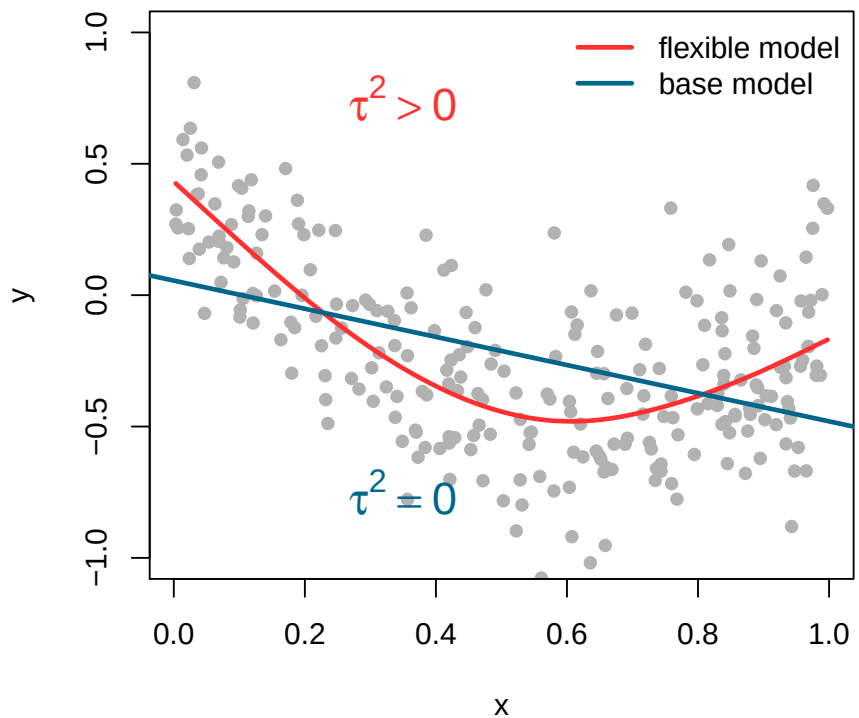
Scale-Dependent Priors for Variance Parameters in Distributional Regression

Nadja Klein
Melbourne Business School, University of Melbourne
November 15th, 2017

Bayes on the Beach 2017, Gold Coast

Base Models and Increased Complexity

- Nested structure inherent in many model components.
- Hyperparameters τ^2 to determine the deviation of a flexible alternative from the base model.
- Occurs also in additive approaches to Bayesian models with blocks like i.i.d. Gaussian random effects, splines of continuous variables, etc.



In particular, . . .

- Exploit the nested structure to construct proper prior distributions for variance parameters in structured additive distributional regression.
- Use idea of divergence-based priors.
- Transfer principles of penalised complexity (PC) priors for incidence type design matrices and within INLA*
- Add user-defined scale parameters that allow to penalise the complexity resulting from deviations from the simpler model.

*Simpson, D. P., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H. (2017): Penalising model component complexity: A principled, practical approach to constructing priors, *Statistical Science*, 32(1), 1-28

Structured Additive Distributional Regression

- Observed data pairs $(\boldsymbol{y}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_n)$.
- Model assumption 1: Conditional distribution $F(\boldsymbol{y}_i|\boldsymbol{x}_i)$ given $\boldsymbol{x}_i$, $i = 1, \ldots, n$ is from pre-specified class of K-parametric densities

$$p(\boldsymbol{y}_i|\vartheta_{i1}, \ldots, \vartheta_{iK}).$$

- Model assumption 2: Each parameter ϑ_{ik}, $k = 1, \ldots, K$ is related to a regression predictor $\eta_{ik} = \eta_k(\boldsymbol{\nu}_i)$: $\vartheta_{ik} = h_k(\eta_{ik})$ and $\eta_{ik} = h_k^{-1}(\vartheta_{ik})$, where

$$\eta_{ik} = \beta_{0,k} + \sum_{j=1}^{p_k} f_{j,k}(\boldsymbol{x}_i).$$

$f_{j,k}$ modelled as linear combination of p_k basis functions $b_{1,k}, \ldots, b_{p_k,k}$, $f_{j,k}(x) = \sum_{j=1}^{p_k} \beta_{j,k} b_{j,k}(x)$, such that

$$\boldsymbol{f}_{j,k} = \boldsymbol{Z}_{j,k}\boldsymbol{\beta}_{j,k}.$$

Bayesian Regularisation

Employ multivariate Gaussian priors $\boldsymbol{\beta}_{j,k} \sim \mathsf{N}(\mathbf{0}, \tau^2_{j,k}\boldsymbol{K}^-_{j,k})$ shrinkage priors,

$$p(\boldsymbol{\beta}_{j,k}|\tau^2_{j,k}) \propto \exp\left(-\frac{1}{2\tau^2_{j,k}}\boldsymbol{\beta}'_{j,k}\boldsymbol{K}_{j,k}\boldsymbol{\beta}_{j,k}\right),$$

and where

- $\boldsymbol{K}_{j,k}$ is the precision of the normal distribution,
- $\boldsymbol{K}^-_{j,k}$ the generalised inverse of $\boldsymbol{K}_{j,k}$,
- $\tau^2_{j,k}$ represents the inverse smoothing parameter.

