

The case for a fully robust hierarchical Bayesian statistical analysis of clinical trials

Luis Raúl Pericchi Guerra*, María Eglée Pérez*

Department of de Mathematics
Universidad de Puerto Rico, Río Piedras Campus
Biostatistics and Bioinformatics Core BBC of the
Comprehensive Cancer Center of the Universidad de Puerto
Rico*

O'Bayes09, June 2009

Content

In Fúquene, Cook and Pericchi (2008) <http://www.bepress.com/mdandersonbiostat/paper44> for two clinical trials (past and present), conjugate priors are compared with (Robust) Cauchy- Student-t and what we call Berger's prior. The behavior of the latter Robust Bayesian methods is qualitatively different from Conjugate Bayesian Methods and arguably much more reasonable and acceptable to the practitioner and regulatory agencies.

Extension to multicenter trials

Here we extend the previous work to several Clinical Trials or Hospitals, in a hierarchical fashion. The Hierarchical Modeling with Conjugate Priors and Exponential Family likelihoods, is myopic with respect to large deviations of a single group from the bulk of the groups. This leads to potential excessive (and spurious) shrinkage. On the other hand, Robust Hierarchical Modeling leads to “local shrinkage”; that is, only groups which are consistent affect each other, but groups which are outliers have a smaller influence on the not outlying groups. In that sense, outlying groups are discounted under the Robust Hierarchical Bayesian Model. The methods are illustrated with both simulated and real data.

Extension to multicenter trials

Here we extend the previous work to several Clinical Trials or Hospitals, in a hierarchical fashion. The Hierarchical Modeling with Conjugate Priors and Exponential Family likelihoods, is myopic with respect to large deviations of a single group from the bulk of the groups. This leads to potential excessive (and spurious) shrinkage. On the other hand, Robust Hierarchical Modeling leads to “local shrinkage”; that is, only groups which are consistent affect each other, but groups which are outliers have a smaller influence on the not outlying groups. In that sense, outlying groups are discounted under the Robust Hierarchical Bayesian Model. The methods are illustrated with both simulated and real data.

Extension to multicenter trials

Here we extend the previous work to several Clinical Trials or Hospitals, in a hierarchical fashion. The Hierarchical Modeling with Conjugate Priors and Exponential Family likelihoods, is myopic with respect to large deviations of a single group from the bulk of the groups. This leads to potential excessive (and spurious) shrinkage. On the other hand, Robust Hierarchical Modeling leads to “local shrinkage”; that is, only groups which are consistent affect each other, but groups which are outliers have a smaller influence on the not outlying groups. In that sense, outlying groups are discounted under the Robust Hierarchical Bayesian Model. The methods are illustrated with both simulated and real data.

Particular Findings

- ▶ In order to simplify complex relationships in Hierarchical Models the Theory of Regular Variation and Polynomial Bounded Tails is be applied to check robustness.
- ▶ A Beta Type 2 distribution appears naturally as a distribution of the variances.

Is Robust Bayes also Objective Bayes?

In Clinical Trials there is usually plenty of prior information, and prior data.

Although there is historical information, we consider the methods objective since:

1. prior information is not (entirely) subjective and
2. Robust Priors are discounted for location parameters and partially discounted for scale parameters when there is a conflict with sample information.

What ARE (2) of the types of Robustness in a Bayesian Framework?

Robustness = Bounded Influence

- ▶ **Robustness with respect to the Prior Information:**
Conflict between Likelihood and Prior then Prior is discounted.
- ▶ **Robustness with respect to trials or hospitals:** If a trial is an outlier in conflict with the rest then the group is discounted.

Robustness in Hierarchical Models involve BOTH kinds of robustness.

Relationships between Polynomial Bounded Tails and Functions of Regular Variation

Two lines of recent advances on (parametric) Bayesian Robustness are: Functions of Regular Variation (Andrade and O'Hagan) and Polynomial Bounded Tails (Cook, Fúquene, LRP).

