n(\hat{\theta}) - \ell_n(\theta_0)\} \rightarrow_d \chi_p^2$, and explain how this may be used to test hypotheses and form confidence regions for the model parameter.

Ex. 13.17 *Waiting times in a two-state Markov chain.* As in Ex. 13.3, consider a chain on $\{0, 1\}$ with transition probability matrix

$$P = \begin{pmatrix} p_{0,0} & p_{0,1} \\ p_{1,0} & p_{1,1} \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

(a) Find the waiting time probabilities $f_{0,0}^{(n)}$, for $T_0 = \min\{n \geq 1: X_n = 0\}$, after having started in 0. Find the generating function $F_{0,0}(s)$, along with ξ_0 and τ_0^2 , the mean and variance.

(b) Find the limit distribution for $\sqrt{n}(N_0/n - \pi_0)$ with $N_0 = \sum_{i=1}^n I(X_i = 0)$ counting the number of visits to state 0 in the course of the first n times. Verify the connection between σ_0 and τ_0 found in Ex. 13.13.

Ex. 13.18 *A symmetric random walk with two reflecting barriers.* Consider the Markov chain on $\{0, 1, \dots, k\}$ with transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ p & q & p & 0 & \cdots & 0 \\ 0 & p & q & p & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix},$$

where $2p + q = 1$. The chain takes steps to the right or to the left, with a probability q of staying put, and there are reflecting barriers at 0 and k .

(a) From $\pi P = \pi$, show that the stationary distribution is given by

$$\pi_0 = \pi_k = \frac{p}{k-1+2p}, \quad \pi_i = \frac{1}{k-1+2p} \quad \text{for } i = 1, \dots, k-1.$$

(b) Assume the chain $X_0, X_1, \dots, X_n$ has been observed, with $N_{i,i-1}$, $N_{i,i}$, $N_{i,i+1}$ the number of transitions $(i, i-1)$, (i, i) , $(i, i+1)$. Show that the log-likelihood becomes

$$\ell_n(p) = \sum_{i=1}^{k-1} \{(N_{i,i-1} + N_{i,i+1}) \log p + N_{i,i} \log(1-2p)\} = A_n \log p + B_n \log(1-2p),$$

with $A_n = \sum_{i=1}^{k-1} (N_{i,i-1} + N_{i,i+1})$ and $B_n = \sum_{i=1}^{k-1} N_{i,i}$. Show that the ML estimator is $\hat{p} = \frac{1}{2} A_n / (A_n + B_n)$.

(c) Show, using results from Ex. 13.13, that

$$A_n/n \rightarrow_{\text{pr}} 2(k-1)p/(k-1+2p), \quad B_n/n \rightarrow_{\text{pr}} (k-1)q/(k-1+2p),$$

and that this entails consistency of $\hat{p}$ for p . For the Fisher information, see Ex. 13.16, show that $J = 2(k-1)/\{(pq)(k-1+2p)\}$, and that indeed $\sqrt{n}(\hat{p} - p) \rightarrow_d N(0, \tau^2)$ with variance $pq(k-1+2p)/\{2(k-1)\}$.

(d) Make a program to simulate such chains, for given state space size $k + 1$ and parameters $(p, q) = (p, 1 - 2p)$. Check that (A_n, B_n) are close to their limits. For one such chain, with $k = 5$ and length $n = 250$, we got the transition counts $N_{a,b} = \sum_{i=1}^n I(x_{i-1} = a, x_i = b)$ below. Estimate p and q and find 90 percent confidence intervals.

	0	1	2	3	4	5
0	0	14	0	0	0	0
1	14	9	17	0	0	0
2	0	17	6	24	0	0
3	0	0	24	11	31	0
4	0	0	0	32	8	21
5	0	0	0	0	22	0

Markov chain Monte Carlo methods

Ex. 13.19 *The Metropolis and Metropolis–Hastings MCMC algorithms.* We learn in Ex. 13.2 that a given Markov chain has its equilibrium distribution; if X follows this (π_j) distribution, then $X_n \rightarrow_d X$. The table can be turned: for a given perhaps complicated X , we can construct a perhaps simple Markov chain having this X is its equilibrium. Let (π_i) be a probability distribution over some large sample space. The task is to simulate realisations from this distribution.

(a) Suppose $X_0, X_1, X_2, \dots$ come from a Markov chain with transition probability matrix $p_{i,j} = \Pr(X_n = j \mid X_{n-1} = i)$. Show that if these are constructed such that

$$\pi_i p_{i,j} = \pi_j p_{j,i} \quad \text{for all } i, j, \quad (13.3)$$

and also that the chain is irreducible with period 1, then the stationary (or equilibrium) distribution for the chain is actually (π_i) . The (13.3) are called the *detailed balance conditions*.

(b) There ought to be quite some elbow room for many different $p_{i,j}$ constructions that obey the conditions above. The Metropolis method of 1953 uses

$$p_{i,j} = q_{i,j} \min(1, \pi_j / \pi_i) \quad \text{for } j \neq i,$$

where $q_{i,j} = \Pr(X' = j \mid X = i)$ is the so-called proposal distribution, assumed here to be symmetric ($q_{i,j} = q_{j,i}$). Show that the condition of (a) really is in force with such $p_{i,j}$ constructions. Explain also that the Markov chain generated in this fashion can be represented as follows, where we also write $\pi_j = \pi(j)$. For an i.i.d. sequence $U_1, U_2, \dots$ from the uniform, create dependent coin flips $J_n = I(U_n \leq r_n)$, with dynamically evolving acceptance probabilities $r_n = \min\{1, \pi(X_{n,\text{prop}}) / \pi(X_{n-1})\}$, and then let

$$X_n = (1 - J_n)X_{n-1} + J_n X_{n,\text{prop}} \quad \text{for } n \geq 1.$$

(c) Sometimes it is however practical, or even necessary, to employ $q_{i,j}$ that are not symmetric in (i, j) . Let $q_{i,j}$ be a potentially non-symmetric proposal distribution that from the present i proposes a j . Attempt using an accept probability of the type $\min(1, s_{i,j} \pi_j / \pi_i)$, i.e. $p_{i,j} = q_{i,j} \min(1, s_{i,j} \pi_j / \pi_i)$. Show that this really works, provided $s_{i,j} = q_{j,i} / q_{i,j}$! This amounts to the [Hastings \(1970\)](#) generalisation of the Metropolis algorithm: propose j from a symmetric or non-symmetric $q_{i,j}$, and accept with probability $\min(1, (q_{j,i} / q_{i,j}) \pi_j / \pi_i)$.

Ex. 13.20 *Metropolis for a distribution for 10^{12} numbers.* Consider a probability distribution across all natural numbers from zero up to a million million, defined by

$$\pi(x_1, \dots, x_{12}) = \frac{1}{Z(\lambda)} \exp\left\{-\lambda \sum_{j=1}^{12} (x_j - \bar{x})^2\right\} \quad \text{for } (x_1, \dots, x_{12}) \in \{0, \dots, 9\}^{12}.$$

Here $Z(\lambda)$ is the required summation constant, that perhaps not even HAL could manage to compute accurately, and $\bar{x} = (x_1 + \dots + x_{12})/12$. We may think of an outcome $x = (x_1, \dots, x_{12})$ as a random telephone number, in a country employing 12-digit telephone numbers.

(a) What kind of numbers will be preferred by this distribution, i.e. what type of x are likely and what type less likely? Describe some aspects of outcomes, for situations where λ is respectively negative, close to zero, moderate, and large.

(b) How can one manage to sample say 10,000 random telephone numbers from this distribution, for a given λ ? Set up and implement a Metropolis algorithm for achieving this, and discuss how well it works.

(c) Let $\xi(\lambda) = E_\lambda U(X) = E_\lambda \sum_{j=1}^{12} (X_j - \bar{X})^2$, the mean of the random variance, as a function of the underlying λ . Set up simulations in a loop across λ values from -3 to 3 , to find numerical approximations for $\xi(\lambda)$, and plot the resulting curve. Check directly that $\xi(0) = 90.75$.

(d) We've got hold of $n = 200$ telephone numbers from the country in question, and computed $U(x) = \sum_{j=1}^{12} (x_j - \bar{x})^2$ for each of these. Their average value turns out to be $\bar{U} = 11.111$. Estimate the λ parameter. Show that maximum likelihood estimation is equivalent to solving $\xi(\lambda) = 11.111$, and use simulation to show that its solution is $\hat{\lambda} \doteq 0.488$. For this trick, simulating a curve to see when it crosses a certain point, in various cases therefore computing the ML estimate of complicated models, see Story #90 for an illustration.

(e) How can you supplement the $\hat{\lambda}$ parameter estimate you found above with a confidence interval, or a standard deviation estimate?

Ex. 13.21 *MCMC for continuous state spaces.* Crucially, the MCMC apparatus developed in Ex. 13.19 for discrete sample spaces can be modified and properly lifted from the discrete to the continuous case, as already spelt out and exemplified in our Bayesian Ch. 7, see Ex. 7.16-7.17 for relevant statements. This is also used in Stories (xx fill in xx). Our book is largely fully self-contained, in that all methods and statements can be worked through and proven on its basis alone. This is one of very few instances where it is most practical for us, in view of pages and their volume, to push you to other bookshelves, if you need to sift through full detailed proofs of the general-level statements. (xx we point perhaps in N & P to the MCMC book, to Robert Casella 2011, to Harris recurrence, conditions, ergodic averages $m^{-1} \sum_{j=1}^m h(x_j)$ will converge to $E_\pi h(X) = \int h(x)\pi(x) dx$. xx)

check Robert and Casella (2011), ...

(a) Nevertheless, work through the following simple illustration. Suppose it is desired to simulate from the Cauchy density $f(x) = (1/\pi)/(1+x^2)$ via MCMC Metropolis. Do this, in two related ways. (i) Follow the general scheme, as per Ex. 7.16, with proposals $x^{\text{prop}} = x^{\text{old}} + \varepsilon$, where $\varepsilon \sim N(0, 0.123^2)$. (ii) Discretise and follow Ex. 13.19, creating a Markov chain on the gridded state space $[-100, 100]$ with grid-fineness 0.001; Here do $x^{\text{prop}} = x^{\text{old}} + \varepsilon$ with ε the $N(0, 0.123^2)$ but rounded off to 3rd decimal. For both setups, simulate say 10^5 realisations, and compare histograms of density estimates. Make precise

in other books

statements of what is achieved. The moral of the story is that the human eye barely sees any difference; path (ii) has a solid established apparatus behind it, via the exercises above, whereas path (i) requires more pages, theorems, and proofs.

(b) The key statements, to have such on board, are these. Let $\pi(x)$ be a target density on $\mathcal{X}$, a subset of $\mathbb{R}^k$, w.r.t. Lebesgue measure, and with the $\pi(x)$ being positive and continuous over a connected support (the support cannot be split into two disjoint open sets). With the chain having reached x_n at step n , let $q(x, \cdot)$ be a proposal distribution, often symmetric in its two arguments, from which we draw a proposal x^{prop} . Then accept the proposal with probability

$$\text{pr}_{\text{accept}} = \min\left\{1, \frac{\pi(x^{\text{prop}})}{\pi(x_n)} \frac{q(x_n, x^{\text{prop}})}{q(x^{\text{prop}}, x_n)}\right\};$$

if not accepted, then $x_{n+1} = x_n$. Assume the chain, with its proposals, can reach any subset of $\mathcal{X}$ with positive π measure. Then the chain is so-called Harris recurrent, and there is ergodic convergence; for any bounded g , the average $n^{-1} \sum_{i=1}^n g(x_i)$ tends a.s. to the limit $\int g(x)\pi(x) dx$. (xx still missing: light tail condition for the proposal. xx)

Autocorrelation and Gaussian processes

Ex. 13.22 *When is a Gaussian process Markovian?* Consider a Gaussian zero-mean process $Y = \{Y(t) : t \geq 0\}$, with covariance function $K(s, t) = \text{cov}\{Y(s), Y(t)\}$. We have seen various such processes in Ch. 10.

(a) For time points $s < t < u$, write down the conditional distributions (i) for $Y(u)$ given $Y(t)$ and (ii) for $Y(u)$ given both $Y(s)$ and $Y(t)$. Show from this that the Y process is Markovian if and only if $K(s, t)K(t, u) = K(s, u)K(t, t)$ for all $s \leq t \leq u$.

(b) Consider the Brownian motion and Brownian bridge processes worked with in Ch. 10, see e.g. Ex. 10.3. Show that both are Markovian. Explain that the same goes for the Ornstein–Uhlenbeck process, worked with in Ex. 10.5.

(c) Then inspect the correlation function $\exp\{-a|s - t|^\gamma\}$, with $\gamma \in (0, 2)$. Show that a Gaussian process having this correlation function is Markovian only for the case of $\gamma = 1$, which is the Ornstein–Uhlenbeck process.

Ex. 13.23 *An autocorrelation model.* Consider a simple Gaussian time series $Y_1, \dots, Y_n$ with mean zero, variance one, and autocorrelations $\text{corr}(Y_i, Y_j) = \rho^{|j-i|}$, for some $\rho \in (-1, 1)$.

(a) Show that the full observation vector Y has distribution $N_n(0, A)$, where A is the matrix with elements $\rho^{|j-i|}$. Show also that $Y_1, Y_2, \dots$ forms a Markov process.

(b) Write down the joint distribution for $Y_1, \dots, Y_n$ in two ways: First as $(2\pi)^{-n/2} |A|^{-1/2} \exp(-\frac{1}{2} y^t A^{-1} y)$, and then as $f(y_1) \prod_{i=2}^n f(y_i | y_{i-1})$. Deduce from equating these expressions (i) that the determinant is $|A| = (1 - \rho^2)^{n-1}$, and (ii) that the inverse is a band matrix, with elements

$$A^{i,i} = \frac{1 + \rho^2}{1 - \rho^2} \text{ for } 2 \leq i \leq n-1, \quad A^{1,1} = A^{n,n} = \frac{1}{1 - \rho^2}, \quad A^{i-1,i} = -\frac{\rho}{1 - \rho^2}.$$

(c) Give an expression for the log-likelihood function $\ell_n(\rho)$. Show that it is a function of the two statistics $S_0 = \sum_{i=1}^{n-1} Y_i^2$ and $S_1 = \sum_{i=1}^{n-1} Y_i Y_{i+1}$, along with the end point Y_n , even though the model has merely one parameter.

Ex. 13.24 *Gaussian models with correlation functions.* Suppose in general terms that $Y_1, \dots, Y_n$ is a series of observations, with mean function $E Y_i = m_i(\beta)$, constant variance σ^2 , and correlations $\rho_{i,j} = \text{corr}(Y_i, Y_j)$. There would often be further assumptions, concerning the form of the trend function $m_i(\beta)$, how $\rho_{i,j}$ depends on the interdistance from position i to position j , etc. If the observations form an ordinary time series, the times are equidistant, but in this generic setup they do not have to be. We also postulate that the joint distribution is multinormal.

you may check
with Story #37

(a) Show that the full observation vector $Y \sim N_n(m(\beta), \sigma^2 A)$, where A is the matrix of $\rho_{i,j}$ and $m(\beta)$ the vector of $m_i(\beta)$. Show that the log-likelihood function can be written

$$\ell(\beta, \sigma) = -n \log \sigma - \frac{1}{2} \log |A| - \frac{1}{2} \{y - m(\beta)\}^t A^{-1} \{y - m(\beta)\} / \sigma^2 - \frac{1}{2} n \log(2\pi).$$

(b) Assume first that the correlations are known, making A a fixed known matrix (a special case is that of independence, for which $A = I_n$, the identity matrix). Show that the ML estimator for β is the minimiser of $Q(\beta) = \{y - m(\beta)\}^t A^{-1} \{y - m(\beta)\}$, i.e. corresponding to least weighted squares. For the linear mean case of $m_i(\beta) = z_i^t \beta$, with a covariate vector z_i , of dimension say p , show that

$$\hat{\beta} = (Z^t A^{-1} Z)^{-1} Z^t A^{-1} y,$$

where Z is the $n \times p$ matrix with z_i^t as its i th row. Show also that $\hat{\beta}$ is multinormal, unbiased, with variance matrix $\sigma^2 (Z^t A^{-1} Z)^{-1}$. This generalises a classic linear regression result, see Ex. 3.33, from independence to a general correlation structure.

(c) Show next that the ML estimator for σ^2 is $\hat{\sigma}^2 = Q_0/n$, with $Q_0 = \min_{\text{all } \beta} Q(\beta)$. Show, extending the classic result for the case of independence, that $\hat{\sigma}^2 \sim \sigma^2 \chi_{n-p}^2$, assuming that $Z^t A^{-1} Z$ has full rank p . Show also that the maximised log-likelihood is $\ell_{\max} = -n \log \hat{\sigma} - \frac{1}{2} \log |A| - \frac{1}{2} n - \frac{1}{2} n \log(2\pi)$.

(d) More realistically, and interestingly, the correlations will not be known, and need to be modelled and in its turn estimated. Assume $\text{corr}(Y_i, Y_j) = \rho_{i,j}(\lambda)$, say, in terms of a parameter λ of perhaps low dimension. This leads to correlation matrix $A = A(\lambda)$. Show that the log-likelihood function for this extended model becomes

$$\ell(\beta, \sigma, \lambda) = -n \log \sigma - \frac{1}{2} \log |A(\lambda)| - \frac{1}{2} \{y - m(\beta)\}^t A(\lambda)^{-1} \{y - m(\beta)\} / \sigma^2 - \frac{1}{2} n \log(2\pi),$$

and that the methods above lead to the profiled log-likelihood function

$$\ell_{\text{prof}}(\lambda) = -n \log \hat{\sigma}(\lambda) - \frac{1}{2} \log |A(\lambda)| - \frac{1}{2} n - \frac{1}{2} n \log(2\pi),$$

where one of each λ computes first $\hat{\beta}(\lambda) = (Z^t A(\lambda)^{-1} Z)^{-1} Z^t A(\lambda)^{-1} y$ and then $\hat{\sigma}^2 = Q_0(\lambda)/n$. Here $Q_0(\lambda) = Q(\hat{\beta}(\lambda), \lambda)$ with $Q(\beta, \lambda) = (y - Z\beta)^t A(\lambda)^{-1} (y - Z\beta)$.

(e) (xx can then find ML for λ , via the profiling, and then for β and σ . give an example to work through, perhaps the NewHaven dataset, for some correlations. xx)

(f) Consider the three-parameter stationary normal model for $Y_1, \dots, Y_n$, with mean μ , standard deviation σ , and autocorrelations $\rho^{|i-j|}$, for some correlation $\rho \in (-1, 1)$. The $A(\rho)$ matrix then has 1 on the diagonal and $\rho^{|i-j|}$ outside, as in Ex. 13.23. Give the log-likelihood function $\ell(\mu, \sigma, \rho)$, using the expressions from the exercise pointed to for the inverse and determinant of $A(\rho)$. Also give an expression for the profiled log-likelihood function $\ell_{\text{prof}}(\rho)$.

(g) This particular stationary case with autocorrelations $\rho^{|i-j|}$ is a classic time series model, where formulae above may be used to give simplified expressions for the log-likelihood $\ell(\mu, \sigma, \rho)$ and the profiled $\ell_{\text{prof}}(\rho)$. Using this will simplify and speed up computations considerably, for high n , but is not strictly necessary. Note however that the general model formulation, with $Y \sim N_n(\mu \mathbf{1}, \sigma^2 A(\rho))$, and the more general formulae above, can be used also in cases with gaps in the time series, or irregular sampling. (xx perhaps enough for now. we point to Story #37. it has a four-parameter model, with linear mean, a variance, and an autocorrelation. xx)

Ex. 13.25 *AR(1) series, I.* A classical and crucial result from traditional statistics is that if $Y_1, \dots, Y_n$ are independent with the same distribution, then $\text{Var } \bar{Y}_n = \sigma^2/n$ for the data average $\bar{Y}_n$, where σ^2 is the variance of a single observation. This is rather different for models with dependence. Suppose now that $Y_1, \dots, Y_n$ is a stationary sequence, with $\text{cov}(Y_i, Y_{i+h}) = \sigma^2 \rho(|h|)$, for some correlation function $\rho(h) = \text{corr}(Y_i, Y_{i+h})$; in particular, each Y_i has variance σ^2 .

(a) Show that

$$\text{Var } \bar{Y}_n = \frac{\sigma^2}{n} \left\{ 1 + 2 \sum_{h=1}^n (1 - h/n) \rho(h) \right\} = \frac{\sigma^2}{n} \sum_{h=-n}^n (1 - |h|/n) \rho(h).$$

(b) For the special case of $\rho(h) = \rho^h$, called autocorrelation of order 1, typically written AR(1), show that

$$\text{Var } \bar{Y}_n = \frac{\sigma^2}{n} \left\{ 1 + 2 \sum_{h=1}^{n-1} \rho^h - n^{-1} \sum_{h=1}^{n-1} h \rho^h \right\} = \frac{\sigma^2}{n} \left\{ \frac{1+\rho}{1-\rho} + O(1/n) \right\}.$$

With a positive autocorrelation, therefore, the variance of $\bar{Y}_n$ becomes clearly bigger than under independence.

