

# Identifying Consistent Statements about Numerical Data with **Dispersion-Corrected Subgroup Discovery**

**Mario Boley**, Bryan Goldsmith, Luca Ghiringhelli, Jilles Vreeken  
[mboley@mpi-inf.mpg.de](mailto:mboley@mpi-inf.mpg.de)

Max Planck Institute for Informatics and Saarland University  
Fritz Haber Institute of the Max Planck Society

# Subgroup discovery (with binary target)

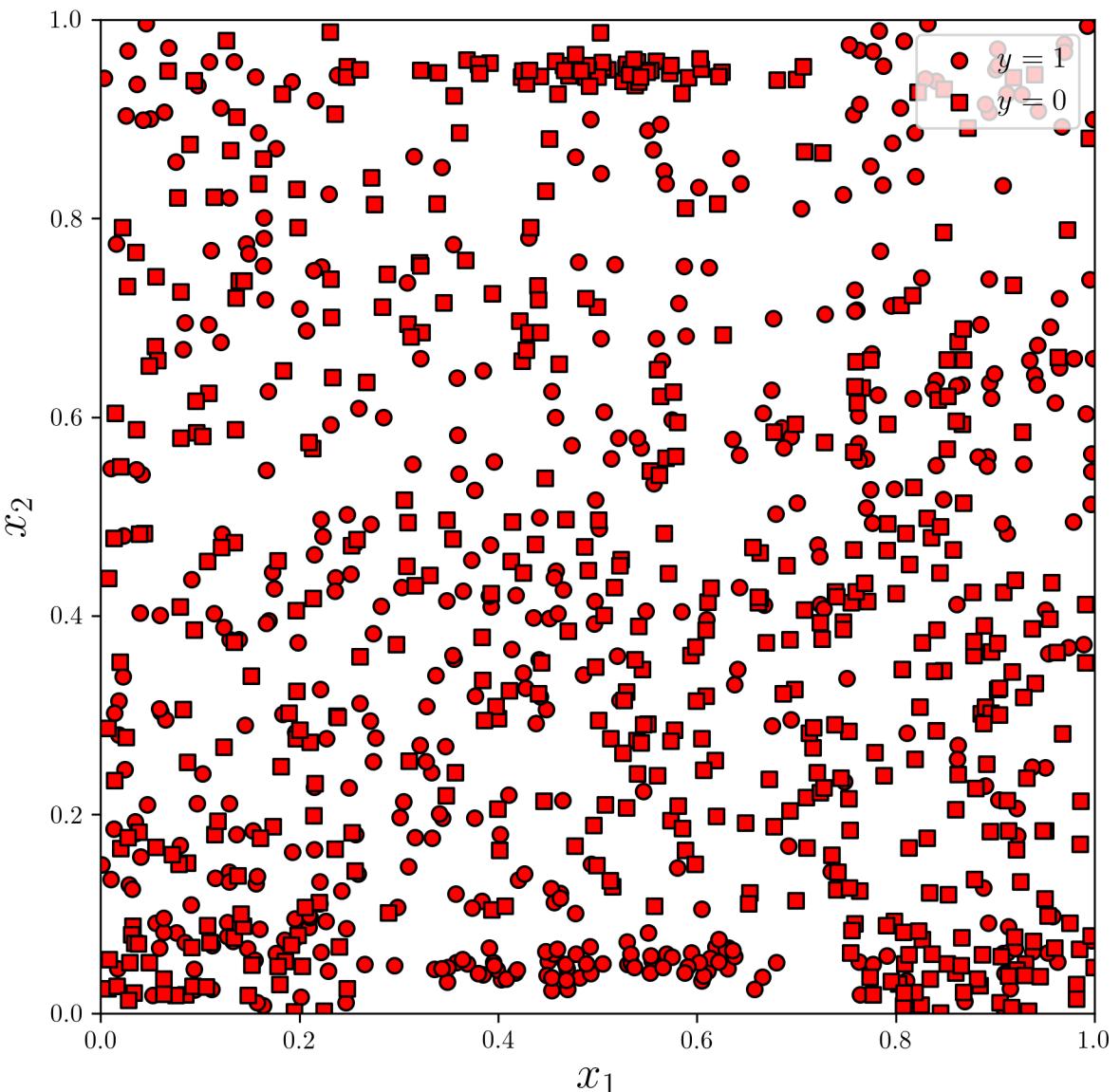
## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \{\oplus, \ominus\}$

Description variables  $x_j: P \rightarrow X_j$

[Klösgen, 1996; Wrobel, 1997; Duivesteijn et al., 2008]



# Subgroup discovery (with binary target)

## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \{\oplus, \ominus\}$

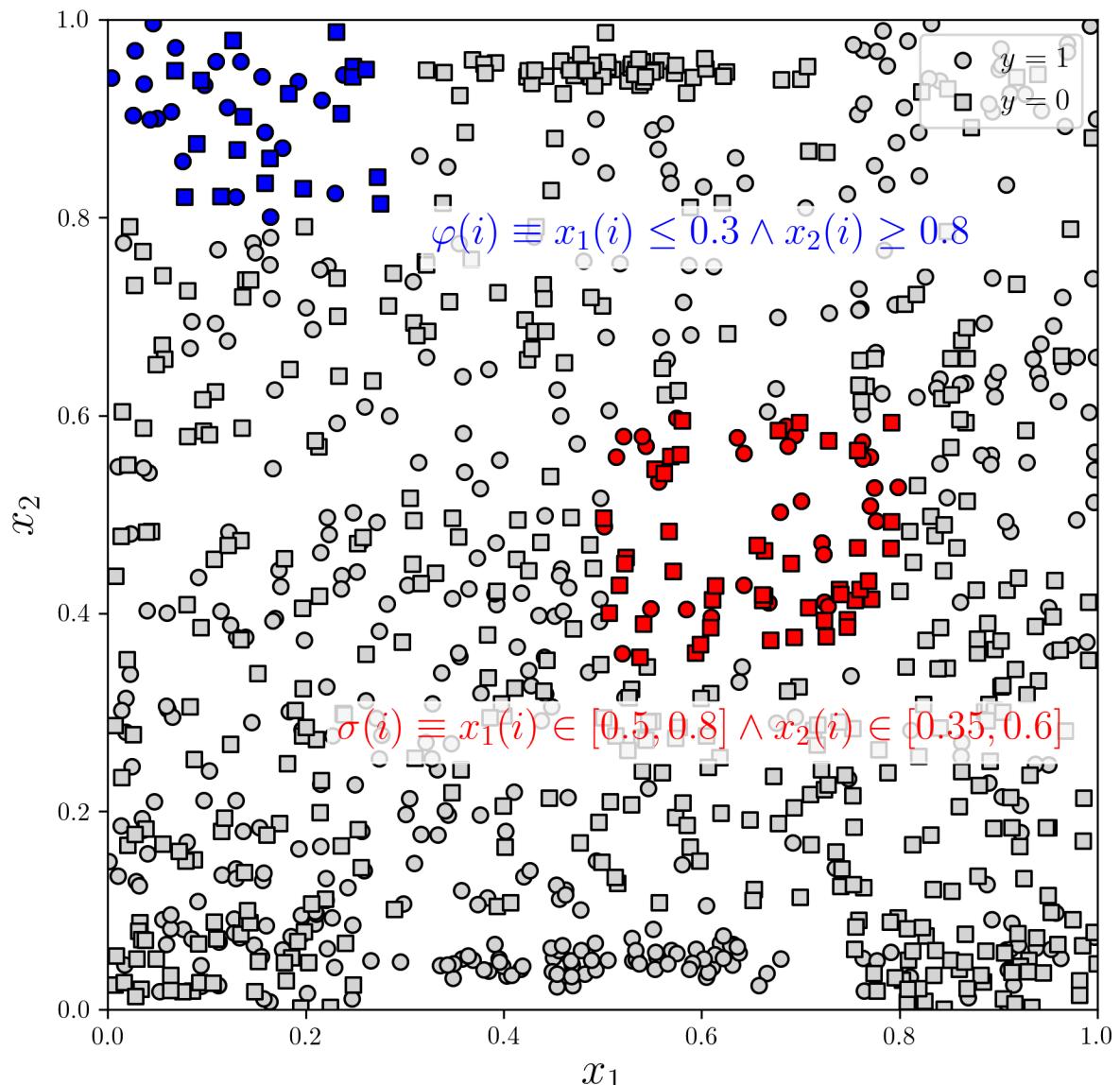
Description variables  $x_j: P \rightarrow X_j$

## Define

Selection language  $\mathcal{L}_x \subseteq \{\perp, \top\}^P$

$(\sigma \in \mathcal{L}_x \text{ defines } \text{ext}(\sigma) = \{i \in S: \sigma(i) = \top\} \subseteq S)$

[Klösgen, 1996; Wrobel, 1997; Duivesteijn et al., 2008]



# Subgroup discovery (with binary target)

## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \{\oplus, \ominus\}$

Description variables  $x_j: P \rightarrow X_j$

## Define

Selection language  $\mathcal{L}_x \subseteq \{\perp, \top\}^P$

$(\sigma \in \mathcal{L}_x \text{ defines } \text{ext}(\sigma) = \{i \in S: \sigma(i) = \top\} \subseteq S)$

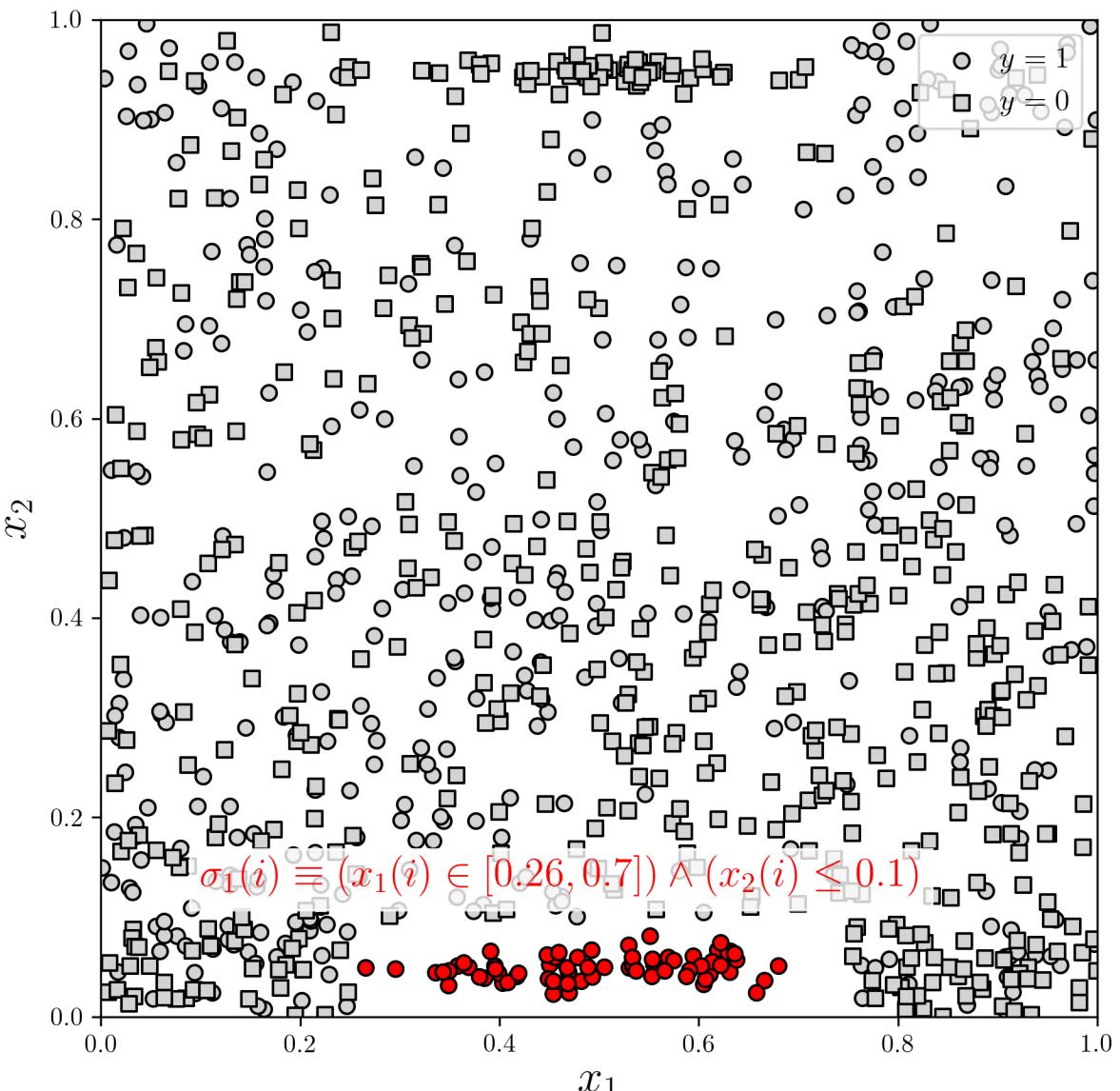
## Optimize

$$f(Q) = \text{cov}(Q)^\gamma \text{eff}(Q)_+$$

with

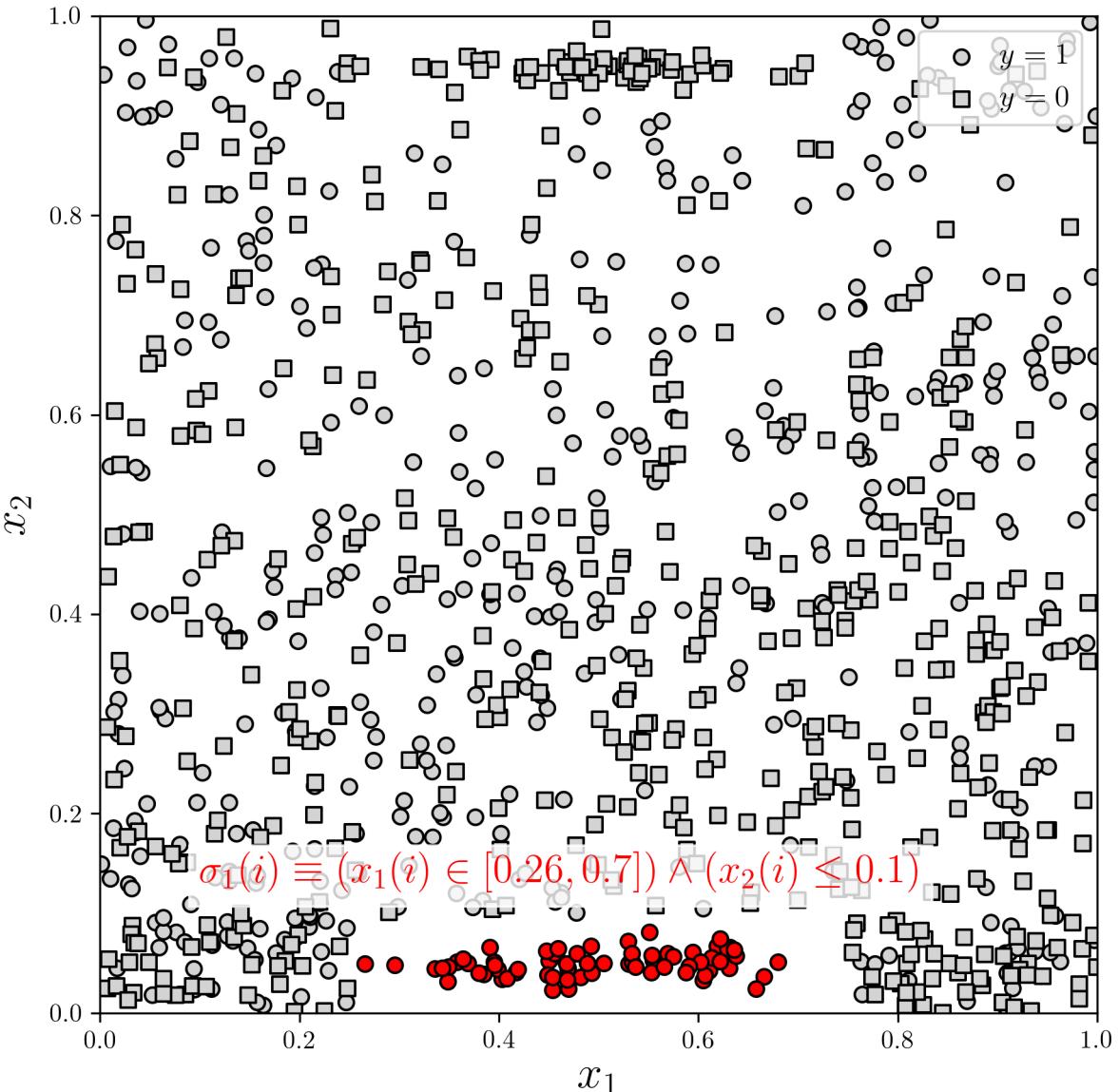
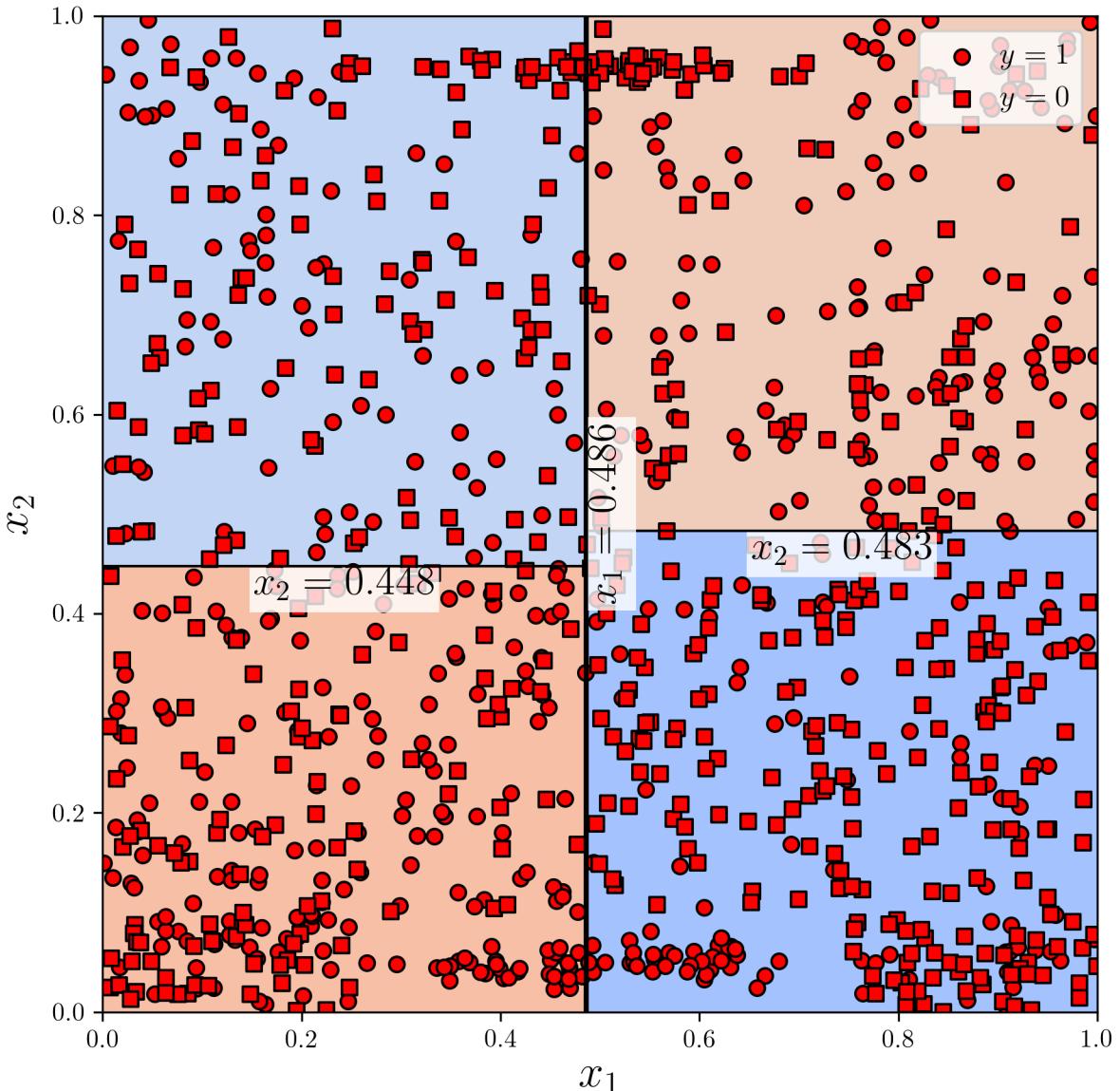
- $Q = \text{ext}(\sigma)$
- $\text{cov}(Q) = |Q|/|S|$  **coverage**
- $\text{eff}(Q) = \tilde{y}(Q) - \tilde{y}(S)$  **effect**
- $\tilde{y}(Q) = \{i \in Q: y(i) = \oplus\}/|Q|$  **pos. prob.**

