Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Corso di Laurea Magistrale in Ingegneria Matematica

INFERENCE IN FUNCTIONAL DATA ANALYSIS FRAMEWORK: SIMULATION STUDIES AND CODE OPTIMIZATION

Relatore: Prof. Simone VANTINI
Correlatore: Ing. Alessia PINI

Tesi di laurea di:
Marc Georges Giovanni LAMBLIN Matr. 799302

Anno Accademico 2014–2015

"The best thing about being a statistician is that you get to play in everyone's backward"

John Wilder Tukey

Ringraziamenti

In primis, vorrei ringraziare il prof. Simone Vantini per avermi dato la possibilità di lavorare a questa tesi. Sono molto riconoscente anche verso Alessia Pini che mi ha su(o)pportato e mi ha dedicato molto tempo. Ringrazio Aymeric Stamm per avermi fornito informazioni utili in materia di dottorati. Penso che l'aver svolto una tesi in ambito prettamente statistico abbia completato in maniera soddisfacente la mia formazione universitaria e possa costituire una buona base per il mio futuro lavorativo. Ringrazio anche a tutte le persone che hanno ideato e che mantengono la didattica del corso di studi in Ingegneria Matematica.

Sono profondamente grato nei confronti dei miei genitori che hanno creduto in me. Non mi avete mai fatto mancare nulla. Ancora più importante è stato il vostro costante supporto senza il quale non sarei mai riuscito a raggiungere questo traguardo.

Un grazie speciale è rivolto a Cristina. Forse senza di te non sarei nemmeno riuscito ad accedere in Magistrale, figuriamoci a concluderla. Grazie per essermi stata vicina anche nei momenti più improbabili.

Non posso non ringraziare i miei compagni del liceo Vale, Edo ed Ago che sono sempre stati al mio fianco e penso che lo saranno sempre. Un sentito ringraziamento va alla mia amica storica Ceci che per me è una certezza. Ringrazio Albi e Jorio che ho avuto modo di conoscere durante gli anni universitari e con cui ho stretto amicizia fin dal principio. Mi ricorderò delle giornate di studio con il magister Albi nella campagna veneta e delle discussioni con Jorio intorno all'Analisi che terminavano ad orari inverosimili. Spero di continuare a rimanere in contatto con voi nonostante la texanizzazione in atto e l'ovvia conseguente distanza geografica. Volevo ringraziare anche Mauro che per qualche anno avevo perso di vista. Sei ricomparso all'improvviso per puro caso e sempre per puro caso è nata un'amicizia imprevedibile. Thanks Ramin for your sincere friendship and for introducing me to a culture different from mine. *one-way* is still possible. Infine, volevo ringraziare tutte le persone incluse le vecchie guardie del Politecnico Open unix Labs per la qualità del loro operato di diffusione dell'open source e del software libero di primaria importanza per la mia formazione. In particolare, sono debitore nei confronti di Andrea Gussoni per avermi recuperato dei dati sensibili. Scusate per il passaggio da macrocosmo a microcosmo.

È doveroso ringraziare il bomber Mila che attualmente rischia di diventare il barista di fiducia e che è sempre fonte di grandi emozioni e risate. Ringrazio il mitico Morse che con il suo senso dell'umorismo e con il suo *argot* mi ha sempre fatto morire dal ridere (grintoso benessere!). Volevo ringraziare Valenta che ha sempre garantito serate senza tempo per cui mi scordavo di tornare a casa ed ha sempre offerto un buon caffé. Grazie anche al compare Guarna. Grazie a Paru che mi tiene aggiornato su startup, sharing economy, bitcoin e management stuff che al momento, da come mi esprimo, evidentemente non è pane per i miei denti. Ma come è possibile che per mesi ci siamo incontrati svariate volte in quattro biblioteche diverse? Ringrazio Mario per la sua cucina sicula che è stata fondamentale per il conseguimento della laurea ma anche per il suo indubbio spessore umano. Infine, ringrazio Marisa di cui apprezzo la schiettezza calorosa e la capacità di farti sentire come se fossi a casa tua.

Abstract

This work focuses on inferential methods for functional data. An overview about inferential methods able to select the statistically significant intervals of the domain is provided, especially focusing on permutation solutions. The *Interval Testing Procedure* properties are explored through simulations. The investigation of the Smoothing effects on the inferential analysis is all-important. A simulation study is conducted where the *Interval Testing Procedure* is compared with *Benjamini-Hochberg*, *Bonferroni-Holm* and *Bonferroni* multiple testing procedures. The chosen metrics are the *Family Wise Error Rate*, the *Rejection Rate of the False null hypotheses*, the *Rejection Rate of the True null hypotheses* and the *Power*. The hypothesis testing problem is the two-sided distributional comparison between two independent populations of functions. The differences between populations in terms of mean are localized in an interval located in the center of the domain. The B-spline basis expansion is used throughout the simulations. Both the *Regression* and *Smoothing splines methods* are considered. The parameters of interest related to Smoothing are the order of the basis elements, the number of basis elements and the smoothing parameter. The parameters of interest determining the data set are the number of evaluations, the standard deviation of the additive Normal noise and the number of statistical units. It is of relevant interest to explore the differences in terms of the ability to make true discoveries between the *Interval Testing Procedure* and the *Benjamini-Hochberg* procedure, knowing that the former procedure controls the *Family Wise Error Rate* on intervals and the latter procedure ensures only a weak control of the *Family Wise Error Rate.* Best practices can be deduced from the simulation results such as the optimality of the cubic splines with a sufficiently high number of basis elements for the *Interval Testing Procedure* in the case of discontinuos functional data. In these scenarios, the performances of the *Interval Testing Procedure* and the *Benjamini-Hochberg* procedure are equivalent in terms of the *Rejection Rate of the False null hypotheses.* The *Rejection Rate of the False null hypotheses* is a more precise measure of the ability to make true discoveries than the *Power*. Finally, in general for *Interval Testing Procedure* it is better to choose the number of basis elements relatively high. The code for the simulations has been implemented in R. The *fdatest* R package has been used modifying the source code. The most important update is the implementation in C of the combining matrix construction which is the most computationally expensive task in the *Interval Testing Procedure* algorithm for the Two-population framework. The used implementation of the *Interval Testing Procedure* directly works on an object of the *functional data* class. Hence, the Smoothing is entrusted to the user avoiding subjective choices which had to be taken automatically in the original version of *fdatest*. These features involve a significant gain in terms of execution time and a simplification of the interface.

Keywords: Functional Data Analysis, Inference, Interval Testing Procedure, Permutation Tests, Domain Selection, B-splines, *fdatest* R package

Sommario

Questo lavoro si focalizza sui metodi inferenziali per dati funzionali. Viene data una visione d'insieme riguardo ai metodi inferenziali in grado di selezionare gli intervalli del dominio statisticamente significativi, in particolare concentrandosi su soluzioni permutazionali. Le proprietà dell'*Interval Testing Procedure* vengono esplorate per via simulativa. Lo studio degli effetti dello Smoothing sull'analisi inferenziale è di primaria importanza. Uno studio di simulazione viene effettuato dove l'*Interval Testing Procedure* viene confrontata con le correzioni di molteplicità *Benjamini-Hochberg*, *Bonferroni-Holm* e *Bonferroni*. Le metriche scelte sono il *Family Wise Error Rate*, il *Tasso di Rifiuto delle ipotesi nulle False*, il *Tasso di Rifiuto delle ipotesi nulle Vere* e la *Potenza*. Il test di ipotesi è il confronto distribuzionale bilatero tra due popolazioni indipendenti di funzioni. Le differenze tra le popolazioni in termini di media sono localizzate in un intervallo situato nel centro del dominio. L'espansione in base B-spline è usata in tutte le simulazioni. Vengono considerati entrambi i metodi *Regression splines* e *Smoothing splines*. I parametri di interesse legati allo Smoothing sono l'ordine degli elementi della base, il numero degli elementi della base e lo smoothing parameter. I parametri di interesse che determinano il data set sono il numero di valutazioni, la deviazione standard del rumore Normale additivo ed il numero di unità statistiche. È di rilevante interesse esplorare le differenze in termini dell'abilità di effettuare vere scoperte tra l'*Interval Testing Procedure* e la procedura *Benjamini-Hochberg*, sapendo che la prima procedura controlla il *Family Wise Error Rate* per intervalli e la seconda procedura garantisce solo un controllo debole del *Family Wise Error Rate*. Dai risultati delle simulazioni si possono dedurre best practices come l'ottimalità delle spline cubiche con un numero sufficientemente elevato di elementi della base per l'*Interval Testing Procedure* nel caso di dati funzionali discontinui. In questi scenari, le prestazioni dell'*Interval Testing Procedure* e della procedura *Benjamini-Hochberg* sono equivalenti in termini del *Tasso di Rifiuto delle ipotesi nulle False*. Il *Tasso di Rifiuto delle ipotesi nulle False* costituisce una misura più precisa dell'abilità di effettuare vere scoperte rispetto alla *Potenza*. Infine, in generale per l'*Interval Testing Procedure* è meglio scegliere il numero degli elementi della base sufficientemente elevato. Il codice per le simulazioni è stato implementato in R. Il pacchetto R *fdatest* è stato usato modificandone il codice sorgente.

L'aggiornamento più importante è l'implementazione in C della costruzione della combining matrix che è l'operazione più costosa nell'algoritmo per l'*Interval Testing Procedure* nel caso del confronto distribuzionale tra due popolazioni di funzioni. L'implementazione usata dell'*Interval Testing Procedure* opera direttamente su un oggetto della classe *functional data*. Pertanto, lo Smoothing è affidato all'utente evitando scelte soggettive che dovevano essere prese in automatico nella versione originale di *fdatest*. Queste caratteristiche comportano un significativo guadagno in termini di tempo d'esecuzione ed una semplificazione dell'interfaccia.

Parole chiave: Analisi Dati Funzionali, Inferenza, Interval Testing Procedure, Test di Permutazione, Domain Selection, B-splines, pacchetto R *fdatest*

Contents

List of Figures

Introduction

The statistical analysis of functions is a noteworthy research area commonly called Functional Data Analysis (Ramsay and Silverman, 2002; Silverman and Ramsay, 2005; Cuevas, 2013). The observed functions are conceived as elements of a functional space which is typically Hilbert separable. A well-known example is given by the space of the square integrable functions L^2. Other types of spaces have been also considered such as the metric spaces (Ferraty and Vieu, 2006; Ferraty and Romain, 2011). Several results have been obtained as far as concerns the data description (e.g., Smoothing methods) and the data exploration (e.g., Functional Principal Component Analysis). Many techniques are available and well-established (De Boor, 2001; Silverman and Ramsay, 2005; Schumaker, 2007).

On the contrary, Inference is a recent and challenging topic in Functional Data Analysis. Inferential methods have been proposed specifically for the Functional Analysis of Variance problem in Zhang (2013) and under parametric and/or asymptotic assumptions in Horváth and Kokoszka (2012). In the case of functional data, the classical parametric assumptions (e.g., homoscedasticity, normality) can be unrealistic or assumed for mere convenience. Moreover, the normality assumption is not verifiable in the Functional Data Analysis framework. In fact, it implies that the projections of the functional data over every element of the functional space is normally distributed (Tarpey and Kinateder, 2003). Alternatively, we can opt for nonparametric methods. The nonparametric inference is generally based on permutation tests (Corain et al., 2014) or on bootstrap techniques (Hall and Tajvidi, 2002; Hall and Keilegom, 2007).

The overwhelming majority of nonparametric and parametric available methods for hypothesis testing in Functional Data Analysis are global and, hence, if the null hypothesis is rejected, they are not able to impute such a rejection on a particular part of the domain. The selection of the intervals where the null hypothesis is false is a desired property (*domain selection)*, especially in applications. The methodologies proposed in Pini and Vantini (2013) (Interval Testing Procedure) and in Pini and Vantini (2015a) (Interval-wise test) guarantee the selection of the significant domain intervals. The same holds for the test proposed in Vsevolozhskaya et al. (2014) for a set of a priori selected sub-intervals.

From a methodological point of view, we focus on the study of inferential methods provided with the *domain selection* property in the Functional Data Analysis framework. Additionally, we want to explore the Interval Testing Procedure properties through a simulation study. Therefore, in this work we use the permutation tests since in the Interval Testing procedure the Nonparametric Combination methodology is applied; the Nonparametric Combination is a fundamental theoretical tool for multiple testing in the permutation approach (Pesarin and Salmaso, 2010).

The Interval Testing Procedure, as a first step, requires the discretization of the data through an appropriate basis expansion. The univariate permutation tests are expressed in terms of the basis coefficients. Then, from the univariate p-values, the multivariate tests, expressed in terms of the coefficients and associated with the family of all possible consecutive hypotheses, are performed using the Nonparametric Combination methodology. Finally, with a maximization operation on the p-values obtained from the previous step, the univariate p-values are adjusted. The Interval Testing Procedure is characterized by an interval-wise control of the *Family Wise Error Rate* (i.e., the control of the *Family Wise Error Rate* is guaranteed for every set of consecutive true null hypotheses). Such a control is intermediate between the strong control (i.e., the control of the *Family Wise Error Rate* is guaranteed for all possible sets of true null hypotheses) and the weak control (i.e., the control of the *Family Wise Error Rate* is guaranteed only when the set of the true null hypotheses is composed by all null hypotheses).

A permutation test is essentially based on a family of transformations which preserve the likelihood under the null hypothesis (admissible permutations) and on a suitable test statistic which is stochastically larger under the alternative hypothesis than under the null hypothesis. The p-value of the test is given by the proportion of permuted scenarios in which the test statistic evaluated on the permutations is greater than or equal to the value of the test statistic applied on the observed data. Since the set of all possible permutations is often too large to be explored in its entirety, only a subset of the permutations is explored through a Conditional Monte Carlo algorithm. The permutation tests are conditional procedures of inference. The conditioning is on a set of sufficient statistics under the null hypothesis. This kind of conditioning and the assumed existence of likelihood-invariant transformations under the null hypothesis (exchangeability condition under the null hypothesis) imply the independence of the permutation tests from the likelihood model associated with the population distribution. A fundamental property implied by these conditions is the exactness of the permutation tests (Pesarin and Salmaso, 2010).

The purpose of the simulation study is the evaluation of the Interval Testing Procedure performances with particular interest in Smoothing effects on the inferential analysis. The Interval Testing Procedure is compared with the multiple testing procedures Benjamini-Hochberg (Benjamini and Hochberg, 1995), Bonferroni-Holm

(Holm, 1979) and Bonferroni based on the coefficients of the same basis expansion. The selected metrics are the *Family Wise Error Rate* (i.e., the probability of rejecting at least one null hypothesis among the hypotheses to be accepted), the *Rejection Rate False null hypotheses* (i.e., the expected rate of rejected null hypotheses among the hypotheses to be rejected), the *Rejection Rate True null hypotheses* (i.e., the expected rate of rejected null hypotheses among the hypotheses to be accepted) and the *Power* (i.e., the probability of rejecting at least one null hypothesis among the hypotheses to be rejected).

It is of relevant interest to explore the differences in terms of the ability to make true discoveries between the Interval Testing Procedure and the Benjamini-Hochberg procedure, knowing that the former procedure controls the *Family Wise Error Rate* on intervals and the latter procedure ensures only a weak control of the *Family Wise Error Rate*.

The hypothesis testing problem chosen for the simulations is the two-sided distributional comparison between two independent populations of functions (unpaired case). The tested types of synthetic data set are populations of constant and step functions (data set *const-step)* and populations of constant and tricube functions (data set *const-tricube*). For each data type, the differences of the two populations in terms of mean (fixed deterministic effect) are localized in an interval located in the center of the domain. We have chosen these types of data set for the totally different degree of regularity that distinguishes them; the data set *const-step* is not regular since it exhibits discontinuities and, instead, the data set const-tricube is rather regular since the tricube kernel belongs to $\mathcal{C}^9$. In order to avoid fictitious differences in terms of true discovery rate, we have imposed that the L^2 distance between the means of the first population composed by constant functions and of the second population is the same for the data sets *const-step* and *const-tricube*.

The B-spline basis expansion is used throughout the simulations. Both the *Regression* and *Smoothing splines methods* are considered. The parameters of interest related to Smoothing are the order m of the B-spline basis elements, the number of basis elements p and the smoothing parameter λ (*Smoothing parameters*). The parameters determining the data set are the number of evaluations n_{eval}, the standard deviation σ_{noise} of the additive Normal noise and the number of statistical units n_{units} (*Data parameters*).

In brief, the purpose of the simulation study is the evaluation of the Interval Testing Procedure performances compared with the multiple testing procedures Benjamini-Hochberg, Bonferroni-Holm and Bonferroni ones by varying:

- The data set type (*const-step* or *const-tricube*)
- The parameters m and p, λ (*Smoothing parameters*), n_{eval} and σ_{noise}, n_{units} (*Data parameters*)
- The Smoothing method: *Regression* or *Smoothing splines* using the B-spline basis expansion

The performances of the methods are evaluated by means of the metrics *Family Wise Error Rate*, *Rejection Rate False null hypotheses*, *Rejection Rate True null hypotheses* and *Power*.

The code for the simulations has been implemented in R 3.2.0 using *fdatest* 2.1 and *fda* 2.4.4 packages (Team, 2015; Pini and Vantini, 2015b; Ramsay et al., 2014). Smoothing has been performed using *fda* package. *fdatest* package has been used for applying the Interval Testing Procedure in the Two-population framework using the B-spline basis expansion.

The p-values obtained with the Interval Testing Procedure are stored in $p \times p$ matrix called combining matrix which is the input of the p-values adjustement phase. p is the number of the univariate p-values and it coincides with the number of basis elements. The construction of the combining matrix is the most computationally expensive task in the Interval Testing Procedure algorithm for the Two-population framework. It costs p^2.

Hence, the *fdatest* source code has been conveniently modified. In order to improve the execution times, the construction of the combining matrix has been implemented in C. It has been used the .C interface to R. The new implementation has provided a speedup over the original implementantion equal to around 60x. This result is useful both for simulations, where the Interval Testing Procedure must be applied several times for each value of the parameter under analysis, and for scenarios where it is reasonable to choose a high number of basis elements p.

Finally, the used implementation of the *Interval Testing Procedure* directly works on an object of the *functional data class*. Therefore, the Smoothing is performed by the user avoiding subjective choices which had to be taken automatically in the original version of *fdatest*.

Chapter 1 provides an overview about Inference in the Functional Data Analysis Framework focusing on methods characterized by the *domain selection* property. The Interval-wise test (Pini and Vantini, 2015) and the Vsevolozhskaya-Greenwood-Holodov test (Vsevolozhskaya et al., 2014) are explained.

In Chapter 2 the adopted methodologies in the simulation study are illustrated. The multiple testing procedures Bonferroni, Bonferroni-Holm and Benjamini-Hochberg are briefly described. Then, the working principle of the Interval Testing Procedure is explained for the Two-Population, Multi-Population, One-Population frameworks. Finally, a brief discussion about the *Smoothing splines method* is reported.

Chapter 3 describes the simulation setting. More details are provided about the tested data sets and the evaluation metrics.

In Chapter 4 the simulation results are reported and discussed. The presentation of the simulation results is divided in two parts determined by the *Smoothing parameters* (m and p for *Regression splines*, λ for *Smoothing splines*) and by the *Data parameters* (n_{eval} and σ_{noise}, n_{units}). m and p are the order and the number of the B-spline basis elements, respectively. λ is the smoothing parameter. n_{eval} and σ_{noise} are the number of evaluations and the standard deviation of the additive Normal noise, respectively. n_{units} is the number of statistical units.

Finally, a discussion about the main simulation study results is reported and possible future developments are suggested.

Chapter 1

State of the Art

1.1 Introduction

In recent decades great progress has been made in the Functional Data Analysis (FDA) framework, i.e., the statistical analysis of functions (Silverman and Ramsay, 2005; Ferraty and Vieu, 2006; Ferraty and Romain, 2011; Cuevas, 2013). Several methods have been developed for Smoothing (i.e., the conversion from observed data to functional data) and dimensionality reduction (e.g., functional principal components analysis) or more in general for exploratory analysis (Ramsay and Silverman, 2002; Silverman and Ramsay, 2005). In FDA, the statistical units belong to a functional space which is typically a Hilbert separable space. For instance, the well-known L^2 space is Hilbert separable and it constitutes the natural extension of Euclidean Geometry. So far, the methods mentioned are essentially used for the description of the data.

Nowadays one challenging area is the inference in FDA. Inference in the FDA framework poses several methodological problems. For instance, it can be shown that a probability density function does not exist for a generic functional space (Delaigle and Hall, 2010). On the contrary, it is possible to define a cumulative distribution function for random functions but this notion is scarcely useful, especially from a practical point of view.

Moreover, it is often common that the inferential tools of the Multivariate Analysis are not straightforwardly generalizable in the FDA framework since the number of features is by far bigger than the number of statistical units. A representative example is the Hotelling's theorem. This theorem holds only if the number of features is lower than the number of units which is obviously not the case in FDA (Hotelling, 1931). The Hotelling's theorem is the most used tool for testing the mean of normal populations in the multivariate framework.

One way to deal with these issues is the use of methods based on parametric assumptions and/or asymptotic results (Horváth and Kokoszka, 2012). This approach, however, implies assumptions (e.g., homoscedasticity, normality and random sampling) that could be difficult to justify or unrealistic. Furthermore, the normality assumption implies that the projections of the functional data over every element of the functional space must be normally distributed (Tarpey and Kinateder, 2003), condition practically impossible to verify even if a basis expansion is used. For these reasons, it could be more convenient to take into account nonparametric methods such as permutation tests (Corain et al., 2014; Hall and Keilegom, 2007).

A permutation test is essentially based on a family of transformations which preserve the likelihood under the null hypothesis H_0 (i.e., it is satisfied the so called exchangeability condition under the null hypothesis) and on a suitable test statistic T which is stochastically larger under the alternative hypothesis H_1 than under H_0. The family of transformations depends on the assumptions of independence in the model.

Under the null hypothesis H_0, the information content provided by an admissible (i.e, likelihood-invariant) permutation coincides with the information content provided by the observed data. Hence, if the null hypothesis H_0 is true, the difference between the test statistic applied on the observed data T_0 and the test statistic applied on a permutation of the observed data T^* is small. In theory under the null hypothesis H_0, we can compute the discrete distribution of the test statistic applied on the permuted data T^* exploring in toto the set of all possible permutations of the observed data. The p-value of the test is given by the proportion of permuted scenarios in which the test statistic evaluated on the permutations T^* is greater than or equal to the value of the test statistic applied on the observed data T_0 . In practice, the set of all possible permutations is often too large to be explored in its entirety. Hence, only a subset of the permutations is explored through a Conditional Monte Carlo (CMC) algorithm whose function is to simulate, under the null hypothesis H_0, the distribution of the test statistic applied on the permuted data T^*. The CMC algorithm is reported in 2.2.1.4.

In general, the permutation tests are conditional procedures of inference. The conditioning is on a set of sufficient statistics under the null hypothesis. In most practical situations, the conditioning is on the observed data set which is often the unique sufficient and minimal function of the sample in a nonparametric framework. This kind of conditioning and the assumed exchangeability with respect to groups under the null hypothesis (i.e., in general there exist likelihood-invariant transformations under the null hypothesis) imply the independence of the permutation tests from the likelihood model associated with the population distribution. A fundamental property implied by these conditions is the exactness of the permutation tests. If the exchangeability condition under the null hypothesis is violated, then the permutation tests are not exact. If it is difficult to obtain approximate permutation solutions, an alternative solution is the use of nonparametric bootstrap techniques which require less assumptions but at the same time they are provided with less theoretical properties such as, for instance, the exactness (Pesarin and Salmaso, 2010).

From a review on the literature, it has been found that the overwhelming majority of nonparametric and parametric methods for hypothesis testing in FDA are global in the sense that they test a null hypothesis globally on the whole domain of the functions. If the null hypothesis is rejected, they are not able to impute such a rejection on a particular part of the domain. An example of such kind of tests is illustrated in Corain et al. (2014), where the global distributional comparison between two populations of functions is performed basing inference on the Nonparametric Combination (NPC). NPC is a fundamental theoretical tool for multiple testing in the permutation framework and it constitutes an important part of the Interval Testing Procedure (Pini and Vantini, 2013). Moreover, with this methodology it is possible to solve hypothesis testing problems where the number of statistical units are smaller than the number of features and the univariate tests are possibly dependent. These features are fundamental in the FDA framework.

The NPC methodology is applied in the following procedures:

- The permutation and combination based time-to-time method (time-to-time analysis).
- The permutation and combination by derived variable approach (the derived variables are the coefficients of a basis expansion).

In the time-to-time method, the global hypothesis is decomposed in a number of sub-hypotheses equal to the number of measurements. It is assumed that the statistical units share a common domain where the evaluations are uniformly distributed. In the derived variable approach the rationale is the same. Thanks to the NPC, the global hypothesis is decomposed in a number of sub-hypotheses equal to the dimension of the basis expansion used. With the NPC, a global test that combines the information given by the univariate tests associated with the sub-hypotheses is obtained.

Another example of global test model-based is given by the functional Hotelling's theorem (Pini et al., 2015). The finite-dimensional approximation of the functional Hotelling's T^2 coincides with the high-dimensional p-asymptotic counterpart proposed in Secchi et al. (2013) where p is the number of features; this approximation is computable in closed form; it does not require the resolution of an optimization problem.

The use of a global test could be scarcely useful from a practitioner point of view. For instance, suppose to have two sets of functions. The aim is to test if the populations differ in distribution. Hence, the null hypothesis is the equality between the population distributions. If a global test is applied and the null hypothesis is rejected, it is not possible to localize the intervals of the domain which are sources of distributional difference. The selection of the intervals where the null hypothesis is false is a desired property for practitioners. If with a procedure it is possible to select the significant intervals of the domain, the procedure used is said to be endowed with the *domain selection* property which can be seen as the functional counterpart of the *feature selection*.

The methodologies proposed in Pini and Vantini (2013) (Interval Testing Procedure) and in Pini and Vantini (2015a) (Interval-wise test) guarantee the selection of the significant domain intervals. The same applies for the test proposed in Vsevolozhskaya et al. (2014) for a set of a priori selected sub-intervals (discretization of the domain).

The ITP, as a first step, requires the discretization of the data through a suitable basis expansion. The univariate permutation tests are expressed in terms of the basis coefficients. From the univariate p-values, the multivariate tests, expressed in terms of the coefficients and pertaining the family of all possible consecutive hypotheses, are performed using the NPC methodology. Finally, with a maximization operation on the p-values obtained from the previous step, the univariate p-values are adjusted. The ITP is characterized by a control on intervals of the Family Wise Error Rate $FWER$ (i.e., the probability of rejecting at least one null hypothesis belonging to the set of the true null hypotheses). Controlling the $FWER$ on intervals means that the control of the $FWER$ is guaranteed for every set of consecutive true null hypotheses. Such a control is intermediate between the strong control (i.e., the control of the $FWER$ is guaranteed for all possible sets of true null hypotheses) and the weak control (i.e., the control of the $FWER$ is guaranteed only when the set of the true null hypotheses is composed by all null hypotheses). The ITP is explained in Chapter 2 and its properties are explored through simulations. The simulation setting and the simulations results are reported in Chapter 3 and Chapter 4, respectively.

In the next sections the Interval-wise test and the Vsevolozhskaya-Greenwood-Holodov test are described in more detail.

1.2 Interval-wise Testing for Functional Data

1.2.1 Introduction

The Interval-wise test is a nonparametric procedure able to select the significant intervals of the domain, i.e., the intervals where the null hypothesis is rejected (Pini and Vantini, 2015a). Unlike the Interval Testing Procedure (ITP) (Pini and Vantini, 2013) and the Vsevolozhskaya-Greenwood-Holodov test (VGH test) (Vsevolozhskaya et al., 2014), this procedure does not require the discretization of the data by means of a basis expansion as required by the ITP and the discretization of the domain by specifying a priori a partition of the domain as required by the VGH test. Therefore, the Interval-wise test is a totally data-driven inferential procedure.

We assume that the functional data belong to the Hilbert space L^2. The hallmark of the Interval-wise test is the introduction of the unadjusted $p(t)$ and adjusted $\tilde{p}(t)$ p-value functions, where t is a generic time instant (it can be a position or a frequency or more in general a continuos independent variable). A point-wise p-value function is not trivially defined in L^2 since in this functional space each function is an equivalence class defined on the equivalence relation infinitely often with respect to the Lebesgue measure. By thresholding the unadjusted p-value function $p(t)$ with a fixed significance level α, we select the significant intervals of the domain controlling the point-wise error rate, i.e., given any point of the domain where the null hypothesis is not violated, the probability of wrongly selecting it as significant is controlled. By thresholding the adjusted p-value function $\tilde{p}(t)$ with a fixed significance level α, we select the significant intervals of the domain controlling the interval-wise error rate, i.e., given any interval of the domain where the null hypothesis is not violated, the probability of wrongly selecting it as significant is controlled. Moreover, the p-value functions $p(t)$ and $\tilde{p}(t)$ are consistent. In detail, the unadjusted p-value function $p(t)$ is point-wise consistent, i.e., given any point of the domain where the null hypothesis is violated, the probability of selecting it as significant goes to one as the sample size goes to infinity and the adjusted p-value function $\tilde{p}(t)$ is interval-wise consistent, i.e., given any interval of the domain where the null hypothesis is almost everywhere violated the probability of selecting it as significant goes to one as the sample size goes to infinity.

In the following, in order to clarify how the Interval-wise test works, we detail its basic steps in the Two-Population framework. The Interval-wise test procedure can be easily extended to more general frameworks such as the Multi-Population framework.

1.2.2 Interval-wise test in the Two-Population framework

We want to test the difference in terms of mean between two functional populations. We observe the functional data $y_{ij} \in L^2(\mathcal{D})$ where $j = 1, 2$ (population index), $i = 1, \ldots, n_j$ (unit index) and $\mathcal{D} = (a,\ b) \subset \mathbb{R}$. The hypotheses of the global test are:

$$H_0 : \mu_1 = \mu_2 \ against \ H_1 : \mu_1 \neq \mu_2 \tag{1.1}$$

where the equality $\mu_1 = \mu_2$ is intended in the L^2 sense, i.e.,

$$\mu_1 = \mu_2 \iff \int_{\mathcal{D}} (\mu_1(t) - \mu_2(t))^2 \, dt = 0$$

Interval-wise testing

Let $I \subseteq \mathcal{D} = (a,\ b) \subset \mathbb{R}$ be any interval of the domain. The hypotheses of the partial test on I are:

$$H_0^I : \mu_1^I = \mu_2^I \ against \ H_1^I : \mu_1^I \neq \mu_2^I \tag{1.2}$$

We denote with p^I the p-value of the functional test (1.2). The restriction on I of the chosen test statistic T is used to evaluate p^I. The choice of the test statistic for (1.2) is not unique and it is not necessary to opt for a permutation solution. In the permutation framework, a natural choice is the test statistic proposed in Hall and Tajvidi (2002) given by

$$T^I = \frac{\int_I \left(\overline{y}_1(t) - \overline{y}_2(t)\right)^2 dt}{|I|} = \frac{||\overline{y}_1 - \overline{y}_2||^2_{L^2(I)}}{|I|}$$

where $\overline{y}_j(t) = \frac{\sum_{i=1}^{n_j} y_{ji}(t)}{n_j}$ for $j = 1, 2$.

An exact permutation test for the test on I can be obtained by evaluating the test statistic T^I over all possible permutations of the observed data over the sample units. It is implictly assumed that the two populations are independent (unpaired case). If this assumption is not satisfied, then the permutations of the observed data over the sample units are not likelihood-invariant transformations under the null hypothesis H_0^I and, hence, the exchangeability condition under the null hypothesis H_0^I, which is necessary for the exactness of the partial test, is violated. If the two populations are dependent (paired case), we have to restrict the set of the permutations over the sample units. Under H_0^I, the exchangeability is just between and within couples. In both cases, the p-value p^I of the test (1.2) is the proportion of the test statistics evaluated on the permuted data T^{I*} greater than or equal to the test statistic evaluated on the observed data T_0^I.

The multiplicity correction would involve a family of infinite tests whose cardinality is the cardinality of the continuum. Despite this, in practice the number of tests is limited by the number of functional evaluations n_{eval} at disposal. For sake of simplicity, we assume that the statistical units are evaluated on the same grid. In detail, denote with $\{t_k\}_{k=1}^{n_{eval}} \in \mathcal{D}$ the common set of points where the observed data $\{y_{ji}\}_{i=1}^{n_j} \in L^2(\mathcal{D})$ with $j = 1, 2$ (population index) are sampled. We suppose that, the generic unit y_{ij} assume the same value $y_{ij}(t_k)$ in the interval $[t_k,\ t_{k+1})$. In the particular case $k = n_{eval}$, we have the degenerate interval coincident with the point t_{neval}. We perform a permutation test for every interval of the form $[t_k,\ t_{k+1})$ and also for all possible their unions constituted by consecutive intervals (interval-wise testing) using the sum of the related univariate tests statistics (direct combination approach; Pesarin and Salmaso, 2010). The number of possible unions of consecutive intervals are limited by the number of evaluations n_{eval}. However, the number of tests remains high excluding the scenarios with very low number of evaluations n_{eval}. Therefore, it does not generally make sense to use classical multiple testing procedures such as Bonferroni-Holm (Holm, 1979) or Benjamini-Hochberg (Benjamini and Hochberg, 1995) since the adjusted p-values obtained with these procedures tend to be unitary due to the high number of tests.

Definition of the p-value functions

From the results of the previous step, we define the unadjusted p-value function $p(t)$ and the adjusted p-value function $\tilde{p}(t)$ and we report in detail their theoretical properties.

The unadjusted p-value function $p(t)$ is given by

$$p(t) = \underset{I \to t}{limsup}\ p^I$$

where the notation $I \to t$ means that the extremes of the intervals I converge to t. The adjusted p-value function $\tilde{p}(t)$ is given by

$$\tilde{p}(t) = \underset{t \ni I}{sup}\ p^I$$

The unadjusted p-value function $p(t)$ is provided with a control of the point-wise error rate, i.e, fixed the significance level α, for every $t \in \mathcal{D}$ such that it exists an interval $I \subseteq \mathcal{D}$ that includes t and with associated null hypothesis H_0^I true, we have

$$\mathbb{P}(\{p(t) \leq \alpha\}) \leq \alpha$$

The adjusted p-value function $\tilde{p}(t)$ is provided with a control of the interval-wise error rate, i.e, fixed the significance level α, we have

$$\forall I \subseteq \mathcal{D} : H_0^I\ is\ true \Rightarrow \mathbb{P}(\{\forall t \in I,\ \tilde{p}(t) \leq \alpha\}) \leq \alpha$$

It can be proved (Pini and Vantini, 2015a) that, for almost every $t \in \mathcal{D}$ (i.e., infinitely often with respect to the Lebesgue measure), the unadjusted p-value function $p(t)$ coincides with the p-value of the univariate permutation test given by $\underset{I \to t}{lim}\ p^I$ where $I \to t$ means that the extremes of the interval I converge to t. Consequently, since each univariate permutation test is assumed to be exact, fixed the significance level α, for almost every $t \in \mathcal{D}$ such that it exists an interval $I \subseteq \mathcal{D}$ that includes t and with associated null hypothesis H_0^I true, we have

$$\mathbb{P}(\{p(t) \leq \alpha\}) = \alpha$$

Furthermore, both the unadjusted p-value $p(t)$ and the adjusted p-value $\tilde{p}(t)$ functions are consistent. In detail, the unadjusted p-value function $p(t)$ is point-wise consistent, i.e., fixed the significance level α, we have

$$\forall t \in \mathcal{D} \textit{ subject to } \nexists I \subseteq \mathcal{D} : t \in I \textit{ and } H_0^I \textit{ is true } \Rightarrow \mathbb{P}\left(\{p(t) \leq \alpha\}\right) \overset{n \to \infty}{\rightarrow} 1$$

where $n = n_1 + n_2$.

The adjusted p-value function $\tilde{p}(t)$ is interval-wise consistent, i.e., fixed the significance level α, we have

$$\forall I \subseteq \mathcal{D} \textit{ subject to } \nexists J \subseteq I : H_0^J \textit{ is true } \Rightarrow \mathbb{P}\left(\{\forall t \in I,\ \tilde{p}(t) \leq \alpha\}\right) \overset{n \to \infty}{\rightarrow} 1$$

It is worth noticing that such theoretical properties hold in general, and not only in the Two-population framework. In general, the assumptions required for the validity of the theoretical properties of the p-value functions are:

- The test statistic T is *real-valued.*

- The test of H_0^I *against* H_1^I is *exact*, i.e.,

  $$\forall I \subseteq \mathcal{D} : H_0^I \textit{ is true } \Longrightarrow \mathbb{P}\left(\left\{p^I \leq \alpha\right\}\right) = \alpha$$

- The test of H_0^I *against* H_1^I is *consistent*, i.e.,

  $$\forall I \subseteq \mathcal{D} : H_0^I \textit{ is false } \Longrightarrow \mathbb{P}\left(\left\{p^I \leq \alpha\right\}\right) \overset{n \to \infty}{\rightarrow} 1$$

  where $n = n_1 + n_2$ is the sample size.

Domain Selection

The intervals of the domain presenting a significant mean difference between the two populations are selected by thresholding the p-value functions computed in the previous step. In detail, the significant intervals of the domain obtained by controlling the point-wise (interval-wise) error rate are selected by thresholding the p-value function $p(t)$ ($\tilde{p}(t)$) with the chosen significance level α.

It is worth mentioning that, in the same way as the Interval Testing Procedure, the adjustement is such that the control of the interval-wise error rate proper of $\tilde{p}(t)$ is guaranteed for every interval and for the complementary set of the interval itself (recycled version of the p-values adjustement implementation; for details refer to Pini and Vantini, 2013).

1.3 Vsevolozhskaya-Greenwood-Holodov test

1.3.1 Introduction

The Vsevolozhskaya-Greenwood-Holodov test (VGH test) is an inferential procedure for pairwise comparison of population means in the Functional Analysis of Variance framework (Vsevolozhskaya et al., 2014)

The VGH test is divided in two parts: firstly, the significant intervals are selected from a a priori specified partition of the domain; then, pairwise comparisons between population means are performed for the significant intervals previously identified. In order to control the Family Wise Error Rate $FWER$, the closure multiple testing procedure is applied (Marcus et al., 1976).

1.3.2 Methodological aspects

Functional Analysis of Variance model

The Functional Analysis of Variance (FANOVA) model is described by the equation:

$$y_{ji}\left(t\right) = \mu_j\left(t\right) + \epsilon_{ji}\left(t\right) \tag{1.3}$$

where $\mu_j\left(t\right)$ is the functional mean of group j at time t with $j = 1, \ldots,\ k$ (population index), $i = 1, \ldots,\ n$ (unit index), $t \in \mathcal{D} = [a,\ b]$ and $\epsilon_{ji}\left(t\right)$ is the additive noise function. Each $\epsilon_{ji}\left(t\right)$ has null mean and it is an independent Normal stochastic process. The hypotheses of the FANOVA model (1.3) are:

$$H_0 : \mu_1\left(t\right) = \ldots = \mu_k\left(t\right) \ \ against\ H_1 : \exists i, j \in \{1, \ldots, k\} \ \ and\ \exists t \in \mathcal{D} : \mu_i\left(t\right) \neq \mu_j\left(t\right) \tag{1.4}$$

Selection significant intervals

Firstly, the common domain $\mathcal{D} = [a,\ b]$ is a priori divided in m mutually exclusive sub-intervals of the form $[a_i,\ b_i]$, such that $[a,\ b] = \cup_{i=1}^{m}[a_i,\ b_i]$. Then, the null hypothesis H_0 is tested on every sub-interval. The test statistic used T_i is the numerator of the test statistic of the global (i.e, the differences are detected for the entire domain $\mathcal{D}$ and not for specific intervals) test proposed in Shen (2004), which is for the i-th interval:

$$\mathcal{F}_i = \frac{\int_{a_i}^{b_i} \sum_{j=1}^{k} n_j \left(\hat{\mu}_j\left(t\right) - \hat{\mu}\left(t\right)\right)^2 dt}{\int_{a_i}^{b_i} \sum_{j=1}^{k} \sum_{s=1}^{n} \left(y_{js}\left(t\right) - \hat{\mu}_j\left(t\right)\right)^2 dt} \cdot \frac{(n-k)}{(k-1)},\ i = 1, \ldots,\ m \tag{1.5}$$

where k is the number of groups and, for sake of simplicity, the experiment is assumed to be balanced, i.e., $n_1 = \ldots = n_k = n$. In the same way as the Snedecor F test statistic in the classical analysis of variance framework, the numerator and the denominator of the test statistic (1.5) are measures of the variability between groups and within groups, respectively. The test statistic (1.5) is the functional counterpart of the Snedecor F test statistic.

The combining function chosen is the sum of the univariate test statistics:

$$T = \sum_{i=1}^{m} T_i$$

A p-value for the global null hypothesis H_0 (all marginal null hypotheses are true; the i-th marginal null hypothesis $H_0^{[a_i,\ b_i]}$ is the equality between the means of the populations in the interval $[a_i,\ b_i]$) can be based on parametric assumptions (i.e., it is known the distribution of T under the global null hypothesis H_0) or on a permutation approach.

Afterwards, the closure multiple testing procedure is applied (Marcus et al., 1976), i.e., the null hypothesis is tested on every possible union of sub-intervals. The null hypothesis $H_0^{[a_i,\ b_i]}$ is rejected if all tested hypotheses implying $H_0^{[a_i,\ b_i]}$ are rejected. This procedure guarantees a strong control of the Family Wise Error Rate $FWER$ on the intervals a priori selected and on their unions.

The closure multiple testing procedure is computationally intensive if the number of sub-intervals m is high, since it implies the execution of $2^m - 1$ tests. If the computational cost is too high, a possible solution consists in the use of shortcut versions of the closure procedure. In Vsevolozhskaya et al. (2014) several shortcut versions of the closure procedure are proposed: adjustement based on the ordered test statistics shortcut, adjustement based on the ordered unadjusted p-values shortcut and adjustement obtained combining the first two shortcuts. For details refer to Vsevolozhskaya et al. (2014).

Via a simulation study, a comparison has been made between these shortcuts and the closure procedure without modifications. It has been performed a permutation test based on the test statistics $\{T_i\}_{i=1}^{m}$, where $m = 5$ intervals have been chosen. In each interval, 1000 permutations have been sampled. The experiment has been repeated 1000 times. It has been observed that the third shortcut (combination of the adjustement based on the ordered test statistics shorcut and of the adjustement based on the ordered unadjusted p-values shortcut) is the best one in terms of the ratio between the adjusted p-values underestimated (i.e., not adjusted enough) and the total number of p-values (all null hypotheses were true). Moreover, the type I error rates under the shortcut with best performances and the closure procedure are coincident.

Another issue is the fact that the VGH procedure and all associated shortcuts become very conservative if m is high.

Pairwise comparisons in the significant intervals

At this stage, the aim is to identify the pairs of functional means that are different in the significant intervals previously selected. The steps are:

- For the generic significant interval $[a_i,\ b_i]$, where there is statistical evidence to affirm that the means of the populations differ, the (unadjusted) p-value associated with the null hypothesis of no difference among the generic pair of population means is initialized to the adjusted p-value associated with the test on the interval $[a_i,\ b_i]$.

- For every pair of populations, compute the pairwise statistics

$$U_i^{(h,\ g)} = \int_{a_i}^{b_i} \sum_{j \in \{h,\ g\}} n_j \left(\hat{\mu}_j\left(t\right) - \hat{\mu}\left(t\right)\right)^2 dt$$

  where h and g are the identifiers of the populations to be compared ($h, g \in \{1, \ldots,\ k\}$), $\hat{\mu}_j\left(t\right) = \frac{\sum_{i=1}^{n_j} y_{ji}(t)}{n_j}$ and $\hat{\mu}\left(t\right) = \frac{\sum_{j \in \{h,\ g\}} \hat{\mu}_j(t)}{2}$.

- For every pair of populations, perform the standard permutation test, i.e., the p-value associated with test comparison between the populations with identifier h and g, is given by

$$p_i^{(h,\ g)} = \frac{\mathbb{I}\left(\left\{U_i^{(h,\ g)*} \geq U_i^{(h,\ g)}\right\}\right)}{|\chi_{|Y}|}$$

  where $U_i^{(h,\ g)*}$ is the statistic applied on the generic likelihood-invariant under the null hypothesis permutation and $|\chi_{|Y}|$ is the cardinality of the set of all likelihood-invariant under the null hypothesis permutations. Finally, the family of p-values obtained for every pair-wise comparison are adjusted by means of the closure multiple testing procedure applied to every possible set of groups.

Chapter 2

Methodology

2.1 Introduction

Inference is one of the most challenging research topic in Functional Data Analysis (FDA). In particular, we are interested in inferential methods provided with the *domain selection* property, i.e., the ability to select the statistically significant intervals of the domain. The *domain selection* property can be seen as the functional counterpart of the *feature selection* property. At the same time, we require a control on false positives. More precisely, both the Interval Testing Procedure (Pini and Vantini, 2013) and the Interval-wise test (Pini and Vantini, 2015a) control the Family Wise Error Rate $FWER$ (i.e., the probability of rejecting at least one null hypothesis belonging to the set of the true null hypotheses) on intervals(i.e., the control of the $FWER$ is guaranteed for every set of consecutive true null hypotheses). The Vsevolozhskaya-Greenwood-Holodov test (Vsevolozhskaya et al., 2014) controls the $FWER$ on the intervals a priori chosen and on their unions. The Interval-wise test and Vsevolozhskaya-Greenwood-Holodov test are explained in Chapter 1.

In order to deal with inferential problems in the FDA framework, several parametric and/or asymptotic solutions have been proposed (Horváth and Kokoszka, 2012). Alternatively nonparametric methods can be used. Nonparametric methods require less assumptions (for example, normality and homoscedasticity). For instance, it is possible to adopt permutation solutions (Corain et al., 2014).

We have opted for the nonparametric permutation approach since the Interval Testing Procedure, whose performances are explored in simulations (refer to Chapter 3 and Chapter 4), is based on the Nonparametric Combination (NPC), a methodology for multiple testing in the permutation framework (Pesarin and Salmaso (2010)). In detail in simulations, it is performed the adjustment of the univariate p-values associated with a suitable set of permutation tests by using the Interval Testing Procedure. For general concepts about the permutation approach refer to Chapter 1.

In this work we analyze the performances of the Interval Testing Procedure by carrying out a comparison with the multiple testing procedures Bonferroni, Bonferroni-Holm (Holm, 1979) and Benjamini-Hochberg (Benjamini and Hochberg, 1995).

We denote with α the significance level which is the probability of I error type, i.e, the rejection of the null hypothesis H_0 when the null hypothesis H_0 is true. In hypothesis testing, the multiple testing procedures are used in order to avoid the inflation of the I error type. In general if we choose the same significance level α for every univariate test, it is not guaranteed a control of the I error type with probability α for the set of all univariate tests. In order to obtain this kind of control, we need to apply a multiple testing procedure whose effect is a significance level reduction for each univariate test with respect to the starting significance level α. A decrease of the significance level α involves an increase of the univariate p-values.

In the following, we describe the multiple testing procedures Bonferroni, Bonferroni-Holm, Benjamini-Hochberg and Interval Testing Procedure. In the Bonferroni procedure, every unadjusted p-value is multiplied by the number of tests performed n (adjusted p-values larger than 1 are set to 1). Equivalently, we reject the null hypothesis of the univariate test with unadjusted p-value p_{unadj} if it holds

$$p_{unadj} \leq \frac{\alpha}{n} \tag{2.1}$$

where α is the significance level and n is the number of tests. The control of the Family Wise Error Rate $FWER$ provided by the procedure Bonferroni is strong (i.e., the control of the $FWER$ is guaranteed for all possible sets of true null hypotheses).

