Implementing discrete approximations to continuous mixture distributions

Christian Röver

Department of Medical Statistics
University Medical Center Göttingen

December 5, 2014

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

Overview

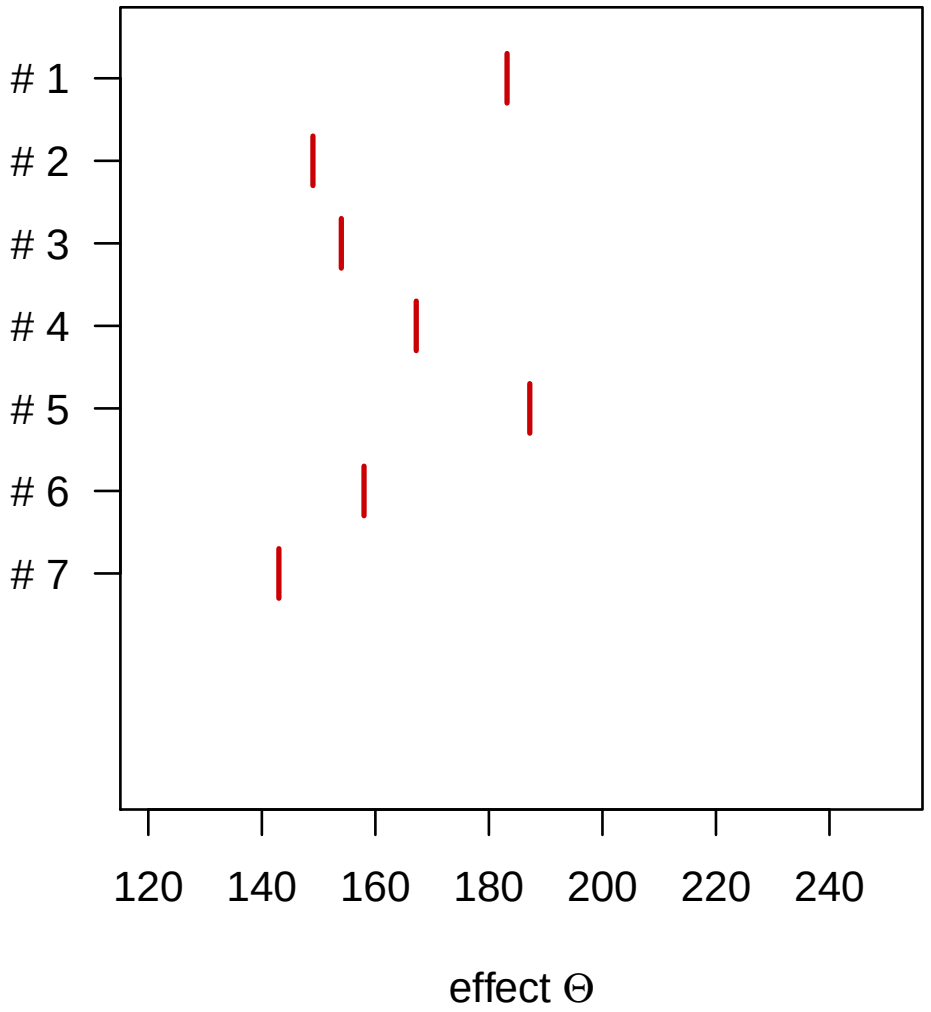
- mixture distributions
- meta analysis example
- discrete ‘grid’ approximations
- design strategy / algorithm
- example application

Mixture distributions

- 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* distribution of y 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
  - …

Meta analysis

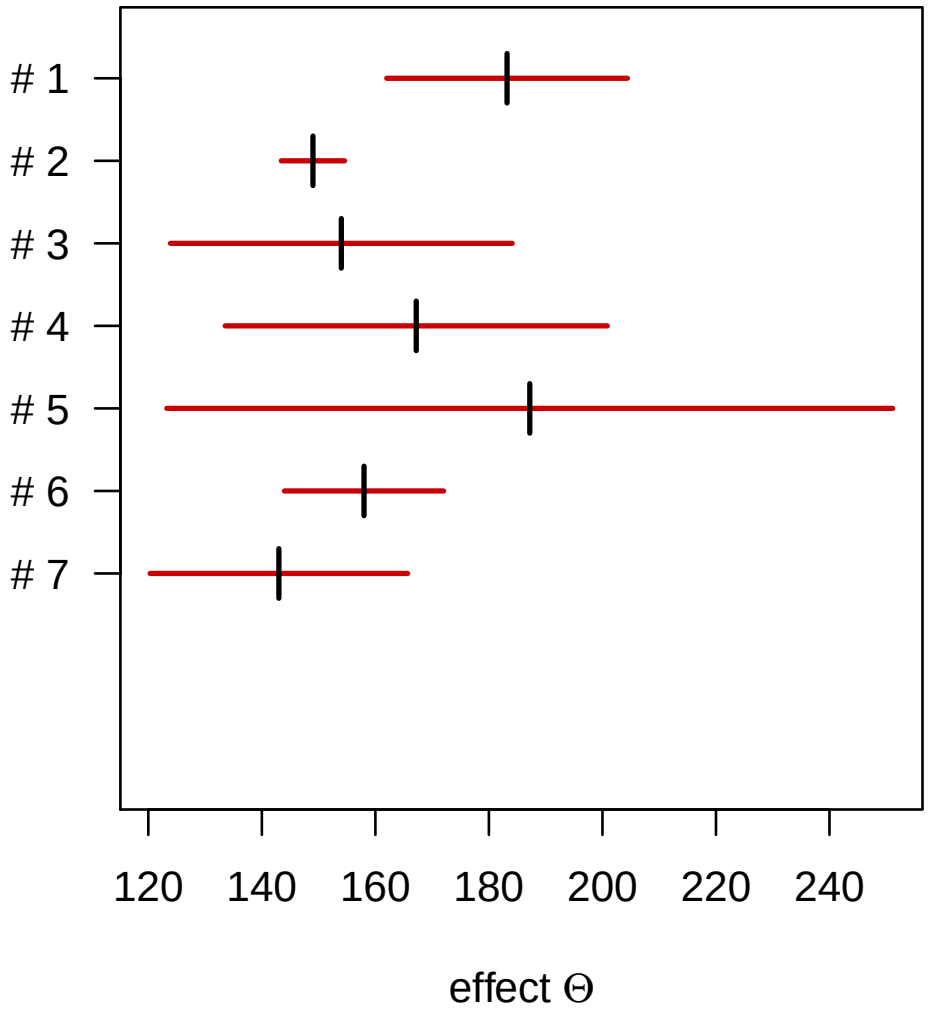
Context: random-effects meta-analysis



- have:
  - estimates y_i
  - standard errors σ_i
- want:
  - combined estimate $\hat{\Theta}$

Meta analysis

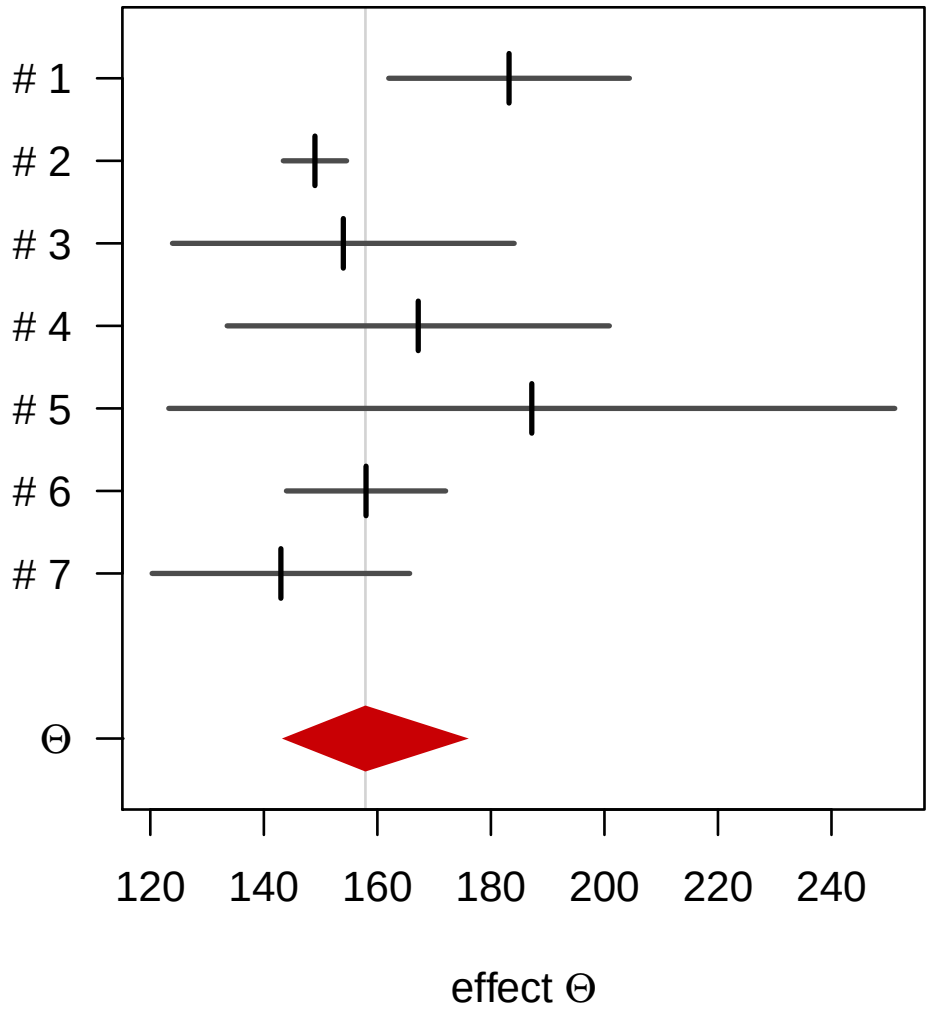
Context: random-effects meta-analysis



- have:
  - estimates y_i
  - standard errors σ_i
- want:
  - combined estimate $\hat{\Theta}$

Meta analysis

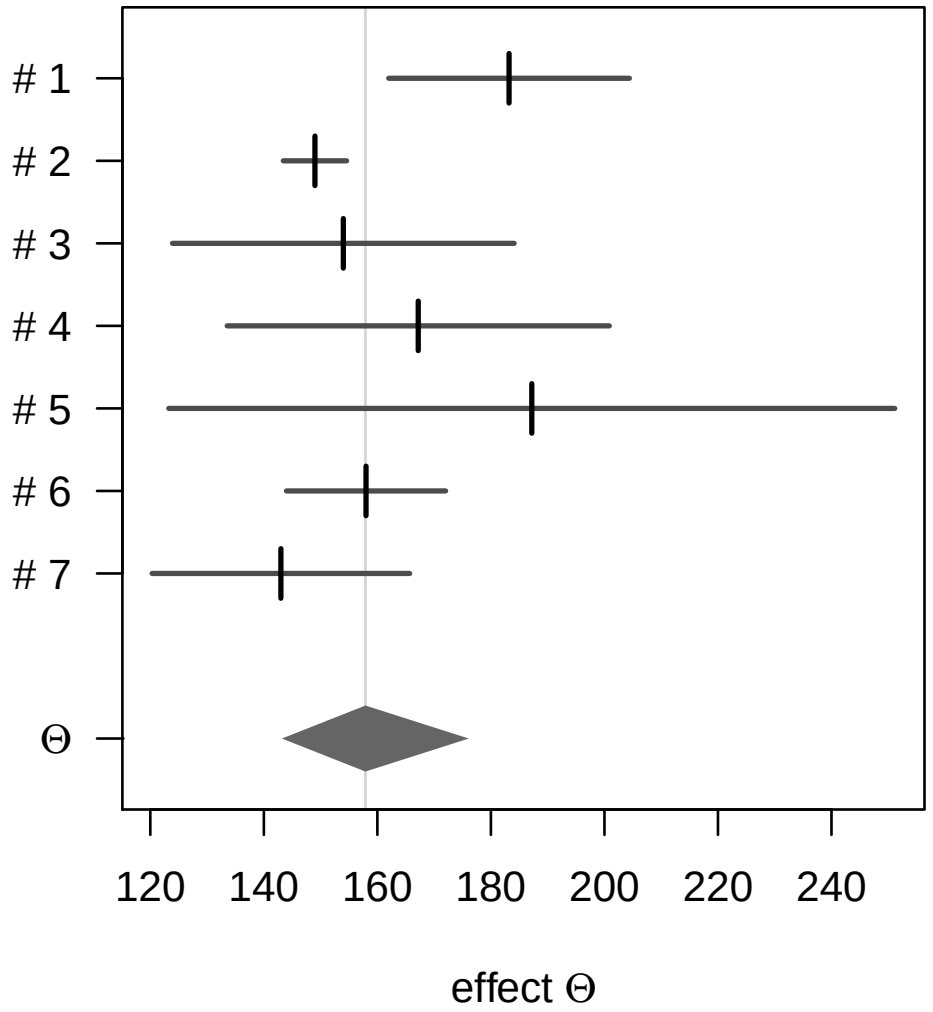
Context: random-effects meta-analysis



- have:
  - estimates y_i
  - standard errors σ_i
- want:
  - combined estimate $\hat{\Theta}$

Meta analysis

Context: random-effects meta-analysis



- have:
  - estimates y_i
  - standard errors σ_i
- want:
  - combined estimate $\hat{\Theta}$

Meta analysis

The random effects model

- assume:

$$y_i \sim \text{Normal}(\Theta, {\sigma_i}^2 + \tau^2)$$

Meta analysis

The random effects model

- assume:

$$y_i \sim \text{Normal}(\Theta, {\sigma_i}^2 + \tau^2)$$

- ingredients:

Data:

- estimates y_i
- standard errors σ_i

Parameters:

- true parameter value Θ
- heterogeneity τ

Meta analysis

The random effects model

- assume:

$$y_i \sim \text{Normal}(\Theta, {\sigma_i}^2 + \tau^2)$$

- ingredients:

Data:

- estimates y_i
- standard errors σ_i

Parameters:

- true parameter value Θ
- heterogeneity τ

Meta analysis

The random effects model

- assume:

$$y_i \sim \text{Normal}(\Theta, {\sigma_i}^2 + \tau^2)$$

- ingredients:

Data:

- estimates y_i
- standard errors σ_i

Parameters:

- true parameter value Θ
- heterogeneity τ

Meta analysis

The random effects model

- assume:

$$y_i \sim \text{Normal}(\Theta, {\sigma_i}^2 + \tau^2)$$

- ingredients:

Data:

- estimates y_i
- standard errors σ_i

Parameters:

- true parameter value Θ
- heterogeneity τ

Meta analysis

The random effects model

- assume:

$$y_i \sim \text{Normal}(\Theta, {\sigma_i}^2 + \tau^2)$$

- ingredients:

Data:

- estimates y_i
- standard errors σ_i

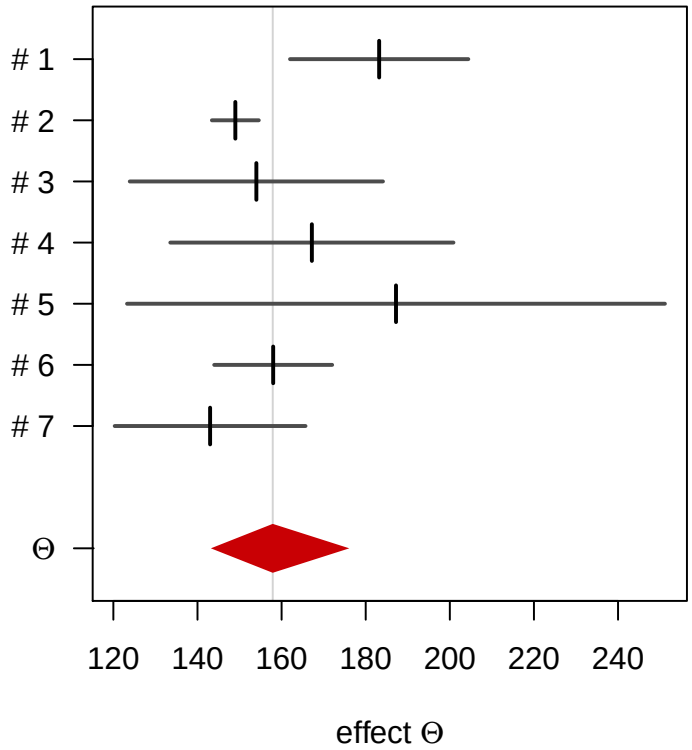
Parameters:

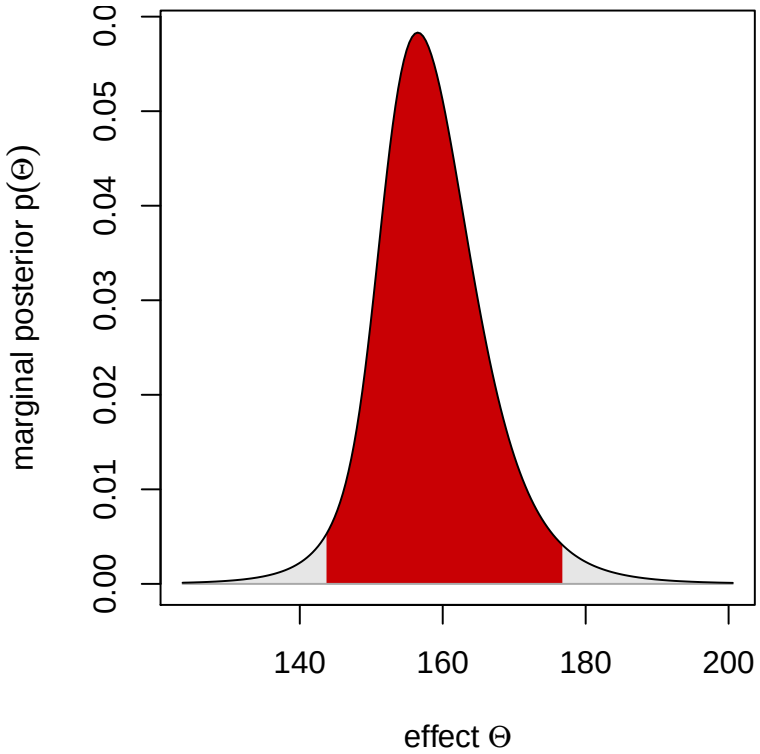
- true parameter value Θ
- heterogeneity τ

- $\Theta \in \mathbb{R}$ of primary interest
- $\tau \in \mathbb{R}^+$ nuisance parameter: account for (potential) incompatibility