These allow for data-driven shrinkage $\Rightarrow$ smooth, flexible estimate of $f_{j,k}$.

Bayesian Regularisation

Employ multivariate Gaussian priors $\boldsymbol{\beta}_{j,k} \sim \mathsf{N}(\mathbf{0}, \tau^2_{j,k}\boldsymbol{K}^-_{j,k})$ shrinkage priors,

$$p(\boldsymbol{\beta}_{j,k}|\tau^2_{j,k}) \propto \exp\left(-\frac{1}{2\tau^2_{j,k}}\boldsymbol{\beta}'_{j,k}\boldsymbol{K}_{j,k}\boldsymbol{\beta}_{j,k}\right),$$

and where

- $\boldsymbol{K}_{j,k}$ is the precision of the normal distribution,
- $\boldsymbol{K}^-_{j,k}$ the generalised inverse of $\boldsymbol{K}_{j,k}$,
- $\tau^2_{j,k}$ represents the inverse smoothing parameter.

These allow for data-driven shrinkage $\Rightarrow$ smooth, flexible estimate of $f_{j,k}$.

A 'Usual Way' for τ^2

Assign inverse gamma priors to τ^2:

$$p(\tau^2) \propto \frac{1}{(\tau^2)^{a+1}} \exp\left(-\frac{b}{\tau^2}\right).$$

Proper for	$a > 0,\ b > 0$	Flat prior for $\log(\tau^2)$, with $a = b = \varepsilon$ small (Jeffreys' prior)
		Flat prior for precision $1/\tau^2$ with $a = 1$, b small.
Improper for	$b = 0,\ a = -1$	Flat prior for variance τ^2,
	$b = 0,\ a = -\frac{1}{2}$	Flat prior for standard deviation τ.

Advantages:

- Gibbs sampler possible.
- Sufficient conditions for proper posteriors can be provided.

Disadvantages:

- In extreme situations (such as flat likelihood, small n, close to flat priors, high total rank deficiencies) propriety questionable.
- No axiomatic meaningful reason for the choice.
- Prior parameters hardly elicitable.

Scale-Dependent Hyperpriors for τ^2

Notation:

- $\frac{1}{\tau^2}\boldsymbol{K}$ the precision matrix of the flexible model $p(\boldsymbol{\beta}|\tau^2)$ for a vector of regression coefficients $\boldsymbol{\beta}$.
- For $\tau^2 \to 0$ results the base model $p_b(\boldsymbol{\beta}|\tau^2 = 0)$ representing the nullspace of $\boldsymbol{K}$.

Principle 1: *Occam's Razor*

- The hyperprior should invoke the principle of parsimony.
- Simple base model for each effect is preferred unless the data provide convincing evidence for more complex modelling.

Principle 2: *Measure of Complexity*

- The increased complexity is measured by the Kullback-Leibler divergence

$$\text{KLD}(p||p_b) = 2\int p(u)\log\left(\frac{p(u)}{p_b(u)}\right)\mathrm{d}u$$

  for the base model p_b and an alternative flexible model p.
- Gives a measure of the information loss when the base model is used to approximate the more flexible models.
- Define

$$d(p||p_b) = \sqrt{2\text{KLD}(p||p_b)}$$

  as the unidirectional 'distance' from the flexible model to the base model.

Principle 3: *Constant Rate Penalisation*

- Constant rate penalisation implies an exponential prior on the distance scale, i.e.

$$p_d(d) = \lambda \exp(-\lambda d), \quad \lambda > 0$$

  with mode at $d = 0$.
- Constant rate of decay in the distance prior from p_b to stronger deviations from p_b.
- λ determines the rate of penalisation.

Resulting Prior for τ^2

- Change of variable theorem gives

$$p(\tau^2) = \lambda \exp(-\lambda d(\tau^2)) \left| \frac{\partial d(\tau^2)}{\partial \tau^2} \right| \text{ with } d(\tau^2) = \sqrt{2\text{KLD}}.$$

- Finally gives a Weibull prior for τ^2

$$p(\tau^2|\theta) = \frac{1}{2\theta} \left(\frac{\tau^2}{\theta} \right)^{-1/2} \exp\left(- \left(\frac{\tau^2}{\theta} \right)^{1/2} \right).$$

