Song S. Qian

Environmental and Ecological Statistics with R

In memory of my grandmother 张一贯，mother 仲泽庆, and father 钱拙.

Contents

Chapter 11

Evaluating Models Based on Statistical Significance Testing

11.1 Introduction

Applications of statistics can be grouped into two categories: model development and model evaluation. Model development is to propose a hypothesis or model and model evaluation is to assess the validity of the model. We discussed the differences between Fisher and Neyman-Pearson in Chapter 4. Throughout the book, I followed Fisher's hypothetical deductive approach because I believe the Fisherian approach is consistent with scientific methods. The hypothetical deductive approach, however, requires knowledge in both statistics and subject matter knowledge. Statistics helps us to propose feasible probabilistic distribution assumptions, while the subject matter knowledge ensures that the proposed model is reasonable.

When developing a hypothesis, I consult subject matter experts on the likely model forms. In addition, I follow Tukey's advice in conducting a thorough exploratory data analysis (EDA). Some EDA tools have been described in this book. A thorough EDA often allows me to provide a good summary of the data. But more importantly, a model derived based on EDA results is more likely consistent with data. After necessary revision of the initial model based on discussions with subject matter experts, the model will be fit to the data. Once a model is developed, we start the process of model evaluation. In addition to the usual model assessment steps described in, for example,

Chapter 5, I often use simulation to evaluate whether a mode can capture known features of the data.

In this chapter, I present a detailed critique of a model proposed for detecting ecological thresholds. This example illustrates how statistical simulation can be used to characterize a hypothesis test-based model, and how such characterization can be used for model assessment. With readily available computing power, computation intensive methods such as CART and bootstrapping are increasingly accessible. However, hypothesis testing using these computation intensive methods can often lead to an unintended "multiple comparison trap," where the result is selected among many comparisons or repeated hypothesis testing. For example, I presented a threshold estimation method borrowing the CART strategy of searching for the break point resulting in the largest reduction of response variable deviance [Qian et al., 2003a]. In that paper, I suggested that the statistical significance of the resulting "threshold" can be tested using a χ^2 test. The null hypothesis of the test is that the response variable distribution does not change along the gradient. The χ^2 test can be simplified to a t-test when the response variable distribution is normal. However, the process of finding the change point searches all possible splits to find the one resulting in the largest difference (in deviance) between the two groups. In other words, the estimated threshold is based on repeated comparisons of two groups. A t-test on the threshold is then misleading because the true type I error probability is much larger than the declared significance level. The type I error probability can be estimated by using a simple simulation.

To estimate the probability of making a type I error using simulation, we repeatedly draw data from a distribution specified by the null hypothesis and carry out the test for each set of simulated data. The probability of making a type I error is the fraction of time when the null is rejected. For this case, we can use the function `chngp` in Section 9.3.2 to calculate the change point and conduct a t-test to see if the difference in means is statistically different. Using a significance level of 0.05, we should reject the null hypothesis about 5% of the time.

```
#### R Code ####
set.seed(123)
reject <- 0
n.sims <- 50000
for (i in 1:n.sims){
   temp <- data.frame(X=runif(30), Y=rnorm(30))
   split <- chngp(temp)
   if (split==min(temp$X) | split==max(temp$X))
        p.value=0.5
   else
        p.value <- t.test(Y~I(X<split), data=temp,
                          var.equal=T)$p.value
   reject <- reject + (p.value<0.05)/n.sims
```

```
}
print(reject)
> [1] 0.1568
```

When using a sample size of 30, the type I error probability is 0.1568, not the declared 0.05. The inflated type I error probability is expected; and the larger the sample size is, the higher the type I error probability is. In other words, using a significance test on a CART model result can be misleading. I did not recognize the problem while writing the paper because my emphasis was on another change point model described in the same paper. The significance test of the change point was added to the paper during the review process.

A type I error in the context of identifying ecological threshold is a false positive result. With the increased popularity of computation intensive methods, the likelihood of inflated type I error probability becoming a problem is also increasing. In the rest of this chapter, I will focus on a more complicated model for detecting ecological threshold. The statistical issue behind the model is the same: the significance level of a test is not what is declared because of the multiple comparison trap. Furthermore, the rejection of the null hypothesis does not imply support to a specific alternative hypothesis. When we are to conclude a step function as the alternative model, rejecting the null of no change is not enough.

11.2 Evaluating TITAN

Model evaluation is a difficult task requiring knowledge in both statistics and the subject matter science. We have discussed model evaluation extensively in this book, with a focus on evaluating model assumption compliance. In this example, I will focus on the assessment of a model or method based on statistical significance test. A statistical significance test is often characterized by the probabilities of making type I and type II errors. A properly designed test (based on the Neyman–Pearson lemma) sets the probability of type I error to be a small constant (the significance level α) and minimizes the type II error probability. Briefly, a test starts with a definition of the null hypothesis and a test statistic. The test statistic has a known probability distribution under the null hypothesis, known as the null distribution. A rejection region is defined based on the null hypothesis such that the type I error probability is limited to α. The lemma shows that a test following this procedure has the smallest type II error probability among all possible tests.

Accordingly, when evaluating a significant test-based method, we should present (1) the null hypothesis, (2) the test statistic, (3) the null distribution, and (4) the rejection region. If this information can be clearly presented and

verified, the Newman–Pearson lemma guarantees that the test is optimal, in that the probability of making a type I error is limited to α and the probability of making a type II error is minimized. In scientific research, questions facing a scientist are often more complicated than the questions addressed by typical significant tests described in statistics textbooks. As a result, significant test-based methods developed to address scientific questions are often a combination of more than one "textbook test." Consequently, the evaluation of these methods can be difficult. However, the basic concepts of a significant test apply. We should still be able to characterize a test-based method using the probabilities of making type I and type II errors.

As in the simulation for comparing a nonparametric test and the t-test in Section 4.5.4 (page 106), we examine the test by evaluating the type I error probability under the condition of the null hypothesis model. Comparing this probability with the declared significance level (α) we can describe the performance of the test when the null hypothesis is true. Furthermore, we examine the behavior of the test under selected alternative models by estimating the test's statistical power.

