Measuring nuisance parameter effects in Bayesian inference

Alastair Young

Imperial College London

WHOA-PSI-2017

Acknowledgements: Tom DiCiccio, Cornell University; Daniel Garcia Rasines, Imperial College London; Todd Kuffner, WUSTL.

Background

- Inferences drawn from Bayesian data analysis depend on the assumed prior distribution.

Background

- Inferences drawn from Bayesian data analysis depend on the assumed prior distribution.
- Substantial Bayesian robustness literature, quantifying degree to which posterior quantities depend on prior.

Background

- Inferences drawn from Bayesian data analysis depend on the assumed prior distribution.
- Substantial Bayesian robustness literature, quantifying degree to which posterior quantities depend on prior.
- Especially important with hierarchical models, interpretability deeper in hierarchy becomes challenging.

Approaches

- Global: all possible inferences arising from class of priors considered, hope resulting class of inferences is small, so robust to prior. Limited by difficulty of computing range of posterior quantity as prior varies.

Approaches

- Global: all possible inferences arising from class of priors considered, hope resulting class of inferences is small, so robust to prior. Limited by difficulty of computing range of posterior quantity as prior varies.

- Local: based on infinitesimal perturbations to prior, differentiation of posterior quantities wrt prior.

Typical inferential problem

Let $Y = \{Y_1, \ldots, Y_n\}$ be random sample from underlying distribution F_θ, indexed by $d+1$-dimensional parameter $\theta = (\theta^1, \ldots, \theta^{d+1}) = (\psi, \chi)$.

Have $\psi = \theta^1$ a scalar parameter of interest, χ a d-dimensional nuisance parameter.

Key question

How sensitive is the (marginal) posterior inference on ψ to presence in model of nuisance parameter χ and assumed form of prior distribution?

Purpose

- Motivated by corresponding frequentist ideas, provide machinery, based on Bayesian asymptotics, to measure extent of the influence that nuisance parameters have on the inference.

Purpose

- Motivated by corresponding frequentist ideas, provide machinery, based on Bayesian asymptotics, to measure extent of the influence that nuisance parameters have on the inference.

- Offer direct, practical interpretation of measures of influence in terms of posterior probabilities (in 'full' and 'reduced' models).

Key (frequentist) statistic

For testing $H_0 : \psi = \psi_0$ against a one-sided alternative the signed root likelihood ratio statistic is

$$R \equiv R(\psi_0) = \mathrm{sgn}(\hat{\psi} - \psi_0)\big[2\big\{l(\hat{\theta}) - l(\psi_0, \hat{\chi}_0)\big\}\big]^{1/2},$$

where $l(\theta) = l(\psi, \chi)$ is the log-likelihood function, $\hat{\theta} = (\hat{\psi}, \hat{\chi})$ is the MLE of θ, and $\hat{\chi}_0$ is the constrained MLE of χ given the value $\psi = \psi_0$ of the interest parameter ψ.

Bayesian posterior approximation: 'full model'

For continuous prior density $\pi(\theta) = \pi(\psi, \chi)$, the posterior tail probability given the data $Y = (Y_1, \ldots, Y_n)$ is

$$P(\psi \geq \psi_0 | Y) = \Phi(R^*) + O(n^{-3/2}),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution,

$$R^* \equiv R^*(\psi_0) = R + R^{-1} \log\big(T/R\big),$$

$$T \equiv T(\psi_0) = I_\psi(\psi_0, \hat{\chi}_0) \frac{|-I_{\chi\chi}(\psi_0, \hat{\chi}_0)|^{1/2}}{|-I_{\theta\theta}(\hat{\theta})|^{1/2}} \frac{\pi(\hat{\theta})}{\pi(\psi_0, \hat{\chi}_0)},$$

and $\hat{\psi} - \psi_0$ is of order $O(n^{-1/2})$.

A factorization

We may factorize

$$T = T_{\mathrm{INF}}\, T_{\mathrm{NP}},$$

with

$$T_{\mathrm{INF}} \equiv T_{\mathrm{INF}}(\psi_0) = \frac{\pi^{\psi}(\hat{\psi})}{\pi^{\psi}(\psi_0)} I_{\psi}(\psi_0, \hat{\chi}_0)\{-I^{\psi\psi}(\hat{\theta})\}^{1/2},$$

$$T_{\mathrm{NP}} \equiv T_{\mathrm{NP}}(\psi_0) = \frac{|-I_{\chi\chi}(\psi_0, \hat{\chi}_0)|^{1/2}}{|-I_{\chi\chi}(\hat{\theta})|^{1/2}} \frac{\pi^{\chi|\psi}(\hat{\theta})}{\pi^{\chi|\psi}(\psi_0, \hat{\chi}_0)},$$

with $\pi^{\psi}(\psi)$ the marginal prior density for ψ and $\pi^{\chi|\psi}(\theta)$ the conditional prior density for χ given ψ.

