Marginal likelihood for parallel series

Peter McCullagh

Department of Statistics
University of Chicago

Wermuth conference
Thorskogs Slott, July/August 2008

www.stat.uchicago.edu/~pmcc/reports/profile.pdf
www.stat.uchicago.edu/~pmcc/reports/distance.pdf

Outline

A simple model for parallel series

Observation consists of k series in parallel
- $Y = (Y^{(1)}, \ldots, Y^{(k)})$ matrix of order $n \times k$
- Each series short; k fairly large
- Each series has same or similar distribution (AR1, AR2,...)
- (common autocorrelation coefficient β)
- Series might not be independent

Model: $Y \sim N(0, \Gamma \otimes \Sigma)$; $\quad \mathrm{cov}(Y_{ir}, Y_{js}) = \Gamma_{ij}\Sigma_{rs}$
- $\mathrm{cov}(Y_{ir}, Y_{jr}) = \Gamma_{ij}\Sigma_{rr}$
- Γ is the parameter of interest (for this talk)

Extended version $Y \sim N(X\theta, \Gamma \otimes \Sigma)$ with $\theta \in \mathcal{R}^{pk}$

Some examples

Example I: Longitudinal study

one series per subject

$i \mapsto x_i$ is observation time (or location in space)

AR1: $\Gamma_{ij} = e^{-\beta|x_i - x_j|}$

Example II: Phylogenetic relationships

r is locus/site/gene on genome

i represents individual or species (unordered)

Y_{ir} is measurement (phenotype) for individual i at site r

$\Gamma \in RT_n$ is the ancestral tree: $\Gamma_{ij} \geq \min(\Gamma_{ik}, \Gamma_{jk})$

How to pool information about Γ from k short series

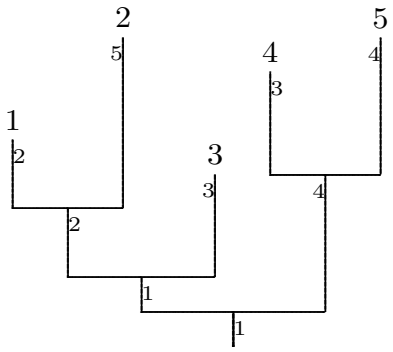
Digression on rooted trees

A non-negative symmetric matrix $\Gamma_{ij} \geq \min(\Gamma_{ik}, \Gamma_{jk})$

Interpretation: Γ_{ij} is the distance from the root to the junction at which i and j occur on separate branches.
Γ_{ii} is distance from root to leaf i

Tree inequality implies graphical tree representation

$$\Gamma = \begin{pmatrix} 6 & 4 & 2 & 1 & 1 \\ 4 & 9 & 2 & 1 & 1 \\ 2 & 2 & 5 & 1 & 1 \\ 1 & 1 & 1 & 8 & 5 \\ 1 & 1 & 1 & 5 & 9 \end{pmatrix} =$$



Models for Σ

Model I: $\Sigma = \sigma^2 I_k$

series iid; group $\mathcal{G} = \mathcal{R}^+$ (positive scalar multiples)

Model II: $\Sigma = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_k^2\}$

series independent and similarly distributed;

Group $\mathcal{G} = (\mathcal{R}^+)^k$ positive $k \times k$ diagonal matrices

Model III: First-order Markov: $\Sigma_{rs} = a_r b_s$ for $r \leq s$

Σ^{-1} is tri-diagonal; no group

Model IV: Σ arbitrary in PD_k

Series similarly distributed but not independent

Group $\mathcal{G} = GL(\mathcal{R}^k)$ ($k \times k$ invertible matrices)

Group action: $Y \mapsto Yg$ multiplication on right

Profile log likelihood

$$
\begin{aligned}
l(\Gamma, \Sigma; Y) &= -\tfrac{1}{2} \log \det(\Gamma \otimes \Sigma) - \tfrac{1}{2} \operatorname{tr}(Y'\Gamma^{-1} Y \Sigma^{-1}) \\
&= -\frac{k}{2} \log |\Gamma| - \frac{n}{2} \log |\Sigma| - \tfrac{1}{2} \operatorname{tr}(Y'\Gamma^{-1} Y \Sigma^{-1}),
\end{aligned}
$$

