Statistical Modelling

Anthony Davison and Jon Forster

http://stat.epfl.ch, http://www.s3ri.soton.ac.uk

1. Model Selection

Overview

1. Basic ideas
2. Linear model
3. Sparse variable selection
4. Bayesian inference
5. Bayesian variable selection

Basic Ideas

Why model?

George E. P. Box (1920–):

> All models are wrong, but some models are useful.

- ☐ Some reasons we construct models:
  - to simplify reality (efficient representation);
  - to gain understanding;
  - to compare scientific, economic, ... theories;
  - to predict future events/data;
  - to control a process.
- ☐ We (statisticians!) rarely believe in our models, but regard them as temporary constructs subject to improvement.
- ☐ Often we have several and must decide which is preferable, if any.

Criteria for model selection

- ☐ Substantive knowledge, from prior studies, theoretical arguments, dimensional or other general considerations (often qualitative)
- ☐ Generalisability of conclusions and/or predictions
- ☐ Sensitivity to failure of assumptions (prefer models that are robustly valid)
- ☐ Quality of fit—residuals, graphical assessment (informal), or goodness-of-fit tests (formal)
- ☐ Prior knowledge in Bayesian sense (quantitative)

Motivation

Even after applying these criteria (but also before!) we may compare many models:

- ☐ linear regression with p covariates, there are 2^p possible combinations of covariates (each in/out), before allowing for transformations, etc.— if $p = 20$ then we have a problem;
- ☐ choice of bandwidth $h > 0$ in smoothing problems
- ☐ the number of different clusterings of n individuals is a Bell number (starting from $n = 1$): 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, ...
- ☐ we may want to assess which among 5×10^5 SNPs on the genome may influence reaction to a new drug;
- ☐ ...

For reasons of economy we seek 'simple' models.

Albert Einstein (1879–1955)

'Everything should be made as simple as possible, **but no simpler**.'

William of Occam (?1285–1347/9)

Occam's razor: **Pluralitas non est ponenda sine necessitate**: **entities should not be multiplied beyond necessity**.

Setting

- ☐ To focus and simplify discussion we will consider parametric models, but the same ideas generalise to semi-parametric and non-parametric settings
- ☐ We will take generalised linear models (GLMs) as example of moderately complex parametric models:
  - Normal linear model has three key aspects:
    - ▷ *structure for covariates*: *linear predictor* $\eta = x^{\mathrm{T}}\beta$;
    - ▷ *response distribution*: $y \sim N(\mu, \sigma^2)$; and
    - ▷ relation $\eta = \mu$ between $\mu = \mathrm{E}(y)$ and η.
  - GLM extends last two to
    - ▷ y has density

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y; \phi)\right\},$$

      where θ depends on η; *dispersion parameter* ϕ is often known; and
    - ▷ $\eta = g(\mu)$, where g is monotone *link function*.

Logistic regression

- Commonest choice of link function for binary reponses:

$$\Pr(Y=1)=\pi=\frac{\exp(x^{\mathrm{T}}\beta)}{1+\exp(x^{\mathrm{T}}\beta)},\quad \Pr(Y=0)=\frac{1}{1+\exp(x^{\mathrm{T}}\beta)},$$

  giving linear model for log odds of 'success',

$$\log\left\{\frac{\Pr(Y=1)}{\Pr(Y=0)}\right\}=\log\left(\frac{\pi}{1-\pi}\right)=x^{\mathrm{T}}\beta.$$

- Log likelihood for β based on independent responses $y_1,\ldots,y_n$ with covariate vectors $x_1,\ldots,x_n$ is

$$\ell(\beta)=\sum_{j=1}^{n}y_j x_j^{\mathrm{T}}\beta-\sum_{j=1}^{n}\log\left\{1+\exp(x_j^{\mathrm{T}}\beta)\right\}$$

- Good fit gives small deviance $D=2\left\{\ell(\tilde{\beta})-\ell(\widehat{\beta})\right\}$, where $\widehat{\beta}$ is model fit MLE and $\tilde{\beta}$ is unrestricted MLE.

Nodal involvement data

Table 1: Data on nodal involvement: 53 patients with prostate cancer have nodal involvement (r), with five binary covariates age etc.

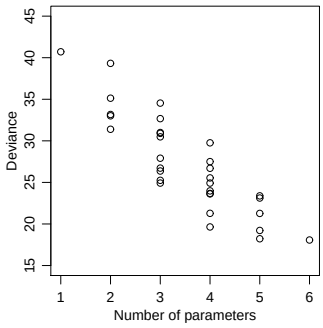
m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	
1	1	0	0	1	0	1
1	0	0	0	0	1	1
1	0	0	0	0	1	0

Nodal involvement deviances

Deviances D for 32 logistic regression models for nodal involvement data. $+$ denotes a term included in the model.

age	st	gr	xr	ac	df	D	age	st	gr	xr	ac	df	D
					52	40.71	+	+	+			49	29.76
+					51	39.32	+	+		+		49	23.67
	+				51	33.01	+	+			+	49	25.54
		+			51	35.13	+		+	+		49	27.50
			+		51	31.39	+		+		+	49	26.70
				+	51	33.17	+			+	+	49	24.92
+	+				50	30.90		+	+	+		49	23.98
+		+			50	34.54		+	+		+	49	23.62
+			+		50	30.48		+		+	+	49	19.64
+				+	50	32.67			+	+	+	49	21.28
	+	+			50	31.00	+	+	+	+		48	23.12
	+		+		50	24.92	+	+	+		+	48	23.38
	+			+	50	26.37	+	+		+	+	48	19.22
		+	+		50	27.91	+		+	+	+	48	21.27
		+		+	50	26.72		+	+	+	+	48	18.22
			+	+	50	25.25	+	+	+	+	+	47	18.07

Nodal involvement



- Adding terms
  - always increases the log likelihood $\widehat{\ell}$ and so reduces D,
  - increases the number of parameters,

  so taking the model with highest $\widehat{\ell}$ (lowest D) would give the full model
- We need to trade off quality of fit (measured by D) and model complexity (number of parameters)

Log likelihood

- Given (unknown) **true model** $g(y)$, and **candidate model** $f(y;\theta)$, Jensen's inequality implies that

$$\int \log g(y) g(y)\, dy \geq \int \log f(y;\theta) g(y)\, dy, \tag{1}$$

  with equality if and only if $f(y;\theta) \equiv g(y)$.
- If θ_g is the value of θ that maximizes the expected log likelihood on the right of (1), then it is natural to choose the candidate model that maximises

$$\overline{\ell}(\widehat{\theta}) = n^{-1} \sum_{j=1}^{n} \log f(y;\widehat{\theta}),$$

  which should be an estimate of $\int \log f(y;\theta) g(y)\, dy$. However as $\overline{\ell}(\widehat{\theta}) \geq \overline{\ell}(\theta_g)$, by definition of $\widehat{\theta}$, this estimate is biased upwards.
- We need to correct for the bias, but in order to do so, need to understand the properties of likelihood estimators when the assumed model f is not equal to the true model g.

Wrong model

Suppose the true model is g, that is, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, but we assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$. The log likelihood $\ell(\theta)$ will be maximised at $\widehat{\theta}$, and

$$\overline{\ell}(\widehat{\theta}) = n^{-1}\ell(\widehat{\theta}) \xrightarrow{\text{a.s.}} \int \log f(y;\theta_g) g(y)\, dy, \quad n \to \infty,$$

where θ_g minimizes the Kullback–Leibler discrepancy

$$KL(f_\theta, g) = \int \log \left\{ \frac{g(y)}{f(y;\theta)} \right\} g(y)\, dy.$$

θ_g gives the density $f(y;\theta_g)$ closest to g in this sense, and $\widehat{\theta}$ is determined by the finite-sample version of $\partial KL(f_\theta, g)/\partial\theta$, i.e.

$$0 = n^{-1} \sum_{j=1}^{n} \frac{\partial \log f(y_j;\widehat{\theta})}{\partial \theta}.$$

Wrong model II

Theorem 1 *Suppose the true model is g, that is, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, but we assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$. Then under mild regularity conditions the maximum likelihood estimator $\widehat{\theta}$ satisfies*

$$\widehat{\theta} \;\dot{\sim}\; N_p\left\{\theta_g, I(\theta_g)^{-1}K(\theta_g)I(\theta_g)^{-1}\right\}, \tag{2}$$

where f_{θ_g} is the density minimising the Kullback–Leibler discrepancy between f_θ and g, I is the Fisher information for f, and K is the variance of the score statistic. The likelihood ratio statistic

$$W(\theta_g) = 2\left\{\ell(\widehat{\theta}) - \ell(\theta_g)\right\} \;\dot{\sim}\; \sum_{r=1}^{p} \lambda_r V_r,$$

where $V_1, \ldots, V_p \overset{\text{iid}}{\sim} \chi^2_1$, and the λ_r are eigenvalues of $K(\theta_g)^{1/2} I_g(\theta_g)^{-1} K(\theta_g)^{1/2}$. Thus $\mathrm{E}\{W(\theta_g)\} = \mathrm{tr}\{I(\theta_g)^{-1}K(\theta_g)\}$.

Under the correct model, θ_g is the 'true' value of θ, $K(\theta) = I(\theta)$, $\lambda_1 = \cdots = \lambda_p = 1$, and we recover the usual results.

Note: 'Proof' of Theorem 1

Expansion of the equation defining $\widehat{\theta}$ about θ_g yields

$$\widehat{\theta} \doteq \theta_g + \left\{ -n^{-1} \sum_{j=1}^{n} \frac{\partial^2 \log f(y_j;\theta_g)}{\partial\theta\partial\theta^{\mathrm{T}}} \right\}^{-1} \left\{ n^{-1} \sum_{j=1}^{n} \frac{\partial \log f(y_j;\theta_g)}{\partial\theta} \right\}$$

and a modification of the usual derivation gives

$$\widehat{\theta} \;\dot{\sim}\; N_p\left\{\theta_g, I(\theta_g)^{-1} K(\theta_g) I(\theta_g)^{-1}\right\}, \tag{3}$$

where the *information sandwich* variance matrix depends on

$$\begin{aligned} K(\theta_g) &= n \int \frac{\partial \log f(y;\theta)}{\partial\theta} \frac{\partial \log f(y;\theta)}{\partial\theta^{\mathrm{T}}} g(y)\, dy, \\ I_g(\theta_g) &= -n \int \frac{\partial^2 \log f(y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} g(y)\, dy. \end{aligned} \tag{4}$$

If $g(y) = f(y;\theta)$, so that the supposed density is correct, then θ_g is the true θ, then

$$K(\theta_g) = I_g(\theta_g) = I(\theta),$$

and (2) reduces to the usual approximation.
In practice $g(y)$ is of course unknown, and then $K(\theta_g)$ and $I_g(\theta_g)$ may be estimated by

$$\widehat{K} = \sum_{j=1}^{n} \frac{\partial \log f(y_j;\widehat{\theta})}{\partial\theta} \frac{\partial \log f(y_j;\widehat{\theta})}{\partial\theta^{\mathrm{T}}}, \quad \widehat{J} = -\sum_{j=1}^{n} \frac{\partial^2 \log f(y_j;\widehat{\theta})}{\partial\theta\partial\theta^{\mathrm{T}}}; \tag{5}$$

the latter is just the observed information matrix. We may then construct confidence intervals for θ_g using (2) with variance matrix $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$.
Similar expansions lead to the result for the likelihood ratio statistic.

Out-of-sample prediction

- We need to fix two problems with using $\overline{\ell}(\widehat{\theta})$ to choose the best candidate model:
  - upward bias, as $\overline{\ell}(\widehat{\theta}) \geq \overline{\ell}(\theta_g)$ because $\widehat{\theta}$ is based on $Y_1, \ldots, Y_n$;
  - no penalisation if the dimension of θ increases.
- If we had another independent sample $Y_1^+, \ldots, Y_n^+ \overset{\text{iid}}{\sim} g$ and computed

$$\overline{\ell}^+(\widehat{\theta}) = n^{-1} \sum_{j=1}^{n} \log f(Y_j^+; \widehat{\theta}),$$

then both problems disappear, suggesting that we choose the candidate model that maximises

$$\mathrm{E}_g \left[\mathrm{E}_g^+ \left\{ \overline{\ell}^+(\widehat{\theta}) \right\} \right],$$

where the inner expectation is over the distribution of the Y_j^+, and the outer expectation is over the distribution of $\widehat{\theta}$.

Information criteria

- Previous results on wrong model give

$$\mathrm{E}_g \left[\mathrm{E}_g^+ \left\{ \overline{\ell}^+(\widehat{\theta}) \right\} \right] \doteq \int \log f(y; \theta_g) g(y)\, dy - \frac{1}{2n} \mathrm{tr}\{I_g(\theta_g)^{-1} K(\theta_g)\},$$

where the second term is a penalty that depends on the model dimension.

- We want to estimate this based on $Y_1, \ldots, Y_n$ only, and get

$$\mathrm{E}_g \left\{ \overline{\ell}(\widehat{\theta}) \right\} \doteq \int \log f(y; \theta_g) g(y)\, dy + \frac{1}{2n} \mathrm{tr}\{I_g(\theta_g)^{-1} K(\theta_g)\},$$

- To remove the bias, we aim to maximise

$$\overline{\ell}(\widehat{\theta}) - \frac{1}{n} \mathrm{tr}\{\widehat{J}^{-1} \widehat{K}\},$$

where

$$\widehat{K} = \sum_{j=1}^{n} \frac{\partial \log f(y_j; \widehat{\theta})}{\partial \theta} \frac{\partial \log f(y_j; \widehat{\theta})}{\partial \theta^{\mathrm{T}}}, \quad \widehat{J} = -\sum_{j=1}^{n} \frac{\partial^2 \log f(y_j; \widehat{\theta})}{\partial \theta \partial \theta^{\mathrm{T}}};$$

the latter is just the observed information matrix.

Note: Bias of log likelihood

To compute the bias in $\overline{\ell}(\widehat{\theta})$, we write

$$\begin{aligned} \mathrm{E}_g\left\{\overline{\ell}(\widehat{\theta})\right\} &= \mathrm{E}_g\left\{\overline{\ell}(\theta_g)\right\} + \mathrm{E}\left\{\overline{\ell}(\widehat{\theta}) - \overline{\ell}(\theta_g)\right\} \\ &= \mathrm{E}_g\left\{\overline{\ell}(\theta_g)\right\} + \frac{1}{2n}\mathrm{E}\left\{W(\theta_g)\right\}, \\ &\doteq \mathrm{E}_g\left\{\overline{\ell}(\theta_g)\right\} + \frac{1}{2n}\mathrm{tr}\{I_g(\theta_g)^{-1}K(\theta_g)\}, \end{aligned}$$

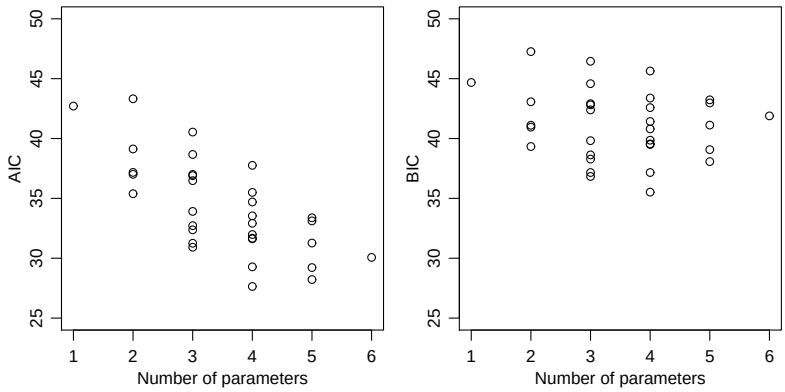
where E_g denotes expectation over the data distribution g.

Information criteria

- Let $p = \dim(\theta)$ be the number of parameters for a model, and $\widehat{\ell}$ the corresponding maximised log likelihood.
- For historical reasons we choose models that **minimise** similar criteria
  - $2(p - \widehat{\ell})$ (AIC—Akaike Information Criterion)
  - $2\{\mathrm{tr}(\widehat{J}^{-1}\widehat{K}) - \widehat{\ell}\}$ (NIC—Network Information Criterion)
  - $2(\frac{1}{2}p\log n - \widehat{\ell})$ (BIC—Bayes Information Criterion)
  - $\mathrm{AIC_c}$, $\mathrm{AIC_u}$, DIC, EIC, FIC, GIC, TIC, ...
  - Mallows $C_p = RSS/s^2 + 2p - n$ commonly used in regression problems, where RSS is residual sum of squares for candidate model, and s^2 is an estimate of the error variance σ^2.

Nodal involvement data

AIC and BIC for 2^5 models for binary logistic regression model fitted to the nodal involvement data. Both criteria pick out the same model, with the three covariates `st`, `xr`, and `ac`, which has deviance $D = 19.64$. Note the sharper increase of BIC after the minimum.



Theoretical aspects

- We may suppose that the true underlying model is of infinite dimension, and that by choosing among our candidate models we hoe to get as close as possible to this ideal model, using the data available.
- If so, we need some measure of distance between a candidate and the true model, and we aim to minimise this distance.
- A model selection procedure that selects the candidate closest to the truth for large n is called **asymptotically efficient**.
- An alternative is to suppose that the true model is among the candidate models.
- If so, then a model selection procedure that selects the true model with probability tending to one as $n \to \infty$ is called **consistent**.

Properties of AIC, NIC, BIC

- We seek to find the correct model by minimising $\mathrm{IC} = c(n,p) - 2\widehat{\ell}$, where the penalty $c(n,p)$ depends on sample size n and model dimension p
- Crucial aspect is behaviour of differences of IC.
- We obtain IC for the true model, and IC_+ for a model with one more parameter. Then

$$\begin{aligned}\Pr(\mathrm{IC}_+ < \mathrm{IC}) &= \Pr\left\{c(n,p+1) - 2\widehat{\ell}_+ < c(n,p) - 2\widehat{\ell}\right\} \\ &= \Pr\left\{2(\widehat{\ell}_+ - \widehat{\ell}) > c(n,p+1) - c(n,p)\right\}.\end{aligned}$$

  and in large samples

$$\begin{aligned}\text{for AIC, } c(n,p+1) - c(n,p) &= 2 \\ \text{for NIC, } c(n,p+1) - c(n,p) &\dot{\sim} 2 \\ \text{for BIC, } c(n,p+1) - c(n,p) &= \log n\end{aligned}$$

- In a regular case $2(\widehat{\ell}_+ - \widehat{\ell}) \dot{\sim} \chi^2_1$, so as $n \to \infty$,

$$\Pr(\mathrm{IC}_+ < \mathrm{IC}) \to \begin{cases} 0.16, & \mathrm{AIC}, \mathrm{NIC}, \\ 0, & \mathrm{BIC}. \end{cases}$$

  Thus AIC and NIC have non-zero probability of over-fitting, even in very large samples, but BIC does not.

