Regression and variable selection

Regression

Large fraction of statistical analyses dealing with representation of dependences between several variables, rather than marginal distribution of each variable

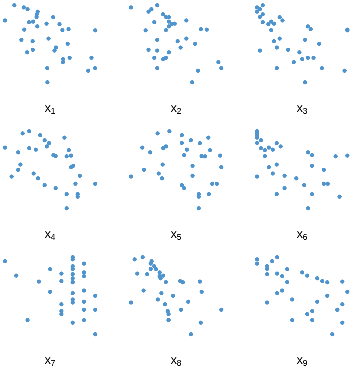
Pine processionary caterpillars

Pine processionary caterpillars

Pine processionary caterpillar colony size influenced by

- x_1 altitude
- x_2 slope (in degrees)
- x_3 number of pines in the area
- x_4 height of the central tree
- x_5 diameter of the central tree
- x_6 index of the settlement density
- x_7 orientation of the area (from 1 [southbound] to 2)
- x_8 height of the dominant tree
- x_9 number of vegetation strata
- x_{10} mix settlement index (from 1 if not mixed to 2 if mixed)

Pine processionary caterpillars



Goal of a regression model

From a statistical point of view, find a proper representation of the distribution, $f(y|\theta, x)$, of an observable variable y given a vector of observables x, based on a sample of $(x, y)_i$'s.

Linear regression

Linear regression: one of the most widespread tools of Statistics for analysing (linear) influence of some variables or some factors on others

Linear regression

Linear regression: one of the most widespread tools of Statistics for analysing (linear) influence of some variables or some factors on others

Aim

To uncover explanatory and predictive patterns

Regressors and response

Variable of primary interest, y, called the *response* or the *outcome* variable [assumed here to be continuous]

E.g., number of Pine processionary caterpillar colonies

Regressors and response

Variable of primary interest, y, called the *response* or the *outcome* variable [assumed here to be continuous]

E.g., number of Pine processionary caterpillar colonies

Covariates $x = (x_1, \ldots, x_k)$ called *explanatory variables* [may be discrete, continuous or both]

Regressors and response

Variable of primary interest, y, called the *response* or the *outcome* variable [assumed here to be continuous]

E.g., number of Pine processionary caterpillar colonies

Covariates $x = (x_1, \ldots, x_k)$ called *explanatory variables* [may be discrete, continuous or both]

Distribution of y given x typically studied in the context of a set of *units* or experimental *subjects*, $i = 1, \ldots, n$, for instance patients in an hospital ward, on which both y_i and $x_{i1}, \ldots, x_{ik}$ are measured.

Regressors and response cont'd

Dataset made of the conjunction of the vector of outcomes

$$y = (y_1, \ldots, y_n)$$

Regressors and response cont'd

Dataset made of the conjunction of the vector of outcomes

$$y = (y_1, \dots, y_n)$$

and of the $n \times (k+1)$ matrix of explanatory variables

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & x_{31} & x_{32} & \dots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

Linear models

Ordinary normal linear regression model such that

$$y|\beta, \sigma^2, X \sim \mathscr{N}_n(X\beta, \sigma^2 I_n)$$

Linear models

Ordinary normal linear regression model such that

$$y|\beta,\sigma^2,X \sim \mathscr{N}_n(X\beta,\sigma^2 I_n)$$

and thus

$$\begin{aligned}\mathbb{E}[y_i|\beta,X] &= \beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}\\ \mathbb{V}(y_i|\sigma^2,X) &= \sigma^2\end{aligned}$$

Categorical variables

- There is a difference between finite valued regressors like x_7 in caterpillar [orientation of the area] and *categorical* variables (or *factors*), which are also taking a finite number of values but whose range has no numerical meaning.

Categorical variables

- There is a difference between finite valued regressors like x_7 in caterpillar [orientation of the area] and *categorical* variables (or *factors*), which are also taking a finite number of values but whose range has no numerical meaning.

Example

If x is the socio-professional category of an employee, this variable ranges from 1 to 9 for a rough grid of socio-professional activities, and from 1 to 89 on a finer grid.

The numerical values are not comparable

Categorical variables (cont'd)

Makes little sense to involve x directly in the regression: replace the single regressor x [in $\{1, \ldots, m\}$, say] with m indicator (or *dummy*) variables

$$x_1 = \mathbb{I}_1(x), \ldots, x_m = \mathbb{I}_m(x)$$

Categorical variables (cont'd)

Makes little sense to involve x directly in the regression: replace the single regressor x [in $\{1,\ldots,m\}$, say] with m indicator (or *dummy*) variables

$$x_1 = \mathbb{I}_1(x), \ldots, x_m = \mathbb{I}_m(x)$$

Convention

Use of a different constant β_i for each class categorical variable value:

$$\mathbb{E}[y_i|\beta, X] = \ldots + \beta_1 \mathbb{I}_1(x) + \ldots + \beta_m \mathbb{I}_m(x) + \ldots$$

Identifiability

Identifiability issue: For dummy variables, sum of the indicators equal to one.

Convention

Assume that X is of full rank:

$$\text{rank}(X) = k + 1$$

[X is of full rank if and only if $X^{\mathrm{T}}X$ is invertible]

Identifiability

Identifiability issue: For dummy variables, sum of the indicators equal to one.

Convention

Assume that X is of full rank:

$$\text{rank}(X) = k + 1$$

[X is of full rank if and only if $X^{\mathrm{T}}X$ is invertible]

E.g., for dummy variables, this means eliminating one class

Likelihood function & estimator

The likelihood of the *ordinary normal linear model* is

$$\ell\left(\beta,\sigma^2|y,X\right)=\left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}(y-X\beta)^{\mathrm{T}}(y-X\beta)\right]$$

Likelihood function & estimator

The likelihood of the *ordinary normal linear model* is

$$\ell\left(\beta,\sigma^2|y,X\right)=\left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}(y-X\beta)^{\mathrm{T}}(y-X\beta)\right]$$

The MLE of β is solution of the least squares minimisation problem

$$\min_{\beta}\,(y-X\beta)^{\mathrm{T}}(y-X\beta)=\min_{\beta}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_{i1}-\ldots-\beta_kx_{ik}\right)^2\ ,$$

namely

$$\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

Least square estimator

- $\hat\beta$ is an unbiased estimator of β.
- $\mathbb{V}(\hat\beta|\sigma^2, X) = \sigma^2(X^{\mathrm{T}}X)^{-1}$
- $\hat\beta$ is the *best* linear unbiased estimator of β: for all $a \in \mathbb{R}^{k+1}$,

$$\mathbb{V}(a^{\mathrm{T}}\hat\beta|\sigma^2, X) \leq \mathbb{V}(a^{\mathrm{T}}\tilde\beta|\sigma^2, X)$$

  for any unbiased linear estimator $\tilde\beta$ of β.
- Unbiased estimator of σ^2

$$\hat\sigma^2 = \frac{1}{n-k-1}(y - X\hat\beta)^{\mathrm{T}}(y - X\hat\beta) = \frac{s^2}{n-k-1},$$

