Tree based diagnostic procedures following a smooth test of goodness-of-fit

Gilles R. Ducharme, Walid Al Akhras

Noname manuscript No.
(will be inserted by the editor)

Tree based Diagnostic Procedures Following a Smooth Test of Goodness-of-Fit

Gilles R. Ducharme,Walid Al Akhras

the date of receipt and acceptance should be inserted later

Abstract This paper introduces a statistical procedure, to be applied after a goodness-of-fit test has rejected a null model, that provides diagnostic information to help the user decide on a better model. The procedure goes through a list of departures, each being tested by a local smooth test. The list is organized into a hierarchy by seeking answers to the questions *"Where is the problem ?"* and "*What is the problem there ?*. This hierarchy allows to focus on finer departures as the data becomes more abundant. The procedure controls the family-wise Type 1 error rate. Simulations show that the procedure can succeed in providing useful diagnostic information.

1 Introduction

Let X be a continuous random variable with unknown density f over support $\mathcal{X}$. We consider the problem of testing $H_0 : f = f_0$ where f_0 is entirely specified. There is no loss of generality in setting $\mathcal{X} = [0, 1]$ and f_0 as the uniform $U(0, 1)$ density. A sample of independent copies $\mathcal{E} = \{X_1, ..., X_n\}$ is available to assess whether H_0 holds.

This is accomplished by goodness-of-fit (GoF) tests. When a GoF test fails to reject H_0, with the risk of both types of error deemed adequate, the null model f_0 may be acted upon with some confidence. However if the GoF test rejects, a better model, say f_1, must be determined.

One problem with many GoF tests is that after rejection, they provide little information as to what aspects of f_0 are contradicted by the data. In some cases, qualitative informations about the GoF test (e.g. "this tests is good at detecting asymmetry") or plots of an estimator of f can be exploited to suggest some features f_1 should possess.

Institut Montpellierain Alexander Grothendieck, CC051, Université de Montpellier, Place Eugène Bataillon, 34095, Montpellier, Cedex 5, France.
Corresponding author : gilles.ducharme@umontpellier.fr
Tel : 33 (0)4 67 14 35 69 FAX : 33 (0)4 67 14 35 58

A GoF test that can provide solid (i.e. controls the risk of errors) and valid (i.e. not misleading) information about the aspects of f_0 contradicted by the data is said to possess diagnostic (Dx) capabilities. One such test, popular in the applied literature, is Pearson's chi-square test which can help in identifying the parts of $\mathcal{X}$ where the data contradict f_0. Another test having Dx capabilities is Neyman's (Rayner & Best 1989) smooth test; after rescaling, its components yield solid and valid Dx information about the moments of f and thus clues about how its shape departs from f_0.

The present work introduces a Diagnostic Extraction Procedure (DxEP) to get solid, valid and more informative Dx information following the rejection of H_0. This DxEP goes further than the above existing approaches. It proceeds by seeking answers to the questions *: "Where (in $\mathcal{X}$) are the GoF problems ?*" and then "*What causes the GoF problem there ?*", i.e. localize and then identify the nature of the local departures, to identify the aspects of f_0 that need repairing. It tests these departures using local versions of the smooth test. To control the Type 1 family-wise error rate while reducing the resulting power loss, it uses the closure principle and organizes the null hypotheses and their test into a tree of trees hierarchical structure that can be adapted to the available information : as the size of $\mathcal{E}$ increases, finer areas of $\mathcal{X}$ can be explored and more subtle departures can be investigated. The DxEP can thus provide information leading to a better choice for f_1 while controlling the risk of repairing features of f_0 that do not need to be repaired.

The paper is organized as follows. In Section 2, the Dx information provided by Pearson's chi-square and Neyman's smooth tests are reviewed. Section 3 introduces the local versions of the smooth test and explains how to derive valid Dx information from the magnitude of its components. Section 4 presents the tree of trees structure of null hypotheses and explains how to control the Type 1 family-wise error rate. Section 5 presents some simulations to understand the behavior of the DxEP. It is seen that our procedure can provide solid, valid and informative Dx information that help in suggesting a better f_1. A conclusion closes the paper.

2 Two existing diagnostic extraction procedures (DxEP)

Let $0 = a_0 < a_1 < \cdots < a_K = 1$ generate the partition $\{P_k = (a_{k-1}, a_k), k = 1, ..., K\}$ of $\mathcal{X}$. Unless specified otherwise and to avoid notation problems, intervals are noted in parenthesis, leaving the user decide whether they are open, closed, etc. To test $H_0 : X \sim f_0 = U(0,1)$, Pearson's chi-square test statistic takes the form

$$\chi^2 = \sum_{k=1}^{K} \frac{(N_k - np_k)^2}{np_k} = \sum_{k=1}^{K} C_k^2, \tag{2.1}$$

where $N_k = \sum_{i=1}^{n} \mathbb{I}\{X_i \in P_k\}$, $\mathbb{I}\{A\}$ is the indicator of event A and $p_k = \mathbb{P}_{H_0}[X \in P_k]$. Under H_0, χ^2 is asymptotically χ^2_{K-1}, the chi-square distribution with $K-1$ degrees of freedom.

The Dx capabilities of Pearson's statistic are sometimes exploited in the applied literature, e.g. von Eye & Bogat (2004). The component C_k^2 can help in answering the question *"Where is the GoF problem ?"* : a large C_k^2 indicates an unexpected amount of data in P_k under H_0. But it is difficult to extract more Dx information. For instance, it is wrong to compare C_k^2 to a χ_1^2. Controlling Type 1 errors is impaired by the dependence between the C_k^2. This points to a subtle problem with a DxEP based on these C_k^2 : consider $P_1 = (0, \frac{1}{2}), P_2 = (\frac{1}{2}, 1)$ and $f(x) = (2 - \frac{8}{3}x)\mathbb{I}\{x \in P_1\} + \frac{2}{3}\mathbb{I}\{x \in P_2\}$. If n is large enough, C_1^2 and C_2^2 will lead to the Dx that the $U(0,1)$ fits the data nowhere. However, only the part of f_0 in P_1 needs repairing, as on P_2, $f(x) = \frac{2}{3} \neq 1$ because a density must integrate to one. This shows that a DxEP should aim at identifying the P_k where the conditional density differs from uniformity on P_k.

Variants of C_k^2 solve some of these problems. Rayner and Best (1989) give a decomposition of χ^2 into $\sum_{k=1}^{K-1} V_k^2$, where the V_k^2 are asymptotically independent χ_1^2 under H_0. This allows control over the Type 1 errors, but being linear combinations of the p_k, the V_k^2 detect departures from the moments of the joint distribution of the N_k under H_0. Thus, they provide answers to the global question " *What causes the GoF problem ?*", as moments of a distribution are related to its global properties.

If moments are to produce Dx information, it seems better to use the raw data in $\mathcal{E}$ instead of the categorized N_k. This leads to Neyman's smooth test for $H_0 : X \sim U(0,1)$ (for details and an explanation regarding the term "smooth", see Rayner and Best 1989). The test statistic takes the form

$$\mathcal{R}_M = \mathcal{L}_1^2 + \mathcal{L}_2^2 + ... + \mathcal{L}_M^2 \tag{2.2}$$

and, under H_0, $\mathcal{R}_M$ is asymptotically χ_M^2. In (2.2), $\mathcal{L}_m = n^{-1/2} \sum_{i=1}^n L_m(X_i)$ where $L_m(\cdot)$ is the orthonormalized on (0,1) Legendre polynomial of degree m. In particular, $\mathcal{L}_1 = \sqrt{12n}\left(\bar{X} - \frac{1}{2}\right)$ so a large $\mathcal{L}_1^2$ contradicts H_0 because $\bar{X}$ is far from the value $\frac{1}{2}$ expected under H_0. Likewise $\mathcal{L}_2 = 6\sqrt{5n}\left(S^2 - \frac{1}{12}\right)$, where $S^2 = n^{-1}\sum_{i=1}^n (X_i - \frac{1}{2})^2$. A large $\mathcal{L}_2^2$ does not support H_0 because the empirical variance departs from $\frac{1}{12}$. Similarly, $\mathcal{L}_3^2$ is related to the skewness coefficient and thus carries Dx information about the asymmetry of f. Finally, $\mathcal{L}_4^2$ is related to the kurtosis and allows detection of departures in the tails of distributions. Thus, one heuristic is that large values of $\mathcal{R}_M$ are indicative of non-uniformity because some empirical moments differ significantly from those expected under H_0. We refer to such insights about the form of f as moment-based Dx information. However using these $\mathcal{L}_m^2$ to provide answers to the question " *What causes the GoF problem* ?" is trickier. Henze & Klar (1996), Henze (1997) and Klar (2000) have showed that to provides valid Dx information, the $\mathcal{L}_m^2$ must be rescaled; see Section 3.2.

Using the components of χ^2 or $\mathcal{R}_M$ in a DxEP involves multiple comparisons to reference distributions and this increases the Type 1 error rate. Moreover, large components attract the attention but because they are identified after observing the data, the Type 1 error is distorted (see Henze 1997). Thus a form of family-wise error rate must be kept under control.

