Diagnostic et surveillance des processus complexes par réseaux bayésiens

Sylvain Verron

UNIVERSITÉ D'ANGERS

Année 2007
N° 857

DIAGNOSTIC ET SURVEILLANCE DES PROCESSUS COMPLEXES PAR RÉSEAUX BAYÉSIENS

THÈSE DE DOCTORAT

Spécialité : Sciences de l'ingénieur

ÉCOLE DOCTORALE D'ANGERS

Présentée et soutenue publiquement
Le 13 décembre 2007
À l'Institut des Sciences et Techniques de l'Ingénieur d'Angers

Par Sylvain VERRON

Devant le jury ci-dessous :

Patrice AKNIN	Rapporteur	Directeur de Recherche à l'Institut National de Recherche sur les Transports et leur Sécurité
Jean-Marc THIRIET	Rapporteur	Professeur à l'Université Joseph Fourier de Grenoble
Paul MUNTEANU	Examinateur	Président du Directoire de la société BAYESIA de Laval
Daniel NOYES	Examinateur	Professeur à l'École Nationale d'Ingénieurs de Tarbes
Philippe WEBER	Examinateur	Maître de conférences à l'École Supérieure des Sciences et Technologies de l'Ingénieur de Nancy
Abdessamad KOBI	Examinateur	Professeur à l'Institut des Sciences et Technologies de l'Ingénieur d'Angers
Teodor TIPLICA	Examinateur	Maître de conférences à l'Institut des Sciences et Technologies de l'Ingénieur d'Angers

Directeur de thèse : **Abdessamad KOBI**
Co-encadrant : **Teodor TIPLICA**

Laboratoire : **Laboratoire en Sûreté de fonctionnement, Qualité et Organisation**
62, avenue Notre Dame du Lac
49000 ANGERS

ED 363

Remerciements

J'exprime mes profonds remerciements à mon directeur de thèse, le professeur Abdessamad Kobi, qui m'as permis d'effectuer cette thèse. Son soutien, sa confiance et sa disponibilité m'ont permis de m'épanouir sereinement tout au long de mes travaux.

J'adresse également mes vifs remerciements à Teodor Tiplica pour son aide, sa disponibilité, ses judicieux conseils pendant toute la durée de ma thèse. Sa patience et sa pédagogie sont exemplaires, son oeil critique m'a été très précieux pour guider et structurer ces trois années de travail.

Je remercie Monsieur Patrice Aknin, Directeur de Recherche à l'Institut National de Recherche sur les Transports et leur Sécurité, ainsi que Monsieur Jean-Marc Thiriet, professeur à l'Université Joseph Fourier de Grenoble, d'avoir accepté de rapporter mon mémoire et pour l'intérêt qu'ils ont bien voulu porter à ce travail.

Mes remerciements s'adressent également à Monsieur Paul Munteanu, Président du Directoire de la société BAYESIA, à Monsieur Daniel Noyes, Professeur à l'École Nationale d'Ingénieurs de Tarbes, ainsi qu'à Philippe WEBER, Maître de conférences à l'École Supérieure des Sciences et Technologies de l'Ingénieur de Nancy, pour avoir accepté de prendre part au jury.

Je remercie la communauté d'agglomération angevine "Angers Loire Métropole", qui a financé mes recherches durant ces trois années.

Mes remerciements vont à tout le personnel du laboratoire LASQUO qui m'a accueilli durant ces trois années. L'ambiance chaleureuse est propice à un travail efficace.

Je tiens tout particulièrement à remercier tous les doctorants du laboratoire (Razvan, Alin, Sorin, Florina, Amel, Radouane, et les autres ...), pour l'ambiance studieuse mais sympathique qu'ils ont réussi à instaurer.

Je remercie infiniment Lucie et Mahé pour la compréhension des impératifs qu'entraîne un tel travail, et pour leurs encouragements réguliers. Finalement, je tiens à remercier ma famille et mes amis pour leur soutien moral.

"Le doute est le commencement de la sagesse."
(Aristote)

"Trop de Nord, tue le Nord."

Je dédie cette thèse
à Lucie et Mahé.

Table des matières

Table des figures

Introduction générale

De nos jours, les procédés sont de plus en plus complexes, automatisés, ou tout du moins informatisés. De plus, les enjeux économiques induits par la production de biens sont de plus en plus importants. Durant toute la durée de vie d'un procédé, on recherche à ce que celui-ci fonctionne de manière sécurisée vis à vis de son environnement humain et matériel, ainsi que vis à vis de sa propre intégrité. Bien entendu, on recherche également à utiliser le procédé de manière optimale, afin de réduire différents coûts ou temps d'exécution et atteindre une viabilité économique du procédé.

La conception et la mise en service d'un procédé influencent fortement les objectifs cités plus haut. Cependant, bien que nécessaires, ces activités ne sont pas suffisantes pour assurer le bon fonctionnement d'un procédé. En effet, une fois le procédé en service, il faut s'assurer à chaque instant de son bon fonctionnement, et si ce n'est pas le cas, remédier au problème afin de retrouver les conditions normales d'utilisation du procédé. Ceci est le but de la surveillance des procédés. On distingue généralement deux phases : la détection et le diagnostic. La première, la détection, permet de statuer sur la présence ou non d'une faute dans le procédé, entraînant un dysfonctionnement de celui-ci. La seconde phase, le diagnostic, permet de conclure sur la nature de la faute présente dans le procédé. Une méthode couramment employée pour le diagnostic est la classification supervisée. Celle-ci, en se basant sur un historique du procédé, permet de classer une faute détectée dans une des classes de fautes prédéfinies. L'objectif de cette thèse est de développer une méthode de surveillance complète, basée sur un outil de classification particulier : le réseau bayésien.

Le premier chapitre permet d'exposer le concept de la variabilité des procédés. Pour cela, nous allons présenter les causes de la variabilité des procédés, et ce que celles-ci impliquent sur la qualité de la production, amenant au besoin de surveillance. Nous exposerons alors les différents points clés de la surveillance des procédés, en étudiant les diverses approches pour réaliser celle-ci. Nous nous focaliserons sur les méthodes permettant le suivi du procédé, méthodes basées sur l'exploitation des données issues du procédé. Ainsi, nous présenterons de manière non-exhaustive des méthodes supervisées et non-supervisées de détection et de diagnostic. Nous terminerons ce premier chapitre par

le choix d'un classifieur adapté à la surveillance : le réseau bayésien.

Le second chapitre est l'objet d'une présentation plus approfondie des réseaux bayésiens. Ainsi, dans un premier temps, nous allons nous intéresser à présenter le principe des réseaux bayésiens grâce à un exemple très simple. Nous étudierons alors les différentes relations pouvant exister entre les nœuds de ce type de réseaux, puis nous nous intéresserons aux extensions possibles et intéressantes de ce genre d'outil dans le contexte de la surveillance des procédés. Dans un second temps, nous établirons un état de l'art des méthodes de surveillance ou de diagnostic basées sur les réseaux bayésiens. Nous verrons alors les différents points pouvant être améliorés.

Le troisième chapitre est consacré aux contributions apportées au domaine de la surveillance des procédés par réseaux bayésiens. La première partie porte sur la réalisation de la détection dans un réseau bayésien. Pour cela, nous démontrerons mathématiquement l'équivalence entre une analyse discriminante (modélisée par réseaux bayésiens) et une carte de contrôle multivariée. Dans un second temps, nous exposerons comment effectuer une étape de diagnostic supervisé par réseaux bayésiens. Pour cela, nous proposerons tout d'abord un algorithme de sélection de variables basé sur un nouveau résultat concernant l'information mutuelle, puis nous expliquerons comment discerner une faute non présente dans l'historique des données. Ensuite, nous proposerons une méthode de diagnostic non-supervisée en proposant une amélioration d'une méthode déjà existante. Enfin, dans une dernière partie, nous présenterons la structure complète d'un réseau bayésien dédié à la surveillance des procédés. Ce réseau va permettre la détection de faute, ainsi que le diagnostic supervisé et non-supervisée.

Le quatrième et dernier chapitre est une application de la méthode proposée sur un exemple classique : le procédé Tennessee Eastman. Ce procédé comporte 53 variables, et peut être soumis à 20 types de fautes. Nous étudierons les performances en détection et diagnostic (supervisé et non-supervisé) du réseau bayésien face à ce procédé complexe.

Chapitre 1

Surveillance des procédés

Sommaire

1.1 Introduction

Ce premier chapitre permet d'introduire différents concepts liés à la maîtrise (ou surveillance) des procédés. La maîtrise des procédés a pour principal objectif de garantir le fonctionnement d'un procédé en assurant une production sûre pour les hommes et/ou le matériel. Bien entendu, tout le monde se souvient de la catastrophe nucléaire de Tchernobyl (accident nucléaire le plus grave jamais survenu) le 26 avril 1986. Plus de 20 ans après, on est en droit d'attendre des systèmes de contrôle ne permettant plus aucune catastrophe de ce type. Mais ce n'est pas le cas. Par exemple, le 9 août 2004, à Fukui, à 320 km au nord-ouest de Tokyo, un accident dans la centrale nucléaire de Mihama provoque la mort de cinq personnes et fait sept blessés. La cause de l'accident est une fuite de vapeur non-radioactive dans un bâtiment abritant les turbines du réacteur. Les victimes ont été prises dans ces jets de vapeurs brûlantes et l'opérateur de la centrale reconnaît un défaut de surveillance de ses installations. Un autre exemple en Suède en 2006, la défaillance d'un système de secours de la centrale de Forsmark a engendré la fermeture de deux de ses réacteurs, et le constructeur de la centrale a déclaré : "C'est le hasard qui a évité qu'une fusion du cœur ne se produise".

Au vu de ces événements et de bien d'autres, on comprend que la maîtrise des systèmes est un thème de recherche primordial. Ainsi, nous allons tout d'abord présenter la variabilité, et ce qu'elle entraîne sur la qualité des produits. Nous verrons ensuite ce qu'est la maîtrise des procédés d'une manière générale. Nous présenterons alors les principales approches permettant la surveillance des procédés. Ensuite, dans le contexte de la surveillance se basant uniquement sur les données, nous exposerons les principaux outils pour la détection de faute dans un procédé. Suite à cela, extrapolant le problème du diagnostic à un problème de classification supervisée, nous étudierons les classifieurs les plus connus. Enfin, une dernière section donnera les raisonnements permettant d'effectuer la surveillance des procédés grâce à un classifieur, ainsi qu'une synthèse sur les classifieurs étudiés, et le choix de l'un d'entre-eux : les réseaux bayésiens.

1.2 Variabilité des procédés

Cette partie a pour but d'expliciter ce qu'est la variabilité des procédés et pourquoi sa réduction est importante dans un contexte économique toujours plus tendu. Ainsi, nous allons tout d'abord définir la notion de procédé. Ensuite, nous expliquerons d'où provient la variabilité des procédés et ses conséquences sur la production.

1.2.1 Définition d'un procédé

Selon Montgomery [104], un procédé se modélise comme sur la figure 1.1, où :
- x_0 est l'entrée du procédé (ex : matière première, composants),
- les x_i sont des facteurs contrôlables (ex : réglages machine, matière, opérateur),
- les z_i sont des facteurs non-contrôlables (ex : micro-vibrations, température, humidité, perturbations électro-magnétiques),
- la sortie $\boldsymbol{Y}$ est le produit fini (ex : une vis, une ou plusieurs caractéristiques qualité du produit fini).

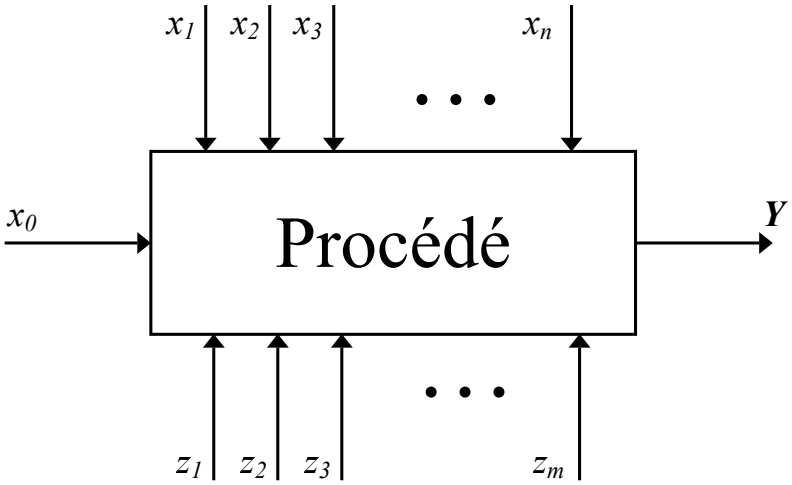


FIG. 1.1 – Modélisation d'un procédé

La sortie $\boldsymbol{Y}$ est un ensemble de caractéristiques qualité : $\boldsymbol{Y} = f(y_1, y_2, \ldots, y_p)$ (ex : longueur de la vis, diamètre de la vis). Bien entendu, il peut exister certaines corrélations entre les y_i.

Chaque caractéristique y_i est soit une caractéristique mesurable (par exemple une longueur, une température, etc), soit une caractéristique non mesurable liée à un attribut (par exemple une surface rayée ou non, présence ou absence de poussière sur une surface, nuance d'un coloris). Ces deux types de caractéristiques ne vont pas se traiter exactement de la même manière.

Dans la section suivante, nous présentons le phénomène de variabilité des procédés et l'implication de cette variabilité sur la qualité de la production. Pour faciliter la compréhension, seul le cas d'un procédé univarié ($\boldsymbol{Y} = f(y_1)$) est présenté, mais les mêmes conclusions peuvent être étendues au cas des procédés multivariés (on suit toutes les variables).

1.2.2 Les causes de variabilité des procédés

Dû à sa complexité, le procédé fait intervenir un grand nombre de paramètres pouvant être assimilés à des variables aléatoires. On peut également considérer que les caractéristiques qualité d'intérêt sont directement ou indirectement influencées par l'ensemble de ces paramètres (définissant les conditions opérationnelles nominales pour le procédé). On peut alors appliquer un théorème fort connu en statistique, le Théorème Central Limite (TCL), énoncé ci-dessous [127].

Théorème 1 (Théorème Central Limite) *Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires indépendantes. Supposons que $E(X_k) = \mu$ et $Var(X_k) = \sigma^2$ existent. Si $S_n = X_1 + X_2 + \cdots + X_n$, alors la loi de probabilité de la somme réduite*

$$S_n^* = (S_n - n\mu)/\sigma\sqrt{n}$$

converge vers une loi normale centrée réduite, c'est à dire que pour tout a, b $(a < b)$ et lorsque $n \to \infty$, on a :

$$P(a \leq S_n^* \leq b) \to \Phi(b) - \Phi(a) \tag{1.1}$$

où P signifie probabilité, et où Φ est la fonction de répartition de la loi normale centrée réduite. Sous ces termes mathématiques, si une variable aléatoire S_n est composée d'une somme de plusieurs variables aléatoires indépendantes $(X_1 + X_2 + \cdots + X_n)$, alors S_n suit une distribution s'approchant d'autant plus de la loi normale que le nombre n de facteurs composant la somme est grand, et ceci est valable quelle que soit la distribution des facteurs composant cette somme.

C'est pour cette raison que nous considérons en pratique que Y suit une loi normale $N(\mu_0, \sigma_0^2)$. En effet, si le procédé était parfait (aucun des paramètres d'entrées ne varie et pas d'influence indésirable subie), nous aurions la distribution de la figure 1.2 pour Y.

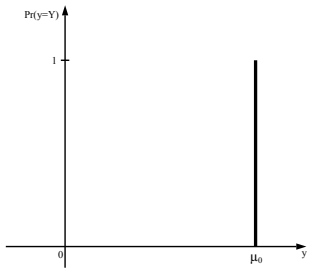


FIG. 1.2 – Sortie d'un procédé parfait ayant pour cible μ_0

Mais, le procédé n'est pas parfait et même si les entrées contrôlables x_i sont maintenues fixes (procédé maîtrisé), les variations des entrées non contrôlables z_i induisent une dispersion de la distribution généralement suivant une loi normale. Nous obtenons donc la distribution de la figure 1.3 pour Y. Ces entrées non contrôlables sont couramment appelées "causes aléatoires" ou "bruit" du procédé [120].

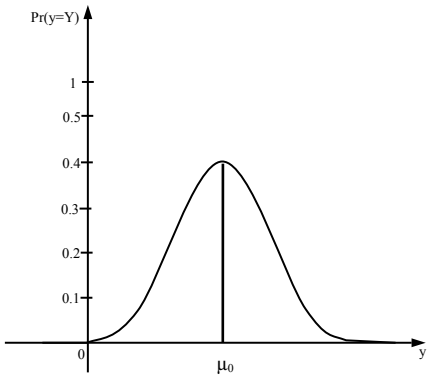


FIG. 1.3 – Sortie d'un procédé maîtrisé

De plus, si les entrées supposées contrôlées ne le sont plus, alors nous allons assister soit à un décentrage de la distribution précédente, soit à une augmentation de la dispersion, soit les deux. On peut voir sur la figure 1.4 des exemples de ces déréglages.

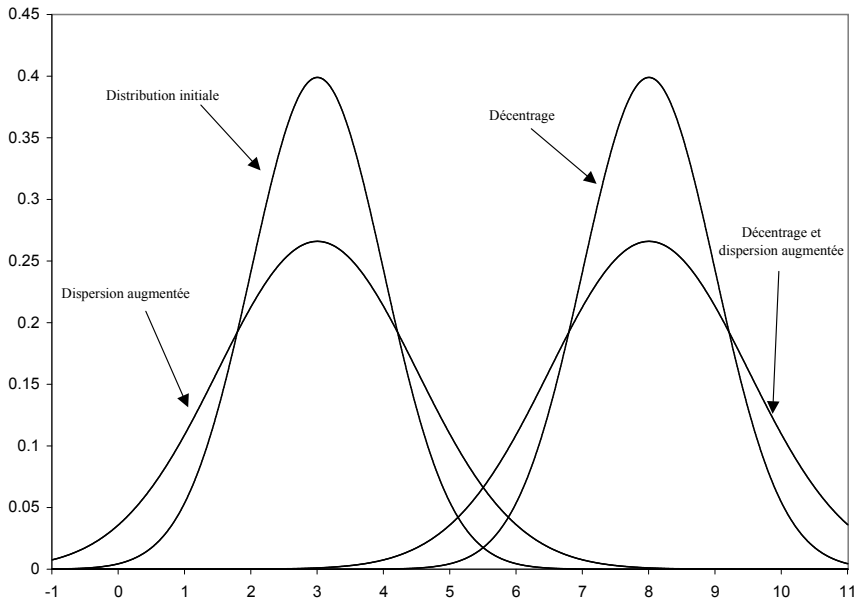


FIG. 1.4 – Déréglage de la sortie d'un procédé

C'est là qu'intervient la maîtrise statistique des procédés. Le but est de détecter l'apparition de ces décentrages ou bien de ces augmentations de la dispersion. Mais, une fois ce problème détecté, il faut également pouvoir identifier les causes de ce déréglage et pouvoir remettre le procédé en fonctionnement normal, et ce, le plus rapidement possible. En effet, l'impact de ces différents déréglages est loin d'être négligeable. Nous allons donc voir, toujours sur le cas d'un procédé univarié, les implications des déréglages d'un procédé sur la qualité d'une production.

1.2.3 Variabilité des procédés et qualité de production

Nous avons vu que la sortie d'un procédé suit généralement une loi normale de moyenne μ_0 et d'écart-type σ_0. Mais, regardons plutôt la sortie de notre procédé sur une durée plus longue. Dans le cas où notre procédé n'est soumis qu'à des causes aléatoires, nous obtenons la situation de la figure 1.5, où nous voyons que la sortie à court ou à long terme est la même.

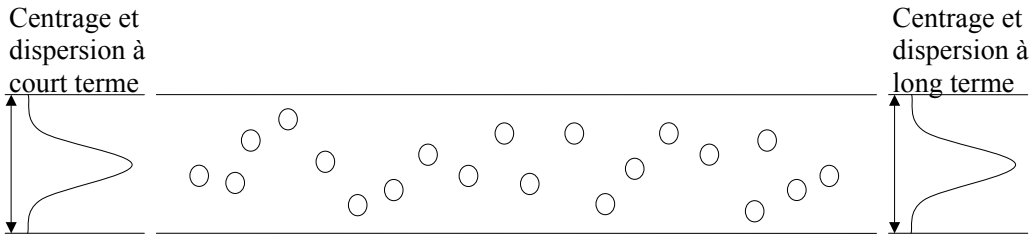


FIG. 1.5 – Evolution d'un procédé stable

Plaçons à présent les spécifications de notre caractéristique sur la sortie de notre procédé (voir figure 1.6). Les limites supérieures et inférieures de tolérance sont respectivement notées LST et LIT.

On remarque alors que les zones hachurées représentent un certain pourcentage de la production produite pour lequel la caractéristique qualité n'est pas conforme aux exigences demandées. On a l'habitude d'exprimer cette proportion de production non conforme en ppm (pièce par million). Dans le cas d'un procédé stable, on voit que plus la variabilité à court terme est forte, plus la production de non-conforme augmente. Ainsi, une production de qualité est obligatoirement associée à une variabilité à court terme la plus faible possible. Cependant, cela n'est pas suffisant. En effet, supposons à présent que les entrées contrôlées ne le soient plus, alors nous assistons soit à un décentrage de la distribution précédente, soit à une augmentation de la dispersion, soit les deux. On dit alors que l'évolution dans le temps du procédé est soumise à des "causes spéciales" ou "causes assignables" [66], ce sont des causes de variation du procédé qui ne sont pas aléatoires

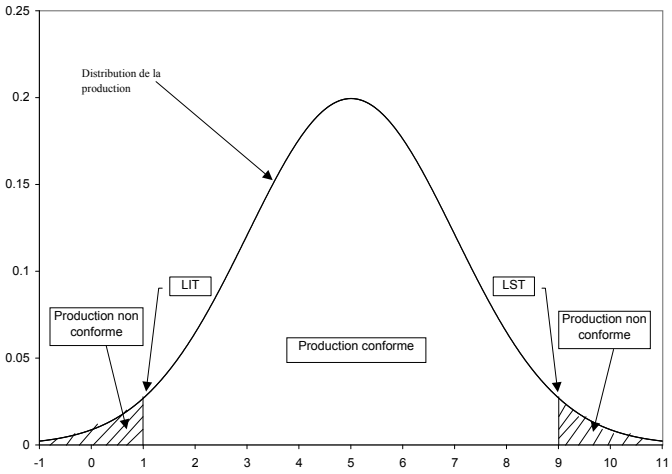


FIG. 1.6 – Production avec spécifications

et sur lesquelles il est possible d'agir (ex : usure d'outil). Ces causes agissent de deux manières sur le procédé, elles engendrent : soit des décentrages à court terme, soit une augmentation (ou diminution mais le cas est rare) de la dispersion à court terme. Si le centrage du procédé est soumis à des variations, l'évolution de la sortie du procédé est représentée sur la figure 1.7. On observe alors que des décentrages à court terme influencent le centrage, ainsi que la dispersion, à long terme. De même, si le procédé est soumis à des causes spéciales faisant varier la dispersion à court terme, nous voyons sur la figure 1.8 que la dispersion à long terme augmente également.

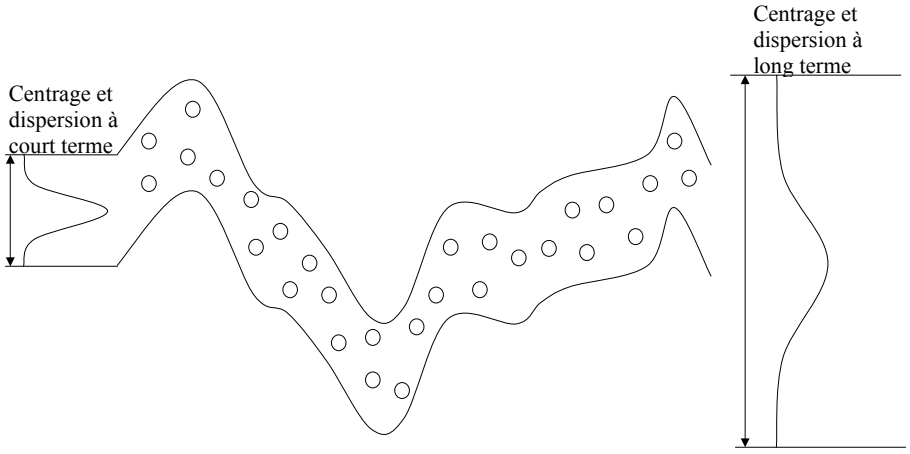


FIG. 1.7 – Variations du centrage

Les deux dernières figures (figures 1.7 et 1.8) nous démontrent bien que plus un procédé est soumis à des déréglages (soit de centrage, soit de dispersion, soit les deux), et plus la dispersion à long terme de la production est importante, engendrant alors un nombre de produits non conformes plus élevé.

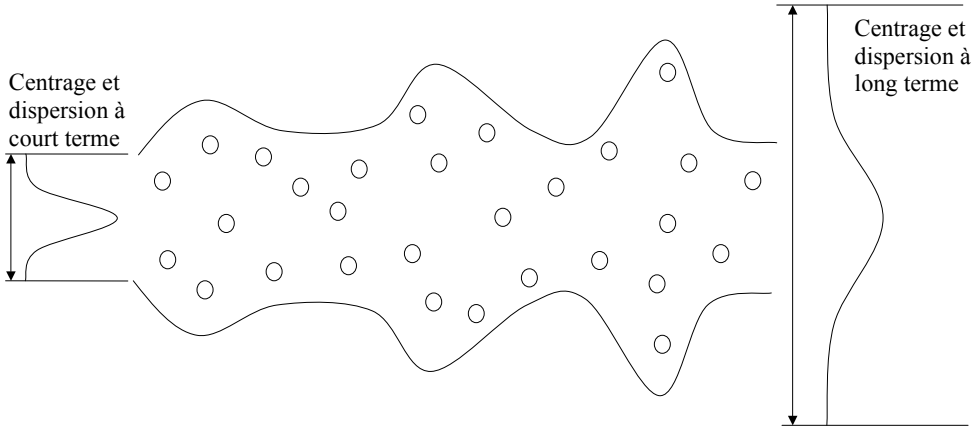


FIG. 1.8 – Variations de la dispersion

Ainsi, pour produire peu de produits non conformes, deux aspects sont essentiels. Premièrement, il faut que la variabilité à court terme soit la plus faible possible. Afin d'abaisser la variabilité à court terme d'un procédé, on utilise généralement des plans d'expériences permettant de régler au mieux le procédé [84, 132]. Deuxièmement, il faut que la variabilité à long terme soit la plus faible possible. Or, nous avons vu que si le procédé n'est pas stable, la variation à long terme est élevée, entraînant alors une augmentation de produits non conformes. Le but de la maîtrise des procédés est précisément de détecter au plus tôt ces problèmes de variations (de centrage et de dispersion) dans le procédé. Ainsi, la maîtrise des procédés permet d'obtenir une meilleure stabilité de la production et ainsi une production de pièces non conformes beaucoup moindre. On peut alors affirmer que plus la variabilité d'une production est faible, plus la qualité de cette production est importante. En effet, Montgomery [104] définit la qualité par la phrase suivante : "La qualité est inversement proportionnelle à la variabilité". Ainsi, la production d'un procédé soumis à une forte variabilité n'est pas de bonne qualité comparativement à la production d'un procédé soumis à une faible variabilité. Donc, pour obtenir une production de qualité, un objectif majeur est la réduction de la variabilité des procédés. Or, le rôle de la maîtrise des procédés, et notamment des cartes de contrôle, est de détecter la présence de causes spéciales (variations du centrage et variations de la dispersion à court terme), afin de pouvoir les éliminer et retrouver ainsi le régime nominal. L'objectif de la maîtrise des procédés est donc la réduction de la variabilité, afin d'obtenir une qualité correspondant à des critères donnés.

1.3 Maîtrise des procédés

L'objectif de la maîtrise des procédés porte sur la réduction ou la disparition des causes spéciales (ou fautes) dans le procédé. Dans un premier temps, nous allons présenter le schéma de fonctionnement global de la maîtrise des procédés, puis nous présenterons les différents types d'approches permettant de réaliser celle-ci.

1.3.1 Les grandes étapes de la maîtrise des procédés

La maîtrise des procédés peut se décomposer en plusieurs grandes étapes [17]. En effet, quelle que soit l'approche employée afin de maîtriser le procédé, il est toujours possible d'identifier trois principales étapes : la détection, le diagnostic et la reconfiguration. La combinaison de ces étapes permet d'ôter du procédé toutes causes spéciales étant apparues dans celui-ci.

1.3.1.1 Détection

La première étape de la maîtrise des procédés consiste en la détection de causes spéciales. Cette étape a pour but de détecter si le procédé est soumis à l'effet d'une cause spéciale impliquant un accroissement de la variabilité du procédé. Cette première étape de la maîtrise des procédés est essentielle. En effet, si une faute n'est pas détectée, la production engendrée peut ne plus devenir conforme aux spécifications exigées.

L'objectif de performance de la détection se situe sur la vitesse de détection. En effet, plus une faute est présente dans le procédé, plus elle engendre de production erronée. Il est donc essentiel pour tout système de détection d'être capable de conclure sur la présence ou non de tout type de faute, et ce le plus rapidement possible. Dès lors, lorsque nous parlerons de système de détection nous aurons deux critères significatifs : l'aptitude à détecter différents types de fautes dans le procédé, ainsi que la vitesse de détection associée au type de faute.

Tout système de détection statue sur la présence ou non de faute dans le procédé. Mais, le fait de savoir qu'une faute est apparue dans le procédé ne permet pas de savoir quelle est la nature de cette faute : ceci va être le rôle de la phase de diagnostic.

1.3.1.2 Diagnostic

L'étape de diagnostic est le fait de désigner la faute qui est apparue, c'est à dire déterminer la cause de la détection d'une situation hors-contrôle (faute dans le procédé). On peut étendre cette définition en diagnostiquant la faute (son type) mais également

son emplacement, son amplitude, sa durée. Il est courant de séparer cette étape en deux phases : l'identification et la décision. La décision est la phase de diagnostic proprement dite, aboutissant sur l'attribution d'une cause physique affectant le procédé et ayant impliqué son fonctionnement anormal. La phase d'identification, quant à elle, intervient en amont de la phase de décision. Son rôle est d'identifier un ensemble de variables actives dans l'apparition de la faute, et ainsi fournir au système de décision un espace plus restreint pour la recherche du diagnostic. Cette phase préliminaire d'identification n'est pas obligatoire, mais dans certains cas elle s'avère quasiment indispensable afin de diagnostiquer correctement une situation hors-contrôle [17, 43].

Un système de diagnostic statue sur l'origine (la cause) de l'apparition d'une faute dans le procédé. Mais, avoir identifié la cause d'un problème sur un procédé ne signifie pas que le problème est résolu : ceci est le but de l'étape de reconfiguration.

1.3.1.3 Reconfiguration

La reconfiguration est l'étape dans laquelle l'entité (opérateur, ingénieur, automate,...) chargée de la bonne marche du système doit remédier à la faute apparue [55]. On peut voir cette étape comme un retour aux conditions nominales de fonctionnement du procédé. Ce retour aux conditions nominales est différent suivant les types de problèmes rencontrés. Il peut s'agir d'actions correctives sur des composants physiques du procédé et/ou d'adaptations de réglages sur la commande du procédé.

1.3.1.4 La boucle de surveillance

Au vu des étapes décrites auparavant, nous établissons le schéma de la maîtrise des procédés sur la figure 1.9 [17, 43].

Sur le schéma de la figure 1.9, nous retrouvons le procédé, couplé à son organe de commande. On voit que nous sommes en présence d'un procédé régulé par boucle de retour. Cette condition de régulation par boucle fermée n'est pas obligatoire et le principe de maîtrise du procédé est le même pour un procédé commandé en boucle ouverte.

Nous considérons que notre système est muni d'un système d'instrumentation, au sens large, permettant l'acquisition des différentes données émanant du procédé ainsi que de sa commande. Le système d'instrumentation est vu comme un ensemble de capteurs, de mesures relevées par un opérateur ou par un automate de mesure, etc. Ce système transmet les données acquises vers l'organe de surveillance.

L'organe de surveillance est composé des deux premières étapes de la maîtrise des procédés, la détection et le diagnostic (identification et décision), ainsi que d'un modèle

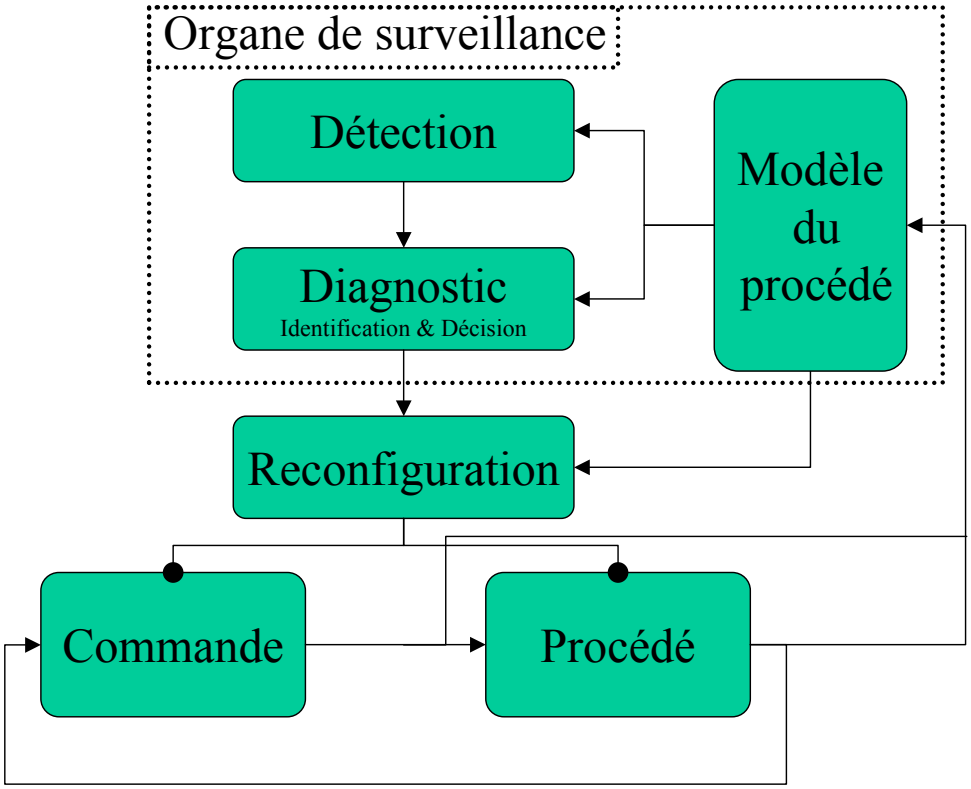


FIG. 1.9 – Schéma de la maîtrise des procédés

du procédé. Nous entendons par modèle toute information que l'on possède sur le procédé. Suivant les informations contenues dans le modèle, nous distinguons alors plusieurs types d'approches concernant la surveillance des procédés. Ces types d'approches sont développés dans la section suivante.

Suite à l'organe de surveillance, on retrouve la dernière étape fondamentale de la maîtrise des procédés : la reconfiguration. On voit sur le schéma que la reconfiguration est considérée pouvoir agir à la fois sur le procédé en lui-même, mais également sur sa commande.