The concept of ecological threshold is appealing to environmental managers as it implies that a critical point, beyond which the ecosystem in question may be irrevocably changed, can be identified. Consequently, knowing the threshold would help the manager to set a goal for the protection of the ecosystem. In the Everglades example, we see the use of the concept in developing an environmental standard for total phosphorus. For example, Richardson et al. [2007] used a step function model to study how several ecological indicators change as a function of total phosphorus concentration. The step function is a type of threshold model. It assumes that the specific feature of the Everglades ecosystem measured by the ecological indicator does not change as phosphorus concentration increases. Once the concentration exceeds a threshold the indicator will jump to a different level and stay constant again as phosphorus concentration continues to increase. In Fisher's hypothetical deduction framework, the model is a hypothesis to be tested. I discussed issues related to model evaluation of this type of models in Qian [2014a]. In this section, I present an evaluation of a more complicated hypothesis testing based model for estimating ecological threshold.

11.2.1 A Brief Description of TITAN

Baker and King [2010] presented a program known as the threshold indicator taxa analysis (TITAN) for calculating community "thresholds." TITAN is often applied to find the level of disturbance beyond which a significant change at a community level is expected. Data used by TITAN are species abundances along an environmental or disturbance gradient. For each species (or taxon), the program finds a splitpoint along the gradient to divide the samples into two groups. The splitpoint is selected based on an indicator value that describes a taxon's association with a number of existing clusters

[Dufrêne and Legendre, 1997]. When there are only two clusters, the indicator value (IV) is the product of the taxon's relative abundance and its frequency of occurrence:

$$IV_i = A_i B_i \tag{11.1}$$

where $i = 1, 2$ is the cluster index, A_i is the relative abundance (fraction of individuals of the taxon in cluster i) calculated as the ratio of the mean abundance in cluster i (a_i) over the sum of the cluster means ($A_i = a_i/(a_1 + a_2)$), and B_i is the frequency of occurrence (fraction of non-zero observations) in cluster i.

While IV was developed to describe the association of a given taxon to an existing cluster, TITAN uses IV to define clusters along a disturbance gradient. Observations along the gradient are successively divided into two groups by moving a dividing line along the gradient; an IV is calculated for each group at each potential splitpoint. Baker and King [2010] defines the IV for each potential splitpoint as the larger of the two IVs and selects the splitpoint as the one with the maximum IV. The splitpoint selected under this definition is the same as the splitpoint with the largest difference between the two IVs based on the definition of Dufrêne and Legendre [1997]. This process searches for a maximum of the indicator value differences between the two groups. The gradient value associated with the maximum IV value is identified as an ecological threshold. Because the calculation of IV requires two clusters (or groups in this case), TITAN starts the search at some distance from the low end of the gradient to allow a pre-determined number of data points to be included in the "left group" and ends also at a distance from the upper end of the gradient to allow the same minimum number of data points in the "right group." This calculation is equivalent to a truncated variable transformation (truncating the data at both ends of the disturbance gradient and transforming the total abundance data into IV).

Two statistical inference methods were used on the maximum. One is the permutation test for "statistical significance" of the identified threshold. When the maximum is statistically "significant," the location of the maximum along the gradient is used as the estimate of the threshold. The other inference is about the uncertainty of the estimated threshold, using the bootstrapping method. This process of threshold identification (the permutation test) and estimation (bootstrapping) is repeatedly applied to each taxon separately. Thresholds for individual taxa are combined through the use of normalized IV values to derive the "community" threshold. I will focus on the evaluation of the permutation test because it is the basis of the subsequent analyses.

An obvious problem of TITAN is the use of the permutation test on the maximum of IV along the gradient. This problem is similar to the multiple comparisons in an ANOVA problem. That is, when multiple variables are drawn from the same population and we only compare the pair of variables with the largest sample mean difference, the t-test will reject the null hypothesis of no difference more often than the declared significance level α (the probability of making a type I error). This is because the two samples are no

longer simple random samples. Likewise, when applying the permutation test on the two groups with the largest IV value, the data are not simple random samples. Often the violation of the independence assumption is not obvious. As a result, we should evaluate the method by its probabilities of making type I and type II errors. A type I error is erroneously rejecting the null hypothesis when the null is true. A type II error is the failure to reject the null hypothesis when the alternative hypothesis is true. Type I error probability can be estimated by carrying out the test using data from the null hypothesis model, which requires that we know the null hypothesis model and are able to draw random data from the model. As in Section 4.5.4 (page 106), we will use simulation to evaluate the probability of making a type I error. A type II error (and the power) is associated with a specific alternative hypothesis. As a result, I will specify a number of relevant alternative models. Unlike the simulations in Section 4.5.4 where the null hypothesis model is a simple normal distribution, the null hypothesis model in this case is not defined in the TITAN literature. Instead, a vague description of one class of potential alternative hypothesis models is mentioned.

11.2.2 Hypothesis Testing in TITAN

TITAN derives the threshold based on a series of hypothesis tests. First, for a given taxon at a given splitpoint, TITAN uses a permutation test. The data distribution assumption was not clearly defined, hence the preference of a "distribution-free" test. The data used to calculate the IVs are the taxon abundance values observed along an environmental gradient. Using hypothesis testing, the TITAN authors want to learn whether the estimated IV is "statistically significant." When a splitpoint is "significant," the respective gradient value is deemed as a potential threshold. When we used the term "statistically significant" in Chapter 4, we mean that the observed data show strong evidence against the null hypothesis model. That is, statistical significance is relative to the null hypothesis model. What is, then, the null hypothesis in the permutation test incorporated in the TITAN program?

The null hypothesis was never clearly stated in a way that can be used to formulate a proper model. But based on what a permutation test does, I will try to deduce what is the null. The observed IV is calculated from two subsets of taxon abundance data separated by the splitpoint along the disturbance gradient. For a specific splitpoint, the sample sizes of the two subsets (n_1 and n_2) are known. The permutation test, in theory, tabulates all possible permutations of splitting the data into two subsets of sample sizes n_1 and n_2. For each permutation, an IV is calculated. The collection of these IVs is used to form an empirical distribution, and the p-value is calculated as the fraction of these IVs exceeding the observed IV. In other words, the distribution of these permutation-derived IVs is taken as the null distribution. The null hypothesis, with respect to taxon abundance, must be that taxon abundance distribution is not affected by the disturbance gradient. Because

abundance is a count variable, we can approximate its distribution using the Poisson distribution or its overdispersion variants; the null hypothesis can be simplified to be a constant mean abundance along the gradient.

