Beyond Empirical - Advanced Estimation Methods for Optimal Cutpoints

Christian Thiele & Gerrit Hirschfeld

03 September 2018

Hochschule Osnabrück
University of Applied Sciences

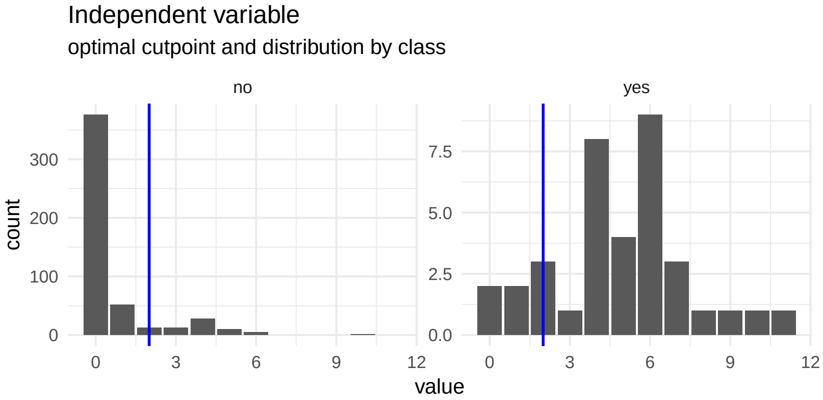
Outline

1. What do we want "optimal" cutpoints for?
2. Problems with optimal cutpoints
3. Cutpoint estimation methods:
    - Nonparametric Empirical
    - Normal and Transformed Normal
    - Kernel
    - LOESS
    - Splines
    - Generalized Additive Models
4. Simulation to assess estimation quality
5. Conclusion

What do we want ‘optimal’ cutpoints for?

An “optimal” cutpoint c^* allows for binary classification via:

- Biological markers
- Psychological scores
- Model predictions



Problems with 'optimal' cutpoints

- Selecting the "optimal" cutpoint $\hat{c}^*$ by trying out all possible ones leads to
  - highly variable cutpoints
  - overestimation of accuracy
  - = "overfitting"
- However, this "traditional" empirical method (EMP) is the most popular one

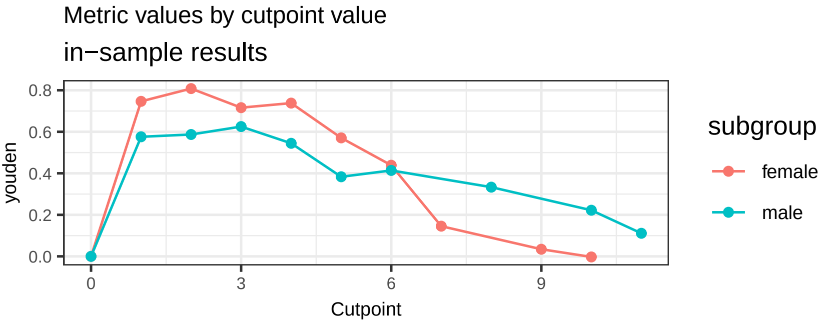
(Apparently) most popular metric: The Youden-Index J with

$$\begin{aligned} J &= max\{\frac{TP(c)}{TP(c)+FN(c)} + \frac{TN(c)}{TN(c)+FP(c)} - 1\} \\ &= max\{(1 - G_D(c)) + F_H(c) - 1\} \\ &= max\{Se(c) + Sp(c) - 1\} \end{aligned}$$

Empirical method (EMP)

1. Sort unique values of predictor x
2. Use these values as cutoffs c to split data in positive / negative class
3. Calculate metric $m(c)$ at every one of those values
4. Maximize $m(c)$ numerically

```
cutpointr(suicide, dsi, suicide, gender, metric = youden) %>%
    plot_metric() + theme_bw() + theme_smallertext
```



Possible solutions

The $\hat{c}^*$ obtained from EMP are unbiased but suffer from high variance.

