♠ Conjugate Priors on Regression Parameters

- Model in matrix form:

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Use notation $\mathbf{H} = \mathbf{X}'$ for comparability with traditional statistics notation: $\mathbf{H}$ is $n \times p$ design matrix

$$\mathbf{y} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}) = N(0, \phi^{-1}\mathbf{I})$$

♠ LSE formulæ:

-

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\mathbf{H}'\mathbf{y} \quad \text{and} \quad \mathbf{V}^{-1} = \mathbf{H}'\mathbf{H}$$

♠ (Simple) Standard Shrinkage Priors

- Conjugate normal prior centred at zero: shrinkage towards zero
- Genes centred: zero is canonical "null" hypothesis of no (partial) association
- Scale: scale of gene expression, standardised scales

$$\boldsymbol{\beta}|\phi \sim N(0, \tau\phi^{-1}\mathbf{I})$$

- Intercept: include in $\boldsymbol{\beta}$ - different status: different scale?

$$\boldsymbol{\beta}|\phi \sim N(0, \phi^{-1}\mathbf{C}^{-1})$$

- Prior precison matrix $\mathbf{C}$ (up to constant ϕ) is diagonal – common value τ^{-1} for genes, possibly different (larger) for intercept
- Prior on residual variance/precision: $\phi \sim Ga(a/2, b/2)$ with prior variance estimate b/a and degree of fredom a. Vague reference prior is special case $a, b \to 0$

♠ Features of Posterior Under Shrinkage Prior

- $\boldsymbol{\beta}|\phi, \mathbf{y} \sim N(\mathbf{b}, \phi^{-1}\mathbf{B}^{-1})$
- $\mathbf{B} = \mathbf{C} + \mathbf{H}'\mathbf{H}$ and $\mathbf{b} = \mathbf{B}^{-1}\mathbf{H}'\mathbf{y}$
- Special limiting cases: vague prior (zero precision) - $\mathbf{B} = \mathbf{H}'\mathbf{H}$ and $\mathbf{b} = \hat{\boldsymbol{\beta}}$
- Otherwise, shrinkage towards zero induced by prior: $||\mathbf{b}|| < ||\hat{\boldsymbol{\beta}}||$
- Decision Theory: Bayes's estimates improve expected performance in estimating $\boldsymbol{\beta}$
  - quadratic (or other convex) loss functions
  - Bayesian and frequentist measures of risk in estimation
- In practical terms, shrinkage generally stabilises/regularises estimation and improves robustness in predictions
- Large p questions: LSE/likelihood methods inapplicable – need for regularisation
- Log posterior (conditional on ϕ) = log likelihood + log prior: a constant plus a term proportional to

$$(\mathbf{y} - \mathbf{H}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{H}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{C}\boldsymbol{\beta}$$

- Posterior mean/mode $\mathbf{b}$ represents shrunken/regularised MLE (=LSE)

♠ Marginal Posteriors for b and ϕ

- $\phi|\mathbf{y} \sim Ga(a'/2, b'/2)$ with $a' = a + n$ and $b' = b + q$ where $q > 0$ can be written in one of several useful forms:
  - simplest for computation: $q = \mathbf{y}'\mathbf{e}$ where $\mathbf{e} = \mathbf{y} - \mathbf{Hb}$ is the estimated residual vector
  - quadratic form representation: $q = \mathbf{y}'\mathbf{P}'\mathbf{y}$ where $\mathbf{P} = \mathbf{I} - \mathbf{HB}^{-1}\mathbf{H}'$
  - interesting limiting case of vague reference prior: we know $\mathbf{b} \to \hat{\boldsymbol{\beta}}$ and now also $q \to$ usual residual sum of squares in LSE regression with $\mathbf{P} \to \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$
- Common point estimate of $\sigma^2 = \phi^{-1}$ is reciprocal of posterior mean of ϕ, namely $s = b'/a'$
- $\boldsymbol{\beta}|\mathbf{y} \sim T_{a'}(\mathbf{b}, s\mathbf{B}^{-1})$
  - credible intervals for elements of $\boldsymbol{\beta}$ are derived from Student T distribution on a' degrees of freedom
- (see matlab code: mregbayes)
- Role of $\mathbf{C}$, Key example:
  - $C_{1,1}$ =large (intercept term, vague prior)
  - $C_{i,i} = \tau^{-1}$ otherwise – common precision for (common scale) genes
  - Marginal likelihood function for τ (see matlab code, examples) to assess τ. Can be shown that marginal likelihood $p(\mathbf{y}|\tau)$ implies

$$\log(p(\mathbf{y}|\tau)) = \text{constant} + 0.5\log(|\mathbf{C}|/|\mathbf{B}|) - 0.5n\log(\mathbf{y}'\mathbf{e})$$

♠ Shrinkage Priors More Generally: Considerations

- Different degrees of shrinkage across variables (genes) via different diagonal elements in $\mathbf{C}$
- General question of learning from data about shrinkage parameters: inference on $\mathbf{C}$ generally
- Orthogonal designs: e.g., Factor regressions and other models: design is orthogonal (by design!)
  - $\mathbf{H}'\mathbf{H}$ is diagonal
  - so that $\mathbf{B}$ is diagonal when $\mathbf{C}$ is
  - elements of $\boldsymbol{\beta}$ are uncorrelated under the posterior as well as the prior (dependent only through scale ϕ)