Approximating mixture distributions using finite numbers of components

Christian Röver and Tim Friede

Department of Medical Statistics
University Medical Center Göttingen

March 17, 2016

InSPiRe
Innovative methodology for small populations research

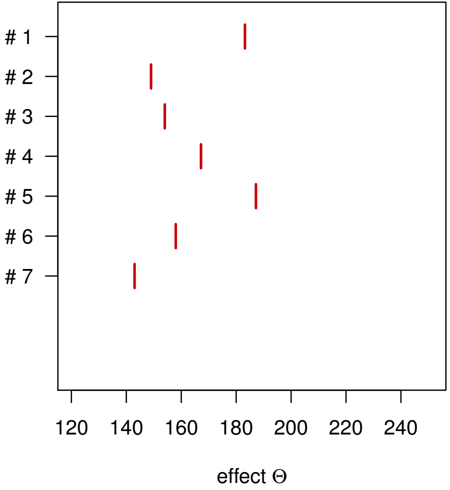
This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number FP HEALTH 2013-602144.

Overview

- meta analysis example
- general problem: mixture distributions
- discrete ‘grid’ approximations
- design strategy & algorithm
- application to meta analysis problem

Meta analysis

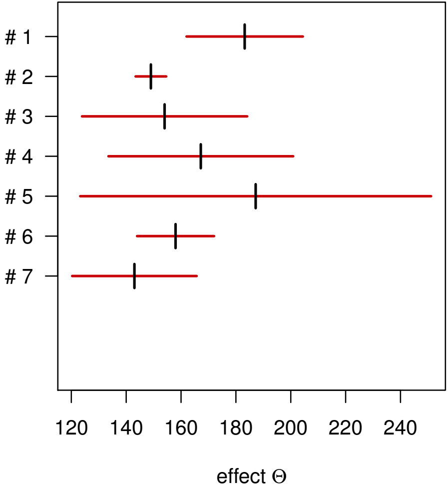
Context: random-effects meta-analysis



- have:
  - estimates y_i
  - standard errors σ_i
- want:
  - combined estimate $\hat{\Theta}$
- nuisance parameter:
  - between-trial heterogeneity τ

Meta analysis

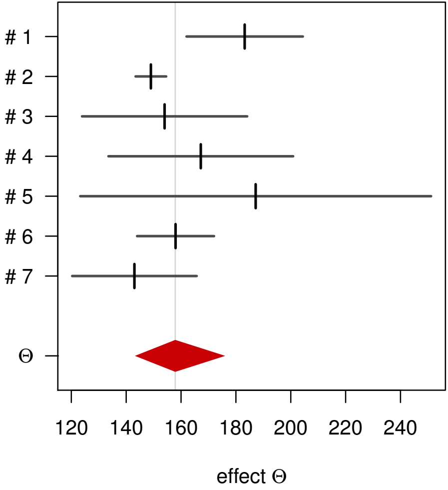
Context: random-effects meta-analysis



- have:
  - estimates y_i
  - standard errors σ_i
- want:
  - combined estimate $\hat{\Theta}$
- nuisance parameter:
  - between-trial heterogeneity τ

Meta analysis

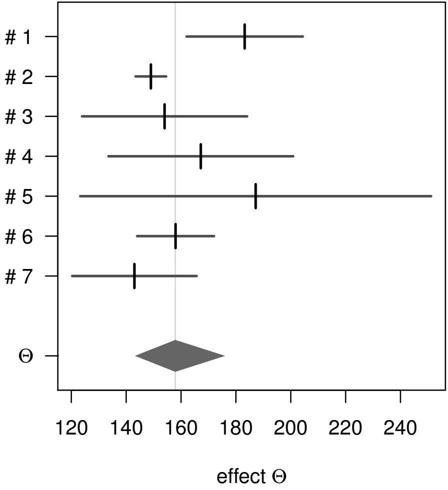
Context: random-effects meta-analysis



- have:
  - estimates y_i
  - standard errors σ_i
- want:
  - combined estimate $\hat{\Theta}$
- nuisance parameter:
  - between-trial heterogeneity τ