The Bonferroni-Holm procedure is a Bonferroni sequentially rejective multiple testing procedure. The statistical power (i.e., the probability of rejecting at least one null hypothesis belonging to the set of the false null hypotheses) of the Bonferroni-Holm procedure is greater than or equal to the statistical power of the Bonferroni procedure. At the same time in the same way as the Bonferroni procedure, the Bonferroni-Holm procedure is provided with the strong control of the Family Wise Error Rate $FWER$. n denotes the number of tests. Firstly, we perform the univariate tests saving their p-values $p^1, \dots, p^n$. Then, we order these p-values in ascending order and we denote them with $p^{(1)}, \dots,\ p^{(n)}$ with associated null ordered hypotheses $H_0^{(1)}, \dots,\ H_0^{(n)}$. We apply to the first p-value $p^{(1)}$ in the unadjusted p-values ordering the Bonferroni procedure for n tests (scaling factor of the significance level α equal to n in (2.1)). If the test is not significant (i.e., we obtain $p^{(1)} > \frac{\alpha}{n}$), we don't reject the current null hypothesis $H_0^{(1)}$ and the null hypotheses associated with the subsequent tests $H_0^{(2)}, \dots,\ H_0^{(n)}$, and the procedure stops. If the test is significant (i.e., we obtain $p^{(1)} \leq \frac{\alpha}{n}$), the second p-value in the unadjusted p-values ordering $p^{(2)}$ is adjusted with the Bonferroni procedure for $n-1$ tests and the rationale is the same as in the previous step. In general if the null ordered hypotheses $H_0^{(1)},\ H_0^{(2)}, \dots,\ H_0^{(i-1)}$

have been rejected, the i-th p-value $p^{(i)}$ is adjusted with the Bonferroni procedure for $n-(i-1)$ tests (scaling factor of the significance level α equal to $n-(i-1)$ in (2.1)). If the test is not significant (i.e., we obtain $p^{(i)} > \frac{\alpha}{n-(i-1)}$), we don't reject the current null hypothesis $H_0^{(i)}$ and the null hypotheses associated with the subsequent tests $H_0^{(i+1)}, \ldots,\ H_0^{(n)}$, and the procedure stops. If the test is significant (i.e., we obtain $p^{(i)} \leq \frac{\alpha}{n-(i-1)}$), we reject the current i-th test and the same procedure is performed for the next p-value in the unadjusted p-values ordering $p^{(i+1)}$ or the procedure stops if $p^{(i)}$ is the last p-value in the ordering of the unadjusted p-values.

The Benjamini-Hochberg procedure is a sequentially rejective multiple testing procedure. The steps of the Benjamini-Hochberg procedure are:

- Perform the univariate tests and save their corresponding p-values $p^1, \ldots,\ p^n$. Order these p-values in ascending order obtaining the set of p-values $p^{(1)}, \ldots,\ p^{(n)}$ with associated null ordered hypotheses $H_0^{(1)}, \ldots,\ H_0^{(n)}$.

- Set a desired value q for the False Discovery Rate FDR (i.e., the mean value of the proportion of rejected true null hypotheses or equivalently the expected rate of false discoveries). Denoting with V and R the random variables number of false discoveries and number of rejections respectively, this metric is given by

$$FDR = \mathbb{E}\left[\frac{V}{R}\right]$$

- Search the largest p-value $\tilde{p}$ in the set $p^{(1)}, \ldots,\ p^{(n)}$ such that it is satisfied the constraint

$$\tilde{p} \leq k\left(\frac{q}{n}\right)$$

  where k is the position of $\tilde{p}$ in the set $p^{(1)}, \ldots,\ p^{(n)}$.

- The p-value $\tilde{p}$ and the p-values smaller than it are significant, i.e., we reject the set of ordered null hypotheses $H_0^{(1)}, \ldots,\ H_0^{(k)}$.

The Benjamini-Hochberg procedure controls the FDR, i.e, $FDR \leq q$ for every possible configuration of true null hypotheses.

The Interval Testing Procedure, as a first step, requires the discretization of the data through a suitable basis expansion. The univariate permutation tests are expressed in terms of the basis coefficients. From the unadjusted p-values, the multivariate tests, expressed in terms of the coefficients and pertaining the family of all possible consecutive hypotheses, are performed using the NPC methodology. Finally, with a maximization operation on the p-values obtained from the previous step, the univariate p-values are adjusted. The Interval Testing Procedure is characterized by a control of the Family Wise Error Rate $FWER$ on intervals (i.e., the control of the $FWER$ is guaranteed for every set of consecutive true null hypotheses). Such a control is intermediate between the strong control and the weak control (i.e., the control of the $FWER$ is guaranteed only when the set of the true null hypotheses is composed by all null hypotheses).

The Interval Testing Procedure is explained in a more detailed fashion in the next section. Finally, a brief discussion about the Smoothing splines method is reported. In this work we have used both the Smoothing B-splines and the regression B-splines.

2.2 The Interval Testing Procedure

Firstly, the Interval Testing Procedure (ITP) is presented for the two populations case, i.e. for testing differences in mean between two functional populations (Two-population framework). At a later stage, its extensions to the multiple populations and one population frameworks are described (Multi-population and One-population frameworks). For the ITP extension to functional linear models with functional responses and parameters and scalar covariates refer to Pini (2014).

2.2.1 The ITP in the Two-Population Framework

Let $y = \{y_{11}, \dots,\ y_{n_1 1},\ y_{12}, \dots,\ y_{n_2 2}\}$ be the data set, where n_1 (n_2) is the number of elements belonging to the first (second) population and

$$y_{ij} \in L^2\left(T\right),\ i = 1, \dots, n_j;\ j = 1, 2;\ T = (a,\ b)$$

Denote with $n = n_1 + n_2$ the total number of statistical units. The hypotheses are:

$$H_0 : Y_1 \overset{d}{=} Y_2 \ \mathit{against}\ H_1 : Y_1 \overset{d}{\neq} Y_2 \tag{2.2}$$

where Y_1 and Y_2 are two random functions. The two samples are drawn from these random functions, i.e.,

$$\{y_{ij}\}_{i=1}^{n_j} \overset{i.i.d}{\sim} Y_j,\ j = 1, 2$$

Depending on the assumptions on Y_1 and Y_2, we have two possible cases:

1. Unpaired case, i.e. Y_1 and Y_2 are independent.
2. Paired case, i.e. Y_1 and Y_2 are dependent.

We here describe the ITP for testing (2.2) in both cases. The steps of the ITP are:

1. Use of a basis expansion.
2. Joint permutation univariate tests (expressed in terms of the basis coefficients).
3. Interval-wise combination and adjustment of the univariate p-values obtained from the previous step.

2.2.1.1 Basis Expansion

First of all, a suitable basis expansion has to be chosen. Let $\{\phi^{(k)}\}_{k=1}^{p}$ be the basis, where p is the tunable finite dimension of the functional space. Hence, we have

$$y_{ij}(t) = \sum_{k=1}^{p} c_{ij}^{(k)} \phi^{(k)}(t) \tag{2.3}$$

Every statistical unit $y_{ij}(t)$ is identified by the coefficients $\left\{c_{ij}^{(k)}\right\}_{k=1}^{p}$.

In the unpaired case, the model is

$$c_{11}^{(k)}, \dots, c_{n_1 1}^{(k)} \overset{i.i.d}{\sim} C_1^{(k)},\ c_{12}^{(k)}, \dots, c_{n_2 2}^{(k)} \overset{i.i.d}{\sim} C_2^{(k)},\ \forall k \in \{1, \dots,\ p\}$$

and $C_1^{(k)}$ and $C_2^{(k)}$ are independent. In the paired case we have $n_1 = n_2$, and the model is

$$\left(c_{11}^{(k)},\ c_{12}^{(k)}\right), \dots, \left(c_{n_1 1}^{(k)},\ c_{n_2 2}^{(k)}\right) \overset{i.i.d}{\sim} \left(C_1^{(k)},\ C_2^{(k)}\right),\ \forall k \in \{1, \dots,\ p\}$$

Here we use the B-spline basis expansion, and two methods are available for Smoothing:

- Regression splines. The observed data are smoothed by means of a linear combination of the form (2.3) whose coefficients are generally computed according to a least squares approach. In the case of the B-spline basis expansion, the tunable parameters are the number of basis elements p and the order m. Unlike the method Smoothing splines, no penalty term is used.

- Smoothing splines. For a discussion about this Smoothing method refer to Section 2.3.

2.2.1.2 Joint Permutation Univariate Tests

p univariate permutation tests for the coefficients of the basis expansion are performed. The hypotheses (2.2) are reformulated in terms of the basis coefficients:

$$H_0^{(k)} : C_1^{(k)} \overset{d}{=} C_2^{(k)}\ against\ H_1^{(k)} : C_1^{(k)} \overset{d}{\neq} C_2^{(k)},\ \forall k \in \{1, \dots,\ p\} \tag{2.4}$$

Hence, the global testing problem (2.2) is decomposed in p univariate sub-problems. The k-th univariate permutation test is essentially based on a family of transformations which preserve the likelihood under the null hypothesis $H_0^{(k)}$ (i.e., exchangeability under the null hypothesis) and on a suitable test statistic which is stochastically larger under $H_1^{(k)}$ than under $H_0^{(k)}$. In general the family of transformations depends on the assumptions of independence between the two populations. The choice of the test statistic T is determined by the type of test and by the basis expansion used.

In this framework it is reasonable to choose as a test statistic the modulus of the difference between the population sample means for both the paired and unpaired cases. It is implicitly assumed that, if there is a distributional difference between the two populations, this difference is due to a fixed or random effect (i.e., the means of the populations are different) which is either positive or negative (not both).

In the unpaired case, the family of likelihood-invariant transformations under $H_0^{(k)}$ is composed by any permutation over the sample units of the basis coefficients. Instead, in the paired scenario we have to restrict the set of the permutations over the sample units. Under $H_0^{(k)}$, the exchangeability is just between and within couples. If it is used a test statistic of type $T(Y) = S_1(Y_1) - S_2(Y_2)$ where S_i are symmetric functions having the same form (e.g., sample means), the permutations between couples are ineffective. Hence, in these cases only the permutations within couples are considered. It is worth observing that the permutations of the coefficients must be jointly performed since, in general, the coefficients $\left\{c_{ij}^{(k)}\right\}_{k=1}^{p}$ are dependent with respect to index k. The underlying dependence relations are implicitly taken into account by the permutation approach due to its nonparametric nature.

Therefore, from tests (2.4), we obtain p univariate p-values which have to be adjusted to take into account the multiplicity. The adjustement can be made using classical multiple testing procedures such as the Bonferroni, Bonferroni-Holm (Holm, 1979). These multiple testing are applicable since the data have been discretized by means of a basis expansion. In the ITP, the adjustement is performed in a different way in order to take into account the structure of the functional data.

2.2.1.3 Interval-wise Combination and Correction of the Univariate Tests

At this stage, we need to construct suitable combinations of the univariate test statistics in order to obtain an interval-wise control of the Family Wise Error Rate $FWER$ (i.e., the probability of rejecting at least one null hypothesis belonging to the set of the true null hypotheses). This task is carried out combining the p univariate test statistics by means of multivariate nonparametric combinations, i.e., the Nonparametric Combination (NPC) methodology is applied (Pesarin and Salmaso, 2010). The NPC is a theoretical tool that is used to obtain multivariate permutation tests by combining the results of possibly dependent univariate permutation tests. The method will be described in 2.2.1.4 for the global hypothesis testing problem $H_0 : \cap_{k=1}^{p} C_1^{(k)} \overset{d}{=} C_2^{(k)}$ *against* $H_1 : \cup_{k=1}^{p} C_1^{(k)} \overset{d}{\neq} C_2^{(k)}$. For a rich literature about NPC refer to Pesarin and Salmaso (2010). In the ITP the same procedure described in 2.2.1.4 is applied for each set of subsequent basis coefficients obtaining a family of tests with their associated p-values (Interval-wise Combination).

Lastly for $k \in \{1, \dots, p\}$, the k-th ITP adjusted p-value is obtained calculating the maximum p-value over the p-values of the previous tests with null hypothesis implying $H_0^{(k)}$ (Correction of the Univariate Test). We get the significant components by applying to the adjusted p-values a threshold equal to the significance level α.

The ITP is an inferential procedure provided with a control of the Family Wise Error Rate $FWER$ on any interval of components. As a result also the weak control of the $FWER$ (i.e., the control of the $FWER$ is guaranteed only when the set of the true null hypotheses is composed by all null hypotheses) and the control of the Component Wise Error Rate $CWER$ (i.e., the probability of rejecting a true null hypothesis) are guaranteed, since these are control of the $FWER$ on particular intervals: the entire set of components (global test) and the single component (marginal test), respectively. The procedures Bonferroni and Bonferroni-Holm (Holm, 1979) are provided with a strong control of the $FWER$ (i.e., it is guaranteed the control of the $FWER$ for all possible sets of true null hypotheses). Since the multiple testing procedure Benjamini-Hochberg (Benjamini and Hochberg, 1995) controls the False Discovery Rate FDR (i.e., the mean value of the proportion of rejected null hypotheses that are true or equivalently the expected rate of false discoveries), this procedure is only provided with a weak control of $FWER$ since, if all null hypotheses are true, then the metrics $FWER$ and FDR coincide.

The interval-wise control of the $FWER$ is a desired property in the FDA framework since the functional basis are characterized by an ordered structure. For instance, with the B-spline basis we have localization in space and the ITP can be used to select the significant intervals of the domain (intervals where there is statistical evidence to affirm that the population distributions are different). For example, with the Fourier basis we have localization in frequency and the ITP can be used to select the significant band of frequencies in a data-driven fashion.

It can be proved that the "interval" power (i.e, the probability of rejecting at least one $H_0^{(k)}$ when at least one of the sub-hypotheses is false) of the ITP is greater than the "interval" power of the Closed Testing Procedure (CTP), i.e the procedure whose tests are all possible multivariate tests which are $2^p - 1$ and its k-th adjusted p-value is obtained by computing the maximum p-value over the p-values of all tests whose null hypothesis implies $H_0^{(k)}$ (Marcus et al., 1976). The CTP guarantees the strong control of the $FWER$. However, its computational cost is unaffordable in the functional framework due to the too large number of tests required. The "interval" power of the ITP is smaller than the "interval" power of the Global Testing Procedure (GTP), i.e., the procedure whose test is the global one; the p-value of the global test is obtained from the formula (2.7) which constitutes the final step of the NPC application described in 2.2.1.4. The GTP provides only a weak control of the $FWER$ and, in case of rejection unlike the ITP and CTP, it does not give any information about how the data differ in distribution. For details refer to Pini and Vantini (2013) in which these theoretical properties are proved, and a simulation study is reported.

2.2.1.4 Details on the implementation

In applications, the set of possible permutations is often too large to be explored in its entirety. Hence, only a subset of the permutations is explored through a Conditional Monte Carlo (CMC) algorithm whose steps are:

Algorithm 2.1 Conditional Monte Carlo for sampling the distribution of the test statistic T applied on permuted data under the null hypothesis H_0

1. Evaluate the test statistic on the observed data $T_0 = T(Y)$.
2. Repeat these steps B times: sample a permutation of the observed data Y_b^* and calculate the test statistic on the permuted data $T_b^* = T(Y_b^*)$.
3. Compute the p-value of the test as the proportion of permuted scenarios in which the test statistic evaluated on the simulated permutations is greater than or equal to T_0:
$$\hat{\lambda} = \frac{\#\{T_b^* \geq T_0\}}{B}$$

In the following, the working principle of the NPC is detailed for the Two-Population framework. For ease of notation, the unit index for the second population starts from $n_1 + 1$ where n_1 is the number of statistical units of the first population. The hypotheses of the global test are:

$$H_0 : \cap_{k=1}^{p} C_1^{(k)} \stackrel{d}{=} C_2^{(k)} \ against \ H_1 : \cup_{k=1}^{p} C_1^{(k)} \stackrel{d}{\neq} C_2^{(k)} \tag{2.5}$$

Let $T_0^{(1)}, \ldots,\ T_0^{(p)}$ be the values assumed by the test statistic T applied on the sets of the observed coefficients

$$\left\{c_{11}^{(k)}, \ldots, c_{n_1 1}^{(k)},\ c_{n_1+1,2}^{(k)}, \ldots, c_{n_1+n_2,2}^{(k)}\right\}_{k=1}^{p}$$

Let $T_b^{(1)*}, \ldots,\ T_b^{(p)*}$ be the values assumed by the test statistic T applied on the sets of the coefficients jointly permuted

$$\left\{c_{\pi_1^{(b)}1}^{(k)*}, \ldots, c_{\pi_{n_1}^{(b)}1}^{(k)*},\ c_{\pi_{n_1+1}^{(b)}2}^{(k)*}, \ldots, c_{\pi_{n_1+n_2}^{(b)}2}^{(k)*}\right\}_{k=1}^{p}$$

where $\left(\pi_1^{(b)}, \ldots,\ \pi_{n_1}^{(b)},\ \pi_{n_1+1}^{(b)}, \ldots,\ \pi_{n_1+n_2}^{(b)}\right)$ is a likelihood-invariant random permutation of the unit indexes $(1, \ldots,\ n_1, \ldots,\ n_1 + n_2)$ and $b \in \{1, \ldots,\ B\}$ (B is the number of iterations of the CMC algorithm and it coincides with the number of permutations sampled). Denote with $\hat{L}_l^{(k)}$ the empirical marginal survival function of the test statistics $T^{(k)}$ computed in $T_l^{(k)*}$ with $l \in \{0,\ 1, \ldots,\ B\}$ and $k \in \{1, \ldots,\ p\}$.

$\hat{L}_l^{(k)}$ is given by

$$\hat{L}_l^{(k)} = \frac{\sum_{q=1}^{B} \mathbb{I}\left(\left\{T_q^{(k)*} \geq T_l^{(k)*}\right\}\right) + \frac{1}{2}}{B+1}, \; l \in \{0,\ 1, \dots,\ B\},\ k \in \{1, \dots,\ p\} \quad (2.6)$$

Note that $\left\{\hat{L}_0^{(k)}\right\}_{k=1}^{p}$ are the univariate p-values (empirical survival function evaluated in the realizations of the test statistic applied on the observed coefficients). Then, we calculate the quantities

$$T_0^{(1,\dots,\ p)} = \psi\left(\hat{L}_0^{(1)}, \dots,\ \hat{L}_0^{(p)}\right);\ T_b^{(1,\dots,\ p)*} = \psi\left(\hat{L}_b^{(1)}, \dots,\ \hat{L}_b^{(p)}\right),\ b \in \{1, \dots,\ B\}$$

where $\psi : [0,\ 1]^p \mapsto \mathbb{R}$ is a combining function, i.e., a continuous non-increasing function which is symmetric on its arguments and attains its maximum value when at least one argument tends to zero. An example is given by the *Fisher omnibus* combining function $\psi_{Fisher} : [0,\ 1]^n \mapsto \mathbb{R}$ whose analytical expression is:

$$\psi_{Fisher} = -2 \sum_{i=1}^{n} log\left(x_i\right)$$

Finally, we compute the p-value associated with the global test (2.5):

$$\hat{L}_0^{(1,\dots,\ p)} = \frac{\sum_{q=1}^{B} \mathbb{I}\left(\left\{T_q^{(1,\dots,\ p)*} \geq T_0^{(1,\dots,\ p)}\right\}\right) + \frac{1}{2}}{B+1} \quad (2.7)$$

The formulas (2.6) and (2.7) are theoretically justified by the following result: conditioned to the observed data, every point in the space of likelihood-invariant permutations is equally likely. Under the null hypothesis H_0, a random admissible permutation and the observed data provide the same information content (exchangeability under H_0). Therefore under H_0, conditioning on the available data is equivalent to conditioning on the set of admissible permutations. The admissible permutations are the transformations of data which are likelihood-invariant under H_0.

The terms 1/2 and 1 added respectively to numerator and denominator of the estimators (2.6), (2.7) involve survival function estimates in the open interval (0,1). In this way the transformations by inverse cumulative distribution function of continuous distribution are well defined and continuous. But since these transformations can be not required for the analysis and B is generally chosen sufficiently big, this modification of the classical empirical survival estimator is irrelevant in practice.

2.2.2 The ITP in more general Hypothesis Testing problems

The ITP can be extended to different hypothesis testing problems. Specifically, we illustrate the basic elements of the ITP in the framework of testing difference between more than two populations (Multi-population framework), and in the case of testing the center of symmetry of one population (One-population framework).

2.2.2.1 The ITP in the Multi-Population Framework

We have $g > 2$ sets of functions $y = \left\{y_{11}, \ldots,\ y_{n_1 1}, \ldots,\ y_{1g}, \ldots,\ y_{n_g g}\right\}$ and we want to perform the following test:

$$H_0 : Y_1 \overset{d}{=} Y_2 \overset{d}{=} \ldots \overset{d}{=} Y_g \ against \ H_1 : \exists l,\ s \in \{1, \ldots, g\} :\ Y_l \overset{d}{\neq} Y_s$$

where $\{y_{ij}\}_{i=1}^{n_j} \overset{i.i.d}{\sim} Y_j,\ j \in \{1, \ldots, g\}$. Under the null hypothesis H_0, the g populations share the same distribution. Under the alternative hypothesis H_1, it exists at least one couple of populations with different distributions. First of all, we need to compute for each statistical unit the p coefficient of its basis expansion (2.3). Consequently, we have for the k-th basis component the vector

$$c^{(k)} = \left(c_{11}^{(k)}, \ldots,\ c_{n_1 1}^{(k)}, \ldots,\ c_{1g}^{(k)}, \ldots,\ c_{n_g g}^{(k)}\right),\ k \in \{1, \ldots,\ p\}$$

and it holds

$$c_{ij}^{(k)} \sim C_j^{(k)},\ j \in \{1, \ldots,\ g\}\,;\ i \in \{1, \ldots,\ n_j\}$$

We denote with n the total number of statistical units. The sub-hypotheses on basis coefficients are:

$$H_0^{(k)} : C_1^{(k)} \overset{d}{=} C_2^{(k)} \overset{d}{=} \ldots \overset{d}{=} C_g^{(k)} \ against \ H_1^{(k)} : \exists i, j : C_i^{(k)} \overset{d}{\neq} C_j^{(k)},\ k \in \{1, \ldots,\ p\}$$

The set of permutations likelihood-invariant under the null hypothesis is determined by the model assumptions. In the independent case (Functional Analysis of Variance), the family of transformations are the simple permutations of units. In the dependent case (Functional Repeated Measurements), the likelihood-invariant transformations are the within and between permutations of the observed units.

In the unpaired scenario, we can use the Fisher's test statistic for the univariate test:

$$T\left(c_{11}^{(k)*}, \ldots,\ c_{n_1 1}^{(k)*}, \ldots,\ c_{1g}^{(k)*}, \ldots,\ c_{n_g g}^{(k)*}\right) = \frac{\sum_{j=1}^{g} n_j \left(\bar{c}_j^{(k)*} - \bar{c}^{(k)*}\right)^2}{\sum_{j=1}^{g} \sum_{i=1}^{n_\tau} \left(c_{ij}^{(k)*} - \bar{c}^{(k)*}\right)} \frac{(n-g)}{(g-1)} \tag{2.8}$$

where $\bar{c}_j^{(k)*}$ is the sample mean of the permuted coefficient belonging to the population j. $\bar{c}^{(k)*}$ is the sample mean of the permuted coefficients and it coincides with the sample mean of the observed coefficients $\bar{c}^{(k)}$. These quantities are said to be permutationally equivalent. The concept of permutational equivalence can be exploited in order to simplify the analytical expression of the test statistic used obtaining a gain in terms of computational cost (Pesarin and Salmaso, 2010).

In the paired scenario, we can use the Hotelling's T^2:

$$T^2\left(c_{11}^{(k)*},\ldots,\ c_{n_1 1}^{(k)*},\ldots,\ c_{1g}^{(k)*},\ldots,\ c_{n_g g}^{(k)*}\right) = n_1\left(\Delta\bar{c}^{(k)*}\right)'\left(\Delta S_c^*\Delta'\right)^{-1}\left(\Delta\bar{c}^{(k)*}\right) \tag{2.9}$$

where $\Delta \in \mathbb{R}^{(g-1)\times g}$ is a contrast matrix, $\bar{c}^{(k)*}$ is the vector of the sample means of the k-th permuted coefficients and S_c^* is the sample variance-covariance matrix of the permuted coefficients. Observe that (2.8), (2.9) are the test statistics used in the classical parametric tests. It occurs in permutation tests that the test statistic coincides with the test statistic of the parametric counterpart. This is a sound characteristic of the permutation tests since it permits a direct comparison between the permutation procedures and the parametric methods.

2.2.2.2 The ITP in the One-Population Framework

Let $y = \{y_1,\ldots,\ y_n\}$ be the data and we have $y_1,\ldots,\ y_n \overset{i.i.d}{\sim} Y \in \mathbb{R}$. We want to test the center of symmetry of the functional population Y, i.e., we want to verify if there is statistical evidence to claim that the center of symmetry of the population Y is equal to a function μ_0. The hypotheses are:

$$H_0 : \mathbb{E}\left[Y\right] = \mu_0 \ against \ H_1 : \mathbb{E}\left[Y\right] \neq \mu_0$$

As a first step, we have to consider the basis expansions of the data y and of the mean under the null hypothesis μ_0:

$$y_i\left(t\right) = \sum_{k=1}^{p} c_i^{(k)}\phi^{(k)}\left(t\right),\ \mu_0\left(t\right) = \sum_{k=1}^{p} c_0^{(k)}\phi^{(k)}\left(t\right)$$

Consequently, for each unit and for μ_0 we have p coefficients. In particular, we have

$$c_1^{(k)},\ldots,c_n^{(k)} \overset{i.i.d}{\sim} C^{(k)},\ k \in \{1,\ldots,\ p\}$$

Afterwards, the functional hypotheses are expressed in terms of the coefficients:

$$H_0^{(k)} : center\left[C^{(k)}\right] = c_0^{(k)} \ against \ H_1^{(k)} : center\left[C^{(k)}\right] \neq c_0^{(k)}$$

In order to find a set of likelihood-invariant transformations under the null hypothesis, the distribution from which y is sampled is assumed to be symmetric. If the mean of the population distribution exists, it is equal to the center of symmetry and, hence, the previous test is also a test for the mean. In this framework, the likelihood-invariant transformations are the reflections of the function y_i with respect to μ_0. These transformations, in terms of the basis coefficients, coincide with the simultaneous reflections through $\left\{c_0^{(k)}\right\}_{k=1}^{p}$ of all coefficients $\left\{c_i^{(k)}\right\}_{k=1}^{p}$ of the same unit $i \in \{1, \ldots,\ n\}$. The test statistic depends on the basis expansion used. If the B-spline basis is used, a reasonable choice is a suitable divergence test statistic between the sample mean of the transformed k-th coefficients associated with the data y and the k-th coefficient associated with the mean under the null hypothesis μ_0:

$$T^{(k)*} = |\frac{\sum_{s=1}^{n} c_s^{(k)*}}{n} - \mu_0^{(k)}|,\ k = 1, \ldots,\ p$$

If the mean does not exist, other empirical measures of the center of symmetry can be used such as the median or an existing trimmed mean.

2.3 Smoothing splines

A peculiar aspect of the Functional Data Analysis (FDA) is the possibility to extract information from the derivatives of the functions. Thus, the Smoothing should be done carefully in order to obtain an accurate reconstruction of the desired derivatives combining linearly the derivatives of the basis elements with the coefficients associated with the original functions.

This aim can be achieved with the Smoothing splines method. In this framework, the number of the basis elements p is equal to the maximum value coinciding with the number of evaluations n_{eval} (interpolation). A penalization is imposed in order to avoid overfitting. The cost function chosen for the simulations is the Residual Sum of Squares with penalization on the second derivative which is given by:

$$RSS\,(f,\ \lambda) = \sum_{i=1}^{n_{eval}} \{y_i - f\,(x_i)\}^2 + \lambda \int \left\{f"\,(t)\right\}^2 dt \tag{2.10}$$

where λ is the smoothing parameter, n_{eval} is the number of evaluations, $\{y_i\}_{i=1}^{n_{eval}}$ are the observed data and f is the unknown function. The data fitting is evaluated with the term $\sum_{i=1}^{N} \{y_i - f\,(x_i)\}^2$. The term $\lambda \int \left\{f"\,(t)\right\}^2 dt$ penalizes the second derivative (curvature). In a Bayesian perspective the penalty term can summarize the a priori information. If $\lambda = 0$, f is the interpolating spline. If $\lambda = \infty$, the second derivative of f must be zero and, hence, we are considering the least squares line fit.

It can be proved that the function f minimizing the Residual Sum of Squares RSS belongs to a finite dimensional space (in general RSS is defined on a functional space with infinite dimension), it is unique and it can be obtained in closed form. This solution is given by the natural (null second and third derivatives at the endpoints) cubic splines with knots in the evaluation points $\{x_i\}_{i=1}^{n_{eval}}$ of the observed data (Friedman et al., 2001).

Therefore, λ is the unique parameter that requires tuning. However, it could be interesting to control if different trends of the tested procedures can be observed by violating one hypothesis of this functional minimization result.

Chapter 3

Simulation setting

3.1 Introduction

The purpose of the simulation study is the evaluation of the Interval Testing Procedure (ITP) performances. A comparison is carried out with the Benjamini-Hochberg (BH) (Benjamini and Hochberg, 1995), Bonferroni-Holm (BFH) (Holm, 1979) and Bonferroni (BF) multiple testing procedures. The ITP, BH, BFH, BF procedures are described in Chapter 2.

The evaluation metrics are:

- The Family Wise Error Rate $FWER$, i.e., the probability of rejecting at least one null hypothesis among the hypotheses to be accepted.
- The Rejection Rate of the False null hypotheses ρ, i.e., the expected rate of rejected null hypotheses among the hypotheses to be rejected,
- The Rejection Rate of the True null hypotheses γ, i.e., the expected rate of rejected null hypotheses among the hypotheses to be accepted.
- The Power π, i.e., the probability of rejecting at least one null hypothesis among the hypotheses to be rejected.

For a description of these metrics refer to Section 3.3.

It is of relevant interest to explore the differences in terms of the ability to make true discoveries between the ITP and the BH, knowing that the former procedure controls the $FWER$ on intervals and the latter procedure ensures only a weak control of the $FWER$, i.e., the control of the $FWER$ is guaranteed only when the set of the true null hypotheses is composed by all null hypotheses.

The tested types of synthetic data set are:

1. Populations of constant and step functions (data set const-step).
2. Populations of constant and tricube functions (data set const-tricube)

For each type of data set, the differences of the two populations in terms of mean (fixed deterministic effect) are localized in the interval denoted by $[h_{inf},\ h_{sup}]$. For details about the data generation mechanism refer to Section 3.2.

Both the Regression and Smoothing splines methods are used. In detail, we have used both the Smoothing B-splines and the Regression B-splines.

The parameters included in the simulation study and related to Smoothing are:

- Order m of the B-spline basis elements.
- Number of basis elements p.
- Smoothing parameter λ. In the Smoothing splines method, λ is the unique parameter that requires tuning (for details about Smoothing splines method refer to Section 2.3). Regarding the regression splines method, the regularity of the Smoothing is determined by the order m (the degree of the basis elements is $m-1$). By augmenting m, we obtain a more regular functional approximation of the observed data.

The other selected parameters, which determine the data set, are:

- Number of evaluations n_{eval}. The statistical units share the same domain $\mathcal{D}$ where the evaluations are uniformly distributed.
- Standard deviation σ_{noise} of the additive Normal noise.
- Number of statistical units n_{units}.

Hence, the aim of the simulation study is the evaluation of ITP performances compared with the BH, BFH and BF ones by varying:

- The data set type (const-step or const-tricube)
- The parameters of interest m and p, λ (Smoothing parameters), n_{eval} and σ_{noise}, n_{units} (Data parameters)
- The Smoothing method: Regression or Smoothing splines using the B-spline basis expansion

The performances of the methods are evaluated by means of the metrics Family Wise Error Rate $FWER$, Rejection Rate False null hypotheses ρ, Rejection Rate True null hypotheses γ and Power π.

The hypothesis testing problem chosen for the simulations is the two-sided distributional comparison between two independent populations of functions (unpaired case). The populations differ in distribution in the interval $[h_{inf},\ h_{sup}] \subset \mathcal{D} = [0,\ 2]$ for a fixed effect (difference in mean) where $\mathcal{D}$ is the domain. With the same notation of Chapter 2, $y_{ij}(t)$ is the statistical unit i in the population j at instant t, with $j = 1,\ 2$, $i = 1, \dots,\ n_j$.

For $i = 1, \dots,\ n_j$, we have

$$y_{i,1} \overset{i.i.d}{\sim} Y_1,\ y_{i,2} \overset{i.i.d}{\sim} Y_2$$

and Y_1 and Y_2 are independent random functions. The hypothesis testing problem is:

$$H_0 :\ Y_1 \overset{d}{=} Y_2 \ against\ H_1 :\ Y_1 \overset{d}{\neq} Y_2$$

The B-spline basis expansion is used throughout the simulations. Hence, each statistical unit y_{ij} is identified by the set of the coefficients $\left\{c_{ij}^{(k)}\right\}_{k=1}^{p}$ associated with the B-spline basis elements.

The univariate p-value λ_k, with $k = 1, ...,\ p$, is obtained using the permutation solution based on:

- The test statistic $T = |\frac{1}{n_1}\sum_{i=1}^{n_1} c_{i,1} - \frac{1}{n_2}\sum_{i=1}^{n_2} c_{i,2}|$. In this setting, it is reasonable to choose the test statistic T because the means of the population distributions are different in the interval $[h_{inf},\ h_{sup}]$ (T is stochastically larger under the alternative hypothesis with respect to the null hypothesis).

- The family of transformations likelihood-invariant under the null hypothesis is composed by the permutations over the sample units of the basis coefficients (exchangeability condition under the null hypothesis).

The univariate p-values $\{\lambda_k\}_{k=1}^{p}$ are approximated through the Conditional Monte Carlo (CMC) algorithm whose number of iterations B is set to 1000 if not specified. The CMC algorithm is described in 2.2.1.4.

Finally as far as concerns the ITP, the Interval-wise Combination and Correction of the univariate tests are performed using the Nonparametric Combination (NPC) methodology with Fisher combining function (Pesarin and Salmaso, 2010). The control of the Family Wise Error Rate $FWER$ is guaranteeed on intervals. For details about the ITP and the NPC refer to Chapter 2.

The code for the simulations has been implemented in R 3.2.0 using *fdatest* 2.1 and *fda* 2.4.4 packages (Team, 2015; Pini and Vantini, 2015b; Ramsay et al., 2014). Smoothing has been performed using *fda* package. *fdatest* package has been used for applying the Interval Testing Procedure in the Two-population framework using the B-spline basis expansion.

The p-values obtained with the Interval Testing Procedure are stored in $p \times p$ matrix called combining matrix which is the input of the p-values adjustement phase. p is the number of the univariate p-values and it coincides with the number of basis elements. The construction of the combining matrix is the most computationally expensive task in the Interval Testing Procedure algorithm for the Two-population framework. It costs p^2.

Hence, the *fdatest* source code has been conveniently modified. In order to improve the execution times, the construction of the combining matrix has been implemented in C. It has been used the .C interface to R. The new implementation has provided a speedup over the original implementantion equal to around 60x. This result is useful both for simulations, where the Interval Testing Procedure must be applied several times for each value of the parameter under analysis, and for scenarios where it is reasonable to choose a high number of basis elements p.

Finally, the used implementation of the *Interval Testing Procedure* directly works on an object of the *functional data class.* Therefore, the Smoothing is performed by the user avoiding subjective choices which had to be taken automatically in the original version of *fdatest.*

In the subsequent sections more details are provided about the tested types of data set const-step and const-tricube and the adopted metrics $FWER$, ρ, γ, π. The simulation results are reported in Chapter 4.

3.2 Data sets

The synthetic data sets chosen for the analysis are:

1. Population constituted by constant functions and population constituted by step functions (data set const-step). The support of the step is $[h_{inf},\ h_{sup}]$ and its value v_{CS} is a fixed deterministic effect.

2. Population constituted by constant functions and population constituted by tricube functions (data set const-tricube). The tricube functions are constituted by a symmetric tricube kernel with support $[h_{inf},\ h_{sup}]$. We denote with v_{CT} the maximum value of the tricube kernel.

We have chosen these types of data set for the totally different degree of regularity that distinguishes them; the data set const-step is not regular since it exhibits discontinuities in the jump points; on the contrary, the data set const-tricube is rather regular since the tricube kernel belongs to $\mathcal{C}^9$. The parameters with subscript or superscript CS are associated with the data set const-step. The parameters with subscript or superscript CT are associated with the data set const-tricube

The value assumed by each constant function is sampled from a Normal distribution with mean $\mu_{model} = 0$ and standard deviation $\sigma_{model} = 0.15$. Then, a local Normal noise with mean $\mu_{noise} = 0$ and standard deviation σ_{noise} is added to each constant function. In the same way, the constant component is computed for the elements belonging to the non-constant population where in scenarios with data set:

1. const-step, for each unit the fixed effect v_{CS} is added in $[h_{inf},\ h_{sup}]$.

2. const-tricube, for each unit the tricube symmetric kernel with maximum value v_{CT} is added with mid point of its support $[h_{inf},\ h_{sup}]$ coincident with the mid point d of the domain $\mathcal{D}$.

For sake of simplicity and without loss of generality, we assume that the interval $[h_{inf},\ h_{sup}]$, where the null hypothesis H_0 (populations share the same distribution) is false, is symmetric with respect to the mid point d of the domain $\mathcal{D}$ and it is the same in both scenarios with data set const-step and const-tricube. Hence, we have that $h_{inf}^{CS} = h_{inf}^{CT} = h_{inf}$ and $h_{sup}^{CS} = h_{sup}^{CT} = h_{sup}$. Instances of the data sets of type const-step and const-tricube with null standard deviation of noise σ_{noise} are reported in Figure 3.1.

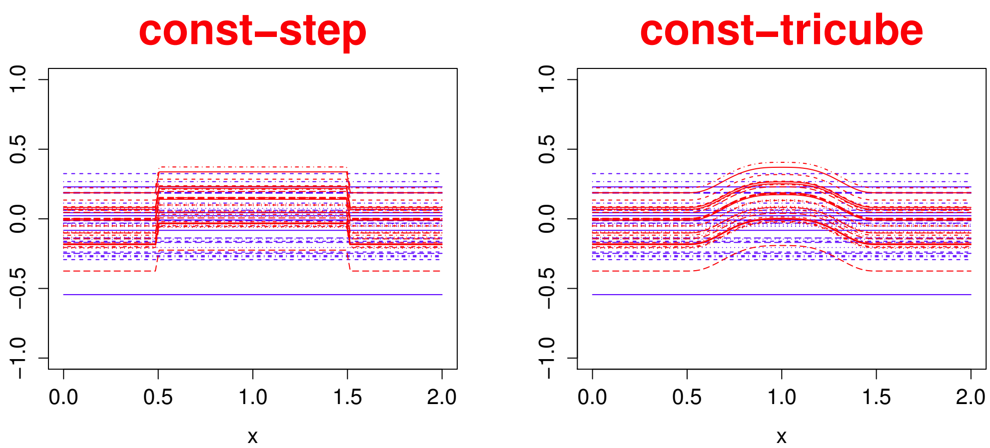

Figure 3.1: Data sets const-step and const-tricube with $\sigma_{noise} = 0$

In order to avoid fictitious differences in terms of true discovery rate, we impose that the L^2 distance between the means of the first population composed by constant functions Y_1 and of the second population Y_2 is the same for the data sets const-step and const-tricube, i.e., we have to satisfy the constraint:

$$\int_{h_{inf}}^{h_{sup}} \left(\mathbb{E}\left[Y_2^{CS}\right] - \mathbb{E}\left[Y_1^{CS}\right]\right)^2 dt = \int_{h_{inf}}^{h_{sup}} \left(\mathbb{E}\left[Y_2^{CT}\right] - \mathbb{E}\left[Y_1^{CT}\right]\right)^2 dt \tag{3.1}$$

where $\mathbb{E}\left[Y_1^{CS}\right] = \mu_{model} = 0$, $\mathbb{E}\left[Y_1^{CT}\right] = \mu_{model} = 0$ (mean values of the populations composed by constant functions associated with data sets const-step and const-tricube, respectively), $\mathbb{E}\left[Y_2^{CS}\right] = v_{CS}$ and

$$\mathbb{E}\left[Y_2^{CT}\right] = \mu_{CT}(t) = v_{CT}\left(1 - |t|^3\right)^3 \mathbb{I}\left(\{|t| \leq 1\}\right) \tag{3.2}$$

μ_{CT} is the tricube kernel with maximum value v_{CT}. If we center the tricube kernel in the origin as in the expression (3.2), we have that

$$t = \frac{x}{d - h_{inf}} = \frac{x}{h_{sup} - d}, \; x \in [h_{inf} - d, \; h_{sup} - d]$$

where $d - h_{inf} = h_{sup} - d$ since the tricube kernel has support $[h_{inf}, \; h_{sup}]$ and it is symmetric with respect to d. The constraint (3.1) can be expressed in the following way:

$$\int_{h_{inf}}^{h_{sup}} (v_{CS})^2 \, dt = \int_{h_{inf}-d}^{h_{sup}-d} (v_{CS})^2 \, dt = v_{CS}^2 \, (h_{sup} - h_{inf}) = \int_{h_{inf}-d}^{h_{sup}-d} (\mu_{CT}\,(x))^2 \, dx$$

Finally, the parameters $h_{inf}, \; h_{sup}, \; v_{CS}^2, \; v_{CT}^2$ must satisfy the following equation:

$$v_{CS}^2 \, (h_{sup} - h_{inf}) = v_{CT}^2 \int_{h_{inf}-d}^{h_{sup}-d} \left(\left(1 - | \frac{x}{d - h_{inf}} |^3 \right)^3 \right)^2 dx \tag{3.3}$$

Observe that, since $\mathbb{E}\left[Y_1^{CS}\right] = \mathbb{E}\left[Y_1^{CT}\right] = \mu_{model} = 0$, the constraint (3.1) is equivalent to impose that the L^2 norms of the means of the data sets const-step and const-tricube coincide.

3.3 Metrics

We here introduce the metrics that will be used in the simulation study to compare the procedures ITP, BH, BFH and BF. We consider a multiple testing problem, where the number of tests is equal to the number of basis elements p. We denote with p_0 the number of true null hypotheses. If the extremes of the interval $[h_{inf}, \; h_{sup}]$ (where H_0 is false) and the number of basis elements p are fixed, p_0 depends only on the order m of the B-splines. The notation used for hypothesis testing is reported in Table 3.1.

Table 3.1: Notation for Hypothesis Testing

	Accept H_0	Reject H_0	Total
True null hypotheses	U	V	p_0
False null hypotheses	T	S	$p - p_0$
	$p - R$	R	p

We denote with n_{data} the number of independent data sets generated. A seed has been fixed to 14091990 for reproducibility of the simulated data sets. The significance level α has been always set to the standard value 0.05.

The metrics and their empirical counterparts are:

- Family Wise Error Rate $FWER$, i.e., the probability of rejecting at least one null hypothesis among the hypotheses to be accepted:

$$FWER = \mathbb{P}\left(\{V > 0\}\right); \; \widehat{FWER} = \frac{\sum_{i=1}^{n_{data}} \mathbb{I}\left(\left\{\exists l \in A : p_l^i \leq \alpha\right\}\right)}{n_{data}}$$

  where $A = A\left(m, \; p, \; h_{inf}, \; h_{sup}\right)$ is the set constituted by the identifiers of the basis coefficients with null hypothesis true, p_l^i is the p-value associated with the coefficient with identifier l when the i-th data set is used.

- Rejection Rate of the False null hypotheses ρ, i.e., the expected rate of rejected null hypotheses among the hypotheses to be rejected:

$$\rho = \frac{\mathbb{E}\left[S\right]}{p - p_0}; \; \hat{\rho} = \frac{\sum_{i=1}^{n_{data}} \frac{s_i}{p - p_0}}{n_{data}}$$

  where s_i is the number of true discoveries when the i-th data set is used.

- Rejection Rate of the True null hypotheses γ, i.e., the expected rate of rejected null hypotheses among the hypotheses to be accepted:

$$\gamma = \frac{\mathbb{E}\left[V\right]}{p_0}; \; \hat{\gamma} = \frac{\sum_{i=1}^{n_{data}} \frac{v_i}{p_0}}{n_{data}}$$

  where v_i is the number of false discoveries when the i-th data set is used.

- Power π, i.e., the probability of rejecting at least one null hypothesis among the hypotheses to be rejected:

$$\pi = \mathbb{P}\left(\{S > 0\}\right); \; \hat{\pi} = \frac{\sum_{i=1}^{n_{data}} \mathbb{I}\left(\left\{\exists l \in C : p_l^i \leq \alpha\right\}\right)}{n_{data}}$$

  where $C = C\left(m, \; p, \; h_{inf}, \; h_{sup}\right)$ is the set constituted by the identifiers of the basis coefficients with null hypothesis false, p_l^i is the p-value associated with the coefficient with identifier l when the i-th data set is used.

In the explored scenarios, the populations share the same distribution in the union of intervals $[0, \; h_{inf}) \cup (h_{sup}, \; 2]$. Due to the ITP p-values adjustment implementation (recycled version of the family composed by the interval-wise tests explained in Pini and Vantini, 2013), this set for the ITP can be considered as a single interval. Consequently, since the ITP is characterized by an interval-wise control of $FWER$, in $[0, \; h_{inf}) \cup (h_{sup}, \; 2]$ this procedure controls the $FWER$. For a detailed explanation of the ITP working principle refer to Section 2.2.

The selected metrics assume values in $[0, \; 1]$. Therefore, in order to quantify the maximum variability of the estimated metric, we can use the Normal approximation of the confidence interval for the Binomial proportion which is given by:

$$\hat{\xi} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_{data}} \hat{\xi} \left(1 - \hat{\xi}\right)}$$

where $\hat{\xi}$ is the metric approximation, α is the significance level, $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-quantile of the standard Normal distribution. This choice is reasonable because n_{data} is generally chosen sufficiently big in a simulation. If ξ is close to 0 or 1, we are not generally interested in quantifying the variability of the estimate with greater accuracy. These scenarios are not appealing for the simulations. However if n is not big and/or ξ tends to assume the minimum or the maximum values, different confidence intervals can be used such as, for instance, the Wilson and Clopper-Pearson confidence intervals. With $\alpha = 0.05$, we obtain the following upper bound for the variability of the estimate:

$$z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n_{data}} \hat{\xi} \left(1 - \hat{\xi}\right)} \leq 2 \sqrt{\frac{1}{n_{data}} \hat{\xi} \left(1 - \hat{\xi}\right)} \leqslant 2 \sqrt{\frac{1}{n_{data}} \frac{1}{4}} = \sqrt{\frac{1}{n_{data}}} = \sigma_{estimate}$$

Therefore, if $n_{data} = 100$, we have $\sigma_{estimate} = 0.1$. If $n_{data} = 1000$, the upper bound is $\sigma_{estimate} \approx 0.032$, which is an acceptable value for the simulations.

Chapter 4

Simulation results

4.1 Introduction

In simulations, the used procedures are:

- Interval Testing Procedure (ITP).
- Benjamini-Hochberg (BH).
- Bonferroni-Holm (BFH).
- Bonferroni (BF).

These procedures are explained in Chapter 2.

The parameters of interest related to Smoothing are:

- Order m of the B-spline basis elements.
- Number of basis elements p.
- Smoothing parameter λ.

The other selected parameters, which determine the data set, are:

- Number of evaluations n_{eval}. The statistical units share the same domain $\mathcal{D} = [0,\ 2]$ where the evaluations are uniformly distributed.
- Standard deviation σ_{noise} of the additive Normal noise.
- Number of statistical units n_{units}.

The tested types of synthetic data are:

1. Populations of constant and step functions (data set const-step).
2. Populations of constant and tricube functions (data set const-tricube)

Instances of the data sets of type const-step and const-tricube with null standard deviation of noise σ_{noise} are reported in Figure 4.1. The parameters with subscript or superscript CS (CT) are associated with the data set const-step (const-tricube). The generation mechanism of these data types is described in Section 3.2.

The considered hypothesis testing problem is the two-sided distributional comparison between two independent populations of functions (unpaired case). The populations differ in distribution in the interval $[h_{inf},\ h_{sup}] \subset \mathcal{D} = [0,\ 2]$ for a fixed effect (difference in mean) where $\mathcal{D}$ is the domain. The hypothesis testing problem and its solution in the permutation framework are explained in Chapter 3.

In order to evaluate the performances of the used procedures, we use the following metrics in detail defined in Section 3.3:

- Family Wise Error Rate $FWER$, i.e., the probability of rejecting at least one null hypothesis among the hypotheses to be accepted.
- Rejection Rate False null hypotheses ρ, i.e., the expected rate of rejected null hypotheses among the hypotheses to be rejected.
- Rejection Rate True null hypotheses γ, i.e., the expected rate of rejected null hypotheses among the hypotheses to be accepted.
- Power π, i.e., the probability of rejecting at least one null hypothesis among the hypotheses to be rejected.