(c) Suppose you observe such a stationary time series $y_1, \dots, y_n$, with autocorrelation function $\rho(h) = \rho^h$ for $h = 1, 2, 3, \dots$, and with unknown mean ξ , variance σ^2 , and autocorrelation parameter $\rho \in (-1, 1)$. If you do the traditional $\bar{y}_n \pm 1.96 \hat{\sigma}/\sqrt{n}$ interval for ξ , recommended in a hundred statistics books, with $\hat{\sigma}$ the empirical standard deviation, what will be its confidence coverage level? How should it be repaired? (xx to answer this point well one should know that $\bar{y}_n$ is becoming normal, which is becoming clearer in the following exercise. pointer to Story #38. xx)

(d) (xx take care, checking other parts of kiosk, to avoid repetition. but make the connection to the Ex. 13.23. xx)

Ex. 13.26 *AR(1) series, II.* (xx nils rant at the moment. will be combined with the above if sufficiently simple. but we will also connect to log-likelihoods under normality. seeing y_n in terms of innovations; normality via Lindeberg; t tests with corrected denominators. xx) Here we present a somewhat different view of AR(1) series, complementing what we have seen in Ex. 13.25 and 13.23. For simplicity of presentation we take the zero-mean case first. Instead of simply postulating the autocorrelation function $\text{corr}(Y_i, Y_{i+h}) = \sigma^2 \rho^{|h|}$, we build the model in terms of i.i.d. error variables ε_i , zero mean and variance σ_0^2 , and then write $Y_0 = 0$ and $Y_i = \rho Y_{i-1} + \varepsilon_i$ for $i = 2, 3, \dots$. We assume here that $\rho \in (-1, 1)$, which will be seen to imply that the series grows to its stationary stage, with variances and covariances stabilising.

(a) Show that $E(Y_i | Y_{i-1}) = \rho Y_i$ and $\text{Var}(Y_i | Y_{i-1}) = \sigma_0^2$, and deduce from this that $E Y_i = 0$ and that $\text{Var} Y_i = \sigma_0^2 + \rho^2 \text{Var} Y_{i-1}$. When the process has reached stationarity, show that $\text{Var} Y_i = \sigma^2 = \sigma_0^2 / (1 - \rho^2)$.

(b) More carefully, without assuming that the process has reached stationarity, show that $Y_n = \rho^{n-1} \varepsilon_1 + \rho^{n-2} \varepsilon_2 + \cdots + \rho \varepsilon_{n-1} + \varepsilon_n$, and that $\text{Var} Y_n = \sigma_0^2 (1 - \rho^{2n-1}) / (1 - \rho^2)$. For growing n we indeed discover the $\sigma^2 = \sigma_0^2 / (1 - \rho^2)$ relation. Find also a formula for $\text{cov}(Y_{n-1}, Y_n)$, and show that it tends to $\sigma^2 \rho$. Generalise.

(c) Writing $a_0 = 1$ and $a_m = 1 + \rho + \cdots + \rho^m = (1 - \rho^{m+1}) / (1 - \rho)$ for $m \geq 1$, show that

$$\begin{aligned} \sum_{i=1}^n Y_i &= a_{n-1} \varepsilon_1 + a_{n-2} \varepsilon_2 + \cdots + a_2 \varepsilon_{n-2} + a_1 \varepsilon_{n-1} + a_0 \varepsilon_n \\ &= (1 - \rho)^{-1} \{ (1 - \rho^n) \varepsilon_1 + (1 - \rho^{n-1}) \varepsilon_2 + \cdots + (1 - \rho^2) \varepsilon_{n-1} + (1 - \rho) \varepsilon_n \}. \end{aligned}$$

Use this to demonstrate that the distribution becomes normal in the limit, regardless of the distribution of the ε_i ; specifically that (xx check all this and make it clearer xx) $n^{-1/2} \sum_{i=1}^n Y_i \rightarrow_d N(0, \tau^2)$, with

$$\tau^2 = \sigma_0^2 / (1 - \rho)^2 = \sigma^2 (1 + \rho) / (1 - \rho).$$

Explain how this connects with Ex. 13.25.

(d) Then consider the case where the Y_i have mean ξ ; write $Y_0 = \xi$ and then $Y_i - \xi = \rho(Y_{i-1} - \xi) + \varepsilon_i$, with the ε_i i.i.d. with zero mean and variance σ_0^2 . Explain from the above that $\sqrt{n}(\bar{Y}_n - \xi) \rightarrow_d N(0, a^2 \sigma^2)$, with the extra factor $a^2 = (1 + \rho) / (1 - \rho)$. Comment on these aspects: (i) that this amounts to a CLT for autocorrelated observations, with a correction factor compared to the ordinary case of independence; and (ii) that we are able to prove the CLT by representing $\sum_{i=1}^n Y_i$ via the i.i.d. variables ε_i . Then explain how we may construct confidence intervals for ξ from autocorrelated observations $Y_1, \dots, Y_n$, and that the usual t test for say $\xi = 0$ needs a correction factor, like $t = \sqrt{n}(\bar{Y}_n - \xi_0) / (\hat{\sigma} \hat{a})$. How can one construct such an estimator $\hat{a}$ for a ?

Ex. 13.27 *Running t tests for an AR(1) process.* (xx to be done. needed for Kola and Hjort index story. xx) Suppose $Y_1, \dots, Y_n$ are observed, with $Y_i = \xi + \sigma \varepsilon_i$ and a zero-mean unit-variance AR(1) process for the ε_i . How can we monitor the constancy of the mean, with running t tests, comparing left to right?

(a) Before coming to the running t tests, consider a setup with two samples $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$, each following AR(1) processes as formulated above, with their separate $(\xi_1, \sigma_1, \rho_1)$ and $(\xi_2, \sigma_2, \rho_2)$. Argue that a natural test ratio, for assessing the hypothesis $\xi_1 = \xi_2$ of equal means, is

$$t = (\bar{Y} - \bar{X}) / \hat{\kappa}, \quad \text{with } \hat{\kappa}^2 = \hat{\kappa}_1^2 + \hat{\kappa}_2^2 = \frac{\hat{\sigma}_1^2}{n_1} \frac{1 + \hat{\rho}_1}{1 - \hat{\rho}_1} + \frac{\hat{\sigma}_2^2}{n_2} \frac{1 + \hat{\rho}_2}{1 - \hat{\rho}_2}.$$

Show indeed that for growing n_1 and n_2 , this becomes a standard normal under $\xi_1 = \xi_2$. (xx more care needed here; need to properly define $\hat{\sigma}$ and $\hat{\rho}$. xx)

(b) As in Ex. 10.46, for each potential changepoint τ , divide data into left and right, and compute averages $\bar{y}_L$ and $\bar{y}_R$, over respectively $y_1, \dots, y_\tau$ and $y_{\tau+1}, \dots, y_n$. Then construct $M_n(\tau) = (\bar{y}_R - \bar{y}_L) / (\hat{\kappa}_L^2 + \hat{\kappa}_R^2)^{1/2}$, with the appropriate denominator, from the above. Show both that $M_n(\tau) \rightarrow_d N(0, 1)$, for $\tau = j/n$ with $j/n \rightarrow s$, and also that there is full process convergence to $W^0(s) / \{s(1-s)\}^{1/2}$. (xx polish and round off. xx)

(c)

Ex. 13.28 *CLT with autocorrelation, and consequences for estimators.* (xx nils ranting at the moment. might be needed to round off Story #38. might beed a few details to carry this out nicely just by representing $\sum_{i=1}^n Y_i$ via the i.i.d. ε_i , i.e. without needing a CLT for dependent data per se. xx)

(a) Consider a stationary AR(1) sequence $Y_1, \dots, Y_n$, with mean ξ and covariances $\text{cov}(Y_i, Y_{i+h}) = \sigma^2 \rho^{|h|}$. We have seen in Ex. 13.26 that $\sqrt{n}(\bar{Y}_n - \xi) \rightarrow_d N(0, a^2 \sigma^2)$, with the extra scaling factor $a^2 = (1 + \rho)/(1 - \rho)$ caused by the autocorrelation. Generalise to the multivariate case, where Y_i with mean ξ is of dimension p , and the same autocorrelation parameter is at work for the full vector, i.e. $E(Y_i - \xi)(Y_{i+h} - \xi)^t = \rho^{|h|} \Sigma$ for $h = 0, \pm 1, \pm 2, \dots$. Show that $\sqrt{n}(\bar{Y}_n - \xi) \rightarrow_d N_p(0, a^2 \Sigma)$. Explain how the delta method yields a similar result, for each smooth function $\phi = g(\xi)$, with estimator $\hat{\phi}_n = g(\bar{Y}_n)$. In detail, $\sqrt{n}(\hat{\phi}_n - \phi) \rightarrow_d N(0, a^2 \tau^2)$, where $\tau^2 = c^t \Sigma c$ and $c = \partial g(\xi)/\partial \xi$. So the autocorrelation aspect of the Y_i follows through for smooth functions, and for estimators constructed based on such vector series.

(b) (xx hm, this is slightly messy, if there are different AR(1) parameters for the different components. if $Y_1, \dots, Y_n$ are autocorrelated from $\text{Gam}(a, b)$, then things happen for the autocorrelations for $(Y_i, \log Y_i)$, which is what we need to have on board for the $(\hat{a}, \hat{b})$. xx) (xx not easy to write down in simple fashion what happens to the empirical correlation coefficient $\hat{\phi}$, for an AR(1) series $(X_1, Y_1), (X_2, Y_2), \dots$ xx)

Other stuff & leftovers

Notes and pointers

must point to clear references for MCMC for continuous lemmas.

mention [Markov \(1906, 1913\)](#) with Evgenij Onegin 1913. a bit in [Hjort and Varin \(2008\)](#). point to [Billingsley \(1961\)](#).

mention [Cox and Miller \(1965\)](#), with sentence on simulations.

point to Gibbs Sampler and reversible jumps.

mention [Stoltenberg and Hjort \(2021\)](#) for uses in health inspection processes.

mention [Geman and Geman \(1984\)](#), [Casella and George \(1992\)](#), [Cox and Miller \(1965\)](#), [Shumway and Stoffer \(2016\)](#), ...

mention Tatyana and Paul Ehrenfest 1907 and point to OU and Story Pedro.

Bayesian nonparametrics

In Ch. 7 we went through the Bayesian basics, but there essentially limited to finite-parametric models. Conceptually speaking, the same ideas should carry over to bigger and infinite-dimensional models. This needs to involve probability models over distributions, regression functions, etc., which implies heavier lifting, both mathematically and algorithmically. Fuller analysis needs perhaps book-length treatments, but in the present chapter we go through Dirichlet processes for random distributions, Beta processes for random cumulative hazard functions, Gaussian processes for random curves and surfaces, illustrating their uses. These fat priors with large supports also lend themselves to showing nonparametric minimaxity and admissibility.

Key words: Bayesian kriging, Bayesian extensions of Nelson–Aalen and Kaplan–Meier, Beta processes, Dirichlet processes, gamma processes, infinite-dimensional parameters, nonparametric updating, Poisson processes, time-discrete survival analysis

A perhaps simplistic but nevertheless intriguing and informative two-by-two table is the following, meant to comprise the essence of most of statistical inference methodology:

	parametric	nonparametric
frequentist	I	II
Bayesian	III	IV = (I $\cup$ II $\cup$ III) ^c

The idea is that once one to a moderate extent has pinned down that is contained in boxes I, II, III, ‘everything else’ should be in box IV, tentatively defining Bayesian nonparametrics.

(I) ‘Frequentist parametrics’ encompasses the core of classical statistics, involving methods associated primarily with maximum likelihood, developed in the 1920ies and onwards; cf. our chapters 2, 3, 4, 5. Such methods relate to various optimum tests, with calculation of p-values, optimal estimators, confidence intervals, multiple comparisons, etc. Some of the procedures stem from exact probability calculations for models that are sufficiently amenable to mathematical derivations, while others relate to the application of large-sample techniques (central limit theorems, delta methods, higher-order corrections involving expansions or saddle-point approximations, etc.).

(II) ‘Bayesian parametrics’ correspondingly comprises classic methodology for prior and posterior distributions in models with a finite (and often low) number of parameters, as in our Ch. 7. Such methods, starting from the premise that uncertainty about

model parameters somehow may be represented in terms of probability distributions, have arguably been in existence for more than a hundred years (since the basic theorem that drives the machinery simply says that the posterior density is proportional to the product of the prior density with the likelihood function, which again relates to the Bayes theorem of ca. 1763), but were naturally quite limited to a short list of sufficiently simple statistical models and priors. The applicability of Bayesian parametrics widened significantly with the advent and availability of modern computers, say from ca. 1975 and onwards, and then with the development of further numerical methods and software packages pertaining to numerical integration and Markov chain Monte Carlo (MCMC) simulations, say from ca. 1990 and onwards.

(III) ‘Frequentist nonparametrics’ is a somewhat mixed bag, covering various different areas of statistics. The term has historically been associated with various test procedures that are or asymptotically become ‘distribution free’, leading also to nonparametric confidence intervals and bands, etc.; for methodology related to statistics based on ranks (cf. Lehmann, 1975); then progressively with estimation of probability densities, regression functions, link functions etc., without parametric assumptions; and also with specific computational techniques such as the bootstrap. Again, asymptotics plays an important role, both for developing fruitful approximations and for understanding and comparing properties of performance.

(IV) What ostensibly appears to remain for our fourth category, then, to be handled in the present chapter, are models and methods characterised by (a) big parameter spaces (unknown density and regression functions, link and response functions, etc.) and (b) construction of probability measures over these spaces. Typical examples include Bayesian set-ups for density estimation (in any dimension), nonparametric regression with a fixed error distribution, hazard rate and survival function estimation for survival analysis, without or with covariates, etc. The division between ‘small’ and ‘moderate’ and ‘big’ for parameter spaces is not meant to be very sharp, and the scale is interpreted flexibly.

Crucially, the Bayesian principles laid out and developed in Ch. 7 continue to hold, regarding loss functions, decision rules, risk functions, the Master Recipe; the Bayes solution remains that of minimising the posterior expected risk $E\{L(P, a) | \text{data}\}$. What is different in Bayesian nonparametrics, and more demanding, is that the space of models P is bigger; the construction of both the loss function and of adequate priors for such P is harder; and finding the posterior $P | \text{data}$ sometimes much harder. We do manage, though, in a series of cases, and the MCMC machinery of Ch. 7 is again helpful, when it comes to posterior simulation of $\theta(P)$ for vital parameters.

(xx a sentence here or below saying that we partly use P and F interchangeably, working nonparametrically with a distribution. xx)

Methods of Ch. 10 are useful when it comes to defining and working with full processes here, as opposed to the finite-parameter models of Ch. 7. Material covered below include gamma processes, often used as prior processes for Poisson data; Dirichlet processes, priors for unknown c.d.f.s; Beta processes, as priors for unknown cumulative hazard functions in survival and event history settings; and Gaussian processes, priors for trend functions. (xx point to various stories where BNP is used. xx)

Gamma processes and Poisson processes

Ex. 14.1 *Gamma and Poisson: from a fine grid to processes.* We have seen the basic

machinery of gamma priors for Poisson data in Ex. 7.2; so we can deal with a sequence $(\theta_1, y_1), \dots, (\theta_m, y_m)$. In various settings there is a need to extend this to full processes.

(a) Suppose first that $Y_1, \dots, Y_m$ are independent Poissons, with parameters $w_1\theta_1, \dots, w_m\theta_m$. With independent $\theta_i \sim \text{Gam}(a_i, b_i)$ as priors for these, show that $\theta_i | \text{data} \sim \text{Gam}(a_i + y_i, b_i + w_i)$. For $G_j(t) = \sum_{i \leq j} \theta_i$, show that

$$\mathbb{E}\{G_j(t) | \text{data}\} = \sum_{i \leq j} \frac{a_i + y_i}{b_i + w_i}, \quad \text{Var}\{G_j(t) | \text{data}\} = \sum_{i \leq j} \frac{a_i + y_i}{(b_i + w_i)^2}.$$

(b) The idea now is to construct ‘fine limits’ of the gamma and Poisson above. We start properly with the gamma part. We have seen how gamma processes can be constructed in Ex. 10.21, obtained as limits Z in $D[0, 1]$ of well-defined fine-grid processes Z_n . This is now generalised. First, if $V \sim \text{Gam}(A, b)$, show that its Laplace transform becomes

$$\mathbb{E} \exp(-uV) = b^A / (b + u)^A = \exp\{-A \log(1 + u/b)\}.$$

Consider $A(t) = \int_0^t a(s) ds$, for some nonnegative function $a(s)$, and define $G_m(t) = \sum_{i \leq [mt]} \theta_i$, with independent small Gamma contributions $\theta_i \sim \text{Gam}(a(i/m)(1/m), b)$. Show first that $\mathbb{E} G_m(t) \rightarrow A(t)/b$ and $\text{Var} G_m(t) \rightarrow A(t)/b^2$, and then that the Laplace transform converges,

$$\mathbb{E} \exp\{-uG_m(t)\} = \prod_{i/m \leq t} \exp\{-a(i/m)(1/m) \log(1 + u/b)\} \rightarrow \exp\{-A(t) \log(1 + u/b)\}.$$

Show as for the simpler case of $\alpha(s) = 1$ and $b = 1$, dealt with in Ex. 10.21, that there is finite-dimensional convergence for G_m , that it is tight, implying convergence $G_m \rightarrow_d G$ to a Gamma process with independent increments $G(t) - G(s) \sim \text{Gam}(A(t) - A(s), b)$.

Ex. 14.2 Extended Gamma processes. The construction used in Ex. 14.1 led to well-defined Gamma processes, of the type $Z(t) \sim \text{Gam}(A(t), b)$, with $A(t) = \int_0^t a(s) ds$. This may be extended further to the case of non-constant shape parameter b .

(a) Start with independent $\theta_i \sim \text{Gam}(a(i/m)(1/m), b(i/m))$, for continuous positive functions $a(s), b(s)$, and define $G_m(t) = \sum_{i/m \leq t} \theta_i$ for $t \geq 0$. Show that the mean and variance converge,

$$\begin{aligned} \mathbb{E} G_m(t) &= \sum_{i/m \leq t} \frac{a(i/m)(1/m)}{b(i/m)} \rightarrow \int_0^t \frac{a(s)}{b(s)} ds, \\ \text{Var} G_m(t) &= \sum_{i/m \leq t} \frac{a(i/m)(1/m)}{b(i/m)^2} \rightarrow \int_0^t \frac{a(s)}{b(s)^2} ds. \end{aligned}$$

(b) Then show that there is convergence of the Laplace transform,

$$\mathbb{E} \exp\{-uG_m(t)\} = \prod_{i/m \leq t} \mathbb{E} \exp(-uV_i) \rightarrow \exp\left\{-\int_0^t a(s) \log(1 + u/b(s)) ds\right\}.$$

Use this to argue that there is finite-dimensional convergence in distribution for G_m .

(c) Demonstrate tightness of G_m via Ex. 10.21. This leads to the clear existence of such an extended Gamma process, with mean and variance found above. Note however that increments $Z(t) - Z(s)$ are not Gamma distributed; show that this would require b to be constant over that interval.

Ex. 14.3 *Poisson process observations of a Gamma process.* As in Ex. 14.2, consider an extended gamma process $G_m(t) = \sum_{i/m \leq t} \theta_i$, with independent summands $\theta_i \sim \text{Gam}(a(i/m)(1/m), b(i/m))$. Given the G_m , let now $H_m(t) = \sum_{i/m \leq t} Y_i$, where the Y_i are mostly small Poissons, with means θ_i .

(a) We have seen that $G_m \rightarrow_d G$, a well-defined extended gamma. Explain also that given G_m , H_m is tending to a Poisson process H , with marginal mean function tending to $\int_0^t a(s)/b(s) ds$. Show that $G_m | H_m$ is a sum of updated gammas, with $\theta_i | H_m \sim \text{Gam}(a(i/m)(1/m) + y_i, b(i/m) + 1)$. In particular, show

$$\hat{G}_m(t) = E\{G_m(t) | \text{data}\} = \sum_{i/m \leq t} \frac{a(i/m)(1/m) + y_i}{b(i/m) + 1} \rightarrow \int_0^t \frac{a(s) ds + dH(s)}{b(s) + 1}.$$

(b) Via conditional mean and variance for the H_m , show next that

$$\begin{aligned} E H_m(t) &= \sum_{i/m \leq t} \frac{a(i/m)(1/m)}{b(i/m)} \rightarrow \int_0^t \frac{a(s)}{b(s)} ds, \\ \text{Var } H_m(t) &= \sum_{i/m \leq t} \left\{ \frac{a(i/m)(1/m)}{b(i/m)} + \frac{a(i/m)(1/m)}{b(i/m)^2} \right\} \rightarrow \int_0^t \left\{ \frac{a(s)}{b(s)} + \frac{a(s)}{b(s)^2} \right\} ds. \end{aligned}$$

For $v \geq 0$, show also first that

$$E[\exp\{-vH_m(t)\} | G_m] = \prod_{i/m \leq t} \exp[-\theta_i\{1 - \exp(-v)\}],$$

and from this that

$$\begin{aligned} E \exp\{-vH_m(t)\} &= \prod_{i/m \leq t} \exp[-a(i/m)(1/m) \log(1 + \{1 - \exp(-v)\}/b(i/m))] \\ &\rightarrow \exp\left[-\int_0^t a(s) \log(1 + \{1 - \exp(-v)\}/b(s)) ds\right]. \end{aligned}$$

(c) Show that the process (G_m, H_m) in $D([0, 1]^2)$ has independent increments, and identify the limit of $E \exp\{-uG_m(t) - vH_m(t)\}$ for nonnegative u, v . Using details from Ex. 10.21, show that there is a well-defined limit process (G, H) to which (G_m, H_m) converges.