Principle 4: *User-Defined Scaling*

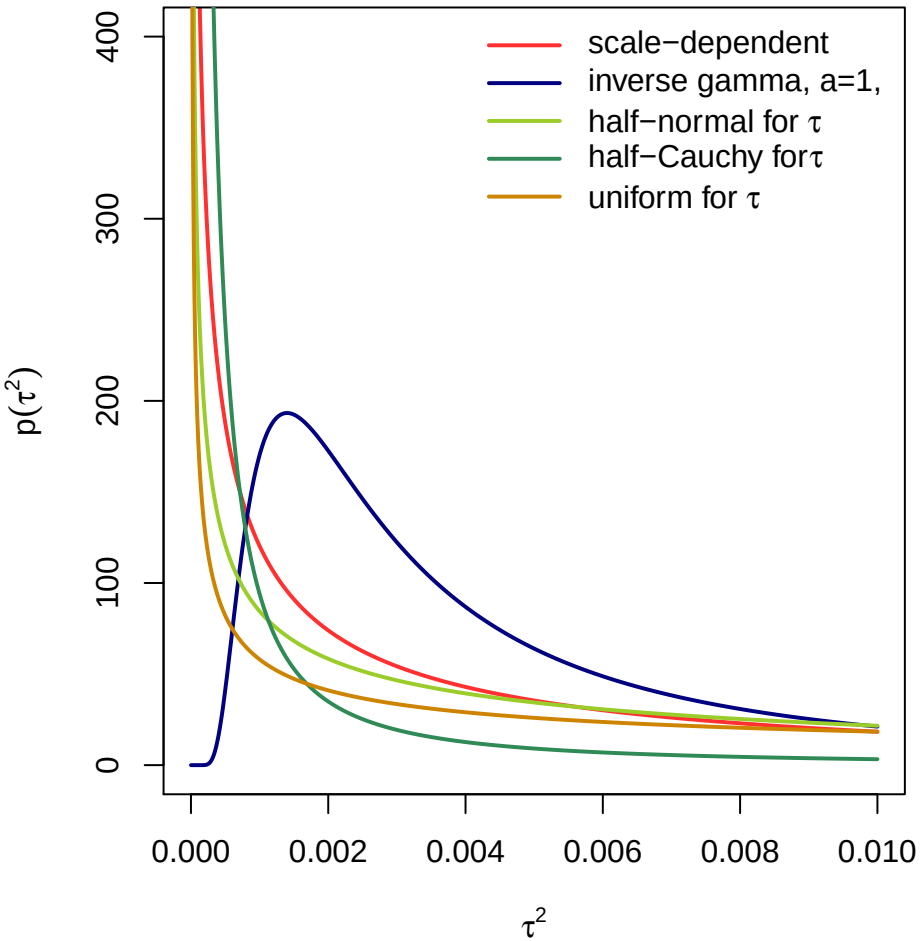
- The decay rate $\exp(-\lambda)$ can be controlled by some condition

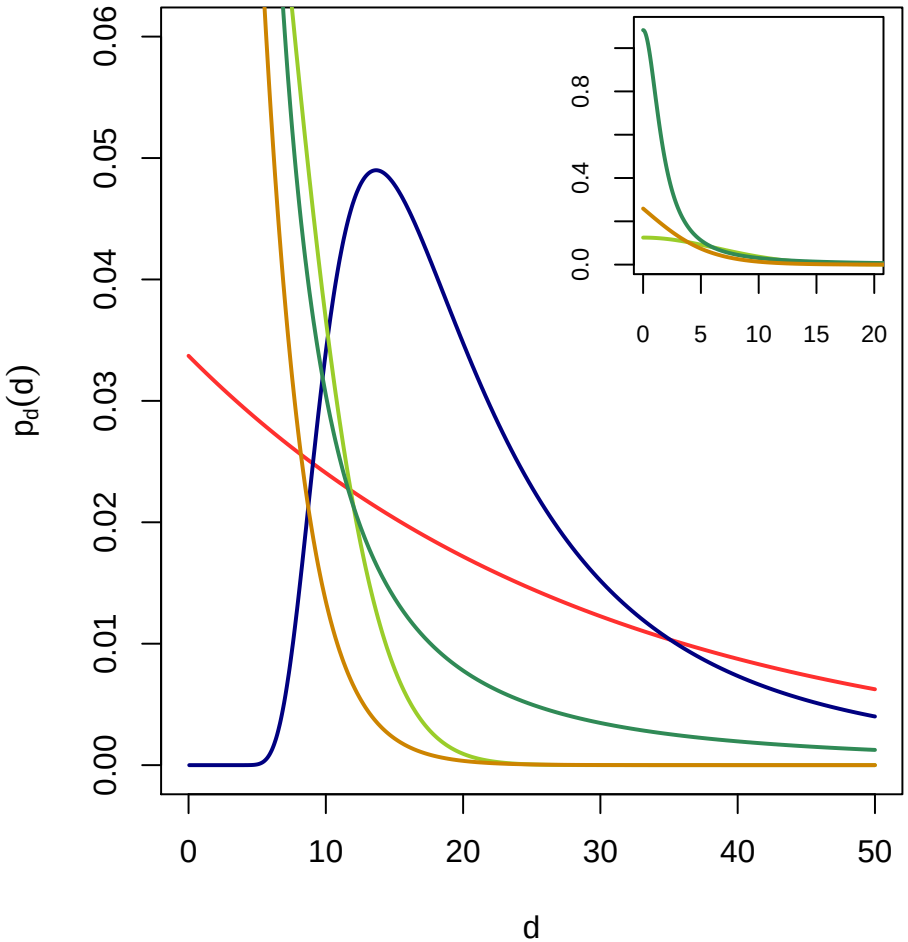
$$\mathbb{P}(q(\tau^2) \leq c) = 1 - \alpha$$

  for transformation $q(\cdot)$ of τ^2 and user-defined values c and α.
- Prior knowledge about the scale of functional effects f allows to specify a certain interval with high marginal probability:

$$\mathbb{P}(|f(\boldsymbol{x})| \leq c \; \forall \boldsymbol{x} \in \mathcal{D}) \geq 1 - \alpha.$$

