Competing Sparsity: Modelling with Hierarchical Shrinkage Priors

Phil Brown (joint work with Jim Griffin)

University of Kent
School of Mathematics, Statistics and Actuarial Science
Canterbury, UK

Overview

- Modelling regression in various forms with many variables, $p >> n$
- $n \approx 100, p \in (100, 10000)$
- Fitting a ‘minimal prior’ distribution bringing out sizeable effects without necessarily removing small effects which may be important for following-up designs
- Nominally roots in Bayesian theory but a long stretch from Rev Thomas Bayes (1763)
- Many flavours of Bayes (46656!, IJ Good, ’Good Thinking’) Now even more with MCMC

Introduction

We look at the standard regression model

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$

where response y_i (eg activity), errors $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, and $x_{ij}, j = 1, \ldots, p$, are p measured variables (eg fragments) on $n < p$ observations, $i = 1, \ldots, n$.

Work over the past twenty year has looked at how to estimate α and β if we believe that β is sparse (many of the regressors zero or close to zero) but don't know which. They try to improve on methods of variable selection and regularisation (i.e. using shrinkage).

Linear model with interactions

Suppose that we extend the regression model to include interactions

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j=1}^{p}\sum_{k=1}^{i-1} x_{ij}x_{ik}\gamma_{jk} + \epsilon_i.$$

Potentially many more interaction parameters than main effects, combinatorial 2 from p.
An assumption often made is that an interaction (γ_{jk}) parameter will be close to zero if either main effects (β_j or β_k) is close to zero, often referred to as heredity, strong or weak : strong heredity requires both main effects to be included, eg not near zero. [Blood Glucose example]

General Additive Model (GAM)

The general additive model assumes that

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i.$$

Suppose that $f_j(x_{ij})$ is expressed in some basis function, $g_k()$ (such as a spline) with hyperparameters κ_k eg knots for splines.

$$f_j(x_{ij}) = x_{ij}\theta_j + \sum_{k=1}^{m} \beta_{jk} g_k(x_{ij}, \kappa_k).$$

General Additive Model (GAM)

The general additive model becomes

$$y_i = \sum_{j=1}^{p} x_{ij}\theta_j + \sum_{j=1}^{p}\sum_{k=1}^{m} \beta_{jk} g_k(x_{ij}, \kappa_k) + \epsilon_i$$

There are two-LEVELS of regularization problems here:

- Basis-level: The $\beta_{j1}, \ldots, \beta_{jm}$ have to be regularized.
- Variable-level: Removing a variable corresponds to setting $\theta_j = \beta_{j1} = \cdots = \beta_{jm} = 0$.

[Prostate data Example]

Multivariate extensions

- p variables on each of q responses eg [Activity, Solubility, $C\log P$, Safety]
- Matrix of coefficients $B(p \times q)$
- Perhaps structured as 'main effects' and 'interaction' for shrinkage
- Errors in each regression typically correlated

Leverhulme Emeritus fellowship to exploit correlation, 2018-2020.

Sequential Design

- Sequential design, , perhaps ten observations at a time
- Can use Bayesian predictive distribution to choose the next ten observations
- Maximise expected gain of Shannon information (Chaloner & Verdinelli, 1996; Gramacy & Lee, 2012) for activity or other responses.
- fills in where data is needed
- use an exchange algorithm to optimise (Wynn, 1972)

Hierarchical shrinkage

In all cases, the regression coefficients can be arranged in levels.

Here we have two LEVELS but we can have more levels.

They have the property that:

- The shrinkages at different LEVELS are linked.
- There are different pressures on shrinkage at different LEVELS (*i.e.* there is greater sparsity at higher levels).