(d) Using infinitesimal intuition and notation, the idea is to model an intensity G with Poisson counting H on top, via small independent increments

$$dG(s) \sim \text{Gam}(c(s)a(s) ds, c(s)) \quad \text{and} \quad dH(s) | dG(s) \sim \text{Pois}(dG(s)),$$

Argue that this may be made precise, and that

$$dG(s) | H \sim \text{Gam}(c(s)a(s) ds + dH(s), c(s) + 1).$$

(xx can put in a point on the Emil-Nils Police Twitter model, with covariates and β parameters. xx)

Ex. 14.4 *A Gamma process of a Poisson process.* Let $X = \{X(t): t \geq 0\}$ be a Poisson process with rate λ , and let $Y = \{Y(t): t \geq 0\}$ be a Gamma process with $Y(t) \sim \text{Gam}(at, b)$. From these define the combined process $Z(t) = Y(X(t))$. Find its mean and variance. [xx a bit more. show that it can be reached as a limit of $Y_n(X_n(t))$, with small things placed at i/m for $i = 0, 1, 2, \dots$ also its distribution. do some simulations. xx]

Ex. 14.5 *Gamma processes: their jumps.* The paths of Gamma processes are increasing but not continuous; one may indeed show that there are infinitely many random jumps on each interval.

(a) To work with the size of the biggest jump, consider $X_\varepsilon \sim \text{Gam}(\varepsilon, 1)$, with ε small. Writing Γ_ε for its c.d.f., i.e. $\Gamma_\varepsilon(v) = G(v, \varepsilon, 1)$ in terms of the Gamma c.d.f. with parameters $(\varepsilon, 1)$, show that

$$\Gamma_\varepsilon(v) = 1 - \int_v^\infty \frac{1}{\Gamma(\varepsilon)} x^{\varepsilon-1} \exp(-x) dx = 1 - \varepsilon \int_v^\infty \frac{1}{\Gamma(1+\varepsilon)} x^\varepsilon x^{-1} \exp(-x) dx,$$

and that this may be written $1 - \varepsilon E_1(v)\{1 + O(\varepsilon)\}$, in terms of the exponential integral function $E_1(v) = \int_v^\infty x^{-1} \exp(-x) dx$.

(b) Let now $J_m(t)$ be the biggest jump for Z_m , in the construction above, over the window $[0, t]$. Show that $\Pr(J_m(t) \leq v) = \prod_{i/m \leq t} \Gamma_{a/m}(v)$, and that this converges to $\exp\{-aE_1(v)\}$ for $v > 0$. Show that the median $v_{0.50}$ in this biggest jump distribution is the solution to $E_1(v_{0.50}) = (\log 2)/a$.

The Dirichlet process

Ex. 14.6 *Dirichlet and multinomial: from a fine grid to processes.* In Ex. 14.1 we saw natural limit processes coming out of a sequence of gamma-Poisson pairs, seen on an increasingly fine grid. Here we attempt the same with Dirichlet-multinomial modelling. For distributions on $[0, 1]$, form a fine grid of cells $C_i = ((i-1)/m, i/m]$, and consider the probability vector $p = (p_1, \dots, p_m)$, with these being seen as $p_i = F(C_i)$ for an unknown c.d.f. F .

(a) Give the probability vector p a $\text{Dir}(ap_0)$ prior, with $p_0 = (p_{0,1}, \dots, p_{0,m})$, where $p_{0,i} = F_0(C_i)$ for some prior guess type F_0 . With first a single Y observed from this distribution, landing in cell i with probability p_i , explain that $p | y_{\text{obs}} \sim \text{Dir}(ap_{0,1}, \dots, ap_{0,i} + 1, \dots, ap_{0,m})$, where i is the cell landed in.

(b) With more data, continue updating: with $Y_1, \dots, Y_n$ i.i.d. for given p , landing in cells $1, \dots, m$, with these probabilities, and using the $\text{Dir}(ap_0)$ prior, explain that

$$p | \text{data} \sim \text{Dir}(ap_{0,1} + N_1, \dots, ap_{0,m} + N_m),$$

with N_i the number of datapoints falling in cell i . This is classic Dirichlet-multinomial updating, as per Ex. 7.11.

(c) It then ought to be possible to ‘go to a fine limit’. (i) Argue heuristically that this would mean a random c.d.f. F , with the property that $(p_1, \dots, p_m) \sim \text{Dir}(ap_{0,1}, \dots, ap_{0,m})$, for $p_i = F(C_i)$, where $p_{0,i} = F_0(C_i)$ for some prior guess c.d.f. F_0 . We need this to hold for every selection of cells. (ii) Argue boldly further that with (i) properly constructed, one should have $F | \text{data}$ another random c.d.f. of the same type, but with $p | \text{data}$ as above, with an extra N_i for cell i . If all cells are small, and data are continuous, there will be an extra ‘1’ for the perhaps few cells being visited by the data. – We have now seen heuristic motivation for the Dirichlet process, and tend to (i) below, with (ii) worked with in Ex. 14.10.

(d) To be used in a minute: if $(U, V, W) \sim \text{Dir}(ap, aq, ar)$, with $p + q + r = 1$, show that $EU V = \{a/(a+1)\}pq$; see Ex. 14.7. Regarding (i) in the previous point, consider for given a and F_0 the random c.d.f. F_m in $D[0, 1]$ with value $p_1 + \dots + p_i$ at i/m , and $p_1 + \dots + p_{i-1}$ just to the left, i.e. on $((i-1)/m, i/m]$. where $p = (p_1, \dots, p_m)$ is $\text{Dir}(ap_{0,1}, \dots, ap_{0,k})$,

as above. Show that for $t_1 < \dots < t_k = 1$, there is convergence in distribution of $(F_m(t_1), \dots, F_m(t_k))$ to that of $(D_1, D_1 + D_1, \dots, D_1 + \dots + D_k)$, where $(D_1, \dots, D_k) \sim \text{Dir}(aF_0(t_1), aF_0(t_1, t_2]), \dots, aF_0(t_{k-1}, t_k])$. Also show via (10.3) that the F_m sequence is tight. Conclude that yes, there is proper convergence $F_m \rightarrow_d F$ in $D[0, 1]$, a random c.d.f. with the Dirichlet property that $(F(I_1), \dots, F(I_k)) \sim \text{Dir}(aF_0(I_1), \dots, aF_0(I_k))$ for cells $I_i = (t_{i-1}, t_i]$. We say that F is a Dirichlet process, and write $F \sim \text{Dir}(aF_0)$.

a Dirichlet process on the unit interval

(e) Another and more direct construction is as follows. Let G be a gamma process on $[0, 1]$, with $G(t) \sim \text{Gam}(aF_0(t), 1)$. Show that $F = G/G(1)$ has the Dirichlet property, essentially due to Ex. 1.20. Since a c.d.f. H determines not only interval probabilities $H(t) - H(s)$, but all probabilities $P(A) = \int_A dH(s)$, argue that a Dirichlet process $F \sim \text{Dir}(aF_0)$ similarly determines the joint probability distribution of any $(P(A_1), \dots, P(A_k))$, writing P for the probability measure associated with F .

Ex. 14.7 *The Dirichlet process.* Let in general terms P_0 be a probability measure on some measurable space $(\mathcal{X}, \mathcal{A})$, and with a a positive number. Then there exists a random P , on the space of all probability measures on $(\mathcal{X}, \mathcal{A})$, with the remarkable property that for any partition $A_1, \dots, A_k$, i.e. sets whose disjoint union is $\mathcal{X}$, we have

$$(D_1, \dots, D_k) = (P(A_1), \dots, P(A_k)) \sim \text{Dir}(aP_0(A_1), \dots, aP_0(A_k)). \quad (14.1)$$

the aP_0 is the base measure for the Dirichlet process

We denote such a random P a Dirichlet process, with base measure aP_0 , and write $P \sim \text{Dir}(aP_0)$. Arguments from Ex. 14.6 can be extended from the unit interval to prove the existence of such a $\text{Dir}(aP_0)$ for the full real line, and to Euclidean space; existence is guaranteed also for fully general spaces.

see Notes and Pointers

(a) From the definition above, show that it is logically consistent, in the sense that with sums over small cells, the Dirichlet structure remains. In symbols, with $B_1, \dots, B_m$ being formed as unions over disjoint subsets of the A_j , with $m < k$, we have the property above, for $(P(B_1), \dots, P(B_m))$. Explain also that this summing over smaller cells property is required for any construction of a random P .

(b) Consider any subset A with $p_0 = P_0(A)$. Show that

$$P(A) \sim \text{Beta}(ap_0, a(1 - p_0)) \quad \text{with } E P(A) = p_0, \text{Var } P(A) = p_0(1 - p_0)/(a + 1).$$

(c) For any partition $A_1, \dots, A_k$, and writing $p_{0,j} = P_0(A_j)$, show that

$$E P(A_1)^{j_1} \dots P(A_k)^{j_k} = \frac{\Gamma(a)}{\Gamma(ap_{0,1}) \dots \Gamma(ap_{0,k})} \frac{\Gamma(ap_{0,1} + j_1) \dots \Gamma(ap_{0,k} + j_k)}{\Gamma(a + j_1 + \dots + j_k)}.$$

Via this general formula, suppose $(U, V, W) \sim \text{Dir}(ap, aq, ar)$. Show first that

$$E UV = \frac{\Gamma(a)}{\Gamma(ap)\Gamma(aq)\Gamma(ar)} \frac{\Gamma(ap + 1)\Gamma(aq + 1)\Gamma(ar)}{\Gamma(a + 2)} = \frac{ap aq}{a(a + 1)} = \frac{pq}{a + 1},$$

and from this that $\text{cov}(U, V) = pq(1/(a + 1) - (a + 1)/(a + 1)) = -pq/(a + 1)$.

Ex. 14.8 *Support of the Dirichlet process.* For $\mathcal{D}_{aP_0}$ the probability measure for $P \sim \text{Dir}(aP_0)$, the support of the distribution is the smallest closed set $\mathcal{M}^*$ such that $\mathcal{D}_{aP_0}(\mathcal{M}^*) = 1$. We learn here that the support of the Dirichlet process is big.

(a) The support notion, with the closed sets, depends on the topology being used. Let that topology being defined via weak convergence, so $Q_n \rightarrow Q$ means convergence in distribution. Explain that a basis for neighbourhoods under this topology is the set of

check the Portmanteau theorem of Ch. 2

$$U = U(Q, A_1, \dots, A_k, \varepsilon_1, \dots, \varepsilon_k) = \{Q' : |Q'(A_j) - Q(A_j)| < \varepsilon_j \text{ for } j=1, \dots, k\},$$

with k any integer, the A_j measurable subsets, and the ε_j positive, and the crucial condition that the A_j are Q -continuous, i.e. $Q(\partial A_j) = 0$, with $\partial A_j = \bar{A}_j - A_j^-$ the boundary set. Show

$$\text{supp}(\mathcal{D}_{aP_0}) = \{Q : \text{supp}(Q) \subset \text{supp}(P_0)\}.$$

(b) Then shift the topological perspective to strong convergence, where $Q_n \rightarrow Q$ means $Q_n(A) \rightarrow Q(A)$ for all measurable A . Show $\text{supp}(\mathcal{D}_{aP_0}) = \{Q : Q \ll P_0\}$.

Ex. 14.9 *Sampling from a Dirichlet process.* Consider a sample space $\mathcal{X}$, taken to be a metric space with d the metric, and $\mathcal{A}$ the sigma-algebra of all measurable subsets of $\mathcal{X}$, generated by the open sets. With $P \sim \text{Dir}(aP_0)$ on this $(\mathcal{X}, \mathcal{A})$, it is drawn from a rather complex probability distribution $\mathcal{D}_{aP_0}$ on the big set $\mathcal{M}$ of all probability distributions on $(\mathcal{X}, \mathcal{A})$. Now suppose X is drawn from that P drawn from the Dirichlet distribution.

(a) Show via double expectation that the marginal distribution of X is simply P_0 . In particular, its mean and variance are $\xi_0 = \int x dP_0$ and $\sigma_0^2 = \int (x - \xi_0)^2 dP_0$.

(b) Consider next X, Y an independent pair from a random P drawn from the $\text{Dir}(aP_0)$ prior, and write Q_2 for the marginal distribution of the pair. Show that

$$Q_2(A \times B) = \Pr(X \in A, Y \in B) = \int P(A)P(B) \mathcal{D}_{aP_0}(dP) = \mathbb{E} P(A)P(B),$$

with expectation operator that of the $\text{Dir}(aP_0)$. This becomes an exercise in Dirichlet moments; write $A \cup B$ as the disjoint union $C \cup D \cup E$, with $C = A - B$, $D = A \cap B$, $E = B - A$, and show that

$$Q_2(A \times B) = \mathbb{E} \{P(C) + P(D)\} \{P(D) + P(E)\} = \mathbb{E} (U + V)(V + W),$$

say. Here $(U, V, W, T) \sim \text{Dir}(ap, aq, ar, a - ap - aq - ar)$, with $U + V + W + T = 1$, and $p = P_0(C)$, $q = P_0(D)$, $r = P_0(E)$. Do the algebra to find

$$\begin{aligned} Q_2(A \times B) &= \frac{ap aq}{a(a+1)} + \frac{ap ar}{a(a+1)} + \frac{aq ar}{a(a+1)} + \frac{aq(aq+1)}{a(a+1)} \\ &= \frac{a}{a+1} P_0(A)P_0(B) + \frac{1}{a+1} P_0(A \cap B). \end{aligned}$$

Use this to show that

$$Q_2 = a/(a+1) P_0 \times P_0 + 1/(a+1) P_{1,2},$$

in which $P_{1,2}$ is the probability measure concentrated on the diagonal, with $X \sim P_0$ and then $X = Y$. The first part, with $P_0^2 = P_0 \times P_0$, corresponds to X, Y being independent from P_0 . Explain that this means that with probability $a/(a+1)$, the X, Y are independent from P_0 , but that with probability $1/(a+1)$, $X = Y$, from P_0 . Discuss the implications for large a and small a .

(c) Let Q_n be the marginal distribution of a Dirichlet sample $X_1, \dots, X_n$. Show that

$$Q_n(A_1 \times \dots \times A_n) = \mathbb{E} P(A_1) \dots P(A_n),$$

with the expectation being under $P \sim \text{Dir}(aP_0)$. Show

$$Q_3 = \frac{a^2}{(a+1)(a+2)} P_0^3 + \frac{a}{(a+1)(a+2)} (P_{1,2} + P_{1,3} + P_{2,3}) + \frac{3}{(a+1)(a+2)} P_{1,2,3}.$$

Explain that this can be translated to the following, for the marginal distribution of (X, Y, Z) , when these are independently sampled from the random P : with probability $a^2/(a+1)(a+2)$, they are independent from P_0 ; with probability $a/(a+1)(a+2)$, we have either $X = Y$ then Z , or $X = Z$ then Y , or X then $Y = Z$, with the two independent from P_0 ; and with probability $3/(a+1)(a+2)$, we have $X = Y = Z$, from P_0 .

(d) (xx to come. xx) $Q_{n+1}(A \times B)$ etc.

(e) Assume here that the prior mean distribution P_0 is free of atoms, i.e. $P_0(\{x\}) = 0$ for all x . For $n = 2$, show that $X = Y$ with probability $1/(a+1)$. For $n = 3$, show that the number of distinct values is 3, 2, 1, with probabilities $a^2/(a+1)(a+2)$, $3a/(a+1)(a+2)$, $1/(a+1)(a+2)$. Let D_n be the number of distinct values in the sample $X_1, \dots, X_n$. We have $\Pr(D_n = n) = a^n/(a+1) \cdots (a+n-1)$. More: $D_n = R_1 + \cdots + R_n$, independent Bernoullis, with decreasing probability $\Pr(R_i = 1) = a/(a+i-1)$. $\hat{a} = D_n/\log n$ tends to a , with normality, via Lindeberg, etc.

Ex. 14.10 *The updating scheme for Dirichlet processes.* Consider the setup of Ex. 14.9, with $X_1, \dots, X_n$ sampled from a random P , where $P \sim \text{Dir}(aP_0)$. Here we learn that $P \mid \text{data}$ is another Dirichlet process, with updated base measure $aP_0 + \delta(x_1) + \cdots + \delta(x_n)$, adding a unit point mass at each new data point. As anticipated in Ex. 14.6, this is an infinite-dimensional generalisation of the basic updating result for the Dirichlet-multinomial setup.

so $\delta(x)$ means
the measure
having unit
point mass at
position x

(a) Starting with a single X , we are indeed to prove that when $X \mid P$ has distribution P , and $P \sim \text{Dir}(aP_0)$, then $P \mid (X = x) \sim \text{Dir}(aP_0 + \delta(x))$, where $\delta(x)$ is unit point mass at x . Consider a measurable partition $A_1, \dots, A_k$, and suppose without losing generality that $x \in A_1$. Write $p_j = P(A_j)$ for $j = 1, \dots, k$. Via the defining property (14.1), explain that it suffices to prove that

$$(p_1, \dots, p_k) \mid (X = x) \sim \text{Dir}(aP_0(A_1) + 1, aP_0(A_2), \dots, aP_0(A_k)).$$

View this as the probability vector for a multinomial $(I_1, \dots, I_k)$, say. Argue that the information $x \in A_1$ is the same as $I_1 = 1$, and complete the updating argument. We are essentially learning about the general case via the finite-dimensional Dirichlet updating theorem.

(b) Then suppose that given P , the sample $X_1, \dots, X_n$ are i.i.d. from P . Use induction to show that $P \mid (x_1, \dots, x_n) \sim \text{Dir}(aP_0 + \delta(x_1) + \cdots + \delta(x_n))$. Explain in particular that $P(A) \mid x_1, \dots, x_n$ is a Beta, with parameters $(aP_0(A) + nF_n(A), aP_0(A^c) + nF_n(A^c))$, with F_n the empirical c.d.f. for the data. From this again, show

$$\begin{aligned}\hat{F}_n &= E(F \mid \text{data}) = \frac{a}{a+n} F_0 + \frac{n}{a+n} F_n, \\ \text{Var}(F \mid \text{data}) &= \hat{F}_n(1 - \hat{F}_n)/(n + a + 1).\end{aligned}$$

Ex. 14.11 *Dirichlet process inference via simulation.* It is useful and almost necessary to be able to simulate Dirichlet processes, e.g. for carrying out Bayesian nonparametric inference for a given dataset. For an illustration we are using the simple dataset worked with in Story #18, with alpha waves EEG measurements for two groups of inmates in a Canadian prison, one with normal confinement, the other after a five-day solitary confinement. For the present exercise we use the data $X_1, \dots, X_n$ from the first group; see the story pointed to for more information and analysis.

```
X from F: nonconfined      10.7 10.7 10.4 10.9 10.5 10.3  9.6 11.1 11.2 10.4
Y from G: solitary confinement 9.6 10.4  9.7 10.3  9.2  9.3  9.9  9.5  9.0 10.9
```

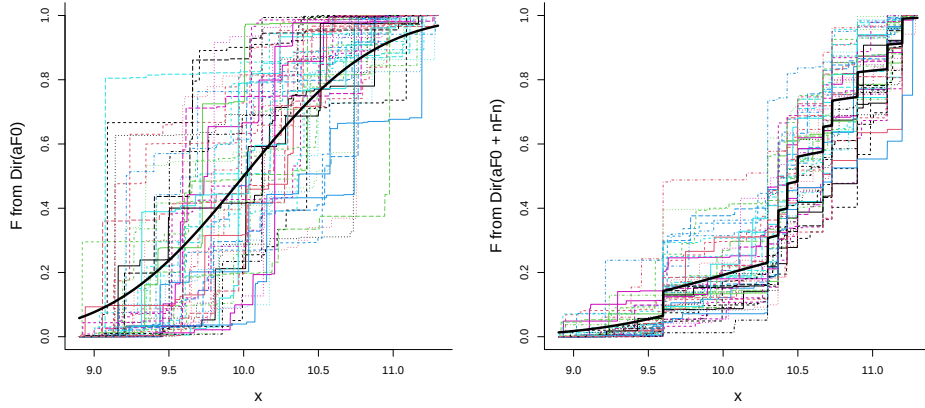


Figure 14.1: *Dirichlet process simulations: 50 simulated c.d.f.s F , from the prior $\text{Dir}(aF_0)$ (left panel) and the posterior $\text{Dir}(aF_0 + nF_n)$ (right panel), for the data $x_1, \dots, x_n$ with $n = 10$. Here $a = 3$ with prior guess $F_0 \sim N(10.0, 0.7^2)$. The fat smooth curves in the middle of the two plots are the mean functions, i.e. F_0 for the prior and $(aF_0 + nF_n)/(a + n)$ for the posterior.*

(a) For the purposes of a clean illustration, avoiding some consequences of ties, we slightly arbitrarily give the two X_i values of 10.7 new values 10.67 and 10.73, and two values of 10.4 new values 10.37 and 10.43. Now for the random $F \sim \text{Dir}(aF_0)$ on the real line, split the interval in question into m small boxes, say $z_j \pm \frac{1}{2}\delta$ for a suitably small δ . Then represent the F approximation as the cumulative sums of $(p_1, \dots, p_m) \sim \text{Dir}(ap_{0,1}, \dots, ap_{0,m})$, with $p_{0,j} = F_0(z_j + \frac{1}{2}\delta) - F_0(z_j - \frac{1}{2}\delta)$. Carry out this for the prior $\text{Dir}(aF_0)$, with tiny boxes covering the interval $[8.90, 11.30]$, from the Dirichlet prior, using prior sample size $a = 3$ and $F_0 = N(10.0, 0.7^2)$, a reasonably informative prior with prior guess parameters suitable for a representative sample for the broader population. Construct a version of Figure 14.1, left panel, with 50 simulated prior F .