3 Dx information and the smooth test

3.1 Local smooth tests

A local version of Neyman's smooth test drives our DxEP. Let $Q = (a, b) \subseteq \mathcal{X}$. Recall that $X \sim f$ with cumulative distribution function (CDF) F and write $F(Q) = F(b) - F(a)$. Let $f^Q(x) = f(x)/F(Q)\mathbb{I}\{x \in Q\}$ be the density of $(X \mid X \in Q)$. Consider the problem of testing

$$H_0^Q : (X \mid X \in Q) \sim U(Q), \tag{3.1}$$

where $U(Q)$ is the uniform distribution over Q. The data available for this test are those in $\mathcal{E}^Q = \{X_i \in \mathcal{E} \bigcap Q\}$ with random sample size N^Q. Our test statistic is based on the $L_m(\cdot)$. Transform X into $X^* = F_0^Q(X)$, where $F_0^Q(\cdot)$ is the CDF of the $U(Q)$. Under H_0^Q, X^* has a mixed density over [0,1] with masses $F(a)$ at $x = 0$ and $1 - F(b)$ at $x = 1$. To get rid of these unknown masses, introduce the version $L_m^*(x) = L_m(x)\mathbb{I}\{0 < x < 1\}$ of the orthonormalized Legendre polynomials and consider

$$\mathcal{L}_m^Q = \frac{1}{\sqrt{N^Q}} \sum_{X_i \in \mathcal{E}^Q} L_m^*(X_i^*). \tag{3.2}$$

In some situations, it may be useful if the endpoints of Q can be defined by some characteristics of the unknown f. For example, let $0 \leq \alpha_1 < \alpha_2 \leq 1$ and suppose that $a = F^{-1}(\alpha_1)$, $b = F^{-1}(\alpha_2)$. Because f is unknown, these need to be estimated, e.g. by $\hat{a} = \hat{F}_n^{-1}(\alpha_1)$ and $\hat{b} = \hat{F}_n^{-1}(\alpha_2)$ where $\hat{F}_n(\cdot)$ is the empirical CDF from $\mathcal{E}$, to yield $\hat{Q} = (\hat{a}, \hat{b})$. To test (3.1) consider

$$\mathcal{L}_m^{\hat{Q}} = \frac{1}{\sqrt{N^{\hat{Q}}}} \sum_{X_i \in \mathcal{E}^{\hat{Q}}} L_m^*(\hat{X}_i^*), \tag{3.3}$$

where $\hat{X_i}^* = F_0^{\hat{Q}}(X_i)$ and $F_0^{\hat{Q}}$ is defined as F_0^Q with (a, b) replaced by $(\hat{a}, \hat{b})$. This involves a two stage procedure where, from the data in $\mathcal{E}$, $\hat{a}$ and $\hat{b}$ are first computed. Then $\mathcal{E}^{\hat{Q}} = \{X_i \in \mathcal{E} \bigcap \hat{Q}\}$ of size $N^{\hat{Q}}$ is identified to get the component $\mathcal{L}_m^{\hat{Q}}$. The following result, whose proof is differed to the Appendix, gives the behavior of the local smooth test statistics based on the components $\mathcal{L}_m^Q$ and $\mathcal{L}_m^{\hat{Q}}$.

Theorem 1 *Suppose that* $(\hat{a}, \hat{b}) = (a, b) + O_p(n^{-1/2})$. *Under* H_0^Q, $\mathcal{R}_M^Q = \sum_{m=1}^M \left(\mathcal{L}_m^Q\right)^2$ *and* $\mathcal{R}_M^{\hat{Q}} = \sum_{m=1}^M \left(\mathcal{L}_m^{\hat{Q}}\right)^2$ *are asymptotically* $\mathcal{X}_M^2$.

Remark 1 Taking $Q = \mathcal{X}$ gives back Neyman's original $\mathcal{R}_M$ and the argument leading to the $\mathcal{X}_M^2$ is standard. $\mathcal{R}_M^Q$ requires a slightly more involved argument as the randomness of N^Q needs to be accommodated. The real difficulty is with $\mathcal{R}_M^{\hat{Q}}$. Theorem 1 states that its asymptotic distribution is not affected by the estimation of Q. This is reminiscent of the behavior of Pearson's chi-square test (2.1) using data-dependent cells (Pollard 1979).

3.2 Dx Information from smooth tests

The $\mathcal{L}_m^2$ in (2.2) must be rescaled to yield valid Dx information. To see this, let $\boldsymbol{\mu} = (\mu_1, ..., \mu_M)^T$ with components $\mu_m = \int_0^1 L_m(x) f(x) dx$. Dx informations arising from the magnitude of $\mathcal{L}_m^2$ indicate whether $\mu_m =$ or $\neq 0$. Define the $M \times M$ matrix $\Sigma = (\sigma_{mm'})$ where $\sigma_{mm'} = \int_0^1 L_m(x) L_{m'}(x) f(x) dx - \mu_m \mu_{m'}$, with $\Lambda = \text{Diag}\{\lambda_1, \ldots, \lambda_M\}$ being the diagonal matrix of its eigenvalues and P the matrix of its orthonormalized eigenvectors. It is shown in Inglot et al. (1994, Theorem 2.1 and eq. (5)) that the power of Neyman's smooth test can be uniformly approximated to within $O(n^{-1/2})$ by the $\sum_{m=1}^M \lambda_m \chi_1^2(\nu_m^2)$ distribution, where the non-centrality parameters ν_m are the components of $\sqrt{n}\Lambda^{-1/2} P \boldsymbol{\mu}$. Specializing, $\mathcal{L}_m^2$ is approximately $\sigma_{mm}\chi_1^2(n\mu_m^2/\sigma_{mm})$.

When σ_{mm} is small, $\sigma_{mm}\chi_1^2(n\mu_m^2/\sigma_{mm})$ may be stochastically closer to 0 than the reference χ_1^2 even if $\mu_m \neq 0$. Thus a small $\mathcal{L}_m^2$ may lead to the wrong Dx that $\mu_m = 0$. By contrast, if σ_{mm} is large but $\mu_m = 0$, then $\mathcal{L}_m^2$ will tend to be large, possibly leading to the wrong Dx that $\mu_m \neq 0$.

To correct this, Klar (2000) considers the rescaled test statistic :

$$\mathcal{K}_M = \boldsymbol{L}_M^T \hat{\Sigma}^{-1} \boldsymbol{L}_M, \tag{3.4}$$

where $\boldsymbol{L}_M = (\mathcal{L}_1, \mathcal{L}_2, \ldots \mathcal{L}_M)^T$ and $\hat{\Sigma}$ is an estimator of Σ. The power of $\mathcal{K}_M$ is approximated by a $\chi_M^2(n\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu})$. For the single component m, this reduces to $\mathcal{K}_{\{m\}} = \mathcal{L}_m^2/\hat{\sigma}_{mm}$, which is approximately $\chi_1^2(n\mu_m^2/\sigma_{mm})$ and thus provides valid Dx information about μ_m, as a $\chi_1^2(n\mu_m^2/\sigma_{mm})$ is stochastically larger than a χ_1^2.

With the local versions of the smooth test, similar Dx information about $f^Q(\cdot)$ can be extracted from the rescaled $\mathcal{L}_m^Q$. A large $\mathcal{K}_{\{m\}}^Q = \left(\mathcal{L}_m^Q/\hat{\sigma}_{mm}^Q\right)^2$, where $(\hat{\sigma}_{mm}^Q)^2 = \frac{1}{N^Q} \sum_{X_i \in \mathcal{E}^Q} \left(L_m^*(X_i^*) - \bar{L}_m^*\right)^2$, indicates that the data do not support $\mu_m^Q = \int_Q L_m^*(F_0^Q(x)) f^Q(x) dx = 0$ under H_0^Q (local moment-based Dx).

From $\mathcal{K}_{\{m\}}^Q$, one can also extract interesting Dx information about the shape of f^Q. Approximate

$$f^Q(x) \approx g_M^Q(\cdot; \boldsymbol{\theta}) = \left(1 + \sum_{m=1}^M \theta_m L_m^*(F_0^Q(x))\right) \tag{3.5}$$

and observe that $\mu_m^Q \simeq \theta_m$. If $\theta_m = 0$ for all $m > m^*$, then f^Q is nearly a polynomial of degree m^*. If $m^* = 1$, this gives the Dx information that f^Q roughly exhibits a linear trend. If $m^* = 2$, f^Q is close to a second degree polynomial, etc. We refer to such Dx information about the shape of f^Q as being projection-based.

Note finally that, as the length of Q decreases, approximation (3.5) becomes more plausible, and the resulting Dx sharper, because the smoothness of f induces f^Q to behave locally as a low order polynomial. See the examples in Subsection 5.2.