Pine processionary caterpillars

```
Residuals:      Min       1Q    Median      3Q       Max
             -1.6989  -0.2731  -0.0003  0.3246    1.7305
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
intercept   10.998412   3.060272   3.594  0.00161 **
XV1         -0.004431   0.001557  -2.846  0.00939 **
XV2         -0.053830   0.021900  -2.458  0.02232 *
XV3          0.067939   0.099472   0.683  0.50174
XV4         -1.293636   0.563811  -2.294  0.03168 *
XV5          0.231637   0.104378   2.219  0.03709 *
XV6         -0.356800   1.566464  -0.228  0.82193
XV7         -0.237469   1.006006  -0.236  0.81558
XV8          0.181060   0.236724   0.765  0.45248
XV9         -1.285316   0.864847  -1.486  0.15142
XV10        -0.433106   0.734869  -0.589  0.56162
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conjugate priors

If [conditional prior]

$$\beta|\sigma^2, X \sim \mathscr{N}_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}),$$

where M $(k+1, k+1)$ positive definite symmetric matrix, and

$$\sigma^2|X \sim \mathscr{I}\mathscr{G}(a, b), \qquad a, b > 0,$$

Conjugate priors

If [conditional prior]

$$\beta|\sigma^2, X \sim \mathscr{N}_{k+1}(\tilde{\beta}, \sigma^2 M^{-1})\,,$$

where M $(k+1, k+1)$ positive definite symmetric matrix, and

$$\sigma^2|X \sim \mathscr{I}\mathscr{G}(a, b), \qquad a, b > 0,$$

then

$$\beta|\sigma^2, \mathbf{y}, X \sim \mathscr{N}_{k+1}\left((M + X^{\mathrm{T}}X)^{-1}\{(X^{\mathrm{T}}X)\hat{\beta} + M\tilde{\beta}\}, \sigma^2(M + X^{\mathrm{T}}X)^{-1}\right)$$

and

$$\sigma^2|\mathbf{y}, X \sim \mathscr{I}\mathscr{G}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\tilde{\beta} - \hat{\beta})^{\mathrm{T}}\left(M^{-1} + (X^{\mathrm{T}}X)^{-1}\right)^{-1}(\tilde{\beta} - \hat{\beta})}{2}\right)$$

Experimenter dilemma

Problem of the choice of M or of c if $M = I_{k+1}/c$

Experimenter dilemma

Problem of the choice of M or of c if $M = I_{k+1}/c$

Example (Processionary caterpillar)

No precise prior information about $\tilde{\beta}$, M, a and b. Take $a = 2.1$ and $b = 2$, i.e. prior mean and prior variance of σ^2 equal to 1.82 and 33.06, and $\tilde{\beta} = 0_{k+1}$.

Experimenter dilemma

Problem of the choice of M or of c if $M = I_{k+1}/c$

Example (Processionary caterpillar)

No precise prior information about $\tilde{\beta}$, M, a and b. Take $a = 2.1$ and $b = 2$, i.e. prior mean and prior variance of σ^2 equal to 1.82 and 33.06, and $\tilde{\beta} = 0_{k+1}$.
Lasting influence of c:

c	$\mathbb{E}^\pi(\sigma^2 \mid \mathbf{y}, X)$	$\mathbb{E}^\pi(\beta_0 \mid \mathbf{y}, X)$	$\mathbb{V}^\pi(\beta_0 \mid \mathbf{y}, X)$
.1	1.0044	0.1251	0.0988
1	0.8541	0.9031	0.7733
10	0.6976	4.7299	3.8991
100	0.5746	9.6626	6.8355
1000	0.5470	10.8476	7.3419

Zellner's informative G-prior

Constraint

Allow the experimenter to introduce information about the location parameter of the regression while bypassing the most difficult aspects of the prior specification, namely the derivation of the prior correlation structure.

Zellner's informative G-prior

Constraint

Allow the experimenter to introduce information about the location parameter of the regression while bypassing the most difficult aspects of the prior specification, namely the derivation of the prior correlation structure.

Zellner's prior corresponds to

$$\begin{aligned}\beta|\sigma^2, X &\sim \mathscr{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^{\mathrm{T}}X)^{-1}) \\ \sigma^2 &\sim \pi(\sigma^2|X) \propto \sigma^{-2}\,.\end{aligned}$$

[Special conjugate]

Prior selection

Experimental prior determination restricted to the choices of $\tilde{\beta}$ and of the constant c.

Note

c can be interpreted as a measure of the amount of information available in the prior relative to the sample. For instance, setting $1/c = 0.5$ gives the prior the same weight as 50% of the sample.

Prior selection

Experimental prior determination restricted to the choices of $\tilde{\beta}$ and of the constant c.

Note

c can be interpreted as a measure of the amount of information available in the prior relative to the sample. For instance, setting $1/c = 0.5$ gives the prior the same weight as 50% of the sample.

↯ There still is a lasting influence of the factor c

Posterior structure

With this prior model, the posterior simplifies into

$$
\begin{aligned}
\pi(\beta,\sigma^2|y,X) &\propto f(y|\beta,\sigma^2,X)\pi(\beta,\sigma^2|X)\\
&\propto \left(\sigma^2\right)^{-(n/2+1)}\exp\left[-\frac{1}{2\sigma^2}(y-X\hat\beta)^{\mathrm{T}}(y-X\hat\beta)\right.\\
&\qquad\left.-\frac{1}{2\sigma^2}(\beta-\hat\beta)^{\mathrm{T}}X^{\mathrm{T}}X(\beta-\hat\beta)\right]\left(\sigma^2\right)^{-k/2}\\
&\qquad\times\exp\left[-\frac{1}{2c\sigma^2}(\beta-\tilde\beta)X^{\mathrm{T}}X(\beta-\tilde\beta)\right],
\end{aligned}
$$

because $X^{\mathrm{T}}X$ used in both prior and likelihood

