

Issues in hierarchical and non hierarchical combining of information

M.J. Bayarri (U Valencia, Duke U, NISS)
... and many collaborators

2008 SAMSI Program on Meta-analysis: Synthesis and appraisal of multiple
sources of empirical evidence. June 2-13, 2008

Outline

We'll briefly consider several issues that arise specially when combining several sources of information. Specifically

- Exchangeable sources (hierarchical combination)
 - Issues arising in the first level (data level): Uncertainty in some aspects of model specification.
 - Issues arising in the second level (information sources level): Inadequacy of prior/metamodel specification
- Disparate sources (non hierarchical combination): combining field data and computer models data.

(Similar issues and worries have been mentioned by previous speakers this week, from different perspectives)

Uncertainty at the data level

- Often there is unavoidable uncertainty about some aspects in the model for the data.
- This uncertainty might have only a mild effect when only one study is considered, but the effect can be highly magnified when combining several studies and simply can not be ignored.
- This is the case in particular when using *weighted distributions*, which are usual models for taking into account unknown publication policies

Meta-analysis of published experiments

(with Degroot, Berger ...)

- Assume that we are combining results from (similar) experiments that have appeared published in the literature.
- Assume all these experiments involved hypothesis testing, and let X_i , $i = 1, \dots, p$ be the test statistics
- For a random experiment, we have a model which we feel reasonably confident about (not strictly a problem of model uncertainty)

$$X_i \mid \theta_i \sim f_i(x_i \mid \theta_i)$$

- **BUT** published experiments are not a random sample from performed experiments. In particular, 'significant' results are more likely to be published than 'non significant' ones (the well known 'File Drawer' problem)

- The actual publication mechanism can be modeled as a weight function $\omega_i(x)$ that distorts the density $f_i(x | \theta_i)$ by which x gets selected, so that the actual density of a **published** x_i is

$$f_i^\omega(x_i | \theta_i) = \frac{\omega_i(x_i) f_i(x_i | \theta_i)}{\nu_i^\omega(\theta_i)}$$

where $\nu_i^\omega(\theta_i) = \mathbb{E}^{X_i|\theta_i}[\omega_i(X_i)]$. Often $\omega_i(x)$ is interpreted as the probability of selecting x_i

- This weight function can have an important impact in the conclusions, with a potentially extremely dramatic effect if many studies are combined.

- The likelihood function of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ for p studies is,

$$L^\omega(\boldsymbol{\theta}) \propto \ell(\boldsymbol{\theta}) \left(\prod_{i=1}^p \nu_i^\omega(\theta_i) \right)$$

where $\ell(\boldsymbol{\theta})$ would be the likelihood function for the unweighed base density. For a (marginal) prior $\pi(\boldsymbol{\theta})$ the posterior is

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto \ell(\boldsymbol{\theta}) \left(\prod_{i=1}^p \nu_i^\omega(\theta_i) \right) \pi(\boldsymbol{\theta})$$

so that at least for moderate and large p , the effect of the weight function can be huge, way larger than that of $\pi(\boldsymbol{\theta})$

Uncertainty in the publication mechanism

- Occasionally, w_i can be taken to be known. For instance, if only results that are significant at the 0.05 level get published in journal i , then if X_i is a standardized (one sided) Normal test statistic, we have

$$w_i(x) = 1_{[1.645, \infty]}(x)$$

but this is very rare.

- Often, parametric classes of weight functions $\omega(x) = x^a$ (when $X > 0$) are assumed, but there is rarely scientific justification for such specific parametric models
- Most often the specification of ω_i is highly uncertain. Two possibilities for dealing with high uncertainty about ω : Robust Bayes and Non-parametric Bayes

Robust Bayes.

Because of the uncertainty in the ω_i it is of particular interest to study the robustness of the analysis to choice of ω_i . In Bayarri and Berger (1998), nonparametric classes of weight functions \mathcal{W} are considered, and range of posterior functionals of interest computed as ω ranges over \mathcal{W}

EXAMPLE:

Assume that studies are published in a journal only if they are

- significant at the 0.05 level
- significant at the .1 level and deemed “exceptionally” important by the editors

then we could consider classes of functions of the form

$$\mathcal{W}_1 = \{\omega : \omega^l(x) \leq \omega(x) \leq \omega^u(x)\}$$

$$\mathcal{W}_2 = \{\text{non decreasing } \omega : \omega^l(x) \leq \omega(x) \leq \omega^u(x)\}$$

with $\omega^l(x) = 1_{[1.645, \infty]}(x)$ and $\omega^u(x) = 1_{[1.282, \infty]}(x)$.

Tightening of the bounds depends on the situation.

- Good news: an appropriate robustness analysis (difficult) can sometimes make the range (upper bound - lower bound) for certain functionals decrease as p increases
- Not so good news: there will always remain unresolved uncertainty (which can be considerable, and unavoidable)

Non parametric Bayes

Berger and Lee (2001) addressed the problem of uncertainty in the weight function by using non-parametric priors for the weight function. In particular, they used (mixtures of) Dirichlet process priors. They found the results to be highly sensitive to choice of the prior distributions, even for moderately large sample sizes, another indication of the difficulty of dealing with uncertainty in the weight functions.

Uncertainty at the ‘combining’ level

(with Berger, Liu)

- Consider the following simplified meta-analysis scenario (random effect model)

$$Y_{ij} = \mu + \delta_i + \epsilon_{ij} \quad j = 1, \dots, n; \quad i = 1 \dots p, \quad p \text{ very large,}$$

$$\epsilon_{ij} \mid \sigma_i^2 \sim N(0, \sigma_i^2),$$

$$\delta_i \mid \tau^2 \sim N(0, \tau^2),$$

- σ_i^2 's and τ^2 are unknown, and for simplicity, we make the argument conditional on μ (fixed for the argument).
- feel confident about the distribution of ϵ (first level o.k.) but not so about δ 's: the second level is usually highly uncertain; Normality is only a ‘default’, ‘convenient’ choice
- This uncertainty can have unanticipated impact in the results

Posterior distribution

With objective priors $\pi(\sigma_i^2) \propto (\sigma_i^2)^{-1}$ and $\pi(\tau^2 | \boldsymbol{\sigma}^2) \propto (\tau^2 + \bar{\sigma}^2/n)^{-1}$ the joint posterior is

$$\pi(\tau^2, \boldsymbol{\sigma}^2 | \bar{\mathbf{y}}, \mathbf{s}^2) \propto \frac{1}{\tau^2 + \bar{\sigma}^2/n} \prod_{i=1}^p \frac{1}{(\sigma_i^2)^{\frac{n+1}{2}}} \exp\left\{-\frac{ns_i^2}{2\sigma_i^2}\right\} \frac{1}{(\tau^2 + \sigma_i^2/n)^{1/2}} \exp\left\{-\frac{1}{2} \frac{(\bar{y}_i - \mu)^2}{\tau^2 + \sigma_i^2/n}\right\}.$$

where $\bar{y}_i = \sum_{r=1}^K y_{ir}$, $s_i^2 = \sum_{r=1}^K (y_{ir} - \bar{y}_i)^2 / n$.

A hint that something is going wrong: Some of the σ_i^2 (variances corresponding to certain ‘effects’) get ‘stuck’ at very large values with the consequence that the corresponding δ_i ’s are estimated as near zero (and this happens for the huge ones!!!)