Meta analysis example

Motivation: background

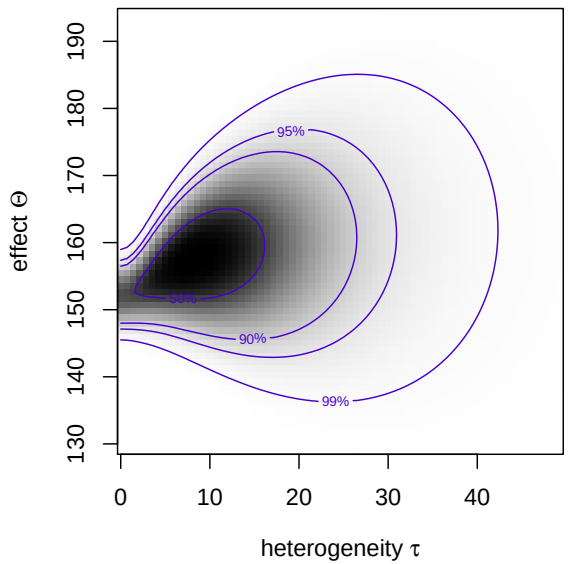


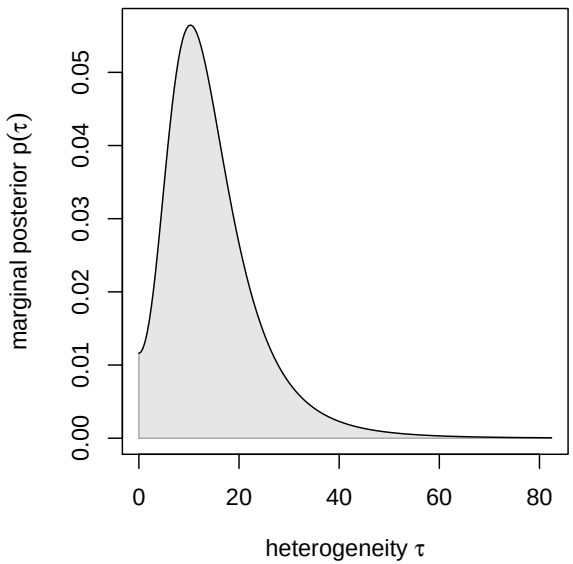


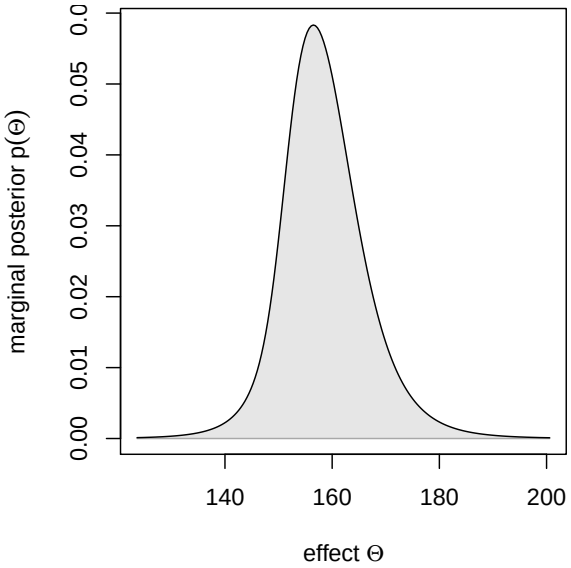
- estimation:
  via marginal posterior distribution of parameter Θ

Meta analysis example

Motivation: two-parameter model & marginals







- two unknowns:
  joint & marginal posterior distributions

Meta analysis example

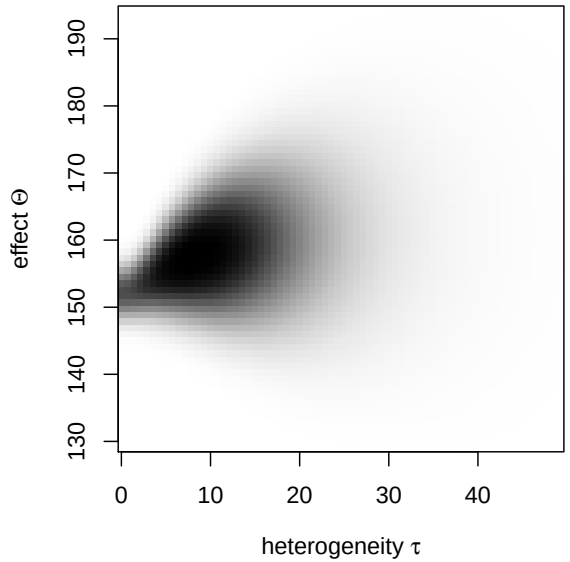
Motivation: two-parameter model, conditionals & marginals

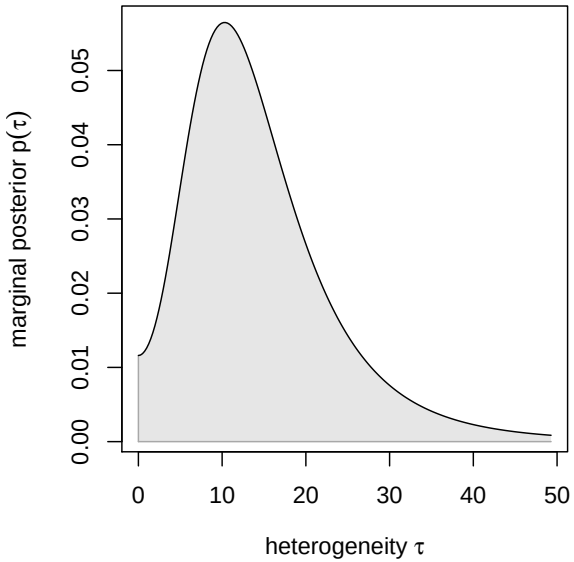
- here:
  easy to derive one of the **marginals**: $p(\tau|y)$
  and **conditional** posteriors $p(\Theta|\tau, y)$
- $p(\tau|y) = \ldots$ (...function of y_i, σ_i,...)
- $p(\Theta|\tau, y) = \text{Normal}(\mu = f_1(\tau), \sigma = f_2(\tau))$

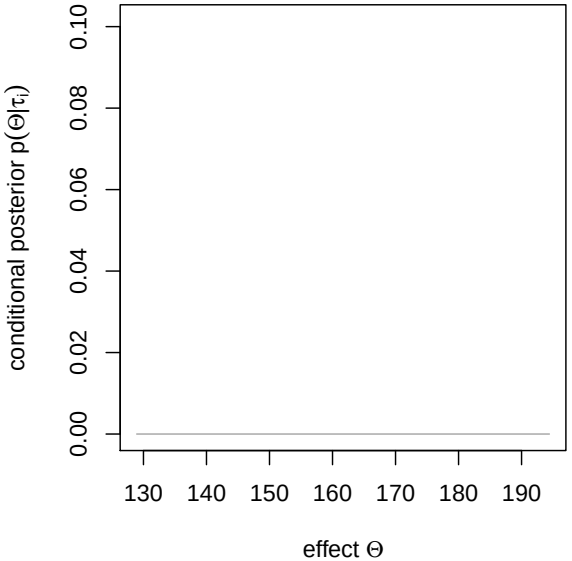
- but main interest in *other* marginal: $p(\Theta|y)$
- $p(\Theta|y) = \int \underbrace{p(\Theta|\tau, y)}_{\text{conditional}} \underbrace{p(\tau|y)}_{\text{marginal}} \, \mathrm{d}\tau$ is a **mixture distribution**

