Caterpillar Regression Example: Conjugate Priors, Conditional & Marginal Posteriors, Predictive Distribution, Variable Selection

Prof. Nicholas Zabaras
University of Notre Dame
Notre Dame, IN, USA

Email: nzabaras@gmail.com
URL: https:/www.zabaras.com/

September 22, 2017

Contents

- The Caterpillar Regression Problem, Linear Regression Model, MLE Estimator, Likelihood, MLE Estimates for our example, Conjugate Priors, Conditional and Marginal Posteriors, Ridge regression, Predictive Distribution, Implementation, Influence of the Conjugate Prior

- Zellner's G Prior, Marginal Posterior Mean and Variance, Predictive Modeling, Credible Intervals

- Jeffreys' non-informative Prior, Credible Intervals, Zellner's G Prior Marginal Distribution of y, Point Null Hypothesis and Bayes Factors

- Variable Selection, Model Competition, Variable Selection-Prior, Stochastic Search for the Most Probable Model, Gibb's Sampling for Variable Selection, Implementation

C. P. Robert, *The Bayesian Core*, *Springer, 2nd edition,* chapter 3 (full text available)

Regression

- Regression refers to statistical analysis that deals with the representation of dependencies between several variables.

- In particular, we want to find a representation of the distribution $f(y|\theta,\boldsymbol{x})$ of an observable variable y given a vector of observables $\boldsymbol{x}$, using samples of $(\boldsymbol{x}_i,y_i)$, $i=1,\ldots,n$.

- Here, we will consider a particular example of modeling dependencies in pine processionary caterpillar colony size.

Linear Regression Models

- In linear regression, we analyze the linear influence of some variables on others.

- In our particular example of pine processionary caterpillar colonies,

  - Response Variable: y is the number of processionary caterpillar colonies

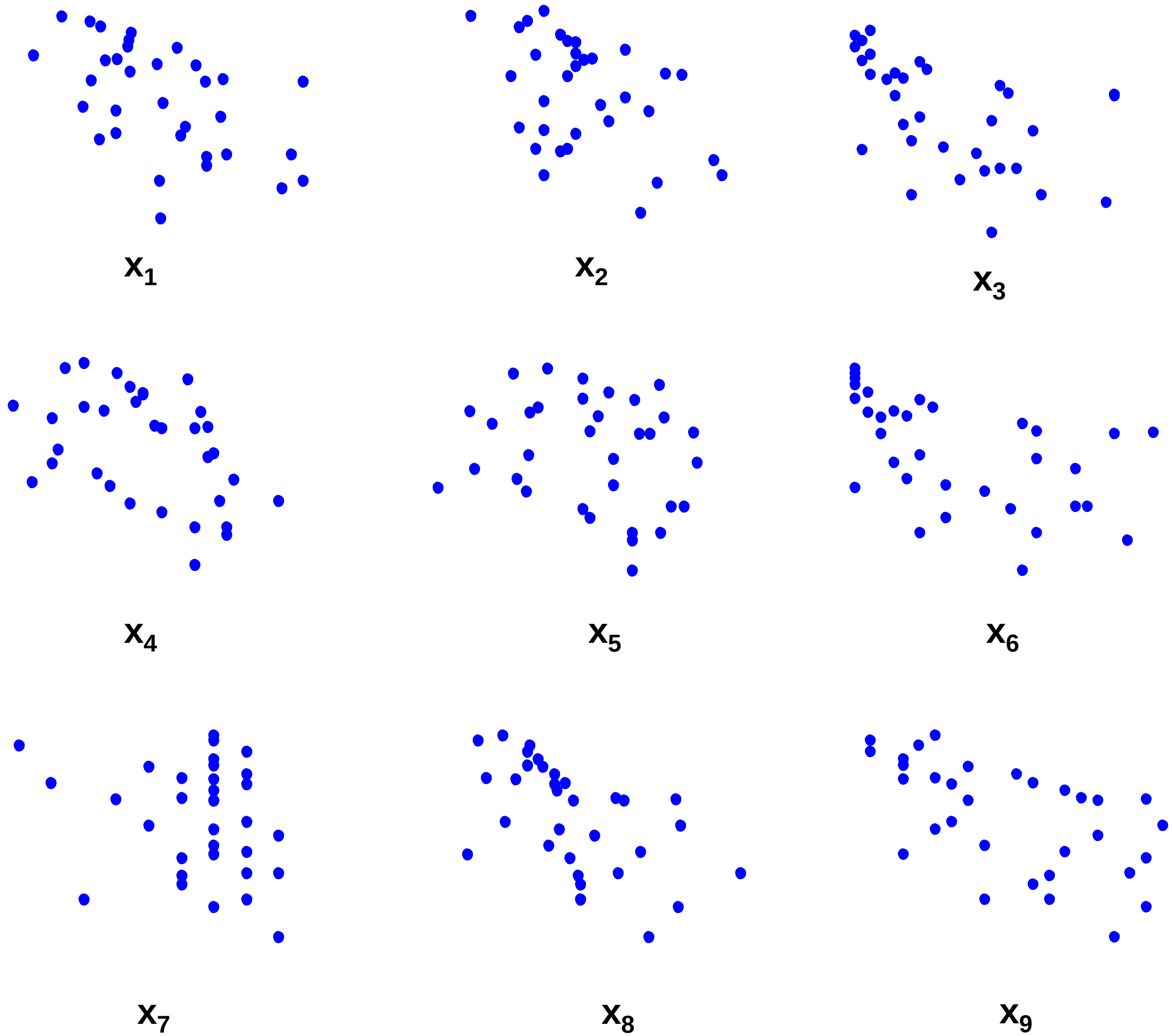
  - Explanatory Variables: Covariates $\boldsymbol{x}=(x_1,x_2,\ldots,x_k)$ as defined next (in general can be continuous, discrete or mixed type)

Caterpillar Regression Problem

The pine processionary caterpillar colony size (y=number of nests) is influenced by the following explanatory variables:

- x_1 is the altitude (in meters)
- x_2 is the slope (in degrees)
- x_3 is the number of pines in the square
- x_4 is the height (in meters) of the tree sampled at the center of the square
- x_5 is the diameter of the tree sampled at the center of the square
- x_6 is the index of the settlement density
- x_7 is the orientation of the square (from 1 if southbound to 2 otherwise)
- x_8 is the height (in meters) of the dominant tree
- x_9 is the number of vegetation strata
- x_{10} is the mix settlement index (from 1 if not mixed to 2 if mixed).

Caterpillar Regression Problem



Semilog-y plot of the data (x_i, y), $i=1,..,9$

An implementation is available
MatLab, C++

Regression

- The distribution of y given $\mathbf{x}$ is considered in the context of a set of experimental data, $i = 1, \ldots, n$, on which both y, and $x_{i1}, \ldots, x_{ik}$ are measured.

- The dataset is made from the outcomes $y = (y_1, \ldots, y_n)$ and of the $n \times (k+1)$ matrix of explanatory variables

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

Linear Regression

- The most common linear regression model is of the form:

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

- From this, we conclude that:

$$\mathbb{E}\left[y_i \mid \boldsymbol{\beta}, \boldsymbol{X} \right] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$
$$Var\left[y_i \mid \sigma^2, \boldsymbol{X} \right] = \sigma^2$$

- Note the difference between finite valued regressors x_i (like in our caterpillar problem) and categorical variables which also take finite number of values but with range that has no numerical meaning.

- Makes no sense to involve x directly in the regression: replace the single regressor x [belonging in {1, . . . ,m}, say] with m indicator (or dummy) variables.

Categorical Variables

- Note the difference between finite valued regressors x_i (like in our caterpillar problem) and categorical variables which also take finite number of values but with range that has no numerical meaning.

- Makes no sense to use x directly in the regression: replace the single regressor x [in $\{1, . . . ,m\}$, say] with m indicator variables

$$x_1 = \mathbb{I}_1(x),\, x_2 = \mathbb{I}_2(x), \ldots, x_m = \mathbb{I}_m(x)$$

- Use different β_i for each class categorical variable value:

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

- *Identifiability* requires eliminating one of the classes (e.g. $\beta_1=0$) since $\sum_i \mathbb{I}_1(x) = 1$

Linear Regression

- Returning back to our linear regression model:

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

- From this, we conclude that:

$$\mathbb{E}\left[y_i \mid \boldsymbol{\beta}, \boldsymbol{X} \right] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

$$Var\left[y_i \mid \sigma^2, \boldsymbol{X} \right] = \sigma^2$$

- Assume that $k+1<n$ and that $\boldsymbol{X}$ is of full rank: $rank(\boldsymbol{X}) = k + 1$
  - $\boldsymbol{X}$ is of full rank if and only if $\boldsymbol{X}^T \boldsymbol{X}$ is invertible

- The likelihood is then:

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

MLE Estimator

- The MLE of β is the solution of the least squares minimization problem

$$\min_{\beta}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = \min_{\beta}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_{i1}-\ldots-\beta_k x_{ik}\right)^2 \Rightarrow \boxed{\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}$$

- Since $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is a linear transform of $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) = \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

- Thus $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator and $\boxed{Var\left(\widehat{\boldsymbol{\beta}}|\sigma^2, \boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}}$

- Also $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of β:

$$\boldsymbol{\alpha} \in \mathbb{R}^{k+1}, Var\left(\boldsymbol{\alpha}^T\widehat{\boldsymbol{\beta}}|\sigma^2, \boldsymbol{X}\right) \leq Var\left(\boldsymbol{\alpha}^T\widetilde{\boldsymbol{\beta}}|\sigma^2, \boldsymbol{X}\right) \textit{ where } \widetilde{\boldsymbol{\beta}} \textit{ is any unbiased linear estimator of } \boldsymbol{\beta}$$

MLE Estimator

- The MLE of σ^2 is the solution of:

$$\min_{\sigma^2} \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{n}{2}\ln\sigma\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = 0 \Rightarrow \boxed{\sigma_{MLE}^2 = \frac{\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}{n}}$$

- Let $\boldsymbol{M} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{X}^T$ be the projection $n \times n$ matrix of the data **y** on the column space of $\boldsymbol{X}$

$$\mathbb{E}(\sigma_{MLE}^2) = \frac{\mathbb{E}\left[\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right]}{n} = \frac{\mathbb{E}[(\boldsymbol{y}-\boldsymbol{M}\boldsymbol{y})^T(\boldsymbol{y}-\boldsymbol{M}\boldsymbol{y})]}{n} =$$

$$= \frac{\mathbb{E}[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}]}{n} = \frac{\mathbb{E}[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}]}{n}$$

- You can easily show that (see this slide):

$$For\ \boldsymbol{Y}\ a\ n-dim\ random\ vector\ with\ \mathbb{E}\left[\boldsymbol{Y}\right] = \boldsymbol{\mu},\ Cov\left[\boldsymbol{Y}\right] = \boldsymbol{V},$$

$$and\ \boldsymbol{A}\ a\ n \times n\ matrix\ \ then: \mathbb{E}\left[\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}\right] = tr\left(\boldsymbol{A}\boldsymbol{V}\right) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$$

MLE Estimator

$$\mathbb{E}\left(\sigma^2_{MLE}\right) = \frac{\mathbb{E}\left[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}\right]}{n}$$

- You can easily show that (see this slide):

$$For\ \boldsymbol{Y}\ a\ n-dim\ random\ vector\ with\ \mathbb{E}\left[\boldsymbol{Y}\right] = \boldsymbol{\mu}, Cov\left[\boldsymbol{Y}\right] = \boldsymbol{V},$$
$$and\ \boldsymbol{A}\ a\ n\times n\ matrix\ \ then: \mathbb{E}\left[\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}\right] = tr\left(\boldsymbol{A}\boldsymbol{V}\right) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$$

- Applying this we have:

$$\mathbb{E}\left[\boldsymbol{y}^T\left(\boldsymbol{I}-\boldsymbol{M}\right)\boldsymbol{y}\right] = tr\left(\left(\boldsymbol{I}-\boldsymbol{M}\right)\overbrace{Cov(\boldsymbol{y})}^{\sigma^2\boldsymbol{I}_n}\right) + \mathbb{E}\left(\boldsymbol{y}\right)^T\left(\boldsymbol{I}-\boldsymbol{M}\right)\mathbb{E}\left(\boldsymbol{y}\right) =$$

$$= \sigma^2 rank\left(\boldsymbol{I}-\boldsymbol{M}\right) + \boldsymbol{\beta}^T\boldsymbol{X}^T\underbrace{\left(\boldsymbol{I}-\boldsymbol{M}\right)\boldsymbol{X}\boldsymbol{\beta}}_{0} = \sigma^2\left(n-k-1\right)$$

