Warwick, Workshop C, April 2008

Likelihood, Pseudo-Likelihood, Composite Likelihood for Markov Chain Models

Nils Lid Hjort
[joint with Cristiano Varin]

* Many models with complex dependencies where full ML is too impractical, e.g. spatial and spatial-temporal models, (hidden) Markov random fields, truncation models, etc.

* May try **PL**: product over local conditionals (CCL)

* May try **CL**: product over local joint likelihoods (CML)

* Difficult [in general] to assess consequences [how much is lost? what does PL or CL do when the model is not correct?]

* **Markov chains**: can do precise analysis

* **Model selection** with CL: the CLIC, the FCLIC ...

* **CL better than PL**: can also lead to new modelling strategies

Марков, А.А. (1906). Распространение закона больших чисел на величины, зависящие друг от друга. Известия Физико-математического обчества при Казанском университете **15** (2-я серия), 124–156.

Markov, A.A. (1913). Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цепь. Известия Академии Наук, Санкт-Петербург **7** (6-я серия), 153–162.

Hjort, N.L. and Varin, C. (2008). ML, QL, PL for Markov chain models. *Scandinavian Journal of Statistics* **35**, 64–82.

0 ML, PL, CL in spatial models

1 Markov chains: ML [classic]

2 CL

3 PL

4 Illustrations; Markov chains for DNA sequences

5 **Model robustness:** When the models are not correct

6 **Model selection:** CLIC, FCLIC

7 CL as **model building** tool; concluding comments

0. Spatial models: examples

(A) Markov Random Fields, defined on lattices:

$$f(x, \beta) = \frac{1}{Z(\beta)} \exp\{\beta_1 H_1(x) + \cdots + \beta_p H_p(x)\},$$

e.g. the Ising (1925) model, with $x_i \in \{-1, 1\}$ and

$$H(x) = \sum_i \#\{x_j \in x_{\partial i} \colon x_j = x_i\}.$$

ML difficult [but now doable]. **Much easier**, Besag (1974, 1975, 1977):

$$\mathrm{PL}(\beta) = \prod_i p_\beta(x_i \,|\, \mathrm{rest}).$$

(B) Hidden Markov Random Fields:

$$y_i = g(x_i) + \varepsilon(x_i) = g(x_i, \beta) + \varepsilon(x_i, \sigma, \phi),$$

perhaps a zero-mean stationary Gaußian noise process [image reconstructions, etc.]. PL ok for $\varepsilon(x)$ white noise process, but difficult in general.

(C) Gaußian Random Fields:

$$y \sim \mathrm{N}_n(X\beta, \sigma^2 R(\phi)).$$

ML doable, but difficult for n big, and properties not well enough understood. **Much easier:** CL (called QL in Hjort and Mohn, 1987, Hjort and Omre, 1994, etc.).

(D) Point process models:

$$f_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\{\theta_1 A_1(\mathbf{x}) + \cdots + \theta_p A_p(\mathbf{x})\},$$

where $\mathbf{x} = \{x_i\}$ is a set of points. ML difficult – as are PL and CL. Might encourage **new [quasi]models** that start with modelling over smaller areas.

(E) Lattice model from **truncated normal processes**:
$y_i = I\{Z(x_i) \geq c\}$, where Z has stationary covariance structure.
Here both ML and PL have difficulties. Easier:

$$\mathrm{CL}(\theta) = \prod_{\text{pairs}} f_\theta(y_i, y_j), \quad \text{or} \quad \mathrm{CL}(\theta) = \prod_{\text{triples}} f_\theta(y_i, y_j, y_k),$$

or bigger local neighbourhoods: Hjort and Omre (1994), Nott and Rydén (1999), Heagerty and Lele (1998), others.

The talk I'm not giving [today]:

926 children in Salvador, Brazil, followed from Oct 2000 to Jan 2002, twice-a-week 0–1 data on infant diarrhoea. Borgan, Henderson, Barreto (2007): event history analysis via variations on Aalen's additive hazard regression model. My approach:

$$y_i(t) = I\{Z_i(t) \geq c_i\} \quad \text{for child } i,$$

with

$$Z_i(t) = x_i(t)^{\mathrm{t}}\beta + \sigma_i \, \mathrm{OU}_i(t).$$

I am using **CL machinery** for estimation and inference.