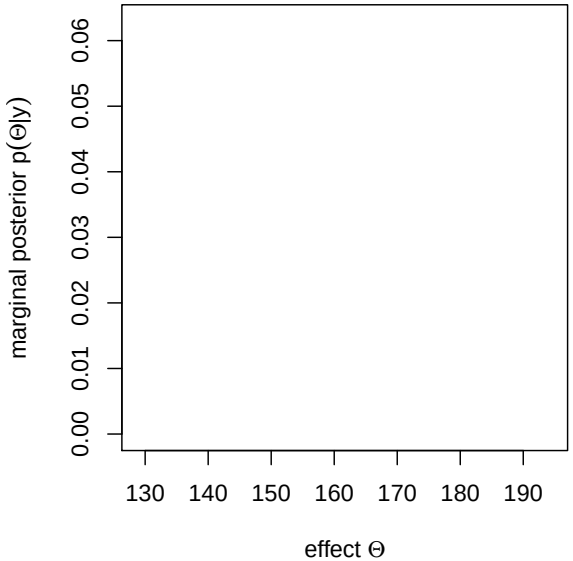
Meta analysis example

Motivation: two-parameter model, conditionals & marginals



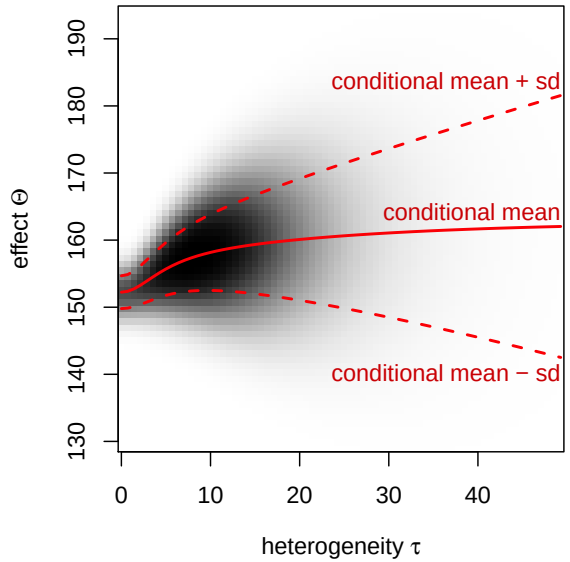


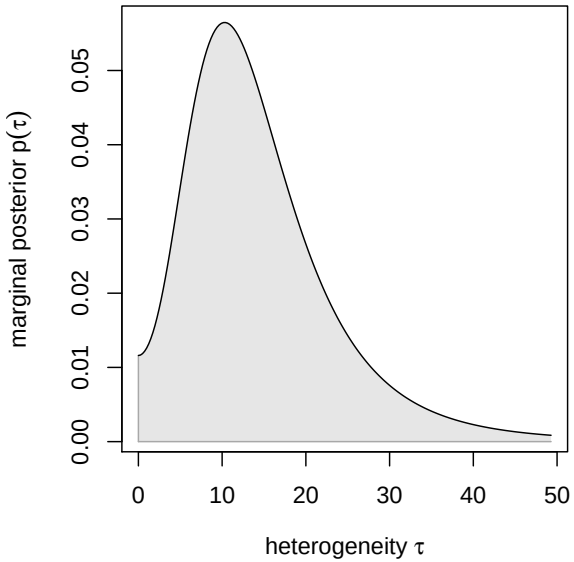


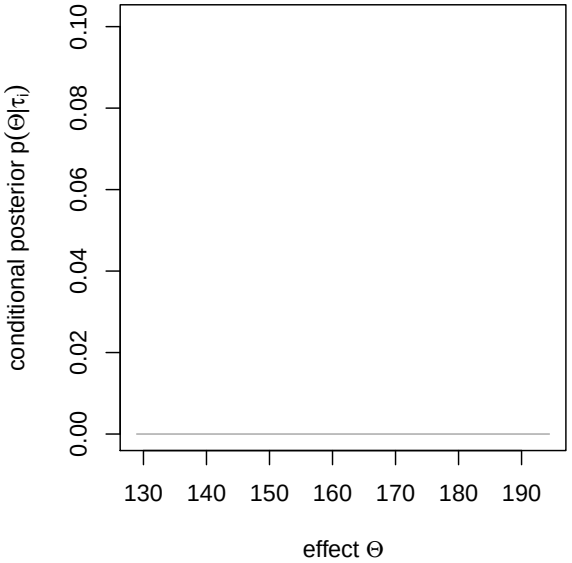


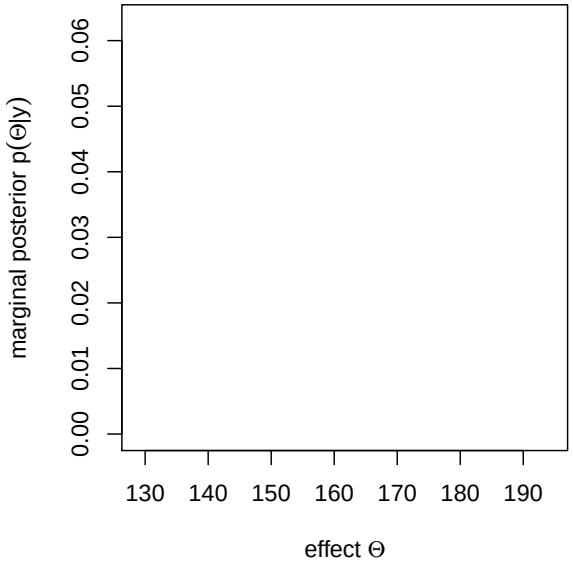
Meta analysis example

Motivation: two-parameter model, conditionals & marginals



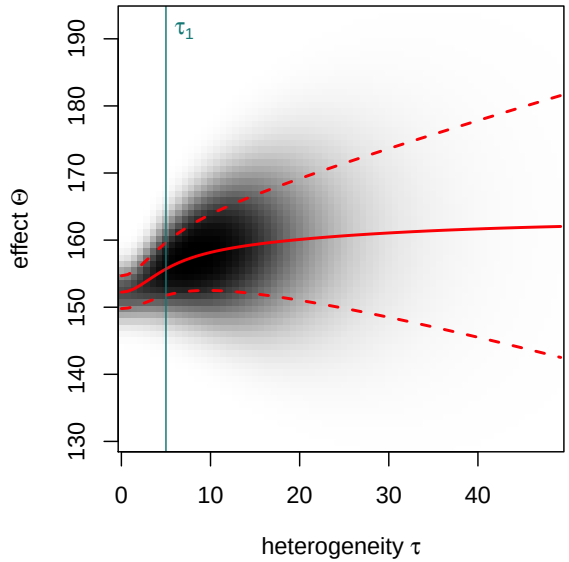


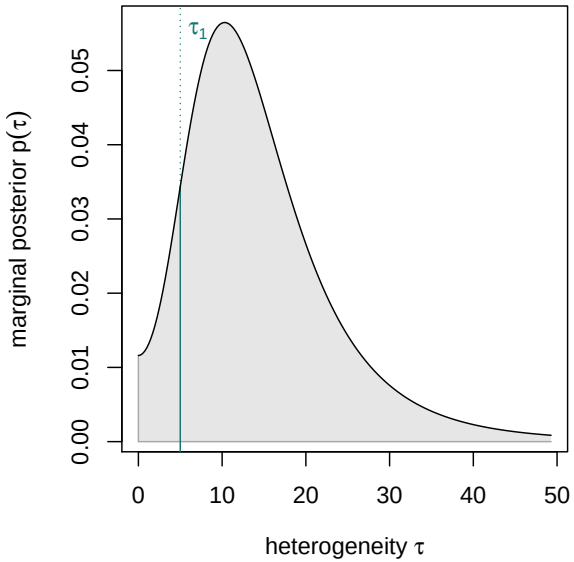


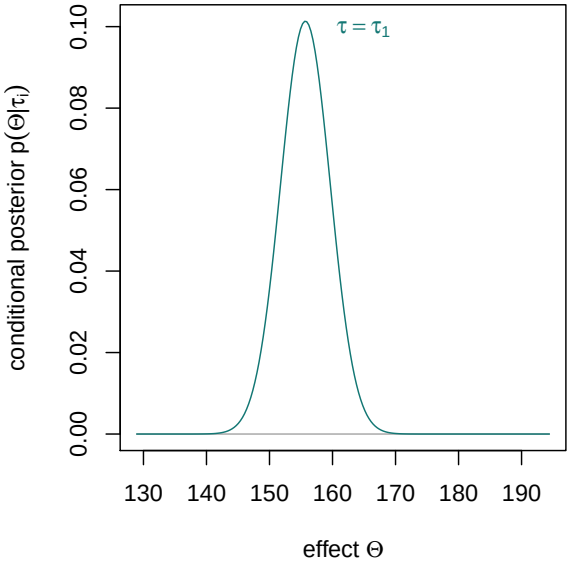


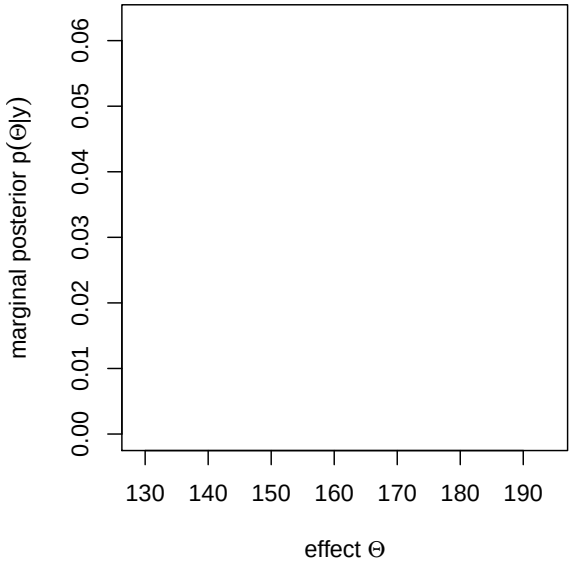
Meta analysis example

Motivation: two-parameter model, conditionals & marginals



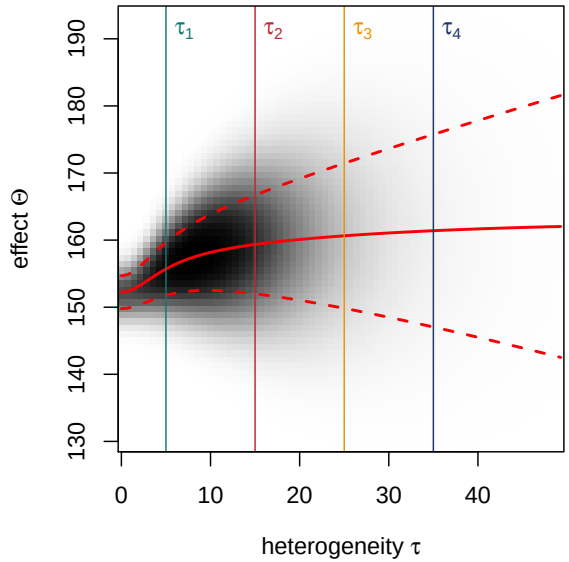


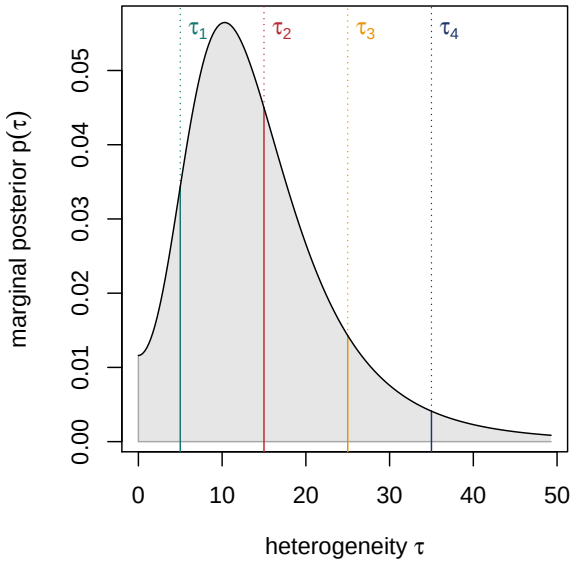


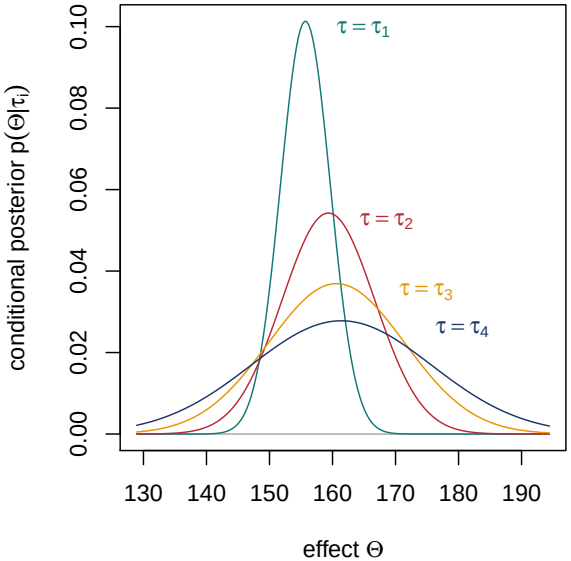


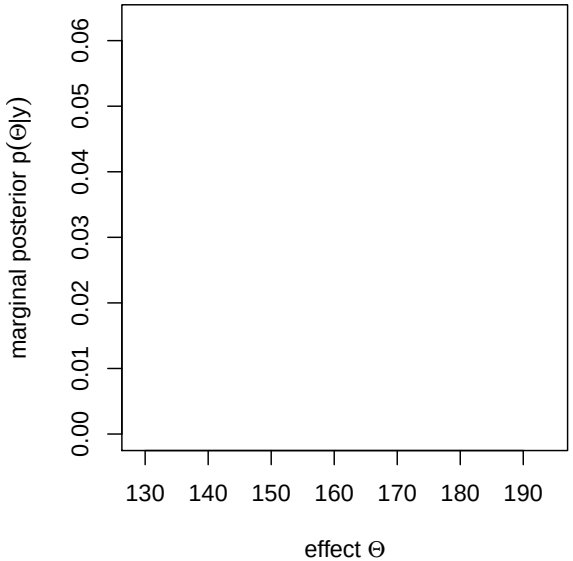
Meta analysis example

Motivation: two-parameter model, conditionals & marginals



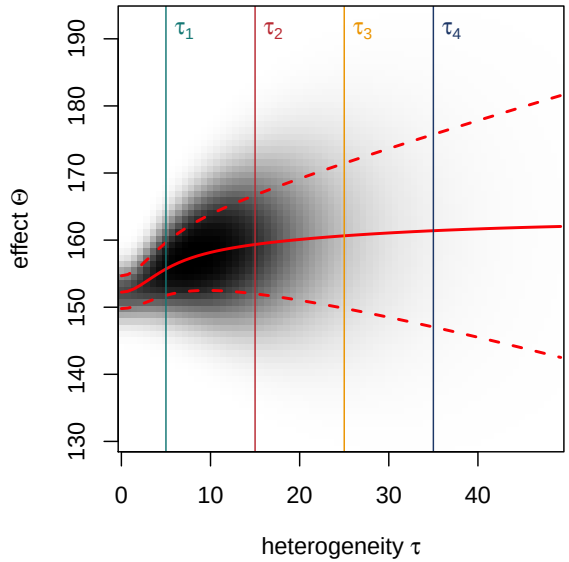


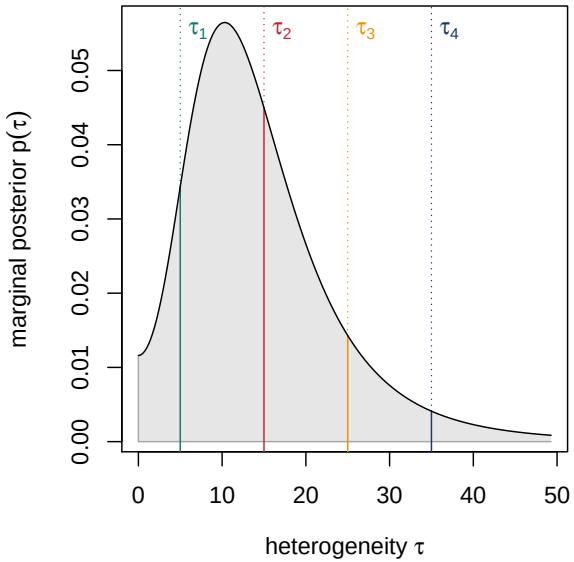


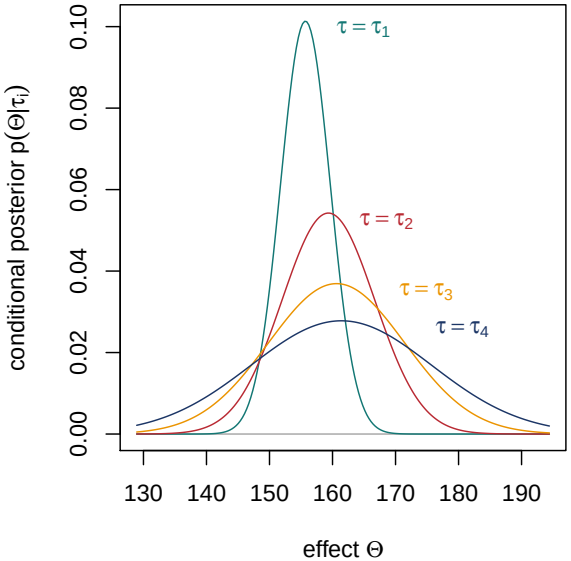


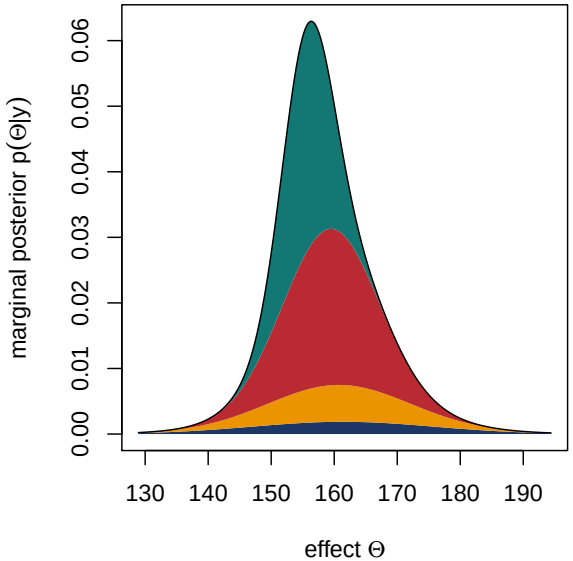
Meta analysis example

Motivation: two-parameter model, conditionals & marginals









Meta analysis example

Questions

- approximating the **continuous** mixture through a **discrete** set of points in $\tau \ldots$
- actual marginal:

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

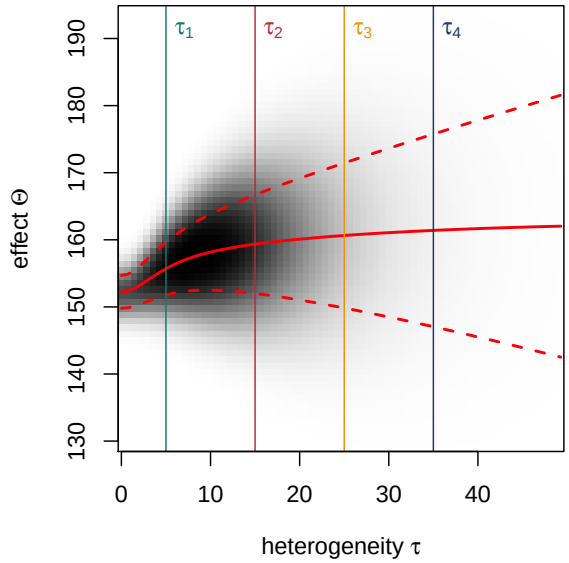
- approximation:

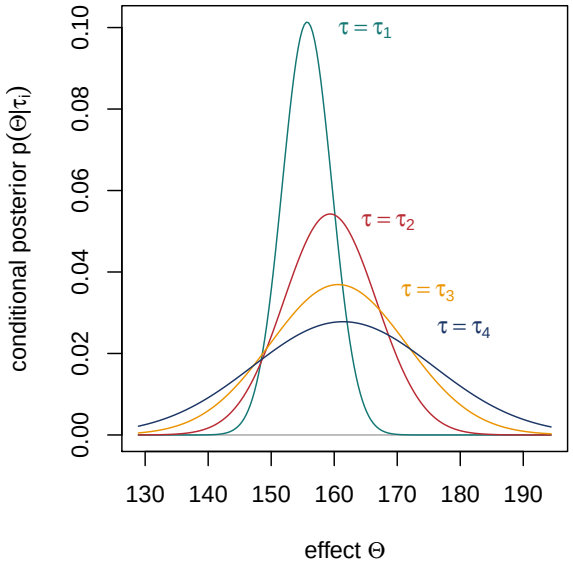
$$p(\Theta) \approx \sum_i p(\Theta|\tau_i)\, \pi_i$$

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

Meta analysis example

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?

Discretizing mixture distributions

Terminology and examples

- random variables X, Y
- joint density $p(x, y) = p(y|x) \times p(x)$
- marginal density $p(y) = \int p(y|x)\, p(x) \mathrm{d}x$
- $p(y)$ “mixture distribution”
- $p(x)$ “mixing distribution”

- Examples:
  - $Y|\lambda \sim \text{Poisson}(\lambda), \quad \lambda \sim \text{Gamma}(\alpha, \sigma)$
    $\Rightarrow \quad Y \sim$ Negative Binomial
  - $Y|p \sim \text{Binomial}(p, N), \quad p \sim \text{Beta}(\alpha, \beta)$
    $\Rightarrow \quad Y \sim$ Beta-Binomial
  - $Y|\sigma \sim \text{Normal}(0, \sigma), \quad \sigma = \sqrt{\frac{\nu}{X}}, \quad X \sim \chi^2_\nu$
    $\Rightarrow \quad Y \sim$ Student-t
  - ...

Discretizing mixture distributions

Setting up a binning

- 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

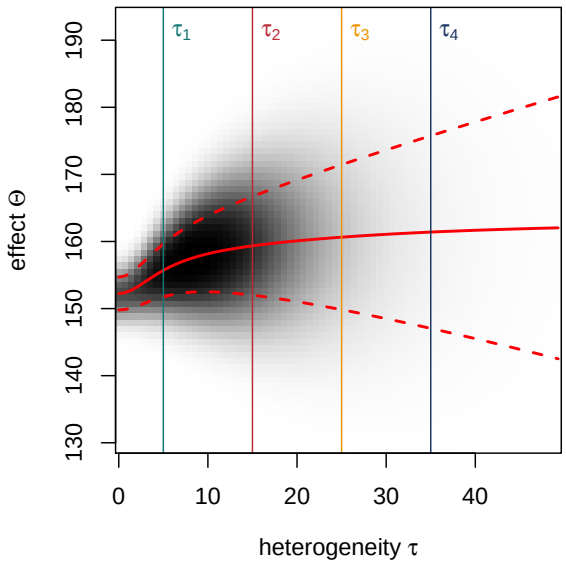
- actual distribution: $p(x, y)$
- discrete approximation: $q(x, y)$
- same marginal (mixing distribution): $q(x) = p(x)$
- but "binned" conditionals:
  $q(y|x) = p(y|x = \tilde{x}_i)$ for $x \in \mathcal{X}_i$.
- q similar to p,
  instead of conditioning on "exact" x,
  conditioning on corresponding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

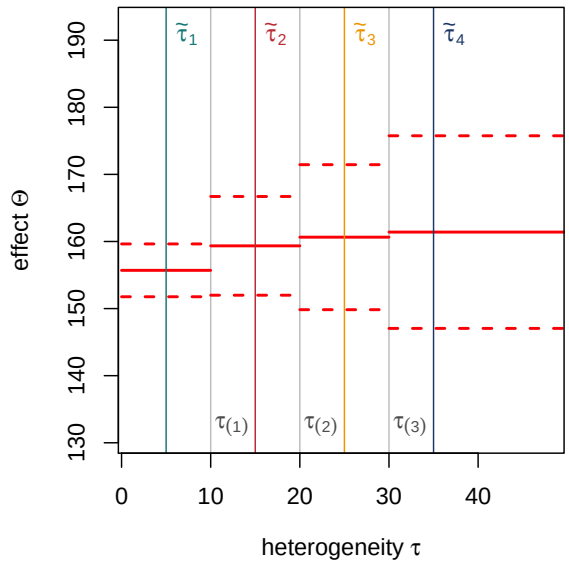
$$\begin{aligned} q(y) &= \int q(y|x)\, q(x)\, \mathrm{d}x \\ &= \sum_i \pi_i\, p(y|\tilde{x}_i) \end{aligned}$$

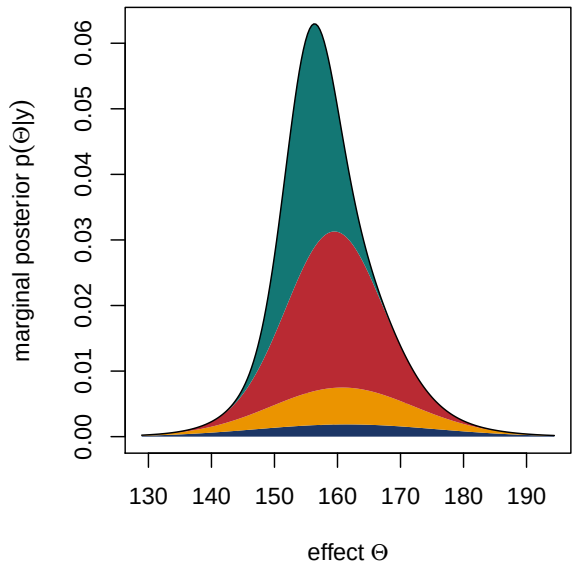
Discretizing mixture distributions

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

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}\big(p(\theta)\big\|q(\theta)\big) &= \int_{\Theta} \log\Big(\frac{p(\theta)}{q(\theta)}\Big)\, p(\theta)\, \mathrm{d}\theta \\ &= \mathrm{E}_{p(\theta)}\left[\log\Big(\frac{p(\theta)}{q(\theta)}\Big)\right] \end{aligned}$$

- the KL-divergence
  - is always positive: $\mathcal{D}_{\mathrm{KL}}\big(p(\theta)\big\|q(\theta)\big) \geq 0$
  - is not symmetric: $\mathcal{D}_{\mathrm{KL}}\big(p(\theta)\big\|q(\theta)\big) \neq \mathcal{D}_{\mathrm{KL}}\big(q(\theta)\big\|p(\theta)\big)$

Similarity / dissimilarity of distributions

Symmetrized KL-divergence

- 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 obviously symmetric: $\mathcal{D}_{\mathrm{s}}\big(p(\theta)\big\|q(\theta)\big) \;=\; \mathcal{D}_{\mathrm{s}}\big(q(\theta)\big\|p(\theta)\big)$
  - bounds the individual directed divergences:
    $\mathcal{D}_{\mathrm{s}}\big(p(\theta)\big\|q(\theta)\big) \;\geq\; \max\big\{\mathcal{D}_{\mathrm{KL}}\big(p(\theta)\big\|q(\theta)\big),\, \mathcal{D}_{\mathrm{KL}}\big(q(\theta)\big\|p(\theta)\big)\big\}$

Divergence

Interpretation

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

Divergence

Interpretation

- Divergence for two normals:

$$\mathcal{D}_{\mathrm{s}}\big(p(\theta|\mu_A,\sigma_A)\big\|p(\theta|\mu_B,\sigma_B)\big) \;=\; \frac{(\mu_A-\mu_B)^2}{\big(\frac{1}{2}(\sigma_A^{-2}+\sigma_B^{-2})\big)^{-1}} + \frac{(\sigma_A^2-\sigma_B^2)^2}{2\,\sigma_A^2\,\sigma_B^2}$$

- obvious special cases:
  - equal variances:

$$\sigma_B = \sigma_A, \quad \mu_B = \mu_A + c\sigma_A \qquad \Rightarrow \qquad \mathcal{D}_{\mathrm{s}}\big(p\|q\) \;=\; c^2$$

  - equal means:

$$\mu_B = \mu_A, \quad \sigma_B = (1+c)\sigma_A \qquad \Rightarrow \qquad \mathcal{D}_{\mathrm{s}}\big(p\|q\) \;=\; \tfrac{c^2(c+2)^2}{2(c+1)^2} \;\approx\; 2c^2$$

Divergence

Bin-wise maximum divergence: definition

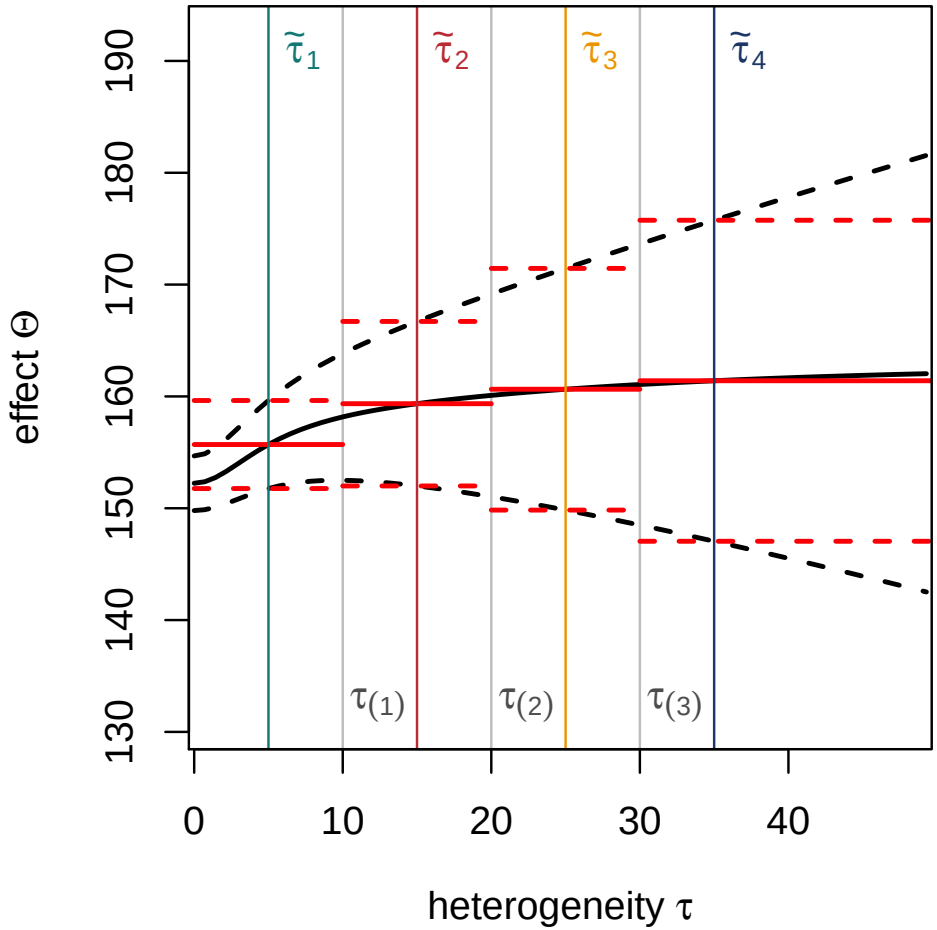
- consider: divergence between reference point and other points *within each bin*
- define:

