Top scoring pair classifiers: asymptotics and applications

Christophe Denis
Laboratoire MAP5, UMR CNRS 8145,
Université Paris Descartes, Sorbonne Paris Cité, Paris, France
christophe.denis@parisdescartes.fr

Abstract

The original *top scoring pair* (TSP) classifier was proposed by Geman *et al.* (2004) for binary classification of diseases based on genetic profiles. We show the consistency of two versions of the TSP classifier and their two cross-validated counterparts relative to two different risks: the classical misclassification risk and an asymmetric version of this risk which gives more weight to the rarer class. A numerical study illustrates our results and sheds further light on the different TSP classification procedures.

Keywords: Top scoring pair classifier; Classification; Cross-validation

1 Introduction

This work is devoted to the study of two top scoring pair (TSP) classifiers inspired by the original TSP classifier defined and studied by Geman *et al.* (2004).

The original TSP classifier was coined to address the classification of different cancers based on gene-expression profiles. The main feature of the original TSP classifier is that it is based on *pairwise-comparisons*. More precisely, the method consists in differentiating between two classes by finding pairs of genes whose expression levels typically invert from one class to the other. A single pair of genes (or sometimes a handfull of them) is selected by maximizing a score. Then the resulting TSP classification rule is based only on the selected pair(s) of genes. Thus, the TSP classifier does not suffer from the lack of interpretability which often arises in the statistical analysis of microarray data. Indeed, it is easy to interpret the fact that the expression level of a gene is larger than the expression level of another gene, even if these expression levels are obtained under different experimental conditions. Hopefully, one can pass the so-called "elevator exam": explaining to one's colleague from the biology department how this classifier works in the elevator between the first and the fourth floors. Moreover, this classification rule is robust to quantization effects and invariant to pre-processing such as normalization methods (Yang *et al.*, 2001).

Furthermore, the TSP classifier is easy to compute, and its implementation requires no tuning parameters. We refer to the `tspair` package for an `R` implementation. Geman *et al.* (2004) also argue that the TSP classifier behaves well in the "small n large p" paradigm, and they show on several real datasets that the TSP classification rule compares favorably with more complex ones. Note that the TSP procedure proved useful in other contexts: for instance, Chambaz and Denis (2012) apply the TSP classification procedure for classifying subjects in terms of postural maintenance.

Various extensions of the TSP classification procedure have been proposed. Tan *et al.* (2005) and Zhou *et al.* (2012) are interested in k-TSP procedures which involve the pairs that achieve the k largest scores rather than the highest score only. Note that the k-TSP procedure of Tan *et al.* (2005) applies to the multi-class framework. Furthermore, Czajkowski and Kretowski (2011) propose a TSP procedure based on decision trees. As far as we know, there is no theoretical study of the TSP classification procedure.

The aim of this work is to provide such a theoretical study. We show that the differences in risks (for two different risks and their cross-validated versions) of the two empirical TSP classifiers that we consider here relative to their theoretical counterparts are $O(\sqrt{\log(M)/N})$, where M is the number of pairs and N is either the sample size n or the product $n\pi$ with π the probability to observe the rarer of the two labels. In particular, the results shed some light on how the empirical TSP classifiers behave in the "small n large p" paradigm.

The article is organized as follows. In Section 2, we define the two TSP classification procedures of interest as two maximizers of two different scores. Their empirical counterparts are defined as maximizers of the corresponding empirical scores in Section 3, where we also carry out their asymptotic study. We introduce the cross-validated versions of the two TSP classification rules in Section 4, where we also investigate their asymptotic behaviors. We present a numerical illustration based on a real dataset in Section 5, where we also summarize the results of a simulation study to the compare the performances of the different TSP procedures. We draw some conclusions and present some perpectives in Section 6. The proofs of the main results are postponed to Section 7.

2 General framework

This section is devoted to the definition of two TSP classifiers. We first introduce some useful notations and definitions in Section 2.1. We define the TSP classifiers in Section 2.2.

2.1 Notations

Let $O = (X, Y)$ be the observed data-structure taking values in $\mathbb{R}^G \times \{0, 1\}$ (for a possibly large integer G). For instance, X can be viewed as the expression levels of G genes while Y can indicate if the subject is healthy or not. The true data-generating distribution is P_0, which is an element of the set $\mathcal{M}$ of all candidate data-generating distributions. We denote $D_n = \left\{O^k = (X^k, Y^k), k = 1, \ldots, n\right\}$ a learning dataset, where $O^1 = (X^1, Y^1), \ldots, O^n =$

(X^n, Y^n) are independent copies of O. We set $\mathcal{J} = \{J = (i,j) \in \{1,\dots,G\}^2\ , i < j\}$ and $Z_J = \mathbf{1}\{X_i < X_j\}$ for all $J \in \mathcal{J}$. Obviously, $\mathrm{card}(\mathcal{J}) = G(G-1)/2$.

We introduce the following notations: $p^1 = P_0(Y=1) = 1 - p^0$, $p = \min(p^1, p^0)$, and for each $J \in \mathcal{J}$, $\alpha_J = P_0(Z_J = 1)$, $\eta_J(Z_J) = P_0(Y = 1|Z_J)$, $p_J(1) = P_0(Z_J = 1|Y = 1)$, and $p_J(0) = P_0(Z_J = 1|Y = 0)$. We assume that $\mathrm{card}(\mathcal{J}) \geq 2$ and $p > 0$.

Let $\mathcal{F}$ be the set of these functions which map $\mathbb{R}^G$ onto $\{0,1\}$, and consider the loss functions $L : \mathbb{R}^G \times \{0,1\} \times \mathcal{F} \to \mathbb{R}_+$ such that

$$L((X,Y),\Psi) = \mathbf{1}\{\Psi(X) \neq Y\} = Y\mathbf{1}\{\Psi(X) \neq 1\} + (1-Y)\mathbf{1}\{\Psi(X) \neq 0\}.$$

The loss function L is the usual loss function in the classification framework. For all $P \in \mathcal{M}$, this yields the risks $R_1^{(P)}, R_2^{(P)} : \mathcal{F} \to \mathbb{R}_+$ characterized by

$$\begin{aligned} R_1^{(P)}(\Psi) &= E_P[L(O,\Psi)] = P(\Psi(X) \neq Y), \quad \text{and} \\ R_2^{(P)}(\Psi) &= E_P[L(O,\Psi)|Y=1] + E_P[L(O,\Psi)|Y=0] \\ &= \frac{1}{p^1} P(\Psi(X) \neq 1, Y = 1) + \frac{1}{p^0} P(\Psi(X) \neq 0, Y = 0). \end{aligned}$$

The risk R_1 is called misclassification risk. We call R_2 a weighted misclassification risk. The risk R_2 is particularly usefull when $p \ll \max(p^1, p^0)$ and it is important to identify the elements of the rare class.

Finally, we define $\mathcal{F}_{\text{pair}} = \bigcup_{\mathcal{J} \in \mathcal{J}} \mathcal{F}_J$ where $\mathcal{F}_J$ is the set of these functions t of Z_J such that $t(Z_J) \in \{0,1\}$. For $J \in \mathcal{J}$, a classifier $t \in \mathcal{F}_J$ is called a pair classifier. Note that $\mathrm{card}(\mathcal{F}_{\text{pair}}) = 4\mathrm{card}(\mathcal{J})$.

2.2 Definition of the TSP classifiers

The two TSP classifiers that we consider here are elements of $\mathcal{F}_{\text{pair}}$. Their definitions involve the risks R_1 and R_2. Of course there is no guarantee a priori that classifying based on basic comparisons as they do will prove efficient. However, they are so simple and so fast that one can try them almost at no cost.

2.2.1 TSP for the misclassifcation risk

We first introduce the TSP classifier for the misclassification risk R_1.