The presentation of the simulation results is divided in two parts determined by the Smoothing parameters (m and p for Regression splines method, λ for Smoothing splines method) and by the Data parameters (n_{eval} and σ_{noise}, n_{units}). m and p are the order and the number of basis elements of the B-splines, respectively. λ is the smoothing parameter. n_{eval} and σ_{noise} are the number of evaluations and the

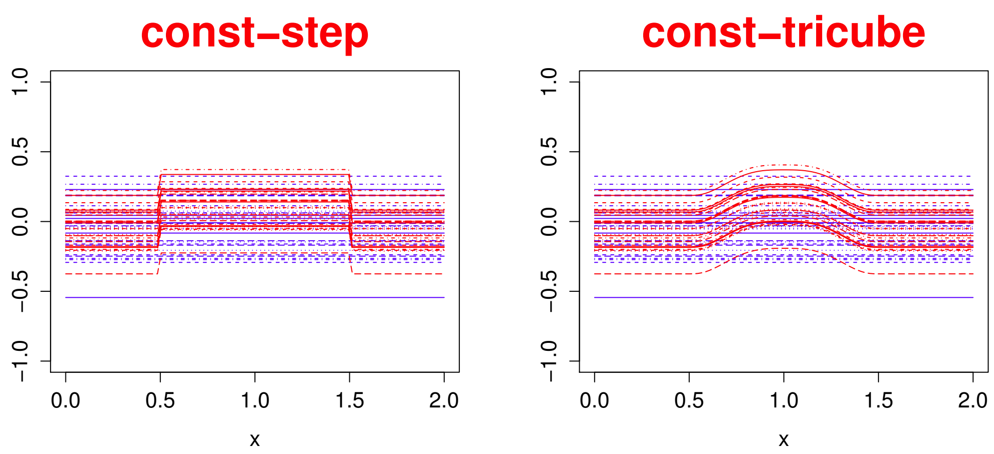

Figure 4.1: Data sets const-step and const-tricube with $\sigma_{noise} = 0$

standard deviation of the additive Normal noise, respectively. n_{units} is the number of statistical units. We assume that the number of statistical units of the first population n_1 coincides with the number of statistical units of the second population n_2. Hence, we have $n_1 = n_2 = n_{units}$.

Summarizing, the main purpose of the simulation study is the evaluation of the ITP performances compared with the BH, BFH and BF ones by varying:

- The data set type (const-step or const-tricube)
- The parameters of interest m and p, λ (Smoothing parameters), n_{eval} and σ_{noise}, n_{units} (Data parameters)
- The Smoothing method: Regression or Smoothing splines using the B-spline basis expansion

The performances of the methods are evaluated by means of the metrics Family Wise Error Rate $FWER$, Rejection Rate False null hypotheses ρ, Rejection Rate True null hypotheses γ and Power π.

Moreover, it is of relevant interest to explore the differences in terms of the ability to make true discoveries between ITP and BH, knowing that the former procedure controls the $FWER$ on intervals and the latter procedure ensures only a weak control of the $FWER$, i.e., the control of the $FWER$ is guaranteed only when the set of the true null hypotheses is composed by all null hypotheses.

In all simulations, the significance level α has been set to the standard value 0.05 and a seed has been set to 14091990 for reproducibility of the simulated data sets. The selected values of the parameters h_{inf}, h_{sup}, v_{CS} (height of the step) and v_{CT} (maximum value of the tricube kernel) guarantee that the L^2 distance between the means of the first population composed by constant functions and of the second population is the same for the data sets const-step and const-tricube.

For details about the hypothesis testing problem considered, the types of data tested, the parameters of interest, the evaluation metrics and in general the simulation setting, consult Chapter 3.

Throughout the whole chapter as far as concerns the adopted metrics, the subscripts ITP, BH, BFH and BF are associated with the Interval Testing Procedure, Benjamini-Hochberg, Bonferroni-Holm, Bonferroni procedures, respectively.

Simulations with Smothing parameters variable

In the first part of this simulation study we investigate the performances of ITP depending on the Smoothing parameters. In Section 4.2, we consider Regression B-splines, and focus on parameters order m and number of basis elements p. In Section 4.3, we consider instead Smoothing B-splines, and focus on the smoothing parameter λ.

4.2 Regression Splines parameters variable

4.2.1 Introduction

In this exploratory analysis we want to evaluate the performances of the Interval Testing Procedure (ITP) by varying the order of the B-splines m and the number of basis elements p. In monographs it is possible to find several methods for Smoothing (Silverman and Ramsay, 2005; Friedman et al., 2001) whereas the opposite is true for the effects on performances due to Smoothing. Hence, the exploration of the parameters m and p is interesting in itself, particularly for practitioners who are not used to deal with basis expansions.

The ability to make true discoveries is often seen only as a function of the number of statistical units even if there are other potentially relevant parameters such as, for instance, m and p. Another important and unknown parameter related to the ability to make true discoveries is the size of the effect to be detected.

4.2.1.1 Values of the parameters

The common parameters in both scenarios with data set const-step and const-tricube are:

- Order B-splines: $m \in \{1,\ 2, \dots,\ 10\}$.
- Number basis elements: $p \in \{10,\ 20, \dots,\ 100\}$.
- Number statistical units: $n_{units} = n_1 = n_2 = 30$.
- Number evaluations: $n_{eval} = 100$. The evaluations are uniformly distributed in the domain $\mathcal{D} = [0,\ 2]$.
- Standard deviation additive Normal noise: $\sigma_{noise} = 0$
- Number generated data sets: $n_{data} = 1000$. The data sets are independent and they are the same for each pair $(m,\ p)$.

The values of the remaining parameters are:

- $h_{inf}^{CS} = h_{inf}^{CT} = h_{inf} = 0.5$
- $h_{sup}^{CS} = h_{sup}^{CT} = h_{sup} = 1.5$
- Height of the step (fixed effect) with support $[h_{inf},\ h_{sup}]$: $v_{CS} = 0.15$
- Maximum value of the tricube kernel with support $[h_{inf},\ h_{sup}]$: $v_{CT} = 0.22$

4.2.1.2 Degenerate cases

The results for the cases with $p = 10$ are degenerate for high m in the sense that we observe high and fictious values of the Power. By increasing the order m, the supports associated with the elements of the basis expansion B-spline expand. In the extreme case $\{m = 10,\ p = 10\}$, all p supports of the basis elements have non-empty intersection with the interval $[h_{inf},\ h_{sup}]$ where the null hypothesis is false. Hence, it is correct to reject all p null hypotheses associated with the basis coefficients.

Therefore, a test with Smoothing parameters m and p chosen in this way is similar to a global test. At the same time, if the ITP is used with low p and high m, we are not taking advantage of its domain selection property from which it is possible to identify the intervals of the domain which are sources of distributional difference. Consequently, by augmenting m, the Power functions π with $p = 10$ increase and the unitary value is essentially reached by all procedures (see Figure 4.2). This Power π improvement is fictious and similar observations hold for the cases with $p = 20$.

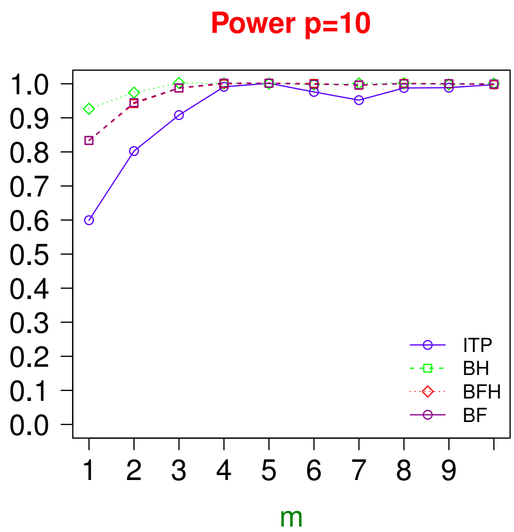

Figure 4.2: Power π as a function of the B-spline order m with number of basis elements $p = 10$. The data type is populations of constant and step functions.

In the following we report the heat maps alternated with a selection of the standard graphics (i.e., metric as a function of m with fixed p or vice versa) and general comments about the observed patterns and the encountered problems. In the standard graphics, we do not report the cases with $p = 10$ since, as discussed previously, these scenarios are degenerate for high m. For completeness, all the standard graphics divided according to the type of data (const-step and const-tricube) are reported in Appendix A.

4.2.2 Results

We have chosen to present the simulation results by means of suitable heat maps which represent the values of a metric by varying both parameters m and p. If need be, the heat maps are alternated with a selection of the standard graphics (i.e., metric as a function of m with fixed p or vice versa).

For each data type (const-step or const-tricube) and metric (Family Wise Error Rate $FWER$, Rejection Rate False null hypotheses ρ, Rejection Rate True null hypotheses γ, Power π), we have a plot with four subgraphs; each subgraph is the representation through an heat map of a metric for one of the tested procedures (Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF)).

When needed, contour lines are added to ease the comparison between the different graphics. In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0. In heat maps of $FWER$ and γ, the squares white opaque are associated with not available numbers, corresponding to the cases when all null hypotheses are false; in these scenarios $FWER$ and γ are not defined. By means of the heat maps, a direct comparison between the scenarios with data sets const-step and const-tricube can be achieved.

4.2.2.1 Family Wise Error Rate

In Figure 4.3 on page 45 we report the heat maps of $FWER$ of ITP, BH, BFH, BF.

Firstly, as expected, $FWER_{BF}$ and $FWER_{BFH}$ assume lower values than $FWER_{BH}$ and $FWER_{ITP}$. The results obtained with BF and BFH are very similar in both scenarios with data sets const-step and const-tricube.

ITP controls the $FWER$ (net of the upper bound of the metric estimate variability $\sigma_{estimate}(\alpha) = \sigma_{estimate} = \frac{1}{\sqrt{n_{data}}} \approx 0.032$; for details refer to Section 3.3). In particular, considering the cases with data type const-step, $FWER_{ITP}$ assumes larger values for the set of orders $\{2,\ 3,\ 4\}$ and for high p. In contrast, in scenarios with data set const-tricube, $FWER_{ITP}$ tends to be globally smaller and we observe low variability of this metric.

BH does not control the $FWER$ with both data sets const-step and const-tricube. In scenarios with data set const-step, $FWER_{BH}$ is greater than α essentially in all extreme cases (low p and high m, low p and low m, high p and high m, high p and low m). The difference $FWER_{BH} - \alpha$ is maximized if the cubic splines are used with very high p, especially with the interpolating spline ($p^{max} = n_{eval} = 100$). In the case of data set const-tricube, we observe less variability of the $FWER_{BH}$ and $FWER_{BH} > \alpha$ for low p and high m and for high p and small m (in the former region $FWER_{BH}$ attains its maximum).

The fact that, in scenarios with data set const-step, for ITP and BH there is more variability of the $FWER$ is coherent with the low degree of regularity of the data type const-step. Since the data set const-step is discontinuous, the assumption of intrinsic regularity of the datum typical of the Functional Data Analysis is violated. Hence, Smoothing is a critical phase in this case as we will see in the next paragraphs. The same observations hold for the Rejection Rate False null hypotheses ρ, particularly for ITP.

In Figure 4.4 we report the $FWER$ as a function of p with the most used orders in applications $m \in \{1,\ 2,\ 3,\ 4\}$ and in scenarios with data type const-step. Firstly for $m = 1$, $FWER_{ITP} \leq \alpha$ for every p. This relation holds for the orders $m \in \{1,\ 6,\ 7,\ 8,\ 9,\ 10\}$. For $m \in \{2,\ 3,\ 4,\ 5\}$, the ITP controls the $FWER$ net of the upper bound of the metric estimate variability $\sigma_{estimate}$ (for details refer to Section 3.3). What it is interesting to observe is that $FWER_{ITP}$ is maximized if the cubic B-splines are used and the same holds for ρ_{ITP} as we will see in the next paragraphs.

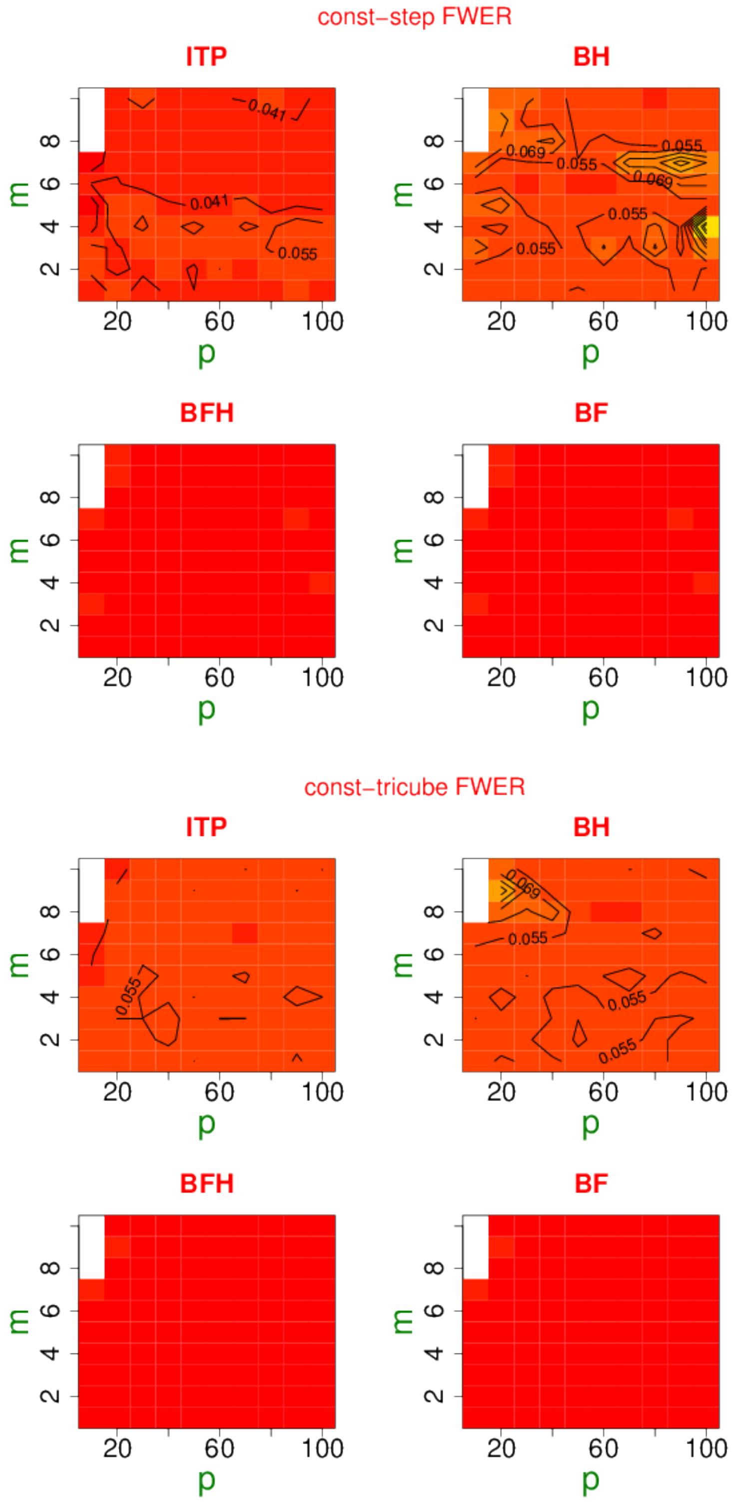

Figure 4.3: Heat maps of the Family Wise Error Rate $FWER$ as a function of p and m for the procedures Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF) and for the data sets const-step (populations of constant and step functions) and const-tricube (populations of constant and tricube functions). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0. The squares white opaque are associated with not available numbers, corresponding to the cases when all null hypotheses are false; in these scenarios $FWER$ is not defined.

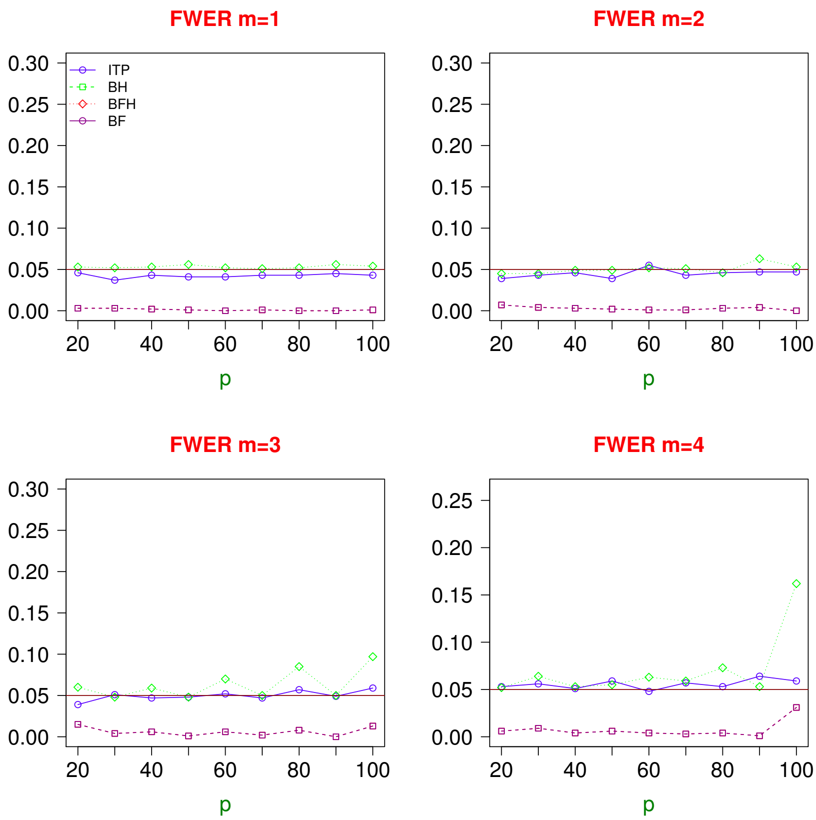

Figure 4.4: Family Wise Error Rate $FWER$ as a function of the number p of B-spline basis elements with orders $m = 1,\ 2,\ 3,\ 4$. The data type is populations of constant and step functions.

4.2.2.2 Rejection Rate False null hypotheses

In Figure 4.5 on page 49 we report the heat maps of the Rejection Rate False null hypotheses ρ. Consider the scenarios with data set const-step. In general ρ_{ITP} assumes maximum values for the orders $m \in \{2,\ 3,\ 4,\ 5\}$ and for high p. In particular for $m \in \{3,\ 4\}$ and for all p, ITP is less conservative than BFH and BF (see Figure A.3 on page 107).

From contour lines of ρ_{ITP} for $m \in \{2,\ 3\}$, we see that the trend of ρ_{ITP} as a function of p is characterized by oscillations which could be due to Smoothing approximations of the step functions near the points h_{inf} and h_{sup}, where the jumps are situated (see the Figure A.3 on page 107). Since using the basis B-spline we obtain localization in space, these approximations can influence in a marked way the p-value univariate estimate. Indeed, in order to obtain the i-th adjusted p-value (associated with the i-th test/coefficient), the maximum is searched in the collection of p-values (concerning sets of consecutive coefficients) whose associated tests have null hypothesis implying $H_0^{(i)}$, where $H_0^{(i)}$ is the null hypothesis of the test with the i-th unadjusted p-value. For details about the ITP p-values adjustement phase refer to 2.2.1.3. For instance, suppose that the step functions are smoothed with rounded up values near h_{inf} and inside $[h_{inf},\ h_{sup}]$ where the null hypothesis is false. The p-values associated with this region can be relatively high. Consequently, due to the nature of the ITP p-values adjustment phase which takes into account the structure of the functional data, this set of p-values might distort the adjustment of the p-values near in space. In scenarios where the noise power σ^2_{noise} is not null, this distortion phenomenon called "edge effect" could be less marked.

With the other procedures by construction, the p-values adjustment phase should not be noticeably affected by overestimates or underestimates of localized sets constituted by univariate p-values. The procedures BH and BFH are sequentially rejective multiple testing procedures characterized by the initial univariate p-values ordering step. Hence, these procedures treat separately each unadjusted p-value without exploiting the B-spline basis ordered structure (in space) as the ITP does. Essentially the same observation holds for the procedure BF with the difference that this procedure operates directly on the significance level with a scaling factor coincident with the number of univariate tests performed; the univariate p-values are adjusted in the same way independently from the properties of the data set.

The ITP reaches its optimum for $m = 4$ and for p relatively high. In these cases, the results of the procedures ITP and BH are very similar (see Figures 4.6 and 4.7). Hence, if a functional datum has discontinuities, for the ITP the best choice would seem the use of the cubic splines with a sufficiently high number of basis elements.

It can be proved that for the Smoothing splines method the optimal splines (minimizer of the Residual Sum of Squares with penalization on the second derivative) are natural (null second and third derivatives at the endpoints) and cubic with knots coincident with the sampling points of the observed data (for details refer to Section 2.3). Therefore, even if we have used the Regression splines method in this simulation, it is reasonable that the cubic splines are optimal in terms of the localized biases that might appear in correspondence of the jump points h_{inf} and h_{sup} (this phenomenon is called "edge effect"). We do not have biases at the endpoints since all the statistical units are constant at the edges and the power of noise σ^2_{noise} is null.

As noted for $FWER$ in scenarios with data set const-step, the Smoothing is critical, particularly for ITP.

It is interesting to observe that the unitary order is optimal for BH and it is exactly the opposite for ITP, BFH and BF (see Figure 4.8 on page 51). BH is on the whole the procedure with the best performances for both cases with data set const-step and const-tricube. ρ_{BH} is increasing in p except for the scenarios with unitary order m. In this case ρ_{BH} is approximately constant. ρ_{BH} gradually decreases by reducing both the order m and the number of basis elements p. We observe the same tendency in scenarios with data sets const-tricube with a slower decay.

The results obtained with BF and BFH are very similar. Furthermore, for these procedures the scenarios with data sets const-step and const-tricube do not present particular differences.

For low p and low m the performances of BF and BFH are similar to the ones of the ITP. From the standard graphics in Figure A.4 on page 108, we note that, for p greater than a certain value roughly equal to half of the maximum possible value $p^{max} = n_{eval} = 100$, we have $\rho_{ITP} \geq \rho_{BFH(BF)}$. A similar situation occurs if the data type is const-tricube although the differences between the procedures ITP, BFH and BF are less evident (see Figure A.12 on page 117).

This latter fact can be explained in terms of the properties of the data set const-tricube. Indeed, the tricube kernel has most part of its mass concentrated in the middle of $[h_{inf},\ h_{sup}]$ (support tricube kernel and region where the null hypothesis is false). Consequently, the false null hypotheses associated with the central part of $[h_{inf},\ h_{sup}]$ are often rejected by all procedures. Therefore, regarding these null hypotheses, it is difficult to observe significant differences among ITP, BF and BFH. Moreover, the kernel tricube has few mass at the endpoints of its support. This property constitutes a disadvantage for the ITP since this procedure tends to maximize the true discovery rate in the middle of $[h_{inf},\ h_{sup}]$, minimizing it at the borders (as already discussed in Pini and Vantini, 2013).

In scenarios with data type const-tricube, for all procedures the performances are generally worse than in scenarios with data type const-step. Moreover, if the data type is const-tricube, the best order m is the unitary one as can been seen in Figure 4.9 on page 51 (all graphics are reported in Figure A.11 on page 116). In particular observe the slope change of ρ: when the order m is low the slope of ρ is roughly null; by increasing m, the curves ρ tend to be increasing in p.

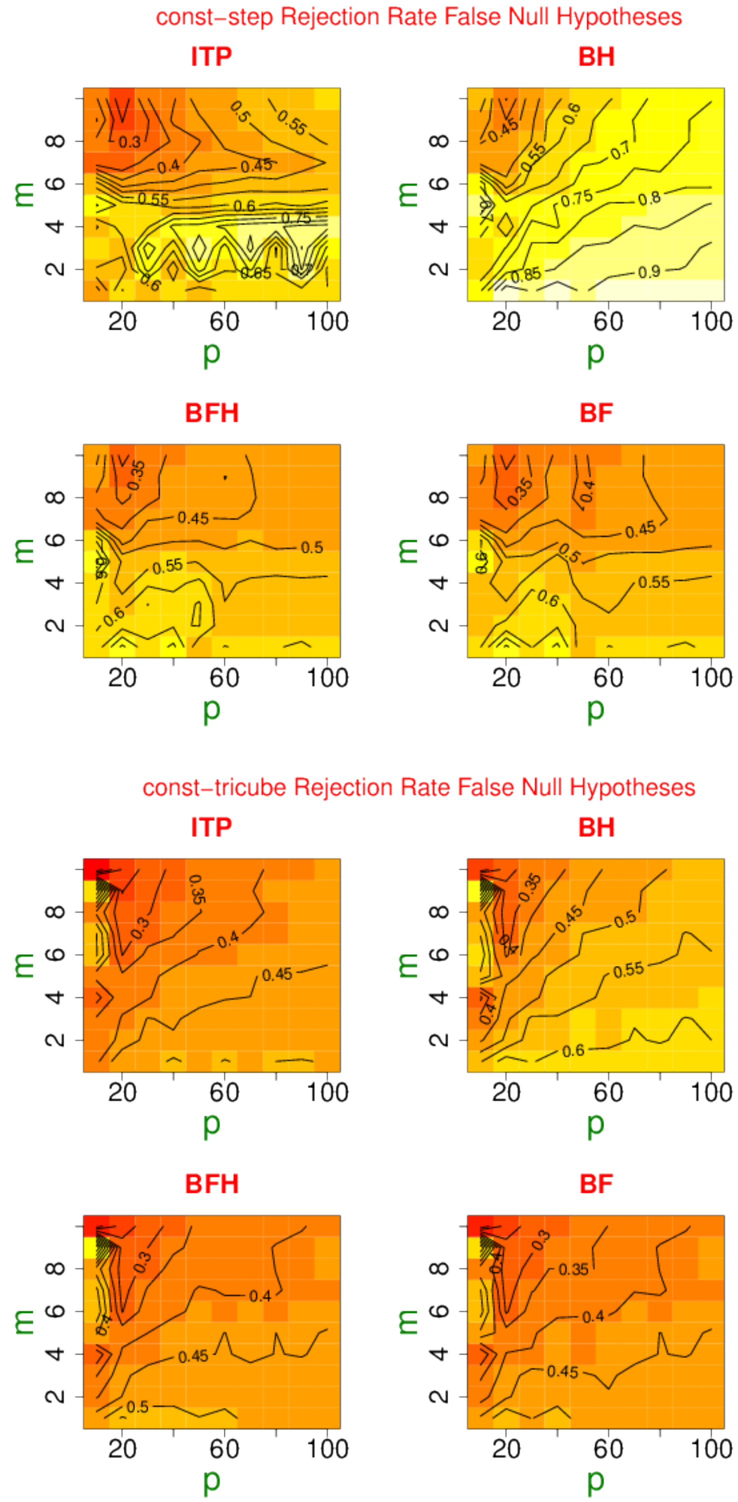

Figure 4.5: Heat maps of the Rejection Rate False null hypotheses ρ as a function of p and m for the procedures Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF) and for the data sets const-step (populations of constant and step functions) and const-tricube (populations of constant and tricube functions). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0.

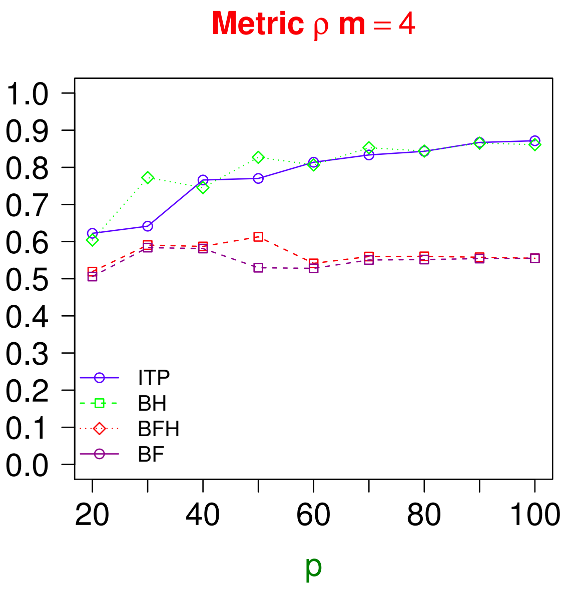

Figure 4.6: Rejection Rate False null hypotheses ρ as a function of the number p of B-spline basis elements with cubic splines. The data type is populations of constant and step functions.

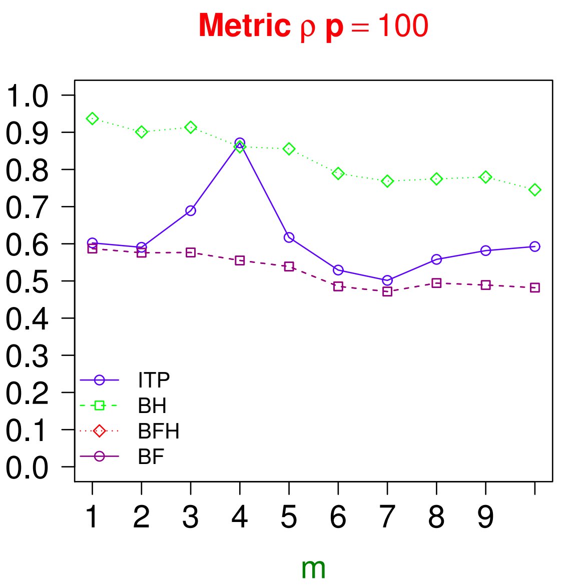

Figure 4.7: Rejection Rate False null hypotheses ρ as a function of the order m B-spline basis with the interpolating spline ($p = 100$). The data type is populations of constant and step functions.

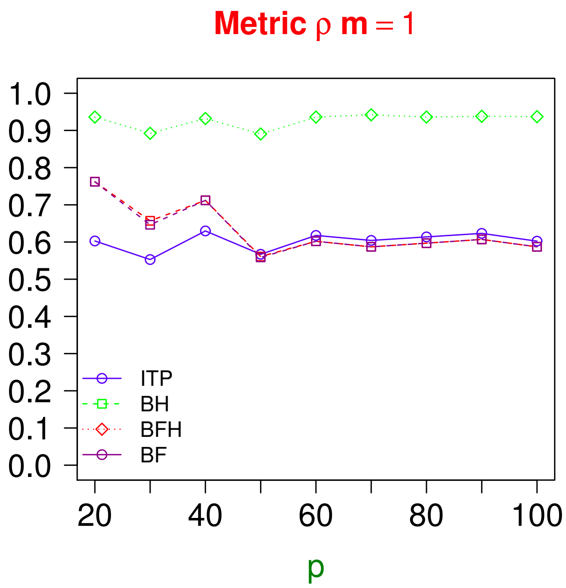

Figure 4.8: Rejection Rate False null hypotheses ρ as a function of the number p of B-spline basis elements with constant splines. The data type is populations of constant and step functions.

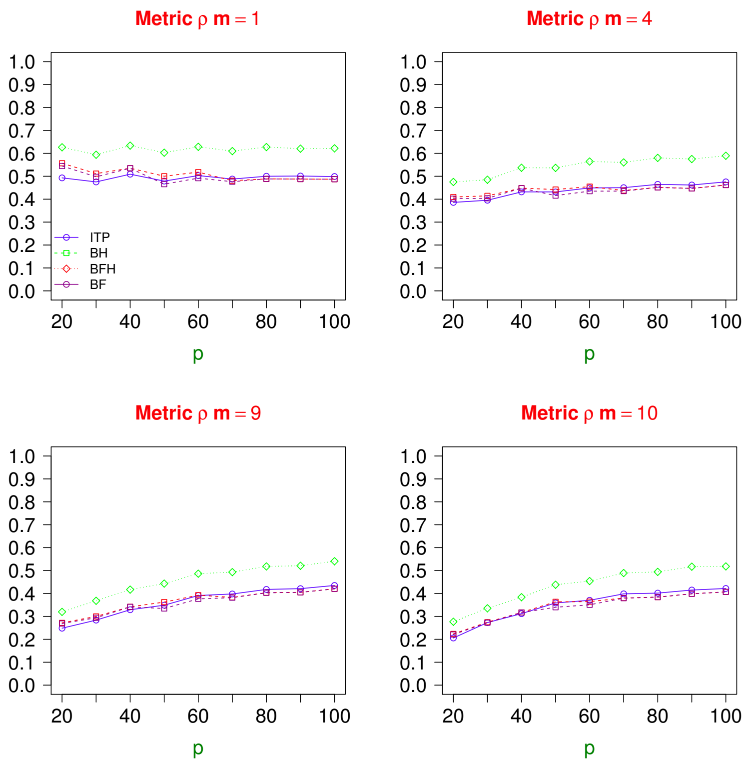

Figure 4.9: Rejection Rate False null hypotheses ρ as a function of the number p of B-spline basis elements with orders $m = 1,\ 4,\ 10$. The data type is populations of constant and tricube functions.

4.2.2.3 Rejection Rate True null hypotheses

In Figure 4.10 on page 53 we report the heat maps of the Rejection Rate True null hypotheses γ. Similarly to the $FWER$, γ_{BF} and γ_{BFH} assume lower values than γ_{ITP} and γ_{BH}. Furthermore, in most cases the approximation $\gamma_{BF} \approx \gamma_{BFH} \approx 0$ is valid for both data sets const-step and const-tricube. $\gamma_{BFH(BF)}$ is positive only for $p = 10$ and for some low values of the order m, and, however, its values are very small.

In scenarios with data set const-step, the function γ_{ITP} is maximized for values of the order m approximately in the set $\{3,\ 4,\ 5\}$ and for relatively high p coherently with the observations pertaining the $FWER$. For really high orders and independently from p, γ_{ITP} is minimized. For very low orders and independently from p, γ_{ITP} assumes intermediate values. Regarding the procedure BH, by augmenting p and by decreasing m, it can be noted that γ_{BH} tends to increase. γ_{BH} is maximized for low p and high m; in the other cases γ_{BH} assumes lower values.

In scenarios with data type const-tricube, we have that γ_{ITP} is maximized in a greater set of orders $m \in \{2,\ 3,\ 4,\ 5\}$ for every number of basis elements and for really high m and for any reasonable values of p ($p > n_1 = n_2 = n_{units} = 30$). The trend of γ_{BH} is similar to the one of γ_{ITP} except for the fact that the set of orders where γ_{BH} is maximized is constituted by more elements (roughly $m \in \{1, \ldots,\ 6\}$). In this set we also observe less variability of γ_{BH}, excluding few peaks.

However, from the heat maps of γ, it is difficult to detect particular tendencies and to observe differences among the tested procedures due to the low values assumed by this metric. In addition, unlike the $FWER$, this metric is not theoretically conserved by the inspected methods. It is only possible to detect the regions where γ is optimized in order to have essentially a set of rules of thumb about how to choose in principle the Smoothing parameters m and p for each procedure and for each data type.

In scenarios with data set const-step, from the standard graphics in Figure A.5 on page 109, we can observe that except for the quadratic and cubic splines ($m = 3, 4$), the following ordering holds

$$\gamma_{BF} \approx \gamma_{BFH} < \gamma_{ITP} < \gamma_{BH}$$

and, in contrast, for $m \in \{3,\ 4\}$ the performances of ITP and BH are equivalent. It is worth to mention that at the same time ρ_{ITP} is coherently maximized for $m = 3, 4$ and p relatively high.

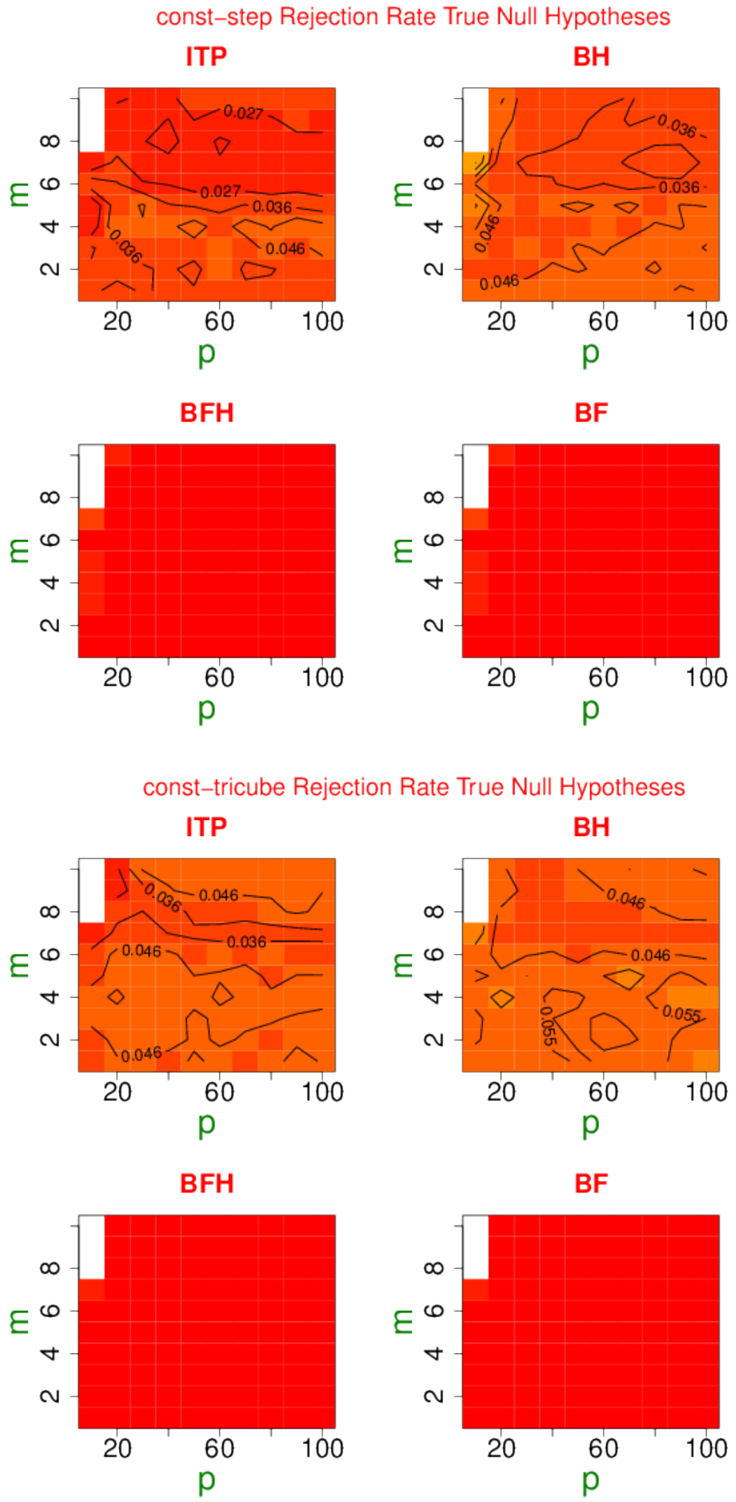

Figure 4.10: Heat maps of the Rejection Rate True null hypotheses γ as a function of p and m for the procedures Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF) and for the data sets const-step (populations of constant and step functions) and const-tricube (populations of constant and tricube functions). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0. The squares white opaque are associated with not available numbers, corresponding to the cases when all null hypotheses are false; in these scenarios γ is not defined.

4.2.2.4 Power

In Figure 4.11 on page 55 we report the heat maps of the Power π. Considering the scenarios with data set const-step, we observe that the BH globally provides the best performances.

The order $m = 1$ is suboptimal for ITP, BF and BFH. By increasing m in the set $\{1,\ 2,\ 3,\ 4\}$, the power improves for all procedures. Furthermore, with the cubic splines π_{BH} essentially assumes unitary values (see Figure 4.12 on page 56).

If the data type is const-tricube, the Power does not provide interesting information about the tested procedures. Indeed, the curves π assume very high values in all cases. The causes are multiple. The tricube kernel mass is mostly concentrated in the middle of its support $[h_{inf},\ h_{sup}]$, reaching higher mean values with respect to the ones of data set const-step. Therefore, it is reasonable that the Power can be not particularly informative since it is likely that in many instances at least one false null hypothesis is rejected by all procedures. Specifically in the simulation setting chosen and in general by construction, the Rejection Rate False null hypotheses ρ measures in a more precise way the rate of true discoveries.

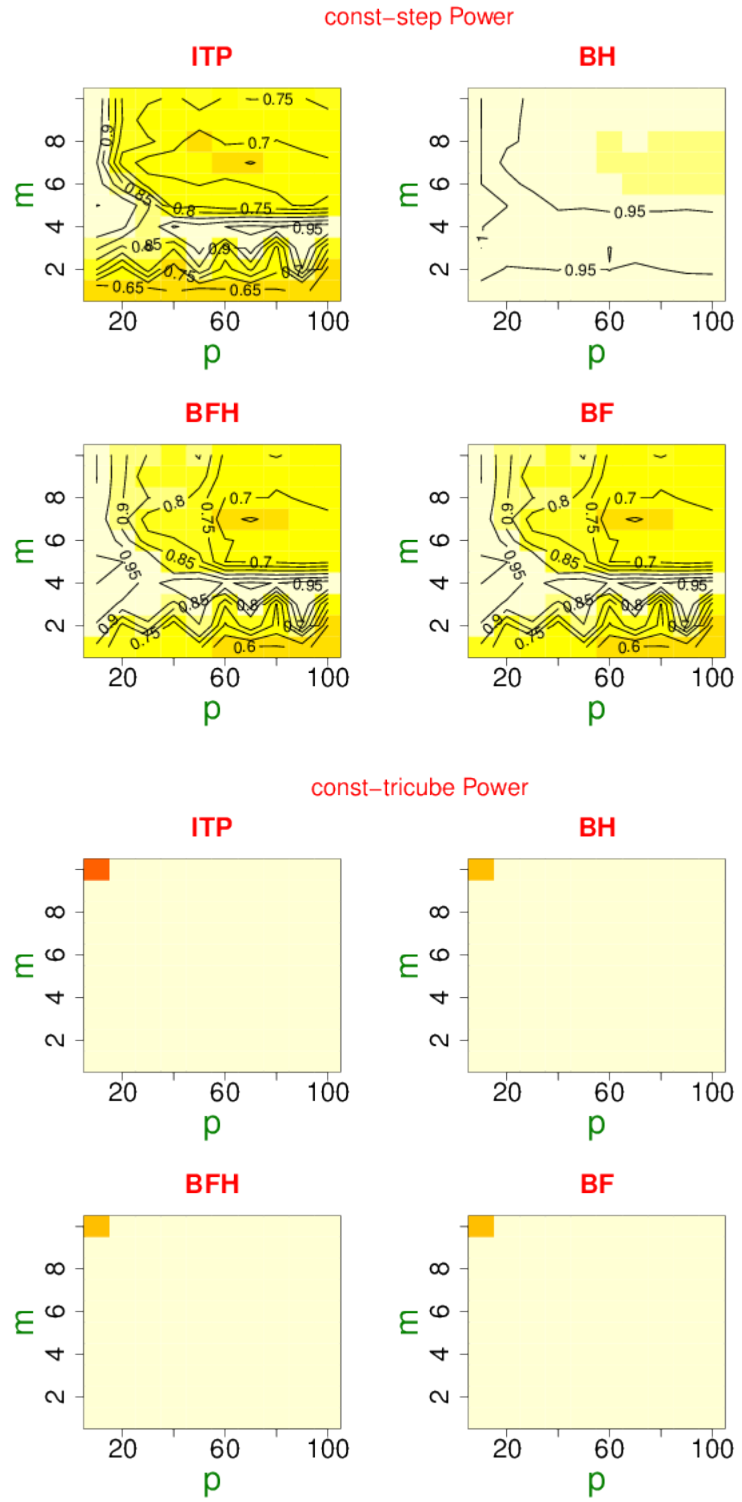

Figure 4.11: Heat maps of the Power π as a function of p and m for the procedures Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF) and for the data sets const-step (populations of constant and step functions) and const-tricube (populations of constant and tricube functions). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0.

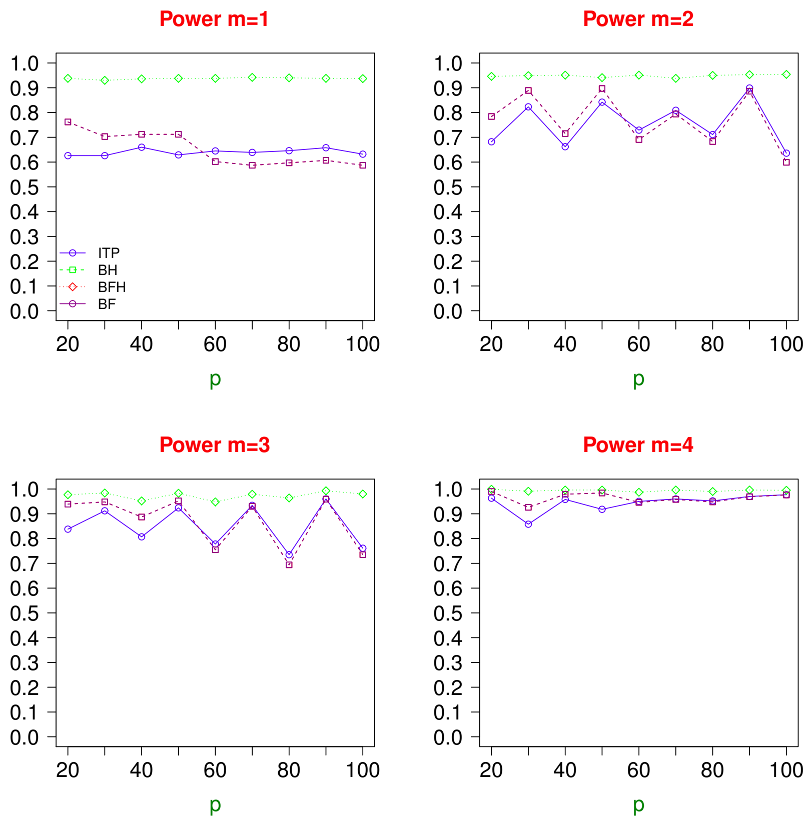

Figure 4.12: Power π as a function of the number p of B-spline basis elements with orders $m = 1,\ 2,\ 3,\ 4$. The data type is populations of constant and step functions.

In all scenarios we have observed that the performances of BFH are better than expected. Indeed, also for high p, BFH is not highly conservative. This result could be a consequence of the fact that, by construction of the univariate permutation tests, the set of attainable p-values estimated with the Conditional Monte Carlo algorithm is discrete (the Conditional Monte Carlo algorithm is described in 2.1). In particular, the probability that a p-value is estimated as 0 is greater than zero. If a p-value is 0, the adjustment of the p-value itself is ineffective. The same applies to the procedure BF.

It is worth to observe that for the procedures BH and BFH, the effective number of tests is less than p. In fact, being both populations constant for $x \notin [h_{inf},\ h_{sup}]$, the p-values associated with the basis elements with null intersection between their supports and $[h_{inf},\ h_{sup}]$ are roughly constant. Consequently since the first step for these procedures is the ordering of the univariate p-values, in practice every set of constant or very similar p-values is considerable as a single p-value.

In conclusion, it is possible to note that for the ITP it is better to choose p relatively high. In particular, this is true also for high m in this simulation. If few basis elements and a low order are chosen, the procedures BF and BFH constitute an equivalent alternative to ITP. For that reason, if the ITP is used, it could be convenient to choose the number of basis elements as high as possible. Moreover, we have observed that for functional data with discontinuities the best choice for the ITP would seem the use of the cubic splines with a sufficiently high number of basis elements. Finally in the scenarios with data set const-tricube, it is suggested to keep low the order and it is not necessary to choose a high number of basis elements.

4.3 Smoothing Splines parameter variable

4.3.1 Introduction

The Smoothing B-splines are studied in the following simulations. We consider also the Smoothing B-splines since they constitute an important tool usable for a representation of the derivatives with specific constraints expressed in terms of a penalty function. We investigate here how the smoothing parameter λ affects the performances of the different procedures.

The objective function chosen here is the Residual Sum of Squares RSS with penalization on the second derivative (curvature) given by:

$$RSS\left(f,\ \lambda\right) = \sum_{i=1}^{N} \left\{y_i - f\left(x_i\right)\right\}^2 + \lambda \int \left\{f''\left(t\right)\right\}^2 dt \tag{4.1}$$

It can be proved that the minimizer f_{opt} of (4.1) is given by the natural cubic splines with knots coincident with the evaluation points of the observed data. For details refer to Section 2.3. For exploratory reasons an hypothesis of this optimization result can be on purpose violated. For instance, in simulations with non-null σ_{noise} the B-splines are not natural.