Because TITAN uses IV as a test statistic, the distribution of IV under the null hypothesis is unknown. Using a permutation test, we can avoid the derivation of a theoretical null distribution of IV. As in all hypothesis testing situations, the null hypothesis is a specific statistical model and the alternative is undefined. Two questions arise:

1. If the permutation test of the IV of one splitpoint has a significance level of $\alpha = 0.05$, what is the type I error probability of the test of the maximum IV?

2. Why is the rejection of the null hypothesis equivalent to a threshold response model?

We address these questions using simulation.

11.2.3 Type I Error Probability

Characterizing a test's type I error probability can be easily done by using simulation. That is, we repeatedly carry out the test on data drawn from a model specified by the null hypothesis. The frequency of rejection is an estimate of the type I error probability. If the frequency is close to the declared significance level of α, the test behaves as expected. If the frequency is very different from α, the test is problematic. A frequency much lower than α suggests that the test has a probability of making a type I error much lower than the declared significance level. A lower than expected type I error probability implies a higher type II error probability (hence a lower statistical power). A test with a type I error probability higher than α is associated with a higher than expected power. At the same time, it is also more likely to reject the null hypothesis when the null is true.

As in Chapter 4, we can easily evaluate a test's type I error probability by using simulation: repeatedly drawing data from the null hypothesis model and carrying out tests on these fake data to record the frequency of rejection. In this example, TITAN's null hypothesis is that a taxon's mean abundance does not change along the gradient of interest. To simulate data likely to happen under the null hypothesis, I will assign a number of sampling points and draw taxon abundances from a Poisson distribution with mean of 20:

1. The simulation starts with a predetermined number (`ns`) of sampling points along a gradient between 0 and 1: `x <- seq(0,1,,ns)`.

2. For each sampling point, a taxon abundance is drawn from a Poisson distribution: `y <- dpois(ns, 20)`.

3. Assuming the resulting `ns` Poisson random variates are taxon abundance

data along the gradient, equation (11.1) is used to calculate IV values for all potential splitpoints, using TITAN's default setup. This step leads to the test statistic – the maximum of these IV values, and the gradient value at which IV is the maximum.

4. A permutation test is carried out (at the splitpoint with the maximun IV) to derive the null distribution and a p-value is calculated. When the p-value is less than $\alpha = 0.05$, the null is rejected.

5. The process is repeated 5000 times to record the number of times the null is rejected.

I repeated the simulation using `ns = 15, 25, 51, 101,` and `201` to evaluate whether the type I error probability is a function of sample size. The estimated type I error probabilities are 0.14, 0.23, 0.31, 0.31, and 0.30, respectively. These numbers are considerably larger than the significance level of $\alpha = 0.05$. Furthermore, it seems that the type I error probability increases as the number of sampling points increases. When discussing ANOVA, we mentioned the difference between family-wise and test-wise type I errors. In an ANOVA setting, a family-wise type I error concerns the rejection of one of many possible comparisons. This concern arises when multiple comparisons are of interest. In a multiple comparisons problem, we used the Tukey's method, where the null hypothesis distribution of the largest difference is derived. TITAN is similar to a multiple comparisons problem. The permutation test used in testing the significance of the maximum IV has a comparison-wise significance level of 0.05. As such, the "family-wise" type I error probability is always higher than 0.05.

Perhaps in an attempt to correct for the considerably higher than expected type I error probability, TITAN also uses a normalized IV calculated based on the permutation test. The normalized IV is called the z-score.

In the permutation test, a number of IVs are calculated at each potential splitpoint, one for each random permutation. These IV values form the null hypothesis distribution of IV. The mean ($\hat{\mu}_i$) and standard deviation ($\hat{\sigma}_i$) of these IV values are used to normalize the observed IV value:

$$z_i = \frac{IV_i - \hat{\mu}_i}{\hat{\sigma}_i}$$

Although the methods section of Baker and King [2010] stated that the gradient associated with the highest IV is used as the estimated threshold, the gradient value associated with the highest z_i value is clearly used as the threshold in the accompanying computer code. In other words, the test statistic is the z-score. Because it is the normalized IV, using z-score as the test statistic appears to assume that the z-score is a standard normal random variable under the null hypothesis. In the previous simulation, I also calculated the z-score and the type I error probability using the z-score as the test statistics, which are 0.14, 0.23, 0.31, 0.31, and 0.30, for $ns = 15$, 25, 51, 101, and 201,

respectively. In other words, the change of test statistic did not make any difference in terms of the type I error probability.

There was little discussion on whether there would be a difference in the estimated threshold value when the test statistic is switched from IV to the z-score. It seems that the authors of TITAN assumed that the two test statistics will result in the same outcome (the same p-value and the same estimated threshold value). But this assumption was not obvious to me. To compare the two statistics under the null hypothesis, I carried out another simulation with `ns=101`. This time, I calculate IV and the z-score for all potential splitpoints. That is, a permutation test is carried out at each potential splitpoint. As before, the gradient is between 0 and 1 and the 101 sampling points are evenly spaced along the gradient. Using this simulation, I illustrate the pattern of IV and z-score along the gradient.

The null hypothesis assumes a constant taxon abundance along the gradient, which is simulated by drawing random count variables along the gradient from the same Poisson distribution with a mean of 20 (Figure 11.1(a)). For each iteration of the simulation, I calculate IV, as well as $\hat{\mu}, \hat{\sigma}$, and z-score, for all potential splitpoints with a minimum of five data points in each group. There are a total of 92 potential splitpoints. After the process is repeated 5000 times, there are 5000 simulated $IV, \hat{\mu}, \hat{\sigma}$, and z-scores at each potential splitpoint to approximate the distributions of these statistics along the gradient. The simulated IV distributions show a distinct pattern (Figure 11.1(b)); the means and standard deviations near both ends of the gradient are higher than the same near the middle of the gradient. The pattern in IV along the gradient suggests that we are more likely to identify split points near both ends as thresholds using IV even when the underlying taxon abundances are the same across the gradient.

Permutation means, as well as standard deviations, show a similar pattern as IV values, high near both ends of the gradient and low near the middle of the gradient (Figure 11.2). However, the estimated z-scores show no discernable pattern along the gradient (Figure 11.1 (c)). The distribution of these simulated z-scores is very close to the standard normal distribution at all potential splitpoints. The type I error probability would be around 0.05 at any given potential splitpoint if we reject the null when $z > 1.96$.