Properties

For Bayesian inference,

$$R^* = R + Z_{\mathrm{NP}} + Z_{\mathrm{INF}},$$

where $Z_{\mathrm{NP}} = R^{-1} \log T_{\mathrm{NP}}$ and $Z_{\mathrm{INF}} = R^{-1} \log\big(T_{\mathrm{INF}}/R\big)$.

Properties

For Bayesian inference,

$$R^* = R + Z_{\mathrm{NP}} + Z_{\mathrm{INF}},$$

where $Z_{\mathrm{NP}} = R^{-1} \log T_{\mathrm{NP}}$ and $Z_{\mathrm{INF}} = R^{-1} \log\big(T_{\mathrm{INF}}/R\big)$.

This decomposition has two important properties:

Properties

For Bayesian inference,

$$R^* = R + Z_{\mathrm{NP}} + Z_{\mathrm{INF}},$$

where $Z_{\mathrm{NP}} = R^{-1} \log T_{\mathrm{NP}}$ and $Z_{\mathrm{INF}} = R^{-1} \log\big(T_{\mathrm{INF}}/R\big)$.

This decomposition has two important properties:

- Both T_{INF} and T_{NP} are invariant under interest-respecting transformations, hence Z_{NP} and Z_{INF} also have this property.

Properties

For Bayesian inference,

$$R^* = R + Z_{\mathrm{NP}} + Z_{\mathrm{INF}},$$

where $Z_{\mathrm{NP}} = R^{-1} \log T_{\mathrm{NP}}$ and $Z_{\mathrm{INF}} = R^{-1} \log\big(T_{\mathrm{INF}}/R\big)$.

This decomposition has two important properties:

- Both T_{INF} and T_{NP} are invariant under interest-respecting transformations, hence Z_{NP} and Z_{INF} also have this property.
- T_{NP} does not appear in the tail probability formula when nuisance parameters are absent, so Z_{NP} also vanishes in this case.

Relationship with posterior mean

The posterior mean of R satisfies

$$E\{R(\psi)|Y\} = -\{g_{\mathrm{NP}} + g_{\mathrm{INF}}\} + O(n^{-1}),$$

where

$$g_{\mathrm{NP}} = \lim_{\psi_0 \to \hat{\psi}} Z_{\mathrm{NP}}(\psi_0), \quad g_{\mathrm{INF}} = \lim_{\psi_0 \to \hat{\psi}} Z_{\mathrm{INF}}(\psi_0).$$

Standard calculations show that $Z_{\mathrm{NP}}(\psi_0) = g_{\mathrm{NP}} + O(n^{-1})$ and $Z_{\mathrm{INF}}(\psi_0) = g_{\mathrm{INF}} + O(n^{-1})$, for ψ_0 in $O(n^{-1/2})$ neighborhood of $\hat{\psi}$.

Key observation

Z_{INF} agrees to error of order $O(n^{-1})$ with the value it would take in the reduced problem where the nuisance parameter χ is set equal to the maximum likelihood estimate $\hat{\chi}$ and Bayesian inference is considered only for the scalar parameter ψ based on the prior density $\pi^{\psi}(\psi)$.

Key observation

Z_{INF} agrees to error of order $O(n^{-1})$ with the value it would take in the reduced problem where the nuisance parameter χ is set equal to the maximum likelihood estimate $\hat{\chi}$ and Bayesian inference is considered only for the scalar parameter ψ based on the prior density $\pi^{\psi}(\psi)$.

It is apparent that Z_{INF}, at least to error of order $O(n^{-1})$, does not account for the presence of nuisance parameters.

Key observation

Z_{INF} agrees to error of order $O(n^{-1})$ with the value it would take in the reduced problem where the nuisance parameter χ is set equal to the maximum likelihood estimate $\hat{\chi}$ and Bayesian inference is considered only for the scalar parameter ψ based on the prior density $\pi^{\psi}(\psi)$.

It is apparent that Z_{INF}, at least to error of order $O(n^{-1})$, does not account for the presence of nuisance parameters.

Nuisance parameters are accounted for by Z_{NP}.

Proposal

Candidate measures to assess the extent of the influence that the nuisance parameters have on the Bayesian inference might be the ratios $g_{\mathrm{NP}}/g_{\mathrm{INF}}$ and $Z_{\mathrm{NP}}/Z_{\mathrm{INF}}$.

Proposal

Candidate measures to assess the extent of the influence that the nuisance parameters have on the Bayesian inference might be the ratios $g_{\mathrm{NP}}/g_{\mathrm{INF}}$ and $Z_{\mathrm{NP}}/Z_{\mathrm{INF}}$.

We offer a practical interpretation of this ratio in terms of posterior probabilities.

Proposal

Candidate measures to assess the extent of the influence that the nuisance parameters have on the Bayesian inference might be the ratios $g_{\mathrm{NP}}/g_{\mathrm{INF}}$ and $Z_{\mathrm{NP}}/Z_{\mathrm{INF}}$.