For each $J \in \mathcal{J}$, let Ψ_J denote the Bayes classifier on the set $\mathcal{F}_J$, defined by

$$\Psi_J(X) = \Psi_J(Z_J) = \mathbf{1}\{\eta_J(Z_J) \geq 1/2\}. \tag{1}$$

The classifier Ψ_J votes for the class with the larger probability conditionally on $X_i < X_j$ or $X_i \geq X_j$. We recall that Ψ_J is also characterized by $\Psi_J \in \arg\min_{t \in \mathcal{F}_J} R_1^{(P_0)}(t)$. We define the score γ_J of the pair J as

$$\gamma_J = \alpha_J |\eta_J(1) - 1/2| + (1-\alpha_J)|\eta_J(0) - 1/2|. \tag{2}$$

The following lemma connects the score of a pair $J \in \mathcal{J}$ to the misclassification risk of Ψ_J.

Lemma 1. *For each $J \in \mathcal{J}$ it holds that $\gamma_J = 1/2 - R_1^{(P_0)}(\Psi_J)$.*

Proof. Set $J \in \mathcal{J}$. We first decompose $R_1^{(P_0)}(\Psi_J)$ as follows:

$$\begin{aligned} R_1^{(P_0)}(\Psi_J) = {} & E_{P_0}\left[\mathbf{1}\{\Psi_J(1) \neq 1\}\mathbf{1}\{Y = 1\}\mathbf{1}\{Z_J = 1\}\right] \\ & + E_{P_0}\left[\mathbf{1}\{\Psi_J(0) \neq 1\}\mathbf{1}\{Y = 1\}\mathbf{1}\{Z_J = 0\}\right] \\ & + E_{P_0}\left[\mathbf{1}\{\Psi_J(1) \neq 0\}\mathbf{1}\{Y = 0\}\mathbf{1}\{Z_J = 1\}\right] \\ & + E_{P_0}\left[\mathbf{1}\{\Psi_J(0) \neq 0\}\mathbf{1}\{Y = 0\}\mathbf{1}\{Z_J = 0\}\right]. \end{aligned}$$

From this decomposition, we deduce that

$$\begin{aligned} R_1^{(P_0)}(\Psi_J) = {} & \alpha_J\left[\mathbf{1}\{\Psi_J(1) \neq 1\}\eta_J(1) + (1 - \mathbf{1}\{\Psi_J(1) \neq 1\})(1 - \eta_J(1))\right] \\ & + (1 - \alpha_J)\left[\mathbf{1}\{\Psi_J(0) \neq 1\}\eta_J(0) + (1 - \mathbf{1}\{\Psi_J(0) \neq 1\})(1 - \eta_J(0))\right]. \end{aligned}$$

Using the facts that, firstly, $\Psi_J(1) \neq 1$ implies $\eta_J(1) < 1/2$ and, secondly, $\Psi_J(0) \neq 1$ implies $\eta_J(0) < 1/2$, we obtain that

$$\begin{aligned} 1/2 - \left[\mathbf{1}\{\Psi_J(1) \neq 1\}\eta_J(1) + (1 - \mathbf{1}\{\Psi_J(1) \neq 1\})(1 - \eta_J(1))\right] &= \left|\eta_J(1) - 1/2\right|, \\ 1/2 - \left[\mathbf{1}\{\Psi_J(0) \neq 1\}\eta_J(0) + (1 - \mathbf{1}\{\Psi_J(0) \neq 1\})(1 - \eta_J(0))\right] &= \left|\eta_J(0) - 1/2\right|. \end{aligned}$$

The last equalities with $1/2 = \alpha_J/2 + (1 - \alpha_J)/2$ completes the proof. □

Lemma 1 teaches us that the larger the score γ_J, the better the classification based only on the pair J. Therefore, the TSP J_1^* is characterized by

$$J_1^* \in \arg\max_{J \in \mathcal{J}} \gamma_J. \tag{3}$$

It yields the TSP classifier for the misclassification risk:

$$\Psi_{J_1^*}(X) = \Psi_{J_1^*}(Z_{J_1^*}) = \mathbf{1}\{\eta_{J_1^*}(Z_{J_1^*}) \geq 1/2\}.$$

By (3) and Lemma 1, one can equivalently characterize this TSP classifier as

$$\Psi_{J_1^*} \in \arg\min_{t \in \mathcal{F}_{\text{pair}}} R_1^{(P_0)}(t),$$

showing that $\Psi_{J_1^*}$ can also be viewed as a risk minimizer–we will draw advantage of this remark later.

2.2.2 TSP for the weighted misclassification risk

We now introduce the TSP classifier for the weighted misclassification risk. It is the original TSP classifier of Geman *et al.* (2004). It can be viewed as weighted counterpart of the

TSP classifier $\Psi_{J_1^*}$ in the sense that it is a minimizer of the weighted misclassification risk over $\mathcal{F}_{\text{pair}}$.

For each $J \in \mathcal{J}$, we introduce the classifier $\Phi_J \in \mathcal{F}_J$ defined by

$$\Phi_J(X) = \Phi_J(Z_J) = \mathbf{1}\{p_J(1) > p_J(0)\}\mathbf{1}\{Z_J = 1\} + \mathbf{1}\{p_J(1) \leq p_J(0)\}\mathbf{1}\{Z_J = 0\}. \tag{4}$$

The classifier Φ_J votes for the class where the observed ordering between X_i and X_j is the more likely. We also introduce the score Δ_J of each $J \in \mathcal{J}$ as

$$\Delta_J = |p_J(1) - p_J(0)|\,. \tag{5}$$

The following lemma teaches us that one can interpret Δ_J as the weighted counterpart of γ_J and Φ_J as the weighted counterpart of Ψ_J.

Lemma 2. *Set $J \in \mathcal{J}$. For all $t \in \mathcal{F}_J$, it holds that*

$$R_2^{(P_0)}(t) - R_2^{(P_0)}(\Phi_J) = \Delta_J \left(\mathbf{1}\{t(1) \neq \Phi_J(1)\} + \mathbf{1}\{t(0) \neq \Phi_J(0)\}\right),$$

which implies that $\Phi_J \in \arg\min_{t \in \mathcal{F}_J} R_2^{(P_0)}(t)$. Moreover, $\Delta_J = 1 - R_2^{(P_0)}(\Phi_J)$.

Proof. Set $J \in \mathcal{J}, t \in \mathcal{F}_J$, and define

$$A_1 = E_{P_0}\left[\mathbf{1}\{t(Z_J) \neq 1\}|Y = 1\right], \text{ and } A_0 = E_{P_0}\left[\mathbf{1}\{t(Z_J) \neq 0\}|Y = 0\right].$$

We can decompose A_1 as

$$\begin{aligned} A_1 &= E_{P_0}\left[\mathbf{1}\{t(1) \neq 1\}\mathbf{1}\{Z_J = 1\}|Y = 1\right] + E_{P_0}\left[\mathbf{1}\{t(0) \neq 1\}\mathbf{1}\{Z_J = 0\}|Y = 1\right] \\ &= p_J(1)\mathbf{1}\{t(1) \neq 1\} + (1 - p_J(1))\mathbf{1}\{t(0) \neq 1\}. \end{aligned} \tag{6}$$

Similarly,

$$\begin{aligned} A_0 &= E_{P_0}\left[\mathbf{1}\{t(1) \neq 0\}\mathbf{1}\{Z_J = 1\}|Y = 0\right] + E_{P_0}\left[\mathbf{1}\{t(0) \neq 0\}\mathbf{1}\{Z_J = 0\}|Y = 0\right] \\ &= p_J(0)\mathbf{1}\{t(1) \neq 0\} + (1 - p_J(0))\mathbf{1}\{t(0) \neq 0\} \\ &= p_J(0)\left(1 - \mathbf{1}\{t(1) \neq 1\}\right) + (1 - p_J(0))\left(1 - \mathbf{1}\{t(0) \neq 1\}\right). \end{aligned} \tag{7}$$

Since $R_2^{(P_0)}(t) = A_1 + A_0$, we deduce from (6) and (7) that

$$R_2^{(p_0)}(t) = 1 + (p_J(0) - p_J(1))\mathbf{1}\{t(0) \neq 1\} + (p_J(1) - p_J(0))\mathbf{1}\{t(1) \neq 1\}. \tag{8}$$

Equation (8) holds in particular when $t = \Phi_J$. Therefore,

$$\begin{aligned} R_2^{(P_0)}(t) - R_2^{(P_0)}(\Phi_J) &= (p_J(0) - p_J(1))\left(\mathbf{1}\{t(0) \neq 1\} - \mathbf{1}\{\Phi_J(0) \neq 1\}\right) \\ &+ (p_J(1) - p_J(0))\left(\mathbf{1}\{t(1) \neq 1\} - \mathbf{1}\{\Phi_J(1) \neq 1\}\right) \\ &= \Delta_J\left(\mathbf{1}\{t(1) \neq \Phi_J(1)\} + \mathbf{1}\{t(0) \neq \Phi_J(0)\}\right), \end{aligned}$$

which is the first stated result. Moreover, a direct application of (8) with $t = \Phi_J$ yields second the result.

