Journal of Data Science 23 (1), 225–242
January 2025

DOI: 10.6339/24-JDS1137
Statistical Data Science

An S-Curve Method for Abrupt and Gradual Changepoint Analysis

Lan Jiang[1],*, Collin Kennedy[1], and Norman Matloff[1]
[1] *University of California, Davis, USA*

Abstract

Changepoint analysis has had a striking variety of applications, and a rich methodology has been developed. Our contribution here is a new approach that uses nonlinear regression analysis as an intermediate computational device. The tool is quite versatile, covering a number of different changepoint scenarios. It is largely free of parametric model assumptions, and has the major advantage of providing standard errors for formal statistical inference. Both abrupt and gradual changes are covered.

Keywords *difference in difference; identifiability; nonlinear regression models; standard errors*

1 Introduction

Here we introduce a new approach to changepoint analysis for means, linear model slopes and intercepts, and many other changepoint models, using nonlinear least squares with an S-curve as a device. The method handles both abrupt and gradual change points, in a unified manner. And somewhat rare among changepoint methodology, our method provides standard errors for the estimates of the changepoint location, the pre- and post-changepoint means and other values of interest, enabling formal statistical inference.

For data $Y_1, \dots, Y_n$, a *changepoint* is an index i such that some statistical property—for example the expected value μ_i—is different for indices before i than afterward. As explained below, our definition will actually be slightly broader than this. Denote the changepoints by τ_j, $j = 1, 2, \dots, \eta$, where η is the unknown number of changepoints. In this paper, we are primarily concerned with the case $\eta = 1$.

The field has a long history, and a wide variety of applications. For example, changepoint methods have been used in finance to identify events that induce significant volatility shifts in foreign markets (Aggarwal et al., 1999), in bioinformatics to identify damaged genes and genomic imbalances (Muggeo and Adelfio, 2010), and in text analyses and forensic linguistics to shed light on authorship debates (Chen and Zhang, 2015). Changepoint methods have also been applied to problems in seismology, climatology, psychometrics, and macroeconomics, among many others.

The literature of changepoint detection methodology is as substantial as the body of research *applying* such methods. For a more extensive exploration of changepoint methodology, Aminikhanghahi and Cook (2017) and Truong et al. (2020) provide excellent bibliographies, as does the website `changepoint.info` (Killick et al., 2012b). Extensive theoretical work has always appeared, such as Song and Chen (2021).

*Corresponding author. Email: lanjiang@ucdavis.edu or cjkennedy@ucdavis.edu or nsmatloff@ucdavis.edu.

© 2025 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.

Received April 18, 2024; Accepted April 19, 2024

The importance of changepoint analysis to data science may also be seen in the numerous R packages that have been developed in this realm, such as Killick and Eckley (2014), Lindeløv (2020), Erdman and Emerson (2007), Liao and Meyer (2023), and many, many others. There is a very useful (though only partial) comparison chart in Lindeløv (2023).

Contributions and Organization of the Present Work

The following are the major contributions of our work to the literature:

- Our method *provides standard errors* for changepoint locations, magnitudes of jumps and so on, thus enabling formal statistical inference, a must in scientific research and the like.
- Our method models not only abrupt changes but *also gradual changes*.
- The S-Curve method is quite *generalizable*. Not only does it deal with the typical changepoint problem of changes in mean, but also changes in slope or intercept in linear models, and so on.
- We do *not assume errors to be Gaussian or have any other parametric distribution*.

Our method is implemented in an R package, `changeS` (Jiang et al., 2024), and we will make occasional references to it in this paper. Note, though, that the main focus of this paper is our S-curve method itself, and development of other implementations would be straightforward. For example, in Python the `curve_fit` function in the SciPy library could be used (Virtanen et al., 2020).

The organization of the remainder of this paper is as follows. After a review of previous literature in Section 2, Section 3 presents the details of the S-Curve method. Section 4 discusses the wide extensibility of the method to determining changepoints in linear, generalized linear and other common statistical models. Then Section 5 demonstrates the method on a variety of real and simulation data sets, along with a comparison to the `segmented` package (Muggeo, 2008) and estimation of coverage probabilities. Section 6 details the implications of our findings and the practical considerations that may arise from using the S-Curve method, followed by concluding remarks in Section 7.

2 Previous Related Work

To set the stage for the remainder of the paper, it is important to compare and contrast the present work to previous literature. We begin by identifying two divergent philosophical paths in research in changepoint methodology. We then discuss other differences in methodology, first parametric versus nonparametric, then abrupt versus gradual change.

2.1 Two Different Statistical Views

The fact that the S-Curve method provides standard errors for the changepoint locations and so on is especially noteworthy, since many changepoint-focused R packages do not do so (Lindeløv, 2023). A method that does offer standard errors is that of Muggeo (2003, 2008, 2017) which assumes normal errors and is based on heuristic workarounds to lack of a differentiable likelihood.

Though our focus is applicability to formal statistical inference in scientific research in the form of confidence intervals and hypothesis tests, we mention that some Bayesian packages such as `bcp` and `mcp` compute posterior *credible intervals* (Wang et al., 2018; DasGupta, 2008).

We will divide the literature on changepoint analysis into two approaches, whose difference is actually much more profound than what might appear at first glance:

- `Approach 1:`
  The changepoint is viewed as an integer, the index i in μ_i. If a changepoint is known to exist, or presumed so, the analyst is then presented with a discrete set of choices as to the identity of i. Each i is tested as the candidate changepoint.
- `Approach 2:`
  Our present work and that of Muggeo view the changepoint as a point in continuous time, which typically takes on continuous real, i.e. noninteger values. This presents a continuous range of possible locations, analyzed as a classical statistical point estimation problem.

Different applications have different needs. Approach 1 is clearly the more appropriate one in applications in which the changepoint is inherently integer-valued, for instance genetic microarray settings. On the other hand, in many cases the changepoint location is continuous, making Approach 2 more appropriate.

The ability to perform formal statistical inference, as noted important for scientific research, is missing in most work on Approach 1. This is actually a consequence of the integer-valued nature of the changepoints, which precludes asymptotic normal inference as in Maximum Likelihood estimation and nonlinear regression models. One might still try producing a confidence interval by, say, inverting the hypothesis test or something similar (Bai and Perron, 2003).