We offer a practical interpretation of this ratio in terms of posterior probabilities.

Asymptotics guiding the relevant finite-sample quantity to examine.

Details

Let the reference value $\hat{\psi}_{1-\alpha}$ be the asymptotic $1-\alpha$ posterior percentage point of ψ, given by $R(\hat{\psi}_{1-\alpha}) = z_\alpha$, where z_α is the α-quantile of the standard normal distribution.

The relative difference in the percentile bias of the asymptotic percentage point $\hat{\psi}_{1-\alpha}$ between the full and the reduced problems is

$$\frac{\{P(\psi \le \hat{\psi}_{1-\alpha}|Y) - (1-\alpha)\} - \{P_{\text{Reduced}}(\psi \le \hat{\psi}_{1-\alpha}|Y) - (1-\alpha)\}}{P_{\text{Reduced}}(\psi \le \hat{\psi}_{1-\alpha}|Y) - (1-\alpha)}$$

$$= \frac{\exp\{z_{1-\alpha} Z_{\text{NP}}(\hat{\psi}_{1-\alpha})\} - 1}{1 - \exp\{-z_{1-\alpha} Z_{\text{INF}}(\hat{\psi}_{1-\alpha})\}},$$

to error of order $O(n^{-1})$.

The relative difference in the percentile bias of the asymptotic percentage point $\hat{\psi}_{1-\alpha}$ between the full and the reduced problems is

$$\frac{\{P(\psi \le \hat{\psi}_{1-\alpha}|Y) - (1-\alpha)\} - \{P_{\text{Reduced}}(\psi \le \hat{\psi}_{1-\alpha}|Y) - (1-\alpha)\}}{P_{\text{Reduced}}(\psi \le \hat{\psi}_{1-\alpha}|Y) - (1-\alpha)}$$

$$= \frac{\exp\{z_{1-\alpha} Z_{\text{NP}}(\hat{\psi}_{1-\alpha})\} - 1}{1 - \exp\{-z_{1-\alpha} Z_{\text{INF}}(\hat{\psi}_{1-\alpha})\}},$$

to error of order $O(n^{-1})$.

In examples, such as high-dimensional linear regression, found to be highly accurate.

Simplifications

To error of order $O(n^{-1})$ we have

$$\frac{\exp\{z_{1-\alpha} Z_{\mathrm{NP}}(\hat{\psi}_{1-\alpha})\} - 1}{1 - \exp\{-z_{1-\alpha} Z_{\mathrm{INF}}(\hat{\psi}_{1-\alpha})\}} = \frac{Z_{\mathrm{NP}}(\hat{\psi}_{1-\alpha})}{Z_{\mathrm{INF}}(\hat{\psi}_{1-\alpha})} = \frac{g_{\mathrm{NP}}}{g_{\mathrm{INF}}}.$$

Hierarchical specification

The prior distribution of the parameter (ψ, χ) that determines the distribution of Y may itself depend on a hyperparameter β which is given a prior distribution.

Hierarchical specification

The prior distribution of the parameter (ψ, χ) that determines the distribution of Y may itself depend on a hyperparameter β which is given a prior distribution.

The entire prior density is of the form

$$\pi(\psi, \chi, \beta) = \pi^{\psi,\chi|\beta}(\psi, \chi, \beta)\pi^{\beta}(\beta),$$

while the log-likelihood function depends only on (ψ, χ).

Hierarchical specification

The prior distribution of the parameter (ψ, χ) that determines the distribution of Y may itself depend on a hyperparameter β which is given a prior distribution.

The entire prior density is of the form

$$\pi(\psi, \chi, \beta) = \pi^{\psi,\chi|\beta}(\psi, \chi, \beta)\pi^{\beta}(\beta),$$

while the log-likelihood function depends only on (ψ, χ).

A factorization $T = T_{\mathrm{INF}}\, T_{\mathrm{NP}}$ can be obtained as before by first integrating the complete prior density $\pi(\psi, \chi, \beta)$ with respect to β to obtain the marginal density $\pi^{\psi,\chi}(\psi, \chi) = \int \pi(\psi, \chi, \beta) d\beta$.

Example: Poisson distribution

$Y_1, \ldots, Y_{d+1}$ are independent, where Y_i has the Poisson distribution with mean $\lambda_i t_i$ and t_i is known, $i = 1, \ldots, d+1$. Thus, the model parameter is $\lambda = (\lambda_1, \ldots, \lambda_{d+1})$; suppose the scalar parameter of interest is a component of λ, say $\psi = \lambda_1$, so the nuisance parameter is $\chi = (\lambda_2, \ldots, \lambda_{d+1})$.