Several possible solutions to lower the variance:

- Distributional assumptions and parametric estimation methods
- Smoothing of $m(c)$

Method based on assuming normal distributions (N)

If the predictor can be assumed to follow a normal distribution in both classes, a low variance method for calculating the 'optimal' cutpoint for maximizing sens + spec (or equivalently the Youden-Index J) is

$$c^* = \frac{(\mu_D \sigma_H^2 - \mu_H \sigma_D^2) - \sigma_H \sigma_D \sqrt{(\mu_H - \mu_D)^2 + (\sigma_H^2 - \sigma_D^2) log(\sigma_H^2 / \sigma_D^2)}}{\sigma_H^2 - \sigma_D^2}$$

Advantages:

- Computable from summary statistics, even if no access to original data
- Low computational burden

Disadvantages:

- Distributional assumptions must be made

Transformed Normal method (TN)

If the predictor is not normally distributed but can be transformed to normality using a Box-Cox type of transformation t

$$t(x) = x^{(\lambda)} = \begin{cases} (x^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ log(x) & \lambda = 0 \end{cases}$$

N can be used and the resulting cutpoint can be transformed back to the original scale by applying the reverse of t.

Bivariate Lambda for TN

- We need a common λ for the back-transformation of c^* into the original scale

Zou et al. assume a binormal model and construct the profile log-likelihood function

$$l(\lambda|x_1,\ldots,x_m,y_1,\ldots,y_n) = -m * log(s'_x) - n * log(s'_y) \\ + (\lambda-1)\Big[\sum_{i=1}^{m} log(x_i) + \sum_{j=1}^{n} log(y_j)\Big] + c$$

where c is a constant, x_i and y_j are the diseased and healthy samples, s'_x and s'_y are the sample standard deviations of the transformed diseased and healthy samples and m and n are the sizes of the diseased and healthy samples respectively.

Bootstrapped c^*

Classical nonparametric bootstrap (B):

1. Resample the data B times (e.g. $B = 1000$) per class with replacement
2. Calculate $\hat{c}_b^*$ in every resample $b = 1, \ldots, B$ via EMP
3. Calculate the mean of $\hat{c}_b^*$ and use it as $\hat{c}^*$

Advantages:

- No tuning parameters
- Suitable for optimization of any metric m (e.g. misclassification cost, $|Se - Sp|$, accuracy, $\ldots$)

Disadvantages:

- Higher computational burden

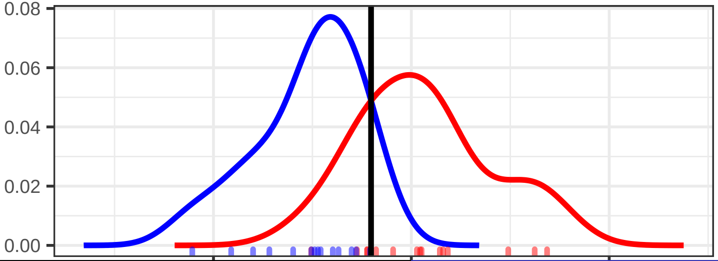
Kernel method (K)

- Nonparametric smoothing of the cdfs of the “diseased” and “healthy” samples
- Gaussian kernel function (following Fluss et al.)
- Bandwidth selection via direct plug-in method
- Here, the implementations from `KernSmooth` are used

Maximize $\hat{J} = max_c\{\hat{F}_H(c) - \hat{G}_D(c)\}$ numerically to find $\hat{c}^*$ that maximizes J where $\hat{F}_H$ and $\hat{G}_D$ are the Kernel density estimates of the cdfs of the healthy and diseased populations respectively.

Advantages: No tuning parameters and low computational burden

Optimal cutpoint based on kernel smoothed densities

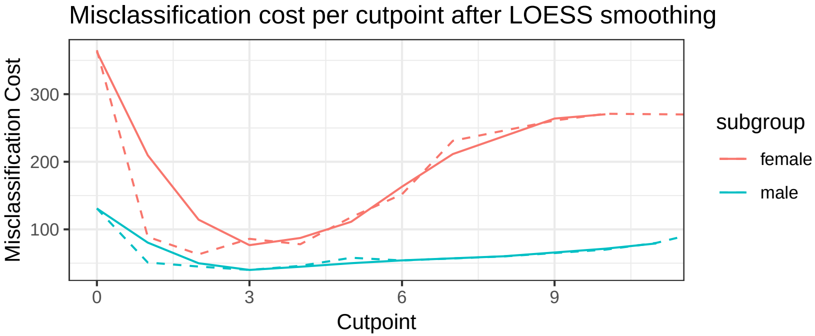


Loess method (L)

- Local polynomial regression to smooth $m(c)$ with automatic smoothing parameter selection

Advantages:

- No manual tuning necessary
- Suitable for any metric m
- Low computational burden



Spline smoothing (S)

Spline smoothing finds a cubic spline g that minimizes

$$\sum_{1}^{n}(y_i - g(x_i))^2 + \lambda \int_{-\infty}^{+\infty} [g^{''}(x)]^2 dx$$

where the smoothing parameter λ is a fixed tuning constant.

