Approche hybride par segmentation aveugle

Sommaire

9.1 Principe

Il est possible d'introduire la dimension temporelle dans le processus en combinant l'approche statique de la classification à une méthode dynamique de découpage du flux audio en segments dont le contenu est acoustiquement homogène. Cette notion de segment rejoint celle introduite dans le chapitre précédent dans la présentation des modèles de segments associés aux HSMM (section 8.4).

Ce découpage, appelé segmentation, se fait non par l'analyse directe du contenu des trames, mais par la recherche des points de changement délimitant les segments successifs. Cette méthodologie est généralement employée dans le domaine du suivi ou de l'identification de locuteurs [128][6][34][74][227][10], de manière à mettre en évidence les tours de paroles ne contenant qu'un locuteur à la fois. La reconnaissance de locuteur est un problème particulier qui dépasse le cadre de cette thèse, parce qu'il implique un nombre très important de classes (de locuteurs), dont il est en général impossible de connaître la totalité, comme par exemple sur des bulletins d'informations radiophoniques, où potentiellement n'importe qui peut être présent dans une interview. Cette contrainte implique donc l'usage de techniques de segmentation dite aveugles (ou non-supervisées) qui ne reposent pas sur l'apprentissage préalable des modèles des classes susceptibles d'être observées. Nous en expliquerons le principe dans la section 9.2.

Le processus de segmentation aveugle nous permet, en reprenant les notations de la section 8.4, de répartir la séquence des trames $i = 1, \ldots, n_k$ du fichier k, en S_k segments successifs dont les indices de trames initiaux sont respectivement $i_1^k, \ldots, i_{S_k}^k$ (avec bien sûr $i_1^k = 1$ et par convention

$i_{S_k+1} = n_k + 1$). Contrairement au post-traitement par HSMM, nous ne connaissons pas les classes $c_1^k, \ldots, c_{S_k}^k$ associées aux segments mais nous supposons ces derniers homogènes par rapport aux classes acoustiques considérées. On remarque que l'on a également substitué aux longueurs, la donnée équivalente des indices de début de trames qui convient mieux dans le cadre présent puisque la segmentation aveugle consiste en la détermination des indices i_s^k.

L'approche hybride peut alors se faire de deux manières différentes :

- Comme **pré-traitement** : c'est la méthode que l'on trouve généralement dans la littérature. Chaque segment est classifié dans son ensemble par une unique prise de décision. Ainsi la conjonction des différentes trames du segment permet d'accroître la confiance dans la décision. Les intégrations temporelles des descripteurs sont ainsi plus fiables puisqu'elles portent sur une période plus large.
- Comme **post-traitement** : c'est l'approche que nous suivons dans ce document. L'information de segmentation intervient ici après la classification, sur la donnée des probabilités a posteriori, comme pour les approches de post-traitement présentées dans le chapitre précédent. On associe à chaque segment s la classe $\hat{c}_s$ maximisant la somme des probabilités sur les trames du segment. Soit :

$$\hat{c}_s = \underset{1 \leq c \leq C}{\arg\max} \sum_{i=i_s^k}^{i_{s+1}^k - 1} p_c(i),$$

où l'on rappelle que $p_c(i)$ est la probabilité a posteriori évaluée sur la trame i pour la classe c, et i_s^k est l'indice de la première trame du segment s. Comme pour l'approche par pré-traitement, la conjonction des décisions sur les trames d'un segment renforce la fiabilité de la décision. Dans le cas d'une erreur, elle peut cependant avoir l'effet contraire et contaminer l'ensemble des trames d'un segment.

Les frontières entre segments étant à priori caractérisées par une transition plus ou moins brusque d'un modèle acoustique à un autre, on détermine celles-ci par des algorithmes de détection de rupture, dont nous présentons le principe dans la section suivante.

9.2 Détection de rupture

Le mécanisme de la détection de rupture est généralement assez simple. On suppose que l'on observe le signal contenu dans une fenêtre d'analyse W de n échantillons représentés par les vecteurs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, et que l'on souhaite examiner l'hypothèse d'une rupture à l'indice t . On exploite pour cela les signaux des sous-fenêtres antérieure et postérieure, respectivement $W_1 = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}]$ et $W_2 = [\boldsymbol{x}_t, \ldots, \boldsymbol{x}_n]$, de tailles respectives $n_1 = t$ et $n_2 = n - t$. La figure 9.1 résume la configuration des fenêtres d'analyse mises en jeu.

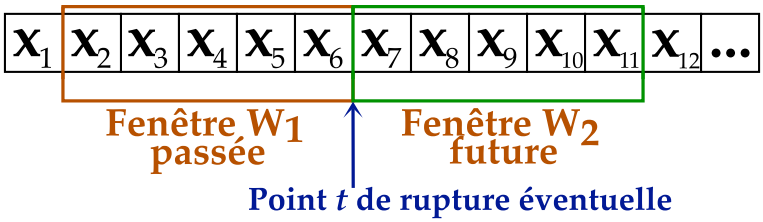


FIGURE 9.1 – Exemple de configuration pour les fenêtres d'analyse passée et future, respectivement W_1 et W_2 (ici de mêmes tailles $n_1 = n_2 = 5$), par rapport à l'instant d'hypothèse t.

Les algorithmes proposés pour la détection de rupture se basent sur une modélisation stochastique des signaux des fenêtres d'analyse. Ainsi on suppose que les exemples de la fenêtre d'analyse W sont les réalisations d'une variable aléatoire gouvernée par la loi de probabilité $P_0(\boldsymbol{X})$; de même, les lois de probabilité $P_1(\boldsymbol{X})$ et $P_2(\boldsymbol{X})$ gouvernent respectivement la réalisation des exemples des fenêtres W_1 et W_2.

Une taxonomie communément admise dans la littérature [49][120][250] distingue les quatre modalités suivantes pour la détection de rupture :

- **Approche énergétique** : en se basant sur la supposition que les tours de parole sont généralement séparés par de courtes périodes de silence entre les locuteurs, certains algorithmes rudimentaires ne basent la segmentation que sur le seuillage d'un critère d'énergie court terme [249]. Cette approche, déjà contestée dans le domaine de la parole pure, n'est pas pertinente sur un signal audio quelconque, où la présence de silences intermédiaires est plutôt l'exception que la règle.
- **Approche par modèles** : consiste à modéliser chacune des classes mises en jeu afin de classifier le contenu des fenêtres W_1 et W_2 sur le critère du maximum de vraisemblance [17]. En plus de ne pas être aveugle, cette approche est le processus exactement inverse de ce que nous recherchons ici puisque la classification est utilisée comme outil de segmentation.
- **Approche métrique** : la détermination des frontières entre segments est basée sur la recherche des maxima locaux d'une métrique qui évalue la similarité entre les modèles des fenêtres W_1 et W_2 [217][89]. C'est l'approche que nous suivrons ici.
- **Approche sur critère d'information** : cette approche est similaire à la précédente mais substitue aux critères métriques, nécessitant un seuil de décision, une mesure d'information appelée Critère d'Information Bayésienne (BIC, pour *Bayesian Information Criterion*), pour laquelle le seuil est implicite [49][45][61], comme nous le verrons par la suite. Ce critère implique en outre un autre paradigme de détection. Tandis que l'approche métrique évalue une mesure de distance entre les deux fenêtres W_1 et W_2, cette approche compare les hypothèses d'absence et de présence de rupture. Nous examinons cette approche plus en détail dans la section suivante.