Note that bridging the gap between these two approaches is not simple:

- One cannot connect Approach 2 to Approach 1 by merely rounding the results of the latter to the nearest integer, as that would invalidate the standard errors.
- As discussed above, one cannot connect Approach 1 to Approach 2 in the usual sense of asymptotic standard errors.

Just as is the case under Approach 1, Approach 2 does *not* presume that a changepoint exists. That issue is handled by suitable interpretation of the parameter being estimated.

The `chngpt` package of Fong et al. (2017, 2019) offers myriad options for various changepoint models. Viewed by the authors as a successor Muggeo's `segmented` package, it in essence allows the user a choice of Approaches 1 and 2 (with Approach 1 preferred by the authors), both assuming normal errors. Approach 1, termed *grid*, follows the typical path of that approach, computing likelihood at all possible changepoints; then choosing the maximizing one; the bootstrap can then be used to form confidence intervals. Citing Zhou and Liang (2008), Approach 2 is similar to our S-curve model, which can be used to compute standard errors.

2.2 Distribution-Free Methods

Another important issue is that many changepoint methods assume the data Y_i are generated from some parametric family, typically normal. Distribution-free methods include those of Killick et al. (2012a) and Fryzlewicz (2014), but neither offers standard errors, for the reasons given above. The latter work does prove asymptotic consistency.

The spirit of being distribution-free would also suggest not assuming homoskedasticity. See Section 3.2.

2.3 General Types of Changes

Though typical changepoint analysis in the literature has concerned a shift in mean, changes in slope or intercept in linear regression models have been considered by a number of authors, such as Chen et al. (2011). In Section 4, we will apply our S-curve method in this setting as well, and note that the method easily generalizes to other settings.

Much of the changepoint literature has been concerned with abrupt change models, but some authors have considered the problem of a gradual change. Consider for example data presented in Pawitan (2005), involving breast cancer in Sweden. It is thought that the onset of menopause is associated with increase in incidence of the disease, i.e. that menopause is a changepoint. This was modeled as an abrupt changepoint in the above paper, but a gradual model may provide better insight, as (a) the effects of menopause on cancer presumably are gradual, and (b) different women attain menopause at different ages.

Thus methodology for gradual models is of interest. For instance, Hušková (1999) proposed modeling change as a power of its argument. In our S-curve method gradual change is modeled as having a logistic form.

Also, Bhaduri et al. (2022) describe a method for identifying gradual changepoints that utilizes a combination of rough sets and fuzzy logic (rough fuzzy sets), along with an accompanying Python package `roufcp` to implement the new approach. The method demonstrates notable improvements over a wide variety of changepoint methods with respect to detecting gradual changes; however, it performs relatively poorly in the case of abrupt changepoints. Other fuzzy changepoint detection algorithms exist, such as the fuzzy changepoint algorithm (Chang et al., 2015), the fuzzy classification maximum likelihood changepoint algorithm (Lu and Chang, 2016), and the fuzzy shift changepoint algorithm (Lu et al., 2016). Also, Wu et al. (2024) describe a method that utilizes Bayesian dynamic linear models (DLMs) to identify both gradual and abrupt changepoints, which they refer to as ‘drift’ and ‘shifts’, respectively.

2.4 Multiple Changepoints

There is also the related problem of detecting multiple changepoints, treated in work such as Killick et al. (2012a), Muggeo (2003) and Fryzlewicz (2014). Yau and Zhao (2015) considered the special case of analysis of a stationary time series.

Yao and Au (1989) proposed a least-squares method based on moving averages of the Y_i. Though they proved asymptotic consistency of the estimated changepoint locations $\widehat{\tau}_j$, and even of their number, η, they did not establish standard errors for the $\widehat{\tau}_j$. As explained above, the integer nature of the τ_j makes computing conventional standard errors for these quantities highly problematic.

This problem is especially challenging in theoretical work, where unusual assumptions are made that may be problematic in practice. Commonly, one must assume that the maximum number of changepoints is not too large, and the minimum spacing between changepoints is not too small (Fryzlewicz, 2014). Bai and Perron (2003)’s treatment of the linear segmented case makes the assumption that as $n \to \infty$, the magnitudes of the changes in slope and intercept from one segment to the next go to 0.

To our knowledge, the multi-changepoint case is still an open question in that regard; no general methods exist in the literature that are distribution-free and offer standard errors for the changepoint locations. One can of course employ *binary segmentation*, recursively partitioning the range of X; each time a changepoint is found, one can subdivide the current range of interest, before and after the changepoint, and then check for changepoints in the two subranges. However, this then would pose a major challenge to finding proper standard errors, as it becomes a matter of *postselection inference*. It is not just a matter of *simultaneous inference* (Hsu, 1996). Much work has been done in recent years in postselection inference (Berk et al., 2013; Kuchibhotla et al., 2022), and in the future this may have applications to doing formal inference in binary segmentation settings. See Section 4 for further discussion of this point, in the context of our package.

3 Methodology

3.1 General Model

As mentioned, our approach is unified, in that the same model handles both abrupt and gradual changepoint analyses. Both cases are handled by the "S-curve," a generalized logistic function, as follows. (This is the logistic *function*, but our model will generally not be a logistic *regression model*.)

Our notation will include not only Y but also X. In simple cases, the latter is simply time, e.g. age in the breast cancer example above or something similar, such as below-surface depth in geological layers. But in more elaborate models, X may also include covariates and so on. $Y|X$ will refer to observing Y conditioned on a specific value of X.

In general changepoint analysis, in the abrupt case, $E(Y|X = x)$ is a step function in x. In our method, we approximate this by a smooth function. Let Y be the observed value of interest, at a given index x. We fit the model,

$$E(Y|X = x) = \alpha_1 + (\alpha_2 - \alpha_1) \cdot \frac{1}{1 + \exp(-[\alpha_4(x - \alpha_3)])}, \tag{1}$$

where the roles of the parameters are:

- α_1: pre-changepoint value of $E(Y|X = x)$
- α_2: post-changepoint value of $E(Y|X = x)$
- α_3: changepoint
- α_4: abruptness approximation parameter

For large $|\alpha_4|$, this approximates a step function at $x = \alpha_3$. The jump is from height α_1 to height α_2.