MLE:

$$
\hat{\Sigma}_\Gamma = \begin{cases} Y'\Gamma^{-1} Y & \text{(Model IV)} \\ \operatorname{diag}(Y'\Gamma^{-1} Y) & \text{(Model II)} \\ \operatorname{tr}(Y'\Gamma^{-1} Y) I_k / k & \text{(Model I)} \end{cases}
$$

Profile log likelihood

$$
l_p(\Gamma; Y) = \begin{cases} -\frac{k}{2} \log |\Gamma| - \frac{nk}{2} \log \operatorname{tr}(Y'\Gamma^{-1} Y) & \text{(I)} \\ -\frac{k}{2} \log |\Gamma| - \frac{n}{2} \log |\operatorname{diag}(Y'\Gamma^{-1} Y)| & \text{(II)} \\ -\frac{k}{2} \log |\Gamma| - \frac{\max(k,n)}{2} \log |Y'\Gamma^{-1} Y| & \text{(IV)} \end{cases}
$$

Bartlett identities

$$2\frac{\partial l_p}{\partial \beta} = \begin{cases} -k\,\mathrm{tr}(WD) + nk\,\mathrm{tr}(Y'AY)/\,\mathrm{tr}(Y'WY) & \mathrm{I} \\ -k\,\mathrm{tr}(WD) + n\sum_{r=1}^{k}(Y_r'AY_r)/(Y_r'WY_r) & \mathrm{II} \\ -k\,\mathrm{tr}(WD) + n\,\mathrm{tr}((Y'WY)^{-1}Y'AY) & \mathrm{IV} \end{cases}$$

where $D = \partial\Gamma/\partial\beta$, $W = \Gamma^{-1}$ and $A = WDW$.

$$E\left(\frac{\mathrm{tr}(Y'AY)}{\mathrm{tr}(Y'WY)}\right) = \frac{k\,\mathrm{tr}(A\Gamma)}{nk} = \frac{\mathrm{tr}(WD)}{n}.$$

implies $E(\partial l_p/\partial\beta) = 0$ for I & II.

$$\mathrm{var}\left(\frac{\partial l_p}{\partial \beta}\right) = -E\left(\frac{\partial^2 l_p}{\partial \beta^2}\right) = \begin{cases} V\frac{k^2}{2(nk+2)} & (I) \\ V\frac{k}{2(n+2)} & (II) \\ V\frac{k(n-k)}{2(n-1)(n+2)} & (IV) \end{cases}$$

Marginal log likelihood

Models I, II and IV are each associated with a group

$$Y \sim N(0, \Gamma \otimes \Sigma) \Rightarrow Yg \sim N(0, \Gamma \otimes g\Sigma g')$$

Maximal invariants: $\check{Y} = Y\tilde{\Sigma}^{-1/2}$

Model I: Y/s, II: $(Y^{(1)}/s_1, \ldots, Y^{(k)}/s_k)$, IV: $\tilde{\Sigma} = Y'Y/n$

Log likelihood based on maximal invariant is

$$l(\Gamma; \check{Y}) = \begin{cases} \frac{k}{2}\log \mathrm{Det}(W) - \frac{nk}{2}\log \mathrm{tr}(Y'WY) & \text{(I)} \\ \frac{k}{2}\log \mathrm{Det}(W) - \frac{n}{2}\log |\mathrm{diag}(Y'WY)| & \text{(II)} \\ \frac{k}{2}\log \mathrm{Det}(W) - \frac{\max(k,n)}{2}\log \mathrm{Det}(Y'WY) & \text{(IV)} \end{cases}$$

where $W = \Gamma^{-1}$.