□

The TSP classifier for the weighted misclassification risk is $\Phi_{J_2^*}$ with J_2^* characterized by

$$J_2^* \in \arg\max_{J\in\mathcal{J}} \Delta_J. \tag{9}$$

By Lemma 2 and (9), it holds that $\Phi_{J_2^*} \in \arg\min_{t\in\mathcal{F}_{\text{pair}}} R_2^{(P_0)}(t)$, showing that $\Phi_{J_2^*}$ can be viewed as a minimizer of the weighted misclassification risk over $\mathcal{F}_{\text{pair}}$.

3 Empirical TSP classifiers

In this section, we introduce our empirical TSP classifiers and study their asymptotic behaviors in terms of risks control. Section 3.1 and Section 3.2 are devoted to the empirical TSP classification procedures for the misclassification risk and the weighted misclassification risk, respectively.

3.1 Empirical TSP classifier for the misclassification risk

The definition of the empirical TSP classifier for misclassification risk relies on estimators of J_1^* and $\eta_{J_1^*}$ that we plug into (1).

For every $t \in \mathcal{F}_{\text{pair}}$, we set $\widehat{R}_1(t) = \frac{1}{n}\sum_{k=1}^n \mathbf{1}\{t(X^k) \neq Y^k\}$, the empirical misclassification risk of t. For each $J \in \mathcal{J}$, let $\widehat{\gamma}_J = \widehat{\alpha}_J|\widehat{\eta}_J(1) - 1/2| + (1-\widehat{\alpha}_J)|\widehat{\eta}_J(0) - 1/2|$ be the empirical score, where $\widehat{\alpha}_J = \frac{1}{n}\sum_{k=1}^n Z_J^k$ and

$$\widehat{\eta}_J(z) = \begin{cases} \frac{1}{n\widehat{\beta}_J(z)} \sum_{k=1}^n \mathbf{1}\{Z_J^k = z\}\mathbf{1}\{Y^k = 1\} & \text{if } \widehat{\beta}_J(z) > 0 \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

with $\widehat{\beta}_J(z) = z\widehat{\alpha}_J + (1-z)(1-\widehat{\alpha}_J)$ (for both $z = 0, 1$). The random variable $\widehat{\eta}_J(z)$ is the empirical version of $\eta_J(z)$. If $\text{card}(\{k,\ Z_J^k = z\}) = 0$, we choose $\widehat{\eta}_J(z) = 1/2$ by convention. The plug-in estimator $\widehat{\Psi}_J(\cdot) = \mathbf{1}\widehat{\eta}_J(\cdot) \geq 1/2$ of Ψ_J implements a majority voting rule:

$$\widehat{\Psi}_J(z) = \begin{cases} 1 \text{ if } \text{card}(\{k,\ Z_J^k = z,\ Y^k = 1\}) \geq \text{card}(\{k,\ Z_J^k = z,\ Y^k = 0\}) \\ 0 \text{ otherwise,} \end{cases}$$

hence

$$\widehat{\Psi}_J \in \arg\min_{t\in\mathcal{F}_J} \widehat{R}_1(t). \tag{10}$$

We illustrate the classification rule $\widehat{\Psi}_J$ in Figure 1. Finally, $\widehat{J}_1 = \arg\max_{J\in\mathcal{J}} \widehat{\gamma}_J$ defines an estimator of the TSP J_1^* which leads to the emprical TSP classifier $\widehat{\Psi}_{\widehat{J}_1}$. A slight adaptation of the proof of Lemma 1 shows the following result:

Lemma 3. *For each $J \in \mathcal{J}$, it holds that $\widehat{\gamma}_J = 1/2 - \widehat{R}_1\left(\widehat{\Psi}_J\right)$.*

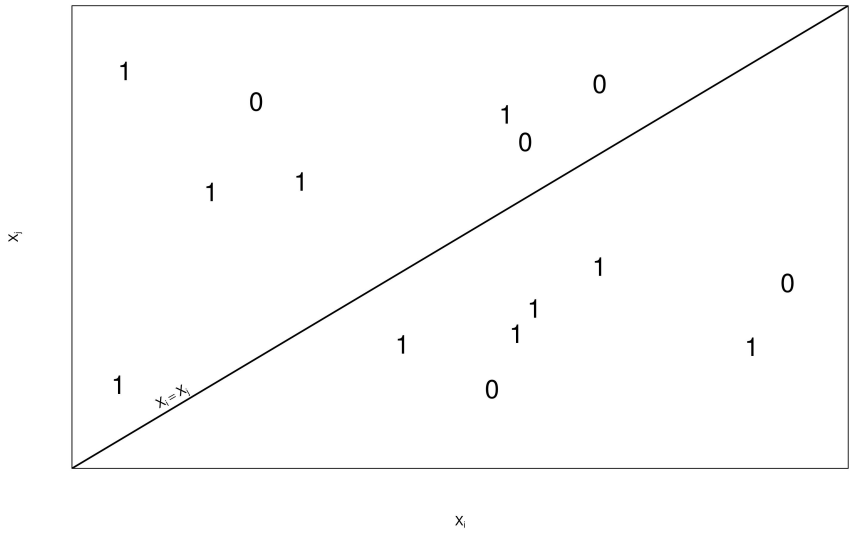

Figure 1: Illustration of the empirical classification rules $\widehat{\Psi}_J$ and $\widehat{\Phi}_J$ for a pair $J = (i, j)$. First, we have $\widehat{\eta}_J(1) = 5/8$ and $\widehat{\eta}_J(0) = 5/7$. Therefore, for a new observation X, $\widehat{\Psi}_J(X) = 1$ if $X_j \leq X_i$ and $\widehat{\Psi}_J(X) = 1$ if $X_j \leq X_i$. Moreover, the score $\widehat{\gamma}_J$ of the pair J is equal to $(8/15)\,|5/8 - 1/2| + (7/15)\,|5/7 - 1/2| = 1/6$. For the computation of $\widehat{\Phi}_J$, $\widehat{p}_J(1) = 1/2$ and $\widehat{p}_J(0) = 3/5$. Therfore, for a new observation X we obtain $\widehat{\Phi}_J(X) = 0$ if $X_j > X_i$ and $\widehat{\Phi}_J(X) = 1$ si $X_j \leq X_i$. Moreover, the score $\widehat{\Delta}_J$ of the pair J is equal to $|1/2 - 3/5| = 1/10$.

Lemma 3 and (10) entail that

$$\widehat{\Psi}_{\widehat{J}_1} \in \arg\min_{t \in \mathcal{F}_{\text{pair}}} \widehat{R}_1(t). \tag{11}$$

This property leads to the following asymptotic result which teaches us that, in the limit, $\widehat{\Psi}_{\widehat{J}_1}$ performs as well as the TSP classifier $\Psi_{J_1^*}$.

Theoreme 1. *It holds that*

$$0 \leq E\left[R_1^{(P_0)}(\widehat{\Psi}_{\widehat{J}_1}) - R_1^{(P_0)}(\Psi_{J_1^*})\right] = O\left(\sqrt{\frac{\log(\text{card}(\mathcal{J}))}{n}}\right).$$

This is the classical rate of convergence that one expects for a classifier which can be viewed as a minimizer, over the set of the classifiers defined on $\mathcal{F}_{\text{pair}}$, of the empirical misclassification risk (Bousquet *et al.*, 2004). We see clearly how the number of pairs affects the rate of convergence.