The Polynomial Tails Theorem includes the following condition:
Let $f(\gamma)$ be any likelihood function such that as $|\gamma| \rightarrow \infty$

$$\int_{|\gamma|>m} f(\gamma) d\gamma = \mathcal{O}(m^{-\nu-1-\varepsilon}). \quad (1)$$

We may say that f has *decay of order faster* than: polynomial of κ degree and it is denoted by $f \in \mathcal{P}_\kappa$ with $\kappa = -\nu - 1 - \varepsilon$. In short we may say that f has decay at most κ . Note the likelihood does not have to be location/scale.

Generalized Polynomial Theorem

If the prior

$$p(\lambda|\mu, \tau) = \frac{1}{\tau} \cdot p\left(\frac{\lambda - \mu}{\tau}\right),$$

as $\lambda \rightarrow \infty$ is $\mathcal{O}(\lambda^{\nu-1})$ and its derivative is $\mathcal{O}(\lambda^{\nu-2})$ then as $\mu \rightarrow \infty$

$$\frac{\pi^P(\lambda|Data)}{\pi^U(\lambda|Data)} = 1, \quad (2)$$

where $\pi^U(\lambda|Data)$ is the posterior obtained using an uniform prior.

Regular Variation: from Extreme Value Theory to Bayesian Robustness

On the other hand, the Regular Variation has the following definition:

$$g \in \mathcal{R}_\rho \text{ if and only if } \frac{g(\lambda x)}{g(x)} \rightarrow \lambda^\rho, \text{ as } x \rightarrow \infty, \forall \lambda > 0. \quad (3)$$

Regular Variation is contained in Polynomial-Bounded Tails

$$\mathcal{R}_\rho \text{ is contained in } \mathcal{P}_\kappa \quad (4)$$

The converse is not quite true: take a function of Regular Variation and multiply it by a sine.

So Polynomial-Bounded tails is more general, but Regular Variation is simpler to work with.

Hierarchical Models

1. Observations Level: Likelihood:

$$f(\mathbf{x}|\theta)$$

2. Prior Level (Structure):

$$g(\theta|\gamma)$$

3. Hyper-Prior Level (tuning):

$$h(\gamma)$$

$$\rightarrow \hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k],$$

are “shrunk” closer together promoting admissibility for $k > 2$
 (“solving” Stein-Paradox).

The Models and the Assumptions

Let us focus in the example on page 137 of Bravo, Llatas and Pérez (2008) (taken from Gould 1998 “Multi-center trial analysis revisited” *Statist. Medicine*). Drug: Finasteride: treatment of benign prostatic hypertrophy: 25 centers in USA and 5 in Canada. 900 patients. Response: Changes on a Score on each patient; Range from 0 to 36. Dose: 5mg/day, and a Placebo. Final Response: Difference of mean score between treatments and controls per center.

Center	Placebo			Finasteride 5 mg		
	N	Mean	St. Dev.	N	Mean	St. Dev.
1	7	0.43	4.58	8	-2.63	3.38
2	11	0.10	4.21	12	-2.21	4.14
3	6	2.58	4.80	7	1.29	7.39
4	10	-2.30	3.86	10	-1.40	2.27
5	10	2.08	6.46	10	-5.13 (-15.00)	3.91
6	6	1.13	3.24	5	-1.59	3.19
7	5	1.20	7.85	5	-1.40	2.61
8	12	-1.21	2.66	12	-4.08	6.32
9	8	1.13	5.28	9	-1.96	5.84
10	9	-0.11	3.62	10	0.60	3.53
11	15	-4.37	6.12	15	-2.14	4.27
12	8	-1.06	5.27	9	-2.03	5.76
13	12	-0.08	3.32	11	-6.22	5.33
14	9	0.00	5.20	7	-3.29	5.12
15	6	1.83	5.85	6	-1.00	2.61
16	14	-4.21	7.53	12	-5.75	5.63
17	13	0.76	3.82	13	-0.63	5.41
18	15	-1.05	4.54	14	-2.80	2.89
19	15	2.07	4.88	15	-3.43	4.71
20	11	-1.46	5.48	10	-6.77	5.19
21	5	0.80	4.21	5	-0.23	4.14
22	11	-2.92	5.42	11	-4.45	6.65
23	9	-3.37	4.73	7	0.57	2.70
24	12	-1.92	2.91	12	-2.39	2.27
25	9	-3.89	4.76	8	-1.23	4.91
26	15	-3.48	5.98	14	-3.71	5.30
27	11	-1.91	6.49	11	-1.52	4.68
28	10	-2.66	3.80	10	-4.70	3.43
29	13	-0.77	4.73	13	-0.47	4.95