[G-prior trick]

Posterior structure (cont'd)

Therefore,

$$
\begin{aligned}
\beta|\sigma^2, y, X &\sim \mathscr{N}_{k+1}\left(\frac{c}{c+1}(\tilde{\beta}/c+\hat{\beta}), \frac{\sigma^2 c}{c+1}(X^{\mathrm{T}}X)^{-1}\right) \\
\sigma^2|y, X &\sim \mathcal{IG}\left(\frac{n}{2}, \frac{s^2}{2}+\frac{1}{2(c+1)}(\tilde{\beta}-\hat{\beta})^{\mathrm{T}}X^{\mathrm{T}}X(\tilde{\beta}-\hat{\beta})\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\beta|y, X \sim \mathscr{T}_{k+1}\Bigg(n, \frac{c}{c+1}\left(\frac{\tilde{\beta}}{c}+\hat{\beta}\right), \\
\frac{c(s^2+(\tilde{\beta}-\hat{\beta})^{\mathrm{T}}X^{\mathrm{T}}X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}(X^{\mathrm{T}}X)^{-1}\Bigg).
\end{aligned}
$$

Bayes estimator

The Bayes estimators of β and σ^2 are given by

$$\mathbb{E}^\pi[\beta|y,X] = \frac{1}{c+1}(\tilde\beta + c\hat\beta)$$

and

$$\mathbb{E}^\pi[\sigma^2|y,X] = \frac{s^2 + (\tilde\beta-\hat\beta)^{\mathrm{T}}X^{\mathrm{T}}X(\tilde\beta-\hat\beta)/(c+1)}{n-2}\,.$$

Bayes estimator

The Bayes estimators of β and σ^2 are given by

$$\mathbb{E}^\pi[\beta|y,X] = \frac{1}{c+1}(\tilde{\beta}+c\hat{\beta})$$

and

$$\mathbb{E}^\pi[\sigma^2|y,X] = \frac{s^2+(\tilde{\beta}-\hat{\beta})^{\mathrm{T}}X^{\mathrm{T}}X(\tilde{\beta}-\hat{\beta})/(c+1)}{n-2}\,.$$

Note: Only when c goes to infinity does the influence of the prior vanish!

Pine processionary caterpillars

β_i	$\mathbb{E}^\pi(\beta_i\mid\mathbf{y},X)$	$\mathbb{V}^\pi(\beta_i\mid\mathbf{y},X)$
β_0	10.8895	6.4094
β_1	-0.0044	2e-06
β_2	-0.0533	0.0003
β_3	0.0673	0.0068
β_4	-1.2808	0.2175
β_5	0.2293	0.0075
β_6	-0.3532	1.6793
β_7	-0.2351	0.6926
β_8	0.1793	0.0383
β_9	-1.2726	0.5119
β_{10}	-0.4288	0.3696

$c = 100$

Pine processionary caterpillars (2)

β_i	$\mathbb{E}^\pi(\beta_i \mid \mathbf{y}, X)$	$\mathbb{V}^\pi(\beta_i \mid \mathbf{y}, X)$
β_0	10.9874	6.2604
β_1	-0.0044	2e-06
β_2	-0.0538	0.0003
β_3	0.0679	0.0066
β_4	-1.2923	0.2125
β_5	0.2314	0.0073
β_6	-0.3564	1.6403
β_7	-0.2372	0.6765
β_8	0.1809	0.0375
β_9	-1.2840	0.5100
β_{10}	-0.4327	0.3670

$$c = 1,000$$

Conjugacy

Moreover,

$$\mathbb{V}^\pi[\beta|y,X] = \frac{c(s^2 + (\tilde\beta - \hat\beta)^\mathrm{T} X^\mathrm{T} X(\tilde\beta - \hat\beta)/(c+1))}{n(c+1)}(X^\mathrm{T} X)^{-1}\,.$$

Conjugacy

Moreover,

$$\mathbb{V}^\pi[\beta|y,X] = \frac{c(s^2 + (\tilde\beta - \hat\beta)^\mathrm{T} X^\mathrm{T} X(\tilde\beta - \hat\beta)/(c+1))}{n(c+1)} (X^\mathrm{T} X)^{-1} .$$

Convenient tool for translating prior information on β: For instance, if $c = 1$, this is equivalent to putting the same weight on the prior information and on the sample:

$$\mathbb{E}^\pi(\beta|y,X) = \left(\frac{\tilde\beta + \hat\beta}{2}\right)$$

average between prior mean and maximum likelihood estimator.

Conjugacy

Moreover,

$$\mathbb{V}^\pi[\beta|y,X] = \frac{c(s^2 + (\tilde\beta - \hat\beta)^{\mathrm{T}} X^{\mathrm{T}} X(\tilde\beta - \hat\beta)/(c+1))}{n(c+1)} (X^{\mathrm{T}}X)^{-1}\,.$$

Convenient tool for translating prior information on β: For instance, if $c = 1$, this is equivalent to putting the same weight on the prior information and on the sample:

$$\mathbb{E}^\pi(\beta|y,X) = \left(\frac{\tilde\beta + \hat\beta}{2}\right)$$

average between prior mean and maximum likelihood estimator.

If, instead, $c = 100$, the prior gets a weight of 1% of the sample.

Predictive

Prediction of $m \geq 1$ future observations from units in which the explanatory variables $\tilde{X}$—but not the outcome variable

$$\tilde{y} \sim \mathcal{N}_m(\tilde{X}\beta, \sigma^2 I_m)$$

—have been observed

Predictive

Prediction of $m \geq 1$ future observations from units in which the explanatory variables $\tilde{X}$—but not the outcome variable

$$\tilde{y} \sim \mathcal{N}_m(\tilde{X}\beta, \sigma^2 I_m)$$

—have been observed

Predictive distribution on $\tilde{y}$ defined as marginal of the joint posterior distribution on $(\tilde{y}, \beta, \sigma^2)$. Can be computed analytically by

$$\int \pi(\tilde{y}|\sigma^2, y, X, \tilde{X})\pi(\sigma^2|y, X, \tilde{X})\,\mathrm{d}\sigma^2\,.$$

Gaussian predictive

Conditional on σ^2, the future vector of observations has a Gaussian distribution with

$$\begin{aligned}
\mathbb{E}^\pi[\tilde{y}|\sigma^2, y, X, \tilde{X}] &= \mathbb{E}^\pi[\mathbb{E}^\pi(\tilde{y}|\beta, \sigma^2, y, X, \tilde{X})|\sigma^2, y, X, \tilde{X}] \\
&= \mathbb{E}^\pi[\tilde{X}\beta|\sigma^2, y, X, \tilde{X}] \\
&= \tilde{X}\frac{\tilde{\beta} + c\hat{\beta}}{c+1}
\end{aligned}$$

independently of σ^2.

Gaussian predictive

Conditional on σ^2, the future vector of observations has a Gaussian distribution with

$$\begin{aligned}
\mathbb{E}^\pi[\tilde{y}|\sigma^2, y, X, \tilde{X}] &= \mathbb{E}^\pi[\mathbb{E}^\pi(\tilde{y}|\beta, \sigma^2, y, X, \tilde{X})|\sigma^2, y, X, \tilde{X}] \\
&= \mathbb{E}^\pi[\tilde{X}\beta|\sigma^2, y, X, \tilde{X}] \\
