Functional uniform prior distributions for nonlinear regression

Björn Bornkamp (Novartis Pharma AG)

1. Motivation
   - Bayesian optimal design
   - IBS dose-response study
2. Functional Uniform Prior Distributions
3. Application
   - IBS dose-response study
   - Simulations
   - Bayesian optimal design
4. Summary

Motivation

- Fisher information matrix $I(d, \boldsymbol{\theta})$ in non-linear models depends on unknown parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (here d is the experimental design)
- Typically "design criterion" $\Phi(I(d, \boldsymbol{\theta}))$ used (*e.g.* for $\Phi(I(d, \boldsymbol{\theta})) = \log(\det(I(d, \boldsymbol{\theta})))$)
- Can optimize $\Phi(I(d, \boldsymbol{\theta}))$ with respect to d only for a given $\boldsymbol{\theta}$ $\rightarrow$ calculated design only "locally" optimal
- In practice need to deal with the uncertainty $\boldsymbol{\theta}$!

  One approach: Use (prior) probability distribution $\pi(\boldsymbol{\theta})$ and optimize

$$\int \Phi(I(d, \boldsymbol{\theta}))\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Fine in case there is a properly elicited prior (*e.g.* as in Peter Thall's talk yesterday) or a prior based on historical data. But what to do if this is not available?

- Traditional non-informative priors (Jeffreys, reference) cannot be used $\rightarrow$ depend on d, which is what we want to calculate!
- Often a reasonable compact subset $K \in \boldsymbol{\Theta}$ of the parameter space is selected and then a uniform prior is used
  $\rightarrow$ Approach crucially depends on chosen parameterization and selected bounds
  $\rightarrow$ Approach often more "informative" than intuitively expected
- Unclear what could be a "reasonable" prior

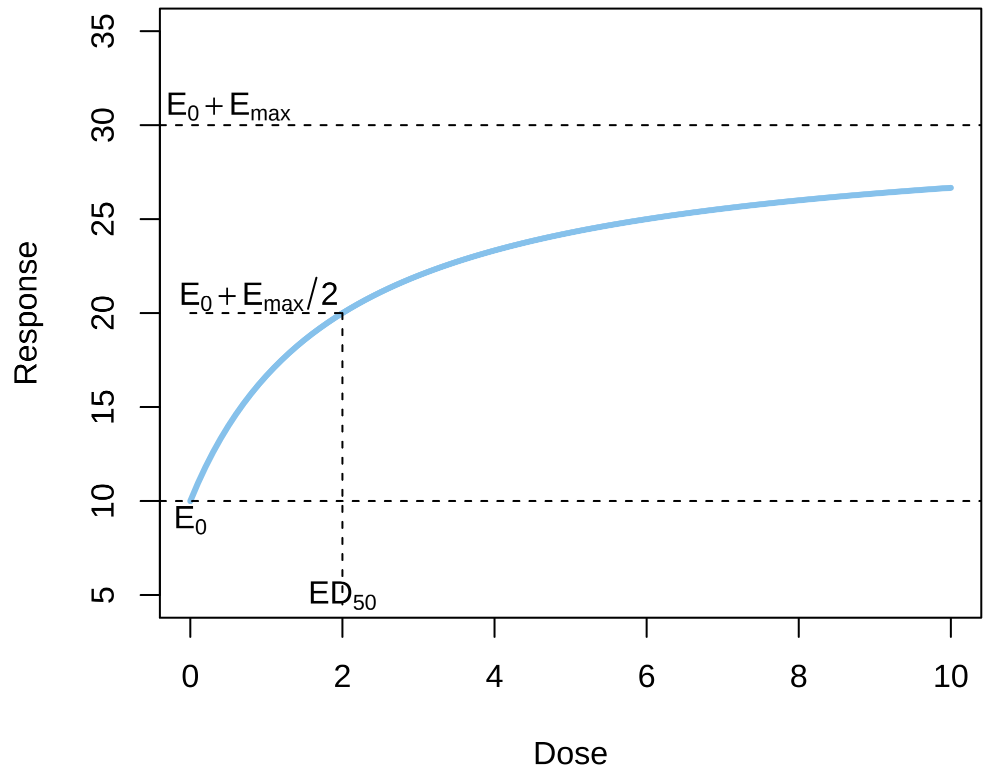
- Data from a dose-response trial for treatment of the irritable bowel syndrome (`IBScovars` data set from `DoseFinding` package)
- In total 369 patients completed the trial with roughly balanced allocations on the 5 treatment arms (placebo and doses 1, 2, 3, 4)
- Endpoint change from baseline in pain score (larger value corresponds to less pain)
- Assume E_{max} model to analyze data, *i.e.*

$$y_{ij} = \mu(x_{ij}, \boldsymbol{\theta}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