Variable selection

- Consider normal linear model

$$Y_{n\times 1} = X^{\dagger}_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n),$$

  where **design matrix** $X^{\dagger}$ has full rank $p < n$ and columns x_r, for $r \in \mathcal{X} = \{1, \ldots, p\}$. Subsets $\mathcal{S}$ of $\mathcal{X}$ correspond to subsets of columns.
- Terminology
  - the **true** model corresponds to subset $\mathcal{T} = \{r : \beta_r \neq 0\}$, and $|\mathcal{T}| = q < p$;
  - a **correct** model contains $\mathcal{T}$ but has other columns also, corresponding subset $\mathcal{S}$ satisfies $\mathcal{T} \subset \mathcal{S} \subset \mathcal{X}$ and $\mathcal{T} \neq \mathcal{S}$;
  - a **wrong** model has subset $\mathcal{S}$ lacking some x_r for which $\beta_r \neq 0$, and so $\mathcal{T} \not\subset \mathcal{S}$.
- Aim to identify $\mathcal{T}$.
- If we choose a wrong model, have bias; if we choose a correct model, increase variance—seek to balance these.

Stepwise methods

- **Forward selection**: starting from model with constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms is significant, stop; otherwise
  3. update the current model to include the most significant new term; go to 1
- **Backward elimination**: starting from model with all terms,
  1. if all terms are significant, stop; otherwise
  2. update current model by dropping the term with the smallest F statistic; go to 1
- **Stepwise**: starting from an arbitary model,
  1. consider 3 options—add a term, delete a term, swap a term in the model for one not in the model;
  2. if model unchanged, stop; otherwise go to 1

Nuclear power station data

```
> nuclear
     cost  date t1 t2  cap pr ne ct bw cum.n pt
1  460.05 68.58 14 46  687  0  1  0  0    14  0
2  452.99 67.33 10 73 1065  0  0  1  0     1  0
3  443.22 67.33 10 85 1065  1  0  1  0     1  0
4  652.32 68.00 11 67 1065  0  1  1  0    12  0
5  642.23 68.00 11 78 1065  1  1  1  0    12  0
6  345.39 67.92 13 51  514  0  1  1  0     3  0
7  272.37 68.17 12 50  822  0  0  0  0     5  0
8  317.21 68.42 14 59  457  0  0  0  0     1  0
9  457.12 68.42 15 55  822  1  0  0  0     5  0
10 690.19 68.33 12 71  792  0  1  1  1     2  0
...
32 270.71 67.83  7 80  886  1  0  0  1    11  1
```


Nuclear power station data

	Full model		Backward		Forward	
	Est (SE)	t	Est (SE)	t	Est (SE)	t
Constant	-14.24 (4.229)	-3.37	-13.26 (3.140)	-4.22	-7.627 (2.875)	-2.66
date	0.209 (0.065)	3.21	0.212 (0.043)	4.91	0.136 (0.040)	3.38
log(T1)	0.092 (0.244)	0.38				
log(T2)	0.290 (0.273)	1.05				
log(cap)	0.694 (0.136)	5.10	0.723 (0.119)	6.09	0.671 (0.141)	4.75
PR	-0.092 (0.077)	-1.20				
NE	0.258 (0.077)	3.35	0.249 (0.074)	3.36		
CT	0.120 (0.066)	1.82	0.140 (0.060)	2.32		
BW	0.033 (0.101)	0.33				
log(N)	-0.080 (0.046)	-1.74	-0.088 (0.042)	-2.11		
PT	-0.224 (0.123)	-1.83	-0.226 (0.114)	-1.99	-0.490 (0.103)	-4.77
s (df)	0.164 (21)		0.159 (25)		0.195 (28)	

Backward selection chooses a model with seven covariates also chosen by minimising AIC.

Stepwise Methods: Comments

- Systematic search minimising AIC or similar over all possible models is preferable—not always feasible.
- Stepwise methods can fit models to purely random data—main problem is no objective function.
- Sometimes used by replacing F significance points by (arbitrary!) numbers, e.g. $F = 4$
- Can be improved by comparing AIC for different models at each step—uses AIC as objective function, but no systematic search.

Prediction error

- To identify $\mathcal{T}$, we fit candidate model

$$Y = X\beta + \varepsilon,$$

  where columns of X are a subset $\mathcal{S}$ of those of $X^\dagger$.
- Fitted value is

$$X\widehat{\beta} = X\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\} = HY = H(\mu + \varepsilon) = H\mu + H\varepsilon,$$

  where $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the **hat matrix** and $H\mu = \mu$ if the model is correct.
- Following reasoning for AIC, suppose we also have independent dataset Y_+ from the true model, so $Y_+ = \mu + \varepsilon_+$
- Apart from constants, previous measure of prediction error is

$$\Delta = n^{-1}\mathrm{E}\ \mathrm{E}_+\left\{(Y_+ - X\widehat{\beta})^{\mathrm{T}}(Y_+ - X\widehat{\beta})\right\},$$

  with expectations over both Y_+ and Y.

Prediction error II

- Can show that

$$\Delta = \begin{cases} n^{-1}\mu^{\mathrm{T}}(I-H)\mu + (1+p/n)\sigma^2, & \text{wrong model,} \\ (1+q/n)\sigma^2, & \text{true model,} \\ (1+p/n)\sigma^2, & \text{correct model;} \end{cases} \tag{6}$$

  recall that $q < p$.
- **Bias**: $n^{-1}\mu^{\mathrm{T}}(I-H)\mu > 0$ unless model is correct, and is reduced by including useful terms
- **Variance**: $(1+p/n)\sigma^2$ increased by including useless terms
- Ideal would be to choose covariates to minimise Δ: impossible—depends on unknowns μ, σ.
- Must estimate Δ

Note: Proof of (6)

Consider data $y = \mu + \varepsilon$ to which we fit the linear model $y = X\beta + \varepsilon$, obtaining fitted value

$$X\widehat{\beta} = Hy = H(\mu + \varepsilon)$$

where the second term is zero if μ lies in the space spanned by the columns of X, and otherwise is not. We have a new data set $y_+ = \mu + \varepsilon_+$, and we will compute the average error in predicting y_+ using $X\widehat{\beta}$, which is

$$\Delta = n^{-1}\mathrm{E}\left\{(y_+ - X\widehat{\beta})^{\mathrm{T}}(y_+ - X\widehat{\beta})\right\}.$$

Now

$$y_+ - X\widehat{\beta} = \mu + \varepsilon_+ - (H\mu + H\varepsilon) = (I - H)\mu + \varepsilon_+ - H\varepsilon.$$

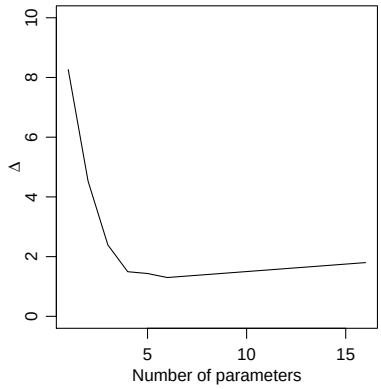
Therefore

$$(y_+ - X\widehat{\beta})^{\mathrm{T}}(y_+ - X\widehat{\beta}) = \mu^{\mathrm{T}}(I - H)\mu + \varepsilon^{\mathrm{T}}H\varepsilon + \varepsilon_+^{\mathrm{T}}\varepsilon_+ + A$$

where $\mathrm{E}(A) = 0$; this gives that

$$\Delta = \begin{cases} n^{-1}\mu^{\mathrm{T}}(I - H)\mu + (1 + p/n)\sigma^2, & \text{wrong model,} \\ (1 + q/n)\sigma^2, & \text{true model,} \\ (1 + p/n)\sigma^2, & \text{correct model.} \end{cases}$$

Example



Δ as a function of the number of included variables p for data with $n = 20$, $q = 6$, $\sigma^2 = 1$. The minimum is at $p = q = 6$:

- ☐ there is a sharp decrease in bias as useful covariates are added;
- ☐ a slow increase with variance as the number of variables p increases.

Cross-validation

- If n is large, can split data into two parts (X', y') and (X^*, y^*), say, and use one part to estimate model, and the other to compute prediction error; then choose the model that minimises

$$\widehat{\Delta} = n'^{-1}(y' - X'\widehat{\beta}^*)^{\mathrm{T}}(y' - X'\widehat{\beta}^*) = n'^{-1}\sum_{j=1}^{n'}(y'_j - x'_j\widehat{\beta}^*)^2.$$

- Usually dataset is too small for this; use **leave-one-out cross-validation** sum of squares

$$n\widehat{\Delta}_{\mathrm{CV}} = \mathrm{CV} = \sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j})^2,$$

  where $\widehat{\beta}_{-j}$ is estimate computed without (x_j, y_j).
- Seems to require n fits of model, but in fact

$$\mathrm{CV} = \sum_{j=1}^{n}\frac{(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2}{(1-h_{jj})^2},$$

  where $h_{11}, \ldots, h_{nn}$ are diagonal elements of H, and so can be obtained from one fit.

Cross-validation II

- Simpler (more stable?) version uses **generalised cross-validation** sum of squares

$$\mathrm{GCV} = \sum_{j=1}^{n}\frac{(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2}{\{1 - \mathrm{tr}(H)/n\}^2}.$$

- Can show that

$$\mathrm{E}(\mathrm{GCV}) = \mu^{\mathrm{T}}(I-H)\mu/(1-p/n)^2 + n\sigma^2/(1-p/n) \approx n\Delta \tag{7}$$

  so try and minimise GCV or CV.
- Many variants of cross-validation exist. Typically find that model chosen based on CV is somewhat unstable, and that GCV or k-fold cross-validation works better. Standard strategy is to split data into 10 roughly equal parts, predict for each part based on the other nine-tenths of the data, and find model that minimises this estimate of prediction error.

Note: Derivation of (7)

We need the expectation of $(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta})$, where $y - X\widehat{\beta} = (I - H)y = (I - H)(\mu + \varepsilon)$, and squaring up and noting that $\mathrm{E}(\varepsilon) = 0$ gives

$$\mathrm{E}\left\{(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta})\right\} = \mu^{\mathrm{T}}(I - H)\mu + \mathrm{E}\left\{\varepsilon^{\mathrm{T}}(I - H)\varepsilon\right\} = \mu^{\mathrm{T}}(I - H)\mu + (n - p)\sigma^2.$$

Now note that $\mathrm{tr}(H) = p$ and divide by $(1 - p/n)^2$ to give (almost) the required result, for which we need also $(1 - p/n)^{-1} \approx 1 + p/n$, for $p \ll n$.

Other selection criteria

- Corrected version of AIC for models with normal responses:

$$\mathrm{AIC_c} \equiv n \log \widehat{\sigma}^2 + n\frac{1 + p/n}{1 - (p + 2)/n},$$

  where $\widehat{\sigma}^2 = \mathrm{RSS}/n$. Related (unbiased) $\mathrm{AIC_u}$ replaces $\widehat{\sigma}^2$ by $S^2 = \mathrm{RSS}/(n - p)$.

- Mallows suggested

$$C_p = \frac{SS_p}{s^2} + 2p - n,$$

  where SS_p is RSS for fitted model and s^2 estimates σ^2.

- Comments:
  - AIC tends to choose models that are too complicated; $\mathrm{AIC_c}$ cures this somewhat
  - BIC chooses true model with probability $\to 1$ as $n \to \infty$, if the true model is fitted.

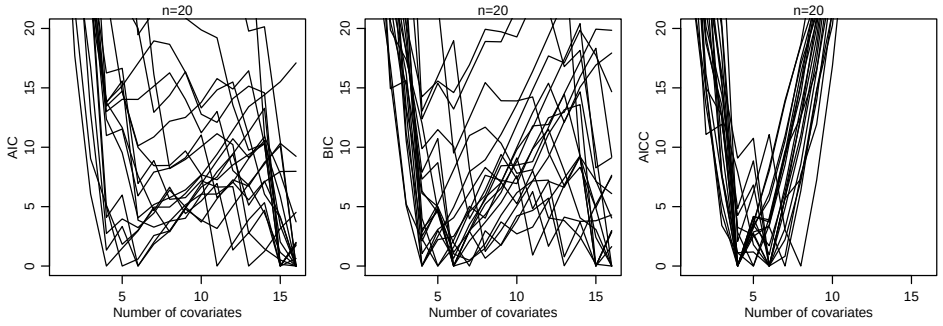
Simulation experiment

Number of times models were selected using various model selection criteria in 50 repetitions using simulated normal data for each of 20 design matrices. The true model has $p = 3$.

n		Number of covariates						
		1	2	3	4	5	6	7
10	C_p		131	504	91	63	83	128
	BIC		72	373	97	83	109	266
	AIC		52	329	97	91	125	306
	$\mathrm{AIC_c}$	15	398	565	18	4		
20	C_p		4	673	121	88	61	53
	BIC		6	781	104	52	30	27
	AIC		2	577	144	104	76	97
	$\mathrm{AIC_c}$		8	859	94	30	8	1
40	C_p			712	107	73	66	42
	BIC			904	56	20	15	5
	AIC			673	114	90	69	54
	$\mathrm{AIC_c}$			786	105	52	41	16

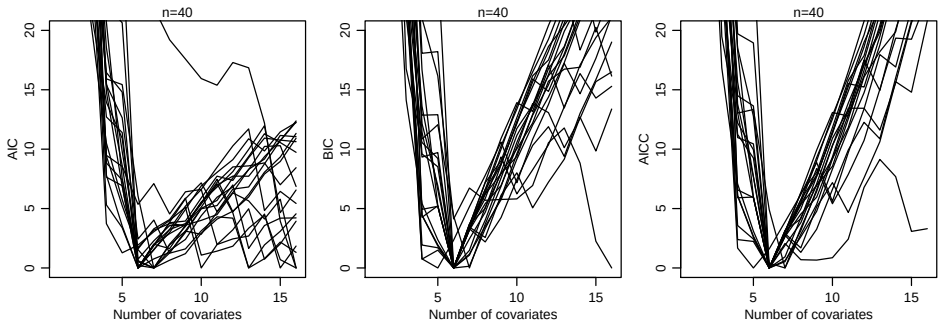
Simulation experiment

Twenty replicate traces of AIC, BIC, and $\mathrm{AIC_c}$, for data simulated with $n = 20$, $p = 1, \ldots, 16$, and $q = 6$.



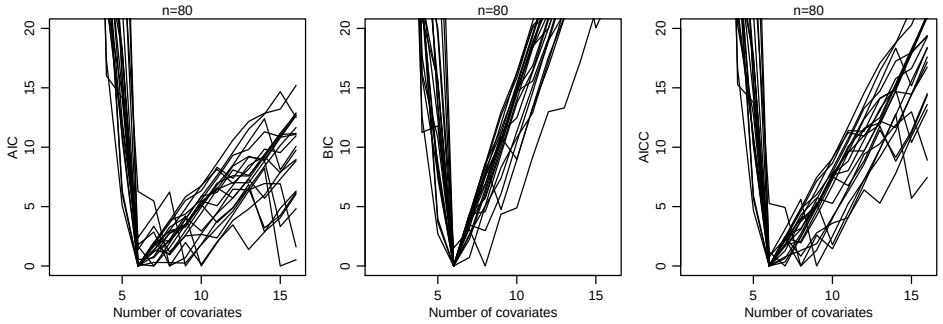
Simulation experiment

Twenty replicate traces of AIC, BIC, and $\mathrm{AIC_c}$, for data simulated with $n = 40$, $p = 1, \ldots, 16$, and $q = 6$.



Simulation experiment

Twenty replicate traces of AIC, BIC, and $\mathrm{AIC_c}$, for data simulated with $n = 80$, $p = 1, \ldots, 16$, and $q = 6$.



As n increases, note how

- ☐ AIC and $\mathrm{AIC_c}$ still allow some over-fitting, but BIC does not, and
- ☐ $\mathrm{AIC_c}$ approaches AIC.

Desiderata

Would like variable selection procedures that satisfy:

- ☐ **near unbiasedness**—the estimators almost provide the true parameters, when these are large and $n \to \infty$;
- ☐ **sparsity**—small estimates are reduced to zero by a threshold procedure; and
- ☐ **continuity**—the estimator is continuous in the data, to avoid instability in prediction.

None of the previous approaches is sparse, and stepwise selection (for example) is known to be highly unstable. To overcome this, we consider a **regularised** (or penalised) log likelihood of the form

$$\tfrac{1}{2}\sum_{j=1}^{n} \ell_j(x_j^{\mathrm{T}}\beta; y_j) - n\sum_{r=1}^{p} p_\lambda(|\beta_r|),$$

where $p_\lambda(|\beta|)$ is a penalty discussed below.

Example: Lasso

- ☐ The **lasso** (least absolute selection and shrinkage operator) chooses β to minimise

$$(y - X\beta)^{\mathrm{T}}(y - X\beta) \text{ such that } \sum_{r=1}^{p} |\beta_r| \leq \lambda,$$

  for some $\lambda > 0$; call resulting estimator $\tilde{\beta}_\lambda$.
- ☐ $\lambda \to 0$ implies $\tilde{\beta}_\lambda \to 0$, and $\lambda \to \infty$ implies $\tilde{\beta}_\lambda \to \widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$.
- ☐ Simple case: orthogonal design matrix $X^{\mathrm{T}}X = I_p$, gives

$$\tilde{\beta}_{\lambda,r} = \begin{cases} 0, & |\widehat{\beta}_r| < \gamma, \\ \mathrm{sign}(\widehat{\beta}_r)(|\widehat{\beta}_r| - \gamma), & \text{otherwise}, \end{cases} \quad r = 1, \ldots, p. \tag{8}$$

- ☐ Call this **soft thresholding**.
- ☐ Generalised version is **least angle regression** (Efron et al., 2004, Annals of Statistics).