1. Markov Chains

Observe chain $X_0, X_1, \ldots,$

$$\pi_{a,b} = \Pr_\theta\{X_i = b \,|\, x_{i-1} = a\} = p_{a,b}(\theta)$$

for $a, b = 1, \ldots, S$. **The Lik:**

$$\mathrm{l}_n(\theta) = \prod_{i=1}^{n} \Pr_\theta\{X_i = x_i \,|\, X_{i-1} = x_{i-1}\} = \prod_{a,b} p_{a,b}(\theta)^{N_{a,b}}.$$

The PL:

$$\begin{aligned} \mathrm{pl}_n(\theta) &= \prod_{i=1}^{n-1} \Pr_\theta\{X_i = x_i \,|\, \mathrm{rest}\} \\ &= \prod_{a,b,c} \left\{ \frac{p_a(\theta) p_{a,b}(\theta) p_{b,c}(\theta)}{p_a(\theta) p^{(2)}_{a,c}(\theta)} \right\}^{N_{a,b,c}}. \end{aligned}$$

The CL:

$$\begin{aligned} \mathrm{cl}_n(\theta) &= \prod_{i=1}^{n} \Pr_\theta\{X_{i-1} = x_{i-1}, X_i = x_i\} \\ &= \prod_{a,b} \{p_a(\theta) p_{a,b}(\theta)\}^{N_{a,b}}. \end{aligned}$$

Higher order versions [bigger windows] can be used for PL and CL.

ML theory: goes back to Anderson and Goodman (1957), Billingsley (1961a, 1961b). To reach result, need to sort out joint limit of

$$\sqrt{n}\{N_{a,b}/n - p_a(\theta)p_{a,b}(\theta)\} \to_d Z_{a,b}.$$

For $a, b, c, d = 1, \ldots, S$:

$$\mathrm{cov}(Z_{a,b}, Z_{c,d}) = p_a p_{a,b}(\delta_{a,c}\delta_{b,d} - p_c p_{c,d}) + p_{a,b}p_{c,d}(p_a\gamma_{a,c} + p_c\gamma_{d,a}),$$

with

$$\gamma_{a,b} = \sum_{k=0}^{\infty}(p_{a,b}^{(k)} - p_b).$$

ML theorem:

$$\sqrt{n}(\widehat{\theta} - \theta) \to_d \mathrm{N}(0, J^{-1}),$$

where

$$J = \sum_a p_a J_a = \sum_{a,b} p_a p_{a,b} u_{a,b} u_{a,b}^{\mathrm{t}},$$

with

$$u_{a,b}(\theta) = \frac{\partial \log p_{a,b}(\theta)}{\partial \theta}.$$

2. CL estimation

We have

$$\log \mathrm{cl}_n(\theta) = \sum_a N_{a,\cdot} \log p_a(\theta) + \sum_{a,b} N_{a,b} \log p_{a,b}(\theta),$$

with $N_{a,\cdot} = \sum_b N_{a,b}$. This is for 2-window CL. For 3-window CL:

$$\log \mathrm{cl}_{n,3}(\theta) = \sum_a N_{a,\cdot,\cdot} \log p_a(\theta) + \sum_{a,b} N_{a,b,\cdot} \log p_{a,b}(\theta) + \sum_{b,c} N_{\cdot,b,c} \log p_{b,c}(\theta),$$

with 2nd and 3rd term almost the same:

$$\log \mathrm{cl}_k(\theta) = \sum_a N_a \log p_a(\theta) + (k-1) \sum_{a,b} N_{a,b} \log p_{a,b}(\theta).$$

With $k \geq 5$ (say), very little difference between ML and CL.

Large-sample theory: Need limit in probability of 2nd derivative of $n^{-1} \log \mathrm{cl}_k(\theta)$ and limit in distribution of 1st derivative of $n^{-1/2} \log \mathrm{cl}_k(\theta)$.