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

  the **bin-wise maximum divergence**
- "worst-case discrepancy" introduced within each bin
- (note: *symmetrized* divergence $\mathcal{D}_{\mathrm{s}}$)

Divergence

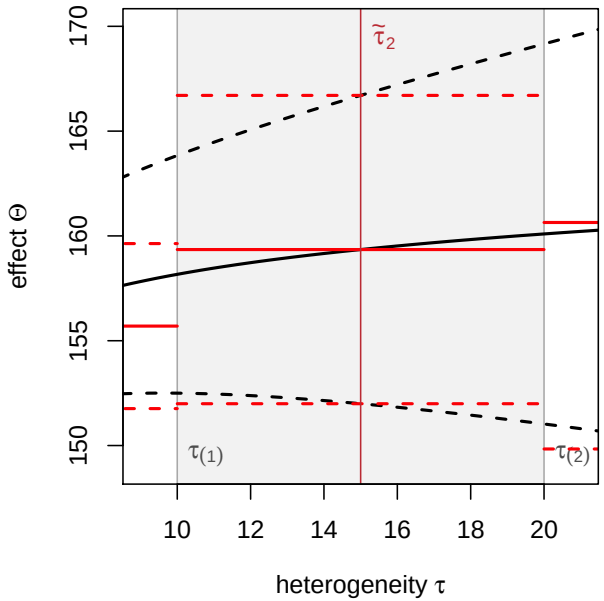
Bin-wise maximum divergence: example

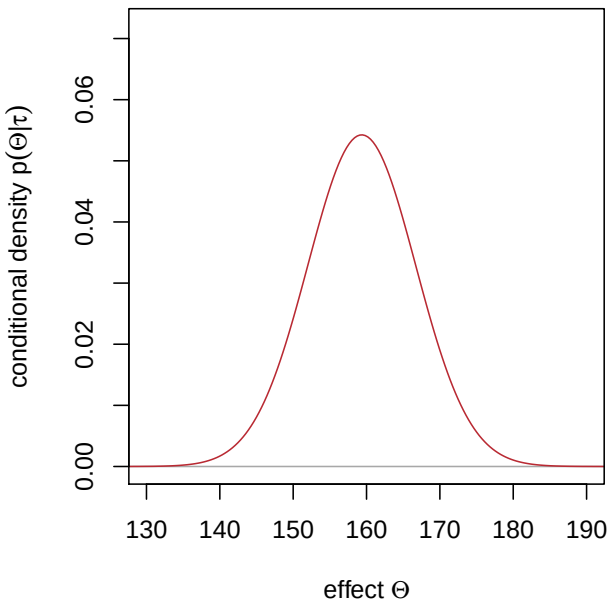


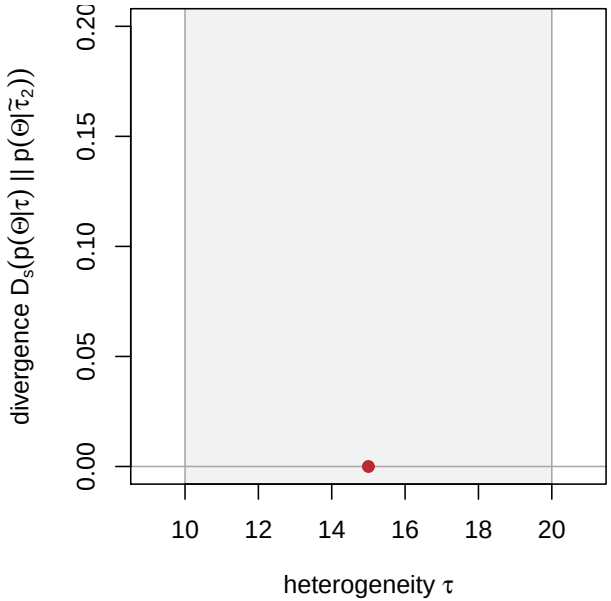
- recall: actual parameters of conditionals $p(y|x)$ (in black) vs. parameters of $q(y|x)$ assumed through binning (in red)

Divergence

Bin-wise maximum divergence: example

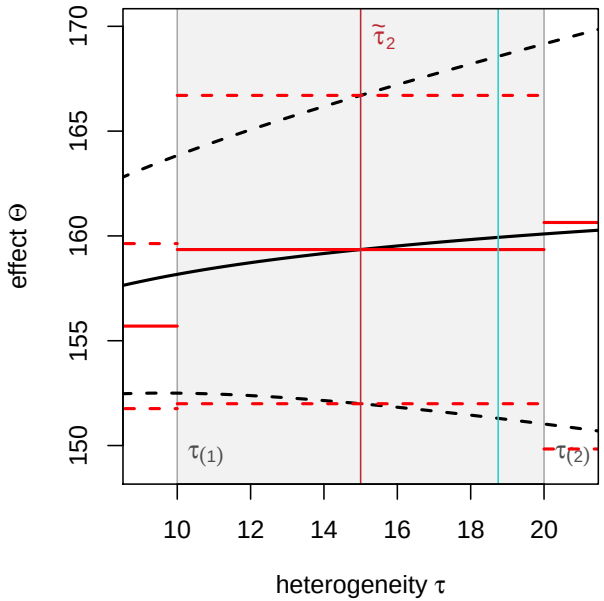


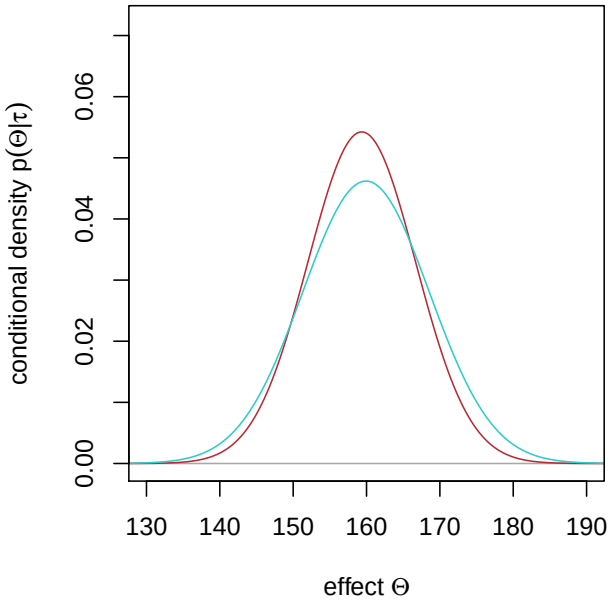


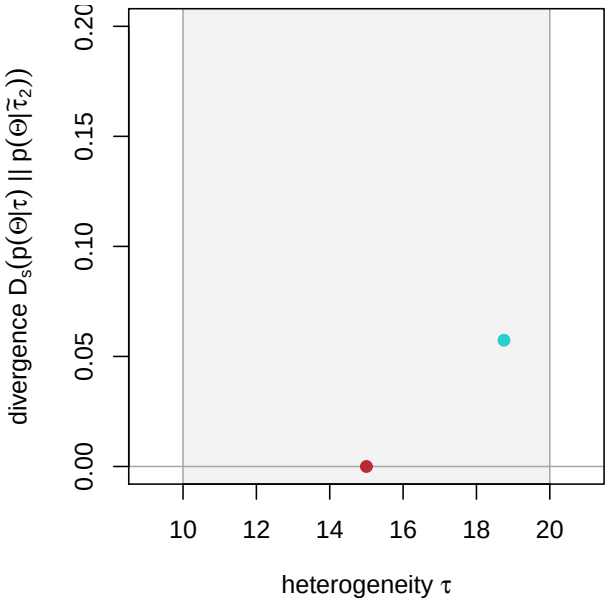


Divergence

Bin-wise maximum divergence: example

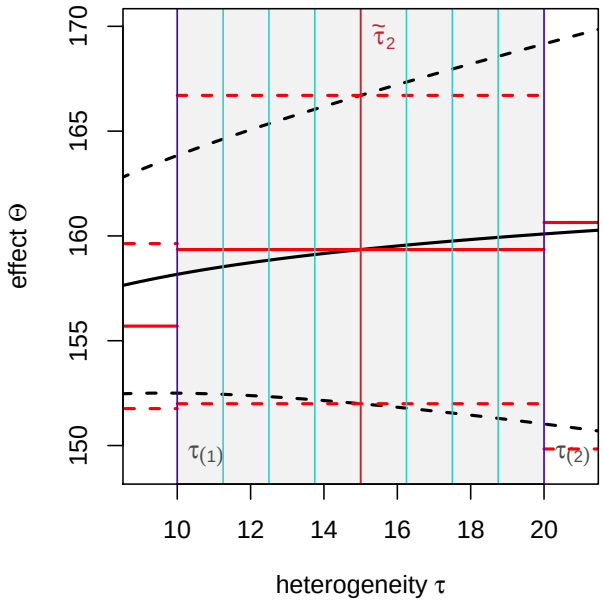


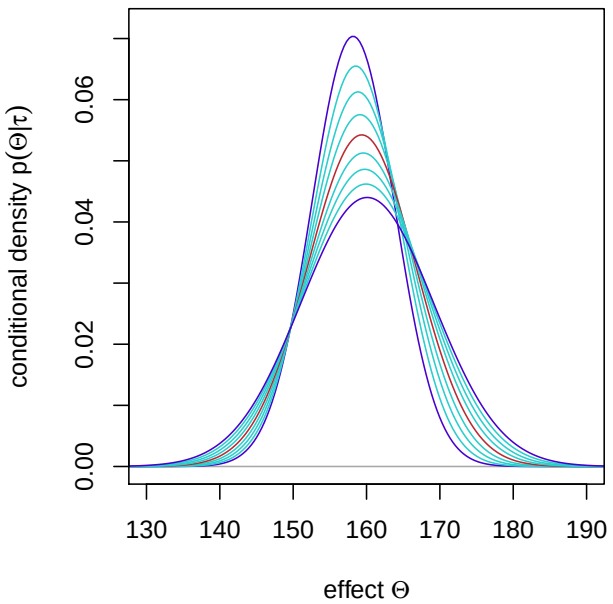


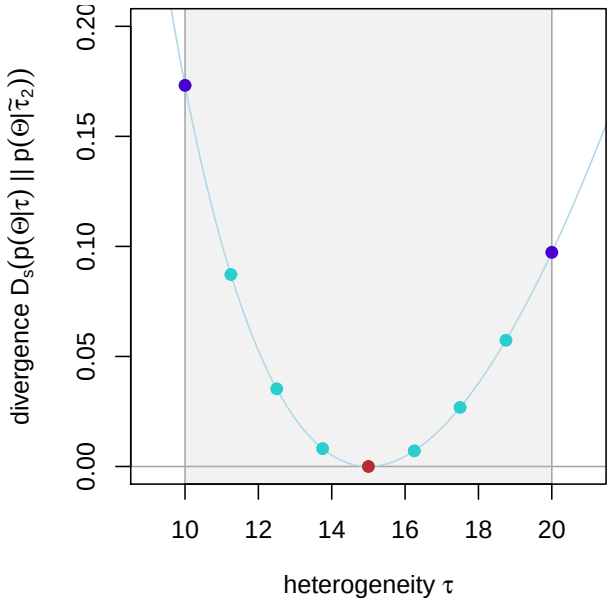


Divergence

Bin-wise maximum divergence: example

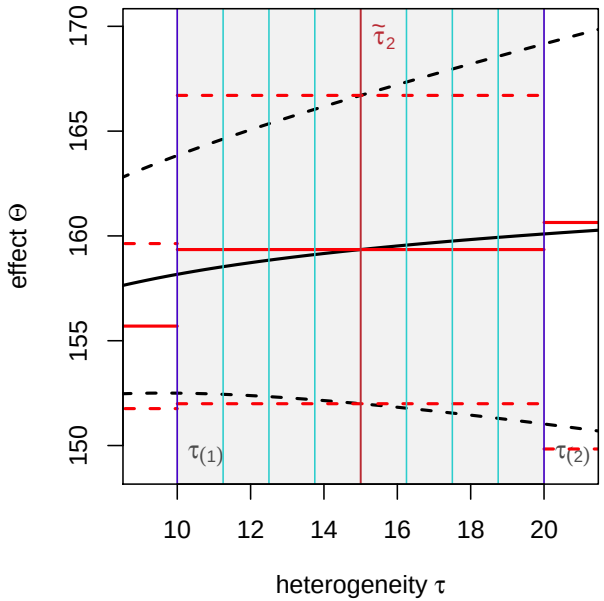


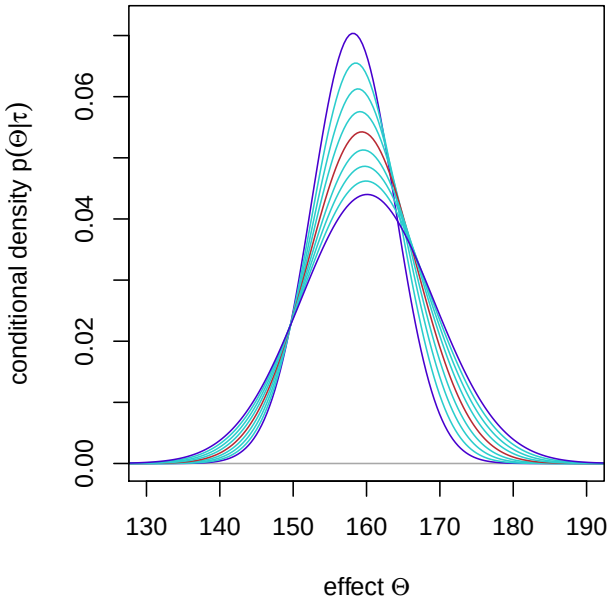


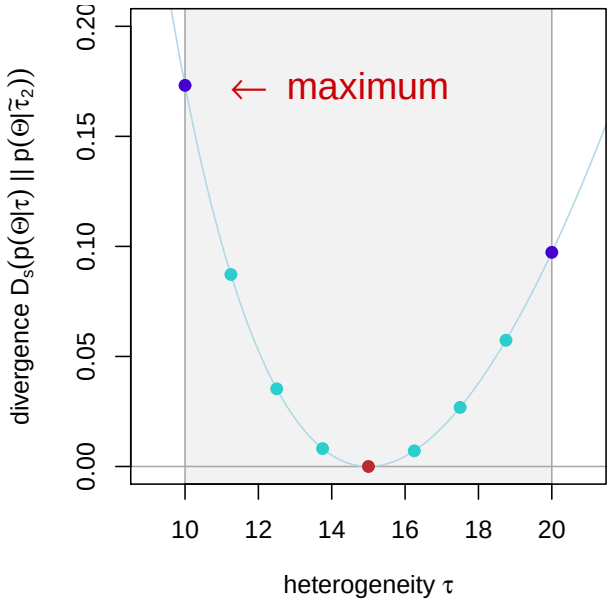


Divergence

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 $\mathcal{D}_s\big(p(y)\big\|q(y)\big)$?
- having the individual *bin-wise* divergences d_i, we can show:

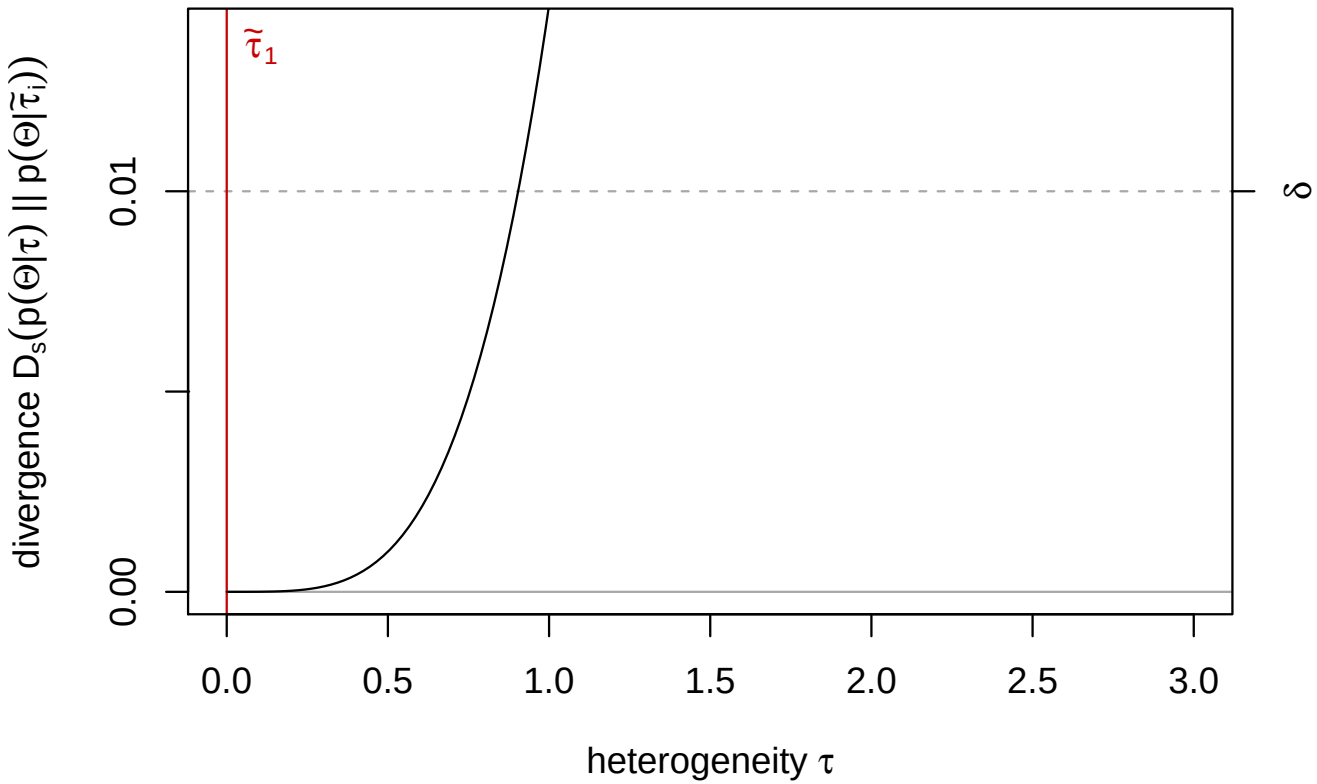
$$\begin{aligned}\mathcal{D}_s\big(p(y)\big\|q(y)\big) &\leq \sum_i \pi_i\, d_i \\ &\leq \max_i d_i\end{aligned}$$