Note: Derivation of (8)

If the $X^{\mathrm{T}}X = I_p$, then with the aid of Lagrange multipliers the minimisation problem becomes

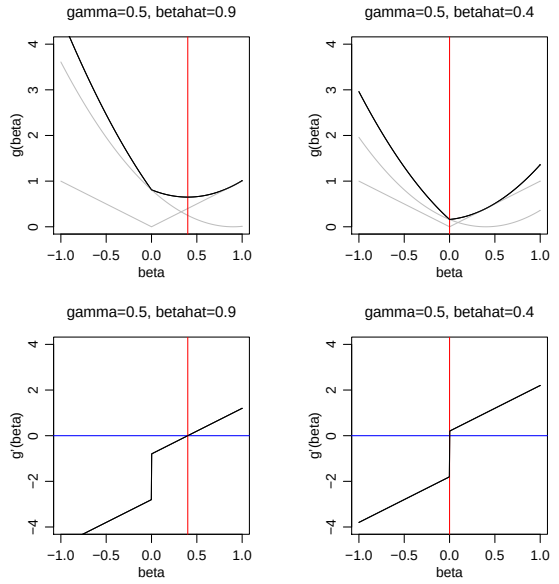
$$\min_{\beta}(y - X\widehat{\beta} + X\widehat{\beta} - X\beta)^{\mathrm{T}}(y - X\widehat{\beta} + X\widehat{\beta} - X\beta) + 2\gamma\left(\sum_{r=1}^{p}|\beta_r| - \lambda\right)$$

and this boils down to individual minimisations of the form

$$\min_{\beta_r} g(\beta_r), \quad g(\beta) = (\beta - \widehat{\beta}_r)^2 + 2\gamma|\beta|.$$

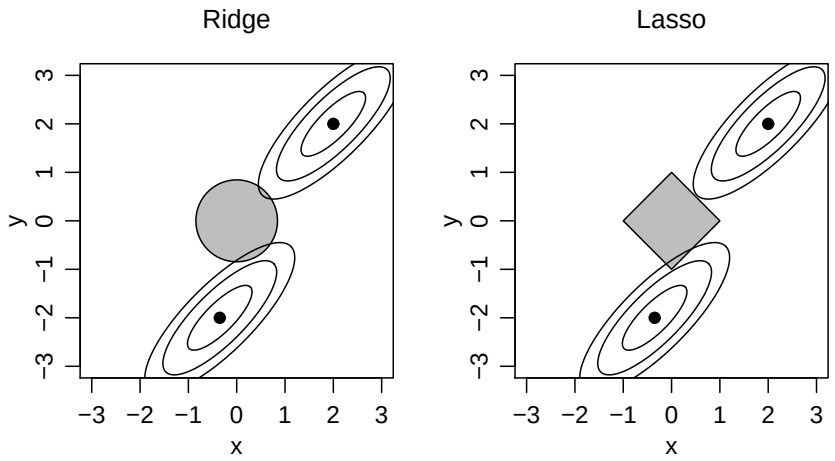
This function is minimised at $\beta = 0$ if and only iff the left and right derivatives there are negative and positive respectively, and this occurs if $|\widehat{\beta}_r| < c$. If not, then $\tilde{\beta} = \widehat{\beta}_r - \gamma$ if $\widehat{\beta} > 0$, and $\tilde{\beta} = \widehat{\beta}_r + \gamma$ if $\widehat{\beta} < 0$. This gives the desired result.

Soft thresholding



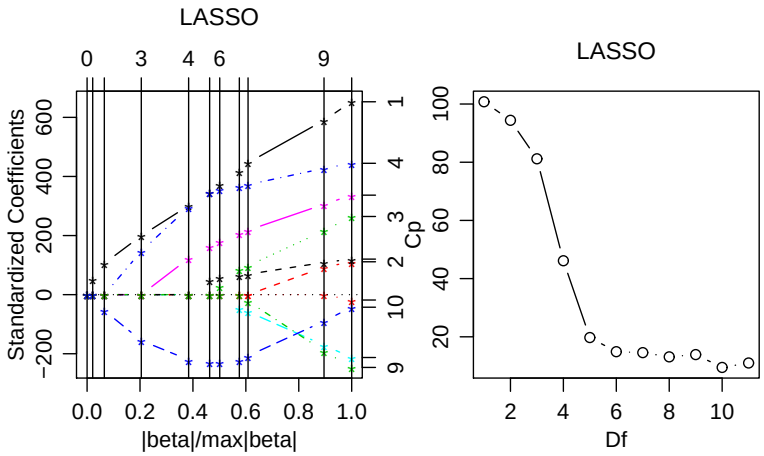
Graphical explanation

In each case aim to minimise the quadratic function subject to remaining inside the shaded region.



Lasso: Nuclear power data

Left: traces of coefficient estimates $\widehat{\beta}_\lambda$ as constraint λ is relaxed, showing points at which the different covariates enter the model. Right: behaviour of Mallows' C_p as λ increases.



Penalties

Some possible penalty functions $p_\lambda(|\beta|)$, all with $\lambda > 0$:

- ☐ **ridge regression** takes $\lambda|\beta|^2$;
- ☐ **lasso** takes $\lambda|\beta|$;
- ☐ **bridge regression** takes $\lambda|\beta|^q$, for $q > 0$;
- ☐ **hard threshold** takes $\lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$;
- ☐ **smoothly clipped absolute deviation (SCAD)** takes

$$\begin{cases} \lambda|\beta|, & |\beta| < \lambda, \\ -(\beta^2 - 2a\lambda|\beta| + \lambda^2)/\{2(a-1)\}, & \lambda < |\beta| < a\lambda, \\ (a+1)\lambda^2/2, & |\beta| > a\lambda, \end{cases}$$

  for some $a > 2$.

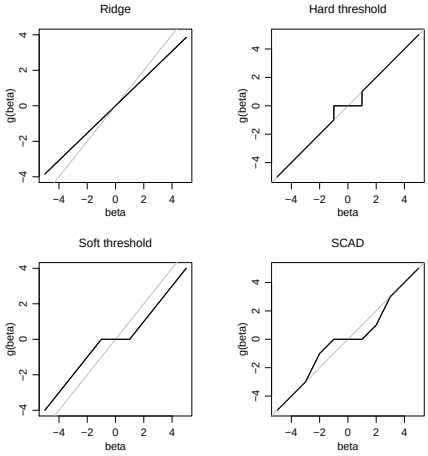
In least squares case with a single observation seek to minimise $\frac{1}{2}(z-\beta)^2 + p_\lambda(|\beta|)$, whose derivative

$$\text{sign}(\beta)\{|\beta| + \partial p_\lambda(|\beta|)/\partial\beta\} - z$$

determines the properties of the estimator.

Some threshold functions

- ☐ Ridge—shrinkage but no selection; hard threshold—subset selection, unstable; soft threshold—lasso, biased; SCAD—continuous, selection, unbiased for large β, but non-monotone.



Properties of penalties

It turns out that to achieve

- **near unbiasedness**, the penalty must satisfy $\partial p_\lambda(|\beta|)/\partial\beta \to 0$ when $|\beta|$ is large, so then the estimating function approaches $\beta - z$;
- **sparsity**, the minimum of the function $|\beta| + \partial p_\lambda(|\beta|)/\partial\beta$ must be positive; and
- **continuity**, the minimum of $|\beta| + \partial p_\lambda(|\beta|)/\partial\beta$ must be attained at $\beta = 0$.

The SCAD is constructed to have these properties, but there is no unique minimum to the resulting objective function, so numerically it is awkward.

Oracle

- Oracle:

  A person or thing regarded as an infallible authority or guide.
- A **statistical oracle** says how to choose the model or bandwidth that will give us optimal estimation of the true parameter or function, but not the truth itself.
- In the context of variable selection, an oracle tells us which variables we should select, but not their coefficients.
- It turns out that under mild conditions on the model, and provided $\lambda \equiv \lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, variable selection using the hard and SCAD penalties has an oracle property: the estimators of β work as well *as if* we had known in advance which covariates should be excluded.
- Same ideas extend to generalised linear models, survival analysis, and many other regression settings (Fan and Li, 2001, JASA).
- Harder: what happens when $p \to \infty$ also?

Bayesian Inference slide 50

Thomas Bayes (1702–1761)

Bayes (1763/4) *Essay towards solving a problem in the doctrine of chances*. Philosophical Transactions of the Royal Society of London.

Bayesian inference

Parametric model for data y assumed to be realisation of $Y \sim f(y;\theta)$, where $\theta \in \Omega_\theta$.
Frequentist viewpoint (cartoon version):

- there is a true value of θ that generated the data;
- this 'true' value of θ is to be treated as an unknown constant;
- probability statements concern randomness in hypothetical replications of the data (possibly conditioned on an ancillary statistic).

Bayesian viewpoint (cartoon version):

- all ignorance may be expressed in terms of probability statements;
- a joint probability distribution for data and all unknowns can be constructed;
- Bayes' theorem should be used to convert prior beliefs $\pi(\theta)$ about unknown θ into posterior beliefs $\pi(\theta \mid y)$, conditioned on data;
- probability statements concern randomness of unknowns, conditioned on all known quantities.

Mechanics

- Separate from data, we have prior information about parameter θ summarised in density $\pi(\theta)$
- Data model $f(y \mid \theta) \equiv f(y;\theta)$
- Posterior density given by Bayes' theorem:

$$\pi(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{\int \pi(\theta) f(y \mid \theta)\, d\theta}.$$

- $\pi(\theta \mid y)$ contains all information about θ, conditional on observed data y
- If $\theta = (\psi, \lambda)$, then inference for ψ is based on **marginal posterior density**

$$\pi(\psi \mid y) = \int \pi(\theta \mid y)\, d\lambda$$

Encompassing model

- Suppose we have M alternative models for the data, with respective parameters $\theta_1 \in \Omega_{\theta_1}, \ldots, \theta_m \in \Omega_{\theta_m}$. Typically dimensions of Ω_{θ_m} are different.
- We enlarge the parameter space to give an **encompassing model** with parameter

$$\theta = (m, \theta_m) \in \Omega = \bigcup_{m=1}^{M} \{m\} \times \Omega_{\theta_m}.$$

- Thus need priors $\pi_m(\theta_m \mid m)$ for the parameters of each model, plus a prior $\pi(m)$ giving pre-data probabilities for each of the models; overall

$$\pi(m, \theta_m) = \pi(\theta_m \mid m)\pi(m) = \pi_m(\theta_m)\pi_m,$$

  say.
- Inference about model choice is based on marginal posterior density

$$\pi(m \mid y) = \frac{\int f(y \mid \theta_m)\pi_m(\theta_m)\pi_m \, d\theta_m}{\sum_{m'=1}^{M} \int f(y \mid \theta_{m'})\pi_{m'}(\theta_{m'})\pi_{m'} \, d\theta_{m'}} = \frac{\pi_m f(y \mid m)}{\sum_{m'=1}^{M} \pi_{m'} f(y \mid m')}.$$

Inference

- Can write

$$\pi(m, \theta_m \mid y) = \pi(\theta_m \mid y, m)\pi(m \mid y),$$

  so Bayesian updating corresponds to

$$\pi(\theta_m \mid m)\pi(m) \mapsto \pi(\theta_m \mid y, m)\pi(m \mid y)$$

  and for each model $m = 1, \ldots, M$ we need
  - posterior probability $\pi(m \mid y)$, which involves the marginal likelihood $f(y \mid m) = \int f(y \mid \theta_m, m)\pi(\theta_m \mid m) \, d\theta_m$; and
  - the posterior density $f(\theta_m \mid y, m)$.
- If there are just two models, can write

$$\frac{\pi(1 \mid y)}{\pi(2 \mid y)} = \frac{\pi_1}{\pi_2} \frac{f(y \mid 1)}{f(y \mid 2)},$$

  so the posterior odds on model 1 equal the prior odds on model 1 multiplied by the **Bayes factor** $B_{12} = f(y \mid 1)/f(y \mid 2)$.

Sensitivity of the marginal likelihood

Suppose the prior for each θ_m is $\mathcal{N}(0, \sigma^2 I_{d_m})$, where $d_m = \dim(\theta_m)$. Then, dropping the m subscript for clarity,

$$\begin{aligned} f(y \mid m) &= \sigma^{-d/2}(2\pi)^{-d/2} \int f(y \mid m, \theta) \prod_r \exp\left\{-\theta_r^2/(2\sigma^2)\right\} \mathrm{d}\theta_r \\ &\approx \sigma^{-d/2}(2\pi)^{-d/2} \int f(y \mid m, \theta) \prod_r \mathrm{d}\theta_r, \end{aligned}$$

for a highly diffuse prior distribution (large σ^2). The Bayes factor for comparing the models is approximately

$$\frac{f(y \mid 1)}{f(y \mid 2)} \approx \sigma^{(d_2 - d_1)/2} g(y),$$

where $g(y)$ depends on the two likelihoods but is independent of σ^2. Hence, *whatever the data tell us about the relative merits of the two models*, the Bayes factor in favour of the simpler model can be made arbitrarily large by increasing σ.
This illustrates **Lindley's paradox**, and implies that we must be careful when specifying prior dispersion parameters to compare models.

Model averaging

- If a quantity Z has the same interpretation for all models, it may be necessary to allow for model uncertainty:
  - in prediction, each model may be just a vehicle that provides a future value, not of interest *per se*;
  - physical parameters (means, variances, etc.) may be suitable for averaging, but care is needed.
- The predictive distribution for Z may be written

$$f(z \mid y) = \sum_{m=1}^{M} f(z \mid y, m) \Pr(m \mid y)$$

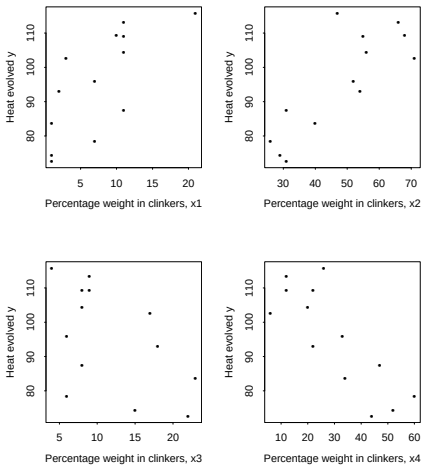
  where

$$\Pr(m \mid y) = \frac{f(y \mid m)\Pr(m)}{\sum_{m'=1}^{M} f(y \mid m')\Pr(m')}$$

- Computational problems can arise if MCMC methods are needed, because jumps between spaces of different dimensions are often required—can be awkward.

Example: Cement data

Percentage weights in clinkers of 4 four constitutents of cement $(x_1, \ldots, x_4)$ and heat evolved y in calories, in $n = 13$ samples.



Example: Cement data

```
> cement
   x1 x2 x3 x4     y
1   7 26  6 60  78.5
2   1 29 15 52  74.3
3  11 56  8 20 104.3
4  11 31  8 47  87.6
5   7 52  6 33  95.9
6  11 55  9 22 109.2
7   3 71 17  6 102.7
8   1 31 22 44  72.5
9   2 54 18 22  93.1
10 21 47  4 26 115.9
11  1 40 23 34  83.8
12 11 66  9 12 113.3
13 10 68  8 12 109.4
```

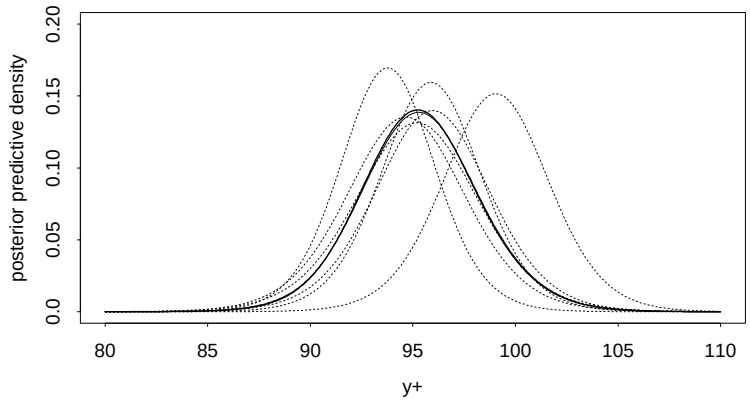
Example: Cement data

Bayesian model choice and prediction using model averaging for the cement data ($n = 13, p = 4$). For each of the 16 possible subsets of covariates, the table shows the log Bayes factor in favour of that subset compared to the model with no covariates and gives the posterior probability of each model. The values of the posterior mean and scale parameters a and b are also shown for the six most plausible models; $(y_+ - a)/b$ has a posterior t density. For comparison, the residual sums of squares are also given.

Model	RSS	$2 \log B_{10}$	$\Pr(M \mid y)$	a	b
– – – –	2715.8	0.0	0.0000		
1 – – –	1265.7	7.1	0.0000		
– 2 – –	906.3	12.2	0.0000		
– – 3 –	1939.4	0.6	0.0000		
– – – 4	883.9	12.6	0.0000		
1 2 – –	57.9	45.7	0.2027	93.77	2.31
1 – 3 –	1227.1	4.0	0.0000		
1 – – 4	74.8	42.8	0.0480	99.05	2.58
– 2 3 –	415.4	19.3	0.0000		
– 2 – 4	868.9	11.0	0.0000		
– – 3 4	175.7	31.3	0.0002		
1 2 3 –	48.11	43.6	0.0716	95.96	2.80
1 2 – 4	47.97	47.2	0.4344	95.88	2.45
1 – 3 4	50.84	44.2	0.0986	94.66	2.89
– 2 3 4	73.81	33.2	0.0004		
1 2 3 4	47.86	45.0	0.1441	95.20	2.97

Example: Cement data

Posterior predictive densities for cement data. Predictive densities for a future observation y_+ with covariate values x_+ based on individual models are given as dotted curves. The heavy curve is the average density from all 16 models.



DIC

- How to compare complex models (e.g. hierarchical models, mixed models, Bayesian settings), in which the 'number of parameters' may:
  - outnumber the number of observations?
  - be unclear because of the regularisation provided by a prior density?
- Suppose model has 'Bayesian deviance'

$$D(\theta) = -2 \log f(y \mid \theta) + 2 \log f(y)$$

  for some normalising function $f(y)$, and suppose that samples from the posterior density of θ are available and give $\overline{\theta} = \mathrm{E}(\theta \mid y)$.
- One possibility is the **deviance information criterion (DIC)**

$$D(\overline{\theta}) + 2p_D,$$

  where the number of associated parameters is

$$p_D = \overline{D(\theta)} - D(\overline{\theta}).$$

- This involves only (MCMC) samples from the posterior, no analytical computations, and reproduces AIC for some classes of models.