The z-score distributions for all potential splitpoints are the same ($N(0, 1)$). Yet, our first simulation resulted in a type I error probability much larger than the significance level. Figure 11.1(c) explains the cause of the inflated type error probability: a test on the maximum IV along the gradient is more likely to be statistically significant, just like the comparison of the maximum difference among the differences of multiple pairs of means in an ANOVA problem.

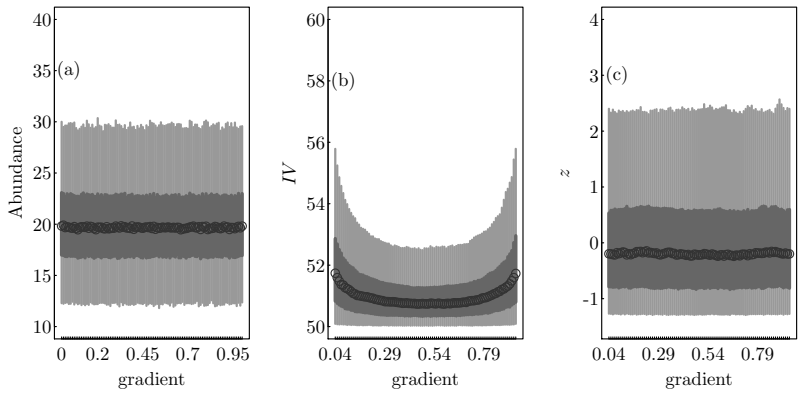

FIGURE 11.1: Simulated data under the null hypothesis model (a), IV (b), and z-score (c) distributions are shown along the gradient. Taxon abundance data at each location along the gradient were drawn from the same Poisson distribution with mean 20. The box plots show a summary of 5000 estimated quantities at each gradient location.

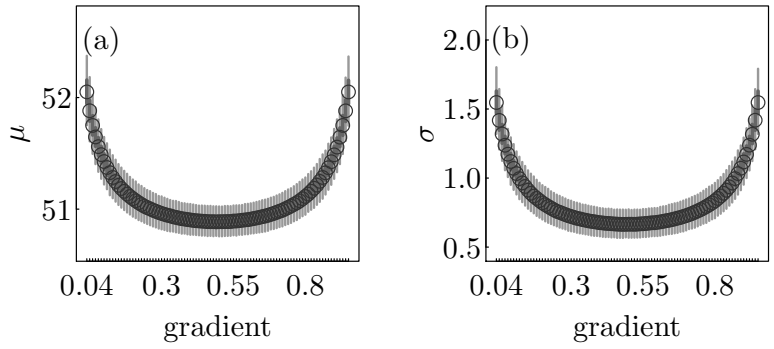

FIGURE 11.2: Permutation estimated mean (μ) and standard deviation (σ) of IV change along the gradient. The box plots show a summary of 5000 estimated quantities μ and σ at each potential splitpoint.

11.2.4 Statistical Power

The objective of TITAN is to find a threshold value. The stated goal of TITAN was "exploring and identifying abrupt changes in both the occurrence frequency and relative abundance of individual taxa along an environmental, spatial or temporal gradient." Using IV as the measure of such changes, Baker and King [2010] assume that the threshold value is the point along the gradient corresponding to the maximum IV value, as they described that when the maximum of IV along a gradient is statistically "significant," a threshold is identified.

A statistically significant result suggests that the null hypothesis is rejected. However, rejecting the null is not the same as evidence supporting a specific alternative hypothesis. There are many "threshold" models, among other possible alternatives. Only one is the "useful" model. The question we need to address after the null hypothesis is rejected should be "which model is supported by the data?" In a t-test, when the null hypothesis of no difference is rejected, we report the observed difference as an estimate, together with the estimated confidence interval of the difference. The confidence interval represents a range of the difference (alternative hypothesis) that is supported by the data. In other words, the confidence interval narrows down the range of all alternative hypotheses. Because the taxon abundance model is a function of the gradient, a likely model should be explored when the null hypothesis is rejected.

The stated goal is vague in that it does not allow a user to write a specific alternative hypothesis model with respect to taxon abundance data. The term abrupt change is not defined. One way to elucidate the meaning is to try several typically used mathematical forms for modeling abrupt change. For each potential alternative model, I will calculate the IV for each potential splitpoint on a set of simulated data without error and see which model results in a peak IV at the known "threshold." These alternative models are:

- Step function (SF) model

  This is the simplest model, stating that the response variable stays constant as we move along the gradient (the x-axis) until reaching a threshold (change point). Once crossing the change point the response variable jumps to a different value and stays constant again (Figure 11.3(a)).

  Mathematically, a step function model for a normal response variable is

  $$y_i = \beta_0 + \delta_1 I(x_i - \phi) + \varepsilon_i \tag{11.2}$$

  We can assume a constant variance ($\varepsilon \sim N(0, \sigma^2)$) or allow the variance to change when crossing the threshold (i.e., $\varepsilon_i \sim N(0, \sigma^2 + \delta_2 I(x_i - \phi))$). The function $I(\theta)$ is a unit step function, taking value 0 when $\theta \leq 0$ and 1 otherwise. The SF model has one discontinuity in the function itself, which is the threshold of interest.

- Hockey stick (HS) model

  This function assumes that the response variable changes as a linear function of the gradient with the slope changes at a threshold. The model resembles two line segments joined at the threshold (Figure 11.3(b)).

$$y_i = \beta_0 + (\beta_1 + \delta_1 I(x_i - \phi))(x_i - \phi) + \varepsilon_i \tag{11.3}$$

  The threshold of interest for the HS model is location of the discontinuity in slope (or the first derivative of the function).

- Disjointed broken stick (dBS) model

  A generalization of both the SF and HS models is the broken stick model. It is a model of two line segments with no constraints on the slope (SF model has two line segments with a common slope of 0) nor on intercept (HS model has two line segments joined at the threshold, or the same intercept with respect to the threshold) (Figure 11.3(c)).

$$y_i = (\beta_0 + \delta_0 I(x_i - \phi)) + (\beta_1 + \delta_1 I(x_i - \phi))(x_i - \phi) + \varepsilon_i \tag{11.4}$$

  The dBS model has discontinuities in the function and the first derivative of the function, and they coincide in the same location (hence the threshold of interest).