Need

$$u_{a,b} = \frac{\partial \log p_{a,b}(\theta)}{\partial\theta} \quad \text{and} \quad v_a = \frac{\partial \log p_a(\theta)}{\partial\theta},$$

and matrices

$$H = \sum_a p_a v_a v_a^{\rm t}, \quad G = \sum_{a,b} p_a \bar\gamma_{a,b} v_a v_b^{\rm t}, \quad L = \sum_{a,b} p_a p_{a,b} u_{a,b} \kappa_b^{\rm t},$$

where

$$\kappa_b = \sum_{k\ge 0}\sum_c (p_{b,c}^{(k)} - p_c) v_c \quad \text{and} \quad \bar\gamma_{a,b} = \sum_{k\ge 1}(p_{a,b}^{(k)} - 1).$$

CL theorem:

$$\sqrt{n}(\widehat\theta - \theta) \to_d {\rm N}(0, J_k^{-1} K_k J_k^{-1}),$$

with

$$J_k = (k-1)J + H,$$
$$K_k = (k-1)^2 J + H + G + G^{\rm t} + (k-1)(L + L^{\rm t}).$$

Proof: 'As expected', keeping track of all terms, still within realm of the limits $Z_{a,b}$ of $\sqrt{n}(N_{a,b}/n - p_a p_{a,b})$.

3. PL estimation

2-step and k-step probabilities enter calculations:

$$\log \mathrm{pl}_n(\theta) = 2\sum_{a,b} N_{a,b}\log p_{a,b}(\theta) - \sum_{a,c} N_{a,\cdot,c}\log p^{(2)}_{a,c}(\theta).$$

In addition to $u_{a,b} = \partial \log p_{a,b}/\partial\theta$, need

$$w_{a,c} = \frac{\partial \log p^{(2)}_{a,c}}{\partial\theta} = \sum_b \frac{p_{a,b}p_{b,c}}{p^{(2)}_{a,c}}(u_{a,b}+u_{b,c}).$$

Also, matrices

$$M = \sum_{a,c} p_a p^{(2)}_{a,c} w_{a,c} w^{\mathrm{t}}_{a,c}, \quad Q = \sum_{a,c,d,f} p_a p_{a,d} p_{d,c} p_{c,f} w_{a,c} w^{\mathrm{t}}_{d,f}.$$

PL theorem:

$$\sqrt{n}(\widehat{\theta}-\theta) \to_d \mathrm{N}(0, J_0^{-1}K_0J_0^{-1}),$$

where

$$J_0 = 2J - M \quad\text{and}\quad K_0 = 4J - 3M + Q + Q^{\mathrm{t}}.$$

Proof: Again 'as expected', but more intricate algebra etc.

Lemma:

$$\sqrt{n}(N_{a,b,c}/n - p_a p_{a,b} p_{b,c}) \to_d Z_{a,b,c},$$

where stamina *&* patience give

$$\begin{aligned}\mathrm{cov}(Z_{a,b,c}, Z_{d,e,f}) = {} & p_a p_{a,b} p_{b,c}(\delta_{a,d}\delta_{b,e}\delta_{c,f} - p_d p_{d,e} p_{e,f}) \\ & + p_a p_{a,b} p_{b,c}(\delta_{b,d}\delta_{c,e} - p_d p_{d,e}) p_{e,f} \\ & + p_d p_{d,e} p_{e,f}(\delta_{e,a}\delta_{f,b} - p_a p_{a,b}) p_{b,c} \\ & + p_a p_{a,b} p_{b,c} \gamma_{c,d} p_{d,e} p_{e,f} \\ & \quad + p_d p_{d,e} p_{e,f} \gamma_{f,a} p_{a,b} p_{b,c}\end{aligned}$$

for $a, b, c, d, e, f = 1, \ldots, S$. Result reached via identifying and working with different contributions from the implied double sum.

Essence of rest of proof:

$$\sqrt{n}(\widehat{\theta} - \theta) \doteq_d \Big\{-\frac{1}{n}\frac{\partial^2 \log \mathrm{pl}_n(\theta)}{\partial\theta\partial\theta^{\mathrm{t}}}\Big\}^{-1} \frac{1}{\sqrt{n}} \frac{\partial \log \mathrm{pl}_n(\theta)}{\partial\theta}.$$

4. Illustrations

For **any parametric Markov model** (and anywhere in the parameter space) we may compute matrices

$$\begin{aligned} J &\quad \text{for ML}, \\ J, H, G, L, J_k, K_k &\quad \text{for CL}, \\ J, M, Q, J_0, K_0 &\quad \text{for PL}, \end{aligned}$$

and compare

$$J^{-1} \quad \text{with} \quad J_k^{-1} K_k J_k^{-1} \quad \text{with} \quad J_0^{-1} K_0 J_0^{-1}.$$

Explicit formulae for a short list of nice models; **numerical results** (in the form of ARE curves etc., directly from transition matrix) for any given model.