Minimum description length

Model selection can also be based on related ideas of **minimum description length (MDL)** or **minimum message length (MML)**, which use ideas from computer science—coding and information theory:

- idea is to choose encoding of data that minimises length of equivalent binary sequence, regarding all data as discrete;
- minimum message includes parameter estimates, data using optimal code based on parameter estimates, (and prior information);
- close links to AIC, BIC, etc.;
- see `http://www.mdl-research.org/` or tutorial on `http://homepages.cwi.nl/~pdg/ftp/mdlintro.pdf` to learn more.

Variable selection

- In Bayesian context, must determine prior probability for the inclusion (or not) of each variable in the model.
- Common to use 'spike and slab' prior for coefficient θ:

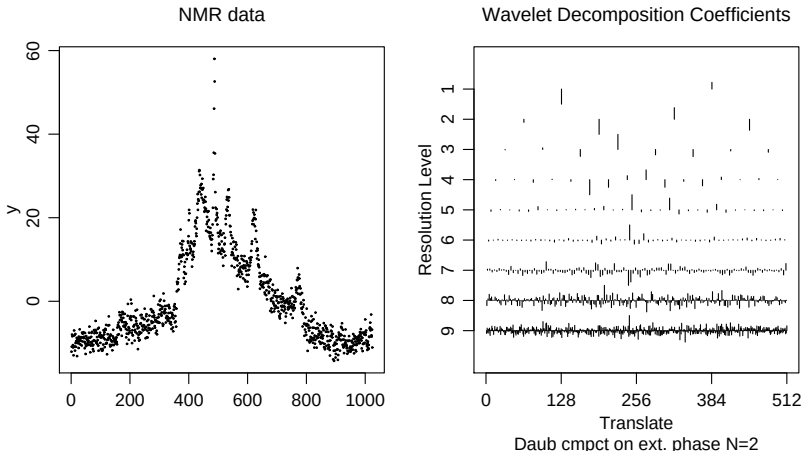
$$\theta = \begin{cases} 0, & \text{with probability } 1-p \\ \mathcal{N}(0,\tau^2), & \text{with probability } p, \end{cases}$$

  corresponding to prior 'density'

$$\pi(\theta) = (1-p)\delta(\theta) + p\tau^{-1}\phi(\theta/\tau), \quad \theta \in \mathbb{R},$$

  where $\delta(\theta)$ is the delta function putting unit mass at $\theta = 0$.
- Now find posterior for β based on data.
- Usually independent priors for each covariate, and typically need clever (dimension-jumping) MCMC.

Example: NMR data



Left: original data, with $n = 1024$
Right: orthogonal transformation into $n = 1024$ coefficients at different resolutions

Orthogonal transformation

- **Model**: original data $X \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$, where signal $\mu_{n\times 1}$ is perturbed by normal noise, giving noisy data $X_{n\times 1}$
- set $Y_{n\times 1} = W_{n\times n} X_{n\times 1}$, where $W^{\mathrm{T}}W = WW^{\mathrm{T}} = I_n$ is orthogonal
- choose W so that $\theta = W\mu$ should be 'sparse' (i.e. most elements of θ are zero)—good choice is wavelet coefficients (mathematical compression properties)
- 'kill' small coefficients of Y, which correspond to noise, giving $\tilde{\theta}_{n\times 1} = \mathrm{kill}(Y) = \mathrm{kill}(WX)$, say, then
- estimate signal μ by

$$\tilde{\mu} = W^{\mathrm{T}}\tilde{\theta} = W^{\mathrm{T}}(\mathrm{kill}(WX)).$$

Posterior

If given θ, $Y \sim \mathcal{N}(\theta, \sigma^2)$, then the posterior 'density' is of form

$$\pi(\theta \mid y) = (1 - p_y)\delta(\theta) + p_y b^{-1}\phi\left(\frac{\theta - ay}{b}\right), \quad \theta \in \mathbb{R},$$

where

$$a = \tau^2/(\tau^2 + \sigma^2), \quad b^2 = 1/(1/\sigma^2 + 1/\tau^2),$$

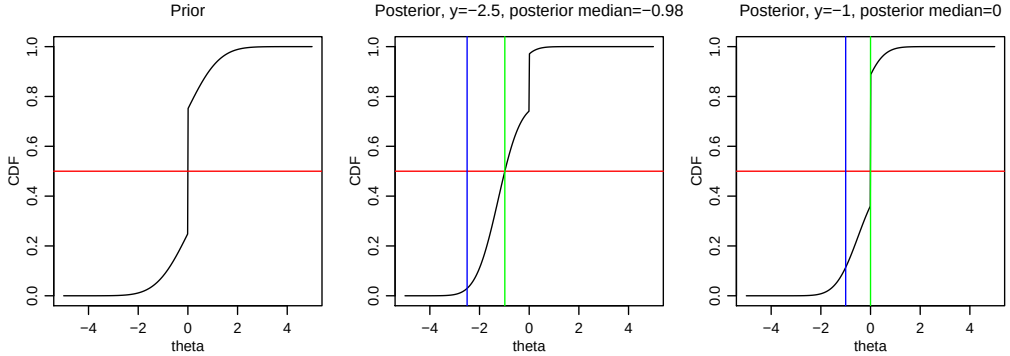
and

$$p_y = \frac{p(\sigma^2 + \tau^2)^{-1/2}\phi\{y/(\sigma^2 + \tau^2)^{1/2}\}}{(1-p)\sigma^{-1}\phi(y/\sigma) + p(\sigma^2 + \tau^2)^{-1/2}\phi\{y/(\sigma^2 + \tau^2)^{1/2}\}}$$

is the posterior probability that $\theta \neq 0$.
Summary statistic: **posterior median** $\tilde{\theta}$, for which $\Pr(\theta \leq \tilde{\theta} \mid y) = 0.5$. For small $|y|$, this gives $\tilde{\theta} = 0$. (Next slide)

Shrinkage

Prior CDF of θ (left), and posterior CDFs when $p = 0.5, \sigma = \tau = 1$, and $y = -2.5$ (centre), and $y = -1$ (right).
Red horizontal line: cumulative probability=0.5
Blue vertical line: data y
Green vertical line: posterior median $\tilde{\theta}$



Empirical Bayes

The parameters p, σ, τ are unknown. We estimate them by empirical Bayes:

- we note that the marginal density of y is

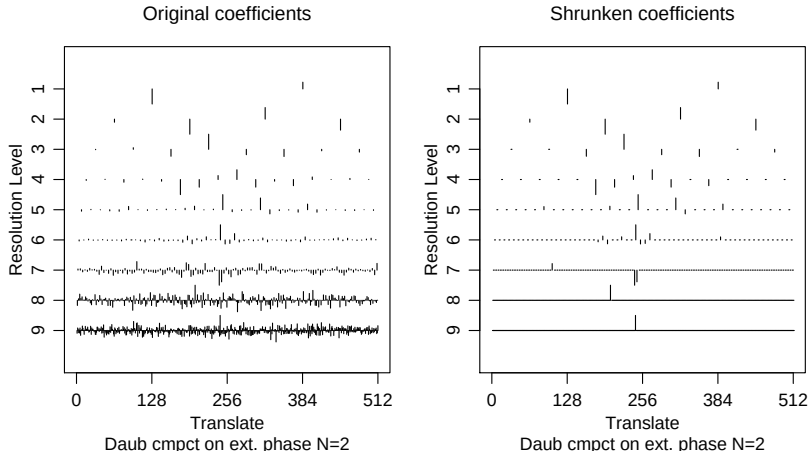
$$f(y) = (1-p)\sigma^{-1}\phi(y/\sigma) + p(\sigma^2+\tau^2)^{-1/2}\phi\{y/(\sigma^2+\tau^2)^{1/2}\}, \quad y \in \mathbb{R},$$

  so if we have $y_1, \ldots, y_n \overset{\text{iid}}{\sim} f$ we estimate p, σ, τ by maximising the log likelihood

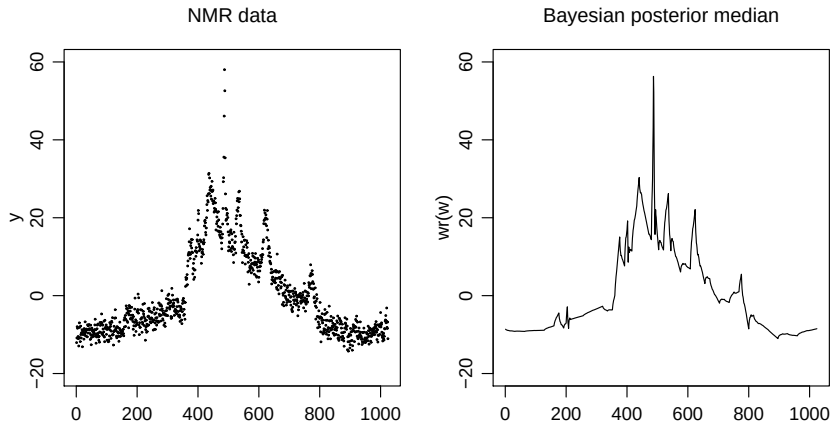
$$\ell(p, \sigma, \tau) = \sum_{j=1}^{n} \log f(y_j; p, \sigma, \tau).$$

- Here we obtain $\tilde{p} = 0.04$, $\tilde{\sigma} = 2.1$, and $\tilde{\tau} = 52.1$.
- Now compute the posterior medians $\tilde{\theta}_j$ corresponding to each y_j.

Example: NMR data



Example: NMR data



Comments

- Large and rapidly-growing literature on Bayesian 'variable' selection, now particularly focused on 'large p, small n' paradigm
- Close relation to classical 'super-efficient' estimation: James–Stein theorem, Hodges–Lehmann estimator, biased (but lower loss) estimation

2. Beyond the Generalised Linear Model

Overview

1. Generalised linear models
2. Overdispersion
3. Correlation
4. Random effects models
5. Conditional independence and graphical representations

Generalised Linear Models

GLM recap

$y_1, \ldots, y_n$ are observations of response variables $Y_1, \ldots, Y_n$ assumed to be independently generated by a distribution of the same exponential family form, with means $\mu_i \equiv \mathrm{E}(Y_i)$ linked to explanatory variables $X_1, X_2, \ldots, X_p$ through

$$g(\mu_i) = \eta_i \equiv \beta_0 + \sum_{r=1}^{p} \beta_r x_{ir} \equiv x_i^{\mathrm{T}} \beta$$

GLMs have proved remarkably effective at modelling real world variation in a wide range of application areas.

GLM failure

However, situations frequently arise where GLMs do not adequately describe observed data.
This can be due to a number of reasons including:

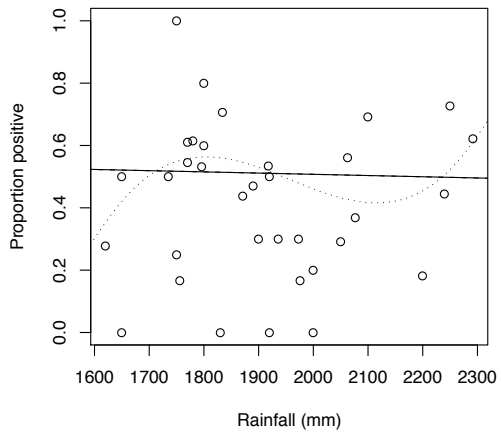
- ☐ The mean model cannot be appropriately specified as there is dependence on an unobserved (or unobservable) explanatory variable.
- ☐ There is excess variability between experimental units beyond that implied by the mean/variance relationship of the chosen response distribution.
- ☐ The assumption of independence is not appropriate.
- ☐ Complex multivariate structure in the data requires a more flexible model class

Example 1: toxoplasmosis

The table below gives data on the relationship between rainfall (x) and the proportions of people with toxoplasmosis (y/m) for 34 cities in El Salvador.

City	y	x	City	y	x	City	y	x
1	5/18	1620	12	3/5	1800	23	3/10	1973
2	15/30	1650	13	8/10	1800	24	1/6	1976
3	0/1	1650	14	0/1	1830	25	1/5	2000
4	2/4	1735	15	53/75	1834	26	0/1	2000
5	2/2	1750	16	7/16	1871	27	7/24	2050
6	2/8	1750	17	24/51	1890	28	46/82	2063
7	2/12	1756	18	3/10	1900	29	7/19	2077
8	6/11	1770	19	23/43	1918	30	9/13	2100
9	33/54	1770	20	3/6	1920	31	4/22	2200
10	8/13	1780	21	0/1	1920	32	4/9	2240
11	41/77	1796	22	3/10	1936	33	8/11	2250
						34	23/37	2292

Example



Toxoplasmosis data and fitted models

Example

Fitting various binomial logistic regression models relating toxoplasmosis incidence to rainfall:

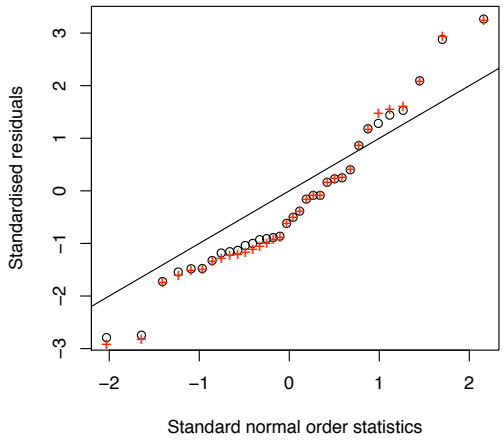
Model	df	deviance
Constant	33	74.21
Linear	32	74.09
Quadratic	31	74.09
Cubic	30	62.62

So evidence in favour of the cubic over other models, but a poor fit ($X^2 = 58.21$ on 30df).

This is an example of **overdispersion** where residual variability is greater than would be predicted by the specified mean/variance relationship

$$\mathrm{var}(Y) = \frac{\mu(1-\mu)}{m}.$$

Example



Toxoplasmosis residual plot

Quasi-likelihood

A quasi-likelihood approach to accounting for overdispersion models the mean and variance, but stops short of a full probability model for Y.

For a model specified by the mean relationship $g(\mu_i) = \eta_i = x_i^{\mathrm{T}}\beta$, and variance $\mathrm{var}(Y_i) = \sigma^2 V(\mu_i)/m_i$, the quasi-likelihood equations are

$$\sum_{i=1}^{n} x_i \frac{y_i - \mu_i}{\sigma^2 V(\mu_i) g'(\mu_i)/m_i} = 0$$

If $V(\mu_i)/m_i$ represents $\mathrm{var}(Y_i)$ for a standard distribution from the exponential family, then these equations can be solved for β using standard GLM software.

Provided the mean and variance functions are correctly specified, asymptotic normality for $\widehat{\beta}$ still holds. The dispersion parameter σ^2 can be estimated using

$$\widehat{\sigma}^2 \equiv \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{m_i(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}$$

Quasi-likelihood for toxoplasmosis data

Assuming the same mean model as before, but $\mathrm{var}(Y_i) = \sigma^2 \frac{\mu_i(1-\mu_i)}{m_i}$, we obtain $\widehat{\sigma}^2 = 1.94$ with $\widehat{\beta}$ (and corresponded fitted mean curves) as before.

Comparing cubic with constant model, one now obtains

$$F = \frac{(74.21 - 62.62)/3}{1.94} = 1.99$$

which provides much less compelling evidence in favour of an effect of rainfall on toxoplasmosis incidence.

Reasons

To construct a full probability model in the presence of overdispersion, it is necessary to consider **why** overdispersion might be present.

Possible reasons include:

- ☐ There may be an important explanatory variable, other than rainfall, which we haven't observed.
- ☐ Or there may be many other features of the cities, possibly unobservable, all having a small individual effect on incidence, but a larger effect in combination. Such effects may be individually undetectable – sometimes described as *natural excess variability between units*.

Reasons: unobserved heterogeneity

When part of the linear predictor is 'missing' from the model,

$$\eta_i^{\text{true}} = \eta_i^{\text{model}} + \eta_i^{\text{diff}}$$

We can compensate for this, in modelling, by assuming that the missing $\eta_i^{\text{diff}} \sim F$ in the population. Hence, given η_i^{model}

$$\mu_i \equiv g^{-1}(\eta_i^{\text{model}} + \eta_i^{\text{diff}}) \sim G$$

where G is the distribution induced by F. Then

$$\mathrm{E}(Y_i) = \mathrm{E}_G[\mathrm{E}(Y_i \mid \mu_i)] = \mathrm{E}_G(\mu_i)$$

$$\mathrm{var}(Y_i) = \mathrm{E}_G(V(\mu_i)/m_i) + \mathrm{var}_G(\mu_i)$$

Direct models

One approach is to model the Y_i directly, by specifying an appropriate form for G.

For example, for the toxoplasmosis data, we might specify a **beta-binomial** model, where

$$\mu_i \sim \mathrm{Beta}(k\mu_i^*, k[1-\mu_i^*])$$

leading to

$$\mathrm{E}(Y_i) = \mu_i^*, \qquad \mathrm{var}(Y_i) = \frac{\mu_i^*(1-\mu_i^*)}{m_i}\left(1 + \frac{m_i - 1}{k+1}\right)$$

with $(m_i - 1)/(k+1)$ representing an overdispersion factor.

Direct models

Models which explicitly account for overdispersion can, in principle, be fitted using your preferred approach, e.g. the beta-binomial model has likelihood

$$f(y \mid \mu^*, k) \propto \prod_{i=1}^{n} \frac{\Gamma(k\mu_i^* + m_i y_i)\Gamma(k(1-\mu_i^*) + m_i(1-y_i))\Gamma(k)}{\Gamma(k\mu_i^*)\Gamma(k(1-\mu_i^*))\Gamma(k+m_i)}.$$

Similarly the corresponding model for count data specifies a gamma distribution for the Poisson mean, leading to a *negative binomial* marginal distribution for Y_i.

However, these models have limited flexibility and can be difficult to fit, so an alternative approach is usually preferred.

A random effects model for overdispersion

A more flexible, and extensible approach models the excess variability by including an extra term in the linear predictor

$$\eta_i = x_i^{\mathrm{T}}\beta + u_i \tag{9}$$

where the u_i can be thought of as representing the 'extra' variability between units, and are called **random effects**.