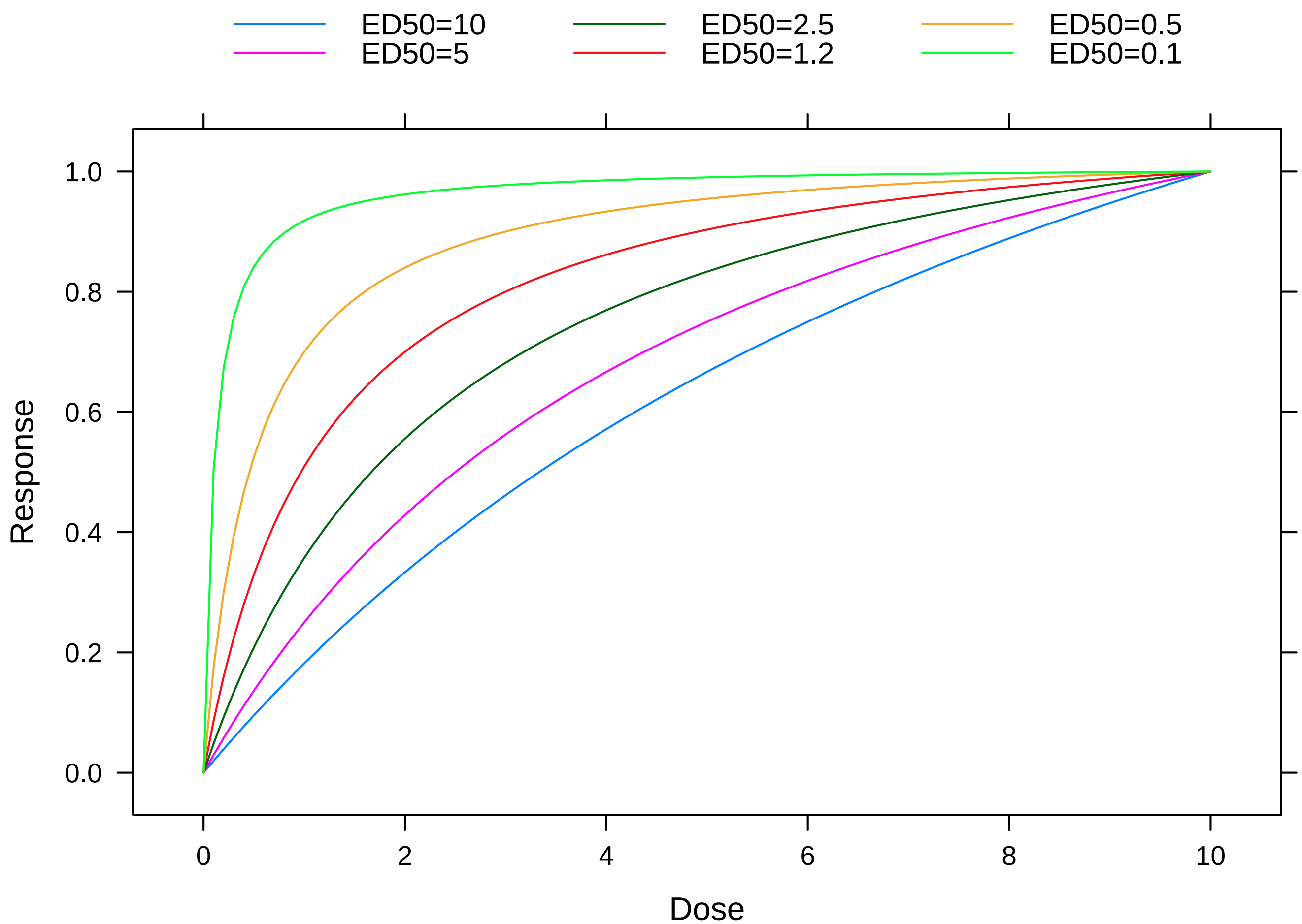
$\mu(x, \boldsymbol{\theta}) = E_0 + E_{\max} x / (ED_{50} + x)$

E_0: placebo effect, $\mathrm{E}_{\max}$: max effect change, ED_{50}: dose at half of max. change