On trouve également dans la littérature plusieurs propositions d'algorithmes hybrides combinant les avantages complémentaires de deux approches ; par exemple la proposition de Kemp et al. [120] qui consiste en une approche par modèles basée sur un premier traitement par approche métrique, ou encore l'algorithme SEQDAC de Cheng et Wang [50]. De nombreux algorithmes hybrides [263][257][61][51] combinent le critère BIC à d'autres métriques comme la statistique T^2 (test basé sur les statistiques de premier et de second ordre de deux modèles probabilistes).

Nous présentons dans la section suivante quelques-uns des algorithmes classiques de détection de rupture, notamment le critère BIC, très largement exploité, afin de montrer leurs implications sur la stratégie de recherche de points de rupture multiples dans un signal. Nous introduirons par la suite certains critères plus récents et plus élaborés tirant parti des résultats sur les espaces à noyaux reproduisants et les machines à noyaux dans les sections 9.4 et 9.5.

9.3 Méthodes classiques

9.3.1 Un exemple d'approche métrique : divergence de Kullback Leibler

Le principe de l'approche métrique consiste à déterminer les points de rupture par une mesure de distance entre les fenêtres voisines W_1 et W_2, soumise à un seuil de détection adéquat. La stratégie de recherche appliquée est généralement celle des *fenêtres glissantes* [257][153][128][34], qui consiste à échantilloner la mesure de distance entre les fenêtres W_1 et W_2 de tailles fixes et égales à M, en glissant itérativement ces dernières d'un pas fixe du début à la fin de la séquence de trames. La figure 9.2 illustre le principe de la recherche par fenêtres glissantes, que nous suivons dans notre cadre expérimental.

Le choix de la métrique est un problème ouvert, et les sections suivantes apporteront diverses propositions pour ce point. Siegler et al. [217] proposent par exemple l'usage de la divergence de Kullback-Leibler, définie par :

$$KL(P \,|\, Q) = E_P \left[\log P(\boldsymbol{X}) - \log Q(\boldsymbol{X})\right],$$

où E_P est l'espérance par rapport à la probabilité $P(\boldsymbol{X})$; soit, sur les densités de probabilités p et

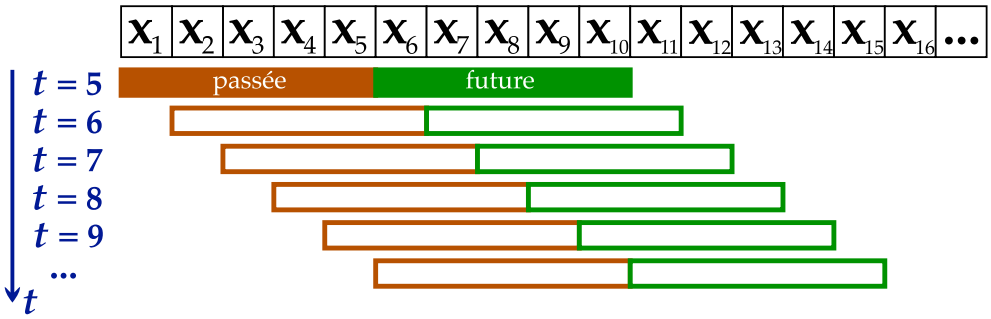


FIGURE 9.2 – Illustration de la recherche de rupture par fenêtres adjacentes glissantes.

q :

$$KL(P\,|\,Q) = \int_{\boldsymbol{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}.$$

Le terme de divergence met en lumière le caractère non symétrique de cette « semi-métrique ». On emploie donc en général la variante symétrisée de la mesure de Kullback-Leibler :

$$\begin{aligned} KL2(P,Q) &= KL(P\,|\,Q) + KL(Q\,|\,P) \\ &= \int_{\boldsymbol{X}} [p(\boldsymbol{x}) - q(\boldsymbol{x})] \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}. \end{aligned}$$

Dans le cas de la détection de rupture, les probabilités P et Q sont estimées par des modèles gaussiens $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ et $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ appris sur les exemples des fenêtres W_1 et W_2. On peut ainsi exprimer analytiquement la métrique $KL2$ dans le cas de distributions gaussiennes :

$$KL2(t) = \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right) (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} \operatorname{tr} \left[\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1 - 2\boldsymbol{I}\right].$$

On obtient donc un signal $KL2(M+1), \ldots, KL2(n-M)$ de distances entre fenêtres, sur lequel on appliquera l'heuristique de recherche de pics maxima présentée dans la section 9.6.

9.3.2 Rapport de vraisemblance généralisé (GLR)

Le critère GLR (*Generalized Likelihood Ratio*), introduit par Gish et Schmidt pour la segmentation aveugle [89], repose sur le test d'hypothèse introduit pour les approches sur critère d'information, à savoir la comparaison entre les deux hypothèses suivantes :

- H_0 : il n'y a pas de rupture. Tous les exemples de la fenêtre W sont donc générés par l'unique modèle $P_0(\boldsymbol{X})$.
- H_1 : il y a une rupture à l'instant t. Les exemples des fenêtres W_1 et W_2 sont donc respectivement générés par les modèles $P_1(\boldsymbol{X})$ et $P_2(\boldsymbol{X})$.

Le critère en question se base donc sur le rapport de vraisemblance entre les deux hypothèses, où l'on considère les échantillons i.i.d. :

$$r(t) = \frac{\prod_{i=1}^{n} P_0(\boldsymbol{x}_i)}{\prod_{i=1}^{t-1} P_1(\boldsymbol{x}_i) \prod_{i=t}^{n} P_2(\boldsymbol{x}_i)}. \tag{9.1}$$

On estime les probabilités par un modèle gaussien évalué sur les exemples des fenêtres considérées. Ainsi, chaque probabilité P_h est modélisée par une loi gaussienne $\mathcal{N}(\boldsymbol{\mu}_h; \boldsymbol{\Sigma}_h)$ de centre $\boldsymbol{\mu}_h$ et de matrice de covariance $\boldsymbol{\Sigma}_h$. En définitive on exploite généralement le logarithme du critère précédent calculé sur les modèles gaussiens estimés [61] :

$$R(t) = -\frac{n}{2} \log |\boldsymbol{\Sigma}_0| + \frac{n_1}{2} \log |\boldsymbol{\Sigma}_1| + \frac{n_2}{2} \log |\boldsymbol{\Sigma}_2|. \tag{9.2}$$

On constate que si l'on utilise un modèle de même complexité pour les trois probabilités P_0, P_1 et P_2, ces deux dernières seront nécessairement plus précises et le rapport de vraisemblance

logarithmique $R(t)$ est donc systématiquement négatif, ce qui implique la nécessité de déterminer un seuil de décision empirique entre les hypothèses H_0 et H_1.

Ajmera et al. proposent [10], pour s'affranchir de la nécessité d'un tel seuil, de modéliser la probabilité P_0 par un mélange de deux gaussiennes, tout en conservant une unique gaussienne pour les modèles P_1 et P_2. Ainsi les deux hypothèses sont de même complexité, ce qui permet de montrer que le seuil de décision se situe naturellement à 0 (pour le rapport logarithmique). Néanmoins, le Critère d'Information Bayésienne est une alternative plus généralement suivie dans la littérature.