[Klösgen, 1996; Wrobel, 1997; Duivesteijn et al., 2008]



# Better than global models in capturing *local* effects

5



# Metric target variables

6

## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \mathbb{R}$

Description variables  $x_j: P \rightarrow X_j$

## Define

Selection language  $\mathcal{L}_x \subseteq \{\perp, \top\}^P$

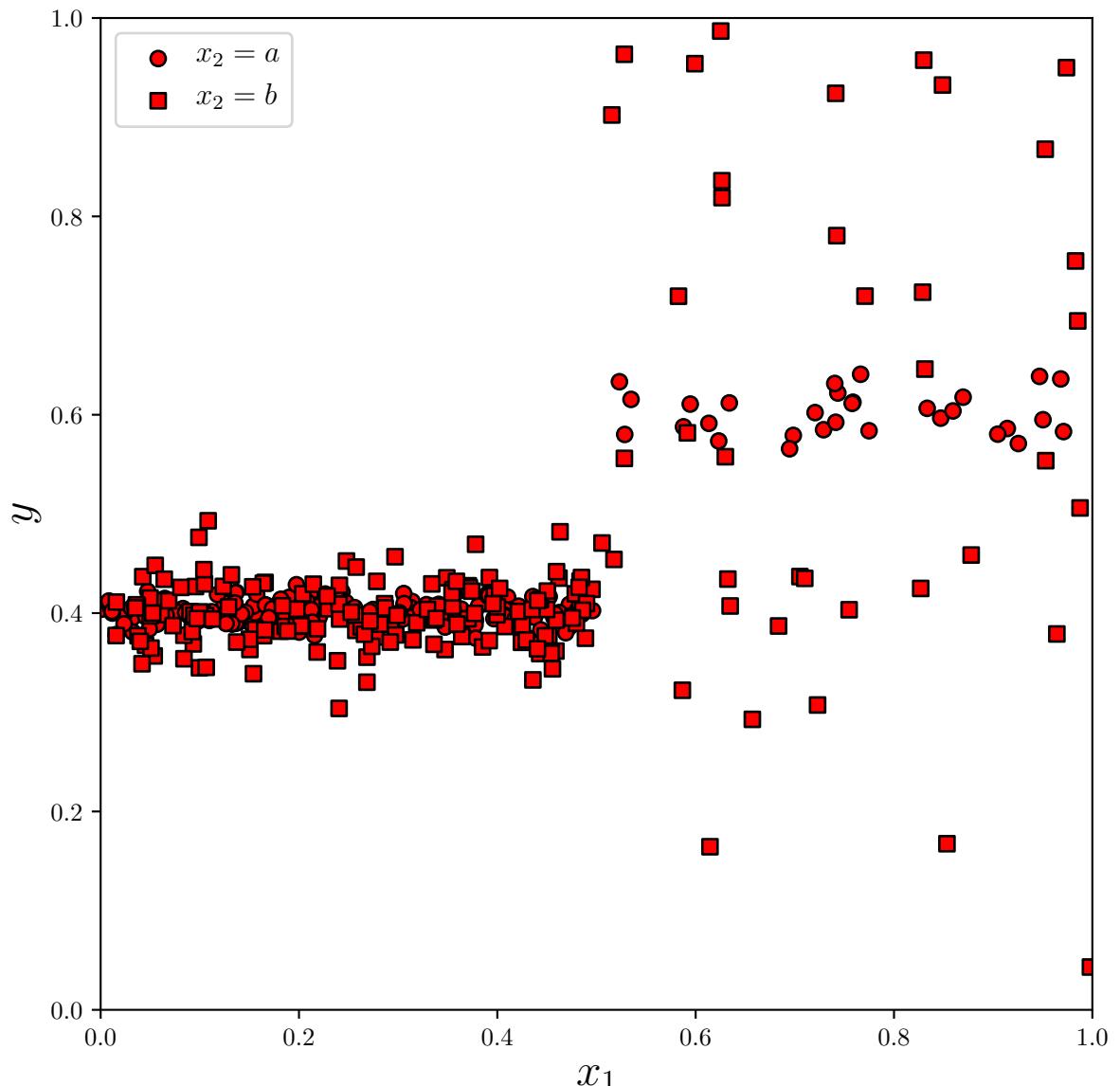
( $\sigma \in \mathcal{L}_x$  defines  $\text{ext}(\sigma) = \{i \in S: \sigma(i) = \top\} \subseteq S$ )

## Optimize

$$f(Q) = \text{cov}(Q)^\gamma \text{eff}(Q)_+$$

with

- $Q = \text{ext}(\sigma)$
- $\text{cov}(Q) = |Q|/|S|$  coverage
- $\text{eff}(Q) = \tilde{y}(Q) - \tilde{y}(S)$  effect
- $\tilde{y}(Q)$  central tendency (mean, median,...)



# Metric target variables

7

## Given

Sample  $S \subseteq P$

Target variable  $y: P \rightarrow \mathbb{R}$

Description variables  $x_j: P \rightarrow X_j$

## Define

Selection language  $\mathcal{L}_x \subseteq \{\perp, \top\}^P$

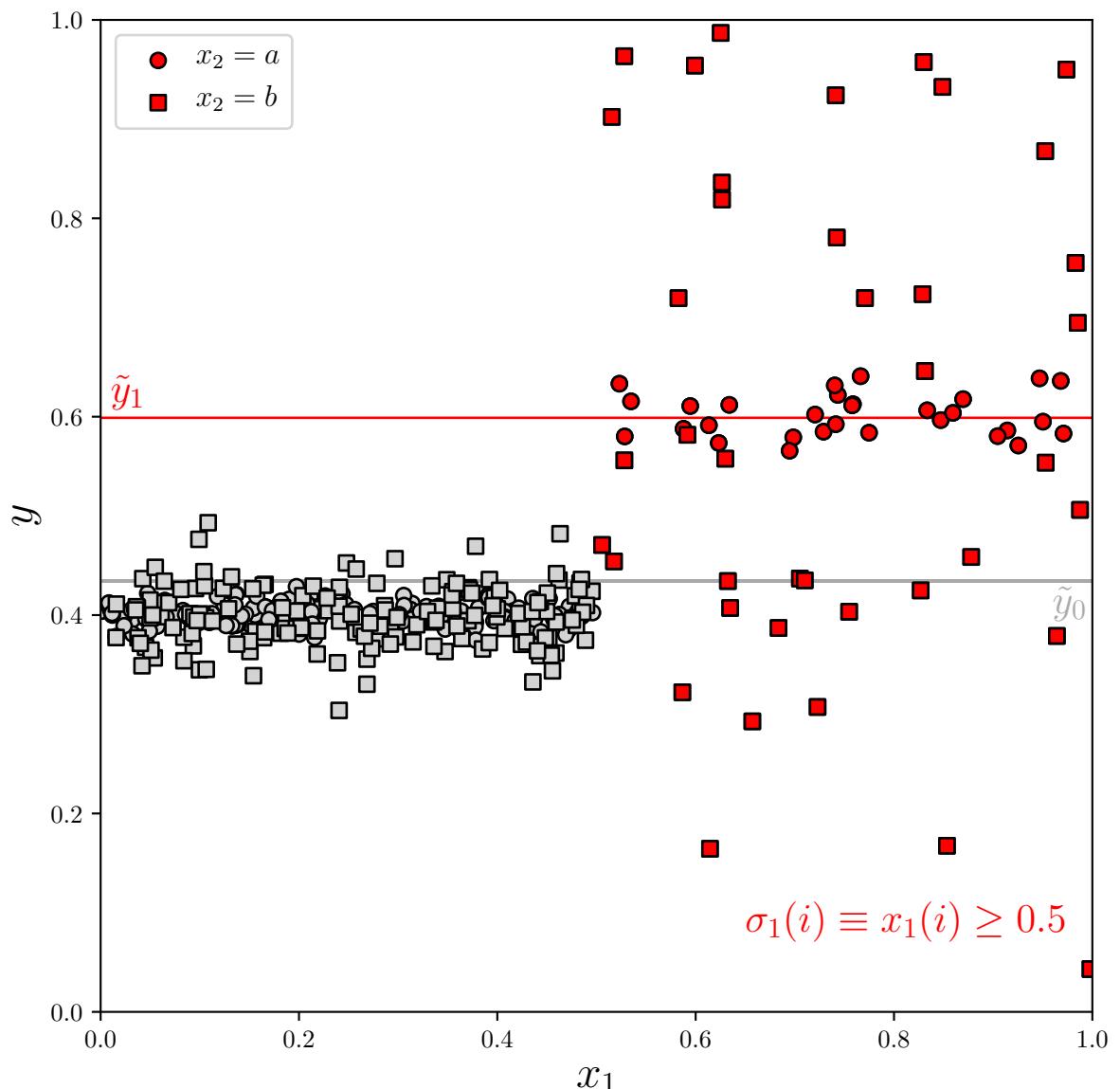
( $\sigma \in \mathcal{L}_x$  defines  $\text{ext}(\sigma) = \{i \in S: \sigma(i) = \top\} \subseteq S$ )

## Optimize

$$f(Q) = \text{cov}(Q)^\gamma \text{eff}(Q)_+$$

with

- $Q = \text{ext}(\sigma)$
- $\text{cov}(Q) = |Q|/|S|$  coverage
- $\text{eff}(Q) = \tilde{y}(Q) - \tilde{y}(S)$  effect
- $\tilde{y}(Q)$  central tendency (mean, median,...)



# Dysfunctionality of pure coverage/effect approach

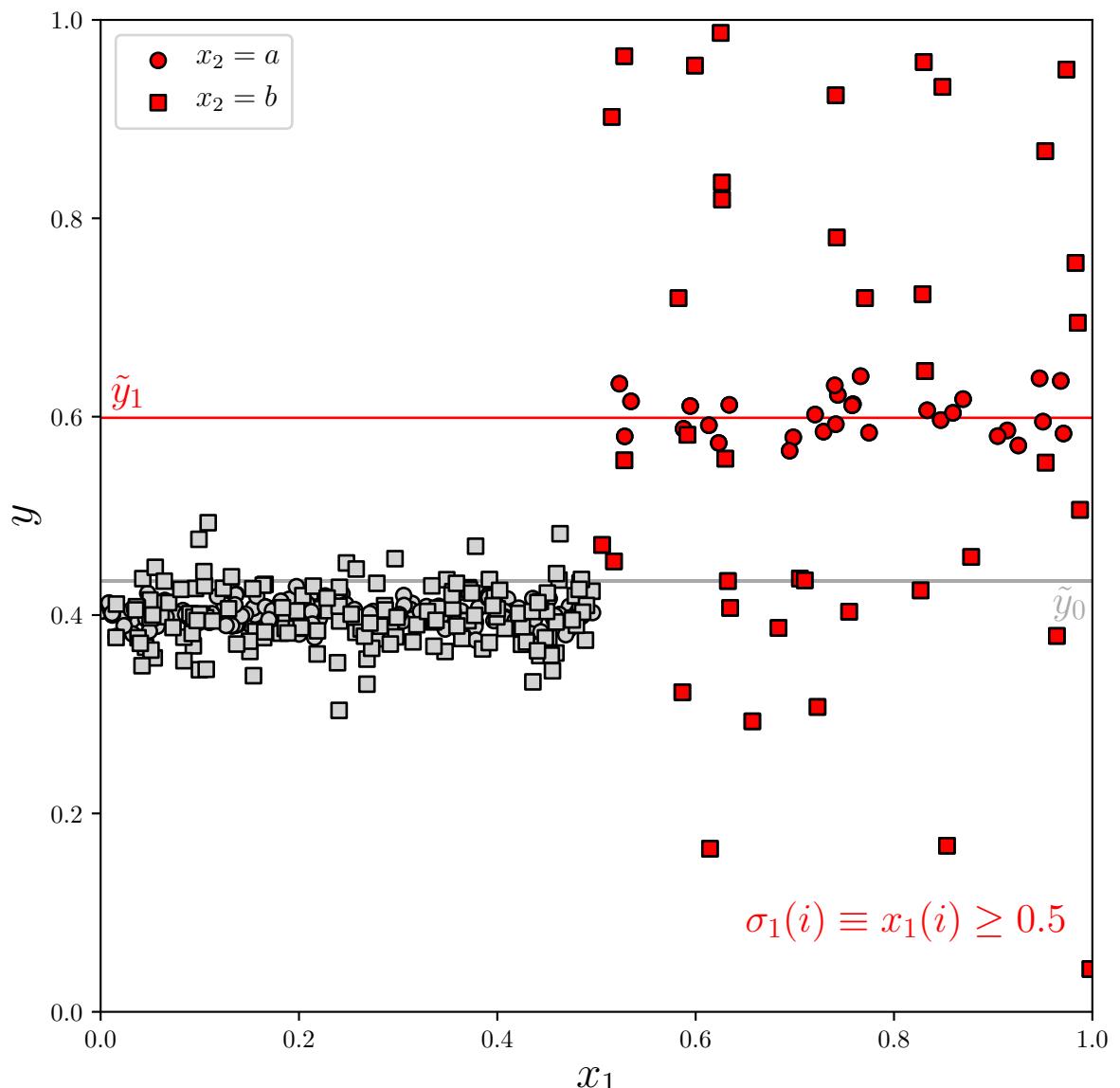
8

## Dispersion

average error  $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case  $\tilde{y}(Q) = \bar{y}(Q)$ :  $e(i) = (\tilde{y}(Q) - y(i))^2$

case  $\tilde{y}(Q) = \text{med}(Q)$ :  $e(i) = |\tilde{y}(Q) - y(i)|$



# Dysfunctionality of pure coverage/effect approach

9

## Dispersion

average error  $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case  $\tilde{y}(Q) = \bar{y}(Q)$ :  $e(i) = (\tilde{y}(Q) - y(i))^2$

case  $\tilde{y}(Q) = \text{med}(Q)$ :  $e(i) = |\tilde{y}(Q) - y(i)|$

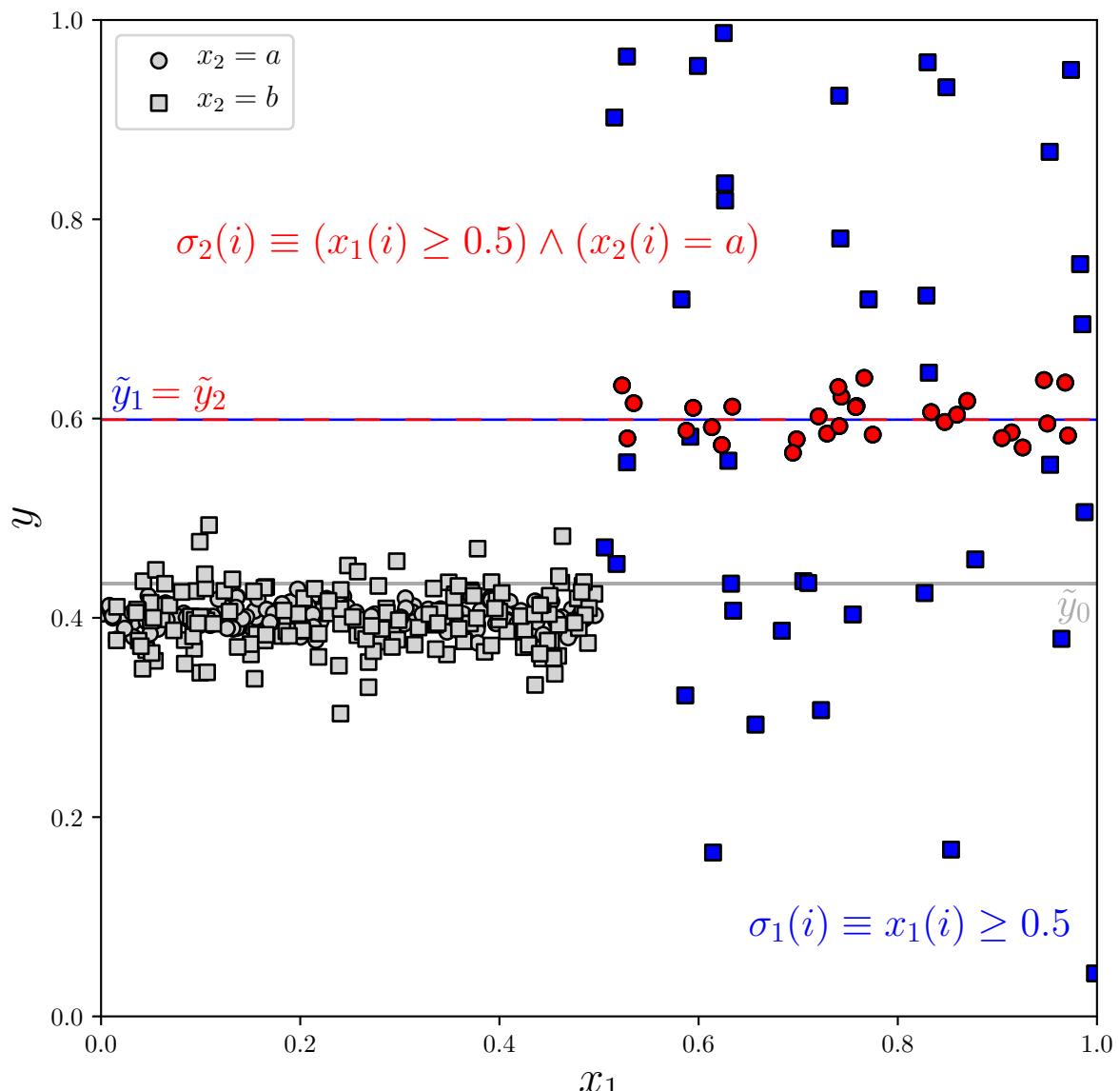
## Selection preference

any

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

monotone in first argument favors groups that

- are not summarized well by  $\tilde{y}$



# Dysfunctionality of pure coverage/effect approach

10

## Dispersion

average error  $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case  $\tilde{y}(Q) = \bar{y}(Q)$ :  $e(i) = (\tilde{y}(Q) - y(i))^2$