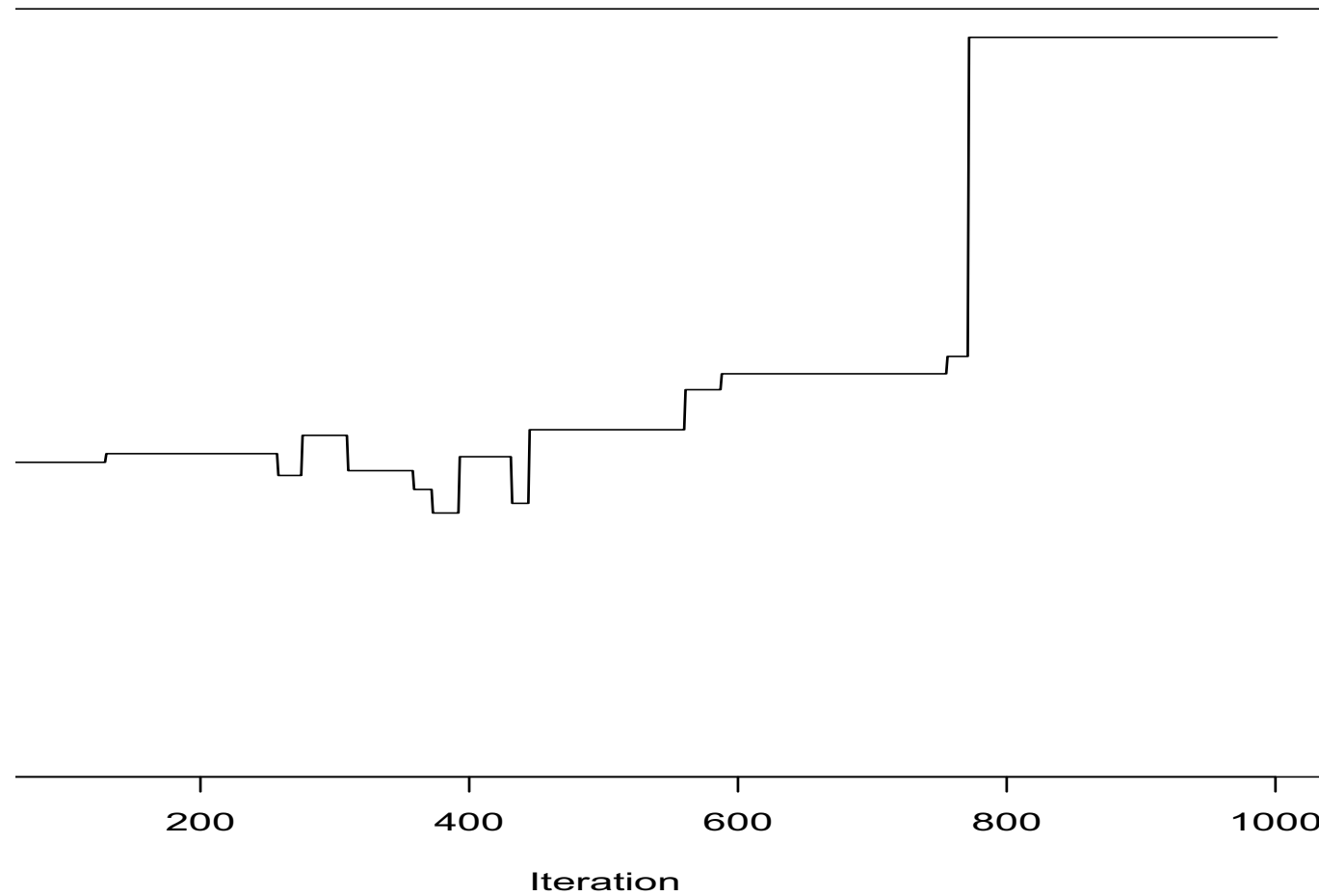


Figure 1: Trace plots for σ_{170}^2 with global Normal analysis;

An explanation:

$$\ell \propto \prod_{i=1}^p \frac{1}{(\sigma_i^2)^{\frac{n-1}{2}}} \exp \left\{ -\frac{ns_i^2}{2\sigma_i^2} \right\} \frac{1}{(\tau^2 + \sigma_i^2/n)^{1/2}} \exp \left\{ -\frac{1}{2} \frac{(\bar{y}_i - \mu)^2}{\tau^2 + \sigma_i^2/n} \right\} .$$

- If one of the random effects, δ_j is huge \leadsto a huge $(\bar{y}_j - \mu)^2$ which should result in an inflated τ^2 (allowing big δ 's).
- MCMC-ing the posterior (or maximizing the LF) can correct for the huge $(\bar{y}_j - \mu)^2$ by either making τ^2 huge or σ_j^2 huge.
- But the penalty for making τ^2 huge ($\approx 1/(\tau^p)$) is much worse than the penalty for making the one single σ_j^2 huge ($\approx \sqrt{n}/(\sigma_j^n)$, with n small compared to p , as in Keith Astronomical situation).
- Hence, the MCMC/likelihood analysis tends to compensate a huge deviation $(\bar{y}_j - \mu)^2$ by making σ_j^2 (and not τ^2) huge.
- It follows that the corresponding effect δ_j will be shrunk toward 0 quite significantly and inappropriately.

A possible solution

- Modeling δ_i as exchangeable normals might not be appropriate because of this, possibly very misleading, handling of the outliers.
- A more robust Bayesian model for the δ_i (that can accommodate extreme realizations) is the Cauchy distribution

$$\pi(\delta_i | \tau^2) \sim \text{Cauchy}(0, \tau^2) .$$

- Using this, our problem disappeared, but computation increases, one has to deal with numerical ‘unpleasantness’ of super heavy tails. Also, in other situations it might not be the solution, or it might not be clear that it is.

The Modular Solution:

This is 'quick and dirty' fix to avoid this type of unanticipated problems caused by 'parts' of the global model specifications about which we are not (we can not really be) very confident about

- Modular approach: prevent unanticipated 'surprises' coming from the 'uncertain' parts of the model to 'contaminate' the 'safe' part.
- Much more about 'modularization' and related approaches in work with coauthors Jim Berger and Fei Liu
- For our example: keep the 'uncertain' normality assumption for the δ_i , but 'separate' the learning for the σ_i^2 from the rest of the Bayesian analysis, i.e., work with the distribution

$$\pi(\{\sigma_i^2\} \mid \{s_i^2\}) \times \pi(\tau^2 \mid \{\sigma_i^2\}, \{\bar{y}_i, s_i^2\})$$

rather than, say, $\pi(\{\sigma_i^2\} \mid \{\bar{y}_i, s_i^2\}) \times \pi(\tau^2 \mid \{\sigma_i^2\}, \{\bar{y}_i, s_i^2\})$.

Checking the second level of a hierarchical model

(with Castellanos)

Assume model

$$X_{ij} | \theta_i \stackrel{i}{\sim} f(x_{ij} | \theta_i) \quad i = 1, \dots, k, \quad j = 1, \dots, n_i$$

$$\theta_i | \boldsymbol{\eta} \stackrel{i}{\sim} \pi(\theta_i | \boldsymbol{\eta}) \quad i = 1, \dots, k$$

$$\boldsymbol{\eta} \sim \pi(\boldsymbol{\eta})$$

Question: model O.K.? \leftrightarrow data compatible with model?

We concentrate on the second level (results for testing the parameter of interest, $\boldsymbol{\eta}$, in the paper, with similar behavior). Want:

- Model check (no comparison) \rightsquigarrow no alternatives
- Objective Bayes \rightsquigarrow no subjective hyper-priors

common approach with no alternative models

- choose a $T = t(\mathbf{X})$ to investigate incompatibility of data with assumed (null) model and compute $t_{obs} = t(\mathbf{x}_{obs})$
- look at some (specified) distribution $f(t)$ of T under the assumed model, and see if t_{obs} is ‘compatible’ with $f(t)$.
- use your favorite method to judge ‘compatibility’ (visual displays, relative height at t_{obs} , p -values ... etc.)
- we use here p -values only for convenience (issues apply to any model checking tools).
- for simplicity: $p\text{-value} = Pr^{f(t)}\{T \geq t_{obs}\}$
- **THE PROBLEM:** if T is not ancillary or nearly so, model checking (both Bayesian and frequentist) is tricky

getting rid of θ the Bayesian way

Bayesian just need to integrate out the parameters to get:

$$m(t) = \int f(t | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

the prior predictive distribution, where

$$\pi(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta} | \boldsymbol{\eta}) \pi(\boldsymbol{\eta}) d\boldsymbol{\eta}$$

- but objective (improper) $\pi(\boldsymbol{\eta}) \rightsquigarrow$ improper $\pi(\boldsymbol{\theta})$
- objective 'Bayes' proposals integrate $\boldsymbol{\theta}$ w.r.t. other (related) distributions for $\boldsymbol{\theta}$

some proposals

We explore three proposals: a usual non-Bayesian, an easy Bayesian, and a more complex Bayesian proposals for $\pi(\boldsymbol{\theta})$