Le principe de fonctionnement de la maîtrise des procédés est donc le suivant. Les données entrant et sortant du procédé sont envoyées à l'organe de surveillance. En se basant sur les informations du modèle, la méthode de détection conclut sur l'état du procédé. Si le procédé est déclaré sous contrôle, aucune action n'est nécessaire. Si le procédé est déclaré hors-contrôle (présence d'une cause assignable), la méthode de diagnostic, tout en s'appuyant sur les informations du modèle, effectue une identification puis statue sur la cause de la faute (décision). Une fois la cause identifiée, l'étape de reconfiguration agit directement sur le procédé et/ou la commande afin de retrouver un régime nominal le plus rapidement possible.

Dans la suite de cette thèse, nous allons travailler sur l'organe de surveillance, nous ferons donc abstraction de la partie de reconfiguration du système. En effet, en pratique cette étape est presque uniquement fonction du procédé réel, ce qui lors d'une approche académique comme la notre ne peut être réellement étudiée [17].

Nous avons vu que l'organe de surveillance, afin de tirer les conclusions nécessaires à la bonne marche du système, se base sur les données qu'il reçoit mais également sur le modèle du système. Il est donc important de comprendre ce que peut être ce modèle, ainsi que les différents types d'approches de détection et diagnostic en découlant.

1.3.2 Les différentes approches

Dans cette section, nous nous intéressons à l'organe de surveillance. Nous avons vu que cet organe de surveillance possède un modèle du procédé, modèle que nous avons pour le moment défini comme étant toute information acquise concernant le procédé. Suivant le type d'information que contient le modèle, il est possible de distinguer trois principaux types d'approches pour la surveillance : les méthodes à base de modèles analytiques [148], les méthodes à base de connaissances [146], et les méthodes basées sur les données [147].

1.3.2.1 Les méthodes à base de modèles analytiques

Les méthodes basées sur des modèles analytiques sont également appelées méthodes à redondance analytique. Ces méthodes utilisent un modèle décrit par des relations mathématiques représentant les différentes relations physiques du procédé. Généralement, ces relations physiques découlent de l'application de lois fondamentales de divers domaines (physique, chimie, électricité, thermodynamique, mécanique, etc). Ainsi, il est possible de créer une modélisation du système qui, en lui appliquant les entrées U similaires au système réel (lois de commande, paramètres du procédé, etc), fournit une réponse estimée du système $\hat{Y}$. Il est alors possible de calculer l'écart entre la réponse réelle du système (Y) et sa réponse estimée ($\hat{Y}$), comme indiqué sur la figure 1.10. Cet écart est usuellement appelé résidu (R).

En d'autres termes, on peut dire que les résidus sont les écarts entre les observations du système et le modèle mathématique. L'objectif de ce type d'approche est de réussir à faire la distinction entre les résidus causés par des fautes (causes assignables) et les résidus causés par les autres sources de variation précédemment citées (causes aléatoires). Les résidus sont relativement élevés lorsqu'une faute est présente dans le procédé, et sont plutôt faibles en l'absence de faute. Dans ce cas, la présence de faute est détectée en appliquant des seuils adéquats sur les résidus.

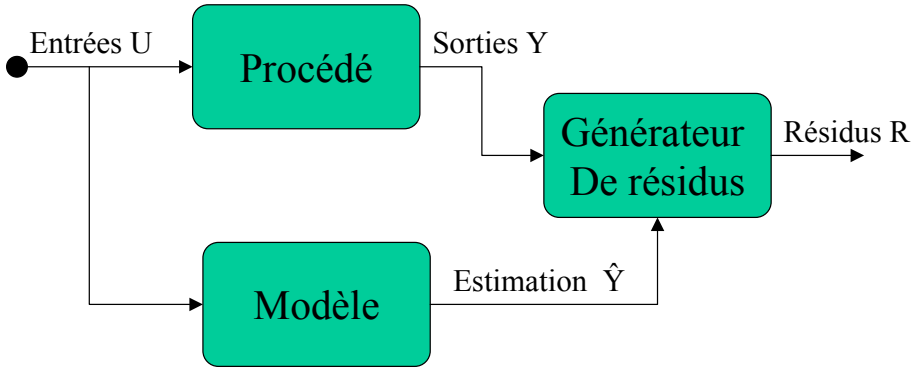


FIG. 1.10 – Génération de résidus

Il existe plusieurs approches de génération de résidus, cependant trois sont vraiment essentielles [17], il s'agit de :

Estimation de paramètres : Les résidus sont la différence entre les paramètres nominaux du modèle et les paramètres estimés du modèle [62].

Observateurs : Les méthodes à base d'observateurs reconstruisent une estimation de la sortie à partir de toutes ou parties des sorties réelles du système. Les résidus sont alors classiquement la différence entre les sorties mesurées et les sorties estimées [48].

Equations de parité : Cette méthode consiste à vérifier l'exactitude des équations mathématiques du modèle en se basant sur les sorties du procédé [52].

Lorsque le modèle mathématique du système est disponible, ces méthodes à base de modèles analytiques sont très performantes. Elles sont généralement intitulées FDI (Fault Detection and Isolation). En effet, alors que pour la notion de détection de fautes toutes les communautés scientifiques partagent la même définition, pour ce qui est du diagnostic beaucoup de divergences apparaissent. Il semble donc important de définir ici ce qui est entendu par isolation de faute. L'isolation de faute est la détermination du lieu exact de la faute afin de déterminer quel(s) composant(s) du système est/sont défectueux. L'isolation de faute fournit plus d'informations que la phase d'identification de variables énoncée au paragraphe 1.3.1, mais moins que la phase de diagnostic toute entière puisque celle-ci comprend également les notions d'emplacement, d'amplitude et de durée de la faute.

Comme nous l'avons déjà signalé, l'approche à base de modèles analytiques donne des résultats supérieurs aux autres méthodes (connaissances ou données). Mais, ceci n'est vrai que lorsque le modèle est bien construit. Or, la construction du modèle pour des systèmes complexes et/ou de grandes envergures, devient presque impossible. De plus, même si l'on arrive à bâtir un modèle, il n'est que rarement assez détaillé et précis pour permettre d'obtenir des résultats satisfaisants.

1.3.2.2 Les méthodes à base de connaissances

Dans le cas où un modèle analytique du procédé n'est pas disponible, une solution est l'exploitation de la connaissance humaine disponible sur le procédé. Il existe alors des méthodes exploitant les connaissances qualitatives que détiennent des experts sur le procédé étudié. On peut notamment citer quelques techniques telles que les systèmes experts [14], l'AMDE (Analyse des Modes de Défaillance et de leurs Effets) [47], l'AMDEC (Analyse des Modes de Défaillance, de leurs Effets et de leurs Criticités) [47], ainsi que les arbres de défaillances [165].

Les systèmes experts sont des techniques d'intelligence artificielle, basés sur les connaissances, permettant d'imiter le raisonnement humain pour la résolution d'un problème. Un système expert bien conçu est capable de représenter l'expertise humaine existante, prendre en compte des bases de données existantes, d'acquérir de nouvelles connaissances, d'effectuer de l'inférence logique, de donner des suggestions, et finalement de prendre des décisions basées sur un raisonnement.

Les 4 composants classiques d'un système expert sont :

- la base de connaissance,
- le moteur d'inférence,
- l'interface avec l'utilisateur,
- l'interface avec l'expert.

On représente un système expert comme sur la figure 1.11.

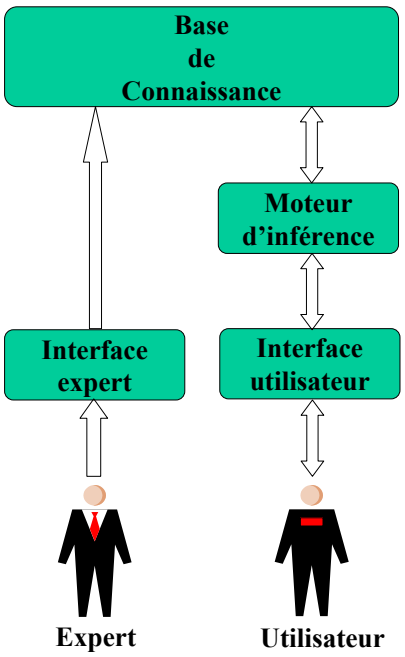


FIG. 1.11 – Schéma d'un système expert

L'avantage principal dans l'utilisation d'un système expert est que l'expert humain n'a plus besoin d'être physiquement présent, puisqu'il est là virtuellement par le biais du système de connaissance. Mais, l'élaboration de la base de connaissance pour des systèmes de grandes tailles est une tâche très ardue, ce qui dans la pratique limite l'application de cette technique à des systèmes avec un nombre d'entrées-sorties restreint. Cependant, certains auteurs [17, 105] affirment que les systèmes experts représentent une approche devant se développer fortement dans le futur.

L'AMDE [47] est une technique issue de la communauté de la sûreté de fonctionnement. Elle permet une analyse systématique et très complète, composant par composant, de tous les modes possibles de défaillance et précise leurs effets sur le système global. L'AMDE consiste à établir sous forme de tableau l'ensemble des différentes défaillances de chaque composant du système, et d'en analyser les conséquences (effets) directes sur le système et son entourage (notamment l'opérateur). Il est possible de renforcer l'AMDE par une étude de la criticité, obtenant ainsi l'AMDEC [47]. L'étude de criticité détermine quels sont les modes de défaillances les plus critiques en prenant en compte les notions de gravité des différents modes couplées à des notions de probabilité (fréquence d'apparition). La table 1.1 montre à quoi ressemble le tableau résultant d'une AMDEC d'une pompe à huile, où f représente la fréquence d'apparition de l'incident, g sa gravité, d sa détection, et où c (la criticité) est la multiplication des trois premiers critères.

	Fonction	Mode de défaillance	Cause de la défaillance	Effet	f	g	d	c
Pompe	Assurer le débit d'huile	Baisse du débit	Usure abrasive des engrenages	Diminution de la durée de vie du système	2	2	4	16
		Irrégularité du débit	Cavitation	Détérioration des parties frottantes	1	2	4	8
		Arrêt du débit	Détérioration du joint a lèvres	Grippage des coussinets	3	4	1	2
			Rupture de la clavette	Grippage des engrenages.	1	4	1	4

Tab. 1.1 – Tableau d'une AMDEC

Une fois l'AMDEC réalisée, on l'utilise afin de diagnostiquer des situations hors contrôle du procédé. Ainsi, en partant des effets observés, on peut remonter rapidement vers la cause de ces effets grâce au tableau réalisé.

Cette méthode est très puissante car dès l'apparition d'effets indésirables sur le procédé elle permet de rapidement remonter vers les causes ayant engendrées ces effets. Mais, plusieurs inconvénients rendent cette démarche non réalisable sur des systèmes trop complexes. En effet, l'établissement d'un tableau AMDEC pour des systèmes de grandes échelles demande un investissement beaucoup trop lourd afin de référencer toutes les dé-

faillances possibles ainsi que les relations causes-effets de celles-ci. De plus, cette méthode ne permet pas la prise en compte de combinaisons de plusieurs défaillances.

Un autre outil issu de la sûreté de fonctionnement sont les arbres de défaillances [90]. Un arbre de défaillance se présente sous la forme d'un diagramme logique où un événement indésirable (une faute précise) est placé au sommet. Ensuite, les causes immédiates de cette faute sont reliées grâce à des connecteurs logiques "ET" et "OU", et ainsi de suite jusqu'à atteindre, à la base, un ensemble d'événements élémentaires (voir figure 1.12). Cet outil présente les mêmes avantages et les mêmes inconvénients que l'AMDEC.

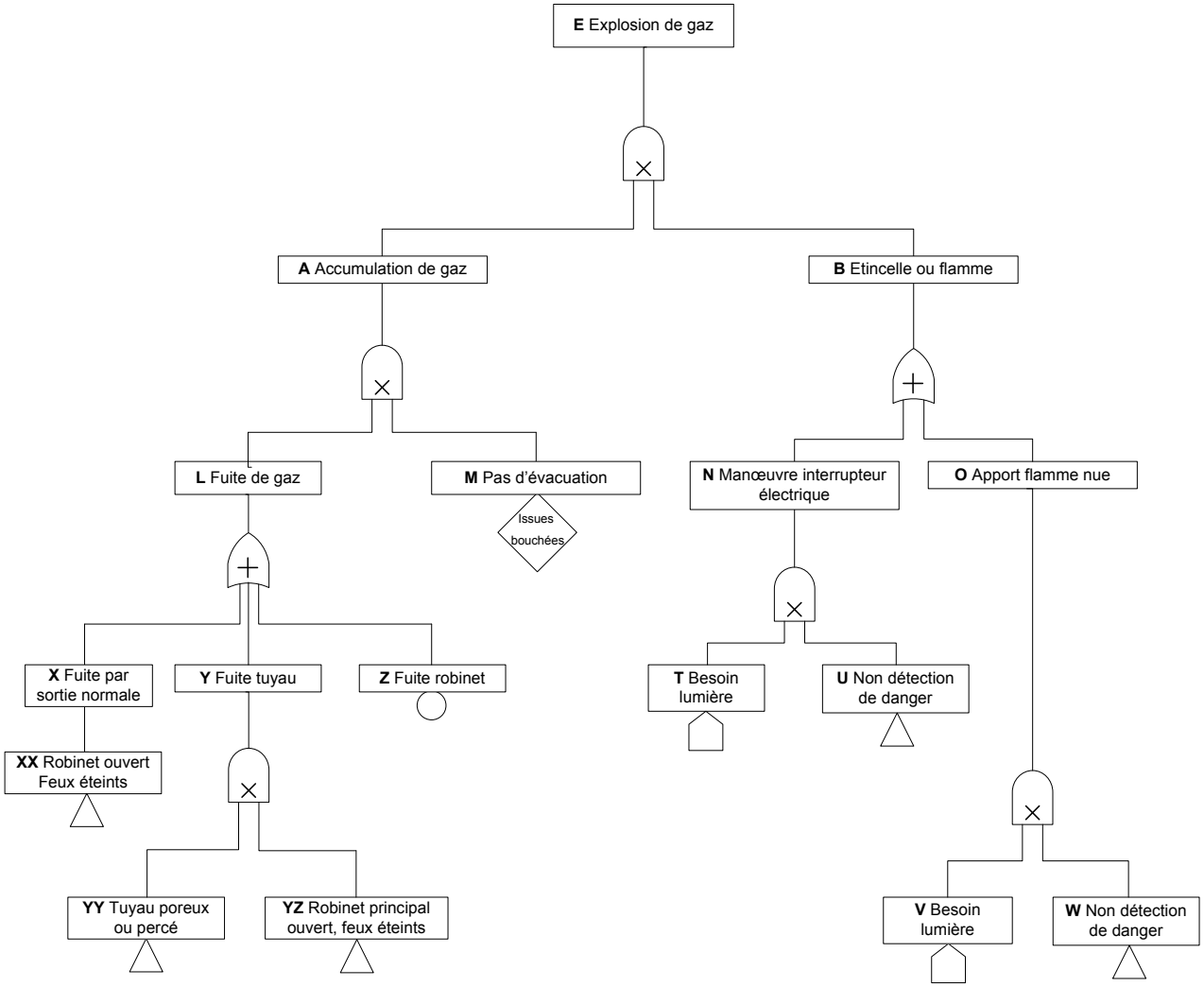


FIG. 1.12 – Exemple d'arbre de défaillances

1.3.2.3 Les méthodes basées sur les données

De nos jours, les procédés sont de plus en plus automatisés, permettant ainsi la récupération d'une quantité non négligeable de données. Il est donc naturel de surveiller le procédé avec des méthodes se basant sur ces données. En effet, la masse de données à traiter est tellement importante qu'un opérateur ne peut pas directement suivre chaque variable du procédé. Nous faisons donc appel à des techniques basées sur les données pour représenter en quelques valeurs judicieuses, l'information exprimée par toutes les variables du procédé. Certaines de ces techniques permettent la détection alors que d'autres s'intéressent au diagnostic. Parmi, ces méthodes, nous pouvons notamment citer les approches par cartes de contrôle [120], par analyse en composantes principales [83] et par projection dans les structures latentes pour la phase de détection, alors que pour la phase de diagnostic, nous retrouvons principalement des outils de classification tels que l'analyse discriminante [42] ou les réseaux de neurones [17, 165].

Du point de vue de la détection, la méthode de surveillance la plus ancienne est sans doute la carte de contrôle. En effet, la première méthode de surveillance basée uniquement sur les données est la carte de contrôle $\overline{X}$ proposée par Shewhart [134]. Cette carte de contrôle est en fait une succession d'un même test d'hypothèse $\mu_0 = \mu_t$, où μ_0 représente la moyenne cible de la variable surveillée et μ_t représente la moyenne du procédé à un instant t. Il existe également d'autres cartes de contrôle permettant de surveiller une seule variable : les cartes R et S pour surveiller la dispersion de la variable [120], ainsi que les cartes EWMA (Exponentially Weighted Moving Average) [123] et CUSUM (CUmulated SUM) [111] pour la détection de faibles sauts dans la moyenne. Le principal inconvénient de ces cartes est qu'elles ne peuvent suivre qu'une variable à la fois. Sur le même principe (test d'hypothèse) que les cartes de contrôle univariées (une seule variable), on peut prendre en compte non plus une mais plusieurs variables grâce aux cartes de contrôle multivariées, notamment la carte du T^2 de Hotelling [60], ainsi que les extensions multivariées des cartes EWMA et CUSUM, à savoir MEWMA (Multivariate EWMA) [93] et MCUSUM (Multivariate CUSUM) [119]. Ces cartes prennent en compte chaque variable, ainsi que la corrélation entre ces variables. Tout comme leurs homologues univariées, les cartes MEWMA et MCUSUM permettent la détection de décentrages de plus faibles amplitudes que la carte T^2. Une autre approche pour la détection est l'utilisation des composantes principales extraites des données. L'Analyse en Composantes Principales (ACP) [46, 83] est une technique permettant de réduire le nombre de variables à étudier de manière significative. En effet, il s'agit d'une transformation linéaire d'un espace de données corrélées en un espace de données non-corrélées. Ainsi, le premier axe de ce nouvel espace est la direction de l'espace expliquant la plus grande partie de la variabilité des données. Puis,

le second axe, orthogonal au premier, est choisi en représentant également un maximum de variabilité, et ainsi de suite. Du fait de cette transformation, les premiers axes de ce nouvel espace expliquent donc à eux seuls la majeure partie de la variabilité des données. Ainsi, une surveillance de quelques premiers axes principaux suffira à détecter une éventuelle faute dans le procédé (notamment au moyen de cartes de contrôle). Il est également possible de surveiller les résidus de la transformation inverse [17]. Dans le même esprit que l'ACP, on trouve la Projection dans les Structures Latentes (PSL) [51, 59], également connue sous la dénomination de Moindres Carrés Partiels (MCP). Cette technique consiste également en une réduction de dimension de l'espace. Elle maximise la covariance entre une matrice de prédicteur et une matrice prédite, et ce, pour chaque composante du nouvel espace. Bien souvent, la matrice prédite regroupe les caractéristiques qualité d'un produit, et toutes les autres variables du système sont placées dans la matrice des prédicteurs. Une fois le nouvel espace décrit, on surveille alors la variabilité des composantes de la même façon que les composantes d'une ACP [17].

Du point de vue du diagnostic, nous nous intéressons aux différentes techniques de classification [44]. En effet, une faute apparue dans un procédé couvre un lieu de l'espace décrit par les variables du procédé. Une autre faute couvre un autre lieu de cet espace ou bien le même lieu mais sous une autre forme ou une autre dispersion [43]. Un exemple bidimensionnel pour 3 classes est proposé sur la figure 1.13.

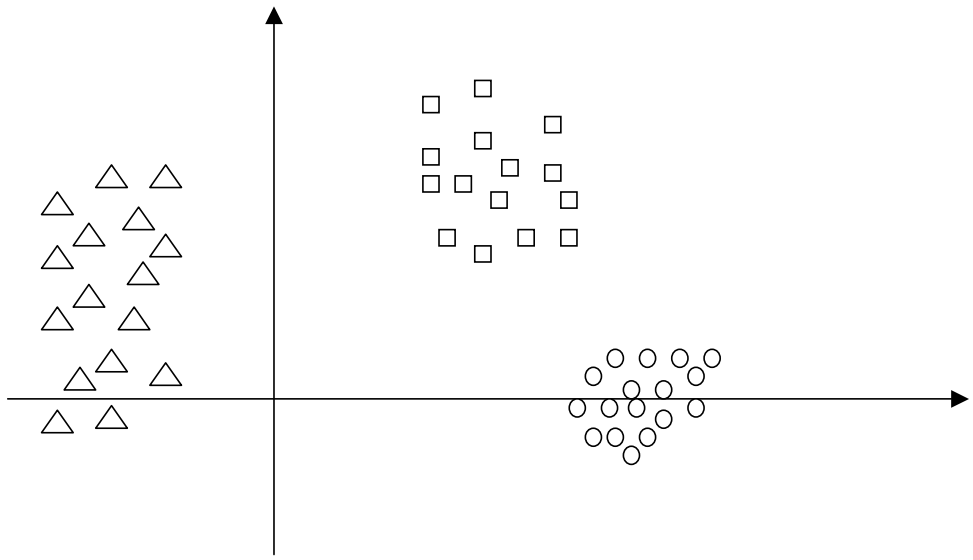


FIG. 1.13 – Différentes classes de fautes

Le diagnostic peut se voir comme la tâche de classer un nouvel individu déclaré hors-contrôle dans une des classes du procédé. Un système de classification évolué doit être capable de traiter plusieurs notions [32] :

– établir les différentes classes du modèle,
– classer correctement les nouveaux individus,
– déceler l'apparition de nouvelles fautes.

La première notion fait appel à ce que l'on nomme l'apprentissage non-supervisé : étant donnés les individus déclarés hors-contrôle, il s'agit d'identifier les différentes classes de fautes à l'intérieur de ces données et d'attribuer ainsi un numéro de faute à chaque classe ainsi qu'aux individus lui appartenant [42]. Nous avons alors à disposition un ensemble d'apprentissage composé d'individus dont la classe de faute est identifiée. Cet ensemble est utilisé en apprentissage supervisé afin de classer correctement de nouveaux individus dans une des classes identifiées. Bien entendu, si un nouveau type de faute apparaît, le système de classification doit être capable de le déceler.

Parmi les classifieurs les plus connus, on peut citer : l'analyse discriminante [49], les k plus proches voisins [27], les arbres de décisions [23], les machines à vecteurs supports [144], les réseaux de neurones [41] ainsi que les réseaux bayésiens [50].

1.4 Méthodes de détection et diagnostic non-supervisées

1.4.1 Détection, diagnostic et classification

Comme nous l'avons vu dans la section précédente, concernant les approches de surveillance basées sur les données, le diagnostic peut être considéré comme une tâche de classification supevisée. Il est également possible de considérer la phase de détection comme une classification supervisée. En effet, la détection peut être vue comme une classification en deux classes : fonctionnement normal du procédé, et respectivement fonctionnement anormal du procédé. Le diagnostic peut être vu comme une extension de la détection pour le cas où le nombre de classes est supérieur à 2. Ainsi, dans le cas d'un fonctionnement anormal détecté, on peut s'intéresser de savoir quelle faute s'est produite. Cependant, les classifications pour le diagnostic se différencient en classification supervisée et non-supervisée. Dans le premier cas, les frontières de séparation des classes sont déduites sur la base d'un échantillon d'individus pour lesquels on connaît l'appartenance aux différentes classes. Dans le cas de la classification non-supervisée, les frontières de séparations sont obtenues suivant certains principes (tels que l'homogénéité des classes ou bien la distance aux différents voisins) à partir des données disponibles et sans aucune information concernant l'appartenance des individus aux différentes classes.

La phase de détection ne peut pas être directement assimilée à ces types de classification. En effet, dans le cas de la détection, bien que l'on suppose posséder des exemples

d'individus de la classe de fonctionnement normal, nous devons pouvoir effectuer la détection même si aucun jeu de données de fonctionnement anormal n'est disponible. Ce type de classification est appelé classification monoclasse ("one-class classification") [137]. Le but de la classification monoclasse est de décrire une classe d'individus, et de distinguer si un nouvel individu appartient ou non à cette classe.

Plusieurs approches existent pour la détection de fautes dans un procédé multivarié. Nous avons vu à la section 1.3.2 que les méthodes à base de modèles analytiques, ainsi que les méthodes à base de connaissances, ne sont pas réellement adaptées pour la surveillance de systèmes complexes. Nous allons dans cette partie nous restreindre à l'étude des méthodes basées sur les données. La plupart de ces méthodes sont basées sur des outils statistiques. Nous allons dans un premier temps présenter les cartes de contrôle multivariées, puis nous passerons aux techniques basées sur l'Analyse en Composantes Principales (ACP) et nous finirons enfin par la présentation de la technique de Projection dans les Structures Latentes (PSL).

1.4.2 Cartes de contrôle multivariées : détection et diagnostic

Dans cette partie, nous présentons les principales cartes de contrôle multivariées [10, 142] : la carte T^2 [60], la carte Multivariate Exponentially Weighted Moving Average (MEWMA) [93] et la carte Multivariate CUmulated SUM (MCUSUM) [119]. Ces différentes cartes permettent la surveillance d'un procédé multivarié. Avant de présenter le principe de chaque carte, nous rappelons tout d'abord quelques éléments de statistique multivariée afin de bien définir les différentes notions utilisées par la suite.

1.4.2.1 Eléments de statistique multivariée

Le vecteur d'observations Le fait de travailler dans un espace à p dimensions implique que nous ne traitons plus une seule variable aléatoire univariée, mais un groupement de variables aléatoires $X_1, X_2, \ldots, X_p$, équivalent à une variable aléatoire multivariée que nous notons $\boldsymbol{X}$. Les observations $x_1, x_2, \ldots, x_p$ des variables aléatoires $X_1, X_2, \ldots, X_n$ sont alors représentées sous la forme d'un vecteur $\boldsymbol{x}$:

$$\boldsymbol{x} = \begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{vmatrix} \tag{1.2}$$

Ce vecteur $\boldsymbol{x}$ est appelé vecteur d'observations, représentant un individu (une obser-

vation) particulier de la distribution de la variable $\boldsymbol{X}$. Ce vecteur représente directement les valeurs des variables d'un procédé, ou alors il représente les moyennes de ces variables pour un échantillon de taille n.

Le vecteur cible Dans le domaine de la statistique multivariée, nous utilisons également le vecteur cible. On le nomme ainsi car il traduit l'objectif du procédé. On le pose comme un vecteur $\boldsymbol{\mu}$ où chaque ligne représente la cible (bien souvent la valeur moyenne) μ_i de la variable aléatoire X_i. De la même manière que pour le vecteur observations, le vecteur cible s'écrit :

$$\boldsymbol{\mu} = \begin{vmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{vmatrix} \tag{1.3}$$

La matrice de variance-covariance Une autre notion importante est celle de la matrice de variance-covariance. En effet, dans le cas univarié, on ne s'intéressait qu'à la variance de la variable aléatoire, mais dans le contexte multivarié, nous nous intéressons à chaque variance (σ_i^2) et à chaque covariance (σ_{ij}). Cette matrice, notée $\boldsymbol{\Sigma}$, s'écrit :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix} \tag{1.4}$$

Loi normale multivariée La loi de distribution la plus courante lorsque l'on traite des données multivariées est la loi normale multivariée ($\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). La distribution normale multivariée possède une fonction de densité de probabilité donnée ci-après en écriture matricielle, où le symbole ^{T} signifie transposition :

$$\phi(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} e^{\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}} \tag{1.5}$$

Pour mieux se rendre compte de ce que peut être une distribution normale multivariée, la figure 1.14 montre la représentation graphique de $\phi(\boldsymbol{x})$ avec deux variables, c'est ce que l'on appelle une distribution normale bivariée. Les cartes de contrôle présentées par la suite font l'hypothèse que le procédé suit une loi normale multivariée $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\mu}$ la cible du procédé et $\boldsymbol{\Sigma}$ sa matrice de variance-covariance.

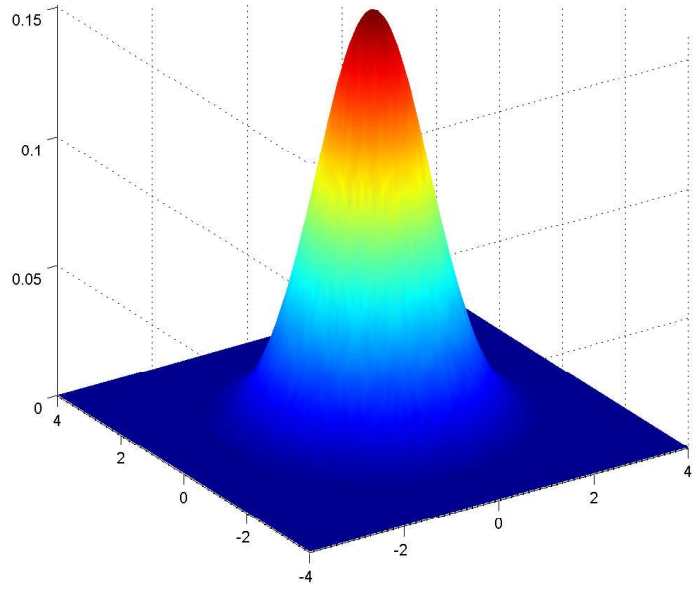


FIG. 1.14 – Distribution bivariée

1.4.2.2 La carte T^2 de Hotelling

Les premiers travaux concernant la carte T^2 datent de 1947 [60]. Hotelling fut le premier à mettre au point un concept de surveillance de procédé multivarié. Hawkins [56] a montré que le principal avantage de cette carte est qu'elle représente le meilleur test statistique pour détecter un déréglage de la moyenne du procédé. Pour un procédé à p variables, et en utilisant les notations matricielles, le T^2 s'écrit sous la forme de la distance statistique suivante :

$$T^2 = n(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{1.6}$$

où :

n : taille de l'échantillon prélevé

$\boldsymbol{x}$: vecteur d'observations à un instant donné,

$\boldsymbol{\mu}$: vecteur cible du procédé,

$\boldsymbol{\Sigma}$: matrice de variance-covariance du procédé.

Le calcul du T^2 donne un scalaire. On peut donc représenter cette mesure dans le temps sous la forme d'une carte de contrôle. Bien entendu, à chaque instant, la mesure du T^2 est comparée à une limite de contrôle supérieure (la limite de contrôle inférieure étant fixée à 0). Cette limite de contrôle, notée LC, permet de respecter un certain taux α de fausses alarmes. Le calcul des limites est différent suivant les cas. En effet, il nous faut distinguer plusieurs hypothèses.

– Premièrement, on peut être en phase d'observation du procédé, phase I (estimation des paramètres avec les données de la carte) ou en phase de surveillance, phase II (les paramètres ont déjà été estimés et sont figés). En phase I, l'estimation est basée sur m échantillons de données que l'on vient de prélever sur le procédé mais sans savoir si celui-ci était sous contrôle lors du prélèvement. Dans la phase II, l'estimation a été faite sur m échantillons alors que le procédé était sous contrôle. Ainsi, les estimateurs trouvés lors de la phase II sont supposés valides pour représenter le fonctionnement normal du procédé.
– Deuxièmement, il nous faut distinguer les limites de contrôle en fonction de la taille des échantillons prélevés. On distinguera donc 2 cas : soit $n > 1$ ou alors $n = 1$.

Le tableau 1.2 répertorie les limites de contrôle dans chaque cas, où B et F représentent respectivement des quantiles des distributions Beta et Fisher. On peut également préciser que si m est très grand ($m > 250$) alors on peut prendre $LC = \chi^2_{\alpha,p}$, où χ^2 représente un quantile de la distribution du chi-deux. Pour les calculs concernant les estimations des paramètres, on peut se reporter à Montgomery [104]. Mais, dans le cas d'observations individuelles (taille d'échantillon de 1), Sullivan [136] propose un comparatif très intéressant entre 5 estimateurs de $\mathbf{\Sigma}$.

	$n = 1$	$n > 1$
Phase I	$LC = \frac{(m-1)^2}{m} B_{\alpha,p/2,(m-p-1)/2}$	$LC = \frac{p(m-1)(n-1)}{nm-m-p+1} F_{\alpha,p,nm-m-p+1}$
Phase II	$LC = \frac{p(m+1)(m-1)}{m^2-mp} F_{\alpha,p,m-p}$	$LC = \frac{p(m+1)(n-1)}{nm-m-p+1} F_{\alpha,p,nm-m-p+1}$

TAB. 1.2 – Les limites de contrôle pour la carte T^2

Ainsi, à chaque instant d'échantillonnage du procédé, nous obtenons le vecteur d'observations $\boldsymbol{x}$. On rappelle que ce vecteur d'observations contient les moyennes de chaque variable pour l'échantillon de taille n prélevé. On calcule alors la valeur du T^2 et on la compare à la limite de contrôle. Si la valeur du T^2 est inférieure à LC, alors le procédé est déclaré sous contrôle, sinon il est déclaré hors-contrôle. Cependant, la carte ne donne aucune indication concernant le diagnostic de la situation hors-contrôle. Pour cela, beaucoup de méthodes ont été proposées [20, 35, 57, 98, 143]. Une étude comparative est effectuée par Tiplica [141]. Cependant, Mason et al. [98] ont démontré que la plupart de ces techniques sont des cas particuliers de la méthode "MYT" [98, 97].

1.4.2.3 Diagnostic par décomposition MYT

Cette décomposition a été mise au point par Mason, Young et Tracy [98], d'où le nom "décomposition MYT". De plus, pour comprendre cette méthode de manière plus intuitive, les auteurs donnent un exemple avec un procédé bivarié [97]. Il est également à préciser que les auteurs ont prouvés que certaines méthodes mises au point peuvent se ramener à des cas particuliers de décomposition MYT [98]. En effet, la décomposition MYT réunit les idées de Hawkins [57], basées sur la régression multiple et l'analyse des résidus, et de Doganaksoy [35] sur la contribution des variables à la statistique de Student.