IG$(1, b)$ has $p_d(0) = 0$

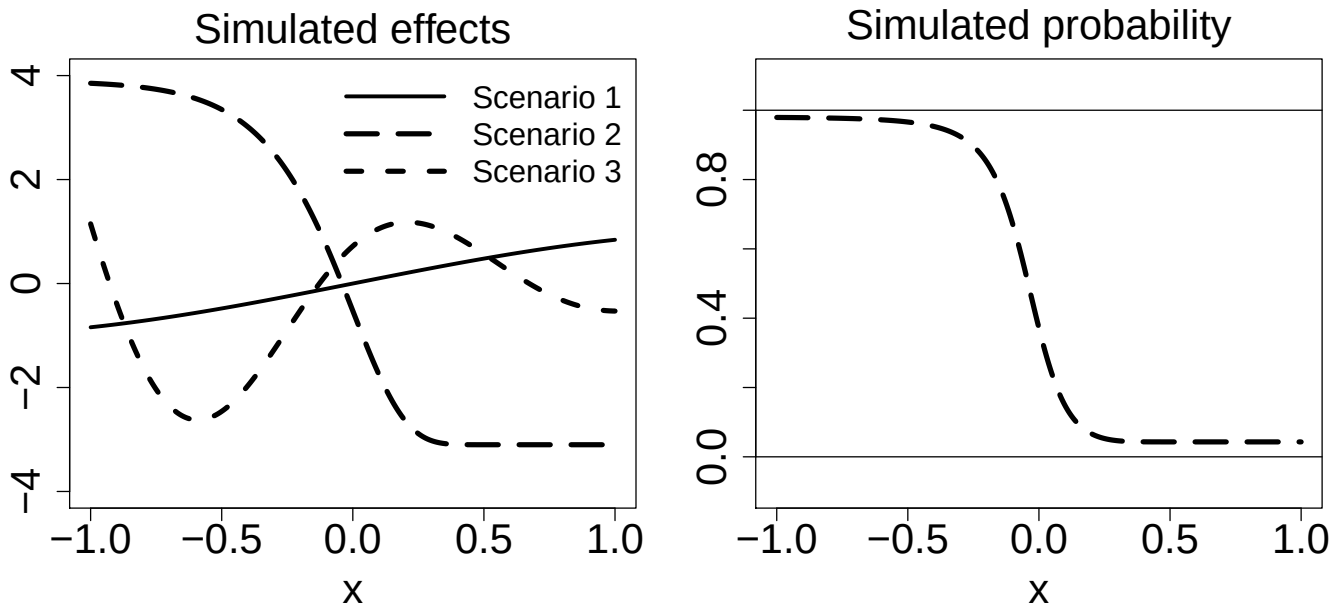




Half-normal, half-Cauchy, uniform* and IG$(1, b)$ are scaled with Principle 4.