case  $\tilde{y}(Q) = \text{med}(Q)$ :  $e(i) = |\tilde{y}(Q) - y(i)|$

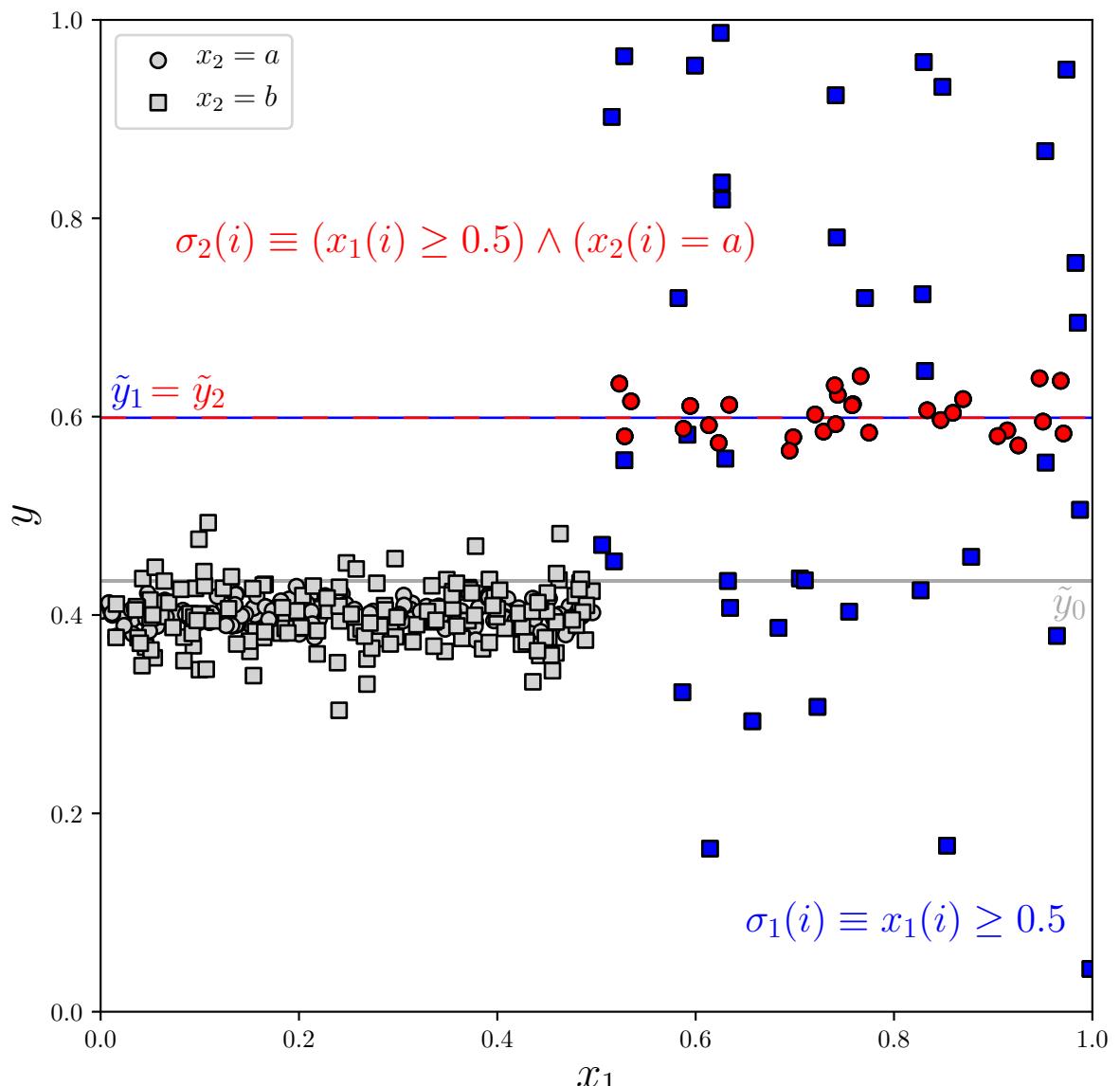
## Selection preference

any

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

monotone in first argument favors groups that

- are not summarized well by  $\tilde{y}$
- contain noise



# Dysfunctionality of pure coverage/effect approach

11

## Dispersion

average error  $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case  $\tilde{y}(Q) = \bar{y}(Q)$ :  $e(i) = (\tilde{y}(Q) - y(i))^2$   
case  $\tilde{y}(Q) = \text{med}(Q)$ :  $e(i) = |\tilde{y}(Q) - y(i)|$

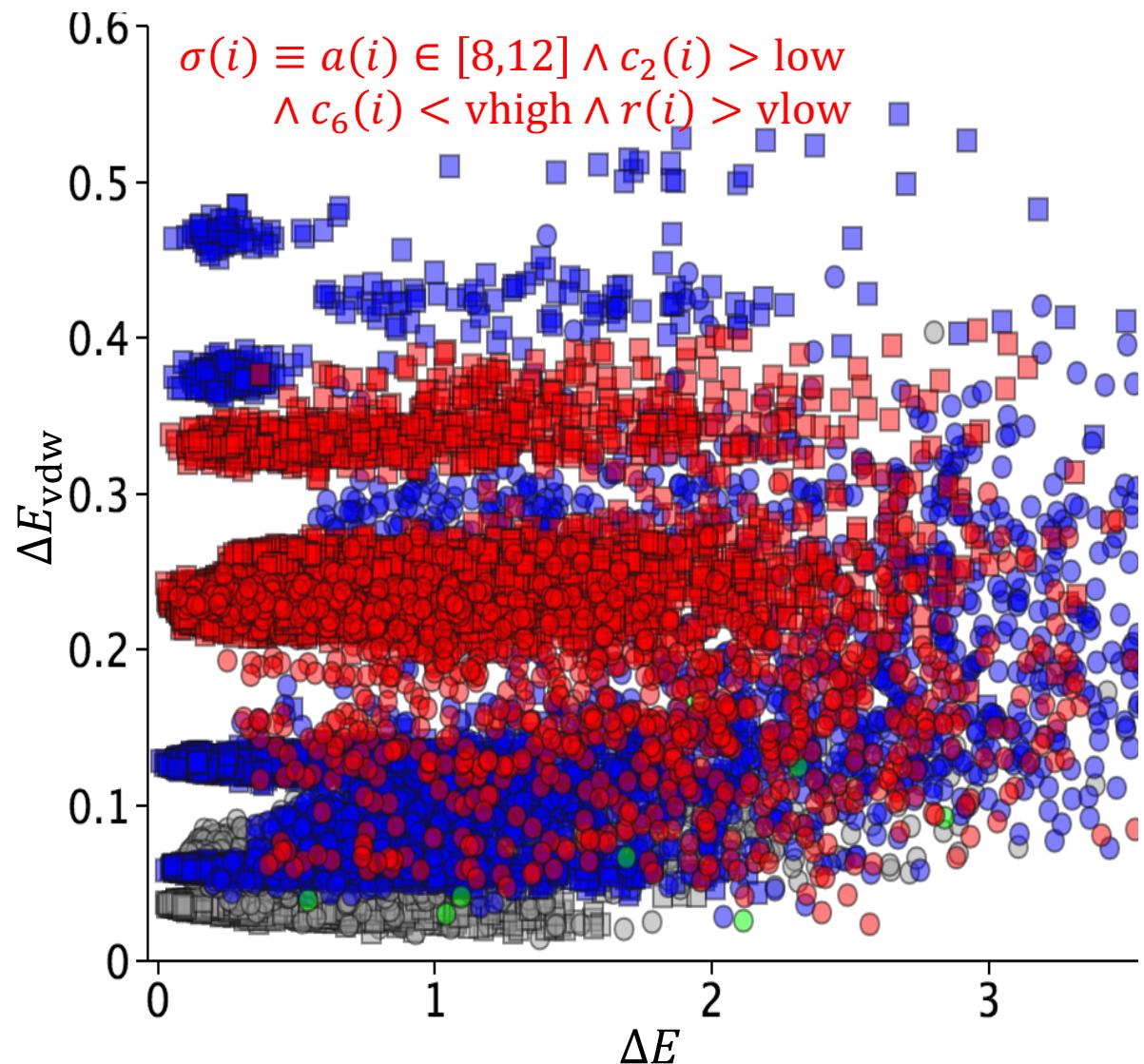
## Selection preference

any

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

monotone in first argument favors groups that

- are not summarized well by  $\tilde{y}$
- contain noise
- are **incoherent**



# Dysfunctionality of pure coverage/effect approach

12

## Dispersion

average error  $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case  $\tilde{y}(Q) = \bar{y}(Q)$ :  $e(i) = (\tilde{y}(Q) - y(i))^2$

case  $\tilde{y}(Q) = \text{med}(Q)$ :  $e(i) = |\tilde{y}(Q) - y(i)|$

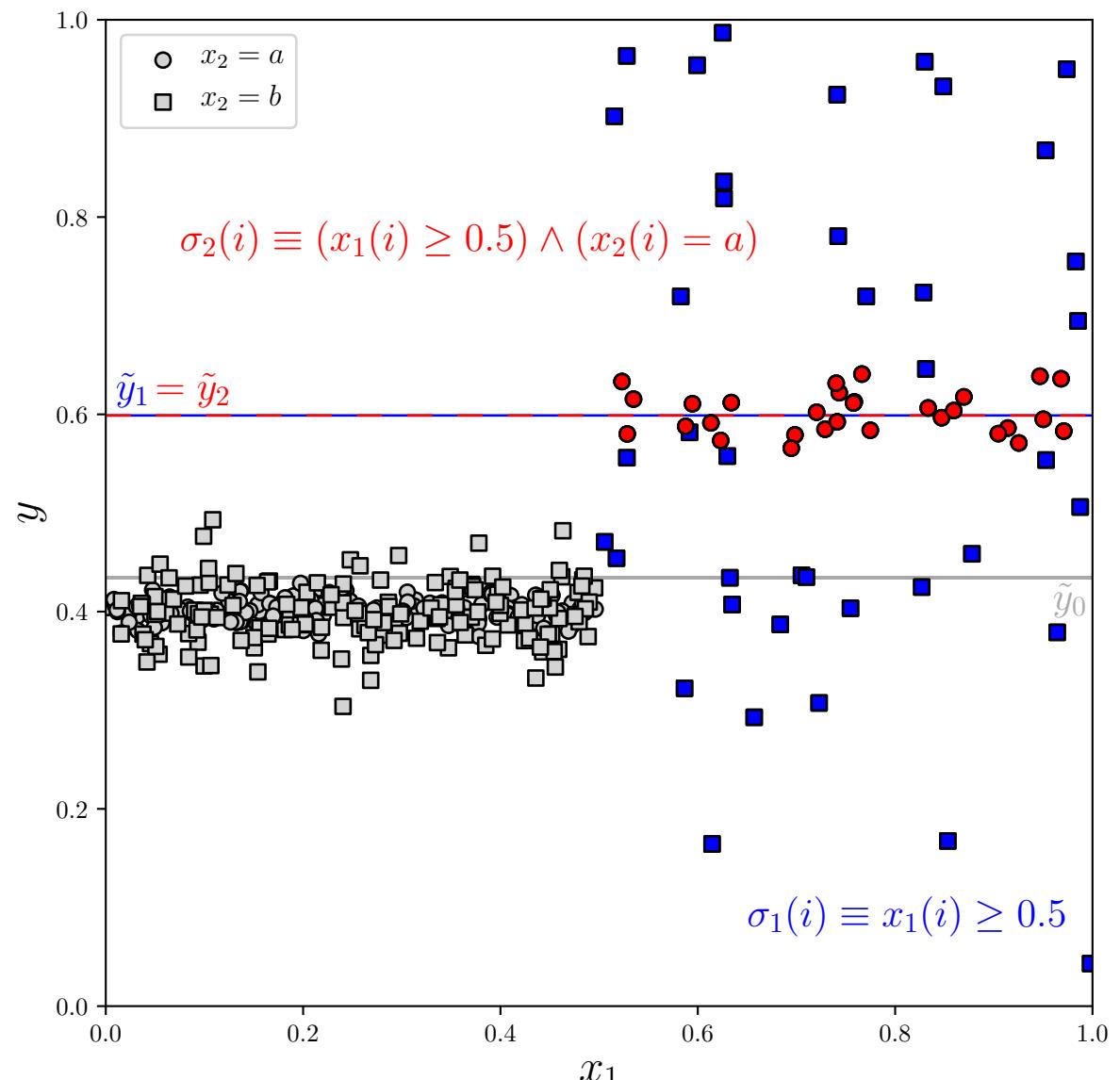
## Selection preference

any

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

monotone in first argument favors groups that

- are not summarized well by  $\tilde{y}$
- contain noise
- are incoherent



# Dysfunctionality of pure coverage/effect approach

13

## Dispersion

average error  $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case  $\tilde{y}(Q) = \bar{y}(Q)$ :  $e(i) = (\tilde{y}(Q) - y(i))^2$

case  $\tilde{y}(Q) = \text{med}(Q)$ :  $e(i) = |\tilde{y}(Q) - y(i)|$

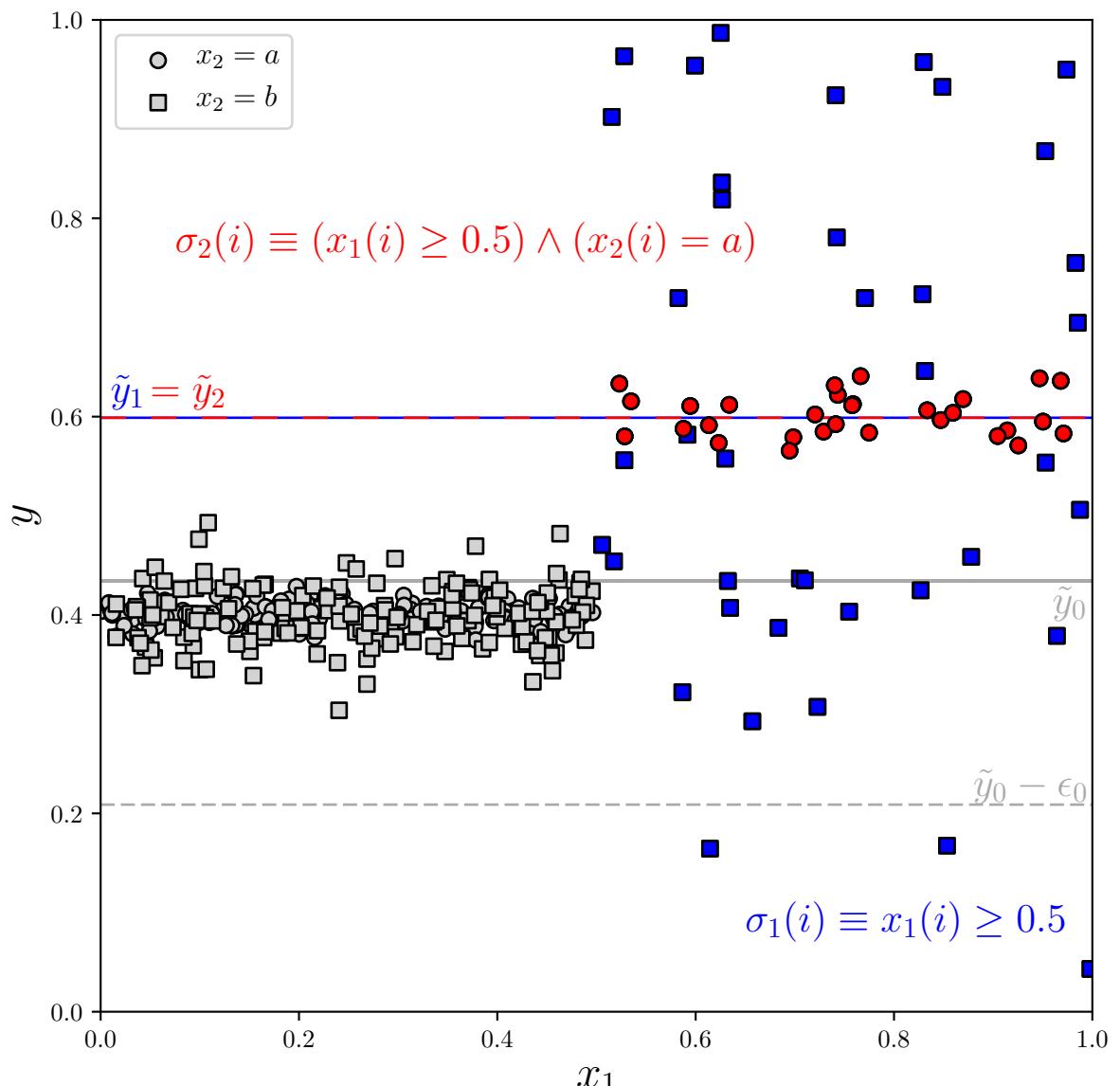
## Selection preference

any

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

monotone in first argument favors groups that

- are not summarized well by  $\tilde{y}$
- contain noise
- are incoherent



# Dysfunctionality of pure coverage/effect approach

14

## Dispersion

average error  $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case  $\tilde{y}(Q) = \bar{y}(Q)$ :  $e(i) = (\tilde{y}(Q) - y(i))^2$