Le principe de la méthode MYT est de décomposer la statistique T^2 dans un nombre limité de composantes orthogonales qui sont également des distances statistiques (et donc surveillables). La décomposition est la suivante :

$$T^2 = T_1^2 + T_{2\bullet 1}^2 + T_{3\bullet 1,2}^2 + T_{4\bullet 1,2,3}^2 + \cdots + T_{p\bullet 1,2,3\cdots p-1}^2 \tag{1.7}$$

où $T_{i\bullet j,k}^2$ représente la statistique T^2 de la régression des variables variables X_j et X_k sur la variable X_i. On voit qu'il existe un nombre important de décompositions différentes ($p!$), et donc qu'il existe un grand nombre de facteur ($p \times 2^{p-1}$) différents. Pour mieux comprendre, sur un procédé de 3 variables, nous obtenons les différentes décompositions suivantes :

$$\begin{aligned}
T^2 &= T_1^2 + T_{2\bullet 1}^2 + T_{3\bullet 1,2}^2 \\
T^2 &= T_1^2 + T_{3\bullet 1}^2 + T_{2\bullet 1,3}^2 \\
T^2 &= T_2^2 + T_{1\bullet 2}^2 + T_{3\bullet 1,2}^2 \\
T^2 &= T_2^2 + T_{3\bullet 2}^2 + T_{1\bullet 2,3}^2 \\
T^2 &= T_3^2 + T_{1\bullet 3}^2 + T_{2\bullet 1,3}^2 \\
T^2 &= T_3^2 + T_{2\bullet 3}^2 + T_{1\bullet 2,3}^2
\end{aligned} \tag{1.8}$$

Le calcul des termes n'est pas détaillé ici, mais on pourra bien entendu se reporter aux travaux de Mason et al. [98, 97]. Il est à noter que les termes T_j^2 sont appelés facteurs non-conditionnés (puisqu'ils ne dépendent pas du tout des autres variables que j), alors que les autres termes sont appelés facteurs conditionnés. Ce qui est intéressant, c'est que chaque facteur suit une distribution de Fisher (à une constante près) :

$$T_{j+1\bullet 1,\cdots j}^2 = \frac{(m+1)(m-1)}{m(m-k-1)} F_{1,m-k-1} \tag{1.9}$$

où k est le nombre de facteurs conditionnés. On pourra donc simplifier cette équation

pour les termes non-conditionnés (k=0) par :

$$T^2_{j+1\bullet 1,\cdots j} \sim \frac{m+1}{m} F_{1,m-1} \tag{1.10}$$

Cela nous permet de détecter un problème sur chacun des facteurs de la décomposition. Par exemple, si l'on s'aperçoit que le facteur $T^2_{2\bullet 1}$ est responsable d'un hors contrôle du procédé, on peut immédiatement aller chercher la cause de l'anomalie sur un réglage physique affectant la corrélation entre ces deux variables. Mais, pour moins de calcul, il suffit d'utiliser une carte T^2 pour la détection de situation hors-contrôle, et si une erreur se produit, alors on utilise la méthode MYT pour déterminer d'où vient l'erreur. L'analyse des facteurs se fait dans l'ordre de niveau (ex : T^2_1, T^2_2, T^2_3, puis $T^2_{2\bullet 1}$, $T^2_{3\bullet 1}$, $T^2_{1\bullet 2}$, $T^2_{3\bullet 2}$, $T^2_{1\bullet 3}$, $T^2_{2\bullet 3}$ puis finalement $T^2_{3\bullet 1,2}$, $T^2_{2\bullet 1,3}$, $T^2_{1\bullet 2,3}$) jusqu'à ce qu'on trouve le facteur ayant causé la détection d'une erreur sur la carte T^2.

L'avantage de la méthode MYT est qu'elle fournie un diagnostic d'une situation hors-contrôle, sans avoir à la comparer à des exemples de fautes préalablement apparues dans le procédé. Ainsi, cette méthode de diagnostic est une méthode non-supervisée. De plus, un autre avantage de cette méthode est qu'elle est basée sur la même démarche que la carte T^2. Les outils statistiques sont les mêmes et on peut penser qu'une implémentation pratique est beaucoup plus compréhensible qu'un mélange de plusieurs techniques.

1.4.2.4 Les autres cartes multivariées

La carte T^2 n'est pas la seule carte de contrôle multivariée. En effet, deux autres cartes sont très connues : la carte MEWMA et la carte MCUSUM. Le diagnostic des fautes détectées par ces cartes peut également être effectué par la décomposition MYT.

La carte MEWMA La carte MEWMA développée par Lowry [93] est l'extension multivariée de la carte univariée EWMA proposée par Roberts en 1959 [123]. On peut traduire carte MEWMA par "carte multivariée pour moyennes mobiles à pondération exponentielle". Cette carte est très adaptée pour le suivi de valeurs individuelles (taille d'échantillon $n = 1$), mais elle est surtout utile pour la détection d'écarts de faibles amplitudes par rapport à la cible. Cette carte est largement employée, mais pas toujours sous cette appellation. En effet, les automaticiens parlent de modèle AR (auto régressif), alors que les électroniciens parle de filtre à réponse impulsionnelle infinie (filtre Rii).

Pour la carte MEWMA [93], il faut calculer de manière récursive pour chaque échan-

tillon la variable $\boldsymbol{y}_t$ (voir équation 1.11), où l'initialisation est faite par $\boldsymbol{y}_0 = \boldsymbol{\mu}$.

$$\boldsymbol{y}_t = \boldsymbol{\lambda}\boldsymbol{x}_t + (\boldsymbol{I} - \boldsymbol{\lambda})\boldsymbol{y}_{t-1} \tag{1.11}$$

Dans l'équation 1.11, $\boldsymbol{x}_t$ est le vecteur observation à l'instant t, $\boldsymbol{I}$ est la matrice identité et $\boldsymbol{\lambda}$ est une matrice diagonale de pondération dont les éléments $\lambda_1, \lambda_2, \ldots, \lambda_p$ seront compris entre 0 et 1 ($0 < \lambda_i < 1$).

Lowry [92, 93] propose d'utiliser $\lambda_i = \lambda$ pour $i = 1, 2, \ldots, p$ si il n'y pas de raison particulière de pondérer les variables différemment. Dans le cas inverse, la carte MEWMA est dite directionnelle puisqu'elle privilégie la détection dans certaines directions de l'espace, mais le traitement de la carte perd alors de sa simplicité. Ainsi, dans la suite de la thèse, nous considérerons le cas de la carte MEWMA non-directionnelle. On peut donc récrire l'équation 1.11 sous la forme de l'équation 1.12

$$\boldsymbol{y}_t = \lambda\boldsymbol{x}_t + (1 - \lambda)\boldsymbol{y}_{t-1} \tag{1.12}$$

Sur la carte de contrôle MEWMA, on trace la statistique suivante :

$$T_t^2 = \boldsymbol{y}_t^T \boldsymbol{\Sigma}_{\boldsymbol{y}_t}^{-1} \boldsymbol{y}_t \tag{1.13}$$

où $\boldsymbol{\Sigma}_{\boldsymbol{y}_t}$ est la matrice de variance-covariance de la variable $\boldsymbol{y}$ pour l'instant t. Elle est définie ainsi :

$$\boldsymbol{\Sigma}_{\boldsymbol{y}_t} = \left\{ \frac{\lambda[1 - (1 - \lambda)^{2t}]}{2 - \lambda} \right\} \boldsymbol{\Sigma} \tag{1.14}$$

Dans le cas où λ n'est pas trop petit ($\lambda > 0.1$), cette matrice approche très rapidement de sa valeur asymptotique [138]. Cette valeur est définie par :

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \left\{ \frac{\lambda}{2 - \lambda} \right\} \boldsymbol{\Sigma} \tag{1.15}$$

Le procédé est déclaré hors-contrôle dès que T_t^2 dépasse la limite de contrôle h_M. Cette limite de contrôle h_M est calculée en fonction de p et de λ, afin de respecter une certaine fréquence donnée de fausses alertes (voir l'article de Lowry et al. [93] pour les valeurs de h_M). Il faut préciser que les performances de la carte MEWMA sont fonction de λ. En effet, plus λ est faible, plus la carte est performante pour des sauts de faibles amplitudes, mais moins elle est performante pour des sauts de fortes amplitudes. Il faut donc choisir λ en fonction de l'amplitude du saut que l'on souhaite détecter. Il est également à signaler

qu'un cas particulier de la carte MEWMA est la carte avec $\lambda = 1$. En effet, dans ce cas, nous obtenons une carte du T^2 de Hotelling (voir paragraphe précédent). La carte MEWMA possède des performances supérieures à la carte T^2 pour la détection de faibles déréglages. Par contre, dans le cas d'un saut d'une forte amplitude, la carte T^2 détectera plus rapidement que la carte MEWMA, car cette dernière possède une inertie que ne possède pas la carte T^2. Plusieurs auteurs se sont penchés sur le bon choix de paramètres pour cette carte. On trouve des conceptions basées sur un choix statistique [121], ou bien sur un choix économico-statistique [103], ou finalement une conception permettant une certaine robustesse [139]. Ces différentes méthodes ont été comparées [138]. Les auteurs montrent que l'utilisation d'une matrice de variance-covariance asymptotique (équation 1.15), ainsi qu'un λ faible ($\approx 0.05 - 0.1$), permet de rendre la carte plus robuste à la non-normalité.

La carte MCUSUM Cette carte est l'extension multivariée de la carte univariée CUSUM développée par Page [111] en 1954. La carte MCUSUM (Multivariate CUmulated SUM) est une carte multivariée utilisant la somme cumulée. Son avantage principal est qu'elle permet, tout comme la carte MEWMA, de détecter plus rapidement qu'une carte T^2 des sauts de faibles amplitudes. Il existe beaucoup de versions de cette carte [28, 119, 129]. La carte la plus performante et la plus simple est celle développée par Crosier [28] : MC1. Le principe est détaillé ci-dessous, il faut commencer par calculer S_n par l'équation 1.16, où $\boldsymbol{S}_0$ est le vecteur nul, et où k est un paramètre fonction du déréglage.

$$C_t = (\boldsymbol{S}_{t-1} + \boldsymbol{X}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{S}_{t-1} + \boldsymbol{X}_t - \boldsymbol{\mu}) \tag{1.16}$$

On calcule ensuite la valeur $\boldsymbol{S}_t$ grâce à la formule suivante :

$$\boldsymbol{S}_t = \begin{cases} 0 & si\ C_t \leq k \\ (\boldsymbol{S}_{t-1} + \boldsymbol{X}_t - \boldsymbol{\mu})/(1 - \frac{k}{C_t}) & si\ C_t > k \end{cases} \tag{1.17}$$

Et finalement, on suit la statistique suivante :

$$MC1_t = \boldsymbol{S}_t^T \boldsymbol{\Sigma}^{-1} \boldsymbol{S}_t \tag{1.18}$$

Le procédé est déclaré hors contrôle dès que $MC1_t$ dépasse une certaine limite fixée [119].

1.4.3 Les approches par ACP et PSL

Cette partie présente un outil très utilisé dans la surveillance des procédés multivariés : l'Analyse en Composantes Principales (ACP). Dans un premier temps, les fondements de l'ACP sont exposés de manière théorique. Dans un second temps, l'application de l'ACP pour la surveillance des procédés est détaillée. Puis finalement, diverses extensions sont présentées. De plus, nous présentons un autre outil comparable à l'ACP : la Projection dans les Structures Latentes (PSL)

1.4.3.1 L'Analyse en Composantes Principales

L'Analyse en Composantes Principales [46, 83], plus souvent dénommée ACP (ou PCA pour Principal Component Analysis), est une technique de recherche d'axes principaux de l'ellipsoïde d'une distribution normale multivariée (voir exemple bivarié de la figure 1.15).

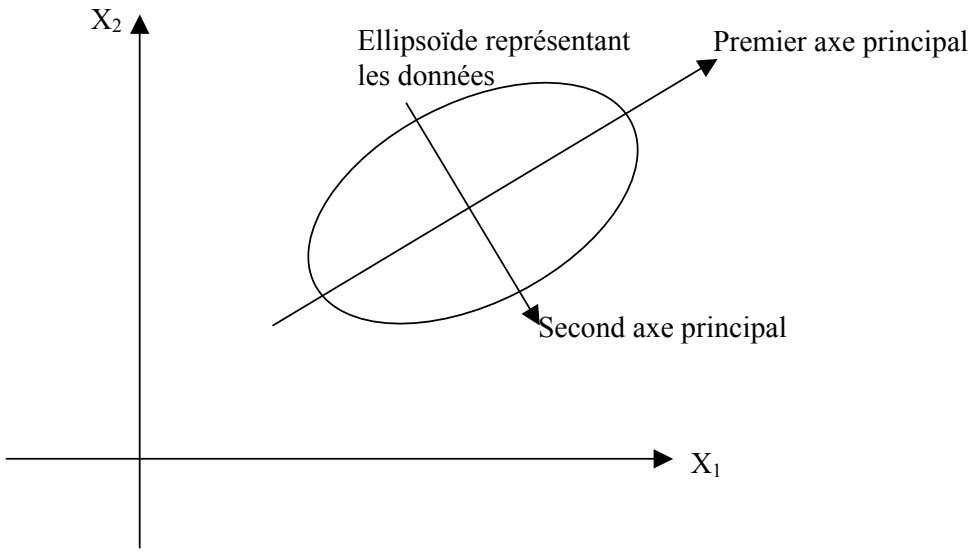


FIG. 1.15 – Exemple des composantes principales

En d'autres termes, on recherche les axes géométriques orthogonaux où la dispersion des données est maximale. L'intérêt de cette technique est qu'un nombre assez réduit de composantes principales permet généralement d'expliquer la quasi-totalité de la variabilité des données. L'ACP est donc une technique linéaire de réduction de dimension, se voulant optimal en terme d'explication (ou capture) de la variabilité d'un jeu de donnée. Les axes recherchés, que l'on nomme axes principaux ou composantes principales, sont ordonnés de l'axe capturant la plus grande variabilité à l'axe capturant le moins de variabilité. Si on appelle $\boldsymbol{X}$ l'ensemble des données (préalablement centrées), de taille $n \times p$ avec n le nombre d'individus de $\boldsymbol{X}$ et p le nombre de variables du procédé, les axes principaux se

calculent en résolvant le problème d'optimisation de l'équation 1.19 suivante.

$$\max_{\boldsymbol{v}\neq 0} \frac{\boldsymbol{v}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} \tag{1.19}$$

Ce problème d'optimisation se résout facilement, et on peut prouver [17] que les axes principaux sont les vecteurs propres $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ de la matrice de variance-covariance $\boldsymbol{\Sigma}$ de $\boldsymbol{X}$, classés en ordre décroissant de valeurs propres associées $(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0)$. On possède alors la matrice permettant le passage de l'espace initial vers l'espace décrit par les axes principaux $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_p]$. Il existe autant d'axes principaux que d'axes initiaux. Le passage d'un individu $\boldsymbol{x}$ dans ce nouvel espace se fait à partir de l'équation 1.20.

$$\boldsymbol{t} = \boldsymbol{x}\boldsymbol{V} \tag{1.20}$$

Nous obtenons alors le vecteur $\boldsymbol{t}$, de dimension $1 \times p$, représentant la projection des données $\boldsymbol{x}$ dans le nouvel espace : $\boldsymbol{t} = [t_1\ t_2\ \ldots\ t_p]^T$. Nous allons maintenant étudier les principes de détection concernant les données projetées dans cet espace.

1.4.3.2 Principes de détection par ACP

Détection dans l'espace réduit L'intérêt particulier d'utiliser une transformation des données par ACP est que les premiers axes de ce nouvel espace capturent la plus grande partie de la variabilité exprimée dans l'espace initial. Ainsi, il est usuel de pratiquer la détection d'une situation hors-contrôle sur les a premières composantes du nouvel espace, avec $a < p$. Dans ce cas, l'espace décrit par les a premières composantes principales est dénommé espace réduit. Le nombre de composantes principales choisies peut nous arranger pour une quelconque raison (les dimensions 2 et 3 sont très pratiques pour la visualisation graphique des données). Sinon, plusieurs méthodes existent afin de fixer correctement a : le test "scree" [12], l'analyse parallèle [78] ou la statistique PRESS [163] (pour ces méthodes, on peut également consulter [17]). Une autre méthode est le test de pourcentage de variabilité [64] dans lequel on se fixe a afin que les axes principaux représentent un certain pourcentage de variabilité pv des données initiales. En effet, $\sum_{i=1}^{p} \lambda_i$ représente la variabilité totale des données. Donc, on peut calculer pour chaque a fixé le pourcentage de variabilité que les a premiers axes principaux représentent, et ce par la formule 1.21.

$$pv_a = \frac{\sum_{i=1}^{a} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{1.21}$$

Nous allons maintenant étudier la détection dans cet espace réduit. Pour cela, nous décomposons tout d'abord la matrice de passage $\boldsymbol{V}$:

$$\boldsymbol{V} = [\boldsymbol{P}\ \boldsymbol{P'}] \tag{1.22}$$

Ainsi, nous obtenons deux matrices de passage. La matrice $\boldsymbol{P}$, de dimension $p \times a$ permet le passage d'un individu $\boldsymbol{x}$ vers les a premières composantes principales, alors que la matrice $\boldsymbol{P'}$, de dimension $(p-a) \times p$ permet le passage d'un individu $\boldsymbol{x}$ vers les $p-a$ dernières composantes de l'espace de l'ACP. Ainsi, on écrit que $\boldsymbol{P} = [\boldsymbol{v_1}\ \boldsymbol{v_2}\ \ldots\ \boldsymbol{v_a}]$, et que $\boldsymbol{P'} = [\boldsymbol{v_{a+1}}\ \boldsymbol{v_{a+2}}\ \ldots\ \boldsymbol{v_p}]$. Il est possible de calculer la projection d'un individu $\boldsymbol{x}$ dans l'espace réduit par :

$$\boldsymbol{t_a} = \boldsymbol{xP} \tag{1.23}$$

où $\boldsymbol{t_a}$ est un vecteur de dimension $1 \times a$ décrit par : $\boldsymbol{t_a} = [t_1\ t_2\ \ldots\ t_a]^T$.

Afin de détecter une faute, il est alors possible d'appliquer une carte multivariée du T^2 de Hotelling sur les observations $\boldsymbol{t_a}$ [63]. Ainsi, on obtient :

$$T^2 = \boldsymbol{t_a}^T \boldsymbol{\Sigma_a}^{-1} \boldsymbol{t_a} \tag{1.24}$$

où $\boldsymbol{\Sigma_a}$ est la matrice de variance-covariance des données dans l'espace réduit. Cette matrice, de dimension $a \times a$, est la matrice diagonale contenant les a plus grandes valeurs propres de la matrice de corrélation $\boldsymbol{R}$ de $\boldsymbol{X}$, soit :

$$\boldsymbol{\Sigma}_a = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_a \end{pmatrix} \tag{1.25}$$

L'équation 1.24 peut aussi s'écrire sous la forme de l'équation 1.26 suivante :

$$T^2 = \frac{t_1^2}{\lambda_1} + \frac{t_2^2}{\lambda_2} + \cdots + \frac{t_a^2}{\lambda_a} = \sum_{i=1}^{a} \frac{t_i^2}{\lambda_i} \tag{1.26}$$

Tout comme dans le cas de la surveillance des données originales avec la carte T^2, la détection d'une faute dans le procédé se fera sur la condition suivante : $T^2 > LC$.

Les mêmes limites de contrôle que dans le cas du traitement des données originales sont appliquées. On pourra donc se reporter à la table 1.2 pour les différentes limites. Du fait que la statistique T^2 ne prend en compte que les axes portant la plus grande variabilité du fonctionnement normal du procédé, le dépassement de la limite de contrôle implique obligatoirement un changement de fonctionnement dans le procédé.

Cependant, la surveillance de l'espace réduit n'est pas suffisante pour surveiller le procédé. En effet, la surveillance des composantes principales permet la détection d'un déréglage dans le procédé, sous l'hypothèse que celui-ci est correctement modélisé par le modèle de l'ACP. Cependant, il peut également se produire un changement de paramètres du procédé rendant alors le modèle ACP incapable de détecter un défaut. Ainsi, nous allons étudier une méthode permettant la détection d'un changement dans le modèle du procédé.

Détection dans l'espace résiduel Afin de détecter un problème affectant le système et ne pouvant pas être détecté par le modèle de l'ACP (projection dans l'espace réduit), un moyen simple est la surveillance des $p - a$ dernières composantes par une carte du T^2. Cependant, ces composantes exprimant très peu de variabilité, elles sont sujettes à beaucoup d'erreur de précision [17]. On préfère alors une méthode plus robuste : l'étude des résidus engendrés par les projections des données dans l'espace réduit [65]. En effet, si le modèle de l'ACP est valide, alors ces résidus sont très faibles. Ils expriment la partie du modèle ACP non pris en compte dans les a composantes principales. Le vecteur des résidus $\boldsymbol{e}$ de la projection de données dans l'espace réduit peut se calculer de deux façons. Tout d'abord, on peut le voir comme la différence entre l'observation $\boldsymbol{x}$ et son estimation $\hat{\boldsymbol{x}}$ après passage dans l'espace réduit. L'estimation de $\boldsymbol{x}$ est tout simplement le résultat de la projection des données de l'espace réduit vers l'espace initial : $\hat{\boldsymbol{x}} = \boldsymbol{t_a}\boldsymbol{P}^T$. Ainsi, nous obtenons la première écriture du vecteur résiduel $\boldsymbol{e}$:

$$\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{t_a}\boldsymbol{P}^T \tag{1.27}$$

On peut également utiliser la matrice $\boldsymbol{P'}$ (matrice permettant le passage d'un individu $\boldsymbol{x}$ vers les $p - a$ dernières composantes de l'espace de l'ACP). Le vecteur résiduel $\boldsymbol{e}$ s'exprime alors ainsi :

$$\boldsymbol{e} = \boldsymbol{x}\boldsymbol{P'}\boldsymbol{P'}^T \tag{1.28}$$

Pour la surveillance dans l'espace résiduel, on utilise l'indice SPE (Squared Prediction Error) [65]. Cet indice est également connu sous le nom de statistique Q. Cette statistique exploite les résidus $\boldsymbol{e}$ de l'ACP et se calcule par l'équation 1.29 suivante :

$$Q = \boldsymbol{e}\boldsymbol{e}^T \tag{1.29}$$

Ainsi, comme on le voit dans l'équation 1.29, Q est un scalaire. La distribution de la statistique Q a été approximée par Jackson et Mudholkar [65] et il est possible d'en calculer des quantiles par l'équation suivante :

$$Q_\alpha = \Theta_1 \left[\frac{h_0 c_\alpha \sqrt{2\Theta_2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{1/h_0} \tag{1.30}$$

où :

$$\Theta_i = \sum_{j=a+1}^{p} \lambda_j^i \tag{1.31}$$

avec λ_j la j^{eme} valeur propre de $\boldsymbol{\Sigma}$, et

$$h_0 = 1 - \frac{2\Theta_1\Theta_3}{3\Theta_2^2} \tag{1.32}$$

Le terme c_α représente la valeur de la loi normale centrée réduite au quantile $1 - \alpha$. En se fixant un α, il est possible de calculer Q_α et de l'utiliser comme limite de contrôle pour la statistique Q. Ainsi, on considére le procédé hors-contrôle (faute) si la condition suivante est vérifiée :

$$Q > Q_\alpha \tag{1.33}$$

La statistique Q permet de détecter des fautes que la statistique T^2 ne détecte pas, et inversement. En effet, comme nous l'avons déjà évoqué, la surveillance dans l'espace réduit n'a de sens que si le modèle ACP utilisé est encore valable à l'instant d'observation. Ainsi, si la statistique Q est supérieure à Q_α, alors une faute s'est produite dans le procédé, faute ayant entraînée un changement de modèle rendant l'espace réduit non-adéquat pour la détection. Cependant, si Q est inférieure à Q_α, alors le modèle de l'ACP est encore valide, et la détection dans l'espace réduit par une carte du T^2 est possible. Ainsi, il est usuel d'utiliser les deux mesures (T^2 et Q) pour la détection de fautes dans un procédé multivarié. Lorsque ces deux statistiques sont utilisées conjointement avec leur limite de contrôle respective, cela produit une région cylindrique de données sous contrôle. Cette région est illustrée pour un exemple en deux dimensions sur la figure 1.16.

Sur cette figure 1.16, les données "x" sont des données obtenues pendant le fonctionnement normal du procédé. Pour ces observations, le modèle ACP est valide et le procédé n'est soumis à aucune faute. Les données "+" représentent des observations violant la li-

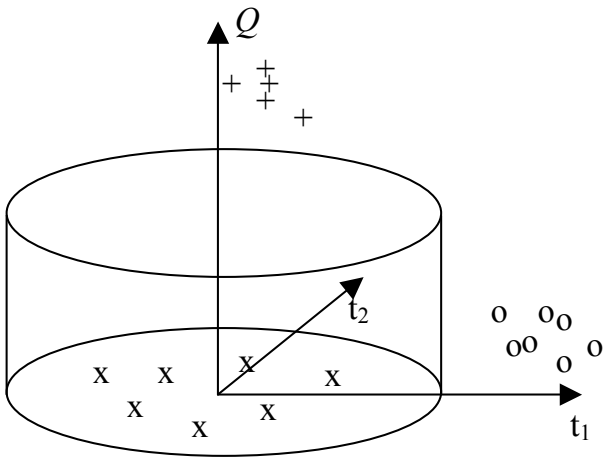


FIG. 1.16 – Illustration de la détection avec la combinaison des statistiques T^2 et Q

mite de contrôle de la statistique Q, signifiant alors qu'une faute présente dans le procédé a changé le modèle ACP de celui-ci. Enfin, les données "o" représentent des observations pour lesquelles le modèle ACP est valide (puisque $Q > Q_\alpha$), mais dont les projections dans l'espace réduit témoigne d'une faute présente dans le procédé.

Notons également que tous les indices de détection que nous avons vus peuvent également être adaptés dans le cas d'un filtrage des données projetées dans les différents espaces de l'ACP.

1.4.3.3 Principes de diagnostic par ACP

Diagnostic dans l'espace réduit Afin de diagnostiquer une faute détectée dans l'espace réduit, on utilise la méthode de contribution des variables originales aux composantes de l'espace réduit [76]. Cette méthode est appliquée en réponse à un dépassement du T^2. Les étapes décrites ci-dessous résument cette méthode.

1. Vérifier, parmi les différents scores normalisés $\frac{t_i^2}{\lambda_i}$ de l'observation $\boldsymbol{x}$ hors-contrôle, les r ($r \leq a$) scores responsables de la situation hors-contrôle. On rappelle qu'un score hors-contrôle se traduit par $\frac{t_i^2}{\lambda_i} > LC$.
2. Calculer la contribution de chaque variable originale X_j aux scores déclarés hors-contrôle t_i par l'équation 1.34 où $p_{i,j}$ est le $(i,j)^{\text{ième}}$ élément de la matrice $\boldsymbol{P}$ de passage dans l'espace réduit.

$$cont_{i,j} = \frac{t_i}{\lambda_i} p_{i,j}(x_j - \mu_j) \tag{1.34}$$

3. Si la contribution $cont_{i,j}$ est négative, alors elle est considérée comme nulle.

4. Calculer la contribution totale de la $j^{\text{ième}}$ variable originale X_j par :

$$CONT_j = \sum_{i=1}^{r} (cont_{i,j}) \tag{1.35}$$

5. Représenter $CONT_j$ pour les p variables originales du procédé.

Les variables possédant une forte contribution totale à la situation hors-contrôle sont suspectées d'être responsables de celle-ci.

Diagnostic dans l'espace résiduel Une autre méthode, développée par Wise et al. [161], permet l'identification des variables responsables d'une situation hors-contrôle, mais cette fois dans l'espace résiduel. Cette méthode est basée sur la quantification de la variation totale de chaque variable originale dans l'espace résiduel. En supposant égales les $p - a$ plus faibles valeurs propres de la matrice de variance-covariance $\mathbf{\Sigma}$, la variance σ^2_{Rj} de chaque variable X_j dans l'espace résiduel peut être estimée par :

$$s^2_{Rj} = \sum_{i=a+1}^{p} p_{i,j}\lambda_i \tag{1.36}$$

Etant données q nouvelles observations, il est possible de tester la variance σ^2_{qj} (estimée par s^2_{qj}) de la variable originale X_j durant ces q observations par le test d'hypothèse suivant :

$$\begin{aligned} H0 &: \sigma^2_{qj} = \sigma^2_{Rj} \\ H1 &: \sigma^2_{qj} \neq \sigma^2_{Rj} \end{aligned} \tag{1.37}$$

Ce test se traduit alors par l'acceptation de $H0$ si les inéquations suivantes sont respectées :

$$\frac{s^2_{qj}}{s^2_{Rj}} < F_{\alpha/2}(q - a - 1, n - a - 1) \tag{1.38}$$

et

$$\frac{s^2_{Rj}}{s^2_{qj}} < F_{\alpha/2}(n - a - 1, q - a - 1) \tag{1.39}$$

où $F_{\alpha/2}(n - a - 1, q - a - 1)$ est un quantile de la distribution de Fisher. De la même façon, on peut réaliser un test d'hypothèse sur les moyennes avec :

$$\begin{aligned} H0 &: \mu_{qj} = \mu_{Rj} \\ H1 &: \mu_{qj} \neq \mu_{Rj} \end{aligned} \tag{1.40}$$

où μ_{Rj} et μ_{qj} représente la moyenne de la variable X_j respectivement pour l'espace résiduel et pour les échantillons de q observations. On accepte l'hypothèse $H0$ si :

$$-t_{\alpha/2}(q+n-2a-2) > \frac{\overline{X}_{qj} - \overline{X}_{Rj}}{s_{Rj}^2\sqrt{\frac{1}{q-a}+\frac{1}{n-a}}} > t_{\alpha/2}(q+n-2a-2) \tag{1.41}$$

où : $\overline{X}_{Rj}$ et $\overline{X}_{qj}$ sont les estimations respectives de μ_{Rj} et μ_{qj}, et $t_{\alpha/2}(q+n-2a-2)$ est un quantile de la distribution de Student.

L'application des deux tests d'hypothèse précédents permettent d'identifier les variables responsables d'une situation hors-contrôle détectée dans l'espace résiduel (en d'autres termes, une situation pour laquelle s'est produite un dépassement de la statistique Q). Cependant, cette méthode implique la prise en compte de plusieurs observations du procédé (q observations). Elle est donc efficace, mais soumise à une certaine inertie lors de déréglage important et soudain. Ainsi, une dernière méthode permet de ne prendre en compte que l'instant t d'observation. Pour cela, on calcule pour chaque variable originale X_j l'erreur normalisée suivante :

$$RES_j = \frac{e_j}{s_{Rj}^2} \tag{1.42}$$

où e_j est la $j^{\text{ième}}$ variable du vecteur résiduel $\boldsymbol{e}$. Ainsi, les variables originales X_j possédant les plus grandes valeurs d'erreur normalisée sont les variables identifiées dans l'apparition de la situation hors-contrôle, permettant alors le diagnostic de la faute.

1.4.3.4 Extensions de l'approche par ACP

Comme nous l'avons vu à la section 1.4.2, les cartes de contrôle EWMA et CUSUM ont été généralisées pour des données multivariées. Ces généralisations (MEWMA et MCUSUM) peuvent être appliquées sur les indices de détection du T^2 impliqués par un modèle d'ACP [17]. L'application de ces méthodes permet d'accroître la sensibilité et la robustesse de la surveillance du procédé. Bien entendu, les remarques générales concernant ces cartes restent valables pour leur application dans une méthode d'ACP. A savoir, ces cartes permettent de détecter des sauts de faibles amplitudes, mais pour des sauts de fortes amplitudes, la détection est alors soumise à un retard dû à l'inertie induite par la prise en

compte des valeurs passées du procédé.

Une autre extension possible de l'ACP est la prise en compte de la production par lots (procédé batch). En effet, l'ACP présentée dans la section précédente fait la supposition que le procédé est strictement continu. Or, dans l'industrie, il est fréquent de trouver des procédés de fabrication fonctionnant par lots. La technique la plus étudiée pour traiter ce genre de problème est l'ACP multiéchelle (multiway PCA) [110, 162]. L'ACP multiéchelle est une extension à trois dimensions de l'ACP classique. Les trois dimensions représentent respectivement les observations, les instants d'observations, et les lots (le raisonnement pour l'ACP classique n'est fait que sur deux dimensions : les observations et les instants d'observations).

L'ACP classique est une transformation linéaire d'un espace initial des données vers un espace représenté par les composantes principales des données. Mais, il se peut qu'il existe des non-linéarités dans les données. Dans ce cas, une technique similaire de transformation, mais non-linéaire, est alors plus performante sur des procédés non-linéaires. Plusieurs approches peuvent être trouvées pour répondre à ce problème. Kramer [77] propose une généralisation de l'ACP classique (linéaire) vers le cas non-linéaire par l'utilisation de réseaux de neurones autoassociatifs. Dong et McAvoy [38] ont également proposé une approche basée sur des réseaux de neurones permettant de produire des composantes principales indépendantes. Ils ont démontré que, dans le cas de certaines non-linéarités, les approches d'ACP non-linéaires par réseaux de neurones permettent de capturer plus de variabilité sur moins de composantes que l'ACP classique.

1.4.3.5 L'approche par PSL

Une autre technique de réduction de dimension est la méthode de Projection dans les Structures Latentes (PSL) [75, 95], également connue sous le nom de méthode des Moindres Carrés Partiels (MCP). Cette technique permet la maximisation de la covariance entre une matrice de prédicteurs $\boldsymbol{X}$ et une matrice prédite $\boldsymbol{Y}$. L'application type de l'utilisation de PSL est de définir la matrice $\boldsymbol{Y}$ contenant uniquement les caractéristiques qualité de la production, alors que la matrice $\boldsymbol{X}$ contient toutes les autres variables du procédé. L'objectif est alors de trouver les espaces de projection pour $\boldsymbol{X}$ et $\boldsymbol{Y}$ pour lesquels la corrélation entre les vecteurs directeurs de chaque espace est la plus importante. Il existe de nombreux algorithmes de calcul des espaces réduits pour la PSL [51, 59, 67, 75, 162], nous allons présenter ici l'algorithme de base dénommé NIPALS (Non Iterative Partial Least Square) [51]. Le but de cet algorithme est le calcul des vecteurs directeurs des nouveaux espaces, ainsi que les projections des différentes données sur celui-ci.

Nous supposons ici que la matrice $\boldsymbol{X}$ est de dimension $n \times p$ où n est le nombre

d'observations, et p est le nombre de variables du procédé. De même, nous supposons que $\boldsymbol{Y}$ est de dimension $n \times m$ où m est le nombre de caractéristiques qualité. Comme dans le cas de l'ACP, il est possible d'écrire :

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} = \sum_{i=1}^{a} \boldsymbol{t}_j \boldsymbol{p}_j^T + \boldsymbol{E} \tag{1.43}$$

où $\boldsymbol{T}$ est la matrice de dimension $n \times a$ des projections dans l'espace réduit de $\boldsymbol{X}$, $\boldsymbol{P}$ est la matrice de dimension $a \times a$ des vecteurs directeurs du nouvel espace, et $\boldsymbol{E}$ est la matrice résiduelle de dimension $n \times p$ de la projection de $\boldsymbol{X}$ dans le nouvel espace.

De même, pour $\boldsymbol{Y}$ il est possible d'écrire :

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F} = \sum_{i=1}^{a} \boldsymbol{u}_j \boldsymbol{q}_j^T + \boldsymbol{F} \tag{1.44}$$

où $\boldsymbol{U}$ est la matrice de dimension $n \times a$ des projections dans l'espace réduit de $\boldsymbol{Y}$, $\boldsymbol{Q}$ est la matrice de dimension $a \times a$ des vecteurs directeurs du nouvel espace, et $\boldsymbol{F}$ est la matrice résiduelle de dimension $n \times p$ de la projection de $\boldsymbol{Y}$ dans le nouvel espace.

L'algorithme employé permet de déterminer les valeurs des matrices $\boldsymbol{T}$, $\boldsymbol{P}$, $\boldsymbol{U}$ et $\boldsymbol{Q}$.

Les auteurs précisent que l'application de l'algorithme proposé ci-dessous (de même que pour l'ACP dans la section précédente) fait l'hypothèse que les données ont été préalablement centrées et réduites.

Les différents vecteurs directeurs ainsi que les différentes valeurs des projections dans les espaces réduits se calculent de manière itérative. Pour cela, nous calculons les différentes matrices résiduelles à chaque itération, nommées $\boldsymbol{E}_j$ et $\boldsymbol{F}_j$. Avant la première itération, aucune donnée n'est encore représentée dans les nouveaux espaces, et donc nous initialisons $\boldsymbol{E}_0 = \boldsymbol{X}$ et $\boldsymbol{F}_0 = \boldsymbol{Y}$.

La première phase de calcul permet de déterminer $\boldsymbol{t}_j$ et $\boldsymbol{u}_j$. Ces vecteurs sont respectivement les vecteurs propres associés aux plus grandes valeurs propres des matrices $\boldsymbol{E}\boldsymbol{E}^T\boldsymbol{F}\boldsymbol{F}^T$ et $\boldsymbol{F}\boldsymbol{F}^T\boldsymbol{E}\boldsymbol{E}^T$.