Only Sufficient Statistics are available

Define $d_i = \bar{x}_{fi} - \bar{x}_{pi}$, $SS_i = (n_{fi} - 1)s_{fi}^2 + (n_{pi} - 1)s_{pi}^2$ the summary quantities for the data.

First we present a common (conjugate or conditionally conjugate) modeling and then 2 robustifications (“beautifications”).

Model I

► Exponential Family Likelihood (Non-Robust)

$$d_i | \delta_i, \sigma_{Wi}^2 \sim N \left(\delta_i, \sigma_{Wi}^2 \left(\frac{1}{n_{fi}} + \frac{1}{n_{pi}} \right) \right), i = 1, \dots, k$$

$$SS_i | \sigma_{Wi}^2 \sim \text{Gamma} \left(\frac{n_{fi} + n_{pi} - 2}{2}, \frac{1}{2\sigma_{Wi}^2} \right), i = 1, \dots, k$$

► Conjugate Prior and Hyper-Prior: Non Robust

$$\delta_i | \Delta, \sigma_B^2 \sim \text{Normal}(\Delta, \sigma_B^2), i = 1, \dots, k$$

$$\sigma_{Wi}^2 \sim \text{InvGamma}(0.01, 0.01), i = 1, \dots, k$$

$$\Delta \sim \text{Normal}(0, 10^5)$$

$$\sigma_B^2 \sim \text{InvGamma}(3, 50)$$

First attempt to Robustness

$$\begin{aligned}d_i|\sigma &\sim St_{\nu_1}(\delta_i, \sigma^2(1/n_1 + 1/n_2)), i = 1, \dots, k \\ \delta_i|\sigma &\sim t_{\nu_2}(0, \sigma^2), \\ \sigma &\sim IGamma(a, b),\end{aligned}$$

with the following tentative assignment: $a = 3$, $b = 50$, $\nu_1 = 7$, $\nu_2 = 9$.

The assessment allows the posterior to reject the prior of σ in favor of the data and the prior $\delta_i|\sigma$. (Andrade and O'Hagan 2006 and 2009). Different combinations of hyper-parameters (to be developed) change the speed and direction of rejection of conflicting information.

A Standard Result: Student as a scale mixture of Normal

We may “*robustify*” the normal by taking a scale mixture and integrating:

$$\text{Student}_\nu(\theta|\mu, \sigma^2) = \int \text{Normal}\left(\theta|\mu, \frac{\sigma^2}{\rho}\right) \cdot \text{Gamma}\left(\rho|\frac{\nu}{2}, \frac{\nu}{2}\right) d\rho.$$

Model II

► Likelihood

$$d_i | \delta_i, \sigma_B^2, \rho_i \sim \text{Normal} \left(\delta_i, \frac{\sigma_B^2}{\rho_i} \left(\frac{1}{n_{fi}} + \frac{1}{n_{pi}} \right) \right), i = 1, \dots, k$$

$$SS_i | \sigma_B^2, \rho_i \sim \text{Gamma} \left(\frac{n_{fi} + n_{pi} - 2}{2}, \frac{1}{2} \frac{\rho_i}{\sigma_B^2} \right), i = 1, \dots, k$$