Example Plot: $E_0 = 10, E_{max} = 20, ED_{50} = 2$

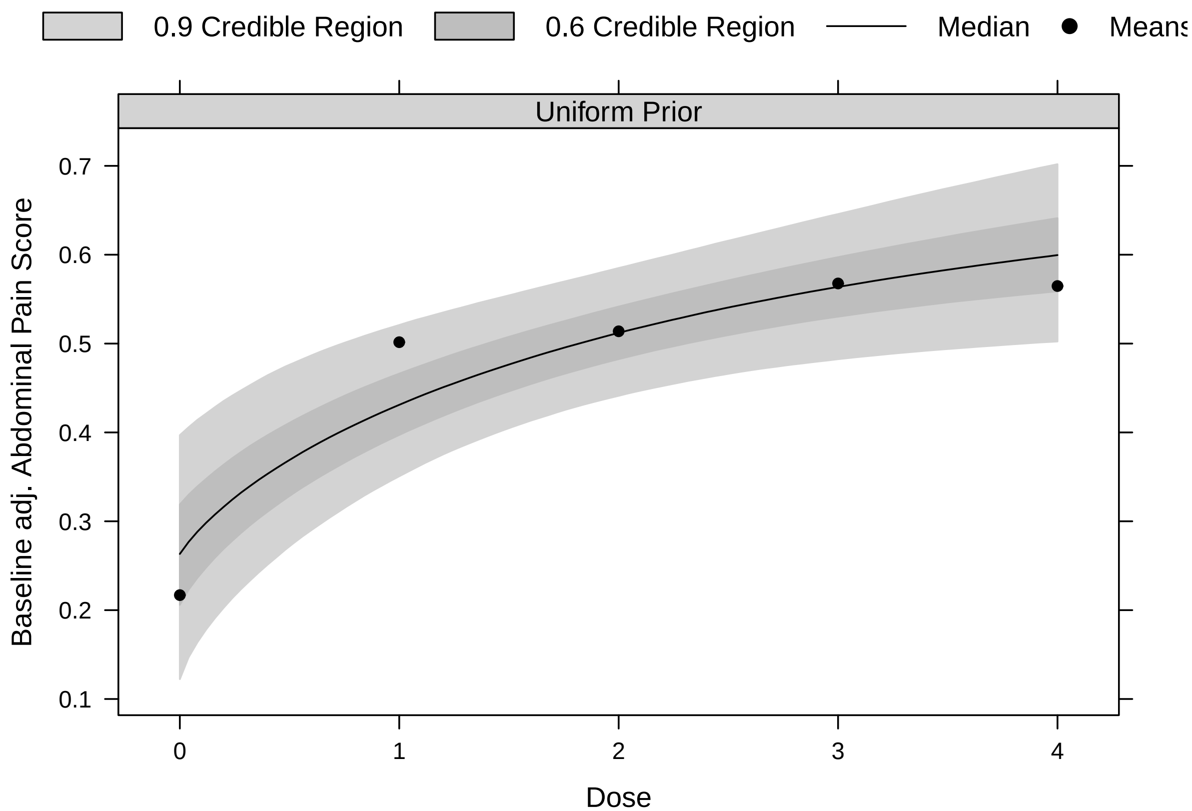


Potential shapes of Emax models*



* setting the E_0 and $E_{\max}$ parameters so that $\mu(0, \boldsymbol{\theta}) = 0$ and $\mu(10, \boldsymbol{\theta}) = 1$

Assume flat priors for E_0 and $E_{\max}$, for nonlinear parameter ED_{50} use uniform prior on $[0, 6]$ (maximum dose is 4)



- When increasing the upper bound
  $\longrightarrow$ even more shrinkage towards linear shape
- Just to re-emphasize: Uniform priors can be more informative than expected (for nonlinear parameters and particularly for low SNR)
- What to use if not uniform priors?
- Jeffreys prior is an option, but methodological problems in adaptive trials (sequential analysis differs from analysis combining all data)

In summary

- Neither uniform priors on the parameter scale nor Jeffreys prior provide a good approach to the problems outlined above
- Either depend strongly on used parameterization and bounds or cannot be stated prior to data collection (needed for calculating design)
- There is a need for a “reasonable”, parameterization invariant prior that can be formulated before data collection!
- Of course, one can find “good” weakly informative priors from experience/experimentation for each particular model, this is an attempt at a more systematic approach