La seconde phase permet alors de calculer $\boldsymbol{p}_j$ par l'équation suivante :

$$\boldsymbol{p}_j = \frac{\boldsymbol{E}_{j-1}^T \boldsymbol{t}_j}{\boldsymbol{t}_j^T \boldsymbol{t}_j} \tag{1.45}$$

ainsi que $\boldsymbol{q}_j$:

$$\boldsymbol{q}_j = \frac{\boldsymbol{F}_{j-1}^T \boldsymbol{t}_j}{||\boldsymbol{F}_{j-1}^T \boldsymbol{t}_j||} \tag{1.46}$$

Une fois $\boldsymbol{u}_j$ et $\boldsymbol{t}_j$ calculés, le coefficient de régression b_j reliant ces deux vecteurs est donné par :

$$b_j = \frac{\boldsymbol{u}_j^T \boldsymbol{t}_j}{\boldsymbol{t}_j^T \boldsymbol{t}_j} \tag{1.47}$$

Les valeurs des nouvelles matrices résiduelles sont calculées par les équations suivantes :

$$\boldsymbol{E}_j = \boldsymbol{E}_{j-1} - \boldsymbol{t}_j \boldsymbol{p}_j^T \tag{1.48}$$

$$\boldsymbol{F}_j = \boldsymbol{F}_{j-1} - b_j \boldsymbol{t}_j \boldsymbol{q}_j^T \tag{1.49}$$

La surveillance du procédé se fait grâce à l'application des cartes T^2 et Q dans les nouveaux espaces [75]. L'avantage de la PSL est que la surveillance se focalise davantage sur des variables qui sont reliées à la qualité finale du produit (composant la matrice $\boldsymbol{Y}$). Il est possible d'obtenir une estimation des caractéristiques qualités en temps-réel, et ce, sans attendre de mesurer concrètement celles-ci. Par exemple, il n'est pas judicieux d'attendre les mesures hors-ligne des différentes concentrations d'un produit chimique pour rectifier le procédé de fabrication défaillant. Donc, la PSL va permettre un gain de temps précieux sur ce type de procédé, où les caractéristiques qualités ne peuvent pas être mesurées directement en ligne.

1.5 Méthodes de classification supervisée pour la détection et le diagnostic

L'objectif de cette partie est de présenter les différents types de classifieurs utilisables pour la détection (classification supervisée à 2 classes) et le diagnostic des systèmes (classification supervisée à k classes). Cependant, il n'existe pas de classifieurs meilleurs que d'autres sur toutes les applications [44]. Il est donc utile de connaître les différents classifieurs utilisables, ainsi que leurs avantages et leurs inconvénients. Nous rappelons tout d'abord le contexte de la classification supervisée, puis nous présentons alors les différents classifieurs.

1.5.1 Classification supervisée

De nos jours, les procédés étant de plus en plus automatisés, ils nous fournissent de plus en plus de données, principalement récupérées par les capteurs. Beaucoup de données

sont récupérées lorsque le procédé est en fonctionnement normal, mais également lorsque le procédé subi une défaillance (ou faute). Lorsque ces défaillances ont été diagnostiquées (la cause de la défaillance a été identifiée), on peut catégoriser les données récoltées suivant les différentes causes associées aux dysfonctionnements. Lorsque les différentes fautes n'ont pas été diagnostiquées, on peut tout de même réaliser une catégorisation par recherche de classes (cluster analysis). Lorsque nous représentons graphiquement les données des différentes fautes, on peut alors chercher à dresser au mieux des frontières entre les différentes classes afin de définir les régions de chaque faute comme illustré sur la figure 1.17.

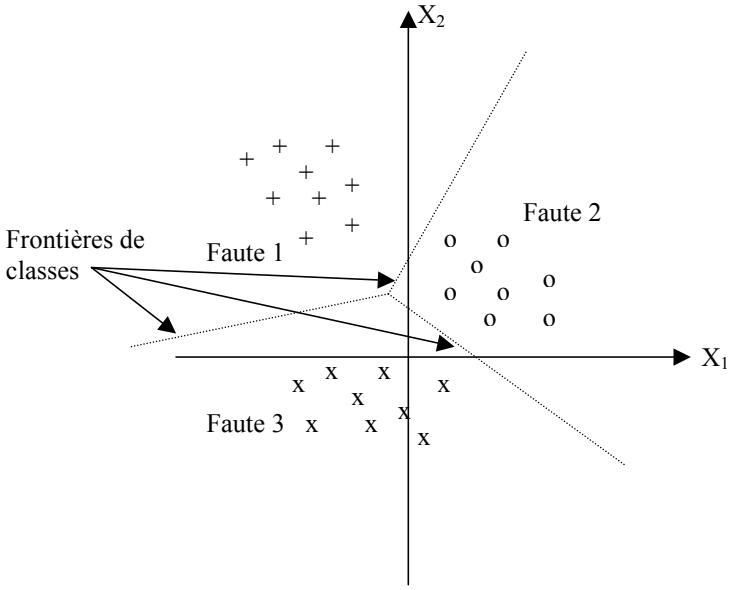


FIG. 1.17 – Exemple de frontières de classes

Lors de l'apparition d'une nouvelle faute (supposée détectée), en la représentant graphiquement, on voit tout de suite à quelle région de faute elle appartient et ainsi diagnostiquer cette nouvelle observation hors-contrôle. L'attribution d'une classe à une nouvelle observation est l'un des buts de la reconnaissance de forme (ou classification). Le système type de reconnaissance de forme se décompose en trois parties : l'extraction de composantes, l'analyse discriminante (ou calcul des coûts) et la sélection [18]. On peut voir le schéma de ce principe de reconnaissance de forme sur la figure 1.18.

L'objectif de l'extraction de composantes est d'accroître la robustesse du système de reconnaissance de forme (ou système de classification). Les fonctions d'extraction f_i permettent de réduire le nombre de dimensions du vecteur d'observations, de telle manière que les nouvelles composantes t_i retiennent l'information de discrimination entre chaque classe. En utilisant cet espace réduit, on calcule alors les coûts K_i d'appartenance du vecteur observation à chaque classe. Chaque coût est calculé grâce à une fonction dis-

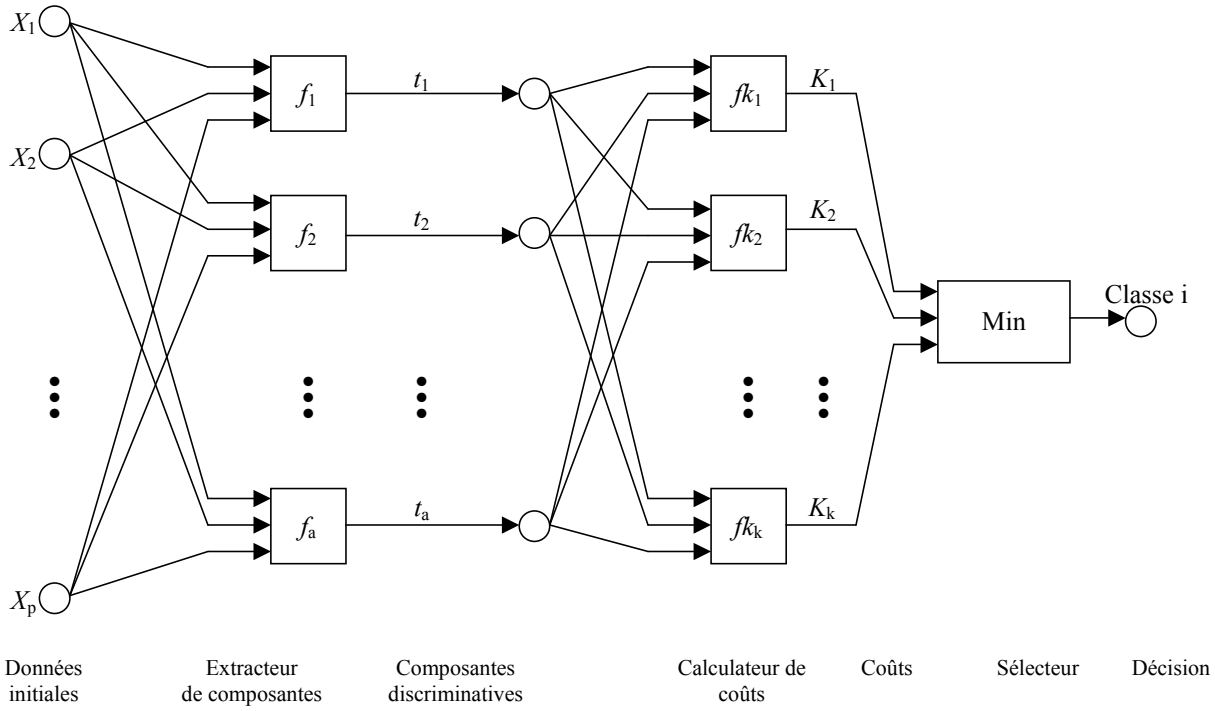


FIG. 1.18 – Système type de reconnaissance de forme

criminante fk_i (ou calculateur de coût). Ce sont ces fonctions discriminantes qui sont à l'origine des différentes frontières de séparation des classes que nous pouvons voir sur la figure 1.17 (les frontières ne sont pas forcément linéaires comme illustré dans l'exemple). Enfin, un sélecteur prend une décision en attribuant au vecteur observation la classe C_i ayant obtenu le plus faible coût K_i. Ce système permet donc de classer les nouvelles observations d'un système.

Il est à noter que la phase de sélection de composantes n'est pas forcément obligatoire. En effet, il est possible de calculer les coûts d'appartenance à chaque classe directement à partir des composantes initiales. Mais, bien souvent, les résultats sont plutôt médiocres et il vaut mieux effectuer cette phase. Même si les composantes discriminantes sont des composantes initiales, le fait de n'en garder que quelques unes permet d'ôter beaucoup de bruit sur l'information discriminative contenue dans l'espace initial.

Bien entendu, à la fois pour déterminer les fonctions d'extraction, ainsi que pour déterminer les fonctions coûts, il faut que l'on ait préalablement des informations sur notre système. En d'autres termes, il faut que l'on ait des exemples des différentes classes du système. Mais, comment savoir que le classifieur va être performant ? L'évaluation de celui-ci est donc essentielle. Mais, voyons tout d'abord ce que veut dire un classifieur performant. Un classifieur performant doit optimiser un ou plusieurs critères de performance. C'est le calcul de ce ou de ces critères qui sont réalisés lors d'une évaluation. On peut trou-

ver beaucoup de critères de performance pour un classifieur : sa complexité, son temps d'apprentissage, le nombre de paramètres à estimer, le temps d'inférence face à un nouvel individu, son taux de mauvaise classification à l'usage, etc. Bien entendu, en dehors de toutes considérations monétaires et/ou temporelles, il paraît juste de citer comme critère de performance primordial : le taux de mauvaises classifications à l'usage. Ce que l'on nomme taux de mauvaises classifications à l'usage est le rapport du nombre de cas futurs qui vont être mal classés sur le nombre total de cas futurs. Le problème important est que nous ne connaissons pas les cas futurs, mais on peut faire l'hypothèse que l'ensemble des cas futurs sera "globalement similaire" à l'ensemble des cas antérieurs (servants pour l'apprentissage). Maintenant que nous avons défini notre critère de performance, il faut que nous possédions une technique d'évaluation de notre classifieur suivant ce critère de performance. Incontestablement, la technique la plus employée pour cela est la validation croisée [25].

La validation croisée divise aléatoirement l'ensemble d'apprentissage initial en m sous-ensembles de même taille $\frac{m}{n}$ (où n est le nombre total d'individus dans l'ensemble d'apprentissage initial). On désigne alors la première partie comme ensemble de test, et l'union des autres parties comme ensemble d'apprentissage. Une fois le classifieur appris, on calcule le taux de mauvaises classifications de l'ensemble de test. Il faut alors réitérer cette procédure m fois (chaque partie va itérativement servir d'ensemble de test). A la fin des m passages, la moyenne des m taux de mauvaises classifications permet de chiffrer la performance du classifieur. Il est à noter que la validation croisée à 10 parties semble être la plus répandue dans la littérature. Cependant, une validation très efficace est celle prenant $m = n$. Bien entendu, cette méthode est très coûteuse en temps de calcul puisque nous devons faire apprendre n fois notre classifieur. Mais, son avantage est qu'elle ne dépend pas du découpage aléatoire de l'ensemble d'apprentissage initial (contrairement à la validation croisée à m parties). Un autre avantage de cette méthode est que chaque apprentissage du classifieur est très proche de l'apprentissage réel du classifieur puisqu'un seul individu est enlevé de l'ensemble d'apprentissage initial.

1.5.2 Généralisation d'un classifieur

Comme nous l'avons déjà vu, l'intérêt d'un classifieur est qu'il donne un taux de mauvaises classifications à l'usage le plus petit possible. C'est en ce sens qu'intervient la notion de généralisation. La généralisation est la capacité à généraliser ses régions de classification, afin d'obtenir un classifieur qui pourra classer correctement des individus futurs. Prenons l'exemple de la figure 1.19, on peut voir les données en 2 dimensions de 2

différentes classes (classe ”carré” et classe ”rond”).

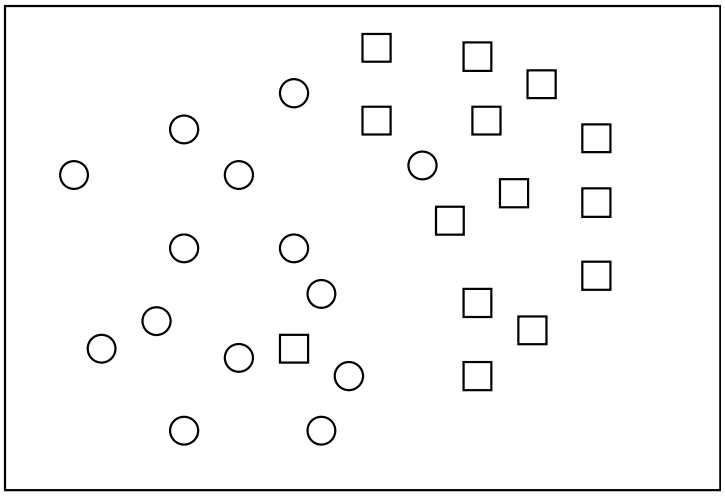


FIG. 1.19 – Représentation de 2 classes en 2 dimensions

Il est toujours possible de trouver un classifieur (notamment avec les réseaux de neurones [41] ou les séparateurs à vaste marge [144]) fournissant une frontière de décision très complexe. Grâce à cette frontière, nous pouvons classer parfaitement chaque individu de l'ensemble d'apprentissage comme sur la figure 1.20.

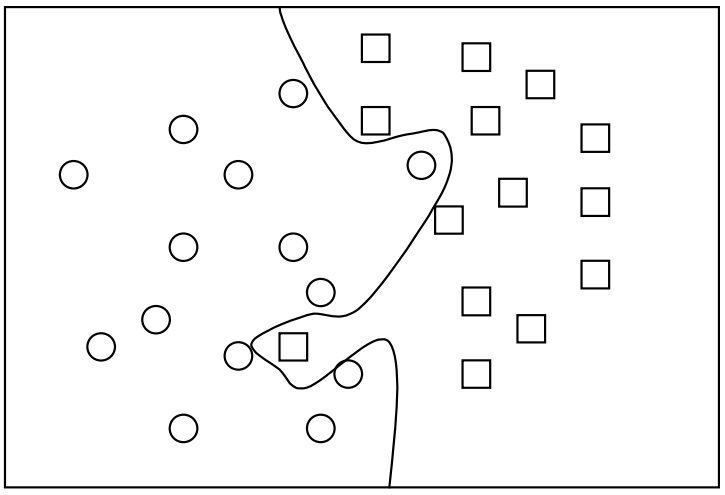


FIG. 1.20 – Frontière de décision parfaite sur données d'apprentissage

Maintenant, si nous soumettons à notre classifieur des individus qu'il n'a jamais vu, comment va-t-il réagir ? Sur la figure 1.21, on peut voir deux nouveaux individus représentés en pointillé. Le classifieur classe alors le rond comme étant un carré, et le carré comme étant un rond.

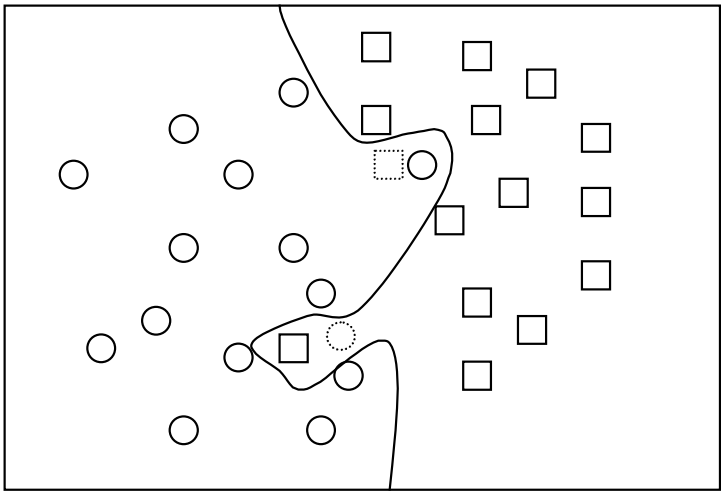


FIG. 1.21 – Nouveaux individus mal classés

Dans le cas de cette figure 1.21, le classifieur a perdu sa capacité de généralisation, il y a eu surapprentissage : il a trop appris l'ensemble d'apprentissage et ne sait reconnaître que celui-ci sans pouvoir le généraliser. Contrairement à la figure 1.21, la figure 1.22 présente un classifieur qui a généralisé son ensemble d'apprentissage. Bien entendu, il conduit lui aussi à quelques erreurs, mais à un taux plus faible que le même classifieur ayant surappris [44].

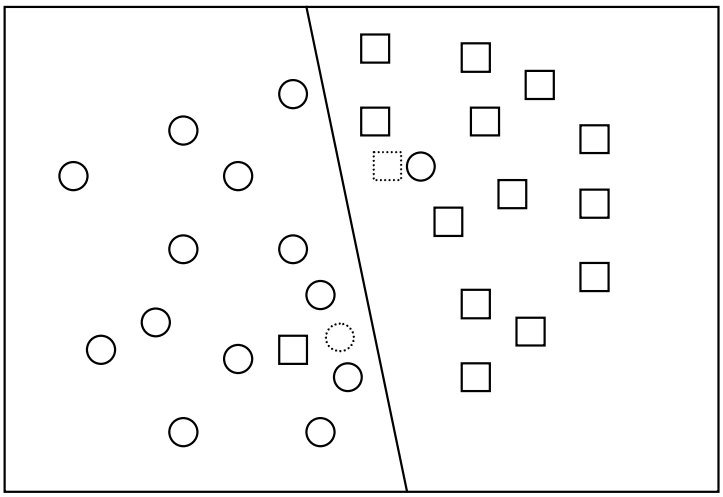


FIG. 1.22 – Nouveaux individus correctement classés

Maintenant que nous avons étudié ce qu'était la classification supervisée, nous allons voir les différentes méthodes (appelées classifieurs) permettant de réaliser une telle tâche. Il n'y a pas de classifieurs meilleurs que d'autres sur toutes les applications [44]. Il est donc utile de connaître les différents classifieurs utilisables, ainsi que leurs avantages et

leurs inconvénients. Bien que la liste des classifieurs présentés ici ne soit pas exhaustive, elle comporte tout de même la majorité des classifieurs les plus performants.

1.5.3 Les séparateurs à vaste marge

Les SVM (Support Vector Machines) ou Machines à Vecteurs Supports, ou bien encore Séparateurs à Vaste Marge, sont des outils modernes permettant la classification et la régression de données [144]. Nous étudions ici leur application à la classification supervisée. Les SVM sont des classifieurs binaires, ils ne peuvent différencier que deux classes d'individus. De plus, les variables descriptives du problème doivent être des variables continues.

Pour un jeu de données avec deux classes, le but d'un séparateur à vaste marge est de trouver un classifieur séparant les données et maximisant la distance entre ces deux classes. Ce classifieur linéaire est appelé hyperplan[1]. Dans la figure 1.23, on détermine un hyperplan séparant les deux ensembles de points. Les points les plus proches sont appelés vecteurs de support. Il est évident qu'il existe une multitude d'hyperplans valides mais la propriété remarquable des séparateurs à vaste marge est que cet hyperplan doit être optimal. Nous cherchons donc parmi les hyperplans valides, celui qui passe "au milieu" des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan "le plus sûr".

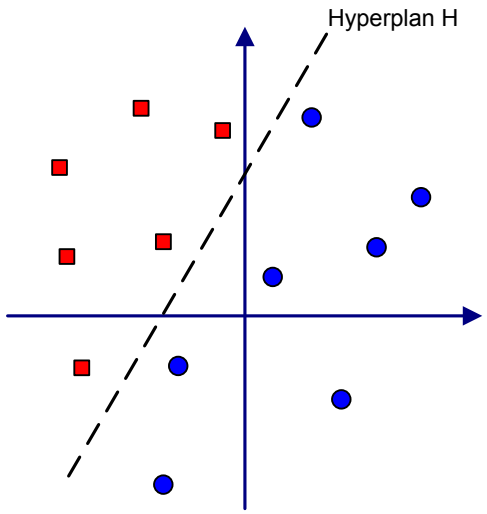


FIG. 1.23 – Séparation des données par l'hyperplan H

[1]dans un espace à une dimension, le séparateur linéaire sera un point. Dans un espace à deux dimensions, le séparateur sera une droite. Dans un espace à trois dimensions, le séparateur sera un plan. Dans un espace de dimension supérieur à 3, le séparateur sera nommé hyperplan. Mais pour plus de simplicité, peu importe la dimension de l'espace, nous appellerons le séparateur hyperplan.

L'hyperplan séparateur optimal est celui maximisant la marge, c'est pourquoi l'on parle de séparateurs à vaste marge [8]. Mais, on se rend bien compte que ce type de technique ne peut réellement traiter qu'un nombre de problèmes restreint puisqu'il faut que les classes puissent être séparables par un hyperplan. Or, dans un grand nombre de problèmes concrets, ce n'est pas le cas. Nous distinguons donc deux types de problèmes pour les séparateurs à vaste marge : les cas de classifications linéairement séparables, et les cas non-linéairement séparables (voir figure 1.24).

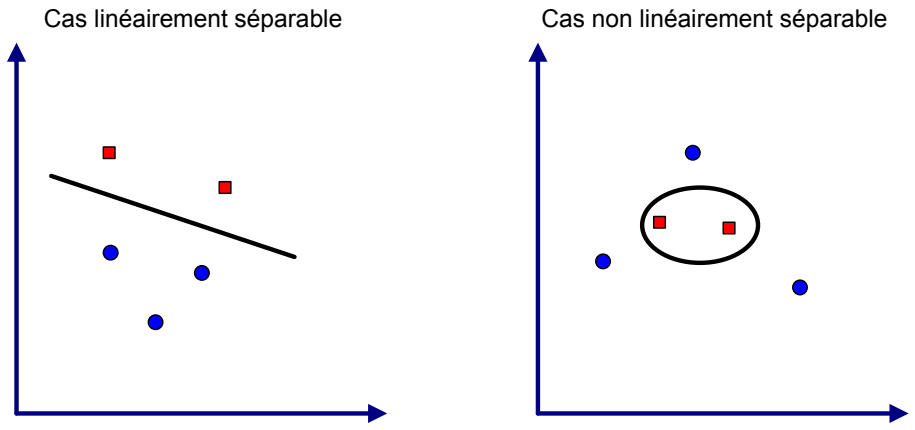


FIG. 1.24 – Illustrations de cas linéairement séparable et non-linéairement séparable

Comme nous l'avons vu, le cas linéairement séparable peut être résolu par un séparateur à vaste marge. Mais, pour le cas non-linéaire ce n'est pas possible. Afin de résoudre ce problème, il faut utiliser une fonction noyau permettant de projeter les données de l'espace initial (où les données sont non-linéairement séparables) vers un nouvel espace, généralement de dimension plus élevée, dans lequel les projections des données sont linéairement séparables. Prenons l'exemple des données de la figure 1.25. Nous avons deux classes de données représentées sur une dimension et qui ne sont pas linéairement séparables.

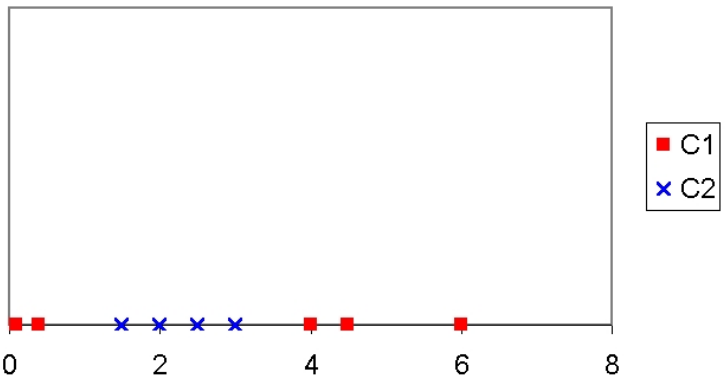


FIG. 1.25 – Exemple de 2 classes non-linéairement séparables

Transformons alors ces données vers un espace à deux dimensions où la première dimension est identique à l'espace initial, mais où la deuxième dimension est une application non-linéaire : $x \rightarrow x^2$. Nous obtenons alors la figure 1.26.

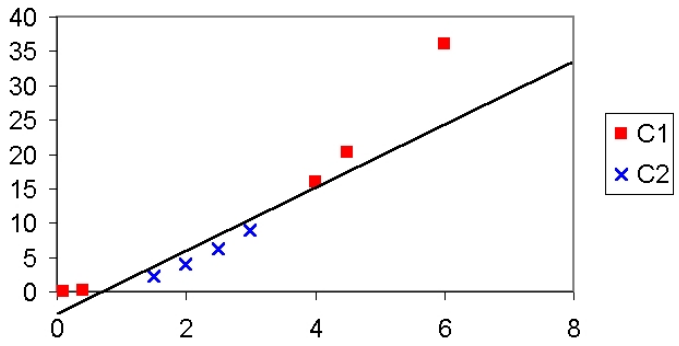


FIG. 1.26 – Même données dans un espace de dimension supérieur par application d'une transformation non-linéaire

Sur cette figure 1.26, on voit que la transformation non-linéaire dans un espace de dimension supérieure permet de rendre le problème linéairement séparable. En pratique, quelques familles de fonctions noyau paramétrables sont connues et il revient à l'utilisateur des séparateurs à vaste marge d'effectuer des tests pour déterminer celle qui convient le mieux pour son application. On peut citer les exemples de noyaux suivants : polynomiale, gaussien, sigmoïde et laplacien [44].

Un séparateur à vaste marge est une méthode de classification qui montre de bonnes performances dans la résolution de problèmes variés. Cette méthode a montré son efficacité dans de nombreux domaines d'applications tels que le traitement d'image, la catégorisation de textes ou le diagnostic médical, et ce même sur des ensembles de données de très grandes dimensions. La réalisation d'un programme d'apprentissage par séparateur à vaste marge se ramène à la résolution d'un problème d'optimisation impliquant un système dans un espace de dimension conséquente. L'utilisation de ces programmes revient à sélectionner une bonne famille de fonctions noyau et à régler les paramètres de ces fonctions. Ces choix sont le plus souvent effectués par une technique de validation croisée, dans laquelle on estime la performance du système en la mesurant sur des exemples n'ayant pas été utilisés en cours d'apprentissage. L'idée est de chercher les paramètres permettant d'obtenir la performance maximale. Si la mise en oeuvre d'un algorithme de séparateur à vaste marge est en général peu coûteuse en temps, il faut cependant compter que la recherche des meilleurs paramètres peut requérir des phases de test assez longues. Chiang et al. [18] ont utilisé les séparateurs à vaste marge afin de discriminer les différentes fautes d'un procédé chimique complexe. Mais sur les mêmes données, nous avons montré qu'une analyse discriminante quadratique pouvait obtenir de meilleurs résultats [150].

1.5.4 Les k plus proches voisins

La méthode des k plus proches voisins (k Nearest Neighborhood), ou kNN, est une technique de discrimination non-paramétrique [27], c'est à dire qu'aucune estimation de paramètres n'est nécessaire à son exécution. Cette technique de classification est plutôt ancienne puisqu'elle date d'environ 1950. Cette méthode s'emploie sur des données continues. Il est également possible de prendre en compte des données binaires (variable discrète à 2 modalités), mais pas multinomial (variable discrète avec plus de 2 modalités).

L'idée de cette méthode est d'observer les k plus proches voisins d'une nouvelle observation afin de décider de la classe d'appartenance de cette nouvelle observation [30]. Pour une nouvelle observation à classer, cet algorithme calcule la distance de cette nouvelle observation à chaque observation présente dans un ensemble d'apprentissage. On sélectionne les k voisins ayant la distance la plus faible avec la nouvelle observation. Au vu des classes d'appartenance des k plus proches voisins, on décide de la classe d'appartenance du nouvel individu. Généralement, on attribue la classe du nouvel individu comme étant la classe la plus représentée parmi ses k plus proches voisins. Pour illustrer cette règle, un exemple de classification de deux classes en deux dimensions est proposé sur la figure 1.27.

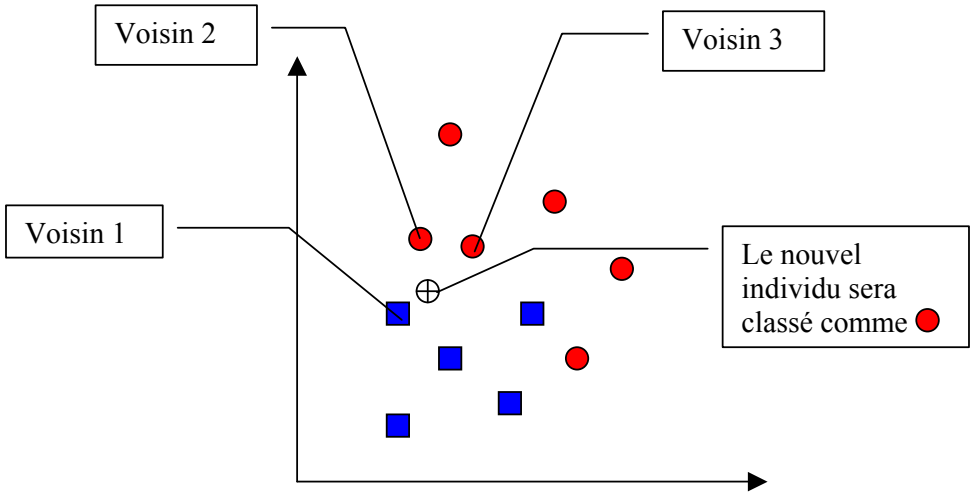


FIG. 1.27 – Exemple d'une attribution avec la règle des 3 plus proches voisins

La règle des k plus proches voisins exige une notification de la notion de voisin. En effet, nous entendons par observation voisine une observation dont la distance au nouvel individu est faible. Dès lors, il faut définir la notion de distance d'un individu à un autre. Généralement, on choisit comme distance la distance euclidienne, mais on peut en utiliser d'autres (distance tangente, distance de Manhattan) [44]. Mais, cela peut être source de problème. En effet, certaines variables peuvent complètement inhiber d'autres variables

lors du calcul de la distance euclidienne. Par exemple, une variable possédant une dispersion très élevée donne une contribution importante à la distance euclidienne, alors qu'une variable avec une dispersion très faible ne contribue presque pas au calcul de la distance. Ainsi, afin d'obtenir des résultats corrects, il est conseillé de toujours appliquer une réduction des données. Cette réduction permet à chaque variable de pouvoir contribuer équitablement à la distance euclidienne et ainsi intervenir dans la discrimination d'un nouvel individu.

Bien que cette approche soit non paramétrique (pas d'estimation de paramètre à partir des données), il reste un paramètre à fixer : le nombre k de plus proches voisins. Une heuristique fréquemment utilisée est de prendre k égal à la dimension de l'espace plus un. Des approches par validation croisée permettent également de tester le comportement du classifieur pour plusieurs valeurs de k, et de choisir ainsi la valeur la plus prometteuse [44].

Un des principaux problèmes de la classification par les k plus proches voisins vient du fait que pour chaque nouvel individu à classer, il faut calculer les distances de ce nouvel individu à chaque individu présent dans la base d'apprentissage. Ce mécanisme peut devenir extrêmement coûteux en calcul et très demandeur en terme de mémoire de stockage.

Pernkopf [117] a démontré sur plusieurs exemples que de simples classifieurs bayésiens permettaient souvent de dépasser les résultats obtenus avec les kNN. En effet, l'auteur prend plusieurs jeux de données (par exemple des données d'inspection de surface à 42 descripteurs) et teste différentes combinaisons de classifieurs à k plus proches voisins contre des classifieurs bayésiens : réseau bayésien naïf et réseau bayésien naïf augmenté par un arbre (voir §1.5.9). L'auteur montre que les classifieurs bayésiens surpassent très souvent les classifieurs à k plus proches voisins, et ce pour des capacités mémoire et des temps de calcul moindres.

1.5.5 Les arbres de décision

Un outil reconnu de discrimination entre plusieurs classes est l'arbre de décision [23]. L'intérêt principal des arbres de décision est qu'ils peuvent aisément se transformer sous forme de règles compréhensibles. Ainsi, le cheminement (la logique) amenant l'arbre à une décision est très clair pour l'utilisateur.

Comme son nom l'indique, un arbre de décision se représente graphiquement sous les traits d'une arborescence (voir figure 1.28). La lecture d'un arbre se fait du haut vers le bas. Dès que l'on croise un nœud, une décision est à prendre, représentée par un test

sur l'un des attributs du système. Pour chaque test, plusieurs décisions sont possibles. Si l'attribut est binaire, nous avons deux décisions possibles, alors que si l'attribut possède k modalités, nous avons k décisions possibles. L'arbre s'étoffe donc en fonction du nombre d'attributs du système, mais également en fonction du nombre de modalités pour chaque attribut. Les nœuds terminaux de l'arbre sont les feuilles de celui-ci, ils représentent la décision finale : la classe d'appartenance pour l'individu dont les observations ont servi aux différents tests de l'arbre.

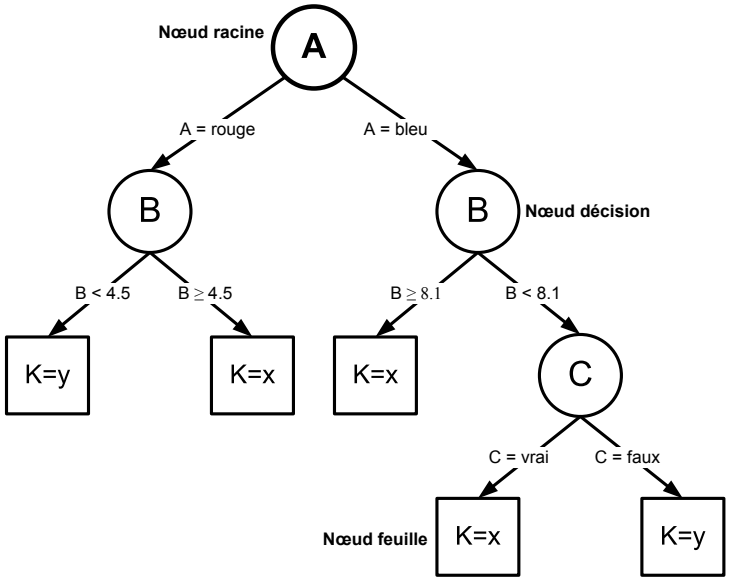


FIG. 1.28 – Exemple d'un arbre de décision

Un arbre particulièrement conséquent peut tout de même être très facilement et rapidement exploité. En effet, l'obtention de la solution n'implique pas l'exécution de tous les tests de l'arbre, mais un chemin parmi les branches de celui-ci jusqu'à une feuille. La rapidité d'exécution est donc une qualité des arbres de décision. Mais, il n'en est pas de même dans l'étape préliminaire : la construction de l'arbre. En effet, bien que l'exécution soit rapide, la construction de l'arbre est quant à elle beaucoup plus coûteuse en temps.