Structure of Talk

- Shrinkage using continuous priors
- Hierarchical shrinkage priors
- Forms of heredity
- Comparative shrinkage propagation
- Example: Linear model with interactions
- Example: GAM

Bayesian variable selection / regularization

There are many proposed priors for Bayesian variable selection and regularization including:

- Spike-and-slab priors (Mitchell and Beauchamp, 1988)
- Lasso prior (cf earlier Lasso with *Penalisation* motivation
- Horseshoe prior (Carvalho *et al*, 2010)
- Normal-Gamma (Caron and Doucet, 2008, Griffin and Brown, 2010)
- Normal-Gamma-Gamma (Armagan *et al*, 2011)

Bayesian variable selection / regularization

These have a common form. The regression coefficients β_i are given the prior

$$\beta_i \sim \mathrm{N}(0, \Psi_i), \qquad \Psi_i \sim F$$

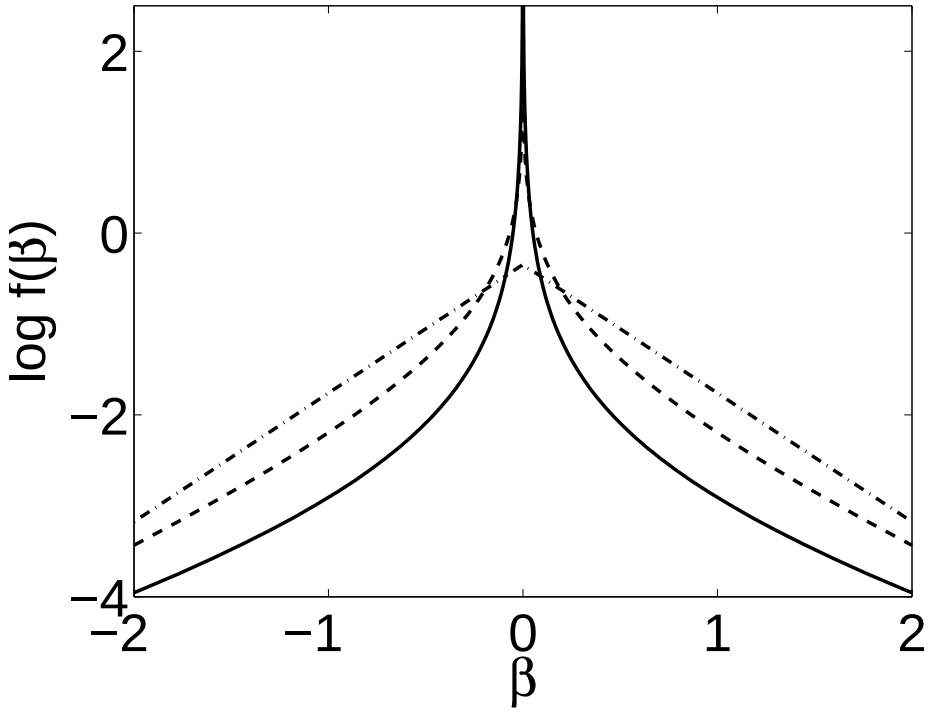
for some distribution F.

The one group priors assume that F is continuous.

Normal-Gamma: $F = \mathrm{Ga}(\lambda, 1/(2\gamma^2))$.

Normal-Gamma-Gamma(λ, c, d): $F = \mathrm{Ga}(\lambda, \gamma_j), \gamma_j = Ga(c, d)$.
c controls heaviness of tails, scale d (Special case Horseshoe)

Log density of the Normal-Gamma prior with variance 2



- $\lambda = 0.1$ (solid line)
- $\lambda = 0.333$ (dot-dashed line)
- $\lambda = 1$ (dashed line)