&= \tilde{X}\frac{\tilde{\beta} + c\hat{\beta}}{c+1}
\end{aligned}$$

independently of σ^2. Similarly,

$$\begin{aligned}
\mathbb{V}^\pi(\tilde{y}|\sigma^2, y, X, \tilde{X}) &= \mathbb{E}^\pi[\mathbb{V}(\tilde{y}|\beta, \sigma^2, y, X, \tilde{X})|\sigma^2, y, X, \tilde{X}] \\
&\quad +\mathbb{V}^\pi[\mathbb{E}^\pi(\tilde{y}|\beta, \sigma^2, y, X, \tilde{X})|\sigma^2, y, X, \tilde{X}] \\
&= \mathbb{E}^\pi[\sigma^2 I_m|\sigma^2, y, X, \tilde{X}] + \mathbb{V}^\pi(\tilde{X}\beta|\sigma^2, y, X, \tilde{X}) \\
&= \sigma^2\left(I_m + \frac{c}{c+1}\tilde{X}(X^{\mathrm{T}}X)^{-1}\tilde{X}^{\mathrm{T}}\right)
\end{aligned}$$

Predictor

A predictor under squared error loss is the posterior predictive mean

$$\tilde{X}\frac{\tilde{\beta}+c\beta}{c+1},$$

Representation quite intuitive, being the product of the matrix of explanatory variables $\tilde{X}$ by the Bayes estimate of β.

Credible regions

Highest posterior density (HPD) regions on subvectors of the parameter β derived from the marginal posterior distribution of β.

Credible regions

Highest posterior density (HPD) regions on subvectors of the parameter β derived from the marginal posterior distribution of β. For a single parameter,

$$\beta_i|y,X \sim \mathscr{T}_1\left(n, \frac{c}{c+1}\left(\frac{\tilde\beta_i}{c}+\hat\beta_i\right), \frac{c(s^2+(\tilde\beta-\hat\beta)^{\mathrm{T}}X^{\mathrm{T}}X(\tilde\beta-\hat\beta)/(c+1))}{n(c+1)}\omega_{(i,i)}\right),$$

where $\omega_{(i,i)}$ is the (i,i)-th element of the matrix $(X^{\mathrm{T}}X)^{-1}$.

T time

If

$$\tau = \frac{\tilde\beta + c\hat\beta}{c+1}$$

and

$$K = \frac{c(s^2 + (\tilde\beta - \hat\beta)^{\mathrm{T}} X^{\mathrm{T}} X (\tilde\beta - \hat\beta)/(c+1))}{n(c+1)} (X^{\mathrm{T}}X)^{-1} = \left(\kappa_{(i,j)}\right) ,$$

the transform

$$\mathfrak{T}_i = \frac{\beta_i - \tau_i}{\sqrt{\kappa_{(i,i)}}}$$

has a standard t distribution with n degrees of freedom.

T HPD

A $1-\alpha$ HPD interval on β_i is thus given by

$$\left[\tau_i - \sqrt{\kappa_{(i,i)}} F_n^{-1}(1-\alpha/2), \tau_i + \sqrt{\kappa_{(i,i)}} F_n^{-1}(1-\alpha/2)\right].$$

Pine processionary caterpillars

β_i	HPD interval
β_0	$[5.7435, 16.2533]$
β_1	$[-0.0071, -0.0018]$
β_2	$[-0.0914, -0.0162]$
β_3	$[-0.1029, 0.2387]$
β_4	$[-2.2618, -0.3255]$
β_5	$[0.0524, 0.4109]$
β_6	$[-3.0466, 2.3330]$
β_7	$[-1.9649, 1.4900]$
β_8	$[-0.2254, 0.5875]$
β_9	$[-2.7704, 0.1997]$
β_{10}	$[-1.6950, 0.8288]$

$$c = 100$$

T marginal

Marginal distribution of y is multivariate t distribution

Proof. Since $\beta|\sigma^2, X \sim \mathscr{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^{\mathrm{T}}X)^{-1})$,

$$X\beta|\sigma^2, X \sim \mathscr{N}(X\tilde{\beta}, c\sigma^2 X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}),$$

which implies that

$$y|\sigma^2, X \sim \mathscr{N}_n(X\tilde{\beta}, \sigma^2(I_n + cX(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})).$$

Integrating in σ^2 yields

$$\begin{aligned} f(y|X) &= (c+1)^{-(k+1)/2}\pi^{-n/2}\Gamma(n/2) \\ &\quad \times \left[y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y - \frac{1}{c+1}\tilde{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\tilde{\beta}\right]^{-n/2}. \end{aligned}$$

Point null hypothesis

If a null hypothesis is $H_0 : R\beta = r$, the model under H_0 can be rewritten as

$$y|\beta^0, \sigma^2, X_0 \overset{H_0}{\sim} \mathscr{N}_n\left(X_0\beta^0, \sigma^2 I_n\right)$$

where β^0 is $(k+1-q)$ dimensional.

Point null marginal

Under the prior

$$\beta^0|X_0,\sigma^2 \sim \mathscr{N}_{k+1-q}\left(\tilde\beta^0, c_0\sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right),$$

the marginal distribution of y under H_0 is

$$\begin{aligned} f(y|X_0,H_0) &= (c+1)^{-(k+1-q)/2}\pi^{-n/2}\Gamma(n/2) \\ &\times\left[y^{\mathrm{T}}y - \frac{c_0}{c_0+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y \right. \\ &\left. -\frac{1}{c_0+1}\tilde\beta_0^{\mathrm{T}}X_0^{\mathrm{T}}X_0\tilde\beta_0\right]^{-n/2}. \end{aligned}$$

Bayes factor

Therefore the Bayes factor is closed form:

$$
\begin{aligned}
B^\pi_{10} &= \frac{f(y|X,H_1)}{f(y|X_0,H_0)} = \frac{(c_0+1)^{(k+1-q)/2}}{(c+1)^{(k+1)/2}} \\
&\left[\frac{y^\mathrm{T}y-\frac{c_0}{c_0+1}y^\mathrm{T}X_0(X_0^\mathrm{T}X_0)^{-1}X_0^\mathrm{T}y-\frac{1}{c_0+1}\tilde\beta_0^\mathrm{T}X_0^\mathrm{T}X_0\tilde\beta_0}{y^\mathrm{T}y-\frac{c}{c+1}y^\mathrm{T}X(X^\mathrm{T}X)^{-1}X^\mathrm{T}y-\frac{1}{c+1}\tilde\beta^\mathrm{T}X^\mathrm{T}X\tilde\beta}\right]^{n/2}
\end{aligned}
$$

Bayes factor

Therefore the Bayes factor is closed form:

$$
\begin{aligned}
B_{10}^{\pi} &= \frac{f(y|X,H_1)}{f(y|X_0,H_0)} = \frac{(c_0+1)^{(k+1-q)/2}}{(c+1)^{(k+1)/2}} \\
&\left[\frac{y^{\mathrm{T}}y - \frac{c_0}{c_0+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y - \frac{1}{c_0+1}\tilde{\beta}_0^{\mathrm{T}}X_0^{\mathrm{T}}X_0\tilde{\beta}_0}{y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y - \frac{1}{c+1}\tilde{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\tilde{\beta}}\right]^{n/2}