- Sigmoidal (SM) model

  A sigmoidal model is a continuous nonlinear model with lower and upper bounds, but without a change point (no parameter change along the gradient) (Figure 11.3(d)). I include the SM model because a threshold is often defined as a rapid change in the response over a short distance of the gradient. A change in one or more model parameters is not required. In other words, an abrupt change does not necessarily imply a discontinuity in the function or its derivatives. It can be simply a "rapid" (but smooth) change. I will use a simple logistic model as an example.

$$y_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \tag{11.5}$$

  The SM model is the "smooth" version of the SF model. It is a continuous function with a continuous first derivative. The slope of the curve (first derivative) reaches the maximum (or minimum) at the inflection point. It is, therefore, natural to consider the inflection point as the threshold of interest.

To show how the test statistic changes along the gradient as a function of taxon abundances, I use simulated data without error. That is, I will assume that the pattern of change in taxon abundances along the gradient can be described by one of the four models and data are observed without error. Using data without error provides information on the behavior of IV as a

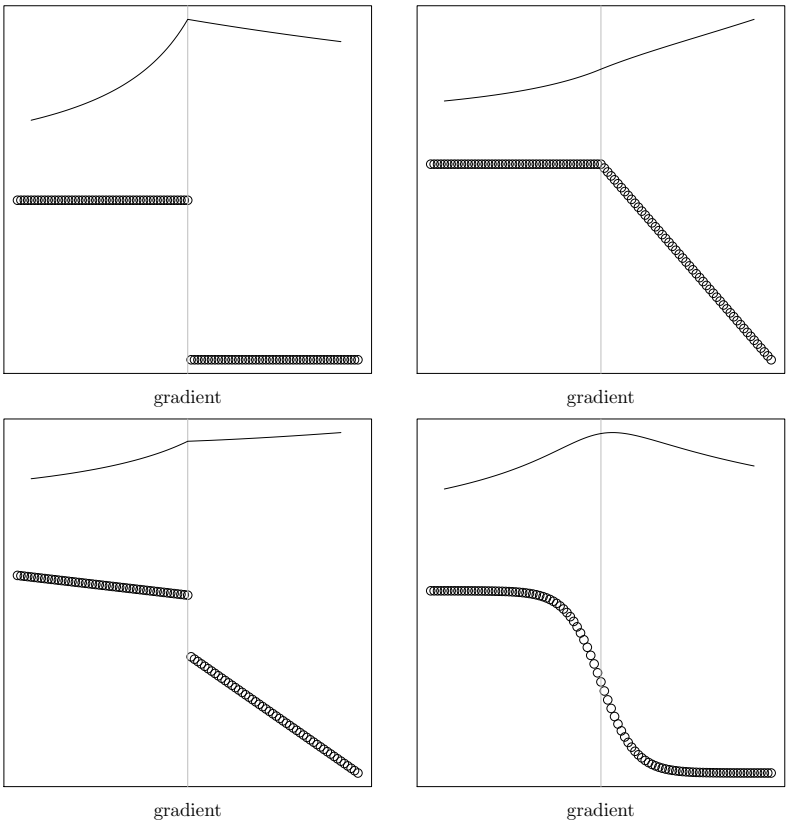

FIGURE 11.3: Data from four potential alternative models describing "abrupt" changes in abundance (open circles) are compared to the calculated IVs along the gradient (solid line).

function of the gradient under the four models. I am particularly interested in the location of the peak IV in comparison to the known "threshold." The simulated data are compared to the calculated IVs (Figure 11.3). It appears that the peak of IV coincides with the known threshold when the underlying model is the SF model only. For the HS model, the estimate IV is a monotonic function of the gradient. For the dBS model, the peak IV is either close to the known threshold or at one end of the gradient, depending on the difference in the slopes and intercepts of the two line segments. When slopes of both line segments are approaching 0, the peak is approaching the known threshold. For the SM model, IV peak is close to, but not at, the inflection point.

To estimate the statistical power of the test, I repeat the simulations in Section 11.2.3, except that the taxon abundance data are drawn from Poisson distributions with means calculated by the alternative models. The `ns` sampling

locations are evenly spaced along the gradient (or `grd <- seq(0,1,,101)`). The taxon abundance data are drawn as follows:

- The SF model: `y <- dpois(ns, 20+(grd>0.5)*10`,
- The HS model: `y <- dpois(ns, 20+(grd>0.5)*20*(grd-0.5)`,
- The SM model: `y <- dpois(ns, 10*invlogit(-5+10*grd)`, and
- The linear model: `y <- 20+10*grd`.

The function `invlogit` is from the package `arm`. Simulation of the dBS model is left as an exercise. The estimated statistical powers for these alternative models are all nearly 1 (between 0.9997 and 0.9999).

These high powers are not surprising because the test has a type I error much higher than the significance level of 0.05. The results show that TITAN will almost surely reject the null hypothesis when the underlying model is different from the null hypothesis model, including the linear model. As usual, when using a hypothesis testing, we are focused on evidence against the null hypothesis, not evidence supporting a specific alternative hypothesis. The power analysis illustrates the point that rejecting the null will not lead to the acceptance of a specific alternative.

Once the null is rejected, TITAN will estimate the threshold. Both IV and z-score are used as the test statistics. To see what is the estimated threshold under the alternative models (Figure 11.3), another round of simulations are carried out. These simulations are designed to characterize the response of IV and z-scores under different alternative models. I will include the liner model as an alternative model.

In these simulations, I derive the distribution of IVs and z-scores for all potential splitpoints. These simulated distributions are presented along the gradient using boxplots.

The linear model assumes that the abundance increases linearly along the gradient (Figure 11.4(a)). This model is different from the null hypothesis model, but without an abrupt change along the gradient. Under the linear model, the IV decreases along the gradient (Figure 11.4(a)), while the permutation estimated means and standard deviations still show the same pattern as before. The resulting z-scores are now shown a clear pattern and we have a statistically "significant" splitpoint near the middle of the gradient (Figure 11.4(c)).

At the low end of the gradient, the range of the middle 95% of the abundance data is between 12 and 30, and the same range at the upper end of the gradient is between 20 and 42. In other words, the null model (e.g., a constant abundance of 25) is within the 95% range of the alternative model. Nevertheless, the statistical power of TITAN for this linear model is still nearly 1, an outcome of the inflated type I error probability due to the multiple comparison trap.