Shrinkage profiles for NG

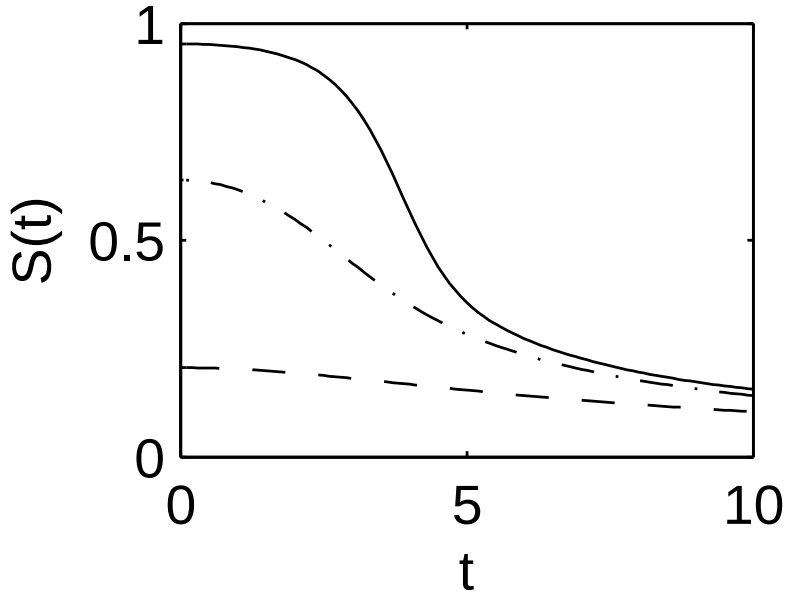


Figure: Shrinkage profiles for an NG prior with $\lambda = 0.1$, (solid line), $\lambda = 1$ (dot-dashed line) and $\lambda = 5$ (dashed line) with $\gamma = 1/SE^2$.

$$t = \hat{\beta}/SE$$

Super efficiency and tail robustness priors

These are illustrated in Polson and Scott (2010).

Kuhlback Leiber super-efficiency relates the the prior having a spike at zero, whereas tail robustness requires the prior to have 'fat' tails (polynomial rather than exponential.)

For certain hyperparmeter values the Normal Gamma (λ, γ) is super efficient since it has a spike at zero when $\lambda \leq 1$ but is not tail robust since it has exponential tails, whereas the NGG can be both super efficient and tail robust, (includes horseshoe prior).

Linear model with interactions

The linear model with interactions can be expressed as

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j=1}^{p}\sum_{k=1}^{i-1} x_{ij}x_{ik}\gamma_{jk} + \epsilon_i$$

Priors for β's and γ's is a scale mixture of normals, *i.e.*

$$\beta_j \sim \mathrm{N}\left(0, \Psi_j\right), \qquad \gamma_{jk} \sim \mathrm{N}\left(0, \Psi_{jk}\right)$$

The following hyperpriors for the Ψ's includes the assumption that γ_{jk} will be close to zero if either β_j or β_k are close to zero.

$$\Psi_j = \eta_j, \qquad \Psi_{jk} = \eta_j\eta_k\eta_{jk}.$$

Products promote strong , sums weak heredity

Shrinkage for products

Gamma priors can be given to all the η's and so generalize the Normal-Gamma prior.

We are interested in priors defined by products: $\Psi_i = \prod_{j=1}^{K_i} \eta_{s_{i,j}}$ where $s_{i,1}, \ldots, s_{i,K_i} \in \{1, 2, \ldots, m\}$, $\eta_1, \ldots, \eta_m$ are independent and $\eta_j \sim \text{Ga}(\lambda_i, b_i)$.

This allows different Ψ's to share the same value of η.

The shrinkage parameter of the product prior is $\min\{\lambda_{s_{i,1}}, \ldots, \lambda_{s_{i,K_i}}\}$. This will allow us to control the shrinkage at different levels of a hierarchical prior.

Sparsity of the priors

Linear model with interactions

$\eta_j \sim \text{Ga}(\lambda, \delta)$ and $\eta_{jk} \sim \text{Ga}(\lambda_{int}, 1)$.

The shrinkage parameter of the main effects is λ and the shrinkage of the interactions is $\min\{\lambda, \lambda_{int}\}$.

GAM

$\eta_j \sim \text{Ga}(\lambda_{group}, \delta)$ and $\eta_{jk} \sim \text{Ga}(\lambda, 1)$.