$Note: \boldsymbol{I}-\boldsymbol{M}$ *has* 1 *as its eigenvalue with multiplicity* $n-k-1$ *and*

$$tr\left(\boldsymbol{I}-\boldsymbol{M}\right) = tr\left(\boldsymbol{I}\right) - tr\left(\boldsymbol{M}\right) = n - tr\left(\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right) = n - tr\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right) = n - tr\left(\boldsymbol{I}_{k+1}\right) = n-k-1$$

MLE Estimator

- The expectation of the MLE of σ^2 is then given as:

$$\mathbb{E}\left(\sigma^2_{MLE}\right) = \frac{\sigma^2}{n}(n-k-1)$$

- The unbiased estimator of σ^2 is thus given:

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \frac{s^2}{n-k-1},\ where{:}\ s^2 \triangleq \left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

- Indeed using the result from the earlier slide we have:

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n-k-1}\mathbb{E}\left[\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right] = \frac{\sigma^2(n-k-1)}{n-k-1} = \sigma^2$$

- To approximate the covariance of $\widehat{\boldsymbol{\beta}}$, in $\widehat{\boldsymbol{\beta}} = \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$ use:

$$Var\left(\widehat{\boldsymbol{\beta}}|\sigma^2,\boldsymbol{X}\right) = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Appendix

$$\textit{For } \boldsymbol{Y} \textit{ a } n-\dim \textit{ random vector with } \mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}, Cov[\boldsymbol{Y}] = \boldsymbol{V},$$
$$\textit{and } \boldsymbol{A} \textit{ a } n \times n \textit{ matrix then: } \mathbb{E}\left[\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}\right] = tr(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu}$$

- We can prove this theorem quite easily as follows:

$$\mathbb{E}\left[(\boldsymbol{Y}-\boldsymbol{\mu})^T \boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{\mu})\right] = \mathbb{E}\left[\boldsymbol{Y}^T \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{Y} - \boldsymbol{Y}^T \boldsymbol{A}\boldsymbol{\mu}\right] \Rightarrow$$

$$\mathbb{E}\left[(\boldsymbol{Y}-\boldsymbol{\mu})^T \boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{\mu})\right] = \mathbb{E}\left[\boldsymbol{Y}^T \boldsymbol{A}\boldsymbol{Y}\right] + \boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{Y}^T \boldsymbol{A}\boldsymbol{Y}\right] - \boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{\mu}$$

- Using the identity $\mathbb{E}\left[tr(\boldsymbol{W})\right] = tr\left[\mathbb{E}(\boldsymbol{W})\right]$ for any random square matrix ***W***, we can write the l.h.s. of the equation above as:

$$\mathbb{E}\left[(\boldsymbol{Y}-\boldsymbol{\mu})^T \boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{\mu})\right] = \mathbb{E}\left[tr\left((\boldsymbol{Y}-\boldsymbol{\mu})^T \boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{\mu})\right)\right] = \mathbb{E}\left[tr\left(\boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T\right)\right] =$$
$$tr\left[\mathbb{E}\left(\boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T\right)\right] = tr\left[\boldsymbol{A}\mathbb{E}\left((\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^T\right)\right] = tr(\boldsymbol{A}\boldsymbol{V})$$

This proves the theorem.

T-Statistic

- We can define the standard t-statistic as follows:

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 \omega_{ii}}} \sim \mathcal{T}(n-k-1,0,1),\ where\text{:}\ \omega_{(i,i)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}|_{(i,i)}$$

- This statistic can be used for hypothesis testing, e.g.

$$accept\ H_0 : \beta_i = 0\ versus\ H_1 : \beta_i \neq 0\ at\ the\ level\ \alpha\ if$$

$$\frac{|\hat{\beta}_i|}{\sqrt{\hat{\sigma}^2 \omega_{ii}}} \leq F_{n-k-1}^{-1}(1-a/2), \text{the } (1-a/2)\text{nd quantile of } \mathcal{T}_{n-k-1}$$

T-Statistic

- The frequentist argument in using this bound is that there is significant evidence against H_0 if the p-value is smaller than α.

$$p_i = P_{H_0}\left(|T_i| > |t_i| = \frac{|\hat{\beta}_i|}{\sqrt{\hat{\sigma}^2\omega_{ii}}}\right) =$$

$$P_{H_0}(T_i < -|t_i|) + P_{H_0}(T_i > |t_i|) = F_{n-k-1}(-|t_i|) + \left(1 - F_{n-k-1}(|t_i|)\right) < \alpha$$

- Finally the statistic T_i can be used to derive the frequentist marginal confidence interval is:

$$\left\{\beta_i \colon |\beta_i - \hat{\beta}_i| \leq \sqrt{\hat{\sigma}^2\omega_{ii}}\, F_{n-k-1}^{-1}(1 - a/2)\right\}$$

MLE (Least Squares) Estimates

	$\hat{\beta}_i$	$\sqrt{\hat{\sigma}^2 \omega_{ii}}$	$t_i = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 \omega_{ii}}}$	$p_i = P_{H_0}\left(\lvert T_i \rvert > \frac{\lvert\hat{\beta}_i\rvert}{\sqrt{\hat{\sigma}^2 \omega_{ii}}}\right)$	
Coefficients	Estimate	Std.Error	t-value	Pr(>|t|)	
intercept	10.998412	3.060892	3.594	0.00161	**
XV1	-0.004431	0.001557	-2.846	0.00939	**
XV2	-0.053830	0.021904	-2.458	0.02232	*
XV3	0.067939	0.099492	0.683	0.50174	
XV4	-1.293636	0.563925	-2.294	0.03168	*
XV5	0.231637	0.104399	2.219	0.03709	*
XV6	-0.356800	1.566782	-0.228	0.82193	
XV7	-0.237469	1.006210	-0.236	0.81558	
XV8	0.181060	0.236772	0.765	0.45248	
XV9	-1.285316	0.865023	-1.486	0.15142	
XV10	-0.433106	0.735018	-0.589	0.56162	

An implementation is available MatLab, C++

Likelihood

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

- Note that the likelihood above can now be written in the following very useful form:

$$\ell(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X}) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\left((\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})+\widehat{\boldsymbol{\beta}}\right)\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\left((\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})+\widehat{\boldsymbol{\beta}}\right)\right)\right) \Rightarrow$$

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

$$= \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}s^2-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)$$

$$where:\ s^2 \triangleq \left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right),$$
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Conjugate Priors

- Observing the form of the likelihood suggests the following conjugate prior:

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{X}\sim\mathcal{N}_{k+1}\left(\widetilde{\boldsymbol{\beta}},\sigma^2\boldsymbol{M}^{-1}\right),\ \boldsymbol{M}\ a\ (k+1)\times(k+1)\ pos.\ def.\ symm.\ matrix$$
$$\sigma^2|\boldsymbol{X}\sim InvGamma(a,b),\ a,b>0$$

- The posterior is then

$$\pi\left(\boldsymbol{\beta},\sigma^2|\widehat{\boldsymbol{\beta}},s^2,\boldsymbol{X}\right)\propto\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}s^2-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)$$
$$\times\frac{1}{(2\pi\sigma^2)^{(k+1)/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)^T\boldsymbol{M}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)\right)\times\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\left(-\frac{1}{\sigma^2}b\right)=$$
$$=\sigma^{-k-1-2a-2-n}\exp\left(-\frac{1}{2\sigma^2}\left\{s^2+2b+\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})+\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)^T\boldsymbol{M}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)\right\}\right)=$$
$$=\sigma^{-k-n-2a-3}\exp\left(-\frac{1}{2\sigma^2}\left\{s^2+2b+\widehat{\boldsymbol{\beta}}^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta}-2\boldsymbol{\beta}^T\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)+\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T(\boldsymbol{X}^T\boldsymbol{X})\widehat{\boldsymbol{\beta}}\right\}\right)$$

Conjugate Priors: Posterior of β given σ^2

$$\pi\left(\boldsymbol{\beta},\sigma^2|\widehat{\boldsymbol{\beta}},s^2,\boldsymbol{X}\right) \propto \sigma^{-k-n-2a-3}\exp\left(-\frac{1}{2\sigma^2}\left\{s^2+2b+\boldsymbol{\beta}^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta}-2\boldsymbol{\beta}^T\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)+\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T(\boldsymbol{X}^T\boldsymbol{X})\widehat{\boldsymbol{\beta}}\right\}\right)=$$

$$\sigma^{-k-n-2a-3}\exp\left(-\frac{1}{2\sigma^2}\left\{\begin{array}{c} s^2+2b+\left(\boldsymbol{\beta}-\left(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right)^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})\left(\boldsymbol{\beta}-(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right) \\ +\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ -\left(M\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-T}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \end{array}\right\}\right)$$

- Based on this, the posterior of β given σ^2 is:

$$\pi(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}) \propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}])^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}-\mathbb{E}[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}])\right)$$

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}\sim\mathcal{N}_{k+1}\left((\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left\{\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right\},\sigma^2(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

$$where$$

$$\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]=(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

$$and$$

$$Var[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]=\sigma^2(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Conjugate Priors: Marginal posterior of σ²

$$\pi\left(\boldsymbol{\beta},\sigma^2|\widehat{\boldsymbol{\beta}},s^2,\boldsymbol{X}\right)\propto$$

$$\sigma^{-k-n-2a-3}\exp\left(-\frac{1}{2\sigma^2}\begin{Bmatrix} s^2+2b+(\boldsymbol{\beta}-\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}])^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}-\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]) \\ +\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ -\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\end{Bmatrix}\right)$$

- Integrating out β gives the following marginal for σ^2:

$$\pi\left(\sigma^2|\widehat{\boldsymbol{\beta}},s^2,\boldsymbol{X}\right)$$

$$\propto\sigma^{-n-2a-2}\exp\left(-\frac{1}{2\sigma^2}\begin{Bmatrix}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+s^2 \\ +2b-\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\end{Bmatrix}\right)$$

- To simplify, we will use the following two identities

$$A:\ \left(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}-\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{M}^{-1}+\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$$

$$B:\ \boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{M}=\left(\boldsymbol{M}^{-1}+\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)^{-1}$$

Conjugate Priors: Marginal Posterior of σ^2

$$\pi\left(\sigma^2|\widehat{\boldsymbol{\beta}},s^2,\boldsymbol{X}\right)\propto\sigma^{-n-2a-2}\exp\left(-\frac{1}{2\sigma^2}\left\{\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+s^2+2b-\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right\}\right)$$

- We can simplify the last term above using identities A & B:

$$\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)=\left(expand\right)$$

$$2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}=$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}-(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}-(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}=$$

$$2\widehat{\boldsymbol{\beta}}^T\underbrace{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{M}}_{(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}}\widetilde{\boldsymbol{\beta}}+\widetilde{\boldsymbol{\beta}}^T\underbrace{\boldsymbol{M}^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{M}}_{\boldsymbol{M}-\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{M}=\boldsymbol{M}-(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\widehat{\boldsymbol{\beta}}=$$

$$2\widehat{\boldsymbol{\beta}}^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\widetilde{\boldsymbol{\beta}}+\widetilde{\boldsymbol{\beta}}^T\boldsymbol{M}\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\widehat{\boldsymbol{\beta}}=$$

$$\widehat{\boldsymbol{\beta}}^T\boldsymbol{M}\widetilde{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)$$

- Finally:

$$\pi\left(\sigma^2|\widehat{\boldsymbol{\beta}},s^2,\boldsymbol{X}\right)\propto\sigma^{-n-2a-2}\exp\left\{-\frac{1}{2\sigma^2}\left(\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)+s^2+2b\right)\right\}$$

Conjugate Priors: Marginal Posterior of σ^2

$$\pi\left(\sigma^2|\widehat{\boldsymbol{\beta}},s^2,\boldsymbol{X}\right) \propto \sigma^{-n-2a-2}\exp\left\{-\frac{1}{2\sigma^2}\left(\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)+s^2+2b\right)\right\}$$

- The marginal posterior is:

$$\sigma^2|\boldsymbol{y},\boldsymbol{X}\sim InvGamma\left(\frac{n}{2}+a, b+\frac{s^2}{2}+\frac{\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)}{2}\right)$$

- The marginal posterior mean for $n \geq 2$ is then:

$$\mathbb{E}^{\pi}[\sigma^2|\boldsymbol{y},\boldsymbol{X}] = \frac{2b+s^2+\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(X^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)}{n+2a-2}$$

Inverse-gamma	$\theta \sim \text{Inv-gamma}(\alpha,\beta)$ $p(\theta) = \text{Inv-gamma}(\theta\|\alpha,\beta)$	shape $\alpha > 0$ scale $\beta > 0$

$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta}, \quad \theta > 0$	$\text{E}(\theta) = \frac{\beta}{\alpha-1}$, for $\alpha > 1$ $\text{var}(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2$ $\text{mode}(\theta) = \frac{\beta}{\alpha+1}$

Conjugate Priors: MLE, Posterior Mean and Ridge Estimator

- Note that setting $\boldsymbol{M} = \boldsymbol{I}_{k+1}/c\,,\ c > 0\ and\ \widetilde{\boldsymbol{\beta}} = \boldsymbol{0}_{k+1}$ in the conditional posterior mean

$$\mathbb{E}[\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}] = (\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{M}\widetilde{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

we obtain the classical Ridge Regression estimate:

$$\mathbb{E}[\beta|\sigma^2, \boldsymbol{y}, \boldsymbol{X}] = \left(\frac{1}{c}\boldsymbol{I}_{k+1} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \left(\frac{1}{c}\boldsymbol{I}_{k+1} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- The general estimator $\mathbb{E}\left[\boldsymbol{\beta}\,|\,\sigma^2, \boldsymbol{y}, \boldsymbol{X}\right]$ can be seen as a weighted average of the prior mean and the MLE.