\end{aligned}
$$

- Means using the *same* σ^2 on both models
- Still depends on the choice of (c_0, c)

Zellner's noninformative G-prior

Difference with informative G-prior setup is that we now consider c as unknown (relief!)

Zellner's noninformative G-prior

Difference with informative G-prior setup is that we now consider c as unknown (relief!)

Solution

Use the same G-prior distribution with $\tilde{\beta} = 0_{k+1}$, conditional on c, and introduce a diffuse prior on c,

$$\pi(c) = c^{-1}\mathbb{I}_{\mathbb{N}^*}(c)\,.$$

Posterior distribution

Corresponding marginal posterior on the parameters of interest

$$\begin{aligned}
\pi(\beta,\sigma^2|y,X) &= \int \pi(\beta,\sigma^2|y,X,c)\pi(c|y,X)\,\mathrm{d}c \\
&\propto \sum_{c=1}^{\infty} \pi(\beta,\sigma^2|y,X,c) f(y|X,c)\pi(c) \\
&\propto \sum_{c=1}^{\infty} \pi(\beta,\sigma^2|y,X,c) f(y|X,c)\, c^{-1}\,.
\end{aligned}$$

Posterior distribution

Corresponding marginal posterior on the parameters of interest

$$\begin{aligned}
\pi(\beta,\sigma^2|y,X) &= \int \pi(\beta,\sigma^2|y,X,c)\pi(c|y,X)\,\mathrm{d}c \\
&\propto \sum_{c=1}^{\infty} \pi(\beta,\sigma^2|y,X,c) f(y|X,c)\pi(c) \\
&\propto \sum_{c=1}^{\infty} \pi(\beta,\sigma^2|y,X,c) f(y|X,c)\, c^{-1}\,.
\end{aligned}$$

and

$$f(y|X,c) \propto (c{+}1)^{-(k+1)/2} \left[y^{\mathrm{T}}y - \frac{c}{c+1} y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \right]^{-n/2}.$$

Posterior means

The Bayes estimates of β and σ^2 are given by

$$\begin{aligned}
\mathbb{E}^\pi[\beta|y,X] &= \mathbb{E}^\pi[\mathbb{E}^\pi(\beta|y,X,c)|y,X] = \mathbb{E}^\pi[c/(c+1)\hat\beta)|y,X] \\
&= \left(\frac{\sum_{c=1}^{\infty} c/(c+1) f(y|X,c)c^{-1}}{\sum_{c=1}^{\infty} f(y|X,c)c^{-1}} \right) \hat\beta
\end{aligned}$$

and

$$\mathbb{E}^\pi[\sigma^2|y,X] = \frac{\sum_{c=1}^{\infty} \frac{s^2 + \hat\beta^{\mathrm{T}} X^{\mathrm{T}} X \hat\beta/(c+1)}{n-2} f(y|X,c)c^{-1}}{\sum_{c=1}^{\infty} f(y|X,c)c^{-1}} .$$

Computational details

- Both terms involve infinite summations on c
- The denominator in both cases is the normalising constant of the posterior

$$\sum_{c=1}^{\infty} f(y|X,c)c^{-1}$$

Computational details (cont'd)

$$
\begin{aligned}
\mathbb{V}^\pi[\beta|y,X] &= \mathbb{E}^\pi[\mathbb{V}^\pi(\beta|y,X,c)|y,X] + \mathbb{V}^\pi[\mathbb{E}^\pi(\beta|y,X,c)|y,Xx] \\
&= \mathbb{E}^\pi\left[c/(n(c+1))(s^2+\hat\beta^{\mathrm{T}}(X^{\mathrm{T}}X)\hat\beta/(c+1))(X^{\mathrm{T}}X)^{-1}\right] \\
&\qquad\qquad +\mathbb{V}^\pi[c/(c+1)\hat\beta|y,X] \\
&= \left[\frac{\sum_{c=1}^\infty f(y|X,c)/(n(c+1))(s^2+\hat\beta^{\mathrm{T}}(X^{\mathrm{T}}X)\hat\beta/(c+1))}{\sum_{c=1}^\infty f(y|X,c)c^{-1}}\right](X^{\mathrm{T}}X)^{-1} \\
&\quad +\hat\beta\left(\frac{\sum_{c=1}^\infty (c/(c+1)-\mathbb{E}(c/(c+1)|y,X))^2 f(y|X,c)c^{-1}}{\sum_{c=1}^\infty f(y|X,c)c^{-1}}\right)\hat\beta^{\mathrm{T}}\,.
\end{aligned}
$$

Marginal distribution

Important point: the marginal distribution of the dataset is available in closed form

$$f(y|X) \propto \sum_{i=1}^{\infty} c^{-1}(c+1)^{-(k+1)/2}\left[y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y\right]^{-n/2}$$

Marginal distribution

Important point: the marginal distribution of the dataset is available in closed form

$$f(y|X) \propto \sum_{i=1}^{\infty} c^{-1}(c+1)^{-(k+1)/2}\left[y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y\right]^{-n/2}$$

$\mathcal{T}$-shape means normalising constant can be computed too.

Point null hypothesis

For null hypothesis $H_0 : R\beta = r$, the model under H_0 can be rewritten as

$$y|\beta^0, \sigma^2, X_0 \overset{H_0}{\sim} \mathscr{N}_n \left(X_0\beta^0, \sigma^2 I_n\right)$$

where β^0 is $(k + 1 - q)$ dimensional.

Point null marginal

Under the prior

$$\beta^0|X_0,\sigma^2,c \sim \mathscr{N}_{k+1-q}\left(0_{k+1-q}, c\sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right)$$

and $\pi(c)=1/c$, the marginal distribution of y under H_0 is

$$f(y|X_0,H_0) \propto \sum_{c=1}^{\infty}(c+1)^{-(k+1-q)/2}\left[y^{\mathrm{T}}y-\frac{c}{c+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y\right]^{-n/2}.$$

Point null marginal

Under the prior

$$\beta^0|X_0,\sigma^2,c \sim \mathscr{N}_{k+1-q}\left(0_{k+1-q}, c\sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right)$$

and $\pi(c)=1/c$, the marginal distribution of y under H_0 is

$$f(y|X_0,H_0) \propto \sum_{c=1}^{\infty}(c+1)^{-(k+1-q)/2}\left[y^{\mathrm{T}}y-\frac{c}{c+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y\right]^{-n/2}.$$

Bayes factor $B_{10}^{\pi}=f(\mathbf{y}|X)/f(\mathbf{y}|X_0,H_0)$ can be computed

Processionary pine caterpillars

For $H_0 : \beta_8 = \beta_9 = 0$, $\log_{10}(B^{\pi}_{10}) = -0.7884$

Processionary pine caterpillars

For $H_0 : \beta_8 = \beta_9 = 0$, $\log_{10}(B^{\pi}_{10}) = -0.7884$