(b) Then, with the same tiny boxes and their midpoints, and perhaps partly using the same computer code, simulate 50 realisations of F | data, where the Dirichlet process parameter is $aF_0 + nF_n$, as per Ex. 14.10, with F_n the empirical c.d.f. for the $n = 10$ datapoints. Construct a version of Figure 14.1, right panel. Comment on the learning jump from the left to the right panel, i.e. from the prior to the posterior. Play a bit with your computer code, to see what happens if the prior used is ‘more wrong’ than in this example.

(c) Check with Story #18 to see how this simulation apparatus may be used to compute and plot the posterior distributions for relevant functionals of F and G , for the two groups, like the median for X minus the median for G . Invent some parameter of interest, read off the posterior from the simulations, and plot.

Ex. 14.12 *The discrete charm of the Dirichlet process.* Consider $P \sim \text{Dir}(aP_0)$ on some space. Even when P_0 is smooth, P itself will be discrete, with probability 1.

(a) With P drawn from the $\text{Dir}(aP_0)$ and then $X | P$ drawn from P , explain via the updating theorem that for any measurable $f(P, x)$, we have

$$\mathbb{E} f(P, X) = \int_{\mathcal{M}} \int_{\mathcal{X}} f(P, x) P(\mathrm{d}x) \mathcal{D}_{aP_0}(\mathrm{d}P) = \int_{\mathcal{X}} \int_{\mathcal{M}} f(P, x) \mathcal{D}_{aP_0 + \delta(x)}(\mathrm{d}P) P_0(\mathrm{d}x).$$

(b) Introduce $A_P = \{x: P\{x\} > 0\}$, the set of atoms for a given P , and define

$$H_\gamma(P) = \mathbb{E}_P P(\{X\})^\gamma = \int P\{x\}^\gamma dP(x) = \sum_{x \in A_P} P\{x\}^{\gamma+1},$$

and its limit $H_0(P)$. So P is fully discrete if and only if $H_0(P) = 1$.

$$\mathbb{E} H_\gamma(P) = \mathbb{E} P\{X\}^\gamma = \frac{\Gamma(1+\gamma)}{\Gamma(1)} \frac{\Gamma(a+1)}{\Gamma(a+1+\gamma)}.$$

Deduce $H_0(P) = 1$ a.s., i.e. P from a Dirichlet process is discrete with probability 1.

Ex. 14.13 *The Dirichlet process as a symmetric limit.* Consider a probability measure P_0 on a sample space $\mathcal{X}$.

(a) For $\xi_1, \dots, \xi_m$ drawn i.i.d. from P_0 , give them random weights $(\beta_1, \dots, \beta_m)$, from the symmetric $\text{Dir}(a/m, \dots, a/m)$. This yields the random probability measure $P_m = \sum_{j=1}^m \beta_j \delta(\xi_j)$. Conditional on the $\xi_1, \dots, \xi_m$, explain that for any subset A , the $P_m(A) = \sum_{\xi_j \in A} \beta_j$ is a $\text{Beta}(a\hat{P}_m(A), a\hat{P}_m(A^c))$, where $\hat{P}_m(A) = (1/m) \sum_{j=1}^m I(\xi_j \in A)$ is the empirical distribution. Show from this that

$$P_m(A) \rightarrow_d \text{Beta}(aP_0(A), aP_0(A^c)) \quad \text{a.s., as } m \rightarrow \infty.$$

(b) Extend the argument to show that for any partition $(A_1, \dots, A_k)$, we have

$$(P_m(A_1), \dots, P_m(A_k)) \rightarrow_d (D_1, \dots, D_m) \sim \text{Dir}(aP_0(A_1), \dots, aP_0(A_k)) \quad \text{a.s.,}$$

i.e. the correct finite-dimensional distributions of the $\text{Dir}(aP_0)$ process.

(c) To illustrate how this can be useful, consider a random Dirichlet mean $\theta = \mathbb{E}_P g(X) = \int g dP$. Show that its distribution is the limit of $\theta_m = m^{-1} \sum_{j=1}^m g(\xi_j) \beta_j$, Dirichlet-symmetric means of the $g(\xi_j)$. This may be used to identify the θ distribution. For a particular case of interest, let P_0 be the standard Cauchy, where we have $\mathbb{E} \{\exp(it\beta_j \xi_j) \mid \beta\} = \exp(-|t|\beta_j)$. Show that the random mean $\theta = \int x dP$ is Cauchy. various other examples in Hjort and Ongaro (2005)

Nonparametric Bayesian estimators via the Dirichlet process

Ex. 14.14 *Bayesian estimation via the Dirichlet process.* With a sample $X_1, \dots, X_n$ from P , how should we estimate the different $\theta(P)$ of interest, with P having a Dirichlet process prior?

(a) With squared error loss $(\hat{\theta} - \theta)^2$, verify that the Bayes solution is $\hat{\theta} = \mathbb{E} \{\theta(P) \mid \text{data}\}$. In view of Ex. 14.10, explain that as long as we have a formula for or a recipe for computing $\mathbb{E} \theta(P)$ for the prior situation, we also have a formula or recipe for $\hat{\theta}$.

(b) Show that the mean of the random mean $\theta(P) = \int x dP(x)$ is $\theta_0 = \int x dP_0(x)$. Then via the updating derive the Bayes estimator $w_n \theta_0 + (1 - w_n) \bar{X}$, with weight $w_n = a/(a+n)$ for the prior mean.

(c) Then work with the variance $\sigma^2(P) = \mathbb{E}_P X^2 - \theta(P)^2$. From Ex. 14.9, show

$$\begin{aligned} \mathbb{E} \theta(P)^2 &= \int xy dP(x) dP(y) = \frac{a}{a+1} \int xy dP_0(x) dP_0(y) + \frac{1}{a+1} \int xy dP_{12}(x, y) \\ &= \frac{a}{a+1} \theta_0^2 + \frac{1}{a+1} \int x^2 dP_0(x), \end{aligned}$$

and derive from this the formula $E \sigma(P)^2 = \{a/(a+1)\} \sigma_0^2$, where $\sigma_0^2 = \int x^2 dP_0 - \theta_0^2$ is the prior variance. Then, via updating, find the Bayes estimator,

$$\hat{\sigma}^2 = E \{\sigma(P^2) | \text{data}\} = \frac{n+a}{n+a+1} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

(d) For a distribution F on the positive half-line, define $\theta(F) = \int x dF / F^{-1}(\frac{1}{2})$, the mean-to-median ratio. Here there is no explicit formula for $E \theta(F)$. Explain how we nevertheless can compute the Bayes estimate, for a given dataset, via simulation, cf. Ex. 14.11.

Ex. 14.15 *Bayesian nonparametric minimax estimation of a c.d.f.* Consider i.i.d. data $X_1, \dots, X_n$ from some c.d.f. F . The task is to estimate this F , with the loss function $L(F, \hat{F}) = \int (\hat{F} - F)^2 dw$, for a suitable nonnegative weight function w .

(a) Show first that with F being given any nonparametric prior, in the space of c.d.f.s on the line, then the Bayes estimator is $\hat{F}(t) = E \{F(t) | \text{data}\}$. With F having the $\text{Dir}(aF_0)$ prior, for some prior guess F_0 and a positive, show that

$$\hat{F}(t) = \{a/(a+n)\} F_0(t) + \{n/(a+n)\} F_n(t),$$

a convex combination of the prior mean and the empirical c.d.f. F_n .

(b) Then show that the risk function for this Bayes estimator can be written

$$R_n(F, \hat{F}) = E_F L(F, \hat{F}) = \frac{n}{(n+a)^2} \int F(1-F) dw + \left(\frac{a}{n+a}\right)^2 \int (F - F_0)^2 dw.$$

(c) Taking the mean of the risk function, with respect to the prior, show that the Bayes risk becomes

$$\begin{aligned} \text{MBR}(\text{Dir}(aF_0)) &= \left\{ \frac{n}{(n+a)^2} \frac{a}{a+1} + \left(\frac{a}{n+a}\right)^2 \frac{1}{a+1} \right\} \int F_0(1-F_0) dw \\ &= \frac{a}{(a+1)(a+n)} \int F_0(1-F_0) dw. \end{aligned}$$

Show, as for associated exercises Ex. 7.8 and 7.15 for finite-dimensional situations, that this Bayes risk is maximised for $a = \sqrt{n}$. Show furthermore that

$$F^*(t) = \frac{1/2}{\sqrt{n}+1} + \frac{\sqrt{n}}{\sqrt{n}+1} F_n$$

has $\sup_F R(F, F^*) = r_n = \frac{1}{4}/(\sqrt{n}+1)^2$. Using the general results of Ex. 9.14, exhibit a sequence of Dirichlet priors $\text{Dir}(\sqrt{n}F_{0,n})$ such that the minimum Bayes risks converge towards r_n . Explain that this proves that F^* is minimax.

Ex. 14.16 *Nonparametric admissibility for the minimax estimator for the mean.* Consider $X_1, \dots, X_n$ i.i.d. from an unknown distribution P on the unit interval, i.e. we only know P comes from the big space $\mathcal{P}$ of all distributions on $[0, 1]$. The task is to estimate the mean $\theta = \int_0^1 x dP(x)$, with squared error loss function.

(a) First, for the estimator $\bar{X}$, show $R(\bar{X}, P) = \sigma(P)^2/n$, with max risk $\frac{1}{4}/n$.

(b) Then consider

$$\hat{\theta} = \frac{\sqrt{n}}{\sqrt{n}+1} \bar{X} + \frac{1}{\sqrt{n}+1} \frac{1}{2}, \text{ and show } R(\hat{\theta}, P) = \frac{1}{(\sqrt{n}+1)^2} (E_P X^2 - E_P X + \frac{1}{4}).$$

Show that its max risk is $\frac{1}{4}/(\sqrt{n}+1)^2$. Prove that it is minimax, via a simple Beta prior for $p = \Pr(X_i = 1)$, on the very tiny subset $\mathcal{P}_0$ of $\mathcal{P}$, those concentrated on $\{0, 1\}$.

(c) The minimaxity proof above is essentially as in [Lehmann \(1950\)](#). He also claims there that the $\hat{\theta}$ is admissible; his argument, op. cit. p. 28, is contained in the passage, oops “The unique, and hence admissible estimate of θ , is [the estimator $\hat{\theta}$ above]”. Explain that this is not sufficient, however; the support of the marginal distribution for the prior he uses, on the tiny $\mathcal{P}_0$, is $\{0, 1\}^n$, so replacing $\bar{Y}$ with $n^{-1} \sum_{i=1}^n X_i^k$, for any k , gives the same estimator for 0-1 data, but not otherwise.

(d) Here, however, we shall give a correct admissibility proof, indeed using Bayesian nonparametrics. We need the Q_n marginal distribution for a Dirichlet sample. Suppose $\hat{\theta}$ is not admissible, so there is a clever competitor θ^* with risk $R(\theta^*, P)$ never above $r_{\max}$ but in some places lower. Explain that there must be points $x^0 = (x_1^0, \dots, x_n^0)$ at which $\hat{\theta}(x^0) \neq \theta^*(x^0)$. For one such x^0 , show that there is a P_0 , with positive masses at positions $x_1^0, \dots, x_n^0$, and with $\int_0^1 x dP_0 = \frac{1}{2}$. They are both minimisers of

$$\text{BR}(\tilde{\theta}, \mathcal{D}_{\sqrt{n}P_0}) = \int \left[\int \{\tilde{\theta}(x) - \theta(P)\}^2 d\mathcal{D}_{\sqrt{n}P_0 + nP_n(x)}(P) \right] dQ_n(x),$$

writing for emphasis $P_n(x)$ for the empirical distribution for the $x = (x_1, \dots, x_n)$. Explain that $Q_n(\{x^0\})$ must be positive. So for this x^0 there can only be one minimiser,

$$\tilde{\theta}(x^0) = \mathbb{E}\{\theta(P) \mid x^0\} = \int \theta(P) d\mathcal{D}_{\sqrt{n}P_0 + nP_n(x^0)}(P) = \frac{1}{\sqrt{n}+1} \frac{1}{2} + \frac{\sqrt{n}}{\sqrt{n}+1} n^{-1} \sum_{i=1}^n x_i^0.$$

Explain that this means there can be no competitor to $\hat{\theta}$, after all, with better risk, and that this implies the admissibility of $\hat{\theta}$.

Ex. 14.17 *Inference for quantiles via Dirichlet processes.* For a c.d.f. F on the real line, its quantile function inverse is defined as $Q(y) = \inf\{t: F(t) \geq y\}$.

(a) For simple illustrations, give the $Q(y)$ functions for the two parametric cases (i) $F \sim N(\xi, \sigma^2)$, (ii) F is Weibull (a, b) , i.e. with $F(t) = 1 - \exp\{-(t/a)^b\}$. Show that $Q(y)$ is always left-continuous, also in cases where F does have discontinuities. Show furthermore that $Q(y) \leq x$ if and only if $y \leq F(x)$.

(b) We use $\text{Be}(y, a, b)$ and $\text{be}(y, a, b)$ to denote the c.d.f. and density of a Beta(a, b). With partial integration, derive the useful identity

$$\text{Be}(c, a, b) - \text{Be}(c, a+1, b-1) = \frac{\text{be}(c, a+1, b)}{a+b} = \frac{\text{be}(1-c, b, a+1)}{a+b}, \quad \text{valid for } b > 1.$$

(c) Let now $F \sim \text{Dir}(aF_0)$, for suitable prior mean c.d.f. F_0 and positive strength parameter a . Show that

$$H_{0,a}(x) = \Pr\{Q(y) \leq x\} = 1 - \text{Be}(y, aF_0(x), a\bar{F}_0(x)) = \text{Be}(1-y, a\bar{F}_0(x), aF_0(x)),$$

where we use $\bar{H}(x) = 1 - H(x)$ for any c.d.f. H . Show that $\text{Be}(y, a, b) = \frac{1}{2}$ demands $a = b$, so $H_{0,a}(x_0) = \frac{1}{2}$ means x_0 being the prior median.

(d) With $x_1, \dots, x_n$ i.i.d. from the random F we know $F \mid \text{data} \sim \text{Dir}(aF_0 + nF_n)$. Explain that

$$H_{n,a}(x) = \Pr\{Q(y) \leq x \mid \text{data}\} = 1 - \text{Be}(y, aF_0(x) + nF_n(x), a\bar{F}_0(x) + n\bar{F}_n(x)).$$

Assume that the data are continuous, with $x_{(1)} < \dots < x_{(n)}$. From this, show that $Q(y)$ has the density

$$h_{n,a}(x) = (\partial/\partial x) \text{Be}(1-y, a\bar{F}_0(x) + n-i, aF_0(x) + i) \quad \text{inside } (x_{(i)}, x_{(i+1)}),$$

and with posterior point mass

$$\begin{aligned}\Delta H_{n,a}(x_{(i)}) &= \text{Be}(y, aF_0(x_{(i)}) + i - 1, a\bar{F}_0(x_{(i)}) + n - i + 1) \\ &\quad - \text{Be}(y, aF_0(X_{(i)}) + i, a\bar{F}_0(x_{(i)}) + n - i) \\ &= (n + a)^{-1} \text{be}(y, aF_0(x_{(i)}) + i, a\bar{F}_0(x_{(i)}) + n - i + 1)\end{aligned}$$

at data point $x_{(i)}$. Show for fixed a that the sum of the probabilities at the data points tends to 1, for growing n .

(e) Let now $a \rightarrow 0$, corresponding to a Dirichlet process with essentially no prior information taken on board, so may view the posterior Dirichlet process, close to $\text{Dir}(\delta(x_{(1)}) + \dots + \delta(x_{(n)}))$, as being the essential information in the data, from an objective Bayesian nonparametric viewpoint. Show that the probabilities between data points go to zero, and that what is left are the probabilities

$$p_n(x_{(i)}) = \text{Be}(y, i - 1, n - i + 1) - \text{Be}(y, i, n - i) = \binom{n-1}{i-1} y^{i-1} (1-y)^{n-i}.$$

Explain that these binomial weights concentrate around y for growing n .

(f) Consider the mean $\hat{Q}_a(y)$ and variance $\hat{V}_a(y)$ of $Q(y)$ given the data. With $a \rightarrow 0$, corresponding to no real information in the prior, show that we reach the following objective Bayesian nonparametrics quantities:

$$\begin{aligned}\hat{Q}_a(y) &\rightarrow \hat{Q}_0(y) = \sum_{i=1}^n \binom{n-1}{i-1} y^{i-1} (1-y)^{n-i} x_{(i)}, \\ \hat{V}_a(y) &\rightarrow \hat{V}_0(y) = \sum_{i=1}^n \binom{n-1}{i-1} y^{i-1} (1-y)^{n-i} \{x_{(i)} - \hat{Q}_0(y)\}^2.\end{aligned}$$

Rather splendidly, quantile confidence bands based on these, of the type $\hat{Q}_0(y) \pm 1.96 \hat{V}_0(y)^{1/2}$ can be trusted, in terms of frequentist coverage; this is a consequence of BvM theorems of Ex. 14.19–14.20. (xx so there is no need to estimate the density in the process, as for other such bands. explain: can make valid confidence bands, without needing the Ch10 results of $\sqrt{n}(Q_n(y) - Q(y)) \rightarrow_d q(y) N(0, y(1-y))$ and the cumbersome estimation of $q(y) = 1/f(F^{-1}(y))$. For an application of this, see Story #23. xx)

(g) (xx briefly: the Doksum shift function, the Parzen quantile comparison function, the Gini index, the Lorenz curve, all using Dir quantiles. point to these methods being used in Story #23. cc)

Ex. 14.18 *An automatic density estimator.* Suppose $x_1, \dots, x_n$ are i.i.d. from some continuous positive density f on the real line, with c.d.f. F . Here we construct an automatic density estimator $\hat{f}$, via the noninformative Dirichlet process prior for F , positive precisely on the range of data and without needing any bandwidth or other fine-tuning parameters.

(a) From the noninformative Dirichlet prior, with $a \rightarrow 0$, we have reached in Ex. 14.17 the intriguing smooth quantile function estimator $\hat{Q}_0(y)$, with binomial type weights for the ordered data $x_{(1)} < \dots < x_{(n)}$. This may be used to construct a density estimator

‘objective’ in the sense of pushing prior information to zero

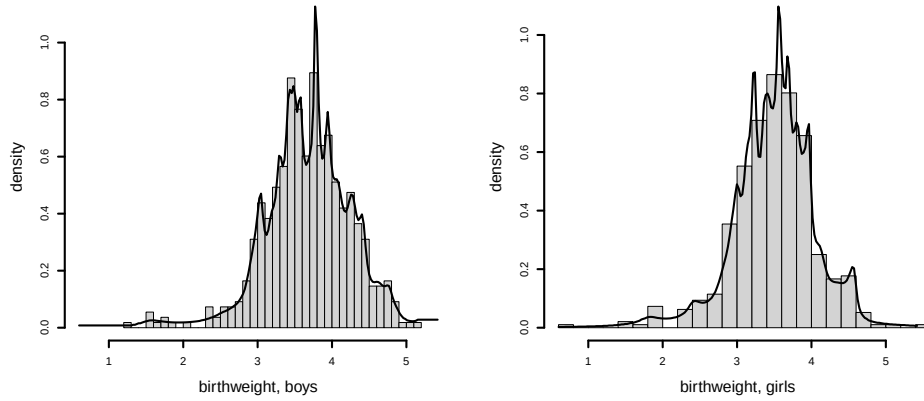


Figure 14.2: Automatic birthweight density estimates, for the boys (left panel) and girls (right panel), along with fine-gridded histograms. The densities are precisely supported on the data range, $[1.275, 5.140]$ for the boys and $[0.601, 5.420]$ for the girls. There is more on both the boys and the girls and on quantile methodologies in Stories [#22](#) and [#23](#).

$\hat{f}_0$, the density associated with $Q'(y) = 1/f(Q(y))$. Start by showing

$$\begin{aligned} \hat{q}_0(y) &= \hat{Q}'_0(y) = \sum_{i=1}^n \frac{n-1}{i-1} y^{i-1} (1-y)^{n-i} \left(\frac{i-1}{y} - \frac{n-i}{1-y} \right) x_{(i)} \\ &= \int_{x_{(1)}}^{x_{(n)}} \text{be}(y, nF_n(x), n\bar{F}_n(x)) dx = \sum_{i=1}^{n-1} (x_{(i+1)} - x_{(i)}) \text{be}(y, i, n-i). \end{aligned}$$

Note that there is no extra smoothing task here, as with a bandwidth; the inherent smoothing comes for free, via the $\hat{Q}_0(y)$.