► Prior

$$\begin{aligned} \rho_i &\sim \Gamma(3.5, 3.5), i = 1, \dots, k \\ \delta_i | \Delta, \sigma_B^2 &\sim t(\Delta, \sigma_B^2, 9), i = 1, \dots, k \\ \Delta &\sim t(0, 10^5, 5) \\ \sigma_B^2 &\sim \text{InvGamma}(3, 50) \end{aligned}$$

Model III

► Likelihood

$$d_i | \delta_i, \sigma_{Wi}^2, \rho \sim \text{Normal} \left(\delta_i, \frac{\sigma_{Wi}^2}{\rho} \left(\frac{1}{n_{fi}} + \frac{1}{n_{pi}} \right) \right), i = 1, \dots, k$$

$$SS_i | \sigma_{Wi}^2, \rho \sim \text{Gamma} \left(\frac{n_{fi} + n_{pi} - 2}{2}, \frac{1}{2} \frac{\rho}{\sigma_{Wi}^2} \right), i = 1, \dots, k$$

► Prior

$$\delta_i | \Delta, \sigma_B^2 \sim t(\Delta, \sigma_B^2, 9), i = 1, \dots, k$$

$$\sigma_{Wi}^2 \sim \text{InvGamma}(0.01, 0.01), i = 1, \dots, k$$

$$\Delta \sim t(0, 10^5, 5)$$

$$\rho \sim \Gamma(3.5, 3.5)$$

$$\sigma_B^2 \sim \text{InvGamma}(3, 50)$$

A New Heavy tailed Distribution for Deviations? Beta of the Second Kind!

$$SS_i | \sigma_B^2, \rho_i \sim \text{Gamma} \left(\frac{n_{fi} + n_{pi} - 2}{2}, \frac{1}{2} \frac{\rho_i}{\sigma_B^2} \right), i = 1, \dots, k$$

$$\rho_i \sim \Gamma(3.5, 3.5), i = 1, \dots, k$$

Implies

$$SS_i | \sigma_B^2 \propto \frac{(SS_i / \sigma_B^2)^{(n_{fi} + n_{pi} - 4)/2}}{(SS_i / \sigma_B^2 + 7)^{(n_{fi} + n_{pi} + 5)/2}}$$

Beta Distribution of the Second Kind

Call $m = \frac{nf_i + np_i - 2}{2}$ and $n = 7/2$ Now since

$$\int_0^{\infty} \frac{x^{(m-1)}}{(a + bx)^{(m+n)}} = \frac{\Gamma(m)\Gamma(n)}{a^n b^m \Gamma(m+n)}, \Rightarrow$$

$$SS_i | \sigma_B^2 \sim \frac{7^n \Gamma(m+n)}{(\sigma_B^2)^m \Gamma(m) \Gamma(n)} \frac{SS_i^{(m-1)}}{(7 + SS_i / \sigma_B^2)^{(m+n)}}.$$

The distribution of the odds in a Beta distribution, flatter tails than a Gamma.

$$E(SS_i) = 7\sigma_B^2 m / (n - 1), \text{ for } n > 1,$$

$$\text{Var}(SS_i) = 7^2 \sigma_B^4 m / ((n - 1)(n - 2)), \text{ for } n > 2.$$

Beta Distribution of the Second Kind

$$p(y|p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \cdot \frac{y^{p-1}}{(1+y)^{(p+q)}}, y > 0,$$

denoted by $y \sim \text{Beta}_2(p, q)$. The result is:

$$\frac{\sigma_B^2}{bL} \sim \text{Beta}_2(a, L/2)$$

Generating $y \sim \text{Beta}_2(p, q)$:

$$z \sim \text{Beta}(p, q),$$

$$y = \frac{z}{1-z}$$

Beta of the Second Kind is of regular variation index

$$\rho = -(q + 1)$$

Definition: Regular Variation Index ρ :

$$\frac{p(\lambda y)}{p(y)} \rightarrow \lambda^\rho, (y \rightarrow \infty), \forall \lambda > 0.$$

$$\frac{Beta_2(\lambda y | p, q)}{Beta_2(y | p, q)} \rightarrow \lambda^{-(q+1)},$$

which for $q > 0$ is a proper prior and is of the same order than a Student-t density with q degrees of freedom.