```
            Estimate  Post. Var.  log10(BF)

(Intercept)  9.2714   9.1164        1.4205 (***)
X1          -0.0037    2e-06        0.8502 (**)
X2          -0.0454   0.0004        0.5664 (**)
X3           0.0573   0.0086       -0.3609
X4          -1.0905   0.2901        0.4520 (*)
X5           0.1953   0.0099        0.4007 (*)
X6          -0.3008   2.1372       -0.4412
X7          -0.2002   0.8815       -0.4404
X8           0.1526   0.0490       -0.3383
X9          -1.0835   0.6643       -0.0424
X10         -0.3651   0.4716       -0.3838

evidence against H0:
(****) decisive, (***) strong, (**) subtantial, (*) poor
```

Markov Chain Monte Carlo Methods

Complexity of most models encountered in Bayesian modelling

Markov Chain Monte Carlo Methods

Complexity of most models encountered in Bayesian modelling

Standard simulation methods not good enough a solution

Markov Chain Monte Carlo Methods

Complexity of most models encountered in Bayesian modelling

Standard simulation methods not good enough a solution

New technique at the core of Bayesian computing, based on *Markov chains*

Markov chains

Markov chain

A process $(\theta^{(t)})_{t\in\mathbb{N}}$ is an *homogeneous Markov chain* if the distribution of $\theta^{(t)}$ given the past $(\theta^{(0)},\ldots,\theta^{(t-1)})$

1. only depends on $\theta^{(t-1)}$
2. is the same for all $t \in \mathbb{N}^*$.

Algorithms based on Markov chains

Idea: simulate from a posterior density $\pi(\cdot|x)$ [or any density] by producing a Markov chain

$$(\theta^{(t)})_{t\in\mathbb{N}}$$

whose stationary distribution is

$$\pi(\cdot|x)$$

Algorithms based on Markov chains

Idea: simulate from a posterior density $\pi(\cdot|x)$ [or any density] by producing a Markov chain

$$(\theta^{(t)})_{t\in\mathbb{N}}$$

whose stationary distribution is

$$\pi(\cdot|x)$$

Translation

For t large enough, $\theta^{(t)}$ is approximately distributed from $\pi(\theta|x)$, no matter what the starting value $\theta^{(0)}$ is *[Ergodicity]*.

Convergence

If an algorithm that generates such a chain can be constructed, the ergodic theorem guarantees that, in almost all settings, the average

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)})$$

converges to $\mathbb{E}^\pi[g(\theta)|x]$, for (almost) any starting value

More convergence

If the produced Markov chains are irreducible [can reach any region in a finite number of steps], then they are both positive recurrent with stationary distribution $\pi(\cdot|x)$ *and* ergodic [asymptotically independent from the starting value $\theta^{(0)}$]

- While, for t large enough, $\theta^{(t)}$ is approximately distributed from $\pi(\theta|x)$ and can thus be used like the output from a more standard simulation algorithm, one must take care of the correlations between the $\theta^{(t)}$'s

Demarginalising

Takes advantage of *hierarchical structures*: if

$$\pi(\theta|x) = \int \pi_1(\theta|x,\lambda)\pi_2(\lambda|x)\,d\lambda\,,$$

simulating from $\pi(\theta|x)$ comes from simulating from the joint distribution

$$\pi_1(\theta|x,\lambda)\;\pi_2(\lambda|x)$$

Two-stage Gibbs sampler

Usually $\pi_2(\lambda|x)$ not available/simulable

Two-stage Gibbs sampler

Usually $\pi_2(\lambda|x)$ not available/simulable

More often, both *conditional posterior distributions*,

$$\pi_1(\theta|x,\lambda) \quad \text{and} \quad \pi_2(\lambda|x,\theta)$$

can be simulated.

Two-stage Gibbs sampler

Usually $\pi_2(\lambda|x)$ not available/simulable

More often, both *conditional posterior distributions*,

$$\pi_1(\theta|x,\lambda) \quad \text{and} \quad \pi_2(\lambda|x,\theta)$$

can be simulated.

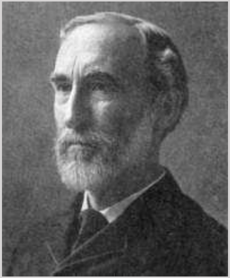
Idea: Create a Markov chain based on those conditionals

Two-stage Gibbs sampler (cont'd)

Initialization: Start with an arbitrary value $\lambda^{(0)}$

Iteration t: Given $\lambda^{(t-1)}$, generate

1. $\theta^{(t)}$ according to $\pi_1(\theta|x, \lambda^{(t-1)})$
2. $\lambda^{(t)}$ according to $\pi_2(\lambda|x, \theta^{(t)})$



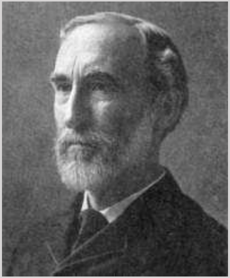
J.W. Gibbs (1839-1903)

Two-stage Gibbs sampler (cont'd)

Initialization: Start with an arbitrary value $\lambda^{(0)}$

Iteration t: Given $\lambda^{(t-1)}$, generate

1. $\theta^{(t)}$ according to $\pi_1(\theta|x, \lambda^{(t-1)})$
2. $\lambda^{(t)}$ according to $\pi_2(\lambda|x, \theta^{(t)})$



J.W. Gibbs (1839-1903)

$\pi(\theta, \lambda|x)$ **is a stationary distribution for this transition**

Implementation

1. Derive efficient decomposition of the joint distribution into simulable conditionals (mixing behavior, acf(), blocking, &tc.)

Implementation

1. Derive efficient decomposition of the joint distribution into simulable conditionals (mixing behavior, acf(), blocking, &tc.)
2. Find when to stop the algorithm (mode chasing, missing mass, shortcuts, &tc.)

Simple Example: iid $\mathscr{N}(\mu, \sigma^2)$ Observations

When $y_1, \ldots, y_n \stackrel{\text{iid}}{\sim} \mathscr{N}(\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate outside a standard family

Simple Example: iid $\mathscr{N}(\mu, \sigma^2)$ Observations

When $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathscr{N}(\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate outside a standard family

But...

$$\mu|\mathbf{y}, \sigma^2 \sim \mathscr{N}\left(\mu \,\middle|\, \tfrac{1}{n}\sum_{i=1}^n y_i, \tfrac{\sigma^2}{n}\right)$$

$$\sigma^2|\mathbf{y}, \mu \sim \mathscr{I}\mathscr{G}\left(\sigma^2 \,\middle|\, \tfrac{n}{2} - 1, \tfrac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right)$$

assuming constant (improper) priors on both μ and σ^2

- Hence we may use the Gibbs sampler for simulating from the posterior of (μ, σ^2)

Gibbs output analysis