Conjugate Priors: Marginal Posterior of β

$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \int \pi(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})\pi(\sigma^2|\boldsymbol{y},\boldsymbol{X})d\sigma^2$$

$$\pi(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}) \propto \mathcal{N}_{k+1}\left((\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\{\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\},\sigma^2(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

$$\pi(\sigma^2|\boldsymbol{y},\boldsymbol{X}) \propto \mathcal{IG}\left(\frac{n}{2}+a, b+\frac{s^2}{2}+\frac{\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)}{2}\right)$$

- To integrate, we use the following transformations:

$$\tau = 1/\sigma^2, \quad d\tau = -\tau^2 d\sigma^2, \quad \widehat{\boldsymbol{\mu}} = \{\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X}\}^{-1}\left[\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right]$$

$$Z = \tau\frac{1}{2}\frac{\left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\mu}})^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\mu}})+2b+s^2+\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)\right]}{G}$$

$$d\tau = \frac{2}{G}dZ$$

Inverse-gamma	$\theta \sim \text{Inv-gamma}(\alpha,\beta)$ $p(\theta) = \text{Inv-gamma}(\theta\|\alpha,\beta)$	shape $\alpha > 0$ scale $\beta > 0$

$$\tau = 1/\sigma^2, \quad d\tau = -\tau^2 d\sigma^2, \quad \widehat{\boldsymbol{\mu}} = \{\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X}\}^{-1}\left[\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right]$$

$$Z = \tau\frac{1}{2}\frac{\left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\mu}})^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\mu}})+2b+s^2+\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)\right]}{G}$$

$$d\tau = \frac{2}{G}dZ$$

Conjugate Priors: Marginal Posterior of β

$$\tau = 1/\sigma^2, \quad d\tau = -\tau^2 d\sigma^2, \quad \widehat{\boldsymbol{\mu}} = \{\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X}\}^{-1}\left[\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right]$$

$$Z = \tau\frac{1}{2}\left[(\boldsymbol{\beta} - \widehat{\boldsymbol{\mu}})^T(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\mu}}) + 2b + s^2 + \left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right)\right]$$

$$d\tau = \frac{2}{G}dZ$$

- The marginal now takes the form:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) \propto \int_0^\infty e^{-Z^2}\tau^{\left(\frac{k+1}{2}\right)+\left(\frac{n}{2}+a+1\right)}d\sigma^2 \sim \int_0^\infty e^{-Z^2}\tau^{\frac{k}{2}+\frac{n}{2}+a-\frac{1}{2}}d\tau$$

$$\sim G^{-\frac{k}{2}-\frac{n}{2}-a-\frac{1}{2}}\int_0^\infty e^{-Z^2}Z^{\frac{k}{2}+\frac{n}{2}+a-\frac{1}{2}}dZ \sim G^{-\frac{k}{2}-\frac{n}{2}-a-\frac{1}{2}}$$

where

$$G = (\boldsymbol{\beta} - \widehat{\boldsymbol{\mu}})^T(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\mu}}) + 2b + s^2 + \left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right)$$

- The last integral above is a scalar quantity.

Conjugate Priors: Marginal Posterior of β

- Thus integrating out σ^2 leads to a multivariate Student's t marginal posterior on β :

$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) \propto \left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\mu}})^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\mu}})+2b+s^2+\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)\right]^{-\frac{k}{2}-\frac{n}{2}-a-\frac{1}{2}}$$

where:

$$\widehat{\boldsymbol{\mu}} = \{\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X}\}^{-1}\left[\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right]$$

- Recall that the density of the multivariate $\mathscr{T}_p(\nu,\theta,\boldsymbol{\Sigma})$ is:

$$\mathscr{T}_p(\nu,\theta,\boldsymbol{\Sigma}) = \frac{\Gamma\left((\nu+p)/2\right)/\Gamma(\nu/2)}{\sqrt{\det(\boldsymbol{\Sigma})\nu\pi}}\left[1+\frac{(t-\theta)^T\boldsymbol{\Sigma}^{-1}(t-\theta)}{\nu}\right]^{-\frac{\nu+p}{2}}$$

Conjugate Priors: Marginal Posterior of β

- We can now see that our marginal

$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) \propto \left[(\boldsymbol{\beta}-\widehat{\boldsymbol{\mu}})^T(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\mu}})+2b+s^2+\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)\right]^{-\frac{k}{2}-\frac{n}{2}-a-\frac{1}{2}}$$

can be written (can check with substitution of the Eqs below) in the form of

$$\mathscr{T}_p(\nu,\theta,\boldsymbol{\Sigma})=\frac{\Gamma\left((\nu+p)/2\right)/\Gamma(\nu/2)}{\sqrt{\det(\boldsymbol{\Sigma})\nu\pi}}\left[1+\frac{(t-\theta)^T\boldsymbol{\Sigma}^{-1}(t-\theta)}{\nu}\right]^{-\frac{\nu+p}{2}}$$

as follows ($p=k+1$, $v=n+2a$):

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}\sim\mathcal{T}_{k+1}\left(n+2a,\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}\right)$$

$$\widehat{\boldsymbol{\mu}}=(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left((\boldsymbol{X}^T\boldsymbol{X})\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right)$$

$$\widehat{\boldsymbol{\Sigma}}=\frac{2b+s^2+\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)}{n+2a}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Multivariate Student's-$\mathcal{T}$ Distribution

- Recall the properties of the multivariate Student's-$\mathcal{T}$ distribution:*

Multivariate Student-t	$\theta \sim t_\nu(\mu, \Sigma)$ $p(\theta) = t_\nu(\theta\|\mu, \Sigma)$ (implicit dimension d)	degrees of freedom $\nu > 0$ location $\mu = (\mu_1, .., \mu_d)$ symmetric, pos. definite $d \times d$ scale matrix Σ

$$p(\theta) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}}|\Sigma|^{-1/2} \times \left(1 + \frac{1}{\nu}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)^{-(\nu+d)/2}$$

$$\mathrm{E}(\theta) = \mu, \text{ for } \nu > 1$$
$$\mathrm{var}(\theta) = \frac{\nu}{\nu-2}\Sigma, \text{ for } \nu > 2$$
$$\mathrm{mode}(\theta) = \mu$$

*Note that in the notation of the tables above the degrees of freedom of the distribution rather than the dimensionality are shown as subscripts, e,g, t_v vs t_d in the earlier slide. Hopefully the notation will be clear from the discussion.

Bayesian Data Analysis, A. Gelman, J. Carlin, H. Stern and D. Rubin, 2004

Conjugate Priors: Predictive Distribution

- For a given *(m,k+1)* explanatory matrix $\tilde{\boldsymbol{X}}$, the outcome $\tilde{\boldsymbol{y}}$ can be inferred through the predictive distribution*

$$\pi\left(\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}\right)$$

- Since $\pi\left(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{X}},\boldsymbol{\beta},\sigma^2\right)\propto\mathcal{N}\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_m\right),$

  and since the posterior of $\boldsymbol{\beta}$ conditional on σ^2 is given as

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

  we can see that:

$$\pi\left(\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}\right)\propto\int\pi\left(\tilde{\boldsymbol{y}}|\boldsymbol{\beta},\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}\right)\pi(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X})d\boldsymbol{\beta}$$

  is a Gaussian.

* We will later on integrate σ^2 out to compute: $\pi\left(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}\right)$

Appendix

- We will make use next of the following conditional expectation equalities:

$$\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}] = \mathbb{E}[\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}]|\boldsymbol{Z}]$$

$$Var[\boldsymbol{X}|\boldsymbol{Z}] = Var[\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}]|\boldsymbol{Z}]+\mathbb{E}[Var[\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}]|\boldsymbol{Z}]$$

- The proof of these is quite straightforward, e.g.

$$\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}] = \int_{\boldsymbol{x}} \boldsymbol{x}p(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{x} = \int_{\boldsymbol{x}} \boldsymbol{x}\int_{\boldsymbol{y}} p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})\boldsymbol{dy}\,d\boldsymbol{x}$$

$$\int_{\boldsymbol{x}} \boldsymbol{x}\int_{\boldsymbol{y}} p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{z})\boldsymbol{dy}\,d\boldsymbol{x} = \int_{\boldsymbol{y}}\left(\int_{\boldsymbol{x}} \boldsymbol{x}p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})\boldsymbol{dx}\right)\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{z})d\boldsymbol{y} = \mathbb{E}[\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}]|\boldsymbol{Z}]$$

Conjugate Priors: Predictive Distribution

- We can compute the mean and the variance of this predictive distribution as follows:

$$\mathbb{E}_\pi\left[\widetilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right]=\mathbb{E}_\pi\left(\mathbb{E}_\pi\left(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta},\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)=$$
$$\mathbb{E}_\pi\left(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)=\widetilde{\boldsymbol{X}}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right)$$

and

$$Var_\pi\left[\widetilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right]=$$
$$\mathbb{E}_\pi\left(Var_\pi\left(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta},\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)+Var_\pi\left(\mathbb{E}_\pi\left(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta},\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)=$$
$$\mathbb{E}_\pi\left(\sigma^2\boldsymbol{I}_m|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)+Var_\pi\left(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)=$$
$$\sigma^2\boldsymbol{I}_m+\widetilde{\boldsymbol{X}}\sigma^2(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\widetilde{\boldsymbol{X}}^T=\sigma^2\left(\boldsymbol{I}_m+\widetilde{\boldsymbol{X}}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\widetilde{\boldsymbol{X}}^T\right)$$

- In conclusion:

$$\widetilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\sim\mathcal{N}_m\left(\widetilde{\boldsymbol{X}}\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}],\sigma^2\left(\boldsymbol{I}_m+\widetilde{\boldsymbol{X}}Var[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]\widetilde{\boldsymbol{X}}^T\right)\right)$$
$$\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]=(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\left\{\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right\}$$

Conjugate Priors: Predictive Distribution

$$\widetilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\sim\mathcal{N}\left(\widetilde{\boldsymbol{X}}\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}],\sigma^2\boldsymbol{I}_m+\widetilde{\boldsymbol{X}}Var[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]\widetilde{\boldsymbol{X}}^T\right)$$

$$\mathbb{E}\left[\widetilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right]=\widetilde{\boldsymbol{X}}\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]=\widetilde{\boldsymbol{X}}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\{\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{M}\widetilde{\boldsymbol{\beta}}\},$$

$$\mathrm{Var}\left[\widetilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right]=\widetilde{\boldsymbol{X}}Var[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]\widetilde{\boldsymbol{X}}^T=\sigma^2\widetilde{\boldsymbol{X}}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\widetilde{\boldsymbol{X}}^T$$