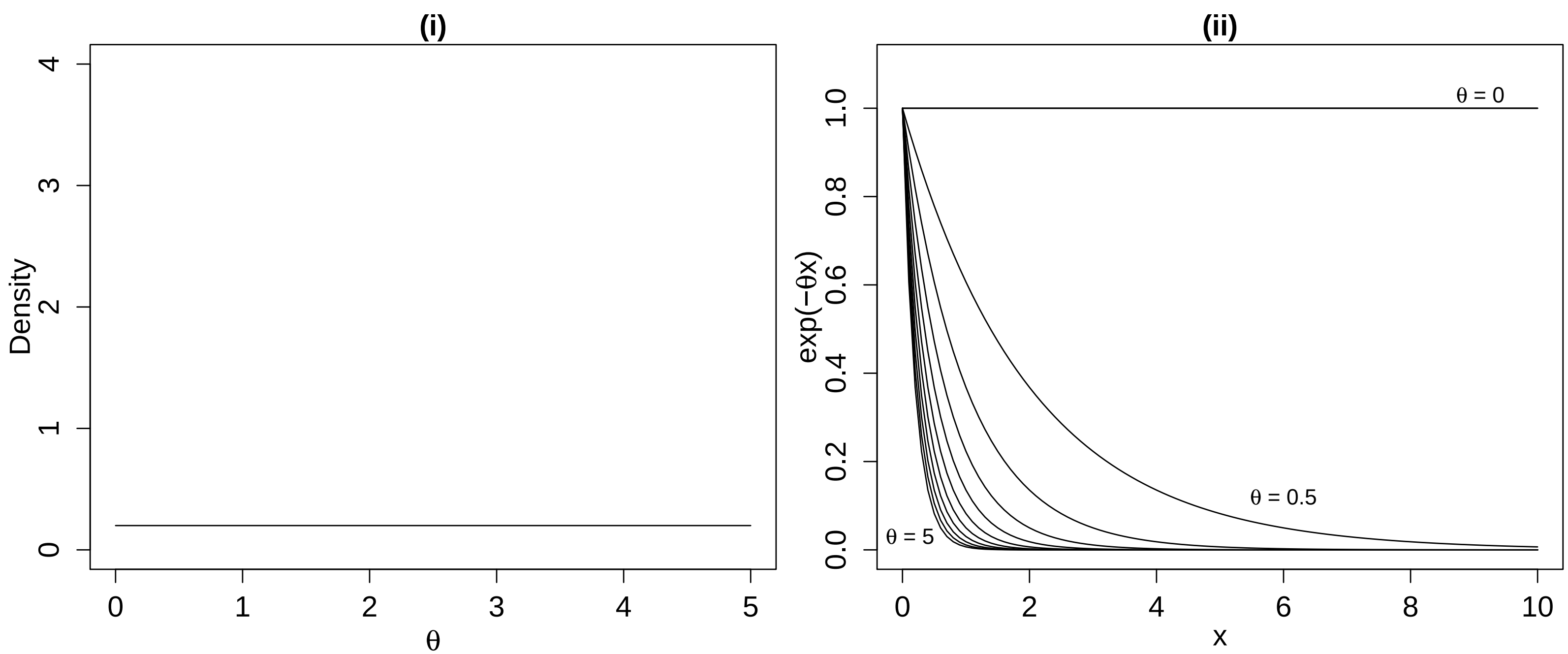
Functional Uniform Priors

Functional Uniform priors: Exponential Regression

- Illustration using the exponential regression model

$$\mu(x, \theta) = \exp(-\theta x), \quad \theta \in [0, 5], x \in [0, 10]$$

- Uniform prior on $[0, 5]$ induces strong prior information towards quickly decreasing shapes, when increasing upper bound even more!



- Essential idea:

  Try to find distribution that is uniform in function space induced by $\exp(-\theta x), \theta \in \boldsymbol{\Theta}$

- Then try to induce a uniform distribution in this space

- This prior is this "uninformative" or "neutral" with regard to the function space, rather than the parameter space

- This will not depend on the chosen parameterization

- Suppose you want to find a distribution for a parameter $\boldsymbol{\theta}$,
- Use bijective function $\phi(\boldsymbol{\theta})$ to map $\boldsymbol{\theta}$ into a metric space (M, d), that defines an "adequate" notion of distance between the parameters
- In this metric space then derive a uniform distribution and transform this distribution back to $\boldsymbol{\theta}$ scale
- In exponential regression example:
  M is space of functions $\exp(-\theta x)$ and distance between θ' and θ'' is then a distance between functions $\exp(-\theta' x)$ and $\exp(-\theta'' x)$

Based on work by Dembski (1990) and Ghosal et al. (1997)

- Dembski (1990) discusses uniform probability on general compact metric spaces (M, d) using the notion of $\epsilon-$lattice, *i.e.* a set $S_\epsilon \subset M$, so that for all $\phi', \phi'' \in S_\epsilon$ holds $d(\phi', \phi'') \geq \epsilon$, and the addition of any point to S_ϵ destroys this property. In addition it needs to have maximum possible cardinality.
- Dembski (1990) defines uniform probability as the limit of a discrete uniform distribution on S_ϵ for $\epsilon \to 0$
- Essentially this is an "equally spaced" grid, where the notion of "equally spaced" depends on (M, d)

- In the exponential regression example choose

$$d(\theta, \theta') = \sqrt{\int (\exp(-x\theta) - \exp(-x\theta'))^2 dx}$$

- Then choose points S_ϵ so that $d(\theta, \theta') \geq \epsilon$ with maximum possible cardinality
- The limit of the discrete uniform distribution on S_ϵ for $\epsilon \to 0$ results in the desired distribution
- Computationally hard to calculate in dimensions > 1

In the finite dimensional case, when one can approximate d in terms of $\boldsymbol{\theta}$

$$d^*(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = c_1 \sqrt{c_2 (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{V}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^k)},$$

for $k \geq 3$ and positive definite $\boldsymbol{V}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}$.

The uniform distribution on M is then given by (in terms on $\boldsymbol{\theta}$)

$$p(\boldsymbol{\theta}) = \frac{\sqrt{\det(\boldsymbol{V}(\boldsymbol{\theta}))}}{\int_{\boldsymbol{\Theta}} \sqrt{\det(\boldsymbol{V}(\boldsymbol{\theta}))} d\boldsymbol{\theta}}.$$

For details see Dembski (1990), Ghosal et al. (1997).

For a different approach, see for example Pennec (2006)

Functional Uniform priors: Examples and special cases

1. When $\phi : \mathbb{R}^p \mapsto \mathbb{R}^p$ and using the Euclidean metric in the transformed space, this is the change of variables theorem

2. When M is space of residual densities and d the empirical Hellinger distance or Kullback-Leibler divergence one obtains Jeffreys prior $\rightarrow$ reasonable "default" prior (uniform prior on the space of residual densities, but conditional on covariates)

Functional Uniform priors: Nonlinear Regression

- Use, *e.g.*, L_2 distance in nonlinear regression

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sqrt{\int_{\mathcal{X}} (\mu(x, \boldsymbol{\theta}) - \mu(x, \boldsymbol{\theta}'))^2 dx}$$

- Now $(\mu(x, \boldsymbol{\theta}) - \mu(x, \boldsymbol{\theta}_0))^2 =$ $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T J_x(\boldsymbol{\theta}_0)^T J_x(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^3)$, where $J_x(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\theta}} \mu(x, \boldsymbol{\theta})$