Profile likelihood = marginal likelihood in each case

Marginal likelihood and regression models

$Y \sim N(X\theta, \Gamma \otimes \Sigma)$ with (θ, Σ) as nuisance parameters

Let $\mathcal{X} \subset \mathcal{R}^n$ be the image (column space) of X of dimension p
Extended group: $\mathcal{X}^{\oplus k} \times \mathcal{G}$

Marginal log likelihood $\neq$ profile log likelihood $n \mapsto (n-p)$

$$l(\Gamma; \check{Y}) = \begin{cases} \frac{k}{2} \log \mathrm{Det}(WQ) - \frac{(n-p)k}{2} \log \mathrm{tr}(Y'WQY) & \text{(I)} \\ \frac{k}{2} \log \mathrm{Det}(WQ) - \frac{n-p}{2} \log |\mathrm{diag}(Y'WQY)| & \text{(II)} \\ \frac{k}{2} \log \mathrm{Det}(WQ) - \frac{\max(k,n-p)}{2} \log \mathrm{Det}(Y'WQY) & \text{(IV)} \end{cases}$$

$W = \Gamma^{-1}$, $Q = I_n - X(X'WX)^{-1}X'W$
$\mathrm{Det}()$ is product of non-zero eigenvalues

(I and II in agreement with Bellhouse (1978); Cruddas, Reid and Cox (1989); Tunnicliffe-Wilson (1989)

Rate of increase of Fisher information

Facts:

- Γ is an AR1 matrix $\Gamma_{ij} = \exp(-\beta|i-j|)$
- Marginal distribution depends only on scalar β
- Two limits: (i) $n \to \infty$, (ii) $k \to \infty$
- How fast does information accumulate in marginal likelihood?

In general if $f_k(\cdot)$ is the density of the maximal invariant
$f_k(y;\beta) = f_{k-1}(y^{(1)},\ldots,y^{(k-1)};\beta) f_k(y^{(k)} \mid y^{(1)},\ldots,y^{(k-1)};\beta)$

$$\mathrm{FI}_k = \mathrm{FI}_{k-1} + \mathrm{var}\left(\frac{\partial \log f_k(y_k \mid y_1,\ldots,y_{k-1};\beta)}{\partial\beta}\right) \geq \mathrm{FI}_{k-1},$$

Implies that FI is increasing in both n and k.

Paradox of decreasing Fisher information

Direct calculation of Fisher information gves

$$\mathrm{var}\left(\frac{\partial l}{\partial \beta}\right) = -E\left(\frac{\partial^2 l}{\partial \beta^2}\right) = \begin{cases} V\frac{k^2}{2(nk+2)} & (I) \\ V\frac{k}{2(n+2)} & (II) \\ V\frac{k(n-k)}{2(n-1)(n+2)} & (IV)\ k \le n \end{cases}$$

where $V = n\,\mathrm{tr}(WDWD) - \mathrm{tr}^2(WD)$ and $D = \partial\Gamma/\partial\beta$.
(for regression effects $W \mapsto WQ$ and $n \mapsto n - p$).

For I and II: FI in increasing in both k and n: $V = O(n^2)$
For III: FI is *decreasing* in k for $k > n/2$

How can we have a one-parameter model with decreasing FI?
Explanation of anomaly

Obfuscations of FI anomaly

Obf.1: Error in calculation of FI

Ans.1: Always a good guess, but no error: verify by simulation

Obf.2: Not surprising: original model has too many parameters $\dim(\Theta) = k(k+1)/2 + 1$ with nk observations

Ans.2: True but doesn't explain the anomaly

Obf.3: Same effect occurs in regression models with $n, p \to \infty$ FI for σ in residuals can decrease with n if $p \uparrow$. Duh!

Ans.3: If p increases with n, you're telling me that the distribution of $Y_1, \ldots, Y_7$ with $n = 7$ is not the same as the distribution of the first 7 components in the sequence $Y_1, \ldots, Y_8$. This is not a process!

In our problem, the distribution of $Y^{(1)}, \ldots, Y^{(k)}$ is the same as the distribution of the first k series in $Y^{(1)}, \ldots, Y^{(k+1)}$. We don't change the law retroactively as further series accumulate.

Explanation of the FI anomaly

Q1: Why does the anomaly occur for (IV) but not for I or II?
Q2: Why does the Fisher information theorem fail for (IV)?
Q3: Does the FI theorem apply to marginal and conditional distributions?