- We now integrate σ^2 against the posterior distribution

$$\sigma^2|\boldsymbol{y},\boldsymbol{X}\sim\mathcal{IG}\left(\frac{n}{2}+a,b+\frac{s^2}{2}+\frac{\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)}{2}\right)$$

to obtain:

$$\pi\left(\widetilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)\propto\int_0^\infty\mathcal{N}\left(\widetilde{\boldsymbol{X}}\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}],\sigma^2\boldsymbol{I}_m+\widetilde{\boldsymbol{X}}Var[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]\widetilde{\boldsymbol{X}}^T\right)$$

$$\times\,\mathcal{IG}\left(\frac{n}{2}+a,b+\frac{s^2}{2}+\frac{\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)}{2}\right)d\sigma^2$$

Conjugate Priors: Predictive Distribution

$$\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}) \propto \int_0^\infty \mathcal{N}\left(\tilde{\boldsymbol{X}}\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}], \sigma^2\boldsymbol{I}_m + \tilde{\boldsymbol{X}}Var[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]\tilde{\boldsymbol{X}}^T\right)$$
$$\times \mathcal{IG}\left(\frac{n}{2}+a, b+\frac{s^2}{2}+\frac{(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})}{2}\right)d\sigma^2$$

- We substitute:

$$\sigma^2\boldsymbol{I}_m + \tilde{\boldsymbol{X}}Var[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}]\tilde{\boldsymbol{X}}^T = \sigma^2\left(\boldsymbol{I}_m + \tilde{\boldsymbol{X}}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^T\right)$$

$$\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}) \propto \int_0^\infty \mathcal{N}\left(\tilde{\boldsymbol{X}}\mathbb{E}[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}], \sigma^2\left(\boldsymbol{I}_m + \tilde{\boldsymbol{X}}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^T\right)\right)$$
$$\times \mathcal{IG}\left(\frac{n}{2}+a, b+\frac{s^2}{2}+\frac{(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})}{2}\right)d\sigma^2$$
$$\propto \int_0^\infty \sigma^{-m-n-2a-2}\exp\begin{pmatrix} -\frac{1}{2\sigma^2}\{2b+s^2+(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})+ \\ (\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}])^T\left(\boldsymbol{I}_m+\tilde{\boldsymbol{X}}\left(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\tilde{\boldsymbol{X}}^T\right)^{-1}(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}])\}\end{pmatrix}d\sigma^2$$

Conjugate Priors: Predictive Distribution

$$\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}) \propto$$

$$\propto \int_0^\infty \sigma^{-m-n-2a-2} \exp\begin{pmatrix} -\frac{1}{2\sigma^2}\{2b+s^2+(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{M}^{-1}+(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})+ \\ (\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}])^T\left(\boldsymbol{I}_m+\tilde{\boldsymbol{X}}\left(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\tilde{\boldsymbol{X}}^T\right)^{-1}(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}])\} \end{pmatrix} d\sigma^2$$

- If we call the term inside {.} as G, then:

$$\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}) \propto \int_0^\infty \sigma^{-m-n-2a-2}\, e^{-\frac{1}{2\sigma^2}G} d\sigma^2 \Rightarrow$$

$$\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}) \propto \int_0^\infty (\sigma^2)^{-\frac{m+n+2a+2}{2}} e^{-\frac{1}{2\sigma^2}G} d\sigma^2 \Rightarrow \pi(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}) \propto G^{-\frac{m+n+2a+2}{2}} \int_0^\infty Z^{\frac{m+n+2a+2}{2}} e^{-Z}\left(-\frac{1}{2}G\frac{1}{Z^2}dZ\right) \sim G^{-\frac{m+n+2a}{2}}$$

$$Z=\frac{1}{2\sigma^2}G$$

$$\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}) \propto \{2b+s^2+(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\left(\boldsymbol{M}^{-1}+\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)^{-1}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})+$$

$$(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}])^T(\boldsymbol{I}_m+\tilde{\boldsymbol{X}}(\boldsymbol{M}+\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^T)^{-1}(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}])\}^{-(m+n+2a)/2}$$

Conjugate Priors: Predictive Distribution

$$\pi(\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}) \propto \{2b + s^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \left(\boldsymbol{M}^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{-1} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) +$$
$$(\tilde{\boldsymbol{y}} - \mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}])^T \left(\boldsymbol{I}_m + \tilde{\boldsymbol{X}}(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^T\right)^{-1} (\tilde{\boldsymbol{y}} - \mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}])\}^{-(m+n+2a)/2}$$

- This corresponds to a Student's t distribution:

$$\mathcal{T}_p(\nu,\theta,\boldsymbol{\Sigma}) = \frac{\Gamma\left((\nu+p)/2\right)/\Gamma(\nu/2)}{\sqrt{\det(\boldsymbol{\Sigma})\nu\pi}}\left[1 + \frac{(t-\theta)^T\boldsymbol{\Sigma}^{-1}(t-\theta)}{\nu}\right]^{-\frac{\nu+p}{2}}$$

with

$$\tilde{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}} \sim \mathcal{T}_m(n+2a,\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$$
$$\hat{\boldsymbol{\mu}} = \mathbb{E}[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{X}}] = \tilde{\boldsymbol{X}}(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{M}\tilde{\boldsymbol{\beta}})$$
$$\hat{\boldsymbol{\Sigma}} = \frac{\left\{2b + s^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{M}^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\right\}\left(\boldsymbol{I}_m + \tilde{\boldsymbol{X}}(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^T\right)}{n+2a}$$

Implementation of the Conjugate Prior

Our Prior:

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{X}\sim\mathcal{N}_{k+1}\left(\tilde{\boldsymbol{\beta}},\sigma^2\boldsymbol{M}^{-1}\right),\ \boldsymbol{M}\ a\ (k+1)\times(k+1)\ pos.\ def.\ symm.\ matrix$$

$$\sigma^2|\boldsymbol{X}\sim\mathcal{IG}(a,b),\ a,b>0$$

- We assume that there is no precise information available about $\tilde{\boldsymbol{\beta}},\boldsymbol{M},a,b$.

- Let us choose as $\boldsymbol{M}=\boldsymbol{I}_{k+1}/c$ and $\tilde{\boldsymbol{\beta}}=\mathbf{0}_{k+1}$.

$$\boldsymbol{\beta}\,|\,\sigma^2,\boldsymbol{X}\sim\mathcal{N}_{k+1}\left(\boldsymbol{0}_{k+1},c\sigma^2\boldsymbol{I}_{k+1}\right)$$

- Let us take a = 2.1 and b = 2, i.e. prior mean and prior variance of σ^2 equal to 1.82 and 33.06.

$$p(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta},\quad\theta>0 \qquad \begin{aligned}&\mathrm{E}(\theta)=\frac{\beta}{\alpha-1},\ \text{for}\ \alpha>1\\&\mathrm{var}(\theta)=\frac{\beta^2}{(\alpha-1)^2(\alpha-2)},\alpha>2\\&\mathrm{mode}(\theta)=\frac{\beta}{\alpha+1}\end{aligned}$$

- The intuition is that if c is large, the prior on β should be more diffuse and have less bearing on the outcome.

- However, this turns out not to be the case. There is lasting influence of c on the posterior means of σ^2 and β_0.

Influence of the Prior Scale c on Bayesian Estimates of β

**************** Conjugate Priors *******************

	$\mathbb{E}_\pi\left(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{X}\right)$	$\mathbb{E}_\pi\left(\boldsymbol{\beta}_0 \mid \boldsymbol{y}, \boldsymbol{X}\right)$	$V_\pi\left(\boldsymbol{\beta}_0 \mid \boldsymbol{y}, \boldsymbol{X}\right)$
c	𝔼[sigma^2 | y, X]	𝔼[beta_0 | y, X]	V[beta_0 | y, X]
0.1	1.0044	0.1251	0.0988
1.0	0.8541	0.9031	0.7733
10.0	0.6976	4.7299	3.8991
100.0	0.5746	9.6626	6.8355
1000.0	0.5470	10.8476	7.3419

An implementation is available MatLab, C++

$$\mathbb{E}^\pi[\sigma^2|\boldsymbol{y},\boldsymbol{X}] = \frac{2b + s^2 + \left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right)^T (\boldsymbol{M}^{-1} + (X^T\boldsymbol{X})^{-1})^{-1}\left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right)}{n + 2a - 2}$$

$$\mathbb{E}^\pi[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}] = (\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X})^{-1}\left((\boldsymbol{X}^T\boldsymbol{X})\widehat{\boldsymbol{\beta}} + \boldsymbol{M}\widetilde{\boldsymbol{\beta}}\right)$$

$$\mathrm{Var}[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}] = \frac{2b+s^2+(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\left(\boldsymbol{M}^{-1}+\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)^{-1}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})}{n+2a}(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Bayesian Estimates of β for c=100

************** Conjugate Priors**************

Bayes estimates of beta for c=100

beta_i	$\mathbb{E}_\pi(\beta_i \mid y, X)$ $\mathbb{E}$[beta_i | y, X]	$V_\pi(\beta_i \mid y, X)$ V[beta_i | y, X]
beta_0	9.6626	6.8355
beta_1	-0.0040	0.0000
beta_2	-0.0516	0.0004
beta_3	0.0418	0.0077
beta_4	-1.2633	0.2615
beta_5	0.2307	0.0090
beta_6	-0.0832	1.9310
beta_7	-0.1917	0.8254
beta_8	0.1608	0.0462
beta_9	-1.2069	0.6127
beta_10	-0.2567	0.4267

An implementation is available MatLab, C++

Influence of the Conjugate Prior

- The value of *c* thus has a significant influence on the estimators and even more on the posterior variance.
- The Bayes estimates stabilize for very large values of *c*.

  Thus the prior associated with a particular c should not be considered as a weak or pseudo-noninformative prior but, on the opposite, associated with a specific proper prior information.

- The dependence on *(a, b)* is equally strong.

  Considering these limitations of conjugate priors on at least the posterior variance, **a more sophisticated noninformative strategy is needed**.

- We first look a middle-ground perspective which settles the problem of the choice of ***M***.

Zellner's Informative G-Prior

- We start with a middle ground solution by introducing only information about the location parameter of the regression but bypassing the selection of the prior correlation structure.

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{X}\sim\mathcal{N}_{k+1}\left(\widetilde{\boldsymbol{\beta}},c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

$$\sigma^2\sim\pi(\sigma^2|\boldsymbol{X})\propto\sigma^{-2}\text{ improper Jeffreys prior}$$

- Zellner, A. (1971). An Introduction to Bayesian Econometrics. John Wiley, New York.
- Zellner, A. (1984). Basic Issues in Econometrics. University of Chicago Press, Chicago.
- Carlin, B. and Louis, T. (1996). Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall, New York.

Zellner's Informative G-Prior

- We start with a middle ground solution by introducing only information about the location parameter of the regression but bypassing the selection of the prior correlation structure.

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{X} \sim \mathcal{N}_{k+1}\left(\tilde{\boldsymbol{\beta}}, c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$
$$\sigma^2 \sim \pi(\sigma^2|\boldsymbol{X}) \propto \sigma^{-2} \; improper\, Jeffreys\, prior$$

- The prior determination is restricted to the choices of $\tilde{\boldsymbol{\beta}}$ and constant *c*. *c* can be interpreted as a measure of the information available in the prior relative to the sample.
  - E.g., we will see that setting 1/c = 0.5 gives the prior the same weight as 50% of the sample.
- There is still strong influence of c.