NOTE: Since Beta of the Second Kind is of Regular Variation \rightarrow Polynomial Tails.

Summary of Results for Regular Variation: Location-Scale Structures, Andrade and O'Hagan 2009

$$\begin{aligned}
 y_n | \mathbf{y}^{(n-1)}, \mu, \sigma &\sim f(y_n | \mu, \theta) = 1/\sigma \cdot h_n[(y_n - \mu)/\theta] \\
 \mu | \mathbf{y}^{(n-1)}, \sigma &\sim p(\mu | \mathbf{y}^{(n-1)}, \sigma) = 1/\sigma \cdot V[\mu/\sigma | \mathbf{y}^{(n-1)}] \\
 \sigma | \mathbf{y}^{(n-1)} &\sim p^*(\sigma | \mathbf{y}^{(n-1)})
 \end{aligned}$$

where all are regularly varying functions with index:

$$V \in \mathcal{R}(-(\alpha + \sum_{k=1}^{n-1} \rho_k)) \text{ and } p^* \in \mathcal{R}(-(\alpha + n - 1)).$$

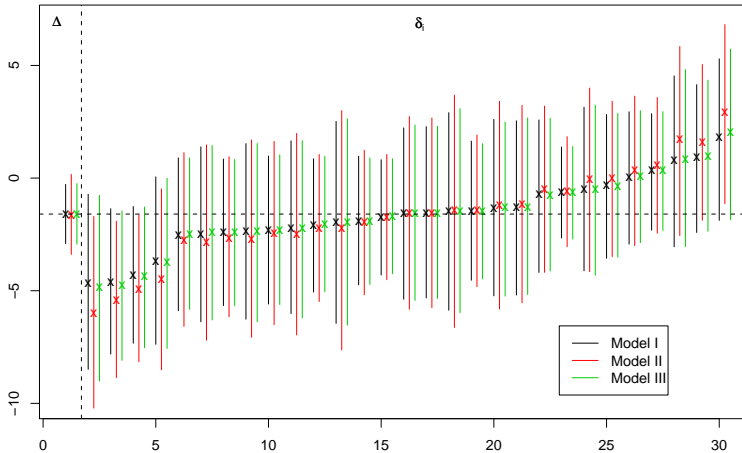
Results:

- 1) As $y_n \rightarrow \infty$ y_n is discarded.
- 2) If the prior location goes to a boundary, it is discarded.
- 3) If prior scale goes to a boundary it is *partially* discarded.

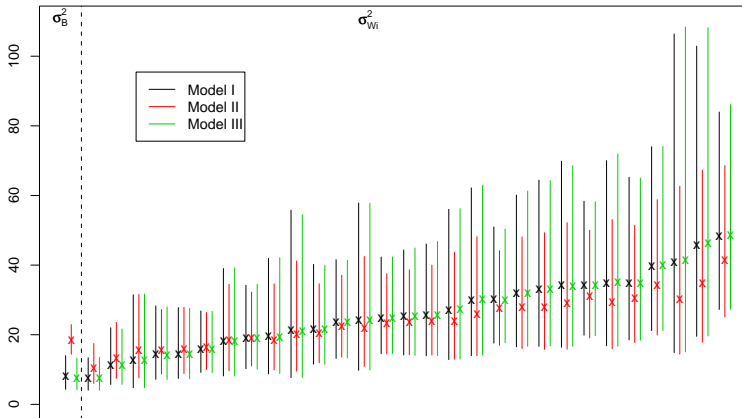
Its not only tail that matters, the origin also matters!

