Recap

- **Vector observation:** \( Y \sim f(y; \theta), \quad Y \in \mathcal{Y} \subseteq \mathbb{R}^m, \quad \theta \in \mathbb{R}^d \)

- sample of independent vectors \( y_1, \ldots, y_n \)

- pairwise log-likelihood
  \[
  \sum_{i=1}^n \sum_{r=1}^m \sum_{s>r} w_{rs} \log f_2(y_{ir}, y_{is}; \theta)
  \]

- weights are often 1

- more generally,
  \[
  c\ell(\theta; y) = \sum_{i=1}^n \sum_{k \in K} w_k \ell_k(\theta; y_i)
  \]
  \[
  \ell_k(\theta; y_i) = \log f(y_i \in \mathcal{A}_k; \theta)
  \]
Inference

- **Sample:** \( Y_1, \ldots, Y_n, \text{i.i.d.}, \ c\ell(\theta; y) = \sum_{i=1}^{n} c\ell(\theta; y_i) \)

- \( \hat{\theta}_{CL} - \theta \sim N\{0, G^{-1}(\theta)\} \)

- \( w(\theta) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\theta)\} \sim \sum_{a=1}^{d} \mu_a Z_a^2 \quad Z_a \sim N(0, 1) \)
  - \( \mu_1, \ldots, \mu_d \) eigenvalues of \( J(\theta)H(\theta)^{-1} = H(\theta)G(\theta)^{-1} \)

- **Nuisance parameters** \( \theta = (\psi, \lambda) \)

- \( \sqrt{n}(\hat{\psi}_{CL} - \psi) \sim N\{0, G^{\psi\psi}(\theta)\} \)

- \( w(\psi) = 2\{c\ell(\hat{\theta}_{CL}) - c\ell(\tilde{\theta}_\psi)\} \sim \sum_{a=1}^{d_0} \mu_a Z_a^2 \)
  - \( \mu_1, \ldots, \mu_{d_0} \) eigenvalues of \((H^{\psi\psi})^{-1} G^{\psi\psi}\)
Example: dichotomized MV Normal

\( Y_{ir} = 1 \{ Z_{ir} > 0 \} \quad Z \sim N(0, R) \quad r = 1, \ldots, m; i = 1, \ldots, \ell \)

\[
\ell_2(\rho) = \sum_{i=1}^{n} \sum_{s<r} \{y_{ir} y_{is} \log P(y_r = 1, y_s = 1) + y_{ir} (1 - y_{is}) \log P_{10} \}
+ (1 - y_{ir}) y_{is} \log P_{01} + (1 - y_{ir}) (1 - y_{is}) \log P_{00} \}
\]

\[
a.\text{var}(\hat{\rho}_{CL}) = \frac{1}{n} \frac{4\pi^2}{m^2} \frac{(1 - \rho^2)}{(m - 1)^2} \text{var}(T) \quad T = \sum_{i} \sum_{s<r} (2y_{ir} y_{is} - y_{ir} - y_{is}) \]

\[
\text{var}(T) = nm^4 (p_{1111} - 2p_{111} + 2p_{11} - p_{11}^2 + \frac{1}{4}) + m^3 (-6p_{1111}) + m^2 (...) + m (...)\]
Numbers changed from last week. Also incorrect in Cox & Reid 2004 Table 1
Example: longitudinal count data

- subjects \( i = 1, \ldots, n \)
- observations counts \( y_{ir}, r = 1, \ldots m_i \)
- model \( y_{ir} \sim \text{Poisson}(u_{ir} x_{ir}^T \beta) \)
- \( u_{i1}, \ldots, u_{im_i} \) gamma-distributed random effects
- but correlated \( \text{corr}(u_{ir}, u_{is}) = \rho^{|r-s|} \)
- joint density has combinatorial number of terms in \( m_i \); impractical
- weighted pairwise composite likelihood

\[
\mathcal{CL}_{\text{pair}}(\beta) = \prod_{i=1}^{n} \frac{1}{m_i - 1} \prod_{r=1}^{m_i} \prod_{s=r+1}^{m_i} f(y_{ir}, y_{is}; \beta)
\]

- weights chosen so that \( \mathcal{L}_{\text{pair}} = \text{full likelihood if } \rho = 0 \)

Henderson & Shimura, 2003
Example: Varin & Czado 2010

- pain severity scores recorded at four time points: morning, noon, evening, bed
- 119 patients; varying number of days per patient
- covariates: personal and weather
- response: pain score 0 1 2 3 4 5