Example (Cauchy posterior)

$$\pi(\mu|\mathscr{D}) \propto \frac{e^{-\mu^2/20}}{(1+(x_1-\mu)^2)(1+(x_2-\mu)^2)}$$

is marginal of

$$\pi(\mu,\boldsymbol{\omega}|\mathscr{D}) \propto e^{-\mu^2/20} \times \prod_{i=1}^{2} e^{-\omega_i[1+(x_i-\mu)^2]} \,.$$

Gibbs output analysis

Example (Cauchy posterior)

$$\pi(\mu|\mathscr{D}) \propto \frac{e^{-\mu^2/20}}{(1+(x_1-\mu)^2)(1+(x_2-\mu)^2)}$$

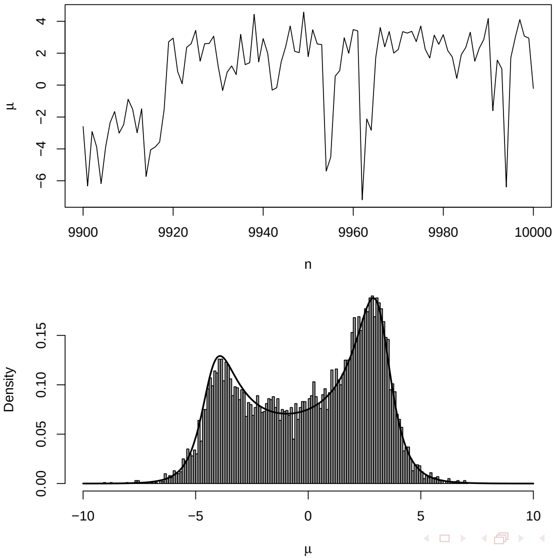
is marginal of

$$\pi(\mu,\boldsymbol{\omega}|\mathscr{D}) \propto e^{-\mu^2/20} \times \prod_{i=1}^{2} e^{-\omega_i[1+(x_i-\mu)^2]} .$$

Corresponding conditionals

$$\begin{aligned}(\omega_1,\omega_2)|\mu &\sim \mathscr{E}xp(1+(x_1-\mu)^2) \otimes \mathscr{E}xp(1+(x_2-\mu))^2) \\ \mu|\boldsymbol{\omega} &\sim \mathscr{N}\left(\sum_i \omega_i x_i/(\sum_i \omega_i + 1/20), 1/(2\sum_i \omega_i + 1/10)\right)\end{aligned}$$

Gibbs output analysis (cont'd)



Generalisation

Consider several groups of parameters, $\theta, \lambda_1, \ldots, \lambda_p$, such that

$$\pi(\theta|x) = \int \ldots \int \pi(\theta, \lambda_1, \ldots, \lambda_p|x)\, d\lambda_1 \cdots d\lambda_p$$

or simply divide θ in

$$(\theta_1, \ldots, \theta_p)$$

The general Gibbs sampler

For a joint distribution $\pi(\theta)$ with full conditionals $\pi_1, \ldots, \pi_p$,

Given $(\theta_1^{(t)}, \ldots, \theta_p^{(t)})$, simulate

1. $\theta_1^{(t+1)} \sim \pi_1(\theta_1 | \theta_2^{(t)}, \ldots, \theta_p^{(t)})$,
2. $\theta_2^{(t+1)} \sim \pi_2(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_p^{(t)})$,

$$\vdots$$

p. $\theta_p^{(t+1)} \sim \pi_p(\theta_p | \theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)})$.

Then $\theta^{(t)} \rightarrow \theta \sim \pi$

Variable selection

Back to regression: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Variable selection

Back to regression: one dependent random variable y and a set $\{x_1, \dots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Variable selection

Back to regression: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Assumption: every subset $\{i_1, \ldots, i_q\}$ of q $(0 \leq q \leq k)$ explanatory variables, $\{\mathbf{1}_n, x_{i_1}, \ldots, x_{i_q}\}$, is a proper set of explanatory variables for the regression of y [intercept included in every corresponding model]

Variable selection

Back to regression: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Assumption: every subset $\{i_1, \ldots, i_q\}$ of q $(0 \leq q \leq k)$ explanatory variables, $\{\mathbf{1}_n, x_{i_1}, \ldots, x_{i_q}\}$, is a proper set of explanatory variables for the regression of y [intercept included in every corresponding model]

Computational issue

2^k models in competition...

Model notations

1. $$X = \begin{bmatrix} \mathbf{1}_n & x_1 & \cdots & x_k \end{bmatrix}$$
   is the matrix containing $\mathbf{1}_n$ and all the k potential predictor variables
2. Each model $\mathfrak{M}_\gamma$ associated with binary indicator vector $\gamma \in \Gamma = \{0,1\}^k$ where $\gamma_i = 1$ means that the variable x_i is included in the model $\mathfrak{M}_\gamma$
3. $q_\gamma = \mathbf{1}_n^{\mathrm{T}}\gamma$ number of variables included in the model $\mathfrak{M}_\gamma$
4. $t_1(\gamma)$ and $t_0(\gamma)$ indices of variables included in the model and indices of variables not included in the model

Model indicators

For $\beta \in \mathbb{R}^{k+1}$ and X, we define β_γ as the subvector

$$\beta_\gamma = \left(\beta_0, (\beta_i)_{i \in t_1(\gamma)}\right)$$

and X_γ as the submatrix of X where only the column $\mathbf{1}_n$ and the columns in $t_1(\gamma)$ have been left.

Models in competition

The model $\mathfrak{M}_\gamma$ is thus defined as

$$y|\gamma, \beta_\gamma, \sigma^2, X \sim \mathscr{N}_n\left(X_\gamma\beta_\gamma, \sigma^2 I_n\right)$$

where $\beta_\gamma \in \mathbb{R}^{q_\gamma+1}$ and $\sigma^2 \in \mathbb{R}_+^*$ are the unknown parameters.

Models in competition

The model $\mathfrak{M}_\gamma$ is thus defined as

$$y|\gamma, \beta_\gamma, \sigma^2, X \sim \mathscr{N}_n\left(X_\gamma\beta_\gamma, \sigma^2 I_n\right)$$

where $\beta_\gamma \in \mathbb{R}^{q_\gamma+1}$ and $\sigma^2 \in \mathbb{R}_+^*$ are the unknown parameters.

Warning

σ^2 is common to all models and thus uses the same prior for all models

Informative G-prior

Many (2^k) models in competition: we cannot expect a practitioner to specify a prior on every $\mathfrak{M}_\gamma$ in a completely subjective and autonomous manner.

Shortcut: We derive *all* priors from a single global prior associated with the so-called *full model* that corresponds to $\gamma = (1, \dots, 1)$.

Prior definitions

(i) For the full model, Zellner's G-prior:

$$\beta|\sigma^2, X \sim \mathscr{N}_{k+1}(\tilde\beta, c\sigma^2(X^{\mathrm{T}}X)^{-1}) \quad \text{and} \quad \sigma^2 \sim \pi(\sigma^2|X) = \sigma^{-2}$$

(ii) For each model $\mathfrak{M}_\gamma$, the prior distribution of β_γ conditional on σ^2 is fixed as

$$\beta_\gamma|\gamma, \sigma^2 \sim \mathscr{N}_{q_\gamma+1}\left(\tilde\beta_\gamma, c\sigma^2\left(X_\gamma^{\mathrm{T}}X_\gamma\right)^{-1}\right),$$

where $\tilde\beta_\gamma = \left(X_\gamma^T X_\gamma\right)^{-1} X_\gamma^T \tilde\beta$ and same prior on σ^2.

Prior completion

The joint prior for model $\mathfrak{M}_\gamma$ is the improper prior

$$\pi(\beta_\gamma, \sigma^2|\gamma) \propto \left(\sigma^2\right)^{-(q_\gamma+1)/2-1} \exp\left[-\frac{1}{2(c\sigma^2)}\left(\beta_\gamma - \tilde{\beta}_\gamma\right)^{\mathrm{T}} (X_\gamma^{\mathrm{T}} X_\gamma)\left(\beta_\gamma - \tilde{\beta}_\gamma\right)\right].$$

Prior competition (2)

Infinitely many ways of defining a prior on the model index γ: choice of uniform prior $\pi(\gamma|X) = 2^{-k}$.