9.3.3 Critère d'Information Bayésienne (BIC)

En 1972, Akaike [11] est le premier à proposer un critère d'information théorique, appelé AIC (*Akaike Information Criterion*) permettant de prendre en compte la complexité du modèle dans les tests d'hypothèses impliquant un rapport de vraisemblance. Il ajoute à la mesure de vraisemblance une pénalité k mesurant le nombre de paramètres libres du modèle, soit pour un modèle donné de loi P :

$$AIC(P) = \log P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - k.$$

Dans le cas d'un modèle gaussien, on dénombre le nombre suivant de paramètres libres pour la moyenne $\boldsymbol{\mu}$ et la matrice de covariance $\boldsymbol{\Sigma}$:

$$k = d + \frac{d(d+1)}{2}.$$

Par la suite, Schwarz propose [215] le Critère d'Information Bayésienne qui pénalise plus fortement les modèles construits sur une large collection d'exemples en ajoutant un facteur multiplicatif $\log n$ au facteur de pénalité ; on y joint généralement un facteur multiplicatif λ de manière à contrôler le compromis entre la vraisemblance et la complexité du modèle, bien que celui-ci soit absent de la proposition originale de Schwartz. Ainsi le critère devient :

$$BIC(P) = \log P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \lambda \frac{k}{2} \log n.$$

Schwarz justifie ce critère par le fait qu'il est asymptotiquement optimal pour le choix de modèle, tandis que le critère AIC à tendance à choisir le modèle le plus complexe lorsque $n \to \infty$. Rissanen montre par ailleurs [198] que, pour $\lambda = 1$, le critère BIC est égal à la MDL (*Minimum Description Length*), grandeur en Théorie de l'Information décrivant le nombre de bits minimum nécessaires pour coder le modèle, ce qui rejoint la notion de complexité du modèle.

Le critère fut plus tard introduit dans le contexte de la détection de rupture [49]. Le test d'hypothèse consiste à évaluer la différence des valeurs BIC entre l'hypothèse de rupture et de non-rupture, sur des modèles gaussiens ; soit :

$$\Delta BIC(t) = R(t) + \frac{1}{2}\lambda \left(d + \frac{d(d+1)}{2} \right) \log n. \tag{9.3}$$

où, $R(t)$ est le critère GLR introduit dans l'équation 9.2. Les auteurs revendiquent la supériorité du critère ΔBIC sur les approches métriques, du fait que celui-ci prend en compte la complexité des modèles et permet ainsi d'appliquer un seuil naturel de décision à 0. Cependant, malgré la valeur théorique $\lambda = 1$, le facteur λ constitue en pratique un paramètre supplémentaire à déterminer [227][183], qui se substitue au seuil de décision.

La taille de la fenêtre d'analyse est un point essentiel dans le comportement du critère BIC et doit fixer un compromis entre les deux contraintes suivantes :

- Une fenêtre trop large est susceptible de contenir plus d'un point de rupture, ce qui affecte directement le taux d'omissions (MD, *Missed Detections*).
- Une fenêtre trop étroite comporte peu d'exemples, ce qui affaiblit l'estimation des modèles.

Cette seconde contrainte constitue l'une des limites du critère BIC. En effet, ce dernier est choisi pour son comportement asymptotique optimal, mais le facteur de pénalité propre au critère ($k \log n$) favorise les modèles les plus simples [263] lorsque le modèle est estimé sur un nombre restreint d'exemples. De plus, le fait qu'il soit exclusivement basé sur des statistiques du second ordre (les matrices de covariances) accroît cette faiblesse, puisque l'estimation du modèle se trouve elle-même pénalisée [227][44]. On trouve ainsi plusieurs propositions d'approches hybrides appliquant une première étape de détection moins fiable mais plus simple, dont le seuil est fixé de manière à minimiser le taux d'omissions, basée par exemple sur le critère GLR [108], la statistique T^2 [264] [263], ou une approche métrique [61][50].

Un autre inconvénient majeur du critère BIC est sa complexité. En effet, le calcul des inverses des matrices de covariances est une opération lourde (de l'ordre de $O(\frac{n^3}{6})$ en utilisant la décomposition de Cholesky). On peut cependant réduire ce coût par des heuristiques, comme la mise à jour incrémentale des matrices de covariance [45][218]. Nous nous contenterons de suivre l'exemple d'Ajmera et al. [10] qui n'exploitent que des matrices de covariance diagonales.

Il est important de préciser que l'algorithme BIC s'accompagne à l'origine [49] d'une stratégie de recherche par maxima locaux par élargissement itératif des fenêtres d'analyse, qui permet d'affiner progressivement les modèles de distributions. Cependant, s'il est théoriquement plus pertinent, cet algorithme n'est pas adapté à un traitement en ligne (ou *online*, c'est-à-dire avec un retard de réponse borné) des données ; aussi nous préférons n'employer que la méthode des fenêtres glissantes, ce qui nous permet en outre de comparer les différents critères de distance sur les mêmes bases méthodologiques.

9.4 Mesures probabilistes dans les espaces RKHS

Les critères présentés dans la section précédente reposent sur des mesures probabilistes basées sur des modèles gaussiens. Nous avons présenté comme exemple la divergence de Kullback-Leibler symétrisée (KL2), mais il existe pléthore de mesures probabilistes alternatives, parmi lesquelles nous nous intéresserons également à la distance de Bhattacharrya [27], définie de la manière suivante :

$$d_B(p_1, p_2) = -\log\left(\int_{\boldsymbol{X}} \sqrt{p_1(\boldsymbol{x})p_2(\boldsymbol{x})}d\boldsymbol{x}\right).$$

Il est également possible d'exprimer analytiquement cette dernière dans le cas gaussien :

$$d_B(p_1, p_2) = \frac{1}{8}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^T \left[\frac{1}{2}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)\right]^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) + \frac{1}{2}\log\frac{\left|\frac{1}{2}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)\right|}{\sqrt{\left|\boldsymbol{\Sigma}_1\right|\left|\boldsymbol{\Sigma}_2\right|}}.$$

Le modèle gaussien est très largement exploité pour ses excellentes propriétés de régularités, sa concision et sa simplicité théorique qui permet généralement d'exprimer analytiquement les mesures probabilistes. Il reste néanmoins très restrictif et peut se révéler inadéquat en présence de données réelles.

Zhou et Chellappa tirent parti [265] des résultats de la théorie des Espaces à Noyaux Reproduisants (RKHS, pour *Reproducing Kernel Hilbert Spaces*), brièvement introduite dans la section 3.5.2, pour étendre le champ des modèles employés sur les fenêtres d'analyse. De la même manière que l'introduction des noyaux permettait de modéliser implicitement des surfaces de décisions plus complexes dans le processus discriminatif des SVM, il est possible d'exploiter le fameux *kernel trick* sur les mesures probabilistes classiques.