- $y_{ij}$ response at time $t_{ij}$ for observation $j$ on subject $i$
- $y^*_{ij}$ a latent variable, continuous: $y^*_{ij} = x_{ij}^T \beta + u_i + \epsilon_{ij}$
- $y_{ij} = k \Leftrightarrow a_{k-1} < y^*_{ij} < a_k$
- if $u_i \sim N(0, \sigma^2)$ and $\epsilon_{ij} \sim N(0, 1)$

$$f(y_{i1}, \ldots, y_{im_i}) =$$
$$\int_{-\infty}^{\infty} \prod_{j=1}^{n_i} \left\{ \Phi(a_{y_{ij}} - x_{ij}^T \beta - u_i) - \Phi(a_{y_{ij}-1} - x_{ij}^T \beta - u_i) \right\} \phi\left(\frac{u_i}{\sigma}\right) du_i$$

- $L(\theta; y) = \prod_{i=1}^{n}$
... pain severity scores

- $y_{ij}^*$ and $y_{ij'}^*$ have constant correlation $\sigma^2 / (\sigma^2 + 1)$
- points nearer in time might be expected to have higher correlation
- change $\epsilon_{ij}$ i.i.d. $N(0, 1)$ to $\text{corr}(\epsilon_{ij}, \epsilon_{ij'}) = \exp(\delta |t_{ij} - t_{ij'}|)$
- now

$$L(\theta; y) = \prod_{i=1}^{n} \int_{\tilde{a}_{y_{i1}}^{-1}}^{\tilde{a}_{y_{im_i}}^{-1}} \phi_{n_i}(z_{i1}, \ldots, z_{imi}; R_i) dz_{i1} \ldots dz_{imi}$$

- pairwise log-likelihood:

$$c\ell(\theta; y) = \sum_{i=1}^{n} \sum_{j<j'} \log f_2(y_{ij}, y_{ij'}; \theta) \mathbb{I}_{[-q,q]}(t_{ij} - t_{ij'})$$

weights are 1 or 0, depending on distance between time points
... pain severity scores

Table 4 Varin & Czado
Supplementary material online
Applications: spatial and space-time data

- conditional approaches seem more natural
- condition on neighbours in space
- condition on small number of lags (in time)
- some form of blockwise components often proposed
  Stein et al, 04; Caragea and Smith, 07

- fMRI time series
  Kang et al, 12

- air pollution and health effects
  Bai et al, 12

- computer experiments: Gaussian process models
  Xi, 12

- spatially correlated extremes
  - joint tail probability known
  - joint density requires combinatorial effort (partial derivatives)
  - composite likelihood based on joint distribution of pairs, triples seems to work well
  Davison et al, (2012); Genton et al., 12, Ribatet, 12
Spatial extremes

- vector observations \((X_{1i}, \ldots, X_{mi}), \ i = 1, \ldots, n\)
- example rainfall at each of \(m\) locations
- component-wise maxima \(Z_1, \ldots, Z_m; Z_j = \max(X_{j1}, \ldots, X_{jn})\)
- \(Z_j\) are transformed (centered and scaled)
- general theory says

\[
\Pr(Z_1 \leq z_1, \ldots, Z_m \leq z_m) = \exp\{-V(z_1, \ldots, z_m)\}
\]

- function \(V(\cdot)\) can be parameterized via Gaussian process models
- example

\[
V(z_1, z_2) = z_1^{-1} \Phi\{(1/2)a(h) + a^{-1}(h) \log(z_2/z_1)\} + z_2^{-1} \Phi\{(1/2)a(h) + a^{-1}(h) \log(z_1/z_2)\}
\]

\[
Z(h) = (z_1, z_2), \ Z(0) = (0, 0), \ a(h) = h^T \Omega^{-1} h
\]
... spatial extremes

- \( \Pr(Z_1 \leq z_1, \ldots, Z_d \leq z_m) = \exp\{-V(z_1, \ldots, z_m)\} \)