(b) From $f(x) = (Q^{-1})'(x)$, valid for smooth densities, argue that the density implied by the $\hat{Q}_0(y)$ is

$$\hat{f}_0(x) = \left[\sum_{i=1}^{n-1} (x_{(i+1)} - x_{(i)}) \text{be}(\hat{F}_0(x), i, n-i) \right]^{-1}.$$

For each x , the equation $\hat{Q}_0(y) = x$ is numerically solved for y to get $\hat{F}_0(x)$. At the endpoints: supply details to get to

$$\begin{aligned} \hat{f}_0(x_{(1)}) &= 1/\hat{q}_0(0) = \{(n-1)(x_{(2)} - x_{(1)})\}^{-1} \\ \hat{f}_0(x_{(n)}) &= 1/\hat{q}_0(1) = \{(n-1)(x_{(n)} - x_{(n-1)})\}^{-1}. \end{aligned}$$

(c) Access the birthweights for Oslo newborns, 548 boys and 480 girls, born in the 2001–2008 period, see Story [#22](#). Compute this automatic density estimate, for the boys and the girls, and construct a version of Figure 14.2. The densities are seen to follow the histograms rather closely, but is by construction smoother. That boys are somewhat bigger than girls at birth, further illustrating methods of this exercise, is examined in Stories [#22](#) and [#23](#).

Nonparametric Bernshtein–von Mises theorems, bootstrapping

BvM

Ex. 14.19 *Bernshtein–von Mises theorems for the Dirichlet process prior.* We have learned in Ch. 7 that Bayesian methods tend to match those associated with likelihood procedures, for growing sample sizes. Specifically, recall from Ex. 7.39 the remarkable Bernshtein–von Mises theorem, for which we now find an equally remarkable nonparametric extension. Suppose $X_1, \dots, X_n$ are i.i.d. from some positive density f_{true} , with c.d.f. F_{true} , which we for simplicity first take to be supported on the unit interval $[0, 1]$. For the empirical c.d.f. F_n we then know from Ex. 10.26 that $Z_n(t) = \sqrt{n}\{F_n(t) - F_{\text{true}}(t)\} \rightarrow_d Z(t) = W^0(F_{\text{true}}(t))$.

(a) Let F be given the Dirichlet process prior aF_0 . Using the updating theorem of Ex. 14.10, we know the posterior mean $\hat{F}_B = w_n F_0 + (1 - w_n)F_n$, with $w_n = a/(a + n)$ tending quickly to zero. Explain that $\hat{F}_B(t) \rightarrow_{\text{pr}} F_{\text{true}}(t)$, actually uniformly in t , via the Glivenko–Cantelli theorem of Ex. 2.60. Show that the process $Z_{n,B}(t) = \sqrt{n}\{F(t) - \hat{F}_B(t)\}$ has

$$\text{cov}[\{Z_{n,B}(s), Z_{n,B}(t)\} | \text{data}] = \frac{n}{n + a + 1} \hat{F}_B(s) \{1 - \hat{F}_B(t)\} \quad \text{for } s \leq t.$$

(b) Show first that the conditional covariance function converges to $F_{\text{true}}(s)\{1 - F_{\text{true}}(t)\}$ for $s \leq t$, and this is the covariance function for $Z(t)$. Use the Dirichlet-normal-approximation lemma of Ex. 7.14. Via this, show the appropriate finite-dimensional convergence,

$$(Z_{n,B}(t_1), \dots, Z_{n,B}(t_k)) | \text{data} \rightarrow_d (Z(t_1), \dots, Z(t_k)) \quad \text{for } t_1 < \dots < t_k.$$

(c) To demonstrate tightness, we use (10.3) of Ex. 10.19. For this purpose, suppose $(U, V, W) \sim \text{Dir}(kp, kq, kr)$, with $p + q + r = 1$. Verify, via patient algebra, that

$$E(U - p)^2(V - q)^2 = \frac{pq}{(k+1)(k+2)(k+3)} \{k - (k-6)(p+q-3pq)\},$$

where the simple upper bound $3pq$ suffices for our needs. From this show indeed that $Z_{n,B}$ is tight, and this implies the desired BvM mirror result:

$$\begin{aligned} Z_n &= \sqrt{n}(F_n - F_{\text{true}}) \rightarrow_d Z = W^0(F_{\text{true}}(\cdot)) \quad (\text{frequentist}), \\ Z_{n,B} &= \sqrt{n}(F - F_n) | \text{data} \rightarrow_d Z = W^0(F_{\text{true}}(\cdot)) \quad (\text{Bayes, given data, a.s.}). \end{aligned} \quad (14.2)$$

Explain that this latter large-sample result holds with both $\sqrt{n}(F - F_n)$ and $\sqrt{n}(F - \hat{F}_B)$. Note also that this holds for any positive a , with the same limit; show indeed that the objective Dirichlet process prior, where $F | \text{data} \sim \text{Dir}(nF_n)$, concentrated on the data points with probabilities $(p_1, \dots, p_n) \sim \text{Dir}(1, \dots, 1)$, leads to $\sqrt{n}(F - F_n) | \text{data} \rightarrow_d Z$.

Ex. 14.20 *Bernshtein–von Mises theorems for inverse Dirichlet processes.* Let the framework be as in Ex. 14.19, with a $\text{Dir}(aF_0)$ prior for the c.d.f. F , and with i.i.d. data drawn from some smooth increasing F_{true} . From the master theorem $\sqrt{n}(F - F_n) | \text{data} \rightarrow_d Z$ there we can half automatically derive various corollaries, via techniques already established in Ch. 10 (xx pointers xx), i.e. we do not need to re-work those to see what they imply in the current Bayesian nonparametrics Dirichlet context.

(a) Write $Q_n = F_n^{-1}$ for the empirical quantiles and $Q_{\text{true}} = F_{\text{true}}^{-1}$ for the underlying quantile function, along with $Q'_{\text{true}}(u) = 1/f_{\text{true}}(Q_{\text{true}}(u))$. Explain that we in fact have

the following, the quantile mirror parallel to (14.2):

$$\begin{aligned} R_n(u) &= \sqrt{n}\{Q_n(u) - Q_{\text{true}}(u)\} \rightarrow_d Z^*(u) = c_{\text{true}}(u)W^0(u) \quad (\text{frequentist}), \\ R_{n,B} &= \sqrt{n}\{Q(u) - Q_n(u)\} | \text{data} \rightarrow_d Z^*(u) \quad (\text{Bayes, given data, a.s.}). \end{aligned}$$

The first result is already contained in 10.xxxx, and the second follows via the same arguments from the main result of Ex. 14.19. Convergence here takes place in each $D[\varepsilon, 1-\varepsilon]$ space; there will be difficulties near the endpoints due to $c_{\text{true}}(u)$ often becoming too big there.

(b) Just as we learned in 10.xxxx, the BvM theorems imply large-sample control and indeed equivalence for large classes of smooth functionals:

$$\sqrt{n}\{T(F_n) - T(F)\} \quad \text{and} \quad \sqrt{n}\{T(F) - T(F_n)\} | \text{data}$$

have the same large-sample behaviour. This in particular includes all smooth functions of a finite number of $F(t_j)$ and $Q(u_k)$. Write out the details for three particular cases: (i) $T_1(F) = E_F X = \int x dF(x)$ the mean; (ii) $T_2(F) = F^{-1}(\frac{1}{2})$ the median; and (iii) $T_3(F) = E_F X / F^{-1}(\frac{1}{2})$ the mean-to-median ratio.

Ex. 14.21 Bayesian bootstrapping. Consider i.i.d. data $X_1, \dots, X_n$ from some P , and with some parameter $\theta = T(P)$ in focus. We saw in Ex. 10.35 that as long as we can invent a machinery, producing say F_n^* given data, with the property that $Z_n^* = \sqrt{n}(F_n^* - F_n) | \text{data}$ has the same large-sample behaviour as does the direct bridge $Z_n = \sqrt{n}(F_n - F)$, then we're in business. The invention discussed there was that of Efron's nonparametric bootstrapping, with F_n^* the empirical c.d.f. for $X_1^*, \dots, X_n^*$ drawn from F_n , with empirically speaking monumental consequences for statistics and data science.

(a) But now we're Bayesian. For the objective Dirichlet prior construction, where $P | \text{data} \sim \text{Dir}(nF_n)$, explain that this corresponds to probabilities $(p_1, \dots, p_n) \sim \text{Dir}(1, \dots, 1)$ at the datapoints $x_1, \dots, x_n$. Via Dirichlet sampling, therefore, you may simulate $B = 10^4$ posterior realisations of $\theta = \theta(P)$. This is essentially the *Bayesian bootstrap*, yielding upon the simulation work the c.d.f. $B(v) = \Pr(\theta(P) \leq v | \text{data})$ as say $B^{-1} \sum_{b=1}^B I(\theta_b^* \leq v)$, with $\theta^* = \theta(P^*)$. For the particular case of $\theta(P) = E_P h(X) = \int h(x) dP(x)$, explain that $\theta(P) = \sum_{i=1}^n p_i h(x_i)$, and show that

$$E\{\theta(P) | \text{data}\} = \bar{h} = n^{-1} \sum_{i=1}^n h(x_i), \quad \text{Var}\{\theta(P) | \text{data}\} = \hat{\sigma}^2 / (n+1),$$

a mild modification away from the customary frequentist $\hat{\sigma}^2/n$, where $\hat{\sigma}^2$ is the sample variance $n^{-1} \sum_{i=1}^n \{h(x_i) - \bar{h}\}^2$.

(b) But does it work? In one direct sense it does, as it's merely a matter of Bayesian computing via simulation for a well-defined posterior. The mean and variance for the random mean above are also promising, by essentially matching the frequentist counterparts. Using arguments as in Ex. 10.35, show in fact that $\sqrt{n}\{\theta(P) - \hat{\theta}\}$ has precisely the same large-sample behaviour as does $\sqrt{n}(\hat{\theta} - \theta_{\text{true}})$. Show further that the Bayesian posterior $\theta(P)$, created via this Bayesian bootstrap, for large n yields the same inference as for the frequentist.

Beta processes

Ex. 14.22 Time discrete Bayesian survival via hazards and Beta products. We work here with the time discrete survival analysis framework of Ex. 11.1, where time is recorded in

$\{0, 1, 2, \dots\}$. A time to event variable T' has probabilities $f_j = (1 - \alpha_0) \cdots (1 - \alpha_{j-1})\alpha_j$ for $j = 0, 1, \dots$, in terms of the hazard rates $\alpha_j = \Pr(T' = j | T' \geq j)$; also, the survival function is $S_j = \Pr(T' \geq j) = (1 - \alpha_0) \cdots (1 - \alpha_j)$. Consider now independent survival data (T_i, δ_i) for $i = 1, \dots, n$, with T_i the possibly censored survival time and δ_i the indicator for non-censoring.

(a) Via efforts in the exercises pointed to, the likelihood function can be expressed as $L = \prod_{j=0}^{\infty} \alpha_j^{dN_j} (1 - \alpha_j)^{Y_j - dN_j}$, in terms of at-risk numbers $Y_j = \sum_{i=1}^n I(T_i \geq j)$ and counting numbers $dN_j = \sum_{i=1}^n I(T_i = j, \delta_i = 1)$. Give now the α_j independent Beta priors, with $\alpha_j \sim \text{Beta}(c_j \alpha_{0,j}, c_j(1 - \alpha_{0,j}))$, with prior means $\alpha_{0,j}$ and strength parameters c_j . Show that the posterior distribution is of the same type, with

$$\alpha_j | \text{data} \sim \text{Beta}(c_j \alpha_{0,j} + dN_j, c_j(1 - \alpha_{0,j}) + Y_j - dN_j),$$

and these being independent.

(b) For the cumulative hazard $A_k = \sum_{j=0}^k \alpha_j$, write $\hat{\alpha}_j = (c_j \alpha_{0,j} + dN_j) / (c_j + Y_j)$, and show that

$$\hat{A}_k = E(A_k | \text{data}) = \sum_{j=0}^k \hat{\alpha}_j, \quad \text{Var}(A_k | \text{data}) = \sum_{j=0}^k \frac{\hat{\alpha}_j(1 - \hat{\alpha}_j)}{c_j + Y_j + 1}.$$

Explain how this becomes a Bayesian extension of the Nelson–Aalen estimator of Ex. 11.2, and that the frequentist estimator corresponds to the $c_j \rightarrow 0$ case.

(c) Kaplan–Meier extension:

(d) (xx product integrals defined as fine limits, Riemann style. notes and pointers: point to Johansen, Gill, Andersen et al. (1993), Hjort (1990b), Gill and Johansen (1990), ... xx) this to be generalised to time-continuous scale below. Check that Ch11 briefly has the time-discrete Cox model: individual i has $1 - \alpha_{i,j} = (1 - \alpha_{0,j})^{\exp(x_i^t \beta)}$. with the $\alpha_{0,i}$ and β unknown parameters.

Ex. 14.23 *Beta priors for survival: from a fine grid to processes.* (xx to be done and polished. xx) Let $a_0(t)$ be some positive continuous function with integral $A_0(t) = \int_0^t a_0(s) ds$, and a positive constant c . Consider the random process $A_m(t) = \sum_{i/m \leq t} p_{m,i}$, with small independent Beta terms,

$$p_{m,i} \sim \text{Beta}(ca_0(i/m)(1/m), c\{1 - a_0(i/m)(1/m)\}),$$

these having moments

$$E p_{m,i} = a_0\left(\frac{i}{m}\right) \frac{1}{m} \quad \text{and} \quad \text{Var } p_{m,i} = \frac{1}{c+1} a_0\left(\frac{i}{m}\right) \frac{1}{m} \left\{1 - a_0\left(\frac{i}{m}\right) \frac{1}{m}\right\}.$$

(a) Show $E A_m(t) \rightarrow A_0(t)$ and $\text{Var } A_m(t) \rightarrow A_0(t)/(c+1)$. then convergence $A_m \rightarrow_d A$. For the Laplace transform, show that

$$E \exp\{-\theta A_m(t)\} = \prod_{i/m \leq t} E \exp(-\theta p_{m,i}) = \prod_{i/m \leq t} \left\{1 + \sum_{j=1}^{\infty} (-1)^j \frac{\theta^j}{j!} E p_{m,i}^j\right\}.$$

Show that

$$E p_{m,i}^j = \frac{\Gamma(c)}{\Gamma(ca_0(i/m)(1/m))} \frac{\Gamma(ca_0(i/m)(1/m) + j)}{\Gamma(c + j)}.$$

Use this to prove that this converges, to $\exp\{-D(\theta, t)\}$, where

$$D(\theta, t) = \int_0^1 \{1 - \exp(-\theta x)\} dL_t(x), \quad dL_t(x) = cx^{-1}(1-x)^{c-1} dx A_0(t),$$

this $L_t(x)$ called the Lévy measure. (xx then tightness. use criteria from Ch10; ok. then need to show that $\mathcal{A}$, the subset of cumulative hazard rates A in $D[0, 1]$, with all jumps $\Delta A(t) \in [0, 1]$, is closed. xx) assume $A_m \rightarrow_d A$, and work with $F_m(t) = 1 - \prod_{[0, t]} \{1 - dA_m(s)\}$. need to show that the limit is a c.d.f. F . Establish $A_m \rightarrow_d A$, which we call a Beta process with parameters c and A_0 .

(b) (xx the basic construction. with $c(s)$ too, not constant. xx)

$$dA(s) \sim \text{Beta}(c(s) dA_0(s), c(s)\{1 - dA_0(s)\}),$$

with

$$E dA(s) = dA_0(s) \quad \text{and} \quad \text{Var } dA(s) = dA_0(s)\{1 - dA_0(s)\}/\{c(s) + 1\}.$$

work out mean and variance for the integral, etc.

(c) (xx fix this. might push ut out. xx) do the skewness too. skew for $\text{Beta}(cp_0, c(1-p_0))$, then updated. with X_i having means zero: $E(\sum_{i=1}^n X_i)^3 = \sum_{i=1}^n E X_i^3$. use formula

$$E(p - p_0)^3 = 2p_0(1 - p_0)(1 - 2p_0)/[(c + 1)(c + 2)].$$

then

$$E\{A(t) - A_0(t)\}^3 = \sum E\{dA(s) - dA_0(s)\}^3 = \sum \frac{2 dA_0(1 - dA_0)(1 - 2dA_0)}{(c + 1)(c + 2)}.$$

check this:

Ex. 14.24 *Updating Beta processes given survival data.* For a certain survival distribution, with c.d.f. $F(t) = 1 - S(t) = 1 - \prod_{[0, t]} \{1 - dA(s)\}$, let that cumulative hazard A be modelled as a Beta process (c, A_0) , with judiciously elicited prior mean function A_0 and strength function c . We then observe survival data from this distribution, of the common form (T_i, δ_i) , see (xx Ch11 xx). Specifically, there are genuine i.i.d. survival times $T_1^0, \dots, T_n^+$ from the F distribution, and these are partially censored, so what we observe are (T_i, δ_i) , with $T_i = \min(T_i^0, z_i)$ and $\delta_i = I(T_i^0 < z_i)$, i.e. indicator functions for non-censoring. Here we learn hwo the distribution of A and hence S change given the data.

(a) We work through the details for one given datapoint (T, δ) , and then apply induction. For the $\delta = 0$ case: find the distribution of A given $T \geq t$. For the $\delta = 1$ case: work with A given $T \in [t, t + \varepsilon]$, and take the limit. (xx pointer here. xx) Verify that this for both cases leads to an *updated Beta process*. Via induction, one new datapoint at a time, your conclusion should be that

$$dA(s) | \text{data} \sim \text{Beta}(c(s) dA_0(s) + dN(s), c(s)\{1 - dA_0(s)\} + Y(s) - dN(s)\}.$$

(b) For illustration, consider the Gehan dataset, with two groups, $x = 0$ for placebo and $x = 1$ for the new medicine in question. Time to remission is recorded, in weeks. For a neutral analysis, we give the survival distributions the same prior, as follows. Take the exponential prior mean with $A_0(t) = \alpha_0 t$, with α_0 set to have prior median 10.0 weeks.

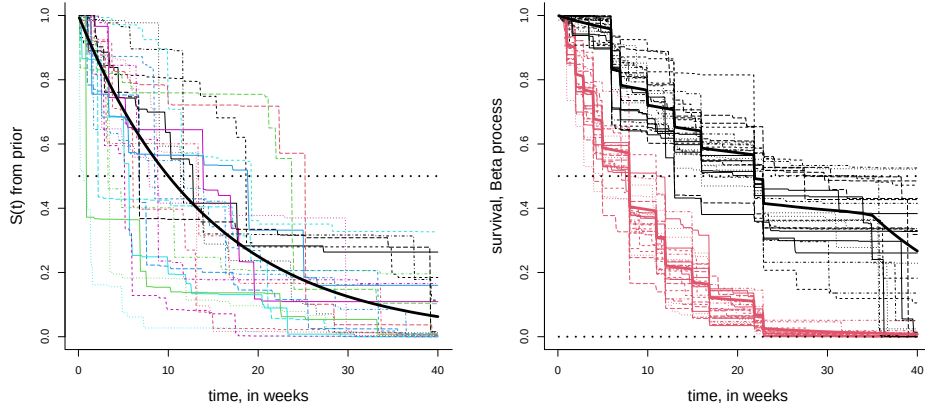


Figure 14.3: Beta process simulations: 25 simulated survival curves $S(t) = \prod_{[0,t]} \{1 - dA(s)\}$, from the common prior (left panel), and then for the posterior distributions for the two groups (right panel). These are tighter, with a clear difference: the treated group does clearly better than the placebo group.

Then use the Beta process with $c(s) = k \exp\{-A_0(t)\}$, using $k = 3.0$ for this illustration. Construct a version of Figure 14.3. The left panel is what we believe, for both groups, before data come in; the right panel is what we believe and understand, after having the seen the data, with group 1 doing considerably better than group 0. Comment on

(c) From the simulated paths $A_0(\cdot)$ and $A_1(\cdot)$ given data, read off $\delta = \text{med}(S_1) - \text{med}(S_2)$, the difference of medians. Use e.g. 10^3 such simulations, give a fine histogram for that posterior distribution, and give its 0.05, 0.50, 0.95 quantiles. How do you conclude?

Ex. 14.25 Beta processes, II. Using results of Ex. 14.24.

$$dA(s) | \text{data} \sim \text{Beta}(c(s) dA_0(s) + dN(s), c(s)\{1 - dA_0(s)\} + Y(s) - dN(s)).$$

(xx basics for posterior mean and variance. xx)

(a) Derive the Bayesian nonparametrics estimators

$$\begin{aligned} \hat{A}(t) &= E\{A(t) | \text{data}\} = \int_0^t \frac{c(s) dA_0(s) + dN(s)}{c(s) + Y(s)}, \\ \hat{S}(t) &= E \prod_{[0,t]} \{1 - dA(s)\} | \text{data} = \prod_{[0,t]} \left\{ 1 - \frac{c(s) dA_0(s) + dN(s)}{c(s) + Y(s)} \right\}. \end{aligned}$$

Explain that these properly generalise the classic Nelson–Aalen and Kaplan–Meier estimators (xx point Ch10 xx), and that these correspond to $c(s) \rightarrow 0$, meaning prior strength zero for the Beta process.