Priors

The prior distribution involves a scalar hyperparameter β: the model parameters $\lambda_1, \ldots, \lambda_{d+1}$ are assumed to be a sample from the gamma distribution with shape parameter ω_0 and scale parameter β, and β is assumed to have the inverse gamma distribution with shape parameter γ_0 and scale parameter δ_0; thus,

$$\pi^{\lambda_1,\ldots,\lambda_{d+1}|\beta}(\lambda_1, \ldots, \lambda_{d+1}, \beta) = \frac{1}{\{\Gamma(\omega_0)\beta^{\omega_0}\}^{d+1}} \Big(\prod_{i=1}^{d+1} \lambda_i\Big)^{\omega_0 - 1} \exp\left(-\sum_{i=1}^{d+1} \frac{\lambda_i}{\beta}\right),$$

and

$$\pi^{\beta}(\beta) = \frac{1}{\Gamma(\gamma_0)\delta_0^{\gamma_0}} \beta^{-(\gamma_0+1)} \exp\left(-\frac{1}{\beta\delta_0}\right).$$

Data example

Here, λ_i failure rates pumps nuclear power plant.

Have 10 observed Poisson variables, so $d = 9$; the observed values are

i	1	2	3	4	5	6	7	8	9	10
Y_i	5	1	5	14	3	19	1	1	4	22
t_i	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

Data example

Here, λ_i failure rates pumps nuclear power plant.

Have 10 observed Poisson variables, so $d = 9$; the observed values are

i	1	2	3	4	5	6	7	8	9	10
Y_i	5	1	5	14	3	19	1	1	4	22
t_i	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

Fix priors parameter values as $(\omega_0, \gamma_0, \delta_0) = (1.802, 0.01, 1)$, as suggested in previous analyses.

Values of g_{NP}, g_{INF}, and the ratio $g_{\mathrm{NP}}/g_{\mathrm{INF}}$ for different interest parameters:

	λ_1	λ_3	λ_5	λ_7	λ_9
g_{INF}	0.6160	0.5971	0.4669	0.5835	0.1398
g_{NP}	-0.01009	-0.01664	-0.3287	-1.164	-1.454
$g_{\mathrm{NP}}/g_{\mathrm{INF}}$	-0.0164	-0.0279	-0.7039	-1.994	-10.401

Dependence on prior

Investigate the susceptibility of the ratio $g_{\mathrm{NP}}/g_{\mathrm{INF}}$ to the choice of $(\omega_0, \gamma_0, \delta_0)$, at least locally, by considering the derivative of the ratio with respect to the components. Table gives derivatives at $(\omega_0, \gamma_0, \delta_0) = (1.802, 0.01, 1)$:

wrt	λ_1	λ_3	λ_5	λ_7	λ_9
ω_0	0.002012	0.003578	0.1595	0.6323	6.924
γ_0	0.03137	0.04655	0.04903	-1.027	-22.884
δ_0	0.05213	0.07496	-0.02642	-1.1910	-15.611