Meta analysis

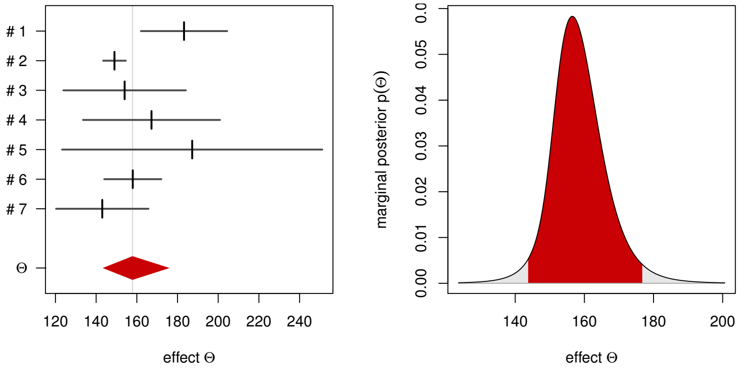
Context: random-effects meta-analysis



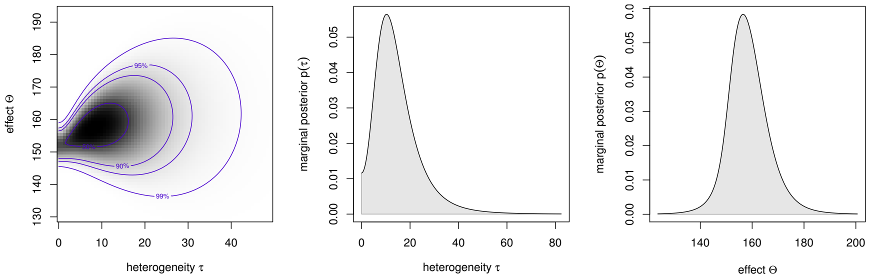
- have:
  - estimates y_i
  - standard errors σ_i
- want:
  - combined estimate $\hat{\Theta}$
- nuisance parameter:
  - between-trial heterogeneity τ

Meta analysis

Context: random-effects meta-analysis



- estimation:
  via marginal posterior distribution of parameter Θ



- two parameters
- parameter estimation with two unknowns:
  joint & marginal posterior distributions

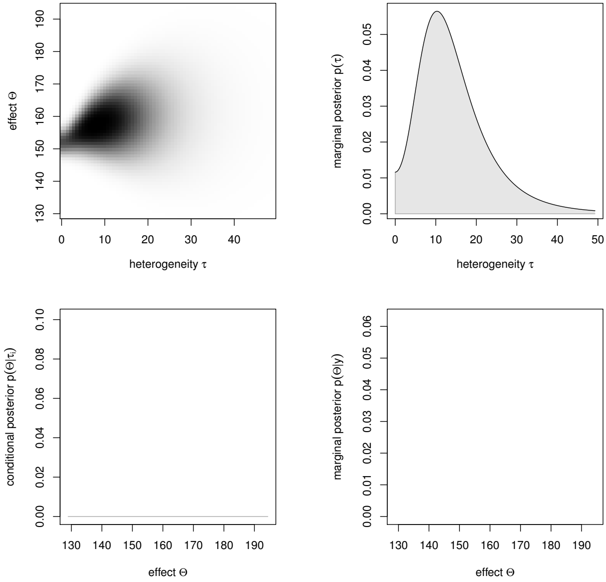
- here:
  easy to derive one of the **marginals**: $p(\tau|y)$
  and **conditional** posteriors $p(\Theta|\tau, y)$
- $p(\tau|y) = \ldots$ (...function of y_i, σ_i,...)
- $p(\Theta|\tau, y) = \text{Normal}(\mu = f_1(\tau), \sigma = f_2(\tau))$

- but main interest in *other* marginal: $p(\Theta|y)$

- $$\begin{aligned} p(\Theta|y) &= \int \overbrace{p(\Theta, \tau, y)}^{\text{joint}}\, \mathrm{d}\tau \\ &= \int \underbrace{p(\Theta|\tau, y)}_{\text{conditional}}\, \underbrace{p(\tau|y)}_{\text{marginal}}\, \mathrm{d}\tau \end{aligned}$$ is a **mixture distribution**