A statistically significant result is desirable because the null hypothesis

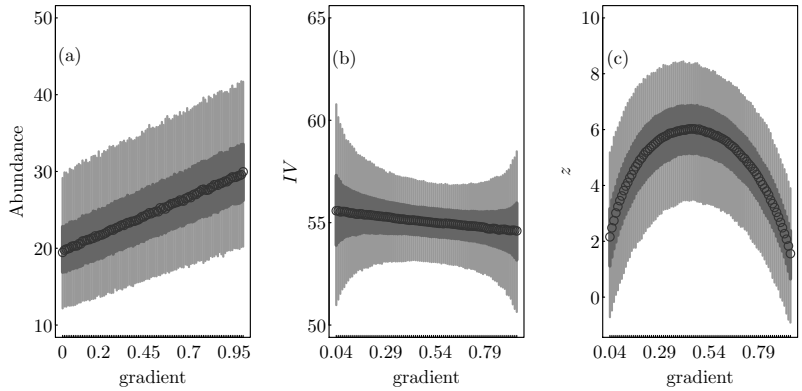

FIGURE 11.4: As in Figure 11.1, but the mean taxon abundance increases along the gradient linearly as shown in panel (a).

model is different from the data (generated from a linear model). But the result does not lead to the conclusion of an abrupt change along the gradient. The data used for the test were drawn from a model with a steady rate of increase. Just as rejecting the null hypothesis in a t-test does not support a specific alternative hypothesis, the rejection of the null hypothesis model of a constant abundance along the gradient cannot be equated to a support for a threshold response model.

Another issue of using TITAN is that the estimated "threshold" value based on IV will be different from the estimated value based on the z-score. The IV peak for the linear model is near the low end of the gradient, while the z-score peak is near the middle of the gradient. This discrepancy is never shown in all applications of TITAN. In a statistical change point problem, if the estimated change point is located near one end of the gradient, we conclude that the change point does not exist. In TITAN, IV is used as an indicator of the presence of a "threshold." If the peak of IV is located near one end of the gradient, we should conclude that no "threshold" is present. In this case, the IV peak is at the low end of the gradient. The standardized IV (the z-score) is calculated by subtracting $\hat{\mu}$ from IVs calculated for a splitpoint and the difference is divided by $\hat{\sigma}$. Because $\hat{\mu}$ is smaller in the middle of the gradient, the difference of $IV - \hat{\mu}$ will be inflated near the middle. Likewise, $\hat{\sigma}$ is lower in the middle, further inflating the difference near the middle of the gradient. As a result, the peak of z-score in this case is likely an artifact of the permutation-based standardization. From a hypothesis testing perspective, this discrepancy is inconsequential because the goal is to evaluate the null hypothesis model of a constant taxon abundance along the gradient. However, because TITAN's authors equate a significant result to the existence of a threshold, the use of the z-score is now misleading.

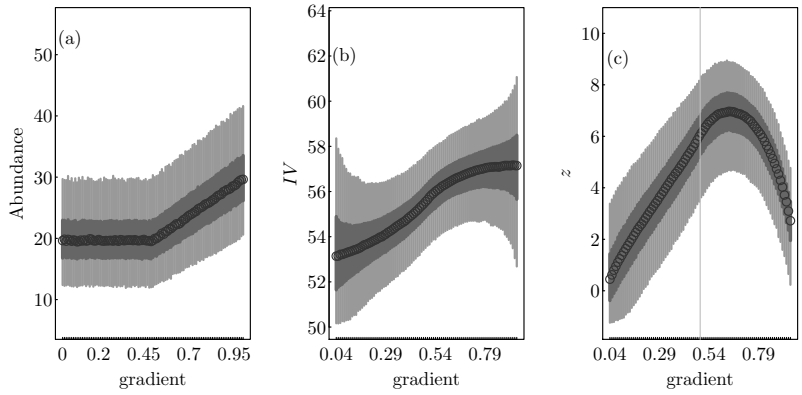

FIGURE 11.5: As in Figure 11.1, but the mean taxon abundance is modeled by a HS model (a).

When the alternative is the HS model (a constant abundance when the gradient is below 0.5 and a linearly increasing abundance after the change point, Figure 11.5(a)), IV increases monotonically along the gradient (Figure 11.5 (b)), while the z-scores show a peak near 0.65 (Figure 11.5 (c)).

Again, the power of the test is practically 1. Although the HS model does have a change point, the estimated "threshold" based on the z-score is not what we expected (0.5); neither is the location of the peak IV.

When the underlying model is the SF model, we have both a power of near 1 and a correctly identified threshold (Figure 11.6).

When the SM model is the underlying model, TITAN will also result in a significant result almost surely. The threshold of interest for the SM model is the inflection point located at 0.5. But the estimated threshold based on the z-score depends on whether the abundance is increasing or decreasing. If the abundance is increasing (as in our simulation), the estimated threshold is below the inflection point. If the abundance is decreasing along the gradient, the estimated threshold is above the inflection point. Furthermore, the test result is also influenced by the rate of change in an SM model. The larger the (maximum) slope is, the closer is the estimated threshold from the inflection point (Figures 11.8 and 11.9).

These simulations suggest that TITAN is effective in terms of rejecting the null hypothesis when the null is not true. However, TITAN has a type I error probability much larger than the declared significance level of 0.05. As a result, it can be overly sensitive to meaningless deviations from the null and a significant result may be practically meaningless. Because a statistical significance test is focused on evidence against the null hypothesis model, a significant result does not imply a threshold response of taxon abundance to disturbance represented by the gradient. Furthermore, TITAN estimated

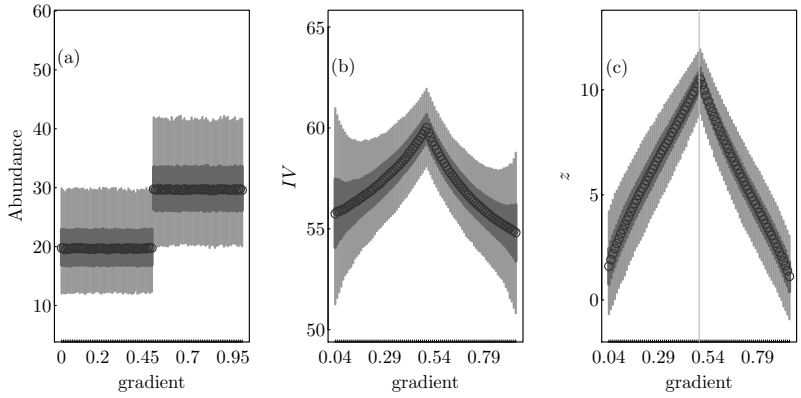

FIGURE 11.6: As in Figure 11.1, but the mean taxon abundance is modeled by a SF model (a).