- ▶ Inverted Gamma is of Regular Variation. But, Inverted Gamma(ϵ, ϵ) has a dangerously high probability near the origin!
- ▶ Beta₂(1, 1) is of the form recommended in Scott and Berger
$$p(\sigma_B^2) \propto \frac{1}{(c + \sigma_B^2)^2}$$
- ▶ When $p, q \rightarrow 0$, Beta₂(p, q) converges to the usual non-informative prior for the variance: $p(\sigma_B^2) \propto \frac{1}{\sigma_B^2}$

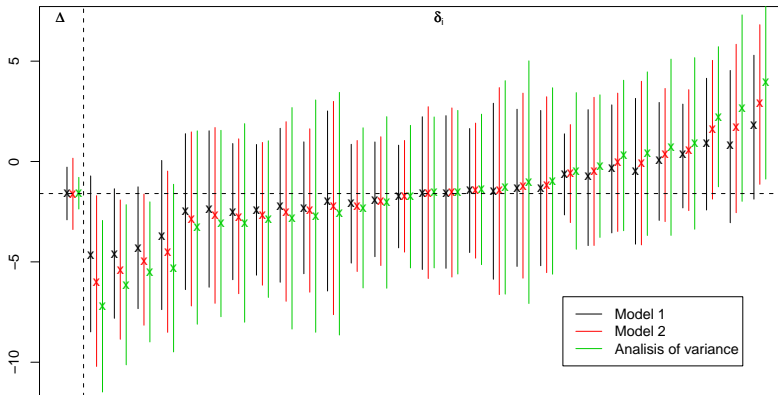
Location parameters, multicenter clinical trial, models I, II and III

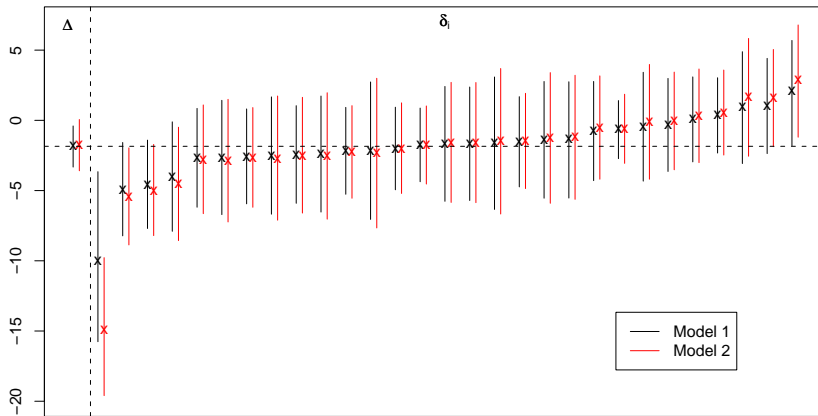


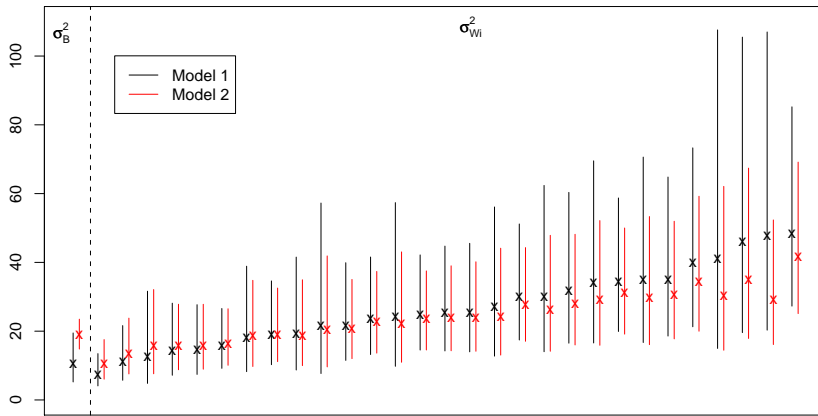
Dispersion parameters, multicenter clinical trial, models I, II and III



Location parameters, multicenter clinical trial, models 1 and 2
and analysis of variance model



**Location parameters, multicenter clinical trial, models 1 and 2,
modified data**

Dispersion parameters, models 1 and 2,
modified data

A Tale of Unwanted Attraction: Or how Hierarchical Bayes pulls up perfect hospitals

One of the uses of Bayesian Hierarchical models is in hospital profiling: Normand and Shahian (2007) “Statistical and clinical aspects of hospital outcomes profiling” *Statistical Science*. Hospitals with no deaths experience large shifts with non-robust hierarchical models. We revisit the data but here do not perform exactly the same analysis as the authors: 1) we do not use the explanatory data \mathbf{x} since it is not available and 2) we enlarge the sample size of one of the hospitals with no deaths to the average size of all hospitals, in order to see more clearly the difference between analyses.