- to compute log-likelihood function, need the density
- combinatorial explosion in computing joint derivatives of \( V(\cdot) \)
- Davison et al. (2012, Statistical Science) used pairwise composite likelihood
- compared the fits of several competing models, using AIC analogue described above
- applied to annual maximum rainfall at several stations near Zurich
FIG. 1. Map of Switzerland showing the stations of the 51 rainfall gauges used for the analysis, with an insert showing the altitude. The 36 stations marked by circles were used to fit the models, and those marked with squares were used to validate the models. Data for the pairs of stations with blue symbols appear in Figure 2.
FIG. 3. Maps of the (predicted) pointwise 25-year return level estimates for rainfall (mm) obtained from the latent variable and max-stable models. The top and bottom rows show the lower and upper bounds of the 95% pointwise credible/confidence intervals. The middle row shows the predictive pointwise posterior mean and pointwise estimates. The left column corresponds to the latent variable model assuming $\text{Gamma}(5, 3)$ prior on $\lambda$. The middle column assumes the less informative priors $\lambda_{\eta} \sim \text{Gamma}(1, 100)$, $\lambda_{\tau} \sim \text{Gamma}(1, 10)$ and $\lambda_{\zeta} \sim \text{Gamma}(1, 10)$. The right column corresponds to the extremal $t$ copula model.
... applications

- time series – a case of large $m$, fixed $n$
  - need new arguments re consistency, asymptotic normality
  - consecutive pairs: consistent, not always asy. normal
  - $AR(1)$: consecutive pairs fully efficient; all pairs terrible (consistent, highly variable)
  - $MA(1)$: consecutive pairs consistent but very inefficient

Davis and Yau (2011)

- genetics: estimation of recombination rate
  - somewhat similar to time series
  - but correlation may not decrease with increasing length
  - suggesting all possible pairs may be inconsistent
  - joint blocks of short sequences seems preferable

- linkage disequilibrium
- family based sampling

Larribe and Fearnhead (2011); Choi and Briollais, 12
Example: Ising model

Ising model:

\[
f(y; \theta) = \exp\left( \sum_{(j,k) \in E} \theta_{jk} y_j y_k \right) \frac{1}{Z(\theta)} \quad j, k = 1, \ldots, K
\]

neighbourhood contributions

\[
f(y_j \mid y_{(-j)}; \theta) = \frac{\exp(2y_j \sum_{k \neq j} \theta_{jk} y_k)}{\exp(2y_j \sum_{k \neq j} \theta_{jk} y_k) + 1} = \exp \ell_j(\theta; y)
\]

penalized CL estimation based on sample \(y^{(1)}, \ldots, y^{(n)}\)

\[
\max_{\theta} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{K} \ell_j(\theta; y^{(i)}) - \sum_{j<k} P_\lambda(|\theta_{jk}|) \right\}
\]

Xue et al., 2012
Ravikumar et al., 2010
Misspecified models: general theory

- $y_1, \ldots, y_n$ independent observations with distribution function $G(\cdot)$ and density $g(\cdot)$

- we fit the **incorrect** model $f(y; \theta)$

- Kullback-Liebler (KL) divergence between $f(y; \theta)$ and $g(y)$ is defined as

\[
KL(\theta) = \int \log \left\{ \frac{g(y)}{f(y; \theta)} \right\} g(y)dy
\]

Wikipedia writes $D_{KL}(G\|F)$ or more precisely $D_{KL}(P_G\|P_F)$

- $KL(\theta) \geq 0$, and $KL(\theta) = 0 \iff f(y; \theta) \equiv g(y)$

- define $\theta^* = \arg \min KL(\theta)$

- $f(y; \theta^*)$ is closest to $G(\cdot)$ in the family $\{f(\cdot; \theta), \theta \in \Theta\}$
... misspecified models

\[ KL(\theta) = \int \log \left\{ \frac{g(y)}{f(y; \theta)} \right\} g(y) dy \]

- \( \theta^* = \arg \min_\theta KL(\theta) \)
- \( \theta^* = \arg \max_\theta \int \log \{ f(y; \theta) \} g(y) dy = \arg \max_\theta E_G \{ \ell(\theta; y) \} \)

leads to a proof that the maximum likelihood estimator converges to \( \theta^* \)

- if \( g(y) = f(y; \theta_0) \), then \( \theta^* = \theta_0 \)
- otherwise \( \theta^* \) is the ‘least false’ parameter value
... misspecified models

- **Example**
  - true model $G$ is log-normal $\log y \sim N(\mu, \sigma^2)$
  - $g(y) = ?$
  - Model $f(y; \theta) = \frac{1}{\theta} \exp(-\frac{y}{\theta})$
  - $E_G\{\ell(\theta; y)\} = -\log \theta + E_G(\frac{y}{\theta})$
  - $\theta^* = E_G(y) = \exp(\mu + \sigma^2/2)$
  - reminder: MLE solves $\partial_\theta \ell(\theta; y) = 0$
  - $y = (y_1, \ldots, y_n)$
  - in the example, $\hat{\theta} = \bar{y}$
  - $\xrightarrow{p} \theta^*$
... misspecified models

- \( \partial_\theta \ell(\hat{\theta}; y) = 0 \)

- inference?