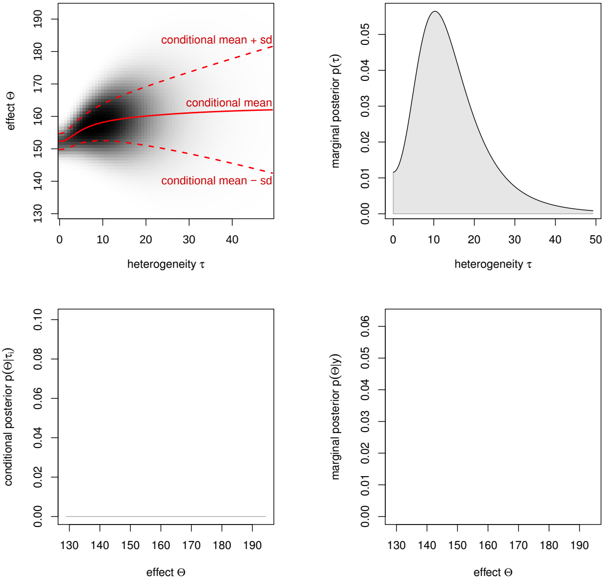
Meta analysis

Context: random-effects meta-analysis

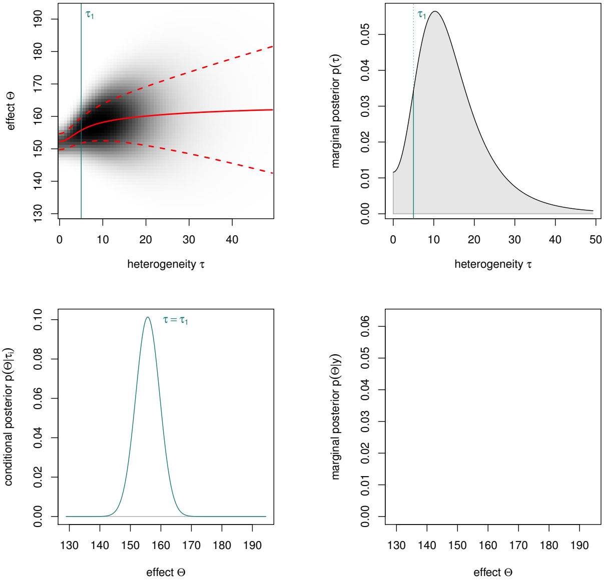


Meta analysis

Context: random-effects meta-analysis

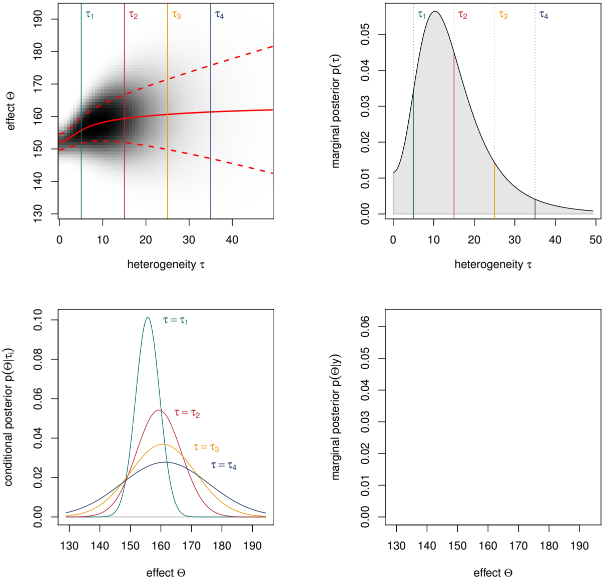


Meta analysis
Context: random-effects meta-analysis



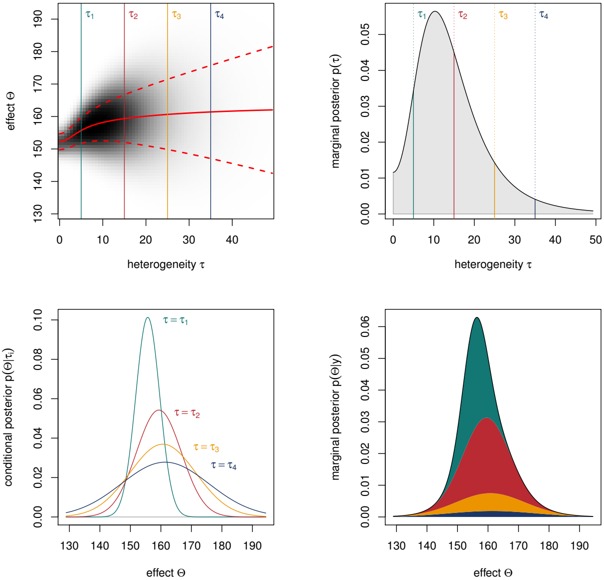
Meta analysis

Context: random-effects meta-analysis



Meta analysis

Context: random-effects meta-analysis



Mixture distributions

The general problem

- mixture distribution:
  - a convex combination of “component” distributions
  - “a distribution whose parameters are random variables”
- (“conditional”) distribution with density $p(y|x)$
- “parameter” x follows a distribution $p(x)$
- *marginal* / *mixture* is $p(y) = \int_X p(y|x)\,\mathrm{d}p(x)$
- x discrete: $p(y) = \sum_i p(y|x_i)\,p(x_i)$
- ubiquitous in many applications
  - Student-*t* distribution
  - negative binomial distribution
  - marginal distributions
  - convolution
  - ...

Mixture distributions

How to approximate?

- approximating the **continuous** mixture through a **discrete** set of points in τ...
- actual marginal:

$$p(\Theta) = \int p(\Theta|\tau)\, p(\tau)\, \mathrm{d}\tau$$

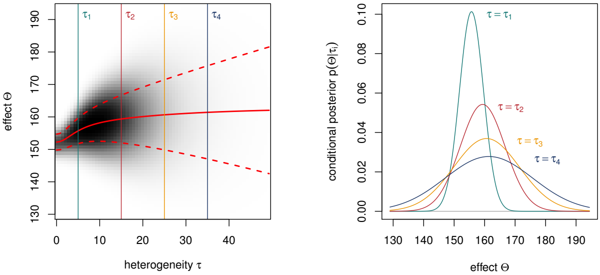
- approximation:

$$p(\Theta) \approx \sum_i p(\Theta|\tau_i)\, \pi_i$$

- Questions:
  - how to set up the discrete grid of points?
  - how well can we approximate?
  - do we have a handle on accuracy?