Advantages:

- Intuitively appealing as the true $m(c)$ is usually expected to be quite smooth and large second derivatives can be directly penalized via λ
- Suitable for any metric function m
- Low computational burden

Disadvantage:

- No value of λ can be generally recommended -> may have to be tuned (e.g. via bootstrapping)
- Most attractive for the case of many unique cutpoints in which a large smoothing parameter can be chosen

GAM smoothing (G)

The default GAM is of the form

$$m_i \sim f(c_i) + \epsilon_i$$

where m are the metric values per cutpoint c, f is a thin plate regression spline and ϵ is i.i.d. $N(0, \sigma^2)$. Smoothing parameter selection by GCV (implementation from package `mgcv`).

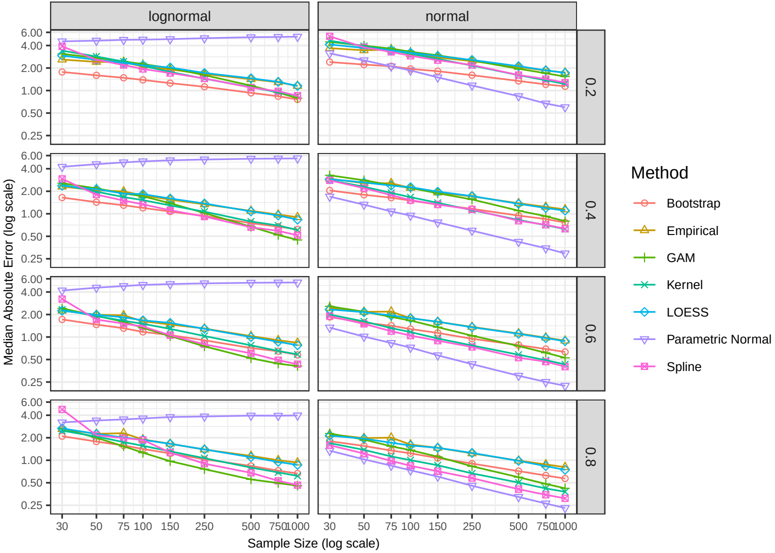
Advantages:

- No tuning necessary
- Suitable for any metric m
- Low computational burden

Simulation setup

- Three types of distributions (normal, log-normal, and gamma)
- Four different levels of separation (Youden-index values in the population of 0.2, 0.4, 0.6 and 0.8) achieved by manipulating the mean of the diseased sample
- In all scenarios the prevalence was held constant at 50 percent without loss of generalizability
- The overall sample sizes were 30, 50, 75, 100, 150, 250, 500, 750 and 1.000
- ...resulting in 108 different scenarios
- 10.000 repetitions of each scenario
- We compared the optimal cutpoint identified by the different methods to the true optimal cutpoint and calculated the median absolute error $MAE = med\{|\hat{c}_i^* - c^*|\}$.
- We always use midpoints instead of exact values (`use_midpoints = TRUE` in `cutpointr`), then all methods are unbiased

Simulation results



Conclusion

- EMP is easy to understand but usually inferior to the other methods
- B is also easy to understand and an improvement. Particularly good with small samples and effect sizes
- GAM is superior to B with large samples (here, also large numbers of unique cutpoints) and effect sizes
- Parametric CIs not available for all methods
- Best tuning method unclear for methods that need tuning
- Reporting of results from smoothing methods not as straightforward as from EMP (report both)

Thank you

Funding: BMBF Indimed

References:

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. Biometrical Journal, 47(4), 458–472.

Leeflang, M. M., Moons, K. G., Reitsma, J. B., & Zwinderman, A. H. (2008). Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clinical Chemistry, (4), 729–738.

Von Glischinski, M., Teisman, T., Prinz, S., Gebauer, J., and Hirschfeld, G. (2017). Depressive Symptom Inventory- Suicidality Subscale: Optimal cut points for clinical and non-clinical samples. Clinical Psychology & Psychotherapy

Appendix

Simulation scenarios

Distribution	μ_h	σ_h	σ_d	μ_d per J			
				0.2	0.4	0.6	0.8
normal	100	10	10	105.05	110.49	108.42	112.82
lognormal	2.5	0.5	0.5	2.76	3.02	3.34	3.78

Table 1: Simulation scenarios.