(b) Give also expressions for the posterior variances. In particular, show that

$$\text{Var}\{A(t) | \text{data}\} = \int_0^t \frac{\hat{A}(s)\{1 - \hat{A}(s)\}}{c(s) + Y(s) + 1}.$$

(c) Two groups, consider the survival ratio

$$\rho(t) = \frac{S_1(t)}{S_2(t)} = \frac{\prod_{[0,t]} \{1 - dA_1(s)\}}{\prod_{[0,t]} \{1 - dA_2(s)\}}.$$

Find expressions for the prior and posterior means. Illustration: the Gehan data above.

Ex. 14.26 *Bayesian semiparametric alternative to Cox regression.* (xx let's see. might have a survival data story with this on board. xx) time-discrete setup: at time position j , individual i has hazard rate $\alpha_i(j)$ given via $1 - \alpha_i(j) = \{1 - \alpha(j)\}^{z_i}$, and with $z_i = \exp(x_i^t \beta)$. The likelihood can be written

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^n L_i = \prod_{i=1}^n \left[\prod_{j=0}^{\infty} (1 - \alpha_i(j))^{Y_i(j) - \Delta N_i(j)} \alpha_i(j)^{\Delta N_i(j)} \right] \\ &= \prod_{j=0}^{\infty} \left[(1 - \alpha(j))^{\sum_{i=1}^n z_i (Y_i(j) - \Delta N_i(j))} \prod_{i=1}^n (1 - (1 - \alpha(j))^{z_i})^{\Delta N_i(j)} \right]. \end{aligned}$$

With prior $\pi(\beta)$ for β , taken independent of a Beta process for the $\alpha(j)$, we should have $\pi(\beta | \text{data}) \propto \pi(\beta) \bar{L}(\beta)$, with

$$\bar{L}(\beta) = \int L(\alpha, \beta) \pi(\alpha) d\alpha = \prod_{j=0}^{\infty} E_0 h(\alpha(j), \beta),$$

where E_0 signals expectation taken with respect to the prior for $\alpha(j)$, and

$$h(\alpha(j), \beta) = \{1 - \alpha(j)\}^{R(j, \beta) - \sum_{i=1}^n z_i \Delta N_i(j)} \prod_{i=1}^n [1 - \{1 - \alpha(j)\}^{z_i}]^{\Delta N_i(j)},$$

featuring also $R(j, \beta) = \sum_{i=1}^n z_i Y_i(j)$. at time position j where there is no jump:

$$h(\alpha(j), \beta) = \{1 - \alpha(j)\}^{R(j, \beta)}, \quad \text{with explicit } E_0 h(\alpha(j), \beta).$$

at time position j with a single jump $\Delta N_i(j) = 1$ for individual i_0 :

$$h(\alpha(j), \beta) = \{1 - \alpha(j)\}^{R(j, \beta) - z_{i_0}} [1 - \{1 - \alpha(j)\}^{z_{i_0}}], \quad \text{with explicit } E_0 h(\alpha(j), \beta).$$

hence full numerical control over $\pi(\beta | \text{data})$. and over $\alpha(j) | (\beta, \text{data})$. this to be used for a given individual with his covariate vector x_{fix} : he has

$$\alpha_{\text{fix}}(j) = 1 - \{1 - \alpha(j)\}^{\exp(x_{\text{fix}}^t \beta)}, \quad \text{for } j = 0, 1, 2, \dots,$$

along with survival function $S_{\text{fix}}(j) = \prod_{j'=0}^j \{1 - \alpha(j')\}^{\exp(x_{\text{fix}}^t \beta)}$.

Ex. 14.27 *The Dirichlet as a special case of the Beta process.* Consider a Beta process with prior mean A_0 and strength function $c(s)$ (xx point up xx). The c.d.f. associated with A_0 is $F_0(t) = 1 - \prod_{[0, t]} \{1 - dA_0(s)\}$. Here we study the associated random c.d.f. $F(t) = 1 - \prod_{[0, t]} \{1 - dA(s)\}$. (xx where to put the minimum of product integral. xx)

(a) For the random $F(t)$, show that its mean is $F_0(t)$, both for smooth and non-smooth cases of the baseline cumulative hazard function A_0 . When $A_0(t)$ is the integral of a continuous hazard rate $\alpha_0(s)$, i.e. $dA_0(s) = \alpha_0(s) ds$, show that this mean function is the simpler and familiar $F_0(t) = 1 - \exp\{-A_0(t)\}$; see (xx point to hazard Ch1 xx).

(b) We first work out a general expression for the Laplace transform of $B = -\log(1 - F)$. Show that this is the product integral

$$E \{1 - F(t)\}^\theta = \prod_{[0, t]} E \{1 - dA(s)\}^\theta = \prod_{[0, t]} \frac{\Gamma(c)}{\Gamma(c(1 - dA_0))} \frac{\Gamma(c(1 - dA_0) + \theta)}{\Gamma(c + \theta)}.$$

For small ε , explain that $\Gamma(c - c\varepsilon)/\Gamma(c) \doteq 1 - c\psi(c)\varepsilon$, with $\psi = \Gamma'/\Gamma$ the log-derivative of the gammafunction. For the case of $A_0(t) = \int_0^t \alpha_0(s) ds$, the integral of a continuous function, show that this leads to

$$\mathbb{E} \{1 - F(t)\}^\theta = \exp \left[- \int_0^t \{ \psi(c(s) + \theta) - \psi(c(s)) \} c(s) dA_0(s) \right],$$

which may be used to characterise distributions of F for different $c(s)$ and A_0 functions.

(c) For the $F_0 = \exp(-A_0)$ smooth case, consider the special case $c(s) = kF_0[s, \infty) = k \exp(-A_0(s))$. Show that

$$\begin{aligned} I_t &= - \int_0^t [\psi(k \exp(-A_0(s)) + \theta) - \psi(k \exp(-A_0(s)))] k \exp(-A_0(s)) \alpha_0(s) ds \\ &= - \int_{k \exp(-A_0(t))}^k \{ \psi(u + \theta) - \psi(u) \} du \\ &= - \log \Gamma(k + \theta) + \log \Gamma(k \exp(-A_0(t)) + \theta) + \log \Gamma(k) - \log \Gamma(k \exp(-A_0(t))). \end{aligned}$$

Verify that this is the same as for the $\text{Dir}(kF_0)$ case, where $F(t)$ has a beta distribution, with

$$\mathbb{E} \{1 - F(t)\}^\theta = \frac{\Gamma(k)}{\Gamma(k(1 - F_0(t)))} \frac{\Gamma(k(1 - F_0(t)) + \theta)}{\Gamma(k + \theta)}.$$

(d) With $c(s) = k \exp\{-A_0(t)\}$, show via generalisations of the above that indeed the random $F = 1 - \prod_{[0, \cdot]} (1 - dA)$ process is a $\text{Dir}(aF_0)$.

(e) (xx round off. we ought to mention empirical Bayes too. xx)

Ex. 14.28 *Beta + Bernoulli = Poisson; Beta + Two Bernoullis = Twin Poissons.* A Beta process has infinitely many jumps, of sizes ranging from very close to 0 to occasionally very close to 1. Adding on Bernoullis with these probabilities, at the jump positions, leads to a Poisson process.

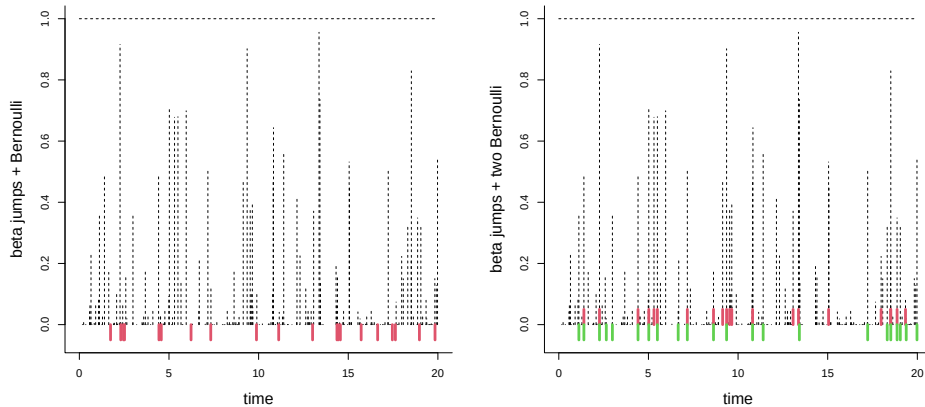


Figure 14.4: Beta process jumps, from $\text{Beta}(0.01, 0.99)$, along with one Bernoulli (left panel, shown as ticks) and two Bernoullis (right panel, shown as up-ticks and down-ticks),

(a) Suppose first that a probability p is drawn from some density on the unit interval, with $\mathbb{E} p = p_0$, and that $X | p \sim \text{binom}(1, p)$. Show that $X = 1$ with probability p_0 , and

hence variance $p_0(1 - p_0)$. this Y has mean p_0 and variance $p_0(1 - p_0)$. Consider then a Beta process on a fine grid, with independent jumps $p_{m,i} \sim \text{Beta}(ca_0(i/m)(1/m), c\{1 - a_0(i/m)(1/m)\})$ at positions i/m , and partial sum process $A_m(t) = \sum_{i/m \leq t} p_{m,i}$, as in (xx point xx). On top of this take $X_{m,i} \sim \text{binom}(1, p_{m,i})$, with sums $X_m(t) = \sum_{i/m \leq t} X_{m,i}$. Find the mean and variance of $X_m(t)$, and show that $X_m \rightarrow_d X$, a full Poisson process with mean $A_0(t) = \int_0^t a_0(s) ds$. Construct a version of Figure 14.4, left panel, in which we have used $c = 1$ and $a_0(s) = 1$, on a fine grid over $[0, 20]$ with grid steps 0.01. (xx hmm: the c disappears in the limit; comment. xx)

(b) Now construct two processes, on top of the Beta jumps, with $X_{m,i}$ and $Y_{m,i}$ independent $\text{binom}(1, p_{m,i})$, followed by cumulative counts $X_m(t) = \sum_{i/m \leq t} X_{m,i}$ and $Y_m(t) = \sum_{i/m \leq t} Y_{m,i}$. Show that (X_m, Y_m) has a well-defined process limit (X, Y) , these being dependent Poisson processes, with correlation $1/(c + 1)$. Construct a version of Figure 14.4, right panel. This could be a model for births, with the occasional twin births.

(c) Now consider $Z_m(t) = \sum_{i/m \leq t} Z_{m,i}$, with $Z_{m,i} | p_{m,i} \sim \text{Pois}(p_{m,i})$. (xx do this. xx) Then $Z_m(t) \rightarrow_d Z(A(t))$. find mean and the extra variance. model for overdispersion. but gamma easier.

(d) (xx then a geometric on top of the $p_{m,i}$. is there a limit? xx) so $Z_m(t) = \sum_{i/m \leq t} Z_{m,i}$, where $Z_{m,i}$ has $(1 - p_{m,i})^{z-1} p_{m,i}$ for $z = 0, 1, \dots$ is there a well-defined limit process? easier with a Poisson, perhaps with the same limit as for X above. geometric, generating function:

$$E(s^{Z_m(t)} | A_m) = \prod_{i/m \leq t} \frac{p_i s}{1 - (1 - p_i)s}.$$

let's see:

$$E(Z_m(t) | A_m) = \sum_{i/m \leq t} \frac{1 - p_{m,i}}{p_{m,i}}, \quad \text{Var}(Z_m(t) | A_m) = \sum_{i/m \leq t} \frac{1 - p_i}{p_i^2},$$

so

$$E Z_m(t) = \sum_{i/m \leq t} \frac{c - 1}{ca_0(i/m)(1/m)}, \quad \text{perhaps } \int_0^t \dots$$

(e) (xx a Bayesian spin. xx)

Gaussian processes and Bayesian Kriging

Ex. 14.29 Bayesian kriging. Suppose there is a continuous process $Z(x)$ on some interval $[c, d]$, which we have observed only in a finite and perhaps small number of locations. How can we estimate $Z(x)$ where we have not seen it, along with a measure of precision?

(a) Let $Z(x)$ be a Gaussian process on $[c, d]$, with constant mean function ξ , and covariance function

$$\text{cov}\{Z(x), Z(x')\} = \sigma^2 K(|x - x'|),$$

where $K(r)$ is the correlation function. The formal requirement for such a K is that $K(0) = 1$ and that the matrix A of $K(|x_i - x_j|)$ should be nonnegative definite, i.e. $\sum_{i,j} c_i c_j K(|x_i - x_j|) \geq 0$ for any $x_1, \dots, x_m$ and any $(c_1, \dots, c_m)$. This means a stationary setup,

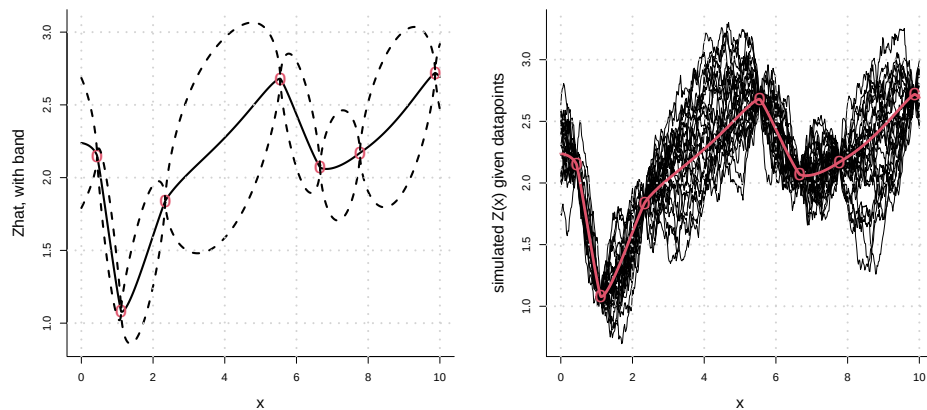


Figure 14.5: *Bayesian kriging*: covariance function $\sigma^2 \exp(-\lambda|x-y|^q)$ with $\sigma = 0.50$, $\lambda = 0.50$, $q = 1.25$, prior level $\xi = 2.00$. Left panel: the spatial interpolator, with pointwise 90 percent confidence band. Right panel: 25 posterior simulations.

where $Z(x)$ and $Z(x+r)$ have a correlation independent of position x . Suppose now that $Z(x_i) = z_i$ have been observed at positions $x_1 < \dots < x_n$. For a given x , using general results from conditioning in multinormal setups, see Ex. 1.44, find the distribution of $Z(x)$ given these $z_1, \dots, z_n$. With Σ the variance matrix of all $K(|x_i - x_j|)$, show in particular that indeed

$$\begin{aligned}\widehat{Z}(x) &= E\{Z(x) \mid \text{data}\} = \xi + k(x)^t \Sigma^{-1} (z_d - \xi \mathbf{1}), \\ \text{Var}\{Z(x) \mid \text{data}\} &= \sigma^2 \{1 - k(x)^t \Sigma^{-1} k(x)\},\end{aligned}$$

with $k(x)$ the vector $(K(|x - x_1|), \dots, K(|x - x_n|))^t$, and $\mathbf{1}$ the vector $(1, \dots, 1)^t$. Verify that this conditional variance is zero at locations $x_1, \dots, x_n$ where $Z(x)$ is actually observed. This way of carrying out spatial interpolation (along with variations and extensions) is called *kriging*, or here *Bayesian kriging*, particularly in context where a, σ, K have been arrived at as with eliciting of priors.

(b) Assume now that the scientific team has come back from their expedition and reports that for $n = 7$ positions of x_i they found $Z(x_i) = z_i$, as below. For this illustration take $\xi = 2$, $\sigma = 0.50$, $K(r) = \exp(-|r|^q)$, with $q = 1.25$. Plot this correlation function. Compute and plot the interpolator $\widehat{Z}(x)$. Construct a version of Figure 14.5, left panel, with the 90 percent pointwise confidence band $\widehat{Z}(x) \pm 1.645 \widehat{\kappa}(x)$.

x: 0.44, 1.11, 2.34, 5.55, 6.66, 7.77, 9.87
z: 2.15, 1.08, 1.84, 2.68, 2.07, 2.17, 2.72

(c) For constructing Figure 14.5, left panel, you should also play with different values of the smoothness parameter q , in the correlation function $K(r) = \exp(-\lambda r^q)$. The geostatistical default value, for various applications, is $q = 1$, with bigger values leading to more smoothness and smaller values to more ruggedness. The mathematically allowed span of such smoothness parameter values is $q \in (0, 2)$.

(d) Above we have reached clear, understandable, implementable, plottable formulae for the mean and standard deviation for each not-observed $Z(x)$ given the data $Z(x_1) = z_1, \dots, Z(x_n) = z_n$ (then used for the tiny set of $n = 7$ data locations). To properly

understand the full conditional process, and to be able to simulate realisations from this Bayesian nonparametrics posterior distribution, we also need the conditional covariance structure. Show indeed that

$$K^*(u_1, u_2) = \text{cov}\{Z(u_1), Z(u_2) \mid \text{data}\} = \sigma^2 \{K(|u_1 - u_2|) - k(u_1)^t \Sigma^{-1} k(u_2)\}.$$

Using the general Gaussian process scheme of Ex. ??, simulate say 25 realisations from the distribution of $Z = \{Z(x) : x \in [0, 10]\}$ given the observed data, and plot these, as with Figure 14.5, right panel.

(e) Add on more more datapoint, at some (x_0, z_0) . Re-do the analysis, and check both how the spatial interpolation curve is changed and how precision is bettered. Check that unless $q = 1$, the interpolation curve changes also outside the interval in which the new datapoint is located.

(f) The above analysis, for $Z(\cdot)$ given values at the sampled locations, needs the prior eliciting of mean level ξ , uncertainty σ , and the correlation function. There are various ways of turning the machinery into empirical Bayes, letting the data inform us of components of the full prior. Spell out one version, and carry out the modified analysis, where ξ is estimated from the data, while keeping the σ and $K(\cdot)$.

(g) Briefly think through how the kriging apparatus above may be extended to 2-dim and even 3-dim.

ToDo & what remains

Details remain for Beta process, e.g. $\mathcal{A}$ is closed in $D[0, 1]$, jumps inside $[0, 1]$, briefly Lévy measure. Details remain for updating Dirichlet and Beta. Should include the after all nice enough $\pi(\beta \mid \text{data})$ for Nils Cox $1 - dA_i(s) = \{1 - dA_0(s)\}^{\exp(x_i^t \beta)}$. Relate to Ch10. Bayesian kriging. A bit Gaussian processes. Perhaps briefly $f = \exp(G) / \int \exp(G)$ random density. BvM for Dirichlet and Beta processes. so Doksum–Dirichlet and Parzen–Dirichlet, via Nils-Sonia. Stick-breaking for Beta. A bit on models with clustering via DP. A bit on dependent DPs via Sethuraman representation. Nils generalisation to one parameter more via stick-breaking. Nils-Emil details in place for Poisson on top of Gamma, perhaps the Police Twitter as a story. Comparing k groups with Beta processes.