case  $\tilde{y}(Q) = \text{med}(Q)$ :  $e(i) = |\tilde{y}(Q) - y(i)|$

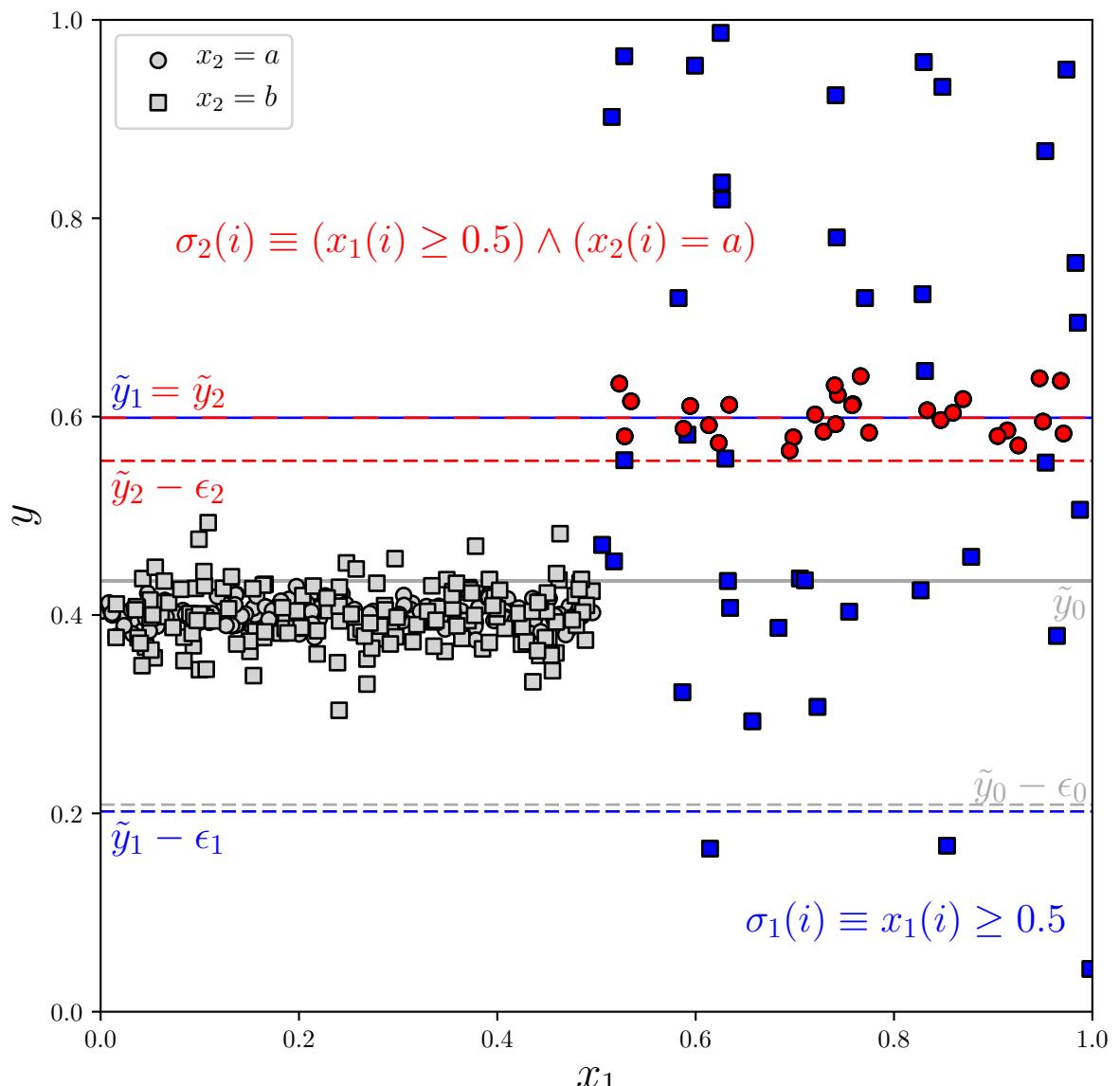
## Selection preference

any

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

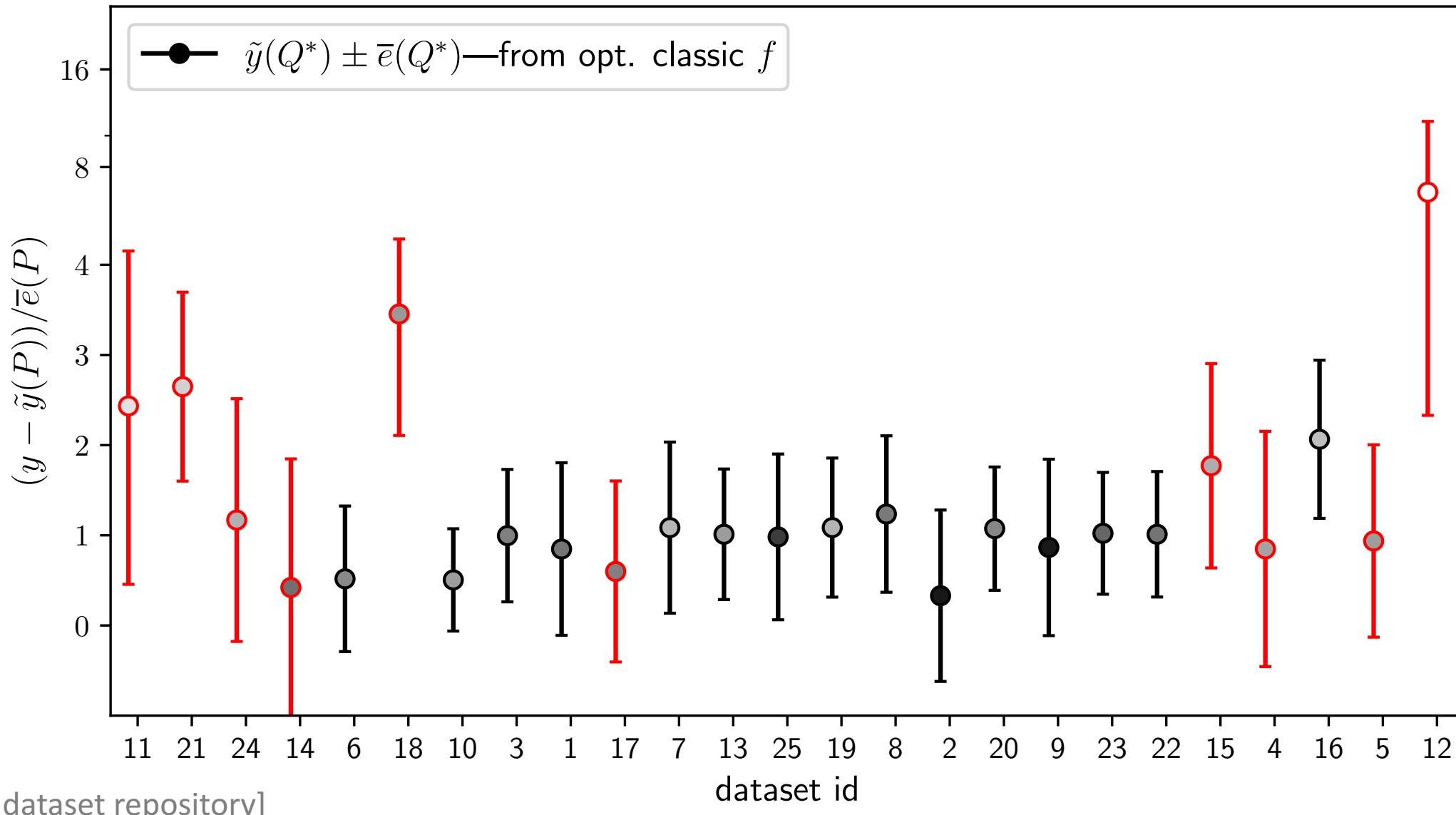
monotone in first argument favors groups that

- are not summarized well by  $\tilde{y}$
- contain noise
- are incoherent
- provide weak *quantitative* guarantees wrt  $P$



# In 10 of 25 datasets local error larger than global

15



[KEEL dataset repository]

# How to fix this?

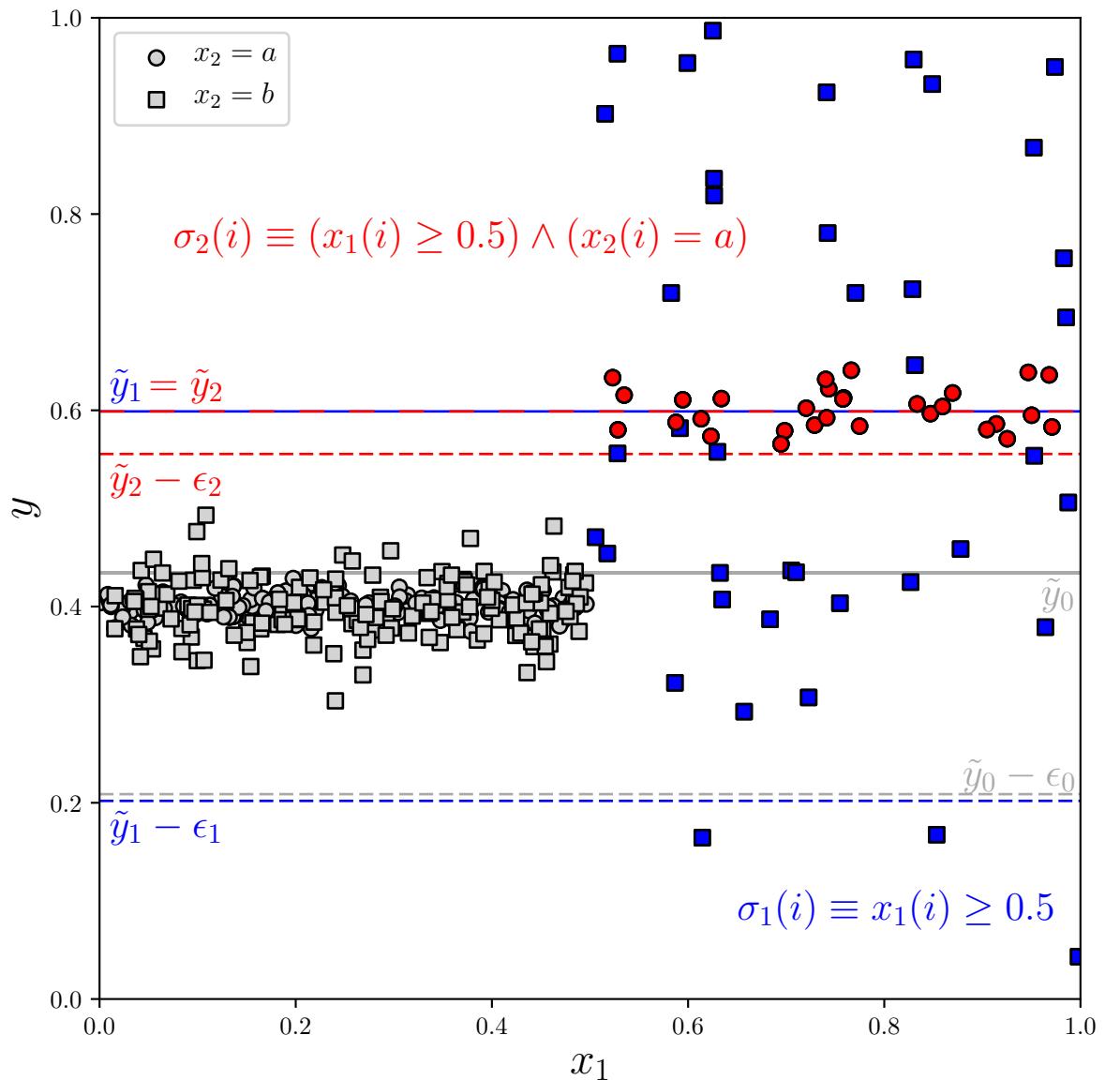
16

**Normalize effect ??**

$$f(Q) = g \left( \text{cov}(Q), \frac{\text{eff}(Q)}{\bar{e}(Q)} \right)$$

- ugly edge cases
- unclear how to optimize

[Klösgen, 2002; Pieters et al., 2010]



# How to fix this?

17

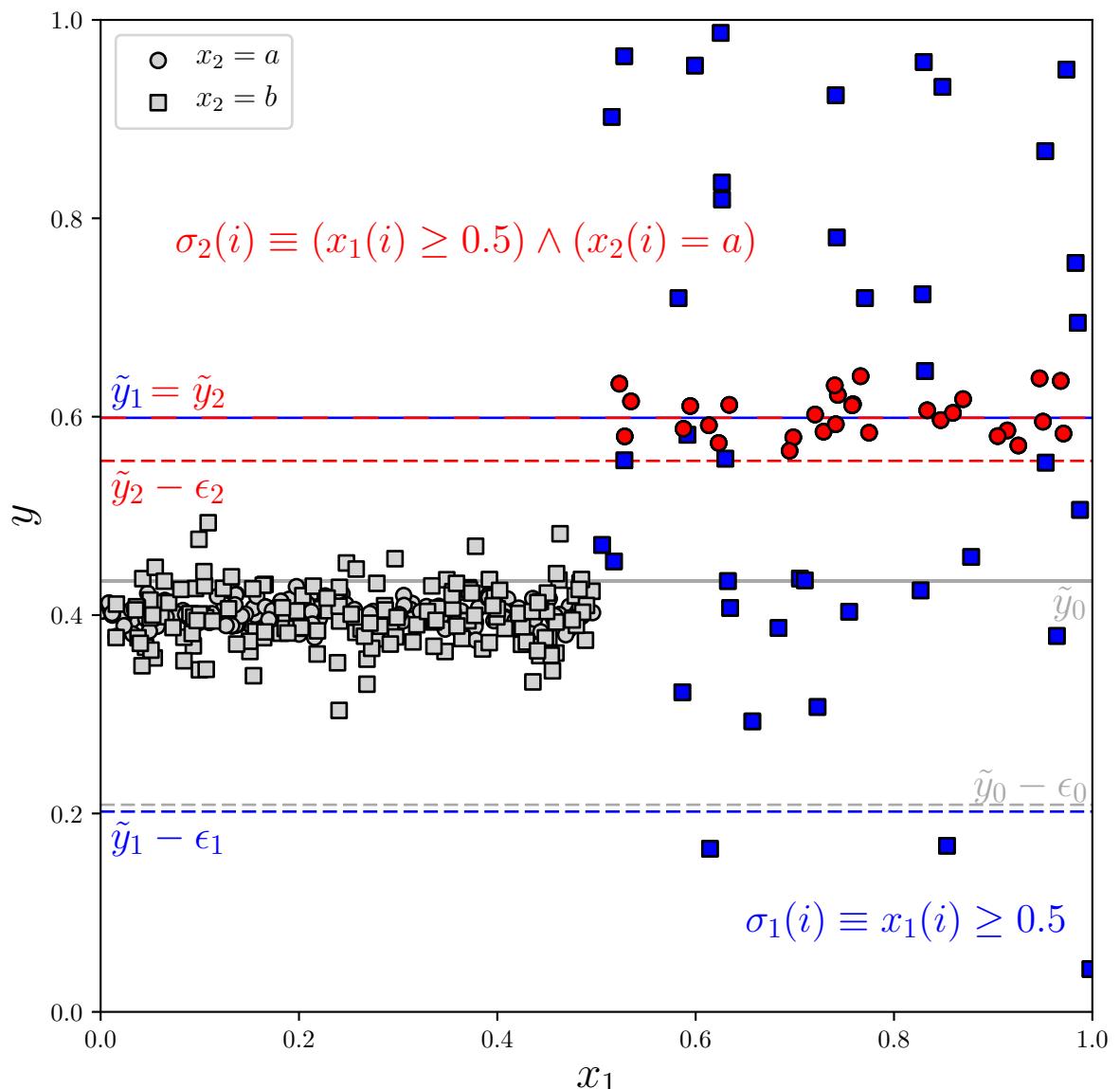
**Normalize effect ??**

$$f(Q) = g \left( \text{cov}(Q), \frac{\text{eff}(Q)}{\bar{e}(Q)} \right)$$

- ugly edge cases
- unclear how to optimize

**Correct coverage**

$$f(Q) = g \left( \text{cov}(Q) \left( \frac{\bar{e}(S) - \bar{e}(Q)}{\bar{e}(S)} \right)_+, \text{eff}(Q) \right)$$



# How to fix this?

18

## Normalize effect ??

$$f(Q) = g \left( \text{cov}(Q), \frac{\text{eff}(Q)}{\bar{e}(Q)} \right)$$

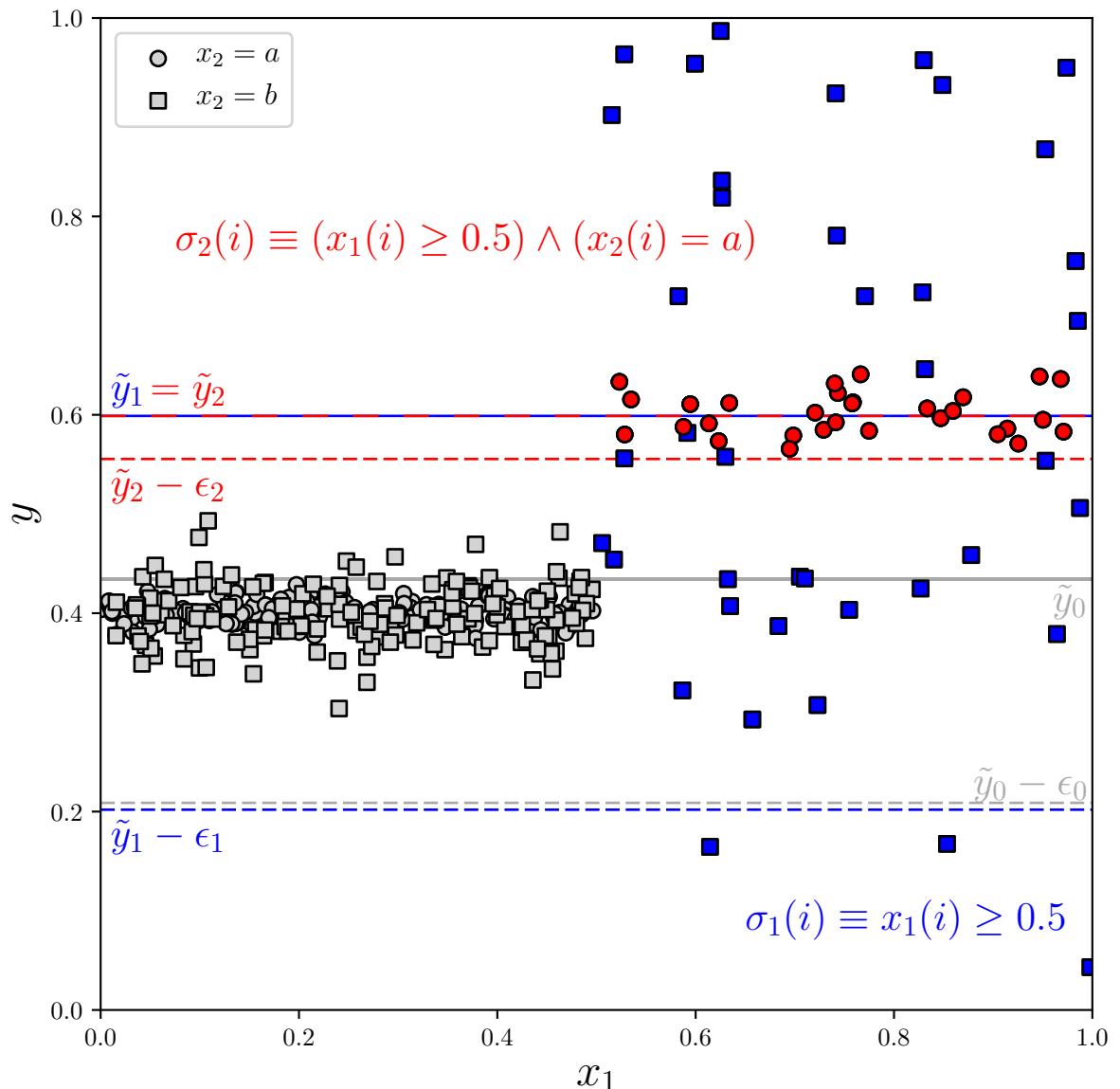
- ugly edge cases
- unclear how to optimize