The model is completed by specifying a distribution F for u_i in the population – almost always, we use

$$u_i \sim N(0, \sigma^2)$$

for some unknown σ^2.
We set $\mathrm{E}(u_i) = 0$, as an unknown mean for u_i would be unidentifiable in the presence of the intercept parameter β_0.

Random effects: likelihood

The parameters of this random effects model are usually considered to be (β, σ^2) and therefore the likelihood is given by

$$\begin{aligned} f(y \mid \beta, \sigma^2) &= \int f(y \mid \beta, u, \sigma^2) f(u \mid \beta, \sigma^2) \mathrm{d}u \\ &= \int f(y \mid \beta, u) f(u \mid \sigma^2) \mathrm{d}u \\ &= \int \prod_{i=1}^{n} f(y_i \mid \beta, u_i) f(u_i \mid \sigma^2) \mathrm{d}u_i \end{aligned} \tag{10}$$

where $f(y_i \mid \beta, u_i)$ arises from our chosen exponential family, with linear predictor (9) and $f(u_i \mid \sigma^2)$ is a univariate normal p.d.f.

Usually no further simplification of (10) is possible, so computation needs careful consideration – we will come back to this later.

Toxoplasmosis example revisited

We can think of the toxoplasmosis proportions Y_i in each city (i) as arising from the sum of binary variables, Y_{ij}, representing the toxoplasmosis status of individuals (j), so $m_i Y_i = \sum_{j=1}^{m_i} Y_{ij}$.
Then

$$\begin{aligned} \mathrm{var}(Y_i) &= \frac{1}{m_i^2}\sum_{j=1}^{m_i}\mathrm{var}(Y_{ij}) + \frac{1}{m_i^2}\sum_{j\neq k}\mathrm{cov}(Y_{ij},Y_{ik}) \\ &= \frac{\mu_i(1-\mu_i)}{m_i} + \frac{1}{m_i^2}\sum_{j\neq k}\mathrm{cov}(Y_{ij},Y_{ik}) \end{aligned}$$

So any positive correlation between individuals induces overdispersion in the counts.

Dependence: reasons

There may be a number of plausible reasons why the responses corresponding to units within a given **cluster** are dependent (in the toxoplasmosis example, cluster = city)

One compelling reason is the unobserved heterogeneity discussed previously.
In the 'correct' model (corresponding to η_i^{true}), the toxoplasmosis status of individuals, Y_{ij}, are independent, so

$$Y_{ij} \perp\!\!\!\perp Y_{ik} \mid \eta_i^{\mathrm{true}} \quad \Leftrightarrow \quad Y_{ij} \perp\!\!\!\perp Y_{ik} \mid \eta_i^{\mathrm{model}}, \eta_i^{\mathrm{diff}}$$

However, in the absence of knowledge of η_i^{diff}

$$Y_{ij} \not\perp\!\!\!\perp Y_{ik} \mid \eta_i^{\mathrm{model}}$$

Hence conditional (given η_i^{diff}) independence between units in a common cluster i becomes marginal dependence, when marginalised over the population distribution F of unobserved η_i^{diff}.

Random effects and dependence

The correspondence between positive intra-cluster correlation and unobserved heterogeneity suggests that intra-cluster dependence might be modelled using random effects, For example, for the individual-level toxoplasmosis data

$$Y_{ij} \overset{\mathrm{ind}}{\sim} \mathrm{Bernoulli}(\mu_{ij}), \quad \log\tfrac{\mu_{ij}}{1-\mu_{ij}} = x_{ij}^{\mathrm{T}}\beta + u_i, \quad u_i \sim N(0,\sigma^2)$$

which implies

$$Y_{ij} \not\perp\!\!\!\perp Y_{ik} \mid \beta, \sigma^2$$

Intra-cluster dependence arises in many applications, and random effects provide an effective way of modelling it.

Marginal models

It should be noted that random effects modelling is not the only way of accounting for intra-cluster dependence.

A **marginal model** models $\mu_{ij} \equiv \mathrm{E}(Y_{ij})$ as a function of explanatory variables, through $g(\mu_{ij}) = x_{ij}^{\mathrm{T}}\beta$, and also specifies a variance relationship $\mathrm{var}(Y_{ij}) = \sigma^2 V(\mu_{ij})/m_{ij}$ and a model for $\mathrm{corr}(Y_{ij}, Y_{ik})$, as a function of μ and possibly additional parameters.

It is important to note that the parameters β in a marginal model have a different interpretation from those in a random effects model, because for the latter

$$\mathrm{E}(Y_{ij}) = \mathrm{E}(g^{-1}[x_{ij}^{\mathrm{T}}\beta + u_i]) \neq g^{-1}(x_{ij}^{\mathrm{T}}\beta) \quad (\text{unless } g \text{ is linear}).$$

- ☐ A random effects model describes the mean response at the subject level ('subject specific')
- ☐ A marginal model describes the mean response across the population ('population averaged')

GEEs

As with the quasi-likelihood approach above, marginal models do not generally provide a full probability model for Y. Nevertheless, β can be estimated using **generalised estimating equations (GEEs)**.

The GEE for estimating β in a marginal model is of the form

$$\sum_i \left(\frac{\partial \mu_i}{\partial \beta}\right)^{\mathrm{T}} \mathrm{var}(Y_i)^{-1}(Y_i - \mu_i) = 0$$

where $Y_i = (Y_{ij})$ and $\mu_i = (\mu_{ij})$

Consistent covariance estimates are available for GEE estimators.

Furthermore, the approach is generally robust to mis-specification of the correlation structure.

For the rest of this module, we focus on fully specified probability models.

Clustered data

Examples where data are collected in clusters include:

- ☐ Studies in biometry where **repeated measures** are made on experimental units. Such studies can effectively mitigate the effect of between-unit variability on important inferences.
- ☐ Agricultural field trials, or similar studies, for example in engineering, where experimental units are arranged within **blocks**
- ☐ Sample surveys where collecting data within clusters or **small areas** can save costs

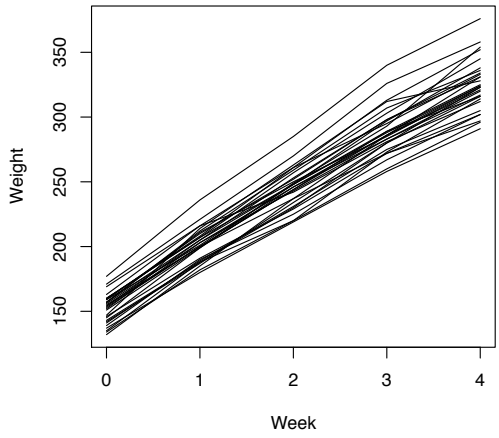
Of course, other forms of dependence exist, for example spatial or serial dependence induced by arrangement in space or time of units of observation. This will be the focus of a later APTS module.

Example 2: Rat growth

The table below is extracted from a data set giving the weekly weights of 30 young rats.

	Week				
Rat	1	2	3	4	5
1	151	199	246	283	320
2	145	199	249	293	354
3	147	214	263	312	328
4	155	200	237	272	297
5	135	188	230	280	323
6	159	210	252	298	331
7	141	189	231	275	305
8	159	201	248	297	338
...	...	...	...	...	...
30	153	200	244	286	324

Example



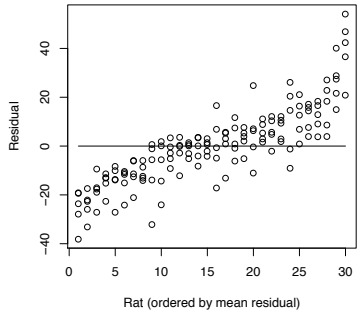
Rat growth data

A simple model

Letting Y represent weight, and X represent week, we can fit the simple linear regression

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

with resulting estimates $\widehat{\beta}_0 = 156.1$ (2.25) and $\widehat{\beta}_1 = 43.3$ (0.92)
Residuals show clear evidence of an unexplained difference between rats

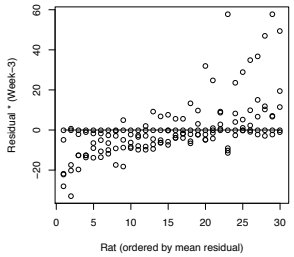


Model elaboration

Naively adding a (fixed) effect for animal

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \epsilon_{ij}$$

Residuals show evidence of a further unexplained difference between rats in terms of dependence on x



More complex cluster dependence required.

Linear mixed models

A linear mixed model (LMM) for observations $y = (y_1, \ldots, y_n)$ has the general form

$$Y \sim N(\mu, \Sigma), \quad \mu = X\beta + Zb, \quad b \sim N(0, \Sigma_b) \tag{11}$$

where X and Z are matrices containing values of explanatory variables. Usually, $\Sigma = \sigma^2 I_n$.
A typical example for clustered data might be

$$Y_{ij} \overset{\text{ind}}{\sim} N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}} b_i, \quad b_i \overset{\text{ind}}{\sim} N(0, \Sigma_b^*) \tag{12}$$

where x_{ij} contain the explanatory data for cluster i, observation j and (normally) z_{ij} contains that sub-vector of x_{ij} which is allowed to exhibit extra between cluster variation in its relationship with Y. In the simplest (random intercept) case, $z_{ij} = (1)$, as in equation (9).

LMM example

A plausible LMM for k clusters with $n_1, \ldots, n_k$ observations per cluster, and a single explanatory variable x (e.g. the rat growth data) is

$$y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})x_{ij} + \epsilon_{ij}, \quad (b_{0i}, b_{1i})^{\mathrm{T}} \overset{\text{ind}}{\sim} N(0, \Sigma_b^*)$$

This fits into the general LMM framework (11) with $\Sigma = \sigma^2 I_n$ and

$$X = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{kn_k} \end{pmatrix}, \quad Z = \begin{pmatrix} Z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_k \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}, \quad b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}, \quad \Sigma_b = \begin{pmatrix} \Sigma_b^* & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_b^* \end{pmatrix}$$

where Σ_b^* is an unspecified 2×2 positive definite matrix.

Variance components

The term **mixed model** refers to the fact that the linear predictor $X\beta + Zb$ contains both fixed effects β and random effects b.
Under an LMM, we can write the marginal distribution of Y directly as

$$Y \sim N(X\beta, \Sigma + Z\Sigma_b Z^{\mathrm{T}}) \tag{13}$$

where X and Z are matrices containing values of explanatory variables.
Hence $\mathrm{var}(Y)$ is comprised of two **variance components**.

Other ways of describing LMMs for clustered data, such as (12) (and their generalised linear model counterparts) are as **hierarchical** models or **multilevel** models. This reflects the two-stage structure of the model, a conditional model for $Y_{ij} \mid b_i$, followed by a marginal model for the random effects b_i.

Sometimes the hierarchy can have further levels, corresponding to clusters nested within clusters, for example, patients within wards within hospitals, or pupils within classes within schools.

Discussion: Why random effects?

It would be perfectly possible to take a model such as (12) and ignore the final component, leading to fixed cluster effects (as we did for the rat growth data).

The main issue with such an approach is that inferences, particularly predictive inferences can then only be made about those clusters present in the observed data.
Random effects models, on the other hand, allow inferences to be extended to a wider population (at the expense of a further modelling assumption).

It also can be the case, as in (9) with only one observation per 'cluster', that fixed effects are not identifiable, whereas random effects can still be estimated. Similarly, some treatment variables must be applied at the cluster level, so fixed treatment and cluster effects are aliased.

Random effects allow 'borrowing strength' across clusters by shrinking fixed effects towards a common mean.

Discussion: A Bayesian perspective

A Bayesian LMM supplements (11) with prior distributions for β, Σ and Σ_b.

In one sense the distinction between fixed and random effects is much less significant, as in the full Bayesian probability specification, both β and b, as unknowns have probability distributions, $f(\beta)$ and $f(b) = \int f(b \mid \Sigma_b) f(\Sigma_b) \mathrm{d}\Sigma_b$

Indeed, prior distributions for 'fixed' effects are sometimes constructed in a hierarchical fashion, for convenience (for example, heavy-tailed priors are often constructed this way).

The main difference is the possibility that random effects for which we have no relevant data (for example cluster effects for unobserved clusters) might need to be predicted.

LMM fitting

The likelihood for $(\beta, \Sigma, \Sigma_b)$ is available directly from (13) as

$$f(y \mid \beta, \Sigma, \Sigma_b) \propto |V|^{-1/2} \exp\left(\tfrac{1}{2}(y - X\beta)^{\mathrm{T}} V^{-1}(y - X\beta)\right) \tag{14}$$

where $V = \Sigma + Z\Sigma_b Z^{\mathrm{T}}$. This likelihood can be maximised directly (usually numerically).

However, mles for variance parameters of LMMs can have large downward bias (particularly in cluster models with a small number of observed clusters).
Hence estimation by **REML** – *REstricted* (or *REsidual*) Maximum Likelihood is usually preferred.

REML proceeds by estimating the variance parameters (Σ, Σ_b) using a *marginal likelihood* based on the residuals from a (generalised) least squares fit of the model $\mathrm{E}(Y) = X\beta$.

REML

In effect, REML maximizes the likelihood of any linearly independent sub-vector of $(I_n - H)y$ where $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the usual hat matrix. As

$$(I_n - H)y \sim N(0, (I_n - H)V(I_n - H))$$

this likelihood will be free of β. It can be written in terms of the full likelihood (14) as

$$f(r \mid \Sigma, \Sigma_b) \propto f(y \mid \widehat{\beta}, \Sigma, \Sigma_b)|X^{\mathrm{T}}VX|^{1/2} \tag{15}$$

where

$$\widehat{\beta} = (X^{\mathrm{T}}V^{-1}X)^{-1}X^{\mathrm{T}}V^{-1}y \tag{16}$$

is the usual generalised least squares estimator given known V.
Having first obtained $(\widehat{\Sigma}, \widehat{\Sigma}_b)$ by maximising (15), $\widehat{\beta}$ is obtained by plugging the resulting $\widehat{V}$ into (16).
Note that REML maximised likelihoods cannot be used to compare different fixed effects specifications, due to the dependence of 'data' r in $f(r \mid \Sigma, \Sigma_b)$ on X.

Estimating random effects

A natural predictor $\tilde{b}$ of the random effect vector b is obtained by minimising the mean squared prediction error $\mathrm{E}[(\tilde{b}-b)^{\mathrm{T}}(\tilde{b}-b)]$ where the expectation is over both b and y.
This is achieved by

$$\tilde{b} = \mathrm{E}(b \mid y) = (Z^{\mathrm{T}}\Sigma^{-1}Z + \Sigma_b^{-1})^{-1}Z^{\mathrm{T}}\Sigma^{-1}(y - X\beta) \tag{17}$$

giving the **Best Linear Unbiased Predictor** (BLUP) for b, with corresponding variance

$$\mathrm{var}(b \mid y) = (Z^{\mathrm{T}}\Sigma^{-1}Z + \Sigma_b^{-1})^{-1}$$

The estimates are obtained by plugging in $(\widehat{\beta}, \widehat{\Sigma}, \widehat{\Sigma}_b)$, and are **shrunk** towards 0, in comparison with equivalent fixed effects estimators.

Any component, b_k of b with no relevant data (for example a cluster effect for an as yet unobserved cluster) corresponds to a null column of Z, and then $\tilde{b}_k = 0$ and $\mathrm{var}(b_k \mid y) = [\Sigma_b]_{kk}$, which may be estimated if, as is usual, b_k shares a variance with other random effects.

Bayesian estimation: the Gibbs sampler

Bayesian estimation in LMMs (and their generalised linear model counterparts) generally proceeds using **Markov Chain Monte Carlo (MCMC)** methods, in particular approaches based on the **Gibbs sampler**. Such methods have proved very effective.

MCMC computation provides posterior summaries, by **generating** a **dependent** sample from the posterior distribution of interest. Then, any posterior expectation can be estimated by the corresponding Monte Carlo sample mean, densities can be estimated from samples etc.

The theory and application of MCMC will be covered in a later APTS module. Here we simply describe the (most basic) Gibbs sampler.

To generate from $f(y_1, \ldots, y_n)$, (where the component y_is are allowed to be multivarate) the Gibbs sampler starts from an arbitrary value of y and updates components (sequentially or otherwise) by generating from the conditional distributions $f(y_i \mid y_{\setminus i})$ where $y_{\setminus i}$ are all the variables other than y_i, set at their currently generated values.

Hence, to apply the Gibbs sampler, we require conditional distributions which are available for sampling.

Bayesian estimation for LMMs

For the LMM

$$Y \sim N(\mu, \Sigma), \quad \mu = X\beta + Zb, \quad b \sim N(0, \Sigma_b)$$

with corresponding prior densities $f(\beta)$, $f(\Sigma)$, $f(\Sigma_b)$, we obtain the *conditional* posterior distributions

$$\begin{aligned} f(\beta \mid y, \text{rest}) &\propto \phi(y - Zb; X\beta, \Sigma) f(\beta) \\ f(b \mid y, \text{rest}) &\propto \phi(y - X\beta; Zb, \Sigma)\phi(b; 0, \Sigma_b) \\ f(\Sigma \mid y, \text{rest}) &\propto \phi(y - X\beta - Zb; 0, \Sigma) f(\Sigma) \\ f(\Sigma_b \mid y, \text{rest}) &\propto \phi(b; 0, \Sigma_b) f(\Sigma_b) \end{aligned}$$

where $\phi(y; \mu, \Sigma)$ is a $N(\mu, \Sigma)$ p.d.f. evaluated at y.

We can exploit **conditional conjugacy** in the choices of $f(\beta)$, $f(\Sigma)$, $f(\Sigma_b)$ making the conditionals above of known form and hence straightforward to sample from. The conditional independence $(\beta, \Sigma) \perp\!\!\!\perp \Sigma_b \mid b$ is also helpful.

See Practical 2 for further details.

Example: Rat growth revisited

Here, we consider the model

$$y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})x_{ij} + \epsilon_{ij}, \quad (b_{0i}, b_{1i})^{\mathrm{T}} \overset{\text{ind}}{\sim} N(0, \Sigma_b)$$

where $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and Σ_b is an unspecified covariance matrix. This model allows for random (cluster specific) slope and intercept.