As an example, with $\boldsymbol{\alpha} = (1.2, 1.7, 5.0, 25.0)$, Figure 1 illustrates how Equation (1) approximates the step function:

$$f(x) = \begin{cases} \alpha_1 = 1.2 & x \leqslant \alpha_3 = 5.0 \\ \alpha_2 = 1.7 & x > \alpha_3 = 5.0. \end{cases} \tag{2}$$

In the abrupt case α_1, α_2 and α_3 are estimated from the data, but the user sets α_4 to some large value to achieve the approximate step function. In our implementation, the default is $\alpha_4 = 10$; much larger values are not recommended, as they may lead to convergence problems or large standard errors.

For a gradual-change model, the algorithm will estimate all four parameters from the data. In the gradual case, α_3 is the inflection point of the S-curve, and α_4 is the slope of the S-curve at that point.

Approximating a step function by a smooth function enables the use of nonlinear least-squares for the estimation process, and to compute standard errors. Our implementation employs the nonlinear regression package `nls.multstart` (Padfield and Matheson, 2023). The data, $(Y_i, x_i), i = 1, \ldots, n$ is fed into the nonlinear least-squares machinery using the model (1).

Again, in the gradual-change model, the algorithm estimates α_4 It is in this manner that the user chooses between abrupt or gradual change; the user specifies the former by setting a large value for α_4, or specifies the latter by allowing the algorithm to estimate this value.

One could adopt a policy of always fitting the gradual model, even in abrupt settings. This of course is the typical modeling question, bias versus variance. If one applies the gradual model, there is an extra parameter (slope of the S-curve at the inflection point) to be estimated, thus

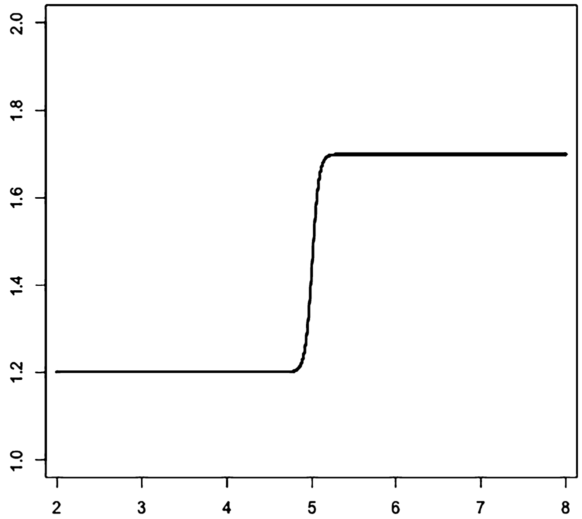

Figure 1: S-curve as a Step Function.

increased standard errors. So, one should use a parsimonious model if it is appropriate for the setting.

We note that a similar nonlinear scheme has been used to model innovation, called *Diffusion of Innovation* (Rogers, 1962). The context is different—one models several different stages of consumer adoption of a new product, and there is no changepoint *per se*—but their setting employs a model similar to (1), with $\alpha_4 < \infty$.

3.2 Standard Errors (SEs)

In most types of statistical methodology, a key part of one's analysis is to perform statistical inference, i.e. confidence intervals and hypothesis tests, via standard errors. This is especially important in publishing scientific research, for example. As noted, it is thus surprising at first that most software packages for changepoint analysis do not compute standard errors (Lindeløv, 2023), say for the location of the changepoint and the magnitude of the jump. However, this lack is natural in light of the explanation in Section 2.1 that classical statistical inference is essentially impossible under Approach 1. Again, our approach does produce a variance-covariance matrix for all estimated parameters in our model, enabling the formation of confidence intervals for the changepoint location and the magnitude of the jump in means.

It should be noted that the use of nonlinear least-squares (NLS) is key to the distribution-free nature of our method. The asymptotic distribution of NLS estimated coefficients is well-known to be multivariate normal (Jennrich, 1969; Matloff, 1981; Wu, 1981; Pollard and Radchenko, 2006), *regardless of the distribution of* $Y|X$, be it normal, gamma or amorphous; all that is needed in terms of distribution is that $Var(Y|X)$ is finite. The SEs are the standard deviations in the limiting normal distribution, and are thus sometimes referred to as *asymptotic standard errors*. Note that these typically differ from the finite-sample standard deviations (Knight, 2000, p. 207).

One then, say, forms confidence intervals (CIs) with the usual $\widehat{\theta} \pm 1.96\ SE(\widehat{\theta})$ computation for estimating a parameter θ. In cases in which θ is some α_i, the SEs are obtained as square roots of the diagonal elements of the associated covariance matrix. But the off-diagonal elements are important as well. For instance, to form a CI for the jump size in (1) above, the population value, $\alpha_2 - \alpha_1$, is estimated by $\widehat{\alpha_2} - \widehat{\alpha_1}$. The latter difference has estimated variance

$$\begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} V \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \tag{3}$$

where V is the estimated asymptotic covariance matrix. Taking the square root yields the appropriate SE.

Note that if one wishes to test the hypothesis that there is no changepoint, that is equivalent to $H_0 : \alpha_1 = \alpha_2$. Thus one can test for there being no changepoint by checking whether the above CI contains 0.

Again, the use of the SEs to form confidence intervals and perform hypothesis tests here is based on the asymptotic behavior of NLS least-squares estimator. No distributional assumption is required for $Y|X$.

However, a reviewer has pointed out that in the case of normal errors, maximizing the likelihood essentially amounts to computing least-squares estimates, since the log-likelihood is the negative sum of squares. In this sense, the reviewer noted, the S-curve method might be viewed as tantamount to assuming normal errors, not distribution-free after all. This is true if one assumes homoskedasticity, but the situation changes markedly in the heteroscedastic setting:

General asymptotic nonlinear least-squares theory does not assume constant variance, nor does it make any other assumptions at all about the structure of the error variance (Wu, 1981; Pollard and Radchenko, 2006). Then least-squares analysis no longer has a likelihood connection; absent a parametric model for $Var(Y|X)$, no likelihood can be defined.