On rappelle que les noyaux respectant la condition de Mercer permettent de construire un espace fonctionnel de Hilbert par la transformation suivante (équation 3.23) :

$$\begin{array}{rcl} \Phi : \mathcal{X} & \rightarrow & \mathbb{R}^{\mathcal{X}} \\ \boldsymbol{x} & \mapsto & k(.\,, \boldsymbol{x}) \end{array}$$

On montre par ailleurs qu'il est possible de définir un produit scalaire sur cet espace qui reproduit le comportement de la fonction noyau :

$$\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle = \langle k(\cdot, \boldsymbol{x}), k(\cdot, \boldsymbol{y}) \rangle = k(\boldsymbol{x}, \boldsymbol{y}).$$

Cette propriété particulière justifie l'appellation d'Espace de Hilbert à Noyaux Reproduisants (RKHS), et montre que l'action d'un noyau de Mercer est équivalente au calcul d'un produit scalaire dans l'espace RKHS, la transformation de l'espace d'origine à ce dernier étant gouvernée par la fonction implicite Φ dont l'expression analytique n'est pas nécessaire. Ce dernier constat constitue ce qu'on appelle le *kernel trick*. On peut en outre montrer que l'espace RKHS est de dimension largement supérieure (voire infinie) à celle de l'espace d'origine, ce qui permet de renforcer la validité de l'hypothèse de gaussianité, comme le montrent les auteurs [265].

On montre que les moyennes et les covariances dans l'espace reproduisant, estimées à partir des exemples des fenêtres W_1 et W_2, prennent les expressions suivantes :

$$\begin{aligned} \hat{\boldsymbol{\mu}}_i &= \boldsymbol{\Phi}_i \, \mathbf{s} \\ \hat{\boldsymbol{\Sigma}}_i &= \boldsymbol{\Phi}_i \, \mathbf{J}\mathbf{J}^T \boldsymbol{\Phi}_i^T, \end{aligned}$$

où l'on a introduit, pour obtenir une expression matricielle, le vecteur des exemples dans l'espace transformé $\boldsymbol{\Phi}_i^T = \left[\Phi(\boldsymbol{x}_{i,1}), \ldots, \Phi(\boldsymbol{x}_{i,n_i})\right]^T$, le vecteur moyennant $\mathbf{s}$ et la matrice de centrage $\mathbf{J}$, définis par :

$$\mathbf{s} = \frac{1}{n_i}\mathbf{1} \qquad\qquad \mathbf{J} = \frac{1}{\sqrt{n_i}} \left(\mathbf{I}_{n_i} - \mathbf{s}\mathbf{1}^T\right).$$

Malheureusement la matrice $\hat{\boldsymbol{\Sigma}}_i$ n'est pas de rang plein puisqu'elle peut être exprimée comme le produit d'un matrice non carrée avec sa transposée ($\boldsymbol{A}\boldsymbol{A}^T$). Or les mesures probabilistes impliquent généralement (c'est le cas des deux mesures considérées ici) l'inverse des matrices de covariance. Zhou et Chellappa proposent ainsi d'approximer $\hat{\boldsymbol{\Sigma}}_i$ par la matrice suivante :

$$\boldsymbol{C}_i = \boldsymbol{\Phi}_i \, \mathbf{J}\mathbf{Q}\mathbf{Q}^T \mathbf{J}^T \boldsymbol{\Phi}_i^T + \rho \, \mathbf{I},$$

où $\mathbf{Q}$ est une matrice de dimension $r \times n_i$. La matrice $\boldsymbol{C}_i$ ainsi est régularisée et inversible. On trouvera dans l'article de Zhou et Cheppalla [265] le développement qui mène à l'expression suivante de l'inverse :

$$\boldsymbol{C}_i^{-1} = \rho^{-1} \left(\mathbf{I}_{n_i} - \mathbf{Q}\mathbf{B}\mathbf{Q}^T\right),$$

avec

$$\mathbf{B} = \rho \mathbf{I}_r + \mathbf{Q}^T \mathbf{J}^T \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i \, \mathbf{J}\mathbf{Q}.$$

La matrice $\mathbf{Q}$ est choisie de manière à ce que $\boldsymbol{C}_i^{-1}$ soit une bonne estimation de $\hat{\boldsymbol{\Sigma}}_i$, c'est-à-dire en conserve les r vecteurs propres principaux. On remarque que la plupart des matrices considérées dans ce développement ($\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\Sigma}}_i$, $\boldsymbol{\Phi}_i$, ...) ne sont pas exprimables en pratique puisque leurs valeurs sont définies dans l'espace transformé, qui peut être de dimension infinie. Mais la forme de l'inverse $\boldsymbol{C}_i^{-1}$ ne dépend que du produit $\boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i = [k(\boldsymbol{x}_{i,k}, \boldsymbol{x}_{i,l})]_{kl} = \boldsymbol{K}$ qui n'est autre que la matrice de Gram définie sur les exemples de la fenêtre i. De même on peut montrer que l'expression des mesures probabilistes considérées s'exprime exclusivement en terme de produits scalaires dans l'espace transformé, et constitue ainsi une démonstration supplémentaire du fameux *kernel trick*.

9.5 SVM à une classe

L'approche précédente tire parti du *kernel trick* en étendant la pertinence du modèle gaussien par son application dans un espace de dimension supérieure. Nous avons vu que les machines à noyaux reposent sur la conjonction du *kernel trick* et du principe de maximisation de la marge, qui implique à la fois la minimisation du Risque Structurel et une sélection parcimonieuse des exemples essentiels pour la fonction de décision. Ce second principe n'est pas appliqué dans l'approche précédente.

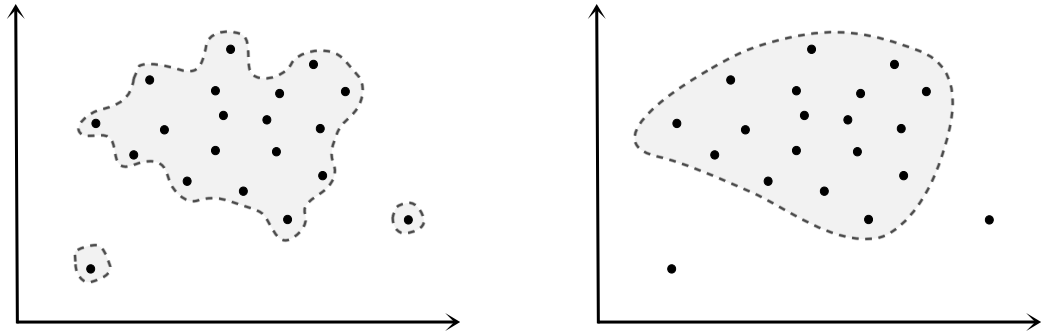
Nous présentons dans cette section les SVM à une classe, qui adaptent le formalisme des SVM pour la caractérisation du support d'une distribution donnée. Après en avoir présenté le principe, nous verrons comment ce dernier peut être exploité pour la détection de rupture.