1. Empirical Bayes (plug-in): $\pi^{EB}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{\eta} = \hat{\boldsymbol{\eta}})$,
replacing the Bayes $m(\boldsymbol{x})$ by the profile (integrated) likelihood
2. An over-killing EB (only for comparison purposes):
$$\pi^{EB}(\boldsymbol{\theta} \mid \boldsymbol{x}_{obs}) \propto f(\boldsymbol{x}_{obs} \mid \boldsymbol{\theta}) \pi^{EB}(\boldsymbol{\theta}),$$
3. Posterior: $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{obs}) = \int \pi(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{x}_{obs}) d\boldsymbol{\eta}$ replacing the
Bayes $m(\boldsymbol{x})$ by the *posterior predictive* distribution $m(\boldsymbol{x} \mid \boldsymbol{x}_{obs})$
4. Partial posterior: $\pi(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{x}_{obs} \setminus t_{obs}) \propto f(\boldsymbol{x}_{obs} \mid t_{obs}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}, \boldsymbol{\eta})$
which (approximately) looks at a particular 'slice' of the Bayes
 $m(\boldsymbol{x})$

A hierarchical normal-normal model

$$X_{ij} | \theta_i \stackrel{i}{\sim} N(\theta_i, \sigma_i^2) \quad j = 1, \dots, n \quad i = 1, \dots, p$$

$$\theta_i | \nu, \tau \stackrel{i}{\sim} N(\nu, \tau^2)$$

$$\pi(\nu, \tau^2) \propto 1/\tau$$

consider $T = \max\{\bar{X}_1, \dots, \bar{X}_k\}$ with p -value $p = Pr^{f(\bullet)}\{T \geq t_{obs}\}$

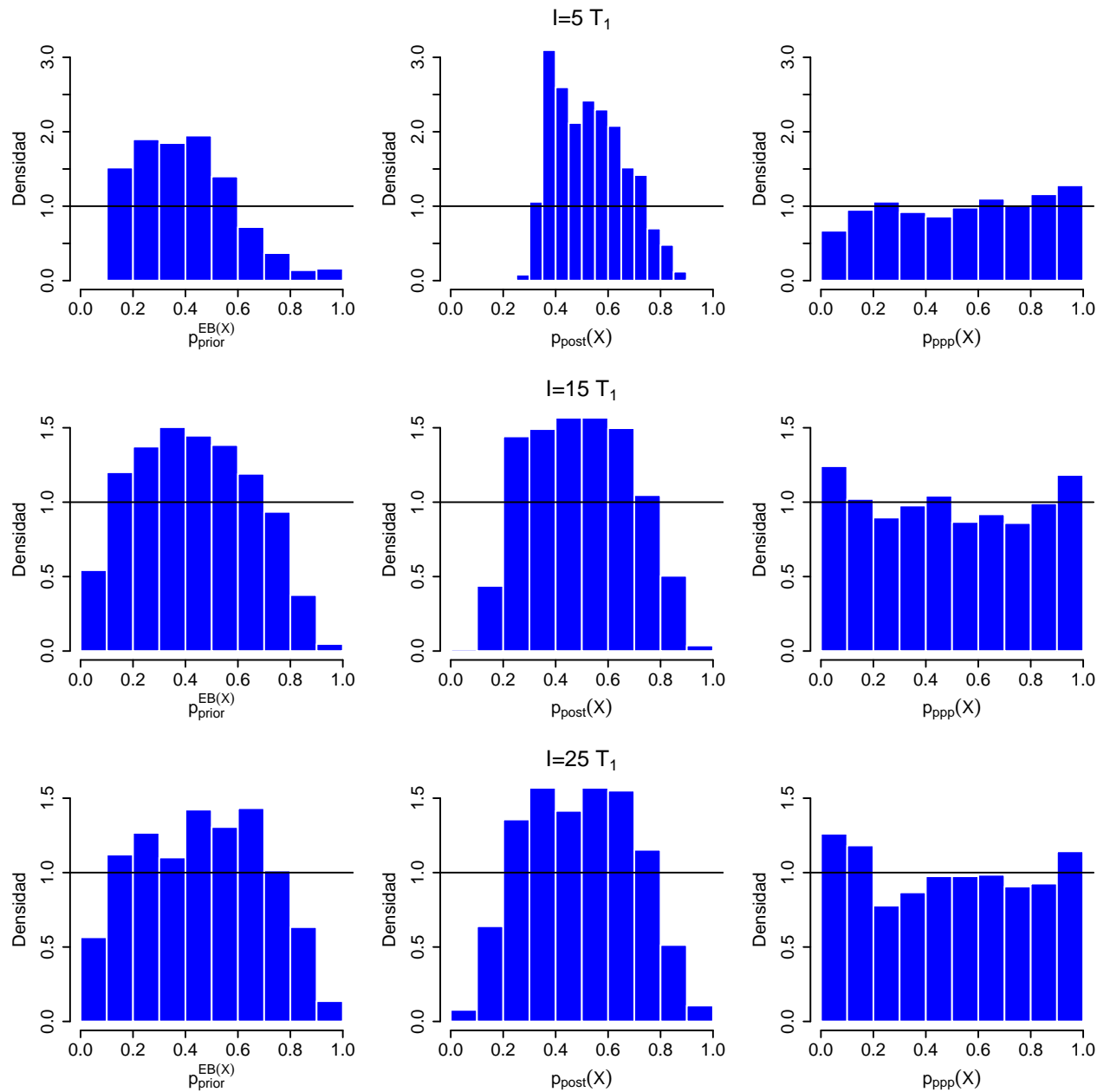
A SIMULATED EXAMPLE with $k = 5$ groups and $n = 8$

$$X_{ij} | \mu_i \sim N(\mu_i, 4), \quad \mu_i \sim N(1, 1), \quad \mu_5 \sim N(5, 1)$$

sample means: 1.56, 0.64, 1.98, 0.01, 6.96 (The mean of the 5th group is 6.65 SD away from the others)