The heteroscedastic case can be easily handled, using the *sandwich estimator* (Boe et al., 2024; Sidik and Jonkman, 2016). In our implementation, one can obtain the estimated covariance matrix for a heteroskedastic setting by calling `sandwich::sandwich()` on the `nlsOut` component in the fitted model.

3.3 Identifiability Issues

There turns out to be a uniqueness issue in Equation (1): For each S-curve fit to the data, there are two sets of parameters with opposite slopes and flipped means that produce that same curve. This makes the curve-fit process problematic, since for any meaningful output, there is another equivalent convergence point. Indeed, it may result in nonconvergence. Also, since α_1 and α_2 are interchangeable, it would be hard to tell which of them carries the mean value before or after the changepoint:

$$\begin{aligned} E(Y|X=x) &= \alpha_1 + \frac{\alpha_2 - \alpha_1}{1 + \exp(-[\alpha_4(x - \alpha_3)])} && (4) \\ &= \alpha_2 + \frac{\alpha_1 - \alpha_2}{1 + \exp(-[-\alpha_4(x - \alpha_3)])}. && (5) \end{aligned}$$

To solve this problem, known in mathematical statistics as lack of *identifiability*, we set a lower bound 0 for α_4 when fitting the curves. Such a maneuver eliminates the symmetric counterpart of each convergence and locks the pre-changepoint mean value and post-changepoint mean value in α_1 and α_2 respectively.

4 Extensibility

One of the major advantages of the S-curve method is that it can be extended to many other settings besides a change in mean. This point is explored in the current section, beginning with an extension which is in our package.

Piecewise Linear Models

The model (1) is easily adapted for detection of a changepoint in slope or intercept in a linear model:

$$\begin{aligned} E(Y) = &\left(\alpha_1 + (\alpha_2 - \alpha_1) \cdot \frac{1}{1 + \exp(-[\alpha_4(x - \alpha_3)])}\right) \cdot x \\ &+ \left(\alpha_5 + (\alpha_6 - \alpha_5) \cdot \frac{1}{1 + \exp(-[\alpha_7(x - \alpha_3)])}\right). \end{aligned} \tag{6}$$

Here α_3 is the changepoint, with $\alpha_2 - \alpha_1$ and $\alpha_6 - \alpha_5$ representing the changes in slope and intercept. In some cases, the analyst will model a change in intercept but with unchanged slope, or *vice versa*. This is achieved by setting $\alpha_1 = \alpha_2$ in the first case, or setting $\alpha_5 = \alpha_6$ in the second. Again for identifiability reasons, the slopes (α_4, α_7) are also forced to be non-negative so the outputs remain consistent at convergence.

Multiple Changepoint Models

Our package does include a binary segmentation function as a convenience to the user, but as with any binary segmentation method, formal statistical inference is not possible. The function is included merely as an exploratory tool.

A possible extension of our S-curve method that would produce true standard errors would be to use the multi-sigmoidal Gompertz curve family, a generalization of the logistic (Román-Román et al., 2019). In essence, it would enable a separate logistic curve for each regime between consecutive changepoints.

Another possibility would be to simply perform a fixed, preset number of iterations of the binary segmentation process, and apply the S-curve method to each resulting subinterval. One could use the Bonferroni Inequality to form multiple confidence intervals. We believe this approach could be made mathematically rigorous, e.g. in the sense of statistical consistency and so on.

DiD Models

Another intriguing area of application of our method may be to `difference in difference (DiD)` models (Angrist and Pischke, 2008; Callaway and Sant'Anna, 2021). DiD is a quasi-experimental method that compares changes in an outcome of interest between two or more groups before and after some treatment, often a policy intervention. This allows researchers to obtain an estimate of the causal impact of said intervention while controlling for time-varying confounders. That impact could be modeled in changepoint terms in our S-curve context.

Moreover, use of the S-curve approach could involve a gradual model, in contrast to standard DiD, in which the change is modeled as abrupt. The impact of a new children's educational reading program, for instance, may come gradually.

Multivariate Response Variable

An extension of the S-curve method to multivariate data, i.e. vector-valued Y, would be possible. In (1), for example, α_3 would be the same for each component, but the values of jumps $\alpha_2 - \alpha_1$, and possibly α_4, would differ according to component. Computation would require a modified `nls.multstart`. Again consider (1). In present form, Y is a scalar, and internally `nls.multstart` minimizes a sum of squares involving the input data (x_i, Y_i). But if now Y is a vector, one could minimize the grand sum of the sum of squares over all components of Y.

Implementation of Extended Models

Our `changeS` package fits S curves, in various forms, to produce nonlinear models that are calculated using the `nls.multstart` package. The latter expects a user-defined function specifying the desired nonlinear model. For example, `changeS::fitS`, the main user interface to `changeS`, implements (1) as an `nls.multstart`-suitable function, while Equation (6) is implemented in another such function. All this is transparent to `changeS` users. However, the latter can develop their own `changeS` functions like `fitS` to implement their own changepoint models, say generalized linear models such as logistic and Poisson regression (details too complex to include here).

5 Data Examples

In this first empirical section of the paper, we illustrate our S-curve method on various datasets. Both abrupt and gradual fits will be shown, as well as an application of the piecewise linear model.

5.1 First Simulation

We first conducted a very simple simulation to demonstrate the method. Here, we considered a sample of $n = 500$ observations, and designated the 334th observation to be the changepoint. Observations Y_i were drawn from a normal distribution, such that for $\{i = 1, \ldots, 333\}$, $Y_i \sim \mathcal{N}(10, 2)$, and for $\{i = 334, \ldots, 500\}$, $Y_i \sim \mathcal{N}(12.5, 2)$, thus an abrupt changepoint. Not surprisingly in this example of a visually obvious change, the S-curve easily identifies the changepoint and does so with relatively high precision as indicated by the small standard error.