The sparsity at the variable level is λ_{group} and at the basis level is $\min\{\lambda, \lambda_{group}\}$

Blood glucose

Previously analysed by Hamada and Wu (1992) and using a LARS-based algorithm by Yuan et al (2007).

The data has one two-level factor and seven three-level factors.

The three-level factors are included as linear and quadratic effects using orthogonal polynomials.

All interactions are included.

Blood glucose: Prior structure

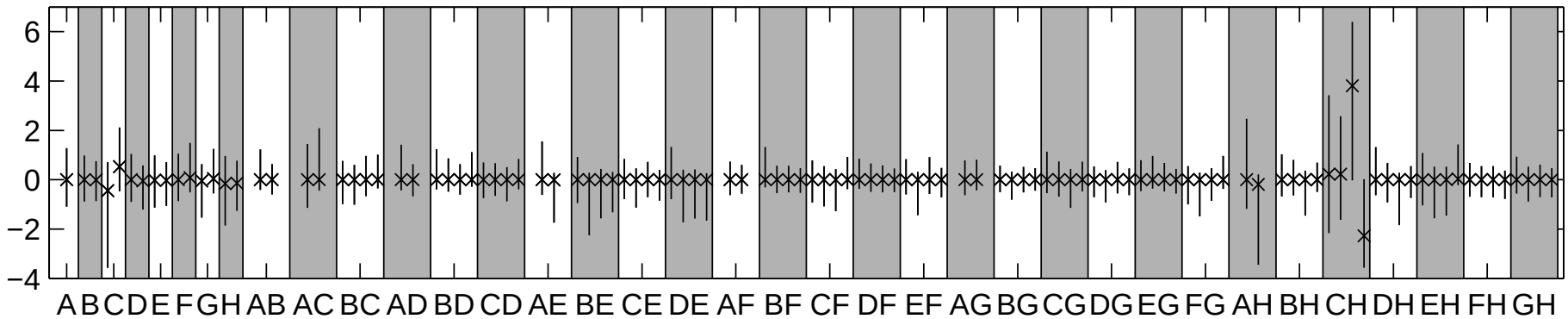
NGG with heavy tail behaviour but finite variance ($c = 2$)