p_{ppp}	p_{prior}^{EB}	p_{post}^{EB}	p_{post}
.015	.195	.371	.405

null distribution of p -values $p(X)$

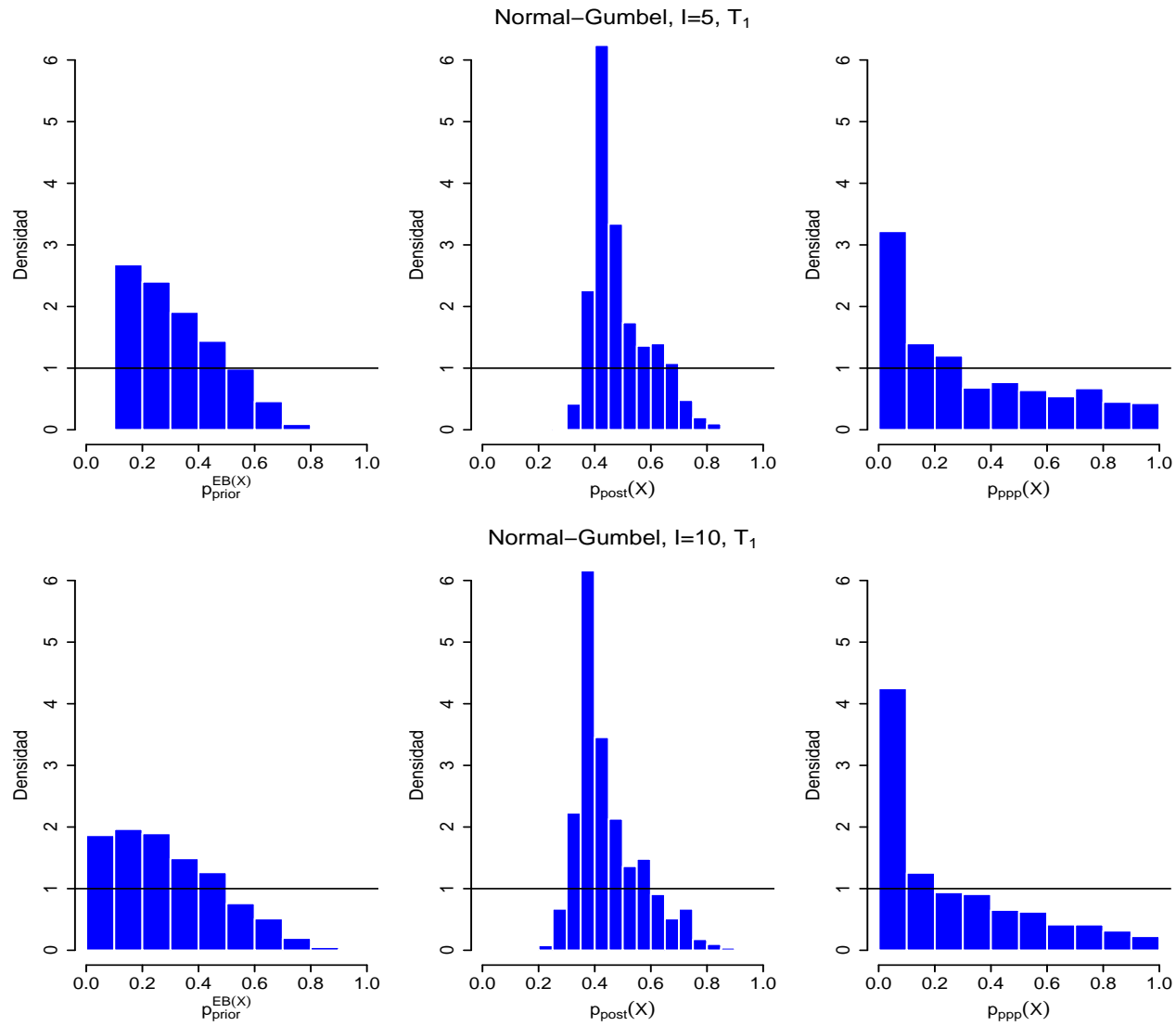


Behavior under alternatives

- “null model”: $X_{ij} | \theta_i \stackrel{i}{\sim} N(\theta_i, \sigma^2), \quad \theta_i | \nu, \tau \stackrel{i}{\sim} N(\nu, \tau^2)$
- Explore behavior of $p_{plug}(\mathbf{X}), p_{post}(\mathbf{X}), p_{ppp}(\mathbf{X})$ when “null model” not true \leadsto POWER
- concentrate in ‘wrong’ second level: simulate X_{ij} from normal and θ_i from non-normal
- First level: $X_{ij} | \mu_i \sim N(\theta_i, 4), \quad n = 8, \quad k = 5, 10$ groups
- second level:
 - $\theta_i \sim Exponential(1)$
 - $\theta_i \sim Gumbel(0, 2)$
 - $\theta_i \sim LogNormal(0, 1)$

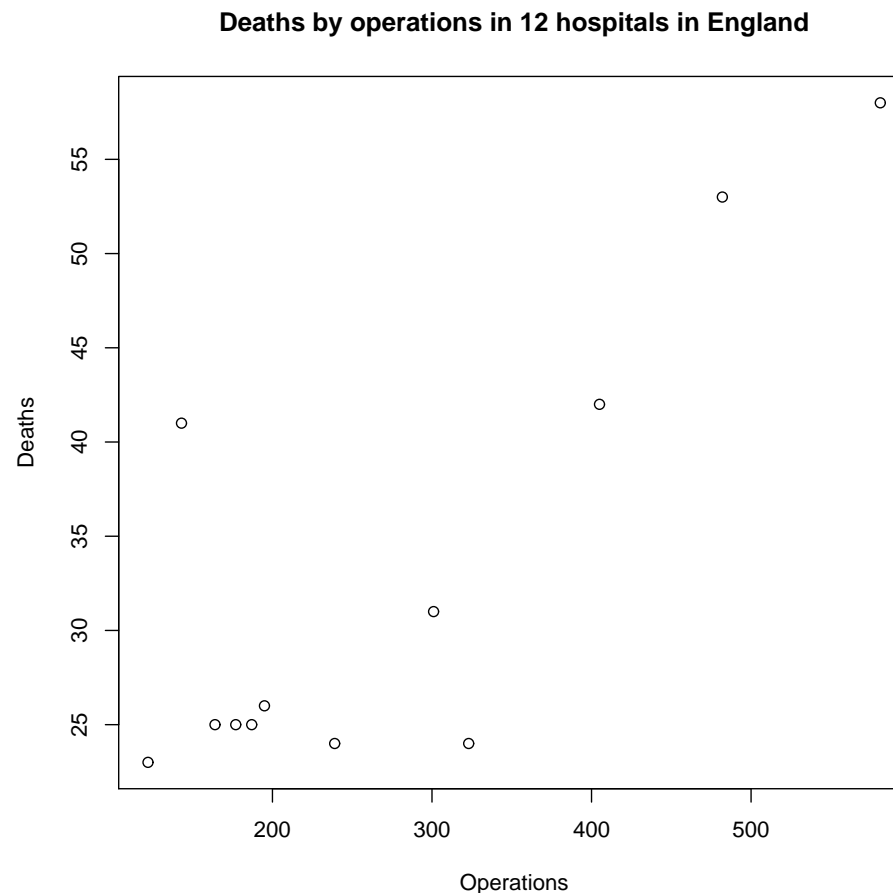
$Pr(p \leq \alpha)$ for some values of α

α	0.02	0.05	0.1	0.2	0.02	0.05	0.1	0.2
Normal-Exponencial								
	k=5				k=10			
p_{ppp}	0.04	0.08	0.15	0.24	0.12	0.20	0.29	0.42
p_{post}	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05
p_{prior}^{EB}	0.00	0.00	0.00	0.23	0.00	0.06	0.18	0.37
Normal-Gumbel								
	k=5				k=10			
p_{ppp}	0.12	0.22	0.32	0.46	0.21	0.31	0.42	0.55
p_{post}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p_{prior}^{EB}	0.00	0.00	0.00	0.23	0.00	0.07	0.19	0.38
Normal-Lognormal								
	k=5				k=10			
p_{ppp}	0.16	0.22	0.31	0.41	0.32	0.42	0.50	0.61
p_{post}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
p_{prior}^{EB}	0.00	0.00	0.00	0.23	0.01	0.06	0.13	0.23



A Binomial-Beta real example: Bristol data

Number n_i of open-heart operations and the corresponding number Y_i of deaths of children under 1 year in 12 hospitals in England, (Spiegelhalter et al. 2002).



Model

$$Y_i | \theta_i \stackrel{i}{\sim} \text{Bin}(\theta_i, n_i), \quad i = 1, \dots, I,$$

$$\pi(\boldsymbol{\theta} | \alpha, \beta) = \prod_{i=1}^I \text{Beta}(\theta_i | \alpha, \beta),$$