Firstly, two simulations with number of independent data sets generated n_{data} equal to 1000 and null standard deviation of the additive Normal noise σ_{noise} have been performed. In the first simulation, the set of knots is different from the grid associated with the observed data while they coincide in the second simulation. Secondly, since in most cases the Smoothing analysis is conceived for the reduction of outer errors (measurement errors), we have considered also scenarios with σ_{noise} positive and variable where n_{data} is equal to 500 for each tested value of σ_{noise}.

4.3.2 First simulation with $\sigma_{noise} = 0$

Before the Smoothing, the observed datum is interpolated with natural splines. In a first moment, the number of samples in output after interpolation has been chosen equal to $2 \cdot n_{eval}$, where n_{eval} is the number of evaluations for each statistical unit uniformly distributed in the domain $\mathcal{D} = [0,\ 2]$. Then, the Smoothing is performed using the basis B-spline. The number of basis elements p is coincident with the number of samples at disposal $2 \cdot n_{eval}$ (Smoothing splines methods) and m is set to 4 (cubic splines). Therefore, the knots are located in the grid $\{\tilde{x}_i\}_{i=1}^{N=2 \cdot n_{eval}}$ composed by uniformly distributed values in [0, 2]. This grid is different from the original one ($\{x_i\}_{i=1}^{N=n_{eval}}$ composed by uniformly distributed values in [0, 2]). One assumption of the result mentioned in 4.3.1 about the minimization of the cost functional RSS is violated: the knots are placed in the starting evaluation points $\{x_i\}_{i=1}^{N=n_{eval}}$. The other hypotheses are satisfied: the splines are natural and cubic. Since in the explored scenarios the standard deviation of noise σ_{noise} is zero and the statistical units are constant at the endpoints, the interpolation preprocessing does not involve issues.

The parameters for the data set const-step are:

- $h_{inf}^{CS} = 0.5$
- $h_{sup}^{CS} = 1.5$
- Fixed effect (height of the step with support $[h_{inf},\ h_{sup}]$): $v_{CS} = 0.15$

The parameters for the data set const-tricube are:

- $h_{inf}^{CT} = 0.5$
- $h_{sup}^{CT} = 1.5$
- Maximum value of the symmetric tricube kernel with support $[h_{inf},\ h_{sup}]$: $v_{CT} = 0.22$

h_{inf}^{CS}, h_{inf}^{CT}, h_{sup}^{CS}, h_{sup}^{CT}, v_{CS} and v_{CT} are chosen in the same manner in Regression and Smoothing B-splines frameworks. The remaining parameters are:

- $n_1 = n_2 = n_{units} = 30$
- Number evaluations uniformly distributed in the common domain $\mathcal{D} = [0,\ 2]$: $n_{eval} = 50$

The values of the smoothing parameter are:

$$\lambda \in \left[10^{-10},\ 1\right]\ with\ 100\ uniformly\ distributed\ samples$$

Family Wise Error Rate

The results in terms of the Family Wise Error Rate $FWER$ are reported in Figure 4.13. The procedure BH does not control the $FWER$. Moreover, $FWER_{BH}$ assumes larger values in scenarios with data set const-step.

It seems that also the procedures ITP, BFH and BF do not control the $FWER$. Denote with $\sigma_{estimate}$ the upper bound of the estimate variability $\frac{1}{\sqrt{n_{data}}}$ obtained from the Normal approximation of the confidence interval for the Binomial proportion with $\alpha = 0.05$ (for details refer to Section 3.3). Actually, the values of λ such that

$$FWER_s - \alpha > \frac{1}{\sqrt{n_{data}}} = \sigma_{estimate} \approx 3\%, \ s \in \{ITP, \ BF, \ BFH\}$$

are very high and meaningless. A functional data reconstruction with λ too big distort the estimate of the p-values and consequently also the computed metrics. This effect is more marked if the regularity of data is low as in scenarios with data set const-step which is discontinuous. Hence, we can confirm that, for reasonable values of the smoothing parameter λ, the procedures ITP, BF and BFH control the $FWER$. It has been chosen such a wide set of values for the smoothing parameter λ for an exploratory purpose.

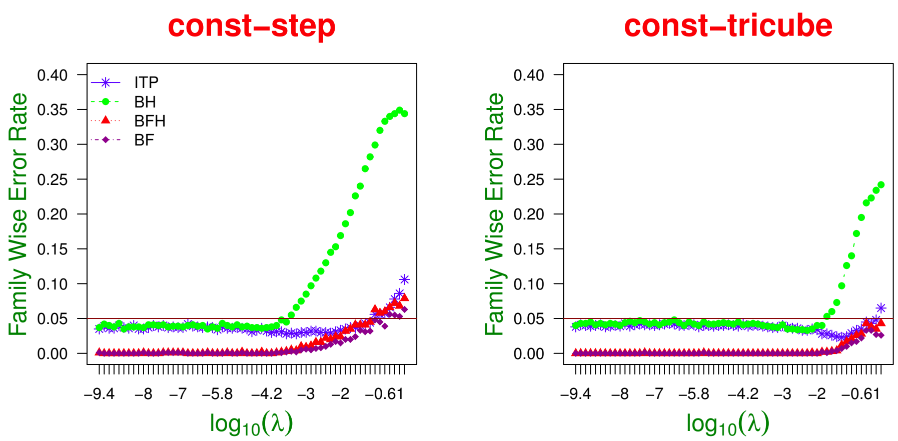

Figure 4.13: Family Wise Error Rate $FWER$ in simulations with $\sigma_{noise} = 0$ and $n_{data} = 1000$ doubling n_{eval} with interpolation.

Rejection Rate False null hypotheses

The Rejection Rate False null hypotheses ρ is reported in Figure 4.14. Independently from the data set type, it holds

$$\rho_{BF} \approx \rho_{BFH}, \ \rho_{BH} > \rho_s, \ s \in \{ITP, \ BF, \ BFH\}$$

In scenarios with data set const-step, we observe:

- For λ approximately lower than 10^{-5}, we have the following ordering:

$$\rho_{BF} \approx \rho_{BFH} < \rho_{ITP} < \rho_{BH}$$

  The slope of the curves ρ is around 0 except for ρ_{ITP} in the last part of the interval $\left(10^{-10},\ 10^{-5}\right)$ where ρ_{ITP} itself starts to decrease. Therefore, we have low variability in this range of the smoothing parameter λ.

- For λ approximately greater than 10^{-5}, the functions ρ are decreasing with parabolic trend. The results of the procedures ITP and BFH (BF) are similar.

If the type of data set is const-tricube, we point out:

- The region $\left[10^{-15},\ \tilde{\lambda}\right]$ ($\tilde{\lambda} \approx 10^{-3}$), where the curves ρ have slope around 0 (also ITP differently from the scenarios with data set const-step), is longer with respect to the scenarios with data set const-step. There is a global maximum of ρ_{BH} near $\hat{\lambda} = 10^{-2}$. For $\lambda \geq \hat{\lambda}$, the functions ρ are decreasing.

- For every value of the smoothing parameter λ, the performances obtained with the procedures ITP and BFH (BF) are very similar.

These results are coherent with the nature of the data sets tested. With the discontinuous data set const-step, it is reasonable that strong penalties (high λ) demote the performances, particularly for the ITP. This effect is less evident for data set const-tricube since this kind of data is more regular (the tricube kernel belongs to $\mathcal{C}^9$). In particular, ρ_{BH} is maximized for a set of high values of the smoothing parameter λ.

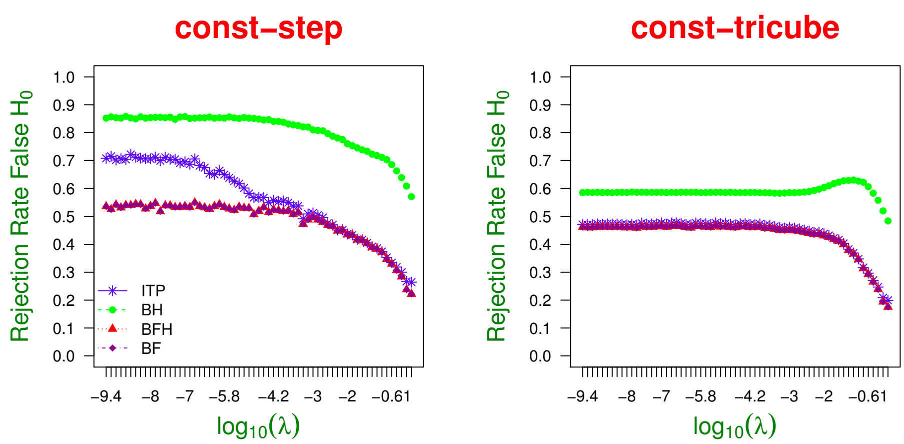

Figure 4.14: Rejection Rate False null hypotheses ρ in simulations with $\sigma_{noise} = 0$ and $n_{data} = 1000$ doubling n_{eval} with interpolation

Rejection Rate True null hypotheses

The results in terms of the Rejection Rate True null hypotheses γ are reported in Figure 4.15. Consider the scenarios with data set const-step. For λ approximately lower than $\tilde{\lambda} = 10^{-3}$, we observe the ordering

$$\gamma_{BF} \approx \gamma_{BFH} < \gamma_{ITP} < \gamma_{BH}$$

and the functions γ are approximately constant.

If $\lambda \geq \tilde{\lambda}$, we note:

1. The curves γ_s with $s \in \{BF,\ BFH,\ BH\}$ are increasing. In particular, this trend characterizes the function γ_{BH}.

2. We have the following ordering:

$$\gamma_{BF} \leq \gamma_{BFH} < \gamma_{ITP} \lll \gamma_{BH}$$

3. The function γ_{ITP} attains its minimum values roughly for a set of high values of λ; moreover, in this set the performances of ITP are very similar to the ones obtained by the procedures BF and BFH.

If the data type is const-tricube, we observe:

1. It exists an interval $\left(10^{-10},\ \tilde{\lambda}\right)$ in which the curves γ have zero slope. This interval is more extended with respect to the one identified in the framework with data set const-step; this observation is valid also for ρ and π and it is coherent with the regularity of the tested types of data set.

2. Similarly to the cases with data type const-step, we have the ordering

$$\gamma_{BF} \leq \gamma_{BFH} < \gamma_{ITP} \lll \gamma_{BH}$$

3. For $\lambda \geq \tilde{\lambda}$, the results are essentially comparable to the ones obtained with data set const-step except for few differences: γ_{BH} assumes smaller values with data set const-tricube and the set of minimum values of γ_{ITP} is more pronounced.

The trend of $FWER$ is similar to the one characterizing the Rejection Rate True null hypotheses γ without the set of minimum values obtained for high λ, particularly in scenarios with data type const-step.

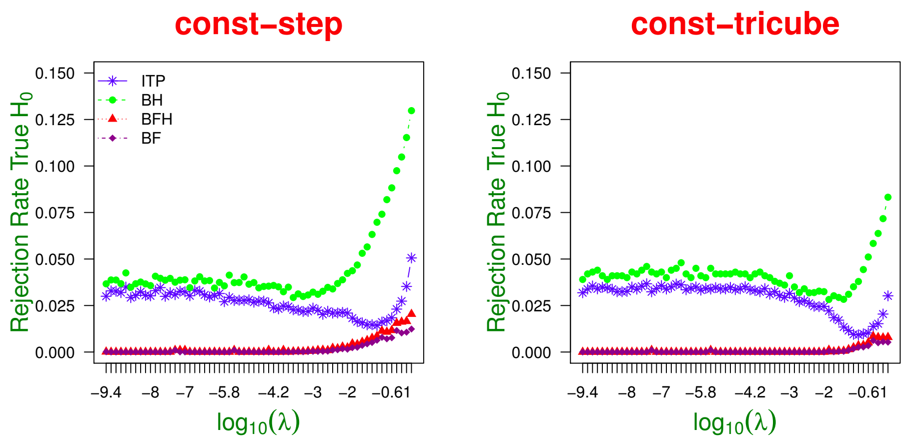

Figure 4.15: Rejection Rate True null hypotheses γ in simulations with $\sigma_{noise} = 0$ and $n_{data} = 1000$ doubling n_{eval} with interpolation

Power

The Power π is reported in Figure 4.16. Firstly, we note that the results obtained with the procedures BFH and BF are coincident.

For the data set const-step, except for really large values of λ, we have the ordering:

$$\pi_{BF} \approx \pi_{BFH} \lesssim \pi_{ITP} < \pi_{BH}$$

For very high values of λ, we have

$$\pi_{BF} \approx \pi_{BFH} \approx \pi_{ITP}$$

Moreover, it exists a common $\hat{\lambda} \approx 10^{-2}$ where π_{BF}, π_{BFH}, π_{ITP} are locally maximized.

In scenarios with data set const-tricube, we observe that the curves π have approximately unitary values until $\lambda \approx 10^{-2}$. Due to the nature of the data set const-tricube, the Power π could be an imprecise index since the mass of the tricube kernel is mainly concentrated in the middle of the interval $[h_{inf},\ h_{sup}]$ where H_0 is false. Therefore, it is plausible that at least one false null hypotheses is rejected by all procedures. For $\lambda \geq 10^{-2}$, the functions are decreasing and in particular

$$\pi_{BH} > \pi_s \ with \ s \in \{ITP,\ BF,\ BFH\}$$

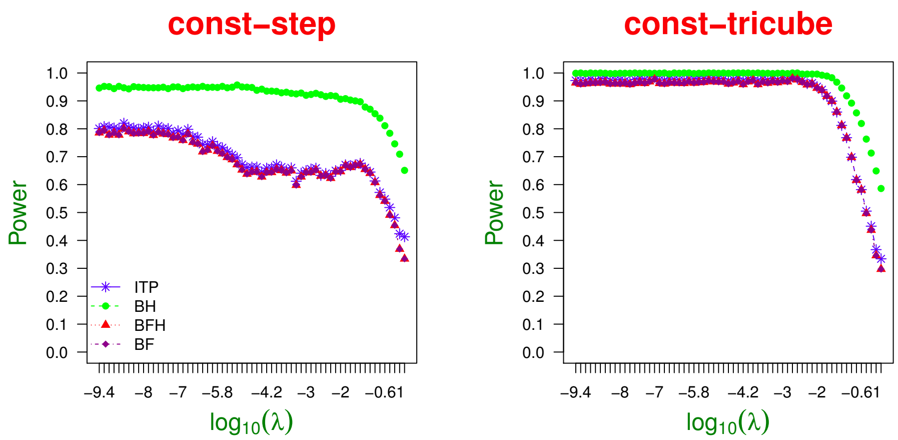

Figure 4.16: Power π in simulations with $\sigma_{noise} = 0$ and $n_{data} = 1000$ doubling n_{eval} with interpolation

4.3.3 Second simulation with $\sigma_{noise} = 0$

For comparison and check, we performed a simulation with the same parameters as the previous one without doubling the number of evaluations n_{eval} (n_{eval} is directly set to 100). In this way, the grid of interpolated data with natural splines is coincident with the grid associated with the original data. Therefore, the hypotheses of the cost functional minimization result reported in Section 2.3 are satisfied.

Family Wise Error Rate

The results in terms of the Family Wise Error Rate $FWER$ are reported in Figure 4.17. Consider the scenarios with data set const-step. Contrary to the preceding simulation, we observe:

- The procedure BH does not control the $FWER$ also for low values of the smoothing parameter λ; this trend is not distinctive of γ even if the control for this metric is not guaranteed.

- Taking into account the variability of the metric estimate, it can be seen that the ITP, BFH and BF control the $FWER$ for almost every λ. In particular roughly for $\lambda \leq 10^{-2}$ we have

$$FWER_s(\lambda) \leq \alpha + \frac{1}{\sqrt{n_{data}}} = \alpha + \sigma_{estimate}, \; s \in \{ITP, \; BFH, \; BF\}$$

- For all procedures, we note a set of minimum values of the $FWER$ approximately in the interval $\left(10^{-7}, \; 10^{-3}\right)$. This trend is less marked for the ITP, in particular with respect to BH.

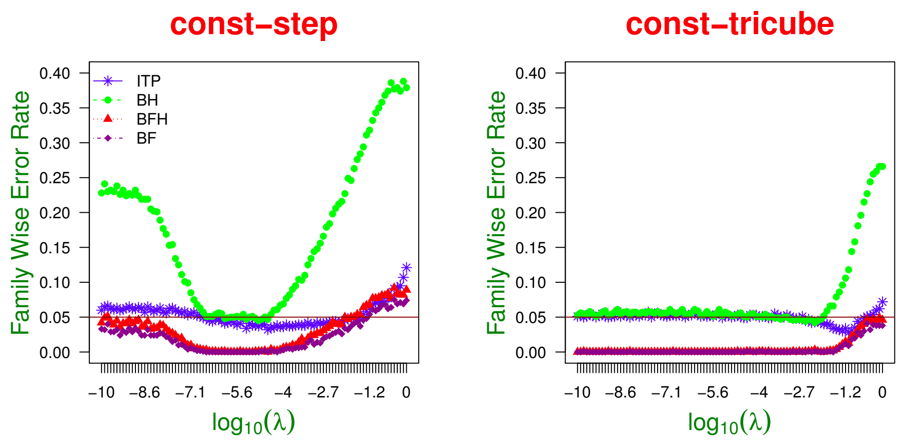

Figure 4.17: Family Wise Error Rate $FWER$ in simulations with $\sigma_{noise} = 0$ and $n_{data} = 1000$ without doubling n_{eval} with interpolation

Rejection Rate False null hypotheses, Rejection Rate True null hypotheses and Power

The Rejection Rate False null hypotheses ρ, Rejection Rate True null hypotheses γ and Power π are reported in the Figure 4.18.

In scenarios with data type const-step, an important difference with respect to the previous simulation is that, in terms of ρ and for low values of λ, the results of the procedures ITP and BH are analogous. The same applies for γ and π; in particular for the Power π the statement includes also the procedures BFH and BF. In particular, for low λ, the power is unitary for all procedures. This outcome is reasonable since the data set const-step is not regular, being discontinuous. Moreover, the standard deviation of noise σ_{noise} is zero.

The performance improvement of the ITP could be explained referring to the result mentioned in 4.3.1 about the minimization of the cost functional (4.1). In the previous simulation, the interpolation has been performed doubling the number of initial evaluations n_{eval}. Hence, the grid for Smoothing was $\{\tilde{x}_i\}_{i=1}^{N=2\cdot n_{eval}}$ composed by uniformly distributed values in [0, 2]. Consequently one assumption of the previously mentioned result is violated: the knots are placed in the starting evaluation points $\{x_i\}_{i=1}^{N=n_{eval}}$ composed by uniformly distributed values in [0, 2]. The ITP seems the most influenced procedure by this hypothesis violation among the tested procedures. This may be due to the fact that the ITP explores the ordered structure in space of the B-spline basis components. Each adjusted p-value is determined by the multivariate p-values with null hypothesis of the associated test implying the null hypothesis related to the adjusted p-value itself. It should be remembered that the tests are expressed in terms of the basis coefficients. We have an univariate test for each coefficient. Moreover, since the B-spline basis expansion is used we have localization in space and, hence, there could be a strong relationship between the functional reconstruction and the trend of the adjusted p-values. In some cases, we have observed a parabolic shape of the ITP adjusted p-values with few of them below the significance level α among the p-values associated with false null hypotheses. Thefore, in general for the ITP the Smoothing would seem crucial.

Regarding the cases with data set const-tricube, there are not important differences compared to the previous simulation.

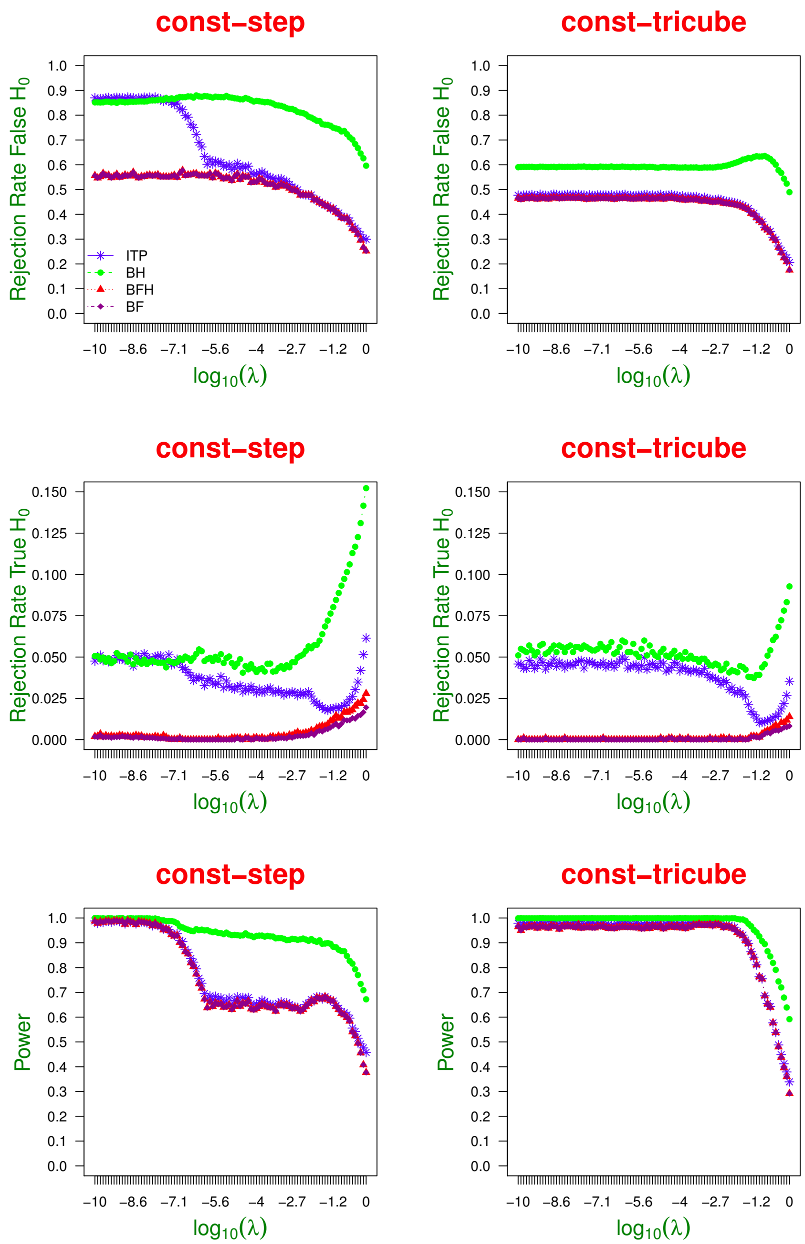

Figure 4.18: Rejection Rate False null hypotheses ρ, Rejection Rate True null hypotheses γ and Power π in simulations with $\sigma_{noise} = 0$ and $n_{data} = 1000$ without doubling n_{eval} with interpolation

4.3.4 Simulations with $\sigma_{noise} > 0$ variable

The problem of the biased functional estimates obtained with Smoothing at the endpoints of the domain (phenomenon called "edge effect") is generally solved using the natural and cubic splines with knots coincidents with the sampling points of the observed data assuming that the cost function is (4.1). However, since we are interested in the comparison among the tested procedures and the statistical units are constant at the endpoints of the domain, we have not used the natural splines for an exploratory purpose. Hence, the Smoothing is directly performed on the discrete data with order $m = 4$ (cubic splines) and the knots are placed in the original grid.

The values of the smoothing parameter λ and of the standard deviation of noise σ_{noise} are:

$$\lambda \in \left[10^{-7},\ 1\right]\ with\, 25\, uniformly\, distributed\, samples;\ \sigma \in \{0.01,\ 0.03,\ 0.06,\ 0.09\}$$

The other parameters are fixed in the same way as in the previous simulation.

4.3.4.1 Family Wise Error Rate

The results in terms of the Family Wise Error Rate $FWER$ are reported in Figure 4.19. Firstly independently from the data type, we observe that the procedure BH does not control the $FWER$ for high λ. It is the same for low λ when σ_{noise} is sufficiently high ($\sigma_{noise} \geq 0.06$) which is reasonable since the Smoothing is fundamentally conceived for the reduction of measurement errors. As noted in the previous simulations, there exists an interval where $FWER_{BH} \leq \alpha$ and, in accordance with the regularity of the data tested, it is more extended on the right side if the data type is const-tricube.

Taking into account the variability of the estimate, it can be stated that ITP, BF and BFH control the $FWER$ in scenarios with data set const-step (excluding high values of λ) and const-tricube (for all λ).

In general, the $FWER$ is higher in scenarios with data set const-step, in particular for $FWER_{BH}$.

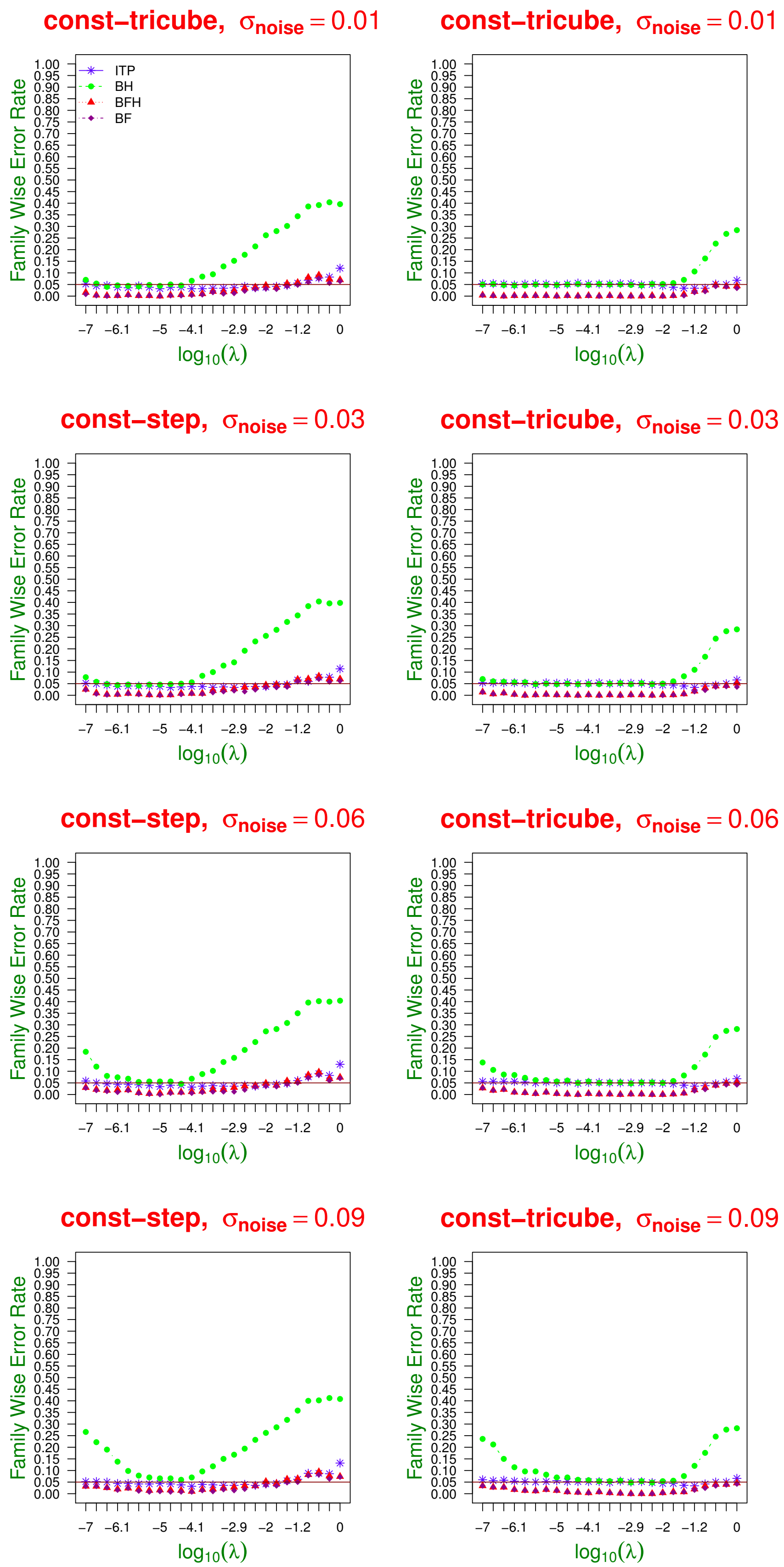

Figure 4.19: Family Wise Error Rate $FWER$ in simulations with $\sigma_{noise} > 0$ variable and $n_{data} = 500$ for each tested value of σ_{noise}

4.3.4.2 Rejection Rate False null hypotheses

The Rejection Rate False null hypotheses ρ is reported in Figure 4.20. Consider the cases with data set const-step. BH is the procedure with the best performances. By increasing σ_{noise}, the curve ρ_{BH} tends to be characterized by a parabolic trend with maximum values for intermediate λ. On the contrary, ρ_{ITP} is always maximized for low λ and it is decreasing in λ. By increasing σ_{noise}, ρ_{ITP} decays in λ more slowly. In general, the following ordering holds:

$$\rho_{BF} \approx \rho_{BFH} \leq \rho_{ITP} \leq \rho_{BH}$$

Differently from the previous simulation, for low λ the performances of ITP and BH are not equivalent, probably because the splines are not natural. Furthermore, we note the same trend of ρ_{BH} for the curves ρ_{BFH} and ρ_{BF}.

As far as concerns the scenarios with data set const-tricube, in the first place we observe that the results of the procedures ITP, BF and BFH are similar; the unique cases in which some differences can be observed are for low λ and high σ_{noise}. The tendency of the functions ρ_{ITP}, ρ_{BFH}, ρ_{BF} is characterized by a first region where the slope is approximately zero (except for low λ and high σ_{noise}) followed by a last part (high λ and hence strong penalties) where it can be observed a decay. The same basically holds for the procedure BH except that there exists a set of maximum values of ρ_{BH} for very high λ independently from the power of noise σ^2_{noise}.

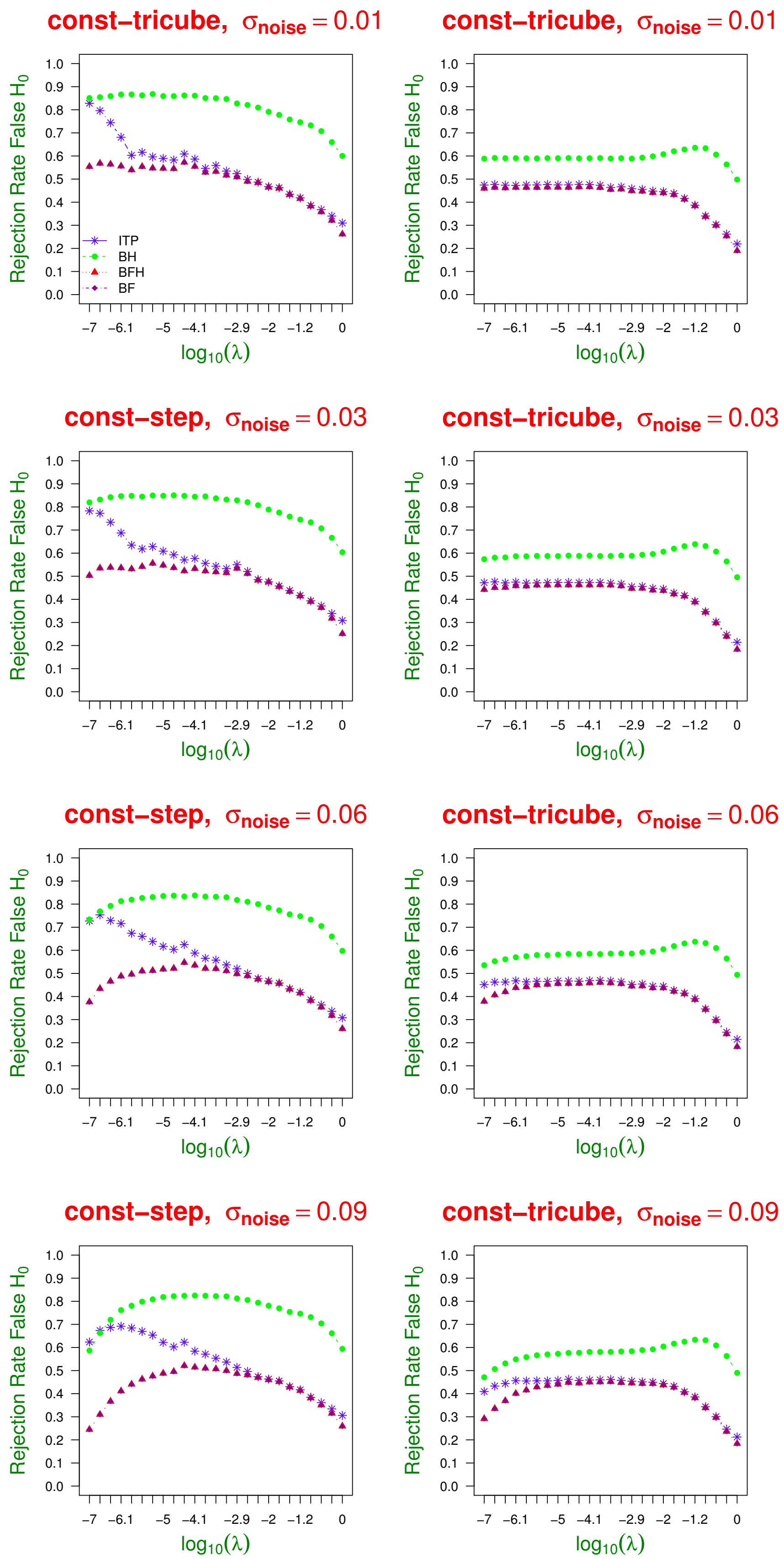

Figure 4.20: Rejection Rate False null hypotheses ρ in simulations with $\sigma_{noise} > 0$ variable and $n_{data} = 500$ for each tested value of σ_{noise}

4.3.4.3 Rejection Rate True null hypotheses

The results in terms of the Rejection Rate True null hypotheses γ are reported in Figure 4.21. We observe that:

- By increasing σ_{noise} and for small and intermediate λ, the curves γ decrease. This result is reasonable since the univariate p-values generally increase by augmenting the standard deviation of noise σ_{noise}.
- The following ordering holds:

$$\gamma_{BF} \approx \gamma_{BFH} \leq \gamma_{ITP} \leq \gamma_{BH}$$

- γ_{ITP} is minimized for high λ with performances comparable to the BFH and BF procedures. The set of the minimum values is more pronounced in scenarios with data set const-tricube.
- The functions γ_{BH}, γ_{BFH}, γ_{BF} are increasing and they tend to assume larger values if the data type is const-step. In particular, the maximum values assumed by the function γ_{BH} are markedly larger in scenarios with data set const-step. Except for really high λ, the curves γ_{BFH}, γ_{BF} assume null values.

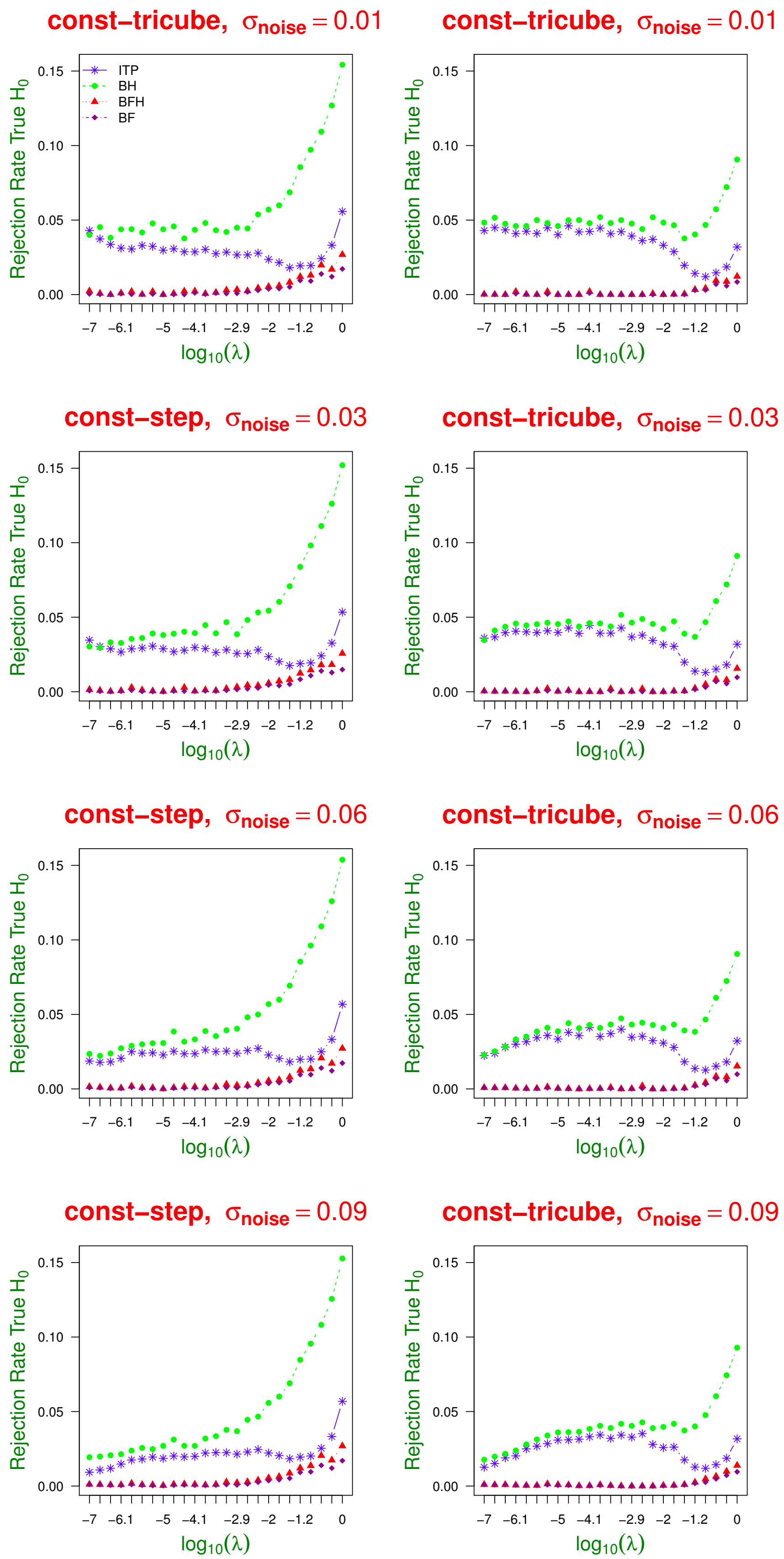

Figure 4.21: Rejection Rate True null hypotheses γ in simulations with $\sigma_{noise} > 0$ variable and $n_{data} = 500$ for each tested value of σ_{noise}

4.3.4.4 Power

The Power π is reported in figure Figure 4.22. Despite the data type, we have $\pi_{BF} \approx \pi_{BFH} \approx \pi_{ITP}$.

In scenarios with data set const-step, the functions π_{ITP}, π_{BFH}, π_{BF} are decreasing and this is less evident increasing σ_{noise}. On the contrary, π_{BH} is decreasing only for really high λ. In the other cases BH is constant independently from σ_{noise}. Furthermore, BH is the most powerful procedure.

For the data type const-tricube, the functions π are constant with unitary value except for very high λ. For very high λ (independently from the power of noise σ^2_{noise}), the functions π are decreasing with $\pi_{BH} > \pi_s$ with $s \in \{ITP,\ BF,\ BFH\}$.

Varying the standard deviation of noise σ_{noise}, does not lead to particular trend differences.

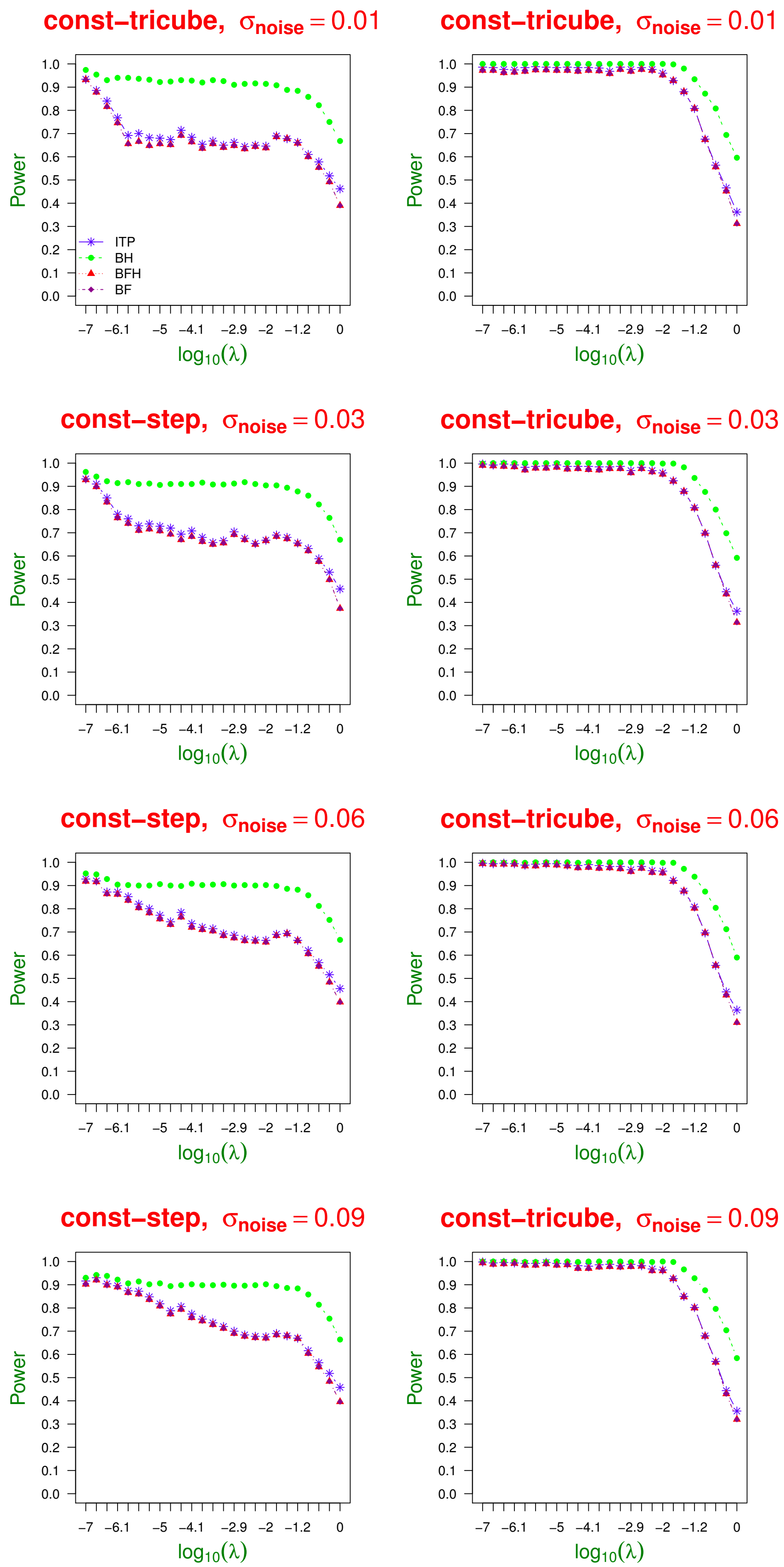

Figure 4.22: Power π in simulations with $\sigma_{noise} > 0$ variable and $n_{data} = 500$ for each tested value of σ_{noise}.

In general it is possible to note that, in terms of ρ, the ITP is the procedure most prone to degradation of the performances due to violations of the assumptions of the cost functional minimization result briefly mentioned in 4.3.1. In detail, the violated assumptions are the placing of the knots in the evaluation points of the observed data (first simulation with $\sigma_{noise} = 0$) and the use of the natural splines (simulations with $\sigma_{noise} > 0$ variable). This result should not be interpreted as a negative fact. Instead, it is an useful result regarding the use of the ITP. If we know a theoretical result such as the one reported in Section 2.3 for Smoothing splines method or guidelines tailored for a particular Smoothing method, for ITP it would seem strongly recommended the application of these results.

On the other hand, in the simulation with parameters m and p (respectively the order and the number of the B-spline basis elements), we have observed that the tuning of the parameters m and p is critical. In fact, for functional data with discontinuities the best choice in terms ρ for the ITP would seem the use of the cubic splines with a sufficiently high number of basis elements. This result was not easily predictable a priori because we expected that the unitary order would have been the best one for the functional reconstruction of a step, as it has been for the BH method.

Finally based on the obtained results, the learned message is that for the ITP it is highly recommended to apply, if available, theoretical results or best practices for the used Smoothing method. Alternatively, it is possible to use the purely nonparametric inferential procedure Interval-wise test (Pini and Vantini, 2015a) which does not require the use of a basis expansion (for details refer to Section 1.2).

Simulations with Data parameters variable

In the second part of this simulation study we investigate the performances of ITP depending on the Data parameters. In Section 4.4, we focus on parameters number of evaluations n_{eval} and standard deviation of the additive Normal σ_{noise}. In Section 4.5, we focus on parameter number of statistical units n_{units}.

4.4 Number of evaluations n_{eval} and standard deviation of noise σ_{noise} variable

4.4.1 Introduction

In the previous simulations, we have analyzed the performances of the tested procedures by varying the Smoothing parameters which are arbitrarily tunable in applications. However, we are also interested in exploring how the performances depend on Data parameters, even if they could not be controllable in real situations.

In this simulation, the variable parameters are the number of functional evaluations n_{eval} and the standard deviation of the additive Normal noise σ_{noise}.

For each value of n_{eval}, n_{data} independent data sets are generated from the const-step and const-tricube models, where for each statistical unit n_{eval} evaluations are uniformly distributed in the domain $\mathcal{D} = [0,\ 2]$. Then, for each value of σ_{noise} and for each data set, a local Normal noise with zero mean and standard deviation σ_{noise} is added independently to each point.

4.4.1.1 Values of the parameters

The common parameters in both scenarios with data set const-step and const-tricube are:

- Number functional evaluations: $n_{eval} = \{50,\ 100, \ldots,\ 500\}$.
- Standard deviation additive Normal noise: $\sigma_{noise} = \{0,\ 0.022, \ldots,\ 0.2\}$.
- Number statistical units: $n_{units} = n_1 = n_2 = 30$.
- B-spline order: $m = 4$.
- Number basis elements: $p = 50$.
- Number data sets generated: $n_{data} = 1000$.

In scenarios with data set const-step, the parameters are:

- $h_{inf}^{CS} = 0.5$
- $h_{sup}^{CS} = 1.5$
- $v_{CS} = 0.15$

In scenarios with data set const-tricube, the parameters are:

- $h_{inf}^{CT} = 0.5$
- $h_{sup}^{CT} = 1.5$
- $v_{CT} = 0.22$

In the following we report the heat maps that, if need be, are alternated with a selection of the standard graphics (i.e., metric as a function of σ_{noise} with fixed n_{eval} or vice versa) and general comments about the observed trends. For completeness, all the standard graphics divided according to the type of data (const-step and const-tricube) are reported in Appendix B.

4.4.2 Results

The simulation results are presented by means of suitable heat maps which represent the values of a metric by varying both parameters σ_{noise} and n_{eval}. In this way a direct comparison between the scenarios with data sets const-step and const-tricube can be achieved.

For each data type (const-step or const-tricube) and metric (Family Wise Error Rate $FWER$, Rejection Rate False null hypotheses ρ, Rejection Rate True null hypotheses γ, Power π), we have a plot with four subgraphs. Each subgraph is the representation through an heat map of a metric for one of the tested procedures (Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF)). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0. When needed, contour lines are added to ease the comparison between the different graphics.

4.4.2.1 Family Wise Error Rate

In Figure 4.23 on page 78 we report the heat maps of the Family Wise Error Rate $FWER$. Independently from the data type, the functions $FWER_{BF}$ and $FWER_{BFH}$ globally assume lower values than $FWER_{ITP}$ and $FWER_{BH}$. Moreover, the results provides by BFH and BF are basically coincident and $FWER_{BFH(BF)}$ assumes higher values in noisy scenarios.

In general for each procedure, the trend of the $FWER$ is independent from the data type.

BF, BFH and ITP control the $FWER$ in all cases (net of the upper bound of the estimate variability $\sigma_{estimate}(\alpha) = \sigma_{estimate} = \frac{1}{\sqrt{n_{data}}} \approx 0.032$; for details refer to Section 3.3). The control of the $FWER$ provided by the ITP is on intervals. This kind of control of the $FWER$ is intermediate between the strong (typical of the BF and BFH) and the weak ones. By construction of the ITP adjustment p-values phase (recycled version of the family composed by all possible interval-wise tests explained in Pini and Vantini, 2013), the union of intervals $[0,\ h_{inf}) \cup (h_{sup}, 2]$, where the null hypothesis is true, can be considered as a unique interval.