9.5.1 Principe des SVM à une classe

Le problème posé par Schölkopf et al. dans [209] et [212] consiste à estimer à partir de réalisations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ le support d'une distribution de probabilité P donnée, c'est-à-dire à déterminer un sous-ensemble S de l'espace d'origine tel qu'on ait idéalement :

$$\begin{aligned} P(\boldsymbol{x}) &> 0 \qquad \forall\, \boldsymbol{x} \in S \\ P(\boldsymbol{x}) &= 0 \qquad \forall\, \boldsymbol{x} \notin S. \end{aligned}$$

Le problème se heurte aux mêmes écueils que le problème de classification. En effet, il est possible d'apprendre « par cœur » la distribution des exemples d'apprentissage (voir figure 9.3(a)) mais on se trouve alors en situation de sur-apprentissage et l'ensemble déterminé ne pourra se généraliser correctement sur des données inconnues. Il est donc nécessaire de lisser la frontière de l'ensemble S en régularisant le problème ; la figure 9.3(b) représente un exemple de solution mieux régularisée. Le principe de minimisation du risque structurel nous permet à nouveau de faire face à cette contrainte.



(a) (b)

FIGURE 9.3 – Estimation du support d'une distribution sur un cas simple à deux dimensions, présentant deux exemples marginaux (*outliers*). Comparaison entre (a) un cas de sur-apprentissage, et (b) un cas correctement régularisé.

On reformule le problème en l'assimilant à une classification *one-vs-all* sur une classe unique, déterminée par les exemples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. On leur associe par convention le label $y_i = +1$ qui correspond au résultat idéal de la fonction de décision f, les exemples hors du support de la distribution sont idéalement associés au résultat $f(\boldsymbol{x}) = -1$. On cherche donc à apprendre une fonction f telle que :

$$\begin{aligned} f(\boldsymbol{x}) &\geq 0 \qquad \forall\, \boldsymbol{x} \in S \\ f(\boldsymbol{x}) &< 0 \qquad \forall\, \boldsymbol{x} \notin S. \end{aligned}$$

La fonction noyau joue ici un rôle essentiel en transposant la distribution dans l'espace transformé. On peut se baser sur la haute dimension de cet espace pour supposer que les exemples sont localisés dans une moitié de l'espace dont l'origine est exclue (cette deuxième supposition est toujours vraie dans le cas du noyau RBF gaussien). Il en résulte que la tâche équivaut à l'apprentissage d'un hyperplan de séparation séparant de manière optimale les exemples et l'origine. On rejoint ainsi le cadre des SVM en exploitant la maximisation de la marge comme critère d'optimalité. La figure 9.4 illustre le problème de séparation dans l'espace transformé.

On transpose aisément le problème de minimisation du modèle SVM (se référer au chapitre 3) dans ce contexte :

$$\begin{aligned} \text{minimiser} \quad & \frac{1}{2}\left\|\boldsymbol{w}\right\|^2 + \frac{1}{\nu n}\sum_i \xi_i - \rho \\ \text{sous les contraintes} \quad & \boldsymbol{w}^T \Phi(\boldsymbol{x}_i) \geq \rho - \xi_i \qquad i = 1, \ldots, n \\ & \xi_i \geq 0, \end{aligned} \tag{9.4}$$

où $\boldsymbol{w}$ est le vecteur normal de l'hyperplan de séparation, ρ l'équivalent de la constante b, et ξ_i sont les variables d'écart pénalisant les erreurs de classification. Le paramètre $\nu \in]0, 1]$ s'inspire

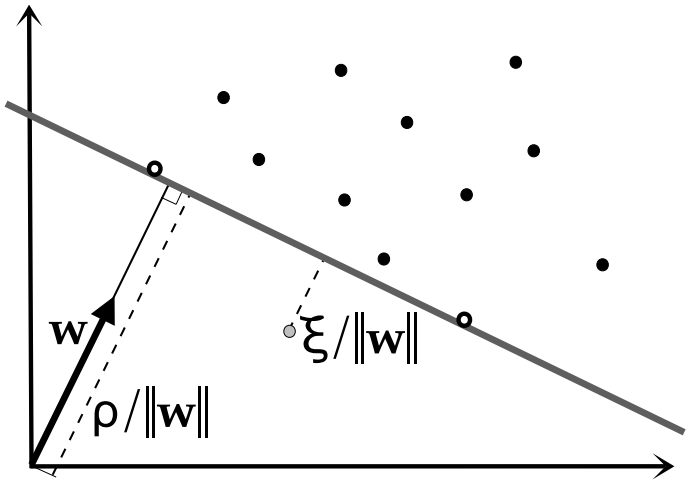


FIGURE 9.4 – Séparation des exemples avec l'origine par l'hyperplan (en gris foncé) défini par le vecteur normal $\boldsymbol{w}$, sur une projection schématique en 2D de l'espace transformé. La figure montre trois vecteurs de support, dont deux à la marge (en blanc) et un autre mal classifié (en gris), dont la distance avec l'hyperplan définit la pénalité ξ.
Cette figure s'inspire d'une figure de l'ouvrage de Schölkopf et Smola [212].

des ν-SVM [211] et permet de contrôler le compromis entre le risque empirique et la complexité du classifieur. On peut par ailleurs montrer [209][212] que ν est à la fois une borne supérieure pour le taux d'erreurs marginales et une borne inférieure pour le taux de Vecteurs de Support dans l'ensemble d'apprentissage, ces deux valeurs convergeant asymptotiquement vers ν lorsque $n \to \infty$.

Nous ne détaillons pas la résolution du problème d'optimisation 9.4. On obtient, de manière similaire aux SVM, un vecteur $\boldsymbol{w}$, moyenne des exemples pondérés par les facteurs de Lagrange α_i (introduits dans la formulation du problème dual). On rappelle que les exemples de facteur non-nul constituent les vecteurs de support du classifieur :

$$\boldsymbol{w} = \sum_i \alpha_i \Phi(\boldsymbol{x}_i).$$

La fonction de décision prend donc l'expression suivante :

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T \Phi(\boldsymbol{x}) - \rho) = \text{sign}\left(\sum_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) - \rho\right).$$

Les SVM à une classe ainsi formulés (que l'on pourra désigner par *SVM1C* dans la suite de ce document) sont par la suite exploités par leurs inventeurs [214] pour la détection de nouveauté, en considérant simplement tout vecteur $\boldsymbol{x}$ comme « nouveau » (c'est-à-dire n'appartenant pas à la classe modélisée) si $f(\boldsymbol{x}) < 0$.

9.5.2 Rapport de vraisemblance par SVM1C (LLR)

Loosli et al. [137] exploitent les SVM à une classe sur la base du Rapport de Vraisemblance Généralisé (GLR), présenté précédemment dans la section 9.3.2. Ils adaptent ce dernier en excluant du test d'hypothèses la fenêtre globale W et son modèle $P_0(\boldsymbol{X})$, et supposent dans tous les cas que les échantillons de la fenêtres W_1 sont décrits par le modèle $P_1(\boldsymbol{X})$. Le test d'hypothèses se résume ainsi à évaluer l'égalité entre les distributions P_1 et P_2, ce qui revient à appliquer une approche métrique.