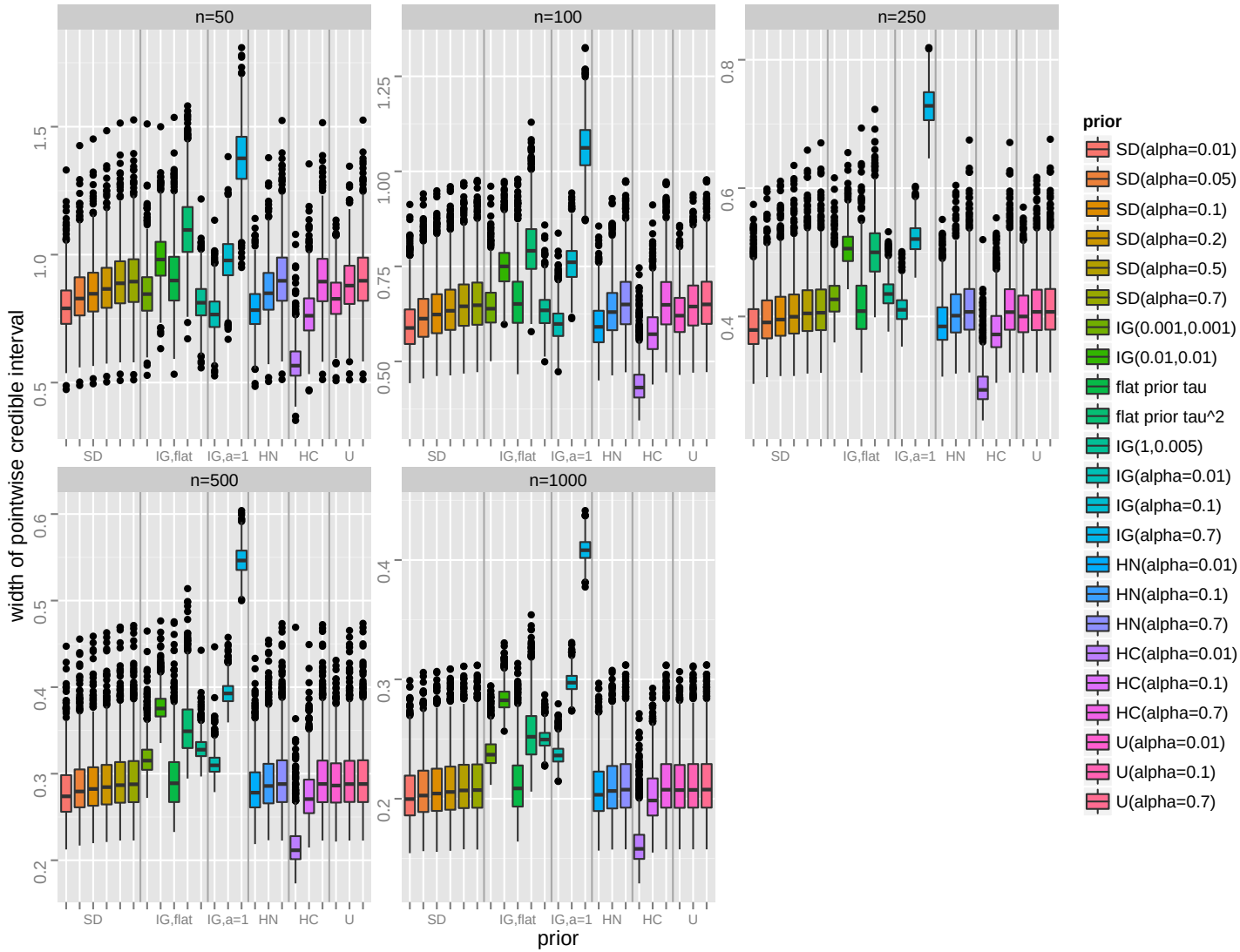
*Gelman, A. (2005). Analysis of variance: why it is more important than ever. *AOS* 33: 1–53.

Simulation Study

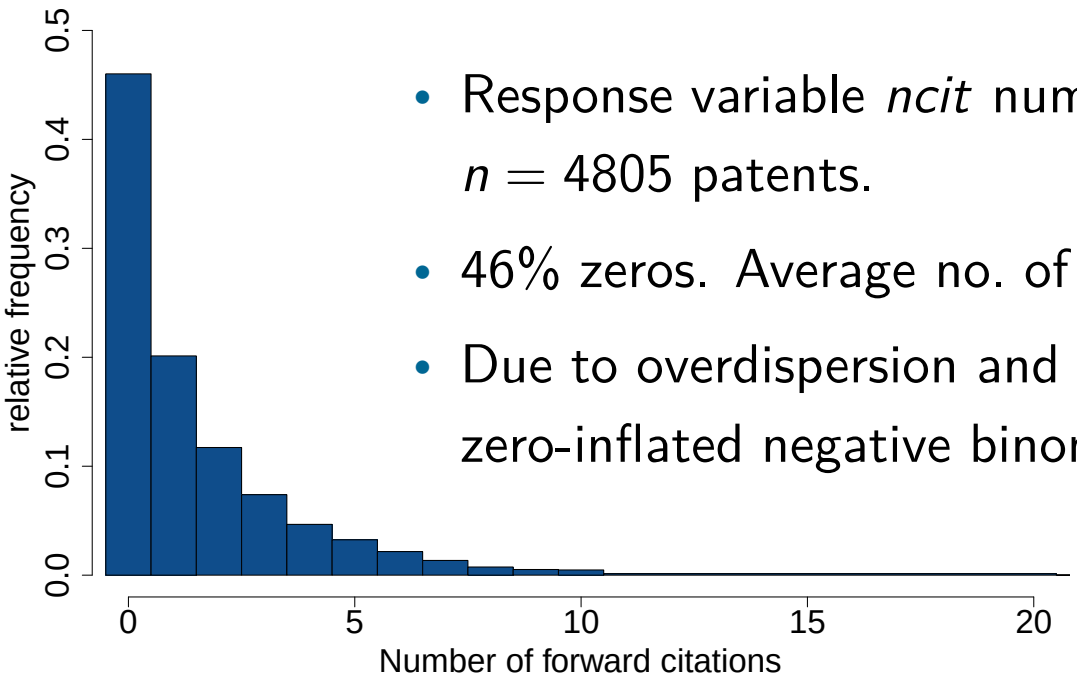


- 3 test functions (f_1 scenario 1, f_2 scenario 2, f_3 scenario 3).
- 2 response distributions:
  - $y \sim \mathrm{N}(f_j(x), 1)$.
  - $y \sim \mathrm{Be}(\pi_j)$, where $\pi_j = \frac{exp(f_j(x))}{1+\exp(f_j(x))}$.
- Varous sample sizes ($n = 50, 100, 200, 500, 1000$).