Leftovers and extra material

Ex. 14.30 A stick-breaking presentation of the Dirichlet process. (xx co be cleaned. Tiwari and Sethuraman. xx) Consider $\xi_1, \xi_2, \dots$ i.i.d. from P_0 , along with B_j i.i.d. from $\text{Beta}(1, a)$. From the latter we build probability weights $w_1 = B_1$ and then $w_j = \bar{B}_1 \cdots \bar{B}_{j-1} B_j$ for $j \geq 2$, with $\bar{B}_j = 1 - B_j$ notation. Below we establish the so-called stick-breaking or Tiwari-Sethuraman representation of the Dirichlet process, that

$$P = \sum_{j=1}^{\infty} w_j \delta(\xi_j) \quad \text{is a Dir}(aP_0). \quad (14.3)$$

(a) Show that the representation above leads to the stochastic equation

$$P(A) =_d BX + \bar{B}P(A), \quad \text{with } X = I(\xi \in A), \ B \sim \text{Beta}(1, a), \text{ and } \bar{B} = 1 - B.$$

so with $\theta = P(A)$, for a given set A , must verify that $\theta =_d BX + \bar{B}\theta$. should verify this for $\theta \sim \text{Beta}(ap_0, a(1-p_0))$, with $p_0 = P_0(A)$. so θ is equal to $\bar{B}\theta'$ with probability $1-p_0$, and $B + \bar{B}\theta$ with probability p_0 .

$$\mathbb{E} h(\theta) = p_0 \mathbb{E} h(B + \bar{B}\theta) + (1-p_0) \mathbb{E} h(\bar{B}\theta).$$

there are more heavy-handed arguments leading to this statement. more direct, satisfying but somewhat messy: show equality of all moments. for the left hand side:

$$m_n = \mathbb{E} \theta^n = \frac{\Gamma(a)}{\Gamma(ap_0)} \frac{\Gamma(ap_0 + n)}{\Gamma(a + n)}.$$

must verify that we have the same result for the right hand side. we have

$$m'_n = \mathbb{E} (BX + \bar{B}\theta)^n = \mathbb{E} (\bar{B})^n m_n + p_0 \sum_{k=0}^{n-1} \binom{n}{k} \mathbb{E} B^{n-k} \bar{B}^k m_k.$$

here, for $B \sim \text{Beta}(1, a)$, density $a(1-b)^{a-1}$, we find

$$\mathbb{E} B^r \bar{B}^s = a \frac{\Gamma(r+1)\Gamma(a+s)}{\Gamma(a+r+s+1)}$$

and $\mathbb{E} \bar{B}^n = a/(a+n)$ (check). so

$$(1 - a/(a+n))m_n = p_0 \sum_{k=0}^{n-1} \binom{n}{k} a \frac{(n-k)!\Gamma(a+k)}{\Gamma(a+n+1)} m_k,$$

recursive moment equation

$$m_n = p_0 \frac{a n!}{n\Gamma(a+n)} \sum_{k=0}^{n-1} \frac{\Gamma(a+k)}{k!} m_k.$$

needed here, to land the moment argument:

$$\sum_{k=0}^{n-1} \frac{\Gamma(a+k)}{k!} = \frac{n}{a} \frac{\Gamma(a+n)}{n!}.$$

(b) (xx let's see. too much, but clarifying. a bit from [Hjort and Ongaro \(2006\)](#). xx) For given $a > 0$ and $p_0 \in (0, 1)$, consider G_1, G_a, G_{1+a} independent gammas with parameters $(1, 1), (a, 1), (1+a, 1)$, and independent of a Bernoulli $Y \sim \text{binom}(1, p_0)$. The task is to verify the interesting claim that $\theta G_{1+a} =_d Y G_1 + G_a \theta$. the rhs is $G_1 + G_a \theta$ and $G_a \theta$ with probabilities p_0 and $1-p_0$. Do this by establishing that the moment sequences are the same for the left and right hand side. with

$$m_j = \mathbb{E} \theta^n = \frac{\Gamma(a)}{\Gamma(ap_0)\Gamma(aq_0)} \frac{\Gamma(ap_0 + j)\Gamma(aq_0)}{\Gamma(a + j)},$$

the lhs has $m_n \Gamma(1+a+n)/\Gamma(1+a)$, and the rhs has

$$\begin{aligned} \mathbb{E} (\text{rhs})^n &= p_0 \sum_{j=0}^n \binom{n}{j} \mathbb{E} G_1^j \mathbb{E} (G_a \theta)^{n-j} + (1-p_0) \mathbb{E} (G_a \theta)^n \\ &= p_0 \sum_{j=0}^n \binom{n}{j} \frac{\Gamma(1+j)}{\Gamma(1)} \frac{\Gamma(a+n-j)}{\Gamma(a)} m_{n-j} + (1-p_0) \frac{\Gamma(a+n)}{\Gamma(a)} m_n \end{aligned}$$

then divide with G_{1+a} to get the Dirichlet thing with $\theta =_d BX + \bar{B}\theta$, with $B = G_a/G_{1+a} \sim \text{Beta}(1, a)$.

(c) (xx clean this. xx) For any subset A , we are to prove that

$$P(A) = \sum_{j=1}^{\infty} w_j I_j, \quad \text{with } I_j = I(\xi_j \in A), \quad \text{is a Beta}(aP_0(A), aP_0(A^c)).$$

First, for a tall n , such that the rest R_n above is small, we work with $P_n(A) = \sum_{j=1}^n w_j I_j$ as a close approximation to the full infinite-sum $P(A)$. For a given ξ_j sequence, use the summing over cells property of the Dirichlet (xx point to Ch1 xx) to show that

$$(P_n(A), P_n(A^c)) \sim \text{Beta}(G_n(A), G_n(A^c)),$$

where $G_n(A) = \sum_{j=1}^n I_j / (n + a)$ and $G_n(A^c) = \{a + \sum_{j=1}^n (1 - I_j)\} / (n + a)$. Then, taking now also the randomness of the $I_j = I(\xi_j \in A)$ into account, show that $G_n(A)$ and $G_n(A^c)$ tend to $P_0(A)$ and $P_0(A^c)$.

(d) Generalise the above arguments to show that $(P_n(A_1), \dots, P_n(A_k)) \rightarrow_d (D_1, \dots, D_k)$, a Dirichlet $(aP_0(A_1), \dots, aP_0(A_k))$. Conclude that P of (14.3) indeed is a $\text{Dir}(aP_0)$. (xx note: discreteness. simulation. xx)

(e) Some more. Some consequences. Easy to simulate, alternative to the fine grid method above. Identities.

(f) (xx round off. perhaps the third moment. xx)

Notes and pointers

point to Nils Paris: $\pi(\mu) \text{EL}_n(\mu)$, with the empirical likelihood, works, and gives the same limit as for $\pi(\mu | \text{data})$ from a Dirichlet aF_0 prior.

Bayesian bootstrap, Rubin (1981) and Hjort (1985a). classic bootstrap, Efron and Tibshirani (1993).

well: Doksum (1974b), Hjort (1976), Phadia (1973), Ghosh and Ramamoorthi (2003), McQueen. Hjort (2003), Hjort and Ongaro (2005), Hjort and Ongaro (2006). Ghosal and van der Vaart (2017), Phadia (1973), Hjort (1976), Hjort (2003) in the Green et al. (2003) book. Sethuraman (1994), Sethuraman and Tiwari (1982), ... Ferguson (1973, 1974), ... Hjort and Petrone (2007), Hjort and Walker (2009), ... mention Kingman, Antoniak, Korwar, Hollander, ... Ghosh et al. (2006)

footnote re Ferguson: the father Thomas invents the Dirichlet process; the son Chris Jesus wins the World Series of Poker.

Dirichlet processes are discrete, Ex. 14.12: there are various other proofs; this is from two pages of Hjort (1976).

rephrase: Lindley (1972) noted in his review of general Bayesian methodology that Bayesians up to then had been ‘embarrassingly silent’ in the area of nonparametric statistics. He pointed out that there were in principle no conceptual difficulties with combining ‘Bayesian’ with ‘nonparametric’, but indirectly acknowledged that the mathematical details in such constructions would have to be more complicated.

Lindley Cromwell’s Rule: ‘I beseech you, in the bowels of Christ, think it possible that you may be mistaken’. Hence reserve an epsilon on that possibility that something else goes on, in your priors.

product integrals: Johansen, Gill, Andersen et al. (1993), Hjort (1990b), Gill and Johansen (1990), ... Aalen, Johansen, Markov processes.

quantile things: Doksum (1974a), Doksum and Stevers (2006), Doss and Gill (1992), Newton (2002), Parzen (1979, 1982, 2004), Sheather and Marron (1990), ...

Part II

100 Statistical Stories

Part III

Appendix

References

- Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726.
- Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. In Klonecki, W., Kozek, A., and Rosiński, J., editors, *Mathematical Statistics and Probability Theory. Lecture Notes in Statistics*, pages 1–25. Springer.
- Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8:907–925.
- Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, 2:951–972.
- Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and Event History Analysis. A process point of view*. Springer, New York.
- Aalen, O. O. and Gjessing, H. K. (2004). Survival models based on the Ornstein–Uhlenbeck process. *Lifetime Data Analysis*, 10:407–423.
- Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Dover, New York.
- Aït-Sahalia, Y. and Jacod, J. (2014). *High-Frequency Financial Econometrics*. Princeton University Press, Princeton.
- Andersen, P. K. and Borgan, Ø. (1985). Counting process models for life history data: A review [with discussion and reply]. *Scandinavian Journal of Statistics*, 12:97–158.
- Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, Berlin.
- Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24:3–30.
- Ashworth, S., Berry, C. R., and de Mesquita, E. B. (2021). *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton University Press, Princeton.
- Aursnes, I., Tvette, I. F., Gåsemyr, J., and Natvig, B. (2005). Suicide attempts in clinical trials with paroxetine randomised against placebo. *BMC Medicine*, xx:1–5.
- Aursnes, I., Tvette, I. F., Gåsemyr, J., and Natvig, B. (2006). Even more suicide attempts in clinical trials with paroxetine randomised against placebo. *BMC Psychiatry*, xx:1–3.
- Ball, P. (1999). *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*. American Academy for the Advancement of Science, Washington.
- Barfort, S., Klemmensen, R., and Larsen, E. G. (2020). Longevity returns to political office. *Political Science Research and Methods*, 9:658–664.
- Barry, D. and Hartigan, J. A. (1987a). Asynchronous distance between homologous dna sequences. *Biometrics*, 43:261–276.

- Barry, D. and Hartigan, J. A. (1987b). Statistical analysis of hominoid molecular evolution. *Statistical Science*, 2:191–210.
- Bartolucci, F. and Lupparelli, M. (2008). Focused Information Criterion for capture-recapture models for closed populations. *Scandinavian Journal of Statistics*, 35:629–649.
- Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC Press, London.
- Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, 15:377–180.
- Bernstein, F. (1924). Ergebnisse einer biostatistischen zusammenfassenden Betrachtung über die erblichen Blutstrukturen des Menschen. *Klinische Wochenschriften*, 3:1467–1495.
- Bickel, P. J. and Doksum, K. A. (2015). *Mathematical Statistics [2nd ed.]*. CRC Press, New York.
- Billingsley, P. (1961). *Statistical Inference for Markov Processes*. Chicago University Press, Chicago.
- Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
- Billingsley, P. (1995). *Probability and Measure. Third Edition*. Wiley, New York.
- Blower, J. G., Cook, L. M., and Bishop, J. A. (1981). *Estimating the Size of Animal Populations*. Allen & Unwin, London.
- Boitsov, V. D., Karsakov, A. L., and Trofimov, A. G. (2012). Atlantic water temperature and climate in the barents sea, 2000–2009. *ICES Journal of Marine Science*, 69:833–840.
- Bolt, U. (2013). *Faster Than Lightning: My Autobiography*. HarperSport, London.
- Borgan, Ø., Fiaccone, R. L., Henderson, R., and Barreto, M. L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in brazil. *Scandinavian Journal of Statistics*, 34:53–69.
- Brazzale, A. R. and Davison, A. C. (2008). Accurate parametric inference for small samples. *Statistical Science*, 23:465–484.
- Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied asymptotics: Case studies in small-sample statistics*. Cambridge University Press, Cambridge.
- Breiman, L. (2001). Statistical modeling: The two cultures [with comments and a rejoinder by the author]. *Statistical Science*, 16:199–231.
- Brown, L. D. (1986). Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *Lecture Notes-Monograph Series*, 9:i–279.
- Brunborg, H., Lyngstad, T. H., and Urdal, H. (2003). Accounting for genocide: How many were killed in Srebrenica? *European Journal of Population*, 19:229–248.
- Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, 84:772–793.
- Casella, G. and Berger, R. L. (2013). *Statistical Inference [2nd ed.]*. Duxbury, Canberra.
- Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *American Statistician*, 46:167–174.
- Claeskens, G., Cunen, C., and Hjort, N. L. (2019). Model selection via Focused Information Criteria for complex data in ecology and evolution. *Frontiers in Ecology and Evolution*, 7:415–428.
- Claeskens, G. and Hjort, N. L. (2003). The focused information criterion [with discussion and a rejoinder]. *Journal of the American Statistical Association*, 98:900–916.
- Claeskens, G. and Hjort, N. L. (2008a). Minimizing average risk in regression. *Econometric Theory*, 24:493–527.

- Claeskens, G. and Hjort, N. L. (2008b). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.
- Claeskens, G. and Hjort, N. L. (2027). The focused information criterion. *Annual Review of Statistics and Its Application*, xx:xx–xx.
- Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances*, 4:1–9.
- Clauset, A. (2020). On the frequency and severity of interstate wars. In Gleditsch, N. P., editor, *Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences*, pages 113–128. Springer, Berlin.
- Clevenson, M. L. and Zidek, J. V. (1975). Simultaneous estimation of the means of independent Poisson laws. *Journal of the American Statistical Association*, 70:698–705.
- Cox, D. R. (1958). Some problems with statistical inference. *The Annals of Mathematical Statistics*, 29:357–372.
- Cox, D. R. (1972). Regression models and life-tables [with discussion]. *Journal of the Royal Statistical Society: Series B*, 34:187–202.
- Cox, D. R. and Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. *Journal of the Royal Statistical Society Series B*, 21:195–200.
- Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman & Hall, London.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.
- Cramér, H. (1976). Half a century with probability theory: some personal reflections. *Annals of Probability Theory*, 4:509–546.
- Cunen, C. (2015). Mortality and Nobility in the Wars of the Roses and Game of Thrones. *FocuStat Blog, University of Oslo*, iv.
- Cunen, C., Hermansen, G. H., and Hjort, N. L. (2018). Confidence distributions for change points and regime shifts. *Journal of Statistical Planning and Inference*, 195:14–34.
- Cunen, C. and Hjort, N. L. (2015). Optimal inference via confidence distributions for two-by-two tables modelled as Poisson pairs: fixed and random effects. In Nair, V., editor, *Proceedings of the 60th World Statistics Congress, ISI Rio*, pages xx–xx. Springer, Rio.
- Cunen, C. and Hjort, N. L. (2022). Combining information from diverse sources: the II-CC-FF paradigm. *Scandinavian Journal of Statistics*, 49:625–656.
- Cunen, C. and Hjort, N. L. (2025). Survival and event history models and methods via Gamma processes. Technical report, University of Oslo. Technical report.
- Cunen, C., Hjort, N. L., and Nygård, H. M. (2020a). Statistical sightings of better angels. *Journal of Peace Research*, 57:221–234.
- Cunen, C., Hjort, N. L., and Schweder, T. (2020b). Confidence in confidence distributions! *Proceedings of the Royal Society, A*, 476:1–5.
- Cunen, C., Walløe, L., and Hjort, N. L. (2020c). Focused model selection for linear mixed models, with an application to whale ecology. *Annals of Applied Statistics*, 14:872–904.
- Dæhlen, I. and Hjort, N. L. (2025). Model robust hybrid likelihood. *Journal of Statistical Planning and Inference*, 241:xx–xx.
- Dæhlen, I., Hjort, N. L., and Hobæk Haff, I. (2024). Accurate bias estimation with applications to focused model selection [to appear]. *Scandinavian Journal of Statistics*, 51:724–759.
- Dagsvik, J. K., Fortuna, M., and Moen, S. H. (2020). How does temperature vary over time?: Evidence on the stationary and fractal nature of temperature fluctuations. *Journal of the Royal Statistical Society A*, pages 883–908.

- Daniell, P. J. (1920). Observations weighted according to order. *American Journal of Mathematics*, pages 222–236.
- Davison, A. (2003). *Statistical Models*. Cambridge University Press, Cambridge.
- De Blasi, P. and Hjort, N. L. (2007). Bayesian survival analysis in proportional hazard models with logistic relative risk. *Scandinavian Journal of Statistics*, 34:229–257.
- DeGroot, M. H. (1970). *Optimal Statistical Decisions*. John Wiley & Sons, Hoboken, N.J.
- Dickter, R. (2014). *Number Time Archetype: The Significance of Magic Squares in China and the West*. Grass Valley, Grass Valley.
- Dickter, R. (2026). *The Secret of the Magic Square: Numerology in Art and Architecture*. Grass Valley, Grass Valley.
- Doksum, K. A. (1974a). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2:267–277.
- Doksum, K. A. (1974b). Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2:183–201.
- Doksum, K. A. and Stevers, G. L. (2006). Plotting with confidence: graphical comparisons of two populations. *Biometrika*, 63:421–434.
- Doss, H. and Gill, R. D. (1992). An elementary approach to weak convergence for quantile processes, with applications to censored survival data. *Journal of the American Statistical Association*, 87:869–877.
- Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge.
- Eddington, A. S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, London.
- Edwards, A. W. F. (1974). The history of likelihood. *International Statistical Review*, 42:9–15.
- Edwards, A. W. F. (2025). *Latin Square: Essays in Defense of R.A. Fisher*. River Cam Books, Cambridge.
- Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, Cambridge.
- Efron, B. (2023). *Exponential Families in Theory and Practice*. Cambridge University Press, Cambridge.
- Efron, B. and Morris, C. (1977). Stein’s paradox in statistics. *Scientific American*, 236:119–127.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Einmahl, J. H. J. and Smeets, S. G. W. R. (2011). Ultimate 100 m world records through extreme-value theory. *Statistica Neerlandica*, 65:32–42.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, London.
- Fagerland, M., Lydersen, S., and Laake, P. (2017). *Statistical Analysis of Contingency Tables*. Chapman and Hall/CRC, New York.
- Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.
- Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, 2:615–629.
- Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

- Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, 80:758–770.
- Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.
- Franklin, B. (1793). *The Autobiography of Benjamin Franklin*. Dover, New York. Reprinted from Dover, New York, 1996.
- Friesinger, A. (2004). *Mein Leben, mein Sport, meine besten Fitness-Tipps*. Goldmann, Berlin.
- Frigessi, A. and Hjort, N. L. (2002). Statistical methods for discontinuous phenomena. *Journal of Nonparametric Statistics*, 14:1–5.
- Gao, L., Shen, H., and Liu, C. (2021). Statistical analysis of songs by composers of the Romantic era: A comparison of F. Schubert and his contemporaries. *Música Hodie*, 21:xx–xx.
- Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Friedrich Perthes und I.H. Besser, Hamburg.
- Geißler, A. (1889). Beiträge zur Frage des Geschlechts verhältnisses der Geborenen. *Zeitschrift des königlichen sächsischen statistischen Bureaus*, 35:1–24.
- Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and Other Stories*. Cambridge University Press, Cambridge.
- Gelman, A., Hill, J., and Vehtari, A. (2022). *Regression and Other Stories*. Cambridge University Press, Cambridge.
- Gelman, A. and Nolan, D. (2002). A probability model for golf putting. *Teaching Statistics*, 24:93–95.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.
- Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge.
- Ghosh, J. K., Hjort, N. L., Messan, C., and Ramamoorthi, R. V. (2006). Bayesian bivariate survival estimation. *Journal of Statistical Planning and Inference*, 136:2297–2308.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Verlag, Berlin.
- Ghosh, M. (2002). Basu’s theorem with applications: a personalistic review. *Sankhya*, 35:721–741. Special issue in memory of D. Basu.
- Gilby, W. H. and Pearson, K. (1911). On the significance of the teacher’s appreciation of general intelligence. *Biometrika*, 8:94–108.
- Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Annals of Statistics*, 18:1501–1555.
- Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17:295–314.
- Gjessing, H. K., Aalen, O. O., and Hjort, N. L. (2003). Frailty models based on Lévy processes. *Advances in Applied Probability*, 35:532–550.
- Gleditsch, N. P. (2020). *Lewis Fry Richardson: His Intellectual Legacy and Influence in the Social Sciences (edited book)*. Springer, Berlin.
- Goudie, I. B. J. and Goudie, M. (2007). Who captures the marks for the Petersen estimator? *Journal of the Royal Statistical Society, Series A*, 170:825–839.
- Green, P. J., Hjort, N. L., and Richardson, S., editors (2003). *Highly Structured Stochastic Systems*. Oxford University Press.

- Grønneberg, S. and Hjort, N. L. (2012). On the errors committed by sequences of estimator functionals. *Mathematical Methods of Statistics*, 20:327–346.
- Groß, J. and Möller, A. (2024). The zero degree of freedom non-central chi squared distribution for ensemble postprocessing. *EMOS Meteorologische Zeitschrift*, xx:1–20.
- Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrics*, 11:1–12.
- Hall, P. (1927). The distribution of means for samples of size N drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika*, 19:240–245.
- Hanche-Olsen, H. and Holden, H. (2010). The Kolmogorov-Riesz compactness theorem. *Expositiones Mathematicae*, 28:385–394.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 9:97–109.
- Heger, A. (2011). Jeg og jordkloden. *Dagsavisen*, Dec. 16.
- Hermansen, G. H. and Hjort, N. L. (2015). Bernstein–von Mises theorems for nonparametric function analysis via locally constant modelling: A unified approach. *Journal of Statistical Planning and Inference*, 166:138–157.
- Hermansen, G. H., Hjort, N. L., and Kjesbu, O. S. (2016). Modern statistical methods applied on extensive historic data: Hjort liver quality time series 1859–2012 and associated influential factors. *Canadian Journal of Fisheries and Aquatic Sciences*, 73:273–295.
- Hindemith, P. (1950). *Johann Sebastian Bach: Ein verpflichtendes Erbe: Rede zum Bachfest in Hamburg am 12. September 1950*. Insel-Verlag, Wiesbaden.
- Hjort, J. (1914). *Fluctuations in the Great Fisheries of Northern Europe, Viewed in the Light of Biological Research*. Conseil Permanent International Pour l’Exploration de la Mer, Copenhagen.
- Hjort, J. and Petersen, C. G. J. (1905). *Short review of the results of the international fisheries investigation*, volume 3. Conseil Permanent International Pour l’Exploration de la Mer, Rapport et Proces-Verbaux des Reunions, København.
- Hjort, N. L. (1976). The Dirichlet Process Applied to Nonparametric Estimation Problems. Technical report, University of Tromsø, University of Tromsø. Cand. real. thesis, Universities of Tromsø and Oslo.
- Hjort, N. L. (1985a). Bayesian nonparametric bootstrap confidence intervals. Technical report, Department of Statistics, Stanford University.
- Hjort, N. L. (1985b). Discussion contribution to P.K. Andersen and Ø. Borgan’s paper ‘Counting process models for life history data: A review’. *Scandinavian Journal of Statistics*, 12:141–151.
- Hjort, N. L. (1986a). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scandinavian Journal of Statistics*, 13:63–85.
- Hjort, N. L. (1986b). *Notes on the Theory of Statistical Symbol Recognition*. Norwegian Computing Centre, Oslo.
- Hjort, N. L. (1988a). The eccentric part of the non-central chi square. *The American Statistician*, 42:130–132.
- Hjort, N. L. (1988b). On large-sample multiple comparison methods. *Scandinavian Journal of Statistics*, 15:259–271.
- Hjort, N. L. (1990a). Goodness of fit tests for life history data based on cumulative hazard rates. *Annals of Statistics*, 18:1221–1258.
- Hjort, N. L. (1990b). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics*, 18:1259–1294.

