Bootstrap Corrected Limits for Rare Signals

Wolfgang A. Rolke
Angel M. López
University of Puerto Rico
at Mayaguez

- **Motivation: we are choosing a cut based on some optimality criterion related to the signal to noise ratio, but this cut is optimal for the data, not the underlying distribution.**
- **Doing a blind analysis solves a large part of this problem, but not all of it.**
- **How large is the bias? Do we need to worry about it?**
- **Can we reduce or eliminate the bias?**

Table of Contents

- **Why there might be a bias**
- **A well-known example: curve fitting**
- **The size of the bias in the search for rare signals**
- **Basic idea for adjusting the bias: split samples**
- **A short reminder of the bootstrap**
- **How we use it here: the double bootstrap**
- **A simulation that shows the effectiveness of the double bootstrap**

Why there might be a bias

- **Step 1: Find optimal cut**

  → without looking at the signal region

  → maximize signal efficiency and minimize predicted background

  → optimization step

- **Step 2: Find the branching ratio based on cut from Step 1**

  → use predicted background, find confidence limits

  → evaluation step

One well known example: curve fitting

- **Problem: Fit a parametric curve to the data**

- **Step 1: Find parameter estimates based on the data**

  $\rightarrow$ optimization step

- **Step 2: Find the χ^2 statistic**

  $\rightarrow$ evaluation step

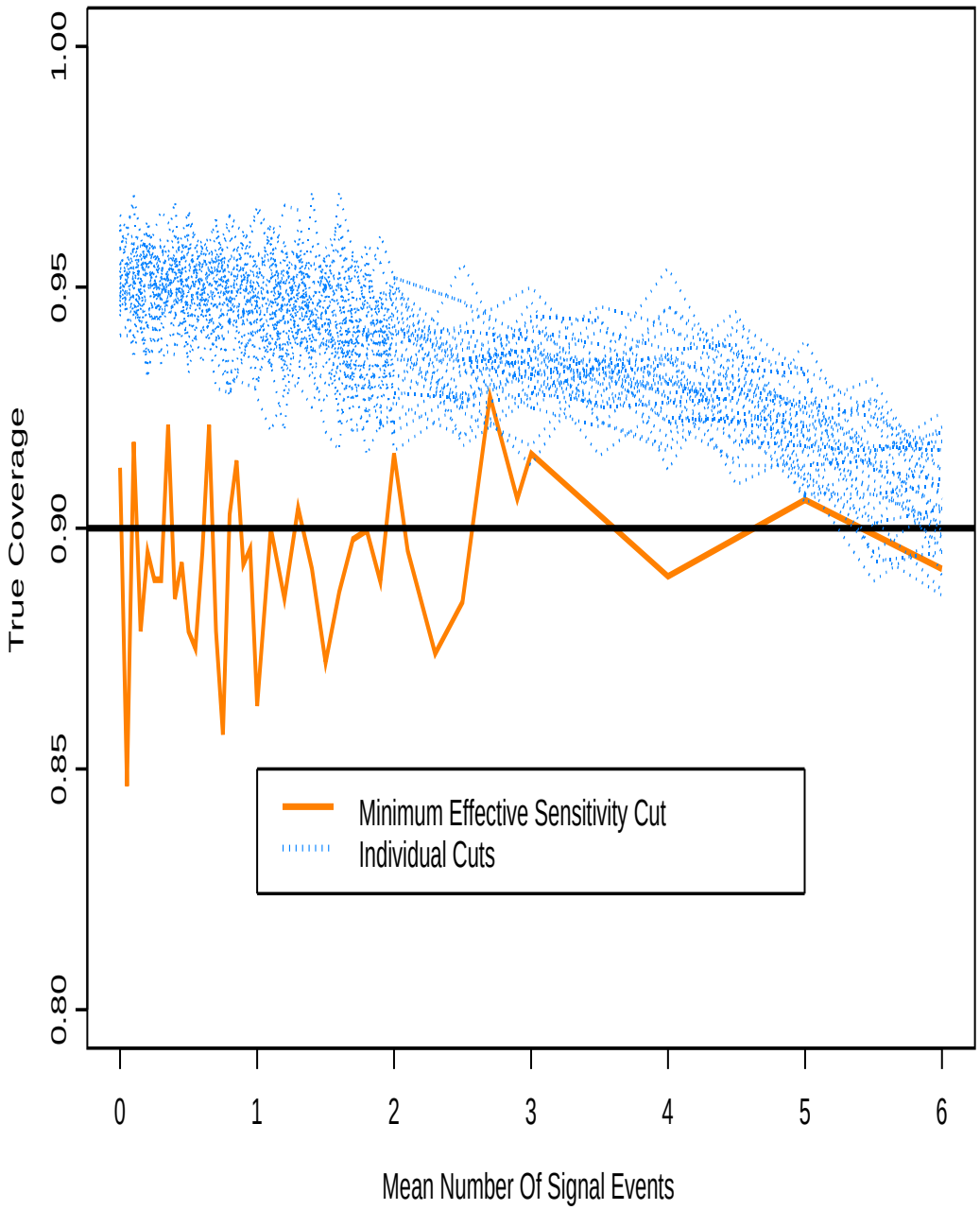
- **How to fix the bias problem in this case: adjust degrees of freedom**