Hjort and Varin (2007, Tech Report): many illlustrations (more than in SJS paper).

Example 1: Let

$$P = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix}.$$

Here ML = CL, estimator $(N_{0,\cdot} + N_{\cdot,1})/n$, while PL uses

$$\widehat{\theta} = \frac{\sqrt{\rho_n}}{\sqrt{\rho_n} + \sqrt{1-\rho_n}},$$

with $\rho_n = (N_{0,1,0} + N_{1,0,1})/(N_{0,\cdot,0} + N_{1,\cdot,1})$:

$$\sqrt{n}(\widehat{\theta}_{\rm ML} - \theta) \to_d \mathrm{N}(0, \theta(1-\theta)) \quad \text{and} \quad \sqrt{n}(\widehat{\theta}_{\rm PL} - \theta) \to_d \mathrm{N}(0, 1/4).$$

Example 2: Markov (1913) took all 20,000 letters from Pushkin's Yevgeniĭ Onegin, and fitted this model:

	гласный	согласный
гласный	p_1	$1-p_1$
согласный	p_2	$1-p_2$

with $p_1 = .128$ and $p_2 = .663$, giving the correct stationary probabilities .432 for vowels and .458 for consonants.

ML and CL are large-sample equivalent; PL does **rather worse**.

HV 2008: comparisons for **2nd order Markov**, for Pushkin data.

Example 3: An equicorrelation chain:

$$P_{i,j} = \begin{cases} (1-\rho)p_j + \rho & \text{if } i = j, \\ (1-\rho)p_j & \text{if } i \neq j. \end{cases}$$

Then $P^k = (1-\rho^k)p + \rho^k p$, so correlation is ρ^k for time interval k.

For $p = (p_1, \ldots, p_S)^{\mathrm{t}}$ known, ML = CL, and PL loses. For both p and ρ unknown: CL loses a little to ML, PL loses rather more.

Example 4: One-dimensional **Ising model**:

$$\begin{aligned} &\Pr\{X_i = x_{i-1} \,|\, x_{i-1}, x_{i+1}\} \\ &\qquad\qquad\qquad \propto \exp\bigl[\beta(I\{x_{i-1} = x_i\} + I\{x_{i+1} = x_i\})\bigr]. \end{aligned}$$

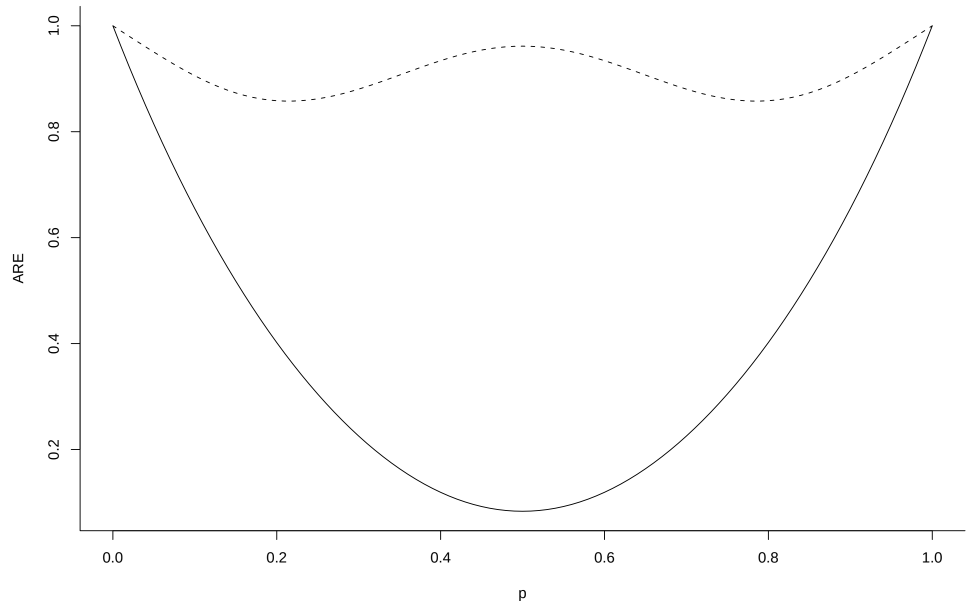
This corresponds to

$$P = \frac{1}{1+\exp(\beta)} \begin{pmatrix} \exp(\beta) & 1 \\ 1 & \exp(\beta) \end{pmatrix}.$$