4 A Tree of trees structured DxEP

4.1 Answering the question "*Where is the problem ?*"

Suppose $H_0 : X \sim U(0,1)$ has been rejected. We proceed to answer the question "*Where is the problem ?*". For this to be make sense, assume that external information ensures that the true f can be segmented into locally approximately uniform and non-uniform parts, with the goal of identifying the non-uniform intervals. Let the user-supplied partition $\{P_k = (a_{k-1}, a_k), k = 1, ..., K\}$ of $\mathcal{X}$ be given, more or less corresponding to these parts (throughout this section we refer to fixed P_k but the results hold for the estimated $\hat{P}_k$). We proceed to test all local hypotheses $H_0^{P_k} : (X \mid X \in P_k) \sim U(P_k)$ using local smooth tests. Note that $H_0 \subseteq \bigcap_{k=1}^{K} H_0^{P_k}$.

In testing this family of hypotheses, errors must be controlled. Our DxEP focusses on the Type 1 family-wise error rate (T1-FWER) which is the probability of falsely rejecting at least one of the $H_0^{P_k}$. Let $\mathcal{K}_0 = \{k \mid H_0^{P_k} \text{ is true}\}$ and denote by $\mathbb{P}_{\mathcal{K}_0}[A]$ the probability of event A computed under all null hypotheses in $\mathcal{K}_0$. This set being unknown, we require a *strong* control of the T1-FWER : for any $\mathcal{K}_0$, $\mathbb{P}_{\mathcal{K}_0}[\text{at least one } H_0^{P_k} \text{ is falsely rejected}] < \alpha$. This can be done via a number of approaches, e.g. the Bonferroni method which tests each $H_0^{P_k}$ at level α/K. However, two main sources of power dilution arise in this family of local tests : one coming from the smaller sample sizes N^{P_k} to test $H_0^{P_k}$ and the other, from the Bonferroni level α/K. Other item of our DxEP that can affect power will be discussed when the theorems controlling its behavior have been introduced.

To reduce power loss in a related context, Ehm, Kornmeier and Heinrich (2010) have proposed to recursively partition $\mathcal{X}$ toward the P_k to create a tree of parts of $\mathcal{X}$. Their approach is a particular case of the closure principle (see Finner and Strassburger 2002, for this and related tools; the problem of testing tree-structured hypotheses have been much explored lately, see Goeman and Finos 2012). Adapted to our context, $H_0 : X \sim U(0,1)$ is tested at the root node while the $H_0^{P_k}$ are tested in the leaves. The nodes in between correspond to tests of $H_0^{Q_S} : (X \mid X \in Q_S) \sim U(Q_S)$ where $Q_S = \bigcup_{k \in S} P_k$ and S are subsets of $\{1, 2, \ldots, K\}$.

Let $\mathcal{T}$ be a set of subsets S of $\{1, 2, \ldots, K\}$. To form a tree, the nodes of $\mathcal{T}$ must satisfy the condition : for any $S, S' \in \mathcal{T}$, either $S \cap S' = \emptyset$, $S \subseteq S'$ or $S' \subseteq S$. We further require that all $\{k\}$ be in $\mathcal{T}$ (the leaves) as well as $\{1, 2, \ldots, K\}$ (the root node). The nodes are arranged in decks or alternatively as branches. We conveniently alternate between referring to the node S, the corresponding interval Q_S and the null hypotheses $H_0^{Q_S}$. Moreover, we set the following identifiability condition : for any $S \in \mathcal{T}, H_0^{Q_S} = \bigcap_{k \in S} H_0^{P_k}$. This imposes restrictions on $\mathcal{T}$ and f; identifiability is achieved, notably, when f is continuous and all Q_S are single intervals. It is crucial in ensuring that when $H_0^{Q_S}$ is true, all its child nodes are also true.

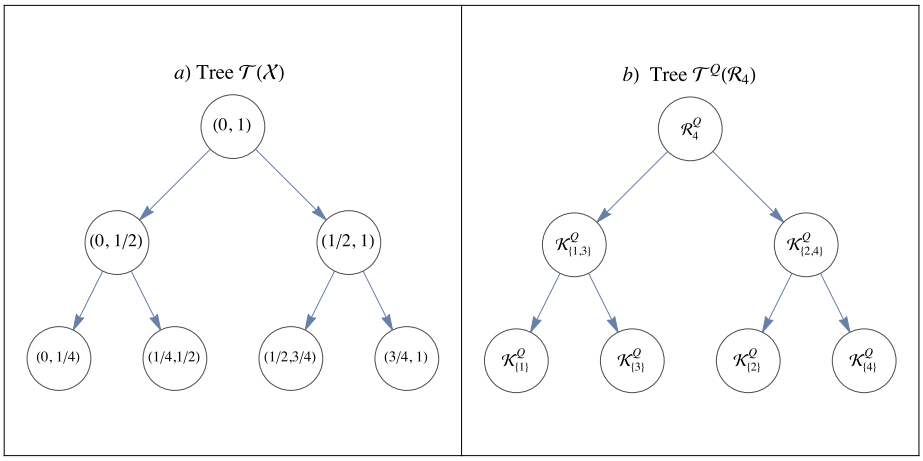

Fig. 4.1 Examples of trees used in the DxEP. Panel *a*) : Tree $\mathcal{T}(\mathcal{X})$ corresponding to the dyadic partition of $\chi = [0,1]$. Panel *b*) : Tree $\mathcal{T}^Q(\mathcal{R}_4)$ corresponding to one decomposition of the local smooth test statistic $\mathcal{R}_4^Q$.

A tree of hypotheses over $\mathcal{X}$ is noted $\mathcal{T}(\mathcal{X})$. As an example, consider the dyadic intervals $P_k = ((k-1)/2^J, k/2^J)$, $k = 1, \ldots, K = 2^J$. Deck 1 is the root node : the test of $H_0 : X \sim U(0,1)$. Deck 2 pertains to the tests of $H_0^{(0,1/2)}$ and $H_0^{(1/2,1)}$; deck 3, to $H_0^{(0,1/4)}$, $H_0^{(1/4,1/2)}$, $H_0^{(1/2,3/4)}$ and $H_0^{(3/4,1)}$, and so on down to the leaves $H_0^{P_k}$. This yields a binary tree but more complex trees can be considered, within identifiability. A representation of this $\mathcal{T}(\mathcal{X})$ for $J = 2$ appears in Panel *a*) of Figure 4.1.

Our DxEP enters $\mathcal{T}(\mathcal{X})$ at the root node to test H_0 via $\mathcal{R}_M$. If rejected, it goes down each branch of the tree testing with $\mathcal{R}_M^Q$ (M could vary) and stops at a node when all tests in its child nodes do not reject, or upon rejecting in a leaf. The process terminates when all branches have been explored. We refer to this as "testing until acceptance". It produces the various nodes Q_S of $\mathcal{T}(\mathcal{X})$ where $H_0^{Q_S}$ is the last rejected hypothesis along a branch.

Going down the tree, dilution of power progressively occurs because of the decreasing sample sizes. One advantage of the tree structure is that even when none of the $H_0^{P_K}$ gets rejected because of poor power, one can still hope that some intermediary $H_0^{Q_S}$ will get rejected, allowing the extraction of less precise but still useful Dx information.

Another advantage is that the T1-FWER for the hypotheses in $\mathcal{T}(\mathcal{X})$ can be strongly controlled while reducing the power dilution coming from the smaller levels. For this, we adapt the procedure in Ehm, Kornmeier and Heinrich (2010) in the following way. For $S \in \mathcal{T}(\mathcal{X})$, let π_S be the p-value obtained for the test of $H_0^{Q_S}$. Call $C \in \mathcal{T}$ the parent of S ($C = pa(S)$) if there does not exist a $C' \in \mathcal{T}$ such that $S \subsetneq C' \subsetneq C$. Then reject $H_0^{Q_S}$ if

1. $\pi_S < \alpha \mathbb{P}_{H_0}[Q_S]$,

2. $H_0^{Q_{pa(S)}}$ has been previously rejected (with $Q_S \neq \mathcal{X}$).

As an example, consider the dyadic intervals $P_k = ((k-1)/2^J, k/2^J)$. The root node H_0 is tested at level α, $H_0^{(0,1/2)}$ and $H_0^{(1/2,1)}$ are tested at level $\alpha/2$, $H_0^{(0,1/4)}$, $H_0^{(1/4,1/2)}$, $H_0^{(1/2,3/4)}$, $H_0^{(3/4,1)}$ at level $\alpha/4$ and so on with the $H_0^{P_k}$ being tested at the Bonferroni level $\alpha/2^J$. Thus, intermediate nodes are rejected at levels $\in \big(\min\{\alpha/2^J\}, \alpha\big)$, limiting this source of power dilution.

Let $\mathcal{T}_{rej} = \{S \in \mathcal{T}; \pi_S < \alpha \mathbb{P}_{H_0}[Q_S]\}$ and $\mathcal{T}_0 = \big\{S \in \mathcal{T}; H_0^{Q_S} \text{ is true}\big\}$. The following theorem, whose proof is in the Appendix, shows that the above procedure strongly controls the T1-FWER asymptotically.