Simulation: Close to the Base Model, Credible Intervals



Patent Citations: Flat Likelihood



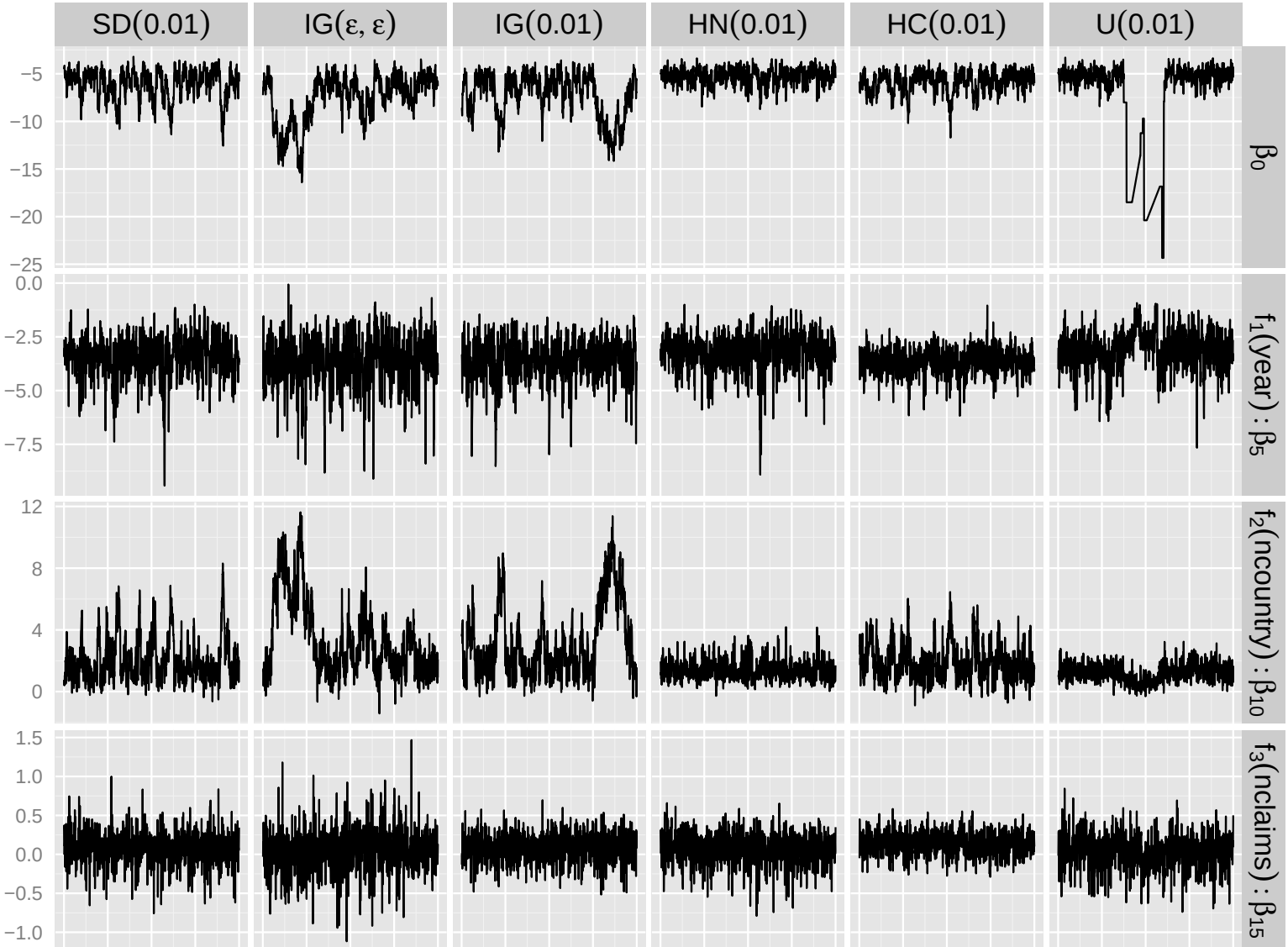
- Response variable *ncit* number of forward citations of $n = 4805$ patents.
- 46% zeros. Average no. of citations: 1.63, variance: 7.35.
- Due to overdispersion and zero-inflation, we assume a zero-inflated negative binomial distribution for y:

$$p(ncit_i|\mu_i, \delta_i, \pi_i) = \pi_i \mathbb{1}_{\{ncit_i=0\}} + (1-\pi_i)\frac{\Gamma(ncit_i+\delta_i)}{\Gamma(ncit_i+1)\Gamma(\delta_i)}\left(\frac{\delta_i}{\delta_i+\mu_i}\right)^{\delta_i}\left(\frac{\mu_i}{\delta_i+\mu_i}\right)^{ncit_i}$$

$$\eta_i = f_1(ncountry_i) + f_2(year_i) + f_3(nclaims_i) + \boldsymbol{x}_i'\boldsymbol{\beta}$$

Problem: zero-inflation only weakly identified for some covariate values.