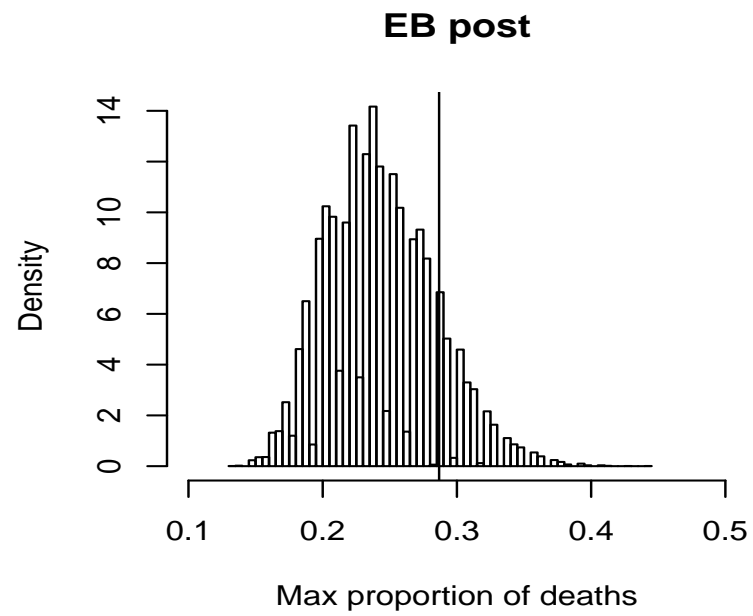
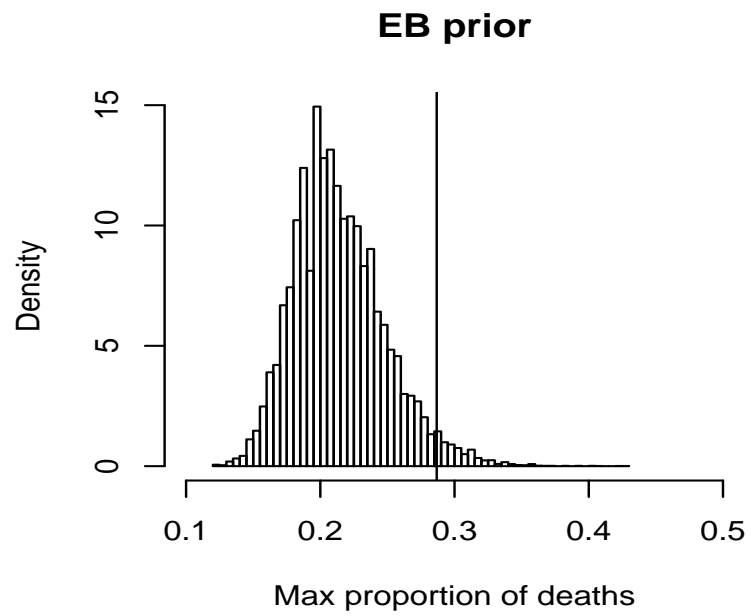
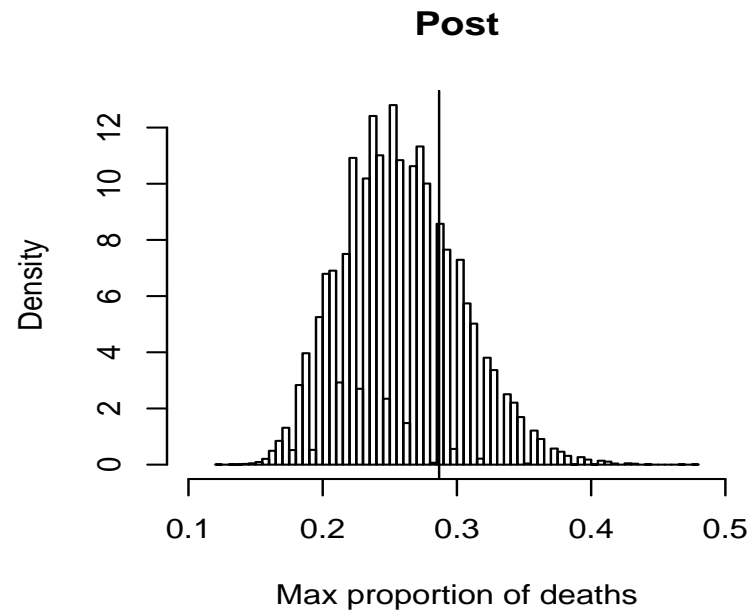
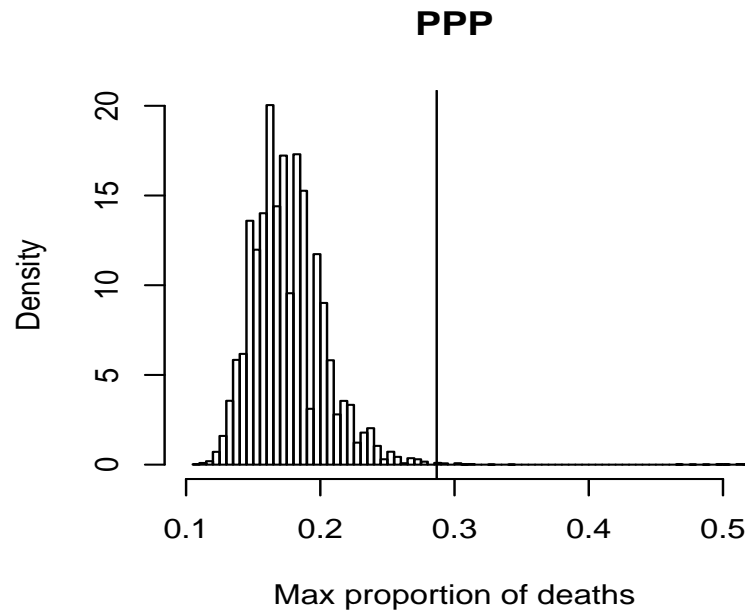
$$\pi(\alpha, \beta) \propto \sqrt{(\psi_1(\alpha) - \psi_1(\alpha + \beta))(\psi_1(\beta) - \psi_1(\alpha + \beta)) - \psi_1(\alpha + \beta)^2},$$

where $\psi_1(x)$ is the trigamma function.

Test statistic $\max \left\{ \frac{y_i}{n_i} \right\}$

p-values

	p_{prior}^{EB}	p_{post}^{EB}	p_{post}	p_{ppp}
max	0.03	0.16	0.23	0.00



Merging disparate sources of data

(with Berger, Liu, Paulo, Sacks and others)

- Meta-analysis goes way beyond combining exchangeable sources of data
- An increasingly popular merging of disparate sources of data for the same phenomena is the merging of data from computer simulators and field data
- Computer simulators are computer implementations of mathematical models, intended as 'surrogates' of reality.
- Useful when physical data is very scarce, very expensive, when extrapolating, etc.

CRASH example

- to evaluate vehicle crashworthiness performance during the vehicle manufacturing process, crashes of vehicles are performed, but this is too expensive for adequate exploration (crashes at different angles, against different barrier types, with different initial velocities ...)
- In addition to these (few) 'field crashes' at different velocities, angles and types, 'virtual' crashes from computer model experiments are performed
- The CRASH computer model simulates the effect of a collision of a vehicle by using a finite element implementation of a non-linear dynamic model requiring extensive geometric information.

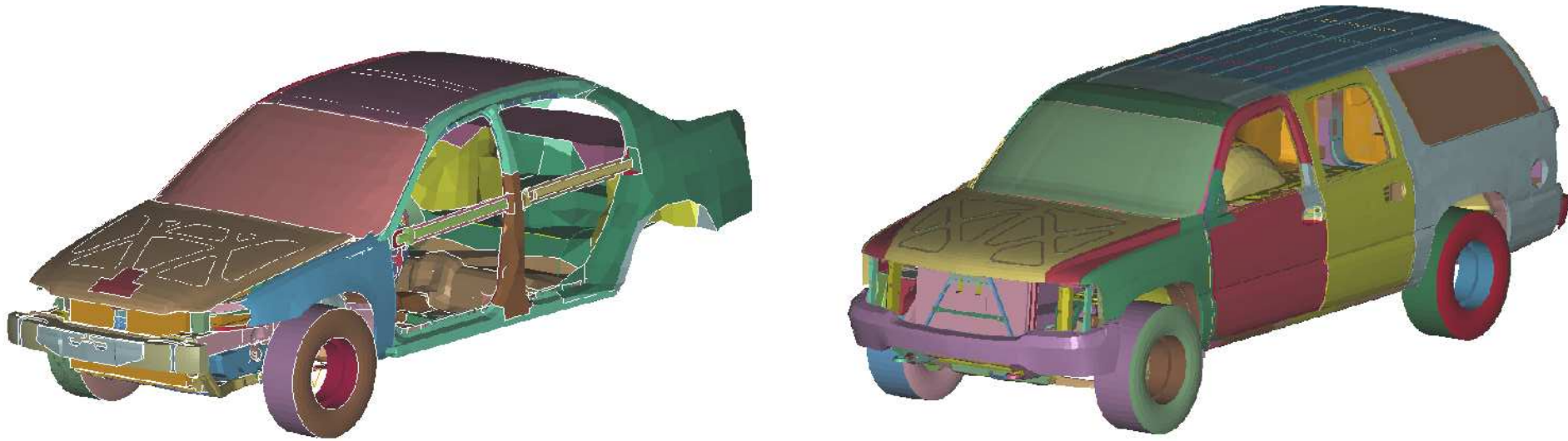


Figure: Typical finite element vehicle models: (a) a car finite element model; (b) a sport utility vehicle model