Theorem 2 *For any $\mathcal{T}_0$,*

$$\mathbb{P}_{\mathcal{T}_0}\left[\mathcal{T}_{rej} \cap \mathcal{T}_0 = \emptyset\right] \geqslant 1 - \alpha + o(1). \tag{4.1}$$

A refinement that further reduces power dilution, comes from Shaffer's correction (Meinshausen 2008), which is applicable when the tree is binary. Define the sibling of S $(= si(S))$ as the children $(= ch(C))$ of $C = pa(S)$ such that $si(S) = ch\left(pa(S)\right) \setminus S$. Define the effective probability of Q_S :

$$\mathbb{P}_{H_0}^{\mathit{eff}}[Q_S] = \begin{cases} \mathbb{P}_{H_0}[Q_S], & \text{if } si(S) \text{ is a leaf,} \\ \mathbb{P}_{H_0}[Q_S] + \mathbb{P}_{H_0}[Q_{si(S)}], & \text{if } si(S) \text{ is not a leaf.} \end{cases}$$

Replace Rule 1. by the new rule 1*. : Reject $H_0^{Q_S}$ if $\pi_S < \alpha \mathbb{P}_{H_0}^{\mathit{eff}}[Q_S]$. Shaffer's correction acts by comparing some p-values to a higher threshold, which increases power. For example, in the tree $\mathcal{T}(\mathcal{X})$ with $P_k = ((k-1)/2^J, k/2^J)$ and $J = 2$, all $H_0^{((k-1)/4,k/4)}$ are tested at level $\alpha/2$ instead of $\alpha/4$. The following theorem, whose proof is in the Appendix, shows that with this correction the T1-FWER is still strongly controlled.

Theorem 3 *For $\mathcal{T}_{rej}$ defined as above, but using the effective probabilities, we have for any $\mathcal{T}_0$,*

$$\mathbb{P}_{\mathcal{T}_0}\left[\mathcal{T}_{rej} \cap \mathcal{T}_0 = \emptyset\right] \geqslant 1 - \alpha + o(1). \tag{4.2}$$

4.2 Answering the question "*What causes the problem there ?*"

The approach of the previous subsection identifies the nodes Q of $\mathcal{T}(\mathcal{X})$ where departures from local uniformity arise using $\mathcal{R}_M^Q$. Now we seek to understand the nature of these departures and, for this, we use the Dx capabilities of

the components of $\mathcal{R}_M^Q$. In view of the assumed smoothness of f and the interpretability (moment and projection based) of these components, it makes sense to choose $M \leq 4$.

Near the root of $\mathcal{T}(\mathcal{X})$ where data is abundant, we may take $M = 4$. Consider one Q for which the corresponding H_0^Q has been rejected using $\mathcal{R}_4^Q$. Partition the components $\{\mathcal{L}_1^Q, \ldots, \mathcal{L}_4^Q\}$ into subsets and combine these to form another tree, noted $\mathcal{T}^Q(\mathcal{R}_4)$, whose root node is H_0^Q and whose leaves correspond to the tests of $H_{0,m}^Q : \mu_m^Q = 0$. On deck 2, one meaningful partition is $\left(\mathcal{L}_1^Q, \mathcal{L}_3^Q\right)$ in the left node and $\left(\mathcal{L}_2^Q, \mathcal{L}_4^Q\right)$ in the right. The first pair provides Dx information regarding symmetry over Q (moment-based Dx), or f^Q having a cubic shape (projection-based Dx). The second pair detects departures from the $U(Q)$ in the "tails" of f^Q (or quartic shape). In order for these to carry valid Dx information, they must be rescaled as in Klar (2000). Thus we use test statistics $\mathcal{K}_{\{1,3\}}^Q$, $\mathcal{K}_{\{2,4\}}^Q$ of (3.4) based on the data $X_i \in \mathcal{E}^Q$. Each is asymptotically χ_2^2 under $H_{0,\{1,3\}}^Q : \mu_1^Q = \mu_3^Q = 0$ (left node) and $H_{0,\{2,4\}}^Q : \mu_2^Q = \mu_4^Q = 0$ (right node). In the leaves, $H_{0,m}^Q$ are tested via the similarly defined $\mathcal{K}_{\{m\}}^Q$. The resulting tree, noted $\mathcal{T}^Q(\mathcal{R}_4)$, appears in Panel *b*) of Figure 4.1.

As the procedure goes down $\mathcal{T}(\mathcal{X})$, the smoothness of f makes f^Q look increasingly like a low-order polynomial. Hence a smaller M could be reasonable near the leaves, and in the sequel we use $M = 2$ in the leaves of $\mathcal{T}(\mathcal{X})$ leading to the tree $\mathcal{T}^Q(\mathcal{R}_2)$ with two leaves pertaining to $\mathcal{K}_{\{1\}}^Q$ (linear) and $\mathcal{K}_{\{2\}}^Q$ (quadratic).

Replacing each node Q_S of $\mathcal{T}(\mathcal{X})$ in Panel a) of Figure 4.1 by the corresponding $\mathcal{T}^{Q_S}(\mathcal{R}_M)$ creates a tree of trees structure as in Figure 4.2 that is noted $\mathcal{T}(\mathcal{X}) \otimes \mathcal{T}^Q(\mathcal{R}_M)$. The structure is used as follows : a sample enters at the root node where a test of H_0 based on $\mathcal{R}_4$ is performed. If H_0 is rejected, the sample goes in the nodes of $\mathcal{T}(\mathcal{X})$ (the root nodes of the various $\mathcal{T}^{Q_S}(\mathcal{R}_M)$) and the corresponding H_0^Q along each branch of $\mathcal{T}(\mathcal{X})$ are tested using the rule "test until acceptance". We refer to $\mathcal{P}(\mathcal{X})$ as the corresponding tree of $\alpha\mathbb{P}_{H_0}^{\mathit{eff}}[Q_S]$ to which the p-values of the test statistics are compared.

Having identified, along each branch of $\mathcal{T}(\mathcal{X})$, the shortest Q_S where local uniformity has been rejected, the trees $\mathcal{T}^{Q_S}(\mathcal{R}_M)$ are entered by their root nodes (which has been rejected) and again, testing until acceptance proceeds using $\mathcal{K}_{\{\ldots\}}^Q$. The thresholds to which their p-values are compared are $\alpha\mathbb{P}_{H_0}^{\mathit{eff}}[Q_S] \times \mathbb{Q}^{\mathit{eff}}[\mathcal{K}_{\{\ldots\}}^{Q_S}]$ where $\mathbb{Q}[\mathcal{K}_{\{\ldots\}}^{Q_S}] = \frac{1}{M} \times$(number of components in $\mathcal{K}_{\{\ldots\}}^{Q_S}$) and $\mathbb{Q}^{\mathit{eff}}$ is computed as $\mathbb{P}_{H_0}^{\mathit{eff}}$ but from the $\mathbb{Q}$. The T1-FWER is strongly controlled throughout all of $\mathcal{T}(\mathcal{X}) \otimes \mathcal{T}^Q(\mathcal{R}_M)$, as can be seen by recursively applying Theorems 2 and 3 at the nodes of $\mathcal{T}^Q(\mathcal{R}_M)$.

In the end, the nodes of each tree $\mathcal{T}^{Q_S}(\mathcal{R}_M)$ where testing until acceptance has stopped are singled out to provide the Dx information. As an example, in Figure 4.2, a sample of size 200 from the density

$$f(x) = 3x\mathbb{I}\{x \in (0, \tfrac{1}{4})\} + \tfrac{3}{4}\mathbb{I}\{x \in (\tfrac{1}{4}, \tfrac{1}{2})\} \\ + (3x - \tfrac{3}{4})\mathbb{I}\{x \in (\tfrac{1}{2}, \tfrac{3}{4})\} + \tfrac{3}{2}\mathbb{I}\{x \in (\tfrac{3}{4}, 1)\}, \tag{4.3}$$

has been put through this DxEP (with $\alpha = 10\%$). The two black nodes indicate the intervals, namely $(0, \frac{1}{4})$ and $(\frac{1}{2}, \frac{3}{4})$, where the DxEP has stopped in $\mathcal{T}(\mathcal{X})$, indicating non uniformity on those intervals (a correct statement). Within the two corresponding $\mathcal{T}^{Q_S}(\mathcal{R}_2)$, the gray nodes indicate where the DxEP has stopped. In both cases, this is $\mathcal{K}^{Q_S}_{\{1\}}$ indicating that the departure from local uniformity is caused by f^{Q_S} having a linear shape there, which again is correct. No Dx information is obtained in the intervals $(\frac{1}{4}, \frac{1}{2})$ and $(\frac{3}{4}, 1)$, where the density is indeed flat. Note that the root tree of this structure has leaves that correspond to the rescaled components of Henze & Klar (1996), Henze (1997) and Klar (2000). Thus our DxEP encompasses their approach, as a sample would have stopped in some nodes of this root tree. Here the sample is allowed to do further down $\mathcal{T}(\mathcal{X})\otimes\mathcal{T}^{Q}(\mathcal{R}_M)$ to extract more precise Dx information.