The proof of Theorem 1 is postponed to Section 7.2.

3.2 Empirical TSP classifier for the weighted misclassification risk

The definition of the empirical TSP classifier for the weighted misclassification risk relies on estimators of J_2^* and $p_{J_2^*}$. It is the original empirical TSP classifier of Geman *et al.* (2004).

Set $I(y) = \{k \leq n, Y^k = y\}$ and $N(y) = \mathrm{card}(I(y))$ for $y \in \{0,1\}$. For each $J \in \mathcal{J}$, the empirical score $\widehat{\Delta}_J$ is defined as $\widehat{\Delta}_J = |\widehat{p}_J(1) - \widehat{p}_J(0)|$, where, for each $y = 0, 1$,

$$\widehat{p}_J(y) = \frac{\mathbf{1}\{N(y) > 0\}}{N(y)} \sum_{k \in I(y)} Z_J^k$$

(with convention $0/0 = 0$). We also define for each $J \in \mathcal{J}$ the empirical counterpart $\widehat{\Phi}_J$ of Φ_J as $\widehat{\Phi}_J(X) = \widehat{\Phi}_J(Z_J) = \mathbf{1}\{\widehat{p}_J(1) > \widehat{p}_J(0)\}\mathbf{1}\{Z_J = 1\} + \mathbf{1}\{\widehat{p}_J(1) \leq \widehat{p}_J(0)\}\mathbf{1}\{Z_J = 0\}$. Finally, $\widehat{J}_2 \in \arg\max_{J \in \mathcal{J}} \widehat{\Delta}_J$ defines an estimator of the TSP J_2^*, which leads to the empirical TSP classifier $\widehat{\Phi}_{\widehat{J}_2}$ for the weighted misclassification risk.

The following asymptotic result shows that $\widehat{\Phi}_{\widehat{J}_2}$ performs as well, in the limit, as $\Phi_{J_2^*}$:

Theorem 2. *It holds that*

$$0 \leq E\left[\left(R_2^{(P_0)}(\widehat{\Phi}_{\widehat{J}_2}) - R_2^{(P_0)}(\Phi_{J_2^*})\right)\mathbf{1}\{0 < N(1) < n\}\right] = O\left(\sqrt{\frac{\log(\mathrm{card}(\mathcal{J}))}{np}}\right). \qquad (12)$$

The above rate of convergence is the same as in Theorem 1 with n replaced by np, the expected number of those observations $O^k = (X^k, Y^k)$ such that $Y^k = y$ where y is the rare outcome (*i.e.*, $p = P_0(Y = y)$). The additional factor $1/\sqrt{p}$ featured in (12) quantifies to what extent working with R_2 instead of R_1 makes the classification problem more difficult.

The proof of Theorem 2 is given in Section 7.3.

4 Cross-validated TSP classifiers

This section parallels Section 3. The main idea is to adopt a different approach to estimate the two TSP classifiers: instead of building the empirical TSP classifiers that we introduce and study in Section 3, we rely here on the cross-validation principle. By doing so, we could possibly achieve a greater stability and greater performances for the resulting estimators. The cross-validation principle has been widely studied both from the theoretical and pratical viewpoints (Dudoit and van der Laan, 2005; Arlot, 2007, and references therein). We define the cross-validated counterparts of R_1 and R_2 in Section 4.1. We introduce the two cross-validated TSP classifiers in Section 4.2, and we study their asymptotic behaviors in Section 4.3.

4.1 Cross-validated risk estimator

We set an integer $V \geq 2$ and a regular partition $(B_v)_{1 \leq v \leq V}$ of $\{1, \ldots, n\}$, *i.e.*, a partition such that, for each $v = 1, \ldots, V$, $\mathrm{card}(B_v) \in \{\lfloor n/V \rfloor, \lfloor n/V \rfloor + 1\}$.

For each $v \in \{1, \ldots, V\}$, we denote $D_n^{(v)}$ (respectively $D_n^{(-v)}$) the dataset $\{O^k, k \in B_v\}$ (respectively $\{O^k, k \notin B_v\}$), and define the corresponding empirical measures

$$\begin{aligned} P_n^{(v)} &= \frac{1}{\mathrm{card}(B_v)} \sum_{k \in B_v} \mathrm{Dirac}(O^k), \quad \text{and} \\ P_n^{(-v)} &= \frac{1}{n - \mathrm{card}(B_v)} \sum_{k \notin B_v} \mathrm{Dirac}(O^k). \end{aligned}$$

Let $\widehat{t}$ be a pair classifier, *i.e.*, a function mapping the empirical distribution to $\mathcal{F}_{\mathrm{pair}}$. Note that $\widehat{t}$ can be viewed simply as a black box algorithm that one applies to data. We characterize the empirical cross-validated risk estimators $\widehat{R}_{1,n}$, for the misclassification risk, and $\widehat{R}_{2,n}$, for the weighted misclassification risk, by

$$\begin{aligned} \widehat{R}_{1,n}(\widehat{t}) &= \frac{1}{V} \sum_{v=1}^{V} R_1^{(P_n^{(v)})} \left(\widehat{t}(P_n^{(-v)}) \right), \quad \text{and} \\ \widehat{R}_{2,n}(\widehat{t}) &= \frac{1}{V} \sum_{v=1}^{V} R_2^{(P_n^{(v)})} \left(\widehat{t}(P_n^{(-v)}) \right) \end{aligned}$$

for all $\widehat{t}$.

For each $v \in \{1, \ldots, V\}$ and $m = 1, 2$, $R_m^{(P_n^{(v)})}(\widehat{t}(P_n^{(-v)}))$ is the empirical estimator of $R_m^{(P_0)}(\widehat{t}(P_n^{(-v)}))$, based on $D_n^{(v)}$ and conditionally on $D_n^{(-v)}$. Obviously, it holds that, for every $v \in \{1, \ldots, V\}$,

$$\begin{aligned} R_1^{(P_n^{(v)})} \left(\widehat{t}(P_n^{(-v)}) \right) &= \frac{1}{\mathrm{card}(B_v)} \sum_{k \in I_v} L(O^k, \widehat{t}(P_n^{(-v)})), \quad \text{and} \\ R_2^{(P_n^{(v)})} \left(\widehat{t}(P_n^{(-v)}) \right) &= \frac{\mathbf{1}\{N_v(1) > 0\}}{N_v(1)} \sum_{k \in I_v(1)} L(O^k, \widehat{t}(P_n^{(-v)})) \\ &\quad + \frac{\mathbf{1}\{N_v(0) > 0\}}{N_v(0)} \sum_{k \in I_v(0)} L(O^k, \widehat{t}(P_n^{(-v)})), \end{aligned}$$

with $I_v(y) = \{k \in B_v, Y^k = y\}$ and $N_v(y) = \mathrm{card}(I_v(y))$ for $y = 0, 1$.

4.2 V-fold cross-validation principle

Let $\widehat{t}_1, \ldots, \widehat{t}_L$ be L pair classifiers (with L a possibly large integer).

We first address the case of the misclassification risk R_1. Each pair classifier can be viewed as a candidate to estimate the TSP classifier $\Psi_{J_1^*}$ for the misclassification risk R_1. One could for instance take $L = \mathrm{card}(\mathcal{J})$ and $\{\widehat{t}_1, \ldots, \widehat{t}_L\} = \{\widehat{\Psi}_J, \ J \in \mathcal{J}\}$. The

goal is to select a pair classifier in the collection $\{\widehat{t}_1, \ldots, \widehat{t}_L\}$, whose risk is the closest to $R_1^{(P_0)}(\Psi_{J_1^*})$. The V-fold cross-validation procedure consists in selecting the pair classifier which minimizes the cross-validated risk $\widehat{R}_{1,n}$. So, we introduce the cross-validated selector $\widehat{\ell}_{1,n} \in \arg\min_{\ell \leq L} \widehat{R}_{1,n}(\widehat{t}_\ell)$. The cross-validated TSP classifier is finally defined as $\widehat{\Psi}_n = \widehat{t}_{\widehat{\ell}_{1,n}}$.