Adaptive shrinkage of main effects exponential mean 1 for λ_1

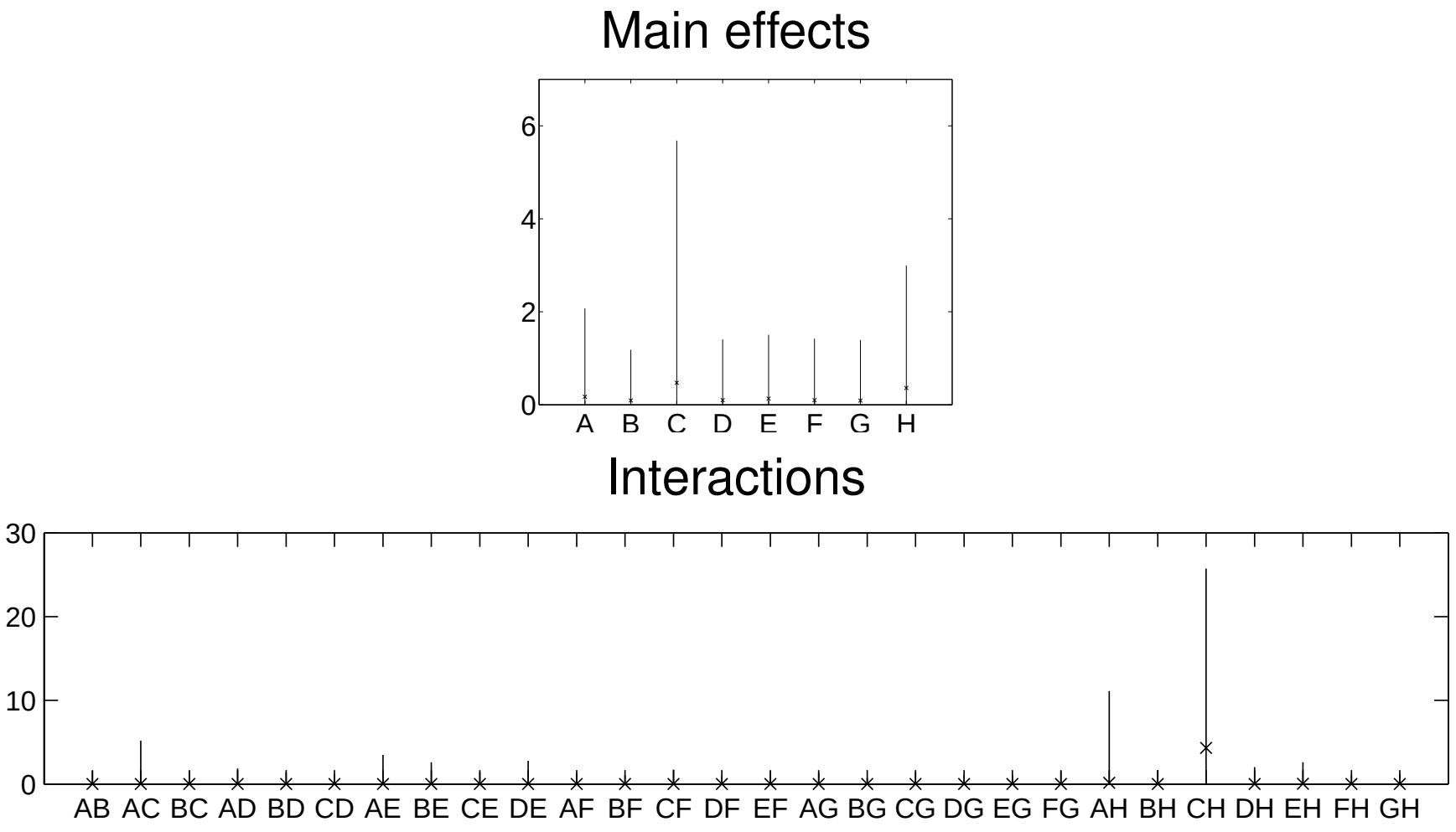
Adaptive shrinkage of interactions, $\lambda_2 = r\lambda_1$ where r is beta mean 1/3: interactions more aggressively shrunk

Scale parameter given a heavy tailed prior with mean 1

Blood glucose: regression coefficients



Blood glucose: Ψ



Blood glucose: posterior hyperparameters

λ_1	0.48 (*0.15, 2.71*)
λ_2	0.054 (*0.018, 0.89*)
d	2.69 (*0.26, 27.9*)

Out-of-sample predictive performance

The results are summarized by the root mean squared error (RMSE) where the posterior predictive mean was used as the estimated prediction; S&S denotes Slab and Spike Bayes

	Blood glucose	Ozone	Boston housing
S&S strong heredity	13.0	4.0390	3.83
S&S weak heredity	11.4	4.0469	3.75
S&S relaxed heredity	12.2	4.0482	3.70
Hierarchical shrinkage	**10.5**	4.0272	**3.40**
Hierarchical lasso	14.4	**4.0181**	3.70

Prostate data

Data from a prostate cancer trial has become a standard example in the regularization literature (Tibshirani, 1996, Zou & Hastie, 2005). We use a GAM.