There are three ingredients on our DxEP that can affect the amount of Dx information extracted. The first is partition's choice. In (4.3), the dyadic intervals coincide with the piecewise expression of f and power to identify the non-uniform parts is maximized. When P_k covers both constant and non-uniform parts of f, some power may be loss. To increase Dx information, the user should attempt to match the P_k with the areas (center, tails etc.) where departures may plausibly occur. To help implementing this, the use of an estimated partition, as allowed by Theorem 1, may be useful; see Subsection 5.3.

A second ingredient is the $\mathbb{P}_{H_0}[Q_S]$ in Theorems 2 and 3. This choice appears reasonable as higher thresholds, and thus powers, are allotted to those Q_S expected under H_0 to contain larger sample sizes. This could be changed without affecting the theorems : $\mathbb{P}[Q_S]$ could be computed under some other density.

A last ingredient is the choice of the initial α, from which derive all the thresholds in $\mathcal{T}(\mathcal{X}) \otimes T^{Q}(\mathcal{R}_M$. To set this level properly, some attention must be given to the goals pursued in performing GoF tests. The goal of our DxEP is to provide information toward a better model f_1, after f_0 has been rejected. Taking α too small may lead to little focussed Dx information, which negates the goal of the procedure. Thus, a case can be made about selecting α at a level higher than the sanctified 5%. A reasonable compromise could be to use $\alpha = 10\%$.

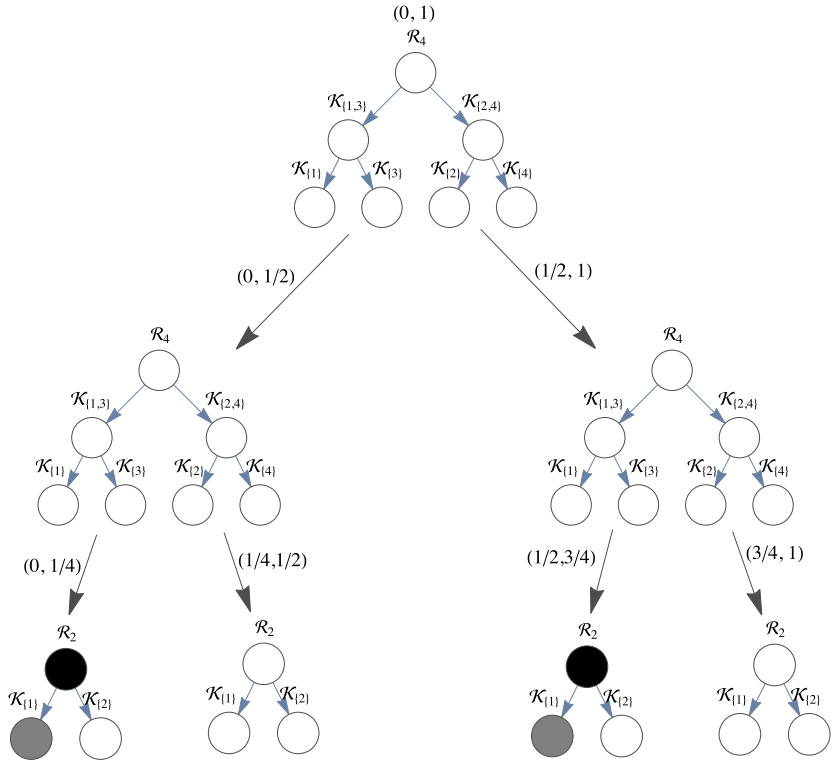

Fig. 4.2 The results of a tree of trees DxEP for a sample of size $n = 200$ from density (4.3) with $\alpha = 10\%$: the black nodes indicate the intervals where the rule "test until acceptance" applied to $\mathcal{T}(\mathcal{X})$ has stopped. In those, the gray nodes point to the components of $\mathcal{R}_2^Q$ where the same rule has stopped.

5 Simulations

5.1 Levels of the DxEP

A first experiment was conducted to see if the tests using the asymptotic χ^2 distribution hold their levels in the nodes of $\mathcal{T}(\mathcal{X}) \otimes \mathcal{T}^Q(\mathcal{R}_M)$ despite the randomness of the sample sizes and the variation coming from an estimated partition. Here $\mathcal{T}(\mathcal{X})$ is the tree in Panel a) of Figure 4.1 with the dyadic $P_k = (q_{(k-1)/4}, q_{k/4})$, where $q_\beta = \beta$ is the β-th quantile of the $U(0,1)$, estimated by $\hat{q}_\beta = \hat{F}_n^{-1}(\beta)$. Its decks 1 and 2 are $\mathcal{T}^Q(\mathcal{R}_4)$, the trees of Panel b) while its leaves are $\mathcal{T}^Q(\mathcal{R}_2)$, to give the structure in Figure 4.2. Samples from the $U(0,1)$ were generated and sent down this structure using $\alpha = 10\%$. Because the interest lies here in the levels, the rule "testing until reaching the leaves" was applied within both $\mathcal{T}(\mathcal{X})$ and the $\mathcal{T}^Q(\mathcal{R}_M)$.

The actual levels of the tests (in %), approximated from 10000 replications, are shown in the top entry of the circles of Figure 5.1 (for $n = 200$). The

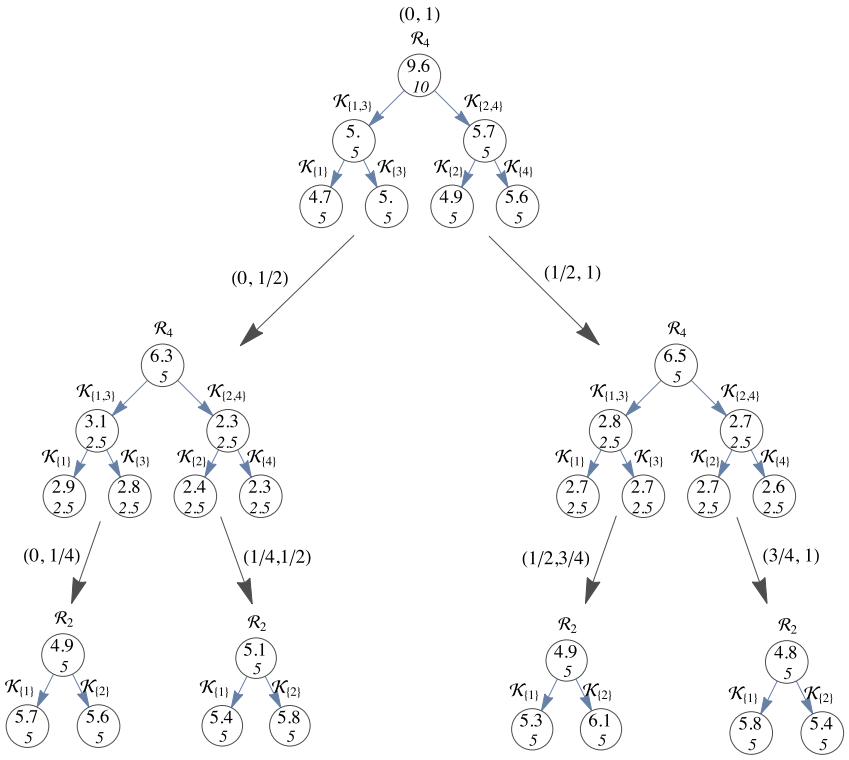

Fig. 5.1 Actual levels (top entry of each circle), approximated from 10000 replications with sample size $n = 200$, of the various tests of our DxEP; italicized bottom entry : nominal levels (with $\alpha = 10\%$ in the root node). The partition of $\mathcal{T}(\mathcal{X})$ is based on the quantiles $q_{\hat{\beta}}$ estimated from the samples ; e.g. $(\frac{1}{4}, \frac{1}{2})$ refers to $(\hat{q}_{1/4}, \hat{q}_{1/2})$.

italicized bottom entries are the nominal levels, computed from the formula in Subsection 4.2. It can be seen that the actual levels are close to nominal and, overall, rather accurate for the purpose at hand. It should be noted that the results obtained using the true quantiles q_β closely resemble those obtained with the $\hat{q}_\beta$ and that other sample sizes as well as other tree of trees structures were explored; the results were rather stable, as H_0 prevails.

5.2 Power and T1-FWER of the DxEP

A second experiment was conducted to analyze the power of our DxEP in producing useful diagnostic information. The general framework was the following : 10000 samples from two alternative distributions having some uniform and

nearly uniform parts were generated and the DxEP applied at level $\alpha = 10\%$ with Shaffer's correction. We took the general structure of $\mathcal{T}(\mathcal{X}) \otimes \mathcal{T}^Q(\mathcal{R}_M)$ in Figures 4.2 and 5.1 so that the nominal levels used for the tests are those of the italicized bottom entries in Figure 5.1.