Want to combine 'physical data' with 'computer model data'

the data

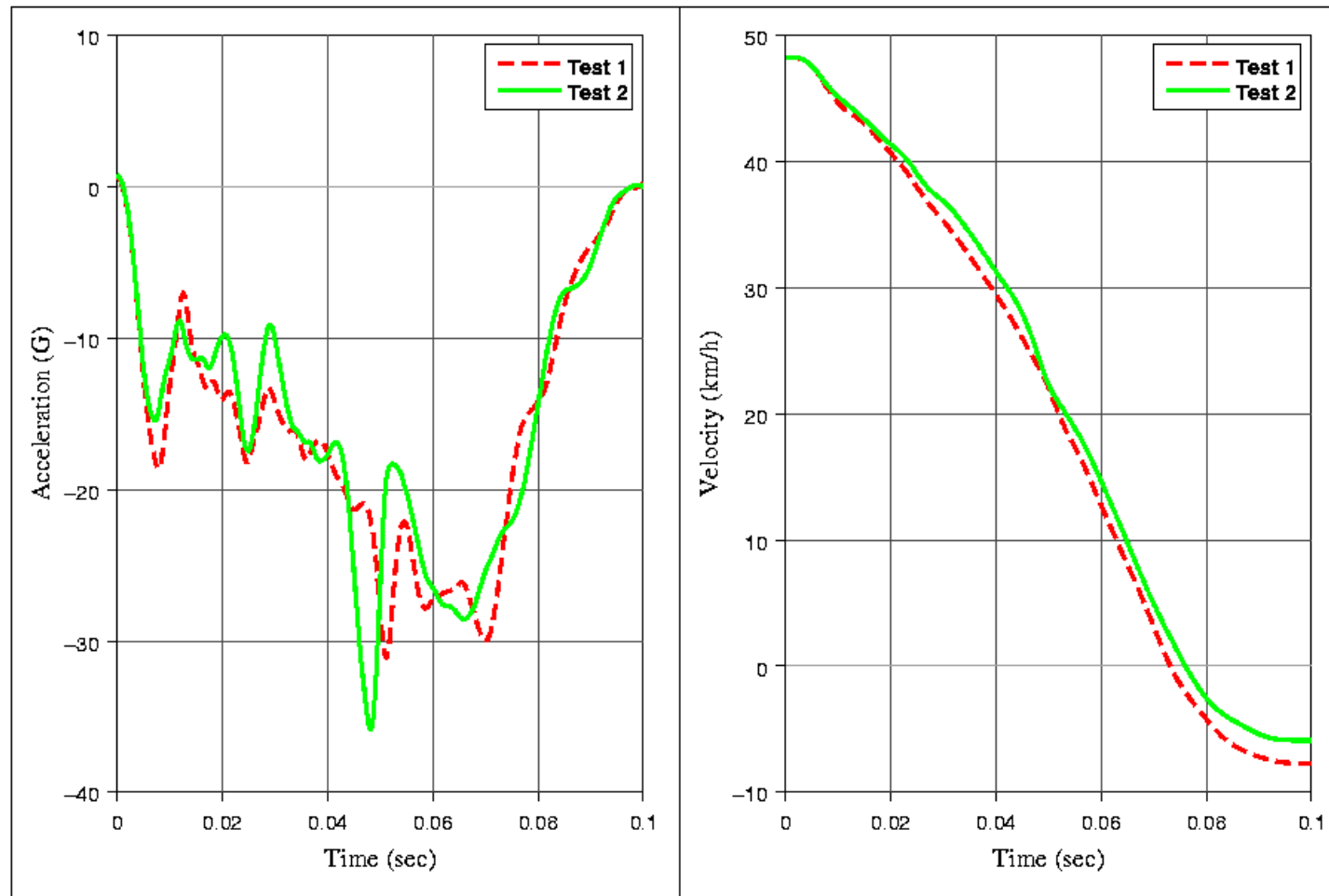


Figure 2: Acceleration and velocity in the occupant sensor from 30mph zero degree rigid barrier frontal impact tests for two production vehicles.

Relating computer model and field data

- output of interest: velocity of the driver's seat sensor (SDM) over the initial 100 millisecond time interval of a frontal barrier collision at impact velocity v
- field = reality + error $\rightsquigarrow y^F(v_i, t) = y^R(v_i, t) + \varepsilon(t)$
- reality = model + bias $\rightsquigarrow y^R(v, t) = y^M(v, t) + b(v, t)$
- Ultimate goal: predict CRITV = velocity calculated 30 ms before the SDM displacement reaches 125 mm (The airbag takes around 30 ms to deploy)
- Scarcity of data. Little data is available
 - Field data is very expensive
 - Computer model runs take days

Bayesian analysis: $y^F(v, t) = y^M(v, t) + b(v, t) + \varepsilon(t)$

- * Conceptually very simple
 - Assess priors to unknowns (functions and parameters)
 - Collect field data $y^F(v_i^*, t)$, $i \in I$ and computer model runs $y^M(v_j, t)$, $j \in J$
 - Compute the posterior distribution of reality $y^R(v, t)$ at impact velocity v , and from it the posterior distribution of CRITV at v

- * along with the unknown functions $b(\cdot, \cdot)$, $\varepsilon(\cdot)$, model $y^M(\cdot, \cdot)$ has to be considered as an unknown function also when the simulator is slow (as in our example). Indeed, $y^M(\cdot, \cdot)$ is known only at the few inputs v_j , $j \in J$, not for the other values of v .

- * We use Gaussian stochastic processes response surfaces as priors for these functions (simplifications were required because of the scarcity of data)
- * The posterior mean of $y^M(\cdot, \cdot)$ is an *emulator*, a fast, statistical approximation to the simulator, that passes through the performed computer model runs and interpolates otherwise, while the variance provides accuracy of the approximation (fast approximation of the simulator needed for the MCMC analysis)

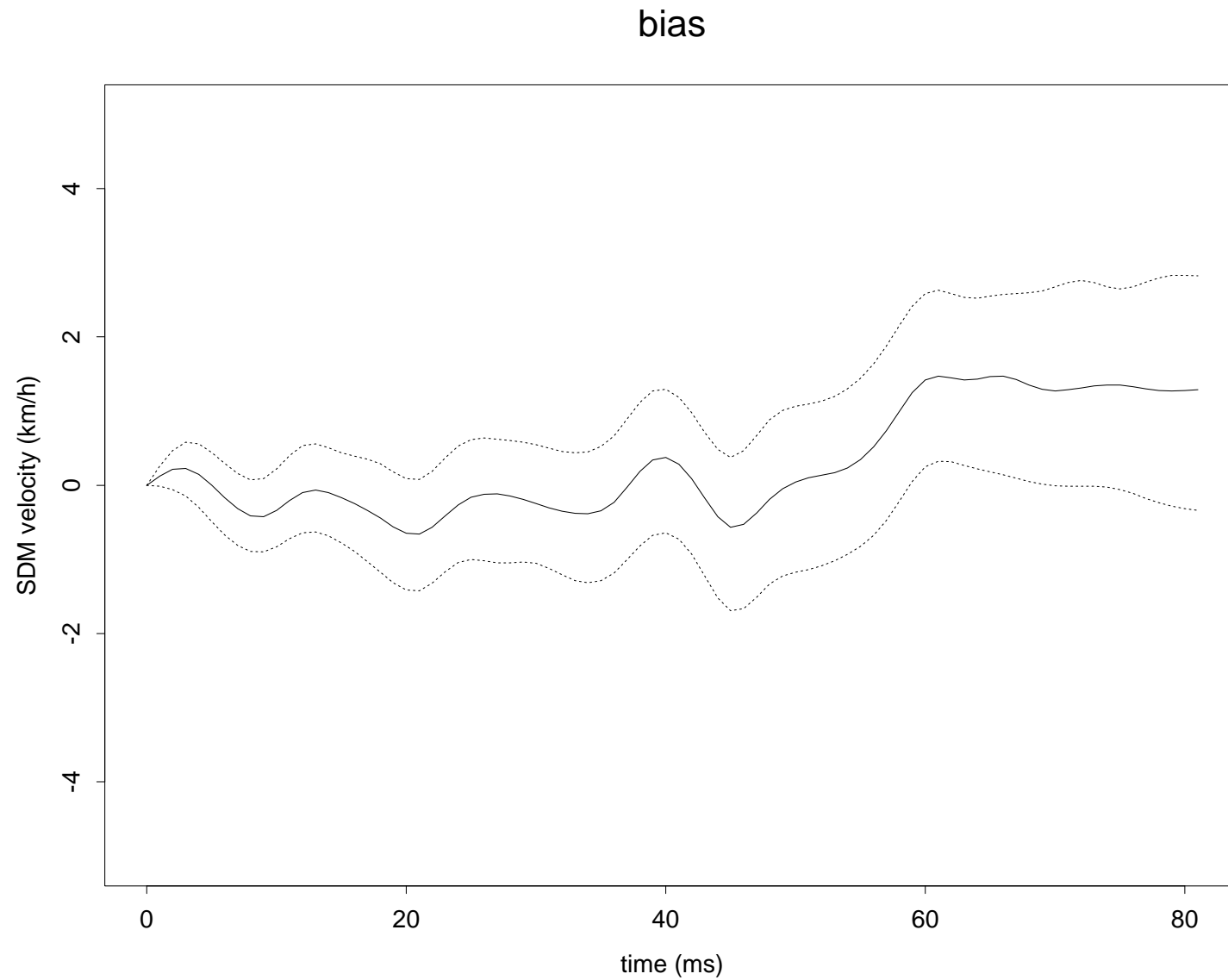


Figure 3: SDM velocity bias at $56.3\text{km}/h$ impact.