- Hence finally one obtains

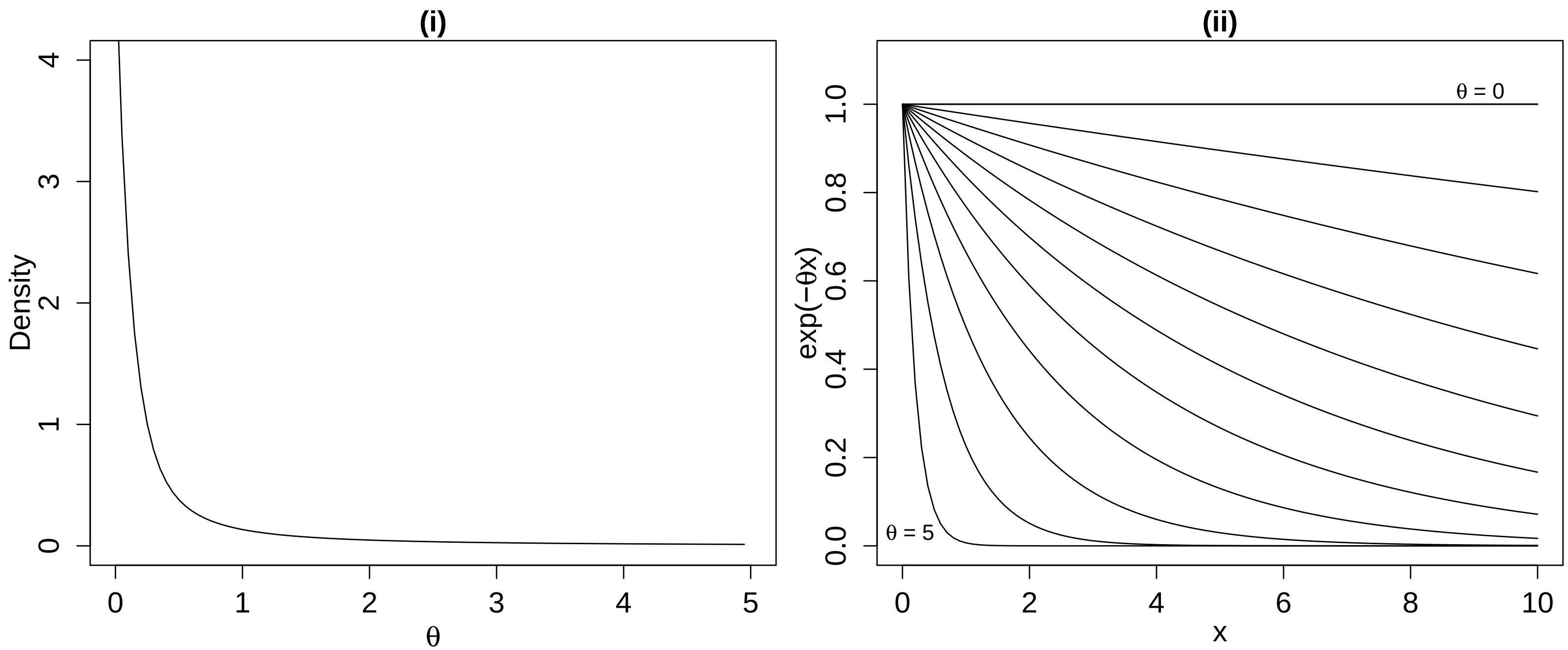
$$p(\boldsymbol{\theta}) \propto \sqrt{\det(\boldsymbol{Z}^*(\boldsymbol{\theta}))},$$

  where $\boldsymbol{Z}^*(\boldsymbol{\theta}) = \int_{\mathcal{X}} J_x(\boldsymbol{\theta})^T J_x(\boldsymbol{\theta}) dx$

- In linear regression one obtains $p(\boldsymbol{\theta}) \propto const$

For the exponential regression example one obtains $J_x(\theta) = -x\exp(-\theta x)$, calculating $\int_{\mathcal{X}} J_x(\theta)^2 dx$ and applying the square root, one obtains

$$p(\theta) \propto \exp(-10\theta)\sqrt{\frac{\exp(20\theta) - 200\theta^2 - 20\theta - 1}{\theta^3}}$$



Functional Uniform priors: Computational Remarks

- In general the functional uniform prior

$$p(\boldsymbol{\theta}) \propto \sqrt{\det(\boldsymbol{Z}^*(\boldsymbol{\theta}))},$$

  can be hard to calculate (as is Jeffreys prior)

- Main advantage here: In most cases needs to be calculated only once and can then be transferred to desired situation (design region and parameter bounds)

- Usually straightforward to approximate in terms of known distributions (*e.g.* mixtures)