There are several ways to report the results. We took the following because of its similarity with standard power analyses of GoF tests. First, the nodes (there could be several) of $\mathcal{T}(\mathcal{X})$ where a sample was stopped by the "test until acceptance" rule were recorded and counts from the 10000 samples were obtained. Normalizing these counts by 10000 provides the empirical probability that a sample yields Dx information regarding the question "*Where is the problem ?*" in the form of statements as : "This is the shortest interval of $\mathcal{X}$ along this branch where the data does not support local uniformity". On the figures, these probabilities appear in the shaded circles (in %).

A sample stopped at node Q of $\mathcal{T}(\mathcal{X})$ was rerouted into $\mathcal{T}^Q(\mathcal{R}_M)$ and its nodes (again there may be several) where this sample had finally stopped were recorded. The resulting counts (divided by 10000) provide empirical probabilities that a sample stops in Q while providing some Dx information regarding the question "*What is the problem there ?* in the form of statements as : "in Q, the density seems to have a linear (quadratic etc.) shape". On the figures, these empirical probabilities appear in the white circles (in %).

Our first alternative density is (4.3). It was chosen 1) to understand the behavior of the DxEp as n increases, 2) to check that strong control is exerted and 3) to show that unexpected behavior can result from the complex relationships between the elements of the power function. Note that the partition defining $\mathcal{T}(\mathcal{X})$ is tailored to the piecewise form of (4.3). Empirical powers are shown in Figure 5.2 for $n = 250$ and $n = 500$ (resp. upper and lower entries of the circles). For both sample sizes, all samples rejected uniformity at the root node and were sent down the structure to extract Dx information with a success that we now comment. First concentrate on the upper numbers ($n = 250$). The root node of $\mathcal{T}(\mathcal{X})$ says that only 3.4% of the samples were unable to yield any Dx information beyond the rejection of H_0. In the left branch of $\mathcal{T}(\mathcal{X})$, 18.0% of the samples rejected the local $U(0, \frac{1}{2})$ but were unable to investigate the smaller intervals. These samples were rerouted to $\mathcal{T}^{(0,1/2)}(\mathcal{R}_4)$. In the end, 17.5 % provided the crudely correct diagnostic information : "in $(0, \frac{1}{2})$, $f^{(0,1/2)}$ seems to possess a linear component ". The other samples went further down and 73.3% were able to provide the correct Dx information that $f^{(0,1/4)}$ has a linear trend.

The corresponding figures on the right branch gives similar Dx information about $(\frac{1}{2}, \frac{3}{4})$, but the empirical probabilities are noticeably smaller. This unexpected behavior, in view of the shape of the density, is explained by the values of the μ_m^Q and σ_{mm}^Q in the elements of the partition. Concentrating on $\mathcal{K}_{\{1\}}^Q$, in $(0, \frac{1}{4})$ we have $(\mu_1, \sigma_{11}) = (0.58,\ 0.67)$ with average sample size 25 leading to a non-centrality parameter around 12.5 whereas in $(\frac{1}{2}, \frac{3}{4})$, one gets (0.19, 0.96) with average sample size 75 and non-centrality parameter about 2.9.

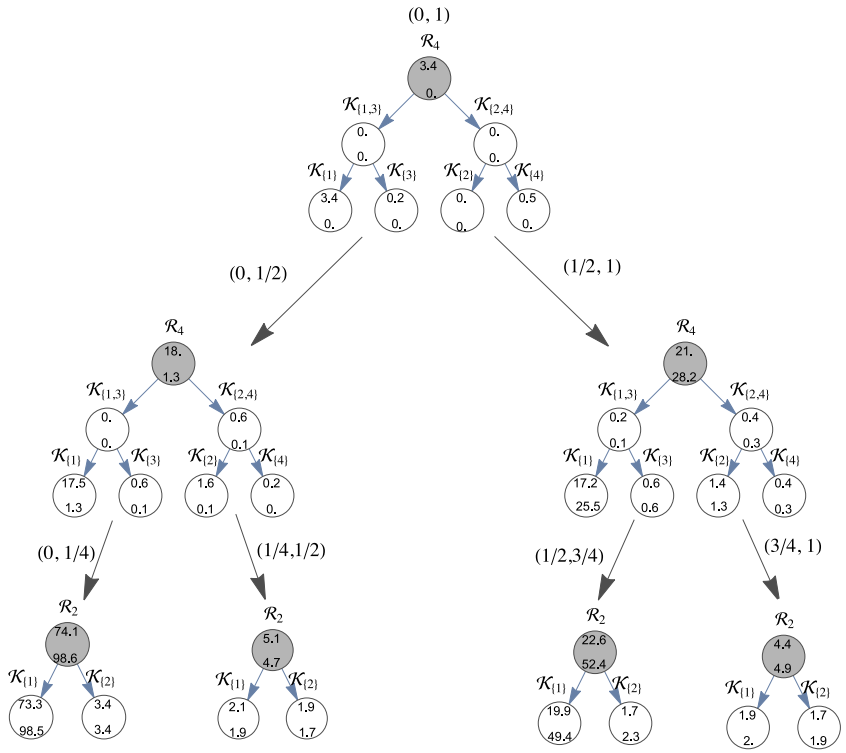

Fig. 5.2 Empirical power ((based on 10000 samples) with $n = 250$ (upper numbers) and $n = 500$ (lower) for alternative (4.3) with α= 10%.

In view of the strong control exerted by the procedure, at most 10% of the samples should reject at least one of $H_0^{(1/4,1/2)}$ and $H_0^{(3/4,1)}$. The figures in these nodes (5.1 and 4.4) gives confirmation that Theorems 2 and 3 are acting on $\mathcal{T}(\mathcal{X})$. In the leaves of $\mathcal{T}^{(1/4,1/2)}(\mathcal{R}_2)$ and $\mathcal{T}^{(3/4,1)}(\mathcal{R}_2)$, the sum of the entries are smaller than 5%, showing that again, strong control is exerted within them.

The numbers in the bottom of the circles ($n = 500$) show the trend when the sample size varies. When n is small, the DxEP tends to stop in the upper nodes of $\mathcal{T}(\mathcal{X}) \otimes \mathcal{T}^{\mathcal{Q}}(\mathcal{R}_M)$ and provide crude diagnostic information. As n increases, the samples are pushed toward the various leaves of $\mathcal{T}(\mathcal{X}) \otimes \mathcal{T}^{\mathcal{Q}}(\mathcal{R}_M)$ where more focussed Dx information can be extracted. In the nodes where uniformity hold, the empirical probabilities are around 2%, showing that wrong Dx informations are rarely produced.

Note finally that the dyadic partition used here coincides with the piecewise expression of this alternative, which results in optimal power. In general, such precise information is unavailable and using, as a default, the dyadic partition can result in less Dx information.

The second alternative density was taken to explore how the use of $\mathcal{R}_M^{\hat{Q}}$ in Theorem 1 can be of some help toward getting a data-adapted partition, as long as some pilot information is available to suggest a good partition, albeit one that must be estimated. We show the gain in power that can be obtained as compared to the use of the default, and here misadapted, dyadic partition in $\mathcal{T}(\mathcal{X}) \otimes \mathcal{T}^{\mathcal{Q}}(\mathcal{R}_M)$ of the previous figures.

This alternative arises from a realistic scenario where, after careful examination of the apparatus generating the data, a user is rather confident that its distribution is locally well modeled by a $N(0,1)$ in its central part. The pilot information is that the user is less confident about the upper and lower 10% tails, where departures are plausible. Accordingly, after applying the probability integral transformation (PIT) to the data, the user chooses the partition $(0, q_{1/10})$, $(q_{1/10}, q_{1/2})$, $(q_{1/2}, q_{9/10})$ and $(q_{9/10}, 1)$. We refer to the resulting tree as $\mathcal{T}^*(\mathcal{X}) \otimes \mathcal{T}^{\mathcal{Q}}(\mathcal{R}_M)$. This partition is estimated via $\hat{F}_n^{-1}$.

Suppose that the alternative density is a $N(0,1)$ contaminated at level 10% by a $N(0, 100)$. After applying the PIT, the resulting density over [0,1] is symmetric and nearly uniform over $(0.07, 0.93) \approx (q_{1/10}, q_{9/10})$ but raises with a faster than linear trend near both boundaries.

Samples (10000) of size $n = 200$ from this contaminated normal were generated. Each sample was submitted to both the dyadic-based and estimated quantiles-based DxEP applied at level $\alpha = 10\%$. The results are shown in Figure 5.3. Because of the symmetry, the left and right branch of each tree of trees are the same, up to random fluctuations and after taking into account the mirror effect at the boundaries caused by symmetry. Hence, to facilitate comparison, only the left branch of $\mathcal{T}(\mathcal{X}) \otimes \mathcal{T}^{\mathcal{Q}}(\mathcal{R}_M)$ and the right branch of the estimated $\mathcal{T}^*(\mathcal{X}) \otimes \mathcal{T}^{\mathcal{Q}}(\mathcal{R}_M)$ are shown on this figure. Also the entries of the root tree nearly coincide and have been removed.