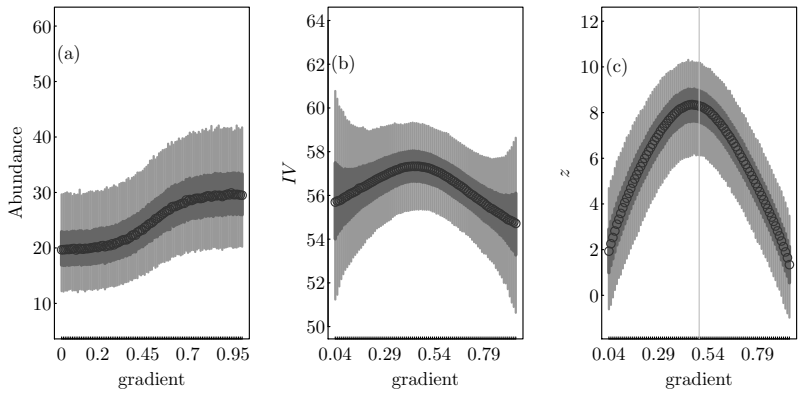

FIGURE 11.7: Same as in Figure 11.1, except that the mean taxon abundance is modeled by a SM model (a).

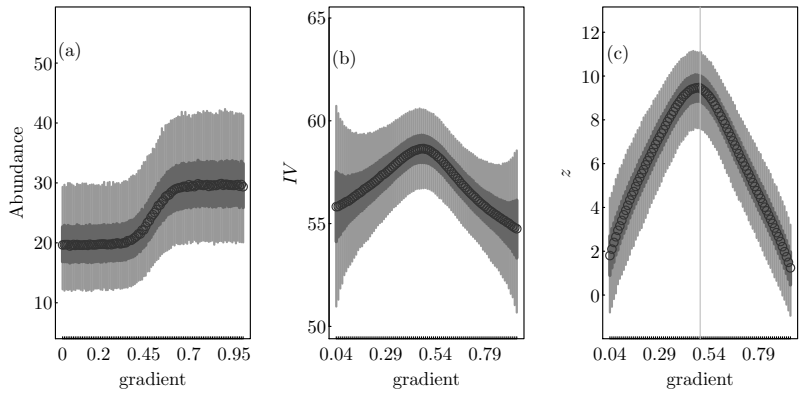

FIGURE 11.8: Same as in Figure 11.7, with a maximum slope twice as large as the same in Figure 11.7.

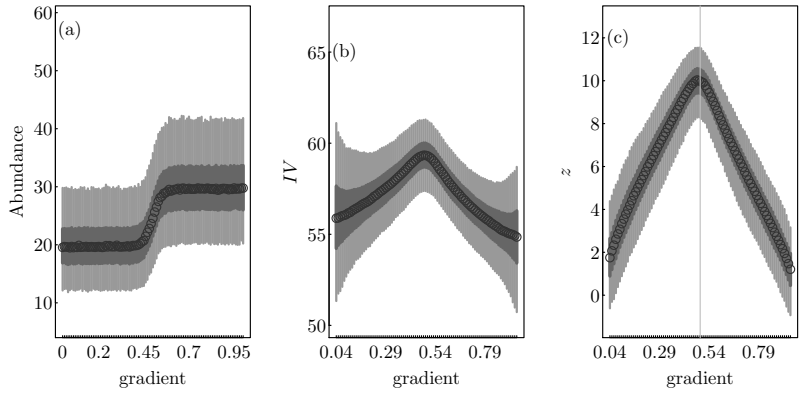

FIGURE 11.9: Same as in Figure 11.7, with a maximum slope 4 times larger than in Figure 11.7.

threshold is correct only when the underlying response pattern is consistent with the SF model.

Because the objective of TITAN is to quantify a specific threshold response model, a significance testing approach is not appropriate. The threshold response model is a specific alternative model. As a result, rejecting the null hypothesis of a constant abundance along the gradient does not translate to evidence supporting any specific alternative model. When a specific model is of interest, we fit the specific model to data and perform model evaluation as outlined in Chapter 9.

In these simulations, the estimated μ and σ have a consistent pattern along the gradient. They tend to be higher near both ends of the gradient than they are in the middle. Because the z-scores are the normalized IVs, z-scores near both ends of the gradient will be lower than z-scores near the middle if we have a constant IV along the gradient. When the null model is true, IVs near both ends are also higher than those near the middle. The net result is a more or less constant z-score along the gradient. When the null model is not the underlying model, for example, when the abundance changes as a linear function of the gradient, the IV will also change monotonically. The patterns in the permutation estimated μ and σ will result in the location of the peak of the z-score being different from the location of the peak of IV, resulting in contradictory estimates of the "threshold." In general, the peak z-score location will be closer to the middle of the gradient than the peak IV location. Apparently, the authors of TITAN did not recognize the potential contradictory results. In the R program accompanying [Baker and King, 2010], the permutation test is used to report the p-value (associated with the maximum of IV) and the peak z-score is used to identify the threshold.

11.2.5 Bootstrapping

TITAN also used bootstrap resampling to calculate the confidence interval of the selected split point. Bootstrapping is a commonly used resampling method for estimating standard deviation and confidence intervals of statistics [Efron and Tibshirani, 1993]. As we discussed in Chapter 9, bootstrapping is a Monte Carlo simulation procedure aimed at obtaining an approximate sampling distribution of the parameter of interest. It substitutes random samples from the target population with random samples of the same size (with replacement) from the existing data. As the sample size of the data increases, bootstrap samples become increasingly closer to random samples from the population. As a result, an empirical distribution of variable calculated from bootstrap samples approximates the true sampling distribution of the variable of interest as sample size increases. The bootstrap method is, however, not appropriate for a splitpoint problem [Bühlmann and Yu, 2002, Banerjee and McKeague, 2007]. Bühlmann and Yu [2002] have shown that the bootstrap estimated standard deviation of the splitpoint is always smaller than the true standard deviation, leading to a narrower confidence interval. In a

splitpoint problem, a sample with k unique gradient values has $k-1$ potential splitpoints. Potential splitpoints in a bootstrap sample are a subset of the same $k-1$ potential splitpoints. In other words, the bootstrapping process repeatedly selects the splitpoints from the same pool of $k-1$ potential splitpoints with the result that the bootstrap estimated standard deviation is much smaller than it should be (and the estimated confidence interval is much narrower than it should be). The simulation carried out in Chapter 9 demonstrates this problem.