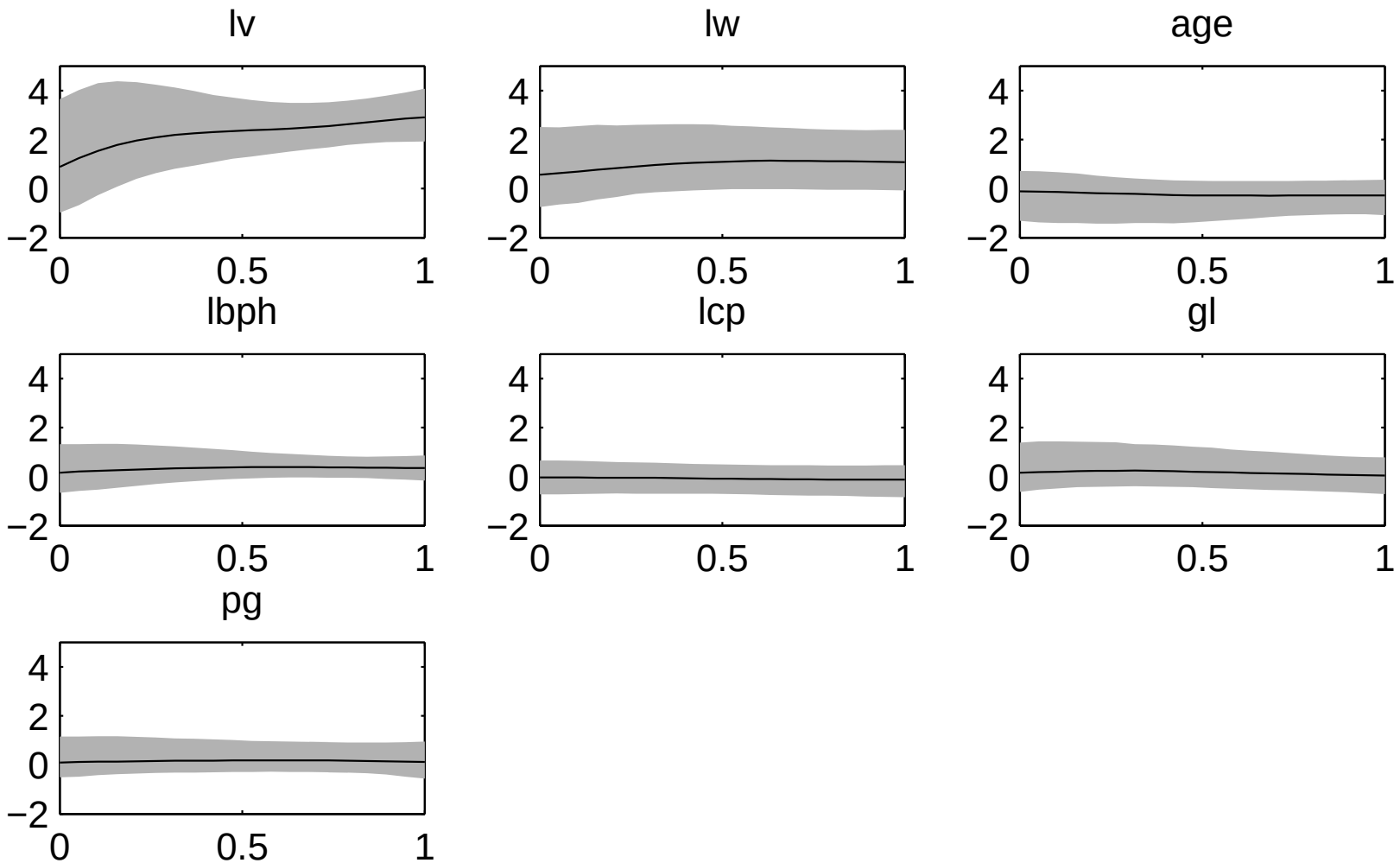
The response is the logarithm of prostate-specific antigen.

There are seven continuous predictors, one binary (*svi*):

- log(cancer volumne) (*lv*)
- log(prostate weight) (*lw*)
- age
- log(benign prostatic hyperplasia) (*lbph*)
- log(capsular penetration) (*lcp*)
- Gleason score (*gl*)
- percentage Gleason score 4 or 5 (*pg*)
- seminal vesicle invasion (*svi*)

Prostate data: regression coefficients

Effect of svi: Posterior medain = 0.58, 95% CI: (*0.08*, *1.06*).