Results can be seen in Figure 2. The output of `summary()` (from the underlying nonlinear least squares library) is

```
          Estimate Std. Error t value Pr(>|t|)
postMean   12.4734     0.1724   72.36   <2e-16 ***
preMean    10.1288     0.1055   95.99   <2e-16 ***
changePt  363.9031     0.4308  844.79   <2e-16 ***
```

5.2 Nile Data

The next data consists of yearly average water flow (m^3/s) of the River Nile, a built-in dataset to the R language. Our model concluded that there is an abrupt changepoint in the middle of year 1898. This matches the fact that the British started the construction of the Aswan Low Dam in that year.

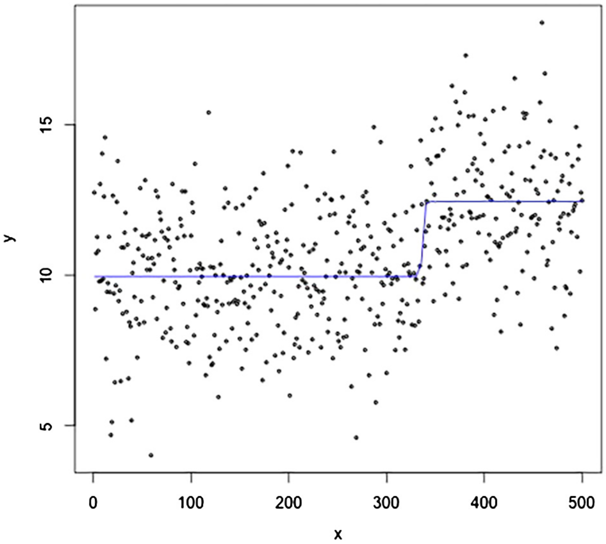

Figure 2: First simulated data example.

Here domain expertise would identify the location of the changepoint, so the main interest is the value of the change. A major drop in water flow is detected around this point, from 1097.75 down to 849.972. The estimated drop, 247.778, has a standard error of 28.93205. Here is the full report:

```
          Estimate Std. Error t value Pr(>|t|)
postMean   849.970     15.126   56.19   <2e-16 ***
preMean   1097.930     24.694   44.46   <2e-16 ***
changePt  1898.381      2.482  764.87   <2e-16 ***
```

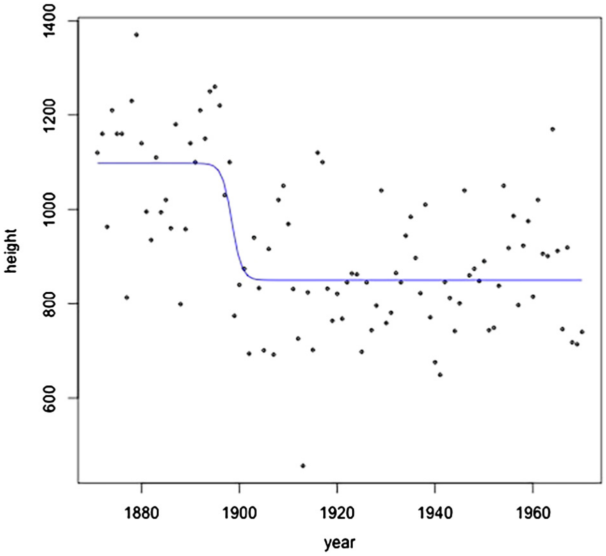

Figure 3: Change in River Nile Flows.

5.3 Breast Cancer Data

We next consider applying the S-Curve approach to data collected in women in Sweden on the rates of breast cancer. There had been speculation that such rates rise with the onset of menopause (Pawitan, 2005).

While that paper considers an abrupt model, the relationship, if one exists, may be gradual. And even if it were abrupt, different women experience menopause at different ages, so that the data would follow a mixture of abrupt changes, thus gradual overall. Thus this is a good example use case for our S-curve approach.

The fitted S-curve, superimposed on the data, can be seen in Figure 4. We find that the inflection point is estimated to be 43.2628, with a standard error of 0.4617. This is somewhat earlier than the reported average menopause age in Sweden of 54.76. It would appear that a revised view is that women's breast cancer risk incurs an inflection point in the years *approaching* menopause. A random effect model for α_3 may be interesting to pursue.

Here is the full report:

```
          Estimate Std. Error t value Pr(>|t|)
postMean    8.9660     0.4091  21.917  < 2e-16 ***
preMean     3.4256     1.0649   3.217  0.00177 **
slope       1.1783     0.6409   1.839  0.06910 .
changePt   43.2628     0.5476  79.003  < 2e-16 ***
```

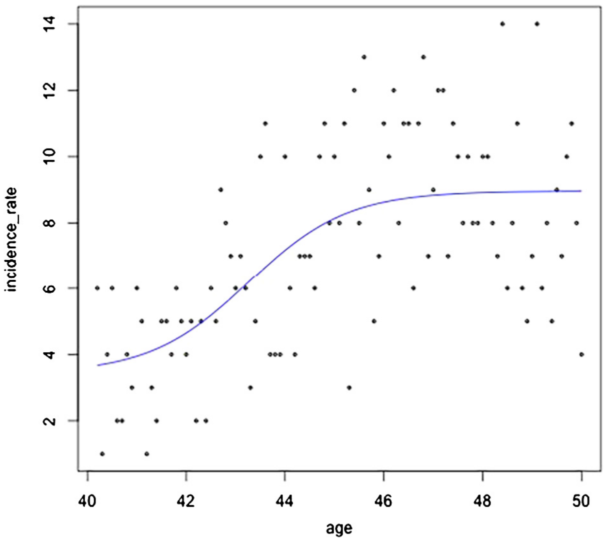

Figure 4: Cancer Rates, Gradual Curve Model.