- in other words:

  by bounding bin-wise divergences (of conditionals)
  we can bound the overall divergence (of marginals)

Discretizing mixtures

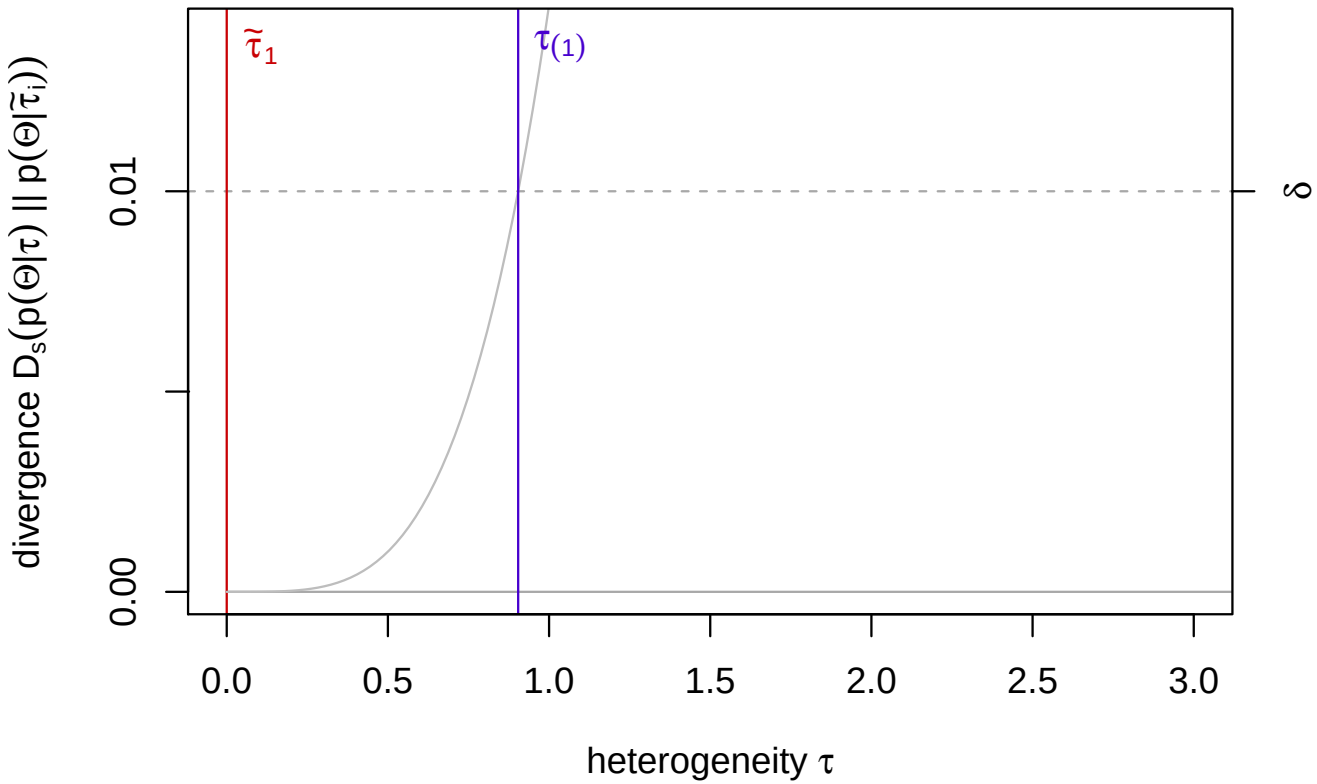
Mixture setup algorithm



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

Discretizing mixtures

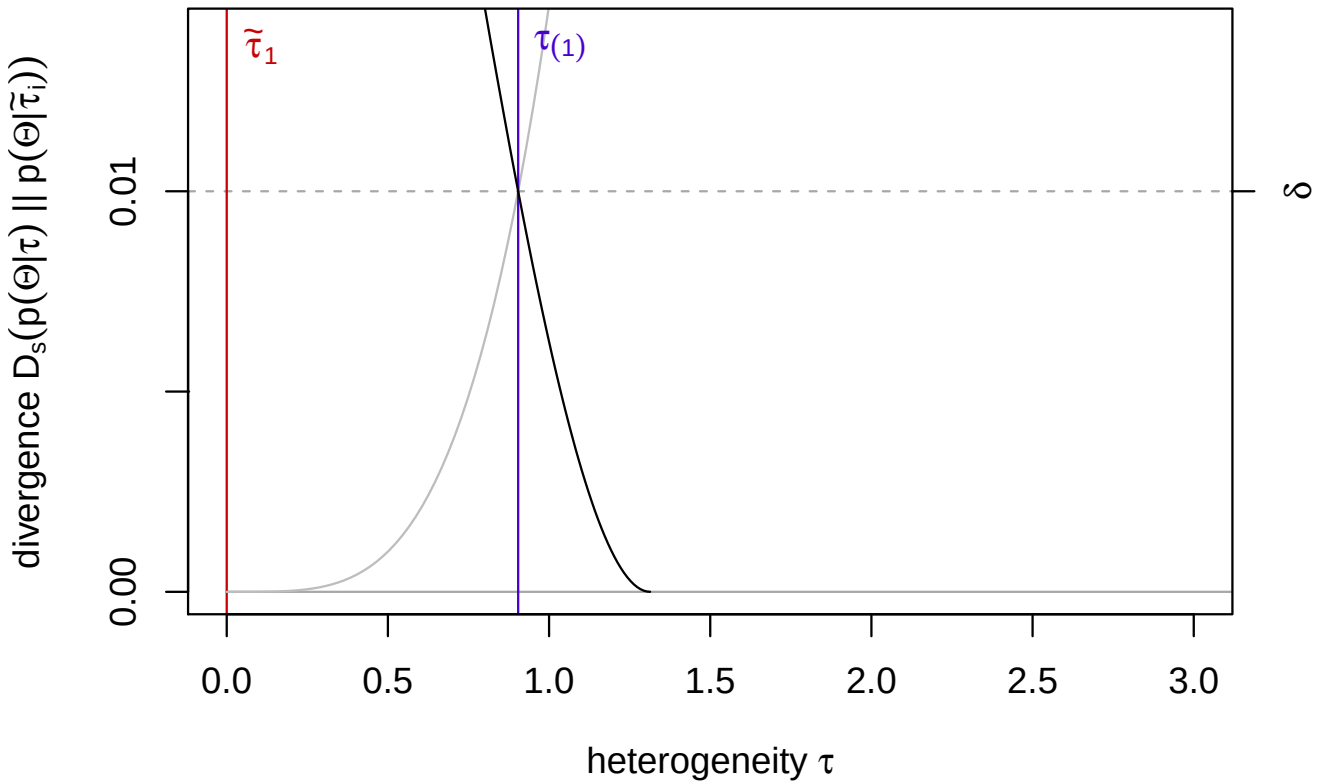
Mixture setup algorithm



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

Discretizing mixtures

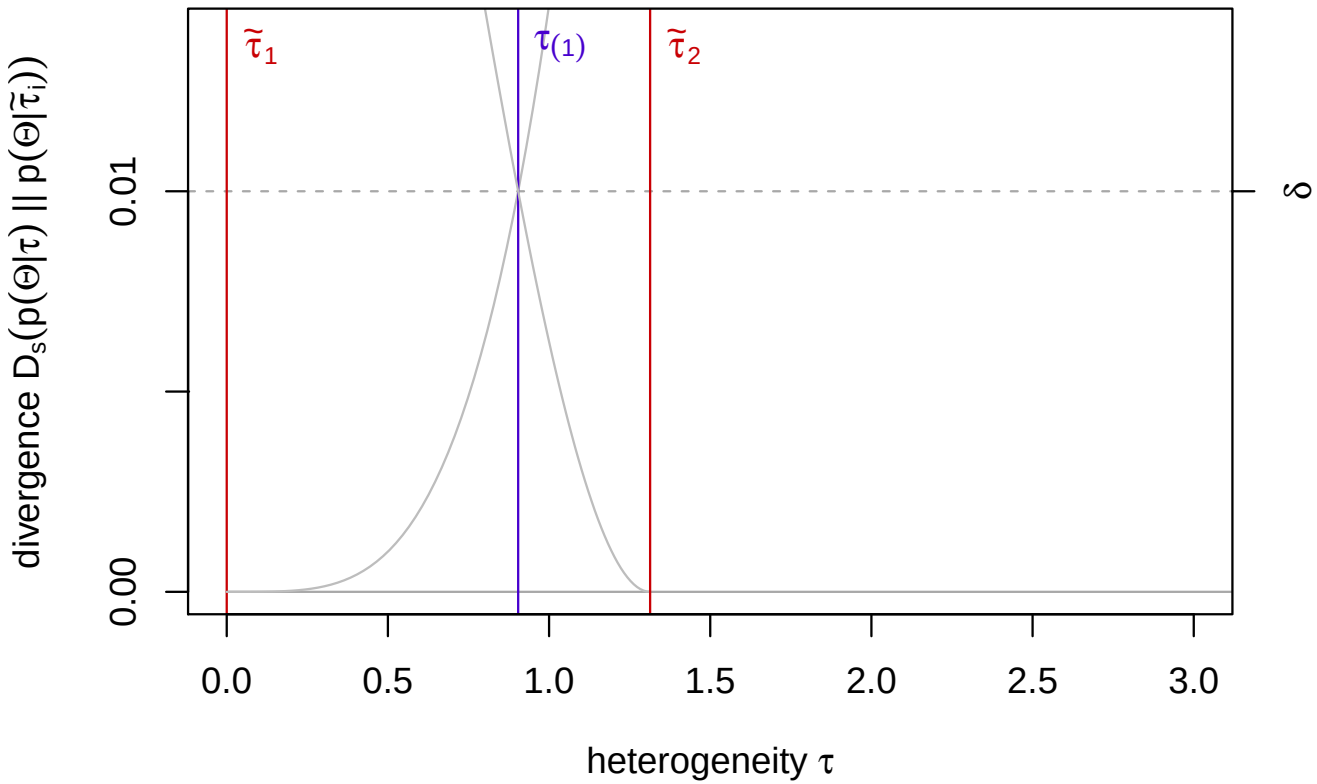
Mixture setup algorithm



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

Discretizing mixtures

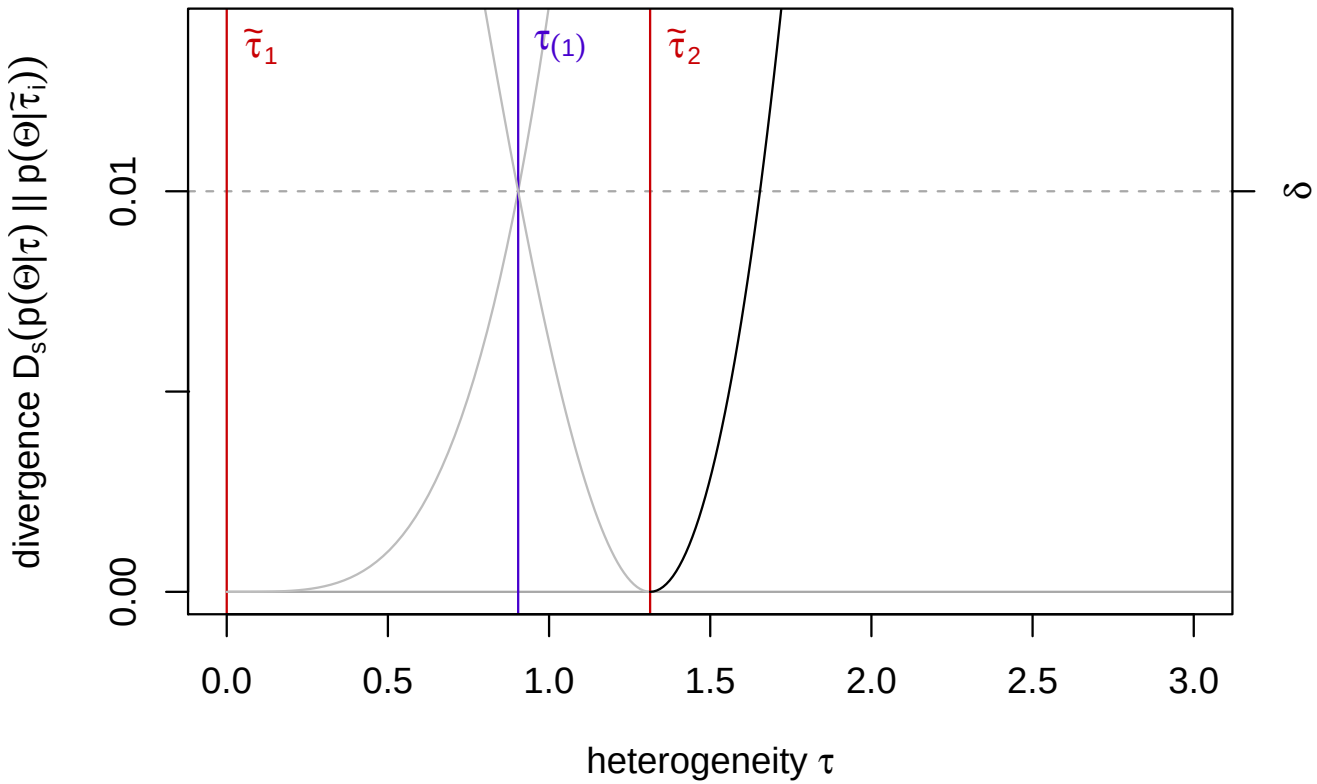
Mixture setup algorithm



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

Discretizing mixtures

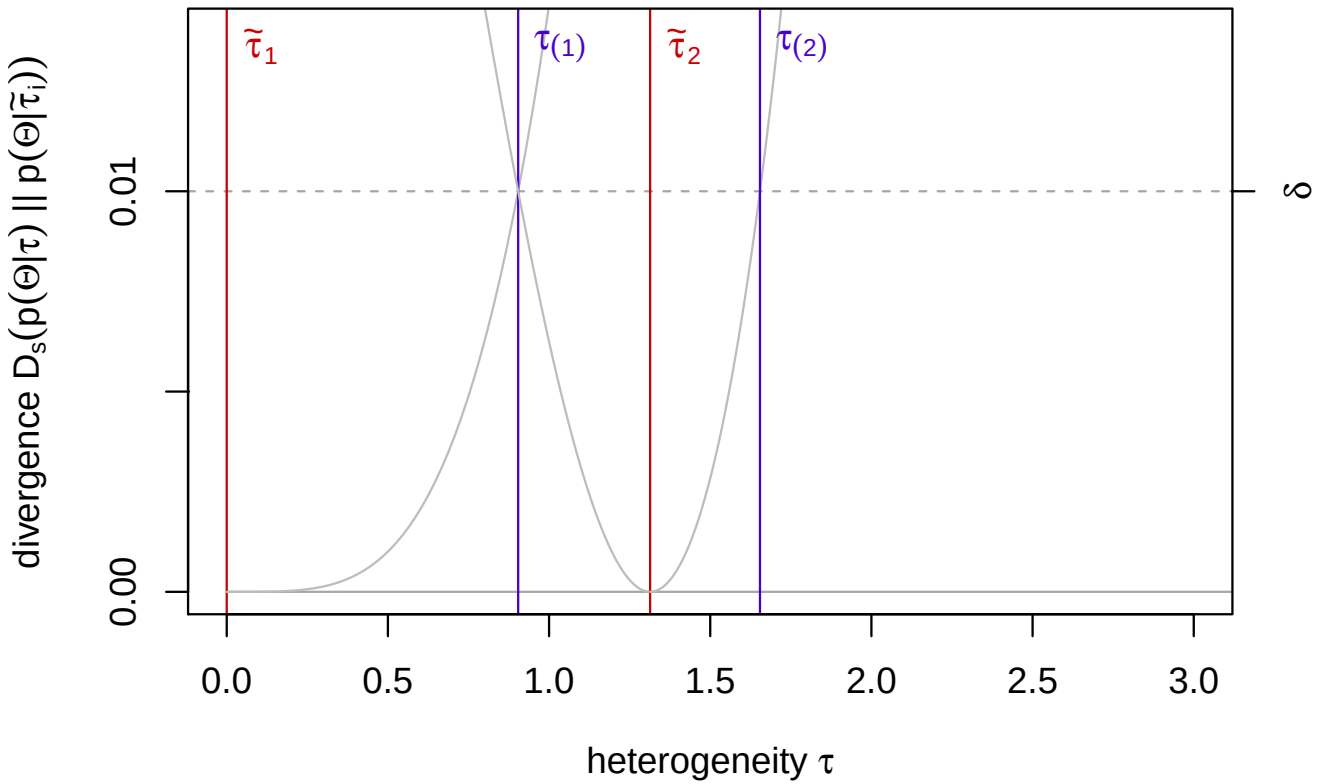
Mixture setup algorithm



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

Discretizing mixtures

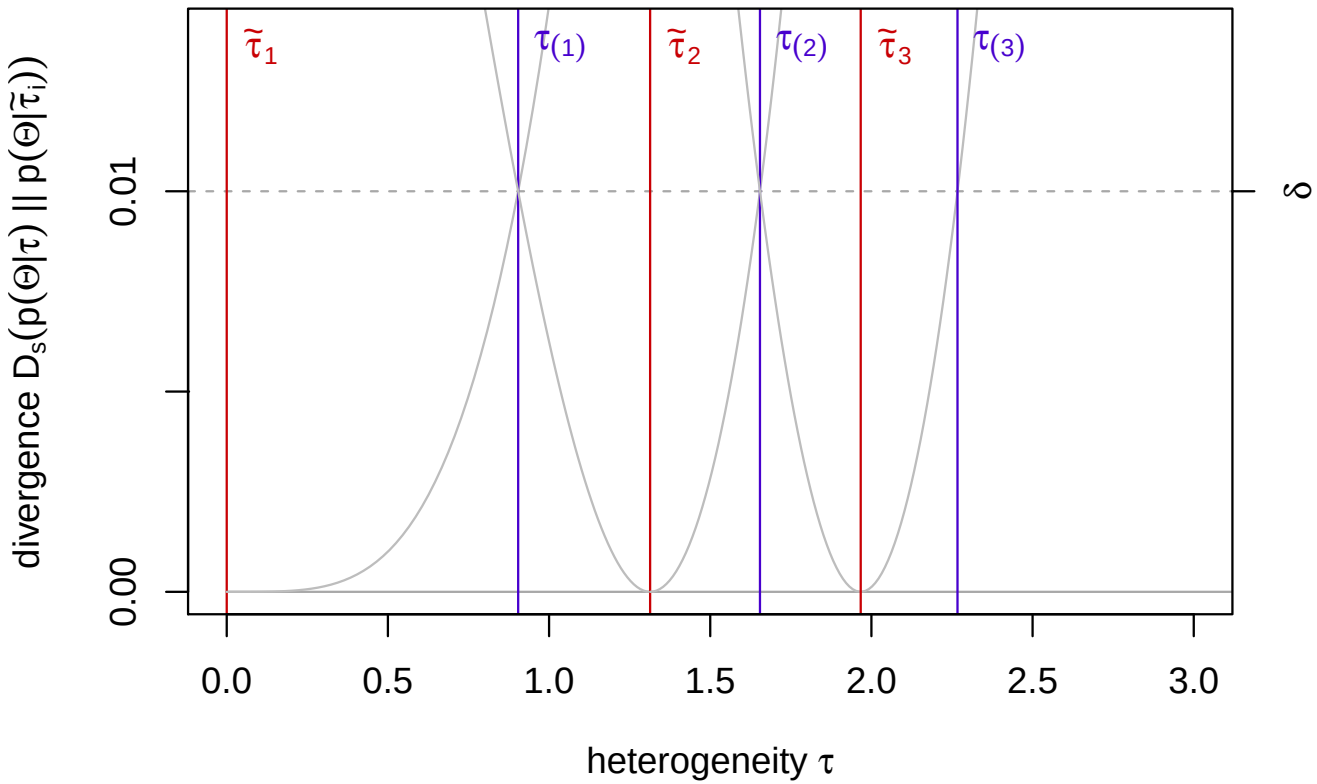
Mixture setup algorithm



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

Discretizing mixtures

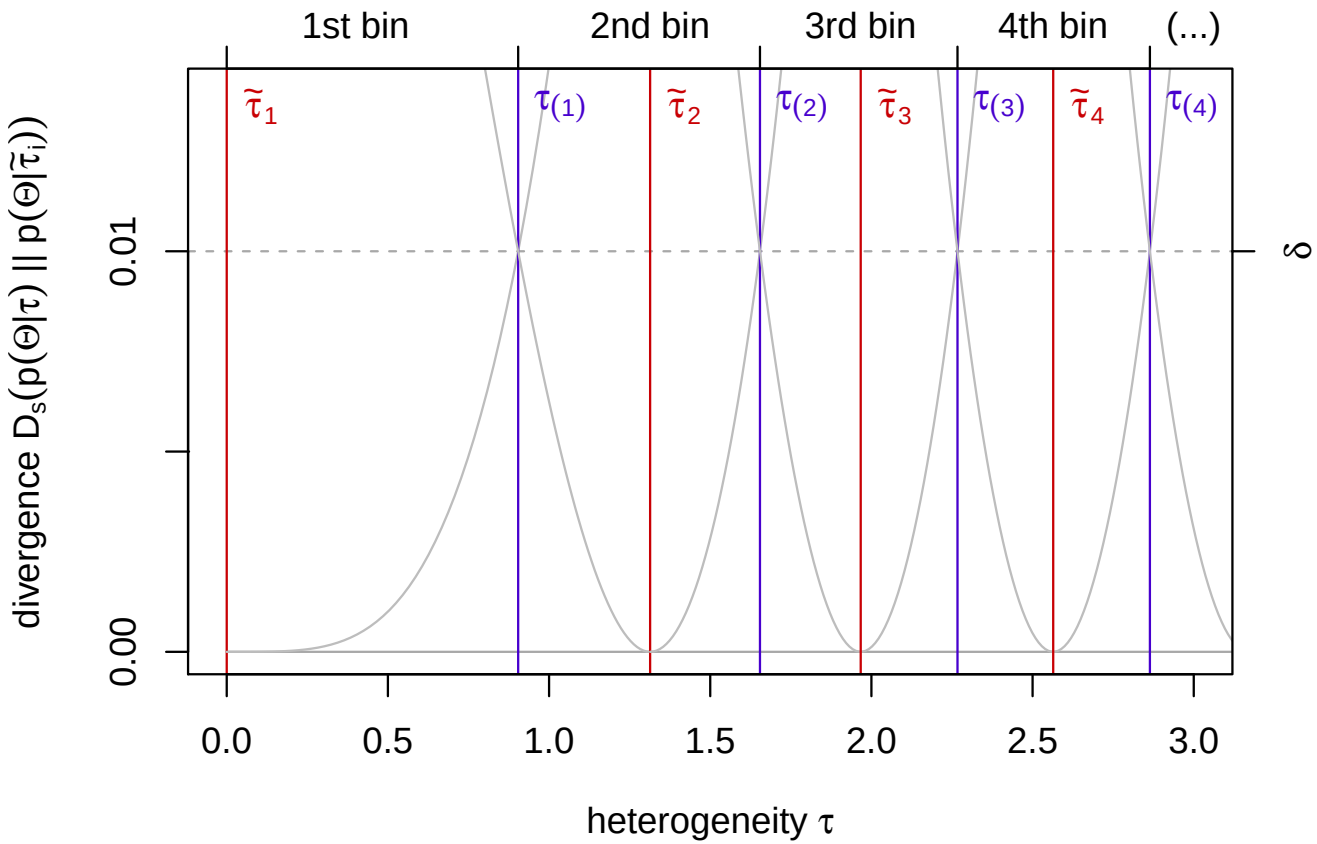
Mixture setup algorithm



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

Discretizing mixtures

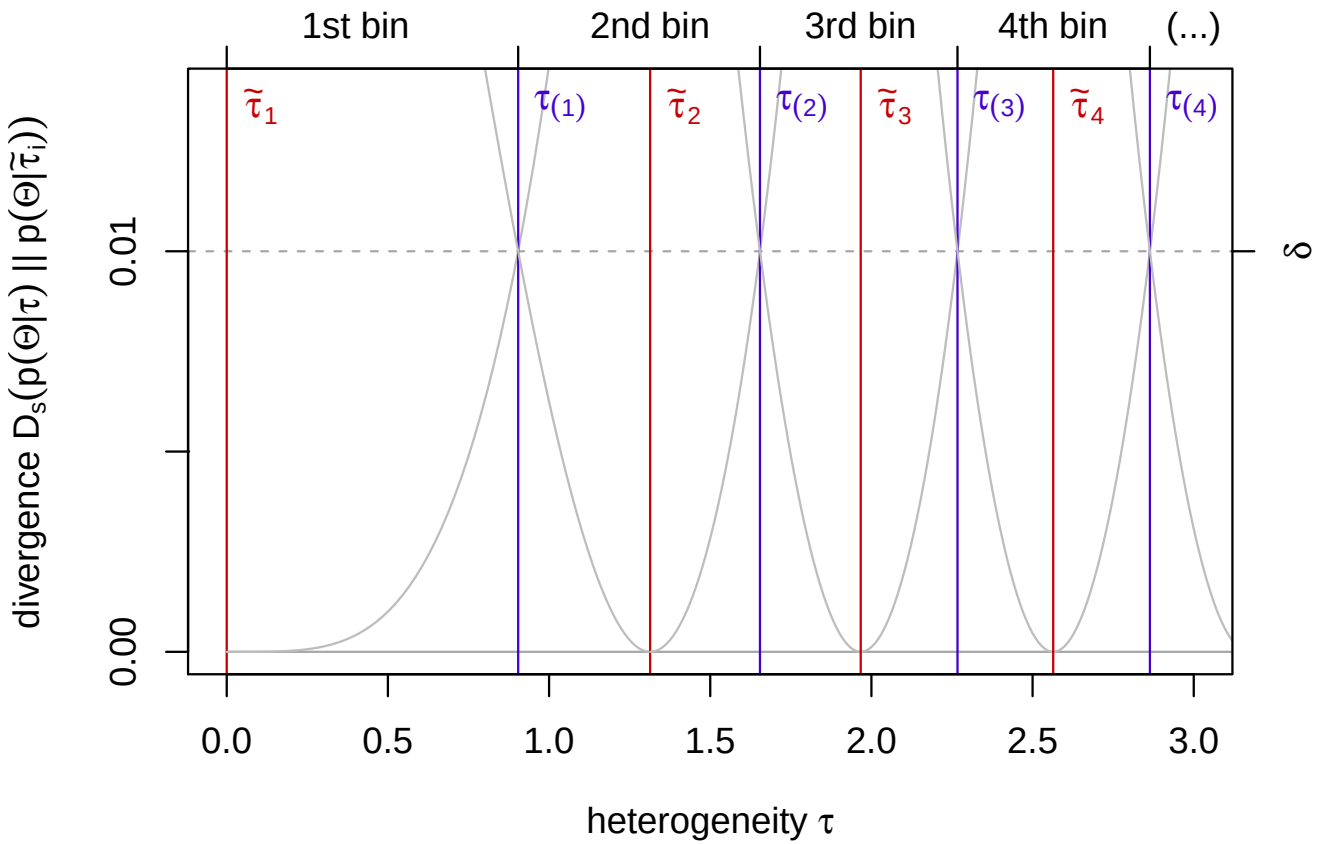
Mixture setup algorithm



- 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

Mixture setup 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*
- (when to stop?)

Discretizing mixtures

General 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, $\mathcal{D}_\mathrm{s}\big(p(y|\tilde{x}_1)\big\|p(y|x^\star)\big) = 0$.
   Now increase $x^\star$ as far as possible while ensuring that $\mathcal{D}_\mathrm{s}\big(p(y|\tilde{x}_1)\big\|p(y|x^\star)\big) \leq \delta$.