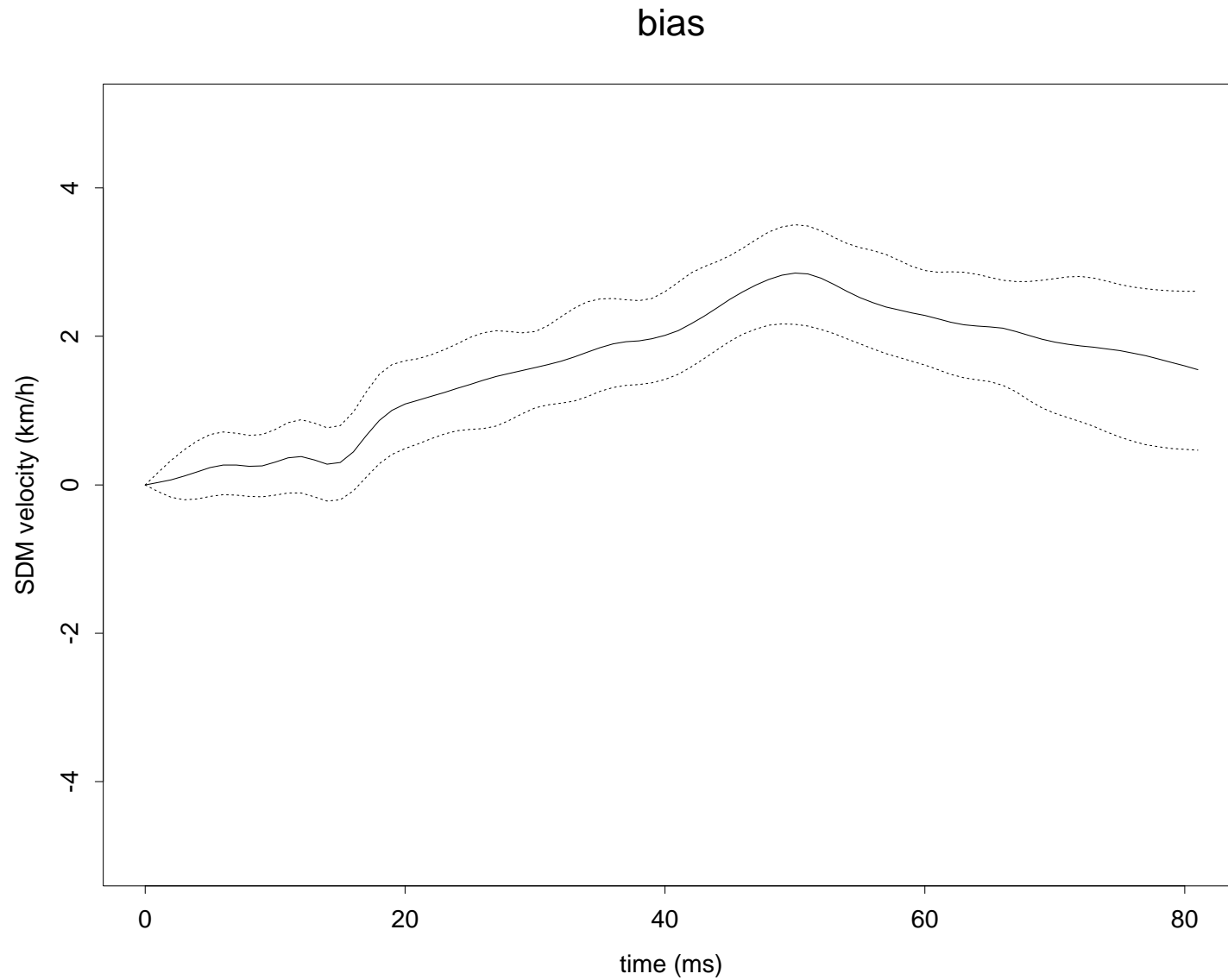
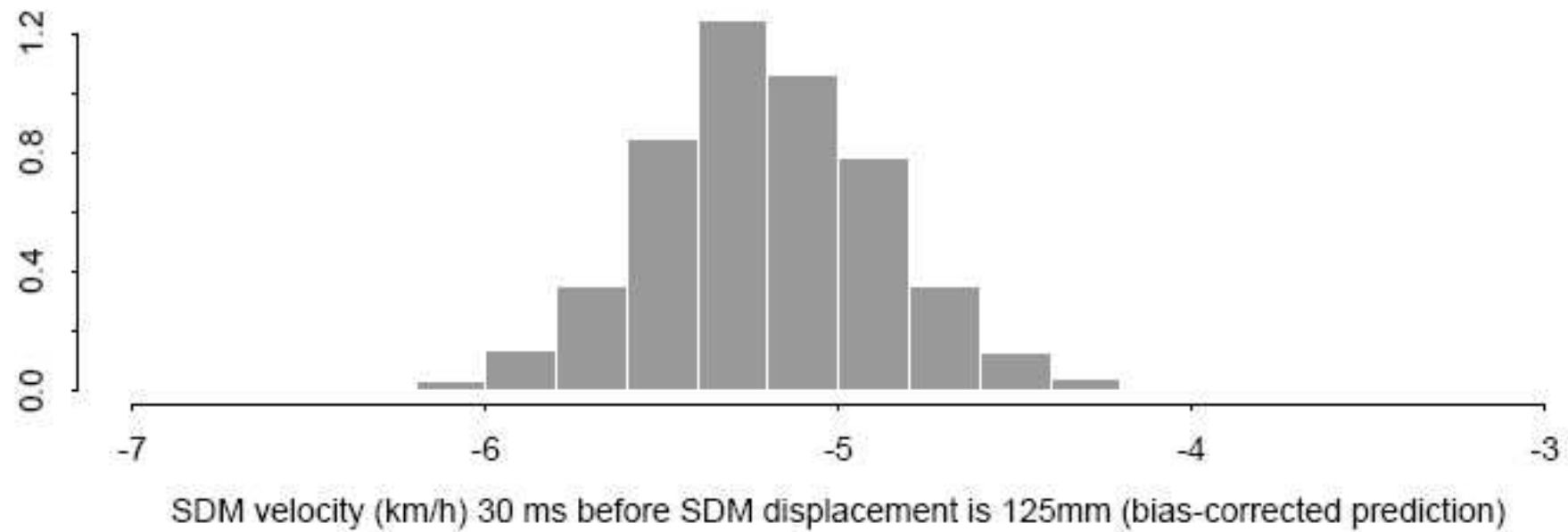


Figure 4: SDM velocity bias at $30\text{km}/h$ impact

Predicting CRITV



'Related' sources of data

- Analysis so far \rightsquigarrow straight frontal, rigid barrier collisions
- There was also (limited) data for right angle, left angle and center pole barrier collisions
- Combine data to borrow information for estimating CRITV
- Severe scarcity of data imposed restrictive assumptions:
 - smoothness of model approximation equal across barriers
 - field and model approximation variances also equal
 - means of the 4 model approximation GASP's come from a two-stage hierarchical model
 - biases for the 4 barriers related through $\log(\lambda_i^b) \sim N(\eta, 4q^2)$ with subjective q . $q = .1$ means that we expect the biases to vary by 10% among the barrier types being considered