FI theorem: *If the distributions $f_k()$ (of MI $y^{(k)}$) are such that*
$f_k(y;\beta) = f_k(y \mid y^{(k-1)};\beta)\, f_{k-1}(y^{(k-1)};\beta)$
then FI is non-decreasing in k.

Corollary: *If the FI is decreasing in k, the factorization fails.*

Crucial question: Is the MI $y^{(k-1)}$ a function of $y^{(k)}$?
Ans: Yes for I and II: no for IV.
Hence factorization fails for IV.

Digression on model comparison using REML

$H_0 : Y \sim N(X\beta, \sigma^2 I_n)$
$H_1 : Y \sim N(X\beta + Z\gamma, \sigma^2 I_n)$
Group: translation by $x \in \mathcal{X}$ plus scalar multiplication

$$g = (\lambda, x): \quad g\colon Y \mapsto x + \lambda Y$$

Distn of maximal invariant depends only on γ (H_0 simple)
Marginal log likelihood is

$$l(\gamma; y) = -\frac{n-p}{2} \log((y - Z\gamma)' Q (y - Z\gamma))$$

$$2l(\hat{\gamma}; y) - 2l(\gamma; y) = (n-p) \log\left(1 + \frac{q\, F_{q,\, n-p-q}}{n-p-q}\right),$$

a function of the standard F-ratio

Note: only one group, which does not include Z-translation

Intermediate models: Model III

Suppose that Σ is Markov of order one.

$$
\begin{aligned}
Y = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(k)}) &\sim N(X\theta, \Gamma \otimes \Sigma) \\
Y^{(1)} &\sim N(X\theta^{(1)}, \Gamma\sigma_{11}) \\
Y^{(r)} \mid Y^{(1)}, \ldots, Y^{(r-1)} &\sim N(X\theta^{(r)} + Y^{(r-1)}\gamma, \Gamma\tau_r^2)
\end{aligned}
$$

coefficient matrix θ of order $p \times k$ unrestricted

Let $Q_1 : \mathcal{R}^n \to \mathcal{R}^n$ be any projection with kernel $\mathcal{X}$
Let $Q_r : \mathcal{R}^n \to \mathcal{R}^n$ be any projection with kernel $\mathcal{X} \oplus Y^{(r-1)}$

$$
\begin{aligned}
Q_1 Y^{(1)} &\sim N(0, Q_1 \Gamma Q_1' \sigma_{11}) \\
Q_r Y^{(r)} \mid Y^{(1)}, \ldots, Y^{(r-1)} &\sim N(0, Q_r \Gamma Q_r' \sigma_{rr.r-1})
\end{aligned}
$$

Markov model contd.

Residual series

$$Q_1 Y^{(1)} \sim N(0, Q_1 \Gamma Q_1' \sigma_{11})$$
$$Q_r Y^{(r)} \mid Y^{(1)}, \ldots, Y^{(r-1)} \sim N(0,\ Q_r \Gamma Q_r' \sigma_{rr.r-1})$$

have structure similar to model II: (Σ diagonal)

The joint density is a product of Gaussian factors
But the $Q_r Y^{(r)}$ are not independent
The 'marginal' log likelihood is

$$\tfrac{1}{2} \sum_r \log \mathrm{Det}(WQ_r) - \tfrac{n-p-1}{2} \log(Y^{(r)\prime} WQ_r Y^{(r)})$$

FI increases approx linearly in k

Conclusions

In the multivariate model $Y \sim N_{nk}(X\theta, \Gamma \otimes \Sigma)$
with (θ, Σ) as unrestricted nuisance parameters
the marginal likelihood based on the maximal invariant

$$\check{l}(\Gamma; Y) = \frac{k}{2} \log \mathrm{Det}(WQ) - \frac{\max(k, n-p)}{2} \log \mathrm{Det}(Y'WQY)$$

is effective for the elimination of nuisance parameters provided that k is small relative to $n - p$.

If $k > (n-p)/2$ it is more efficient to use a sub-model even though the theory may not be so clean.