5.4 T-Bill Data

We also applied the S-Curve method on another real-world dataset, but with an *abrupt* changepoint. This is the data `RealInt` from the R package `bcp` (Wang et al., 2018), consisting of a quarterly time series Treasury Bill, adjusted for inflation, during 1961–1986. Our S-Curve method identifies an upward changepoint during 1980; see Figure 5. The full report is below:

```
          Estimate Std. Error t value Pr(>|t|)
```

```
postMean   5.72181     0.52815   10.834    <2e-16 ***
preMean    0.07861     0.28497    0.276     0.783
changePt  79.93172     0.20335  393.075    <2e-16 ***
```

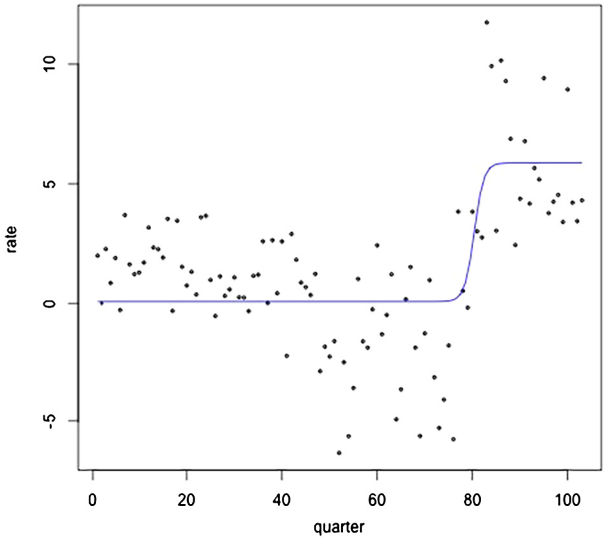

Figure 5: Real Interest Rates.

The identified changepoint aligns with Paul Volcker and the US Federal Reserve's abrupt tightening of the money supply that same year Sablik (2013). Note too another possible changepoint around 1972.

5.5 Linear Model Example: Medicare Data

Here is an example using real data from Medicare, the US medical insurance program for retired people. One nominally qualifies at age 65, though this can occur earlier or later. Here we consider Emergency Room visits in relation to age, using the piecewise linear model (6).

The fitted lines are shown in Figure 6, with summary

```
     Estimate  Std. Error  t value  Pr(>|t|)
b1      9.1749     0.8019   11.442  < 2e-16 ***
h1     18.6334     0.5671   32.859  < 2e-16 ***
c      63.9665     0.1256  509.315  < 2e-16 ***
b2  -133.2359    49.3234   -2.701   0.00795 **
h2  -731.0732    38.0025  -19.237  < 2e-16 ***
```

Before the changepoint at about 63.97, the line had estimated slope and intercept of 9.17 and −133.24, respectively. Afterward, these changed to 18.63 and −731.07. The visit rate per year of age in the population under study appears to increase substantially with the availability of insurance.

5.6 Comparison to Segmented Package

We compared our method's abrupt changepoint-detection capability with that of the package `segmented` (Muggeo, 2008) on several simulated datasets. The package provides tools for fitting

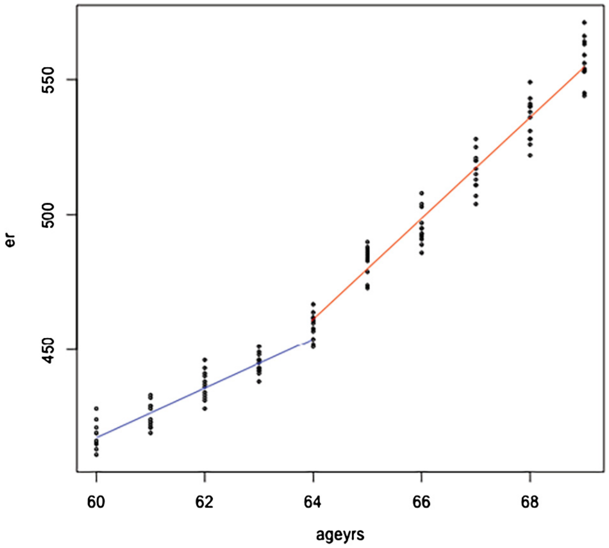

Figure 6: Illustration of the piecewise linear model.

piecewise regression models (with multiple changepoints). As discussed in Section 2.1, we believe that this is the most appropriate package and method for direct comparison, given the similarities between our two approaches. The `segmented` method fits piecewise linear models, but can be used for the change-in-means setting by specifying a model consisting only of an intercept term (via `lm(y ~ 1)`).

We considered three different scenarios, each of which were very similar in underlying structure but varied in terms of the distribution from which data was sampled. At a high level, each simulation was a time series of $n = 1000$ data points, sampled from either a normal, exponential, or fat-tailed (Student–t, 3 df) distribution. Each scenario entailed 100 replications of the respective simulation designed for a particular sampling distribution, and Mean Absolute Error (MAE) was calculated for both methods as the measure of comparison.

- `Normal Case`
  Each simulation considered a time series where $n = 1000$, and random noise was sampled from a normal distribution with $\sigma = 2$. For $i = 1, \ldots 666$, the data has a constant mean $\mu = 10$. From $i = 667, \ldots, n$, the data has a constant mean $\mu = 12.5$.
- `Exponential Case`
  Each simulation considered a time series where $n = 1000$, with the Y_i being exponentially distributed with mean μ_i. The latter quantities were as in the normal case.
- `Fat-Tail Case`
  Each simulation considered a time series where $n = 1000$, and random noise was sampled from a 0-centered t distribution with $df = 3$. From $i = 1, \ldots 666$, the heavy-tail noise was added to a constant $c = 10$. From $i = 667, \ldots, n$, $c = 12.5$.

The results can be seen in Table 1. The sample standard error refers to the accuracy of the second column for 100 replications. For example, a 95% CI for the true expected estimated changepoint value in the first line is $668.22 \pm 1.96 \times 0.04$.

Notably, the average estimated changepoints obtained by our S-Curve method (implemented in our package, `changeS::fitS()`) were consistently more precise than those obtained by `segmented::segmented()` across each of the three sampling distributions considered. While

Table 1: Summary of Comparison: S-Curve vs. segmented (Abrupt Changepoint).

Distribution	Method	True Cpt.	Avg Est. Cpt.	Mean Abs. Dist.	Sample Std. Error
Normal	S-curve	667	668.22	2.40	0.04
Normal	segmented	667	651.88	22.17	0.76
Exponential	S-curve	667	678.75	149.75	3.23
Exponential	segmented	667	592.61	165.08	2.08
Fat-tailed	S-curve	667	667.33	1.25	0.02
Fat-tailed	segmented	667	651.26	21.92	0.83

`segmented` also performed well in each of the three scenarios, it seemed to demonstrate a negative bias in each of the scenarios considered.