As expected, BH does not control the $FWER$ (it controls the False Discovery Rate, implying only a weak control of the $FWER$). The trend of $FWER_{BH}$ as a function of σ_{noise} is parabolic, and by augmenting n_{eval}, the points where $FWER_{BH}$ is maximized increase.

From Figures 4.24, 4.25, we note that it exists a value of the standard deviation of noise $\tilde{\sigma}_{noise}$ above which the $FWER$ of the ITP is lower than the one of BF and BFH procedures. By augmenting n_{eval}, it can be observed that $\tilde{\sigma}_{noise}$ increases. $FWER_{ITP}$ is non-increasing in σ_{noise} and, on the contrary, $FWER_{BFH}$, $FWER_{BH}$ are non-decreasing in σ_{noise}.

In the simulation study in Pini and Vantini, 2013 regarding the Component Wise Probability of Rejection (i.e, the Component Wise Error Rate for components with true null hypothesis and the Component Wise Power for components with false null hypothesis), it has been observed that, if it is identified a significant interval with the procedure BH, no distinction is generally made among the different components. Instead, the ITP maximizes the true discoveries in the middle of a significant interval and it presents a true discovery rate decays at the borders of the significant interval itself. This property is due to the capability of the ITP to explore the ordered structure of the basis expansion used.

We noted that, by augmenting σ_{noise}, the univariate p-values and consequently also the adjusted p-values increase. Due to the ability of the ITP to exploit the ordered structure in space of the B-spline basis expansion, this trend of the p-values can involve the decreasing of the $FWER_{ITP}$ when the power of noise σ^2_{noise} is augmented. On the contrary for BH, for comparable values of n_{eval} and σ_{noise}, there could be difficulties in its ability to detect uniformly a significant interval and, consequently, we observe a parabolic trend of the curve $FWER_{BH}$ as a function of σ_{noise} with maximum value increasing in n_{eval}.

When σ_{noise} is high, it is better to have few evaluations n_{eval} for BH in terms of $FWER$. In contrast, by increasing σ_{noise}, the curve $FWER_{ITP}$ is non-increasing for every n_{eval}.

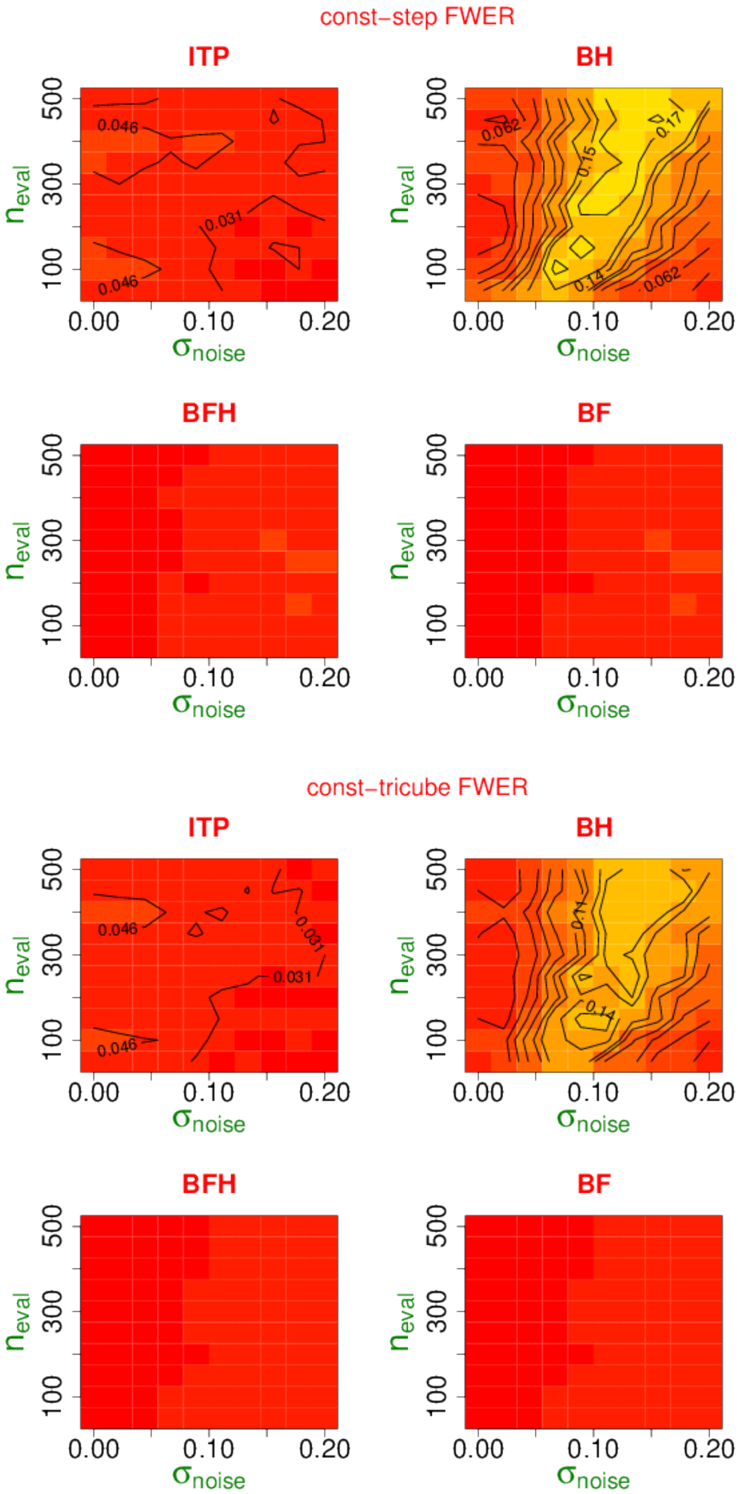

Figure 4.23: Heat maps of the Family Wise Error Rate $FWER$ as a function of σ_{noise} and n_{eval} for the procedures Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF) and for the data sets const-step (populations of constant and step functions) and const-tricube (populations of constant and tricube functions). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0.

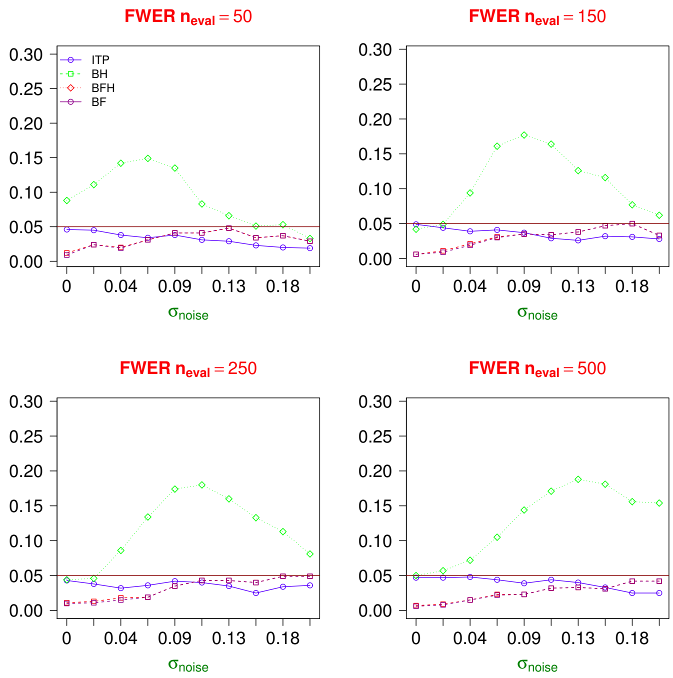

Figure 4.24: Family Wise Error Rate $FWER$ as a function of the standard deviation of noise σ_{noise} with number of evaluations $n_{eval} = 50,\ 150,\ 250,\ 500$. The data type is populations of constant and step functions.

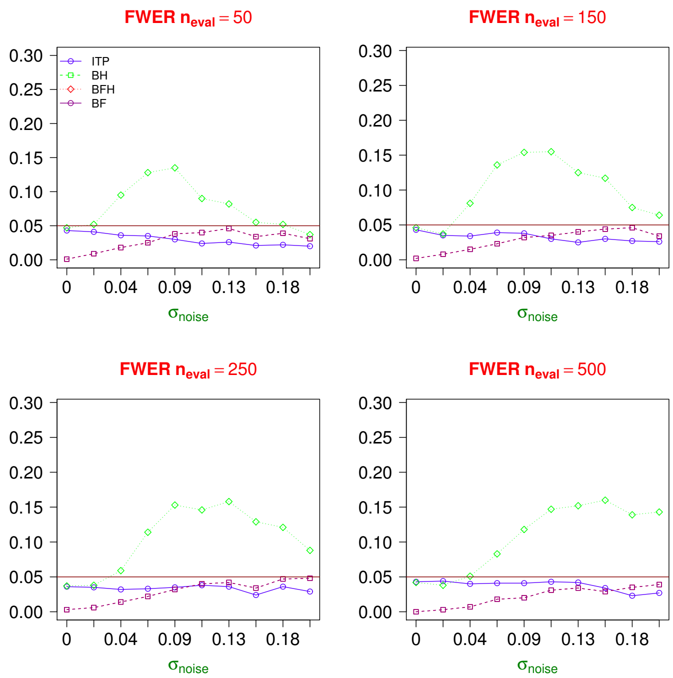

Figure 4.25: Family Wise Error Rate $FWER$ as a function of the standard deviation of noise σ_{noise} with number of evaluations $n_{eval} = 50,\ 150,\ 250,\ 500$. The data type is populations of constant and tricube functions.

4.4.2.2 Rejection Rate False null hypotheses

In Figure 4.26 on page 81 we report the heat maps of the Rejection Rate False null hypotheses ρ. In general for each procedure and for each data type, ρ is characterized by constant values for low σ_{noise}, and by a decay fo intermediate and high σ_{noise}. By increasing n_{eval} the region presenting constant values of ρ tends to be enlarged.

Consider the scenarios with data set const-step. From the countour lines (more clearly from Figure 4.27 on page 82), it is possible to observe the following ordering:

$$\rho_{BF} < \rho_{BFH} < \rho_{ITP} \leq \rho_{BH}$$

For very high values of σ_{noise}, the performances of all procedures are comparable. For high σ_{noise}, we have $\rho_{ITP} \approx \rho_{BH} > \rho_s,\ s \in \{BFH,\ BF\}$. By increasing n_{eval}, the interval of the form $[0,\ \tilde{\sigma}_{noise}]$, where the difference $\rho_{BH} - \rho_{ITP}$ is positive, tends to expand.

The ITP has good performances for low n_{eval} and low σ_{noise}. In particular for $\sigma_{noise} \in (0,\ 0.04)$, the point where ρ_{ITP} is maximized is $n_{eval} = 100$ (see Figure B.4 on page 126). In these scenarios, we have that the performances of ITP and BH are equivalent and in general the slope of the curves ρ are roughly null. Instead by augmenting σ_{noise}, the curves ρ tend to be more and more increasing in n_{eval}. For high σ_{noise}, the procedures with the best performances are ITP and BH, and the results obtained with these procedures are equivalent.

In scenarios with data set const-tricube, for low σ_{noise}, differently from the scenarios with data set const-step, BFH provides higher values of ρ than ITP. However, for high σ_{noise} the procedures with the best performances are still ITP and BH. The results obtained with these procedures are comparable, in particular for very high σ_{noise}. Finally, in general the BH procedure provides the best results.

The quantity $\frac{n_{eval}^2}{\sigma_{noise}^2}$ can be interpreted as the ratio between the Power of Signal and the Power of Noise (Sound to Noise Ratio SNR). When σ_{noise} is low, the SNR is high also for low n_{eval} and all procedures are characterized by good performances. On the contrary, starting from a certain value of σ_{noise}, the SNR is increasing in n_{eval} and all procedures provide better performances by augmenting the number of evaluations n_{eval}. Interpreting the results in terms of SNR is an equivalent way to look at the slope changes of the curves ρ by varying the parameters of interest n_{eval} and σ_{noise}; for low values of σ_{noise}, independently from n_{eval}, the slope of the curves ρ is roughly zero and ρ assumes high values; for intermediate and high σ_{noise}, the curves ρ are non-decreasing in n_{eval} and they generally assume lower values.

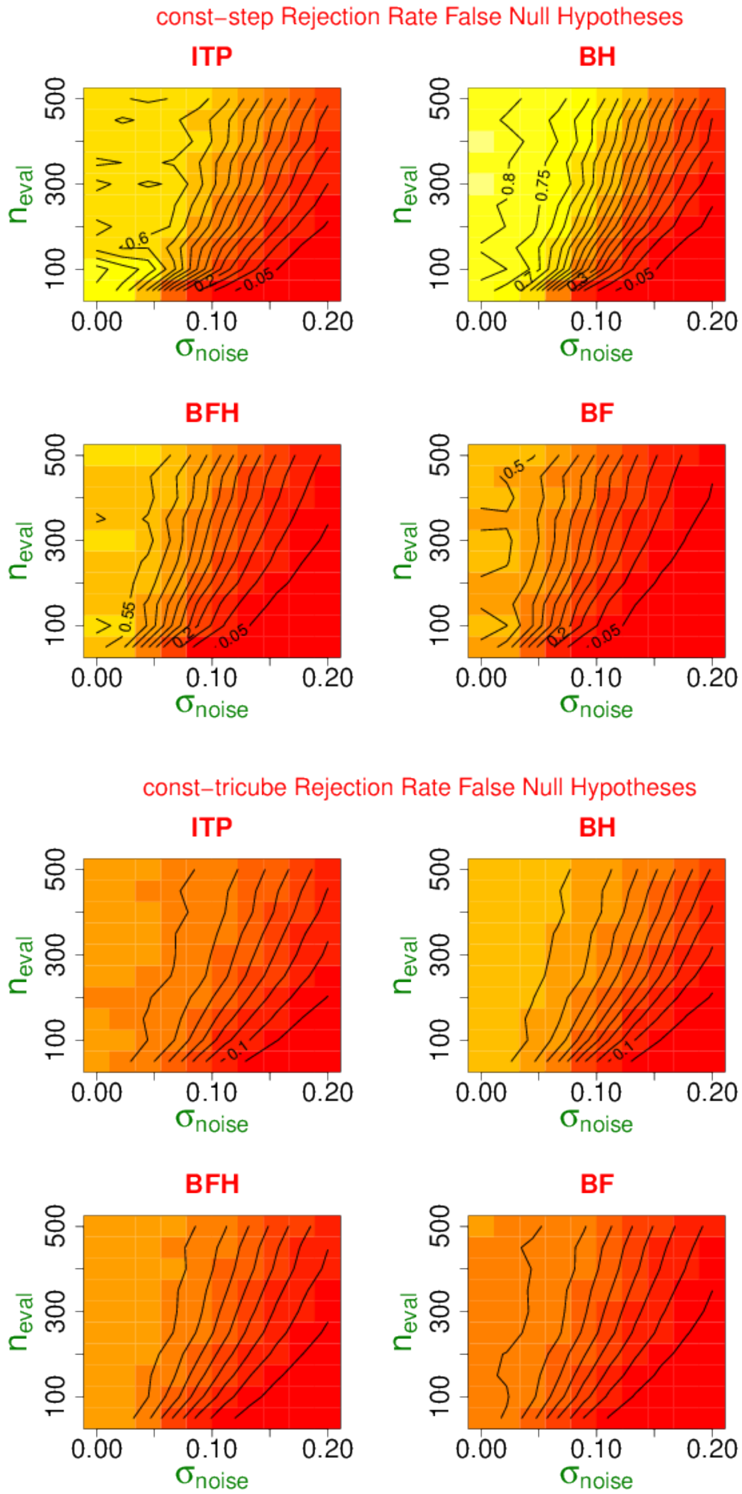

Figure 4.26: Heat maps of the Rejection Rate False null hypotheses ρ as a function of σ_{noise} and n_{eval} for the procedures Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF) and for the data sets const-step (populations of constant and step functions) and const-tricube (populations of constant and tricube functions). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0.

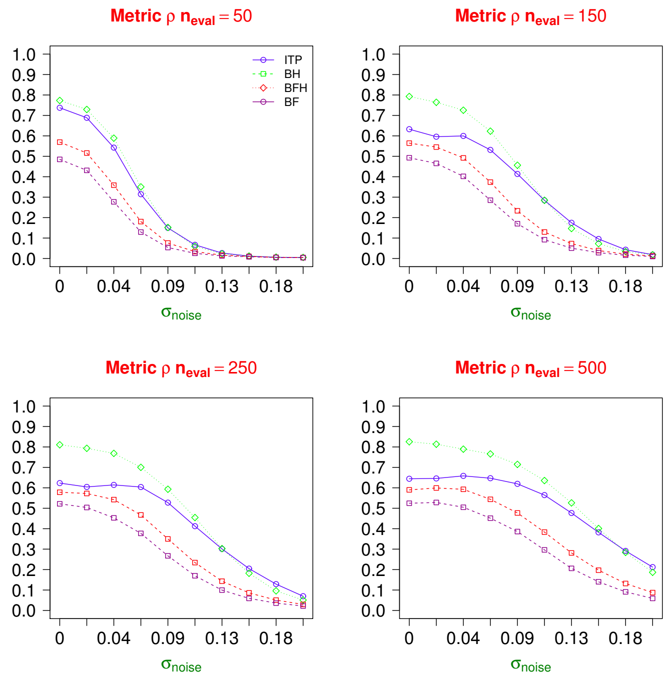

Figure 4.27: Rejection Rate False null hypotheses ρ as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and step functions.

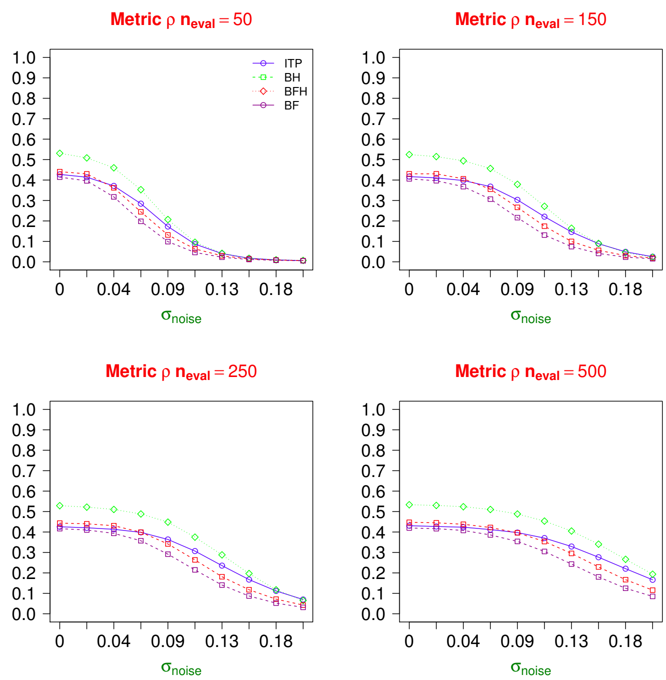

Figure 4.28: Rejection Rate False null hypotheses ρ as a function of the standard deviation of noise σ_{noise} with number of evaluations $n_{eval} = 50,\ 150,\ 250,\ 500$. The data type is populations of constant and tricube functions.

4.4.2.3 Rejection Rate True null hypotheses

In Figure 4.29 on page 84 we report the heat maps of the Rejection Rate True null hypotheses γ. For both data types const-step and const-tricube, γ_{BFH} and γ_{BF} assume very low values around 0 and, for every n_{eval}, γ_{ITP} and γ_{BH} are decreasing in σ_{noise}. The decay of the curves γ_{ITP} and γ_{BH} observable by increasing σ_{noise} is slower by augmenting n_{eval}, trend coherent with the high values of ρ obtained for low and intermediate σ_{noise} and for sufficiently high n_{eval}. Moreover, the functions γ_{ITP} and γ_{BH} decrease by diminishing both n_{eval} and σ_{noise}.

In scenarios with data set const-step from Figure 4.30 on page 85 it can be observed that it exists a value of the standard deviation of noise $\tilde{\sigma}_{noise}$ dependent from n_{eval} beyond which the ITP outperforms BH. By increasing n_{eval}, in the same way as ρ, $\tilde{\sigma}_{noise}$ tends to increase. For all graphics see Figure B.6 on page 128.

We observe the same tendencies in scenarios with data set const-tricube. From the graphics in Figure 4.31 on page 85 by increasing the power of noise σ^2_{noise}, the improvement - absolute and with respect to BH - of the ITP performances is discernible. In particular, for high σ_{noise}, we have $\gamma_{ITP} \approx \gamma_{BF} \approx \gamma_{BFH} \approx 0$. This trend is interesting for ITP since, in the same cases, the performances in terms of ρ of the ITP and BH are comparable. By increasing n_{eval}, γ_{ITP} as a function of σ_{noise} decreases more slowly. For all graphics refer to Figure B.14 on page 137.

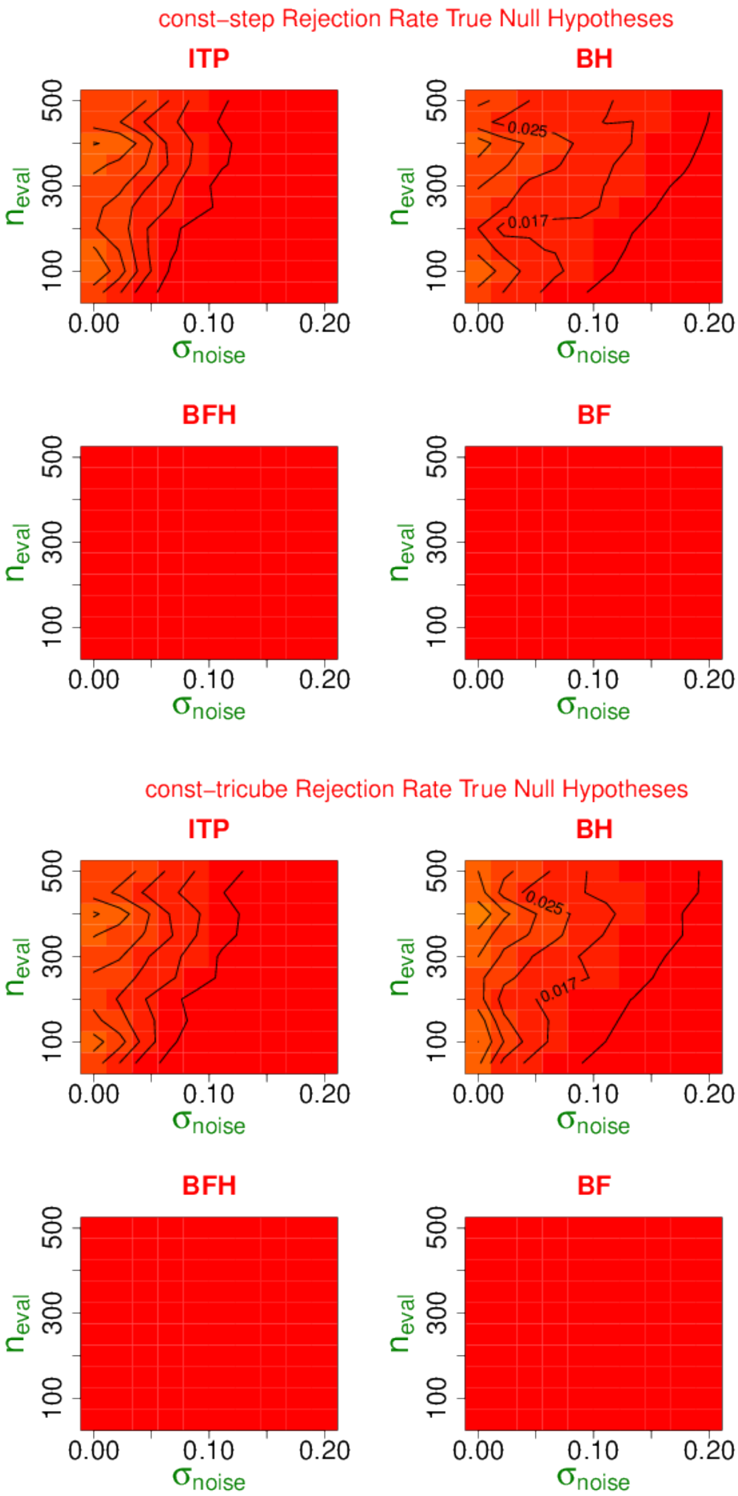

Figure 4.29: Heat maps of the Rejection Rate True null hypotheses γ as a function of σ_{noise} and n_{eval} for the procedures Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF) and for the data sets const-step (populations of constant and step functions) and const-tricube (populations of constant and tricube functions). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0.

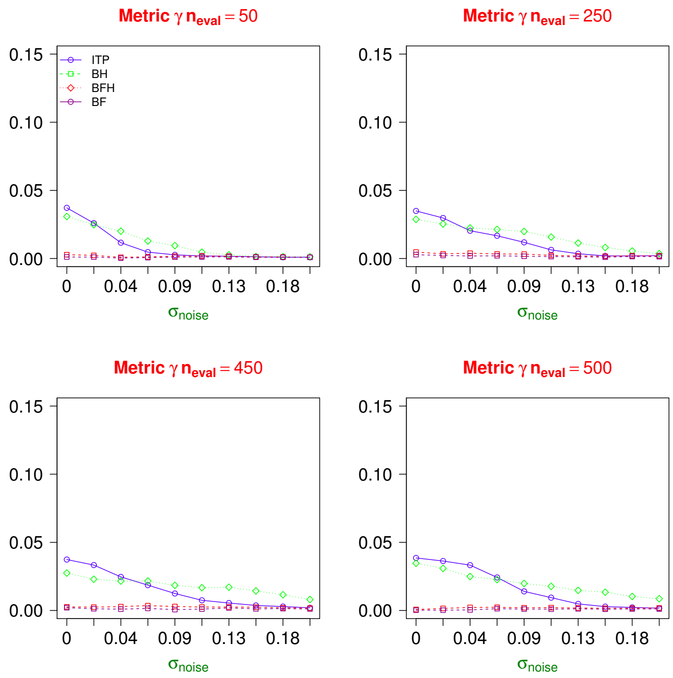

Figure 4.30: Rejection Rate True null hypotheses γ as a function of the standard deviation of noise σ_{noise} with number of evaluations $n_{eval} = 50,\ 250,\ 450,\ 500$. The data type is populations of constant and step functions.

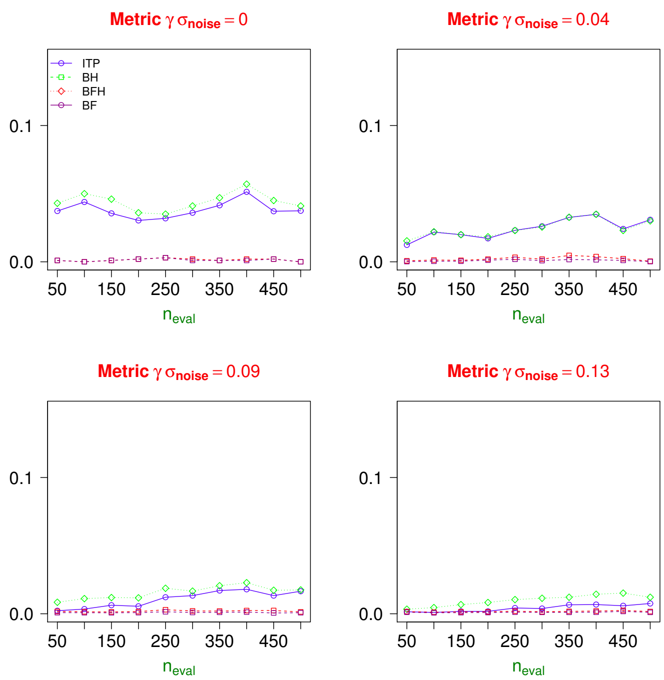

Figure 4.31: Rejection Rate True null hypotheses γ as a function of the number of evaluations n_{eval} with fixed standard deviation of noise $\sigma_{noise} = 0, 0.04, 0.09, 0.013$. The data type is populations of constant and tricube functions.

4.4.2.4 Power

In Figure 4.32 on page 87 we report the heat maps of the Power π. Firstly, we observe that the general tendencies of the curves π_{BF} and π_{BFH} are very similar for both data types const-step and const-tricube.

Consider the scenarios with data set const-step. For $n_{eval} \in (50,\ 100)$, from the countour lines we observe the ordering $\pi_{ITP} < \pi_{BF} \approx \pi_{BFH} \approx \pi_{BH}$. The trend of curves π, similarly to ρ, is characterized by constant values for low σ_{noise} (region 1), and by a decay for intermediate and high σ_{noise} (region 2).

For $n_{eval} > 100$ except for π_{BH} whose trend remains unchanged, the functions π_{ITP}, π_{BFH} and π_{BF} are parabolas with maximum values obtained for intermediate σ_{noise}. Additionally, it holds $\pi_{ITP} < \pi_{BF} \approx \pi_{BFH} \leq \pi_{BH}$ and, in particular, $\pi_{BFH(BF)} < \pi_{BH}$ for low σ_{noise}.

From the simulation results concerning the parameters order m and number p of the B-spline basis elements (consult Section 4.2), we expect that the positive difference $\pi_{BFH(BF)} - \pi_{ITP}$ vanishes for p sufficiently high. Indeed in the simulation with parameters m and p, we observed that for the ITP it is better to choose the number of basis elements p sufficiently high and the cubic splines ($m = 4$). In this simulation we have used the cubic splines and p has been always chosen equal to 50 which is rather low considering that the minimum and the maximum tested values of n_{eval} are $n_{eval}^{min} = 50$ and $n_{eval}^{max} = 500$.

Finally, the results of all procedures are comparable when σ_{noise} is very high. When n_{eval} is very high, the curves π are parabola with small curvature; hence, π have slope roughly null. For all graphics see Figures B.7, B.8.

In scenarios with data type const-tricube, independently from n_{eval} and for low σ_{noise}, the values of the Power are essentially unitary. The trend of π for all procedures is characterized by constant values for low σ_{noise} (region 1) and by a decay for intermediate and high σ_{noise} where we have the ordering $\pi_{ITP} < \pi_{BF} \approx \pi_{BFH} \leq \pi_{BH}$ (region 2). By augmenting n_{eval}, it can be observed an enlargement of region 1, π tends to decrease more slowly in region 2 and the results of all procedures are more and more similar. For all graphics see Figures B.15, B.16.

In several simulations we observed that, by construction of the simulation setting with data type const-tricube and by definition of the Rejection Rate False null hypotheses ρ and of the Power π, π is less precise in evaluating the true discoveries rate and interesting than ρ. Finally, in the same way as ρ, we observe that the results of π can be interpreted in terms of the Sound to Noise Ratio SNR given by $\frac{n_{eval}^2}{\sigma_{noise}^2}$.

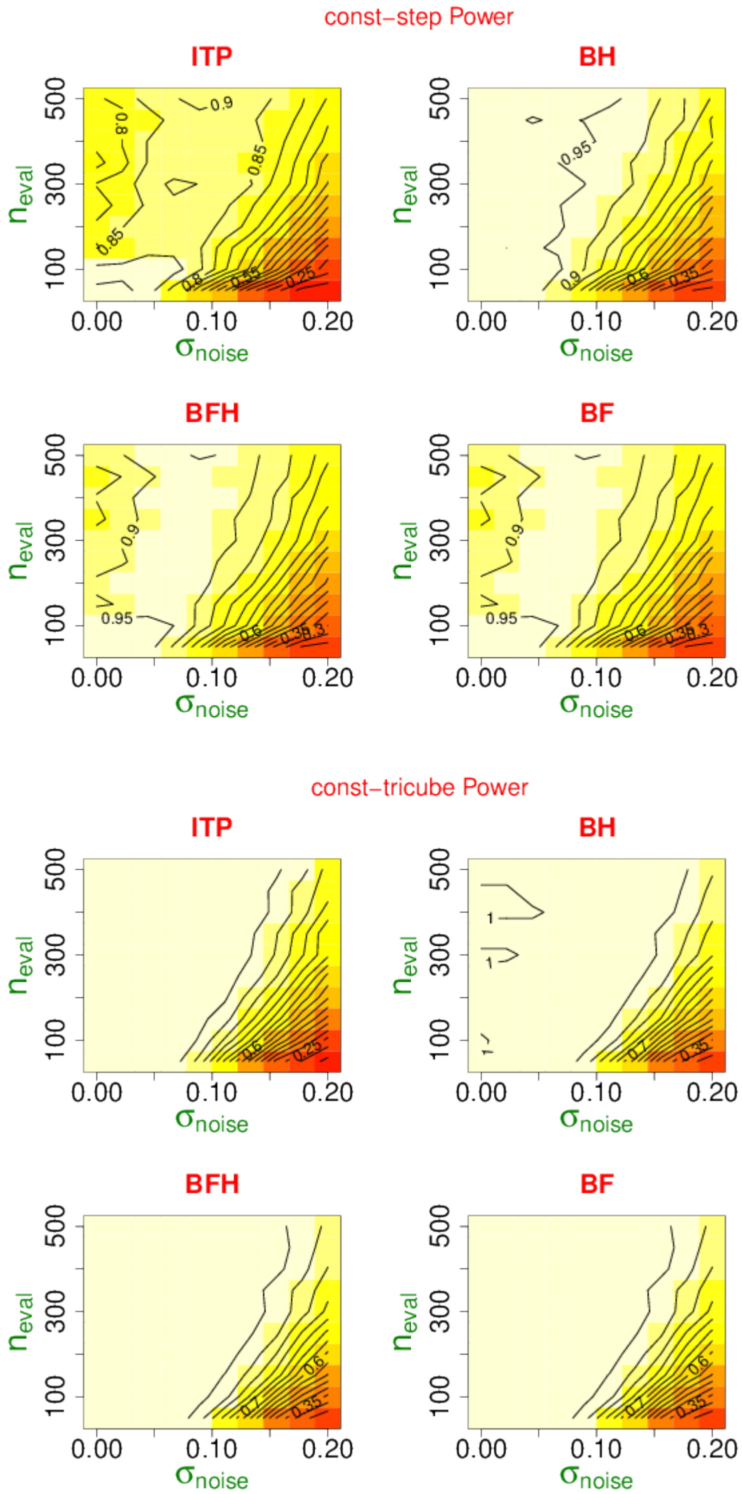

Figure 4.32: Heat maps of the Power π as a function of σ_{noise} and n_{eval} for the procedures Interval Testing Procedure (ITP), Benjamini-Hochberg (BH), Bonferroni-Holm (BFH), Bonferroni (BF) and for the data sets const-step (populations of constant and step functions) and const-tricube (populations of constant and tricube functions). In heat maps the bright white is associated with the maximum value 1 and the red is associated with the minimum value 0.

4.5 Number of statistical units n_{units} variable

4.5.1 Introduction

In this study the Data parameter of interest is the number of statistical units n_{units}. We assume that the the number of statistical units is the same for the two populations which could be of type const-step (populations of constant and step functions) or const-tricube (populations of constant and tricube functions).

n_{data} independent data sets are generated with maximum number of statistical units n_{units}^{max} among the tested values of n_{units}. For each tested value of n_{units} and for each generated data set, the first n_{units} functions of the current data set are selected.

4.5.2 Results

The common parameters in both scenarios with data sets const-step and const-tricube are:

- Number statistical units: $n_{units} = n_1 = n_2 \in \{10, \dots, \ 100\}$.
- Order B-spline basis: $m = 4$.
- Number basis elements : $p = 40$.
- Number functional evaluations: $n_{eval} = 150$. The functional evaluations are uniformly distributed in the domain $\mathcal{D} = [0, \ 2]$.
- Standard deviation additive Normal noise: $\sigma_{noise} = 0.05$
- Number data sets generated: $n_{data} = 1000$.

In scenarios with data set const-step, the parameters are:

- $h_{inf}^{CS} = 0.6$
- $h_{sup}^{CS} = 1.4$
- $v_{CS} = 0.13$

In scenarios with data set const-tricube, the parameters are:

- $h_{inf}^{CT} = 0.6$
- $h_{sup}^{CT} = 1.4$
- $v_{CT} = 0.19$

Family Wise Error Rate

The results in terms of the Family Wise Error Rate $FWER$ are reported in Figure 4.33. In the first place, we observe that BH does not control the $FWER$. In scenarios with data set const-step and for all values of n_{units}, we have $FWER_{BH} > \alpha = 0.05$; $FWER_{BH}$ is markedly non-decreasing; the same trend can be observed in scenarios with data set const-tricube even if it is less marked. Coherently with the theory, the contrary holds for ITP, BF and BFH.

For both data types const-step and const-tricube, we have the ordering

$$FWER_{BF} \leq FWER_{BFH} < FWER_{ITP} < FWER_{BH}$$

If the data set is const-step, we note:

1. The difference $FWER_{BFH} - FWER_{BF} \geq 0$ is negligible.

2. $FWER_{BFH}$, $FWER_{BF}$ are essentially non-decreasing. However, $FWER_{BFH}$ and $FWER_{BF}$ assume low values.

If the data set is const-tricube, we observe:

1. $FWER_{BFH} \approx FWER_{BF}$ and the difference $FWER_{ITP} - FWER_{BFH(BF)}$ is greater with respect to the scenarios with data set const-step.

2. $FWER_{BFH}$ and $FWER_{BF}$ have null slope and assume values near 0.

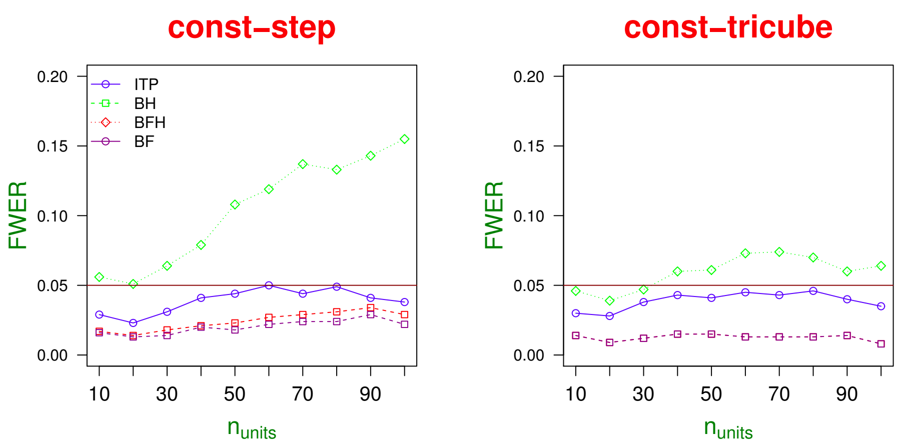

Figure 4.33: Family Wise Error Rate $FWER$ tuning the number of statistical units n_{units} in the first simulation

Rejection Rate False null hypotheses

The Rejection Rate False null hypotheses ρ is reported in Figure 4.34. Firstly, the results obtained with the procedures BFH and BF are coincident.

For all procedures the performances are worse with data set const-tricube, especially for ITP. The tricube kernel has the most part of the mass concentrated in the middle of its support $[h_{inf},\ h_{sup}]$ where the null hypotheses is false. Therefore, it turns out that the ITP is disadvantaged since it tends to maximize the true discoveries in the middle of $[h_{inf},\ h_{sup}]$ and the contrary holds at the edges of the same interval (consult the simulation study in Pini and Vantini, 2013). However, for n_{units} sufficiently big, the performances of ITP and BFH (BF) are equivalent. Furthermore, we observe that BH provides the best performances:

$$\rho_{BH} > \rho_s,\ s \in \{ITP,\ BFH,\ BF\}$$

In scenarios with data type const-step, it exists a value of n_{units} beyond which the ITP outperforms the BF and BFH. Moreover, the performances of ITP and BH are comparable in the set $n_{units} \in (50,\ 70)$.

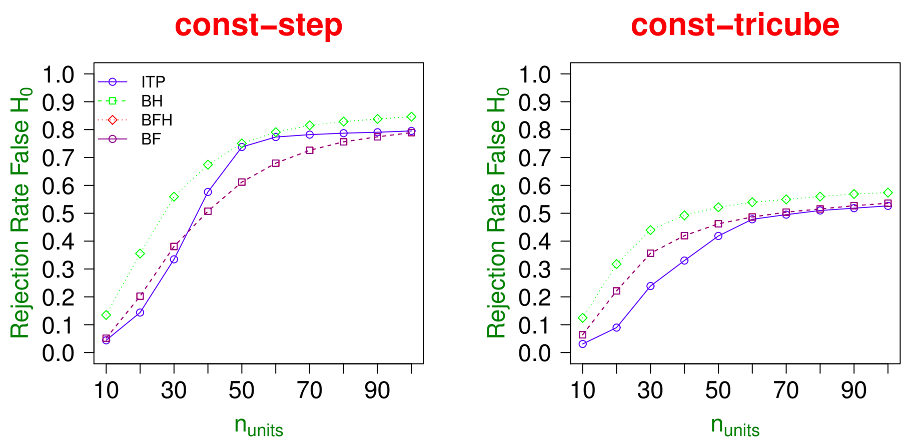

Figure 4.34: Rejection Rate False null hypotheses ρ tuning the number of statistical units n_{units} in the first simulation

Rejection Rate True null hypotheses

The results in terms of the Rejection Rate True null hypotheses γ are reported in Figure 4.35. In scenarios with data set const-step, the following ordering holds:

$$0 \approx \gamma_{BF} \approx \gamma_{BFH} < \gamma_{ITP} < \gamma_{BH}$$

When the data set type is const-tricube, unlike the cases with data set const-step, the results of the procedures ITP and BH are very similar.

In general γ_{ITP} and γ_{BH} tend to be non-decreasing with minimum assumed for $n_{units} = 20$. The ITP performances are better in scenarios with data set const-step.

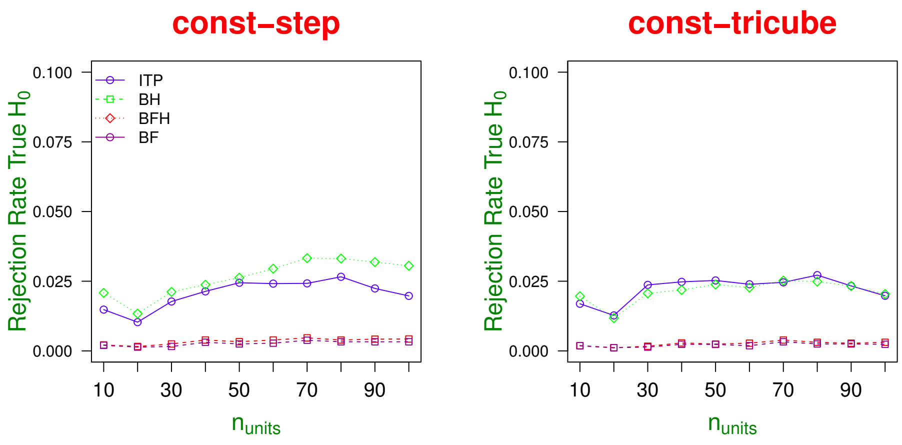

Figure 4.35: Rejection Rate True null hypotheses γ tuning the number of statistical units n_{units} in the first simulation

Power

The Power π is reported in Figure 4.36. There are no particular differences between the cases with data set const-step and const-tricube.

For all procedures, by increasing n_{units}, this metric rapidly reaches the unitary value, especially if the data type is const-tricube. The results obtained with the procedures BH, BFH and BF are essentially coincident and they are better than the ones of the ITP.

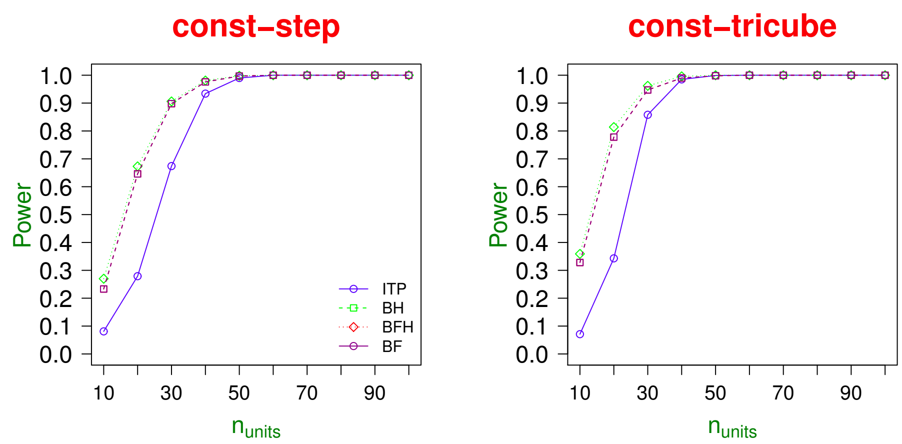

Figure 4.36: Power π tuning the number of statistical units n_{units} in the first simulation

In conclusion, we observe that the number of statistical units is not a critical parameter in the sense that good performances in terms of Rejection Rate False Null hypotheses and especially Power are achieved by all methods with moderate values of this parameter. Particularly significant patterns of the evaluation metrics have not been observed.

Conclusions

The general research interest of this thesis has been the Inference in the Functional Data Analysis framework. In detail, the attention has been focused on inferential methods provided with the *domain selection* property, i.e., methods able to select the statistically significant intervals of the domain.

The contribution of this thesis has been the exploration of the Interval Testing Procedure properties through a suitable simulation study. In short, the aim of the simulation study was the evaluation of the Interval Testing Procedure performances compared with the Benjamini-Hochberg, Bonferroni-Holm and Bonferroni ones by varying:

- The data set type: populations of constant and step functions (data set *const-step)* and populations of constant and tricube functions (data set *const-tricube*).
- The parameters of interest: the order m and the number p of B-spline basis elements, and the smoothing parameter λ (*Smoothing parameters*); the number of functional evaluation n_{eval} and the standard deviation of the additive Normal noise σ_{noise}, and the number of statistical units n_{units} (*Data parameters*).
- The Smoothing method: *Regression* or *Smoothing splines* using the B-spline basis expansion

The performances of the methods have been evaluated by means of the metrics *Family Wise Error Rate*, *Rejection Rate False null hypotheses*, *Rejection Rate True null hypotheses* and *Power.* The fundamental matter was the measure of the ability to make true discoveries between Interval Testing Procedure and Benjamini-Hochberg, knowing that the former procedure controls the *Family Wise Error Rate* on intervals and the latter procedure ensures only a weak control of the *Family Wise Error Rate.* We have adopted a simulation approach because we wanted to gain insights into the use of the B-spline basis expansion considering both the *Regression* and *Smoothing splines methods* in quite different scenarios with data type *const-step* and *const-tricube.* We have chosen these types of data set for the totally different degree of regularity that distinguishes them.

The code for the simulations has been implemented in R using *fdatest* and *fda* packages. The construction of the combining matrix is the most computationally expensive task in the Interval Testing Procedure algorithm for the Two-population framework. In order to improve the execution times, the construction of the combining matrix has been implemented in C, and this result has been integrated in the *fdatest* source code. In this way a speedup over the original implementantion equal to around 60x has been obtained. This result is useful both for simulations, where the Interval Testing Procedure must be applied several times for each value of the parameter under analysis, and for scenarios where it is reasonable to choose a high number of basis elements p. Finally, the used implementation of the *Interval Testing Procedure* directly works on an object of the *functional data class*. Therefore, the Smoothing is entrusted to the user avoiding subjective choices which had to be taken automatically in the original version of *fdatest*.

We have observed that for functional data with discontinuities, for Interval Testing Procedure the best choice would seem the use of the cubic splines with a sufficiently high number of basis elements. In these scenarios, the performances of Interval Testing Procedure and Benjamini-Hochberg are equivalent in terms of *Rejection Rate False null hypotheses*. This result shows that for Interval Testing Procedure the choice of the order and of the number of basis elements can be a critical and manageable problem. Indeed, the optimality of the cubic splines for Interval Testing Procedure was not easily predictable a priori because we expected that the unitary order would have been the best one for the functional reconstruction of a step, as it has been for the Benjamini-Hochberg procedure. However, in several simulations we observed that the *Rejection Rate False null hypotheses* is, essentially due to its definition and to the simulation setting, a more precise measure of the ability to make true discoveries than the *Power*. In general we have also noted that for Interval Testing Procedure it is better to choose the number of basis elements relatively high, to avoid losing information a priori.

If we consider the more regular data set *const-tricube*, the performances in terms of *Rejection Rate False null hypotheses* of all tested procedures are generally worse than in the cases with data set *const-step*. This result has been observed in all types of simulations and it is probably due to the non-uniform distribution of the tricube kernel mass which is mostly concentrated in the center of the support of the tricube kernel itself (the null hypothesis is false in the support of the tricube kernel). The best order is the unitary one (constant splines) for all procedures. In general with data set *const-tricube*, it is suggested to keep low the order and it is not necessary to choose a high number of basis elements in order to achieve optimality in terms of the ability to make true discoveries (metrics *Rejection Rate False null hypotheses* and *Power)*. Also this result was difficult to predict a priori. We expected that higher orders would have been optimal due to the regularity of the tricube kernel.