## Correct coverage

$$f(Q) = g \left( \text{cov}(Q) \left( \frac{\bar{e}(S) - \bar{e}(Q)}{\bar{e}(S)} \right)_+, \text{eff}(Q) \right)$$

$$\text{dcc}(Q) = \left( \frac{|Q|}{|S|} - \frac{e(Q)}{e(S)} \right)_+$$

dispersion-corrected coverage



# How to fix this?

19

**Normalize effect ??**

$$f(Q) = g \left( \text{cov}(Q), \frac{\text{eff}(Q)}{\bar{e}(Q)} \right)$$

- ugly edge cases
- unclear how to optimize

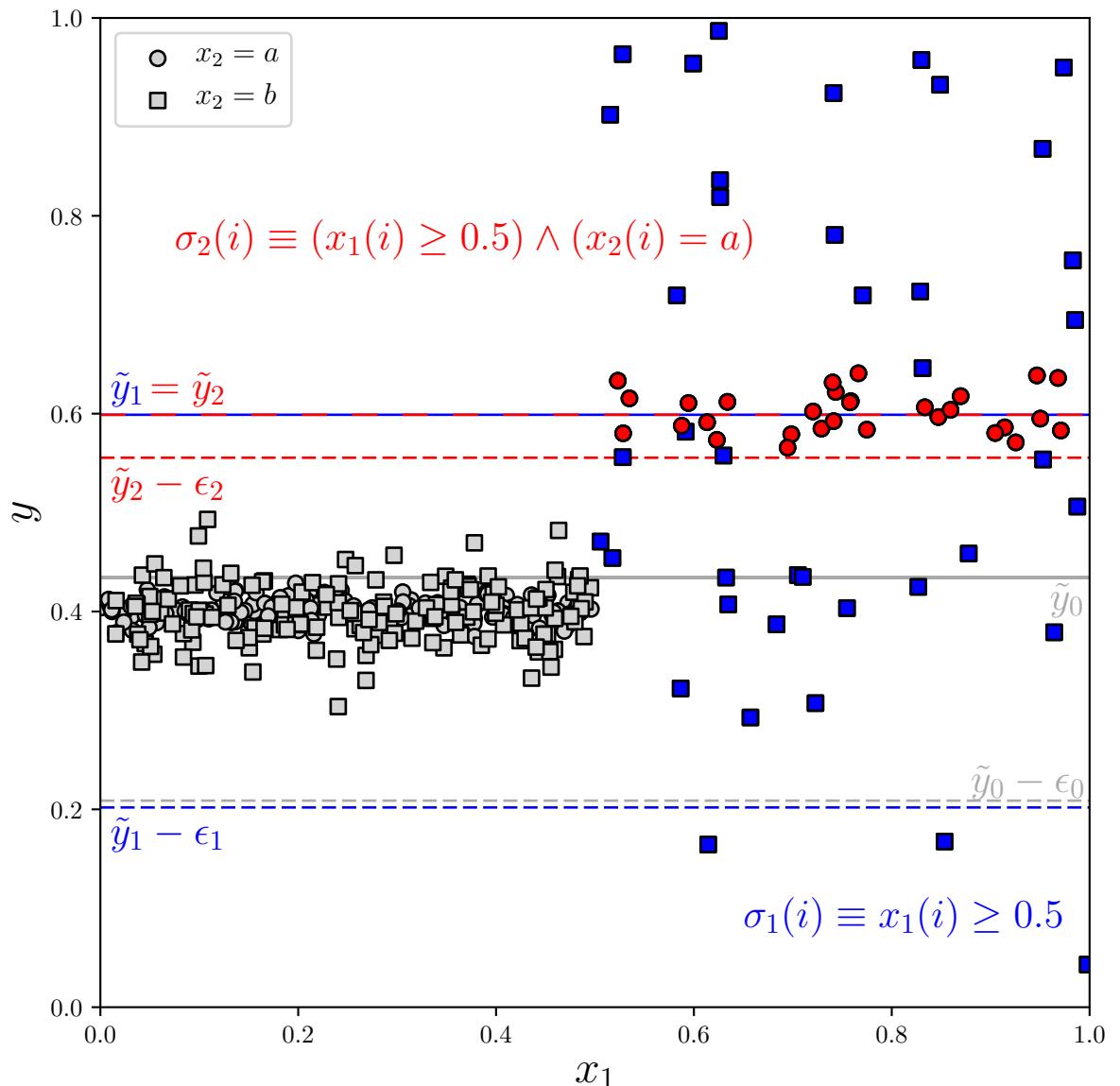
**Correct coverage**

$$f(Q) = g \left( \text{cov}(Q) \left( \frac{\bar{e}(S) - \bar{e}(Q)}{\bar{e}(S)} \right)_+, \text{eff}(Q) \right)$$

$$\text{dcc}(Q) = \left( \frac{|Q|}{|S|} - \frac{e(Q)}{e(S)} \right)_+$$

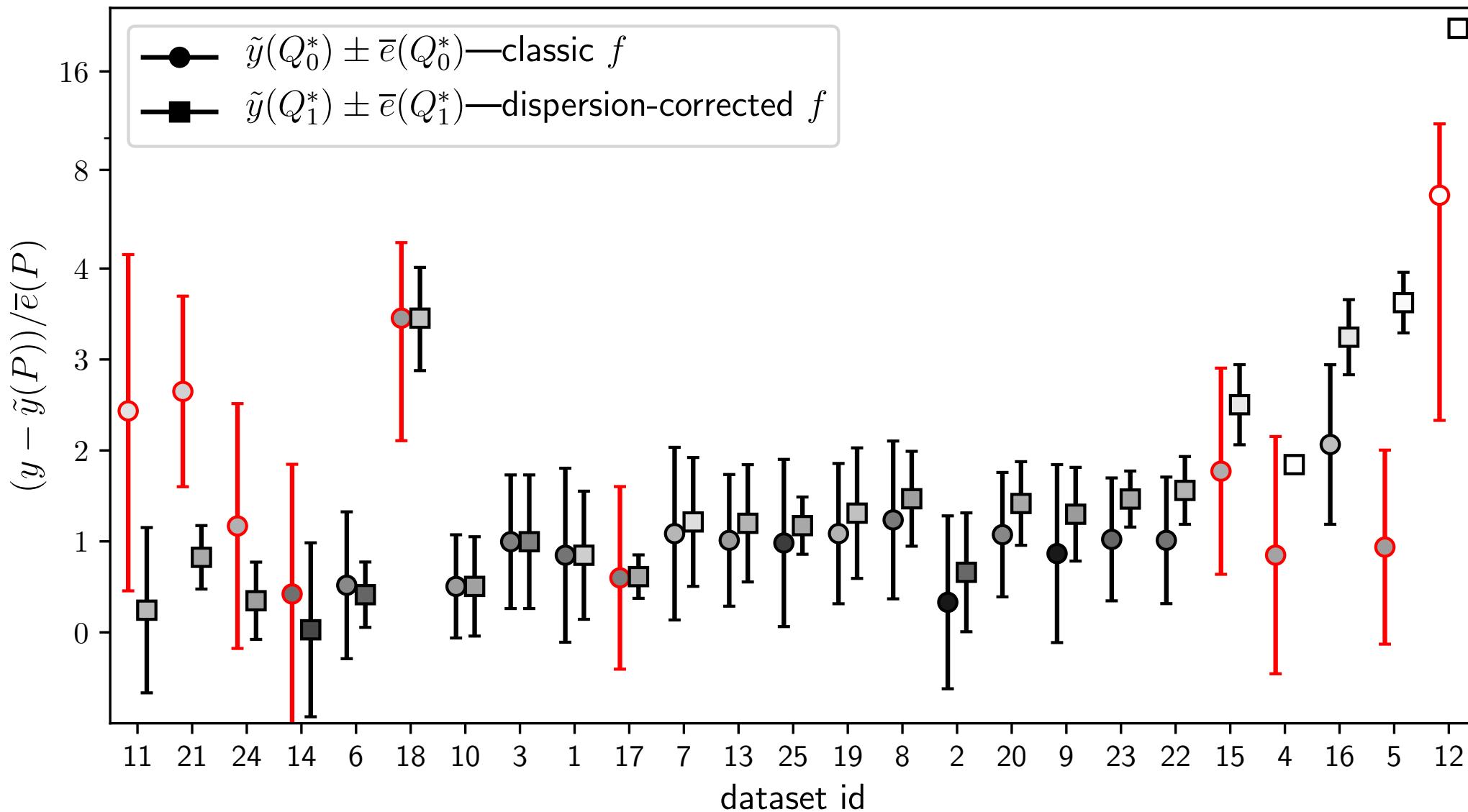
dispersion-corrected coverage

- well-behaved edge cases
- can be optimized w/o loosing performance (median case) !!



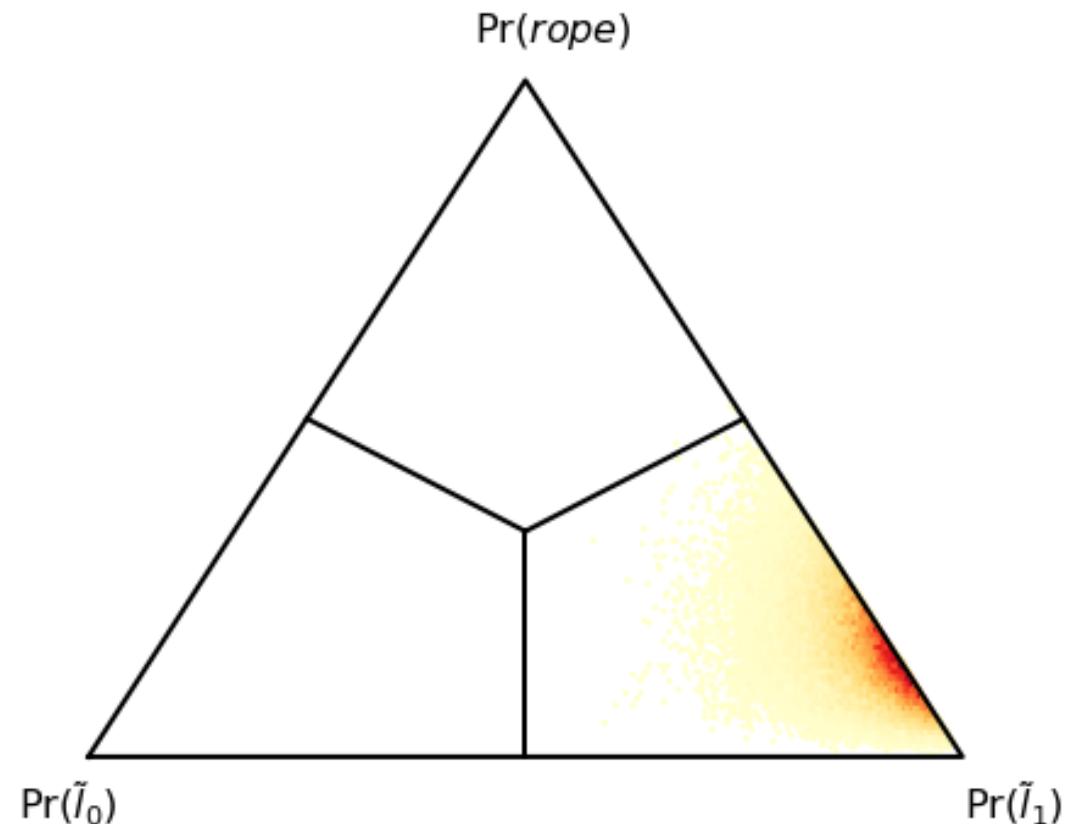
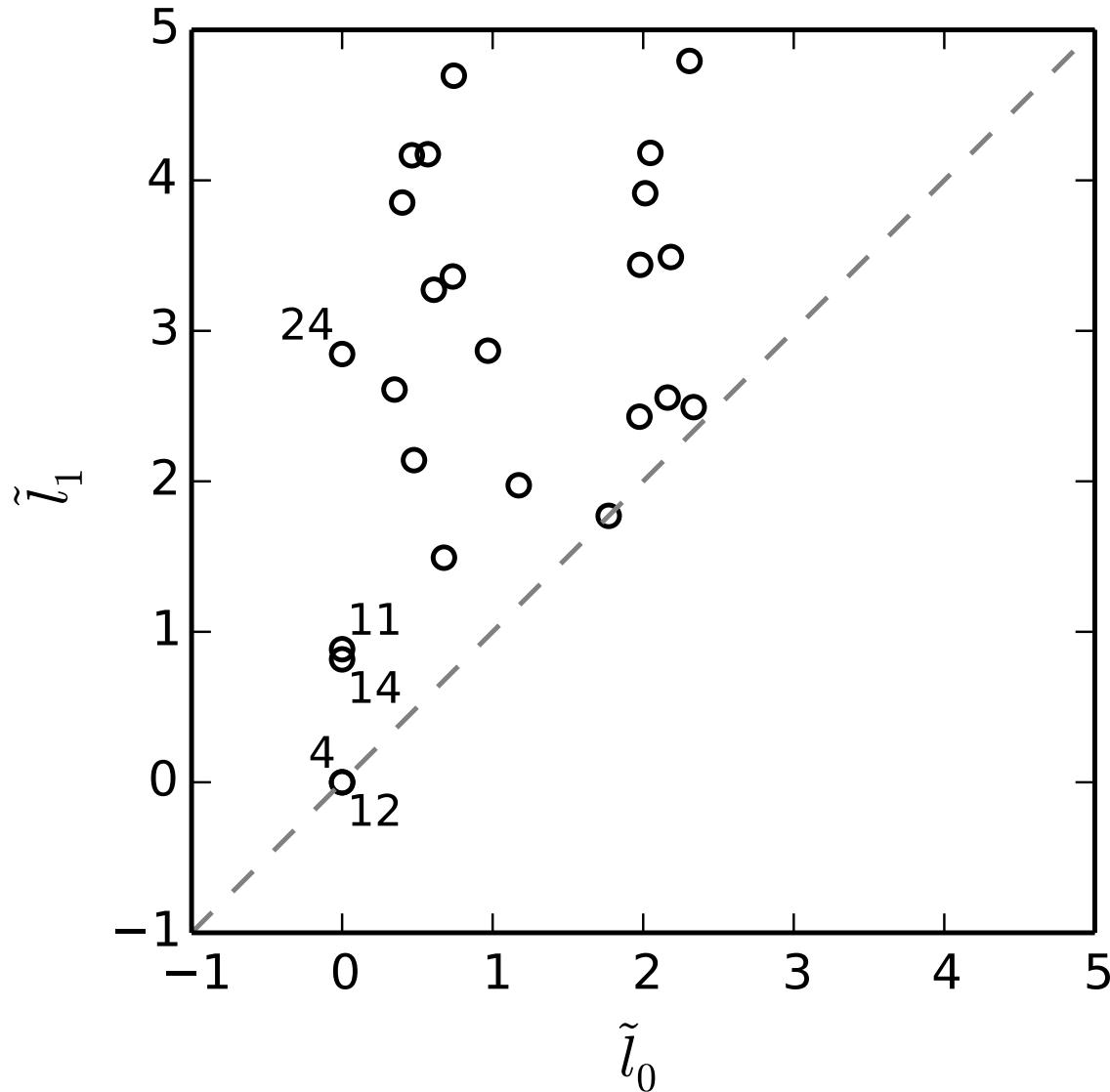
# Dispersion-correction reduces error