One can see that the samples in the left branch (dyadic) have more difficulty reaching the leaves than those on the right branch (estimated quantiles). Moreover, as seen by the power of component $\mathcal{K}_{\{2\}}^{(\hat{q}_{9/10},1)}$, the faster than linear trend in the tails is better detected in $\mathcal{T}^*(\mathcal{X}) \otimes \mathcal{T}^{\mathcal{Q}}(\mathcal{R}_M)$. Thus the exploitation of some pilot information to suggest a data-adapted partition can push a sample toward the leaves, where more precise Dx information is extracted.

6 Conclusion

This paper considers the framework where the model f_0 is entirely specified. This case is important for theoretical reasons and covers some applied problems. It shows the main ideas behind our DxEP in a simplified context. An important evolution of the present methodology would cover those cases where the model depends on unknown parameters and using the above approach in this extended context does not control the risks of errors. Here technical difficulties appears that seem above the case considered here, particularly if these parameters enter in the specification of $\mathcal{T}(\mathcal{X})$. One issue regards the impact

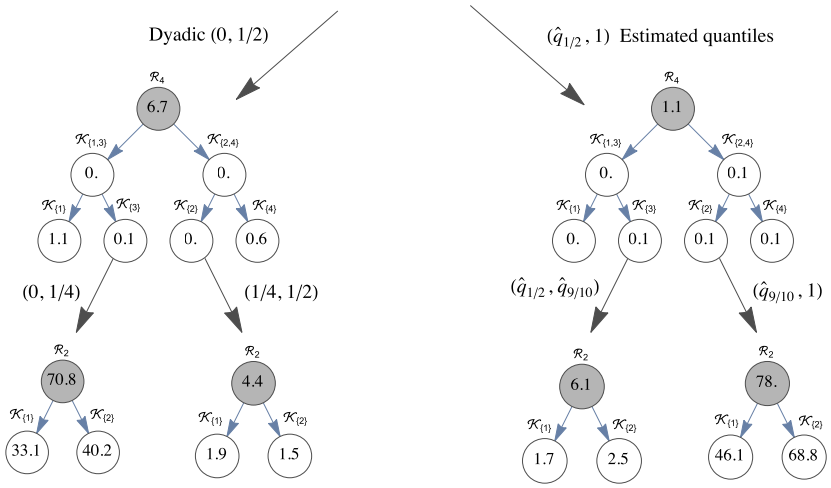

Fig. 5.3 Estimated power (based on 10000 replications with $n = 200$) for the contaminated normal alternative with the tests performed at level α= 10%. The left side corresponds to the dyadic partition in panel a) of Figure 4.1, the right side to the estimated partition $(\hat{q}_{1/2}, \hat{q}_{9/10})$ and $(\hat{q}_{9/10}, 1)$.

of their estimation on the asymptotic distribution of the $\mathcal{R}_M^{\hat{Q}}$. Another issue is invariance, the lack of which could render the Dx information misleading.

Acknowledgements The authors wish to thank the referees for their constructive comments.

References

1. Ehm, W., Kornmeier, J. and Heinrich, S.P. (2010) : Multiple Testing along a Tree. *Electronic Journal of Statistics*, 4, p.462-471.
2. Finner, H., Strassburger, K. (2002) : The partitioning principle : a powerful tool in multiple decision theory. *Ann. Stat.*, 30, p.1194-1213.
3. Goeman, J.J., Finos, L. (2012) : The inheritance procedure : Multiple testing of tree-structured hypotheses. *Statist. Appl. in Genetics and Molecular Biology*, 2, No. 11.
4. Henze, N., Klar, B. (1996) : Properly Rescaled Components of Smooth Tests of Fit are Diagnostic. *Austral. J. Statist.* 38, p. 61-74.
5. Henze, N. (1997) : Do Components of Smooth Tests of Fit have Diagnostic Properties ? *Metrika*, 45, p.121-130.
6. Inglot, T., Kellenberg, W.C., Ledwina, T. (1994) : Power Approximations to and Power Comparison of Smooth Goodness-of-Fit Tests. *Scand. J. of Statist.* 21, p.131-145.
7. Klar B. (2000), Diagnostic Smooth Tests of Fit. *Metrika*, 52, p.237-252.
8. Komlos, J., Major, P. and Tusnady, G. (1976) : An approximation of partial sums of independent rv's and the sample df. II. *Z. Wahrsch verw Gebiete*, 34, p.33–58.
9. Meinshausen, N. (2008) : Hierarchical testing of variable importance. *Biometrika*, 95, p.265-278
10. Pollard. D. (1979) : General Chi-square Goodness-of-fit Test with Data-dependent Cells. *Z. Wahrsch verw Gebiete*, 50, p.317-331.
11. Rayner, J.C.W., Best, D.J. (1989) : *Smooth Tests of Goodness of Fit.* Oxford University Press, New York.

12. von Eye, A., Bogat, G.A. (2004) : Testing the assumption of multivariate normality. *Psychology Science*, 46, p. 243-258.

A Proof of Theorems

We prove Theorem 1and concentrate on the more complex $\mathcal{R}_M^{\hat{Q}}$. To avoid trivial problems, assume that $\hat{b} - \hat{a} > 0$. With $\hat{X}_i^* = F_0^{\hat{Q}}(X_i)$ and $X_i^* = F_0^Q(X_i)$, consider

$$\frac{1}{\sqrt{N^{\hat{Q}}}} \sum_{X_i \in \mathcal{E}^{\hat{Q}}} L_m\left(\hat{X}_i^*\right) = \left(\frac{\sqrt{n}}{\sqrt{N^{\hat{Q}}}}\right) \frac{1}{\sqrt{n}} \sum_{i=1}^n L_m^*(\hat{X}_i^*).$$

Now, $L_m^*(\hat{X}_i^*) - L_m^*(X_i^*) = C(X_i) + D(X_i)$, where $C(x) = \mathbb{I}\{x \in Q\}\left[L_m\left(F_0^{\hat{Q}}(x)\right) - L_m(F_0^Q(x))\right]$ and $D(x) = L_m\left(F_0^{\hat{Q}}(x)\right)\left[\mathbb{I}\left\{x \in \hat{Q}\right\} - \mathbb{I}\{x \in Q\}\right]$. Hence,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n L_m^*(\hat{X}_i^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n L_m^*(X_i^*) + \frac{1}{\sqrt{n}} \sum_{i=1}^n C(X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^n D(X_i). \tag{A.1}$$

Consider the second term on the right hand side of this equation. $F_0^Q(x)$ is differentiable in a, b except at $a = x$ et $b = x$. Because $X_i \neq a, b$ with probability 1, we can Taylor expand $L_m\left(F_0^{\hat{Q}}(x)\right) - L_m\left(F_0^Q(x)\right)$ about (a, b) for $x \in (a, b)$ to get

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n C(X_i) = \left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \in Q\} \begin{pmatrix} \frac{\partial}{\partial a} L_m\left(F_0^Q(X_i)\right) \\ \frac{\partial}{\partial b} L_m\left(F_0^Q(X_i)\right) \end{pmatrix}\right]^T \sqrt{n} \begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix} + o_p(1).$$

By the law of large numbers, the term in bracket converges to its expectation. Because $\mathbb{E}_{H_0^Q}\left(\mathbb{I}\{X_i \in Q\} L_m\left(F_0^Q(X_i)\right)\right) = 0$, Leibniz's integral rule yields

$$\mathbb{E}_{H_0^Q}\left(\mathbb{I}\{X_i \in Q\} \begin{pmatrix} \frac{\partial}{\partial a} L_m\left(F_0^Q(X_i)\right) \\ \frac{\partial}{\partial b} L_m\left(F_0^Q(X_i)\right) \end{pmatrix}\right) = \begin{pmatrix} L_m(0) \\ -L_m(1) \end{pmatrix}.$$

Hence

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n C(X_i) = \begin{pmatrix} L_m(0) \\ -L_m(1) \end{pmatrix}^T \sqrt{n} \begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix} + o_p(1). \tag{A.2}$$

Now, regarding the term involving the $D(X_i)$ in (A.1), it is easy to see that

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n D(X_i) = {} & (L_m(0) + o_p(1)) \operatorname{sgn}(a - \hat{a}) \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{I}\{X_i \in \langle a, \hat{a} \rangle\} \\ & + (L_m(1) + o_p(1)) \operatorname{sgn}(\hat{b} - b) \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{I}\left\{X_i \in \left\langle b, \hat{b} \right\rangle\right\}, \end{aligned}$$

where $\langle a, \hat{a} \rangle = (\min(a, \hat{a}), \max(a, \hat{a}))$ and similarly for $\left\langle b, \hat{b} \right\rangle$. Moreover