Il existe un grand nombre d'algorithmes pour la construction d'un arbre de décision [11], les principaux algorithmes sont CART (Classification And Regression Tree), ID3 et C4.5 [44]. Il est tout de même possible de distinguer l'algorithme classique de création de l'arbre [23] tel que donné ci-dessous.

Procédure : construire-arbre(X)

Si tous les points de X appartiennent à la même classe alors
    Créer une feuille portant le nom de cette classe
Sinon
    Choisir le meilleur attribut a pour créer un nœud
    Le test associé à ce nœud sépare X en deux parties : Xg et Xd
    Construire-arbre (Xg)
    Construire-arbre (Xd)
Finsi

En observant la procédure de construction d'un arbre, on constate qu'une étape pose problème : choisir le meilleur attribut pour créer un nœud. En effet, le but des algorithmes d'arbre de décision est de trouver l'ordre adéquat des décisions à prendre. En d'autres mots, quels attributs doivent être placés dans les premières décisions et quels autres doivent être placés vers le bout de l'arbre (les feuilles). Le but est donc de choisir en premier lieu l'attribut séparant au mieux les données dans l'espace entier d'apprentissage. Ceci équivaut à chercher l'attribut dont l'homogénéité est la plus faible. Afin de résoudre ce problème, les algorithmes cités (CART, ID3 et C4.5) se basent sur la notion d'entropie H [26].

Soient les données suivantes : n exemples, réparties en k classes C_k comportant chacune n_j exemples (avec $\sum_{j=1}^{k} n_j = n$), p attributs binaires notés a_i. Pour un attribut binaire a donné ($a = vrai$ ou $a = faux$), chaque sous-ensemble n_j est divisé en deux parties contenant respectivement v_j exemples où $a = vrai$ et f_j exemples où $a = faux$, avec :

$$v = \sum_{j=1}^{k} v_j \qquad f = \sum_{j=1}^{k} f_j \qquad v + f = n \tag{1.50}$$

L'entropie de l'attribut a est alors calculée par l'équation 1.51.

$$H(C|a) = \frac{v}{n} \sum_{j=1}^{k} \left(\frac{v_j}{v} \log \frac{v_j}{v} \right) + \frac{f}{n} \sum_{j=1}^{k} \left(\frac{f_j}{f} \log \frac{f_j}{f} \right) \tag{1.51}$$

Finalement, on retient l'attribut a_i minimisant l'entropie, soit :

$$i = \underset{i=1,\ldots,p}{\operatorname{argmin}}(H(C|a_i)) \tag{1.52}$$

Une fois l'attribut a choisi, la décision concernant cet attribut coupe l'ensemble de données en deux parties et chacune de ces parties doit alors recommencer cette recherche d'attribut optimal (voir algorithme précédent : construire-arbre(X)).

L'avantage principal des arbres de décision est qu'ils sont facilement transposables sous

forme de conditions interprétables tout en exigeant peu de calcul pour obtenir la classification demandée. Mais, les arbres de décision possèdent tout de même quelques défauts. Le premier d'entre eux est qu'ils ne supportent pas réellement les valeurs continues. Il est toujours possible de les discrétiser mais cela pose alors le problème de la discrétisation optimum (perdant le moins d'information possible par rapport à la variable initiale). De plus, les arbres de décisions sont sensibles au bruit dans les données. Ceci peut alors empêcher une bonne généralisation de l'ensemble d'apprentissage et conduire alors à de fausses conclusions sur les observations futures à classer [44].

1.5.6 Les réseaux de neurones

Les réseaux de neurones artificiels, également appelés réseaux neuromimétiques, constituent une technique non-linéaire de prédiction de données. Cet outil se veut ressemblant au fonctionnement des réseaux de neurones humains qui sont considérés comme les calculateurs les plus puissants qu'ait réalisé la nature [41]. Pour plus de simplicité, nous les nommerons réseaux de neurones. Mais, avant d'étudier un peu plus les réseaux de neurones, regardons ce qui est considéré comme neurone. Un schéma d'un neurone artificiel est présenté sur la figure 1.29.

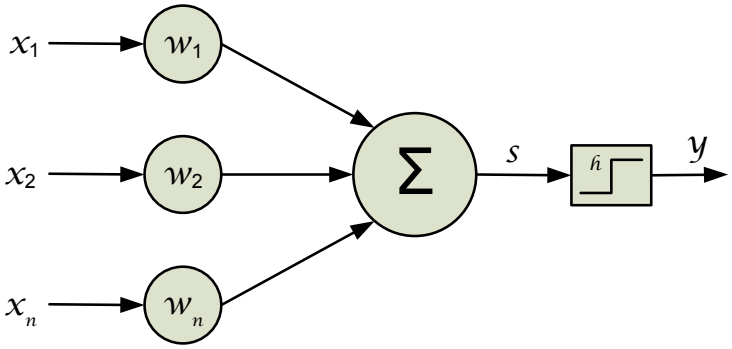


FIG. 1.29 – Un neurone artificiel

Nous pouvons observer qu'un neurone reçoit une information de la part de plusieurs entrées (n en l'occurrence) : $x_1, x_2, \ldots, x_n$. Chaque entrée est pondérée par un poids propre w_j que l'on nomme poids synaptique (en référence aux synapses du neurone naturel). Le neurone effectue la somme de toutes ces entrées pondérées. Nous nommons s cette somme.

$$s = \sum_{i=1}^{n} w_i x_i \tag{1.53}$$

La somme s représente l'état interne du neurone. Elle est transmise à une fonction de transfert nommée fonction d'activation h. La sortie de cette fonction donne la sortie

générale du neurone y. Le fonctionnement du neurone peut donc simplement s'écrire sous la forme de l'équation 1.54.

$$y = h\left(\sum_{i=1}^{n} w_i x_I\right) \tag{1.54}$$

Un neurone permet de modéliser une quantité considérable de comportements suivant les poids synaptiques w_i qu'il possède mais également suivant la fonction d'activation qu'il renferme. Différentes fonctions d'activation peuvent être utilisées [44], mais les principales sont représentées sur la figure 1.30.

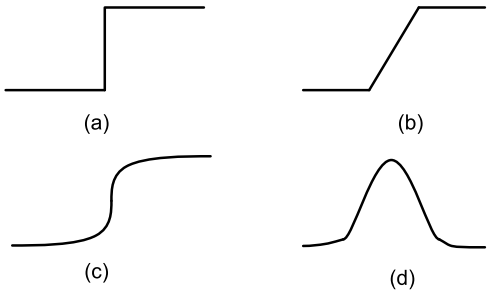


FIG. 1.30 – Les différentes fonctions d'activation h : (a) fonction à seuil, (b) fonction linéaire, (c) fonction sigmoïde, (d) fonction gaussienne

La mise en relation de plusieurs neurones donne naissance à un réseau de neurones. Le réseau possède des entrées venant de l'extérieur connectées à certains neurones, et le réseau fournit alors vers l'extérieur une ou plusieurs sorties (la sortie de un ou de plusieurs neurones du réseau). Le premier réseau de neurones est le Perceptron [126]. Ce réseau est un discriminateur linéaire. Cette linéarité a freiné considérablement le développement des réseaux de neurones pendant de nombreuses années. En effet, les problèmes linéaires pouvant être résolus par des outils plus simples que le Perceptron, cette voie de réseau neuronale ne s'est pas développée rapidement. Un des réseaux le plus connu et le plus exploité est le Perceptron MultiCouche ou PMC (MultiLayer Perceptron ou MLP) [41]. Un perceptron multicouche est un réseau subdivisé en couche de neurones : la sortie d'un neurone d'une couche n'est lié qu'aux neurones de la couche suivante. Il n'y a donc aucune liaison entre les neurones d'une même couche. On nomme généralement la première couche "couche d'entrée" et la dernière couche "couche de sortie". Entre ces deux couches se situe alors une ou plusieurs couches de neurones, nommées couches cachées. Un exemple de perceptron multicouche à quatre entrées, quatre sorties et une couche cachée est montré sur la figure 1.31.

Ce type de réseau est très performant pour les tâches de classification. Mais, il convient tout de même de préciser qu'il n'y a pas de règles fixes pour la création d'un tel réseau.

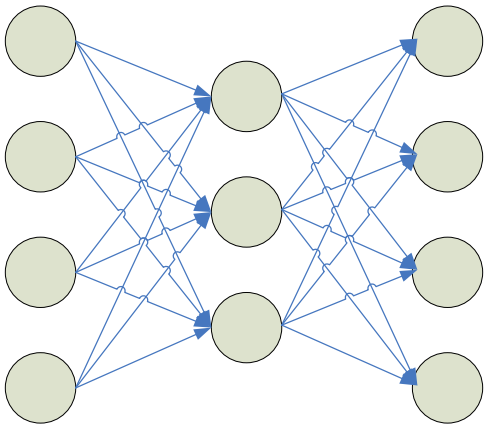


FIG. 1.31 – Exemple de perceptron multicouche

En effet, bien que le nombre de neurones de la couche d'entrée et de sortie soit imposé par le nombre d'entrées du système (pour la couche d'entrée) ainsi que par la codification des différentes classes (pour la couche de sortie), aucune règle mathématique au sens strict ne permet de déterminer, pour un problème donné, le nombre de couches cachées ainsi que le nombre de neurones de chacune de ces couches. Cependant, une pratique courante est l'utilisation d'une seule couche cachée composée d'un nombre de neurones d'environ la moitié du nombre d'entrée du système plus un. De même, pour les fonctions d'activation, il n'y pas de règle stricte permettant de choisir une fonction optimale. Mais, pour la plupart des problèmes, une fonction d'activation sigmoïde permet d'obtenir des résultats corrects. Une fois la structure du réseau fixée, la principale difficulté est l'attribution des différents poids synaptiques dans l'ensemble du réseau. Dans le cas d'un apprentissage supervisé, nous possédons un ensemble d'apprentissage complet. Il est alors possible d'utiliser l'algorithme le plus connu pour ce type de tâche : l'algorithme de rétropropagation du gradient [29, 128]. Cet algorithme calcule une erreur quadratique telle que donnée dans l'équation 1.55 entre la sortie calculée (au vu des entrées) y_i et la réponse k_i attendue pour le jeu d'entrée donné.

$$E = \frac{1}{2}\sum(y_i - k_i)^2 \tag{1.55}$$

L'algorithme procéde à la propagation de cette erreur en sens inverse vers l'avant-dernière couche, en ventilant l'erreur en fonction de l'importance des connexions, permettant un ajustement des poids par une méthode du gradient.

Il faut préciser que l'algorithme de rétropropagation du gradient demande un nombre d'exemples d'apprentissage conséquent afin de trouver des frontières de classification correctes. De plus, bien que le calcul soit rapide pour obtenir une réponse du réseau ayant

appris, la phase d'apprentissage du réseau peut s'avérer extrêmement coûteuse en ressources de calcul ainsi qu'en temps d'exécution. Précisons également que ce type de réseau ne permet pas de prendre en compte de manière directe une variable d'entrée qualitative, et qu'il faut alors considérer cette variable quantitativement. Enfin, bien que les performances atteintes par ce type de réseau soit remarquables, il ne permet pas d'obtenir une représentation explicite de ce qu'il a appris. En effet, la force du réseau réside sur sa structure et sur les poids synaptiques appliqués à ses connexions. On dénomme généralement ces réseaux boites noires, car bien qu'ils arrivent à accomplir la tâche demandée, il est presque impossible de savoir quels sont les mécanismes ou règles sous jacentes à son fonctionnement [17]. Perzick et al. [118] ont comparé les performances d'un simple classifieur naïf de Bayes aux réseaux de neurones sur des données de procédés de fabrication (réelles et simulées). Les auteurs montrent que le plus simple classifieur bayésien (le réseau bayésien naïf) permet dans la plupart des cas d'égaler les réseaux de neurones et même quelques fois de les surpasser. Les auteurs mettent alors l'accent sur les résultats obtenus avec le classifieur bayésien car celui-ci est très simple, demande très peu de ressources de calcul et est très rapide à l'apprentissage. Les auteurs concluent alors sur l'utilité importante que pourrait avoir les réseaux bayésiens dans un contexte industriel.

1.5.7 L'analyse discriminante

L'analyse discriminante est une technique statistique de classification se basant sur la règle de Bayes. En effet, elle affecte à un nouvel individu $\boldsymbol{x}$ la classe C_i qui possède la probabilité a posteriori $P(C_i|\boldsymbol{x})$ maximale d'appartenance sachant la valeur de tous les descripteurs, tel que défini par l'équation 1.56.

$$\boldsymbol{x} \in C_i, si \quad i = \underset{i=1,\ldots,k}{\operatorname{argmax}}\{P(C_i|\boldsymbol{x})\} \tag{1.56}$$

La règle de Bayes permet d'obtenir la valeur de $P(C_i|\boldsymbol{x})$ par la formule de l'équation 1.57, où $P(C_i)$ est la probabilité a priori d'appartenance à la classe C_i.

$$P(C_i|\boldsymbol{x}) = \frac{P(C_i)P(\boldsymbol{x}|C_i)}{P(\boldsymbol{x})} \tag{1.57}$$

On voit que pour chaque classe, le dénominateur de l'équation 1.57 est le même, il n'intervient donc pas dans la fonction discriminante. L'équation 1.56 peut ainsi se récrire sous la forme de l'équation 1.58.

$$\boldsymbol{x} \in C_i, si \quad i = \underset{i=1,\ldots,k}{\operatorname{argmax}}\{P(C_i)P(\boldsymbol{x}|C_i)\} \tag{1.58}$$

Pour plus de facilité nous allons écrire cette règle de décision sous forme de fonction de coût K telle que donnée dans l'équation 1.59

$$K_i(\boldsymbol{x}) = -2log(P(C_i)P(\boldsymbol{x}|C_i)) \tag{1.59}$$

On peut alors écrire la règle d'attribution d'un nouvel individu $\boldsymbol{x}$ à une classe C_i par la règle suivante.

$$\boldsymbol{x} \in C_i, si \quad i = \underset{i=1,\ldots,k}{\operatorname{argmin}}\{K_i(\boldsymbol{x})\} \tag{1.60}$$

1.5.7.1 Application à la loi normale multivariée

La loi normale multivariée est l'extension de la loi normale (ou loi de Gauss) au domaine multivarié. Sa fonction de densité conditionnellement à une classe C_i s'écrit comme indiquée par l'équation 1.61, où $\boldsymbol{\mu}_i$ représente le vecteur des moyennes de la classe C_i et $\boldsymbol{\Sigma}_i$ représente la matrice de variance-covariance de la classe C_i.

$$P(\boldsymbol{x}|C_i) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)) \tag{1.61}$$

Comme les valeurs exactes des paramètres de loi : $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ ne sont pas connus, il faut les estimer à partir des données. Pour cela, on utilise généralement la méthode d'Estimation par Maximum de Vraisemblance (Maximum Likelihood Estimation). L'intérêt premier d'utiliser l'estimation par maximum de vraisemblance est qu'elle possède de bonnes propriétés de convergence lorsque la taille de l'échantillon utilisé pour l'estimation s'accroît. De plus, elle reste la méthode la plus simple d'estimation de paramètres de loi. Dans le cas d'une loi normale multivariée, on peut montrer que l'estimation du vecteur des moyennes $\boldsymbol{\mu}$ revient à l'équation 1.62. Cette estimation est sans biais.

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{1.62}$$

De même, une estimation non-biaisée de $\boldsymbol{\Sigma}$ est :

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T \tag{1.63}$$

Pour une justification approfondie de ces estimations, on peut notamment consulter [44].

L'analyse discriminante quadratique Dans le cas de la règle de décision de l'analyse discriminante appliquée à une loi normale multivariée, l'équation 1.59 du coût s'écrit :

$$K_i(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) - 2log(P(C_i)) + log(|\boldsymbol{\Sigma}_i|) + plog(2\pi) \tag{1.64}$$

Dans l'équation 1.64, on voit que le dernier terme $plog(2\pi)$ est constant à chaque K_i et n'intervient donc pas pour la discrimination. Cette règle se nomme "Analyse Discriminante Quadratique". Elle réalise des séparations quadratiques entre chaque classe. On peut remarquer que l'expression $(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)$ est la distance de Mahalanobis de $\boldsymbol{x}$ pour la classe C_i. Si l'on fait l'hypothèse d'indépendance des variables, alors $\boldsymbol{\Sigma}_i$ est diagonale (toutes les covariances sont nulles) : $diag(\boldsymbol{\Sigma}_i) = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. Cette règle de décision est également connue sous le nom de classifieur de Bayes ou réseau bayésien naïf. En continuité à l'hypothèse d'indépendance des variables, on peut également faire l'hypothèse, pour chaque classe, d'égalité des variances σ^2 des p variables. Dans ce cas, chaque classe possède une forme sphérique mais de taille différente (taille dépendante de σ_i^2). La règle de décision de l'équation 1.64 se simplifie et donne l'équation 1.65 suivante.

$$K_i(x) = \frac{\|\boldsymbol{x} - \boldsymbol{\mu_i}\|^2}{\sigma_i^2} - 2log(P(C_i)) + 2plog(\sigma_i) + plog(2\pi) \tag{1.65}$$

Le problème de l'analyse discriminante quadratique est qu'elle exige l'estimation de beaucoup de paramètres, nécessitant donc beaucoup de données. Les problèmes d'estimation venant principalement des différentes matrices de variance-covariance, une solution consiste en l'analyse discriminante linéaire.

L'analyse discriminante linéaire Pour réaliser une analyse discriminante linéaire, nous faisons l'hypothèse d'égalité des matrices de variance-covariance. C'est à dire que pour toute classe C_i, $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, avec $\boldsymbol{\Sigma}$ la matrice de variance-covariance commune à toutes les classes. Cette matrice est obtenue par l'équation 1.66 dans laquelle les n_i représentent les nombre de cas des classes C_i que contient l'ensemble d'apprentissage et où n est le nombre de cas total (en d'autres termes $n = n_1 + n_2 + \ldots + n_k$).

$$\boldsymbol{\Sigma} = \frac{(n_1 - 1)\boldsymbol{\Sigma}_1 + (n_2 - 1)\boldsymbol{\Sigma}_2 + \ldots + (n_k - 1)\boldsymbol{\Sigma}_k}{n - k} \tag{1.66}$$

Ainsi, nous pouvons de nouveau simplifier la discrimination de l'équation 1.64 puisque le terme $log(|\boldsymbol{\Sigma}_i|)$ devient $log(|\boldsymbol{\Sigma}|)$ et est donc constant pour chaque classe. Ainsi, en posant $log(|\boldsymbol{\Sigma}|) + plog(2\pi) = Cste$ nous obtenons la fonction coût de l'équation 1.67.

$$K_i(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) - 2log(P(C_i)) + Cste \tag{1.67}$$

On peut remarquer que si les probabilités a priori de chaque classe sont égales (même nombre d'exemples pour chaque classe), alors la règle de décision revient à calculer les distances de Mahalanobis pour chaque classe et à attribuer au nouvel individu la classe pour laquelle cette distance est la plus faible.

Cette fonction coût réalise des séparations linéaires (droites, plans, hyperplans) entre les classes. Mais, comme dans le cas de l'analyse discriminante quadratique, nous pourrons également faire l'hypothèse supplémentaire que $\boldsymbol{\Sigma}$ est diagonale ou bien sphérique. L'analyse discriminante linéaire est plutôt robuste face aux fluctuations sur les hypothèses de normalité des classes et d'égalité des matrices de variance-covariance. De ce fait, elle est fréquemment utilisée et doit être considérée comme une méthode de référence. Cependant, bien souvent, l'analyse discriminante linéaire est confondue avec l'analyse factorielle discriminante. Or, l'analyse factorielle discriminante n'est pas une méthode de classification, mais une étape de réduction dimensionnelle.

1.5.7.2 Régularisation de l'analyse discriminante

Le principal problème de l'analyse discriminante est l'estimation des paramètres de loi. En effet, l'estimation par maximum de vraisemblance est optimal lorsque n tend vers l'infini. Mais, dans le cas où l'échantillon d'apprentissage est de taille faible, il peut apparaître quelques problèmes, et notamment les suivants :

- les estimations de matrices de variance-covariance deviennent très variables,
- certains paramètres peuvent ne pas être identifiables,
- certaines matrices de variance-covariance deviennent non-inversibles.

Pour faire face à ces problèmes, plusieurs approches ont été proposées [49, 58, 140]. On appelle ces approches régularisation. Elles ont pour but d'estimer de manière plus juste (donnant de meilleurs résultats lors de la classification) les matrices de variance-covariance de chaque classe lorsque peu de données sont à disposition.

L'approche la plus connue est celle de Friedman [49]. Cette approche propose le calcul d'une matrice de variance-covariance d'une classe comme étant une fonction à deux variables (λ, γ) donnée par :

$$\boldsymbol{\Sigma}_i(\lambda, \gamma) = (1 - \gamma)((1 - \lambda)\boldsymbol{\Sigma}_i + \lambda\boldsymbol{\Sigma}) + \gamma \frac{tr[(1 - \lambda)\boldsymbol{\Sigma}_i + \lambda\boldsymbol{\Sigma}]I}{p} \tag{1.68}$$

Le coefficient γ contrôle l'estimation des valeurs propres des matrices de variance-covariance de chaque classe, alors que le coefficient λ gère la pondération de la matrice de variance-covariance commune à toutes les classes. Il faut donc trouver le couple de valeur (λ, γ) donnant le meilleur résultat en terme de classification. L'évaluation des différentes solutions peut se faire à l'aide de techniques d'évaluation comme la validation croisée. L'avantage de l'analyse discriminante régularisée est qu'elle est directement liée au pourcentage de classification, mais elle demande plus de calculs.

D'autres estimateurs permettent une régularisation de l'analyse discriminante : estimateur LOOC (Leave-One-Out Covariance) [58], estimateur MECS (Maximum Entropy Covariance Selection) [140], etc.

1.5.8 Modèle à mélanges de gaussiennes

La méthode semi-paramétrique la plus connue et la plus utilisée est sans conteste le Modèle à Mélanges de Gaussiennes (Gaussian Mixture Model) [100]. La fonction de densité de probabilité $p(\boldsymbol{x})$ est une addition pondérée de plusieurs fonctions de densité de probabilité de modèles paramétriques, et en l'occurrence de plusieurs modèles gaussiens. On peut définir $p(\boldsymbol{x})$ à l'aide de l'équation 1.69.

$$p(x) = \sum_{j=1}^{d} \alpha_j p(x|D_j) \tag{1.69}$$

Dans cette équation, d représente le nombre de composantes du modèle. En d'autres termes, d représente le nombre de lois normales utilisées pour approximer la fonction de densité de probabilité $p(\boldsymbol{x})$. α_j représente le poids à attribuer à la loi normale représentée par $p(\boldsymbol{x}|D_j)$. Bien entendu, pour tout j on doit avoir $0 < \alpha_j < 1$. De plus, on doit vérifier que $\sum_{j=1}^{d} \alpha_j = 1$. Pour mieux comprendre, une visualisation graphique d'une fonction de densité non normale dans le cas univarié est proposée sur la figure 1.32.

A première vue, il semble plutôt difficile d'obtenir une formule analytique de la fonction de densité de probabilité $p(\boldsymbol{x})$. Or, sur la figure 1.33 on peut remarquer que $p(\boldsymbol{x})$ est tout simplement une addition pondérée de 3 fonctions de densité de probabilité normales : $p(\boldsymbol{x}|D_1)$, $p(\boldsymbol{x}|D_2)$ et $p(\boldsymbol{x}|D_3)$.

Bien entendu, la difficulté majeure est de trouver les paramètres des fonctions $p(\boldsymbol{x}|D_j)$. Pour cela, plusieurs méthodes existent : l'algorithme EM (Expectation-Maximisation) [31], les chaînes de Markov associées à une simulation Monte Carlo [91], ainsi qu'une méthode spectrale [145]. Mais, ces méthodes prennent toujours en compte le fait que le nombre de composante d du modèle soit connu. Or, bien souvent ce n'est pas le cas. Le problème

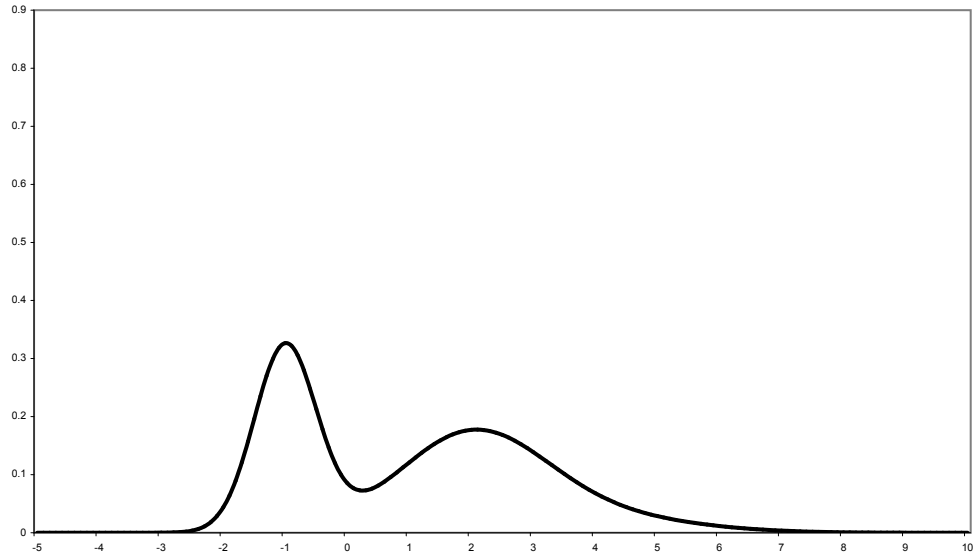


FIG. 1.32 – Exemple de fonction de densité non normale $p(\boldsymbol{x})$

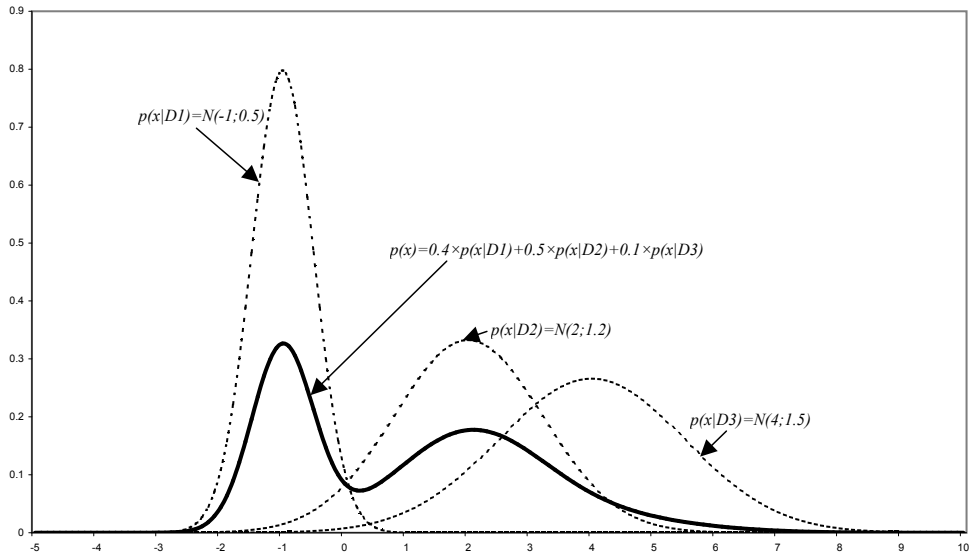


FIG. 1.33 – Mélange de gaussiennes pour définir $p(\boldsymbol{x})$

apparaissant alors est le calcul du nombre optimal pour d. Malheureusement, il n'existe pas pour le moment de réelle solution à ce problème. Il faut donc procéder par itération $d = 1$, $d = 2$, etc, puis effectuer un test statistique entre chaque itération pour voir si l'augmentation de d révèle une amélioration ou non [44].

1.5.9 Les réseaux bayésiens

Un réseau bayésien est un modèle graphique dans lequel les connaissances sont représentées sous forme de variable. Chaque variable est un nœud du graphe et prend ses valeurs dans un ensemble discret ou continu. Le graphe est toujours dirigé et acyclique. Les arcs dirigés représentent un lien de dépendance directe (la plupart du temps il s'agit de causalité). Ainsi un arc allant de la variable X à la variable Y exprimera le fait que X dépende directement de Y. L'absence d'arc ne renseigne que sur la non-existence d'une dépendance directe. Les paramètres expriment les poids donnés à ces relations et sont les probabilités conditionnelles des variables sachant leurs parents (exemple : $P(Y|X)$) ou les probabilités a priori si la variable n'a pas de parents. Pour plus de détails concernant cet outil, on peut consulter [13, 108].

Il est possible de réaliser des classifieurs performants grâce aux réseaux bayésiens [50, 81, 82, 96, 117]. Nous présentons ici les principaux types de structures permettant d'employer les réseaux bayésiens comme classifieurs. Mais, avant cela, voyons plutôt leur dénominateur commun. Un classifieur bayésien d'un problème à p variables a pour particularité de posséder $p + 1$ nœuds. En effet, tous les classifieurs bayésiens modélisent l'appartenance à une classe par un nœud discret nommé "nœud de classe", noté C. C est un nœud discret multinomial à k modalités, où k représente le nombre de classes du problème $(C_1, C_2, \ldots, C_k)$. Ce nœud de classe ne posséde pas de nœud parent. Les autres nœuds, au nombre de p, représentent les variables descriptives du problème et sont notés X_i. Nous étudions les trois principaux types de classifieurs bayésiens : le réseau bayésien naïf, le réseau bayésien naïf augmenté par un arbre, ainsi que le réseau bayésien semi-naïf condensé.

Le réseau bayésien naïf Le classifieur bayésien possédant la structure la plus simple est le réseau bayésien naïf, également appelé classifieur de Bayes. On le qualifie de naïf car il fait l'hypothèse, très forte, que chaque variable descriptive est, conditionnellement à la classe, indépendante des autres. Lorsque toutes les variables descriptives sont incorporées au modèle, on parle de structure naïve complète. Ce classifieur est extrêmement connu car ses performances, notamment dans le cas où toutes les variables sont discrètes, sont surprenantes (au vu de la simplicité de ce classifieur) dans certains domaines et dépassent des techniques beaucoup plus sophistiquées même lorsque l'hypothèse d'indépendance est violée [37]. De ce fait, énormément d'études ont été réalisées sur le réseau bayésien naïf [45, 61, 164]. L'hypothèse d'indépendance des variables permet d'écrire facilement la probabilité a posteriori de chaque classe comme indiqué dans l'équation 1.70.

$$P(C_i|x) = P(C_i) \prod_{j=1}^{p} P(x_j|C_i) \tag{1.70}$$

La structure d'un réseau bayésien naïf peut se représenter comme sur la figure 1.34.

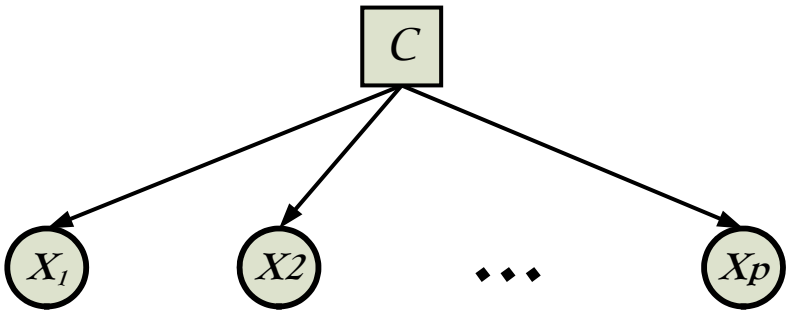


FIG. 1.34 – Réseau bayésien naïf

Le TAN Afin d'améliorer les performances du réseau bayésien naïf, Friedman et al. [50] proposent d'ajouter des arcs entre les différentes variables descriptives du classifieur naïf. Pour cela, ils décident de créer un arbre entre les variables descriptives, à la manière de Chow et Liu [19], afin d'obtenir un TAN (Tree Augmented Naïve Bayes). L'algorithme part d'un réseau bayésien naïf et ajoute un arc entre les variables qui partagent la plus importante information mutuelle. Mais, afin de respecter la topologie de l'arbre, l'algorithme interdit à chaque nœud d'avoir plus de 2 parents (soit un parent en plus du nœud de classe). Ainsi, la solution proposée par Friedman et al. [50] aboutit sur une structure complète, c'est à dire que tous les nœuds sont pris en compte dans le modèle. Il se peut que certaines variables puissent être enlevées du modèle, ainsi que certains arcs. Pour résoudre ce problème, Keogh [72] propose un algorithme basé sur la recherche des structures possibles évaluées au travers du taux de mauvaises classifications. Mais, la complexité et le temps d'exécution de cet algorithme sont beaucoup plus élevés que celui de Friedman et al., et ce, pour une amélioration des résultats qui n'est pas réellement significative [116]. La structure d'un TAN peut se représenter comme sur la figure 1.35.

On peut également citer les travaux de Sahami [130] qui propose, non plus de fixer un maximum d'une relation entre chaque descripteur, mais un nombre k fixé.

Le réseau bayésien semi-naïf condensé Afin de prendre en compte la corrélation entre les différents descripteurs, il a également été proposé les réseaux bayésiens semi-naïfs condensés [74, 113]. On les nomme condensés car ils introduisent une nouvelle sorte de variable : les variables jointes. Ces nouvelles variables jointes représentent un groupement

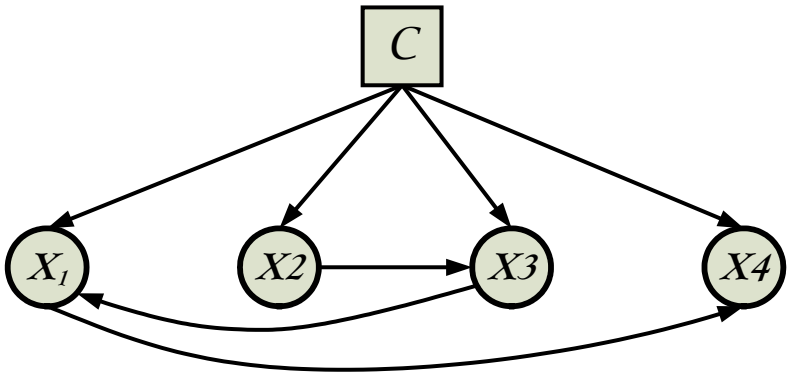


FIG. 1.35 – Réseau bayésien naïf augmenté par un arbre : TAN

de variables descriptives. Bien entendu, une variable descriptive ne peut se trouver que dans une seule variable jointe. Le fait que deux variables se trouvent dans une variable jointe implique que ces deux variables sont corrélées. Une représentation d'un réseau bayésien semi-naïf condensé est donnée sur la figure 1.36.