Here ML = CL again, and

$$\widehat{\beta}_{\mathrm{ML}} = \log \frac{N_{0,0} + N_{1,1}}{N_{0,1} + N_{1,0}} \quad \text{and} \quad \widehat{\beta}_{\mathrm{PL}} = \tfrac{1}{2} \frac{N_{0,0,0} + N_{1,1,1}}{N_{0,1,0} + N_{1,0,1}}.$$

PL suffers serious **efficiency loss** for strong dependence.



Example 5: *The random walk with two reflecting barriers: six states example. The solid line correspond to the ARE for PL, while the dashed one to QL.*

Markov for DNA sequences

	A	G	C	T	total
A	93	13	3	3	112
G	10	105	3	4	122
C	6	4	113	18	141
T	7	4	21	93	125
total	116	126	140	118	500

Summarising evolution of $n = 500$ sites of two homologous DNA sequences. Among various [related] models:

$$\begin{array}{ccccc}
 & A & G & C & T \\
A & 1-2\alpha-\gamma & \gamma & \alpha & \alpha \\
G & \delta & 1-2\alpha-\delta & \alpha & \alpha \\
C & \beta & \beta & 1-2\beta-\gamma & \gamma \\
T & \beta & \beta & \delta & 1-2\beta-\delta
\end{array}$$

One finds equilibrium distribution

$$
\begin{aligned}
p_A &= \frac{\beta}{\alpha+\beta}\frac{\alpha+\delta}{2\alpha+\gamma+\delta}, \qquad & p_G &= \frac{\beta}{\alpha+\beta}\frac{\alpha+\gamma}{2\alpha+\gamma+\delta}, \\
p_C &= \frac{\alpha}{\alpha+\beta}\frac{\beta+\delta}{2\beta+\gamma+\delta}, \qquad & p_T &= \frac{\alpha}{\alpha+\beta}\frac{\beta+\gamma}{2\beta+\gamma+\delta}.
\end{aligned}
$$

Homleid (1995): applied such models to **meteorology**, 'normal weather' split into N1 and N2, 'ugly weather' split into U1 and U2.

Computing all required matrices

* J for the ML;
* H, G, L for the CL;
* M, Q for the PL;

at the 'typical' value $(.027, .041, .122, .126)$, shows once more that **CL is nearly efficient**, while **PL loses a lot**.

This is in agreement with simulation runs (large-sample approximations are effective for small n).

Other parameters: Our results also imply

$$\sqrt{n}(\widehat{\psi}_{\rm ML}-\psi)\to_d {\rm N}(0,\tau_{\rm ML}^2),$$
$$\sqrt{n}(\widehat{\psi}_{\rm CL}-\psi)\to_d {\rm N}(0,\tau_{\rm CL}^2),$$
$$\sqrt{n}(\widehat{\psi}_{\rm PL}-\psi)\to_d {\rm N}(0,\tau_{\rm PL}^2),$$

for any $\psi=\psi(\alpha,\beta,\gamma,\delta)$.

Asynchronous distance between sequences, from Barry and Hartigan (1987): $\Delta=-(1/4)\log|P(\theta)|$. Can work out:

$$\sqrt{n}\{\log|P(\widehat{\theta})|-\log|P(\theta)|\}\to_d {\rm Tr}\{P(\theta)^{-1}V\},$$

in terms of a certain zero-mean normal $V=(V_1,V_2,V_3,V_4)^{\rm t}$ with estimable covariance matrix, for each of ML, CL, PL.

6. When the models are imperfect

Suppose only that there are transition probabilities

$$\pi_{a,b} = \Pr\{X_i = b \,|\, X_{i-1} = a\} \quad \text{for } a, b = 1, \ldots, S.$$

How do estimation methods attempt to get close?