Mixture distributions

Motivation: discretizing a mixture



- Note: conditional distributions $p(\Theta|\tau, y)$ are very **different** for τ_1 and τ_2 and rather **similar** for τ_3 and τ_4.
- idea: may need fewer bins for larger τ values...?
- ...bin spacing based on similarity / dissimilarity of conditionals?

- need: discretization of the mixing distribution $p(x)$.
- domain of X: $\mathbb{R}$ (or subset)
- define **bin margins**: $x_{(1)} < x_{(2)} < \ldots < x_{(k-1)}$
- **bins**:

$$\mathcal{X}_i = \begin{cases} \{x : x \leq x_{(1)}\} & \text{if } i = 1 \\ \{x : x_{(i-1)} < x \leq x_{(i)}\} & \text{if } 1 < i < k \\ \{x : x_{(k-1)} < x\} & \text{if } i = k. \end{cases}$$

- **reference points**: $\tilde{x}_1, \ldots, \tilde{x}_k$, where $\tilde{x}_i \in \mathcal{X}_i$
- **bin probabilities**: $\pi_i = \mathrm{P}\big(x_{(i-1)} < x \leq x_{(i)}\big) = \mathrm{P}\big(x \in \mathcal{X}_i\big)$

Discretizing mixture distributions

Setting up a binned mixture

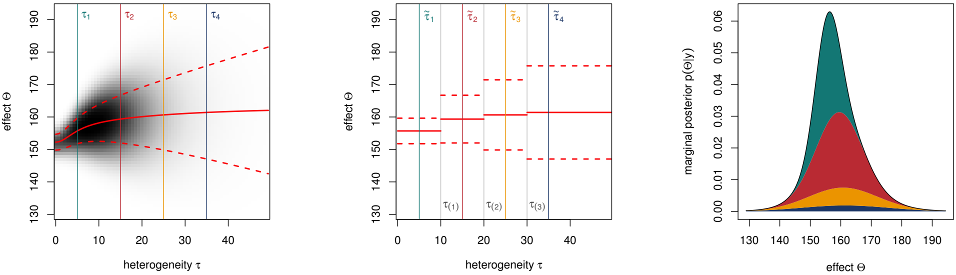
- actual distribution: $p(x, y)$
- discrete approximation: $q(x, y)$
- same marginal (mixing distribution): $q(x) = p(x)$
- but “binned” conditionals:
  $q(y|x) \;=\; p(y|x = \tilde{x}_i) \quad$ for $x \in \mathcal{X}_i$.
- q similar to p,
  instead of conditioning on “exact” x,
  conditioning on corresponding bin’s reference point $\tilde{x}_i$
- marginal:

$$
\begin{aligned}
q(y) &= \int q(y|x)\, q(x)\, \mathrm{d}x \\
&= \sum_i \pi_i\, p(y|\tilde{x}_i)
\end{aligned}
$$

Discretizing mixture distributions

Setting up a binned mixture

- in previous example:
  - bin margins: $\tau_{(1)} = 10$, $\tau_{(2)} = 20$, $\tau_{(3)} = 30$
  - reference points: $\tilde{\tau}_1 = 5$, $\tilde{\tau}_2 = 15$, $\tilde{\tau}_3 = 25$, $\tilde{\tau}_4 = 35$
  - probabilities: $\pi_1 = 0.34$, $\pi_2 = 0.44$, $\pi_3 = 0.15$, $\pi_4 = 0.07$



Similarity / dissimilarity of distributions
Kullback-Leibler divergence

- The **Kullback-Leibler divergence** of two distributions with density functions p and q is defined as

$$\mathcal{D}_{\mathrm{KL}}\big(p(\theta)\big\|q(\theta)\big) \;=\; \int_\Theta \log\Big(\frac{p(\theta)}{q(\theta)}\Big)\, p(\theta)\, \mathrm{d}\theta$$
$$\;=\; \mathrm{E}_{p(\theta)}\Big[\log\Big(\frac{p(\theta)}{q(\theta)}\Big)\Big]$$

- The **symmetrized KL-divergence** of two distributions is defined as

$$\mathcal{D}_{\mathrm{s}}\big(p(\theta)\big\|q(\theta)\big) \;=\; \mathcal{D}_{\mathrm{KL}}\big(p(\theta)\big\|q(\theta)\big) + \mathcal{D}_{\mathrm{KL}}\big(q(\theta)\big\|p(\theta)\big)$$

- the symmetrized KL-divergence…
  - is symmetric: $\mathcal{D}_{\mathrm{s}}\big(p(\theta)\big\|q(\theta)\big) \;=\; \mathcal{D}_{\mathrm{s}}\big(q(\theta)\big\|p(\theta)\big)$
  - is always positive: $\mathcal{D}_{\mathrm{s}}\big(p(\theta)\big\|q(\theta)\big) \;\geq\; 0$