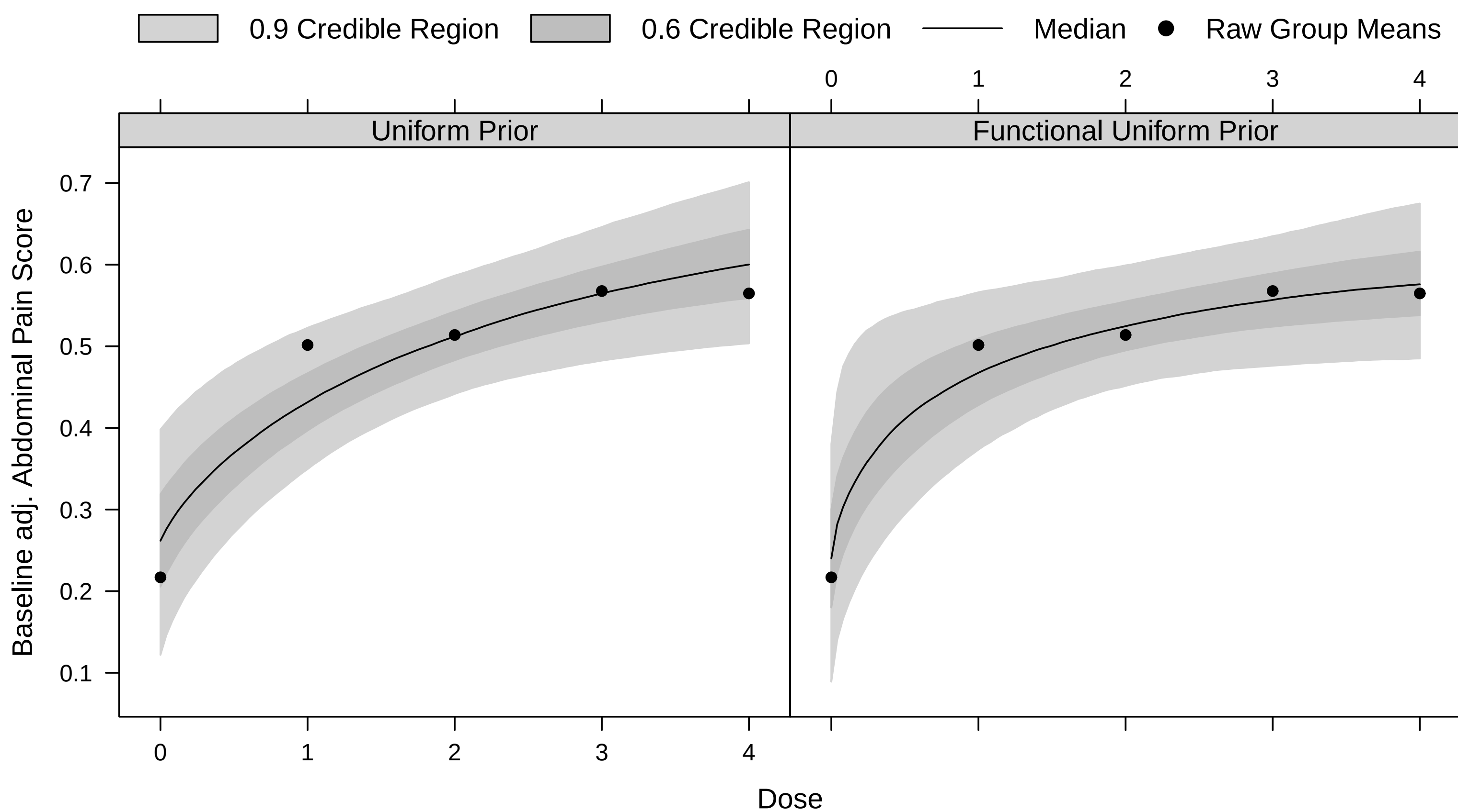
Applications

Assume functional uniform prior on space $\frac{x}{\theta_2+x}$ (other parameter linear), assume same bounds $[0, 6]$ as in uniform parameter scale, and use Euclidean distance in target space.

One obtains $J_x^2 = x^2/(x+\theta_2)^4$, calculating the integral $\int_0^4 J_x^2(\theta_2)\,dx$ and applying the square root leads to

$$p(\theta_2) \propto 1/\sqrt{\theta_2^4 + 12\theta_2^3 + 48\theta_2^2 + 64\theta_2}.$$

IBS dose-response study



- Uncertainty intervals larger for Functional Uniform prior
- Posterior median closer to observed means for Functional Uniform