In general for Interval Testing Procedure, we have noted that the Smoothing must be performed carefully. At the same time, we have seen that suitably chosing the order and the number of basis elements, it is possible to achieve performances equivalent to the ones obtained with Benjamini-Hochberg procedure in terms of ability to make true discoveries. For instance in scenarios with data set *const-step*, chosing the cubic splines and the number of basis elements sufficiently high the performances of Interval Testing Procedure and of Benjamini-Hochberg procedure are equivalent in terms of *Rejection Rate False null hypotheses.*

The evaluation of the Smoothing effects on the inferential analysis is an open issue. Indeed, analytic expression of functional data is rarely directly available. More often, we only observe possibly noisy point-wise evaluations of data. Hence, a Smoothing method has to be applied, even if a basis expansion is not theoretically required by the adopted inferential procedure such as, for instance, the Interval-wise test proposed in Pini and Vantini (2015a). In Functional Data Analysis, the approach called in Zhang and Chen (2007) "Smoothing first, then estimation" is widely adopted. The data used for the analysis are reconstructions of the underlying functional data obtained with a Smoothing method. In the work of Zhang and Chen (2007), the authors report mild conditions under which the information loss, due to substitution of the underlying functional data with their reconstructions obtained with the Local Polynomial Kernel Smoothing method, is negligible. In Vsevolozhskaya et al. (2014), authors affirm that a low quality approximation of the observed data with smooth functions can entail a loss of statistical power and they report general Smoothing guidelines which are contributions from other works (Rice and Wu, 2001; Griswold and Gomulkiewicz, 2008). In Corain et al. (2014), it is claimed that understanding how the application of the Smoothing method affects the performances of the testing procedures is an open and challenging issue about which little is known. In Melas et al. (2014) the unique adressed topic is the optimal choice of the number of empirical Fourier coefficients for comparison of regression curves.

The main problem of the Smoothing is that general aspects are difficult to argue since there is always a certain degree of subjectivity. For example, the results presented in Zhang and Chen (2007) are general only in the Local Polynomial Kernel Smoothing framework. The usual approach is to follow best practices tailored for a specific Smoothing method. The best practices can be obtained via an analytical study or by a simulation study.

A possible future development is the study of the Interval Testing Procedure properties by means of an analytical study in the B-spline basis expansion framework, or of an another simulation study in a new Smoothing framework such as a Kernel Smoothing method. Another future development could be the design of purely inferential methods such as the Interval-wise test (Pini and Vantini, 2015a) which does not require the discretization of the data by means of a basis expansion.

Finally, it could be interesting to compare the Interval Testing Procedure using the B-spline basis expansion with the Vsevolozhskaya-Greenwood-Holodov test (Vsevolozhskaya et al., 2014). The Vsevolozhskaya-Greenwood-Holodov test does not require the discretization of the data by means of a basis expansion. Hence in order to have a fair comparison, the supports of the basis elements for the Interval Testing Procedure must be similar to the intervals selected a priori in Vsevolozhskaya-Greenwood-Holodov test. In most cases this task is not straightforward, except for constant B-splines. In this case, the supports of the basis elements can be easily chosen coincident with an arbitrary set of equal-sized intervals selected a priori. Moreover the number of basis elements must be carefully tuned. If it is too high, the Vsevolozhskaya-Greenwood-Holodov test is disadvantaged in terms of the ability to make true discoveries since it is based on the closure multiple testing procedure (Marcus et al., 1976). At the same time if it is too low, the Interval Testing Procedure is disavantaged since from simulations we have observed that it is generally better to choose the number of basis elements relatively high for this method. However, these extreme cases can be considered for exploratory reasons. The Interval-wise test and the Vsevolozhskaya-Greenwood-Holodov test are compared in Pini and Vantini (2015a).

As far as concerns the simulations with smoothing parameter (*Smoothing splines method*) based on the obtained results, the learned message is that for the Interval Testing Procedure it is highly recommended to apply, if available, theoretical results or best practices for the used Smoothing method. Alternatively in order to bypass partially the handling of the Smoothing, it is possible to apply the purely nonparametric inferential procedure Interval-wise test (Pini and Vantini, 2015a), which does not require the use of a basis expansion.

Regarding the simulations with parameters the standard deviation of noise and the number of evaluations, in general for each procedure the trend of the *Family Wise Error Rate* is independent from the data type. As expected, Benjamini-Hochberg method does not control the *Family Wise Error Rate*. In particular, we have observed a parabolic trend of the *Family Wise Error Rate* as a function of the standard deviation of noise and, by augmenting the number of evaluations, the points where this metric is maximized increase. Bonferroni, Bonferroni-Holm and Interval Testing Procedure methods control the *Family Wise Error Rate* in all cases (in few cases net of the upper bound of the estimate variability which has been obtained using the Normal approximation of the confidence interval for a Binomial proportion with the standard significance level $\alpha = 0.05$). It exists a value of the standard deviation of noise above which the *Family Wise Error Rate* of Interval Testing Procedure is lower than the one of Bonferroni and Bonferroni-Holm procedures. The *Family Wise Error Rate* of Interval Testing Procedure is non-increasing in the standard deviation of noise and the contrary holds for Bonferroni and Bonferroni-Holm procedures. However, the values of the *Family Wise Error Rate* of Bonferroni and Bonferroni-Holm are generally lower than the *Family Wise Error Rate* of Interval Testing Procedure and of Benjamini-Hochberg. When the standard deviation of noise is high, it is better to have few evaluations for Benjamini-Hochberg. In contrast, by increasing the standard deviation of noise, the *Family Wise Error Rate* of Interval Testing Procedure is non-increasing for every value of the number of evaluations.

In general for each procedure and for each data type, the *Rejection Rate False null hypotheses* is characterized by constant values for low standard deviations of noise, and by a decay for intermediate and high standard deviations of noise. By increasing the number of evaluations, the region presenting constant values tends to be enlarged. This predictable trend can be interpreted in terms of the quantity $\frac{n_{eval}^2}{\sigma_{noise}^2}$ which can be seen as the ratio between the Power of Signal and the Power of Noise (Sound to Noise Ratio SNR). When the standard deviation of noise is low, the SNR is high also for low number of evaluations, and all procedures are characterized by good performances in terms of the number of true discoveries. On the contrary, starting from a certain value of the standard deviation of noise, the SNR is increasing in the number of evaluations, and all procedures provide better performances in terms of the number of true discoveries by augmenting the number of evaluations. Interpreting the results in terms of SNR is an equivalent way to look at the slope changes of the *Rejection Rate False null hypotheses* by varying the standard deviation of noise and the number of evaluations. For low values of the standard deviation of noise, independently from the number of evaluations, the slope of the *Rejection Rate False null hypotheses* is roughly zero and this metric assumes high values; for intermediate and high standard deviations of noise, the *Rejection Rate False null hypotheses* is non-decreasing in the number of evaluations, and this metric assumes lower values. Also the parabolic trend of the *Family Wise Error Rate* of Benjamini-Hochberg as a function of the number of evaluations can be interpreted in terms of the ratio $\frac{n_{eval}^2}{\sigma_{noise}^2}$.

In scenarios with data set *const-step*, the *Rejection Rate False null hypotheses* of Bonferroni-Holm is higher than the *Rejection Rate False null hypotheses* of Bonferroni. In simulations with parameters the order and the number of B-spline basis elements, the performances of these procedures were generally very similar probably for the null standard deviation of noise. Bonferroni-Holm and Bonferroni are more conservative than Interval Testing Procedure and Benjamini-Hochberg. The interval Testing Procedure provides good performances in terms of *Rejection Rate False null hypotheses* with respect to Benjamini-Hochberg for high values of the standard deviation of noise. The Interval Testing Procedure has good performances in terms *Rejection Rate False null hypotheses* also for low values of the standard deviation of noise and of the number of evaluations. In scenarios with data set *const-tricube*, for low standard deviations of noise, differently from scenarios with data set *const-step*, Bonferroni-Holm provides higher values of the *Rejection Rate False null hypotheses* than Interval Testing Procedure. However, for high standard deviations of noise the procedures with the best performances in terms of *Rejection Rate False null hypotheses* are still Interval Testing Procedure and Benjamini-Hochberg. The results obtained with these procedures are comparable, in particular for very high values of the standard deviation of noise.

Benjamini-Hochberg generally provides the best results in terms of *Rejection Rate False null hypotheses*. However in terms of this metric, the difference between Interval Testing Procedure and Benjamini-Hochberg is not very marked, especially in scenarios with high values of the standard deviation of noise. Furthermore, contrarily to Interval Testing Procedure, Benjamini-Hochberg does not control the *Family Wise Error Rate* on intervals.

Finally, in general in scenarios with high standard deviations of noise, the Interval Testing Procedure is characterized by good performances as far as concerns both the *Family Wise Error Rate* and the *Rejection Rate False null hypotheses*.

For both data types *const-step* and *const-tricube*, the *Rejection Rate True null hypotheses* of Bonferroni-Holm and Bonferroni are essentially null as expected for their conservative nature. Independently from the number of evaluations, for Interval Testing Procedure and Benjamini-Hochberg the *Rejection Rate True null hypotheses* is decreasing in the standard deviation of noise. The decay of the *Rejection Rate True null hypotheses* for Interval Testing Procedure and for Benjamini-Hochberg, observable by increasing the standard deviation of noise, is slower by augmenting the number of evaluations. This trend is coherent with the high values of the *Rejection Rate False null hypotheses* obtained for low and intermediate standard deviations of noise and for sufficiently high values of the number of evaluations.

In scenarios with data set const-step similarly to the *Family Wise Error Rate*, it exists a value of the standard deviation of noise beyond which the Interval Testing Procedure outperforms Benjamini-Hochberg in terms of *Rejection Rate True null hypotheses*. We have observed the same tendencies in scenarios with data set const-tricube. By increasing the power of noise, the improvement - absolute and with respect to Benjamini-Hochberg - of Interval Testing Procedure performances in terms of *Rejection Rate True null hypotheses* is discernible. In particular it is worth to note that, for high standard deviations of noise, the *Rejection Rate True null hypotheses* is essentially null for Interval Testing Procedure, and, as expected, also for Bonferroni and Bonferroni-Holm. This trend is interesting for Interval Testing Procedure since, in the same cases, the performances in terms of *Rejection Rate False null hypotheses* of this procedure and of Benjamini-Hochberg are comparable.

In terms of *Power*, the general tendencies of Bonferroni and Bonferroni-Holm are very similar for both data types const-step and const-tricube. In scenarios with data set *const-step*, the characteristic patterns of *Power* and of *Rejection Rate False null hypotheses* are essentially similar. Moreover, the Interval Testing Procedure is generally outperformed by the other procedures. However from simulations with parameters the order and the number of B-spline basis elements, we expect that the performances in terms of *Power* of Interval Testing Procedure and Bonferroni-Holm (Bonferroni) would have been similar chosing an higher number of basis elements. Indeed based on the simulation results with parameters the order and the number of basis elements, for Interval Testing Procedure it is better to choose the cubic splines and the number of basis elements rather high. In this simulation we have used the cubic splines and the number of basis elements has been always chosen equal to 50 which is rather low considering that the minimum and the maximum tested values of the numbers of evaluations are 50 and 500. Finally, the results in terms of *Power* of all procedures are comparable for high values of the standard deviation of noise. In scenarios with data type const-tricube, independently from the number of evaluations and for low standard deviation of noise, for all procedures the *Power* is essentially unitary except for a decay of this metric for very high standard deviations of noise. Finally contrarily to the simulations with parameters order and number of the B-spline basis elements, we have observed that the data type has not been relevant.

As far as concerns the simulation with parameter the number of statistical units, particular patterns of the *Family Wise Error Rate* have not been observed. The Benjamini-Hochberg procedure does not control the *Family Wise Error Rate* in both scenarios with data *const-step* and *const-tricube*. The contrary holds for the other tested procedures.

In scenarios with data type *const-step*, it exists a value of the number of statistical units beyond which the Interval Testing Procedure outperforms Bonferroni and Bonferroni-Holm in terms of *Rejection Rate False null hypotheses*. In terms of this metric, the results obtained with Bonferroni-Holm and Bonferroni are very similar. Moreover, the performances in terms of *Rejection Rate False null hypotheses* of Interval Testing Procedure and Benjamini-Hochberg are very similar for relatively high values of the number of statistical units. For all procedures the performances in terms of *Rejection Rate False null hypotheses* are worse with data set *const-tricube*, especially for Interval Testing Procedure. The tricube kernel has the most part of the mass concentrated in the middle of its support where the null hypotheses is false. Therefore, it turns out that the Interval Testing Procedure is disadvantaged since it tends to maximize the true discoveries in the middle of this support and the contrary holds at the edges of the same interval (simulation study in Pini and Vantini 2013). However, for sufficiently high numbers of statistical units, the performances in terms of *Rejection Rate False null hypotheses* of Interval Testing Procedure and of Bonferroni-Holm (Bonferroni) are equivalent. Benjamini-Hochberg provides the best performances in terms *Rejection Rate False null hypotheses.*

In scenarios with data set *const-step*, for Bonferroni-Holm and Bonferroni the *Re-*

jection Rate True null hypotheses is essentially null. The Interval Testing Procedure is more conservative than Benjamini-Hochberg. Instead in scenarios with data type *const-tricube*, the results of Interval Testing Procedure and Benjamini-Hochberg are similar in terms of *Rejection Rate True null hypotheses.*

In general for Interval Testing Procedure and Benjamini-Hochberg, the *Rejection Rate True null hypotheses* tends to be non-decreasing. In terms of this metric, the Interval Testing Procedure performances are better in scenarios with data set *const-step.*

In terms of *Power*, there are no particular differences between the scenarios with data type *const-step* and *const-tricube.* For all procedures, by increasing the number of statistical units, this metric rapidly reaches the unitary value, especially if the data type is *const-tricube.*

In conclusion, we observe that the number of statistical units is not a critical parameter in the sense that good performances in terms of *Rejection Rate False Null hypotheses* and especially *Power* are achieved by all methods with moderate values of this parameter. Particularly significant patterns of the evaluation metrics have not been observed.

Bibliography

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

L. Corain, V. B. Melas, A. Pepelyshev, and L. Salmaso. *New insights on permutation approach for hypothesis testing on functional data*, volume 8. Springer, 2014.

A. Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2013.

C. De Boor. A practical guide to splines, revised edition, vol. 27 of applied mathematical sciences, 2001.

A. Delaigle and P. Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, pages 1171–1193, 2010.

F. Ferraty and Y. Romain. *The Oxford Handbook of Functional Data Analaysis.* Oxford University Press, 2011.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice.* Springer Science & Business Media, 2006.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

C. K. Griswold and N. Gomulkiewicz, R.and Heckman. Hypothesis testing in comparative and experimental studies of function-valued traits. *Evolution*, 62(5):1229–1242, 2008.

P. Hall and I. Keilegom. Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, 17(4):1511, 2007.

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

H. Hotelling. The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360–378, 08 1931. doi: 10.1214/aoms/1177732979. URL `http://dx.doi.org/10.1214/aoms/1177732979`.

R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

V. B. Melas, A. Pepelyshev, P. Shpilev, L. Salmaso, L. Corain, and R. Arboretti. On the optimal choice of the number of empirical fourier coefficients for comparison of regression curves. *Statistical Papers*, pages 1–17, 2014.

F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

A. Pini. *Global and component-wise distribution-free inference for functional data: methods and applications*. PhD thesis, Politecnico di Milano, December 2014.

A. Pini and S. Vantini. The interval testing procedure: inference for functional data controlling the family wise error rate on intervals. *MOX-Report*, 13:2013, 2013.

A. Pini and S. Vantini. Interval-wise testing for functional data. *MOX-Report*, 2015a.

A. Pini, A. Stamm, and S. Vantini. Hotelling's T^2 in functional Hilbert spaces. *MOX-Report*, No. 6 31/2015, 2015.

Alessia Pini and Simone Vantini. *fdatest: Interval Testing Procedure for Functional Data*, 2015b. URL `http://CRAN.R-project.org/package=fdatest`. R package version 2.1.

J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer New York, 2002.

J. O. Ramsay, Hadley Wickham, Spencer Graves, and Giles Hooker. *fda: Functional Data Analysis*, 2014. URL `http://CRAN.R-project.org/package=fda`. R package version 2.4.4.

J. A. Rice and C. O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, pages 253–259, 2001.

L. Schumaker. *Spline functions: basic theory*. Cambridge University Press, 2007.

P. Secchi, A. Stamm, S. Vantini, et al. Inference for the mean of large p small n data: A finite-sample high-dimensional generalization of hotelling's theorem. *Electronic Journal of Statistics*, 7:2005–2031, 2013.

J. Shen, Q.and Faraway. An f test for linear models with functional responses. *Statistica Sinica*, 14(4):1239–1258, 2004.

B. W. Silverman and J. O. Ramsay. *Functional Data Analysis*. Springer, 2005.

T. Tarpey and K. K. J. Kinateder. Clustering functional data. *Journal of classification*, 20(1):093–114, 2003.

R. Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015. URL `https://www.R-project.org/`.

O. Vsevolozhskaya, M. Greenwood, D. Holodov, et al. Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *The Annals of Applied Statistics*, 8(2):905–925, 2014.

J. Zhang. *Analysis of variance for functional data*. CRC Press, 2013.

J. Zhang and J. Chen. Statistical inferences for functional data. *The Annals of Statistics*, 35(3):1052–1079, 2007.

Appendix A

Graphics of the evaluation metrics as a function of the Regression Smoothing parameters

Since the explored parameters are m and p, the number of possible dimensions is two. In order to identify typical trends, it is interesting to observe a metric as a function of p by maintaining m fixed and vice versa.

In the following for completeness we report all the standard graphics (i.e., metric as a function of p with fixed m or vice versa) divided according to the data type (populations of constant and step functions const-step or populations of constant and tricube functions const-tricube).

A.1 Graphical results data set const-step

In Figures A.1, A.2 it is reported the $FWER$ as a function of p with m fixed and as a function of m with p fixed, respectively.

In Figures A.3, A.4 it is reported the Rejection Rate False null hypotheses ρ as a function of p with m fixed and as a function of m with p fixed, respectively.

In Figures A.5, A.6 it is reported the Rejection Rate True null hypotheses γ as a function of p with m fixed and as a function of m with p fixed, respectively.

In Figures A.7, A.8 it is reported the Power π as a function of p with m fixed and as a function of m with p fixed, respectively.

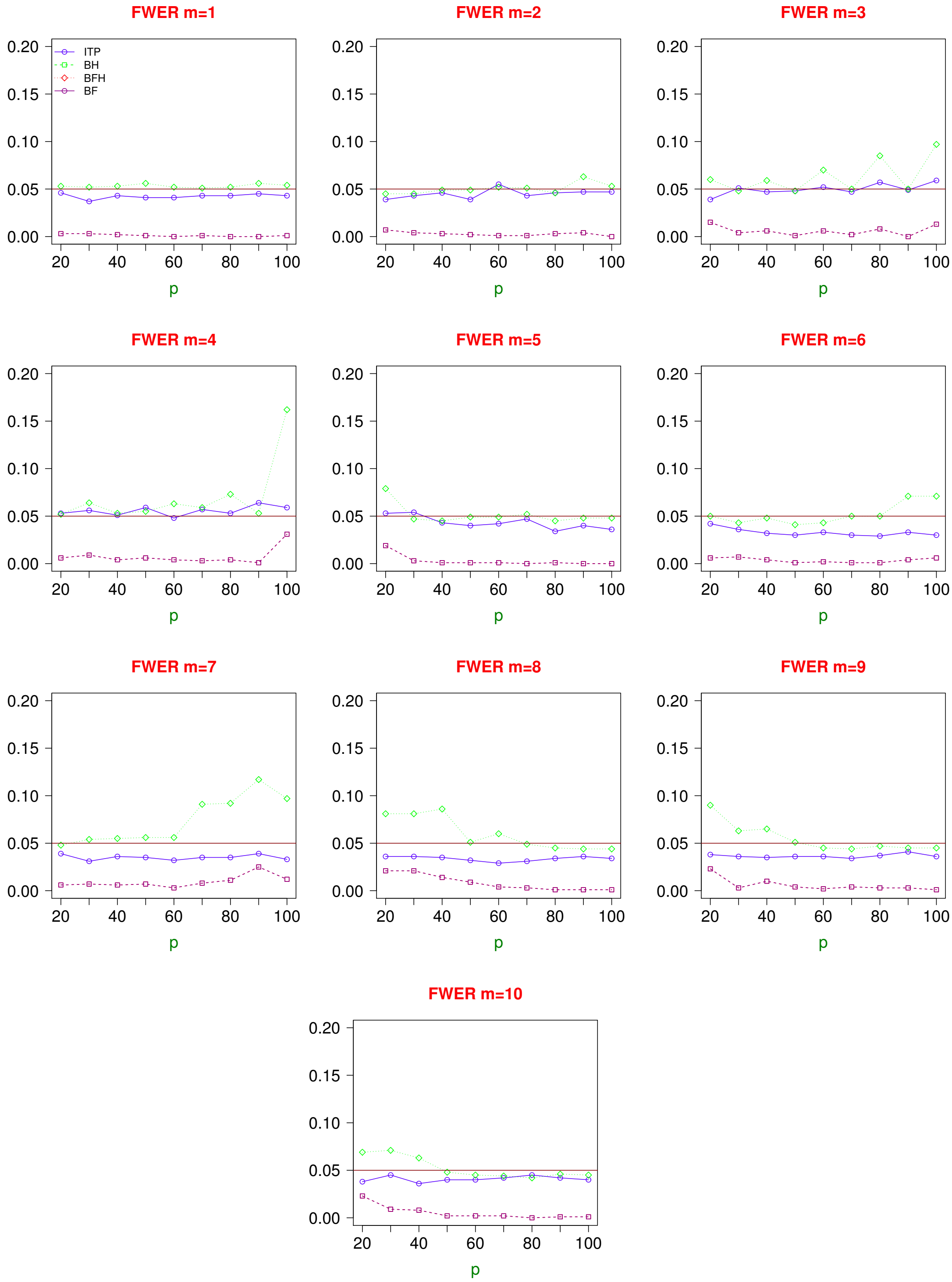

Figure A.1: Family Wise Error Rate $FWER$ as a function of the number p of B-spline basis elements with order m fixed. The data type is populations of constant and step functions.

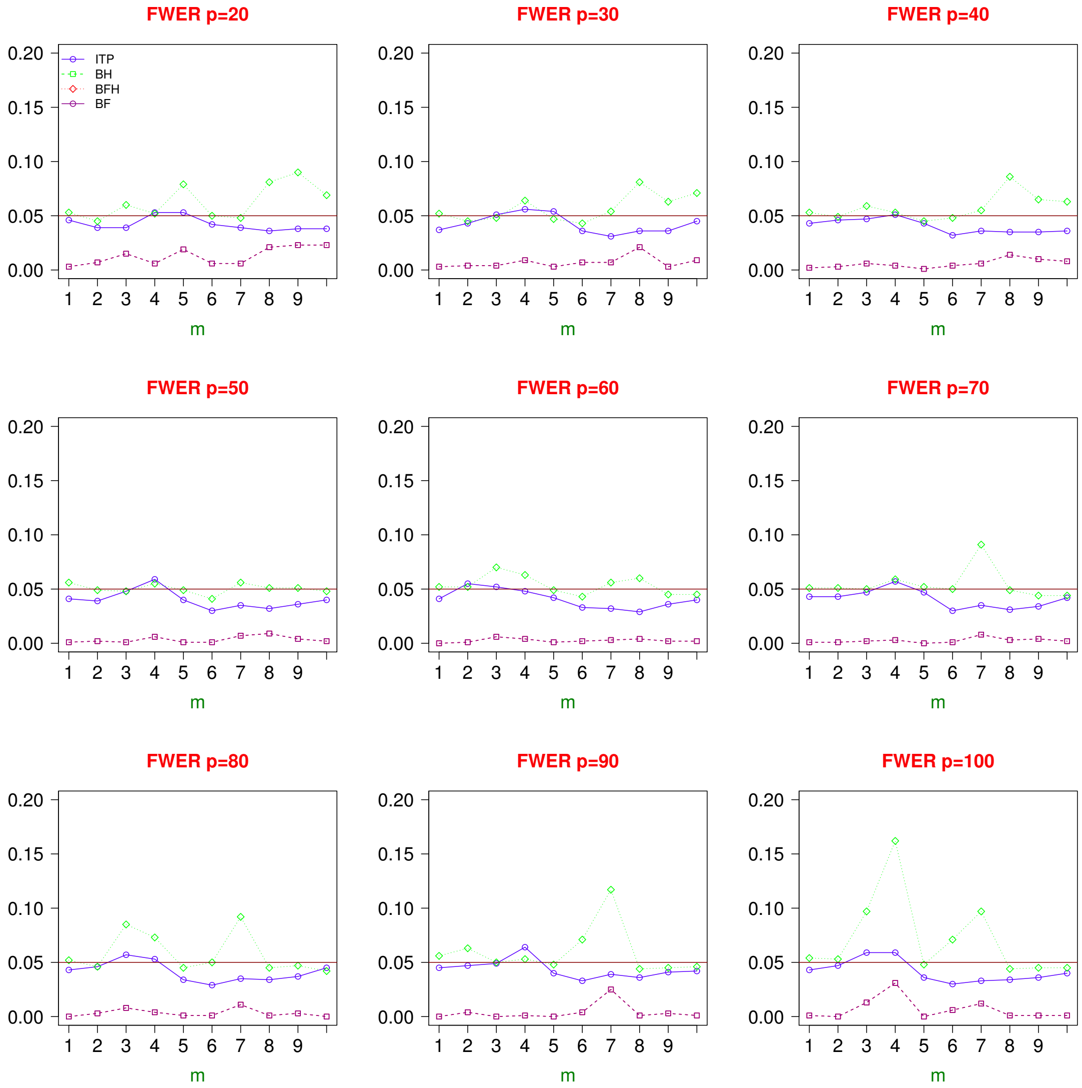

Figure A.2: Family Wise Error Rate $FWER$ as a function of the order m B-spline basis with number basis elements p fixed. The data type is populations of constant and step functions.

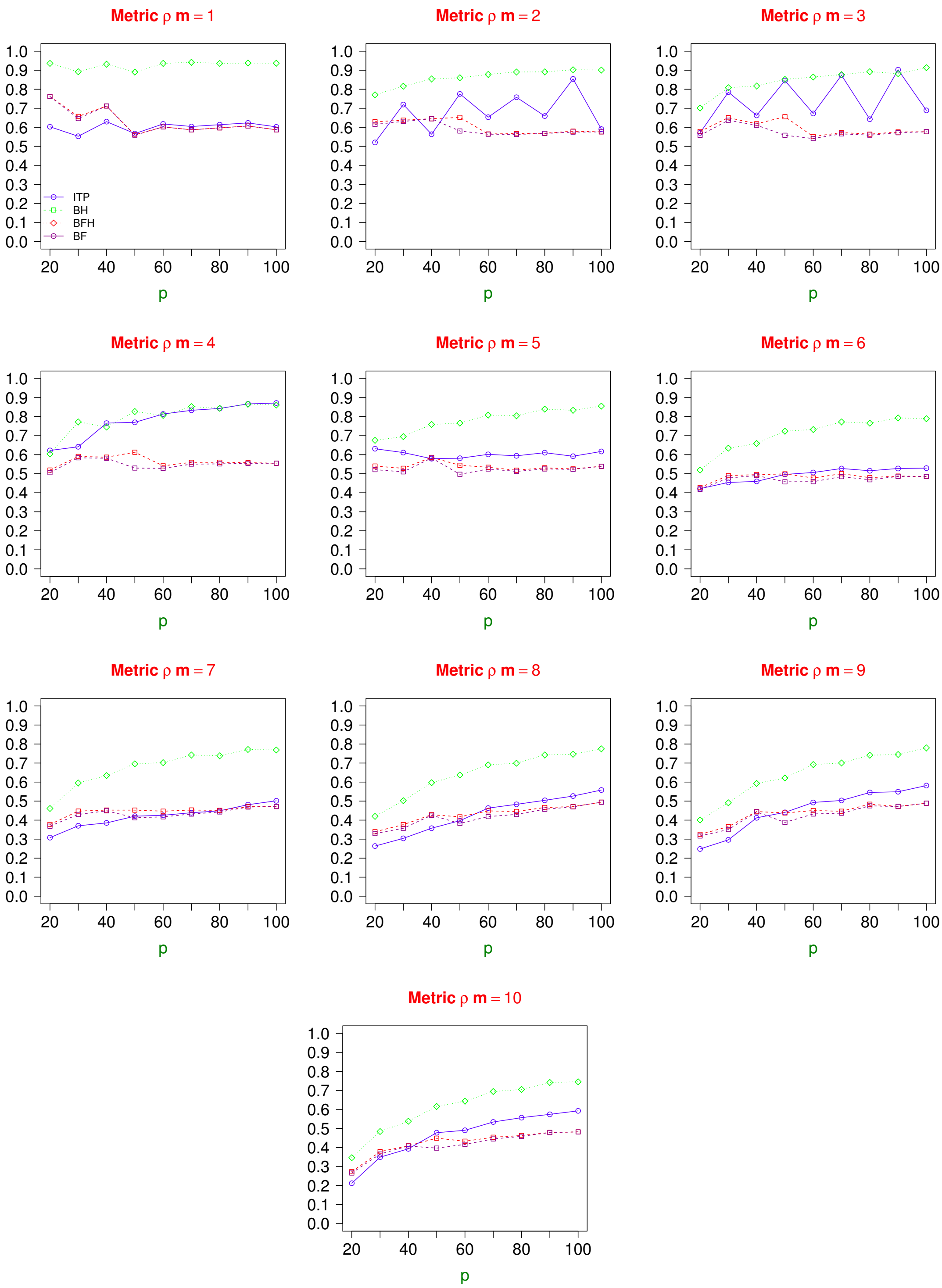

Figure A.3: Rejection Rate False null hypotheses ρ as a function of the number p of B-spline basis elements with order m fixed. The data type is populations of constant and step functions.

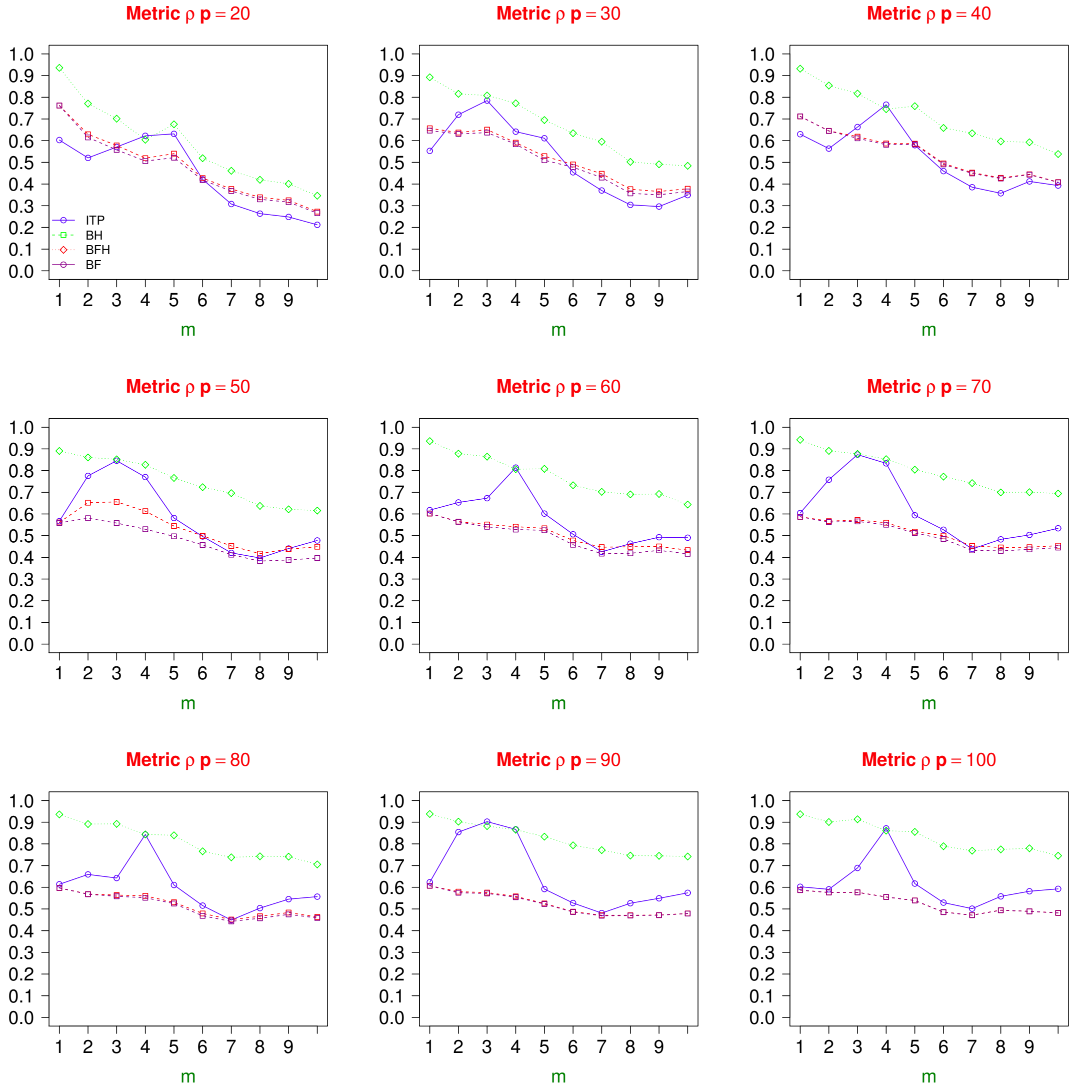

Figure A.4: Rejection Rate False null hypotheses ρ as a function of the order m B-spline basis with number basis elements p fixed. The data type is populations of constant and step functions.

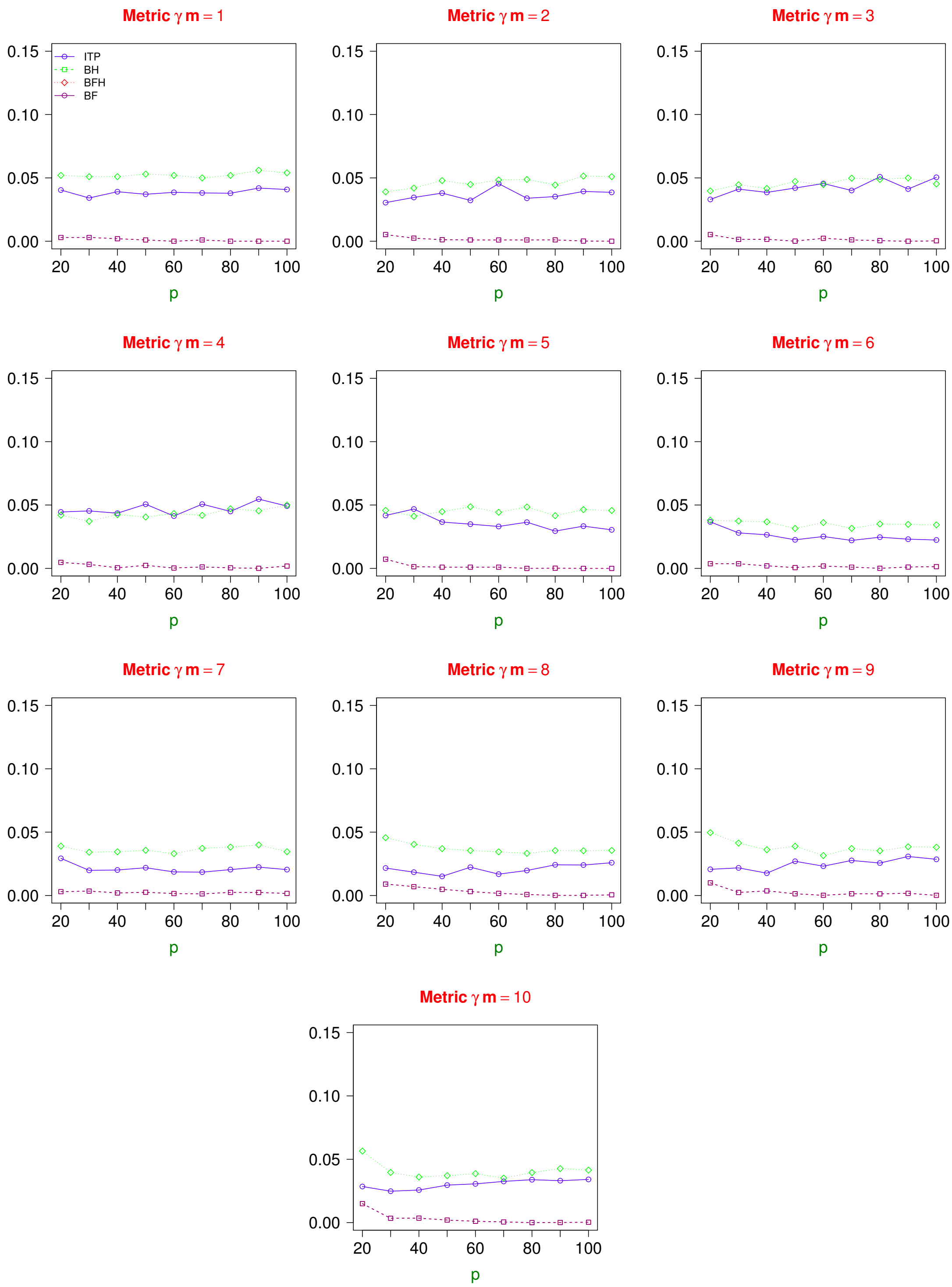

Figure A.5: Rejection Rate True null hypotheses γ as a function of the number p of B-spline basis elements with order m fixed. The data type is populations of constant and step functions.

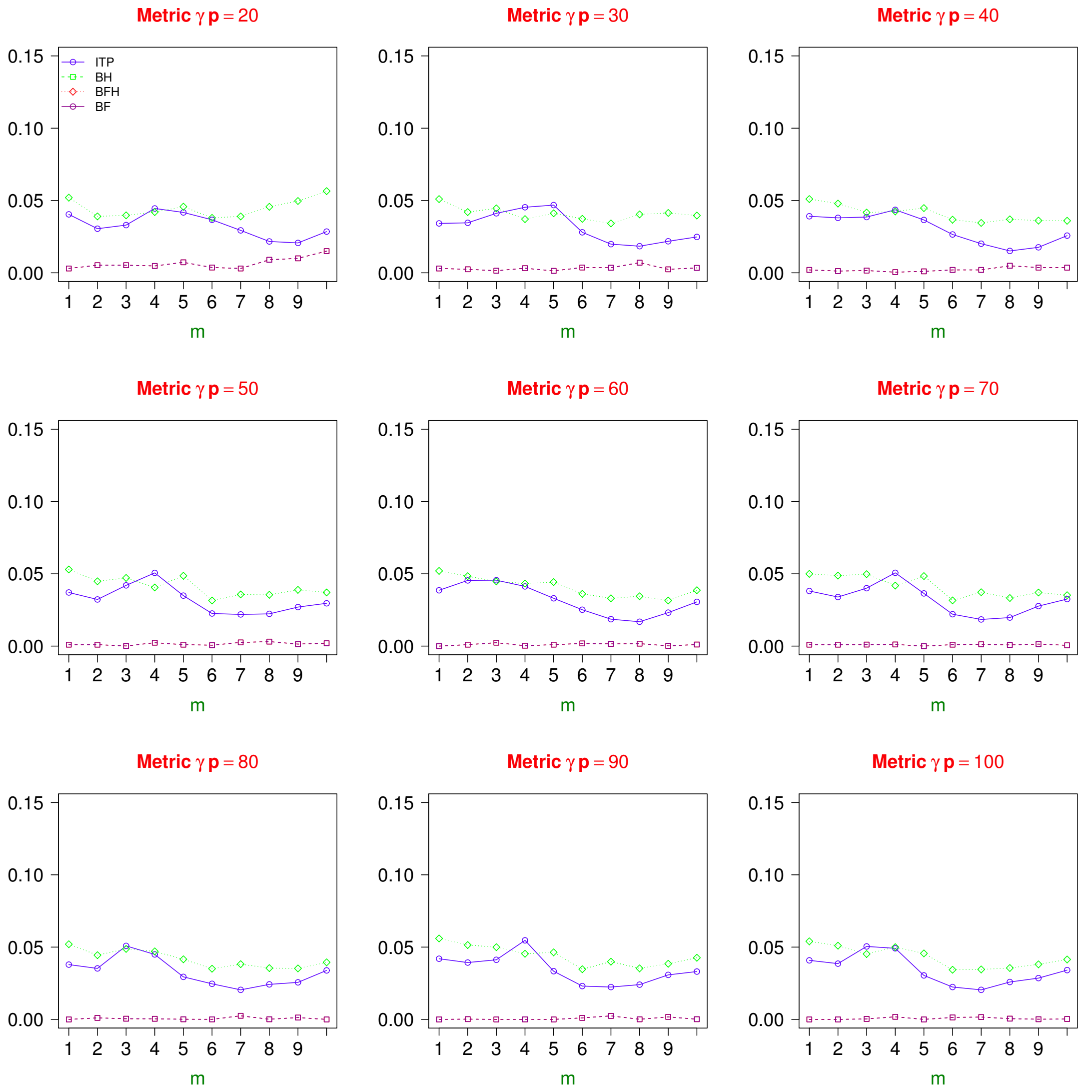

Figure A.6: Rejection Rate True null hypotheses γ as a function of the order m B-spline basis with number basis elements p fixed. The data type is populations of constant and step functions.

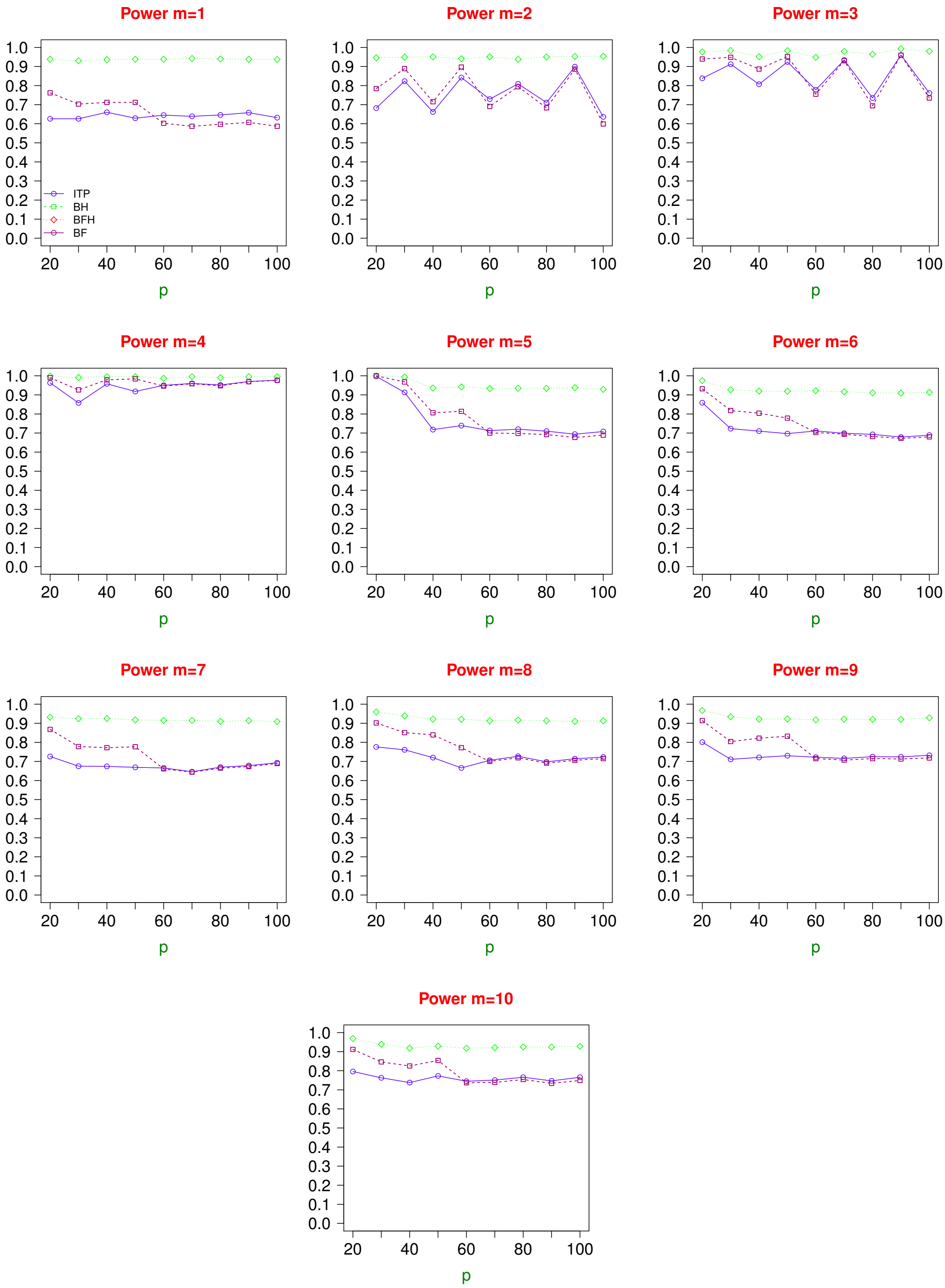

Figure A.7: Power π as a function of the number p of B-spline basis elements with order m fixed. The data type is populations of constant and step functions.

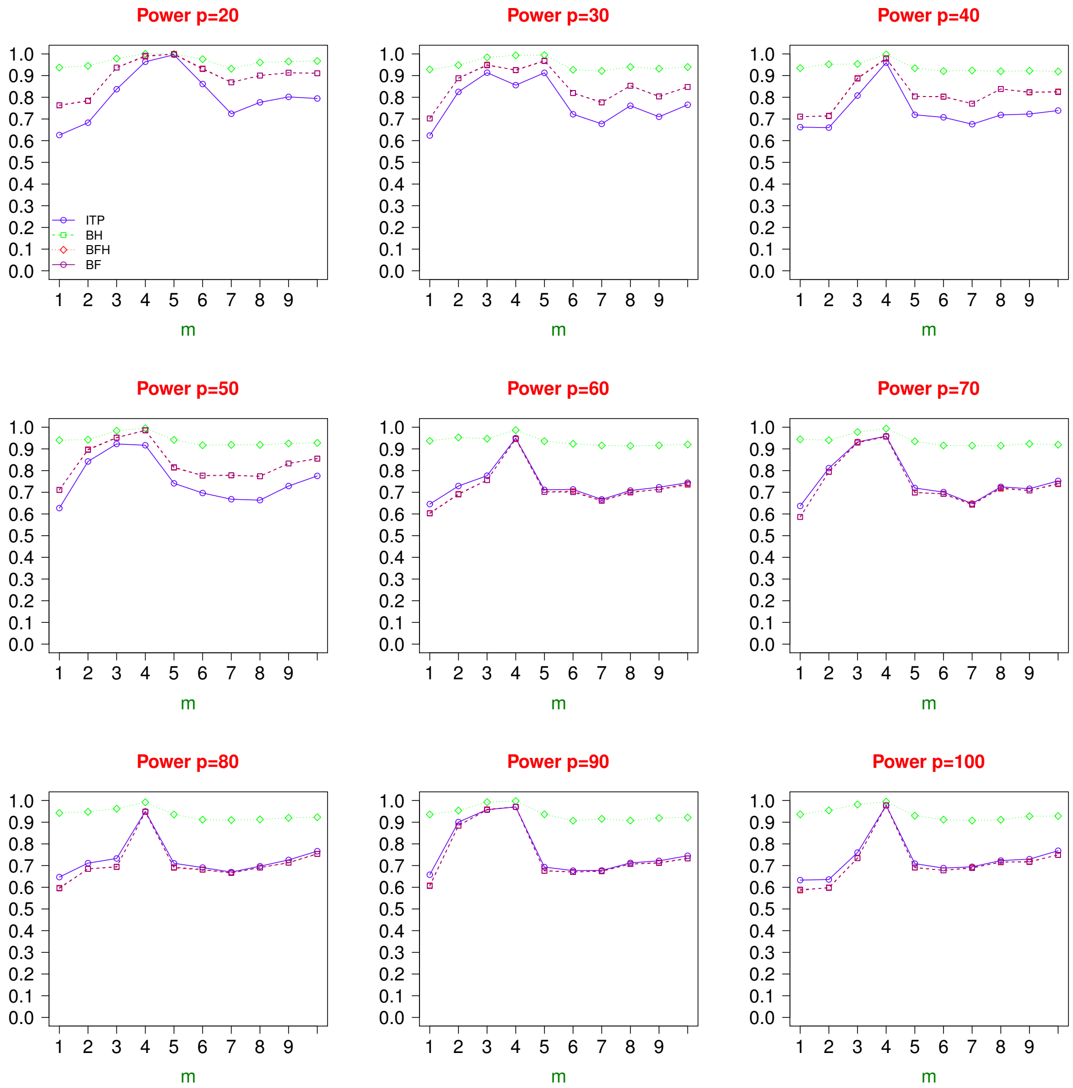

Figure A.8: Power π as a function of the order m B-spline basis with number basis elements p fixed. The data type is populations of constant and step functions.