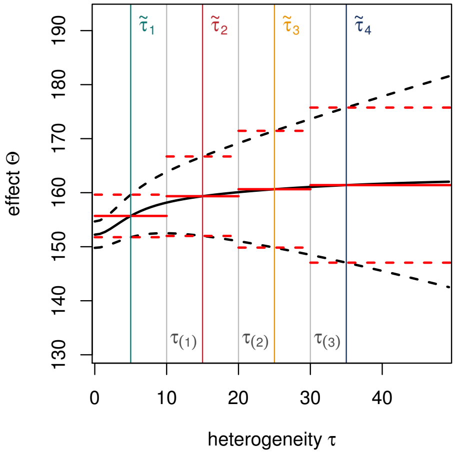
- How to interpret divergences?
- measure of “discrepancy” between distributions
- heuristically: expected log ratio of densities...
  – relevant case here: $p(x) \approx q(x)$.
- $\mathcal{D}_{\mathrm{KL}}\big(p(x), q(x)\big) = 0 \quad$ for $\quad p = q$
- $\mathcal{D}_{\mathrm{KL}}\big(p(x), q(x)\big) = 0.01$
  corresponds to (expected) $\approx 1\%$ difference in densities

- consider: divergence between reference point and other points *within each bin*
- define:

$$d_i \;=\; \max_{x \in \mathcal{X}_i}\Big\{\mathcal{D}_{\mathrm{s}}\big(p(y|x)\big\|p(y|\tilde{x}_i)\big)\Big\} \;=\; \max_{x \in \mathcal{X}_i}\Big\{\mathcal{D}_{\mathrm{s}}\big(p(y|x)\big\|q(y|x)\big)\Big\},$$

  the **bin-wise maximum divergence**
- "worst-case discrepancy" introduced within each bin

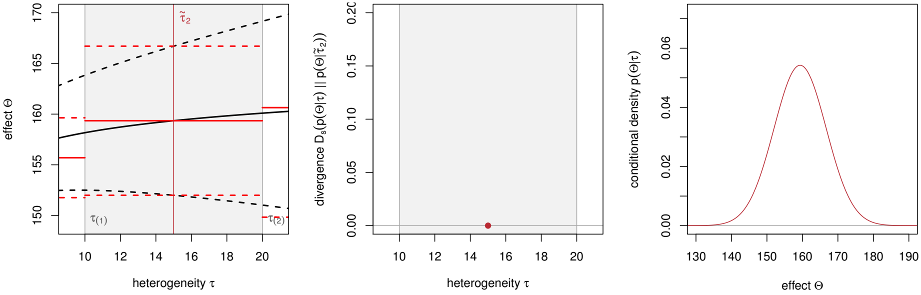
Bin-wise maximum divergence: example



- recall: actual parameters of conditionals $p(y|x)$ (in black) vs. parameters of $q(y|x)$ assumed through binning (in red)