20



# Increases conservative mean estimates significantly

21



# Branch-and-bound subgroup optimization

22

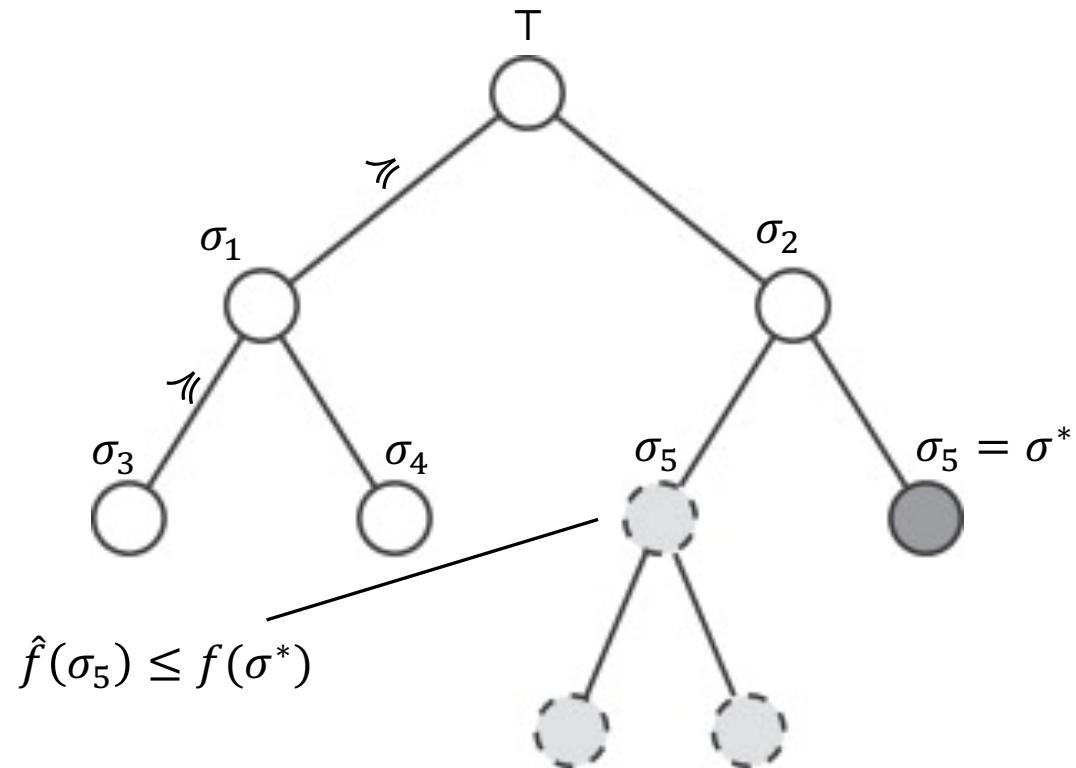
## Branch

$$\mathbf{r}: \mathcal{L}_x \rightarrow 2^{\mathcal{L}_x}$$

$$\begin{aligned}\varphi \in \mathbf{r}(\sigma) &\Rightarrow \sigma \leqslant \varphi \\ &\Rightarrow \text{ext}(\sigma) \supseteq \text{ext}(\varphi)\end{aligned}$$

## Bound

$$\hat{f}(\sigma) \geq \max\{f(\varphi) : \varphi \geqslant \sigma\}$$



[Wrobel 1997; Grosskreutz et al. 2008; Boley and Grosskreutz 2009]

# Branch-and-bound subgroup optimization

23

## Branch

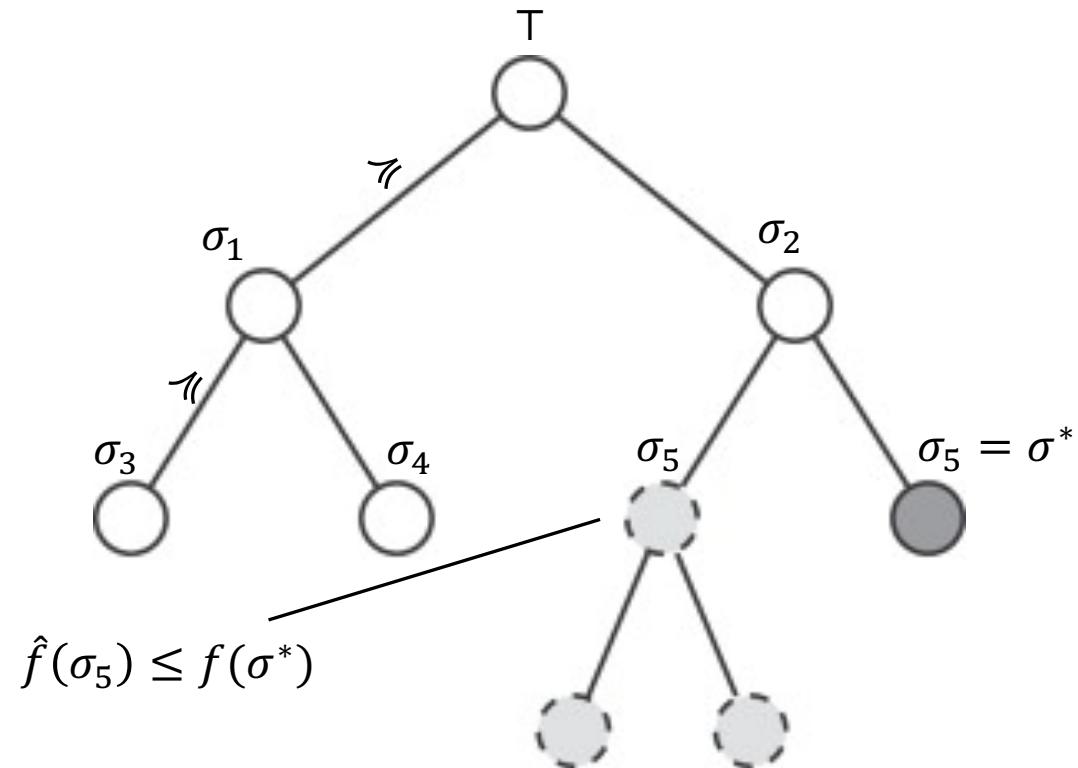
$$r: \mathcal{L}_x \rightarrow 2^{\mathcal{L}_x}$$

$$\begin{aligned}\varphi \in r(\sigma) &\Rightarrow \sigma \leq \varphi \\ &\Rightarrow \text{ext}(\sigma) \supseteq \text{ext}(\varphi)\end{aligned}$$

## Bound

$$\begin{aligned}\hat{f}(\sigma) &= \max\{f(R) : R \subseteq \text{ext}(\sigma)\} \\ &\geq \max\{f(\varphi) : \varphi \geq \sigma\}\end{aligned}$$

tight optimistic estimator



[Wrobel 1997; Grosskreutz et al. 2008; Boley and Grosskreutz 2009]

# How to compute tight opt est in linear time?

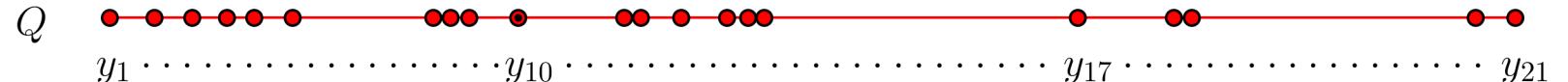
24

**Given**

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

$$Q = \{y_1, \dots, y_m\} \text{ st } y_1 \leq \dots \leq y_m$$

[cf. Lemmerich et al 2016]



# Linear size Pareto front through “top sequence”

25

Given

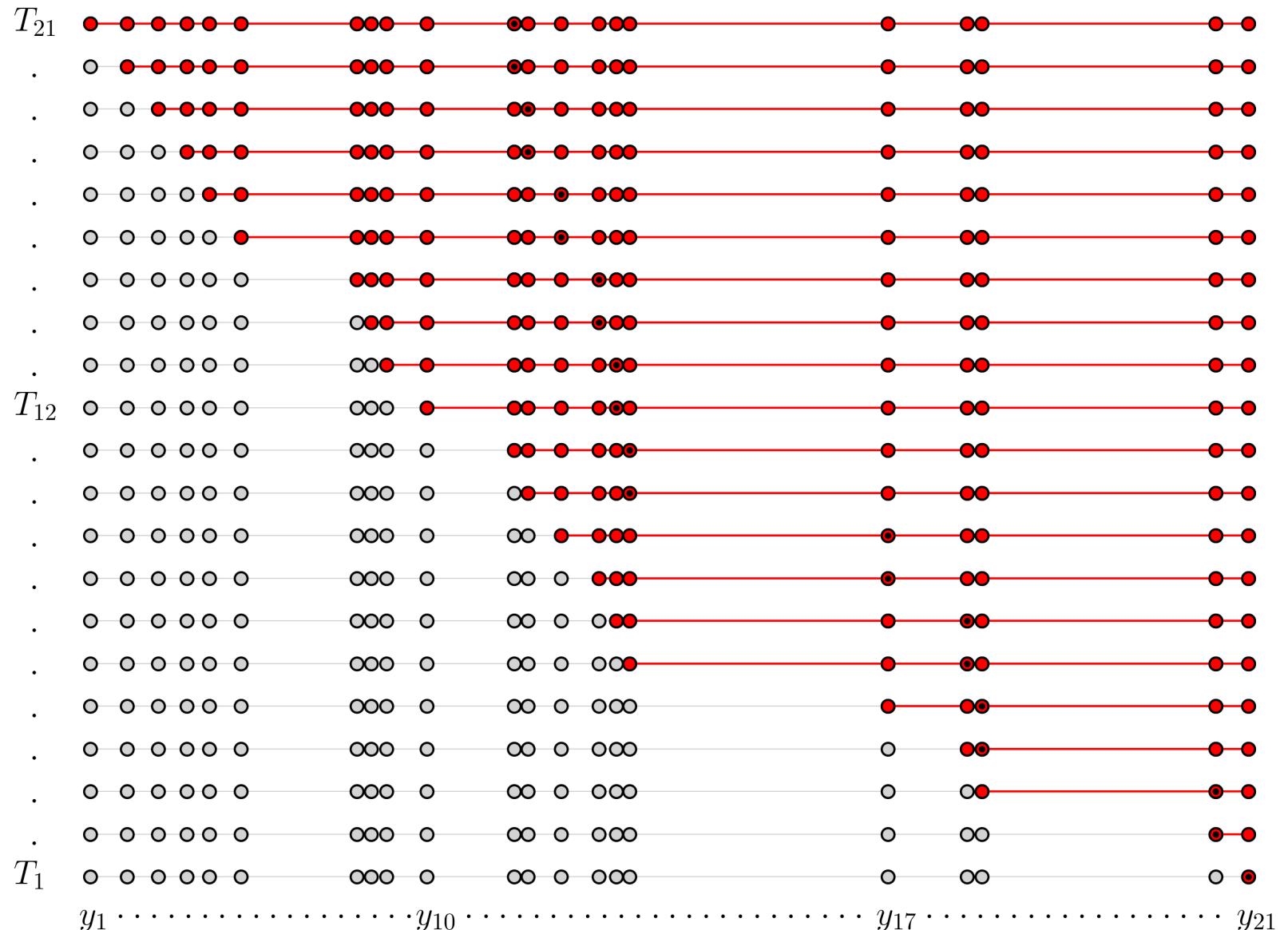
$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

$$Q = \{y_1, \dots, y_m\} \text{ st } y_1 \leq \dots \leq y_m$$

Compute  $\hat{f}$  in time  $O(m)$

$$\hat{f}(Q) = \max\{f(T_l) : 1 \leq l \leq m\}$$

$$T_l = \{y_{m-l+1}, \dots, y_m\}$$



[cf. Lemmerich et al 2016]

# Linear size Pareto front through “top sequence”

26

Given

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

$$Q = \{y_1, \dots, y_m\} \text{ st } y_1 \leq \dots \leq y_m$$

Compute  $\hat{f}$  in time  $O(m)$

$$\hat{f}(Q) = \max\{f(T_l) : 1 \leq l \leq m\}$$

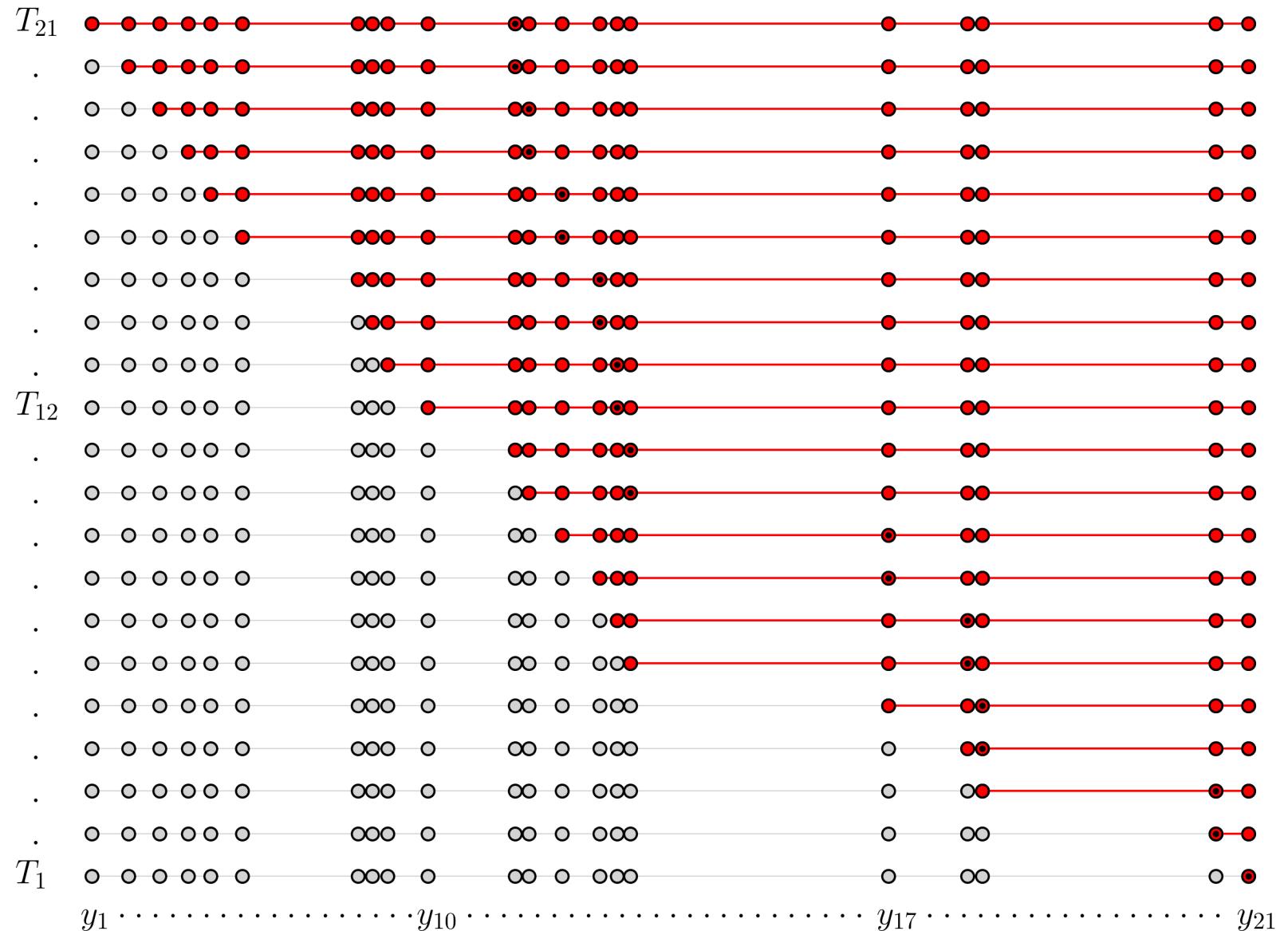
$$T_l = \{y_{m-l+1}, \dots, y_m\}$$

Using incremental  $O(1)$  ops

$$\text{cov}(T_l) = l/m$$

$$\bar{y}(T_{l+1}) = \frac{l\bar{y}(T_l) + y_{m-l}}{l+1}$$

[cf. Lemmerich et al 2016]



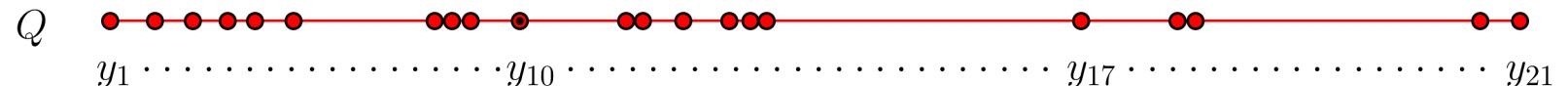
# How can we extend idea to dcc?

27

**Given**

$$f(Q) = g(\text{dcc}(Q), \text{eff}(Q))$$

$$Q = \{y_1, \dots, y_m\} \text{ st } y_1 \leq \dots \leq y_m$$



# Linear size Pareto front through median index

28

**Given**

$$f(Q) = g(\text{dcc}(Q), \text{eff}(Q))$$

$$Q = \{y_1, \dots, y_m\} \text{ st } y_1 \leq \dots \leq y_m$$

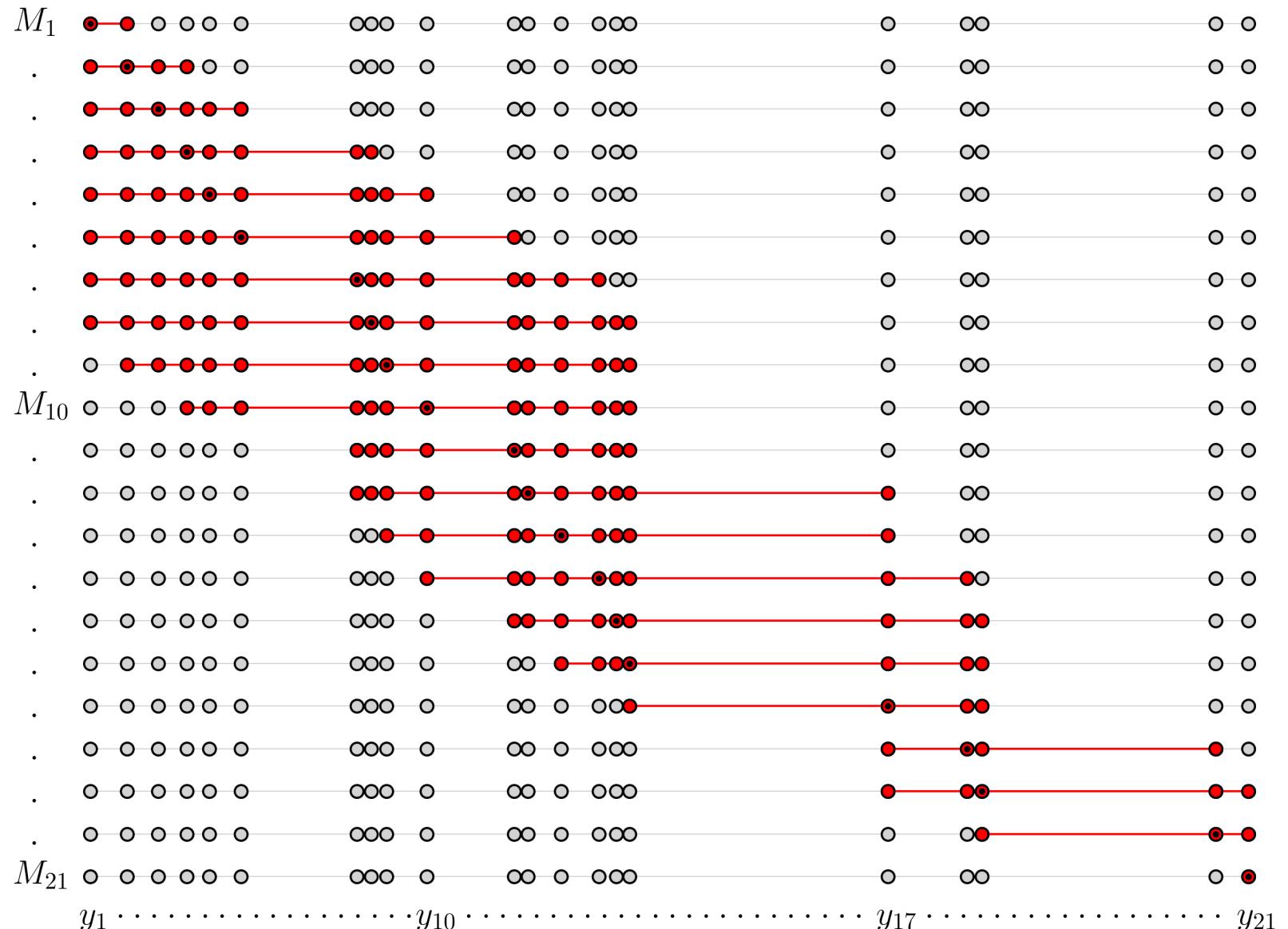
**Compute  $\hat{f}$  in time  $O(m)$**

$$\hat{f}(Q) = \max\{f(M_z) : 1 \leq z \leq m\}$$

$$M_z = M_z^{k_z^*}$$

$k_z^*$  maximizing  $\text{dcc}(M_z^k)$

$$M_z^k = \left\{ y_{z-\lceil \frac{k}{2} \rceil}, \dots, y_z, \dots, y_{z+\lceil \frac{k}{2} \rceil} \right\}$$