Le rapport de vraisemblance de l'équation 9.2 devient donc :

$$r(t) = \frac{\prod_{i=1}^{n} P_1(\boldsymbol{x}_i)}{\prod_{i=1}^{t-1} P_1(\boldsymbol{x}_i) \prod_{i=t}^{n} P_2(\boldsymbol{x}_i)} = \prod_{i=t}^{n} \frac{P_1(\boldsymbol{x}_i)}{P_2(\boldsymbol{x}_i)}.$$

Le dénominateur mesure la vraisemblance des exemples de la fenêtre W_2 sur la distribution calculée sur ces mêmes exemples. On peut donc considérer celui-ci comme constant, ou du moins de variations négligeables par rapport au numérateur, et simplifier ainsi le critère :

$$r(t) = \prod_{i=t}^{n} P_1(\boldsymbol{x}_i). \tag{9.5}$$

On remarque au passage que ce critère ne nécessite que l'apprentissage d'un unique modèle de probabilité, au lieu des trois modèles impliqués dans les critères GLR et BIC. On estime la distribution des exemples de la fenêtre W_1 au moyen d'un SVM à une classe. Les auteurs couplent la fonction de décision au modèle de famille exponentielle afin d'obtenir une estimation de la probabilité P_1 :

$$\hat{P}_1(\boldsymbol{x}) = \exp\left(\sum_{i=1}^{t-1} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) - g(\theta_0)\right),$$

où $g(\theta_0)$ est la fonction de log-partition, mais ne joue aucun rôle ici. En effet, le test de décision se limite à comparer le logarithme du critère $r(t)$ (équation 9.5) à un seuil s, qui inclut de fait la constante $g(\theta_0)$:

$$\sum_{j=t}^{n} \left(\sum_{i=1}^{t-1} \alpha_i k(\boldsymbol{x}_j, \boldsymbol{x}_i)\right) \gtrless s.$$

On a ainsi défini une mesure de vraisemblance des exemples de la fenêtre W_2 par rapport à la distribution du modèle P_1, qui s'exprime simplement à partir de la matrice de Gram.

9.5.3 Kernel Change Detection (KCD)

Désobry et al. proposent un autre algorithme de détection de rupture basé sur les SVM à une classe [63][74], qu'ils nomment *Kernel Change Detection* (KCD). Ils partent pour cela de l'hypothèse que le noyau est normalisé, c'est-à-dire respecte la condition $k(\boldsymbol{x}, \boldsymbol{x}) = 1 \ \forall \boldsymbol{x}$, sans perte de généralité puisque l'on peut normaliser n'importe quel noyau par la relation suivante :

$$k'(\boldsymbol{x}, \boldsymbol{y}) = \frac{k(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{k(\boldsymbol{x}, \boldsymbol{x})\, k(\boldsymbol{y}, \boldsymbol{y})}}.$$

On peut montrer que le noyau k' respecte également la condition de Mercer. La normalisation a pour effet de restreindre la position des exemples sur la sphère unité dans l'espace transformé ($\|\Phi(\boldsymbol{x}_i)\|^2 = k(\boldsymbol{x}, \boldsymbol{x}) = 1$). L'apprentissage d'un SVM à une classe détermine donc la position d'un hyperplan séparant une section de la sphère et son centre, comme l'illustre la figure 9.5 sur une projection plane simplifiée.

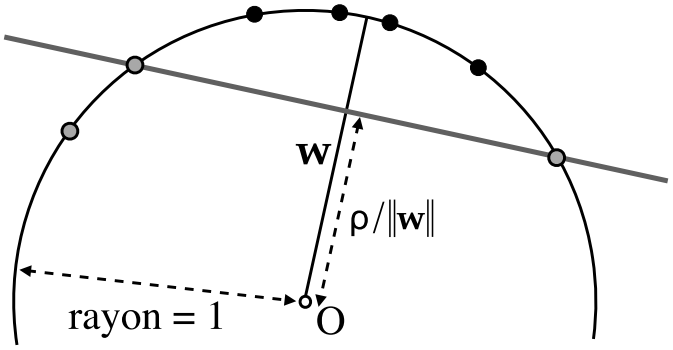


FIGURE 9.5 – La figure montre un exemple d'application de SVM à une classe. La contrainte de normalisation du noyau implique que, dans l'espace transformé, les exemples sont situés sur l'hypersphère de rayon 1. On peut y voir 3 vecteurs de support (en gris), dont deux sont à la marge, et le troisième mal classifié. *Cette figure s'inspire d'une figure de l'article de Desobry et al.* [63].

En suivant le paradigme de l'approche métrique, on apprend deux SVM à une classe modélisant chacun les exemples de l'une des fenêtres W_1 et W_2, en déterminant les hyperplans de séparation de vecteurs normaux respectifs $\boldsymbol{w}_1$ et $\boldsymbol{w}_2$. Il existe nécessairement un plan dans l'espace engendré par les deux vecteurs normaux ; l'intersection de la sphère unité avec ce dernier est un cercle $\mathcal{S}$ de rayon 1 et dont le centre est à l'origine $\boldsymbol{O}$, comme le montre la figure 9.6.

Les auteurs proposent de mesurer la dissimilarité entre les deux modèles sur la base de la distance d'arc entre les points $\boldsymbol{c}_1$ et $\boldsymbol{c}_2$, définis comme les intersections respectives des radiales sur les axes des vecteurs normaux $\boldsymbol{w}_1$ et $\boldsymbol{w}_2$, avec le cercle $\mathcal{S}$ (voir figure 9.6). Néanmoins, une telle mesure n'a de sens que si l'on prend en compte l'étalement des exemples d'une classe autour du « centre » $\boldsymbol{c}_i$ sur le cercle $\mathcal{S}$. Ainsi les auteurs introduisent un dénominateur normalisant la distance précédente par la distance entre les « centres » $\boldsymbol{c}_i$ et les points d'intersection $\boldsymbol{p}_i$ entre les hyperplans $\boldsymbol{H}_i$ et le cercle $\mathcal{S}$, s'inspirant ainsi, de leur propre aveu, du rapport de Fisher entre une statistique du premier ordre (distance entre les moyennes) et du second ordre (déterminant des matrices de covariance). Le critère de dissimilarité ainsi défini a donc l'expression suivante :

$$d_{KCD} = \frac{\widehat{\boldsymbol{c}_1\boldsymbol{c}_2}}{\widehat{\boldsymbol{c}_1\boldsymbol{p}_1} + \widehat{\boldsymbol{c}_2\boldsymbol{p}_2}},$$

où $\widehat{\boldsymbol{x}\boldsymbol{y}}$ représente la distance d'arc entre les points $\boldsymbol{x}$ et $\boldsymbol{y}$ situés sur le cercle $\mathcal{S}$. Ce dernier étant de rayon unitaire, la distance est égale à l'angle $\widehat{\boldsymbol{x}\boldsymbol{O}\boldsymbol{y}}$ exprimé en radian, défini comme l'arccosinus du produit scalaire entre les deux points. Soit :

$$\widehat{\boldsymbol{x}\boldsymbol{y}} = \widehat{\boldsymbol{x}\boldsymbol{O}\boldsymbol{y}} = \arccos k(\boldsymbol{x}, \boldsymbol{y}).$$