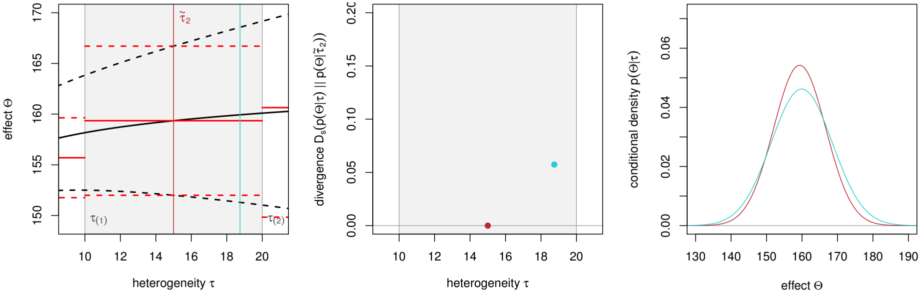
Divergence

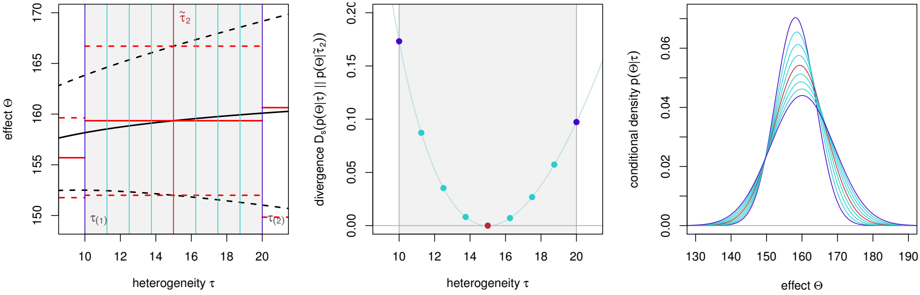
Bin-wise maximum divergence: example



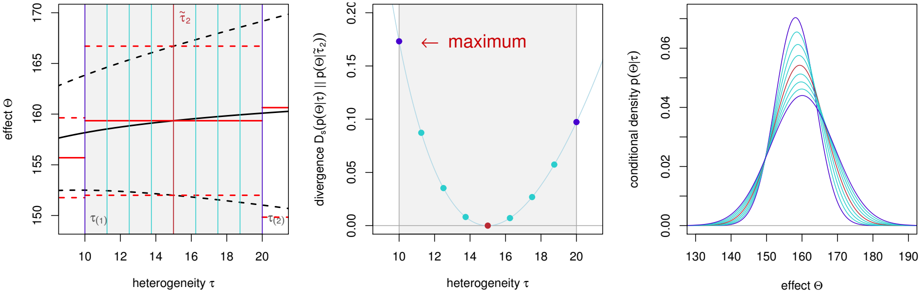
Divergence

Bin-wise maximum divergence: example




effect Θ
heterogeneity τ
$\tilde{\tau}_2$
$\tau_{(1)}$
$\tau_{(2)}$
divergence $D_s(p(\Theta|\tau) \,||\, p(\Theta|\tilde{\tau}_2))$
heterogeneity τ
conditional density p(Θ|τ)
effect Θ

Bin-wise maximum divergence: example



- determine maximum d_i for each bin i (usually at bin margin)

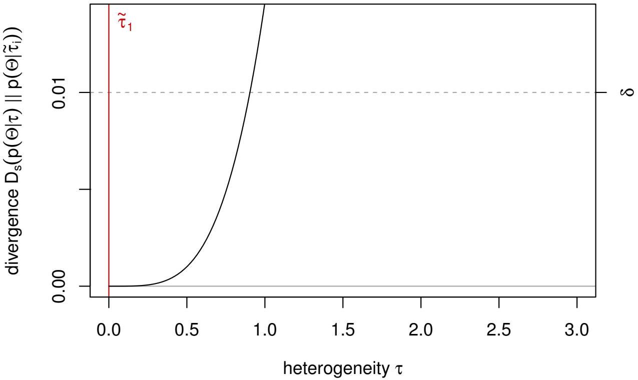
Bounding divergence
Idea

- now consider divergences
  of true and approximate **marginals** $p(y)$ and $q(y)$

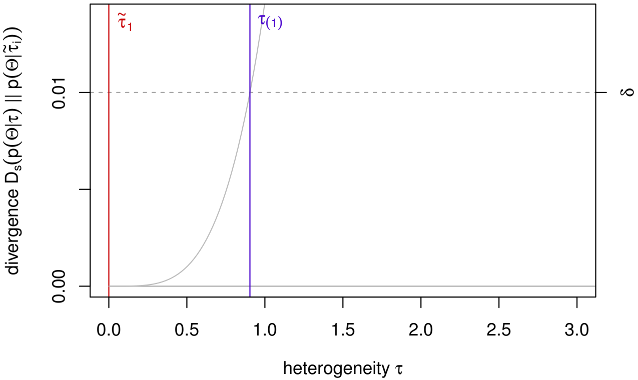
  (*not* the conditionals!)
- what about $\mathcal{D}_{\mathrm{s}}\big(p(y)\big\|q(y)\big)$?
- having the individual *bin-wise* divergences d_i, we can show:

$$\begin{aligned}\mathcal{D}_{\mathrm{s}}\big(p(y)\big\|q(y)\big) &\leq \sum_i \pi_i\, d_i \\ &\leq \max_i d_i\end{aligned}$$

- in other words:
  by bounding bin-wise divergences (of conditionals)
  we can bound the overall divergence (of marginals)
- "DIRECT (Divergence Restricted Conditional Tesselation)" method



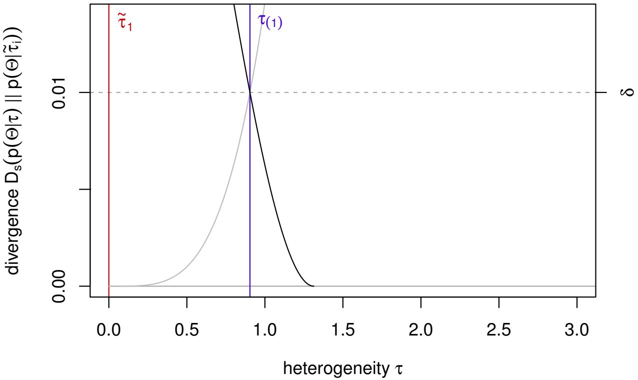
- 1st reference point $\tilde{\tau}_1$ at zero



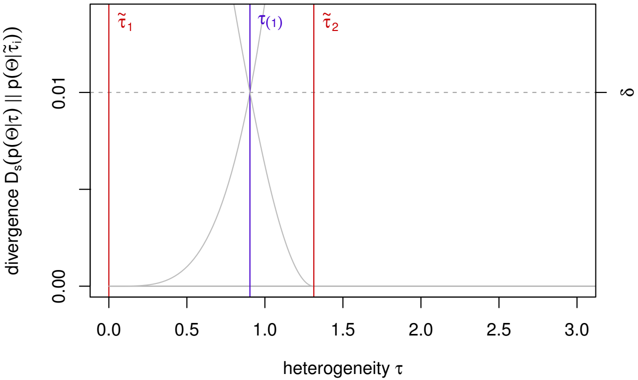
- 1st reference point $\tilde{\tau}_1$ at zero, first margin $\tau_{(1)}$ at 0.904