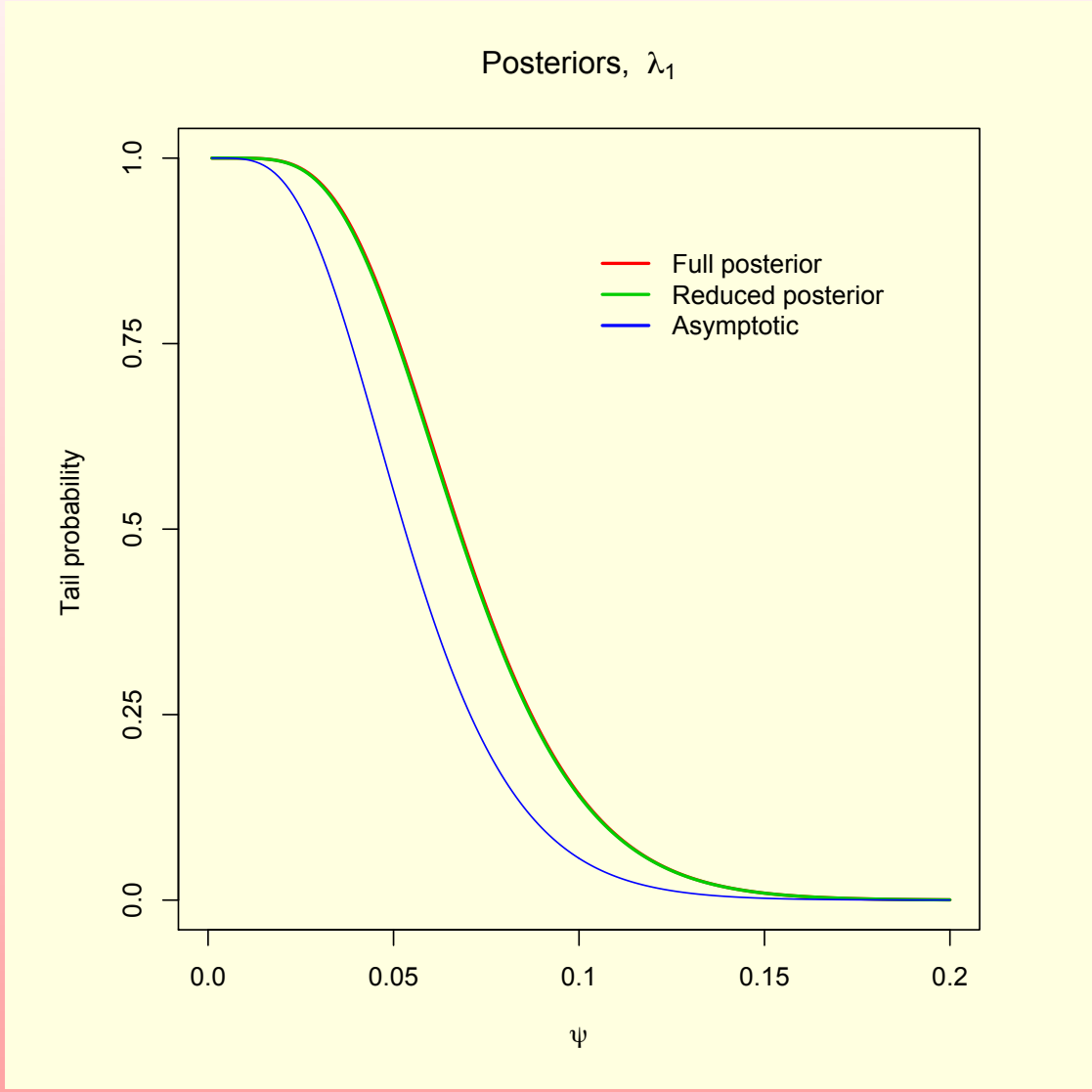
Dependence on prior

Investigate the susceptibility of the ratio $g_{\mathrm{NP}}/g_{\mathrm{INF}}$ to the choice of $(\omega_0, \gamma_0, \delta_0)$, at least locally, by considering the derivative of the ratio with respect to the components. Table gives derivatives at $(\omega_0, \gamma_0, \delta_0) = (1.802, 0.01, 1)$:

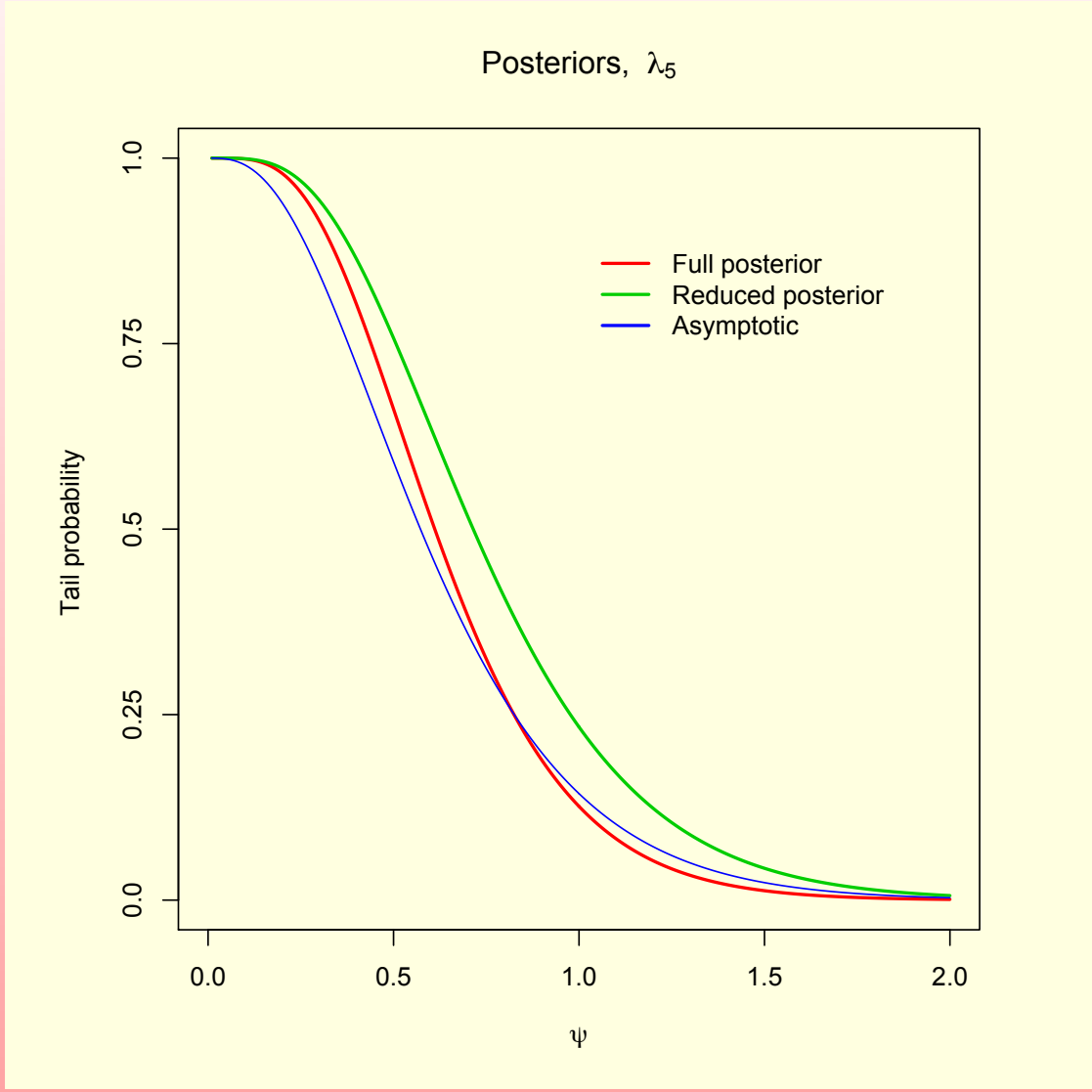
wrt	λ_1	λ_3	λ_5	λ_7	λ_9
ω_0	0.002012	0.003578	0.1595	0.6323	6.924
γ_0	0.03137	0.04655	0.04903	-1.027	-22.884
δ_0	0.05213	0.07496	-0.02642	-1.1910	-15.611