Zellner's Informative G-Prior: Conditional Posterior of β

- The joint posterior now takes the form (note $\boldsymbol{X}^T\boldsymbol{X}$ is used in both likelihood and prior):

$$\pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X}) \propto (\sigma^2)^{-(n/2)}\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)(\sigma^2)^{-(k+1)/2}$$
$$\times\exp\left(-\frac{1}{2c\sigma^2}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)\right)(\sigma^2)^{-1}$$

- We can now compute the following:

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}\sim\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)-\frac{1}{2c\sigma^2}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\right)\right)$$
$$\sim\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right)\right)^T\frac{c+1}{c\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right)\right)\right)\Rightarrow$$

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}\sim\mathcal{N}_{k+1}\left(\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right),\frac{c\sigma^2}{c+1}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$$

Zellner's Informative G-Prior: Posterior Marginal of σ^2

- Starting again with the posterior

$$\pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X}) \propto (\sigma^2)^{-(n+k+3/2)}\exp\left(-\frac{1}{2\sigma^2}s^2 - \frac{1}{2\sigma^2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)$$

$$\times\exp\left(-\frac{1}{2c\sigma^2}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})\right) =$$

$$\sim(\sigma^2)^{-(k+1)/2}\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right)\right)^T\frac{c+1}{c\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right)\right)\right)$$

$$\times(\sigma^2)^{-(n/2+1)}\exp\left(-\frac{1}{2\sigma^2}\left(s^2+\frac{1}{c+1}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})\right)\right)$$

- We can now compute the following posterior marginal with integration in β:

$$\sigma^2|\boldsymbol{y},\boldsymbol{X}\sim InvGamma\left(\frac{n}{2},\frac{s^2}{2}+\frac{1}{2(c+1)}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})\right)$$

$$p(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta},\ \theta>0 \qquad \mathrm{E}(\theta)=\frac{\beta}{\alpha-1},\ \text{for } \alpha>1 \quad \mathrm{var}(\theta)=\frac{\beta^2}{(\alpha-1)^2(\alpha-2)},\alpha>2 \quad \mathrm{mode}(\theta)=\frac{\beta}{\alpha+1}$$

$$\mathbb{E}[\sigma^2|\boldsymbol{y},\boldsymbol{X}] = \frac{s^2+\frac{1}{(c+1)}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})}{n-2}$$

Zellner's Informative G-Prior: Posterior Marginal of β

- Starting again with the posterior

$$\pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},X) \propto (\sigma^2)^{-(n+k+3/2)}\exp\{-\frac{1}{2\sigma^2}[\left(\boldsymbol{\beta}-\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right)\right)^T\frac{c+1}{c}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right)\right)+ s^2+\frac{1}{c+1}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})]\right\}$$

- We can now compute the following posterior marginal with integration in σ^2:

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}\sim\left\{\left(\boldsymbol{\beta}-\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right)\right)^T\frac{c+1}{c}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right)\right)+s^2+\frac{1}{c+1}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})\boldsymbol{\beta}\right\}^{-(n+k+1/2}$$

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}\sim\mathcal{T}_{k+1}\left(n,\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right),\frac{c\left(s^2+(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)/(c+1)\right)}{n(c+1)}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

Multivariate Student-t	$\theta\sim t_\nu(\mu,\Sigma)$ $p(\theta)=t_\nu(\theta\|\mu,\Sigma)$ (implicit dimension d)	degrees of freedom $\nu>0$ location $\mu=(\mu_1,\ldots,\mu_d)$ symmetric, pos. definite $d\times d$ scale matrix Σ	$p(\theta)=\frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}}\|\Sigma\|^{-1/2}$ $\times(1+\frac{1}{\nu}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu))^{-(\nu+d)/2}$	$\mathrm{E}(\theta)=\mu$, for $\nu>1$ $\mathrm{var}(\theta)=\frac{\nu}{\nu-2}\Sigma$, for $\nu>2$ $\mathrm{mode}(\theta)=\mu$

Zellner's Informative G-Prior: Bayes Estimates

- The Bayes estimate of $\boldsymbol{\beta}$ can be derived from:

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X} \sim \mathcal{T}_{k+1}\left(n, \frac{c}{c+1}\left(\frac{\tilde{\boldsymbol{\beta}}}{c}+\hat{\boldsymbol{\beta}}\right), \frac{c\left(s^2+(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})/(c+1)\right)}{n(c+1)}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

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

- The posterior variance of $\boldsymbol{\beta}$ is also given from above as:

$$V_\pi[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}] = \frac{c}{c+1}\frac{\left(s^2+(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})/(c+1)\right)}{n-2}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$p(\theta) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}}|\Sigma|^{-1/2} \times\left(1+\frac{1}{\nu}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)^{-(\nu+d)/2}$$

$$\mathrm{E}(\theta) = \mu, \text{ for } \nu > 1$$
$$\mathrm{var}(\theta) = \frac{\nu}{\nu-2}\Sigma, \text{ for } \nu > 2$$
$$\mathrm{mode}(\theta) = \mu$$

Zellner's Informative G-Prior: Bayes Estimates

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

- If c = 1, you can see from the above equation that it is like putting the same weight on the prior information and on the sample:

$$\mathbb{E}[\beta|\boldsymbol{y},\boldsymbol{X}] = \frac{1}{2}\left(\tilde{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}\right)$$

  which is the average of the prior mean and the MLE estimator.

- If c=100, the prior weights 1% of the sample.

Zellner's Informative G-Prior: Bayes Estimates

- The Bayes estimate of σ^2 (use the mean of InvGamma) can be derived from

$$\sigma^2|\boldsymbol{y},\boldsymbol{X} \sim \mathcal{IG}\left(\frac{n}{2},\frac{s^2}{2}+\frac{1}{2(c+1)}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)\right) \Rightarrow$$

as

$$\mathbb{E}[\sigma^2|\boldsymbol{y},\boldsymbol{X}] = \frac{s^2+\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right)/(c+1)}{n-2}$$

- Note that only when c goes to infinity the influence of the prior vanishes.

$$p(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta},\quad \theta>0$$

$$\mathrm{E}(\theta)=\frac{\beta}{\alpha-1},\ \text{for } \alpha>1$$
$$\mathrm{var}(\theta)=\frac{\beta^2}{(\alpha-1)^2(\alpha-2)},\ \alpha>2$$
$$\mathrm{mode}(\theta)=\frac{\beta}{\alpha+1}$$

Zellner's Informative G-Prior: Marginal Posterior Mean and Variance of β

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

$$V_\pi[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}] = \frac{c}{c+1}\frac{\left(s^2 + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T\boldsymbol{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})/(c+1)\right)}{n-2}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

**************** Zellner's G-Prior ****************

Posterior mean and variance of beta for c=100

beta_i	𝔼[beta_i|y, X] $\mathbb{E}_\pi(\beta_i \mid y, X)$	V[beta_i|y, X] $V_\pi(\beta_i \mid y, X)$	log10(BF) $\log_{10}(BF)$
(Intercept)	10.8895	6.8229	2.1873 (****)
X1	-0.0044	0.0000	1.1571 (***)
X2	-0.0533	0.0003	0.6667 (**)
X3	0.0673	0.0072	-0.8585
X4	-1.2808	0.2316	0.4726 (*)
X5	0.2293	0.0079	0.3861 (*)
X6	-0.3533	1.7877	-0.9860
X7	-0.2351	0.7373	-0.9849
X8	0.1793	0.0408	-0.8225
X9	-1.2726	0.5449	-0.3461
X10	-0.4288	0.3934	-0.8949

See here for BF calculation

An implementation is available MatLab, C++

Evidence against H_0:

(****) decisive
(***) strong
(**) substantial
(*) poor

Zellner's Informative G-Prior: Marginal Posterior Mean and Variance of β

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

$$V_\pi[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}] = \frac{c}{c+1}\frac{\left(s^2 + (\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})/(c+1)\right)}{n-2}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

*************** Zellner's G-Prior ****************

Posterior mean and variance of beta for c=1000

$\mathbb{E}_\pi(\boldsymbol{\beta}_i \mid \boldsymbol{y},\boldsymbol{X})$ $V_\pi(\boldsymbol{\beta}_i \mid \boldsymbol{y},\boldsymbol{X})$ $\log_{10}(BF)$

See here for BF calculation

beta_i	𝔼[beta_i|y, X]	V[beta_i|y, X]	log10(BF)
(Intercept)	10.9874	6.6644	1.7973 (***)
X1	-0.0044	0.0000	0.7375 (**)
X2	-0.0538	0.0003	0.2313 (*)
X3	0.0679	0.0070	-1.3506
X4	-1.2923	0.2262	0.0307
X5	0.2314	0.0078	-0.0588
X6	-0.3564	1.7461	-1.4834
X7	-0.2372	0.7202	-1.4822
X8	0.1809	0.0399	-1.3130
X9	-1.2840	0.5323	-0.8177
X10	-0.4327	0.3843	-1.3885

An implementation is available MatLab, C++

Evidence against H_0:

(****) decisive
(***) strong
(**) substantial
(*) poor

Zellner's Informative G-Prior: Predictive Modeling

- We want to predict (m ≥ 1) future observations when the explanatory variables $\tilde{X}$ but not the outcome variables

$$\tilde{\boldsymbol{y}} \sim \mathcal{N}_m\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_m\right)$$

have been observed.

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

$$\pi\left(\tilde{\boldsymbol{y}} | \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}\right) \propto \int \pi\left(\tilde{\boldsymbol{y}} | \sigma^2, \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}\right) \pi\left(\sigma^2 | \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}\right) d\sigma^2$$

Zellner's Informative G-Prior: Predictive Modeling

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X}\sim\mathcal{N}_{k+1}\left(\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right),\frac{c\sigma^2}{c+1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

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

$$\mathbb{E}_\pi\left(\widetilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)=\mathbb{E}_\pi\left[\mathbb{E}_\pi\left(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta},\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right]$$
$$=\mathbb{E}_\pi\left[\widetilde{\boldsymbol{X}}\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right]=\widetilde{\boldsymbol{X}}\mathbb{E}_\pi\left[\boldsymbol{\beta}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right]\Rightarrow$$

$$\boxed{\mathbb{E}_\pi\left(\widetilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y},\boldsymbol{X},\widetilde{\boldsymbol{X}}\right)=\widetilde{\boldsymbol{X}}\frac{\widetilde{\boldsymbol{\beta}}+c\widehat{\boldsymbol{\beta}}}{c+1}\ (independent\,of\,\sigma^2)}$$

- This representation is quite intuitive, being the product of the matrix of explanatory variables $\widetilde{\boldsymbol{X}}$ by the Bayes estimate of β.

Zellner's Informative G-Prior: Predictive Modeling

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}_{k+1}\left(\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c} + \widehat{\boldsymbol{\beta}}\right), \frac{c\sigma^2}{c+1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

- Similarly, we can compute:

$$\begin{aligned} V_\pi\left(\widetilde{\boldsymbol{y}}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}\right) = \mathbb{E}_\pi\left[V\left(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}\right)|\sigma^2, \boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}\right] \\ + V_\pi\left[\mathbb{E}_\pi\left(\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}\right)|\sigma^2, \boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}\right] = \\ = \mathbb{E}_\pi\left[\sigma^2\boldsymbol{I}_m|\sigma^2, \boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}\right] + V_\pi\left[\widetilde{\boldsymbol{X}}\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}\right] \end{aligned}$$

$$\boxed{V_\pi\left(\widetilde{\boldsymbol{y}}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}\right) = \sigma^2\left(\boldsymbol{I}_m + \frac{c}{c+1}\widetilde{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\widetilde{\boldsymbol{X}}^T\right)}$$

- Here, we are interested on the highest posterior density (HPD) regions on subvectors of the parameter $\boldsymbol{\beta}$ derived from the marginal posterior distribution of $\boldsymbol{\beta}$.

- For a single parameter,

$$\beta_i|\boldsymbol{y},\boldsymbol{X}\sim\mathcal{T}_1\left(n,\frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}_i}{c}+\widehat{\boldsymbol{\beta}}_i\right),\frac{c\left(s^2+(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\widetilde{\boldsymbol{\beta}}-\hat{\beta})/(c+1)\right)}{n(c+1)}\omega_{ii}\right)$$

where ω_{ii} is the (i,i) element of $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$

- Let us define

$$\boldsymbol{\tau} = \frac{\widetilde{\boldsymbol{\beta}} + c\widehat{\boldsymbol{\beta}}}{c+1}$$

- Also

$$\boldsymbol{K} = \frac{c\left(s^2 + (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})/(c+1)\right)}{n(c+1)} (\boldsymbol{X}^T \boldsymbol{X})^{-1} = (K_{ij})$$

- The variable

$$\zeta_i = \frac{\boldsymbol{\beta}_i - \tau_i}{\sqrt{K_{ii}}}$$

has a $\mathcal{T}$-distribution with n degrees of freedom.

- A $1-\alpha$ HPD interval on β_i is thus given by (see also here)

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

where F_n is the CDF of $\mathcal{T}(\nu=n)$.