   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 $\mathcal{D}_\mathrm{s}\big(p(y|x_{(i-1)})\big\|p(y|x^\star)\big) = \delta$. Use this point as the next reference point: $\tilde{x}_i = x^\star$.
4. Increase $x^\star$ again until $\mathcal{D}_\mathrm{s}\big(p(y|\tilde{x}_i)\big\|p(y|x^\star)\big) = \delta$. 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)
  (is there a way to define a criterion "jointly"?)

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

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

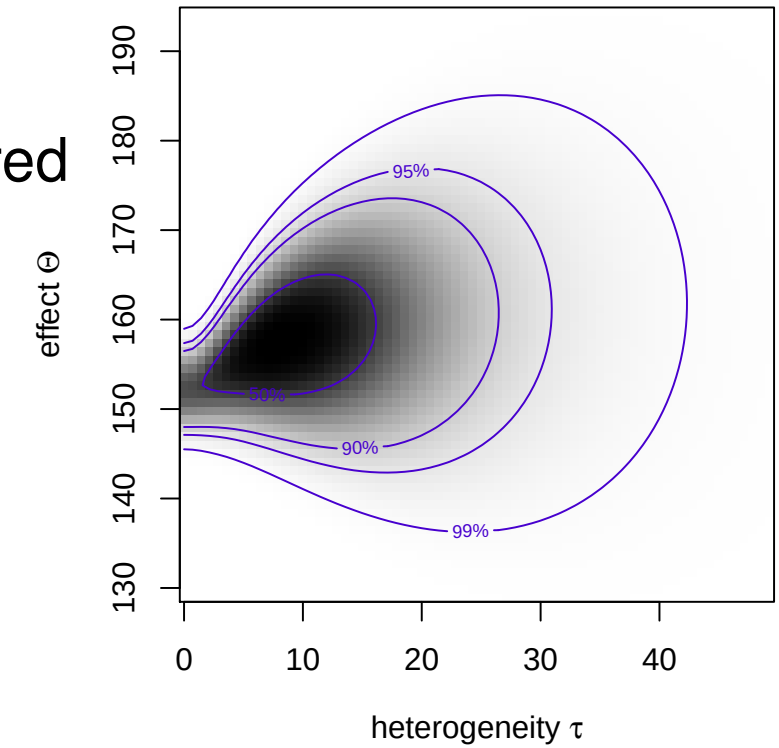
Discretizing mixtures

General algorithm

- remaining issue: ignored $\epsilon > 0$ tail probability
  (usually: problems at domain's margins)
  (is there a way to define a criterion "jointly"?)

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

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



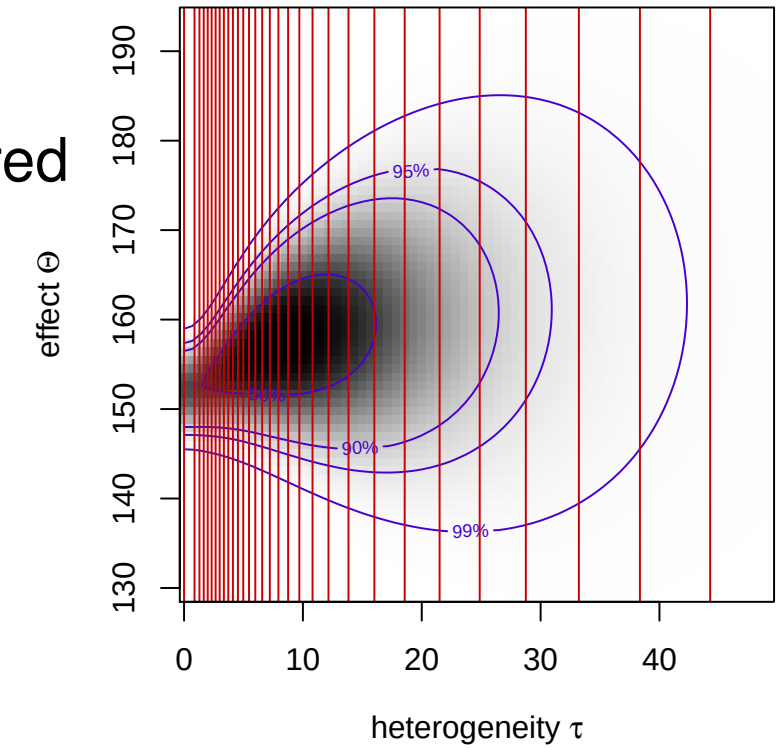
Discretizing mixtures

General algorithm

- remaining issue: ignored $\epsilon > 0$ tail probability
  (usually: problems at domain's margins)
  (is there a way to define a criterion "jointly"?)

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

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



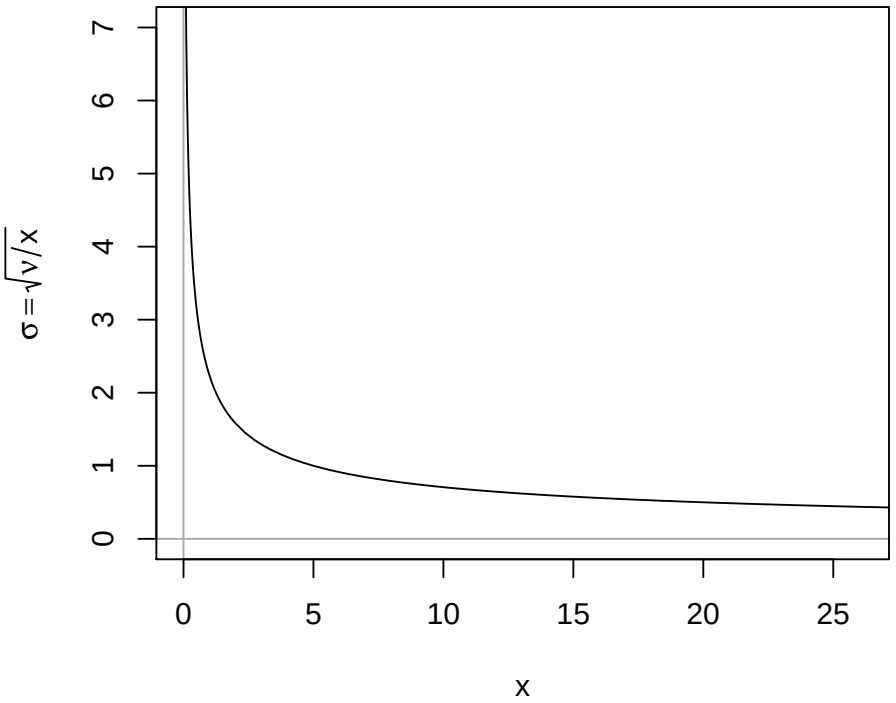
Discretizing mixtures

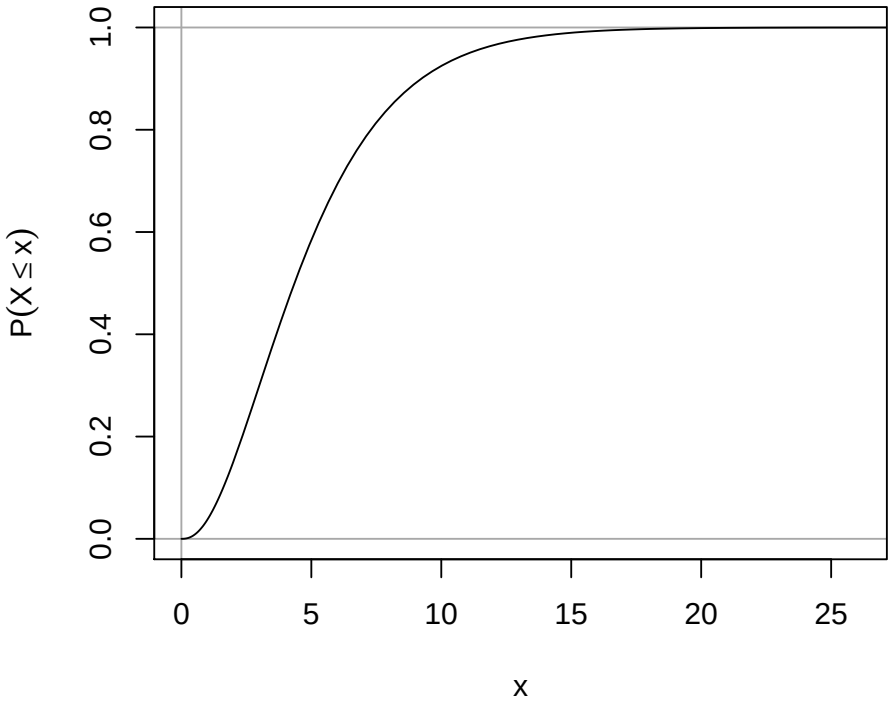
Example application: Student-*t* distribution

- Student-*t* distribution also arises as a mixture distribution:
  - draw X from a χ^2_ν-distribution
  - calculate $\sigma = \sqrt{\frac{\nu}{X}}$
  - draw $Y|\sigma$ from a Normal$(0, \sigma^2)$-distribution
  - marginal of Y is Student-*t* (with ν d.f.)
- set:
  - aimed for divergence: $\delta = 0.01$
  - neglected tail probability: $\epsilon = 0.001$
  - first reference point: $\tilde{x}_1 = \frac{\epsilon}{2}$-quantile of χ^2_ν-distribution
- iterate...