Prostate data: Ψ

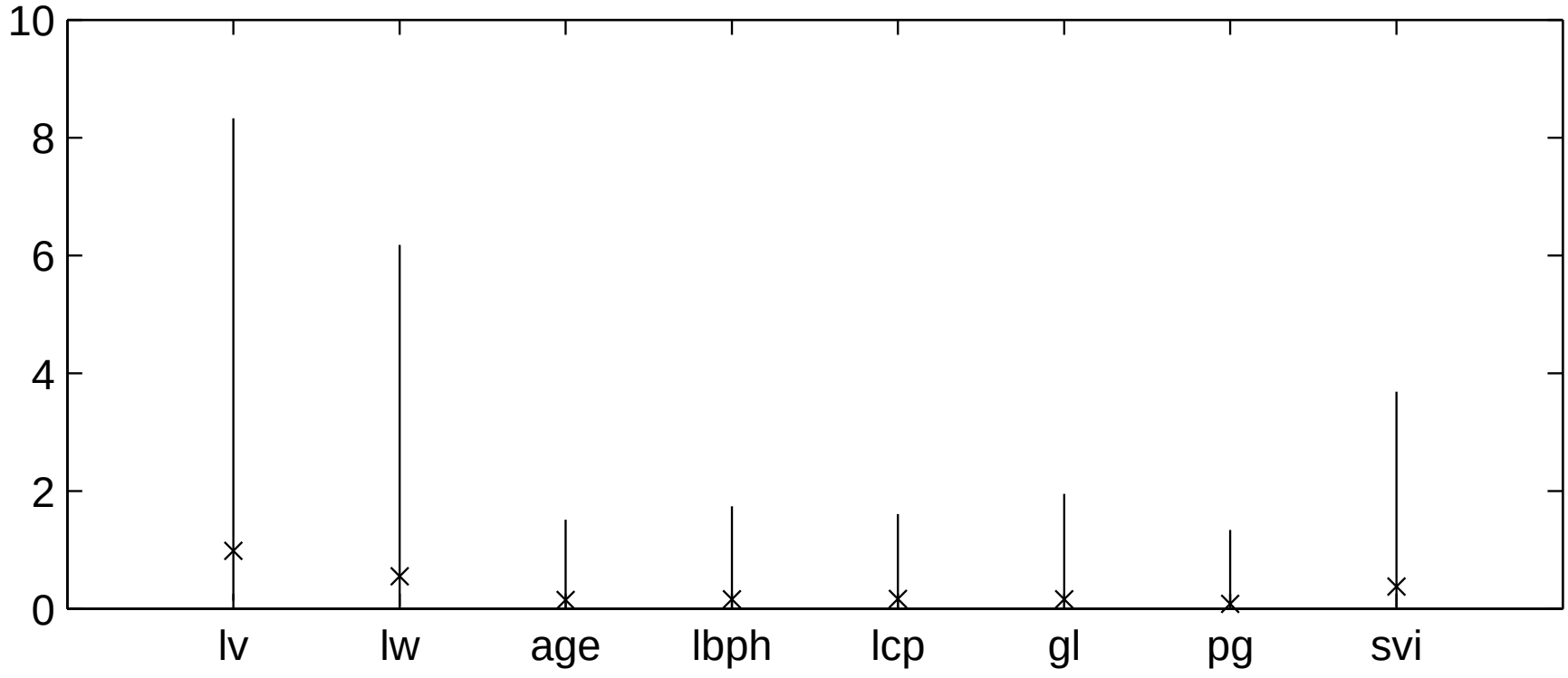


Figure: Prostate cancer data – posterior distribution of Ψ with posterior median (cross) and 95% credible interval (solid line)

Discussion

- Hierarchical shrinkage priors can be defined using products of Gamma or Gamma-Gamma random variables.
- These priors allow dependence of the shrinkage across levels of the hierarchy and different pressures on shrinkage at different levels.
- An aid to sequential design
- An Application to GAMs with interactions is included in the published paper in Bayesian Analysis (2017)
- Multivariate exploiting correlation, yet to develop funded by Leverhulme Emeritus Fellowship