- Note that these HPD are different from the frequentist confidence intervals defined earlier as:

$$\beta_i|\boldsymbol{y},\boldsymbol{X} \sim \mathcal{T}\left(n-k-1, \hat{\beta}_i, \omega_{(i,i)}\, s^2/(n-k-1)\right)$$

$$\frac{\beta_i - \hat{\beta}_i}{\omega_{(i,i)}\, s^2/(n-k-1)} \sim \mathcal{T}(\nu = n-k-1, 0, 1),\ where\text{: } \omega_{(i,i)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}|_{(i,i)}$$

$$\left\{\beta_i\text{: } |\beta_i - \hat{\beta}_i| \le F_{n-k-1}^{-1}(1-a/2)\sqrt{\omega_{(i,i)}\, s^2/(n-k-1)}\right\}$$

$$s^2 \triangleq \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

Zellner's Informative G-Prior: 95% High Posterior Density (HPD) Intervals for β_i's

beta_i	HPD Interval
beta_0	[4.6518, 17.3450]
beta_1	[-0.0077, -0.0012]
beta_2	[-0.0992, -0.0084]
beta_3	[-0.1384, 0.2742]
beta_4	[-2.4629, -0.1244]
beta_5	[0.0152, 0.4481]
beta_6	[-3.6054, 2.8918]
beta_7	[-2.3238, 1.8489]
beta_8	[-0.3099, 0.6720]
beta_9	[-3.0789, 0.5083]
beta_10	[-1.9571, 1.0909]

c=100

An implementation is available MatLab, C++

$$\left\{\beta_i \colon |\beta_i - \hat{\beta}_i| \leq F^{-1}_{n-k-1}(1-a/2)\sqrt{\omega_{(i,i)}\, s^2/(n-k-1)}\right\}$$

beta_i	HPD Interval
beta_0	[5.7435, 16.2533]
beta_1	[-0.0071, -0.0018]
beta_2	[-0.0914, -0.0162]
beta_3	[-0.1029, 0.2387]
beta_4	[-2.2618, -0.3255]
beta_5	[0.0524, 0.4109]
beta_6	[-3.0466, 2.3330]
beta_7	[-1.9649, 1.4900]
beta_8	[-0.2254, 0.5875]
beta_9	[-2.7704, 0.1998]
beta_10	[-1.6950, 0.8288]

$\{\beta_i : |\beta_i - \hat{\beta}_i|$

c=100

An implementation is available MatLab, C++

Note: The results given in ``C. P. Robert, *The Bayesian Core*, *Springer, 2nd edition,* chapter 3 '' refer to 90% HPD intervals rather than 95% as posted. These results agree with what is shown above.

Zellner's Informative G-Prior: Marginal Distribution of y

- The marginal distribution of $\boldsymbol{y}$ (evidence) is a multivariate $\mathcal{T}$.
- Since $\boldsymbol{\beta}|\sigma^2, \boldsymbol{X} \sim \mathcal{N}_{k+1}(\widetilde{\boldsymbol{\beta}}, c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, the linear transform of $\boldsymbol{\beta}$ satisfies:

$$\boldsymbol{X\beta}|\sigma^2, \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{X}\widetilde{\boldsymbol{\beta}}, c\sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)$$

which implies

$$\boldsymbol{y}|\sigma^2, \boldsymbol{X} \sim \mathcal{N}_n\left(\boldsymbol{X}\widetilde{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{I}_n + c\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\right)$$

- Integration in σ^2 with $\pi(\sigma^2) = 1/\sigma^2$ gives (see Appendices):

$$f(\boldsymbol{y}|\boldsymbol{X}, c) = \int_0^\infty \frac{1}{(2\pi)^{n/2}} \frac{1}{(c+1)^{(k+1)/2}} (\sigma^2)^{-n/2-1} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}})^T\left(\boldsymbol{I}_n + c\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)^{-1}(\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}})} \, d\sigma^2$$

$$f(\boldsymbol{y}|\boldsymbol{X}, c) = (c+1)^{-(k+1)/2}\pi^{-n/2}\Gamma(\frac{n}{2})\left[\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{c+1}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}} - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right]^{-n/2}$$

Appendix

- It can be easily shown that

$$\left(\boldsymbol{I}_n + c\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right)^{-1} = \boldsymbol{I}_n - \frac{c}{c+1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$$

- Indeed:

$$\left(\boldsymbol{I}_n - \frac{c}{c+1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right)\left(\boldsymbol{I}_n + c\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right) = \boldsymbol{I}_n - \frac{c}{c+1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T + c\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$$

$$-\frac{c^2}{c+1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{I}_n + \left(c - \frac{c}{c+1} - \frac{c^2}{c+1}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{I}_n$$

- This can simplify the term in the exponential in the earlier slide:

$$(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^T\left(\boldsymbol{I}_n + c\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) = (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^T\left(\boldsymbol{I}_n - \frac{c}{c+1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) =$$

$$\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}}$$

$$-\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} + \widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}} + \frac{c}{c+1}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \frac{c}{c+1}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}} =$$

$$= \boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{c+1}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}} - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}}$$

Appendix

- Let us revisit the matrix $\boldsymbol{I}_n + c\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$
- Note that for any $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$:

$$\left[\boldsymbol{I}_n + c\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right]\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} + c\boldsymbol{X}\boldsymbol{\beta} = (1+c)\boldsymbol{X}\boldsymbol{\beta}$$

which implies that $\boldsymbol{X}\boldsymbol{\beta}$ is an eigenvector with eigenvalue *(1+c).* There are *(k+1)* of those.

- Finally note that for any *z* in the null space of X^T, $\boldsymbol{X}^T z = 0,$

$$\left[\boldsymbol{I}_n + c\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right]z = z + c\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T z = z$$

which implies that these *z* (*n-k-1* in number) are eigenvectors of our matrix with *1* as the eigenvalues.

- The determinant of the matrix is then $(1+c)^{k+1}1^{n-k-1} = (1+c)^{k+1}$. This explains the derivation in the earlier slide.

Zellner's Informative G-Prior: Point Null Hypothesis

- If a null hypothesis is $H_0 : R\boldsymbol{\beta} = r$, R being a q x $(k+1)$ matrix, the model under H_0 can be rewritten as

$$\boldsymbol{y} \mid \boldsymbol{\beta}^0, \sigma^2, \boldsymbol{X}_0 \overset{H_0}{\sim} \mathcal{N}_n(\boldsymbol{X}_0\boldsymbol{\beta}^0, \sigma^2\boldsymbol{I}_n)$$

where $\boldsymbol{\beta}^0$ is $(k + 1 - q)$ dimensional.

- Under the prior

$$\boldsymbol{\beta}^0|\sigma^2, \boldsymbol{X}_0 \sim \mathcal{N}_{k+1-q}\left(\widetilde{\boldsymbol{\beta}}^0, c_0\sigma^2\left(\boldsymbol{X}_0^T\boldsymbol{X}_0\right)^{-1}\right)$$

- The marginal distribution of **y** under H_0 is:

$$\boldsymbol{f}(\boldsymbol{y}|\boldsymbol{X}_0, H_0) = (c_0+1)^{-(k+1-q)/2}\pi^{-n/2}\Gamma\left(\frac{n}{2}\right)$$

$$\times\left[\boldsymbol{y}^T\boldsymbol{y} - \frac{c_0}{c_0+1}\boldsymbol{y}^T\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T\boldsymbol{y} + \frac{1}{c_0+1}\widetilde{\boldsymbol{\beta}}_0^T\boldsymbol{X}_0^T\boldsymbol{X}_0\,\widetilde{\boldsymbol{\beta}}_0 - \frac{2}{c_0+1}\boldsymbol{y}^T\boldsymbol{X}_0\,\widetilde{\boldsymbol{\beta}}_0\right]^{-n/2}$$

Zellner's Informative G-Prior: Bayes Factor

- The Bayes factor is then given in analytical form as:

$$B_{10}^{\pi} = \frac{f(\boldsymbol{y}|\boldsymbol{X})}{f(\boldsymbol{y}|\boldsymbol{X}_0, H_0)} = \frac{(c_0+1)^{(k+1-q)/2}}{(c+1)^{(k+1)/2}} \times$$

$$\left[\frac{\boldsymbol{y}^T\boldsymbol{y} - \frac{c_0}{c_0+1}\boldsymbol{y}^T\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T\boldsymbol{y} + \frac{1}{c_0+1}\widetilde{\boldsymbol{\beta}}_0^T\boldsymbol{X}_0^T\boldsymbol{X}_0\,\widetilde{\boldsymbol{\beta}}_0 - \frac{2}{c_0+1}\boldsymbol{y}^T\boldsymbol{X}_0\,\widetilde{\boldsymbol{\beta}}_0}{\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{c+1}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\,\widetilde{\boldsymbol{\beta}} - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}\,\widetilde{\boldsymbol{\beta}}}\right]^{n/2}$$

- Note that we use the same σ^2 in both models

- The Bayes factor depends on c_0 and c.

- This calculation can be used to evaluate the inclusion or not of each of the terms β_i, $i = 1, .., k+1$ in our regression model.

Note: $\boldsymbol{X}_0 = \boldsymbol{X}(:, setdiff(1{:}k+1, j))$

Noninformative Prior Analysis: Jeffreys' Prior

- Considering the robustness issues of the two priors examined earlier in the case of a complete lack of prior information, we consider now the non-informative Jeffreys' prior.

- The Jeffreys' prior is a flat prior on $(\boldsymbol{\beta}, \log \sigma^2)$

$$\pi^J(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{X}) \propto \sigma^{-2}$$

- The posterior is then given as:

$$\pi^J(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^{-2})^{n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) - \frac{1}{2\sigma^2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right) \times \sigma^{-2}$$

$$= (\sigma^{-2})^{(k+1)/2} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right) \times$$

$$(\sigma^{-2})^{(n-k-1)/2+1} \exp\left(-\frac{1}{2\sigma^2} s^2\right)$$

Noninformative Prior Analysis: Jeffreys' Prior

$$\pi^J(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^{-2})^{(k+1)/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right) \times$$
$$(\sigma^{-2})^{(n-k-1)/2+1} \exp\left(-\frac{1}{2\sigma^2} s^2\right)$$

- From this joint posterior, we can immediately evaluate the following conditional and marginal posteriors:

$$\pi^J(\sigma^2 | \boldsymbol{y}, \boldsymbol{X}) \propto \left(\sigma^2\right)^{-(n-k-1)/2-1} \exp\left(-\frac{1}{2\sigma^2} s^2\right) = \mathcal{IG}\left(\frac{(n-k-1)}{2}, \frac{s^2}{2}\right)$$

$$\pi^J(\boldsymbol{\beta} | \sigma^2, \boldsymbol{y}, \boldsymbol{X}) \propto (\sigma^{-2})^{(k+1)/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right)$$
$$= \mathcal{N}_{k+1}\left(\widehat{\boldsymbol{\beta}}, \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}\right)$$

- From the 1st of these eqs., the Bayes' estimate of σ^2 is:

$$\mathbb{E}^\pi(\sigma^2 | \boldsymbol{y}, \boldsymbol{X}) = \frac{s^2}{n-k-3}$$

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \quad \theta > 0 \qquad \begin{aligned} &\mathrm{E}(\theta) = \frac{\beta}{\alpha-1}, \text{ for } \alpha > 1 \\ &\mathrm{var}(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2 \\ &\mathrm{mode}(\theta) = \frac{\beta}{\alpha+1} \end{aligned}$$

- This estimate is larger (more pessimistic) than earlier estimates

Noninformative Prior Analysis: Jeffreys' Prior

$$\pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{X}) \propto (\sigma^{-2})^{n/2+1}\exp\left(-\frac{s^2+\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{2\sigma^2}\right)$$

- To compute the marginal posterior of β, we need to integrate in σ^2. Using symbolic integrator:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) \propto \int (\sigma^2)^{-n/2-1}\exp\left(-\frac{s^2+\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{2\sigma^2}\right)d\sigma^2$$

$$\propto \left(s^2+\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)\right)^{-n/2} = \left(1+\frac{\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)^T\left(\frac{s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{n-k-1}\right)^{-1}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)}{n-k-1}\right)^{-n/2}$$

- This is a Students' $\mathcal{T}$ distribution:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) \propto \mathcal{T}_{k+1}(\nu=n-k-1,\boldsymbol{\theta},\boldsymbol{\Sigma}),\ with{:}\ \boldsymbol{\theta}=\widehat{\boldsymbol{\beta}},\ \boldsymbol{\Sigma}=\frac{s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{n-k-1}$$

- Thus the Bayes' estimate of β is: $\mathbb{E}^{\pi}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \widehat{\boldsymbol{\beta}}$

- The HPD intervals coincide with the frequentist confidence intervals.