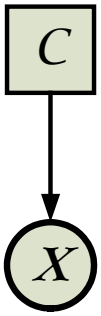


FIG. 1.36 – Réseau bayésien semi-naïf condensé

Il existe trois possibilités pour la variable jointe : elle est composée soit de variables discrètes [74, 113], soit de variables continues [116], soit de variables discrètes et continues. La dernière possibilité n'a, à notre connaissance, jamais été traitée ou évoquée. La possibilité dans laquelle toutes les variables sont discrètes implique que les modalités du nœud joint sont le produit cartésien des modalités des variables le constituant. Ce type de variable est très peu utilisé car le nombre de modalités étant très élevé, on peut arriver à des paramètres erronés ou instables qui faussent les résultats du réseau. Par contre, le cas où une variable jointe est formée par le regroupement de plusieurs variables continues est beaucoup plus simple à traiter. En effet, on peut faire l'hypothèse qu'un regroupement de p variables continues suit une loi normale multivariée et est donc représentée par un seul nœud continu de dimension p.

1.6 Choix d'un classifieur pour la surveillance des procédés

Les phases de détection et de diagnostic ne reposent que sur un seul et même outil : un classifieur. Il est donc intéressant de trouver un classifieur permettant à la fois la détection mais également le diagnostic, et cela de manière efficace pour les deux phases. L'outil idéal doit permettre de réaliser une analyse des données afin de distinguer les différentes classes de fautes du procédé (apprentissage non-supervisé). Suite à cela, il doit pouvoir traiter correctement le cas de discrimination lorsque peu de données sont présentes (typiquement le cas lors du début d'activité d'un procédé). Il doit également pouvoir réaliser une sélection des variables importantes pour la discrimination entre les fautes. Enfin, cet outil doit être résistant aux situations d'informations manquantes (capteur défectueux d'une ou plusieurs variables), situation arrivant fréquemment dans les bases de données industrielles.

Nous avons vu que la détection et le diagnostic pouvaient se considérer comme des tâches de classification. Ainsi, nous établissons la synthèse des outils de classification vu à la section 1.5. Pour cela, nous avons réalisé un tableau de synthèse (voir table 1.3) permettant d'évaluer chaque classifieur sur des critères distincts.

Suite à l'étude des principaux classifieurs supervisés utilisables dans le cadre du diagnostic de fautes à base de données, nous pouvons tirer certaines conclusions.

Nous avons vu que certains classifieurs pouvaient traiter des variables discrètes, des variables continues (sous diverses hypothèses) et des variables continues discrétisées. La plupart des classifieurs traitent uniquement des variables continues (séparateur à vaste marge, réseau de neurones, analyse discriminante, modèle à mélanges de gaussiennes, ainsi que les k plus proches voisins). Dans le cadre du diagnostic ou de la reconnaissance de forme, il est usuel d'avoir un grand nombre de variables quantitatives. Ces classifieurs permettent donc de traiter une grande majorité des cas. Cependant, il est légitime de se poser la question suivante : face à un problème de classification donné, n'y a t-il vraiment aucune variable discrète mise en jeu, ou bien sont elles évincées du système de classification du fait de la non possibilité de les prendre en compte dans le classifieur ? Ne pouvant répondre à cette question, nous prenons ainsi la décision suivante : si le procédé à surveiller possède des variables discrètes, elles doivent absolument être prises en compte dans le système de diagnostic. En effet, il est commun de trouver des variables discrètes lors de contrôle qualité sur les produits. Bien entendu, une variable discrète peut être prise en compte dans un système de classification ne traitant que des variables continues, mais dans ce cas les différentes modalités de la variable discrète sont codées par des

Classifieurs	Séparateurs à vaste marge	k plus proches voisins	Arbres de décision	Réseaux de neurones	Analyses discriminante	Mélanges de gaussiennes	Réseaux bayésiens
Gère les variables discrètes	☹	😐	🙂	☹	☹	☹	🙂
Gère les variables continues	🙂	🙂	😐	🙂	🙂	🙂	🙂
Gère un grand nombre de classe	☹	😐	😐	😐	😐	😐	😐
Nbre d'observations nécessaire faible	☹	🙂	😐	☹	😐	☹	😐
Temps d'apprentissage correct	☹	🙂	😐	☹	😐	😐	😐
Temps de calcul admissible	😐	☹	🙂	🙂	🙂	🙂	😐
Nombre de paramètre à régler faible	☹	😐	🙂	☹	😐	☹	😐
Gère des données manquantes	😐	😐	😐	🙂	😐	😐	🙂
Gère un grand nombre de variables	😐	😐	😐	😐	😐	😐	😐
Corrélation entre les variables	🙂	🙂	🙂	🙂	🙂	🙂	🙂
Gère certaines non-linéarités	🙂	🙂	🙂	🙂	🙂	🙂	🙂
Tolérance correcte au bruit	🙂	🙂	☹	🙂	🙂	🙂	🙂
Information redondante	😐	😐	😐	😐	😐	😐	🙂
Prise en compte du temps	😐	😐	😐	😐	😐	😐	😐
Adaptabilité : facile de réadapter l'outil	😐	😐	😐	😐	😐	😐	🙂

🙂 : signifie que le classifieur permet de prendre en compte ce critère

😐 : signifie que le classifieur permet de prendre en compte ce critère sous certaines conditions

☹ : signifie que le classifieur ne permet pas de prendre en compte ce critère

TAB. 1.3 – Tableau comparatif des différents classifieurs

attributs numériques, ce qui n'a pas réellement de sens et peut conduire à des conclusions erronées. Nous pensons donc qu'un système de classification optimal doit pouvoir traiter les variables discrètes sans les quantifier. La remarque que nous venons de faire concernant les classifieurs travaillant avec des variables continues peut également être fait pour les classifieurs travaillant avec des variables discrètes. En effet, les classifieurs à variables discrètes sont théoriquement incapables de travailler avec des variables continues. Pourtant, il existe une solution pratique très employée : la discrétisation des données.

Il existe énormément de méthodes de discrétisation : discrétisation par k intervalles de même amplitude, discrétisation par k intervalles de même fréquence, discrétisation par minimisation de l'entropie, etc. Yang et Webb [164] ont comparé plusieurs méthodes de discrétisation. Mais, la discrétisation fait perdre une certaine quantité d'information sur la variable discrétisée, ce qui peut également mener à des conclusions erronées. Un outil permet de traiter à la fois le cas des variables discrètes et celui des variables continues : le réseau bayésien. Un réseau bayésien peut contenir des variables discrètes (ou bien continues discrétisées) multinomiales, mais également des variables continues. Cependant, seul le cas d'une variable continue possédant une fonction de densité gaussienne est traitable par les algorithmes d'inférence. Comme pour les modèles à mélanges de gaussiennes, une variable continue peut se modéliser sous forme de mélanges de plusieurs variables gaussiennes. Ainsi, le réseau bayésien peut traiter les deux types possibles de variables, et ce, sans forcement passer par une discrétisation des variables continues.

En ce qui concerne le nombre de classes, les différents classifieurs peuvent en supporter une grande quantité exception faite du séparateur à vaste marge. En effet, comme nous l'avons vu, ce classifieur ne peut prendre une décision qu'entre deux classes. Pour effectuer une classification parmi k classes, il devra être mis en place un système de plusieurs classifieurs prenant une décision par vote. Deux stratégies sont envisageables : "un contre un" ou "un contre tous". Dans la stratégie "un contre un", on construit autant de classifieurs binaires que de couples possibles de deux classes, résultant en $k(k-1)/2$ classifieurs. Dans la stratégie "un contre tous", on construit un classifieur par classe se mesurant alors aux autres classes regroupées dans une seule, soit k classifieurs. Au vu de la complexité calculatoire d'un séparateur à vaste marge, on comprend que la construction de plusieurs d'entre eux demande une ressource de calcul très importante. De plus, avoir plusieurs classifieurs implique un réglage des paramètres sur chacun, ce qui peut devenir très rapidement insolvable. En pratique, les séparateurs à vaste marge ne seront que très rarement utilisés pour des problèmes ayant plus d'une dizaine de classes ($k = 10$), contrairement aux autres classifieurs, comme les réseaux bayésiens, qui permettent la prise en compte d'un nombre de classes très élevé.

Il est important de rappeler que tous ces classifieurs basent leur apprentissage sur un jeu de données. Il est donc légitime de s'interroger sur la quantité de données nécessaire pour réaliser un apprentissage. Nous faisons l'hypothèse que les données récupérées sont vraiment représentatives des différentes classes de fonctionnement du système. Bien que cela dépende fortement du problème spécifique à traiter, on peut tout de même dire que les réseaux de neurones, les séparateurs à vaste marge ainsi que les modèles à mélanges de gaussiennes ont besoin d'un nombre de données suffisamment important afin de fournir

des résultats satisfaisants. Cela est également le cas pour les réseaux bayésiens, l'analyse discriminante ainsi que les arbres de décision, mais dans une moindre mesure que les trois premiers cités. Enfin, le classifieur des k plus proches voisins est le plus à même de classer correctement une nouvelle observation avec un ensemble d'apprentissage faible. En effet, ne faisant pas de regroupement des données de même classe afin de calculer des paramètres de cette classe (paramètres bien souvent non représentatifs en cas de nombre insuffisant de données), son raisonnement n'est pas biaisé par une estimation incorrecte. Il est donc le plus à même de traiter le cas où peu de données sont disponibles pour l'apprentissage. Il faut également ajouter que certaines structures de réseaux bayésiens, notamment la structure naïve, permettent d'obtenir de très bons résultats même lorsque le jeu de données pour l'apprentissage est faible.

Un autre point important à prendre en compte lors du choix d'un classifieur est sa vitesse. Dans le cadre du diagnostic, la vitesse d'un classifieur contraint son utilisation à une exploitation hors-ligne ou bien en ligne. Il convient de distinguer deux éléments principaux : la vitesse d'apprentissage du classifieur et sa vitesse d'inférence. La vitesse d'apprentissage est le temps mis par le classifieur pour fixer et ajuster ses paramètres internes en fonction du jeu de données d'apprentissage qu'on lui a fourni. La vitesse d'inférence est le temps mis par le classifieur ayant déjà appris pour fournir la classe d'attribution d'une nouvelle observation du système. Afin d'utiliser un classifieur en ligne (ce qui a le plus d'intérêt), il convient que sa vitesse d'inférence soit la plus grande possible afin de diagnostiquer au plus vite une situation hors-contrôle du procédé. Nous pensons que la phase d'apprentissage du classifieur, pouvant s'avérer longue ou même très longue pour certains, est alors faisable hors-ligne. Ainsi, un classifieur fortement désavantagé sur ce point est la méthode des k plus proches voisins. Cette méthode n'est pas un classifieur comme les autres puisqu'il n'apprend pas de paramètre, son temps d'apprentissage peut donc être considéré comme nul. Par contre, il prend une décision sur une nouvelle observation en prenant en compte à chaque fois tout l'ensemble d'apprentissage à sa disposition. Son temps d'inférence est alors beaucoup plus long que pour d'autres classifieurs. Il faut également préciser que plus le jeu de données d'apprentissage est conséquent, plus l'exactitude de ce classifieur augmente, mais plus le temps d'inférence croît. Nous pensons que la méthode des k plus proches voisins n'est pas un classifieur idéal pour la classification en ligne de fautes dans un procédé. De plus, au vu de la progression des moyens informatiques, certains classifieurs comme les réseaux bayésiens autrefois jugés un peu lent sont désormais très compétitifs grâce à des algorithmes d'inférence très rapides.

Concernant le nombre de paramètres de réglage du classifieur, les séparateurs à vaste marge et les réseaux de neurones sont plutôt mal situés. Ces deux méthodes demandent

tout d'abord de définir les fonctions non-linéaires associées (fonction d'activation pour les réseaux de neurones, et fonction noyau pour les séparateurs à vaste marge). De plus, elles demandent des paramètres d'arrêt pour l'apprentissage. Or, si ces paramètres sont mal réglés, cela peut engendrer un surapprentissage de l'ensemble de données, conduisant à un taux d'erreur plus important. L'arbre de décision, quant à lui, ne demande aucun paramètre particulier dans sa version de base. Par contre, une phase d'élagage de l'arbre permettant une amélioration de ses performances contraint alors à l'utilisation de paramètres d'élagage. Concernant les modèles à mélanges de gaussiennes et les k plus proches voisins, ces classifieurs ont chacun besoin d'un paramètre : le nombre de voisins à prendre en compte pour les k plus proches voisins, et le nombre de classes pour les modèles à mélanges de gaussiennes. Concernant les réseaux bayésiens, beaucoup de méthodes d'apprentissage peuvent être utilisées [108], certaines demandant un ou plusieurs paramètres, et certaines n'en demandant aucun.

Un point intéressant concernant les réseaux de neurones, ainsi que les réseaux bayésiens, est la gestion des données manquantes. En effet, pour tous les classifieurs étudiés, dans le cas de données manquantes, il est possible d'utiliser la moyenne arithmétique de la variable. Mais, cela n'est pas réellement une solution efficace. Les réseaux bayésiens, en absence de données pour certaines variables permettent le calcul des différentes probabilités en prenant en compte la distribution théorique de la variable, et non pas juste sa moyenne. Pour les réseaux de neurones, l'enchevêtrement des connexions non-linéaires à l'intérieur du réseau permet généralement une meilleure prise en compte d'une donnée manquante. Précisons que cette remarque est également vraie pour les réseaux bayésiens en phase d'apprentissage. En effet, l'algorithme Expectation Maximization [31] permet de trouver l'estimation de paramètres par maximum de vraisemblance dans un modèle possédant des variables cachées ou partiellement cachées (données manquantes).

Tous les classifieurs étudiés permettent de traiter des problèmes de classification non-linéaire (non-linéairement séparables) ainsi que des problèmes de classification linéaire, et cela en étant capable de gérer un nombre important de variables. Pour cela, tous ces classifieurs permettent de prendre en compte la corrélation entre les variables descriptives du problème, ainsi que la prise en compte d'un certain bruit sur ces variables (légère contre performance des arbres de décision sur ce point). Tous ces classifieurs sont sensibles à la redondance d'information entre les variables (exceptés certains réseaux bayésiens tel que le réseau bayésien naïf). Par contre, ces classifieurs possèdent tous un même inconvénient, leurs performances sont diminuées en présence de variables descriptives non-informatives. En effet, si certaines des variables du problème sont importantes pour la classification (variables informatives), certaines sont totalement non-informatives et contribuent à ajouter

un bruit sur l'espace des données. Or, plus il y a de bruit dans l'espace des données, moins le classifieur est performant. Une phase essentielle avant l'utilisation de ces différents classifieurs est l'identification et la suppression de ces variables non-informatives, permettant ainsi d'améliorer les performances du classifieur. Cette phase d'identification est généralement appelée sélection de composantes (feature selection) [44].

Un avantage certain des réseaux bayésiens par rapport aux autres classifieurs est leur adaptabilité. En effet, outre le fait que leurs représentations visuelles permettent des changements plus rapides du modèle du système, il est également possible d'utiliser des réseaux bayésiens orientés objet [160]. On modélise alors un système par un ensemble de sous-systèmes (sous-réseaux) possédant des interactions, permettant une réadaptation du système entier par remplacement de sous-systèmes. De plus, les réseaux bayésiens autorise la prise en compte de l'aspect temporel grâce aux réseaux bayésiens dynamiques. Nous précisons que la structure et les paramètres du réseau n'évoluent pas avec le temps, mais il est possible de modéliser des relations entre les variables entre plusieurs intervalles de temps. Un autre avantage est que les réseaux bayésiens sont capables d'incorporer des nœuds utilité ou décision, devenant ainsi un réel outil d'aide à la décision [160]. Enfin, nous avons déjà cité plusieurs travaux [117, 118] mettant en évidence que, sur une tâche de classification de données industrielles, les réseaux bayésiens sont vraiment capables de concurrencer les réseaux de neurones ou bien les k plus proches voisins.

1.7 Conclusion

Ce premier chapitre a permis de présenter le contexte de la variabilité des procédés, cela en définissant le terme procédé puis en expliquant les différentes sources de variabilité (causes communes et causes spéciales) et leurs implications sur la qualité d'une production. Nous avons alors présenté le contexte de la maîtrise des procédés dont l'objectif principal est la réduction de la variabilité. Suite à cela, une présentation des différentes approches pour la surveillance des procédés a été proposée. Dans le contexte de la surveillance se basant sur les données, nous avons étudié plusieurs outils. Dans un premier temps, des outils pour effectuer la détection, puis dans un deuxième temps des outils pour le diagnostic. Les outils présentés pour le diagnostic sont les principaux classifieurs exploités pour la discrimination entre plusieurs classes de données. De plus, comme la détection peut être vue comme un problème de discrimination à deux classes, l'utilisation d'un classifieur peut permettre la détection et le diagnostic avec un seul outil. Ainsi, suite à la comparaison des différents classifieurs, nous avons conclu que les réseaux bayésiens sont les plus adaptés aux problèmes de la surveillance à base de données.

Chapitre 2

Réseaux bayésiens

Sommaire

2.1 Introduction

Dans le chapitre précédent, nous avons présenté l'utilité et les grands principes de la maîtrise des procédés. Nous avons également étudié les différents classifieurs possibles pour le diagnostic de fautes. Les réseaux bayésiens possèdent un fort potentiel puisqu'ils sont capables de combiner l'aspect statistique, probabiliste, avec des aspects décisionnels, et des aspects de gestion de connaissances.

La formalisation des réseaux bayésiens a débuté il y a 20 ans environ grâce notamment aux travaux de Pearl [114]. Depuis, l'intérêt des réseaux bayésiens, dans la communauté

de l'intelligence artificielle tout d'abord, puis dans toutes les autres communautés scientifiques, n'a cessé de croître. Les développements apportés jusqu'ici aux réseaux bayésiens portent principalement sur trois points essentiels : les algorithmes d'inférence, l'apprentissage de la structure du réseau, et l'apprentissage des paramètres du réseau. À l'heure actuelle, les algorithmes d'inférence permettent de calculer les différentes probabilités d'un réseau bayésien en un temps relativement acceptable. Qu'il s'agisse d'algorithmes d'inférence exactes ou approchés, ceux-ci sont opérationnels pour les réseaux bayésiens statiques [107]. De même, les algorithmes d'apprentissage de structure [87, 115, 135] ou de paramètres [69] ont fait l'objet de beaucoup de recherches durant les 10 dernières années. À présent, on peut dire que ces algorithmes sont opérationnels dans le cas des réseaux bayésiens classiques (statiques), que les données d'apprentissage soient complètes ou incomplètes [87]. Par contre, le cas de ces algorithmes pour les réseaux bayésiens dynamiques reste encore une voie où les recherches ne sont pas finalisées. De même, l'apprentissage de la structure causale d'un réseau n'est pas encore abouti. En effet, face à un système complexe, il existe dans presque tous les cas des variables latentes inhérentes au système, définissant ainsi un modèle causal semi-markovien. L'apprentissage de telles structures reste une piste à approfondir. De même, il est envisageable d'étendre ce type de modèle à un diagramme d'influence, définissant ainsi les processus de décision de Markov partiellement observés. Là encore, l'apprentissage de la structure de tels modèles n'est pas résolue.

Un autre champ de recherche concerne la modélisation des variables continues. En effet, comme nous l'avons présenté, les variables continues ne peuvent être modélisées que sous l'hypothèse de normalité. Bien entendu, ceci permet alors d'exprimer tout type de variables continues grâce notamment au mélange de gaussiennes. D'autres alternatives commencent à émerger comme par exemple les modèles à mélange d'exponentielles tronquées [21]. De même, il est intéressant d'envisager des méthodes à base de noyaux pour modéliser une variable continue [54].

Enfin, une autre piste intéressante est liée aux travaux de Jordan et al. [69] qui ont mis en avant la notion plus générale de modèle graphique probabiliste, unifiant ainsi des approches développées auparavant de façon concurrente comme les réseaux bayésiens, les modèles de Markov cachés, les filtres de Kalman ou les champs aléatoires de Markov.

Tous ces travaux de recherches ont permis de démocratiser l'utilisation des réseaux bayésiens. Il est désormais possible de trouver des applications de réseaux bayésiens dans un ensemble de domaine plus vaste et varié. Ainsi, proche de l'intelligence artificielle, un secteur utilisant énormément les réseaux bayésiens est le secteur de la reconnaissance. Les réseaux bayésiens permettent de prendre en compte nombre d'incertitudes quant à

la reconnaissance d'un élément. Par exemple, les travaux de Jonquieres [68] traitent de la reconnaissance d'objet en 3D. La reconnaissance consiste à identifier un objet puis à le localiser. Le principe est fondé sur l'utilisation de réseaux bayésiens afin de représenter les incertitudes. Les données sensorielles, fournies par une caméra, sont traitées afin d'en extraire les informations pertinentes pour la reconnaissance. Les objets, supposés connus, sont représentés par des modèles polyédriques définis interactivement. Deviren et al. [33] se proposent d'améliorer la reconnaissance vocale. Les auteurs utilisent le formalisme des réseaux bayésiens dynamiques pour construire des modèles acoustiques qui sont capables d'apprendre la structure de dépendance entre le processus caché et observé de la parole. Les auteurs présentent une méthodologie pratique pour utiliser de tels modèles dans la reconnaissance de la parole continue. L'approche permet l'utilisation de modèles à différentes structures pour différents mots dans le vocabulaire.

Un autre secteur utilisant beaucoup les réseaux bayésiens est le secteur de la médecine. En effet, en plus de toutes les applications réalisées en aide au diagnostic de maladies (en prenant les symptômes comme variables aléatoires), on peut notamment citer deux exemples d'utilisation des réseaux bayésiens. Bellot [6] présente une nouvelle approche de la fusion de données et l'applique à la surveillance médicale. La contribution de ce travail se situe au niveau de l'application des réseaux bayésiens dynamiques au diagnostic en télémédecine pour réguler, à distance, l'état physiologique d'un patient. Cette approche a servi de cadre général pour la modélisation et le diagnostic médical en télémédecine et a permis de monitorer l'état d'hydratation de personnes souffrant d'insuffisance rénale. Les réseaux bayésiens dynamiques permettent ici de modéliser des connaissances incertaines et dynamiques grâce aux probabilités. Ce travail théorique aboutit sur l'implémentation d'un moteur d'inférence bayésienne et sur la réalisation d'un système aidant le néphrologue dans ses décisions thérapeutiques. Le travail de Labatut [79] se situe dans la modélisation du traitement de l'information dans des réseaux cérébraux à grande échelle et l'interprétation des données de neuroimagerie. Il est basé sur les réseaux bayésiens dynamiques. L'auteur considère le cerveau comme un ensemble de régions fonctionnelles anatomiquement interconnectées, chaque région étant un centre de traitement de l'information modélisable par un nœud du réseau bayésien.

Enfin, proche du secteur du diagnostic des systèmes, le domaine de la sûreté de fonctionnement a également étudié cet outil. Boudali et al. [9] explorent l'utilisation des réseaux bayésiens afin de modéliser la fiabilité et analyser des systèmes dynamiques. Les composants dynamiques du système montrent des comportements complexes et des interactions, rendant les modèles combinatoires inadéquats pour les résoudre. Les chaînes de Markov ont été largement répandues pour modéliser de tels systèmes. Cependant, le

problème d'explosion combinatoire limite considérablement leur application. Les auteurs proposent donc un cadre d'analyse basé sur les réseaux bayésiens dynamiques. Weber et al. [158] affirment que les processus de fabrication complexes doivent être modélisés et commandés dynamiquement pour optimiser le diagnostic et les stratégies de maintenance. Ils présentent une méthodologie pour développer les réseaux bayésiens dynamiques afin de formaliser de tels modèles dynamiques complexes. Un système de valves est employé pour comparer les évaluations de fiabilité obtenues par le modèle proposé de réseaux bayésiens dynamiques et ceux obtenues par la chaîne de Markov classique. D'un point de vue fiabilité des logiciels, Bai et al. [3] propose d'appréhender une théorie de sûreté de fonctionnement logicielle basée sur des réseaux bayésiens. Le réseau bayésien est un outil puissant pour résoudre le problème de suppositions liées aux modèles, car il montre de fortes capacités à s'adapter dans les problèmes impliquant des facteurs variables complexes. Enfin, Corset [24] traite de l'application des réseaux bayésiens en maintenance et propose une méthodologie de construction à partir d'avis d'experts. Les actions de maintenance sont intégrées comme nouveaux nœuds du graphe. Une intégration du retour d'expérience est proposée par une inférence bayésienne, en quantifiant la confiance attribuée aux avis d'experts.

Dans la suite de ce chapitre, nous présentons plus en détail ce que sont les réseaux bayésiens. Par la suite, une seconde section porte sur une revue de l'application des réseaux bayésiens pour le diagnostic des systèmes. On pourra alors déterminer quels sont les différents points à prendre en compte pour établir une approche par réseaux bayésiens permettant la surveillance des procédés à partir des données.

2.2 Présentation des réseaux bayésiens

2.2.1 Généralités

Un réseau bayésien peut se définir comme un modèle graphique probabiliste. Il porte également d'autres appellations comme réseaux probabilistes ou réseaux de croyances. Un réseau bayésien est un outil complet permettant la visualisation de variables et de leurs dépendances (ou indépendances). Il permet également de décrire quantitativement le fonctionnement d'un système grâce aux différents calculs de probabilités concernant les variables du système. Généralement, on modélise les variables aléatoires comme étant des nœuds. On peut alors dresser un arc entre certaines variables du système. Les arcs tracés peuvent rendre compte d'un phénomène de causalité entre les variables reliées (réseaux causaux), mais ce n'est pas obligatoirement le cas.

Le fait d'indiquer un arc entre deux variables implique une dépendance directe entre

ces deux variables : l'une est le parent, et l'autre l'enfant. Il faut fournir le comportement de la variable enfant au vu du comportement de son ou ses (s'il y en a plusieurs) parents. Pour cela, chaque nœud du réseau possède une table de probabilités conditionnelles. Une table de probabilités conditionnelles associée à un nœud permet de quantifier l'effet du ou des nœuds parents sur ce nœud : elle décrit les probabilités associées aux noeuds enfants suivant les différentes valeurs des nœuds parents. Pour les nœuds racines (sans parents), la table de probabilité n'est plus conditionnelle et fixe alors des probabilités a priori concernant les valeurs de la variable.

Les réseaux bayésiens interdisent les dépendances enfant vers parents. Ainsi, l'ensemble de variables et des arcs vont former un graphe dirigé (les arcs possèdent un sens), et acyclique (pas de cycle dans le graphe).

De manière formelle, un réseau bayésien [108] est défini par :

- un graphe acyclique orienté G, $G = (V, E)$, où V est l'ensemble des nœuds de G, et E est l'ensemble des arcs de G,
- un espace probabilisé (Ω, Z, P), avec Ω un ensemble fini non-vide, Z un ensemble de sous-espaces de Ω, et P une mesure de probabilité sur Z avec $P(\Omega) = 1$,
- un ensemble de variables aléatoires associées aux nœuds du graphe G et défini sur (Ω, Z, P), tel que :

$$P(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} P(V_i | C(V_i)) \tag{2.1}$$

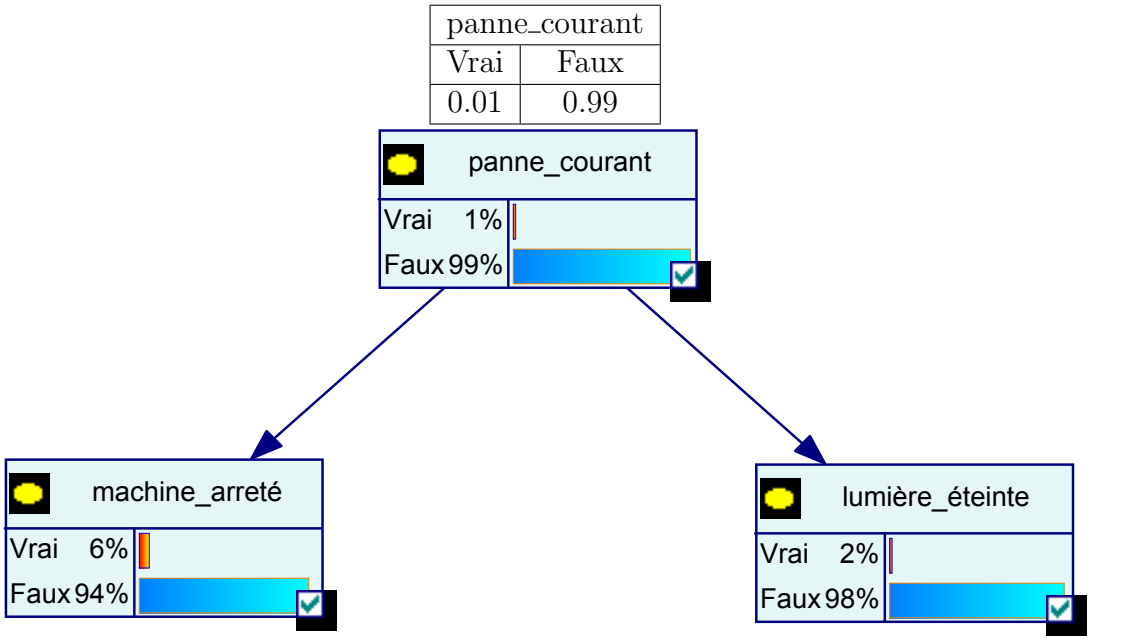
où $C(V_i)$ est l'ensemble des parents (ou causes) de V_i dans le graphe G.

Les calculs dans un réseau bayésien se nomment inférence. L'inférence permet de calculer les probabilités a posteriori de chacune des variables du réseau. Généralement, l'inférence est lancée dès qu'une information nouvelle concernant une ou plusieurs variables est disponible. Cet apport d'information est appelé évidence. Une évidence peut être dure (il pleut, c'est sûr à 100%) ou bien douce (il pleut, c'est sûr à 80%). Une fois l'information indiquée, celle-ci est propagée dans le réseau par le moteur d'inférence. Pour illustrer cette présentation des réseaux bayésiens et comprendre un peu mieux les calculs, un exemple très simple est considéré par la suite.

2.2.2 Exemple d'un réseau bayésien

L'exemple suivant (figure 2.1) représente le raisonnement que l'on peut faire sur une machine dans un atelier. Cette machine est arrêtée en moyenne dans 5% des cas pour

différentes raisons (maintenance, panne, etc). Mais, cette machine est également arrêtée lors d'une panne de courant (ce qui se produit environ 1 fois tous les 100 jours). Près de la machine se trouve une lumière qui est toujours allumée, exceptée lorsque l'ampoule grille (ce qui se produit environ dans 1% des cas), ou bien lorsqu'il y a une panne de courant. On peut alors grâce à ces données remplir les différentes tables de probabilités conditionnelles. Ces tables permettent de formaliser de manière simple les distributions de probabilités conditionnelles associées à chaque variable en fonction de ses parents. Dans cet exemple, les tables de probabilités conditionnelles sont visibles à côté des nœuds représentés.



panne_courant	
Vrai	Faux
0.01	0.99

	machine_arrete	
panne_courant	Vrai	Faux
Vrai	1	0
Faux	0.01	0.99

	lumiere_eteinte	
panne_courant	Vrai	Faux
Vrai	1	0
Faux	0.05	0.95

FIG. 2.1 – Exemple de réseau bayésien classique

Sur ce système, on peut alors se poser différentes questions. Par exemple, quelle est la probabilité que la machine fonctionne encore lorsque l'on voit que la lumière est éteinte ? Pour cela, nous allons intégrer une observation (ou évidence) au réseau. Cette observation est "la lumière est éteinte", la probabilité que "lumiere_éteinte" soit vraie est alors de 100%. Par la loi de Bayes (d'où le terme réseau bayésien), on recalcule alors toutes les probabilités de chacune des modalités de chaque variable du réseau (voir figure 2.2).

On obtient le résultat suivant : suite à l'observation "la lumière est éteinte", on a 50% de chance qu'il y ait une panne de courant. La probabilité d'avoir une panne de courant a

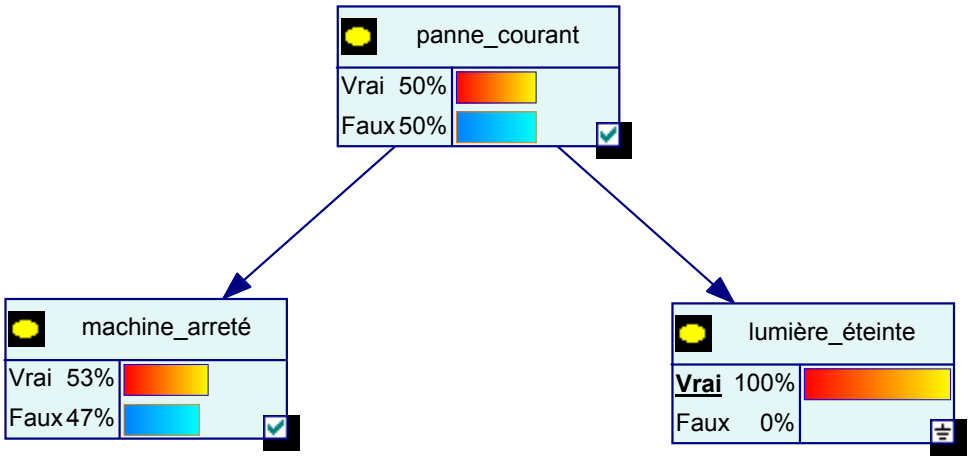


FIG. 2.2 – Exemple de réseau bayésien classique avec évidence

augmenté et cette augmentation est répercutée sur la variable "machine_arretée". Le fait de voir la lumière éteinte fait augmenter la probabilité d'avoir un arrêt de la machine de 6% à 53%.

Cet exemple de réseau bayésien illustre les calculs faits dans ce type de réseau. Tous les calculs sont effectués avec la loi de Bayes[2] qui est donnée par l'équation 2.2 où X et Y sont deux variables aléatoires.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{2.2}$$

Dans notre cas, nous ne manipulons que des variables binaires pouvant être "$vraies$" ou "$fausses$". Pour plus de simplicité, nous écrivons $P(a = vrai)$ par $P(a)$ et $P(a = faux)$ par $P(\overline{a})$. Calculons la probabilité qu'il y ait une panne de courant sachant que l'on voit la lumière éteinte :

$$\begin{aligned} P(panne\_courant|lumiere\_eteinte) &= \frac{P(lumiere\_eteinte|panne\_courant) \times P(panne\_courant)}{P(lumiere\_eteinte)} \\ &= \frac{1 \times 0.01}{0.02} = 0.5 \end{aligned}$$

Nous voyons donc que nous avons bien retrouvé le résultat obtenu sur l'exemple précédent : si la lumière est éteinte, alors la probabilité d'avoir une panne de courant passe à 50%. Calculons à présent la probabilité que la machine soit arretée sachant que l'on a 50% de chance qu'il n'y ait plus de courant :

[2]Rappel : la notation $P(X)$ représente la probabilité d'occurrence de l'événement X, alors que la notation $P(X|Y)$ représente la probabilité d'occurrence de l'événement X sachant que l'événement Y s'est produit.

$$\begin{aligned} P(machine\_arrete|panne\_courant = 0.5) &= P(panne\_courant) \times \\ &\quad P(machine\_arrete|panne\_courant) + P(machine\_arrete|\overline{panne\_courant}) \\ &= 0.5 \times 1.05 = 52.5\% \end{aligned}$$

Cet exemple montre les types de calculs qui sont effectués dans le réseau. On voit que ces calculs sont simples, mais pour de grands réseaux, ils peuvent rapidement devenir très complexes et coûteux en temps. Cependant, pour des réseaux avec quelques centaines de nœuds, un ordinateur actuel peut réussir à donner des résultats dans la seconde.