Consider now the case of the weighted misclassification risk R_2. In that case, each pair classifier can be viewed as a candidate to estimate the TSP classifier $\Phi_{J_2^*}$ for the misclassification risk R_2. One could for instance take $L = \text{card}(\mathcal{J})$ and $\{\widehat{t}_1, \ldots, \widehat{t}_L\} = \{\widehat{\Phi}_J, J \in \mathcal{J}\}$. Similarly, we set $\widehat{\ell}_{2,n} \in \arg\min_{\ell \leq L} \widehat{R}_{2,n}(\widehat{t}_\ell)$ and $\widehat{\Phi}_n = \widehat{t}_{\widehat{\ell}_{2,n}}$.

4.3 Asymptotic perfomances of the cross-validated TSP classifiers

The asymptotic results that we obtain for the cross-validated TSP classifiers defined in Section 4.2 results are similar in nature to those of Dudoit and van der Laan (2005). They are expressed as comparisons to the oracle counterparts of the cross-validated TSP classifiers in terms of risks. Accordingly, define $\widetilde{R}_{1,n}$ and $\widetilde{R}_{2,n}$ the oracle counterparts of $\widehat{R}_{1,n}$ and $\widehat{R}_{2,n}$: for any $\widehat{t}$,

$$\begin{aligned}
\widetilde{R}_{1,n}(\widehat{t}) &= \frac{1}{V}\sum_{v=1}^{V} R_1^{(P_0)}\left(\widehat{t}(P_n^{(-v)})\right), \quad \text{and} \\
\widetilde{R}_{2,n}(\widehat{t}) &= \frac{1}{V}\sum_{v=1}^{V} E_{P_0}[L(O, \widehat{t}(P_n^{(-v)}))|Y=1]\mathbf{1}\{N_v(1) > 0\} \\
&\qquad + E_{P_0}[L(O, \widehat{t}(P_n^{(-v)}))|Y=0]\mathbf{1}\{N_v(0) > 0\}.
\end{aligned}$$

They yield the oracle counterparts $\widetilde{\ell}_{1,n} = \arg\min_{\ell \leq L} \widetilde{R}_{1,n}(\widehat{t}_\ell)$ and $\widetilde{\ell}_{2,n} = \arg\min_{\ell \leq L} \widetilde{R}_{2,n}(\widehat{t}_\ell)$ of $\widehat{\ell}_{1,n}$ and $\widehat{\ell}_{2,n}$, which yield in turn the oracle couterparts $\widetilde{\Psi}_n = \widehat{t}_{\widetilde{\ell}_{1,n}}$ and $\widetilde{\Phi}_n = \widetilde{\Phi}_n = \widehat{t}_{\widetilde{\ell}_{2,n}}$ of $\widehat{\Psi}_n$ and $\widehat{\Phi}_n$. We obtain the following result:

Theoreme 3. *It holds that*

$$E\left[\tilde{R}_{1,n}(\widehat{\Psi}_n) - \tilde{R}_{1,n}(\tilde{\Psi}_n)\right] = O\left(\sqrt{\frac{\log(L)}{\lfloor n/V \rfloor}}\right), \quad \textit{and} \tag{13}$$

$$E\left[\tilde{R}_{2,n}(\widehat{\Phi}_n) - \tilde{R}_{2,n}(\tilde{\Phi}_n)\right] = O\left(\sqrt{\frac{\log(L)}{\lfloor n/V \rfloor p}}\right). \tag{14}$$

As usual when one deals with cross-validated estimators, the theorem compares $\widehat{\Psi}_n$ and $\widehat{\Phi}_n$ to their oracle counterparts in terms of the oracle cross-validated risks. The theorem teaches us that, in the limit, $\widehat{\Psi}_n$ and $\widehat{\Phi}_n$ perform as well as $\widetilde{\Psi}_n$ and $\widetilde{\Phi}_n$.

If we choose $\{\widehat{t}_1, \ldots, \widehat{t}_L\}$ equal to $\{\widehat{\Psi}_J, J \in \mathcal{J}\}$ or $\{\widehat{\Phi}_J, J \in \mathcal{J}\}$, then the results in Theorem 3 are similar to those in Theorems 1 and 2. However, the rates of convergence in Theorem 3 are slightly slower than those of Theorems 1 and 2 due to the factor $\sqrt{V}$.

Equation (13) directly stems from (Dudoit and van der Laan, 2005, Theorem 2). The proof of (14) is postponed to Section 7.4.

5 Numerical study

We gather here the presentations of the application to a real dataset, and the results of a simulation study. The `R` (R Core Team, 2012) coding of our original TSP procedures was eased by the `tspair` package of Leek (2012).

5.1 Application on a real dataset

The different versions of the TSP classifier were applied to the Central Nervous System (CNS) cancer dataset. Originally used by Pomeroy *et al.* (2002) for a study of *medulloblastoma* (a brain tumor), this dataset is included in the `R`-package `stepwiseCM`. The CNS dataset consists of the 60 vectors of gene expression measurements of 7128 genes of 60 patients who received a treatment of medulloblastomas. Twenty-one patients died within two years after the end of their treatment. We tackle the classification problem of recovering whether the patient died or survived based on the gene expression measurements. More specifically, we evaluate the risks R_1 and R_2 achieved by the different versions of the TSP classifiers and the stepwise classification rule (implemented in the `R`-package `stepwiseCM`). We actually provide two different evaluations, relying either on the leave-one-out rule or on the validation hold-out rule. The training and validation sets (respectively made of 40 and 20 patients) are defined in the package.

We refer the reader to Table 1 for a succinct presentation of each classifier, and to Tables 2 and 3 for the evaluations of their performances (by leave-one-out in Table 2 and by validation hold-out in Table 3).

	classifier
`tsp1`	empirical TSP classifier for R_1
`tsp2`	empirical TSP classifier for R_2
`ctsp1(2)`	2-fold cross-validated TSP for R_1
`ctsp1(5)`	5-fold cross-validated TSP for R_1
`ctsp2(2)`	2-fold cross-validated TSP for R_2
`ctsp2(5)`	5-fold cross-validated TSP for R_2
`stepwise`	stepwise classifier

Table 1: Description of the different classifiers involved in the numerical study.

Four features of Table 2 and 3 are specially worth commenting on.

First, it appears that `tsp1` performs better than `tsp2` in terms of the misclassification error rate $\widehat{R}_1$ for both performances evaluations. Note that a TSP for R_1 (R_2, respectively) is not necessarily a TSP for R_2 (R_1, respectively).

leave-one-out rule

	`tsp1`	`tsp2`	`ctsp1(2)`	`ctsp1(5)`	`ctsp2(2)`	`ctsp2(5)`	`stepwise`
$1-\widehat{R}_1$	0.78	0.38	0.80	0.82	0.50	0.38	0.68
$1-\widehat{R}_2$	1.40	0.69	1.52	1.42	0.63	0.96	1.22

Table 2: Performances of the different versions of the TSP classifier on the dataset *CNS*, with leave-one-out evaluation.

validation hold-out rule

	`tsp1`	`tsp2`	`ctsp1(2)`	`ctsp1(5)`	`ctsp2(2)`	`ctsp2(5)`	`stepwise`
$1-\widehat{R}_1$	0.80	0.65	0.80	0.80	0.65	0.65	0.85
$1-\widehat{R}_2$	1.43	1.12	1.43	1.43	1.43	1.43	1.59

Table 3: Performances of the different versions of the TSP classifier on the dataset *CNS*, with validation hold-out evaluation.

Second, and perhaps disappointingly at first glance, we also see that `tsp2` does not perform better than `tsp1` in terms of the weighted misclassification error rate $\widehat{R}_2$, although the definition of `tsp2` relies on the weighted misclassification risk R_2, and although Theorem 2 guarantees its R_2-consistency. This may be a numerical illustration of the fact that its rate of convergence is $O(1/\sqrt{np})$, with p the proportion of the rarer class (and not $O(1/\sqrt{np})$). In the numerical example, 21 observations out of 60 belong to the rarer class.