Influence of the nuisance parameters is relatively unaffected by the choice of ω_0, γ_0 and δ_0 for $\lambda_1, \ldots, \lambda_6$, moderately affected for λ_7 and λ_8, significantly affected for λ_9 and λ_{10}.

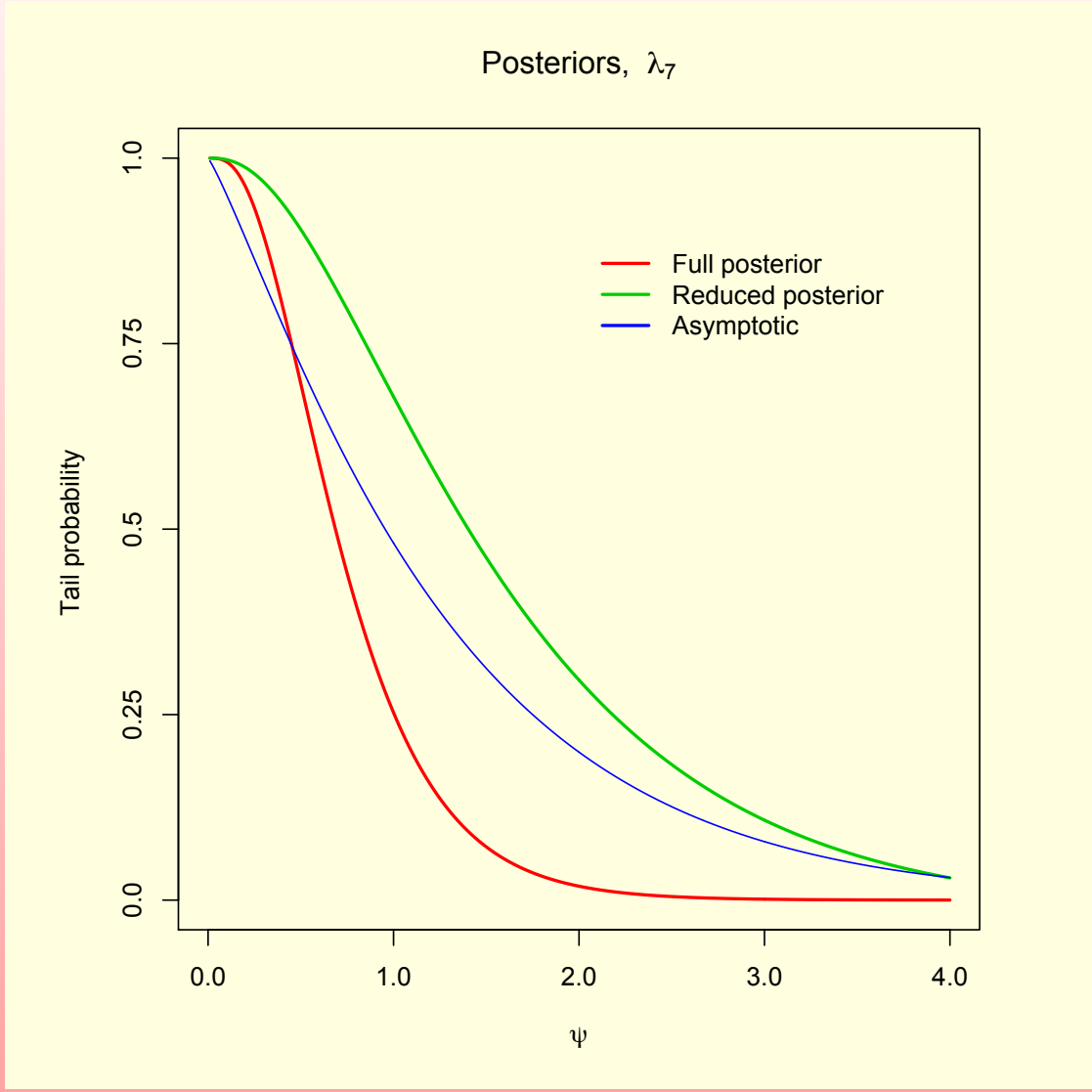
Comparisons, λ_1



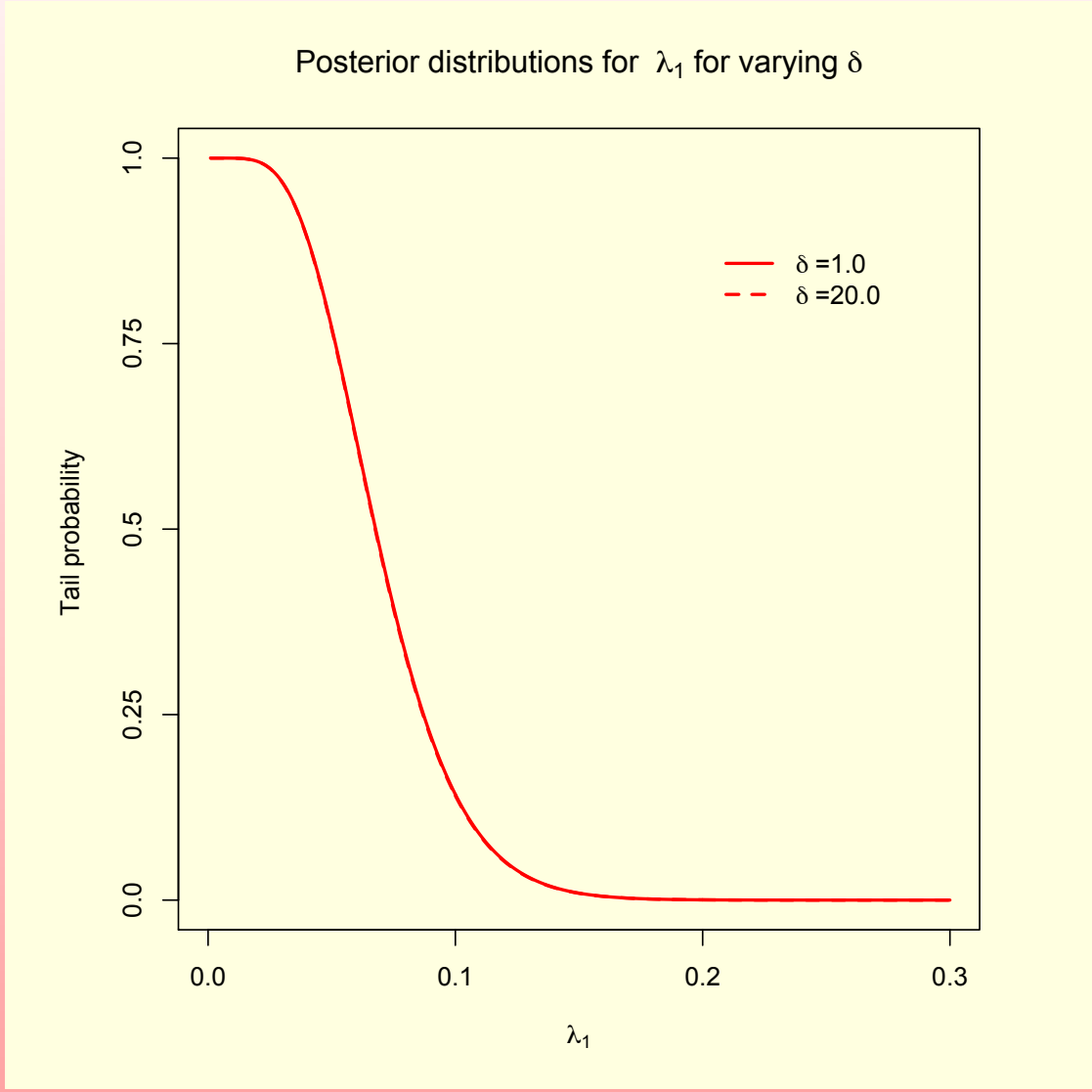
Comparisons, λ_5



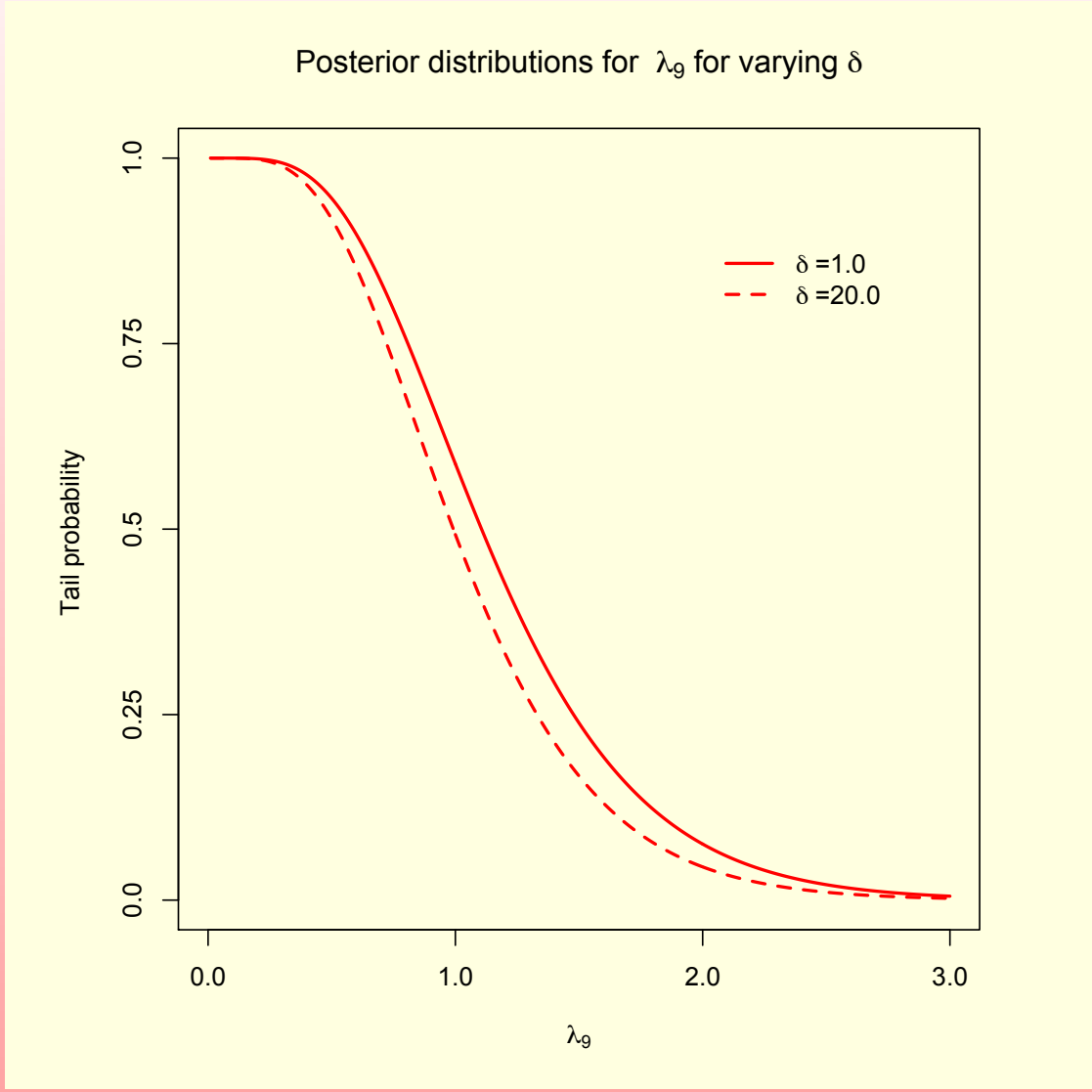
Comparisons, λ_7



Varying δ, interest parameterλ_1



Varying δ, interest parameterλ_9



Summary/Conclusions

Summary/Conclusions

- Decomposition of Bayesian version of adjusted signed root likelihood ratio statistic provides a simple computational machinery for analysis of effects of prior assumptions on nuisance parameters on a marginal inference on an interest parameter.

Summary/Conclusions