5.7 Coverage Probability Estimation

To assess the validity of confidence intervals constructed using standard errors computed in `changeS`, we estimated coverage probabilities for gradual changepoints in data drawn from normal and fat-tail distributions. Coverage probabilities for the Muggeo method were presented in his original paper, Muggeo (2003).

- `Normal Case`
  Here we simulated a time series of n points, normally distributed with $\sigma = 2$. The mean at each point was 10 for the points $\{i = 1, \dots, i_{\text{beginning}}\}$, and 12.5 for the points $\{i_{\text{end}}, \dots, n\}$. A linear change from 10 to 12.5 occurs during $\{i_{\text{beginning}} + 1, \dots, i_{\text{end}} - 1\}$ with slope $(12.5 - 10)/(i_{\text{end}} - i_{\text{beginning}} + 1)$. Random noise was then added to each point, drawn from a normal distribution with $\mu = 0$ and $\sigma = 2$. The sub-indices *beginning* and *end* denote where the gradual change begins and ends.
  We consider two scenarios, with $n = 100$ and $n = 1000$. In both scenarios, we performed 500 replications, and computed asymptotic confidence intervals. Nominal CI levels were 95%. The sandwich estimator was not used.
  Our S-Curve method demonstrated robust performance in the normal case, with 92% and 95% of the estimated confidence intervals containing the true changepoint for the cases ($n = 100$ and $n = 1000$), respectively. For 500 replications, the standard error of the estimated coverage probability is about $\sqrt{0.95 \times 0.05/500}$, around 0.01.
- `Fat-Tail Case (`$t_{df=3}$`)`
  For the fat-tail case, the simulation design is identical to that of the normal case, except that random noise was drawn from a t distribution with three degrees of freedom ($df = 3$). Here the estimated coverage probabilities were about 87% and 94%.

6 Discussion

Our method allows for both abrupt and gradual changepoint modeling. In the case of the Swedish cancer data, the original analysis Pawitan (2005) featured an abrupt model. Our method, using a gradual model, seems to fit the data better, and seems consistent with the biological points we noted.

Our method's use of an underlying nonlinear regression function has the additional benefit that standard model-assessment methods for regression analysis can be used in the model-fitting process. For instance, graphical methods for model fit can be used (Faraway, 2016; Matloff, 2017).

The other usual model-development principles also apply. If the analyst has domain knowledge indicating an abrupt change, she can employ this model. Otherwise, the analyst can first fit a gradual model, and say, use the standard error information to form a confidence interval for α_4; if the interval consists of very large (in absolute value) numbers, one may opt to refit the simpler abrupt model. Similarly, that interval may include or be very near 0, in which case the analyst may conclude that there is little or no evidence for there being any changepoint at all. As mentioned earlier, though, a two-stage method like this technically invalidates the standard errors as with any statistical method. Again, the abrupt model is not only more parsimonious, but also will produce smaller standard errors if the model is appropriate.

As outlined earlier in Section 4, various extensions are possible. Future work is extension to multivariate-Y settings, as is an exploration of applications to DiD.

With some exceptions, such as Bai and Perron (2003) and Kim (1996), most changepoint research has assumed that the observations Y_i are independent. Extensions of our method in this direction could be made, though it may be difficult to maintain the parametric distribution-free nature of the present work.

Following up with the point in Section 2.4 regarding standard errors for the S-Curve method in the multiple-changepoint setting, extension to the multi-sigmoidal Gompertz curve is another possible approach. The theoretical work proposed in that extension will be investigated as well.

7 Conclusions

Our S-curve approach has three main benefits: (a) It enables the analyst to form confidence intervals or perform hypothesis tests on changepoint locations and jump magnitudes. (b) It allows modeling of both abrupt and gradual changepoints. (c) In contrast to the heuristic used in Muggeo (2003), our method has a solid theoretical basis. We have shown the effectiveness of the method, and have investigated the impact of the asymptotic nature of the standard errors on accuracy of statistical inference. Promising extensions will be pursued in future work.

Supplementary Material

The ZIP file contains all code needed to reproduce the figures and results of the experiments.

Acknowledgements

The authors highly appreciate the many comments and suggestions of the reviewers, Associate Editor and Editor-in-Chief, greatly improving the content and clarity of the paper.

References

Aggarwal R, Inclan C, Leal R (1999). Volatility in emerging stock markets. *Journal of Financial and Quantitative Analysis*, 34(1): 33–55. https://doi.org/10.2307/2676245

Aminikhanghahi S, Cook D (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51: 339–367. https://doi.org/10.1007/s10115-016-0987-z

Angrist J, Pischke J (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bai J, Perron P (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1): 1–22. https://doi.org/10.1002/jae.659

Berk R, Brown L, Buja A, Zhang K, Zhao L (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2): 802–837. https://doi.org/10.1214/12-AOS1077

Bhaduri R, Roy S, Pal S (2022). Rough-fuzzy cpd: a gradual change point detection algorithm. *Journal of Data, Information and Management*, 4: 1–24. https://doi.org/10.1007/s42488-022-00077-3

Boe LA, Lumley T, Shaw PA (2024). Practical considerations for sandwich variance estimation in two-stage regression settings. *American Journal of Epidemiology*, 193(5): 798–810. https://doi.org/10.1093/aje/kwad234

Callaway B, Sant'Anna PH (2021). did: Difference in differences. R package version 2.1.2.

Chang ST, Lu KP, Yang MS (2015). Fuzzy change-point algorithms for regression models. *IEEE Transactions on Fuzzy Systems*, 23(6): 2343–2357. https://doi.org/10.1109/TFUZZ.2015.2421072

Chen C, Chan J, Gerlach R, Hsieh W (2011). A comparison of estimators for regression models with change points. *Statistics and Computing*, 21: 395–414. https://doi.org/10.1007/s11222-010-9177-0

Chen H, Zhang N (2015). Graph-based change-point detection. *The Annals of Statistics*, 43(1): 139–176.