Discretizing mixtures

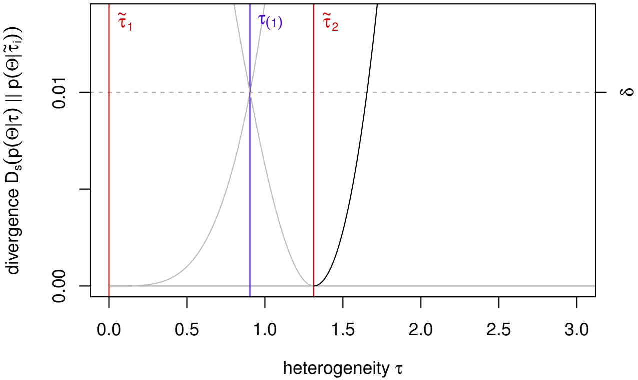
Sequential DIRECT algorithm



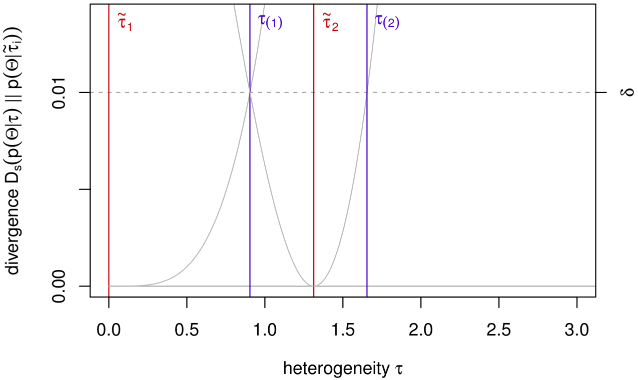
- 1st reference point $\tilde{\tau}_1$ at zero, first margin $\tau_{(1)}$ at 0.904 (...)



- 1st reference point $\tilde{\tau}_1$ at zero, first margin $\tau_{(1)}$ at 0.904 (...)



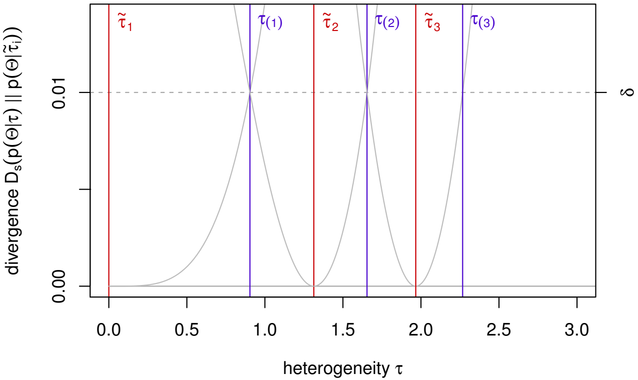
- 1st reference point $\tilde{\tau}_1$ at zero, first margin $\tau_{(1)}$ at 0.904 (...)



- 1st reference point $\tilde{\tau}_1$ at zero, first margin $\tau_{(1)}$ at 0.904 (...)

Discretizing mixtures

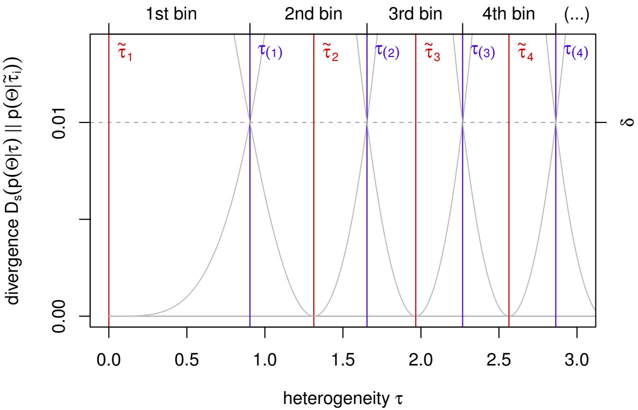
Sequential DIRECT algorithm



- 1st reference point $\tilde{\tau}_1$ at zero, first margin $\tau_{(1)}$ at 0.904 (...)