Discretizing mixtures

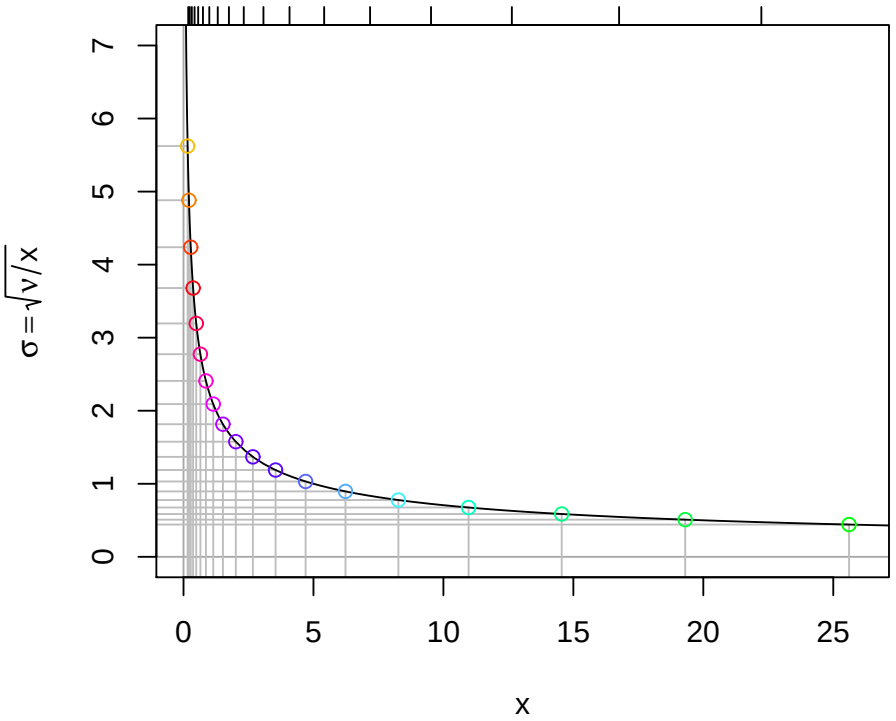
Example application: Student-*t* distribution

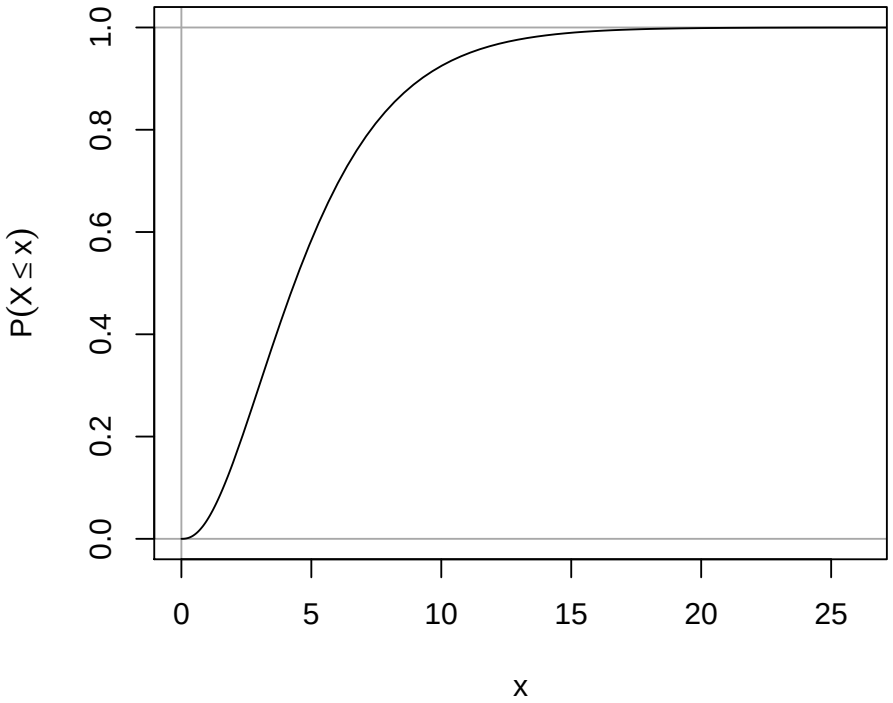




Discretizing mixtures

Example application: Student-*t* distribution

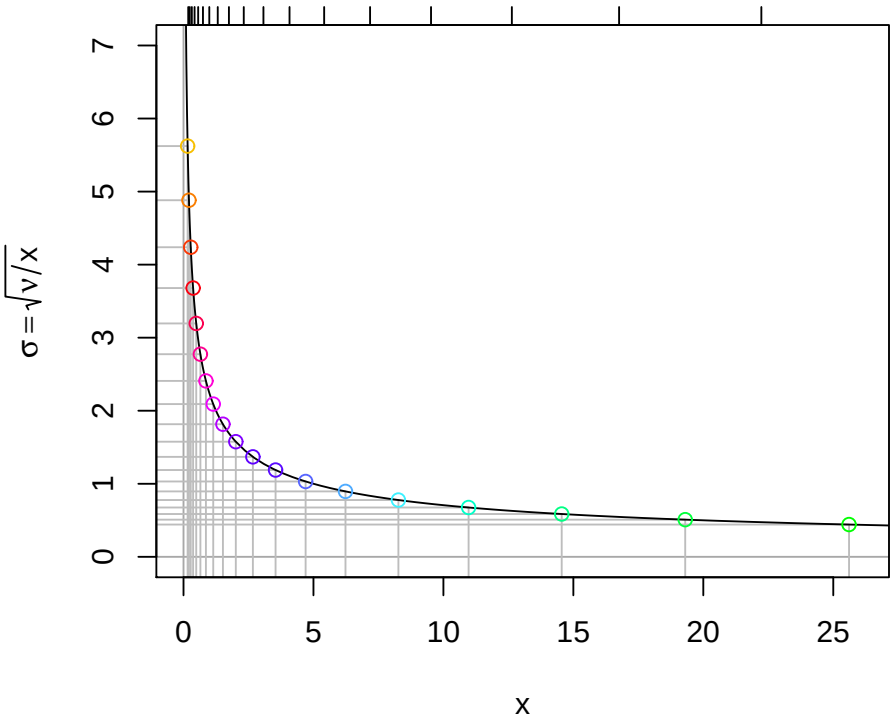


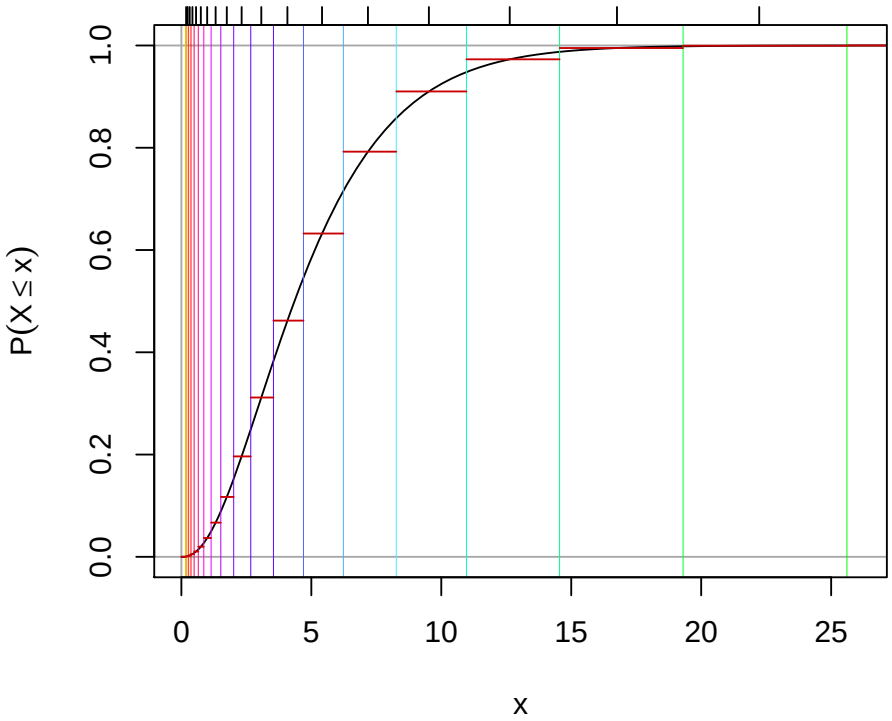


- algorithm yields 19 reference points $\tilde{x}_i$ ($\delta = 0.01$, $\epsilon = 0.001$)

Discretizing mixtures

Example application: Student-*t* distribution

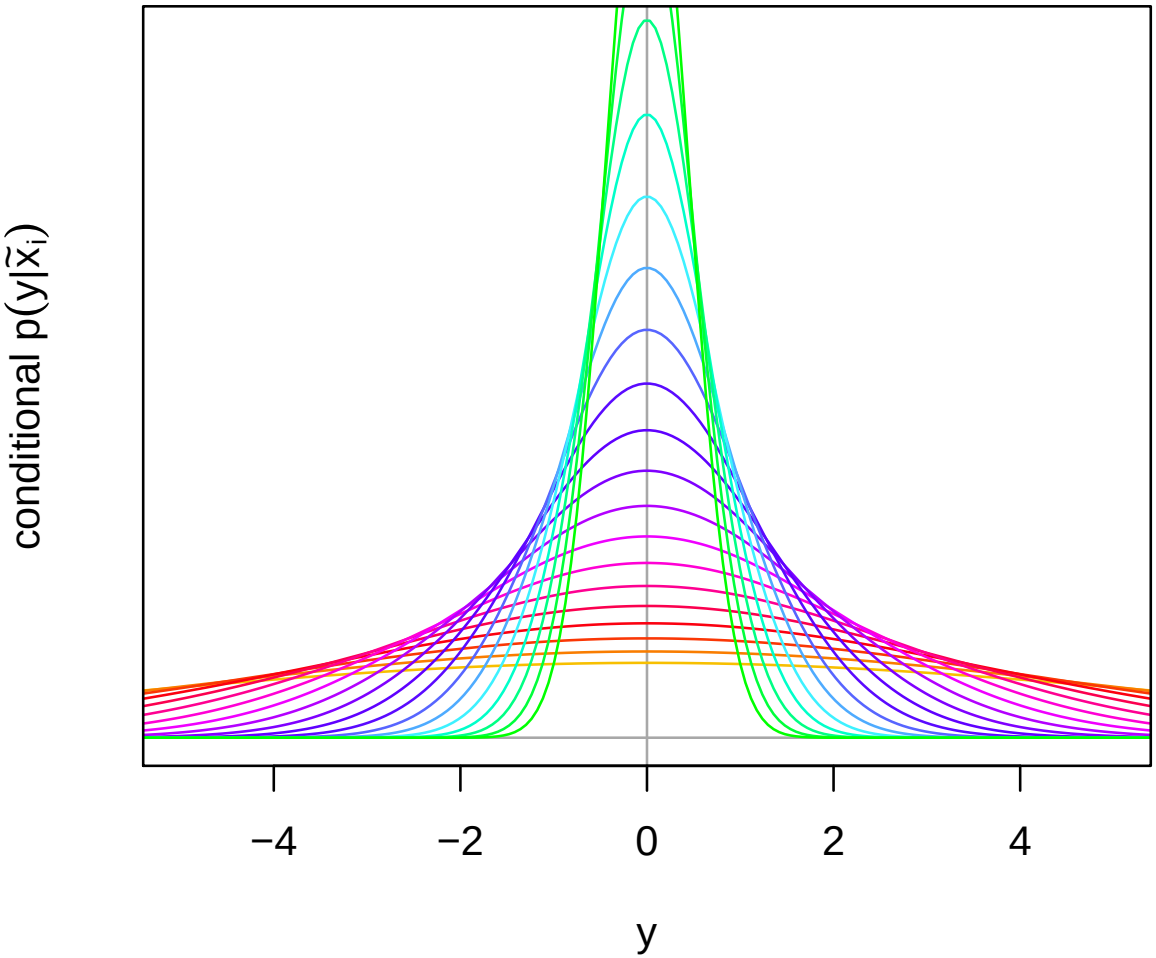




- algorithm yields 19 reference points $\tilde{x}_i$ ($\delta = 0.01$, $\epsilon = 0.001$)

Discretizing mixtures

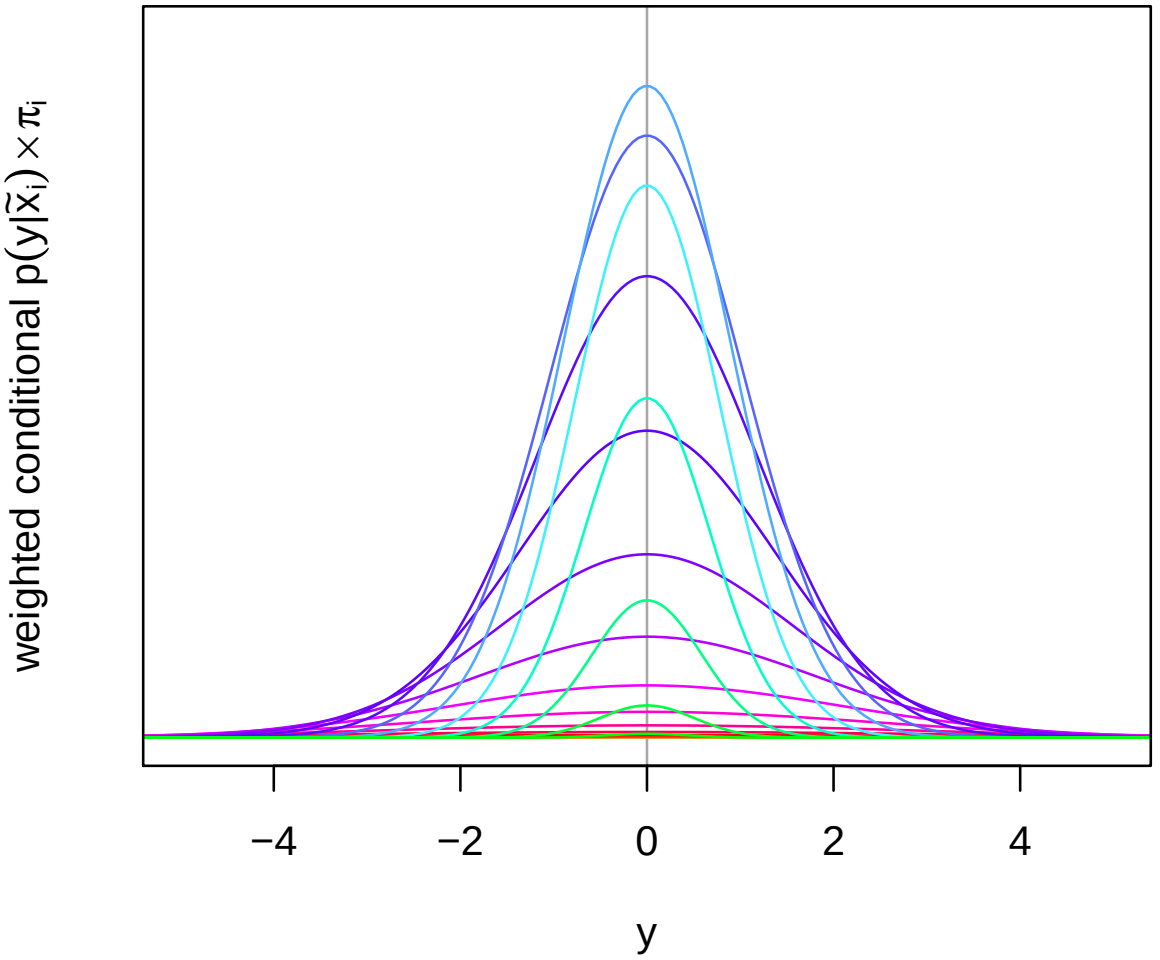
Example application: Student-*t* distribution



- 19 conditionals $p(y|\tilde{x}_i)$

Discretizing mixtures

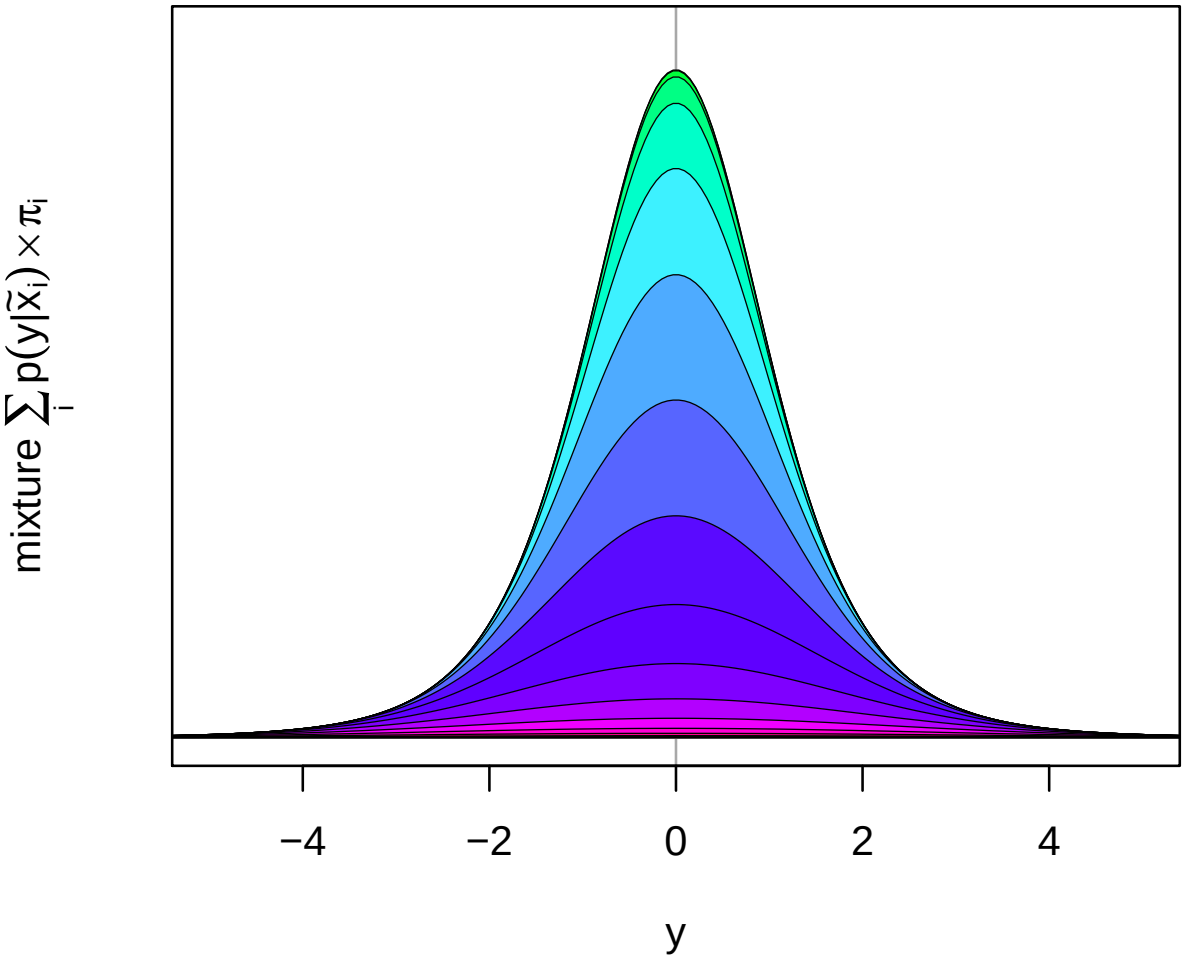
Example application: Student-*t* distribution