DasGupta A (2008). *Asymptotic Theory of Statistics and Probability* Springer Texts in Statistics. Springer, New York.

Erdman C, Emerson JW (2007). bcp: An R package for performing a Bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3): 1–13. https://doi.org/10.18637/jss.v023.i03

Faraway J (2016). *Linear Models with R*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Fong Y (2019). Fast bootstrap confidence intervals for continuous threshold linear regression. *Journal of Computational and Graphical Statistics*, 28: 466–470. https://doi.org/10.1080/10618600.2018.1537927

Fong Y, Huang Y, Gilbert P, Permar S (2017). chngpt: Threshold regression model estimation and inference. *BMC Bioinformatics*, 18: 454. https://doi.org/10.1186/s12859-017-1863-x

Fryzlewicz P (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6): 2243–2281. https://doi.org/10.1214/14-AOS1245

Hsu J (1996). *Multiple Comparisons: Theory and Methods*. CRC Press.

Hušková M (1999). Gradual changes versus abrupt changes. *Journal of Statistical Planning and Inference*, 76(1): 109–125. https://doi.org/10.1016/S0378-3758(98)00173-6

Jennrich RI (1969). Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, 40(2): 633–643. https://doi.org/10.1214/aoms/1177697731

Jiang L, Kennedy C, Matloff N (2024). changeS: S-curve fit for changepoint analysis. R package version 1.0.1.

Killick R, Eckley IA (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3): 1–19. https://doi.org/10.18637/jss.v058.i03

Killick R, Fearnhead P, Eckley I (2012a). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107: 1590–1598.

https://doi.org/10.1080/01621459.2012.737745

Killick R, Nam CF, Aston J, Eckley I (2012b). changepoint.info: The changepoint repository.

Kim HJ (1996). Change-point detection for correlated observations. *Statistica Sinica*, 6(1): 275–287.

Knight K (2000). *Mathematical Statistics*. Chapman & Hall/CRC Press.

Kuchibhotla AK, Kolassa JE, Kuffner TA (2022). Post-selection inference. *Annual Review of Statistics and Its Application*, 9(1): 505–527. https://doi.org/10.1146/annurev-statistics-100421-044639

Liao X, Meyer MC (2023). ShapeChange: Change-point estimation using shape-restricted splines.

Lindeløv JK (2020). mcp: An R package for regression with multiple change points. *OSF Preprints*.

Lindeløv JK (2023). An overview of change point packages in R. https://lindeloev.github.io/mcp/articles/packages.html.

Lu KP, Chang ST (2016). Detecting change-points for shifts in mean and variance using fuzzy classification maximum likelihood change-point algorithms. *Journal of Computational and Applied Mathematics*, 308: 447–463. https://doi.org/10.1016/j.cam.2016.06.006

Lu KP, Chang ST, Yang MS (2016). Change-point detection for shifts in control charts using fuzzy shift change-point algorithms. *Computers & Industrial Engineering*, 93: 12–27. https://doi.org/10.1016/j.cie.2015.12.002

Matloff N (1981). Use of regression functions for improved estimation of means. *Biometrika*, 68(3): 685–689. https://doi.org/10.1093/biomet/68.3.685

Matloff NS (2017). *Statistical Regression and Classification: From Linear Models to Machine Learning*. CRC Press.

Muggeo VM (2017). Interval estimation for the breakpoint in segmented regression: A smoothed score-based approach. *Australian & New Zealand Journal of Statistics*, 59(3): 311–322. https://doi.org/10.1111/anzs.12200

Muggeo VMR (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19): 3055–3071. https://doi.org/10.1002/sim.1545

Muggeo VMR (2008). segmented: An R package to fit regression models with broken-line relationships. *R News*, 8(1): 20–25.

Muggeo VMR, Adelfio G (2010). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 27(2): 161–166. https://doi.org/10.1093/bioinformatics/btq647

Padfield D, Matheson G (2023). nls.multstart: Robust non-linear regression using AIC scores. R package version 1.3.0.

Pawitan Y (2005). Change-point problem. In: *Encyclopedia of Biostatistics* (P Armitage, T Colton, eds.), John Wiley & Sons, Ltd.

Pollard D, Radchenko P (2006). Nonlinear least-squares estimation. *Journal of Multivariate Analysis*, 97(2): 548–562. https://doi.org/10.1016/j.jmva.2005.04.002

Rogers EM (1962). *Diffusion of Innovations*. Free Press.

Román-Román P, Serrano-Pérez J, Torres-Ruiz F (2019). A note on estimation of multi-sigmoidal Gompertz functions with random noise. *Mathematics*, 7(6): 541. https://doi.org/10.3390/math7060541

Sablik T (2013). Recession of 1981–82. https://www.federalreservehistory.org/essays/recession-of-1981-82.

Sidik K, Jonkman JN (2016). A comparison of the variance estimation methods for heteroscedastic nonlinear models. *Statistics in Medicine*, 35(26): 4856–4874. https://doi.org/10.1002/sim.7024

Song H, Chen H (2021). Asymptotic distribution-free changepoint detection for data with repeated observations. *Biometrika*, 109(3): 783–798. https://doi.org/10.1093/biomet/asab048

Truong C, Oudre L, Vayatis N (2020). Selective review of offline change point detection methods. *Signal Processing*, 167: 107299. https://doi.org/10.1016/j.sigpro.2019.107299

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17: 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang X, Erdman C, Emerson JW (2018). *bcp: Bayesian Analysis of Change Point Problems*.

Wu CF (1981). Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, 9(3): 501–513.

Wu H, Schafer TLJ, Ryan S, Matteson DS (2024). *Drift vs Shift: Decoupling Trends and Changepoint Analysis*.

Yao YC, Au ST (1989). Least-squares estimation of a step function. *Sankhya. The Indian Journal of Statistics*, 51(3): 370–381.

Yau CY, Zhao Z (2015). Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 78(4): 895–916. https://doi.org/10.1111/rssb.12139

Zhou H, Liang KY (2008). On estimating the change point in generalized linear models. In: *IMS Collections Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (N Balakrishnan, EA Peña, MJ Silvapulle, eds.), 305–320. Institute of Mathematical Statistics.