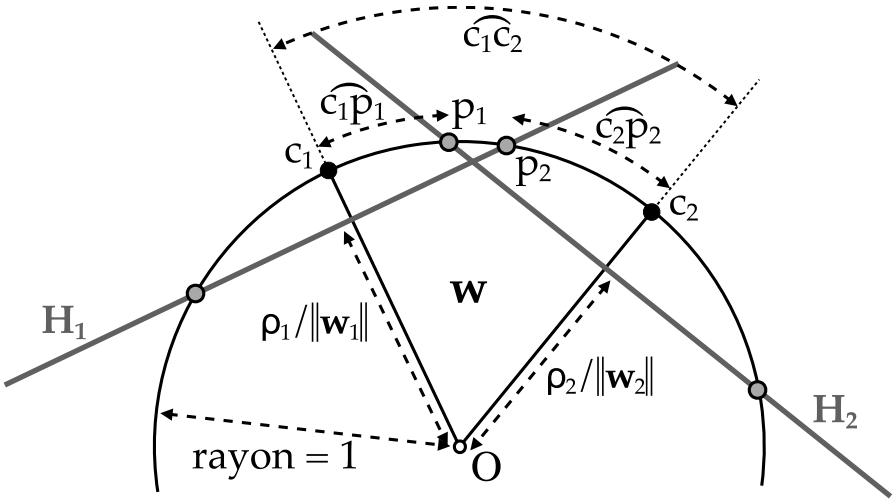


FIGURE 9.6 – Configuration dans l'espace transformé des hyperplans de séparation $\boldsymbol{H}_1$ et $\boldsymbol{H}_2$ pour les fenêtres d'analyse W_1 et W_2. Les distances d'arc $\widehat{\boldsymbol{c}_1\boldsymbol{c}_2}$, $\widehat{\boldsymbol{c}_1\boldsymbol{p}_1}$ et $\widehat{\boldsymbol{c}_2\boldsymbol{p}_2}$, définies à partir des points d'intersection des hyperplans avec l'hypersphère de rayon 1, permettent le calcul de la métrique KCD. *Cette figure s'inspire d'une figure de l'article de Desobry et al.* [63].

De l'expression du point $\boldsymbol{c}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$, on déduit :

$$\widehat{\boldsymbol{c}_1\boldsymbol{c}_2} = \arccos\left(\frac{k(\boldsymbol{w}_1, \boldsymbol{w}_2)}{\sqrt{k(\boldsymbol{w}_1, \boldsymbol{w}_1)\, k(\boldsymbol{w}_2, \boldsymbol{w}_2)}}\right).$$

L'angle $\widehat{\boldsymbol{c}_i\boldsymbol{O}\boldsymbol{p}_i}$ ayant pour cosinus la valeur $\frac{\rho_i}{\|\boldsymbol{w}_i\|}$, on déduit de même :

$$\widehat{\boldsymbol{c}_i\boldsymbol{p}_i} = \arccos\left(\frac{\rho_i}{\sqrt{k(\boldsymbol{w}_i, \boldsymbol{w}_i)}}\right).$$

Toutes les composantes du critère d_{KCD} sont ainsi exprimées en termes de produits scalaires via la fonction noyau, ce qui rend son calcul possible en employant les valeurs de la matrice de Gram.

9.5.4 Mise à jour incrémentale des SVM à une classe

Les deux algorithmes précédents emploient un ou deux SVM à une classe appris itérativement sur les exemples d'une fenêtre glissante. On peut remarquer que, dans le cas d'un pas réduit à un échantillon (ce qui est notre cas), une fenêtre W_i conserve $n_i - 2$ exemples en commun entre les instants t et $t+1$, en supprimant l'exemple $\boldsymbol{x}_t$ et en rajoutant l'exemple $\boldsymbol{x}_{t+n_i}$. Ceci implique que la structure du SVM varie peu puisqu'elle ne diffère au pire que de deux vecteurs de support. Il est possible de tirer parti de ce constat en n'effectuant pas à chaque itération l'apprentissage total du SVM1C.

Le problème d'optimisation des SVM consiste en la minimisation d'un critère sous contrainte, par le biais des multiplicateurs de Lagrange α_i qui, en définitive, déterminent la solution du problème, ainsi que le sous-ensemble des vecteurs de support. Il est donc possible pour l'apprentissage à l'instant $t+1$ de conserver les $\alpha_{i,t}$ relatifs au SVM1C de l'instant t, à l'exception de celui correspondant au vecteur supprimé, et d'initialiser le coefficient du nouveau vecteur à 0. Il en résulte un gain important en nombre d'itérations (et donc en temps de calcul) dans la procédure d'optimisation. On trouvera plus de détail sur cette question dans les articles annexes des auteurs du KCD [94][58].

9.6 Recherche de maxima pour la détection de rupture

Nous avons présenté plusieurs approches de détection de rupture applicables sur une recherche par fenêtres adjacentes glissantes. On suppose ici les fenêtres W_1 et W_2 de même longueur n_w. En exploitant l'une des métriques présentées, on obtient donc à partir de la séquence de N exemples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, une séquence de mesures de distance $[d(1), \ldots, d(n_d)]$ entre fenêtres antérieures et postérieures (avec $n_d = N - 2n_w$), la mesure $d(i)$ d'indice i correspondant à l'instant $i + n_w$, en raison du retard impliqué par la fenêtre antérieure.

La détection de points de rupture se fait généralement par la recherche de maxima locaux dépassant un seuil donné. Cependant, dans le contexte de programmes radiophoniques par exemple, les conditions acoustiques peuvent largement évoluer au sein d'un même fichier (généralement d'une étendue d'une heure) ; aussi il est profitable de prendre en compte les conditions d'enregistrement locales. De plus, la présence de pics secondaires au voisinage des maxima locaux peut parasiter la recherche. On reprend donc pour cela l'algorithme de filtrage non-linéaire proposé par Gillet [88], qui adapte des techniques usuelles en traitement d'image. Celui-ci consiste en 3 étapes successives (illustrées par la figure 9.7, page 142) :

1. **Filtrage médian** : On soustrait dans un premier temps au signal le résultat d'un filtrage médian à large échelle, calculé sur une fenêtre glissante centrée sur l'échantillon concerné, de manière à annuler d'éventuels *offsets* constants locaux sur la métrique. On choisit dans notre cas une fenêtre d'une minute, ce qui correspond à $n_{\text{filt}} = 120$ trames environ, soit :

$$d_{\text{med}}(i) = d(i) - \text{med}\left(d(j_{i,1}), \ldots, d(j_{i,2})\right),$$

avec

$$\begin{aligned} j_{i,1} &= \max(1, i - n_{\text{filt}}/2) \\ j_{i,2} &= \min(n_d, i + n_{\text{filt}}/2 - 1), \end{aligned}$$

où med désigne le filtrage non-linéaire médian. Les variables $j_{i,1}$ et $j_{i,2}$ servent uniquement à s'assurer que les fenêtres de filtrage médian sont correctement définies.

2. **Variance homogène** : Le signal $d_{med}(i)$ est ensuite divisé par la déviation standard locale calculée sur la même fenêtre glissante, afin d'équilibrer la balance des dynamiques sur l'ensemble du signal. On a donc :

$$d_{\text{var}}(i) = d_{\text{med}}(i) - \text{std}\left(d_{\text{med}}(j_{i,1}), \ldots, d_{\text{med}}(j_{i,2})\right),$$

où std désigne le filtrage non-linéaire de calcul de déviation standard.