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{I}\{X_i \in \langle a, \hat{a} \rangle\} = \operatorname{sgn}(\hat{a} - a)\left(\alpha_n(\hat{a}) - \alpha_n(a)\right) + \operatorname{sgn}(\hat{a} - a)\sqrt{n}\left(\hat{a} - a\right),$$

where $\alpha_n(\cdot)$ is the stochastic process $\sqrt{n}\left(\hat{F}_n(\cdot) - \cdot\right)$. From the Komlos-Major-Tusnady (1976) strong approximation, there exists a sequence of Brownian bridges $B_n(\cdot)$ uniformly approximating $\alpha_n(\cdot)$ to the order $O\left(\frac{\log n}{\sqrt{n}}\right)$. Thus

$$\operatorname{sgn}(a-\hat{a})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{I}\left\{x \in \langle a, \hat{a}\rangle\right\} = (B_n(a) - B_n(\hat{a})) - \sqrt{n}(\hat{a}-a) + O\left(\frac{\log n}{\sqrt{n}}\right).$$

The term $B_n(a) - B_n(\hat{a}) = o_p(1)$ by the continuity of Brownian bridges. Hence,

$$(L_m(0) + o_p(1))\operatorname{sgn}(a-\hat{a})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{I}\left\{X_i \in \langle a, \hat{a}\rangle\right\} = -L_m(0)\sqrt{n}(\hat{a}-a) + o_p(1),$$

$$(L_m(1) + o_p(1))\operatorname{sgn}(\hat{b}-b)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{I}\left\{X_i \in \langle b, \hat{b}\rangle\right\} = L_m(1)\sqrt{n}(\hat{b}-b) + o_p(1).$$

Regrouping, we get :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D(X_i) = \begin{pmatrix} -L_m(0) \\ L_m(1) \end{pmatrix}^T \sqrt{n}\begin{pmatrix} \hat{a}-a \\ \hat{b}-b \end{pmatrix} + o_p(1), \tag{A.3}$$

Going back to (A.1), we find after combining (A.2) and (A.3)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} L_m^*(\hat{X}_i^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} L_m^*\left(X_i^*\right) + o_p(1).$$

Finally, as a byproduct of the above, $\frac{N^{\hat{Q}}}{n} = \frac{N^Q}{n} + o_p(1) = F(Q) + o_p(1)$ because N^Q has the binomial distribution $B\left(n, F(Q)\right)$. Hence the asymptotic behavior of the $\mathcal{L}_m^{\hat{Q}}$ are the same as that of the $\mathcal{L}_m^Q$ which are easily shown to independent χ_1^2 . □

Next, we prove Theorem 2. Suppose for simplicity that Q_S are single intervals. We consider the more complex case where Q_S are estimated by $\hat{Q}_S$. Define the adjusted p-value as $\pi_S^{(adj)} = \frac{1}{\mathbb{P}_{H_0}[\hat{Q}_S]}\,\pi_S$, where π_S pertains to $\mathcal{R}_M^{\hat{Q}_S}$. Let $\mathcal{T}_0$ be as in Theorem 2. Define the hierarchical adjusted p-value as

$$\pi_S^{(h,adj)} = \max_{S \subseteq C \in \mathcal{T}} \pi_C^{(adj)}. \tag{A.4}$$

Let $\mathcal{T}_{rej} = \left\{S \in \mathcal{T}; H_0^{Q_S} \text{ is rejected by the rule : } \pi_S^{(h,adj)} < \alpha\right\}$. It is easy to see that the null hypotheses rejected using the hierarchical adjusted p-value coincide with those rejected using the $\pi_S^{(adj)}$. In particular, no null hypothesis gets rejected if its parent has not been rejected because $\pi_S^{(h,adj)} \geq \pi_{pa(S)}^{(h,adj)}$. The probability of a family-wise Type 1 error can be written as

$$\mathbb{P}_{\mathcal{T}_0}\left[\mathcal{T}_{rej} \cap \mathcal{T}_0 \neq \emptyset\right] = \mathbb{P}_{\mathcal{T}_0}\left[\exists S \in \mathcal{T}_0 : \pi_S^{(h,adj)} < \alpha\right].$$

Let $\widetilde{\mathcal{T}}_0$ be a subset of $\mathcal{T}$ maximal in the sense that $\widetilde{\mathcal{T}}_0 := \{S \in \mathcal{T}_0 : \nexists C \in \mathcal{T}_0 \text{ with } S \subset C\}$. Obviously $\widetilde{\mathcal{T}}_0 \subseteq \mathcal{T}_0$. Also the definition of $\pi_S^{(h,adj)}$ implies that a falsely rejected $S \in \mathcal{T}_0 - \widetilde{\mathcal{T}}_0$, implies a falsely rejected $S' \in \widetilde{\mathcal{T}}_0$, where $S \subset S'$. Thus, we need only to look at the probability of committing a Type 1 error in $\widetilde{\mathcal{T}}_0$. But because $\pi_S^{(h,adj)} \geqslant \pi_S^{(adj)}$,

$$\mathbb{P}_{\mathcal{T}_0}\left[\exists S \in \mathcal{T}_0 : \pi_S^{(h,adj)} < \alpha\right] \leq \sum_{S \in \widetilde{\mathcal{T}}_0} \mathbb{P}_{\mathcal{T}_0}\left[\pi_S^{(adj)} < \alpha\right],$$

by Bonferroni's inequality. Notice that $\mathbb{P}_{H_0}[\hat{Q}_S] = \mathbb{E}_{H_0}(\hat{b}-\hat{a}) = \mathbb{P}_{H_0}[Q_S]+o(1)$. Now writing $G_n^S(\cdot)$ and $G_\infty^S(\cdot)$ for the exact and asymptotic CDF under $\widetilde{\mathcal{T}}_0$ of test statistic $\mathcal{R}_M^{\hat{Q}_S}$, we have

$$\begin{aligned}\mathbb{P}_{\mathcal{T}_0}\left[\pi_S^{(adj)} < \alpha\right] &= \mathbb{P}_{\mathcal{T}_0}\left[\pi_S < \alpha\mathbb{P}_{H_0}[Q_S] + o(1)\right] \\ &= 1 - G_n^S\left(G_\infty^{S^{-1}}(1 - \alpha P_{H_0}[Q_S] + o(1))\right) \\ &= \alpha\mathbb{P}_{H_0}[Q_S] + o(1).\end{aligned}$$

Hence, $\sum_{S\in\widetilde{\mathcal{T}}_0}\mathbb{P}_{\mathcal{T}_0}\left[\pi_S^{(adj)} < \alpha\right] < \alpha\sum_{S\in\widetilde{\mathcal{T}}_0}\mathbb{P}_{H_0}[Q_S] + o(1)$. It only remains to show that $\sum_{S\in\widetilde{\mathcal{T}}_0}\mathbb{P}_{H_0}[Q_S] \leqslant 1$. But by the construction of $\widetilde{\mathcal{T}}_0$, $\forall S \neq S' \in \widetilde{\mathcal{T}}_0 : S \cap S' = \emptyset$. Hence $\bigcup_{S\in\widetilde{\mathcal{T}}_0} S \subseteq \{1,\ldots,K\}$. Because $\mathbb{P}_{H_0}[Q_S] = \mathbb{P}_{H_0}\left[X \in \bigcup_{k\in S} P_k\right]$, we have

$$\sum_{S\in\widetilde{\mathcal{T}}_0}\mathbb{P}_{H_0}[Q_S] \leqslant \mathbb{P}_{H_0}\left[X \in \bigcup_{k=1}^{K} P_k\right] = 1,$$

and this completes the proof. □

Finally, we prove Theorem 3 with given Q_S for simplicity. From the above, it suffices to show that $\sum_{S\in\widetilde{\mathcal{T}}_0}\mathbb{P}_{H_0}^{eff}[Q_S] \leq 1$. For simplicity, assume there exists only one $S^* \in \widetilde{\mathcal{T}}_0$ such that $\mathbb{P}_{H_0}^{\text{eff}}[Q_{S^*}] > \mathbb{P}_{H_0}[Q_{S^*}]$. Because the tree is binary,

$$\sum_{S\in\widetilde{\mathcal{T}}_0}\mathbb{P}_{H_0}^{\text{eff}}[Q_S] = \sum_{S\in\widetilde{\mathcal{T}}_0\setminus S^*}\mathbb{P}_{H_0}[Q_S] + \mathbb{P}_{H_0}[Q_{S^*}] + \mathbb{P}_{H_0}[si(Q_{S^*})].$$

Now, identifiability along with the relation $\mathbb{P}_{H_0}^{\text{eff}}[S^*] > \mathbb{P}_{H_0}[S^*]$ imply that $si(S^*) \notin \widetilde{\mathcal{T}}_0$, for otherwise $pa(S^*) \in \widetilde{\mathcal{T}}_0$, which in turn implies $S^* \notin \widetilde{\mathcal{T}}_0$. The conclusion follows from

$$\sum_{S\in\widetilde{\mathcal{T}}_0\setminus S^*}\mathbb{P}_{H_0}[Q_S] + \mathbb{P}_{H_0}[Q_{S^*}] \leq 1 - \mathbb{P}_{H_0}[si(Q_{S^*})].\ \square$$