- Decomposition of Bayesian version of adjusted signed root likelihood ratio statistic provides a simple computational machinery for analysis of effects of prior assumptions on nuisance parameters on a marginal inference on an interest parameter.

- Measures have direct interpretation in terms of posterior probabilities.

- Provide practically useful, readily computed means of assessing nuisance parameter effects in Bayesian analysis.

- Provide practically useful, readily computed means of assessing nuisance parameter effects in Bayesian analysis.

- Calibration: what constitutes a large ratio? Best considered by comparing measures for different priors. Calculation across range of priors allows a straightforward approach to sensitivity analysis/ evaluation of robustness.

- Provide practically useful, readily computed means of assessing nuisance parameter effects in Bayesian analysis.

- Calibration: what constitutes a large ratio? Best considered by comparing measures for different priors. Calculation across range of priors allows a straightforward approach to sensitivity analysis/ evaluation of robustness.

- Permits identification of what components of prior specification have substantial effect on posterior marginal inference of interest. Especially valuable within hierarchical framework.

Some relevant references

Bayesian asymptotics:

DiCiccio & Martin (1993), JRSSB, **55**, 305-316.

DiCiccio & Young (2010), Biometrika, **97**, 497-504.

DiCiccio, Kuffner & Young (2012), Biometrika, **99**, 675-686.

Frequentist asymptotics:

DiCiccio, Kuffner & Young (2015), JSPI, **165**, 1-12.

DiCiccio, Kuffner, Young & Zaretzki (2015), Stat. Sinica, **25**, 1355-1376.

DiCiccio, Kuffner & Young (2017), JSPI, to appear.