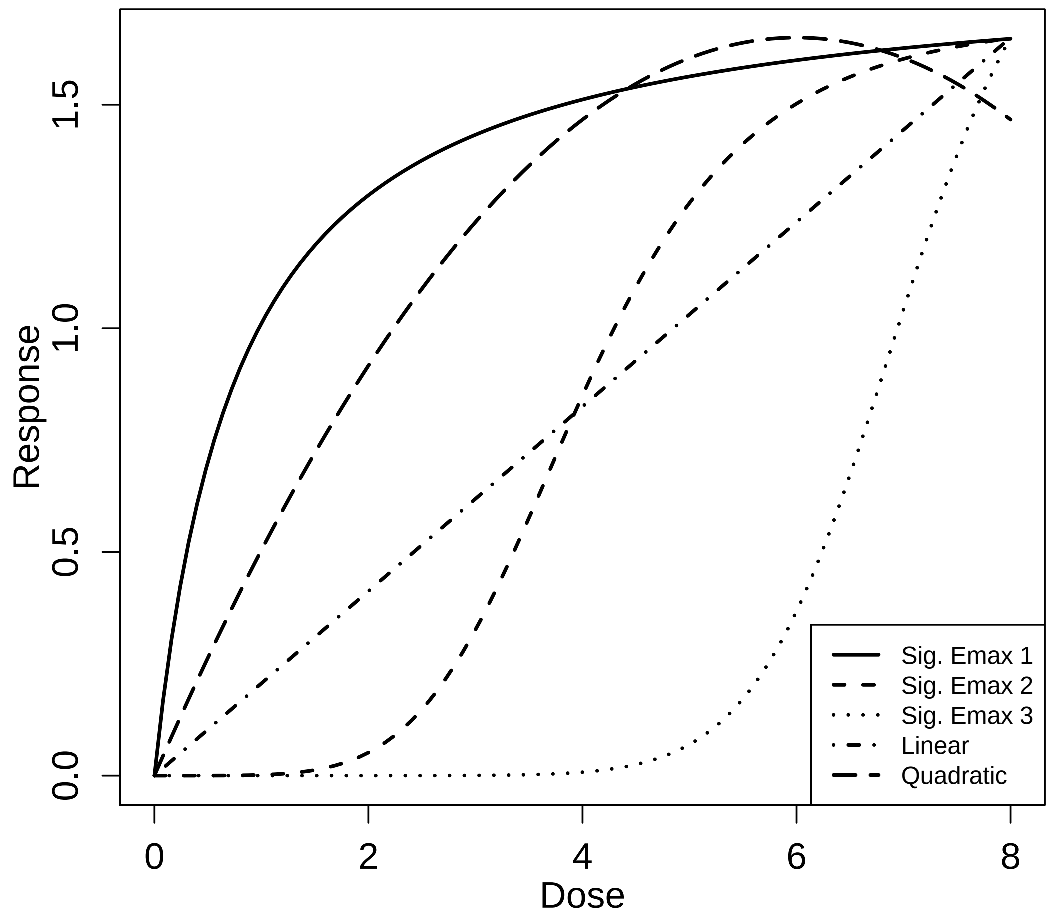
- Setup of PhRMA Adaptive DR working group (2nd round, lead by Vlad), with doses placebo, 2, 4, 6, 8 and 25 or 50 patients per group. Residual distribution is normal with standard deviation $\sqrt{4.5}$.
- Assume sigmoid Emax model

$$\mu(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 x^{\theta_4} / (x^{\theta_4} + \theta_3^{\theta_4})$$

  with the functional uniform prior on θ_3 and θ_4 and flat priors for θ_1 and θ_2. Compare this to uniform priors

- Both approaches use bounds on θ_2 and θ_3

Simulation Scenarios



		Uniform Prior				Functional Uniform Prior			
Scenario	N	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
Sig. Emax 1	125	0.44	0.40	2.72	11.99	0.36	0.45	1.63	4.93
Sig. Emax 2		0.27	0.50	1.36	7.04	0.26	0.53	0.79	2.84
Sig. Emax 3		0.20	0.53	0.59	4.22	0.22	0.52	0.85	2.60
Sig. Emax 1	250	0.29	0.35	1.70	11.24	0.27	0.35	1.20	4.38
Sig. Emax 2		0.19	0.51	1.20	5.88	0.19	0.48	0.84	2.77
Sig. Emax 3		0.14	0.64	0.59	4.25	0.14	0.40	0.33	3.22

		Uniform Prior			Functional Uniform Prior		
Scenario	N	MAE	CP	ILE	MAE	CP	ILE
Sig. Emax 1	125	0.256	0.903	1.098	0.230	0.914	1.028
Sig. Emax 2		0.278	0.895	1.144	0.258	0.909	1.089
Sig. Emax 3		0.243	0.902	1.014	0.251	0.898	1.030
Linear		0.266	0.901	1.100	0.241	0.918	1.057
Quadratic		0.272	0.880	1.109	0.242	0.898	1.038
Sig. Emax 1	250	0.185	0.908	0.818	0.167	0.920	0.768
Sig. Emax 2		0.196	0.908	0.850	0.187	0.910	0.811
Sig. Emax 3		0.174	0.913	0.738	0.170	0.912	0.744
Linear		0.200	0.891	0.831	0.189	0.900	0.794
Quadratic		0.201	0.881	0.839	0.185	0.886	0.782

- Conclusions: Functional uniform prior outperforms uniform prior in terms of frequentist properties, in particular for estimation of nonlinear parameters, because the uniform priors lead to rather informative priors in the function space

- Additional simulations: For binary data also compared uniform, functional uniform and Jeffreys prior