# Linear size Pareto front through median index

29

**Given**

$$f(Q) = g(\text{dcc}(Q), \text{eff}(Q))$$

$$Q = \{y_1, \dots, y_m\} \text{ st } y_1 \leq \dots \leq y_m$$

**Compute  $\hat{f}$  in time  $O(m)$**

$$\hat{f}(Q) = \max\{f(M_z) : 1 \leq z \leq m\}$$

$$M_z = M_z^{k_z^*}$$

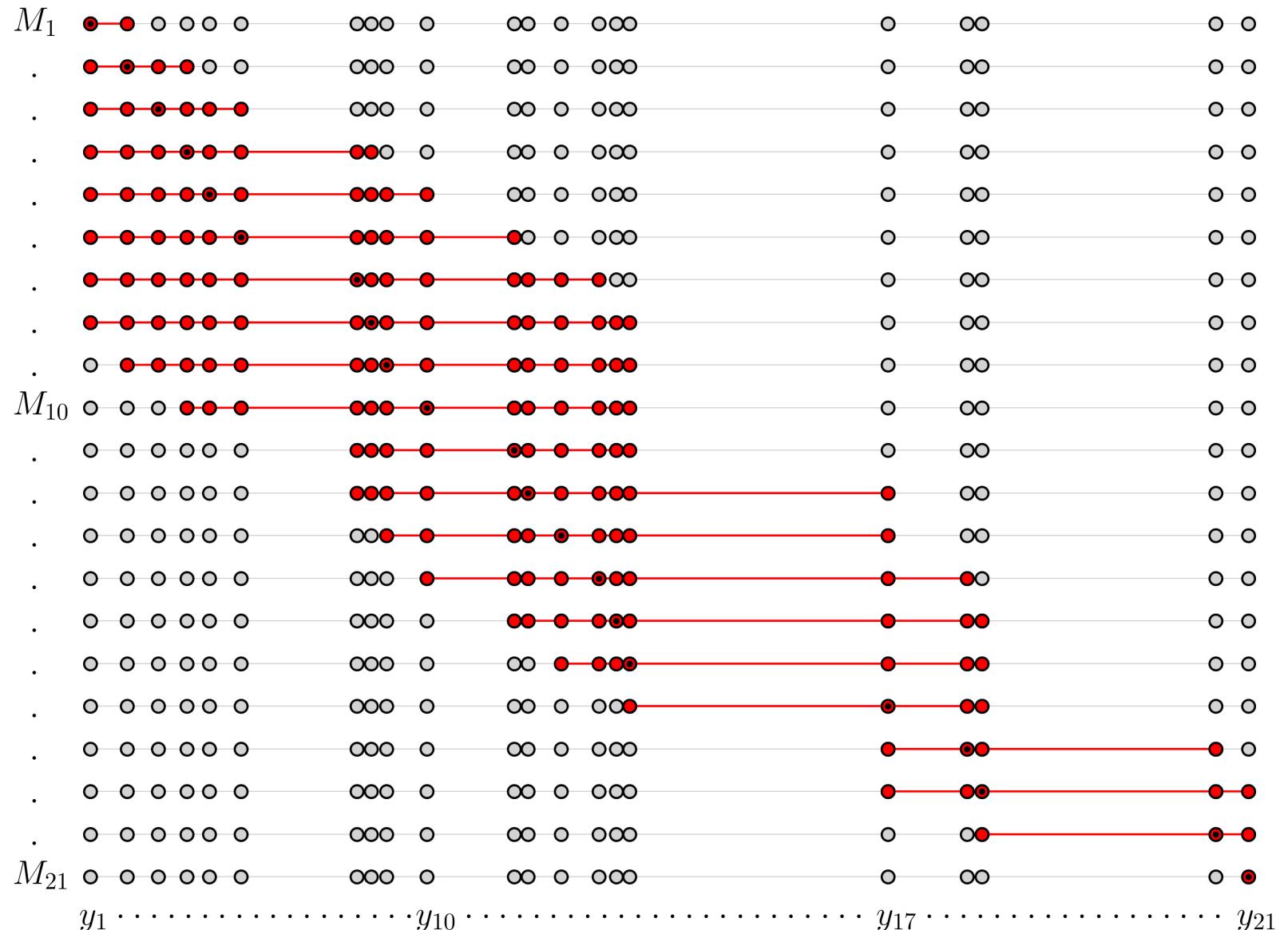
$k_z^*$  maximizing  $\text{dcc}(M_z^k)$

$$M_z^k = \left\{ y_{z-\lceil \frac{k}{2} \rceil}, \dots, y_z, \dots, y_{z+\lceil \frac{k}{2} \rceil} \right\}$$

**Using incremental  $O(1)$  ops**

$$\tilde{y}(M_z) = y_z$$

$$\text{dcc}(M_z) ???$$



# Incremental computation of $k_z^*$ in $O(1)$

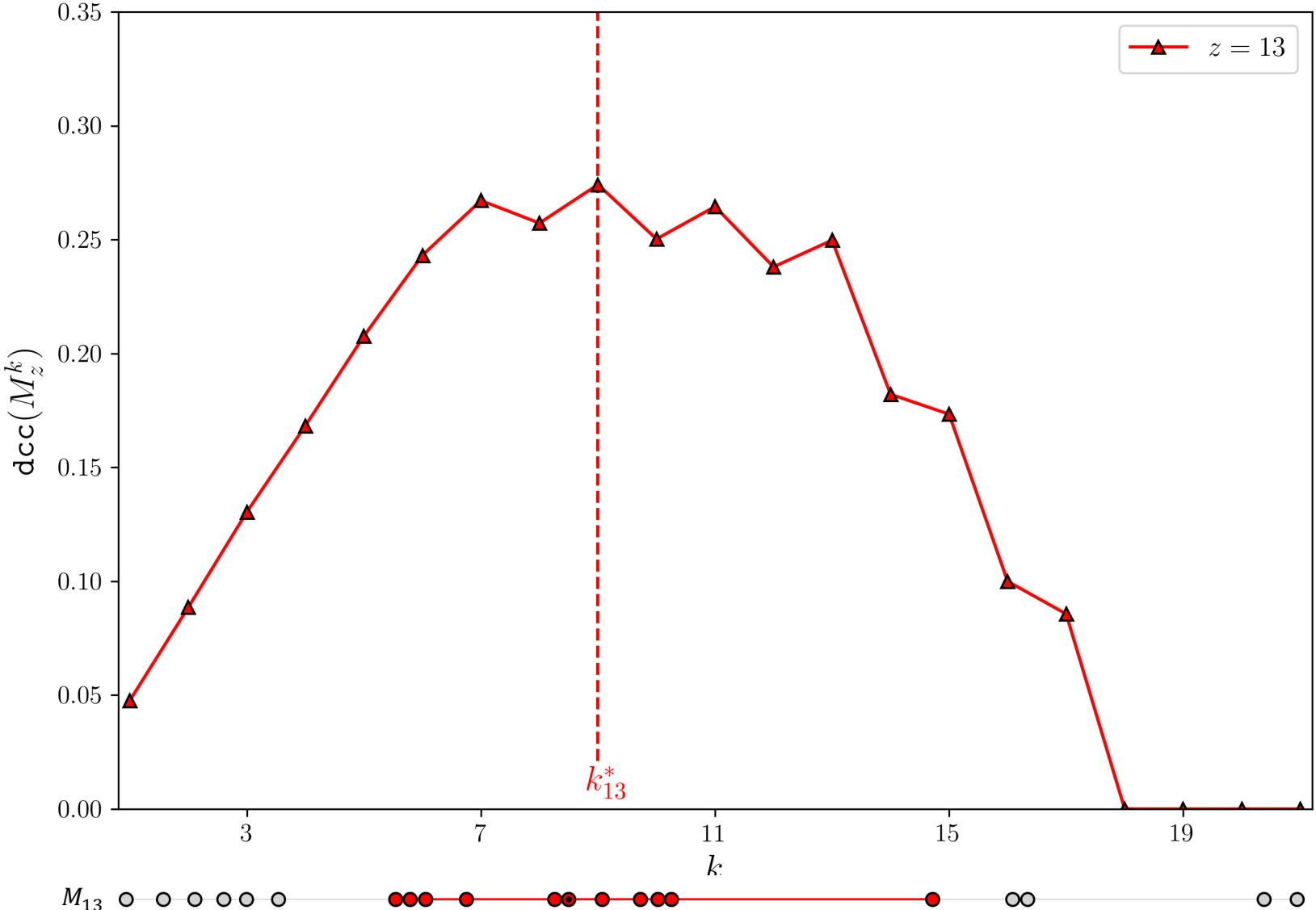
30

## Idea

Analyze functions

$$h_z: k \mapsto \text{dcc}(M_z^k)$$

$$\Delta h_z: k \mapsto (h_z(k) - h_z(k - 1))$$



# Incremental computation of $k_z^*$ in $O(1)$

31

## Idea

Analyze functions

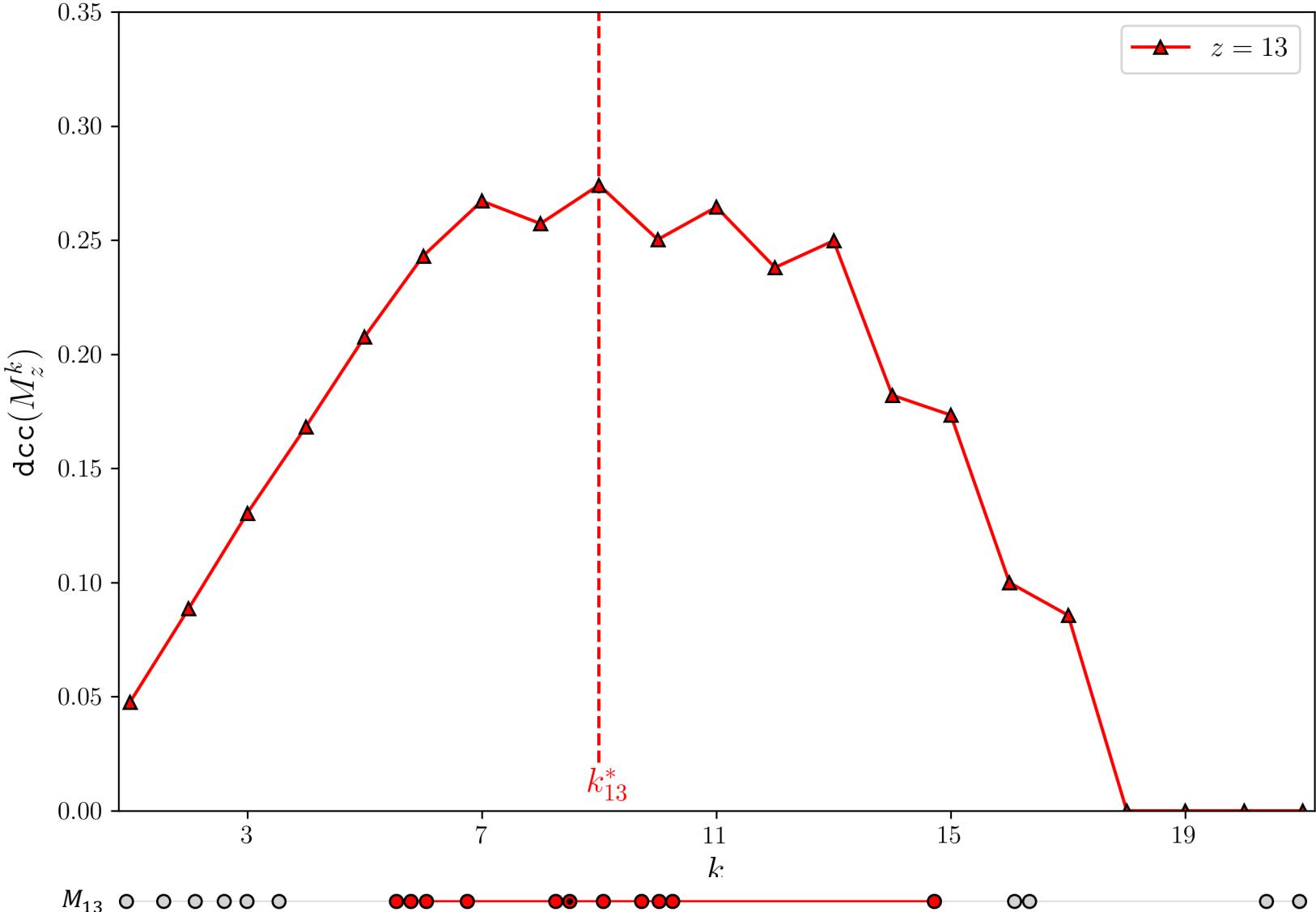
$$h_z: k \mapsto \text{dcc}(M_z^k)$$

$$\Delta h_z: k \mapsto (h_z(k) - h_z(k - 1))$$

## Observations

$h_z$  alternating concave, i.e.,

$$\begin{aligned} \Delta h_z(k + 1) + \Delta h_z(k) \\ \leq \Delta h_z(k) + \Delta h_z(k - 1) \end{aligned}$$



# Incremental computation of $k_z^*$ in $O(1)$

32

## Idea

Analyze functions

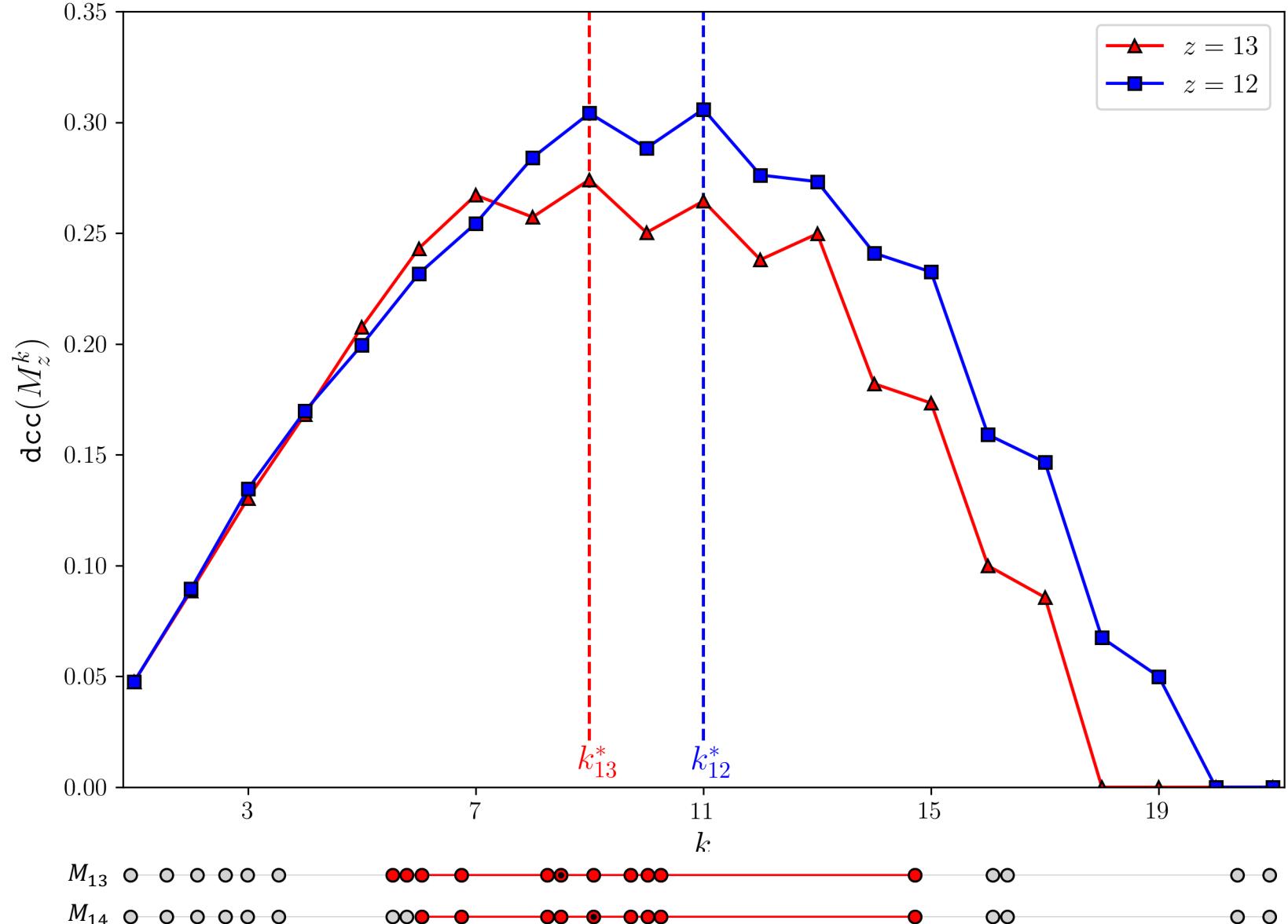
$$h_z: k \mapsto \text{dcc}(M_z^k)$$

$$\Delta h_z: k \mapsto (h_z(k) - h_z(k-1))$$

## Observations

$h_z$  alternating concave, i.e.,

$$\begin{aligned} \Delta h_z(k+1) + \Delta h_z(k) \\ \leq \Delta h_z(k) + \Delta h_z(k-1) \end{aligned}$$