Third, let us comment on the interest of the cross-validated versions of `tsp1` and `tsp2`. On the one hand, we note that both cross-validated versions of `tsp1` perform at least as well as `tsp1` in terms of $\widehat{R}_1$ and for both performances evaluations. On the other hand, we note that except for `ctsp2(2)` the cross-validated versions of `tsp2` perform better than `tsp2` in terms of $\widehat{R}_2$ and for both performances evaluations.

Fourth, comparing what can be compared, `tsp1` performs better than `stepwise` in terms of leave-one-out evaluation (5 more correct labellings) of the performances relative to $\widehat{R}_1$, but slightly worse in terms of validation hold-out evaluation relative to $\widehat{R}_1$ (one less correct labelling).

5.2 Simulation study

In light of the third comment above, we now undertake a simulation study of the influence of the sample size and true probability of the rarer class on the performances of `tsp2` relative to those of `tsp1`. The simulation scheme relies on the dataset CNS. To lessen the computational burden, we only consider the gene expression measurements of the first 100 genes of the original dataset. The simulation of an observation (X, Y) meets the following constraints:

(i) The label Y is drawn from the Bernoulli law with parameter $1-p = 0.8$. Thus, the true probability of the rarer class equals $p = 0.2$.

(ii) The vector of gene expression measurements X is subsequently drawn conditionally on Y from a slightly pertubed version of the empirical conditional distribution of X given Y in the CNS dataset.

We rely on the leave-one-out rule to evaluate and compare the performances of `tsp1` and `tsp2`. More specifically, we repeat independently $B = 100$ times the following steps:

1. simulate a dataset of sample size $n = 60$ (hence $np = 12$);
2. compute the performances of `tsp1` and `tsp2` (leave-one-out rule) over the rare and the frequent classes separately.

From these results, we compute the mean and standard deviation of the perfomances obtained by each classifier. The results are presented in Table 4.

	classification performances (leave-one-out rule)		
	rare+frequent class	rare class	frequent class
`tsp1`	0.75 (0.11)	0.13 (0.19)	0.89 (0.09)
`tsp2`	0.61 (0.17)	0.45 (0.31)	0.65 (0.19)

Table 4: Classification performances of `tsp1` and `tsp2` on simulated data with $np = 12$ ($p = 0.2$, $n = 60$). We report the empirical mean (and standard deviation, between parentheses) of the performances of each classifier over both classes (first column), the rare class (second column) and the frequent class (third column).

Although the standard deviations are rather large (especially for `tsp2`), we can draw interesting conclusions from Table 4. (Note that in each column, the standard deviations are larger for `tsp2` than for `tsp1`. This may be due to the fact that the rate of convergence of `tsp2` is $O(1/\sqrt{np})$ and not $O(1/\sqrt{n})$—more on this later). First, we see again that `tsp1` performs better than `tsp2` in terms of R_1. Inspecting the second and third columns of the table confirms the intuition that this happens because `tsp1` does a good job on the frequent class and a poor one (at low cost for R_1) on the rare class. By construction, `tsp2` outperforms `tsp1` as far as the rare class is concerned.

We now take a closer look at the influence on `tsp2` of the sample size for fixed $p = 0.2$. For that sake, we repeat independently $B = 100$ times the above two-step simulation scheme with $n = 300$ and for `tsp2` only. The results are presented in Table 5.

It is striking that the performances of `tsp2` on both classes and on the frequent class alone are almost identical in Tables 4 and 5. (In particular, this suggests that the larger standard deviations attached to `tsp2` relative to `tsp1` are not due to the difference in rates of convergence.). On the contrary, increasing the sample size does seem to enhance the performances of `tsp2` over the rare class (both in mean and standard deviation).

6 Discussion

The TSP procedures for binary classification that we have studied here involve only one TSP. In future work, we will extend our results to TSP procedures for multi-class clas-

	classification performance (leave-one-out rule)		
	rare+frequent class	rare class	frequent class
`tsp2`	0.64 (0.13)	0.54 (0.22)	0.66 (0.17)

Table 5: Classification performances of `tsp2` only on simulated data with $np = 60$ ($p = 0.2$, $n = 300$). We report the empirical mean (and standard deviation, between parentheses) of the performances computed over both classes (first column), the rare class (second column) and the frequent class (third column).

sification that may involve several TSPs, in the spirit of (Tan *et al.*, 2005; Zhou *et al.*, 2012).

Obviously, the TSP procedures do not lead in general to optimal classification rules. In future work, we will characterize and study the families of distibutions for which the TSP procedures lead to (near) optimal classifiers.

Acknowledgments

The author thanks warmly his supervisor A.Chambaz for his helpful suggestions throughout this work.

7 Proof

This section gathers the proofs of the Theorems 1, 2 and 3.

7.1 Two useful lemmas

Lemma 4. *Set two positive integers N, M and introduce the function f defined on the set of non-negative real numbers by $f(x) = \min(1, \exp(\log(2M) - 2Nx^2))$. The following inequality holds:*

$$\int_0^{+\infty} f(x)\mathrm{d}x \leq \sqrt{\frac{\log(2M)}{2N}} + \frac{\sqrt{\pi}}{2\sqrt{2N}}.$$

Proof. For all $x \geq 0$, we have $f(x) = \exp(-(2Nx^2 - \log(2M))_+)$. Therefore

$$\int_0^{+\infty} f(x)\mathrm{d}x = \sqrt{\frac{\log(2M)}{2N}} + \int_{x \geq \sqrt{\frac{\log(2M)}{2N}}} \exp(-(2Nx^2 - \log(2M)))\mathrm{d}x. \tag{15}$$

Since $a^2 - b^2 \geq (a-b)^2$ for $a \geq b \geq 0$, note that:

$$\begin{aligned} \int_{x \geq \sqrt{\frac{\log(2M)}{2N}}} \exp(-(2Nx^2 - \log(2M)))\mathrm{d}x &\leq \int_{x \geq \sqrt{\frac{\log(2M)}{2N}}} \exp\left(-2N\left(x - \sqrt{\frac{\log(2M)}{2N}}\right)^2\right)\mathrm{d}x \\ &= \frac{1}{\sqrt{2N}} \int_0^{+\infty} \exp(-x^2)\mathrm{d}x = \frac{\sqrt{\pi}}{2\sqrt{2N}}. \end{aligned} \tag{16}$$

Finally, Equation (15) and Equation (16) yield the result. □

Lemma 5. *Let $Z \overset{\mathcal{L}}{=} \mathcal{B}(n,p)$ be a binomial random variable. Then*

$$E\left[\frac{\mathbf{1}\{Z>0\}}{\sqrt{Z}}\right] \leq \sqrt{\frac{2}{(n+1)p}}.$$

Proof. By the Cauchy-Schwartz inequality, it holds that

$$\left(E\left[\frac{1}{\sqrt{Z+1}}\right]\right)^2 \leq E\left[\frac{1}{Z+1}\right] = \sum_{k=0}^{n} \frac{1}{k+1}\binom{n}{k}p^k(1-p)^{n-k} = \int_0^1 (xp+1-p)^n \mathrm{d}x \leq \frac{1}{(n+1)p}.$$

Now, since $1/\sqrt{k} \leq \sqrt{2}/\sqrt{k+1}$ for all $k \geq 1$, we obtain

$$E\left[\frac{\mathbf{1}\{Z>0\}}{\sqrt{Z}}\right] = \sum_{k=1}^{n} \frac{1}{\sqrt{k}}\binom{n}{k}p^k(1-p)^{n-k} \leq \sqrt{2}\sum_{k=1}^{n} \frac{1}{\sqrt{k+1}}\binom{n}{k}p^k(1-p)^{n-k} \leq \sqrt{2}E\left[\frac{1}{\sqrt{Z+1}}\right] \leq \sqrt{\frac{2}{(n+1)p}},$$

which is the stated result. □

7.2 Proof of Theorem 1

The proof of Theorem 1 relies on the characterization (11) of the empirical TSP classifier. We have:

$$0 \leq R_1^{(P_0)}(\widehat{\Psi}_{\widehat{J}_1}) - R_1^{(P_0)}(\Psi_{J_1^*}) = \left(R_1^{(P_0)}(\widehat{\Psi}_{\widehat{J}_1}) - \widehat{R}_1(\widehat{\Psi}_{\widehat{J}_1})\right) + \left(\widehat{R}_1(\widehat{\Psi}_{\widehat{J}_1}) - R_1^{(P_0)}\left(\Psi_{J_1^*}\right)\right).$$