We focus on the log-odds and on the probability of death θ_i .
The interesting feature here is: The outliers are small numbers
(zero or close to zero) rather than large numbers.

Non-Robust Model I

$$\begin{aligned}y_i &\sim \text{Bin}(n_i, \theta_i) \\ \log(\theta_i / (1 - \theta_i)) = \beta_{0i} &\sim N(\mu, \sigma^2) \\ \mu &\sim N(0, 10^3) \\ \sigma^2 &\sim \text{IGamma}(0.001, 0.001)\end{aligned}$$

For MODEL II: the only change is:

$$\sigma^2 \sim \text{Beta}_2(1, 1)$$

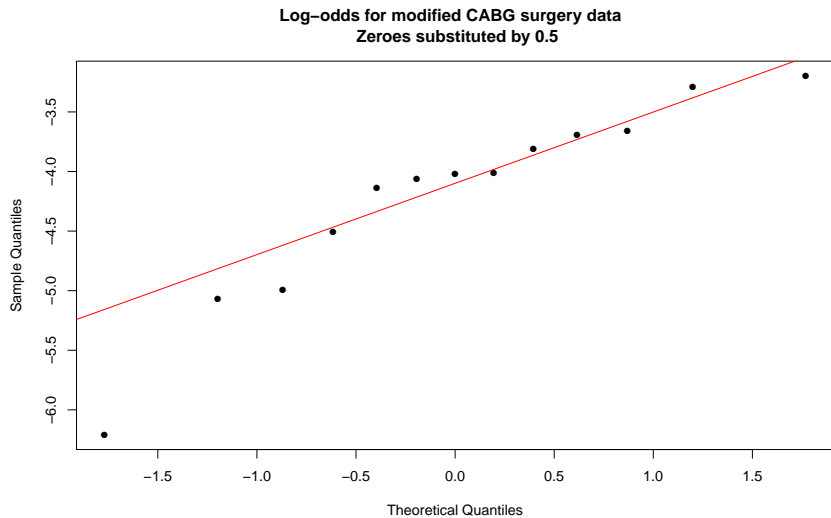
A Robustified Model

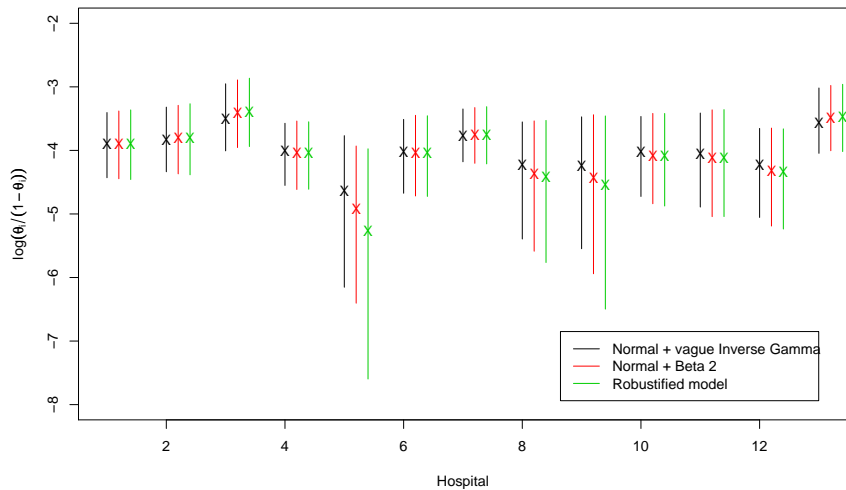
$$\begin{aligned}y_i &\sim \text{Bin}(n_i, \theta_i) \\ \log(\theta_i/(1 - \theta_i)) = \beta_{0i} &\sim N(\mu, \sigma_B^2/\rho_i) \\ \rho_i &\sim \text{Ga}(3.5, 3.5) \\ \mu_i &\sim t_4(M, \sigma_B^2) \\ M &\sim t_2(0, 10^6) \\ \sigma^2 &\sim \text{Beta}_2(1, 1)\end{aligned}$$