# Incremental computation of $k_z^*$ in $O(1)$

33

## Idea

Analyze functions

$$h_z: k \mapsto \text{dcc}(M_z^k)$$

$$\Delta h_z: k \mapsto (h_z(k) - h_z(k-1))$$

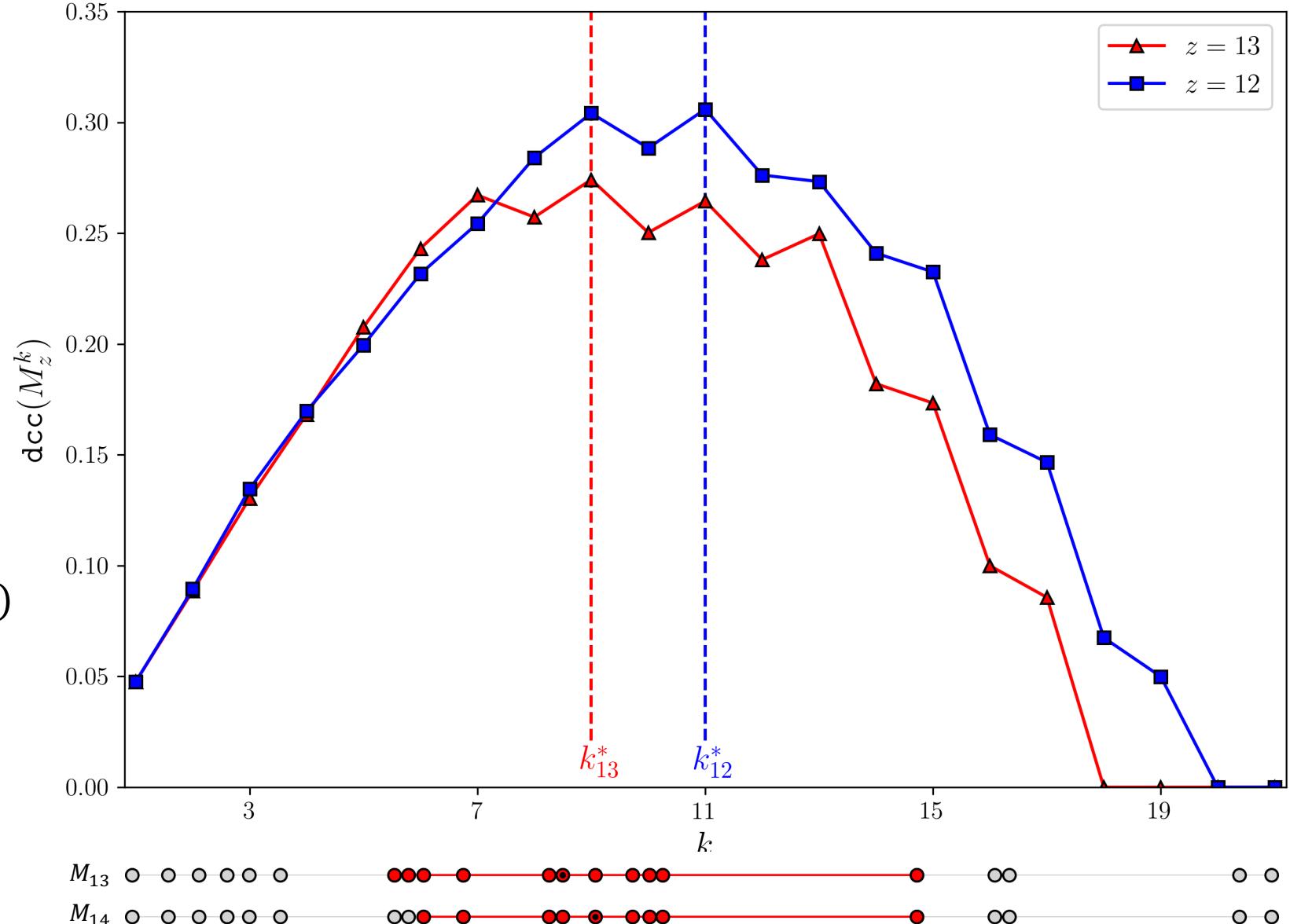
## Observations

$h_z$  alternating concave, i.e.,

$$\begin{aligned} \Delta h_z(k+1) + \Delta h_z(k) \\ \leq \Delta h_z(k) + \Delta h_z(k-1) \end{aligned}$$

$\Delta h_{z+1}$  and  $\Delta h_z$  are coupled, i.e.,

$$\begin{aligned} \Delta h_{z-1}(k) + \Delta z_{-1}(k-1) \\ \leq \Delta h_z(k-2) + \Delta z(k-3) \end{aligned}$$



# Incremental computation of $k_z^*$ in $O(1)$

34

## Idea

Analyze functions

$$h_z: k \mapsto \text{dcc}(M_z^k)$$

$$\Delta h_z: k \mapsto (h_z(k) - h_z(k-1))$$

## Observations

$h_z$  alternating concave, i.e.,

$$\begin{aligned} \Delta h_z(k+1) + \Delta h_z(k) \\ \leq \Delta h_z(k) + \Delta h_z(k-1) \end{aligned}$$

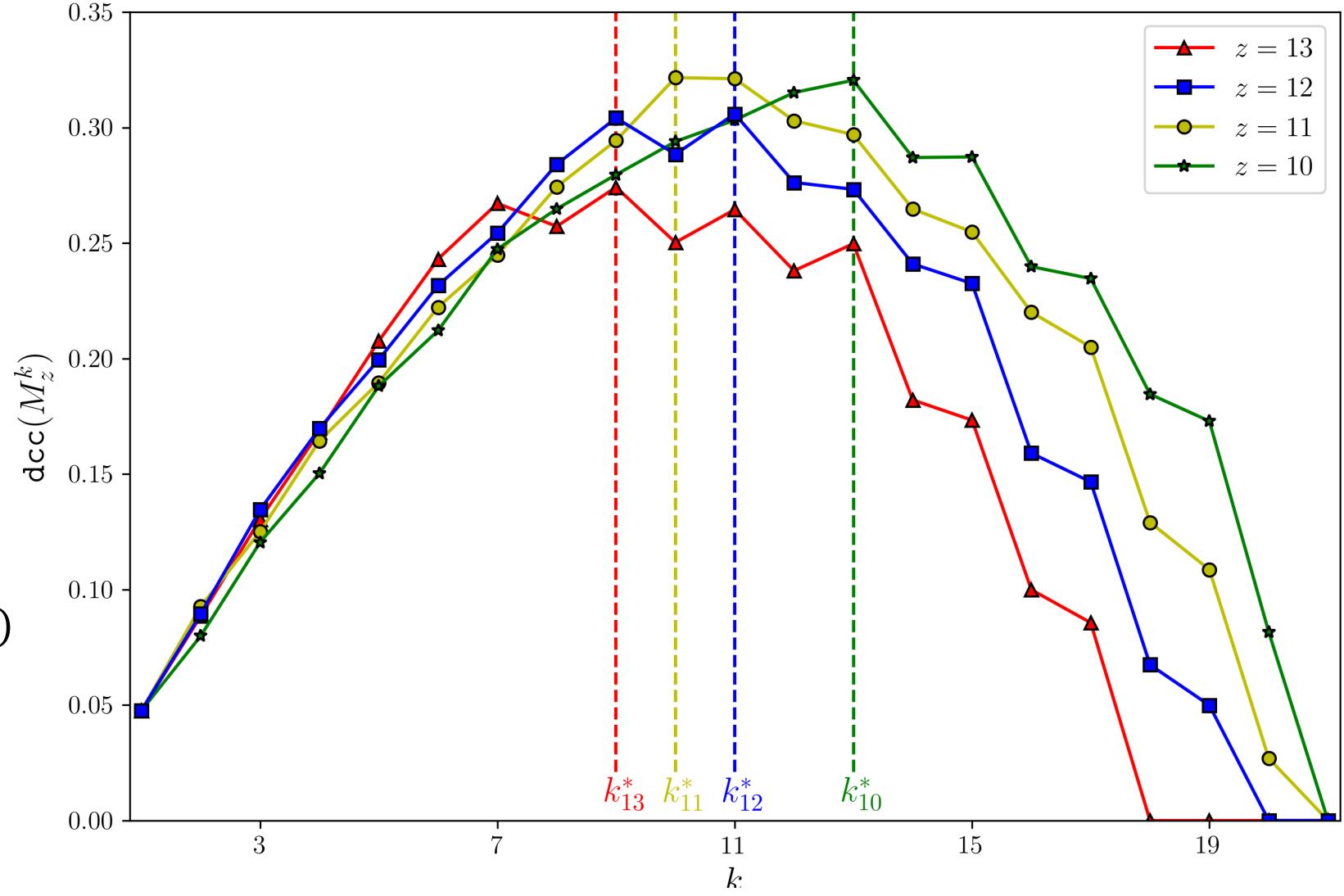
$\Delta h_{z+1}$  and  $\Delta h_z$  are coupled, i.e.,

$$\begin{aligned} \Delta h_{z-1}(k) + \Delta z_{-1}(k-1) \\ \leq \Delta h_z(k-2) + \Delta z(k-3) \end{aligned}$$

## Conclusion

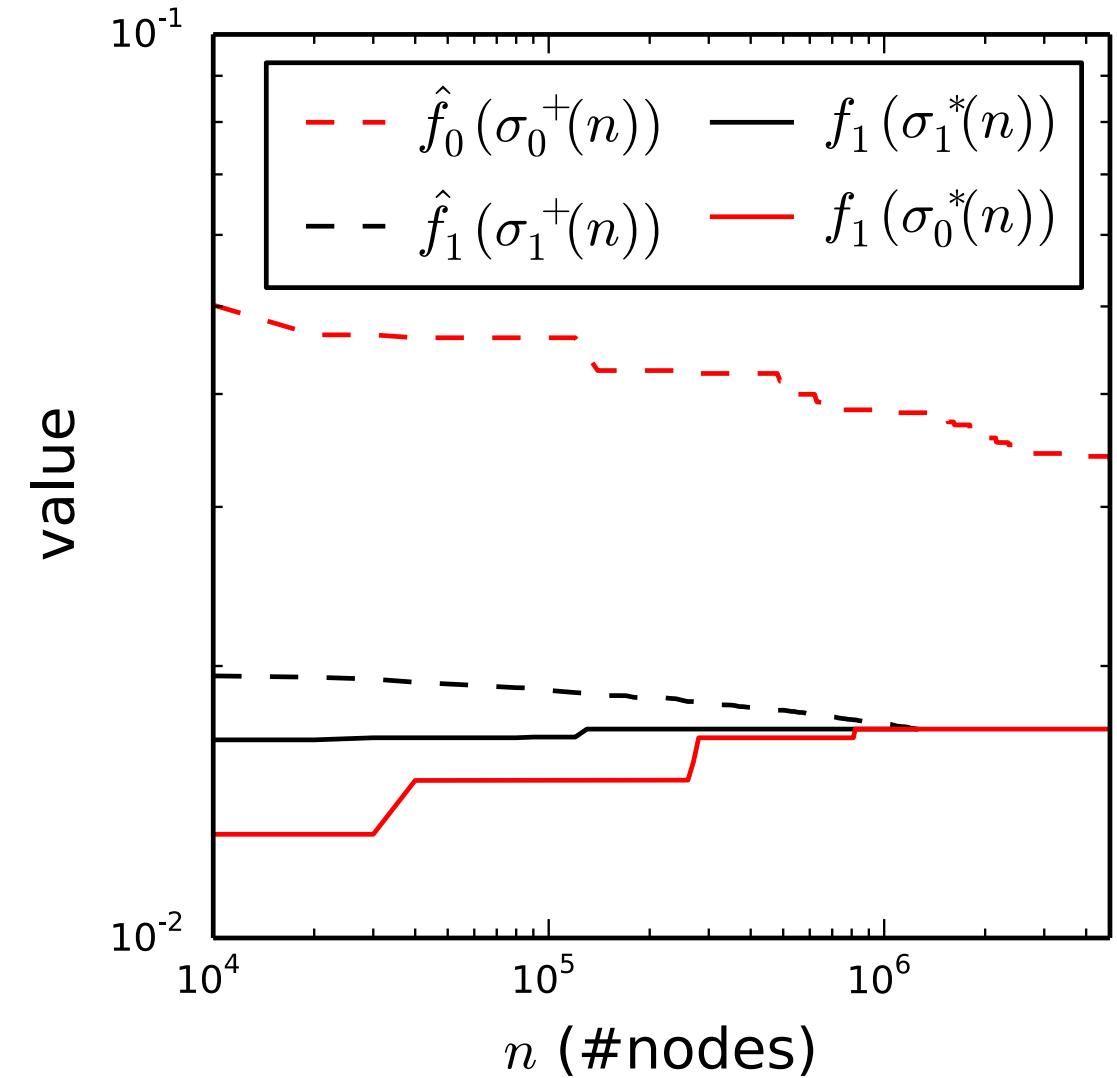
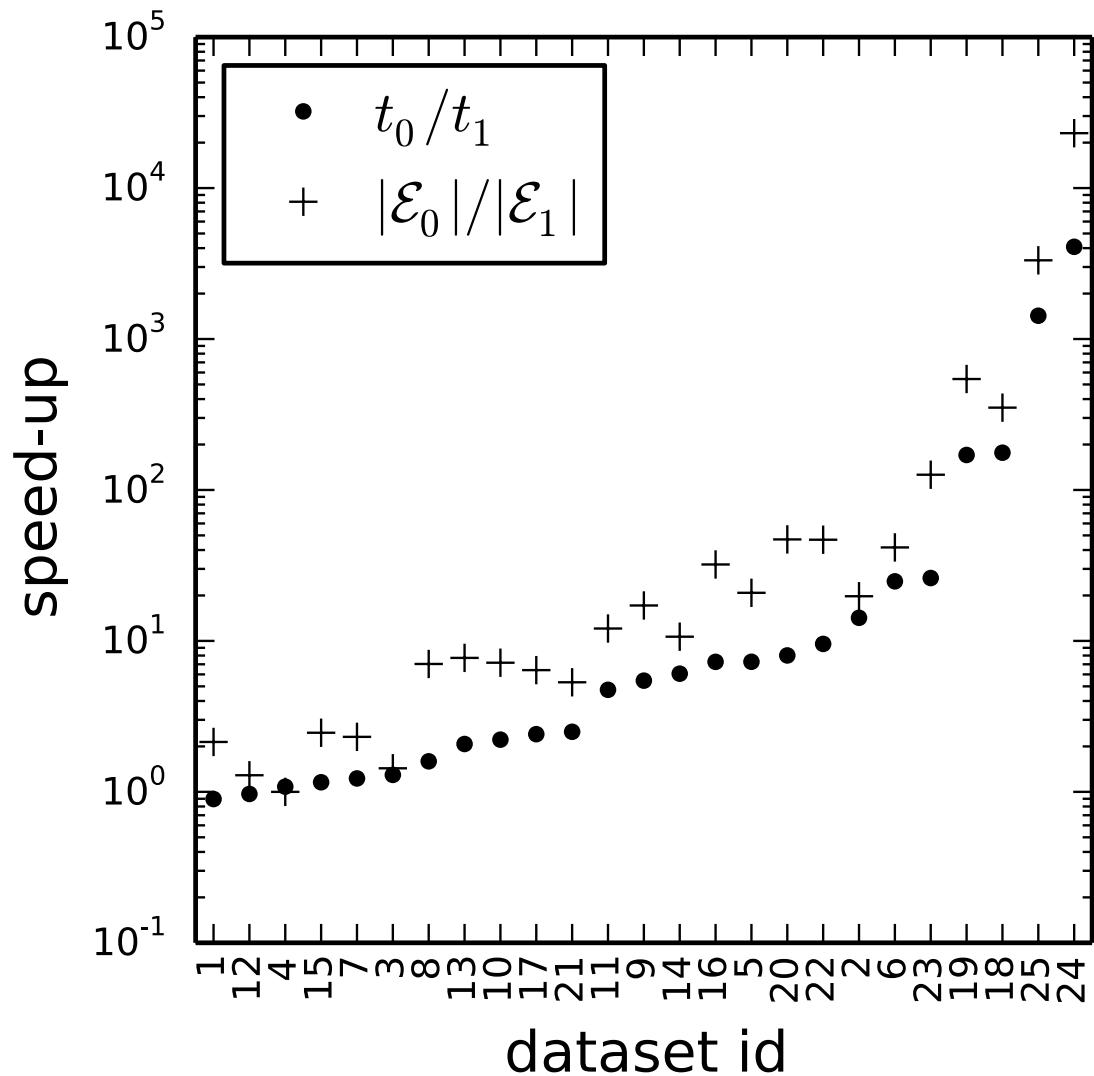
$$k_z^* \in \{k_{z+1}^* - 3, \dots, k_{z+1}^* + 3\}$$

-> incremental  $O(1)$  computation



# Gains of tight estimator over top-sequence-based

35



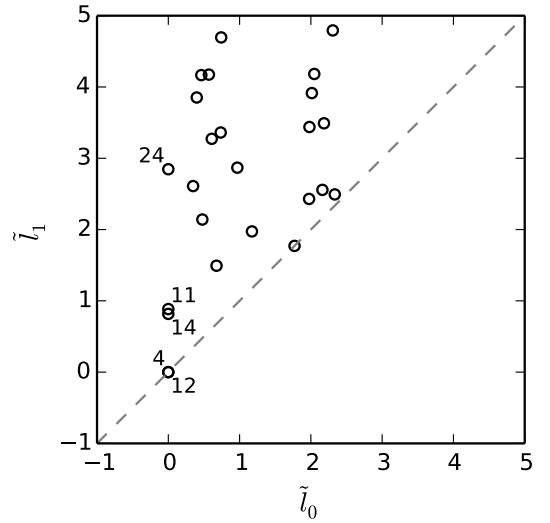
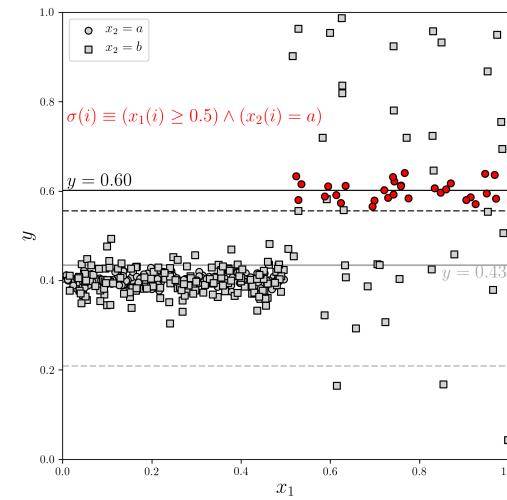
# Conclusion

36

## Summary

dispersion/error is an issue

dispersion-corrected coverage addresses it  
can be optimized effectively (median case)



## Directions

immediate: what about mean case?

bigger picture: complex models / multiple targets