Small π, Sampling Paths



Summary and Conclusion

- In contrast to IG-prior no longer Gibbs-sampler possible, but MCMC can simply be done via Metropolis-Hastings updates.
- Empirical Pros:
  - Hyperprior elicitation and robustness with respect to α and c.
  - Clear improvement in MSE when true model is close to a simpler base model.
  - More narrow credible intervals (while simultaneously maintaining the coverage probability).
  - Higher numerical stability in situations with flat likelihood.
- Implementation provided in software:
  - BayesX (free download at `www.bayesx.org`) with default θ.
  - Scale parameters for given $\boldsymbol{Z}$, $\boldsymbol{K}$, α, c with `sdPrior` on CRAN.

Nadja Klein and Thomas Kneib (2016): Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression, *Bayesian Analysis*, (11) 1071-1106, online at `https://projecteuclid.org/euclid.ba/1448323525`, doi:10.1214/15-BA983.

How to Scale in Practice

- The marginal density of $f(x_p) = \boldsymbol{z}_p'\boldsymbol{\beta}$ is

$$p(\boldsymbol{z}_p'\boldsymbol{\beta}) = \int_0^\infty p(\boldsymbol{z}_p'\boldsymbol{\beta}, \tau^2)\mathrm{d}\tau^2 = \int_0^\infty p(\boldsymbol{z}_p'\boldsymbol{\beta}|\tau^2)p(\tau^2|\theta)\mathrm{d}\tau^2$$

- θ can be chosen such that

$$\left(1 - \int_{-c}^{c}\int_0^\infty p_{\boldsymbol{z}_p'\boldsymbol{\beta}}(u|\tau^2)p(\tau^2|\theta)\mathrm{d}\tau^2\mathrm{d}u\right) = \alpha$$

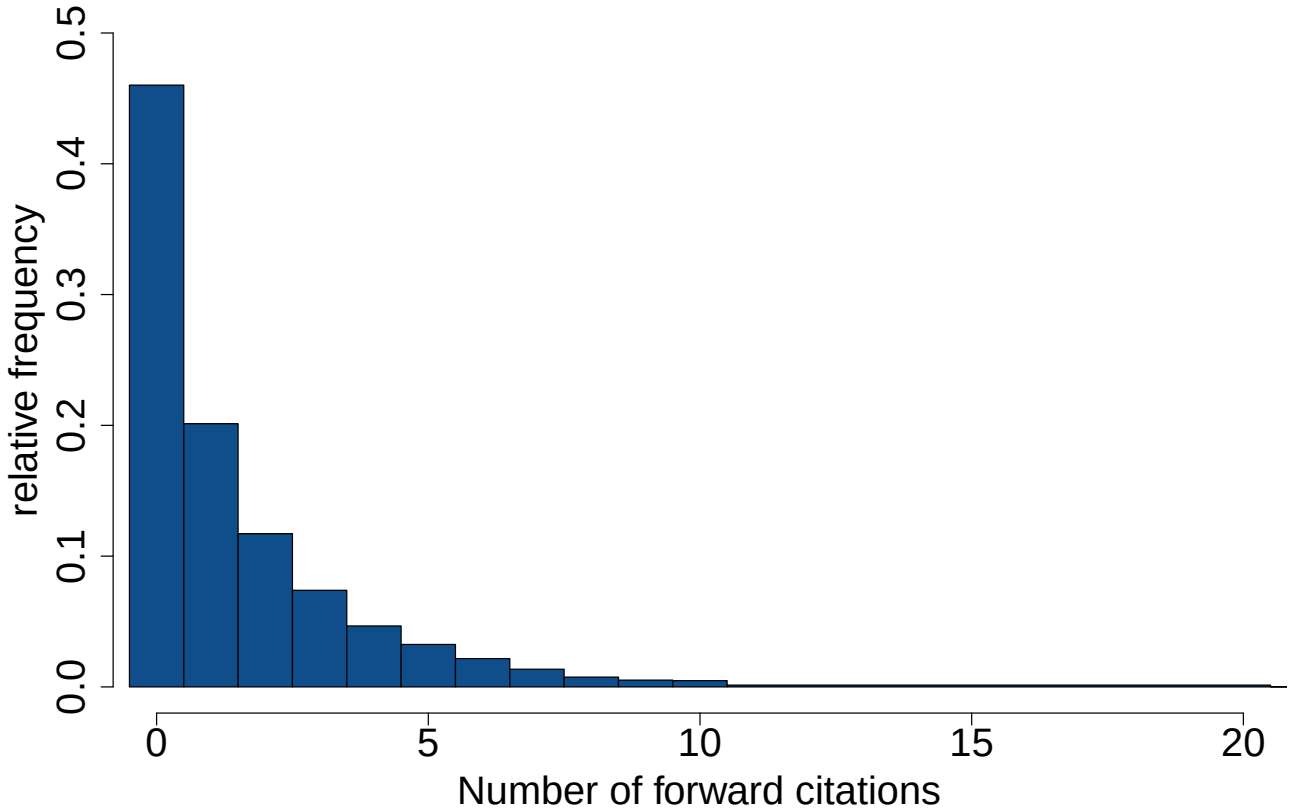
- Simulations to suggest default values for spline models with fixed $\boldsymbol{K}$.
- Functions for general $\boldsymbol{Z}$ and $\boldsymbol{K}$ provided in R-package `sdPrior`.