By (11), this yields that

$$0 \leq R_1^{(P_0)}(\widehat{\Psi}_{\widehat{J}_1}) - R_1^{(P_0)}(\Psi_{J_1^*}) \leq 2\sup_{t\in\mathcal{F}_{\text{pair}}}\left|R_1^{(P_0)}(t) - \widehat{R}_1(t)\right|.$$

Therefore

$$0 \leq E\left[R_1^{(P_0)}(\widehat{\Psi}_{\widehat{J}_1}) - R_1^{(P_0)}(\Psi_{J_1^*})\right] \leq 2E\left[\sup_{t\in\mathcal{F}_{\text{pair}}}\left|R_1^{(P_0)}(t) - \widehat{R}_1(t)\right|\right]. \tag{17}$$

Next, we provide an upper bound for the right-hand side expectation.

By the Bonferroni inequality, we have for all $h \geq 0$,

$$P\left(\sup_{t\in\mathcal{F}_{\text{pair}}}\left|R_1^{(P_0)}(t) - \widehat{R}_1(t)\right| \geq h\right) \leq \min\left(1, \sum_{t\in\mathcal{F}_{\text{pair}}} P\left(\left|R_1^{(P_0)}(t) - \widehat{R}_1(t)\right| \geq h\right)\right).$$

Since for each $t \in \mathcal{F}_{\text{pair}}$, $\widehat{R}_1(t)$ is an empirical mean of i.i.d Bernouilli random variables with common mean $R_1^{(P_0)}(t)$, we deduce from Hoeffding's inequality that:

$$P\left(\sup_{t\in\mathcal{F}_{\text{pair}}}\left|R_1^{(P_0)}(t)-\widehat{R}_1(t)\right|\geq h\right)\leq \min\left(1,\exp\left(\log(2\text{card}(\mathcal{F}_{\text{pair}}))-2nh^2\right)\right).$$

Now, with $\text{card}(\mathcal{F}_{\text{pair}})=4\text{card}(\mathcal{J})$,

$$\begin{aligned}E\left[\sup_{t\in\mathcal{F}_{\text{pair}}}\left|R_1^{(P_0)}(t)-\widehat{R}_1(t)\right|\right] &= \int_0^{+\infty} P\left(\sup_{t\in\mathcal{F}_{\text{pair}}}\left|R_1^{(P_0)}(t)-\widehat{R}_1(t)\right|\geq h\right)\mathrm{d}h\\ &\leq \sqrt{\frac{\log(8\text{card}(\mathcal{J}))}{2n}}+\frac{\sqrt{\pi}}{2\sqrt{2n}},\end{aligned}$$

by Lemma 4. Then (17) yields the theorem.

7.3 Proof of Theorem 2

We have:

$$\begin{aligned}0\leq R_2^{(P_0)}(\widehat{\Phi}_{\widehat{J}_2})-R_2^{(P_0)}(\Phi_{J_2^*}) &= \left(R_2^{(P_0)}(\widehat{\Phi}_{\widehat{J}_2})-R_2^{(P_0)}(\Phi_{\widehat{J}_2})\right)+\left(R_2^{(P_0)}(\Phi_{\widehat{J}_2})-R_2^{(P_0)}(\Phi_{J_2^*})\right)\\ &= \left(R_2^{(P_0)}(\widehat{\Phi}_{\widehat{J}_2})-R_2^{(P_0)}(\Phi_{\widehat{J}_2})\right)+\left(\Delta_{J_2^*}-\Delta_{\widehat{J}_2}\right)\\ &= \left(R_2^{(P_0)}(\widehat{\Phi}_{\widehat{J}_2})-R_2^{(P_0)}(\Phi_{\widehat{J}_2})\right)+\left(\Delta_{J_2^*}-\widehat{\Delta}_{J_2^*}\right)+\left(\widehat{\Delta}_{J_2^*}-\Delta_{\widehat{J}_2}\right)\\ &\leq \left(R_2^{(P_0)}(\widehat{\Phi}_{\widehat{J}_2})-R_2^{(P_0)}(\Phi_{\widehat{J}_2})\right)+\left(\Delta_{J_2^*}-\widehat{\Delta}_{J_2^*}\right)+\left(\widehat{\Delta}_{\widehat{J}_2}-\Delta_{\widehat{J}_2}\right),\end{aligned}$$

by definition of $\widehat{J}_2$.

To complete the proof, it remains to control $E\left[\mathbf{1}\{0<N(1)<n\}\sup_{J\in\mathcal{J}}\left|\Delta_J-\widehat{\Delta}_J\right|\right]$ and $E\left[\mathbf{1}\{0<N(1)<n\}\sup_{J\in\mathcal{J}}\left|R_2^{(P_0)}(\Phi_J)-R_2^{(P_0)}(\widehat{\Phi}_J)\right|\right]$, by relying on Lemmas 6 and 7.

Lemma 6. *For all $J\in\mathcal{J}$, it holds that*

$$R_2^{(P_0)}(\widehat{\Phi}_J)-R_2^{(P_0)}(\Phi_J) \leq 2\left(|\widehat{p}_J(1)-p_J(1)|+|\widehat{p}_J(0)-p_J(0)|\right), \quad \textit{and} \tag{18}$$

$$\left|\widehat{\Delta}_J-\Delta_J\right| \leq |\widehat{p}_J(1)-p_J(1)|+|\widehat{p}_J(0)-p_J(0)|. \tag{19}$$

Proof. Inequality (18) is a by-product of Lemma 2 and the fact that, for each $y\in\{0,1\}$, $\left(\widehat{\Phi}_J(y)\neq\Phi_J(y)\right)$ implies $\Delta_J=|p_J(1)-p_J(0)|\leq|\widehat{p}_J(1)-p_J(1)|+|\widehat{p}_J(0)-p_J(0)|$. To show this implication, we just check one of the four different cases that can arise (the others can be addressed similarly). For instance, if $y=1$ and $\widehat{\Phi}_J(1)=0$ then $\widehat{p}_J(0)\geq\widehat{p}_J(1)$ and $p_J(0)<p_J(1)$. Thus,

$$\begin{aligned}\Delta_J=|p_J(1)-p_J(0)|=p_J(1)-p_J(0) &= (p_J(1)-\widehat{p}_J(1))+(\widehat{p}_J(1)-p_J(0))\\ &\leq (p_J(1)-\widehat{p}_J(1))+(\widehat{p}_J(1)-p_J(0))\\ &\leq |\widehat{p}_J(1)-p_J(1)|+|\widehat{p}_J(0)-p_J(0)|.\end{aligned}$$

Inequality (19) relies on a direct application of the reverse triangle inequality. □

Lemma 7. *For each $y \in \{0,1\}$, it holds that*

$$E_{P_0}\left[\mathbf{1}\{N(y)>0\}\sup_{J\in\mathcal{J}}|\widehat{p}_J(y)-p_J(y)|\right] \leq \sqrt{\frac{2\log(2\mathrm{card}(\mathcal{J}))}{np}}+\sqrt{\frac{\pi}{2np}}. \tag{20}$$