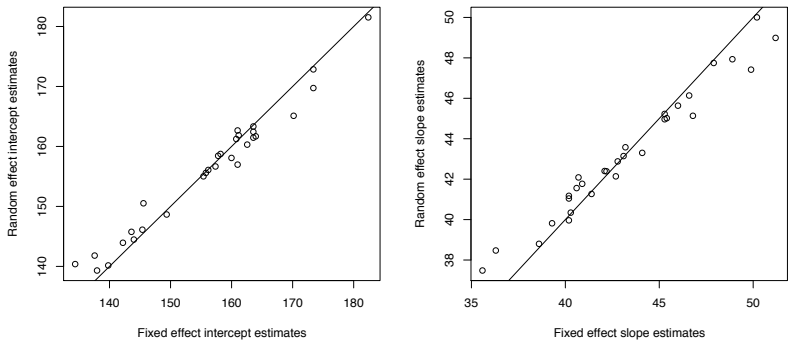
Estimates obtained by REML (ML in brackets) are

Parameter	Estimate	Standard error
β_0	156.05	2.16 (2.13)
β_1	43.27	0.73 (0.72)
$\Sigma_{00}^{1/2} = s.d.(b_0)$	10.93 (10.71)	
$\Sigma_{11}^{1/2} = s.d.(b_1)$	3.53 (3.46)	
$Corr(b_0, b_1)$	0.18 (0.19)	

As expected ML variances are smaller, but not by much.

Example: Fixed v. random effect estimates

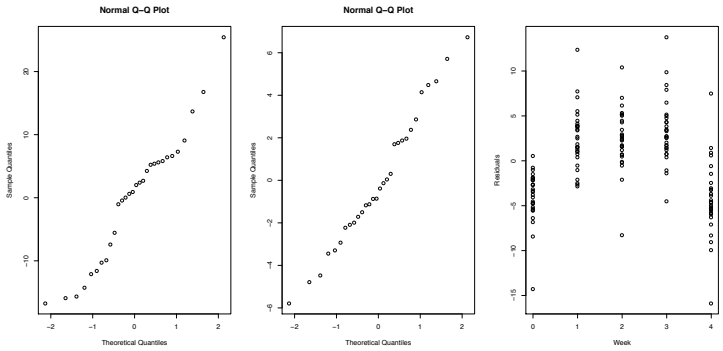
The shrinkage of random effect estimates towards a common mean is clearly illustrated.



Random effects estimates 'borrow strength' across clusters, due to the Σ_b^{-1} term in (17). Extent of this is determined by cluster similarity. This is usually considered to be a desirable behaviour.

Example: Diagnostics

Normal Q-Q plots of intercept (panel 1) and slope (panel 2) random effects and residuals v. week (panel 3)



Evidence of a common quadratic effect, confirmed by AIC (1036 v. 1099) and BIC (1054 v. 1114) based on full ML fits. AIC would also include a cluster quadratic effect (BIC equivocal).

Generalised linear mixed models

Generalised linear mixed models (GLMMs) generalise LMMs to non-normal data, in the obvious way:

$$Y_i \overset{\text{ind}}{\sim} F(\cdot \mid \mu_i, \sigma^2), \quad g(\mu) \equiv \begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_n) \end{pmatrix} = X\beta + Zb, \quad b \sim N(0, \Sigma_b) \tag{18}$$

where $F(\cdot \mid \mu_i, \sigma^2)$ is an exponential family distribution with $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = \sigma^2 V(\mu)/m$ for known m. Commonly (e.g. Binomial, Poisson) $\sigma^2 = 1$, and we shall assume this from here on.

It is not necessary that the distribution for the random effects b is normal, but this usually fits. It is possible (but beyond the scope of this module) to relax this.

GLMM example

A plausible GLMM for binary data in k clusters with $n_1, \ldots, n_k$ observations per cluster, and a single explanatory variable x (e.g. the toxoplasmosis data at individual level) is

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_i), \quad \log \tfrac{\mu_i}{1-\mu_i} = \beta_0 + b_{0i} + \beta_1 x_{ij}, \quad b_{0i} \overset{\text{ind}}{\sim} N(0, \sigma_b^2) \tag{19}$$

[note: no random slope here]. This fits into the general GLMM framework (18) with

$$X = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{kn_k} \end{pmatrix}, \quad Z = \begin{pmatrix} Z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_k \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

$$\beta = (\beta_0, \beta_1)^{\mathrm{T}}, \quad b = (b_{01}, \ldots, b_{0k})^{\mathrm{T}}, \quad \Sigma_b = \sigma_b^2 I_k$$

[or equivalent binomial representation for city data, with clusters of size 1.]

GLMM likelihood

The marginal distribution for the observed Y in a GLMM does not usually have a convenient closed-form representation.

$$\begin{aligned} f(y \mid \beta, \Sigma_b) &= \int f(y \mid \beta, b, \Sigma_b) f(b \mid \beta, \Sigma_b) \mathrm{d}b \\ &= \int f(y \mid \beta, b) f(b \mid \Sigma_b) \mathrm{d}b \\ &= \int \prod_{i=1}^{n} f\left(y_i \mid g^{-1}([X\beta + Zb]_i)\right) f(b \mid \Sigma_b) \mathrm{d}b. \end{aligned} \tag{20}$$

For **nested** random effects structures, some simplification is possible. For example, for (19)

$$f(y \mid \beta, \sigma_b^2) \propto \prod_{i=1}^{n} \int \frac{\exp\left(\sum_j y_{ij}(\beta_0 + b_{0i} + \beta_1 x_{ij})\right)}{\left(1+\exp\left(\sum_j y_{ij}(\beta_0 + b_{0i} + \beta_1 x_{ij})\right)\right)^{n_k}} \phi(b_{0i}; 0, \sigma_b^2) \mathrm{d}b_{0i}$$

a product of one-dimensional integrals.

GLMM fitting: quadrature

Fitting a GLMM by likelihood methods requires some method for approximating the integrals involved.

The most reliable when the integrals are of low dimension is to use Gaussian quadrature (see APTS: Statistical computing). For example, for a one-dimensional cluster-level random intercept b_i we might use

$$\begin{aligned} &\int \prod_j f\left(y_{ij} \mid g^{-1}(x_i^{\mathrm{T}}\beta + b_i)\right) \phi(b_i \mid 0, \sigma_b^2) \mathrm{d}b_i \\ &\qquad \approx \sum_{q=1}^{Q} w_q \prod_j f\left(y_{ij} \mid g^{-1}(x_i^{\mathrm{T}}\beta + b_{iq})\right) \end{aligned}$$

for suitably chosen weights $(w_q, q = 1, \ldots, Q)$ and quadrature points $(b_{iq}, q = 1, \ldots, Q)$

Effective quadrature approaches use information about the mode and dispersion of the integrand (can be done adaptively).

For multi-dimensional b_i, quadrature rules can be applied recursively, but performance (in fixed time) diminishes rapidly with dimension.

GLMM fitting: Penalised quasi-likelihood

An alternative approach to fitting a GLMM uses penalised quasi-likelihood (PQL).

The most straightforward way of thinking about PQL is to consider the adjusted dependent variable v constructed when calculating mles for a GLM using Fisher scoring

$$v_i = (y_i - \mu_i)g'(\mu_i) + \eta_i$$

Now, for a GLMM,

$$\mathrm{E}(v \mid b) = \eta = X\beta + Zb$$

and

$$\mathrm{var}(v \mid b) = W^{-1} = \mathrm{diag}\left(\mathrm{var}(y_i)g'(\mu_i)^2\right),$$

where W is the weight matrix used in Fisher scoring.

GLMM fitting: PQL continued

Hence, approximating the conditional distribution of z by a normal distribution, we have

$$v \sim N(X\beta + Zb, W^{-1}), \quad b \sim N(0, \Sigma_b) \tag{21}$$

where v and W also depend on β and b.

PQL proceeds by iteratively estimating β, b and Σ_b for the linear mixed model (21) for v, updating v and W at each stage, based on the current estimates of β and b.

An alternative justification for PQL is as using a Laplace-type approximation to the integral in the GLMM likelihood.

A full Laplace approximation (expanding the complete log-integrand, and evaluating the Hessian matrix at the mode) is an alternative approach, which itself is a one-point Gaussian quadrature.

GLMM fitting: discussion

Using PQL, estimates of random effects b come 'for free'. With Gaussian quadrature, some extra effort is required to compute $\mathrm{E}(b \mid y)$ – quadrature is an obvious possibility.

There are drawbacks with PQL, and the best advice is to use it with caution.

- ☐ It can fail badly when the normal approximation that justifies it is invalid (for example for binary observations)
- ☐ As it does not use a full likelihood, model comparison should not be performed using PQL maximised 'likelihoods'

Likelihood inference for GLMMs remains an area of active research and vigorous debate. Recent approaches include HGLMs (hierarchical GLMs) where inference is based on the h-likelihood $f(y \mid \beta, b)f(b \mid \Sigma)$.

Bayesian estimation for GLMMs

Bayesian estimation in GLMMs, as in LMMs, is generally based on the Gibbs sampler. For the GLMM

$$Y_i \overset{\text{ind}}{\sim} F(\cdot \mid \mu), \quad g(\mu) = X\beta + Zb, \quad b \sim N(0, \Sigma_b)$$

with corresponding prior densities $f(\beta)$ and $f(\Sigma_b)$, we obtain the *conditional* posterior distributions

$$\begin{aligned} f(\beta \mid y, \text{rest}) &\propto f(\beta) \prod_i f(y_i \mid g^{-1}(X\beta + Zb)) \\ f(b \mid y, \text{rest}) &\propto \phi(b; 0, \Sigma_b) \prod_i f(y_i \mid g^{-1}(X\beta + Zb)) \\ f(\Sigma_b \mid y, \text{rest}) &\propto \phi(b; 0, \Sigma_b) f(\Sigma_b) \end{aligned}$$

For a conditionally conjugate choice of $f(\Sigma_b)$, $f(\Sigma_b \mid y, \text{rest})$ is straightforward to sample from. The conditionals for β and b are not generally available for direct sampling, but there are a number of ways of modifying the basic approach to account for this.

Toxoplasmosis revisited

Estimates and standard errors obtained by ML (quadrature), Laplace and PQL for the individual-level model

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\mu_i), \quad \log \tfrac{\mu_i}{1-\mu_i} = \beta_0 + b_{0i} + \beta_1 x_{ij}, \quad b_{0i} \overset{\text{ind}}{\sim} N(0, \sigma_b^2)$$

Parameter	Estimate (s.e.)		
	ML	Laplace	PQL
β_0	-0.1343 (1.440)	-0.1384 (1.488)	-0.150 (1.392)
β_1 ($\times 10^6$)	5.930 (745.7)	7.215 (770.2)	-5.711 (721.7)
σ_b	0.5132	0.5209	0.4911
AIC	65.75	65.96	'65.98'

Toxoplasmosis continued

Estimates and standard errors obtained by ML (quadrature), Laplace and PQL for the extended model

$$\log \tfrac{\mu_i}{1-\mu_i} = \beta_0 + b_{0i} + \beta_1 x_{ij} + \beta_1 x_{ij}^2 + \beta_1 x_{ij}^3.$$

Parameter	Estimate (s.e.)		
	ML	Laplace	PQL
β_0	-335.5 (136.6)	-335.0 (136.3)	-330.8 (140.7)
β_1	0.5238 (0.2118)	0.5231 (0.2112)	0.5166 (0.2180)
β_2 $(\times 10^4)$	-2.710 (1.089)	-2.706 (1.086)	-2.674 (1.121)
β_3 $(\times 10^8)$	4.463 (1.857)	4.636 (1.852)	4.583 (1.910)
σ_b	0.4232	0.4171	0.4508
AIC	63.84	63.97	'64.03'

So for this example, a good agreement between the different computational methods. Some evidence for the cubic model over the linear model.

Conditional independence and graphical representations slide 127

The role of conditional independence

In modelling clustered data, the requirement is often (as in the toxoplasmosis example above) to construct a model to incorporate both non-normality and dependence. There are rather few 'off-the shelf' models for dependent observations (and those that do exist, such as the multivariate normal, often require strong assumptions which may be hard to justify in practice).

The 'trick' with GLMMs was to model dependence via a series of **conditionally independent** sub-models for the observations y given the random effects b, with dependence induced by marginalising over the distribution of b.

De Finetti's theorem provides some theoretical justification for modelling dependent random variables as conditionally independent given some unknown parameter (which we here denote by ϕ).

De Finetti's theorem

De Finetti's theorem states (approximately) that any $y_1, \ldots, y_n$ which can be thought of as a finite subset of an **exchangeable** infinite sequence of random variables $y_1, y_2 \ldots$, has a joint density which can be written as

$$f(y) = \int f(\phi) \prod_{i=1}^{n} f(y_i \mid \phi) \mathrm{d}\phi$$

for some $f(\phi)$, $f(y_i \mid \phi)$. Hence the y_i can be modelled as conditionally independent given ϕ.

An *exchangeable* infinite sequence is one for which any finite subsequence has a distribution which is invariant under permutation of the lablels of its components.

We can invoke this as an argument for treating as conditionally independent any set of variables about which our prior belief is symmetric.

Complex stochastic models

In many applications we want to model a multivariate response and/or to incorporate a complex (crossed or hierarchically nested) cluster structure amongst the observations.

The same general approach, splitting the model up into small components, with a potentially rich conditional independence structure linking them facilitates both model construction and understanding, and (potentially) computation.

Conditional independence graphs

An extremely useful tool, for model description, model interpretation, and to assist identifying efficient methods for computation is the **directed acyclic graph (DAG)** representing the model.

Denote by $Y = (Y_1, \ldots, Y_\ell)$ the collection of elements of the model which are considered random (given a probability distribution). Then the model is a (parametric) description of the joint distribution $f(y)$, which we can decompose as

$$f(y) = f(y_1) f(y_1 \mid y_2) \cdots f(y_\ell \mid y_1, \ldots, y_{\ell-1}) = \prod_i f(y_i \mid y_{<i})$$

where $y_{<i} = \{y_1, \ldots, y_{i-1}\}$. Now, for certain orderings of the variables in Y, the model may admit conditional independences, exhibited through $f(y_\ell \mid y_1, \ldots, y_{\ell-1})$ being functionally free of y_j for one or more $j < i$. This is expressed as

$$Y_i \perp\!\!\!\perp Y_j \mid Y_{<i \setminus j}$$

where $Y_{<i \setminus j} = \{Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_{i-1}\}$.

DAGs

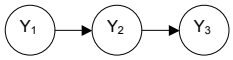
The directed acyclic graph (DAG) representing the probability model, decomposed as

$$f(y) = \prod_i f(y_i \mid y_{<i})$$

consists of a vertex (or node) for each Y_i, together with an directed edge (arrow) **to** each Y_j **from** each $Y_i, i < j$ such that $f(y_j \mid y_{<j})$ depends on y_i. For example, the model

$$f(y_1, y_2, y_3) = f(y_1) f(y_2 \mid y_1) f(y_3 \mid y_2)$$

is represented by the DAG



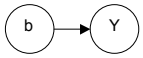
The conditional independence of Y_1 and Y_3 given Y_2 is represented by the absence of a directed edge from Y_1 to Y_3.

DAG for a GLMM

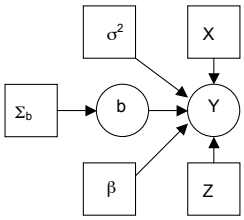
The DAG for the general GLMM

$$Y_i \overset{\text{ind}}{\sim} F(\cdot \mid \mu_i, \sigma^2), \quad g(\mu) = X\beta + Zb, \quad b \sim N(0, \Sigma_b)$$

consists, in its most basic form of two nodes, one for Y and one for b:



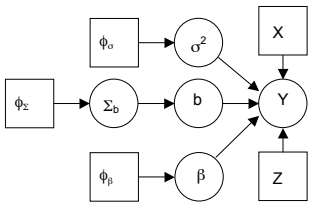
It is generally more informative, to include the model parameters and explanatory data in the DAG. Such fixed (non-stochastic) quantities are often denoted by a different style of vertex in the DAG



DAG for a Bayesian GLMM

A Bayesian model is a full joint probability model, across both the variables treated as stochastic in a classical approach, and any unspecified model parameters. The marginal probability distribution for the parameters represents the prior (to observing data) uncertainty about these quantities.

The appropriate DAG for a Bayesian GLMM reflects this, augmenting the DAG on the previous slide to:



where ϕ_σ, ϕ_Σ and ϕ_β are *hyperparameters* – fixed inputs into the prior distributions for σ^2, σ_b and β respectively.

DAG properties

Suppose we have a DAG representing our model for a collection of random variables $Y = (Y_1, \ldots, Y_\ell)$ where the ordering of the Y_is is chosen such that all edges in the DAG are from lower to higher numbered vertices. [This must be possible for an acyclic graph, but there will generally be more than one possible ordering]. Then the joint distribution for Y factorises as

$$f(y) = \prod_i f(y_i \mid pa[y_i])$$

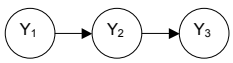
where $pa[y_i]$ represents the subset of $\{y_j, j < i\}$ with edges **to** y_i. Such variables are called the **parents** of y_i.

The local Markov property

A natural consequence of the DAG factorisation of the joint distribution of Y is the **local Markov property for DAGS**. *This states that any variable Y_i is conditionally independent of its non-descendents, given its parents.*

A descendent of Y_i is any variable in $\{Y_j, j > i\}$ which can be reached in the graph by following a sequence of edges **from** Y_i (respecting the direction of the edges).

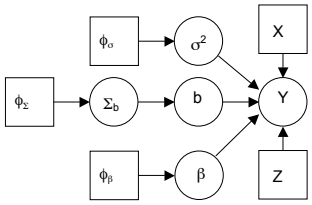
For example, for the simple DAG above



the conditional independence of Y_3 and Y_1 given Y_2 is an immediate consequence of the local Markov property.

The local Markov property – limitations

Not all useful conditional independence properties of DAG models follow immediately from the local Markov property. For example, for the Bayesian GLMM



the posterior distribution is conditional on observed Y, for which the local Markov property is unhelpful, as Y is not a parent of any other variable.

To learn more about conditional independences arising form a DAG, it is necessary to construct the corresponding **undirected conditional independence graph**.