3. **Détection des pics :** la détection de pics maxima est soumise à deux contraintes : un pic local doit être au delà d'un seuil donné τ et éloigné des pics voisins d'une durée minimale donnée, définie par un nombre de trames n_{max}. On choisit $n_{max} = 10$ dans notre cas. On répond à ces deux contraintes par l'intermédiaire du signal « plateau » d_{plat}, défini de la manière suivante à partir du signal d_{var} :

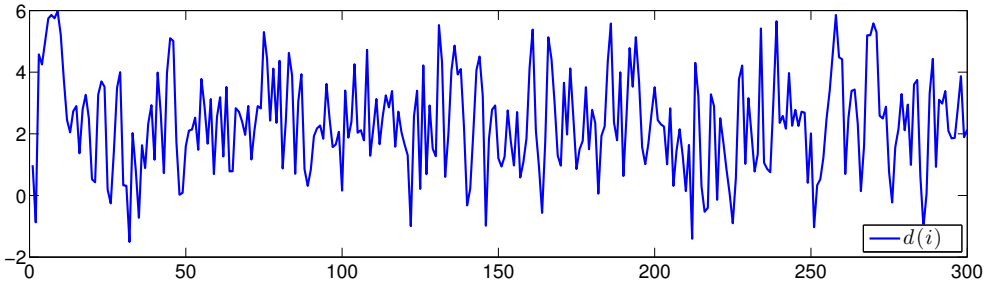
$$d_{\text{plat}}(i) = \max\left(d_{\text{var}}(k_{i,1}), \ldots, d_{\text{var}}(k_{i,2}), \tau\right),$$

où

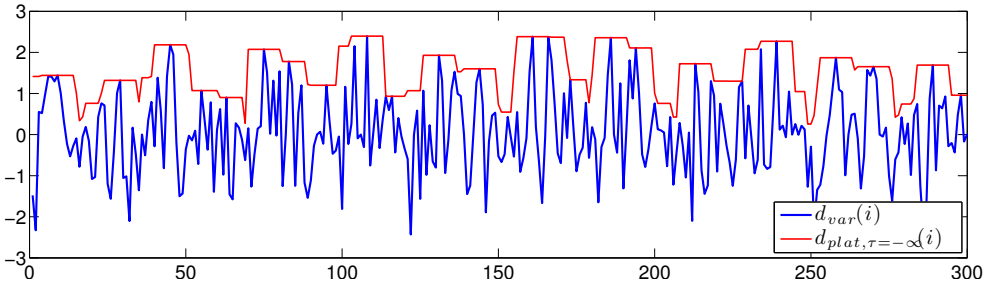
$$\begin{aligned} k_{i,1} &= \max(1, i - n_{\max}/2) \\ k_{i,2} &= \min(n_d, i + n_{\max}/2 - 1). \end{aligned}$$

Les variables $k_{i,1/2}$ jouent le même rôle que les $j_{i,1/2}$ introduits précédemment, pour la longueur de fenêtre $n_{\max}$. On détecte un maximum local quand $d_{\text{var}}(i) = d_{\text{plat}}(i)$.

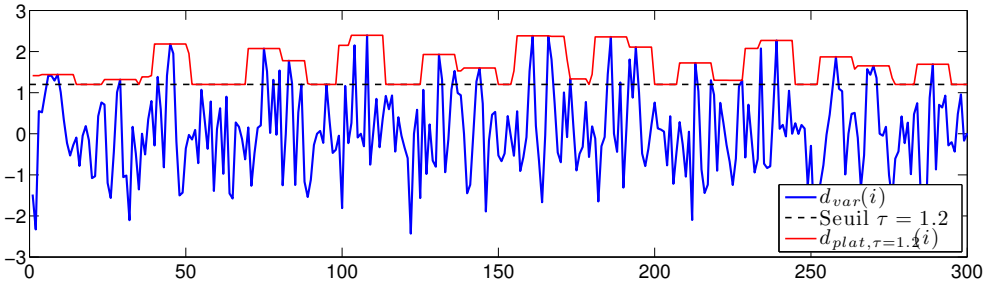
Nous avons introduit dans ce chapitre plusieurs méthodes de détection de rupture, dont la plupart sont étroitement liées à la théorie des Noyaux ou des Machines à Vecteurs de Support. Le chapitre suivant, qui traite de l'évaluation de nos propositions sur des corpus audio publics, comprendra une étude, en section 10.4, dans laquelle nous comparerons les différentes métriques proposées. En particulier, la détermination du seuil τ reste un point essentiel de la détection, qui détermine le compromis entre le nombre de frontières fausses et manquées. Ce point sera abordé de manière pratique en suivant la procédure classique d'estimation empirique de la valeur optimale sur un ensemble de validation.



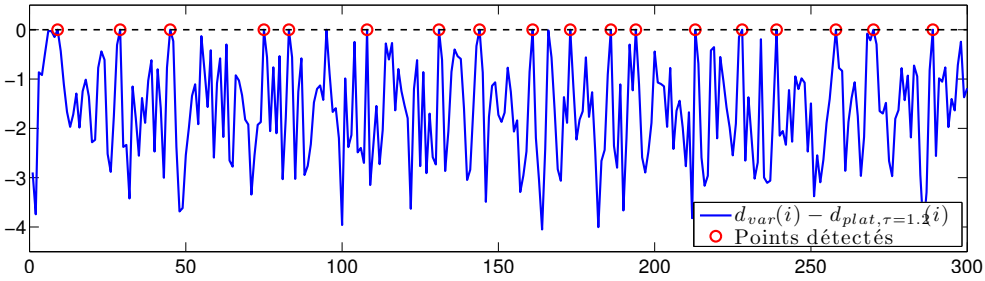
(a) Mesure de nouveauté $d(i)$ brute.



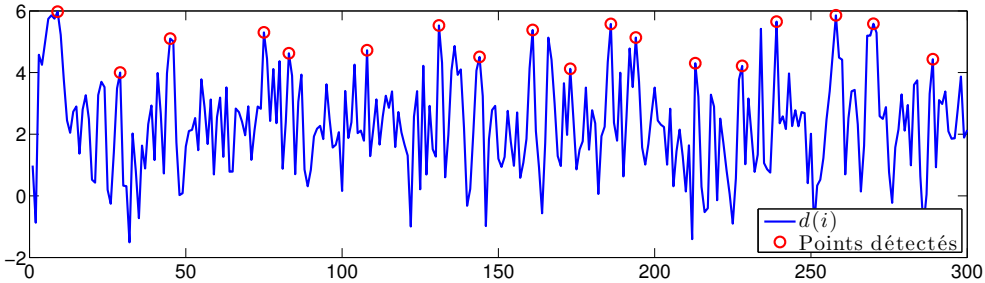
(b) Mesure de nouveauté $d_{var}(i)$ après normalisation, signal plateau (en rouge) $d_{plat,\tau=-\infty}(i)$ sans seuil pour illustrer le principe de filtrage des pics secondaires.



(c) L'application d'un seuil (ici $\tau = 1$) permet de réduire le nombre de maxima détectés.



(d) On observe que les maxima locaux sont détectés aux points où $d_{var}(i) - d_{plat,\tau}(i) = 0$.



(e) Représentation des maxima détectés sur la mesure originale $d_(i)$.

FIGURE 9.7 – Représentation des étapes successives pour la recherche de maxima locaux.