- \( E_G \partial_\theta \ell(\theta; y) = \int \partial_\theta \ell(\theta; y) g(y) dy \equiv 0 \), at \( \theta^* \)

- \( E_G \partial_{\theta\theta} \ell(\theta; y) = \int \partial_{\theta\theta} \ell(\theta; y) g(y) dy \equiv H(\theta) \)

- \( E_G \partial_\theta \ell(\theta; y)^2 = \int \partial_\theta \ell(\theta; y)^2 g(y) dy \equiv J(\theta) \)

- \( (\hat{\theta} - \theta^*) = H(\theta^*)^{-1} U(\theta^*) \{1 + o_p(1)\} \)

- \( \hat{\theta} \sim N\{\theta^*, G^{-1}(\theta^*)\} \)

- \( G(\theta) = H(\theta)J^{-1}(\theta)H(\theta) \)
Examples

- \( y_1, \ldots, y_n \) i.i.d. \( \sim G \); we assume \( f(y; \theta) \) is \( N(\mu, \sigma^2) \)

- \( \hat{\mu} = \bar{y}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \bar{y})^2 \)

- \( \partial_\theta \ell(\theta; y) = \left[ \Sigma(y_i - \mu)/\sigma^2, \quad -(n/2\sigma^2) + \Sigma(y_i - \mu)^2/(2\sigma^4) \right] \)

- \( H(\theta) = n \begin{bmatrix} 1/(\sigma_0^2) & 0 \\ 0 & 1/(2\sigma_0^4) \end{bmatrix} \)

- \( J(\theta) = n \begin{bmatrix} 1/(\sigma_0^2) & \mu_3/(2\sigma_0^6) \\ \mu_3/(2\sigma_0^6) & (\mu_4 - \sigma_0^4)/(4\sigma_0^8) \end{bmatrix} \)

- \( \sigma_0^2 = \text{var}_G(y_1), \mu_3 = \text{E}_G(y_1 - \mu_0)^3, \text{etc.} \)

- \( H(\theta) = -\text{E}_G \ell''(\theta; y); \quad J(\theta) = \text{Var}_G \ell'(\theta; y) \)
... examples

- $y_1, \ldots, y_n$ i.i.d. $\sim G$; we assume $f(y; \theta)$ is $N(\mu, \sigma^2)$

- 
  $$\text{a.var}(\hat{\theta}) = \begin{bmatrix} \sigma_0^2/n & \gamma_1 \sigma_0^3/n \\ \gamma_1 \sigma_0^3/n & \gamma_2 \sigma_0^4/n \end{bmatrix}$$

- $\gamma_3 = E(y_1 - \mu_0)^3/\sigma_0^3$, $\gamma_4 = E(y_1 - \mu_0)^4/(\sigma_0^4) - 1$

- note that $\hat{\mu}$, $\hat{\sigma}^2$ not uncorrelated unless $\gamma_3 = 0$

- $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \{\bar{y}, \Sigma(y_i - \bar{y})^2/n\}$
... examples

- **linear regression model** $G$: $y = X\beta_0 + \epsilon$, $\epsilon \sim (0, \sigma^2 R)$
  linear regression; correlated errors

- **working model** $f(y; \beta) = N(X\beta, \sigma^2 I)$
  uncorrelated errors

- $\hat{\beta} = (X^TX)^{-1}X^Ty$
  least squares estimator; mle under normality

- $E_G(\hat{\beta}) = \beta_0$
  LSE unbiased (consistent)

- $\text{var}_G(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$
  sandwich variance

- working model variance is incorrect

  \[
  \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
  \]
... examples

- linear regression model $G$: $y = X\beta_0 + \epsilon$, $\epsilon \sim (0, \sigma^2 R)$
  linear regression; correlated errors

- $\text{var}_G(\hat{\beta}) = \sigma^2 (X^T X)^{-1} X^T RX (X^T X)^{-1}$
  sandwich variance

- working model variance is incorrect

- single intercept model, constant correlation $\rho$: $\hat{\beta} = \bar{y}$

- true standard error of $\bar{y}$ is $\sigma \{1 + \rho(n - 1)\}^{1/2}/\sqrt{n}$
  compare to $\sigma/\sqrt{n}$

- $\rho = 0.1$, $n = 100$, ratio is 3.3
  variance inflation