ML:

$$n^{-1} \log l_n(\theta) = \sum_{a,b} \frac{N_{a,b}}{n} \log p_a(\theta) \to_p \sum_{a,b} \pi_a \pi_{a,b} \log p_{a,b}(\theta).$$

Maximising this is equivalent to minimising

$$d_{\mathrm{ML}}(\mathrm{truth}, \mathrm{model}) = \sum_a \pi_a \Big\{ \sum_b \pi_{a,b} \log \frac{\pi_{a,b}}{p_{a,b}(\theta)} \Big\}.$$

This is **weighted Kullback–Leibler**, over each row's model.

Similarly for PL and CL: again, weighted versions of (different) Kullback–Leibler distances.

PL:

$$d_{\mathrm{PL}}(\mathrm{truth}, \mathrm{model}) = \sum_{a,c} \pi_a \pi^{(2)}_{a,c} \Big\{ \sum_b \frac{\pi_{a,b}\pi_{b,c}}{\pi^{(2)}_{a,c}} \log \frac{\pi_{a,b}\pi_{b,c}/\pi^{(2)}_{a,c}}{p_{a,b}(\theta)p_{b,c}(\theta)/p^{(2)}_{a,c}(\theta)} \Big\}.$$

CL:

$$d_{\mathrm{CL}}(\mathrm{truth}, \mathrm{model}) = \sum_a \pi_a \log \frac{\pi_a}{p_a(\theta)} + (k-1) \sum_a \pi_a \Big\{ \sum_b \pi_{a,b} \log \frac{\pi_{a,b}}{p_{a,b}(\theta)} \Big\}.$$

Illustration: Using a four-parameter model when a six-parameter model is true. Assume that a Markov chain on the four states A, G, C, T in reality is governed by

$$\begin{pmatrix} 1-2\alpha-\gamma_1 & \gamma_1 & \alpha & \alpha \\ \delta_1 & 1-2\alpha-\delta_1 & \alpha & \alpha \\ \beta & \beta & 1-2\beta-\gamma_2 & \gamma_2 \\ \beta & \beta & \delta_2 & 1-2\beta-\delta_2 \end{pmatrix},$$

but that the four-parameter model Kimura model, assuming $\gamma_1 = \gamma_2$ and $\delta_1 = \delta_2$, is being used for estimation and inference.

One learns:

ML and CL react very similarly, and in a robust way;
PL reacts very differently, and is too sensitive.

7. CLIC and FCLIC: model selection [and averaging]

For a given parametric model:

$$
\begin{aligned}
A_n(\theta) &= n^{-1} \log \mathrm{cl}_n(\theta) \\
&\to_{\mathrm{pr}} A(\theta) = \sum_a \pi_a \log p_a(\theta) + (k-1) \sum_{a,b} \pi_a \pi_{a,b} \log p_{a,b}(\theta)
\end{aligned}
$$

for each θ, and

$$
d_{\mathrm{CL}}(\mathrm{truth}, \mathrm{model}) = \mathrm{const.} - A(\theta).
$$

How good is the model? Answer: size of $A(\widehat{\theta})$.

Model selection idea: estimate $A(\widehat{\theta})$ (almost unbiasedly), for each candidate model.

Convergence of **basic empirical process**:

$$H_n(s) = \log \mathrm{cl}_n(\theta_0 + s/\sqrt{n}) - \log \mathrm{cl}_n(\theta_0)$$
$$\doteq \sqrt{n} U_n^{\mathrm{t}} s - \tfrac{1}{2} s^{\mathrm{t}} J_n s + o_{\mathrm{pr}}(1) \to_d H(s) = s^{\mathrm{t}} U - \tfrac{1}{2} s^{\mathrm{t}} J_k s.$$

Corollary 1:

$$\mathrm{argmax}(H_n) = \sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{argmax}(H) = J_k^{-1} U \sim \mathrm{N}_p(0, J_k^{-1} K_k J_k).$$

Corollary 2:

$$\max H_n = \log \mathrm{cl}_n(\widehat{\theta}) - \log \mathrm{cl}_n(\theta_0) = n\{A_n(\widehat{\theta}) - A_n(\theta_0)\} \to_d \max H = \tfrac{1}{2} Z,$$

for $Z = U^{\mathrm{t}} J_k^{-1} U$. A bit more analysis:

$$A_n(\widehat{\theta}) - A(\widehat{\theta}) = n^{-1} Z_n + \text{variable with mean zero}$$

where $Z_n \to_d Z$. Model selector:

$$\mathrm{CLIC} = \log \mathrm{cl}_{n,\max} - \widehat{p}^*, \quad \text{with } p^* = \mathrm{E}\, Z = \mathrm{Tr}(J_k^{-1} K_k).$$

Can also construct **Focussed CLIC** [following Cleaskens and Hjort].