- Hjort, N. L. (1992). On inference in parametric survival data models. *International Statistical Review*, xx:355–387.
- Hjort, N. L. (1994a). The exact amount of t-ness that the normal model can tolerate. *Journal of the American Statistical Association*, 89:665–675.
- Hjort, N. L. (1994b). Should the Olympic speedskaters run the 500 meter twice? Technical report, Department of Mathematics, University of Oslo. Also on arXiv, 2603.22298.
- Hjort, N. L. (2003). Topics in nonparametric Bayesian statistics. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*. Oxford University Press.
- Hjort, N. L. (2007). And Quiet Does Not Flow the Don: Statistical analysis of a quarrel between Nobel laureates. In Østreng, W., editor, *Conciliation*, pages 134–140. Centre for Advanced Research, Oslo.
- Hjort, N. L. (2008). Discussion of P.L. Davies’ article ‘Approximating data’. *Journal of the Korean Statistical Society*, 37:221–225.
- Hjort, N. L. (2014). Discussion of Efron’s article ‘Estimation and accuracy after model selection’. *Journal of the American Statistical Association*, 110:1017–1020.
- Hjort, N. L. (2017a). Cooling of Newborns and the Difference Between 0.244 and 0.278. *FocuStat Blog, University of Oslo*, xv.
- Hjort, N. L. (2017b). The Semifinals Factor for Skiing Fast in the Finals. *FocuStat Blog, University of Oslo*, xv.
- Hjort, N. L. (2018a). Overdispersed Children. *FocuStat Blog, University of Oslo*, xxi.
- Hjort, N. L. (2018b). Towards a More Peaceful World [insert ‘!’ or ‘?’ here]. *FocuStat Blog, University of Oslo*, xvii.
- Hjort, N. L. (2019a). The Magic Square of 33. *FocuStat Blog, University of Oslo*, xxi.
- Hjort, N. L. (2019b). Sudoku Solving by Probability Models and Markov Chains. *FocuStat Blog, University of Oslo*, xxi.
- Hjort, N. L. (2026). Counting the uncounted: How many were killed in Guatemala, 1978-1995? In de Bin, R., editor, *Proceedings of the 40th International Workshop on Statistical Modelling, Oslo, 2026*, pages 1–10. University of Oslo, Department of Mathematics, Oslo.
- Hjort, N. L. and Claeskens, G. (2003a). Frequentist model averaging [with discussion and a rejoinder]. *Journal of the American Statistical Association*, 98:879–899.
- Hjort, N. L. and Claeskens, G. (2003b). Rejoinder to the discussion of the Hjort and Claeskens and Claeskens and Hjort papers. *Journal of the American Statistical Association*, 98:917–925.
- Hjort, N. L. and Fenstad, G. (1992). On the last time and the number of times an estimator is more than ε from its target value. *The Annals of Statistics*, 20:469–489.
- Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.
- Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24:1619–1647.
- Hjort, N. L. and Koning, A. J. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, 14:113–132.
- Hjort, N. L. and Lumley, T. (1993). Normalised local hazard plots. Technical report, Department of Statistics, University of Oxford, Oxford.
- Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Annals of Statistics*, 37:1079–1111.

- Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2018). Hybrid combinations of parametric and empirical likelihoods. *Statistica Sinica*, 27:2389–2407.
- Hjort, N. L. and Ongaro, A. (2005). Exact inference for random Dirichlet means. *Statistical Inference for Stochastic Processes*, 8:227–254.
- Hjort, N. L. and Ongaro, A. (2006). On the distribution of random Dirichlet jumps. *Metron*, LXIV:61–92.
- Hjort, N. L. and Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. In Nair, V., editor, *Advances in Statistical Modeling and Inference: Essays in Honor of Kjell Doksum*, pages 463–492. World Scientific, New Jersey.
- Hjort, N. L. and Pollard, D. B. (1993). Asymptotics for minimisers of convex processes. Technical report, Department of Mathematics, University of Oslo.
- Hjort, N. L. and Schweder, T. (2018). Confidence distributions and related themes: introduction to the special issue. *Journal of Statistical Planning and Inference*, 195:1–13.
- Hjort, N. L. and Stoltenberg, E. A. (2021). The partly parametric and partly nonparametric additive risk model. *Lifetime Data Analysis*, 27:1–31.
- Hjort, N. L. and Stoltenberg, E. A. (2026). Probability proofs for the Stirling (and more): the ubiquitous rolde of $\sqrt{2\pi}$. *American Statistician*, xx:xx–xx.
- Hjort, N. L. and Varin, C. (2008). ML, PL, QL in Markov chain models. *Scandinavian Journal of Statistics*, 35:64–82.
- Hjort, N. L. and Walker, S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *Annals of Statistics*, 37:105–131.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960.
- Holum, D. (1984). *The Complete Handbook of Speed Skating*. High Peaks Cyclery, Lake Pacid.
- Hosmer, D. W. and Lemeshow, S. (1999). *Applied Logistic Regression*. Wiley, New York.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Annals Mathematical Statistics*, 53:73–101.
- Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York.
- Hveberg, K. (2019). *Lene din ensomhet langsomt mot min*. Aschehoug, Oslo.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge.
- Inlow, M. (2010). A moment generating function proof of the Lindeberg–Lévy central limit theorem. *American Statistician*, 64:228–230.
- Irwin, J. O. (1927). On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson’s type II. *Biometrika*, 19:225–239.
- Jacod, J. and Protter, P. (2004). *Probability Essentials. Second Edition*. Springer, Berlin.
- Jacod, J. and Shiryaev, A. (2013). *Limit Theorems for Stochastic Processes. Second Edition*. Springer, Berlin.
- James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379.
- Jamtveit, B., Jacobsen, A. U., and Wyller, T. B. (2018). Utvikling i andel administrativt personale i norske helseforetak. *Samfunnsøkonomen*, 6:17–21.

- Jamtveit, B., Jettestuen, E., and Mathiesen, J. (2009). Scaling properties of European research units. *Proceedings of the National Academy of Sciences*, 106:13160–13163.
- Jansen, D. (1994). *Full Circle*. Villard Books, New York.
- Jullum, M. and Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*, 27:951–981.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- Kahneman, D., Sibony, O., and Sunstein, C. R. (2020). *Noise: A Flaw in Human Judgment*. William Collins, London.
- Kallenberg, O. (2002). *Foundations of Modern Probability. Second Edition*. Springer, Berlin.
- Keener, R. W. (2011). *Theoretical Statistics: Topics for a Core Course*. Springer, New York.
- Kjesbu, O. S., Opdal, A. F., Korsbrekke, K., Devine, J. A., and Skjæraasen, J. E. (2014). Making use of Johan Hjort’s ‘unknown’ legacy: reconstruction of a 150-year coastal time-series on northeast Arctic cod (*Gadus morhua*) liver data reveals long-term trends in energy allocation patterns. *ICES Journal of Marine Science*, 71:2053–2063.
- Kjetsaa, G., Gustavson, S., Beckman, B., and Gil, S. (1984). *The Authorship of The Quiet Don [also published in Russian]*. Solum/Humanities Press, Oslo.
- Klein, R., Knudtson, M. D., Lee, K. E., Gangnon, R., and Klein, B. E. (2008). The Wisconsin epidemiologic study of diabetic retinopathy: XXII the twenty-five-year progression of retinopathy in persons with type 1 diabetes. *Ophthalmology*, 115:1859–1868.
- Klotz, J. (1972). Markov chain clustering of births by year. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability Theory*, 4:173–185.
- Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. *Annals of Statistics*, 1:373–379.
- Koehler, J. J. and Conley, C. A. (2003). The “hot hand” myth in professional basketball. *Journal of Sport and Exercise Psychology*, 25:253–259.
- Kolmogorov, A. N. (1933a). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer, Berlin.
- Kolmogorov, A. N. (1933b). Sulla determinazione empirica di una legge di distribuzione. *Giorn Ist Ital Attuar*, 4:83–91.
- Kusolitsch, N. (2010). Why the theorem of Scheffé should be rather called a theorem of Riesz. *Periodica Mathematica Hungarica*, 61:225–229.
- Laake, P., Laake, K., and Aaberge, R. (1985). On the problem of measuring the distance between distribution functions: Analysis of hospitalization versus mortality. *Biometrics*, 41:515–523.
- Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les évènements. Technical report, Mémoires de l’Académie Royale des Sciences, Présenté par Divers Savan, vol. 6, Paris.
- Laptook, A. e. a. (2017). Effect of therapeutic hypothermia initiated after 6 hours of age on death and disability among newborns with hypoxic-ischemic encephalopathy: A randomized clinical trial. *Journal of the American Medical Association*, 318:1550–1560.
- Larkey, P. D., Smith, R. A., and Kadane, J. B. (1989). It’s okay to believe in the “hot hand”. *Chance*, 2:22–30.
- Larsen, R. L. and Marx, M. L. (1981). *An Introduction to Mathematical Statistics and Its Applications*. Prentice-Hall, New York.
- Le May Doan, C. (2002). *Going For Gold*. McClelland & Stewart Publisher, Toronto.
- LeCam, L. (1986). The Central Limit Theorem around 1935. *Statistical Science*, 1:78–91.

- Lehmann, E. L. (1950). *Notes on the Theory of Estimation*. Berkeley University Press, Berkeley. Notes recorded by Colin Blyth.
- Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.
- Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.
- Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.
- Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer-Verlag, Berlin.
- Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation [2nd ed.]*. Springer, Berlin.
- Leike, A. (2001). Demonstration of the exponential decay law using beer froth. *European Journal of Physics*, 23:1–21.
- Lessing, D. (1997). *Walking in the Shade: Volume Two of My Autobiography, 1949 to 1962*. xx, xx.
- Li, G., Reid, N., and Varin, C. (2027). *Likelihood and Its Extensions*. CRC Press, New York.
- Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15:211–225.
- Lindley, D. V. (1972). *Bayesian Statistics: A Review*. Society for Industrial and Applied Mathematics, Philadelphia.
- Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model [with discussion]. *Journal of the Royal Statistical Society Series B*, 34:1–41.
- Lindqvist, B. H. (1978). A note on Bernoulli trials with dependence. *Scandinavian Journal of Statistics*, 5:205–208.
- Linero, A. and Rosalsky, A. (2013). On the Toeplitz lemma, convergence in probability, and mean convergence. *Stochastic Analysis and Applications*, 31:684–694.
- Lum, K., Price, M. E., and Banks, D. (2013). Applications of multiple systems estimation in human rights research. *American Statistician*, 24:191–200.
- Mardia, K. V., Kent, J. T., and Taylor, C. C. (1979). *Multivariate Analysis*. Academic Press, London.
- Markov, A. A. (1906). Распространение закона больших чисел на величины, зависящие друг от друга [Extending the law of large numbers for variables that are dependent of each other]. *Известия Физико-математического общества при Казанском университете (2-я серия)*, 15:124–156.
- Markov, A. A. (1913). Пример статистического исследования над текстом “Евгения Онегина”, иллюстрирующий связь испытаний в цепь [Example of a statistical investigation illustrating the transitions in the chain for the ‘Evgenii Onegin’ text]. *Известия Академии Наук, Санкт-Петербург (6-я серия)*, 7:153–162.
- Marshall, E. C. and Spiegelhalter, D. J. (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis*, 2:409–444.
- McCullagh, P. (2002). What is a statistical model? [with discussion]. *Annals of Statistics*, 30:1225–1310.
- McCullagh, P. (2022). *Ten Projects in Applied Statistics*. Springer, Berlin.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models [2nd. ed.]*. Chapman and Hall, London.
- Miller, J. B. and Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, 86:2019–2047.
- Miller, J. B. and Sanjurjo, A. (2021). Is it a fallacy to believe in the hot hand in the NBA three-point contest? *European Economic Review*, 138:103771.

- Mykland, P. A. and Zhang, L. (2012). The econometrics of high frequency data. In Kessler, M., Lindner, A., and Sørensen, M., editors, *Statistical Methods for Stochastic Differential Equations*, pages 109–190. CRC Press.
- Mykland, P. A., Zhang, L., and Chen, D. (2019). The algebra of two scales estimation, and the S-TSRV: High frequency estimation that is robust to sampling times. *Journal of Econometrics*, 208:101–119.
- Newton, M. J. (2002). A conversation with Emanuel Parzen. *Statistical Science*, 17:357–378. Correction, op. cit., 467.
- Neyman, J. and Pearson, E. (1933). On the problem of the most efficient statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 68:289–337.
- Normand, S.-L. T. (1999). Tutorial in biostatistics: Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18:321–359.
- Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.
- Parzen, E. (1979). Nonparametric statistical data modeling [with discussion]. *Journal of the American Statistical Association*, 74:101–131.
- Parzen, E. (1982). Data modeling using quantile and density-quantile functions. In *Some Recent Advances in Statistics: Symposium Lisbon 1980*, pages 23–52. xx.
- Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science*, 19:652–662.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series*, 5(302):157–175.
- Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika*, 1:261–264.
- Petersen, C. G. J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, 6:5–84.
- Peterson, A. V. (1975). Nonparametric estimation in the competing risks problem. Technical report, Department of Statistics, Stanford University.
- Phadia, E. G. (1973). Minimax estimation of a cumulative distribution function. *Annals of Statistics*, 6:1149–1157.
- Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Books, Toronto.
- Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
- Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, 11:453–466.
- Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletins of the Calcutta Mathematical Society*, pages 81–91.
- Reeves, R. V. (2022a). *Of Boys and Men: Why the Modern Male is Struggling, Why it Matters, and What to Do About It*. Brookings Institution Press, Washington, D.C.
- Reeves, R. V. (2022b). Redshirt the boys. *The Atlantic*, October.
- Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, Berlin.
- Robert, C. and Casella, G. (2011). A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26:102–115.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

- Royden, H. L. and Fitzpatrick, P. M. (2010). *Real Analysis [4th ed.]*. Pearson Education Asia, Beijing.
- Rubin, H. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9:130–134.
- Rydén, J. (2020). On features of fugue subjects: A comparison of J.S. Bach and later composers. *Journal of Mathematics and Music*, pages 1–20.
- Sala-i Martin, X. (1997). I just ran two million regressions. *The American Economic Review*, 87:179–183.
- Saleh, J. H. (2019). Statistical reliability analysis for a most dangerous occupation: Roman emperor. *Palgrave Communication*, 5:1–7.
- Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43:142–1542.
- Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18:434–438.
- Schervish, M. J. (1995). *Theory of Statistics*. Springer, New York.
- Schömig, A., Mehili, J., de Waha, A., Seyfarth, M., Pahce, J., and Kastrati, A. (2008). A meta-analysis of 17 randomized trials of a percutaneous coronary intervention-based strategy in patients with stable coronary artery disease. *Journal of the American College of Cardiology*, 52:894–904.
- Schweder, T. (1980). Scandinavian statistics, some early lines of development. *Scandinavian Journal of Statistics*, 7:113–129.
- Schweder, T. (1999). Early statistics in the Nordic countries – when did the Scandinavians slip behind the British? *Bulletin of the International Statistical Institute*, 58:1–4.
- Schweder, T. and Hjort, N. L. (2002). Likelihood and confidence. *Scandinavian Journal of Statistics*, 29:309–322.
- Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.
- Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.
- Sethuraman, J. and Tiwari, R. C. (1982). Convergence of Dirichlet measures and interpretation of their parameter. *Statistical Decision Theory and Related Topics, III*, 2:305–315.
- Shao, J. (1991). Second-order differentiability and jackknife. *Statistica Sinica*, 1:185–202.
- Sheather, S. J. and Marron, J. S. (1990). Kernel quantile estimation. *Journal of the American Statistical Association*, 80.
- Shiryayev, A. N. (1996). *Probability. Second edition*. Springer, Berlin.
- Shumway, R. H. and Stoffer, D. S. (2016). *Time Series Analysis and Its Applications [4th ed.]*. Springer, Heidelberg.
- Silver, N. (2012). *The Signal and the Noise: Why so Many Predictions Fail, but Some Don't*. Penguin.
- Simpson, R. J. S. and Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3:1243–1246.
- Sims, C. A. (2012a). Appendix: inference for the Haavelmo model. Technical report, Public Policy & Finance, Princeton University, Princeton, NJ.
- Sims, C. A. (2012b). Statistical modeling of monetary policy and its effects [Sveriges Riksbank Prize in Memory of Alfred Nobel lecture]. *American Economic Review*, xx:1–22.

- Singh, K., Xie, M., and Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *Annals of Statistics*, 33:159–183.
- Slud, E. (1989). Clipped Gaussian processes are never M-step Markov. *Journal of Multivariate Analysis*, 29:1–14.
- Smith, T. D. (1994). *Scaling Fisheries: The Science of Measuring the Effects of Fishing 1855–1955*. Cambridge University Press, Cambridge.
- Spiegelberg, W. (1901). *Aegyptische und Griechische Eigennamen aus Mumientiketten der Römischen Kaiserzeit*. Greek Inscriptions, Cairo.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206.
- Stigler, S. M. (1973). Studies in the history of probability and statistics. xxxii: Laplace, Fisher and the discovery of the concept of sufficiency. *Biometrika*, 60:439–445.
- Stigler, S. M. (1977). Do robust estimators work with real data? *Annals of Statistics*, 27:1055–1098.
- Stigler, S. M. (1983). Who discovered Bayes’s Theorem? *American Statistician*, 37:290–296.
- Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, 5:147–155.
- Stigler, S. M. (2006). How Ronald Fisher became a mathematical statistician. *Mathematics and Social Sciences*, 44:23–30.
- Stoltenberg, E. A. (2019). An MGF proof of the Lindeberg theorem. Technical report, Department of Mathematics, University of Oslo.
- Stoltenberg, E. A. and Hjort, N. L. (2020). Multivariate estimation of Poisson parameters. *Journal of Multivariate Analysis*, 175:1–19.
- Stoltenberg, E. A. and Hjort, N. L. (2021). Models and inference for on-off data via clipped Ornstein–Uhlenbeck processes. *Scandinavian Journal of Statistics*, 48:908–929.
- Stout, W. F. (1974). *Almost Sure Convergence*. Academic Press, New York.
- Student (1908). The probable error of a mean. *Biometrika*, 6:1–25.
- Tronstad, E. (2025). Window of change: Extensions of change-point analysis. Technical report, University of Oslo. Master Thesis.
- Tversky, A. and Gilovich, T. (1989). The cold facts about the “hot hand” in basketball. *Chance*, 2:16–21.
- Van Bavel, K., Moreels, S., Van de Putte, B., and Matthijs, K. (2011). Family size and intergenerational social mobility during the fertility transition: Evidence of resource dilution from the city of Antwerp in nineteenth century Belgium. In *Social Mobility and Demographic Behaviour: A Long-Term Perspective*, volume 24, pages xx–xx. Holland Holland.
- Varian, H. R. (1975). Distributive justice, welfare economics, and the theory of fairness. *Philosophy and Public Affairs*, 4:223–247.
- Voldner, B., Frøslie, K. F., Haakstad, L., Hoff, C., and Godang, K. (2008). Modifiable determinants of fetal macrosomia: role of lifestyle-related factors. *Acta Obstetrica et Gynecologica Scandinavica*, 87:423–429.
- von Bahr, B. (1965). On the convergence of moments in the central limit theorem. *Annals of Mathematical Statistics*, xx:808–818.
- von Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*. B.G. Teubner, Berlin.
- Walløe, L., Hjort, N. L., and Thoresen, M. (2019a). Major concerns about late hypothermia study. *Acta Paediatrica*, 108:588–589.

- Walløe, L., Hjort, N. L., and Thoresen, M. (2019b). Why results from Bayesian statistical analyses of clinical trials with a strong prior and small sample sizes may be misleading: The case of the NICHD Neonatal Research Network Late Hypothermia Trial. *Acta Paediatrica*, 108:1190–1191.
- Wardrop, R. L. (1995). Simpson’s paradox and the hot hand in basketball. *The American Statistician*, 49:24–28.
- Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge.
- Wissner-Gross, Z. (2020). Can you feed the hot hand? <https://fivethirtyeight.com/features/can-you-feed-the-hot-hand/>. Accessed: December 12, 2020.
- Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review [with discussion and a rejoinder]. *International Statistical Review*, 81:3–39.
- Zabriskie, B. N., Corcoran, C., and Senchaudhuri, P. (2021). A comparison of confidence distribution approaches for rare event meta-analysis. *Statistics in Medicine*, 40:5276–5297.
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81:446–451.
- Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100:1394–1411.

Subject index

at-risk process, 356, 360

counting process, 356, 360
counting process martingale, 356
cumulative hazard, 356

gamma-weibull model, 401

hazard rate, 355

intensity process, 360

Kaplan–Meier estimator, 358

Nelson–Aalen estimator, 365

predictable process, 356
predictable quadratic variation, 356
previsible quadratic variation process, 362

right-censoring, 356, 359

sandwich matrix, 392
stochastic process, 355
survival function, 355, 359

uninformative censoring, 356