Priors for τ^2

Name	**Density**	**Information**
HN(α)	$p(\tau^2) \propto (\tau^2)^{1/2-1} \exp\left(-\tau^2/(2\theta^2)\right)$	gamma prior for τ^2/ half-normal prior for τ
HC(α)	$p(\tau^2) \propto \left(1+\tau^2/\theta^2\right)^{-1} \left(\tau^2/\theta^2\right)^{-1/2}$	generalised beta prime prior for τ^2/ half-Cauchy prior for τ
UD(α)	$p(\tau^2) \propto (\tau^2)^{-1/2} \left(1 - \frac{\exp\left((\tau^2)^{1/2}\tilde{c}/\theta - \tilde{c}\right)}{1+\exp\left((\tau^2)^{1/2}\tilde{c}/\theta - \tilde{c}\right)}\right)$	approximate uniform prior for τ^2/ proper uniform prior for τ
IG(α)	$p(\tau^2) \propto (\tau^2)^{-2} \exp\left(-\theta/\tau^2\right)$	flat prior for $1/\tau^2$ for $\theta \to 0$
IG(ϵ, ϵ)	$p(\tau^2) \propto (\tau^2)^{-\epsilon-1} \exp\left(-\epsilon/\tau^2\right)$	'Jeffreys'/flat prior on log-scale for $\epsilon \to 0$
IG($-1, 0$)	$p(\tau^2) \propto const$	flat prior for τ^2
IG($-\frac{1}{2}, 0$)	$p(\tau^2) \propto 1/\sqrt{\tau^2}$	flat prior for τ

- In contrast to IG-prior no longer Gibbs-sampler possible, but MCMC can simply be done via Metropolis-Hastings updates.
- Advantages:
  - Clear improvement in MSE when true model is close to a simpler base model.
  - More narrow credible intervals (while simultaneously maintaining the coverage probability).
  - Higher numerical stability in situations with flat likelihood.
- Implementation provided in software.

Patent Citations: Flat Likelihood



- 46% zeros (many patents are never cited).
- Average no. of citations: 1.63, variance: 7.35.

Continuous covariates

	description	mean	std	min/max
year	grant year	1991		1980/1997
ncountry	no. of designated states in Europe	7.77	4.12	1/17
nclaims	no. of claims in the patent	12.33	8.13	1/50

Binary covariates

	description	categories	rel freq
biopharm	patent from biotech/pharma sector	yes=1	43.9%
ustwin	U.S. twin exists	yes=1	61.3%
patus	patentholder of the patent from U.S.	yes=1	33.2%
patgsgr	owner from Switzerland, Ger or GB	yes=1	23.7%
opp	oppositions	yes=1	41.1%

- Response variable *ncit* number of forward citations of $n = 4805$ patents.
- Due to overdispersion and zero-inflation, we assume a zero-inflated negative binomial distribution

$$p(ncit_i|\mu_i, \delta_i, \pi_i) = \pi_i \mathbb{1}_{\{ncit_i=0\}} + (1-\pi_i)\frac{\Gamma(ncit_i + \delta_i)}{\Gamma(ncit_i + 1)\Gamma(\delta_i)} \left(\frac{\delta_i}{\delta_i + \mu_i}\right)^{\delta_i} \left(\frac{\mu_i}{\delta_i + \mu_i}\right)^{ncit_i}$$

$$\eta_i = f_1(ncountry_i) + f_2(year_i) + f_3(nclaims_i) + \boldsymbol{x}_i'\boldsymbol{\beta}$$

- Problem: zero-inflation only weakly identified for some covariate values.

- Basic idea of zero-inflated count data regression models: Zeros may arise from either
  - structural zeros, i.e. observations that are 'always zero' or
  - zeros arising from the count data distribution.
-

$$y_i = \kappa_i \tilde{y}_i$$

where κ_i is a binary indicator for structural zeros, i.e.

$$\kappa_i \sim \text{Be}(1 - \pi_i)$$

and $\tilde{y}_i$ follows a standard count data distribution. item Mixed density for the responses y_i:

$$p(y_i) = \pi_i \mathbb{1}_{\{0\}}(y_i) + (1 - \pi_i)\tilde{p}(y_i)$$

where $\tilde{p}(y_i)$ is the density of the count data distribution.

Nonlinear Effects for π