Undirected graphs

An undirected conditional independence graph for Y consists of a vertex for each Y_i, together with a set of undirected edges (lines) between vertices such that **absence** of an edge between two vertices Y_i and Y_j implies the conditional independence

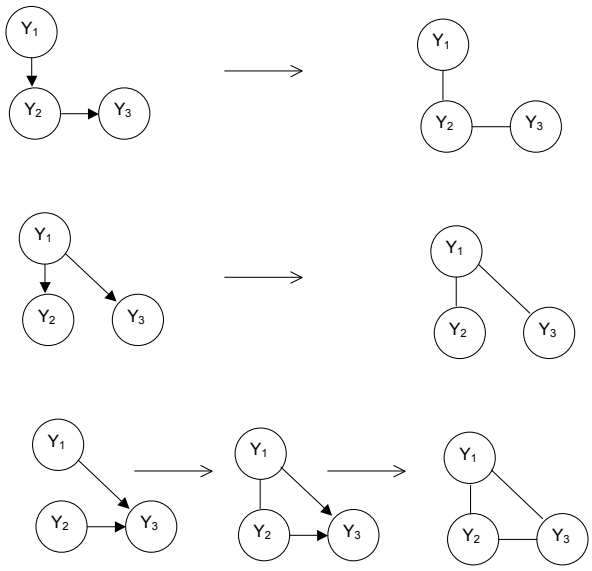
$$Y_i \perp\!\!\!\perp Y_j \mid Y_{\setminus\{i,j\}}$$

where $Y_{\setminus\{i,j\}}$ is the set of varables excluding Y_i and Y_j.

From a DAG, we can obtain the corresponding undirected conditional independence graph via a two stage process

- First we *moralise* the graph by adding an (undirected) edge between ('marrying') any two vertices which have a *child* in common, and which are not already joined by an edge.
- Then we replace all directed edges by undirected edges.

Undirected graphs: examples



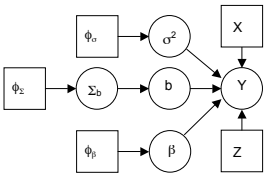
Global Markov property

For an undirected conditional independence graph, the **global Markov property** states that any two variables, Y_i and Y_j say, are conditionally independent given any subset Y_{sep} of the other variables which *separate* Y_i and Y_j in the graph.

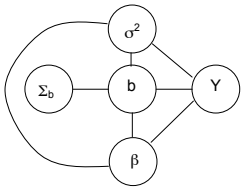
We say that Y_{sep} separates Y_i and Y_j in an undirected graph if any path from Y_i to Y_j via edges in the graph must pass through a variable in Y_{sep}.

Undirected graph for Bayesian GLMM

The DAG for the Bayesian GLMM



has corresponding undirected graph (for the stochastic vertices)



The conditional independence of (β, σ^2) and Σ_b given b (and Y) is immediately obvious.

Markov equivalence

Any moral DAG (one which has no 'unmarried' parents) is **Markov equivalent** to its corresponding undirected graph (i.e. it encodes exactly the same conditional independence structure) .

Conversely, the vertices of any **decomposable** undirected graph (one with no chordless cycles of four or more vertices) can be numbered so that, replacing the undirected edges by directed edges from lower to higher numbered vertices produces a Markov equivalent DAG.

Such a numbering is called a **perfect** numbering for the graph, and is not unique.

It immediately follows that the Markov equivalence classes for DAGs can have (many) more than one member, each of which implies the same model for the data (in terms of conditional independence structure)

The class of DAGs is clearly much larger than the class of undirected graphs, and encompasses a richer range of conditional independence structures.

A genuinely complex model

In the APTS lecture, a practical example from the recent literature will be briefly discussed.

3. Missing Data and Latent Variables: slide 144

Overview

1. Missing data
2. Latent variables
3. EM algorithm

APTS: Statistical Modelling April 2008 – slide 145

Missing Data slide 146

Example 1: Birthweight and smoking

Data from the Collaborative Perinatal Project

		Birth weight (known)		Birth weight (unknown)	
		< 2500	≥ 2500	< 2500	≥ 2500
Mother smokes? (known)	Y	4512	21009		1049
	N	3394	24132		1135
Mother smokes? (unknown)	Y				
	N	142	464		1224

APTS: Statistical Modelling April 2008 – slide 147

Example 2: Political opinions

Data extracted from the British General Election Panel Survey

Sex	Social class	Intention known				Intention unknown			
		Con.	Lab.	Lib.	Other	Con.	Lab.	Lib.	Other
M	1	26	8	7	0				11
	2	87	37	30	6				64
	3	66	77	23	8				77
	4	14	25	15	1				12
	5	6	6	2	0				7
F	1	1	1	0	1				2
	2	63	34	32	2				68
	3	102	52	22	4				77
	4	10	32	10	2				38
	5	20	25	8	2				19

1=Professional, 2=Managerial and technical, 3= Skilled, 4=Semi-skilled or unskilled, 5=Never worked.

APTS: Statistical Modelling April 2008 – slide 148

Introduction

Missing data arises in many practical applications. Typically, our data might appear (with missing data indicated by $*$) as

Unit (i)	Variable (j)				
	1	2	3	$\cdots$	p
1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1p}
2	y_{21}	$*$	y_{23}	$\cdots$	y_{2p}
3	$*$	y_{32}	$*$	$\cdots$	$*$
4	$*$	y_{42}	y_{43}	$\cdots$	y_{4p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	$*$	y_{n3}	$\cdots$	y_{np}

Variables which have missing obervations in our data frame are said to be subject to **item nonresponse**. When there are units with no available data whatsoever, that is referred to as **unit nonresponse**.

If the variables can be ordered so that, for any unit i, (y_{ij} is missing) $\Rightarrow$ (y_{ik} is missing for all $k > j$), the missing data pattern is said to be **monotone** (e.g. longitudinal dropout, special cases like Example 2).

Issues

Missing data creates two major problems for analysis.

- ☐ Suppose that we have a model $f(y \mid \theta)$, for which the likelihood is tractable. When certain y_{ij} are missing, the likelihood for inference must be based on the observed data distribution

$$f(y_{\text{obs}} \mid \theta) = \int f(y_{\text{obs}}, y_{\text{mis}} \mid \theta) \mathrm{d}y_{\text{mis}} \tag{22}$$

  where the subscripts obs and mis refer to observed and missing components, respectively. It is typically much more difficult to compute (22) than $f(y \mid \theta)$ for fully observed data.

- ☐ Even when it can be computed, the likelihood (22) is only valid for inference about θ under the assumption that the fact that certain observations are missing provides no information about θ.

Models

To formalise this, it is helpful to introduce a series of binary response indicator variables $r_1, \ldots, r_p$, where

$$r_{ij} = 1 \Leftrightarrow y_{ij} \text{ is observed}, \quad i = 1, \ldots, n; \; j = 1, \ldots, p.$$

We factorise the joint distribution of (y, r) into a data model for y and a (conditional) response model for r

$$f(y, r \mid \theta, \phi) = f(y \mid \theta) f(r \mid y, \phi).$$

Then the likelihood for the observed data, (y_{obs}, r) is

$$f(y_{\text{obs}}, r \mid \theta, \phi) = \int f(y_{\text{obs}}, y_{\text{mis}} \mid \theta) f(r \mid y_{\text{obs}}, y_{\text{mis}}, \phi) \mathrm{d}y_{\text{mis}}. \tag{23}$$

In this set-up inference for θ should be based on (23), but there are situations when it is valid to **ignore** the missing data mechanism (and the corresponding variable r) and base inference for θ on the simpler $f(y_{\text{obs}} \mid \theta)$.

Ignorability

If $R \perp\!\!\!\perp Y_{\text{mis}} \mid Y_{\text{obs}}, \phi$ then $f(r \mid y_{\text{obs}}, y_{\text{mis}}, \phi)$ in (23) can be replaced by $f(r \mid y_{\text{obs}}, \phi)$, and (23) is simplified to:

$$f(y_{\text{obs}}, r \mid \theta, \phi) = f(y_{\text{obs}} \mid \theta) f(r \mid y_{\text{obs}}, \phi). \tag{24}$$

Hence, the likelihood for (θ, ϕ) factorises and *provided that θ and ϕ are independent* (in a functional sense for likelihood analysis, and in the usual stochastic sense for Bayesian analysis) inference for θ can be based on $f(y_{\text{obs}} \mid \theta)$.

- ☐ Any missing data model which satisfies the two requirements above, namely $[R \perp\!\!\!\perp Y_{\text{mis}} \mid Y_{\text{obs}}, \phi]$ and [ϕ independent of θ] is said to be **ignorable**. Otherwise, it is nonignorable.
- ☐ Missing data which satisfies $R \perp\!\!\!\perp Y_{\text{mis}} \mid Y_{\text{obs}}, \phi$ is said to be **missing at random (MAR)**
- ☐ Missing data which satisfies the stronger condition $R \perp\!\!\!\perp Y_{\text{mis}}, Y_{\text{obs}} \mid \phi$ is said to be **missing completely at random (MCAR)**. For MCAR data, correct (but potentially highly sub-optimal) inferences can be obtained by *complete case analysis*.

Inference under ignorability

For monotone missing data patterns, it may be possible to deal with $f(y_{\text{obs}} \mid \theta)$ directly.

For example, suppose that the $Y_i = (Y_{i1}, \ldots, Y_{ip})$ are conditionally independent given θ, and furthemore that

$$f(y_i \mid \theta) = \prod_j f(y_{ij} \mid y_{i,<j}, \theta_j)$$

where $\theta = (\theta_1, \ldots, \theta_p)$ is a partition into distinct components. Then

$$f(y_{\text{obs}} \mid \theta) = \prod_i f(y_{i,\text{obs}} \mid \theta) = \prod_i \prod_{j=1}^{k_i} f(y_{ij} \mid y_{i,<j}, \theta_j)$$

where k_i is the 'last' observed variable for unit i. Hence the likelihood for θ factorises into individual components.

Otherwise, methods for inference in the presence of an ignorable missing data mechanism typically exploit the fact the full data analysis, based on $f(y \mid \theta)$ is tractable (assuming that it is!)

Gibbs sampler

For Bayesian analysis, this typically involves generating a sequence of values $\{\theta^t, y^t_{\text{mis}}, t = 1, \ldots\}$ from the joint posterior distribution $f(\theta, y_{\text{mis}} \mid y_{\text{obs}})$ using a Gibbs sampler iteratively sampling

- ☐ the model-based conditional for $Y_{\text{mis}} \mid \theta, y_{\text{obs}}$.
- ☐ the complete data posterior conditional for $\theta \mid Y_{\text{mis}}, y_{\text{obs}}$

Often, both of these are convenient for sampling.

The subsample $\{\theta^t, t = 1, \ldots\}$ may then be considered as being drawn from the marginal posterior for $\theta \mid y_{\text{obs}}$, as required.

This is sometimes referred to as *data augmentation*.

EM algorithm

For maximum likelihood, it is often the case that a corresponding iterative algorithm can be constructed by taking the Gibbs sampler steps above and replacing generation from conditionals with (i) taking expectation (for Y_{mis}) and (ii) likelihood maximisation (for θ), respectively.

- ☐ for the current θ^t construct the expected log-likelihood $\mathrm{E}[\log f(Y_{\text{mis}}, y_{\text{obs}} \mid \theta) \mid y_{\text{obs}}, \theta^t]$
- ☐ maximise this expected log-likelihood w.r.t. θ to obtain θ^{t+1}

This is the EM algorithm, of which more details will be presented shortly. The maximisation (M) step is generally straightforward, and for many models, so is the expectation (E) step.

For examples of both the Gibbs sampler and EM algorithm applied to Example 1, see Practical 3.

Nonignorable models

If considered appropriate, then a nonignorable missing data mechanism can be incorporated in $f(y, r \mid \theta, \phi)$. A **selection model** utilises the decomposition

$$f(y, r \mid \theta, \phi) = f(y \mid \theta) f(r \mid y, \phi).$$

where a nonignorable model incorporates dependence of R on Y_{mis}.

Alternatively, a **pattern mixture model** decomposes $f(y, r \mid \theta, \phi)$ as

$$f(y, r \mid \theta, \phi) = f(y \mid r, \theta) f(r \mid \phi).$$

Pattern mixture models tend to be less intuitively appealing, but may be easier to analyse (particularly for monotone missing data patterns).

Under either specification, inference must be based on the observed data likelihood

$$f(y_{\text{obs}}, r \mid \theta, \phi) = \int f(y_{\text{obs}}, y_{\text{mis}}, r \mid \theta, \phi) \mathrm{d}y_{\text{mis}}.$$

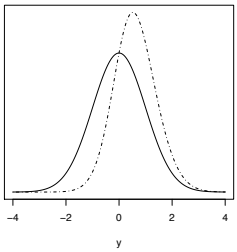
Gibbs sampling or EM can be used for computation, but convergence may be slow.

A simple selection model

Consider the selection model

$$Y \sim N(\theta_1, \theta_2), \qquad P(R = 1 \mid Y = y) = \frac{\exp(\phi_0 + \phi_1 y)}{1 + \exp(\phi_0 + \phi_1 y)}$$

An example of $f(y \mid r = 1)$, the marginal density for y_{obs} is



for $(\theta_1, \theta_2, \phi_0, \phi_1) = (0, 1, 0, 2)$.

The **selection effect** is quite subtle and will clearly be hard to estimate accurately.

Nonignorable model issues

In the previous example, it will be impossible to distinguish, on the basis of observed data only, between the proposed selection model, and an ignorable model where the population distribution of y is naturally slightly skewed.

Generally, nonignorable model inferences are sensitive to model assumptions, and there exist alternative models which cannot be effectively compared on the basis of fit to observed data alone.

Furthermore, inferences from alternative, equally well-fitting models may be very different, as the following (artificial) example illustrates.

	y_2 (Observed)		y_2 (Missing)	
y_1	A	B	A	B
1	6	18		16
2	3	9		8
3	3	27		10

Sensitivity example

Missing data estimates based on the ignorable model $R_2 \perp\!\!\!\perp Y_2 \mid Y_1$

	y_2 (Observed)		y_2 (Missing)	
y_1	A	B	A	B
1	6	18	**4**	**12**
2	3	9	**2**	**6**
3	3	27	**1**	**9**

Missing data estimates based on the nonignorable model $R_2 \perp\!\!\!\perp Y_1 \mid Y_2$

	y_2 (Observed)		y_2 (Missing)	
y_1	A	B	A	B
1	6	18	**14**	**2**
2	3	9	**7**	**1**
3	3	27	**7**	**3**

Potentially very different inferences for the marginal distribution of y_2.

Pragmatic approaches are based on investigating sensitivity to a range of missing data assumptions.

Latent Variables

Basic idea

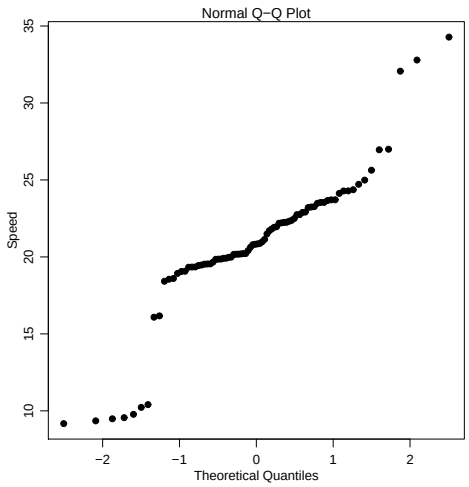
- Many statistical models simplify when written in terms of unobserved **latent variable** U in addition to the observed data Y. The latent variable
  - may really exist, for example, when $Y = I(U > c)$ for some continuous U ('do you earn less than $£c$ per year?');
  - may be imaginary—something called IQ is said to underlie scores on intelligence tests, but is IQ just a cultural construct? ("Mismeasure of man" debate ...);
  - may just be a mathematical/computational device (e.g. in MCMC or EM algorithms).
- Examples include random effects models, use of hidden variables in probit regression, mixture models.

Galaxy data

Velocities (km/second) of 82 galaxies in a survey of the Corona Borealis region. The error is thought to be less than 50 km/second.

9172	9350	9483	9558	9775	10227	10406	16084	16170	18419
18552	18600	18927	19052	19070	19330	19343	19349	19440	19473
19529	19541	19547	19663	19846	19856	19863	19914	19918	19973
19989	20166	20175	20179	20196	20215	20221	20415	20629	20795
20821	20846	20875	20986	21137	21492	21701	21814	21921	21960
22185	22209	22242	22249	22314	22374	22495	22746	22747	22888
22914	23206	23241	23263	23484	23538	23542	23666	23706	23711
24129	24285	24289	24366	24717	24990	25633	26960	26995	32065
32789	34279								

Galaxy data



Mixture density

- Natural model for such data is a p-component mixture density

$$f(y;\theta) = \sum_{r=1}^{p} \pi_r f_r(y;\theta), \quad 0 \le \pi_r \le 1, \sum_{r=1}^{p} \pi_r = 1,$$

  where π_r is the probability that Y comes from the $r\mathrm{th}$ component and $f_r(y;\theta)$ is its density conditional on this event.
- Can represent this using indicator variables U taking a value in $1,\ldots,p$ with probabilities $\pi_1,\ldots,\pi_p$ and indicating from which component Y is drawn.
- Widely used class of models, often with number of components p unknown.
- Aside: such models are non-regular for likelihood inference:
  - non-identifiable under permutation of components;
  - setting $\pi_r = 0$ eliminates parameters of f_r;
  - maximum of likelihood can be $+\infty$, achieved for several θ

Other latent variable models

- Let $[U], D$ denote discrete random variables, and $(U), X$ continuous ones. Then in notation for graphical models:
  - $[U] \to X$ or $[U] \to D$ denotes finite mixture models, hidden Markov models, changepoint models, etc.;
  - $(U) \to D$ denotes data coarsening (censoring, truncation, ...);
  - $(U) \to X$ or $(U) \to D$ denotes variance components and other hierarchical models.
- Binary regression: $U \sim \mathcal{N}(x^{\mathrm{T}}\beta, 1)$ and observed response $Y = I(U \ge 0)$, gives probit regression model, log likelihood contribution

$$Y \log \Phi(x^{\mathrm{T}}\beta) + (1-Y)\log\{1 - \Phi(x^{\mathrm{T}}\beta)\},$$

  and similarly if different continuous distribution is chosen for U (logistic, extreme-value, ...).