A.2 Graphical results data set const-tricube

In Figures A.9, A.10 it is reported the $FWER$ as a function of p with m fixed and as a function of m with p fixed, respectively.

In Figures A.11, A.12 it is reported the Rejection Rate False null hypotheses ρ as a function of p with m fixed and as a function of m with p fixed, respectively.

In Figures A.13, A.14 it is reported the Rejection Rate True null hypotheses γ as a function of p with m fixed and as a function of m with p fixed, respectively.

In Figures A.15, A.16 it is reported the Power π as a function of p with m fixed and as a function of m with p fixed, respectively.

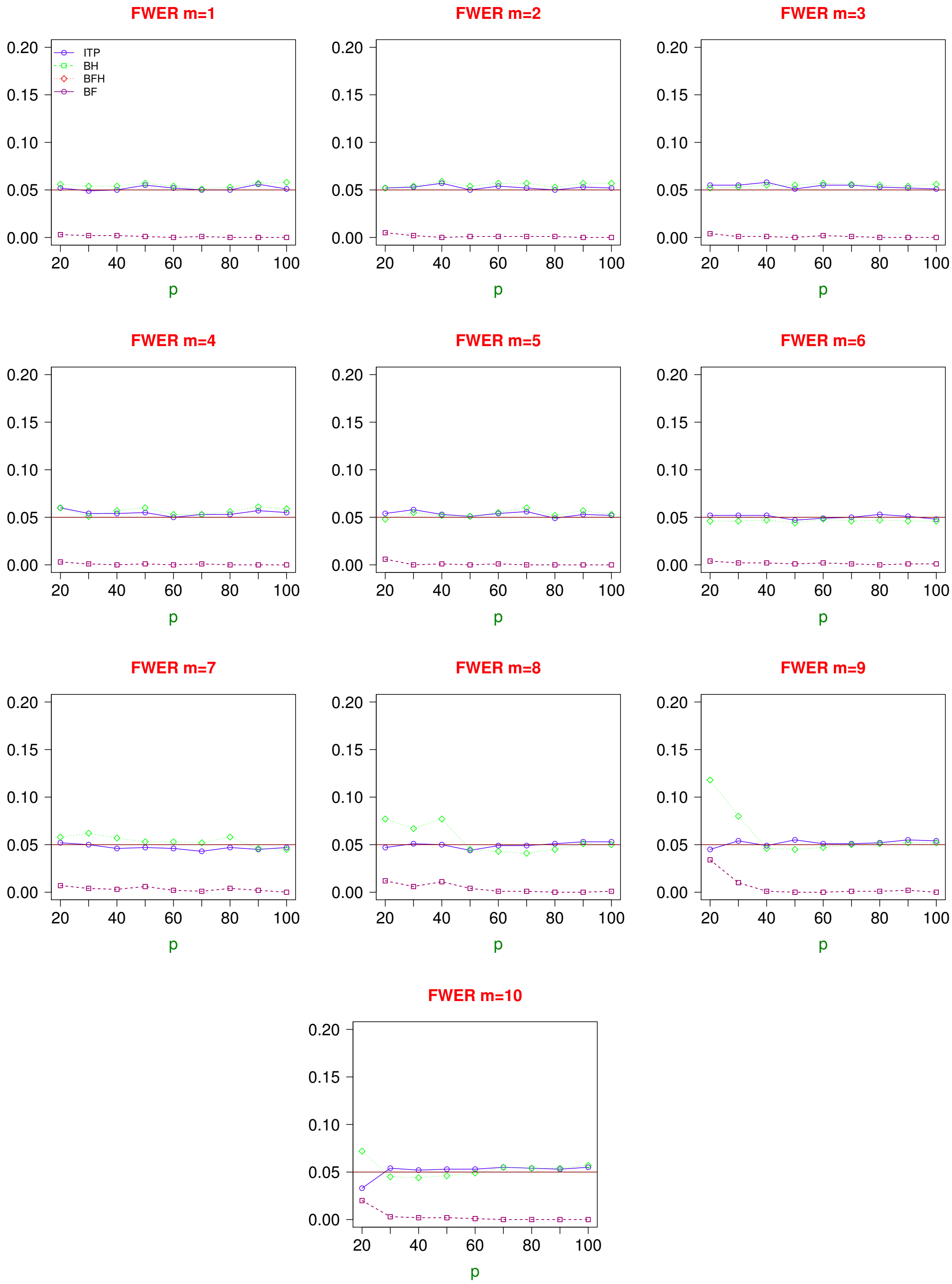

Figure A.9: Family Wise Error Rate $FWER$ as a function of the number p of B-spline basis elements with order m fixed. The data type is populations of constant and tricube functions.

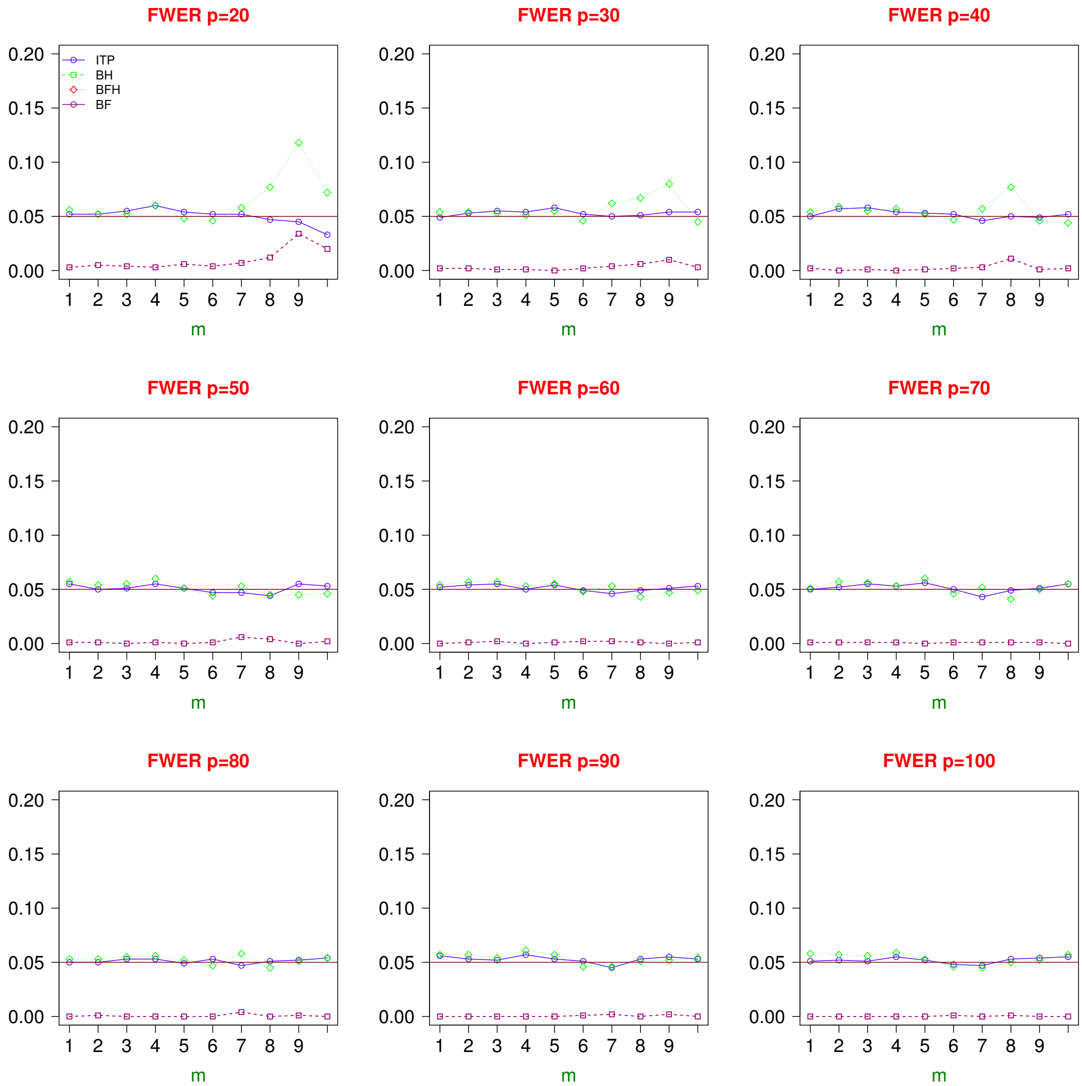

Figure A.10: Family Wise Error Rate $FWER$ as a function of the order m B-spline basis with number basis elements p fixed. The data type is populations of constant and tricube functions.

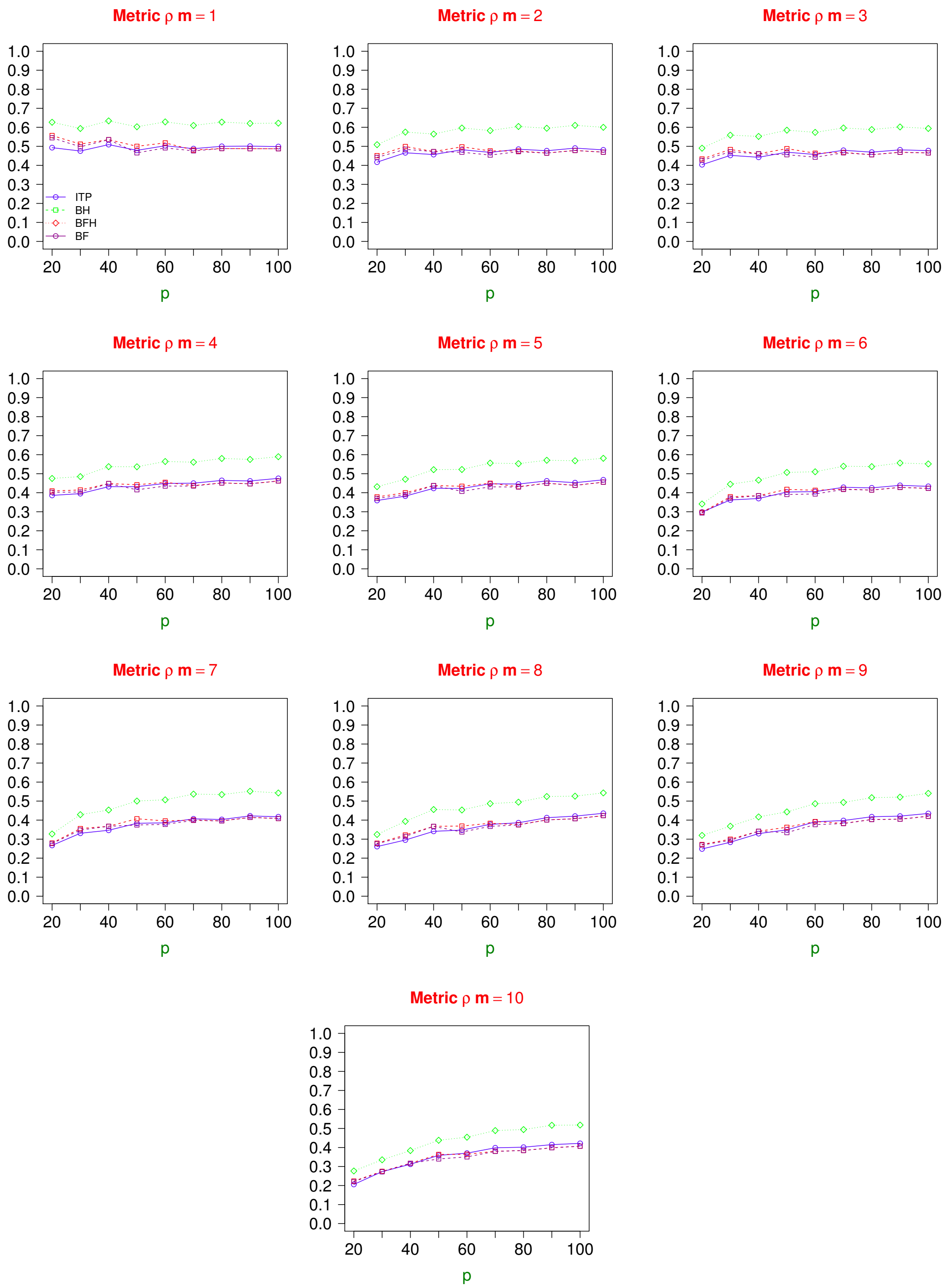

Figure A.11: Rejection Rate False null hypotheses ρ as a function of the number p of B-spline basis elements with order m fixed. The data type is populations of constant and tricube functions.

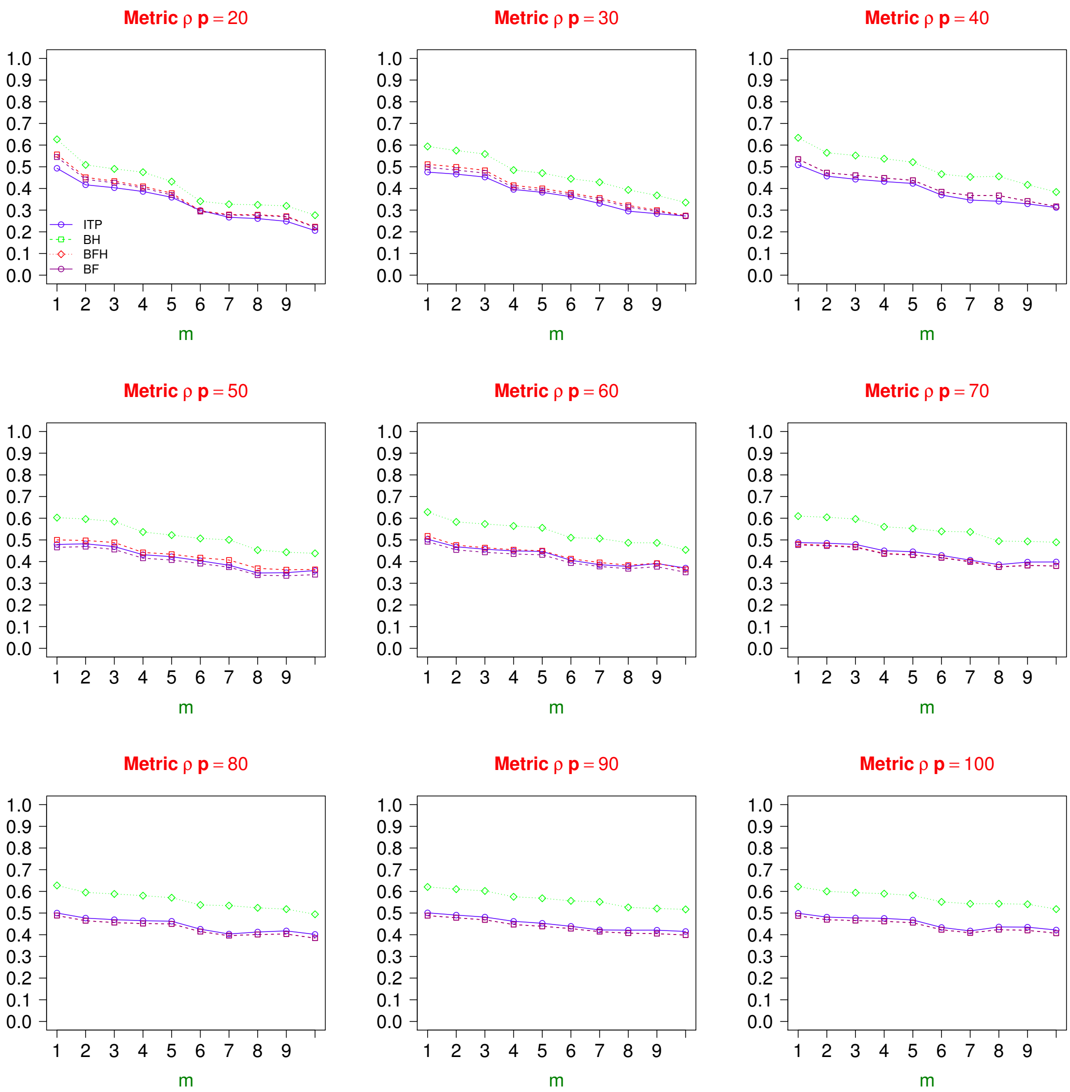

Figure A.12: Rejection Rate False null hypotheses ρ as a function of the order m B-spline basis with number basis elements p fixed. The data type is populations of constant and tricube functions.

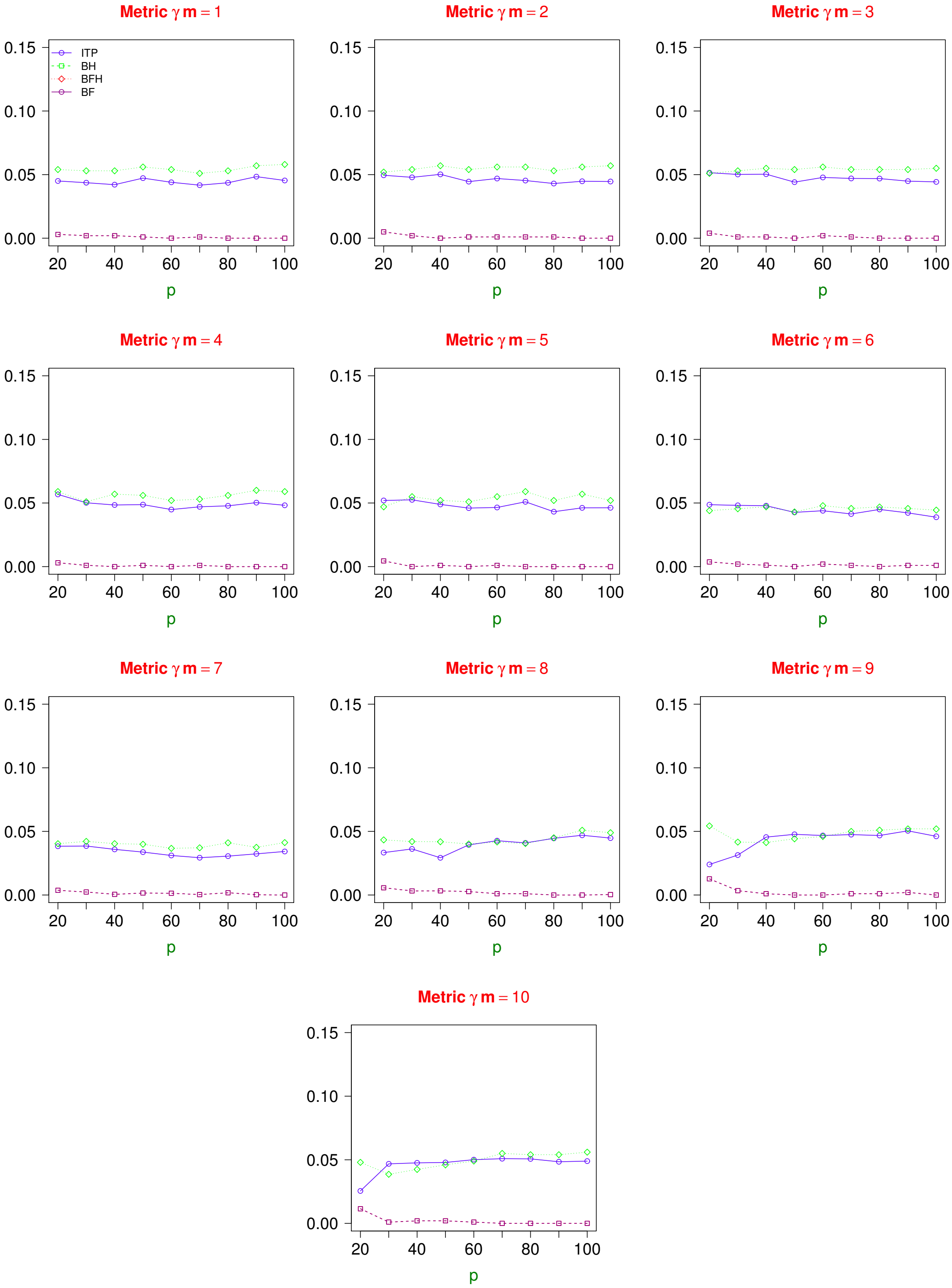

Figure A.13: Rejection Rate True null hypotheses γ as a function of the number p of B-spline basis elements with order m fixed. The data type is populations of constant and tricube functions.

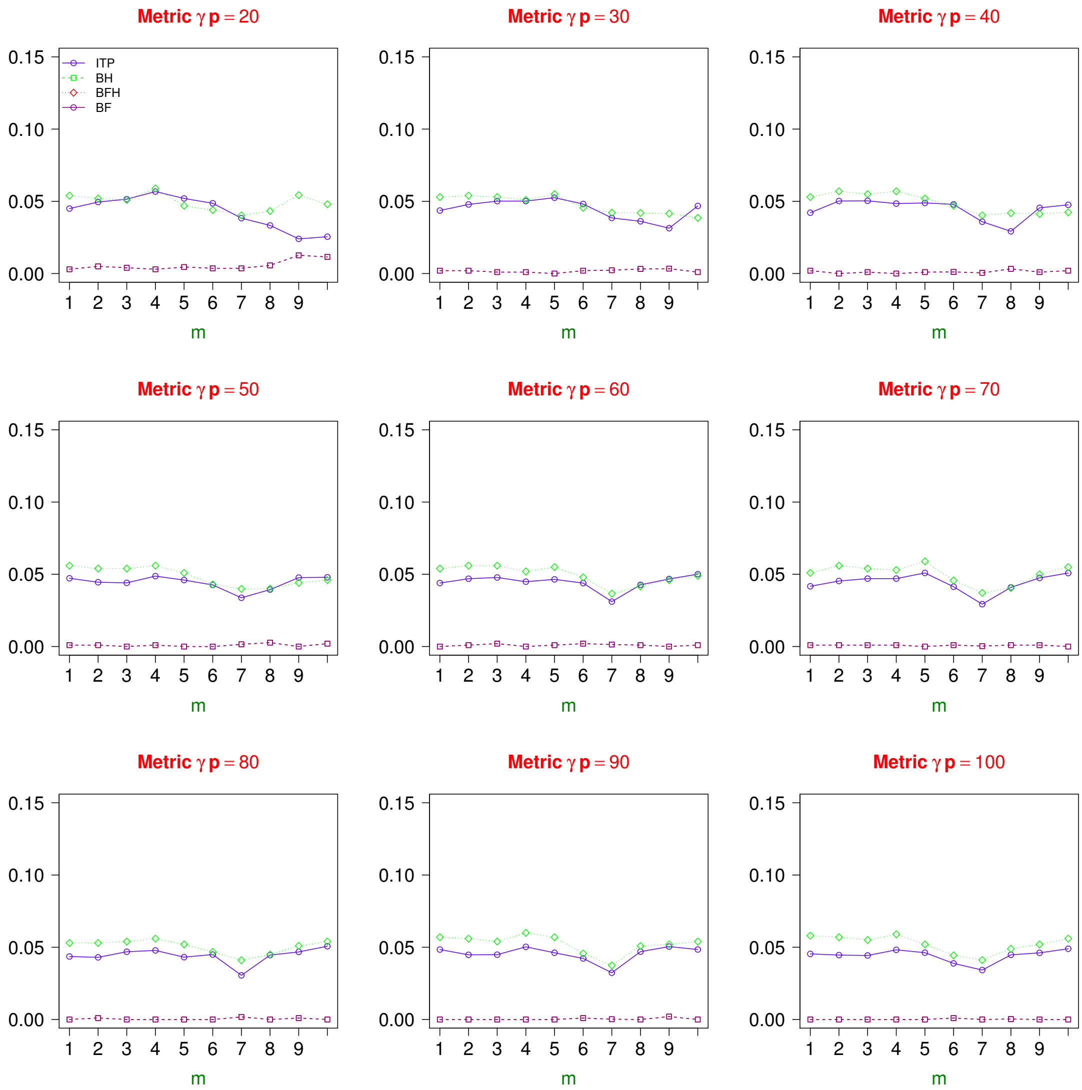

Figure A.14: Rejection Rate True null hypotheses γ as a function of the order m B-spline basis with number basis elements p fixed. The data type is populations of constant and tricube functions.

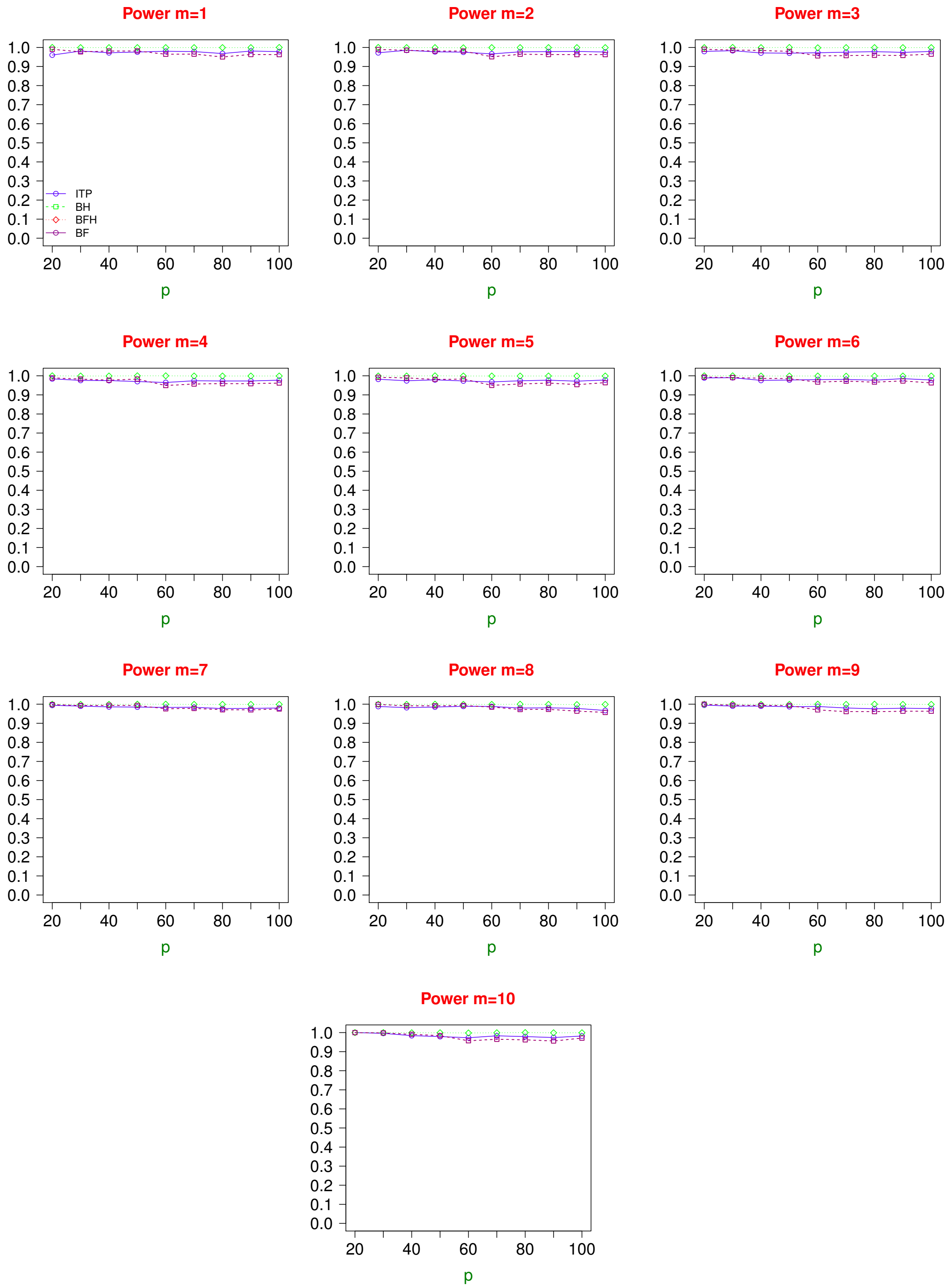

Figure A.15: Power π as a function of the number p of B-spline basis elements with order m fixed. The data type is populations of constant and tricube functions.

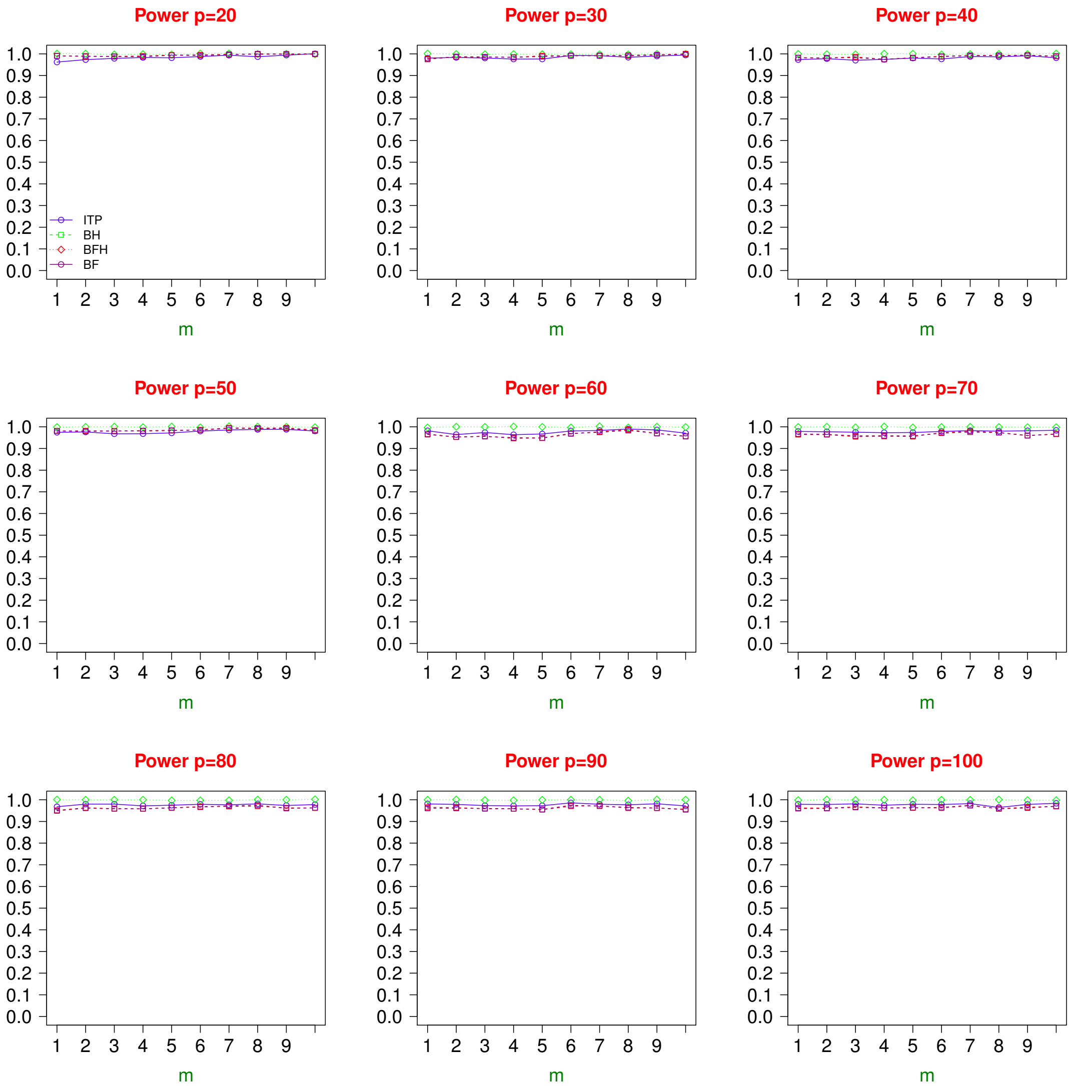

Figure A.16: Power π as a function of the order m B-spline basis with number basis elements p fixed. The data type is populations of constant and tricube functions.

Appendix B

Graphics of the evaluation metrics as a function of σ_{noise} and n_{eval}

Since the explored parameters are σ_{noise} and n_{eval}, the number of possible dimensions is two. In order to identify typical trends, it is interesting to observe a metric as a function of σ_{noise} by maintaining fixed n_{eval} and vice versa.

In the following for completeness we report all the standard graphics (i.e., metric as a function of σ_{noise} with fixed n_{eval} or vice versa) divided according to the data type (populations of constant and step functions const-step or populations of constant and tricube functions const-tricube).

B.1 Graphical results data set const-step

In Figures B.1, B.2 it is reported the $FWER$ as a function of σ_{noise} with n_{eval} fixed and as a function of n_{eval} with σ_{noise} fixed, respectively.

In Figures B.3, B.4 it is reported the Rejection Rate False null hypotheses ρ as a function of σ_{noise} with n_{eval} fixed and as a function of n_{eval} with σ_{noise} fixed, respectively.

In Figures B.5, B.6 it is reported the Rejection Rate True null hypotheses γ as a function of σ_{noise} with n_{eval} fixed and as a function of n_{eval} with σ_{noise} fixed, respectively.

In Figures B.7, B.8 it is reported the Power π as a function of σ_{noise} with n_{eval} fixed and as a function of n_{eval} with σ_{noise} fixed, respectively.

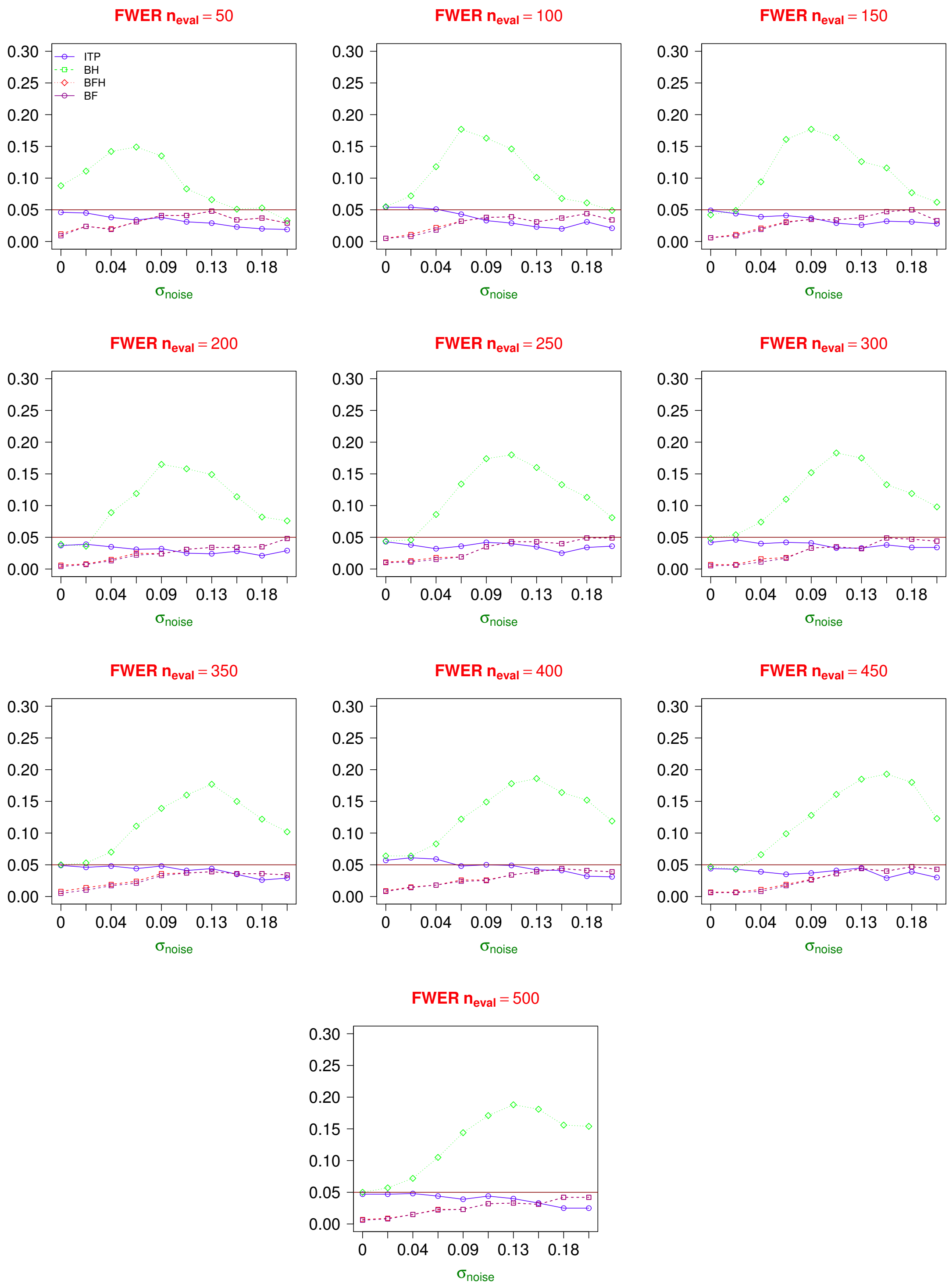

Figure B.1: Family Wise Error Rate $FWER$ as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and step functions.

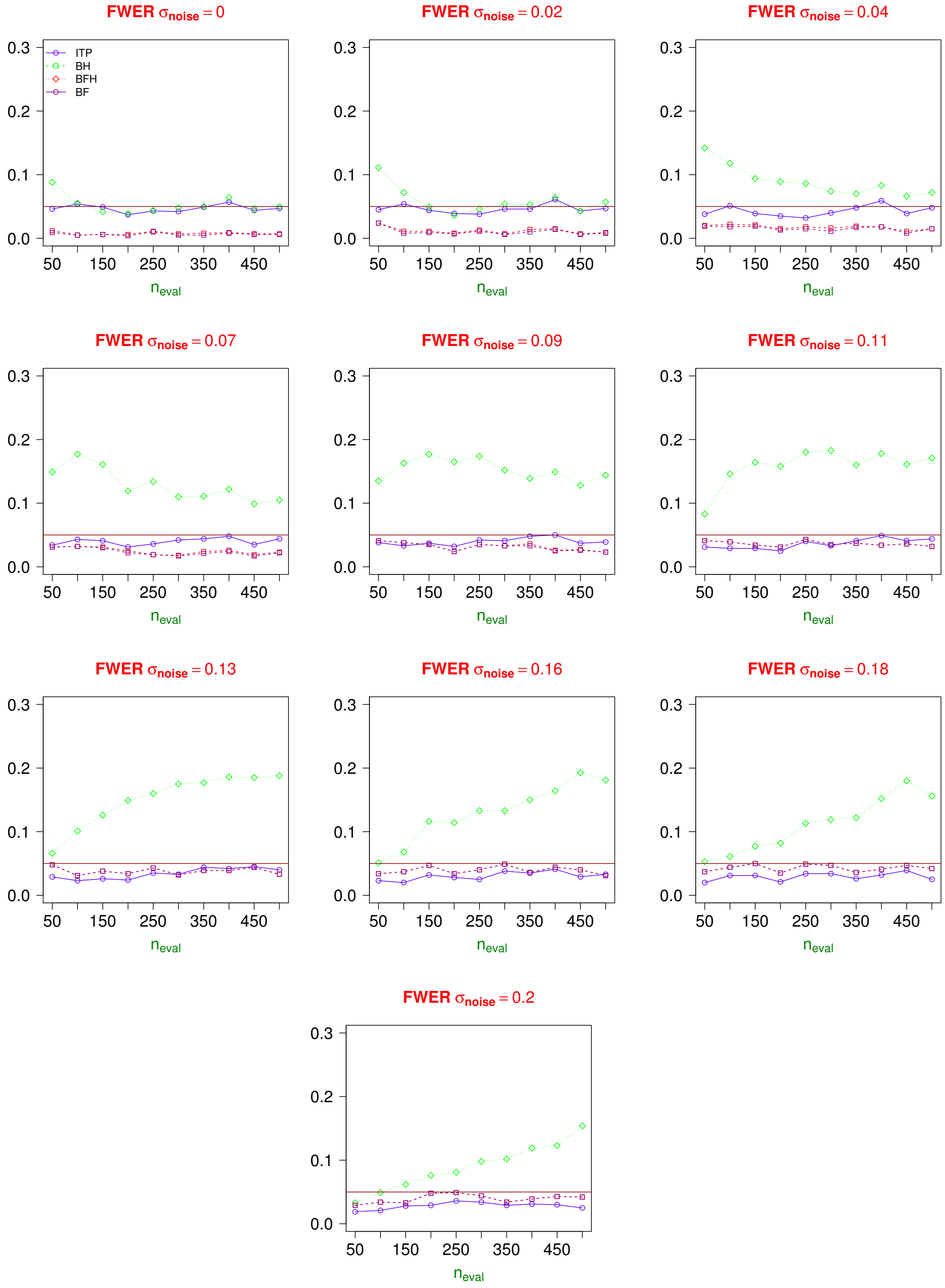

Figure B.2: Family Wise Error Rate $FWER$ as a function of the number of evaluations n_{eval} with fixed standard deviation of noise σ_{noise}. The data type is populations of constant and step functions.

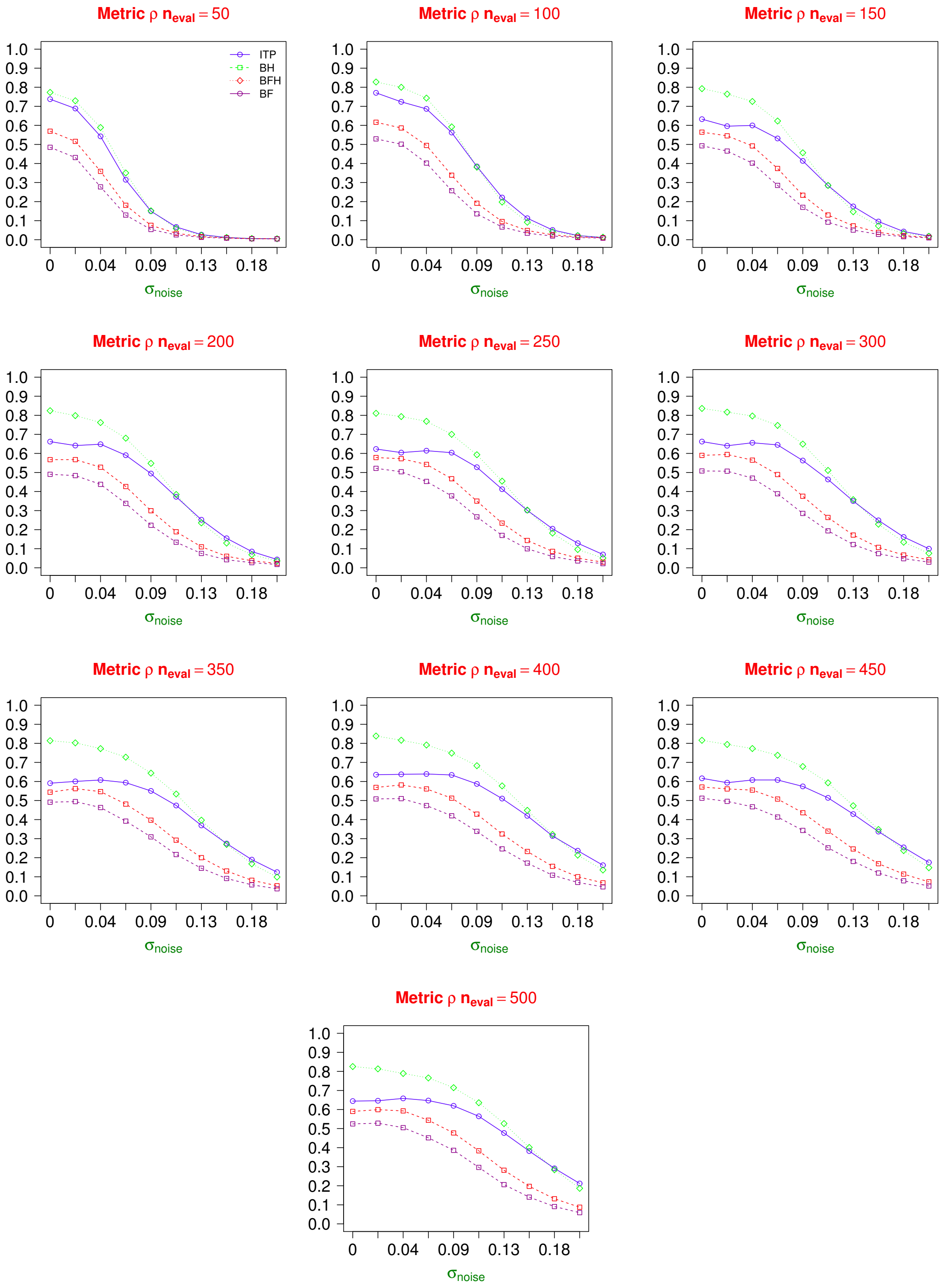

Figure B.3: Rejection Rate False null hypotheses ρ as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and step functions.

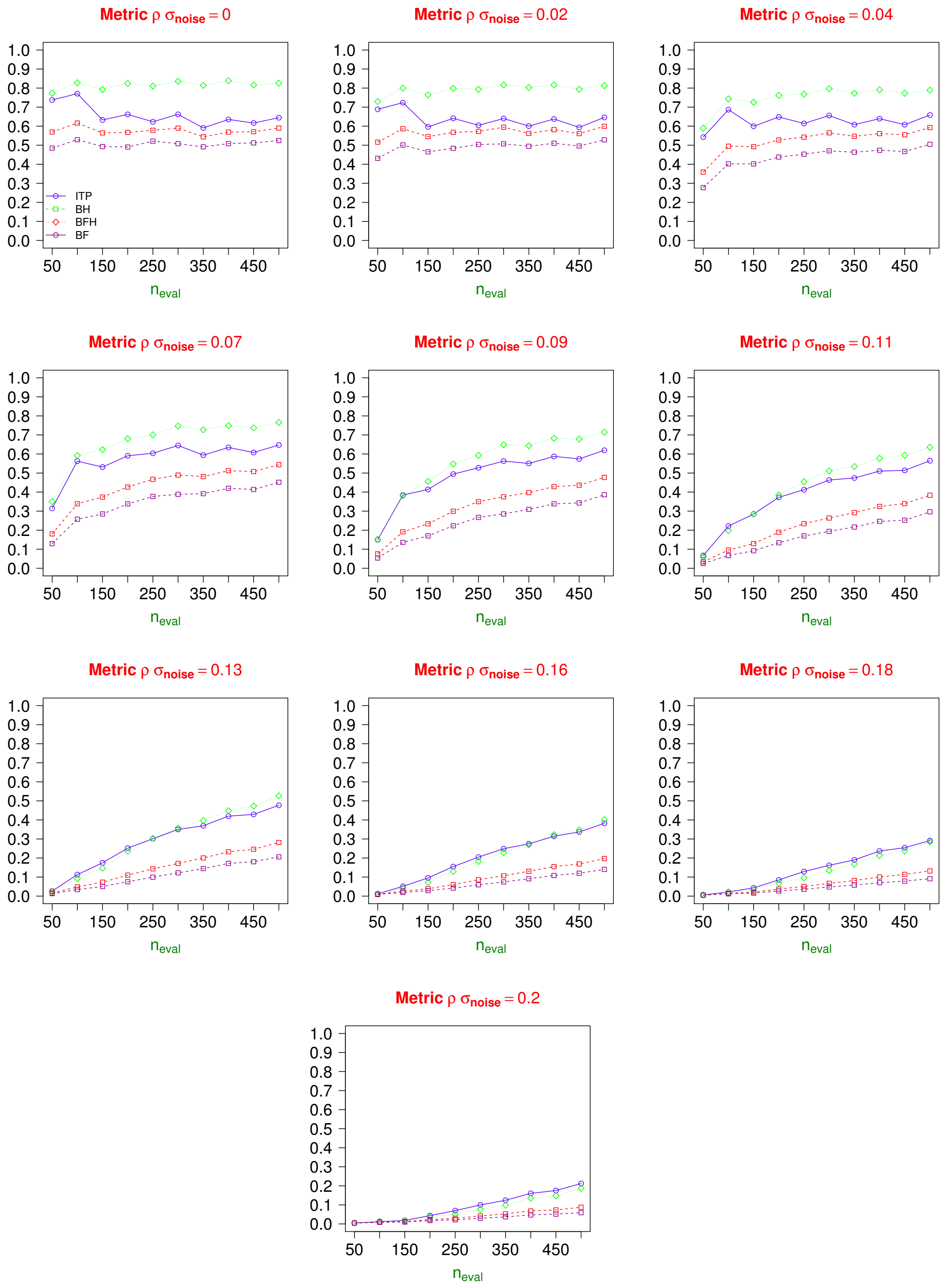

Figure B.4: Rejection Rate False null hypotheses ρ as a function of the number of evaluations $n_{eval}$with fixed standard deviation of noise σ_{noise}. The data type is populations of constant and step functions.