In addition to the problem of estimated confidence interval, using bootstrapping may also have an "edge effect." When using bootstrapping, only a subset of the individual data points is represented in a bootstrap sample. When data points near one or both ends of the gradient are not included, the range of the gradient is further truncated as we still must maintain the minimum number of data points in the two groups separated by a splitpoint. As the permutation test is not used when running bootstrapping in TITAN's R program, TITAN uses IV in each bootstrap simulation to identify the threshold. If the null model is true, IVs tend to be higher on both ends of the gradient. As a result, bootstrapping is likely to identify one or the other end of the gradient as the threshold. Because the range of the gradient in a bootstrap sample is frequently narrower than the range of the data, a bootstrap estimated distribution of the threshold would be shifted towards the middle of the gradient, resulting in an apparent "threshold" distribution slightly away from the end of a gradient.

11.2.6 Community Threshold

The last step of TITAN is to derive the community threshold based on the sum of z-scores calculated for all taxa at all potential splitpoints. This step implies a hypothesis test of the existence of a community threshold. TITAN calculates the sum of z-scores for every potential splitpoint x_i, which is equivalent to testing whether all taxa share the same splitpoint. TITAN's community threshold is implicitly defined as the splitpoint shared by all or most taxa and is selected as the splitpoint with the largest sum of z-scores, through repeated tests for all potential splitpoints. Statistically, such a test is meaningful only if this community threshold definition is meaningful. Under this definition, the sum of z-scores is a test statistic. Z-scores of individual taxa are random samples from the standard normal distribution. If we assume that these z-scores are independent of each other, the sum of these z-scores is a random variable with mean 0 and standard deviation $\sqrt{N}$, where N is the number of taxa. That is, we use the sum as a composite test statistic for the common null hypothesis for all taxa. We note that the sum of squares of the z-score, which has a χ^2 distribution with degrees of freedom of N, is more frequently used. Because this test is repeated for all potential split points, the resulting "community threshold" is likely a result of the type I error (the multiple comparison trap). However, the meaning of the test is moot as the

community threshold defined by this test violates basic ecological principles. Species coexistence forms the conceptual basis for much of community ecology [MacArthur, 1972, Chesson, 2000, Hubbell, 2001] and there is a large body of theoretical work that validates the conclusion of Hutchinson [1959] (species must be different in some way to coexist [Chesson, 1991]). The synchronicity in thresholds reported by King et al. [2011] suggests that coexisting species do not differ in their response (e.g., optima) to changes in environmental resources. Ecological theory (e.g., competition, species packing, resource utilization) suggests that co-occurring species should exhibit differences in species optima, tolerances, and peak abundances across the environmental gradient [Gauch, 1982, Jongman et al., 1995], particularly for species responses that are associated with changes along natural environmental gradients (e.g., elevation, temperature, nutrients, prey abundance). This partitioning of resources should result in species that exhibit different rather than similar response thresholds. Synchronicity might be expected if the environmental change is associated with a toxicant. However, low levels of urbanization or eutrophication are generally not associated with high levels of toxicants; therefore, the synchronicity of thresholds reported by King et al. [2011] cannot be attributed to toxicity. It is far more likely that the apparent synchronicity of thresholds is an artifact of the method used to extract the threshold (e.g., the z-score) rather than an ecological attribute of the community.

11.2.7 Conclusions

TITAN is intended for uncovering discontinuous jumps in taxa abundance data along a disturbance gradient. Instead of formulating specific models about abundance, TITAN's authors used the clustering indicator IV. The resulting program is ambiguous in terms of what kind of threshold was detected. The misuse of the permutation test resulted in a systematic bias in the selected splitpoint based on the z-score towards the center of the data cloud along the gradient. TITAN is written to process large data sets with hundreds of taxa from many sites. As a result, the behavior of the program is opaque. Furthermore, Baker and King [2010] did not give the mathematical and ecological definition of a community threshold, nor the threshold concept at the individual taxon level.

From these simulations, we learned that a statistical test is to assess the evidence against the null hypothesis. In a simple two sample t-test, rejecting the null hypothesis (of a 0 difference of two means) does not provide any evidence in favor of a specific value of the difference. To conclude a specific alternative, evidence supporting the specific alternative model must be provided. Because TITAN's objective is to estimate a threshold, a threshold model is the assumed pattern. We should seek evidence supporting the specific threshold response model. But no specific threshold model was provided. The method packaged in the program implies a null hypothesis model of a constant abundance along

the gradient. Rejecting the null model gives us no evidence of supporting any specific alternative model.

As we discussed in Chapter 4, a hypothesis test using a null of no difference should be used as a "devil's advocate." That is, we present our evidence in supporting the hypothesis of interest (in this case, a specific threshold model) and use the null hypothesis of no change as a last step to show that the data cannot be logically attributed to a model of no change. The null hypothesis test alone is not enough.

11.3 Exercises

1. In evaluating TITAN we used several alternative models to show the pattern of the IVs along an environmental gradient. Another natural pattern discussed by Cuffney and Qian [2013] is the Gaussian response model, where the response curve is similar to a bell-shaped curve. This response pattern is often used to represent the "subsidy-stress" response of a taxon. The initial increase in a pollutant (e.g., nutrient) provides subsidy to the growth of the organism, but the organism is stressed after the pollutant exceeds a threshold. The response pattern can be expressed as a parabola function of the gradient in log-abundance scale:

$$\log(y) = \alpha + \beta x + \gamma x^2$$

where y is the taxon abundance and x is the environmental gradient. A logical threshold is the peak of the quadratic curve. Draw the response curve of IV similar to the curves in Figure 11.3 and discuss whether TITAN is appropriate for this type of threshold response.

2. Often organisms sensitive to pollution are used for setting an environmental standard. Because they are sensitive to environmental disturbance, their abundance can often be 0 near one or the other end of the gradient (e.g., linearly decline and reaching 0 halfway across the gradient). Design a simulation to study the behavior of TITAN when there are multiple zero abundance observations near one or the other end of the gradient.