An example: The search for rare signals

- **We use data from FOCUS E831: $D^0 \rightarrow \mu^+\mu^-$**
- **Simulate background with a sample of "non-muonic" pair events weighted by their momentum-dependent misidentification probability. Use loosest analysis cuts.**
- **Simulate dimuon signal with FOCUS Monte Carlo.**
- **Fake data sets are generated by randomly choosing *n* events from the simulated signal set and *m* events from the simulated background set.**
  - ***n* is Poisson distributed with rate λ.**
  - ***m* is Poisson distributed with rate 25.0 as in the real data.**
- **A large number of fake data sets are generated to study the problem.**

Minimization Criterion

- **Find cut with lowest *effective sensitivity* from the set of all cut combinations (over 13000 in our simulation).**
- **Sensitivity is the average upper limit for an experiment with no signal and the predicted background rate.**
- **Same as Feldman-Cousins, only using Rolke-Lopez limits (background rate is not known, has to be estimated from the sample).**
- ***Effective* means efficiency corrected.**


True Coverage
1.00
0.95
0.90
0.85
0.80
0
1
2
3
4
5
6
Mean Number Of Signal Events
Minimum Effective Sensitivity Cut
Individual Cuts

What can we do about this?

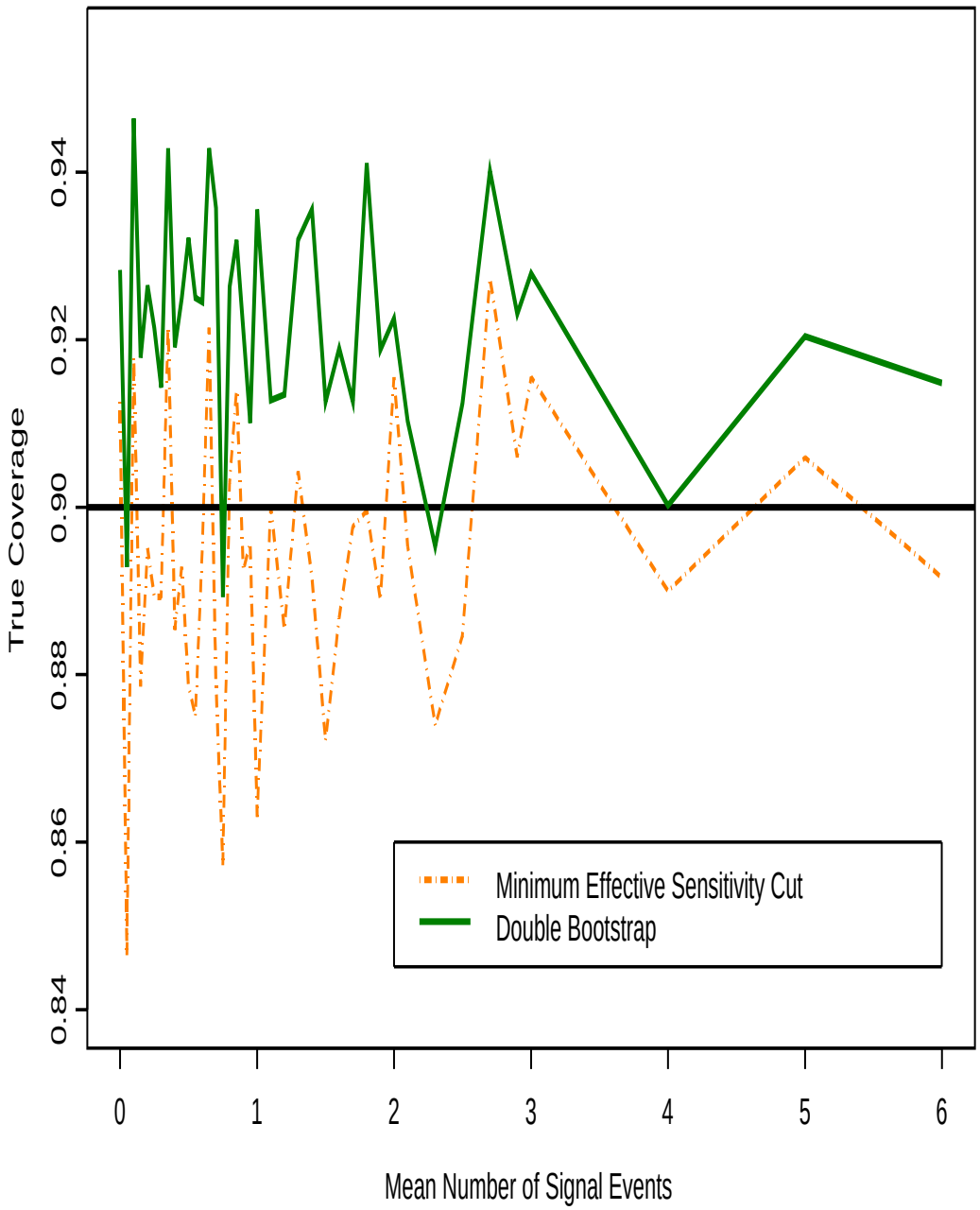
- **Idea: Decouple the optimization and the evaluation step**
- **One way to do this: split samples**
- **But sample size is already very small**
- **Instead: use one bootstrap sample to find best cut**
- **Use a second bootstrap sample to find branching ratio**