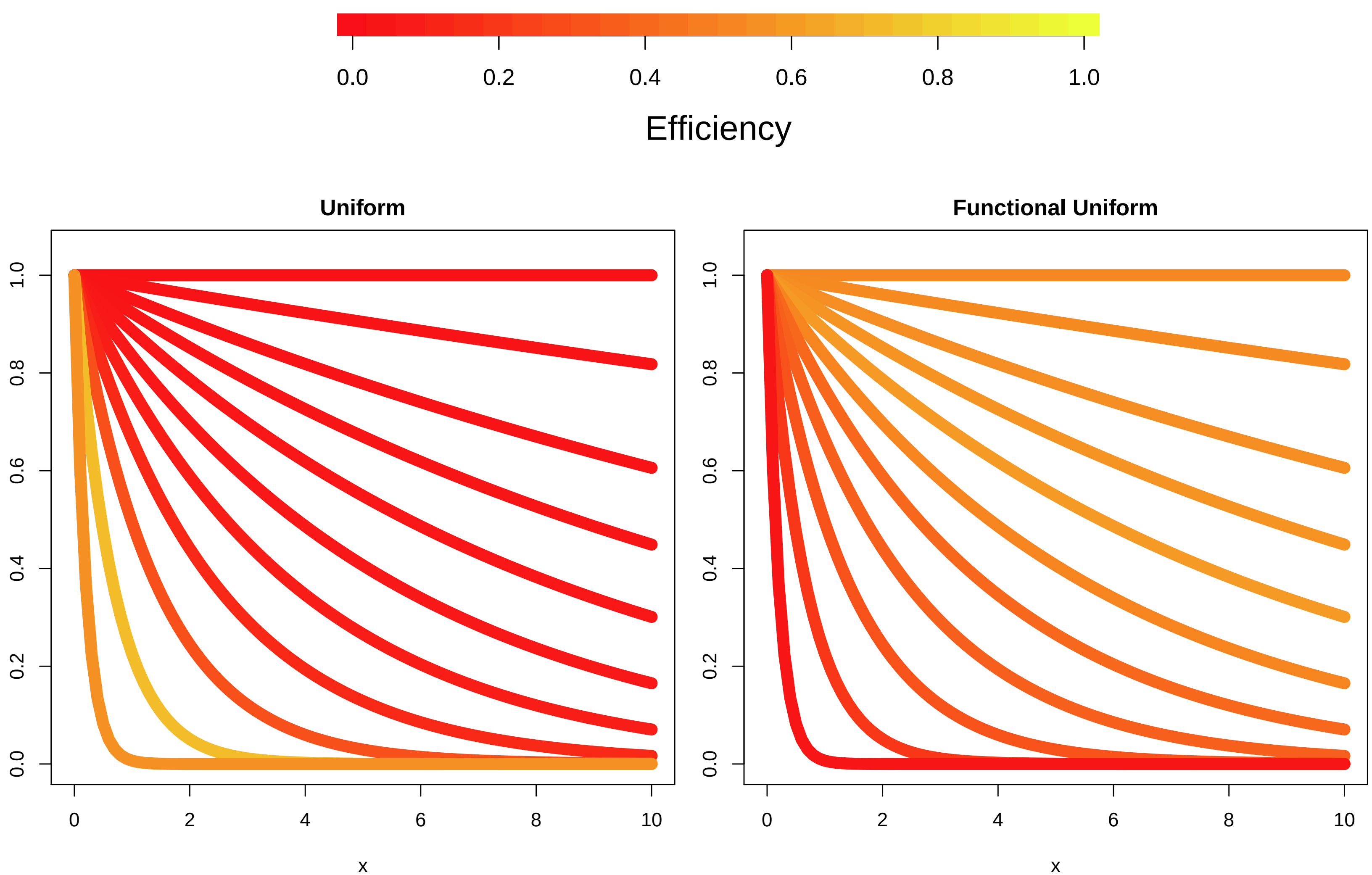
↪ Main conclusion: Functional uniform and Jeffreys perform roughly equally, both perform better in terms of frequentist properties than uniform prior distribution

- Calculate Bayesian D-optimal design for the simple exponential regression model $\exp(-\theta x)$ in case of homoscedastic normal data
- Fisher information is given by $I(d, \theta) \propto \sum w_i x_i^2 \exp(-2\theta x_i)$, the asymptotic variance by $V(d, \theta) \propto 1/I(d, \theta)$
- Minimize $-\int \log(I(d, \theta))\pi(\theta)d\theta$, for uniform and functional uniform prior distribution
- Optimization over restricted design space $x \in [0, 10]$ and only up to 5 design points

- Optimal design under uniform prior concentrates $> 90\%$ on one design point at $x \approx 0.4$
- Optimal design under functional uniform prior has more design points in $[0, 10]$ and allocations are less concentrated
- Compare for different shapes the efficiency

$$\mathrm{Eff}(d, \theta) = \exp(\log(I(d, \theta)) - \log(I(d_{opt}(\theta), \theta)))$$

Bayesian optimal design



Conclusions

- Functional Uniform priors avoid some of the problems of the Jeffreys and distributions uniform in the parameters
  $\longrightarrow$ parameterization invariant and available prior to data collection
- Not motivated by information theoretic consideraions
  $\longrightarrow$ few reason to formally call them uninformative (selection of (M, d) up to the user), but in many situations reasonable
- There are situations were they are not adequate (*e.g.* neural network models, when there is a well-justified prior)

Bornkamp, B., Pinheiro, J. and Bretz, F. (2011) DoseFinding: Planning and Analyzing Dose Finding experiments, R-package `http://cran.r-project.org/package=DoseFinding`

Dembski, W. A. (1990) Uniform probability, Journal of Theoretical Probability, 3, 611–626.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997) Non-informative priors via sieves and packing numbers, in S. Panchapakesan and N. Balakrishnan (eds.), Advances in Statistical Decision Theory and Applications, Birkhäuser, Boston, pp. 119132.

Pennec, X. (2006) Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. Journal of Mathematical Imaging and Vision 25, 127154.