Zellner's Non-Informative G-Prior

- The main difference with informative G-prior setup is that **we now consider *c* as unknown**.
- We use the same G-prior distribution with $\tilde{\boldsymbol{\beta}} = \mathbf{0}_{k+1}$ conditional on c, and introduce a diffuse prior on c,

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

- The corresponding marginal posterior is given as:

$$\pi(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y},\boldsymbol{X}) = \int \pi(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y},\boldsymbol{X},c)\pi(c \mid \boldsymbol{y},\boldsymbol{X})dc$$

$$\propto \sum_{c=1}^{\infty} \pi(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y},\boldsymbol{X},c)\frac{f\left(\boldsymbol{y} \mid \boldsymbol{X},c\right)\pi(c)}{\pi\left(\boldsymbol{y} \mid \boldsymbol{X}\right)}$$

$$\propto \sum_{c=1}^{\infty} \pi(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y},\boldsymbol{X},c) f\left(\boldsymbol{y} \mid \boldsymbol{X},c\right)c^{-1}$$

Zellner's Non-Informative G-Prior

$$\pi(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y},\boldsymbol{X}) \propto \sum_{c=1}^{\infty} \pi(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y},\boldsymbol{X},c)\, f\left(\boldsymbol{y} \mid \boldsymbol{X},c\right) c^{-1}$$

- We have proved earlier (for constant c) that:

$$f(\boldsymbol{y}|\boldsymbol{X},c) = (c+1)^{-(k+1)/2}\pi^{-n/2}\Gamma\left(\frac{n}{2}\right)$$
$$\left[\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right]^{-n/2}$$

- This for our problem with $\tilde{\boldsymbol{\beta}} = \boldsymbol{0}_{k+1}$

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

Zellner's Non-Informative G-Prior: Posterior Mean of β

- Recall that

$$\boldsymbol{y},\boldsymbol{X} \sim \mathcal{T}_{k+1}\left(n, \frac{c}{c+1}\left(\frac{\tilde{\boldsymbol{\beta}}}{c}+\hat{\boldsymbol{\beta}}\right), \frac{c\left(s^2+\left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T\boldsymbol{X}\left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right)/(c+1)\right)}{n(c+1)}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$$

- The Bayes estimates of $\boldsymbol{\beta}$ for $\tilde{\boldsymbol{\beta}} = \mathbf{0}_{k+1}$ is now given by:

$$\mathbb{E}_\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \mathbb{E}_\pi[\mathbb{E}_\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},c)|\boldsymbol{y},\boldsymbol{X}] = \mathbb{E}_\pi\left[\frac{c}{c+1}\hat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{X}\right] = \frac{\sum_{c=1}^{\infty}\frac{c}{c+1}\pi(c|\boldsymbol{y},\boldsymbol{X})}{\sum_{c=1}^{\infty}\pi(c|\boldsymbol{y},\boldsymbol{X})}\hat{\boldsymbol{\beta}} \Rightarrow$$

$$\mathbb{E}_\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \frac{\sum_{c=1}^{\infty}\frac{c}{c+1}f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}{\sum_{c=1}^{\infty}f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}\hat{\boldsymbol{\beta}}$$

Zellner's Non-Informative G-Prior: Posterior Mean of σ^2

- Recall that $\sigma^2|\boldsymbol{y},\boldsymbol{X}\sim\mathcal{IG}\left(\frac{n}{2},\frac{s^2}{2}+\frac{1}{2(c+1)}(\tilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\tilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})\right)\Rightarrow$

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

- The Bayes estimate of σ^2 is given similarly by:

$$\mathbb{E}_{\pi}(\sigma^2|\boldsymbol{y},\boldsymbol{X})=\frac{\sum_{c=1}^{\infty}\frac{s^2+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}/(c+1)}{n-2}f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}{\sum_{c=1}^{\infty}f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}$$

- Both Bayes estimates involve infinite summations on c. The denominator in both cases is the normalizing constant of the posterior $\sum_{c=1}^{\infty}f(y|X,c)c^{-1}$

Zellner's Non-Informative G-Prior: Posterior Variance of β

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X} \sim \mathcal{T}_{k+1}\left(n, \frac{c}{c+1}\left(\frac{\widetilde{\boldsymbol{\beta}}}{c}+\widehat{\boldsymbol{\beta}}\right), \frac{c\left(s^2+(\tilde{\beta}-\widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T\boldsymbol{X}(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}})/(c+1)\right)}{n(c+1)}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

$$V_\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \mathbb{E}_\pi[V_\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},c)|\boldsymbol{y},\boldsymbol{X}] + V_\pi[\mathbb{E}_\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},c)|\boldsymbol{y},\boldsymbol{X}] =$$

$$\mathbb{E}_\pi\left[\frac{c}{(n-2)(c+1)}\left(s^2+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}/(c+1)\right)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right] + V_\pi\left[\frac{c}{(c+1)}\widehat{\boldsymbol{\beta}}|\boldsymbol{y},\boldsymbol{X}\right] =$$

$$\left[\frac{\sum_{c=1}^{\infty}\frac{f(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{c})}{(n-2)(c+1)}\left(s^2+\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}/(c+1)\right)}{\sum_{c=1}^{\infty} f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}\right](\boldsymbol{X}^T\boldsymbol{X})^{-1} +$$

$$\widehat{\boldsymbol{\beta}}\left[\frac{\sum_{c=1}^{\infty}\left(\frac{c}{(c+1)}-\mathbb{E}_\pi\left(\frac{c}{(c+1)}|\boldsymbol{y},\boldsymbol{X}\right)\right)^2 f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}{\sum_{c=1}^{\infty} f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}\right]\widehat{\boldsymbol{\beta}}^T$$

See earlier slide for the term $f(y|X,c)$

Zellner's Non-Informative G-Prior: Marginal Distribution

- The marginal distribution of the dataset is available in closed form

$$f\left(\boldsymbol{y} \mid \boldsymbol{X}\right)=\sum_{c=1}^{\infty} f\left(\boldsymbol{y} \mid \boldsymbol{X}, c\right) c^{-1} \propto$$

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

- The $\mathcal{T}$-shape means that we can also compute the normalizing constant!

Zellner's Informative G-Prior: Posterior Mean and Variance

$$\mathbb{E}_\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \frac{\sum_{c=1}^{\infty} \frac{c}{c+1} f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}{\sum_{c=1}^{\infty} f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}} \widehat{\boldsymbol{\beta}}$$

beta_i	$\mathbb{E}_\pi(\boldsymbol{\beta}_i \mid \boldsymbol{y}, \boldsymbol{X})$ 𝔼[beta_i | y, X]	$V_\pi(\boldsymbol{\beta}_i \mid \boldsymbol{y}, \boldsymbol{X})$ V[beta_i | y, X]
(Intercept)	9.2714	9.6424
X1	-0.0037	0.0000
X2	-0.0454	0.0005
X3	0.0573	0.0092
X4	-1.0905	0.3079
X5	0.1953	0.0105
X6	-0.3008	2.2750
X7	-0.2002	0.9383
X8	0.1526	0.0522
X9	-1.0835	0.7063
X10	-0.3651	0.5020

$$V_\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \left[\frac{\sum_{c=1}^{\infty} \frac{f(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{c})}{(n-2)(c+1)}\left(s^2 + \widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{X} \widehat{\boldsymbol{\beta}}/(c+1)\right)}{\sum_{c=1}^{\infty} f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}\right](\boldsymbol{X}^T\boldsymbol{X})^{-1} + \widehat{\boldsymbol{\beta}}\left[\frac{\sum_{c=1}^{\infty}\left(\frac{c}{(c+1)} - \mathbb{E}_\pi\left(\frac{c}{(c+1)}|\boldsymbol{y},\boldsymbol{X}\right)\right)^2 f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}{\sum_{c=1}^{\infty} f(\boldsymbol{y}|\boldsymbol{X},c)c^{-1}}\right]\widehat{\boldsymbol{\beta}}^T$$

A Matlab implementation can be downloaded here

$$\left(We\ use\ \widetilde{\boldsymbol{\beta}} = \boldsymbol{0}_{11},\ c = 100\right)$$

Zellner's Non-Informative G-Prior: Point Null Hypothesis

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

$$y \mid \boldsymbol{\beta}^0, \sigma^2, \boldsymbol{X}_0 \overset{H_0}{\sim} \mathcal{N}_n(\boldsymbol{X}_0 \boldsymbol{\beta}^0, \sigma^2 \boldsymbol{I}_n)$$

where $\boldsymbol{\beta}^0$ is $(k + 1 - q)$ dimensional.

- Under $\pi(c) = c^{-1}$ and the prior

$$\boldsymbol{\beta}^0 \mid \sigma^2, \boldsymbol{X}_0, c \sim \mathcal{N}_{k+1-q}\left(\boldsymbol{0}_{k+1-q}, c\sigma^2 \left(\boldsymbol{X}_0^T \boldsymbol{X}_0\right)^{-1}\right)$$

the marginal distribution of **y** under H_0 is:

$$\boldsymbol{f}\left(\boldsymbol{y} \mid \boldsymbol{X}_0, H_0\right) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(k+1-q)/2} \left[\boldsymbol{y}^T \boldsymbol{y} - \frac{c}{c+1} \boldsymbol{y}^T \boldsymbol{X}_0 \left(\boldsymbol{X}_0^T \boldsymbol{X}_0\right)^{-1} \boldsymbol{X}_0^T \boldsymbol{y}\right]^{-n/2}$$

- The Bayes factor $\boldsymbol{B}_{10}^{\pi} = \boldsymbol{f}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) / \boldsymbol{f}\left(\boldsymbol{y} \mid \boldsymbol{X}_0, H_0\right)$ can now be computed

Zellner's Non-Informative G-Prior: Posterior Mean and Variance

For $H_0 : \beta_7 = \beta_8 = 0, \log_{10}\left(B_{10}^{\pi}\right) = -0.7884$ (*We use* $\boldsymbol{\beta} = 0_{11}$)

beta_i	$\mathbb{E}_\pi(\boldsymbol{\beta}_i \mid \boldsymbol{y}, \boldsymbol{X})$ 𝔼[beta_i | y, X]	$V_\pi(\boldsymbol{\beta}_i \mid \boldsymbol{y}, \boldsymbol{X})$ V[beta_i | y, X]	$\log_{10}(BF)$ log10(BF)
(Intercept)	9.2714	9.6424	1.4205 (***)
X1	-0.0037	0.0000	0.8503 (**)
X2	-0.0454	0.0005	0.5664 (**)
X3	0.0573	0.0092	-0.3609
X4	-1.0905	0.3079	0.4520 (*)
X5	0.1953	0.0105	0.4007 (*)
X6	-0.3008	2.2750	-0.4411
X7	-0.2002	0.9383	-0.4404
X8	0.1526	0.0522	-0.3383
X9	-1.0835	0.7063	-0.0424
X10	-0.3651	0.5020	-0.3838

Evidence against H_0:

(****) decisive
(***) strong
(**) substantial
(*) poor

A Matlab implementation can be downloaded here

Variable Selection

- Let us return to our regression model with one dependent random variable y and a set of *k* $\{x_1, x_2, ..., x_k\}$ explanatory variables.

- Are all the x_i's needed in the regression?

- We assume that every q-subset $\{i_1, i_2, ..., i_q\}$, $0 \le q \le k$, of the explanatory variables,

$$\{1, x_{i_1}, x_{i_2}, ..., x_{i_q}\}$$

  is a proper set of explanatory variables for the regression of y (as before, the intercept is included in all models).

- We have a total of 2^k models to select from!

Variable Selection

- Following earlier notation, we denote: $\boldsymbol{X} = \left[\mathbf{1}_n \; x_1 \; x_2 \ldots x_k\right]$ as the matrix that contains $\mathbf{1}_n$ and the k potential predictor variables.

- Each model M_γ is 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 M_γ and $\gamma_i=0$ that it is not.

- The number of variables included in the model M_γ is:

$$q_\gamma = \mathbf{1}^T \boldsymbol{\gamma}$$

- The indices of the variables included in the model and not included in the model are denoted, respectively, as: $t_1(\gamma), t_0(\gamma)$

Variable Selection – Models in Competition

- For $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ and $\boldsymbol{X}$, we define $\boldsymbol{\beta}_\gamma$ as the sub-vector

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

- Let $\boldsymbol{X}_\gamma$ be the submatrix of $\boldsymbol{X}$ where only the column $\mathbf{1}_n$ and the columns in $t_1(\gamma)$ have been left.

- The model M_γ is then defined as:

$$\boldsymbol{y} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma, \sigma^2, X \sim \mathcal{N}_n\left(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}_n\right)$$

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

- The σ^2 is common to all models and we use the same prior for all models.

Variable Selection – Models in Competition

- We have a high number 2^k of models in competition.

- We cannot specify a prior on every M_γ in a completely subjective and autonomous manner.