Remember the Bootstrap

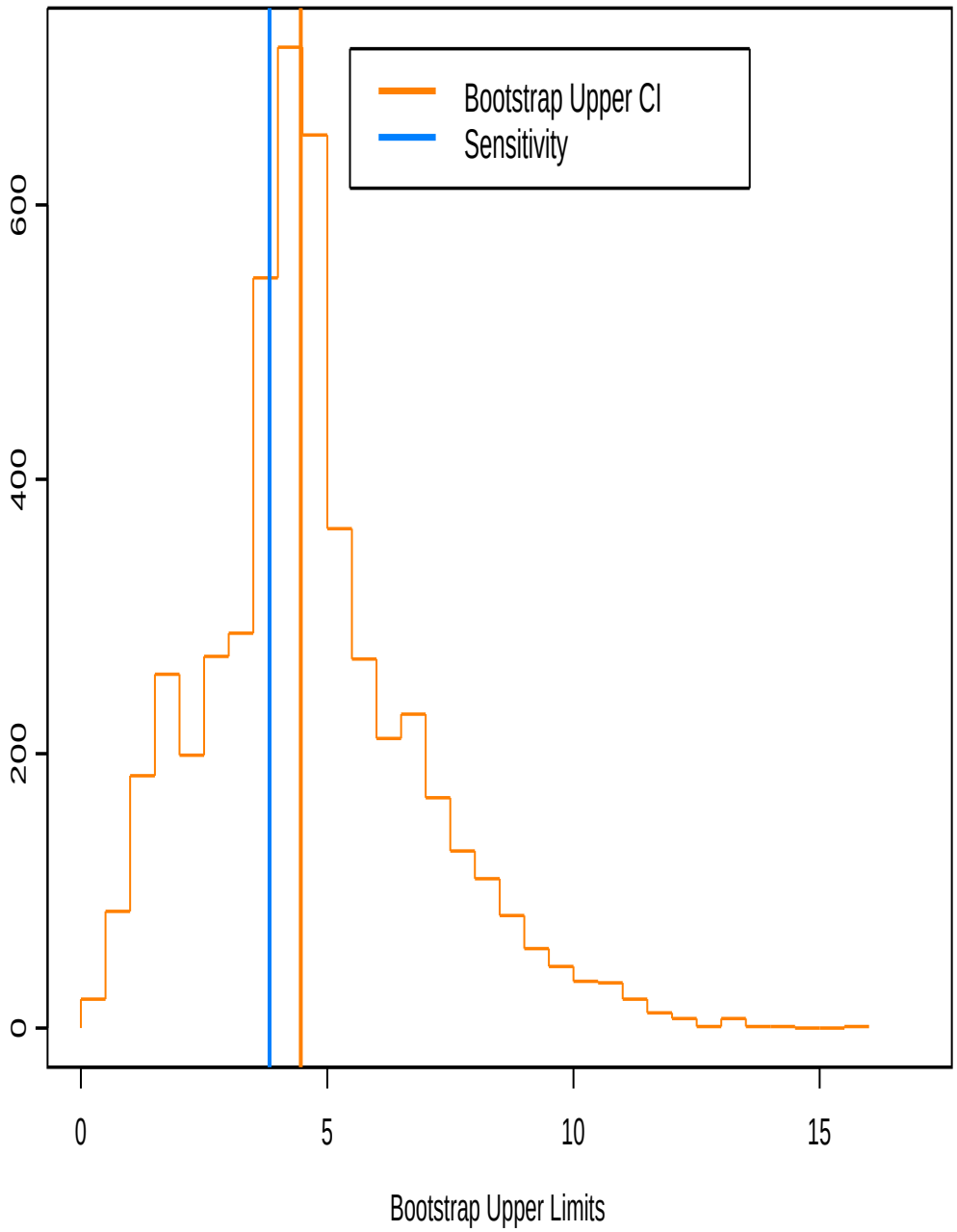
- **Say we have data X=(X_1,…X_n), each X_i an n-tuple**
- **Distribution of X depends on a parameter θ**
- **We want to find an estimate of the parameter θ, say using the estimator T=T(X)**
- **Find a bootstrap sample, that is a sample from X_1,…X_n of size n, taken *with replacement*. Call it X_1^***
- **Find a bootstrap estimate of θ, $T_1^*=T(X_1^*)$**
- **Repeat many times (~B=100 to 10000)**
- **Study T by studying T_1^*,.., T_B^***