- 19 conditionals $p(y|\tilde{x}_i) \rightarrow$ weighted conditionals $p(y|\tilde{x}_i) \times \pi_i$

Discretizing mixtures

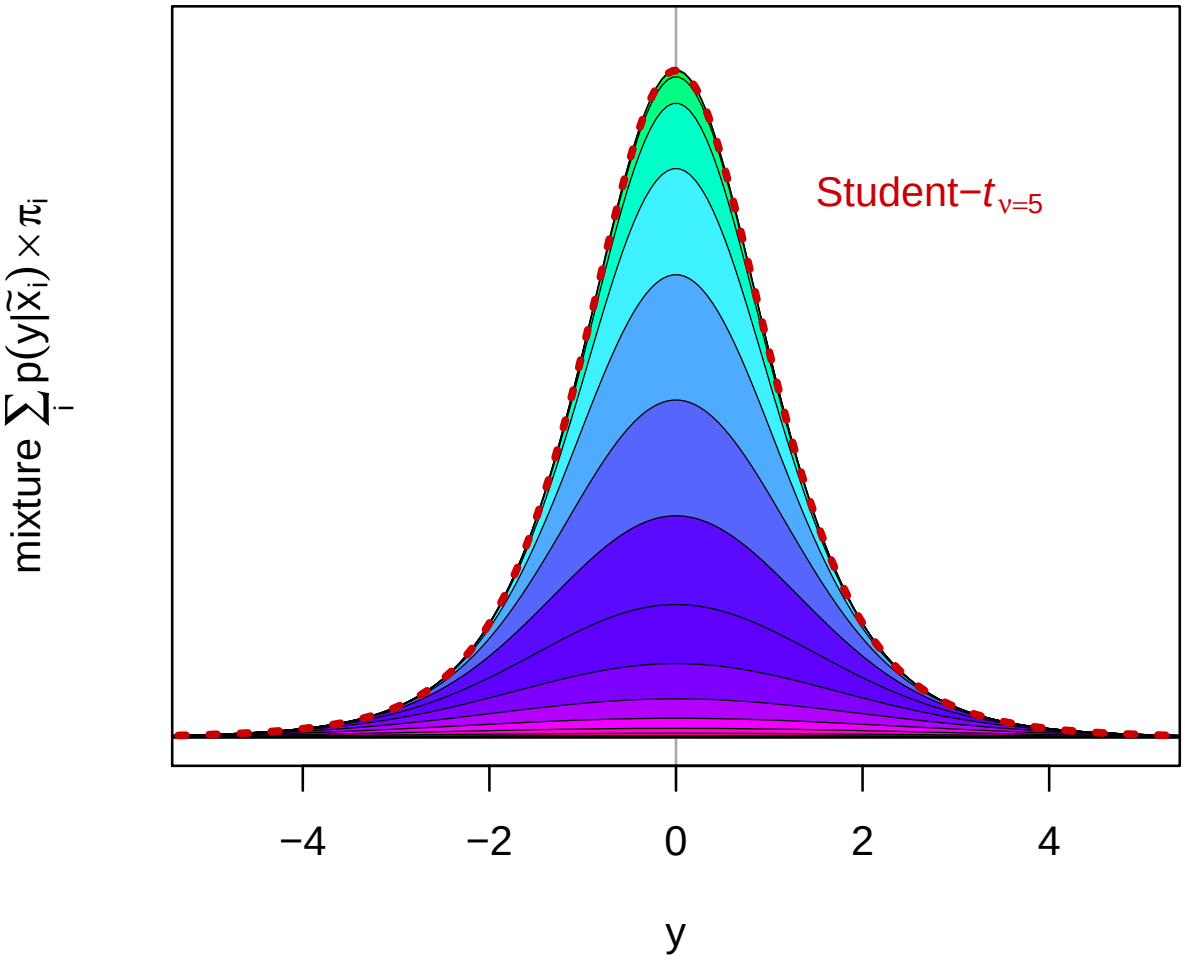
Example application: Student-*t* distribution



- 19 conditionals $p(y|\tilde{x}_i) \rightarrow$ weighted conditionals $p(y|\tilde{x}_i) \times \pi_i$ $\rightarrow$ discrete mixture $\sum_i p(y|\tilde{x}_i) \times \pi_i$

Discretizing mixtures

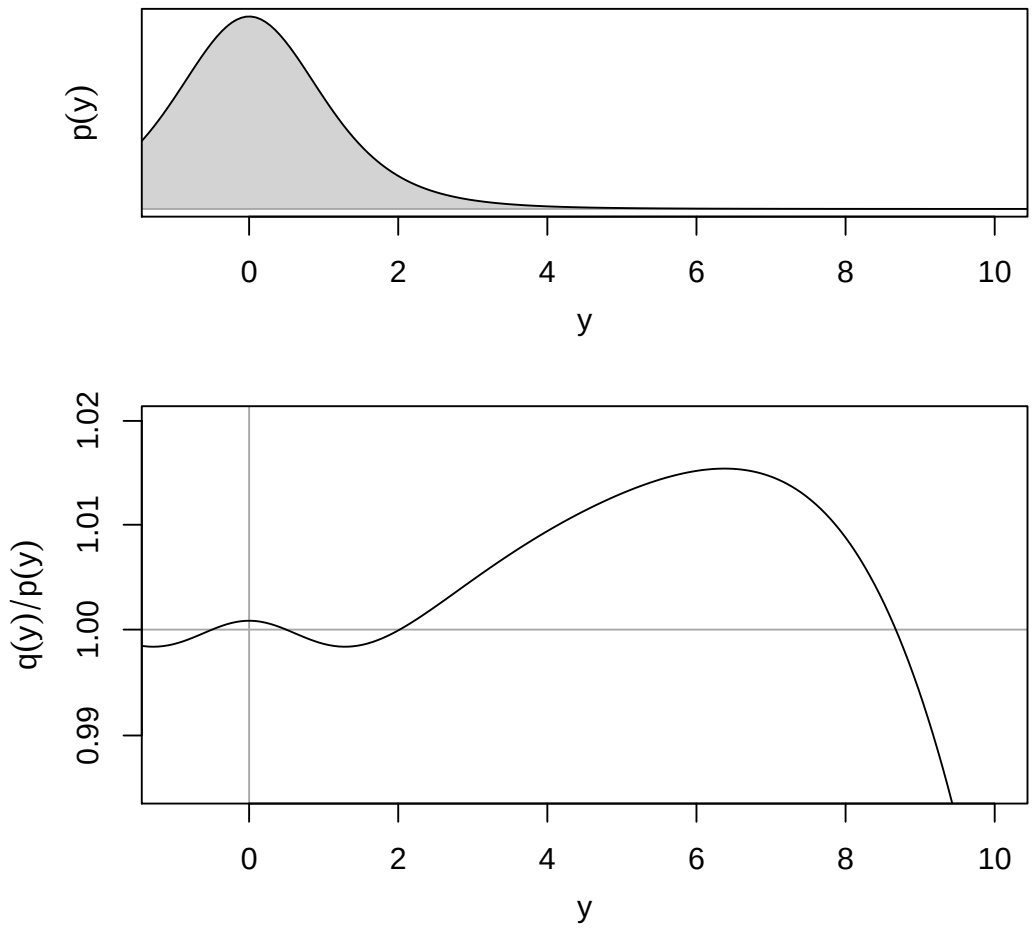
Example application: Student-t distribution



- 19 conditionals $p(y|\tilde{x}_i) \rightarrow$ weighted conditionals $p(y|\tilde{x}_i) \times \pi_i$ $\rightarrow$ discrete mixture $\sum_i p(y|\tilde{x}_i) \times \pi_i$

Discretizing mixtures

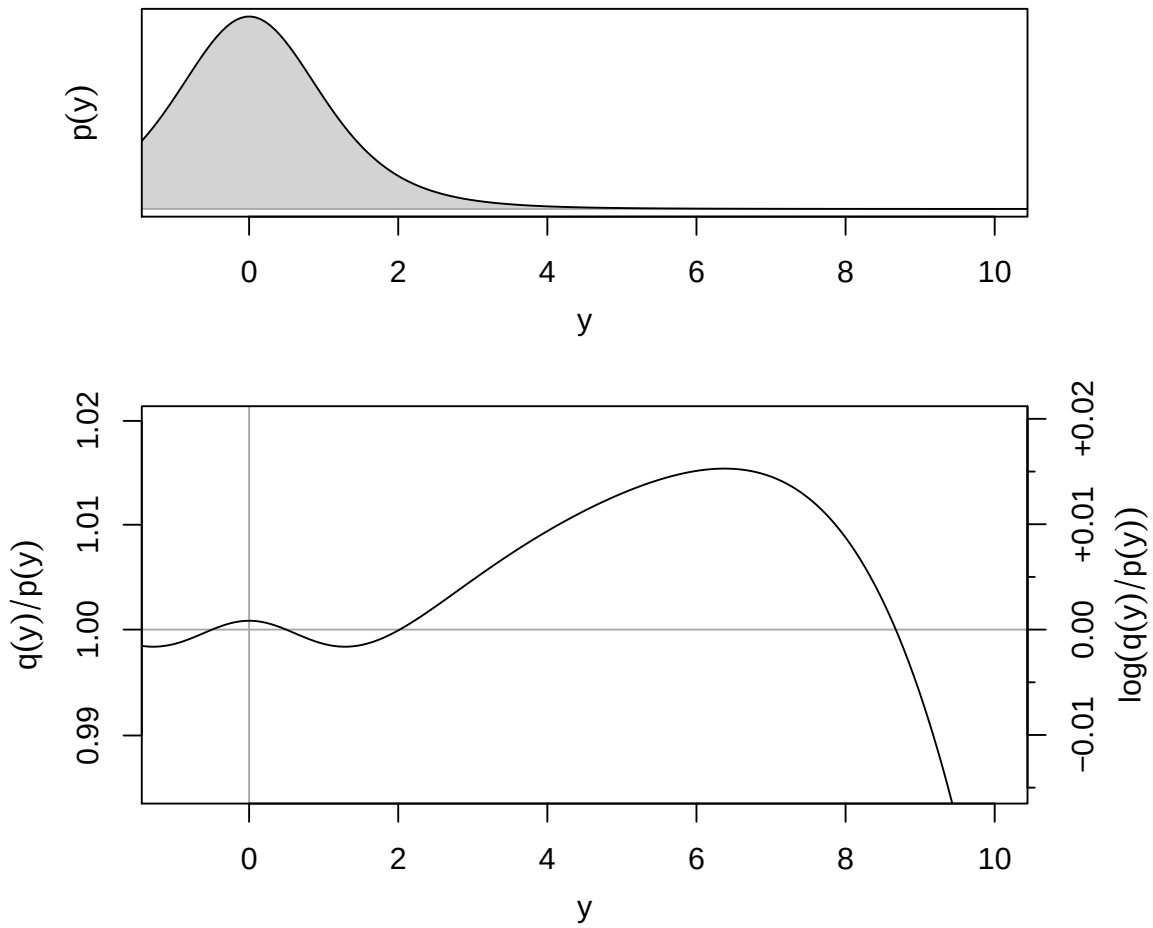
Example application: Student-*t* distribution



- how well do we do?

Discretizing mixtures

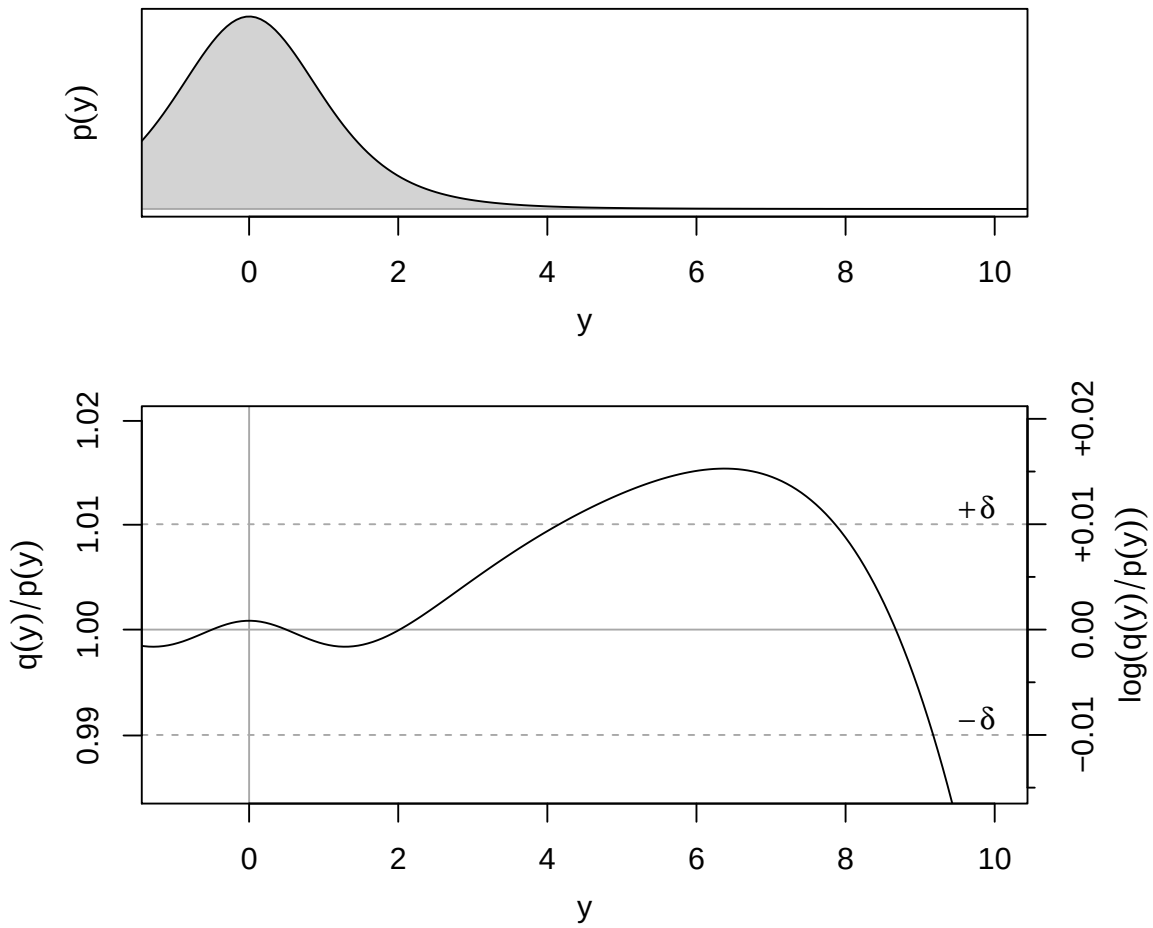
Example application: Student-*t* distribution



- how well do we do?

Discretizing mixtures

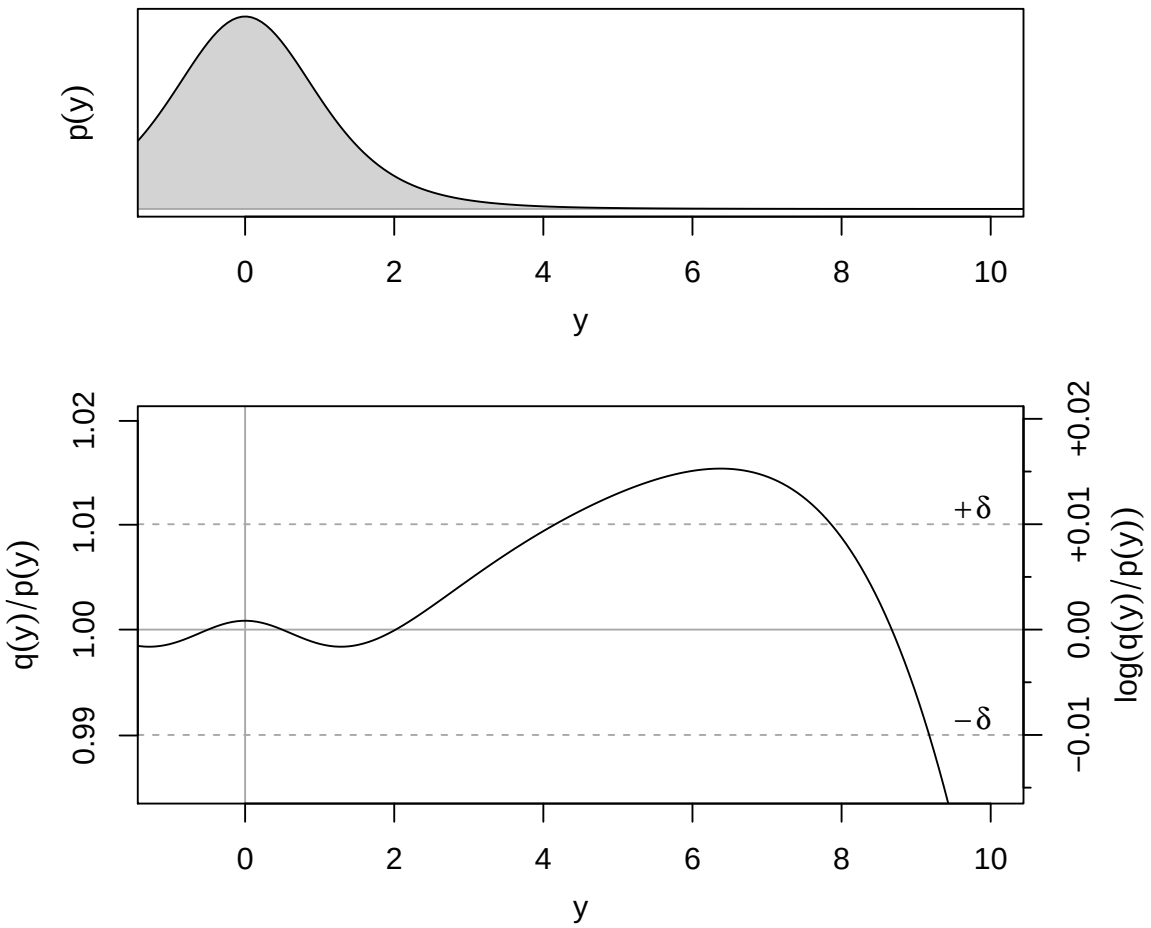
Example application: Student-*t* distribution



- how well do we do?

Discretizing mixtures

Example application: Student-*t* distribution



- how well do we do? $\rightarrow$ compute divergences numerically: $\mathcal{D}_{\mathrm{KL}}\big(p(\theta)\big\|q(\theta)\big) = 0.000035$, $\mathcal{D}_{\mathrm{KL}}\big(q(\theta)\big\|p(\theta)\big) = 0.000013$, $\mathcal{D}_{\mathrm{s}}\big(p(\theta)\big\|q(\theta)\big) = 0.000048$

Conclusions

- discrete 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 makes sense for discrete marginals $p(x)$
- other strategies possible, e.g. aiming not for (bin-wise) *maximum* divergence d_i, but for conditional *expectation*...
- higher dimensions: should work in priciple, probably tricky
- random-effects meta-analysis implemented in `bmeta` R package
- methods paper in preparation