Proof. By symmetry, it suffices to present the proof in the case where $y=1$. Let $\mathcal{Y}$ denotes the σ-field spanned by $\{Y^k, k=1,\ldots,n\}$. We have:

$$E_{P_0}\left[\mathbf{1}\{N(1)>0\}\sup_{J\in\mathcal{J}}|\widehat{p}_J(1)-p_J(1)|\right] = E_{P_0}\left[E\left[\mathbf{1}\{N(1)>0\}\sup_{J\in\mathcal{J}}|\widehat{p}_J(1)-p_J(1)|\,\Big|\mathcal{Y}\right]\right],$$

which equals

$$E_{P_0}\left[\mathbf{1}\{N(1)>0\}\int_0^{+\infty} P\left(\sup_{J\in\mathcal{J}}|\widehat{p}_J(1)-p_J(1)|\geq h|\mathcal{Y}\right)\mathrm{d}h\right].$$

If $N(1)>0$ then conditionally on $\mathcal{Y}$ and for each $J\in\mathcal{J}$, the random variable $\widehat{p}_J(1)$ is an empirical mean of i.i.d Bernoulli random variables with common mean $p_J(1)$. Therefore, by the Bonferroni and Hoeffding inequalities, we obtain for all $h\geq 0$:

$$\begin{aligned}\mathbf{1}\{N(1)>0\}P\left(\sup_{J\in\mathcal{J}}|\widehat{p}_J(1)-p_J(1)|\geq h\Big|\mathcal{Y}\right)&\\ \leq \mathbf{1}\{N(1)>0\}\min\left(1,\exp\left(\log\left(2\mathrm{card}(\mathcal{J})-2N(1)h^2\right)\right)\right.&.\end{aligned}$$

Applying Lemma 4 then gives

$$\begin{aligned}\mathbf{1}\{N(1)>0\}\int_0^{+\infty} P\left(\sup_{J\in\mathcal{J}}|\widehat{p}_J(1)-p_J(1)|\geq h\Big|\mathcal{Y}\right)\mathrm{d}h&\\ \leq \frac{\mathbf{1}\{N(1)>0\}}{\sqrt{2N(1)}}\left(\sqrt{\log\left(2\mathrm{card}(\mathcal{J})\right)}+\frac{\sqrt{\pi}}{2}\right).&\end{aligned} \tag{21}$$

Since $N(1)\overset{\mathcal{L}}{=}\mathcal{B}(n,p_1)$, (21) and Lemma 5 yield the result. □

7.4 Proof of Theorem 3

We recall that (13) directly stems from Dudoit and van der Laan (2005). We now give the proof of (14). By definition of $\tilde{\ell}_n^2$, one has

$$\begin{aligned}0\leq \tilde{R}_{2,n}(\widehat{\Phi}_n)-\tilde{R}_{2,n}(\tilde{\Phi}_n) &= (\tilde{R}_{2,n}(\widehat{\Phi}_n)-\widehat{R}_{2,n}(\widehat{\Phi}_n))+(\widehat{R}_{2,n}(\widehat{\Phi}_n)-\tilde{R}_{2,n}(\tilde{\Phi}_n))\\ &\leq (\tilde{R}_{2,n}(\widehat{\Phi}_n)-\widehat{R}_{2,n}(\widehat{\Phi}_n))+(\widehat{R}_{2,n}(\tilde{\Phi}_n)-\tilde{R}_{2,n}(\tilde{\Phi}_n))\\ &\leq 2\sup_{\ell\in L}\left|\widehat{R}_{2,n}(\widehat{t}_\ell)-\tilde{R}_{2,n}(\widehat{t}_\ell)\right|.\end{aligned}$$

Now, for each $\ell \in \{1, \ldots, L\}$, $\widehat{R}_{2,n}(\widehat{t}_\ell) - \tilde{R}_{2,n}(\widehat{t}_\ell)$ is equal to

$$\frac{1}{V}\sum_{v=1}^{V}\left[\frac{\mathbf{1}\{N_v(1)>0\}}{N_v(1)}\sum_{i\in I_v(1)}\left(L(O^i,\widehat{t}_\ell(P_n^{(-v)})) - E_{P_0}[L(O,\widehat{t}_\ell(P_n^{(-v)}))|Y=1]\right)\right]$$
$$+\frac{1}{V}\sum_{v=1}^{V}\left[\frac{\mathbf{1}\{N_v(0)>0\}}{N_v(0)}\sum_{i\in I_v(0)}\left(L(O^i,\widehat{t}_\ell(P_n^{(-v)})) - E_{P_0}[L(O,\widehat{t}_\ell(P_n^{(-v)}))|Y=0]\right)\right],$$

hence

$$\sup_{\ell\in L}\left|\widehat{R}_{2,n}(\widehat{t}_\ell) - \tilde{R}_{2,n}(\widehat{t}_\ell)\right| \leq \frac{1}{V}\sum_{v=1}^{V}\left(\sup_{\ell\in\{1,\ldots,L\}}\left|H_{\ell,v}^1\right| + \sup_{\ell\in\{1,\ldots,L\}}\left|H_{\ell,v}^0\right|\right), \tag{22}$$

where, for $y = 0, 1$,

$$H_{\ell,v}^y = \frac{\mathbf{1}\{N_v(y)>0\}}{N_v(y)}\sum_{i\in I_v(y)}\left(L(O^i,\widehat{t}_\ell(P_n^{(-v)})) - E_{P_0}[L(O,\widehat{t}_\ell(P_n^{(-v)}))|Y=y]\right).$$

For each $v \in \{1, \ldots, V\}$ and $y \in \{0, 1\}$, conditionally on $D_n^{(-v)}$ and $(Y^i)_{i\in B_v}$, $H_{\ell,v}^y$ is an empirical mean of i.i.d bounded centered variable. Thus, the Bonferroni and Hoeffding inequalities imply that, for all $h \geq 0$,

$$P\left(\sup_{\ell\in\{1,\ldots,L\}}\left|H_{\ell,v}^y\right| \geq h \middle| D_n^{(-v)}, (Y^i)_{i\in B_v}\right) \leq \min(1, \exp(\log(2L) - 2N_v(y)h^2)),$$

so that, for each $v \in \{1, \ldots, V\}$, we deduce by Lemma 4 that

$$E\left[\sup_{\ell\in\{1,\ldots,L\}}\left|H_{\ell,v}^y\right| \middle| D_n^{(-v)}, (Y^i)_{i\in B_v}\right] \leq \frac{\mathbf{1}\{N_v(y)>0\}}{\sqrt{2N_v(y)}}\left(\sqrt{\log(2L)} + \frac{\sqrt{\pi}}{2}\right).$$

Since $N_v(y) \stackrel{\mathcal{L}}{=} \mathcal{B}(n, p_y)$, we complete the proof by applying again Lemma 5 and (22).

References

Arlot, S. (2007). *Rééchantillonage et selection de modèles*. Thèse. Université Paris-Sud, Orsay.

Bousquet, O., Boucheron, S. and Lugosi, G. (2004). Introduction to statistical learning theory. *Advanced Lectures in Machine Learning, Springer* pp. 169–207.

Chambaz, A. and Denis, C. (2012). Classification in postural style. *Annals of Applied Statistics* **6**, 977–993.

Czajkowski, M. and Kretowski, M. (2011). Top scoring pair decision tree for gene expression data analysis. *Software Tools and Algorithms for Biological Systems, Springer* **3**, 27–36.

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.* **2**, 131–154. ISSN 1572-3127.

Geman, D., d'Avignon, C., Naiman, D. Q. and Winslow, R. L. (2004). Classifying gene expression profiles from pairwise mRNA comparisons. *Statistical Applications in Genetics and Molecular Biology* **3**.

Leek, J. T. (2012). *tspair: Top Scoring Pairs for Microarray Classification*. R package version 1.16.0.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E. and Golub, T. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Tan, A., Naiman, D., Xu, L., Winslow, R. and D, G. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* **21**, 3896–3904.

Yang, Y., Dudoit, S., Luu, P., Peng, V., Ngal, J. and Speed, T. (2001). Normalization for cdna microarray data. *Microarrays: Optical Technologies and Informatics* **4266**, 141–152.

Zhou, C., Wang, S., Blanzieri, E. and Liang, Y. (2012). An entropy-based improved k-top scoring pairs (tsp) method for classifying human cancers. *African Journal of Biotechnology* **41**, 10438–10445.