Concluding comments

(A) Why is CL better than PL?

$$\log \mathrm{cl}_n(\theta) = \sum_a N_{a,\cdot} \log p_a(\theta) + \sum_{a,b} N_{a,b} \log p_{a,b}(\theta),$$

with 2nd term equal to ordinary $\log \mathrm{l}_n(\theta)$. The 1st term uses [some] forces to make sure that the equilibrium is well assessed.

So **CL = penalised likelihood**, and can also be seen as an empirical Bayes strategy with a prior of the type

$$g(\theta) \propto \exp\Big\{-\rho \sum_a p_a^0 \log \frac{p_a^0}{p_a(\theta)}\Big\}.$$

This is sensible! – But

$$\log \mathrm{pl}_n(\theta) = 2\sum_{a,b} N_{a,b} \log p_{a,b}(\theta) - \sum_{a,c} N_{a,\cdot,c} \log p_{a,c}^{(2)}(\theta),$$

amounting to a ‘strange penalisation’ of the log-likelihood. Translated to Bayes and empirical Bayes: The PL uses a strange prior, intent on conflict with the ML objectives, and the strength of the prior is proportional to n.

(B) Variations and other models:

Can study **many short chains** instead of one long.

2-step memory length (etc.):
Essentially contained in the 1-step theory.

Markov chain **regression** models:

$$P_i = \begin{pmatrix} 1-\alpha_i & \alpha_i \\ \beta_i & 1-\beta_i \end{pmatrix},$$

with

$$\alpha_i = \frac{\exp(r + sz_i)}{1 + \exp(r + sz_i) + \exp(t + uz_i)},$$
$$\beta_i = \frac{\exp(t + uz_i)}{1 + \exp(r + sz_i) + \exp(t + uz_i)}.$$

Hidden Markov chains: Can do CL. Bickel, Ritov, Rydén (1998): show that the ML works, in principle, but impossible to find formulae for limiting variance matrix $J(\theta)^{-1}$. This appears possible with CL, for at least the simpler HMM models.

(C) Using CL for model building:

Forget (or bypass) transition probabilities, model **joint behaviour of block** directly: e.g.

$$f_{a,b,c,d,e} = A(\rho)\exp\big[-\rho\{h_2(a,c) + h_1(b,c) + h_1(d,c) + h_2(e,c)\}\big].$$

Could be parameters inside 1-neighbour function h_1 and 2-neighbour function h_2. Can use CL to estimate parameters – without writing down the two-step Markov model with transition probabilities etc.

NB: Tempting to use **local models** of type

$$f(x_i, x_{i\pm1}, x_{i\pm2}, x_{i\pm3}) \propto \exp\big[-\rho H_\theta(x_i, x_{i\pm1}, x_{i\pm2}, x_{i\pm3})\big]$$

but only a subset of these correspond to genuine **full models**. Characterisation exercise: derive **local characteristics** from local f and check with Hammersley–Clifford–Besag theorems.

Ok or not [cf. comment from Reid]? Depends on purpose. Local inference: meaningful. But only full models give full insight and predictions.

(D) Time series models: May be handled in reasonable generality. For Zi Jin's AR(1) process: may find explicit limit distributions for ML, PL, CL.

(E) More CL in 2D:

Could invent new spatial models for which

$$\mathrm{cl}_n(\theta) = \prod_{i=1}^{n} p_\theta(x_i, x_{\partial i}) = \prod_{i=1}^{n} p_\theta(x_i) p_\theta(x_{\partial i} \,|\, x_i)$$

works well. This is **turning PL inside out**.

Heagerty and Lele (1998): **pairwise** CL method for model

$$Y(s) = I\{Z(s) \geq c\},$$

a Gaußian truncation model. Can get this to work also with say **quintuple-wise** CL, with data plus four neighbours: needs 'only' a separate function that computes 5-dim-normal probabilities for the 32 quintotants in $\mathcal{R}^5$.