2.2.3 Les relations entre nœuds

2.2.3.1 Les différents types de nœuds

Un réseau bayésien permet de modéliser plusieurs types de nœuds. Dans le cadre des procédés, nous sommes principalement en présence de deux types de nœuds : un nœud représentant une variable discrète que l'on nomme nœud discret, et un nœud représentant une variable continue que l'on nomme nœud continu. Pour une meilleure compréhension, nous représenterons les différents types de nœud comme indiqué sur la figure 2.3.

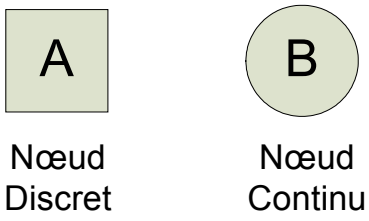


FIG. 2.3 – Représentation des différents types de nœud d'un réseau bayésien

Les réseaux bayésiens traitent de modèles paramétriques. Or, dans le cas discret, un nœud multinomial permet de modéliser toutes fonctions de densité de probabilité d'une variable discrète. En effet, une variable binaire (par exemple Vrai-Faux) peut se représenter grâce à un nœud discret (donc multinomial) de dimension 2 (possédant 2 modalités différentes).

Pour les nœuds continus, il est logiquement possible de pouvoir représenter n'importe quelles fonctions de densité de probabilité d'une variable continue. Mais, à l'heure actuelle, les moteurs d'inférence ne savent traiter qu'une seule fonction de densité de probabilité : celle de la loi normale multivariée de dimension p. Comme nous l'avons déjà vu, toute fonction de densité de probabilité d'une variable continue peut se représenter comme un mélange de plusieurs lois gaussiennes (voir §1.5.8).

Nous allons maintenant aborder les différentes relations entre les nœuds. Il faut tout d'abord énoncer une règle fondamentale : on ne peut pas dresser un arc partant d'un nœud continu vers un nœud discret. Il nous reste tout de même 3 types de relation à étudier : un arc partant d'un nœud discret vers un autre nœud discret, un arc partant d'un nœud discret vers un nœud continu et enfin un arc partant d'un nœud continu vers un autre nœud continu.

2.2.3.2 Arc entre 2 variables discrètes

Prenons le cas de 2 variables discrètes multinomiales A et B de dimension respective a et b (avec $a_1, a_2, \ldots, a_a$ les différentes modalités de A, et $b_1, b_2, \ldots, b_b$ les différentes modalités de B). En dressant un arc partant de A vers B, on doit alors compléter la table de probabilités conditionnelles de B (table 2.1).

	B			
A	b_1	b_2	$\ldots$	b_b
a_1	$P(b_1\|a_1)$	$P(b_2\|a_1)$	$\ldots$	$P(b_b\|a_1)$
a_2	$P(b_1\|a_2)$	$P(b_2\|a_2)$	$\ldots$	$P(b_b\|a_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_a	$P(b_1\|a_a)$	$P(b_2\|a_a)$	$\ldots$	$P(b_b\|a_a)$

TAB. 2.1 – Table de probabilités conditionnelles nœud discret avec parent discret

On voit que l'utilité de la table de probabilités conditionnelles est de répertorier toutes les informations nécessaires à l'inférence dans un réseau. On s'aperçoit également que la taille de cette table est de $a \times b$. Donc, pour des variables avec beaucoup de modalités, elle peut devenir très importante. On voit que le remplissage de cette table peut devenir problématique. En effet, la taille d'une table de probabilités d'un nœud discret X de taille x, ayant p parents (discrets) $Y_1, Y_2, \ldots, Y_p$ de tailles respectives $y_1, y_2, \ldots, y_p$, est de $x \prod_{i=1}^{p} y_i$.

2.2.3.3 Arc entre une variable discrète et une variable continue

Prenons le cas de 2 variables où A est une variable discrète multinomiale de dimension a, et où B est une variable continue de paramètres μ_B et Σ_B. En dressant un arc partant de A vers B, on doit alors compléter la table de probabilités conditionnelles de B comme indiqué dans la table 2.2.

A	B
a_1	$P(B\|a_1) = N(\mu_{a_1}, \Sigma_{a_1})$
a_2	$P(B\|a_2) = N(\mu_{a_1}, \Sigma_{a_2})$
$\vdots$	$\vdots$
a_a	$P(B\|a_a) = N(\mu_{a_a}, \Sigma_{a_a})$

TAB. 2.2 – Table de probabilités conditionnelles nœud continu avec parent discret

La table de probabilités conditionnelles de B se compose de lois conditionnées aux modalités de A. En effet, la table de probabilité d'un nœud continu X, ayant p parents (discrets) $Y_1, Y_2, \ldots, Y_p$ de tailles respectives $y_1, y_2, \ldots, y_p$, est de $\prod_{i=1}^{p} y_i$ lois continues.

2.2.3.4 Arc entre 2 variables continues

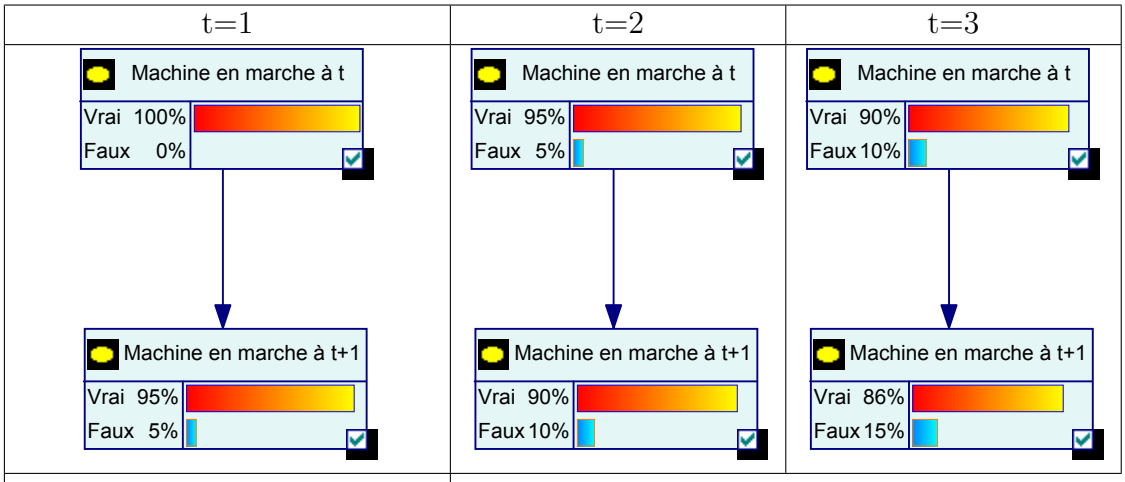
Prenons le cas de 2 variables continues A et B de paramètres respectifs μ_A, Σ_A et μ_B, Σ_B. En dressant un arc partant de A vers B, on effectue alors une régression et l'on peut écrire la loi régissant B pour une valeur a de A comme étant une loi gaussienne de paramètres $(\mu_B + \beta \times a; \Sigma_B)$, où β représente le coefficient de régression.

2.2.4 Extensions des réseaux bayésiens

2.2.4.1 Les réseaux bayésiens dynamiques

Nous présentons ici une classe particulière de réseaux bayésiens : les réseaux bayésiens dynamiques [106, 107]. Les réseaux bayésiens dynamiques sont des réseaux bayésiens intégrant la notion de temps. C'est-à-dire qu'une variable peut influencer sa propre valeur à l'instant suivant (voir figure 2.4).

Si on considère un ensemble de variables $\mathbf{D(t)} = \{D_1(t), D_2(t), \ldots, D_n(t)\}$ évoluant dans le temps, un réseau bayésien dynamique représente la distribution de probabilité jointe de ces variables pour un intervalle borné $[0, T]$. En général, cette distribution peut être codée par un réseau bayésien statique avec T$\times n$ variables. Si le processus est stationnaire, les hypothèses d'indépendance et les probabilités conditionnelles associées sont identiques pour tous les temps t. Dans ce cas, le réseau bayésien dynamique peut être représenté par un réseau bayésien dont la structure est dupliquée pour chaque pas de temps. Un nœud représente donc une variable aléatoire dont la valeur indique l'état occupé à l'instant t.



	Machine en marche à t+1	
Machine en marche à t	Vrai	Faux
Vrai	0.95	0.05
Faux	0	1

FIG. 2.4 – Exemple de réseau bayésien dynamique

2.2.4.2 Réseaux Bayésiens Orientés Objet

Un inconvénient des réseaux bayésiens est le fait qu'ils soient spécifiques. En effet, un réseau bayésien développé pour une application est difficilement transposable vers une autre application. Par analogie à la programmation orientée objet, des chercheurs ont proposé l'utilisation de Réseaux Bayésiens Orientés Objet [73]. Les réseaux bayésiens orientés objet sont de puissants outils de modélisation de la connaissance pour de larges systèmes. Ils permettent la réutilisation de certains éléments du réseau, de même qu'une meilleure visualisation graphique de celui-ci. Les réseaux bayésiens orientés objet permettent de simplifier la représentation graphique d'un réseau bayésien dans le sens où certaines parties du réseau bayésien sont regroupées en un seul objet nommé instance. Une instance contient une partie d'un réseau bayésien, avec des nœuds d'interface : nœuds d'entrée et de sortie. Une instance doit communiquer avec les autres nœuds du réseau bayésien ou bien avec d'autres instances du réseau. Les nœuds d'entrée sont représentés en pointillé, alors que les nœuds de sortie sont représentés en gras. Les autres nœuds de l'instance n'appartiennent qu'à celle-ci et sont donc représentés classiquement lorsque l'on étudie l'instance. Cependant, les nœuds classiques de l'instance sont cachés lorsque l'on représente le réseau bayésien général. Un exemple d'instance est donné sur la figure 2.5.

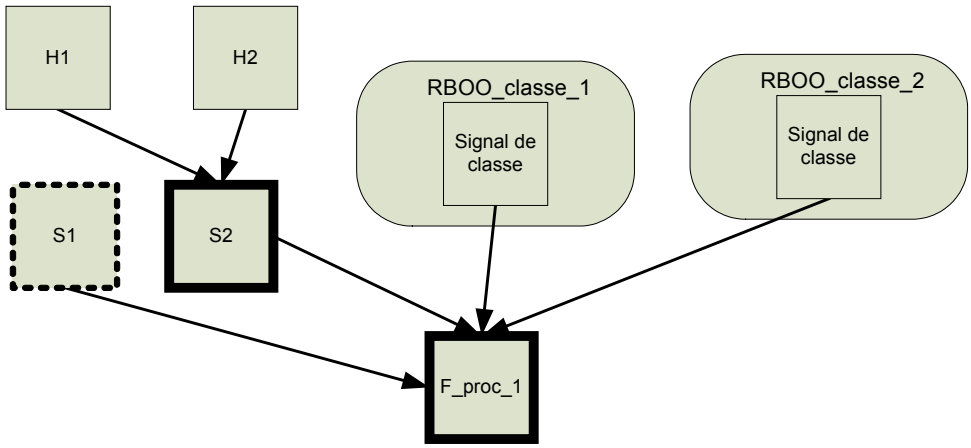


FIG. 2.5 – Exemple d'une instance

Sur la figure 2.5, on s'aperçoit que cette instance est composée elle-même d'autres instances. On remarque également les nœuds classiques, les nœuds de sortie en gras, ainsi que le nœud d'entrée en pointillé. Pour comprendre le concept de réseaux bayésiens orientés objet, la figure 2.6 présente un réseau bayésien avec l'instance présentée précédemment.

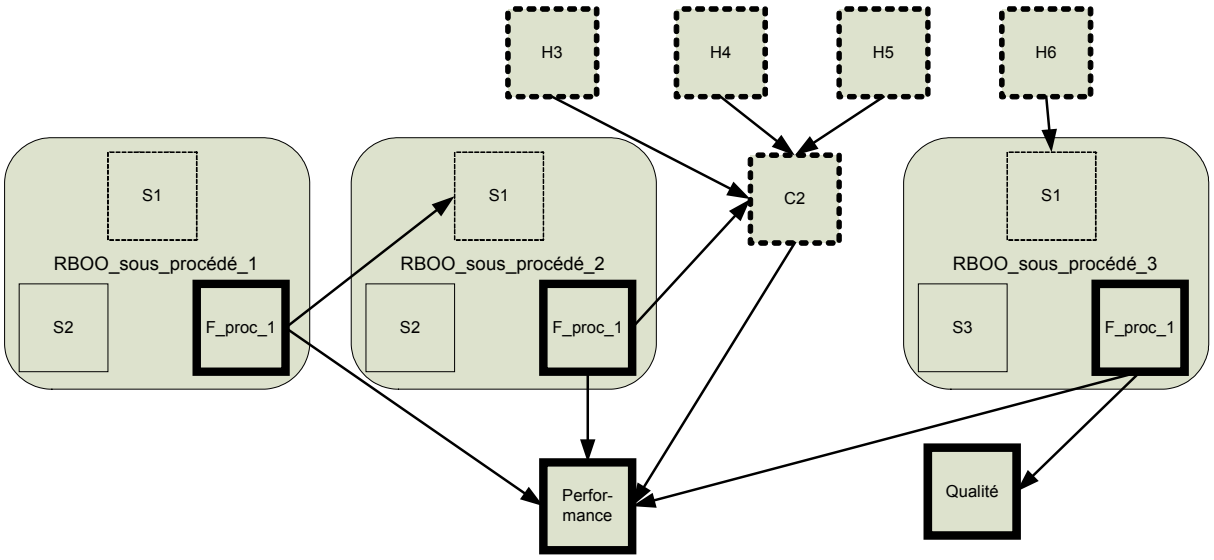


FIG. 2.6 – Exemple d'un RBOO exploitant l'instance de la figure 2.5

La figure 2.6 montre l'utilisation que l'on peut faire des instances dans un réseau. Le réseau bayésien complet mis sous forme de modélisation objet est en réalité une description hiérarchique du système étudié, permettant une meilleure vision des valeurs pertinentes associées au système.

2.2.4.3 Diagramme d'influence

En aide à la décision, un diagramme d'influence est une représentation graphique et mathématique de problèmes d'inférence et de décision. Les diagrammes d'influences sont une généralisation des arbres de décision (voir §1.5.5). Un diagramme d'influence se représente sous la forme d'un graphe acyclique dirigé. Il peut comporter quatre types de nœud : nœud de décision, nœud d'utilité, nœud probabiliste et nœud déterministe. Or, un nœud déterministe n'est autre qu'un nœud probabiliste où l'une des modalités est sûre à 100 %. Ainsi, on comprend rapidement que si l'on ajoute des nœuds utilité et des nœuds décision à un réseau bayésien, ce réseau représente un diagramme d'influence. Un nœud utilité est un nœud permettant d'associer une valeur numérique aux états constitués par la combinaison des différentes modalités de ses parents. Ces valeurs numériques représentent alors la qualité ou le coût de ces états. Un nœud de décision est un nœud multimodal où chaque modalité représente une action influençant le système (et donc le résau). Ainsi, chacune des actions est décrite par l'intermédiaire des tables de probabilités associées aux nœuds enfants.

Un diagramme d'influence est très intéressant car il permet d'étudier les différentes réactions d'un système modélisé par réseaux bayésiens en fonction des actions prises sur le système. Ainsi, grâce à l'utilité, il est possible de comparer les performances de telle ou telle action sur le système. De plus, en combinant des algorithmes d'apprentissage par renforcement et des simulations de type Monte Carlo, il est possible de définir certaines politiques d'apprentissage pour le réseau. Ainsi, il est possible d'optimiser les différentes décisions à prendre sur le système : quelle décision choisir, et à quel instant. La figure 2.7 présente un exemple de diagramme d'influence. Ce réseau bayésien dynamique représente un système de distribution de fluide modélisé par le logiciel BayesiaLab [5]. Les nœuds utilité sont représentés par des losanges, les nœuds de décision sont représentés par des carrés, alors que les nœuds probabilistes (nœuds classiques du réseau) sont représentés par des ronds.

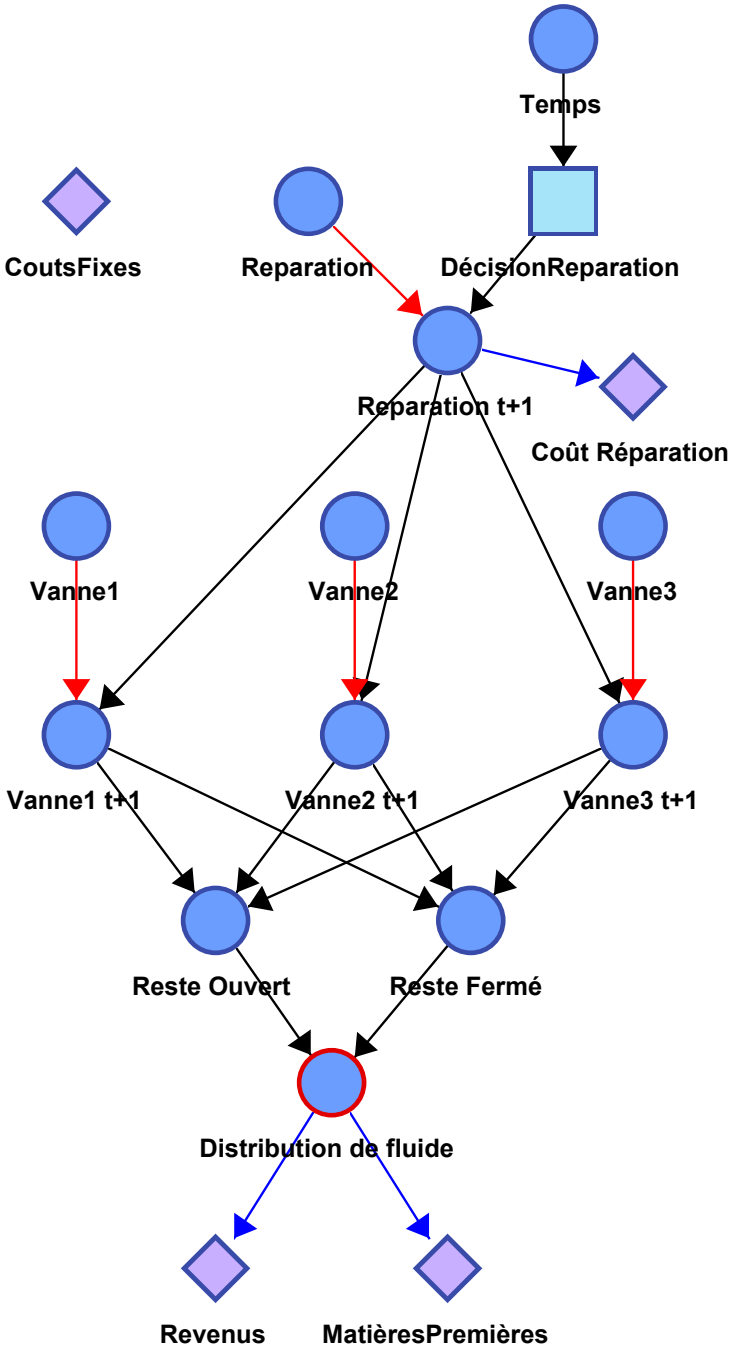


FIG. 2.7 – Exemple d'un diagramme d'inluence sous BayesiaLab

2.3 Réseaux bayésiens et diagnostic : état de l'art

Le but de cette partie est d'étudier les différentes approches proposées dans la littérature, utilisant les réseaux bayésiens dans le cadre du diagnostic des systèmes. Nous avons classé ces approches en trois principales catégories, à savoir : défauts de capteurs, exploitation des données de fautes, exploitation des données de fonctionnement normal.

2.3.1 Méthodes pour les défauts de capteurs

Le diagnostic des défauts de capteurs se situe dans le contexte des procédés complexes. En effet, comme nous l'avons déjà exposé, le contrôle d'un procédé complexe implique la surveillance de plusieurs de ses variables. Afin d'obtenir la mesure correspondant à ces variables, on utilise généralement des capteurs. Un capteur peut se décomposer en plusieurs éléments tels que : un organe sensoriel, un transducteur, un générateur de signal, une interface de communication, etc. Le problème est que les composants des différents capteurs peuvent être défaillants. Dans le cas d'une défaillance d'un des composants d'un capteur, celui-ci fournit alors à l'organe de contrôle une mesure incorrecte de la variable à surveiller. Bien que le procédé puisse être sous-contrôle, une mesure biaisée fournie par un ou plusieurs capteurs défaillants entraîne des décisions inappropriées de la part de l'organe de contrôle. Ces décisions peuvent avoir des impacts minimes sur le procédé (production non-conformes, etc), mais elles peuvent également être à l'origine de conséquences redoutées (mauvais contrôle d'un réacteur nucléaire, etc). Pour répondre à ce problème de capteurs défaillants, il est conseillé de mettre en place une maintenance préventive adaptée et contrôlée. Mais, cela est insuffisant et il est nécessaire d'utiliser un système de redondance. Il est possible d'utiliser une redondance physique, c'est à dire mesurer une variable par plus d'un capteur. Pour la détection de défaillances, une redondance de deux capteurs par variable est suffisante, mais pour le diagnostic, un minimum de trois capteurs est nécessaire. Une autre possibilité est la redondance analytique permettant l'étude de la différence entre l'état du capteur à un instant donné et l'état de celui-ci dans les conditions normales d'utilisation (pas de défaillance des capteurs). Plusieurs outils ont été proposés concernant cette approche : filtres de Kalman [156], méthodes à base d'observateurs [112], relations de parité [53], analyse en composantes principales [162], réseaux de neurones [131]. Nous étudions ici les méthodes utilisant les réseaux bayésiens. Pour cela, nous présentons, dans un premier temps, les différentes structures proposées dans la littérature, puis dans un deuxième temps, nous présentons certaines extensions de ces structures.

2.3.1.1 Différentes structures proposées

Nous présentons quatre structures permettant le diagnostic de défaillances (ou défauts) de capteurs dans un procédé complexe. Les trois premières structures modélisent directement le capteur, alors que la dernière modélise le statut du résidu associé au capteur. Dans leurs travaux Mehranbod et al. [102] dressent la liste de tâches majeures correspondant aux diagnostic de défauts de capteurs :

tâche T1 : détection d'une faute dans le capteur une fois la faute apparue,

tâche T2 : diagnostic du type de faute

tâche T3 : réalisation simultanée des tâches T1 et T2.

On peut alors étudier si les structures proposées permettent ou non d'effectuer ces tâches. Les approches étudiées se basent (implicitement ou explicitement) sur les mêmes hypothèses suivantes :

- La première hypothèse est que le procédé pour lequel le diagnostic de défauts de capteurs est pratiqué soit sous-contrôle. C'est à dire qu'il n'y ait pas de changement dans le procédé (stable), sur les actionneurs et/ou les contrôleurs. En effet, pour surveiller si les capteurs sont défaillants ou non, le procédé lui-même ne doit pas être défaillant, car dans ce cas on ne peut pas conclure si c'est le ou les capteurs qui sont défaillants, ou bien si c'est le procédé.
- La seconde hypothèse concerne la disponibilité des mesures. L'hypothèse est faite qu'il est possible d'obtenir et d'exploiter les données du capteur.
- La troisième hypothèse est la disponibilité d'un modèle du procédé : soit un modèle analytique du procédé, soit un modèle statistique obtenu à partir des données en fonctionnement normal (voir §1.3.2).
- La dernière hypothèse est la supposition qu'un même capteur n'est pas soumis simultanément à plusieurs types de défaillances.

Modèle de Rojas-Guzman et Kramer Ce premier modèle a été proposé par Rojas-Guzman et Kramer en 1993 [125]. Une présentation de ce modèle est fournie sur la figure 2.8.

Sur cette figure, on voit que le modèle de Rojas-Guzman et Kramer [125] est composé de quatre nœuds. Le nœud X_a correspond à la valeur vraie de la variable a. Le nœud B_a représente le biais associé à la variable a. Le nœud N_a représente le bruit associé à la variable a. Enfin, R_a représente la valeur de la variable a fournie par le capteur (valeur disponible). Bien entendu, la valeur vraie de a (X_a) est connue (méthode basée sur un modèle du procédé). De plus, il est supposé que le biais B_a est nul lorsque le capteur n'est pas défaillant, et que le bruit N_a est également connu en l'absence de défaillance. Ainsi, si

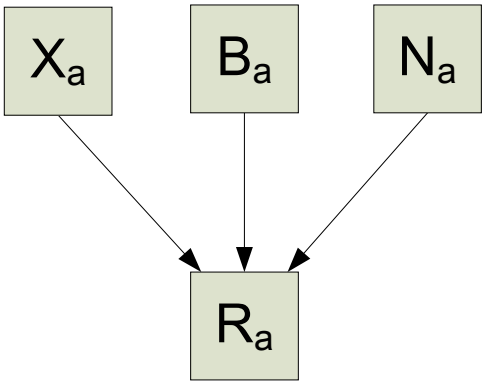


FIG. 2.8 – Modèle proposé par Rojas-Guzman et Kramer

une défaillance se produit sur le capteur, entraînant une valeur mesurée R_a anormale, les nœuds N_a et B_a donnent alors des probabilités plus élevées sur les modalités correspondant aux états défaillants. Les tables de probabilités conditionnelles associées à chaque nœud sont obtenues par simulation Monte Carlo. Ce type de structure permet théoriquement d'effectuer la tâche T3. Cependant, bien que la tâche T1 soit facile sur ce type de modèle, la tâche T2 demande une différenciation très difficile entre les nœuds B_a et N_a.

Modèle d'Aradhye Un deuxième modèle a été proposé par Aradhye en 1997 [1]. Une présentation de ce modèle est donnée sur la figure 2.9.

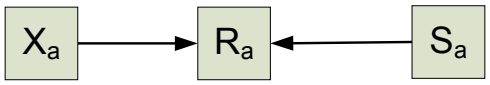


FIG. 2.9 – Modèle proposé par Aradhye

Cette figure 2.9 montre que le modèle proposé par Aradhye est composé de trois nœuds : X_a, R_a et S_a. Les nœuds X_a et R_a correspondent aux même nœuds que ceux du modèle de Rojas-Guzman et Kramer [125], alors que le nœud S_a représente l'état du capteur. Ce nœud est en fait un regroupement des nœuds B_a et N_a du modèle de Rojas-Guzman et Kramer, permettant alors de contourner le problème de différenciation que possédait ce modèle. Le nombre de modalités du nœud S_a proposé par Aradhye dépend de l'application voulue : 2 modalités pour la détection, et plus de 2 pour le diagnostic. Ainsi, afin d'effectuer le diagnostic des défauts de capteurs, Aradhye propose par exemple 4 modalités pour le nœud S_a : 1 modalité pour le mode normal (sans défaillance), et 3 modalités représentant 3 types de défaillances du capteur. La détection (tâche T1) est réalisée sur l'observation de la probabilité a posteriori de la modalité représentant le fonctionnement sans défaillance, alors que le diagnostic est fait en étudiant les probabilités

a posteriori des 3 autres modalités. Ce modèle est alors capable d'effectuer la tâche T3 (T1 et T2 simultanément). Cependant, Aradhye choisit des défaillances qui sont facilement distinguables, alors que dans la réalité ce n'est pas le cas. De plus, il ne traite pas le problème du seuil des différentes fautes : à partir de quelle probabilité de la modalité sans défaillance peut on dire qu'une défaillance est présente, et à partir de quels seuils peut-on dire que la défaillance apparue provient de telle ou telle défaillance spécifique.

Modèle de Mehranbod et al. Le troisième modèle étudié est celui proposé par Mehranbod et al. [102] en 2003. Il s'agit sans doute de la proposition la plus complète et la plus intéressante parmi les trois modélisations de capteur proposées. On peut voir ce modèle sur la figure 2.10.

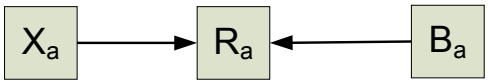


FIG. 2.10 – Modèle proposé par Mehranbod et al.

Le modèle est composé de trois nœuds : X_a, R_a et B_a. Ces trois nœuds sont les mêmes que ceux du modèle de Rojas-Guzman et Kramer [125]. Le nœud B_a joue un rôle essentiel dans l'analyse du capteur. En effet, la détection et le diagnostic des défaillances se fait grâce à l'exploitation des différentes probabilités a posteriori du nœud B_a. La détection (tâche T1) se fait à un instant donné, alors que le diagnostic (tâche T2) nécessite une étude de l'évolution des probabilités a posteriori des différentes modalités de B_a. Ainsi, les tâches T1 et T2 ne peuvent pas être effectuées au même instant (d'où l'incapacité pour la tâche T3). La combinaison de ce modèle pour chaque capteur permet un diagnostic complet des capteurs du procédé. La figure 2.11 représente un réseau bayésien permettant la surveillance de quatre capteurs (F_W, T_j, T et C_m).

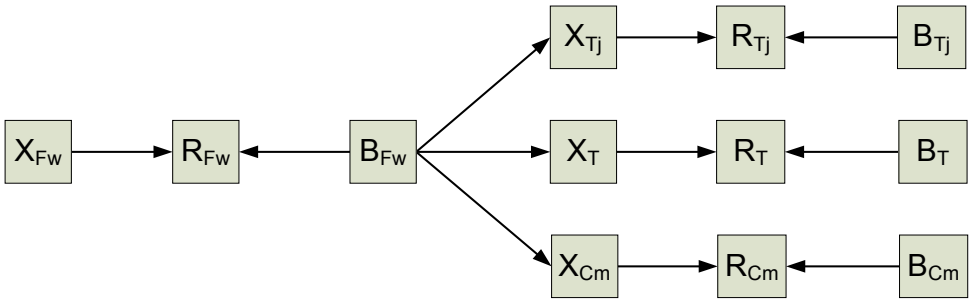


FIG. 2.11 – Structure du réseau bayésien pour un procédé à quatre capteurs

Pour la détection d'éventuelles défaillances sur un capteur, les auteurs introduisent deux mesures. La première est la différence absolue de probabilité : PAD_{ij}. Il s'agit de

la différence, pour une modalité donnée, entre la probabilité a priori et la probabilité a posteriori. Elle est donc définie comme ceci :

$$PAD_{ij} = |P(B_a)_{ij} - P^*(B_a)_{ij}| \tag{2.3}$$

où $P^*(B_a)_{ij}$ et $P(B_a)_{ij}$ représentent respectivement les probabilités a priori et a posteriori de la modalité j du nœud B_a pour le capteur i.

La seconde mesure introduite est la somme des différences absolues de probabilité : S_{PAD_i}. Elle est définie ainsi :

$$S_{PAD_i} = \sum_{j=1}^{m} PAD_{ij} \qquad i = 1, \ldots, n \tag{2.4}$$

où n est le nombre de capteurs du procédé. Dans le cas où $n = 1$, une somme des différences absolues de probabilité non-nulle implique alors que le capteur est soumis à une défaillance. Cependant, dans le cas d'un procédé avec plusieurs capteurs, la corrélation entre les différentes variables implique que la somme des différences absolues de probabilité soit non-nulle même dans le cas où tous les capteurs sont opérationnels (sans défaillance). Pour remédier à ce problème, Mehranbod et al. [102] proposent alors l'indice de détection de faute D_i suivant :

$$D_i = 100 \frac{S_{PAD_i}}{\sum_{j=1}^{n} S_{PAD_i}} \qquad i = 1, \ldots, n \tag{2.5}$$

Le capteur i est déclaré opérationnel si $D_i > T_d$ et défaillant dans le cas contraire, où T_d est un seuil fixé par simulations de différents scénarios de défaillance.

Concernant le diagnostic, Mehranbod et al. [102] attribuent la faute à la modalité de B_a dont les probabilités ont le plus changé par rapport à la situation sans défaillance. Les auteurs précisent alors que la conclusion sur le diagnostic de la défaillance ne peut se faire qu'en surveillant la modalité la plus probable sur plusieurs échantillons, et donnent alors quelques exemples d'évolution des modalités.

Bien que cette approche soit intéressante, on peut tout de même émettre des doutes concernant l'applicabilité de la méthode. En effet, la partie identification est quasi inexistante : bien qu'ils introduisent la notion de modalité la plus probable, ils ne décrivent pas de méthodes concrètes permettant d'exploiter correctement cet aspect. Un autre point négatif est le fait que beaucoup de valeurs du modèle sont déterminées empiriquement : le nombre de modalité de chaque nœud, la largeur de chaque intervalle représentée par une modalité, ainsi que les probabilités a priori. Or, tous ces paramètres influencent fortement les résultats que l'on peut obtenir sur la surveillance des capteurs. De plus, chaque table de

probabilités conditionnelles des différents nœuds est obtenue par simulation Monte-Carlo, sans en connaître davantage les détails.

Modèle de Weber et al. Un autre modèle, proposé par Weber et al. [157], se différencie des trois premiers modèles étudiés. En effet, Weber et al. n'essaient pas de modéliser directement les capteurs du procédé, ils modélisent les différents résidus tirés d'une approche classique à base de modèle analytique du procédé (voir §1.3.2.1). Une fois les résidus obtenus, un test statistique est effectué sur chacun d'entre eux pour savoir si le résidu est statistiquement proche de 0 (valeur pour laquelle le procédé est en fonctionnement normal). Ainsi, pour chaque type de faute du procédé, on liste les différents résidus affectés par celle-ci. On dresse alors une matrice d'incidence comme indiqué sur la table 2.3. Dans cette table, on voit que la faute F_1 implique des résidus non-nuls sur les mesures des capteurs 2 et 3.

	F_1	F_2	F_3
u_1	0	1	0
u_2	1	0	0
u_3	1	0	1

TAB. 2.3 – Exemple de matrice d'incidence

Ainsi, lors de la surveillance du procédé, suite à la génération de résidus et des différents tests statistiques, le vecteur de cohérence U, dont les composantes sont les différents résultats des tests statistiques, est directement comparé aux différents vecteurs de fautes (ou signatures) composant la matrice d'incidence. Par exemple, si le vecteur de cohérence est $U = [1\ 0\ 0]^T$, il est possible de conclure que la faute F_2 s'est produite.

Le principe de modélisation de matrice d'incidence par réseaux bayésiens proposé par Weber et al. [157] est intéressant. Chaque faute est représentée par un nœud F_i à deux modalités : "présente" et "non présente". De plus, chaque statut u_j de résidus (les composantes du vecteur de cohérence U) est représenté par un nœud à deux modalités : "détectée" ou "non détectée". Ensuite, on dresse des arcs partant d'une faute F_i vers les statuts de résidus correspondant à la signature de cette faute. Pour illustrer ceci, la figure 2.12 représente sous forme de réseau bayésien la matrice d'incidence de la table 2.3.

Bien entendu, une fois la structure du réseau bayésien défini, il faut également définir les paramètres de chaque table de probabilités conditionnelles associée à chaque nœud u_i. Ces tables dépendent des taux de fausses alertes α_i et des taux de détections manquées β_i de la faute F_i. Par exemple, on donne la table 2.4 de probabilités conditionnelles associée au nœud u_1 (ne dépendant que du nœud F_2) de la matrice d'incidence de la table 2.3.