Double Bootstrap Method

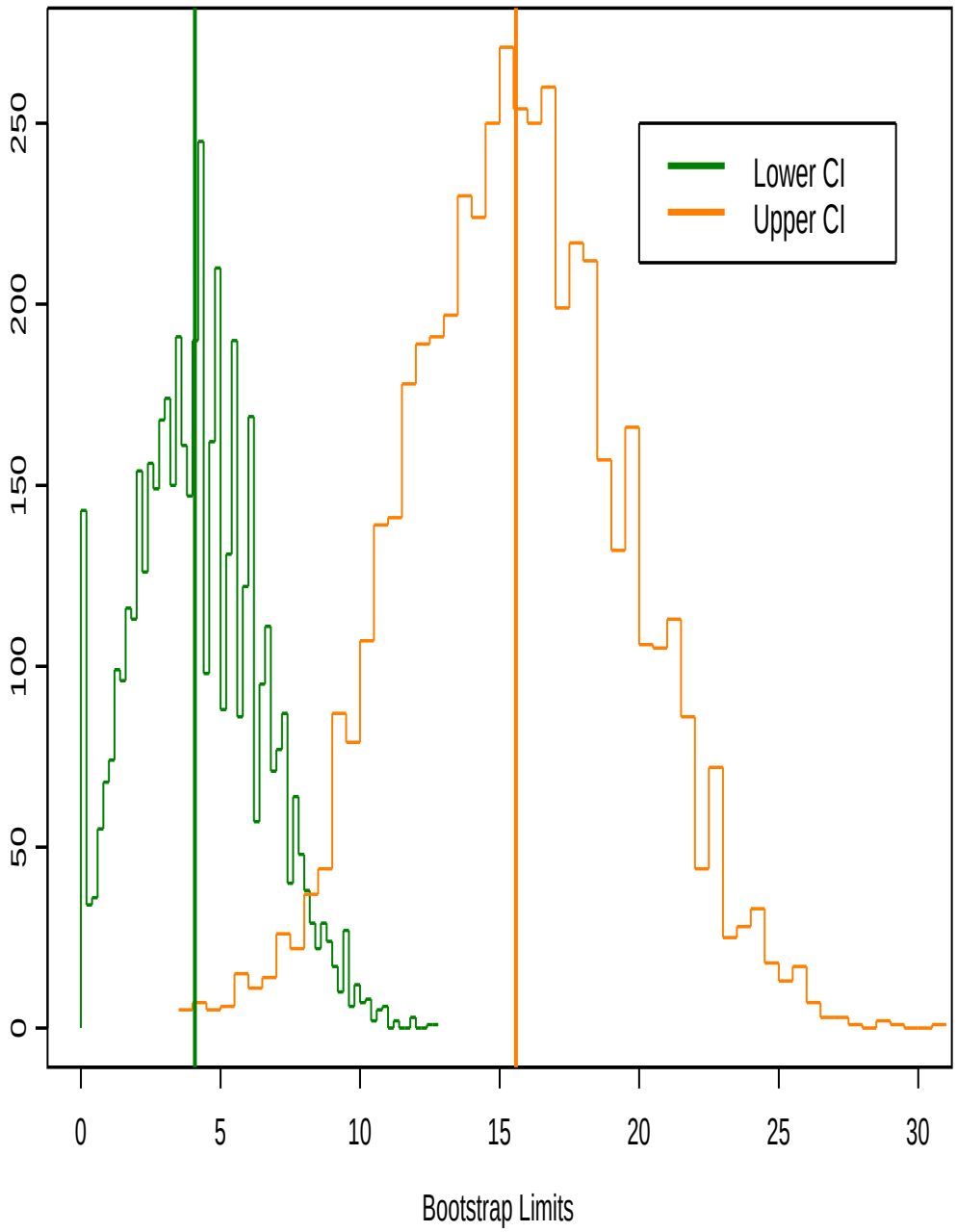
- **Find bootstrap sample, find the cut combination that is optimal for that bootstrap sample.**
- **Find another bootstrap sample, apply the cut combination above to find the number of events in the signal region and the sidebands.**
- **Find limits for the signal rate.**
- **Repeat many times (say ~5000).**
- **Use the medians of the lower and the upper limits as your confidence interval.**


True Coverage
0.84
0.86
0.88
0.90
0.92
0.94
0
1
2
3
4
5
6
Mean Number of Signal Events
Minimum Effective Sensitivity Cut
Double Bootstrap

Case 1: No Signal



Case 2: Signal with Rate 6



Conclusion

- **There is a bias because we choose the cut combination that is optimal for the data, not the distribution.**
- **The double bootstrap method appears to be able to fix that bias and gives confidence intervals with the correct coverage rate.**
- **Paper describing the details of the method will be submitted to *Nuclear Instruments and Methods A*.**
- ***Fortran* routines for computation of Rolke-Lopez limits and for the double bootstrap are available by email from w_rolke@rumac.uprm.edu**