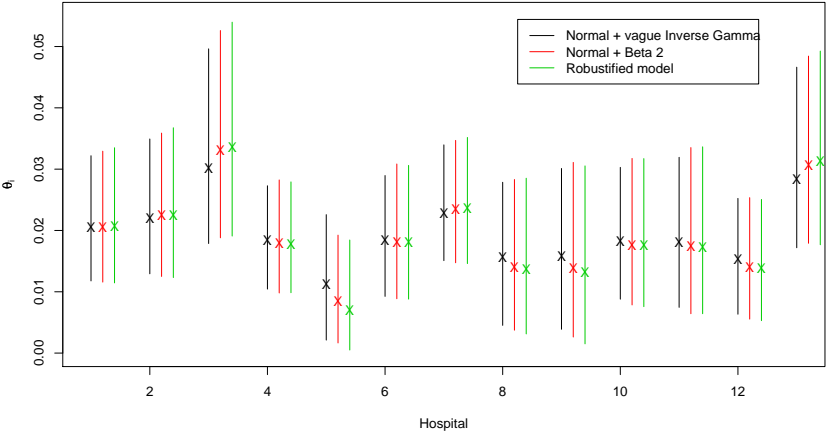
We focus on the log-odds β_{0i} and the probability of death θ_i .

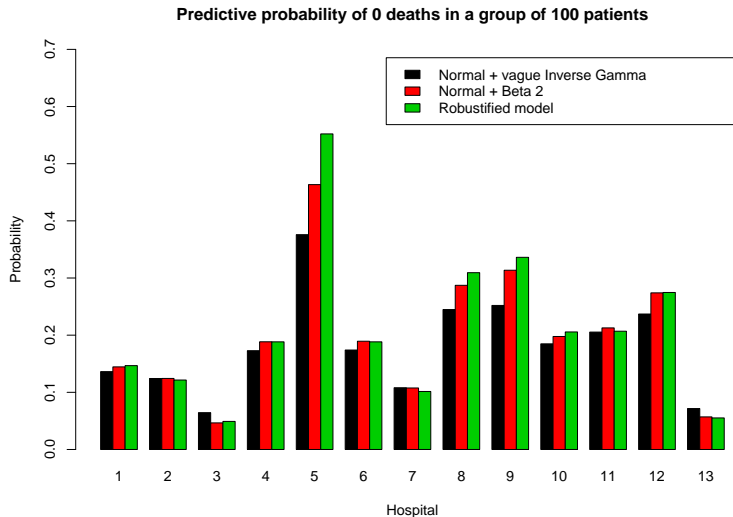
Hospital	Patients	Deaths
1	508	11
2	454	11
3	381	15
4	623	11
5	26 (250)	0
6	393	7
7	718	18
8	149	1
9	80	0
10	296	5
11	191	3
12	365	4
13	419	15

Table: 30-day mortality in 13 nongovernmental hospitals following isolated CABG surgery, Massachusetts, USA (Normand and Shahian, 2007) Hospital 5 was changed from 26 patients to the approximate mean number of patients 250









First Conclusions

- ▶ Regular (Polynomial) Tails in the Likelihood and in the priors allows simultaneous discounting of prior information and of excessive shrinkage in case of conflict (“**unwanted attraction**”)
- ▶ Student for location and Beta of the Second Kind for scale, seem to be a minimal kit to solve “unwanted attraction” and also “unwanted memories”. Both appear naturally as a scale mixtures of Normal Sampling Distributions.