PSfrag replacements
 (λ_i^b)
 μ_i^M

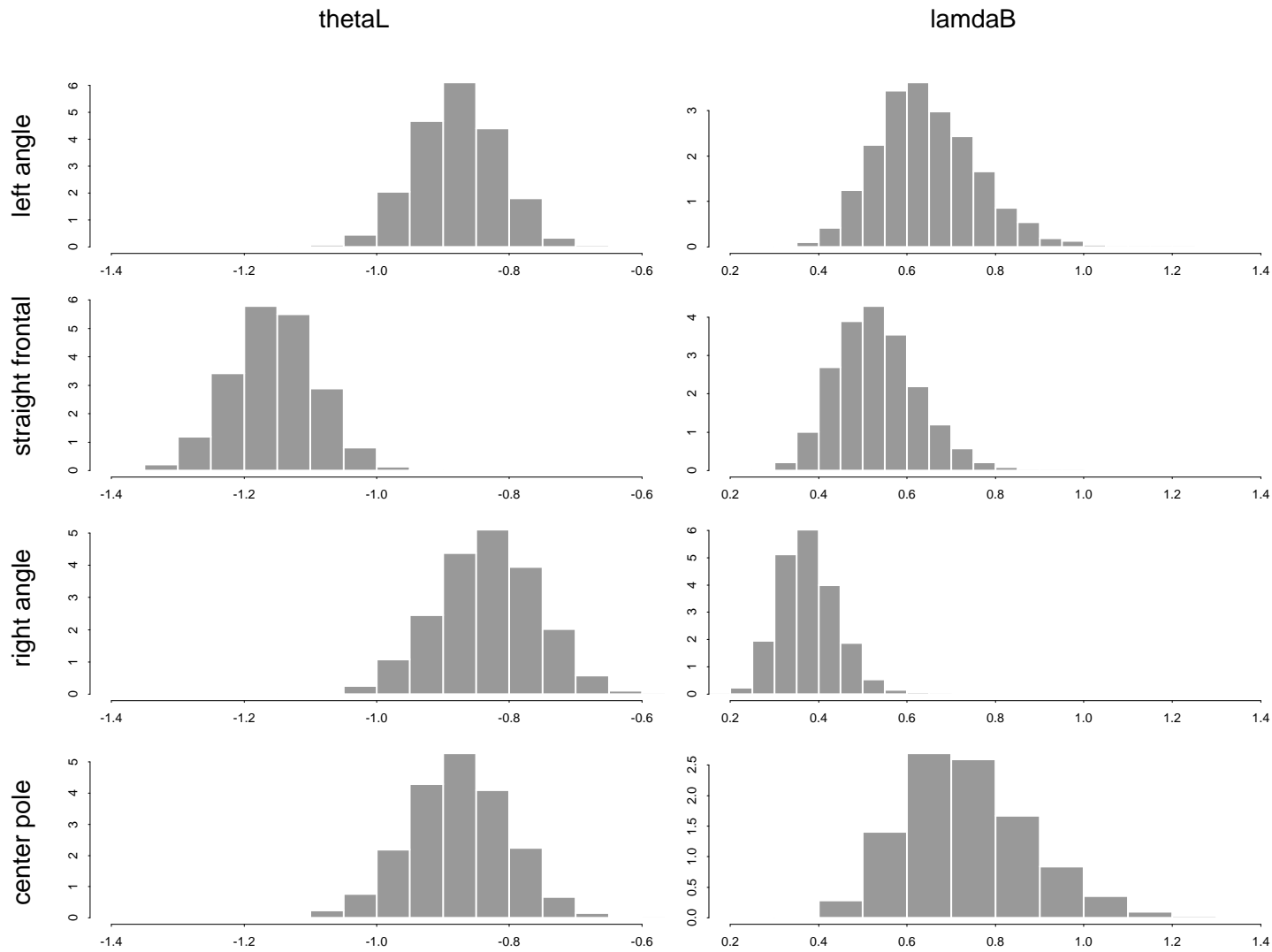


Figure 5: Posterior distributions of μ_i^M and $\log \lambda_i^b$

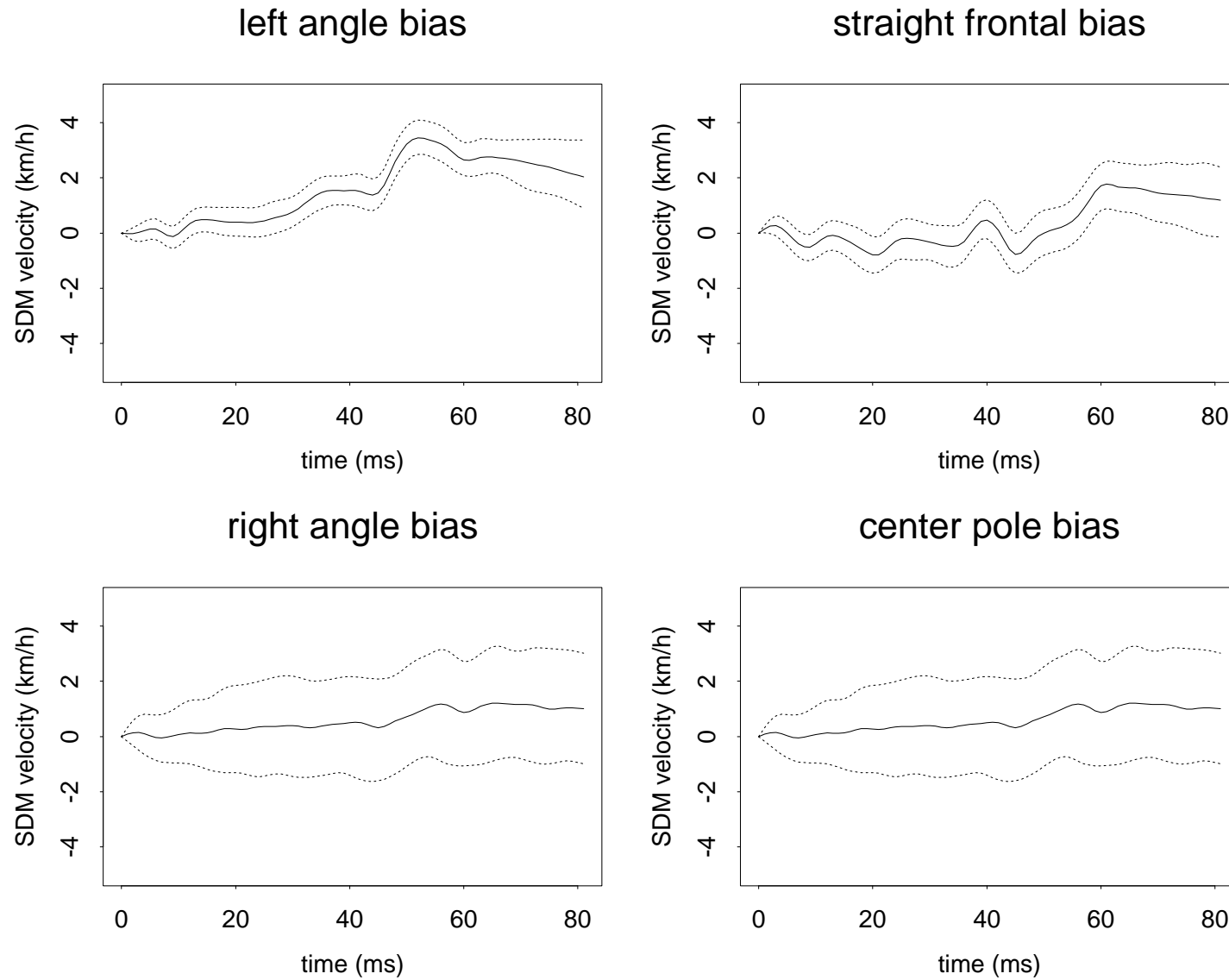


Figure 6: bias at the input velocity 56.3 km/h .

Barrier type	Hierarchical model CRITV	Using frontal data only CRITV
left angle	-6.34 (0.49)	
straight frontal	-5.22 (0.30)	-5.21 (0.33)
right angle	-6.80 (0.96)	
center pole	-6.54 (0.91)	

Table 1: Posterior mean and standard deviation of CRITV, at the input velocity $56.3\text{km}/\text{h}$.

.... and that's all for today

THANKS!!