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

Zellner's Informative Prior: Variable Selection

- For the full model that corresponds to $\gamma = (1, \ldots, 1)$, we use the Zellner's informative G-prior:

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{X} \sim \mathcal{N}_{k+1}\left(\widetilde{\boldsymbol{\beta}}, c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$
$$\sigma^2 \sim \pi(\sigma^2|\boldsymbol{X}) \propto \sigma^{-2} \ improper\,Jeffreys\,prior$$

- For each model M_γ, the prior distribution of $\boldsymbol{\beta}_\gamma$ conditional on σ^2 is fixed as:

$$\boldsymbol{\beta}_\gamma\,|\boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_{q_\gamma+1}\left(\widetilde{\boldsymbol{\beta}}_\gamma, c\sigma^2\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\right)^{-1}\right)$$

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

Zellner's Informative Prior: Variable Selection – Prior

- The joint prior for model M_γ is the improper prior

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

- Infinitely many ways of defining a prior on the model index γ:

  Our choice is a uniform prior $p(\gamma|\boldsymbol{X}) = 2^{-k}$.

- Posterior distribution of γ is central to variable selection since it is proportional to marginal density of $\mathbf{y}$ on M_γ (or evidence of M_γ)

- Posterior distribution of γ is proportional to the marginal density of **y** on M_γ (so it can also be used to compute Bayes factors)

$$\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X}) \propto f(\boldsymbol{y} \mid \gamma, \boldsymbol{X})\pi(\gamma \mid \boldsymbol{X}) \propto f(\boldsymbol{y} \mid \gamma, \boldsymbol{X})$$
$$= \int\left(\int f(\boldsymbol{y} \mid \gamma, \boldsymbol{\beta}, \sigma^2, \boldsymbol{X})\pi(\boldsymbol{\beta} \mid \gamma, \sigma^2, \boldsymbol{X})d\boldsymbol{\beta}\right)\pi(\sigma^2 \mid \boldsymbol{X})d\sigma^2$$

where (see earlier derivation)

$$f(\boldsymbol{y}|\boldsymbol{\gamma},\sigma^2,\boldsymbol{X}) = \int f(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\beta},\sigma^2)\pi(\boldsymbol{\beta}|\boldsymbol{\gamma},\sigma^2)d\boldsymbol{\beta} =$$
$$= (c+1)^{-(q_\gamma+1)/2}(2\pi)^{-n/2}(\sigma^2)^{-n/2} \times$$
$$\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{2\sigma^2(c+1)}\left\{c\boldsymbol{y}^T\boldsymbol{X}_\gamma\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\right)^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{y} - \widetilde{\boldsymbol{\beta}}_\gamma^T\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma + 2\boldsymbol{y}^T\boldsymbol{X}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma\right\}\right)$$

Zellners Informative Prior: Variable Selection – Prior

$$\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X}) \propto \int f(\boldsymbol{y} \mid \gamma, \sigma^2, \boldsymbol{X}) \pi(\sigma^2 \mid \boldsymbol{X}) d\sigma^2 = \int f(\boldsymbol{y} \mid \gamma, \sigma^2, \boldsymbol{X}) \frac{1}{\sigma^2} d\sigma^2$$

- Posterior distribution of γ is then given as:

$$f(\gamma|\boldsymbol{y},\boldsymbol{X}) \propto (c+1)^{-(q_\gamma+1)/2} \times$$
$$\left(\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{(c+1)}\boldsymbol{y}^T\boldsymbol{X}_\gamma\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\right)^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{y} + \frac{1}{c+1}\widetilde{\boldsymbol{\beta}}_\gamma^T\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}_\gamma\widetilde{\boldsymbol{\beta}}_\gamma\right)^{-n/2}$$

- We already have seen this distribution earlier for a fixed c:

$$f(\boldsymbol{y}|\boldsymbol{X},c) = (c+1)^{-(k+1)/2}\pi^{-n/2}\Gamma(\frac{n}{2})\left[\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{c+1}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}} - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right]^{-n/2}$$

Model Selection

- Most likely models ordered by decreasing posterior probabilities using Zellner's informative G-prior with *c=100*.

$t_1(\gamma)$	$\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$
t1_gamma	pi(gamma | y, X)
0 1 2 4 5	0.231543
0 1 2 4 5 9	0.037358
0 1 9	0.034435
0 1 2 4 5 10	0.032975
0 1 4 5	0.030606
0 1 2 9	0.025016
0 1 2 4 5 7	0.024144
0 1 2 4 5 8	0.023784
0 1 2 4 5 6	0.023735
0 1 2 3 4 5	0.023207
0 1 6 9	0.014587
0 1 2 3 9	0.014491
0 9	0.014281
0 1 2 6 9	0.013551
0 1 4 5 9	0.012761
0 1 3 9	0.011712
0 1 2 8	0.011477
0 1 8	0.009519
0 1 2 3 4 5 9	0.009036
0 1 2 4 5 6 9	0.009031

An implementation is available MatLab, C++

Model Selection

- Model M_γ 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.

Model Selection

- For the Zellner's non-informative prior with $\pi(c)=1/c$, we have ($\tilde{\boldsymbol{\beta}} = \mathbf{0}_{k+1}$) :

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{X}) = \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q_\gamma+1)/2}\left[\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}_\gamma\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\right)^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{y}\right]^{-n/2}$$

- Again we have seen this before as

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

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

Model Selection

- Most likely models ordered by decreasing posterior probabilities using Zellner's non-informative G-prior.

$t_1(\gamma)$ $\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$

t1_gamma	pi(gamma | y, X)
0 1 2 4 5	0.092914
0 1 2 4 5 9	0.032553
0 1 2 4 5 10	0.029512
0 1 2 4 5 7	0.023114
0 1 2 4 5 8	0.022843
0 1 2 4 5 6	0.022807
0 1 2 3 4 5	0.022409
0 1 2 3 4 5 9	0.016733
0 1 2 4 5 6 9	0.016725
0 1 2 4 5 8 9	0.013726
0 1 4 5	0.011031
0 1 2 4 5 9 10	0.009933
0 1 2 3 9	0.009698
0 1 2 9	0.009316
0 1 2 4 5 7 9	0.009253
0 1 2 6 9	0.009189
0 1 4 5 9	0.008756
0 1 2 3 4 5 10	0.007933
0 1 2 4 5 8 10	0.007901
0 1 2 4 5 7 10	0.007896

An implementation is available MatLab, C++

Stochastic Search for the Most Likely Model

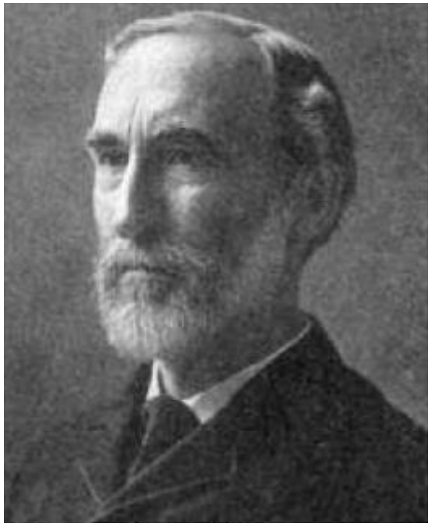
- When k is large, it becomes computationally intractable to compute the posterior probabilities of the 2^k models.
- Need of a tailored algorithm that samples from $\pi(\gamma|\mathbf{y},\mathbf{X})$ and selects the most likely models.
- Can be done by Gibbs sampling*, given the availability of the full conditional posterior probabilities of the γ_i's. If

$$\gamma_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_k) \ (1 \leq i \leq k)$$

then

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

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



* Gibbs and other sampling algorithms will be introduced and discussed in detail in forthcoming lectures.

Gibbs Sampling for Variable Selection

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

Iteration t: Given $\left(\gamma_1^{(t-1)},\ldots,\gamma_k^{(t-1)}\right)$, generate

- $\gamma_1^{(t)}$ *according to* $\pi\left(\gamma_1 \mid \boldsymbol{y},\gamma_2^{(t-1)},\ldots,\gamma_k^{(t-1)},\boldsymbol{X}\right)$
- $\gamma_2^{(t)}$ *according to* $\pi\left(\gamma_2 \mid \boldsymbol{y},\gamma_1^{(t)},\gamma_3^{(t-1)},\ldots,\gamma_k^{(t-1)},\boldsymbol{X}\right)$
- ..
- ..
- $\gamma_k^{(t)}$ *according to* $\pi\left(\gamma_k \mid \boldsymbol{y},\gamma_1^{(t)},\gamma_2^{(t)},\ldots,\gamma_{k-1}^{(t)},\boldsymbol{X}\right)$

Gibbs Sampling for Variable Selection

Question: How to sample γ_1^t according to $\pi\left(\gamma_1 \mid \boldsymbol{y}, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, \boldsymbol{X}\right)$

1. The conditional distribution $\pi\left(\gamma_1 \mid \boldsymbol{y}, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, \boldsymbol{X}\right)$ is proportional to $\pi\left(\gamma \mid \boldsymbol{y}, \boldsymbol{X}\right)$

2. Since γ_1^t only has two possible values which are 0 and 1, we get

$$p_0 \propto \pi\left(\gamma_1 = 0, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)} \mid \boldsymbol{y}, \boldsymbol{X}\right)$$

$$p_1 \propto \pi\left(\gamma_1 = 1, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)} \mid \boldsymbol{y}, \boldsymbol{X}\right)$$

So the probability that $\gamma_1^t = 1$ is p_1, then we can use Gibbs Sampling to approximate the distribution of $\{\gamma_i^t\}$

Gibbs Sampling: Posterior Probabilities

- After $T \gg 1$ MCMC iterations, we approximate the posterior probabilities $p(\gamma|\mathbf{y},\mathbf{X})$ by empirical averages

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

The T_0 first values (burn in) in the MCMC chain are eliminated.

Model Choice Comparison: Gibbs Estimates

First level Informative G-prior model with ($\tilde{\boldsymbol{\beta}} = 0_{11}$, c=100) compared with the Gibbs estimates of the top ten posterior probabilities

$t_1(\gamma)$	$\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$	$\hat{\pi}(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$
t1_gamma	pi(gamma | y, X)	pi_hat(gamma | y, X)
0 1 2 4 5	0.231543	0.239276
0 1 2 4 5 9	0.037358	0.034397
0 1 9	0.034435	0.032397
0 1 2 4 5 10	0.032975	0.030097
0 1 4 5	0.030606	0.029397
0 1 2 9	0.025016	0.025297
0 1 2 4 5 7	0.024144	0.022498
0 1 2 4 5 8	0.023784	0.024898
0 1 2 4 5 6	0.023735	0.023598
0 1 2 3 4 5	0.023207	0.022998

A MatLab implementation is available

Model Choice Comparison: Gibbs Estimates

Non-informative G-prior variable model choice compared with the Gibbs estimates of the top ten posterior probabilities

$t_1(\gamma)$	$\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$	$\hat{\pi}(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$
t1_gamma	pi(gamma | y, X)	pi_hat(gamma | y, X)
0 1 2 4 5	0.092914	0.093391
0 1 2 4 5 9	0.032553	0.033097
0 1 2 4 5 10	0.029512	0.032597
0 1 2 4 5 7	0.023114	0.025097
0 1 2 4 5 8	0.022843	0.023098
0 1 2 4 5 6	0.022807	0.022498
0 1 2 3 4 5	0.022409	0.021698
0 1 2 3 4 5 9	0.016733	0.015998
0 1 2 4 5 6 9	0.016725	0.014899
0 1 2 4 5 8 9	0.013726	0.013399

A MatLab implementation is available

Gibbs Sampling: Probabilities of Inclusion

- An approximation of the probability to include the i-th variable:

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

Probabilities of Inclusion Estimates

Informative ($\tilde{\boldsymbol{\beta}} = 0_{11}$, c=100) and non-informative G-prior variable inclusion estimates (based on the same Gibbs output as in the earlier two tables)

γ_i	$\hat{P}^{\pi}(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{X})$	$\hat{P}^{\pi}(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{X})$
gamma_i	P(gamma_i|y, X)	P(gamma_i|y, X)
gamma_1	0.8733	0.8806
gamma_2	0.7100	0.7789
gamma_3	0.1515	0.2958
gamma_4	0.6842	0.7422
gamma_5	0.6635	0.7234
gamma_6	0.1659	0.2992
gamma_7	0.1343	0.2812
gamma_8	0.1478	0.2740
gamma_9	0.3942	0.5015
gamma_10	0.1135	0.2556

A MatLab implementation is available