Discretizing mixtures

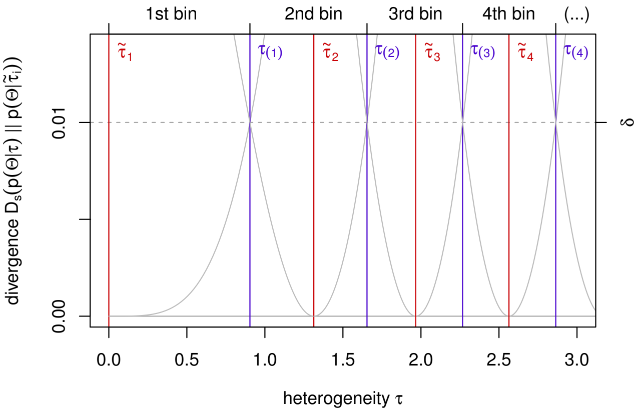
Sequential DIRECT algorithm



- 1st reference point $\tilde{\tau}_1$ at zero, first margin $\tau_{(1)}$ at 0.904 (...)
- result: binning with bounded divergence ($\leq \delta$) *per bin*

Discretizing mixtures

Sequential DIRECT algorithm



- 1st reference point $\tilde{\tau}_1$ at zero, first margin $\tau_{(1)}$ at 0.904 $(\ldots)$
- result: binning with bounded divergence ($\leq \delta$) *per bin*
- (when to stop?)

Discretizing mixtures

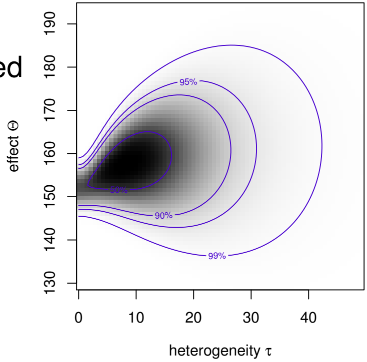
Sequential DIRECT algorithm (variations possible)

1. Specify $\delta > 0$, $0 \leq \epsilon \ll 1$, and starting reference point $\tilde{x}_1$ (e.g. minimum possible value, or $\frac{\epsilon}{2}$-quantile). Define $\epsilon_1 \geq 0$ as $\epsilon_1 := \mathrm{P}(X \leq \tilde{x}_1)$. Set $i = 1$.
2. Set $x^\star = \tilde{x}_1$. Obviously, $\mathcal{D}_\mathrm{s}\big(p(y|\tilde{x}_1)\big|\big|p(y|x^\star)\big) = 0$.
   Now increase $x^\star$ as far as possible while ensuring that $\mathcal{D}_\mathrm{s}\big(p(y|\tilde{x}_1)\big|\big|p(y|x^\star)\big) \leq \delta$.
   Use this point as the first bin margin: $x_{(1)} = x^\star$.
   Compute $\pi_1 = \mathrm{P}(x < x_{(1)})$. Set $i = i + 1$.
3. Increase $x^\star$ until $\mathcal{D}_\mathrm{s}\big(p(y|x_{(i-1)})\big|\big|p(y|x^\star)\big) = \delta$. Use this point as the next reference point: $\tilde{x}_i = x^\star$.
4. Increase $x^\star$ again until $\mathcal{D}_\mathrm{s}\big(p(y|\tilde{x}_i)\big|\big|p(y|x^\star)\big) = \delta$. Use this point as the next bin margin: $x_{(i)} = x^\star$.
5. Compute the bin weight $\pi_i = \mathrm{P}(x_{(i-1)} < X \leq x_{(i)})$.
6. If $\mathrm{P}(X > x_{(i)}) > (\epsilon - \epsilon_1)$, set $i = i + 1$ and proceed at step 3. Otherwise stop.

Discretizing mixtures

General algorithm

- remaining issue: ignored $\epsilon > 0$ tail probability (usually: problems at domain's margins)
- only need to keep track of reference points $\tilde{x}_i$ and probabilities π_i
- meta-analysis example: 35 reference ("support") points required ($\delta = 0.01$, $\epsilon = 0.001$)



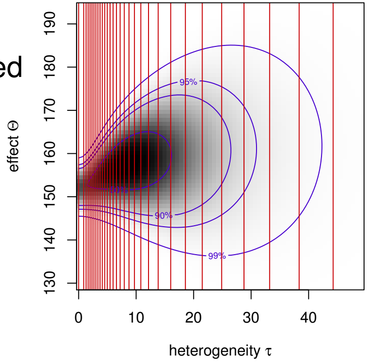
Discretizing mixtures

General algorithm

- remaining issue: ignored $\epsilon > 0$ tail probability (usually: problems at domain's margins)

- only need to keep track of reference points $\tilde{x}_i$ and probabilities π_i

- meta-analysis example: 35 reference ("support") points required ($\delta = 0.01$, $\epsilon = 0.001$)



Conclusions

- approximation allows to compute density, quantiles, moments,...
- algorithm yields quick-and-easy solution
- need to specify error budget in terms of
  - divergence δ
  - tail probability ϵ
- also works for discrete distributions ($p(x)$ or $p(y|x)$)
- meta-analysis application:

  implemented in `bayesmeta` R package
  `http://cran.r-project.org/package=bayesmeta`
- general procedure:

  C. Röver, T. Friede. *Discrete approximation of a mixture distribution via restricted divergence*. (submitted for publication.)
  `http://arxiv.org/abs/1602.04060`