Posterior distribution of γ central to variable selection since it is proportional to marginal density of y on $\mathfrak{M}_\gamma$ (or *evidence* of $\mathfrak{M}_\gamma$)

$$\begin{aligned} \pi(\gamma|y,X) &\propto f(y|\gamma,X)\pi(\gamma|X) \propto f(y|\gamma,X) \\ &= \int \left(\int f(y|\gamma,\beta,\sigma^2,X)\pi(\beta|\gamma,\sigma^2,X)\,\mathrm{d}\beta \right) \pi(\sigma^2|X)\,\mathrm{d}\sigma^2 \,. \end{aligned}$$

$$
\begin{aligned}
f(y|\gamma,\sigma^2,X) &= \int f(y|\gamma,\beta,\sigma^2)\pi(\beta|\gamma,\sigma^2)\,\mathsf{d}\beta \\
&= (c+1)^{-(q_\gamma+1)/2}(2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}y^{\mathrm{T}}y\right. \\
&\quad \left.+\frac{1}{2\sigma^2(c+1)}\left\{cy^{\mathrm{T}}X_\gamma\left(X_\gamma^{\mathrm{T}}X_\gamma\right)^{-1}X_\gamma^{\mathrm{T}}y-\tilde{\beta}_\gamma^{\mathrm{T}}X_\gamma^{\mathrm{T}}X_\gamma\tilde{\beta}_\gamma\right\}\right)
\end{aligned}
$$

this posterior density satisfies

$$
\begin{aligned}
\pi(\gamma|y,X) \propto\ & (c+1)^{-(q_\gamma+1)/2}\left[y^{\mathrm{T}}y-\frac{c}{c+1}y^{\mathrm{T}}X_\gamma\left(X_\gamma^{\mathrm{T}}X_\gamma\right)^{-1}X_\gamma^{\mathrm{T}}y\right. \\
&\left.-\frac{1}{c+1}\tilde{\beta}_\gamma^{\mathrm{T}}X_\gamma^{\mathrm{T}}X_\gamma\tilde{\beta}_\gamma\right]^{-n/2}.
\end{aligned}
$$

Pine processionary caterpillars

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y},X)$
0,1,2,4,5	0.2316
0,1,2,4,5,9	0.0374
0,1,9	0.0344
0,1,2,4,5,10	0.0328
0,1,4,5	0.0306
0,1,2,9	0.0250
0,1,2,4,5,7	0.0241
0,1,2,4,5,8	0.0238
0,1,2,4,5,6	0.0237
0,1,2,3,4,5	0.0232
0,1,6,9	0.0146
0,1,2,3,9	0.0145
0,9	0.0143
0,1,2,6,9	0.0135
0,1,4,5,9	0.0128
0,1,3,9	0.0117
0,1,2,8	0.0115

Pine processionary caterpillars (cont'd)

Interpretation

Model $\mathfrak{M}_\gamma$ with the highest posterior probability is $t_1(\gamma) = (1, 2, 4, 5)$, which corresponds to the variables

- altitude,
- slope,
- height of the tree sampled in the center of the area, and
- diameter of the tree sampled in the center of the area.

Pine processionary caterpillars (cont'd)

Interpretation

Model $\mathfrak{M}_\gamma$ with the highest posterior probability is $t_1(\gamma) = (1, 2, 4, 5)$, which corresponds to the variables

- altitude,
- slope,
- height of the tree sampled in the center of the area, and
- diameter of the tree sampled in the center of the area.

Corresponds to the five variables identified in the R regression output

Noninformative extension

For Zellner noninformative prior with $\pi(c) = 1/c$, we have

$$\begin{aligned}\pi(\gamma|y,X) \quad &\propto \quad \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q_\gamma+1)/2}\left[y^{\mathrm{T}}y - \right.\\ &\qquad \left.\frac{c}{c+1}y^{\mathrm{T}}X_\gamma\left(X_\gamma^{\mathrm{T}}X_\gamma\right)^{-1}X_\gamma^{\mathrm{T}}y\right]^{-n/2}.\end{aligned}$$

Pine processionary caterpillars

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y}, X)$
0,1,2,4,5	0.0929
0,1,2,4,5,9	0.0325
0,1,2,4,5,10	0.0295
0,1,2,4,5,7	0.0231
0,1,2,4,5,8	0.0228
0,1,2,4,5,6	0.0228
0,1,2,3,4,5	0.0224
0,1,2,3,4,5,9	0.0167
0,1,2,4,5,6,9	0.0167
0,1,2,4,5,8,9	0.0137
0,1,4,5	0.0110
0,1,2,4,5,9,10	0.0100
0,1,2,3,9	0.0097
0,1,2,9	0.0093
0,1,2,4,5,7,9	0.0092
0,1,2,6,9	0.0092

Stochastic search for the most likely model

When k gets large, impossible to compute the posterior probabilities of the 2^k models.

Stochastic search for the most likely model

When k gets large, impossible to compute the posterior probabilities of the 2^k models.

Need of a tailored algorithm that samples from $\pi(\gamma|y, X)$ and selects the most likely models.

Can be done by Gibbs sampling, given the availability of the full conditional posterior probabilities of the γ_i's.
If $\gamma_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_k)$ $(1 \leq i \leq k)$

$$\pi(\gamma_i|y, \gamma_{-i}, X) \propto \pi(\gamma|y, X)$$

(to be evaluated in both $\gamma_i = 0$ and $\gamma_i = 1$)

Gibbs sampling for variable selection

Initialization: Draw γ^0 from the uniform distribution on Γ

Gibbs sampling for variable selection

Initialization: Draw γ^0 from the uniform distribution on Γ

Iteration t**:** Given $(\gamma_1^{(t-1)}, \dots, \gamma_k^{(t-1)})$, generate

1. $\gamma_1^{(t)}$ according to $\pi(\gamma_1|y, \gamma_2^{(t-1)}, \dots, \gamma_k^{(t-1)}, X)$
2. $\gamma_2^{(t)}$ according to $\pi(\gamma_2|y, \gamma_1^{(t)}, \gamma_3^{(t-1)}, \dots, \gamma_k^{(t-1)}, X)$

$$\vdots$$

p. $\gamma_k^{(t)}$ according to $\pi(\gamma_k|y, \gamma_1^{(t)}, \dots, \gamma_{k-1}^{(t)}, X)$

MCMC interpretation

After $T \gg 1$ MCMC iterations, output used to approximate the posterior probabilities $\pi(\gamma|y,X)$ by empirical averages

$$\widehat{\pi}(\gamma|y,X) = \left(\frac{1}{T-T_0+1}\right)\sum_{t=T_0}^{T} \mathbb{I}_{\gamma^{(t)}=\gamma}.$$

where the T_0 first values are eliminated as *burnin*.

MCMC interpretation

After $T \gg 1$ MCMC iterations, output used to approximate the posterior probabilities $\pi(\gamma|y,X)$ by empirical averages

$$\widehat{\pi}(\gamma|y,X) = \left(\frac{1}{T-T_0+1}\right)\sum_{t=T_0}^{T} \mathbb{I}_{\gamma^{(t)}=\gamma}\,.$$

where the T_0 first values are eliminated as *burnin*.

And approximation of the probability to include i-th variable,

$$\widehat{P}^{\pi}(\gamma_i=1|y,X) = \left(\frac{1}{T-T_0+1}\right)\sum_{t=T_0}^{T} \mathbb{I}_{\gamma_i^{(t)}=1}\,.$$

Pine processionary caterpillars

γ_i	$\widehat{P}^\pi(\gamma_i = 1\|\mathbf{y}, X)$	$\widehat{P}^\pi(\gamma_i = 1\|\mathbf{y}, X)$
γ_1	0.8624	0.8844
γ_2	0.7060	0.7716
γ_3	0.1482	0.2978
γ_4	0.6671	0.7261
γ_5	0.6515	0.7006
γ_6	0.1678	0.3115
γ_7	0.1371	0.2880
γ_8	0.1555	0.2876
γ_9	0.4039	0.5168
γ_{10}	0.1151	0.2609

Probabilities of inclusion with both informative $(\tilde{\beta} = 0_{11}, c = 100)$ and noninformative Zellner's priors