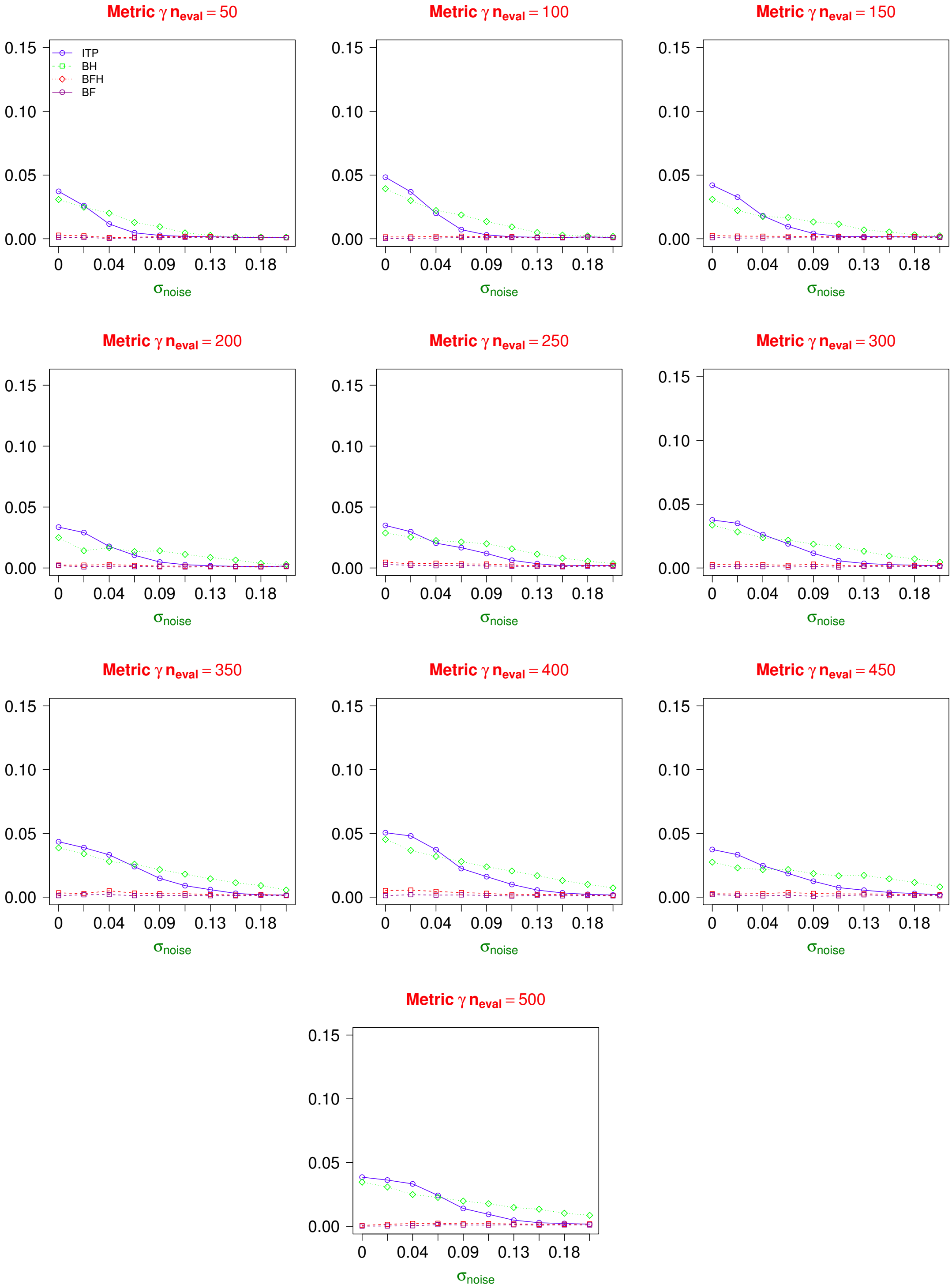

Figure B.5: Rejection Rate True null hypotheses γ as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and step functions.

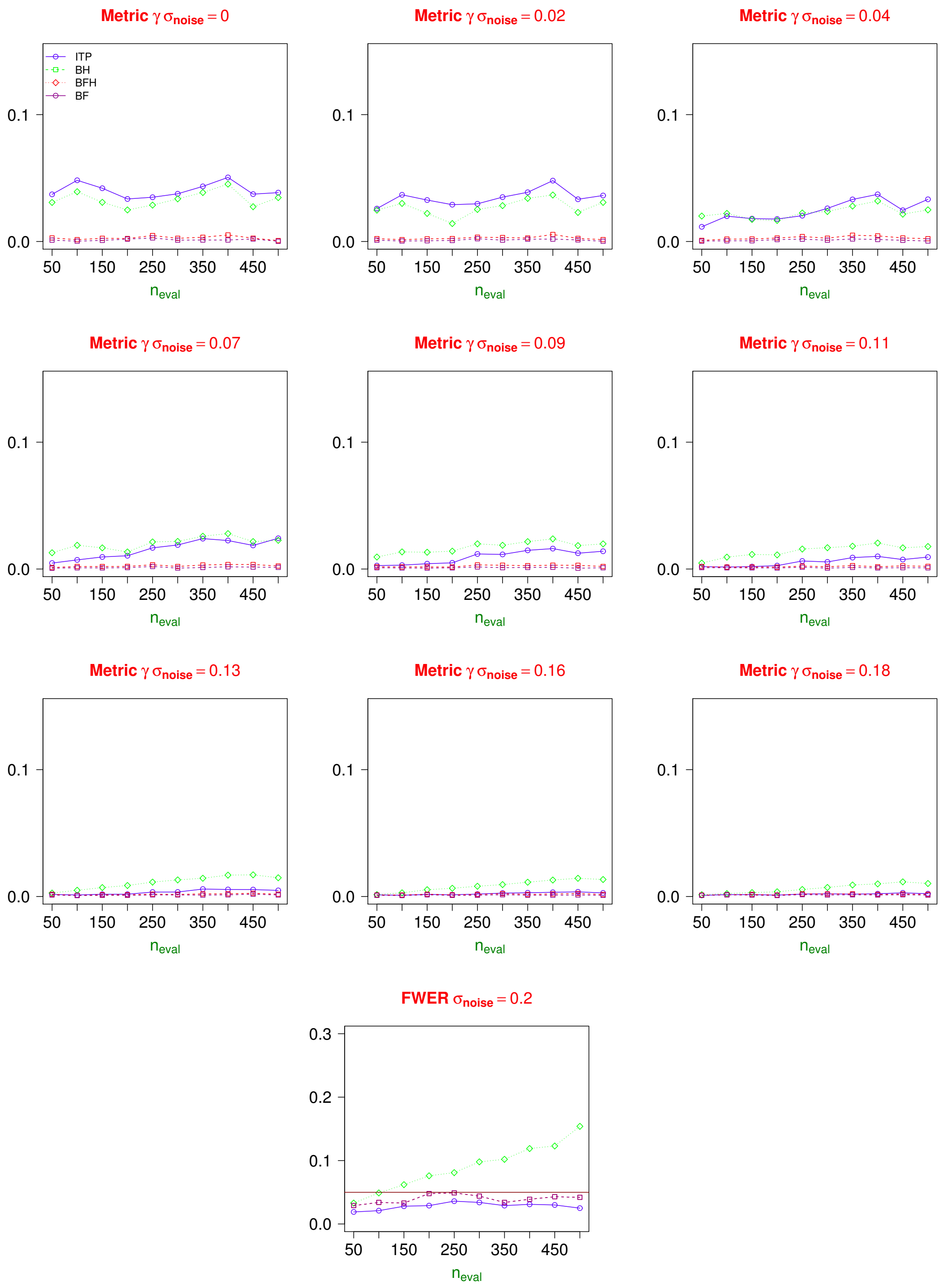

Figure B.6: Rejection Rate True null hypotheses γ as a function of the number of evaluations $n_{eval}$with fixed standard deviation of noise σ_{noise}. The data type is populations of constant and step functions.

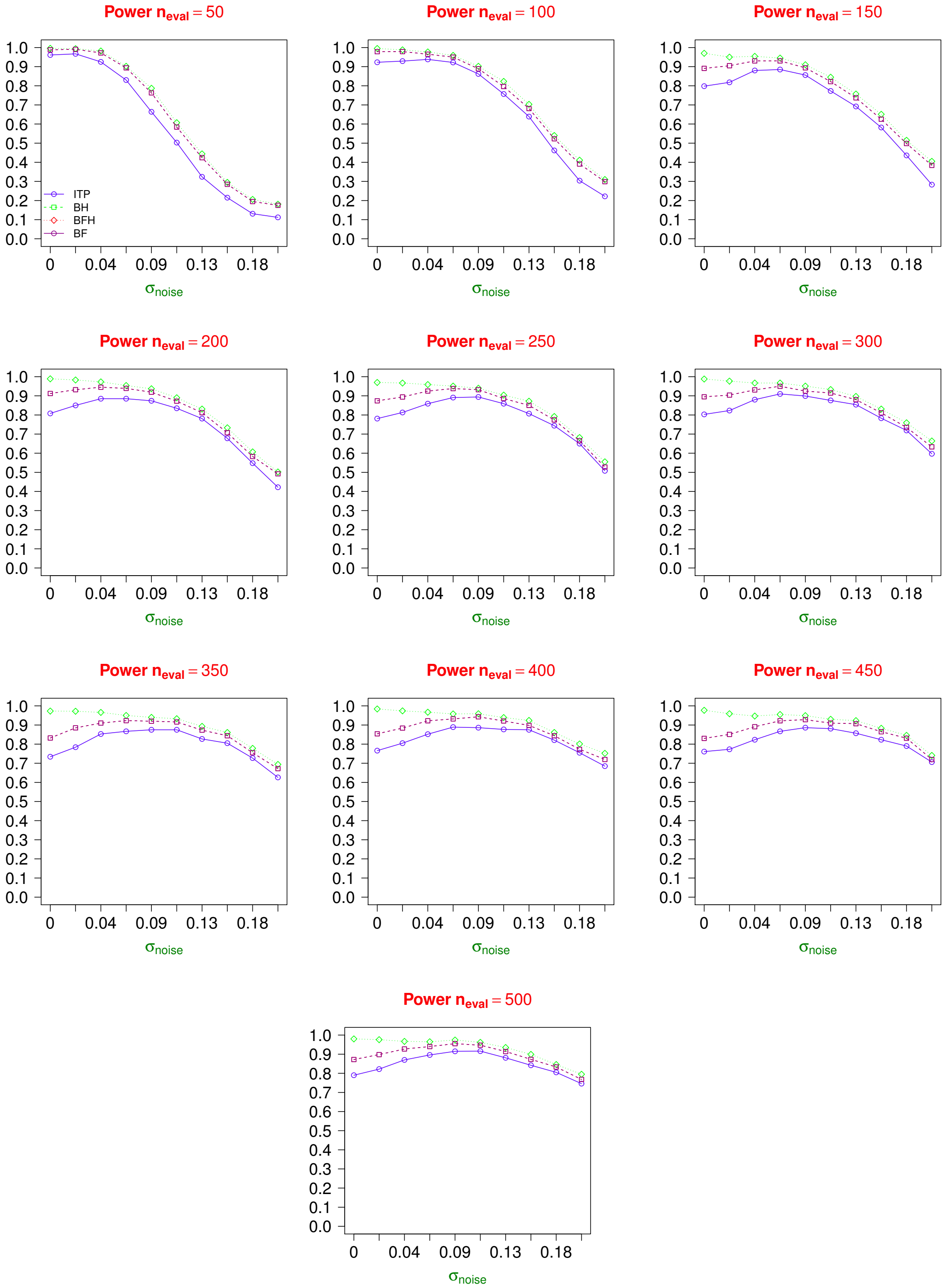

Figure B.7: Power π as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and step functions.

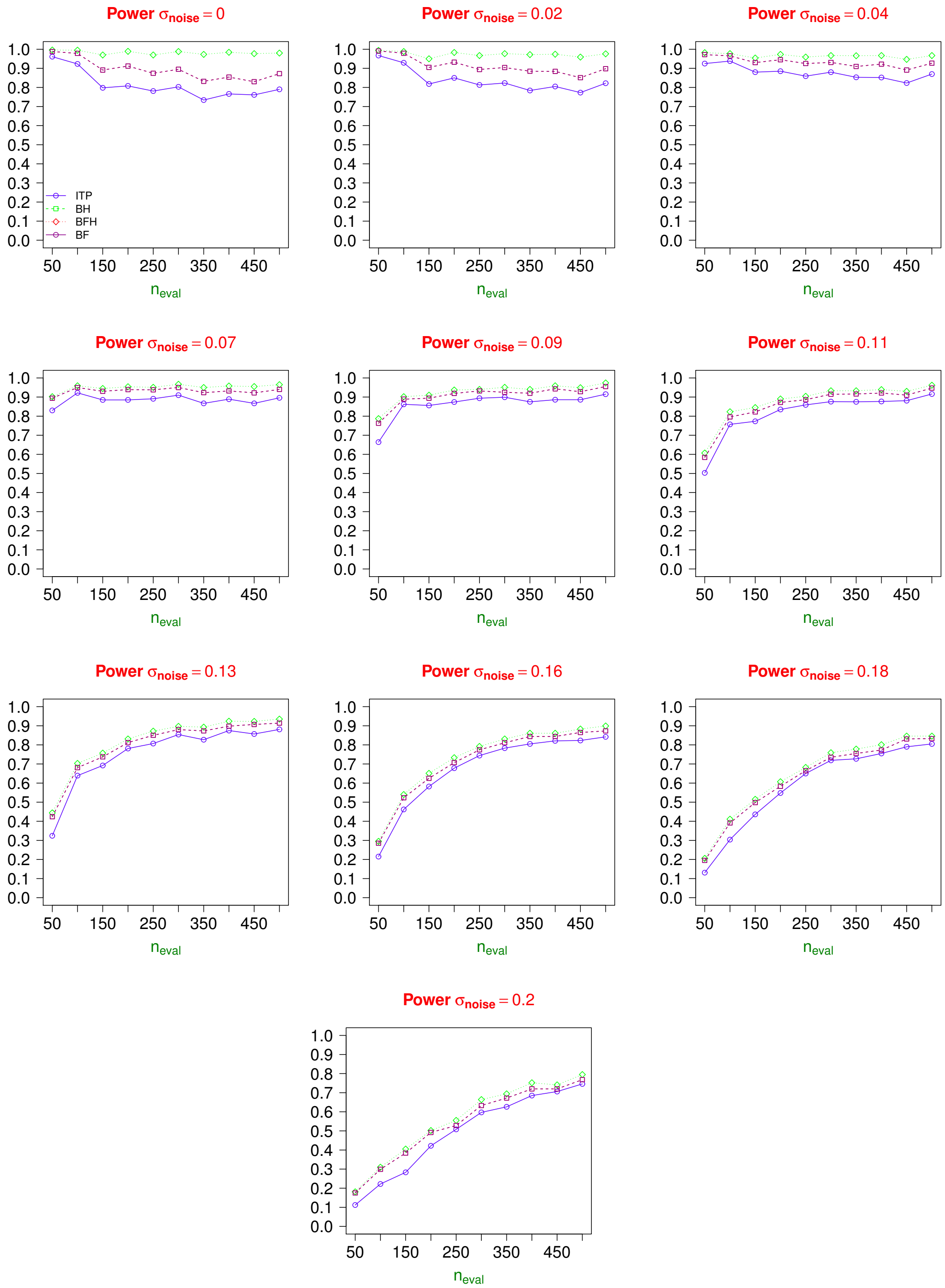

Figure B.8: Power π as a function of the number of evaluations n_{eval} with fixed standard deviation of noise σ_{noise}. The data type is populations of constant and step functions.

B.2 Graphical results data set const-tricube

In Figures B.9, B.10 it is reported the $FWER$ as a function of σ_{noise} with n_{eval} fixed and as a function of n_{eval} with σ_{noise} fixed, respectively.

In Figures B.11, B.12 it is reported the Rejection Rate False null hypotheses ρ as a function of σ_{noise} with n_{eval} fixed and as a function of n_{eval} with σ_{noise} fixed, respectively.

In Figures B.13, B.14 it is reported the Rejection Rate True null hypotheses γ as a function of σ_{noise} with n_{eval} fixed and as a function of n_{eval} with σ_{noise} fixed, respectively.

In Figures B.15, B.16 it is reported the Power π as a function of σ_{noise} with n_{eval} fixed and as a function of n_{eval} with σ_{noise} fixed, respectively.

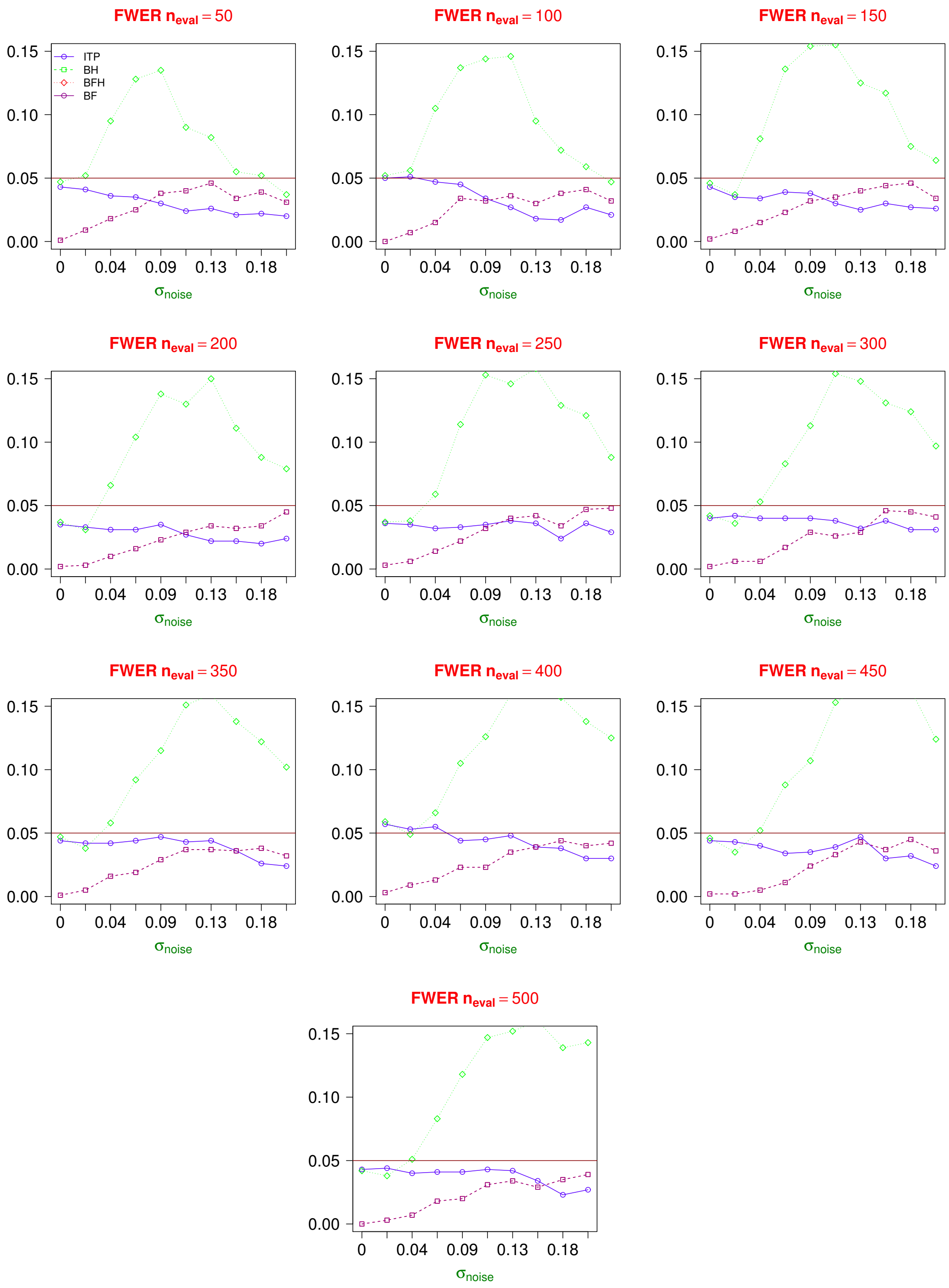

Figure B.9: Family Wise Error Rate $FWER$ as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and tricube functions.

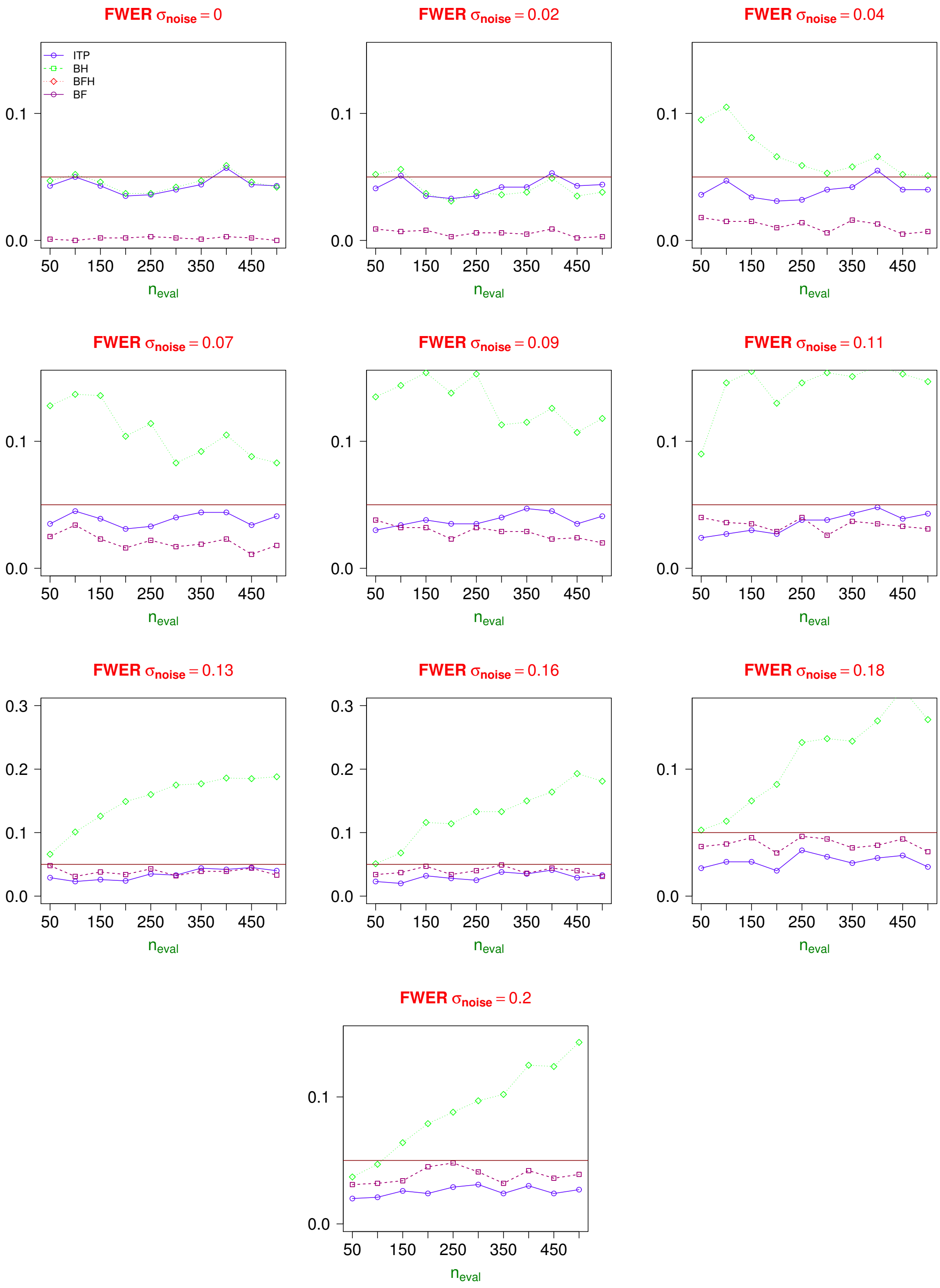

Figure B.10: Family Wise Error Rate $FWER$ as a function of the number of evaluations $n_{eval}$with fixed standard deviation of noise σ_{noise}. The data type is populations of constant and tricube functions.

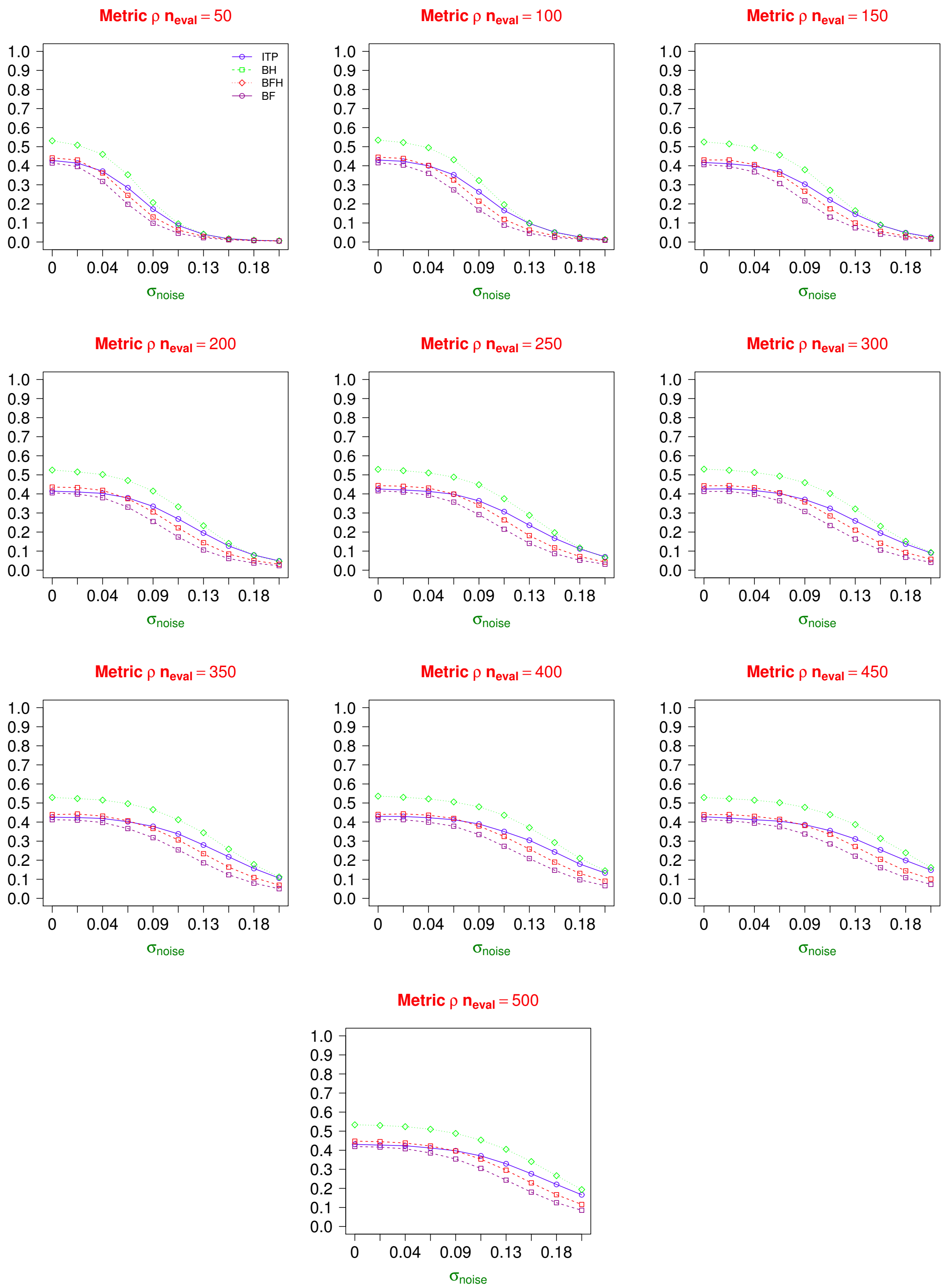

Figure B.11: Rejection Rate False null hypotheses ρ as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and tricube functions.

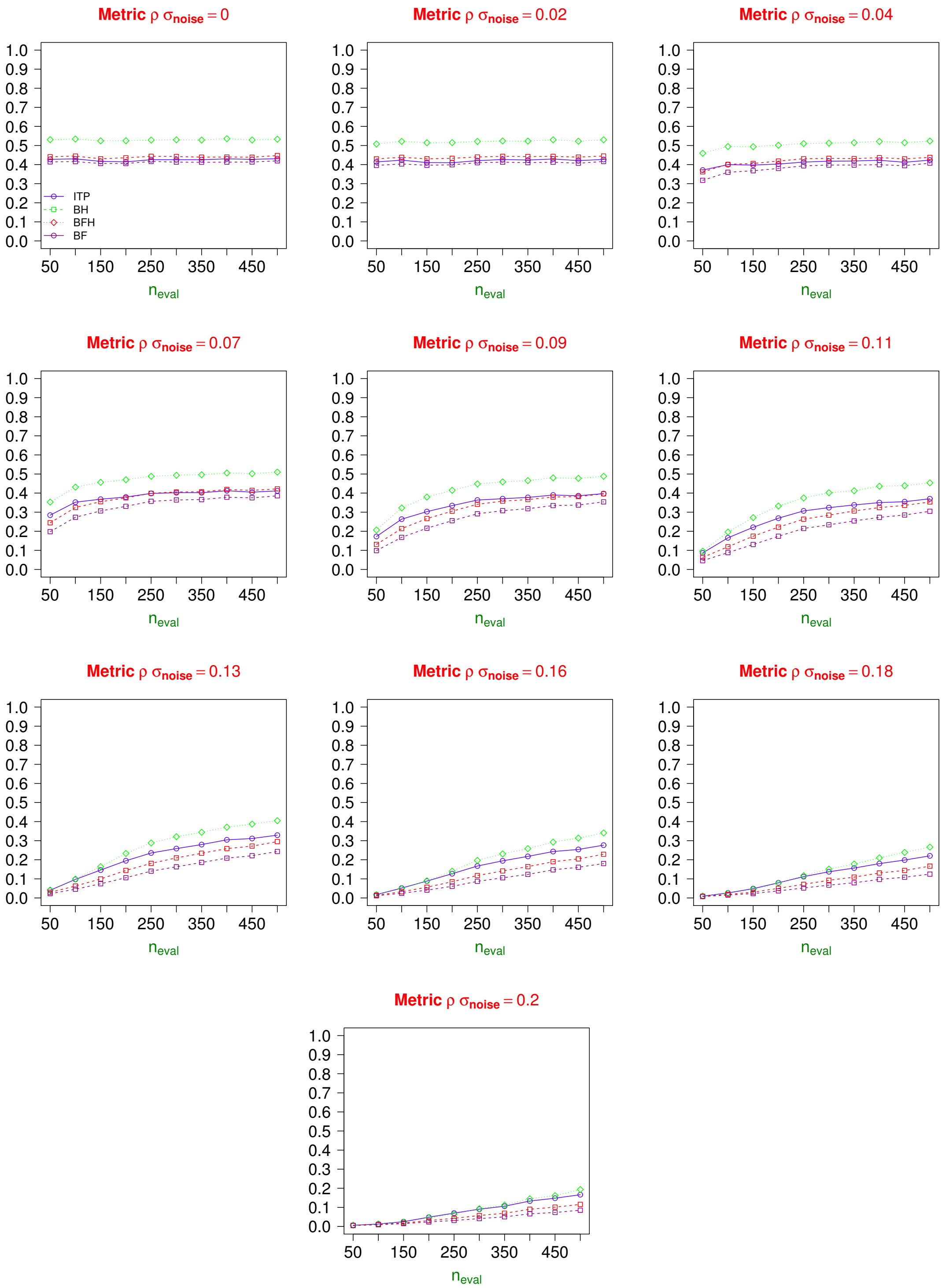

Figure B.12: Rejection Rate False null hypotheses ρ as a function of the number of evaluations $n_{eval}$with fixed standard deviation of noise σ_{noise}. The data type is populations of constant and tricube functions.

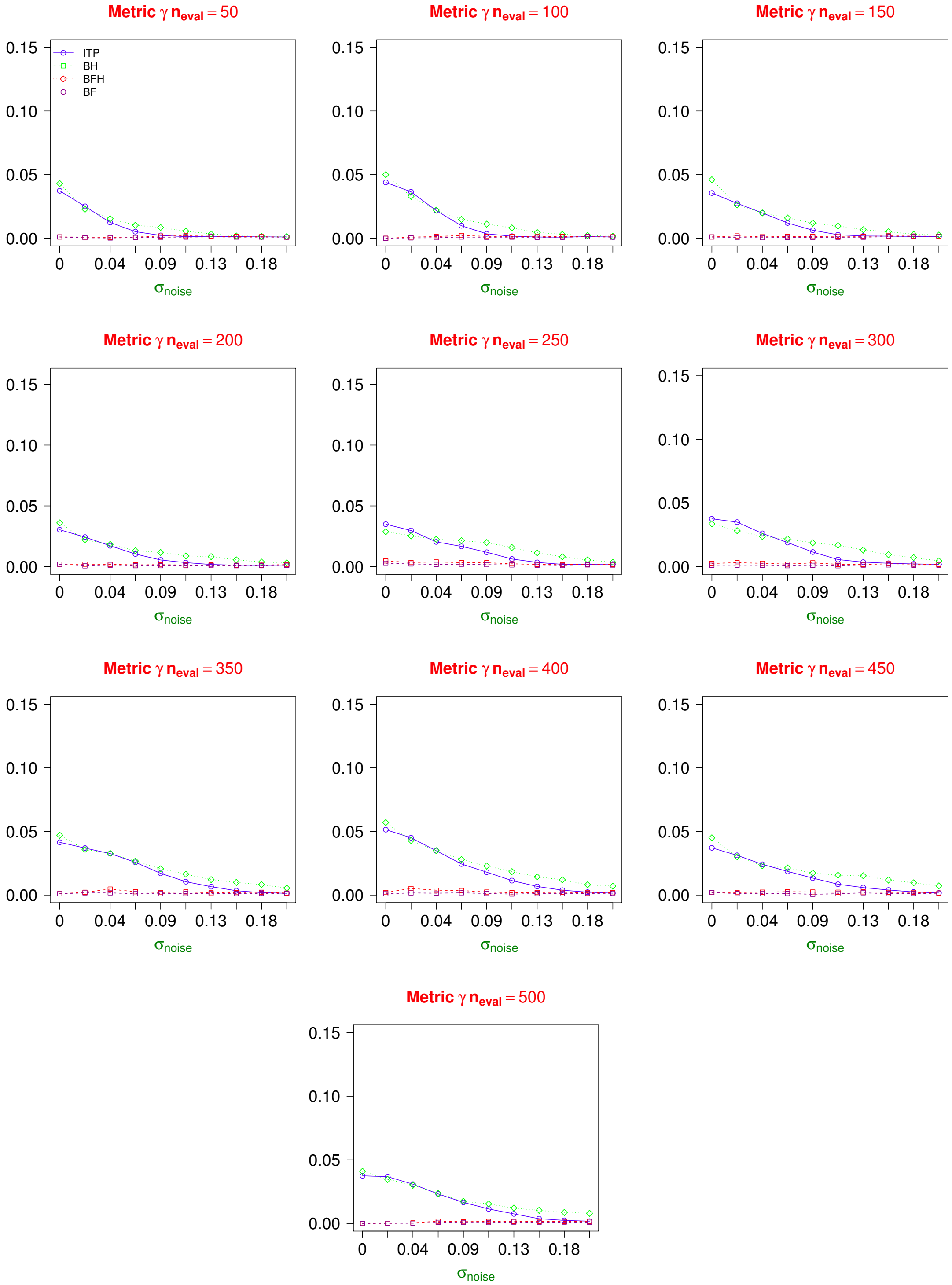

Figure B.13: Rejection Rate True null hypotheses γ as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and tricube functions.

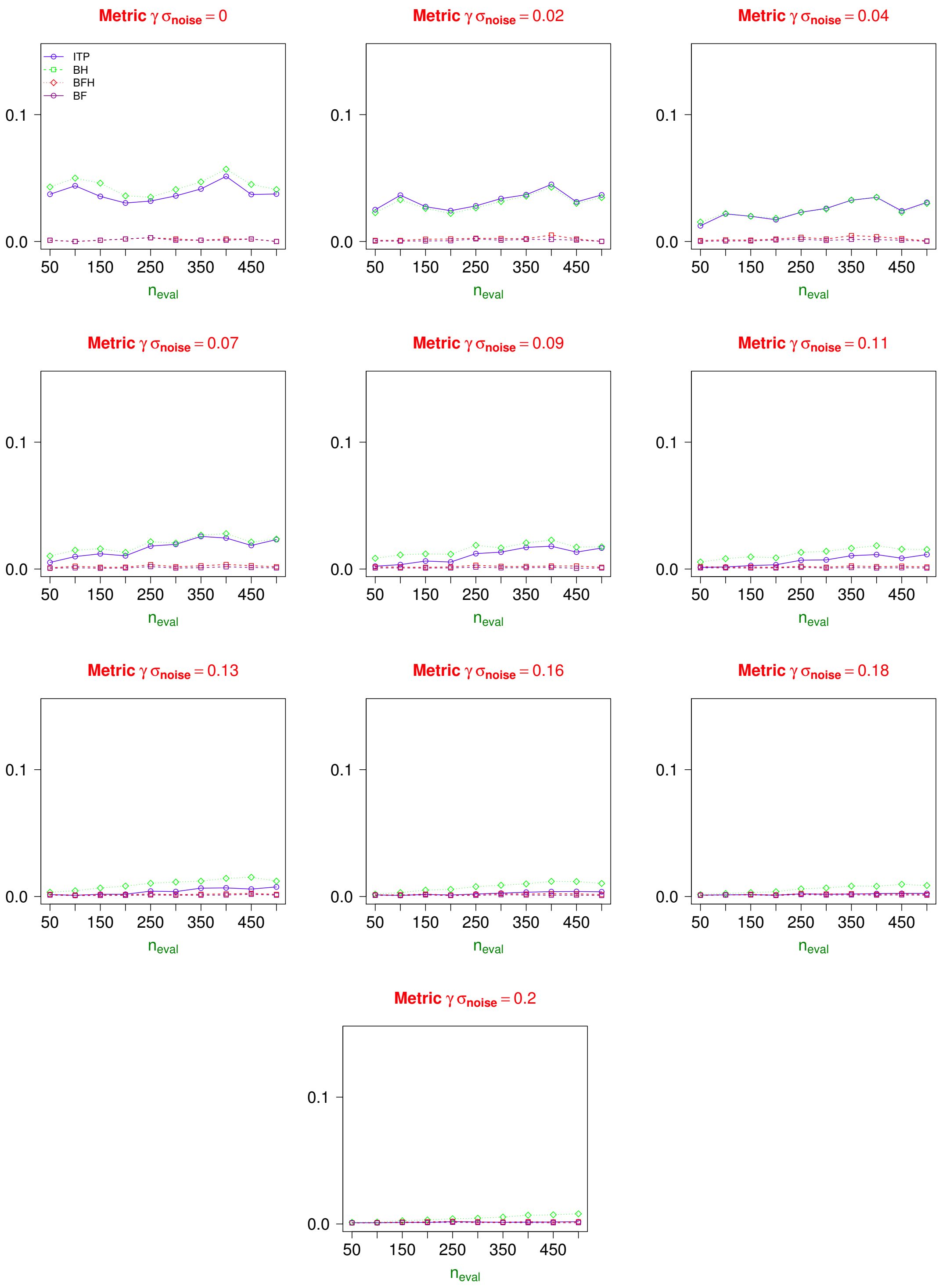

Figure B.14: Rejection Rate True null hypotheses γ as a function of the number of evaluations n_{eval} with fixed standard deviation of noise σ_{noise}. The data type is populations of constant and tricube functions.

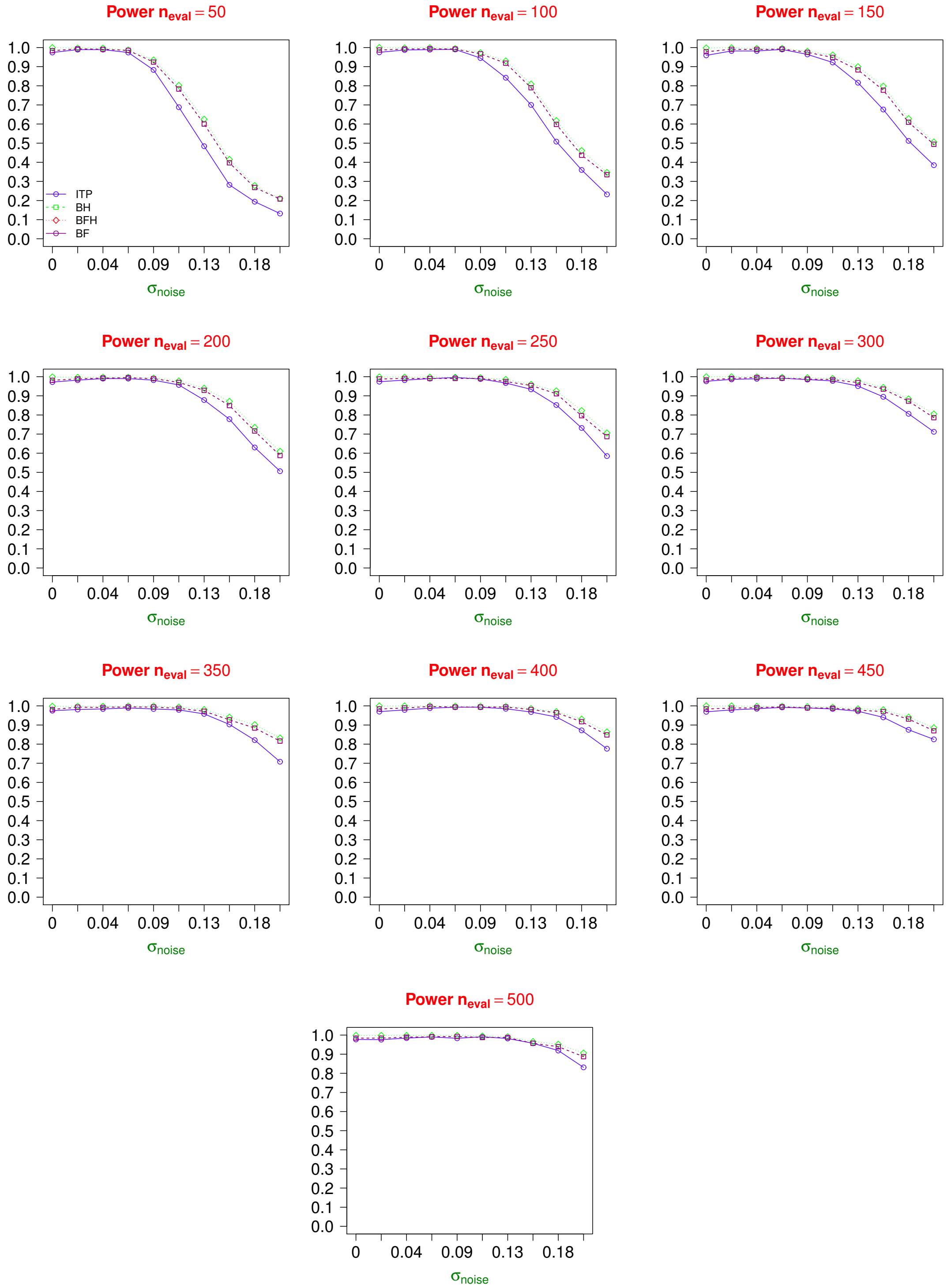

Figure B.15: Power π as a function of the standard deviation of noise σ_{noise} with number of evaluations n_{eval} fixed. The data type is populations of constant and tricube functions.

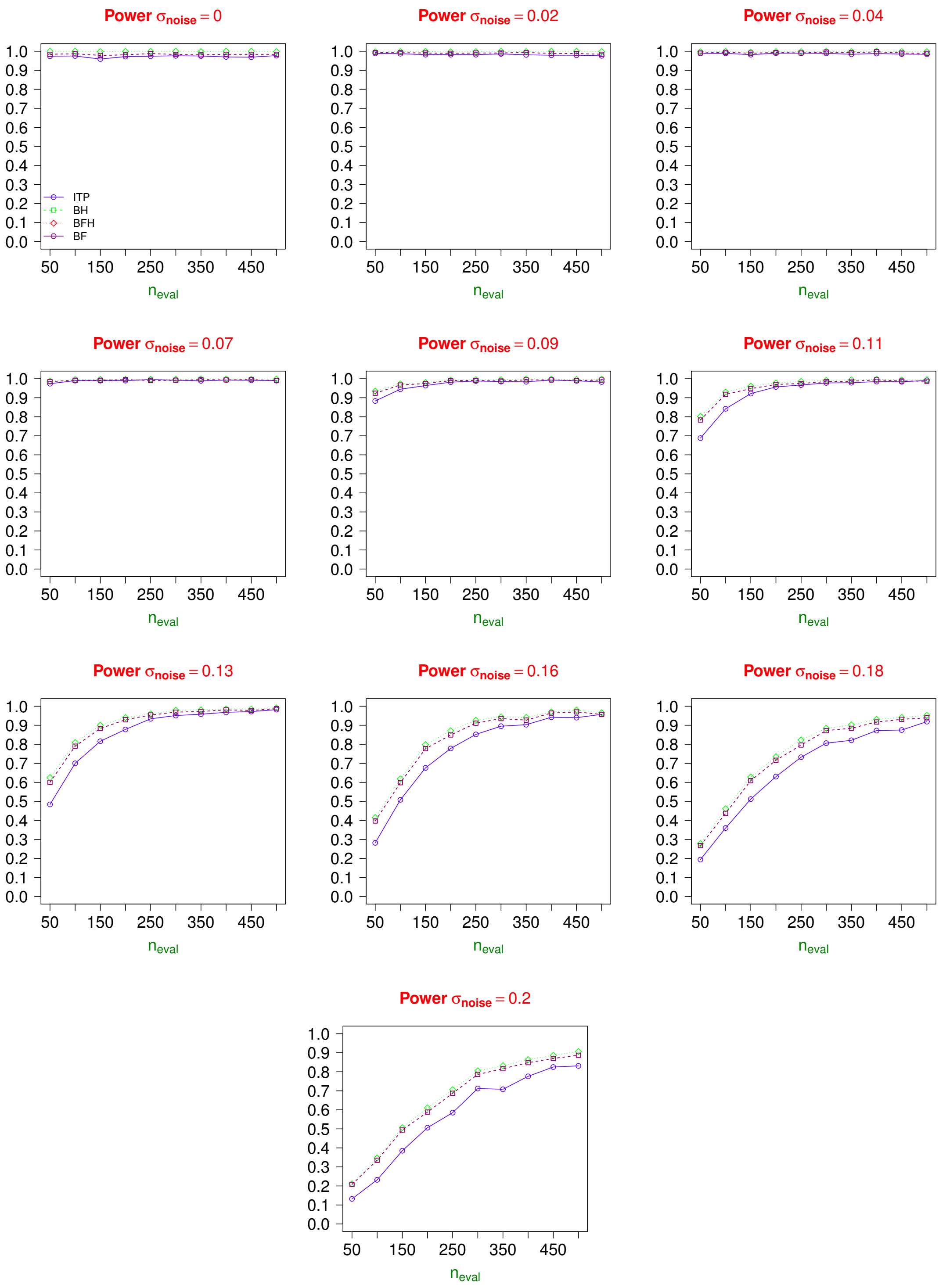

Figure B.16: Power π as a function of the number of evaluations n_{eval} with fixed standard deviation of noise σ_{noise}. The data type is populations of constant and tricube functions.