EM algorithm

- Aim to use observed value y of Y for inference on θ when we cannot easily compute

$$f(y;\theta) = \int f(y \mid u;\theta) f(u;\theta)\, du$$

- The **complete-data log likelihood**

$$\log f(y,u;\theta) = \log f(y;\theta) + \log f(u \mid y;\theta), \tag{25}$$

is based on (U,Y), whereas the **observed-data log likelihood** is

$$\ell(\theta) = \log f(y;\theta).$$

- Take expectation in (25) with respect to $f(u \mid y;\theta')$ to get

$$\mathrm{E}\left\{\log f(Y,U;\theta) \mid Y=y;\theta'\right\} = \ell(\theta) + \mathrm{E}\left\{\log f(U \mid Y;\theta) \mid Y=y;\theta'\right\}, \tag{26}$$

or equivalently $Q(\theta;\theta') = \ell(\theta) + C(\theta;\theta')$.

EM algorithm II

Fix θ' and consider how $Q(\theta;\theta')$ and $C(\theta;\theta')$ depend on θ.

- Note that $C(\theta';\theta') \geq C(\theta;\theta')$, with equality only when $\theta = \theta'$ (Jensen's inequality).
- Thus

$$Q(\theta;\theta') \geq Q(\theta';\theta') \text{ implies } \ell(\theta) - \ell(\theta') \geq C(\theta';\theta') - C(\theta;\theta') \geq 0. \tag{27}$$

- Under mild smoothness conditions, $C(\theta;\theta')$ has a stationary point at $\theta = \theta'$, so if $Q(\theta;\theta')$ is stationary at $\theta = \theta'$, so too is $\ell(\theta)$.
- Hence **EM algorithm**: starting from an initial value θ' of θ,
  1. compute $Q(\theta;\theta') = \mathrm{E}\{\log f(Y,U;\theta) \mid Y=y;\theta'\}$; then
  2. with θ' fixed, maximize $Q(\theta;\theta')$ over θ, giving $\theta^\dagger$, say; and
  3. check if the algorithm has converged, using $\ell(\theta^\dagger) - \ell(\theta')$ if available, or $|\theta^\dagger - \theta'|$, or both. If not, set $\theta' = \theta^\dagger$ and go to 1.

  Steps 1 and 2 are the **expectation (E)** and **maximization (M)** steps.
- The M-step ensures that $Q(\theta^\dagger;\theta') \geq Q(\theta';\theta')$, so (27) implies that $\ell(\theta^\dagger) \geq \ell(\theta')$: the log likelihood never decreases.

Convergence

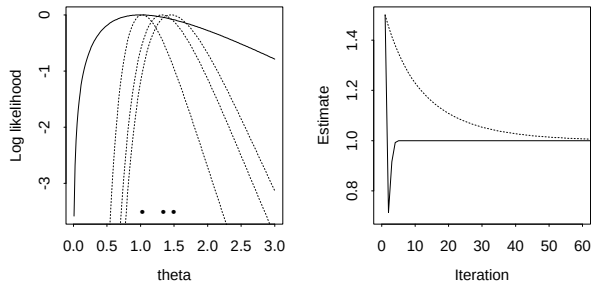
- If $\ell(\theta)$ has
  - only one stationary point, and if $Q(\theta;\theta')$ eventually reaches a stationary value at $\widehat{\theta}$, then $\widehat{\theta}$ must maximize $\ell(\theta)$;
  - otherwise the algorithm may converge to a local maximum of the log likelihood or to a turning point.
- The EM algorithm never decreases the log likelihood so is more stable than Newton–Raphson-type algorithms.
- Rate of convergence depends on closeness of $Q(\theta;\theta')$ and $\ell(\theta)$:

$$-\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} = \mathrm{E}\left\{-\frac{\partial^2 \log f(y,U;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\middle| Y=y;\theta\right\} - \mathrm{E}\left\{-\frac{\partial^2 \log f(U \mid y;\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\middle| Y=y;\theta\right\},$$

  or $J(\theta) = I_c(\theta;y) - I_m(\theta;y)$, giving the **missing information principle**: the observed information equals the complete-data information minus the missing information.
- Rate of convergence slow if largest eigenvalue of $I_c(\theta;y)^{-1}I_m(\theta;y) \approx 1$; this occurs if the missing information is a high proportion of the total.

(Toy) Example: Negative binomial model

Conditional on $U = u$, $Y \sim \mathrm{Poiss}(u)$ and U is gamma with mean θ and variance θ^2/ν. Suppose $\nu > 0$ known and make inference for θ.



EM algorithm for negative binomial example. Left panel: observed-data log likelihood $\ell(\theta)$ (solid) and functions $Q(\theta;\theta')$ for $\theta' = 1.5$, 1.347 and 1.028 (dots, from right). The blobs show the values of θ that maximize these functions, which correspond to the first, fifth and fortieth iterations of the EM algorithm. Right: convergence of EM algorithm (dots) and Newton–Raphson algorithm (solid). The panel shows how successive EM iterations update θ' and $\widehat{\theta}$. Notice that the EM iterates always increase $\ell(\theta)$, while the Newton–Raphson steps do not.

Note: Negative binomial example

For a toy example, suppose that conditional on $U = u$, Y is a Poisson variable with mean u, and that U is gamma with mean θ and variance θ^2/ν. Inference is required for θ with the shape parameter $\nu > 0$ supposed known. Here (25) equals

$$y \log u - u - \log y! + \nu \log \nu - \nu \log \theta + (\nu - 1) \log u - \nu u/\theta - \log \Gamma(\nu),$$

and hence (26) equals

$$Q(\theta; \theta') = (y + \nu - 1)\mathrm{E}(\log U \mid Y = y; \theta') - (1 + \nu/\theta)\mathrm{E}(U \mid Y = y; \theta') - \nu \log \theta$$

plus terms that depend neither on U nor on θ.
The E-step, computation of $Q(\theta; \theta')$, involves two expectations, but fortunately $\mathrm{E}(\log U \mid Y = y; \theta')$ does not appear in terms that involve θ and so is not required. To compute $\mathrm{E}(U \mid Y = y; \theta')$, note that Y and U have joint density

$$f(y \mid u) f(u; \theta) = \frac{u^y}{y!} e^{-u} \times \frac{\nu^\nu u^{\nu-1}}{\theta^\nu \Gamma(\nu)} e^{-\nu u/\theta}, \quad y = 0, 1, \ldots, \ u > 0, \quad \theta > 0,$$

so the marginal density of Y is

$$f(y; \theta) = \int_0^\infty f(y \mid u) f(u; \theta, \nu)\, du = \frac{\Gamma(y + \nu)\nu^\nu}{\Gamma(\nu) y!} \frac{\theta^y}{(\theta + \nu)^{y+\nu}}, \quad y = 0, 1, \ldots$$

Hence the conditional density$f(u \mid y; \theta')$ is gamma with shape parameter $y + \nu$ and mean $\mathrm{E}(U \mid Y = y; \theta') = (y + \nu)/(1 + \nu/\theta')$, and we can take

$$Q(\theta; \theta') \equiv -(1 + \nu/\theta)(y + \nu)/(1 + \nu/\theta') - \nu \log \theta,$$

where we have ignored terms independent of both θ and θ'.
The M-step involves maximization of $Q(\theta; \theta')$ over θ for fixed θ', so we differentiate with respect to θ and find that the maximizing value is

$$\theta^\dagger = \theta'(y + \nu)/(\theta' + \nu). \tag{28}$$

In this example, therefore, the EM algorithm boils down to choosing an initial θ', updating it to $\theta^\dagger$ using (28), setting $\theta' = \theta^\dagger$ and iterating to convergence.

Example: Mixture model

- Consider earlier p-component mixture density $f(y;\theta) = \sum_{r=1}^{p} \pi_r f_r(y;\theta)$, for which likelihood contribution from (y,u) would be $\prod_r \{f_r(y;\theta)\pi_r\}^{I(u=r)}$, giving contribution

$$\log f(y,u;\theta) = \sum_{r=1}^{p} I(u=r)\{\log \pi_r + \log f_r(y;\theta)\}$$

  to the complete-data log likelihood.

- Must compute the expectation of $\log f(y,u;\theta)$ over

$$w_r(y;\theta') = \Pr(U=r \mid Y=y;\theta') = \frac{\pi_r' f_r(y;\theta')}{\sum_{s=1}^{p} \pi_s' f_s(y;\theta')}, \quad r=1,\ldots,p, \tag{29}$$

  the weight attributable to component r if y has been observed.

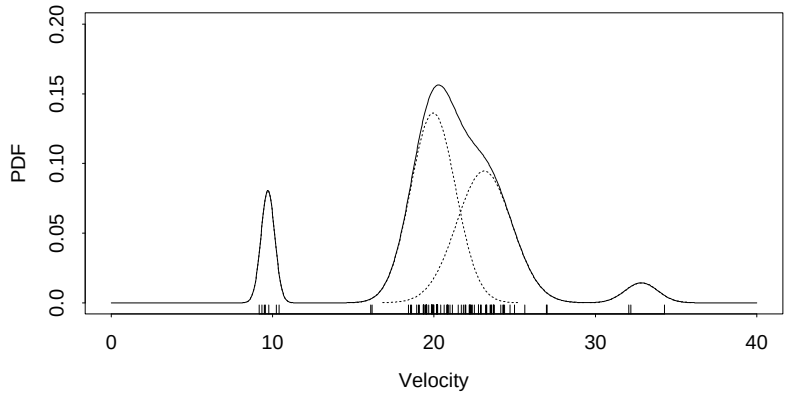
- The expected value of $I(U=r)$ with respect to (29) is $w_r(y;\theta')$, so the expected value of the log likelihood based on a random sample $(y_1,u_1),\ldots,(y_n,u_n)$ is

$$\begin{aligned} Q(\theta;\theta') &= \sum_{j=1}^{n}\sum_{r=1}^{p} w_r(y_j;\theta')\{\log \pi_r + \log f_r(y_j;\theta)\} \\ &= \sum_{r=1}^{p}\left\{\sum_{j=1}^{n} w_r(y_j;\theta')\right\}\log \pi_r + \sum_{r=1}^{p}\sum_{j=1}^{n} w_r(y_j;\theta')\log f_r(y_j;\theta). \end{aligned}$$

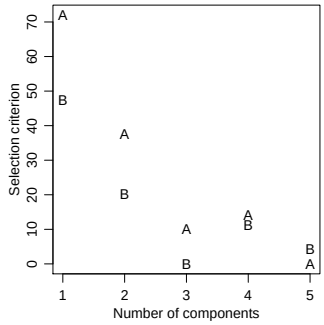
Example: Galaxy data

p	1	2	3	4	5
$\widehat{\ell}$	-240.42	-220.19	-203.48	-202.52	-192.42

Fitted mixture model with $p=4$ normal components:



Galaxy data



AIC and BIC for the normal mixture models fitted to the galaxy data. BIC is minimised for $p = 3$ components, and AIC for $p = 5$ components.

Note: Mixture model

Mixture models arise when an observation Y is taken from a population composed of distinct subpopulations, but it is unknown from which of these Y is taken. If the number p of subpopulations is finite, Y has a p-component mixture density

$$f(y;\theta) = \sum_{r=1}^{p} \pi_r f_r(y;\theta), \quad 0 \le \pi_r \le 1, \sum_{r=1}^{p} \pi_r = 1,$$

where π_r is the probability that Y comes from the $r\text{th}$ subpopulation and $f_r(y;\theta)$ is its density conditional on this event. An indicator U of the subpopulation from which Y arises takes values $1,\ldots,p$ with probabilities $\pi_1,\ldots,\pi_p$. In many applications the components have a physical meaning, but sometimes a mixture is used simply as a flexible class of densities. For simplicity of notation below, let θ contain all unknown parameters including the π_r.
If the value u of U were known, the likelihood contribution from (y,u) would be $\prod_r \{f_r(y;\theta)\pi_r\}^{I(u=r)}$, giving contribution

$$\log f(y,u;\theta) = \sum_{r=1}^{p} I(u=r)\{\log \pi_r + \log f_r(y;\theta)\}$$

to the complete-data log likelihood. In order to apply the EM algorithm we must compute the expectation of $\log f(y,u;\theta)$ over the conditional distribution

$$\Pr(U=r \mid Y=y;\theta') = \frac{\pi'_r f_r(y;\theta')}{\sum_{s=1}^{p} \pi'_s f_s(y;\theta')}, \quad r=1,\ldots,p. \tag{30}$$

This probability can be regarded as the weight attributable to component r if y has been observed; for compactness below we denote it by $w_r(y;\theta')$. The expected value of $I(U=r)$ with respect to (29) is $w_r(y;\theta')$, so the expected value of the log likelihood based on a random sample $(y_1,u_1),\ldots,(y_n,u_n)$ is

$$\begin{aligned} Q(\theta;\theta') &= \sum_{j=1}^{n}\sum_{r=1}^{p} w_r(y_j;\theta')\{\log \pi_r + \log f_r(y_j;\theta)\} \\ &= \sum_{r=1}^{p}\left\{\sum_{j=1}^{n} w_r(y_j;\theta')\right\}\log \pi_r + \sum_{r=1}^{p}\sum_{j=1}^{n} w_r(y_j;\theta')\log f_r(y_j;\theta). \end{aligned}$$

The M step of the algorithm entails maximizing $Q(\theta;\theta')$ over θ for fixed θ'. As the π_r do not usually appear in the component density f_r, the maximizing values $\pi_r^\dagger$ are obtained from the first term of Q, which corresponds to a multinomial log likelihood. Thus $\pi_r^\dagger = n^{-1}\sum_j w_r(y_j;\theta')$, the average weight for component r.
Estimates of the parameters of the f_r are obtained from the weighted log likelihoods that form the second term of $Q(\theta;\theta')$. For example, if f_r is normal with mean μ_r and variance σ_r^2, simple calculations give the weighted estimates

$$\mu_r^\dagger = \frac{\sum_{j=1}^{n} w_r(y_j;\theta')y_j}{\sum_{j=1}^{n} w_r(y_j;\theta')} \quad \sigma_r^{2\dagger} = \frac{\sum_{j=1}^{n} w_r(y_j;\theta')(y_j-\mu_r^\dagger)^2}{\sum_{j=1}^{n} w_r(y_j;\theta')}, \quad r=1,\ldots,p.$$

Given initial values of $(\pi_r,\mu_r,\sigma_r^2) \equiv \theta'$, the EM algorithm simply involves computing the weights $w_r(y_j;\theta')$ for these initial values, updating to obtain $(\pi_r^\dagger,\mu_r^\dagger,\sigma_r^{2\dagger}) \equiv \theta^\dagger$, and checking convergence using the log likelihood, $|\theta^\dagger - \theta'|$, or both. If convergence is not yet attained, θ' is replaced by $\theta^\dagger$ and the cycle repeated.

Note: Galaxy data

We illustrate these calculations using the data above on the velocities at which 82 galaxies in the Corona Borealis region are moving away from our own galaxy. It is thought that after the Big Bang the universe expanded very fast, and that as it did so galaxies formed because of the local attraction of matter. Owing to the action of gravity they tend to cluster together, but there seem also to be 'superclusters' of galaxies surrounded by voids. If galaxies are indeed super-clustered the distribution of their velocities estimated from the red-shift in their light-spectra would be multimodal, and unimodal otherwise. The data given are from sections of the northern sky carefully sampled to settle whether there are superclusters.

Cursory examination of the data strongly suggests clustering. In order to estimate the number of clusters we fit mixtures of normal densities by the EM algorithm with initial values chosen by eye. The maximized log likelihood for $p = 2$ is -220.19, found after 26 iterations. In fact this is the highest of several local maxima; the global maximum of $+\infty$ is found by centering one component of the mixture at any of the y_j and letting the corresponding $\sigma_r^2 \to \infty$. Only the local maxima yield sensible fits, the best of which is found using randomly chosen initial values. The number of iterations needed depends on these and on the number of components, but is typically less than 40. This procedure gives maximized log likelihoods -240.42, -203.48, -202.52 and -192.42 for fits with $p = 1$, 3, 4 and 5. The latter gives a single component to the two observations around 16,000 and so does not seem very sensible. Standard likelihood asymptotics do not apply here, but evidently there is little difference between the 3- and 4-component fits, the second of which is shown in the figure. Both fits have three modes, and the evidence for clustering is very strong.

An alternative is to apply a Newton–Raphson algorithm directly to the log likelihood $\ell(\theta)$ based on the mixture density, but if this is to be reliable the model must be reparametrized so that the parameter space is unconstrained. The effect of the spikes in $\ell(\theta)$ can be reduced by replacing $f_r(y;\theta)$ by $F_r(y+h;\theta) - F_r(y-h;\theta)$, where h is the degree of rounding of the data, here 50 km/second.

Exponential family

- Suppose the complete-data log likelihood is from an exponential family:

$$\log f(y, u; \theta) = s(y, u)^{\mathrm{T}}\theta - \kappa(\theta) + c(y, u).$$

- For EM algorithm, need expected value of $\log f(y, u; \theta)$ with respect to $f(u \mid y; \theta')$. Final term can be ignored, so M-step involves maximizing

$$Q(\theta; \theta') = \mathrm{E}\left\{s(y, U)^{\mathrm{T}}\theta \mid Y = y; \theta'\right\} - \kappa(\theta),$$

  or equivalently solving for θ the equation

$$\mathrm{E}\left\{s(y, U) \mid Y = y; \theta'\right\} = \frac{d\kappa(\theta)}{d\theta}.$$

- Likelihood equation for θ based on the complete data is $s(y, u) = d\kappa(\theta)/d\theta$, so the EM algorithm replaces $s(y, u)$ by its conditional expectation $\mathrm{E}\{s(y, U) \mid Y = y; \theta'\}$ and solves the likelihood equation. Thus a routine to fit the complete-data model can readily be adapted for missing data if the conditional expectations are available.

Comments

- Often E-step requires numerical approximation:
  - simulation from conditional distribution of U given Y;
  - importance sampling;
  - Markov chain algorithm;
- M-step can be performed using Newton–Raphson or similar algorithm, using first and second loglikelihood derivatives (exercise)—may need to be performed in parts, rather than overall
- Can obtain standard errors using these derivatives (exercise)
- In Bayesian analysis, may often be helpful to include latent variables, either
  - because they have useful interpretation in terms of model—*all* parameters are hidden variables, because unobservable in practice
  - to simplify MCMC algorithm—Gibbs sampler is 'Bayesian equivalent' of EM algorithm (exercise)