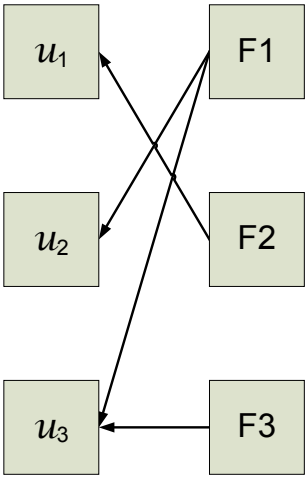


FIG. 2.12 – Réseau bayésien modélisant la matrice d'incidence de la table 2.3

	u_1	
F_2	non détectée	détectée
non présente	$1-\alpha_2$	α_2
présente	β_2	$1-\beta_2$

TAB. 2.4 – Table de probabilités conditionnelles du nœud u_1

Bien que cette table soit relativement simple, la complexité des différentes tables de probabilités conditionnelles augmente avec le nombre de fautes impliquées par un statut de résidus u_j. Toujours sur le même exemple, la table 2.5 de probabilités conditionnelles associée à u_3 est beaucoup plus complexe.

		u_3	
F_1	F_3	non détectée	détectée
non présente	non présente	$1-(\alpha_1+\alpha_3+\alpha_1\alpha_3)$	$\alpha_1+\alpha_3+\alpha_1\alpha_3$
non présente	présente	$\beta_3-\alpha_1\beta_3$	$1-(\beta_3-\alpha_1\beta_3)$
présente	non présente	$\beta_1-\alpha_3\beta_1$	$1-(\beta_1-\alpha_3\beta_1)$
présente	présente	$\beta_1\beta_3$	$1-\beta_1\beta_3$

TAB. 2.5 – Table de probabilités conditionnelles du nœud u_3

Suite à l'introduction d'évidences (valeurs des différents statuts de résidus : u_i) dans le réseau, l'algorithme d'inférence donne alors les probabilités de défaillances de chaque capteur. Cette méthode est intéressante, mais souffre de quelques lacunes. En effet, aucun seuil n'est donné afin de conclure sur les différentes fautes de capteur. Par exemple, une probabilité de défaillance de capteur égale à 50% traduit-elle que le capteur est défectueux ? Une autre remarque, émise par Weber et al. eux-même, est que l'obtention de tous les paramètres (taux de fausses alarmes et taux de détections manquées) est difficile.

2.3.1.2 Extensions

Extension du modèle de Mehranbod et al. Le premier modèle proposé par Mehranbod et al. [102] permet la détection et le diagnostic des défauts de capteurs dans les procédés complexes à états stables. Ces auteurs proposent une extension de leur modèle [101] permettant la prise en compte d'états transitoires dans le procédé.

Les auteurs utilisent les réseaux bayésiens dynamiques afin de modéliser l'aspect transitoire du procédé, ainsi que des nœuds adaptables. Un exemple simple d'utilisation de ce nouveau modèle est un procédé de remplissage de réservoir, où le débit entrant q est fonction de la hauteur de liquide h, mais la hauteur de liquide est elle-même fonction du débit entrant. Pour représenter ce procédé avec le modèle de Mehranbod et al. [102] proposé précédemment, nous aurions la structure de la figure 2.13.

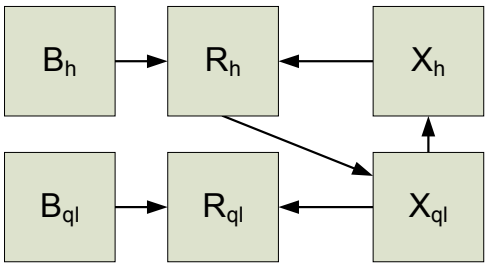


FIG. 2.13 – Structure logique du réseau bayésien pour un procédé de remplissage

On s'aperçoit tout de suite que la structure d'un tel réseau n'est pas viable pour un réseau bayésien : en effet, celle-ci introduit un cycle dans le graphe alors qu'un réseau bayésien doit être défini par un graphe acyclique. Pour représenter ce type d'interactions, Mehranbod et al. [101] proposent alors un réseau bayésien dynamique permettant de supprimer le cycle. Pour l'exemple de la figure 2.13, on peut alors émettre le graphe de la figure 2.14.

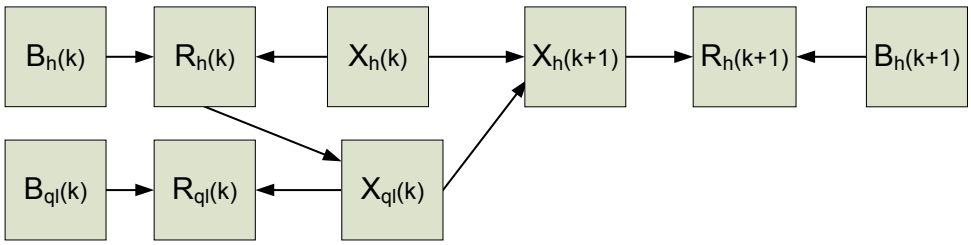


FIG. 2.14 – Structure réelle du réseau bayésien pour un procédé de remplissage

En plus d'utiliser les réseaux bayésiens dynamiques, les auteurs proposent d'utiliser également des nœuds adaptables. Ces nœuds permettent de changer la largeur des différentes modalités et ainsi d'évoluer en même temps que le procédé. Cette extension est

intéressante, elle permet de prendre en compte des phénomènes transitoires dans le procédé. Mais, elle ne comble aucun des défauts cités précédemment pour la première version de leur modèle.

Extension du modèle de Weber et al. Weber et al. [157] se proposent d'ajouter à leur modèle un autre type de connaissance : la fiabilité des composants liés aux différentes fautes (fiabilité des capteurs).

Afin de modéliser la fiabilité par l'intermédiaire d'un réseau bayésien, Weber et al. [157] se basent sur une méthode exposée par Weber et Jouffe [158]. Dans Weber et Jouffe [158], les auteurs démontrent la modélisation de la fiabilité d'un composant au cours du temps grâce à un réseau bayésien. Pour cela, les auteurs s'appuient sur l'équivalence entre une chaîne de Markov en temps discret et un réseau bayésien dynamique. Or, une chaîne de Markov en temps discret permet de calculer la fiabilité d'un composant à un instant donné. Prenons le cas d'un composant ayant deux états, fonctionne (OK) ou hors-service (HS), alors on peut dresser la chaîne de Markov du processus de dégradation du composant grâce à son taux de défaillance λ, telle qu'illustrée sur la figure 2.15.

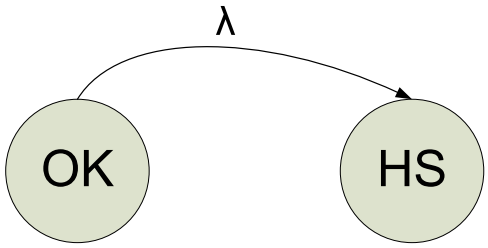


FIG. 2.15 – Chaîne de Markov d'un composant à deux états possibles, avec un taux de défaillance λ

Ainsi, la probabilité d'être dans l'état HS augmente peu à peu avec le temps et l'état OK reflète la fiabilité du composant (probabilité qu'il fonctionne à un instant donné sachant sa probabilité de fonctionner à l'instant précédent). La représentation de ce phénomène par un réseau bayésien dynamique est donnée sur la figure 2.16. On représente un nœud N_t pour un composant. Ce nœud N_t possède deux modalités correspondant aux deux états possibles du composant : OK et HS. Nous précisons qu'un composant peut tout à fait posséder plus de deux états et dans ce cas le nœud le représentant comporte autant de modalités que d'états possibles du composant. Pour obtenir une chaîne de Markov, il suffit de dupliquer ce nœud, mais à l'instant précédent (nous notons ce nœud N_{t-1}). Ensuite, il suffit de dresser un arc entre le nœud N_{t-1} et le nœud N_t puisque l'état du composant à l'instant $t-1$ influence l'état du composant à l'instant t.

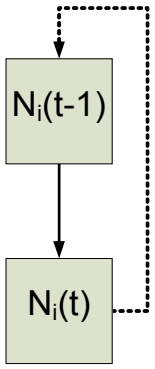


FIG. 2.16 – Réseau bayésien dynamique représentant la chaîne de Markov de la figure 2.15

On peut donner la table 2.6 de probabilités conditionnelles de N_t, qui dépend de N_{t-1}. On remarque que le modèle proposé ne tient pas compte du taux de réparation, mais ceci est également envisageable.

	N_t	
N_{t-1}	OK	HS
OK	$1-\lambda$	λ
HS	0	1

TAB. 2.6 – Table de probabilités conditionnelles du nœud N_t

Le nœud N_t donne les probabilités à un instant donné que le composant soit fonctionnel (OK) ou hors-service (HS). A l'instant suivant, ces probabilités deviennent celles du nœud N_{t-1} (cette copie de probabilité est représentée par un trait pointillé sur la figure 2.16) et les nouvelles probabilités de N_t sont calculées grâce aux probabilités de N_{t-1} et de la table de probabilités conditionnelles de N_t (table 2.6).

Cette nouvelle connaissance concernant la fiabilité du système est intégrée dans le modèle proposé par Weber et al. [157]. L'intégration de la fiabilité de chaque capteur est modélisée comme indiqué sur la figure 2.17.

L'arc entre la fiabilité du composant et la variable représentant la faute associée induit une table de probabilités conditionnelles. Cette table (voir table 2.7) est relativement simple puisque lorsque le composant (capteur) est OK, la faute n'est pas présente, alors que si le composant est HS, la faute est présente.

Cette approche est très intéressante car elle ajoute un nouveau type de connaissance (la fiabilité des différents capteurs) au modèle. Cependant, cette extension ne permet pas de combler les lacunes relevées sur le modèle de base, et notamment le problème de seuil permettant de prendre une décision sur l'état des capteurs.

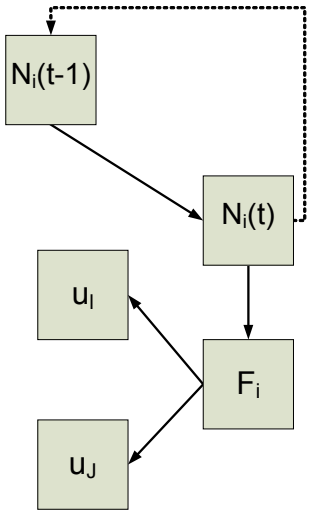


FIG. 2.17 – Modèle de Weber et al. intégrant la fiabilité par réseau bayésien dynamique

	F_i	
N_t	non présente	présente
OK	1	0
HS	0	1

TAB. 2.7 – Table de probabilités conditionnelles du nœud F_i

2.3.2 Méthodes basées sur les exemples de fautes

Dans la littérature, on trouve des approches de diagnostic de système basées sur des jeux de données de fautes. Nous étudions deux approches représentant des cas extrêmes, à savoir : beaucoup de données sont disponibles [89], et peu de données sont disponibles [34].

2.3.2.1 Peu de données disponibles

Des travaux intéressants concernant le diagnostic des procédés par réseaux bayésiens ont été effectués par Dey et Stori [34]. Les auteurs s'intéressent à la surveillance des différents paramètres d'un procédé de production. Plus précisément, Dey et Stori [34] se placent dans le contexte des machines outils (fraiseuse, tour, etc) et s'intéressent particulièrement aux outils de ce type de procédé de fabrication. L'objectif de leurs travaux est de diagnostiquer les causes physiques à l'origine de la variation du procédé, et ce même lorsque plusieurs sources de variations peuvent être à l'origine d'un même problème (par exemple, la dureté du matériau et l'usure de l'outil peuvent être à l'origine d'un même phénomène de production dégradée). Ils exposent de manière très succincte et globale la méthodologie permettant d'effectuer le diagnostic de certaines situations sur

une machine-outil : étudier les relations de causes à effets, construire le réseau, apprendre les paramètres du réseau à partir d'une base de données, acquérir et rentrer de nouvelles évidences (observations) dans le réseau bayésien, propagation et mise à jour des croyances (probabilités).

Afin de bien comprendre les aboutissants de leur méthode, Dey et Stori [34] présentent un exemple en détail. Ils se placent dans le cas d'une pièce à usiner sur une machine outil (une fraiseuse en l'occurrence). La gamme opérationnelle de la pièce à usiner comporte deux phases : un surfaçage et un perçage. Les différentes sources de variations de ce procédé sont : les variations dimensionnelles, les variations de dureté de pièce, les variations dues à l'usure de la fraise, ainsi que les variations dues à l'usure du foret. Ils utilisent un capteur d'émission acoustique, ainsi qu'un capteur de puissance de broche. Chacun de ces capteurs permet d'obtenir plusieurs mesures caractéristiques du procédé : 8 mesures pour le capteur d'émission acoustique, et 7 mesures pour le capteur de puissance de broche.

Afin d'étudier les relations de causes à effets du procédé, Dey et Stori [34] mènent deux plans d'expériences (un pour chaque phase de la gamme opérationnelle), suivi d'une analyse de la variance. Grâce à un ensemble de règles établis par les auteurs, différents résultats de l'analyse de la variance sont exploités afin de dresser un tableau (voir table 2.8) donnant, pour chaque mesure de chaque capteur, les différentes sources de variations pouvant être diagnostiquées par cette mesure. Les différentes causes étudiées sont : les Variations Dimensionnelles (VD), la Dureté des Pièces (DP), ainsi que l'Usure de l'outil en Surfaçage (US) et en Perçage (UP). Une fois le tableau établi, la structure du réseau bayésien est donnée sur la figure 2.18.

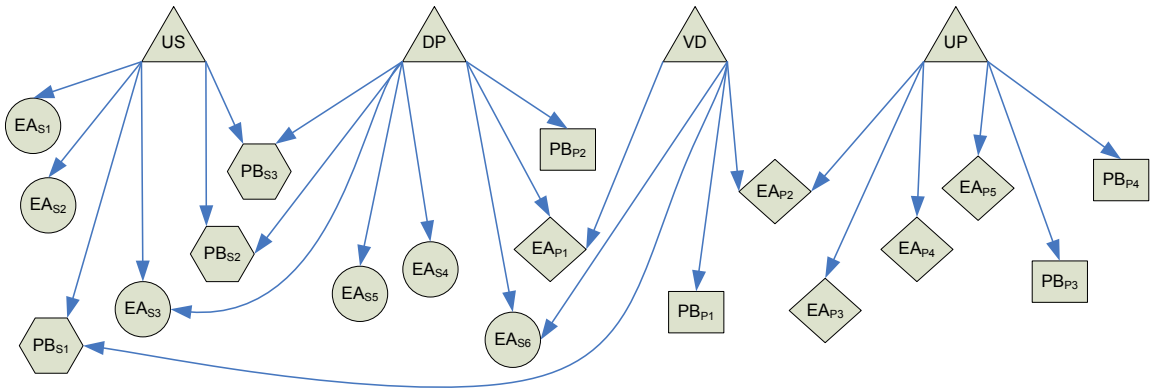


FIG. 2.18 – Structure du réseau bayésien pour le diagnostic des causes premières

Dans la figure 2.18, les 4 nœuds représentant les causes premières de variation sont des variables discrètes à deux modalités : modalité "élevé" et modalité "faible". Les nœuds représentant les mesures des capteurs sont définis comme étant des nœuds discrets avec

Procédé	Capteur	Nom	Métrique	ANAVAR
Surfaçage	Emission Acoustique	EA_{S1}	Ecart-type	US
		EA_{S2}	Pulsation	VD, US
		EA_{S3}	Densité spectrale	VD, US, VDxUS
		EA_{S4}	Moyenne de pics	VD
		EA_{S5}	Fréquence de pics	VD
		EA_{S6}	Moyenne	VD, US, VDxUS
	Puissance Broche	PB_{S1}	Moyenne	DP, US
		PB_{S2}	Ecart-type	VD, US
		PB_{S3}	Moyenne de pics	VD, US
Perçage	Emission Acoustique	EA_{P1}	Moyenne de pics	VD
		EA_{P2}	Pulsation	DPxUP
		EA_{P3}	Moyenne	UP
		EA_{P4}	Densité spectrale	UP
		EA_{P5}	Ecart-type	UP
	Puissance Broche	PB_{P1}	Ecart de moyenne	DP
		PB_{P2}	Moyenne	VD
		PB_{P3}	Ecart-type	UP
		PB_{P4}	Densité spectrale	UP

TAB. 2.8 – Tableau récapitulant l'étude de la variance effectuée

deux ou trois modalités suivant les cas. La structure du réseau étant défini, Dey et Stori [34] proposent l'apprentissage des paramètres en se basant sur les données récoltées lors des deux plans d'expérience, et en utilisant un a priori de Dirichlet [61] sur les nœuds. Dey et Stori [34] évaluent la performance en diagnostic du réseau sur 18 nouveaux essais. A un niveau de confiance de 80%, le réseau bayésien permet de diagnostiquer correctement les causes de variations dans 10 cas sur 18. A 70%, le nombre de bon classement est de 16, et de 17 à 60%.

L'approche proposée par Dey et Stori [34] est intéressante. En effet, l'utilisation de plans d'expériences pour établir la structure et les paramètres du réseau est originale. Des corrélations existent sans doute entre les différents descripteurs, leur prise en compte ne serait donc pas négligeable. Cependant, trouver les corrélations intéressantes, ainsi qu'estimer les paramètres du modèle les prenant en compte, est difficile avec uniquement 32 exemples. Malgré l'originalité de leurs travaux, Dey et Stori [34] laissent trop de points en suspens. En effet, certains nœuds descripteurs sont découpés en 2 modalités et certains en 3, sans en connaître la raison. De plus, aucune explication n'est donnée concernant la discrétisation des nœuds en 2 ou 3 modalités : quelle plage de variations, quel découpage, etc. Finalement, les auteurs ne statuent pas sur les différentes décisions à prendre en fonction des différents pourcentages de croyance obtenus aux 4 nœuds de diagnostic. Or,

il est indispensable de fixer des seuils de probabilités permettant de respecter un certain nombre de fausses alarmes et de diagnostics manqués.

2.3.2.2 Nombre important de données disponibles

Li et Shi [89] ont développé une approche de surveillance par réseaux bayésiens exploitant un historique du procédé. Les auteurs se placent dans le contexte de la surveillance d'une seule caractéristique qualité : un nombre de non-conformités. Une non-conformité peut être, par exemple, une rayure, une tâche, etc. On s'intéresse alors à suivre le nombre de ces non-conformités pour un produit donné.

Plusieurs phases préliminaires précèdent la construction du réseau. Tout d'abord, une sélection de variables est effectuée afin de ne prendre en compte que les variables pouvant potentiellement influencer la caractéristique qualité à surveiller. Cette sélection est faite sur avis d'expert. La seconde phase préliminaire est la discrétisation des données. Pour cela, Li et Shi [89] préconisent l'utilisation de l'algorithme de discrétisation proposé par Dougherty et al. [39]

En faisant la supposition implicite qu'un jeu de données très important est disponible, les auteurs proposent la construction d'une structure causale pour le réseau bayésien. Ils se basent alors sur l'algorithme PC (Peter and Clark) [135], auquel ils apportent quelques modifications pour permettre à un expert du procédé d'intervenir également dans la construction de la structure du réseau. Les paramètres de chaque nœud sont alors appris grâce au jeu de données.

Li et Shi [89] fournissent l'exemple d'un procédé de laminage de barres. Ils s'intéressent alors aux non-conformités de la surface, qui sont des fissures sur la barre laminée. Le jeu de données à disposition pour cet exemple est impressionnant puisque les enregistrements de 100 000 barres laminées sont disponibles pour 22 variables. Une sélection des variables importantes est effectuée par le responsable du procédé, et 7 variables sont retenues comme pouvant provoquer des fissures sur les barres. Suite à la discrétisation des données, l'algorithme de construction de la structure est utilisé, et les paramètres correspondant sont appris (réseau de la figure 2.19).

Li et Shi [89] expliquent que le réseau permet le diagnostic et la prédiction. Ainsi, en entrant comme évidence une modalité du nœud représentant la caractéristique qualité, il est possible d'identifier les modalités des variables les plus responsables. De même, on peut utiliser le réseau en phase de prédiction : en attribuant des évidences sur les variables du procédé, on peut prédire les répercussions sur la caractéristique qualité.

Les travaux de Li et Shi [89] sont intéressants, cependant l'application proposé se restreint dans tous les cas à des valeurs discrètes, de l'information est donc forcément

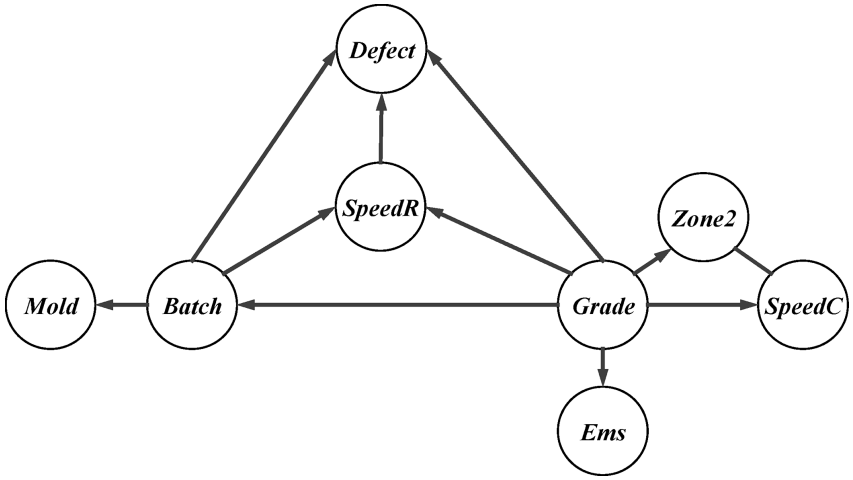


FIG. 2.19 – Structure du réseau bayésien pour le procédé de laminage

perdue durant la discrétisation. De plus, les auteurs se basant sur un jeu de données des différentes fautes, il serait intéressant de baser la sélection de variables importantes sur ces données. Enfin, il faut avouer qu'il est rare d'avoir un jeu de données si important. De plus, pour utiliser ce type d'approche, il faut être sûr que tous les types de fautes du procédé ont déjà été identifiés, et qu'ils sont disponibles dans le jeu de données.

2.3.3 Approches basées sur les données du mode normal

Contrairement à celles étudiées précédemment, les approches présentées dans cette section permettent de prendre en compte les situations suivantes :

- le procédé est d'une telle complexité que les ingénieurs sont incapables de répertorier les différentes fautes possibles du procédé,
- et/ou les fautes sont tellement rares que leurs exemples ne peuvent être utilisés ni pour l'apprentissage de la structure, ni pour l'apprentissage des paramètres.

Dans ce contexte, un diagnostic concret n'est pas envisageable. Le but de ces approches est de fournir un diagnostic non-supervisé : identifier les variables impliquées dans la faute.

2.3.3.1 Variables discrétisées

Nielsen et al. [109] proposent la mise au point d'un réseau bayésien représentant uniquement le mode de fonctionnement normal du procédé. Pour cela, les auteurs se basent sur l'algorithme d'apprentissage causal proposé par Cheng et al. [16]. Tous les nœuds du réseau sont supposés discrets. La discrétisation employée sur les variables continues est une discrétisation par validation croisée, tentant de maximiser l'estimation de la vraisemblance des données.

Nielsen et al. [109] introduisent la notion de conflit de l'évidence. Cette mesure établit le degré d'adéquation d'une évidence (observation à diagnostiquer) par rapport au mode normal de fonctionnement. La mesure de conflit pour une évidence $\boldsymbol{e} = \{e_1, \cdots, e_p\}$, où e_i représente la valeur d'observation de la variable i, est définie par :

$$conf(\boldsymbol{e}) = \log\left(\frac{P(e_1), \cdots, P(e_p)}{P(\boldsymbol{e})}\right) \tag{2.6}$$

Les auteurs précisent alors que pour une observation tirées du mode normal, la valeur de conflit est négative, et que plus la corrélation entre les différentes variables est élevée, et plus la mesure du conflit est négative. Ainsi, si une mesure de conflit prend une valeur positive, cela traduit l'apparition d'une faute dans le procédé. Cependant, les auteurs soulignent également le fait qu'une valeur négative extrêmement faible (beaucoup plus faible que les valeurs de conflit en fonctionnement normal) implique également une faute dans le procédé. Nielsen et al. [109] préconisent alors l'utilisation d'un seuil négatif (en plus du seuil représenté par la valeur 0) permettant, en cas de dépassement, de conclure sur la détection d'une faute dans le procédé. Cependant, les auteurs ne précisent pas comment choisir ce seuil.

Pour le diagnostic, les auteurs recommandent une recherche itérative des différentes observations responsables de la faute détectée : identifier la variable contribuant le plus au dépassement du seuil, l'enlever de l'itération et recommencer ainsi de suite jusqu'à ce que la valeur de conflit repasse sous le seuil fixé.

1. Soit t le seuil d'alerte (normalement fixé à 0).
2. Soit $\boldsymbol{e}$ l'observation déclarée comme hors-contrôle.
3. Répeter
   (a) Sélectionner
   $$e' = \underset{e}{\operatorname{argmax}} \log\left(\frac{P(e)}{P(e|\boldsymbol{e}\backslash\{e\})}\right) \tag{2.7}$$
   (b) Soit $\boldsymbol{e} = \boldsymbol{e}\backslash\{e'\}$
4. Jusqu'à ce que $conf(\boldsymbol{e}) < t$.

L'algorithme précédent permet donc d'identifier les variables responsables de la faute. Ces variables sont celles dont l'évidence a été sélectionnée comme e' dans la troisième étape de l'algorithme. Cette approche possède certains inconvénients. En effet, travailler avec des variables discrétisées (comme toutes les autres approches étudiées jusqu'à présent) impliquent une perte d'information. De plus, le calcul des seuils pour la détection d'une faute est encore une fois très peu défini. Par exemple, le seuil de 0 dans le cas d'un système

avec des variables fortement corrélées n'est, d'après Nielsen et al. [109], pas le plus adapté. De même, Nielsen et al. [109] ne donnent aucun renseignement pour le réglage du seuil négatif.

2.3.3.2 Variables continues

Une approche permet de combler les inconvénients (variables discrètes et seuils non réellement définis) de l'approche de Nielsen et al. [109]. Pour cela, Li et al. [88] proposent une méthode permettant d'améliorer l'efficacité de la décomposition MYT (voir §1.4.2.3), en se basant sur l'utilisation de réseaux bayésiens causaux. Les auteurs étudient la méthode MYT qui décompose un signal T^2 en termes indépendants, rendant alors possible l'identification des variables responsables d'une situation hors-contrôle. Les auteurs soulignent le fait que cette méthode est très intéressante, mais qu'elle est sujette à un inconvénient majeur : le nombre de termes à calculer. En effet, comme nous l'avons déjà expliqué dans la section 1.4.2.3, la méthode MYT impose un nombre de décompositions égale à $p!$ (où p est le nombre de variables du procédé). Or, ces décompositions impliquent alors un total de $p \times 2^{p-1}$ termes distincts à calculer. Par exemple, pour un procédé à 20 variables, plus de 10 millions de termes distincts sont à calculer. Des efforts furent effectués afin de réduire le nombre de termes en appliquant un algorithme en 5 étapes [97] permettant une réduction significative du nombre de termes à calculer. Cependant, Li et al. [88] font la remarque que même avec l'utilisation de l'algorithme, le nombre de termes à calculer est tout de même important (très supérieur à p, notamment en présence de fautes multiples). Les auteurs présentent alors une méthode exploitant les réseaux bayésiens. Un graphe causal représentant le procédé permet alors de réduire le nombre de termes à calculer à p. En plus de la diminution de calcul engendrée par cette méthode, les auteurs précisent que la performance en diagnostic est également améliorée.

L'hypothèse de base de la méthode proposée est que le procédé peut être modélisé sous la forme d'un réseau bayésien causal où chaque variable du procédé est une variable gaussienne univariée. Lorsqu'un réseau bayésien représente uniquement des variables continues normales, il est également appelé modèle linéaire gaussien. Ainsi, pour un procédé à 3 variables, on peut par exemple obtenir le réseau bayésien de la figure 2.20.

Dans le cadre de la modélisation du procédé par un modèle linéaire gaussien, les auteurs font la distinction entre deux types de décomposition MYT : "pour une décomposition du T^2 donnée, s'il existe un terme $T^2_{i\bullet 1,\ldots,i-1}$ tel que l'ensemble de variables $\{X_1, \ldots, X_{i-1}\}$ contient au moins un descendant de X_i, alors cette décomposition est de type A, dans le cas contraire, la décomposition est de type B". Ainsi, nous pouvons classer dans la table 2.9 les différentes décompositions du procédé à 3 variables de la figure 2.20.

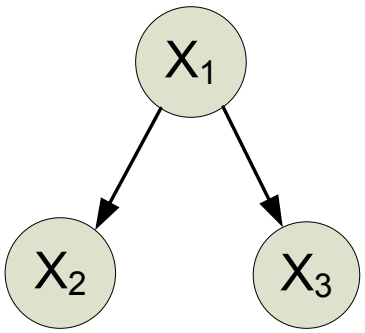


FIG. 2.20 – Exemple d'un modèle causal linéaire gaussien

Décomposition	Type
$T^2 = T^2_1 + T^2_{2\bullet 1} + T^2_{3\bullet 1,2}$	Type B
$T^2 = T^2_1 + T^2_{3\bullet 1} + T^2_{2\bullet 1,3}$	Type B
$T^2 = T^2_2 + T^2_{1\bullet 2} + T^2_{3\bullet 1,2}$	Type A
$T^2 = T^2_2 + T^2_{3\bullet 2} + T^2_{1\bullet 2,3}$	Type A
$T^2 = T^2_3 + T^2_{1\bullet 3} + T^2_{2\bullet 1,3}$	Type A
$T^2 = T^2_3 + T^2_{2\bullet 3} + T^2_{1\bullet 2,3}$	Type A

TAB. 2.9 – Types des décompositions du procédé à 3 variables

Li et al. [88] prouvent, en se basant sur les travaux d'Hawkins [57], que les décompositions de type A permettent un diagnostic moins précis que les décompositions de type B. De plus, ils prouvent également que dans le contexte du modèle linéaire gaussien, toutes les décompositions de type B convergent vers une unique décomposition que les auteurs nomment "causation-based T^2 decomposition". Nous la nommerons décomposition causale du T^2. En effet, chaque décomposition de type B (dans le cas d'un modèle linéaire gaussien causal) converge vers la décomposition causale du T^2 décrite dans l'équation 2.8, où $PA(X_i)$ représentent les parents de la variable X_i sur le graphe causal.

$$T^2 = \sum_{i=1}^{p} T^2_{i\bullet PA(X_i)} \tag{2.8}$$

Ainsi, la décomposition causale du T^2 de l'exemple de la figure 2.20 est la suivante : $T^2 = T^2_1 + T^2_{2\bullet 1} + T^2_{3\bullet 1}$.

Suite à ces différentes démonstrations, les auteurs énoncent alors la procédure de détection et de diagnostic utilisant la nouvelle décomposition causale. Tout d'abord, un réseau bayésien linéaire gaussien est construit afin de représenter les relations causales entre les différentes variables du procédé. Suite à cela, le procédé est surveillé par une carte de contrôle du T^2 (voir §1.4.2.2). Lors de la détection d'une situation hors-contrôle, le T^2 est décomposé par la décomposition causale de l'équation 2.8. Dans cette équation, chaque

$T^2_{i\bullet PA(X_i)}$ est indépendant et, dans le cas où les paramètres du procédé sont connus, suit une distribution du χ^2 à un degré de liberté. On compare alors chaque $T^2_{i\bullet PA(X_i)}$ à la limite $\chi^2_{1,\alpha}$ représentant le quantile à la valeur α (taux de fausses alertes) de la distribution du χ^2 à un degré de liberté. Un $T^2_{i\bullet PA(X_i)}$ significatif (dépassant la limite de contrôle) implique alors que la variable X_i a probablement subi un saut de moyenne. La figure 2.21 représente le diagramme de surveillance du procédé par la méthode énoncée ci-dessus.

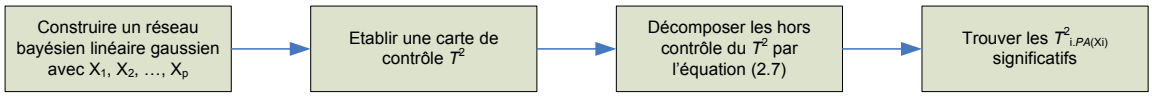


FIG. 2.21 – Surveillance par la méthode de décomposition causale

Afin de démontrer la performance de cette approche, les auteurs utilisent comme exemple un procédé à 5 variables, représenté par le réseau bayésien causal (modèle linéaire gaussien) de la figure 2.22, où la variable X_5 représente une caractéristique qualité, alors que les 4 autres sont des variables du procédé.

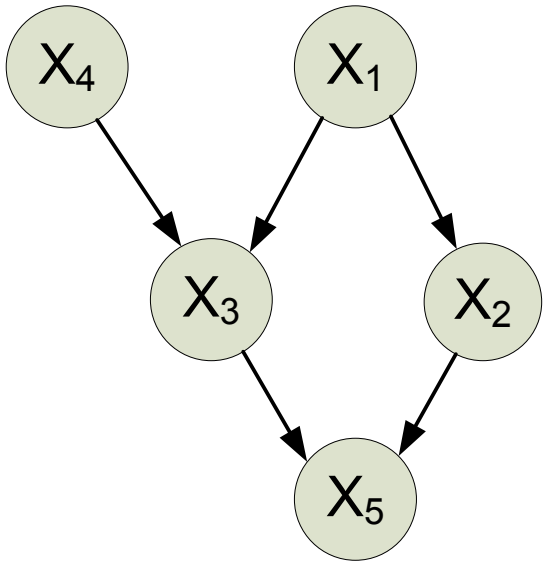


FIG. 2.22 – Modèle linéaire gaussien du procédé

Dans ce contexte (5 variables), il existe alors 31 situations potentielles de fautes : 5 fautes potentielles uniques (1 seule variable incriminée) et 26 fautes potentielles multiples (plusieurs variables incriminées). Chaque faute est représentée comme un saut de moyenne de 3 écart-types sur chaque variable incriminée. Une fois la situation hors-contrôle détectée, elle est diagnostiquée. Si le diagnostic correspond exactement au scénario simulé, alors le diagnostic est considéré comme correct, sinon il est considéré comme erroné. Les 31 diagnostics permettent alors de calculer la performance de la méthode (nombre de

diagnostic correct divisé par le nombre de scénarios). Les auteurs obtiennent, par simulation, la moyenne et l'écart-type de la performance de la méthode. Sur les mêmes données, les auteurs appliquent la procédure MYT enrichie de l'algorithme proposé par Mason et al. [97], afin d'obtenir également le critère de performance de cette méthode. Les résultats sont donnés sur la table 2.10, montrant bien que la performance de la méthode de décomposition causale est beaucoup plus élevée que celle de la décomposition MYT (presque le double).

Méthode	Moyenne	Ecart-type
Décomposition causale	73.6%	7.7%
Décomposition MYT	37.1%	7.4%

TAB. 2.10 – Performances des méthodes

L'approche développée par Li et al. [88] est la seule approche permettant la prise en compte de variables continues. De plus, cette approche exploite des seuils donnés par des quantiles de lois statistiques. Enfin, elle permet d'améliorer considérablement les performances de diagnostic par rapport à la méthode MYT, tout en demandant moins de calcul que celle-ci. Cependant, quelques points seraient à éclaircir. L'approche proposée possède exactement la même idée sous-jacente à l'approche MYT, ou bien à l'approche de Hawkins [56], à savoir la régression des variables du procédé. La différence ici est que l'on apporte une information supplémentaire : les relations causales entre les différentes variables du procédé. Les auteurs se servent du réseau bayésien causal comme base de leur méthode, mais à aucun moment ils ne précisent comment obtenir ce graphe. De plus, cette méthode permet juste l'identification des variables responsables d'une situation hors contrôle, mais elle n'effectue pas la détection par réseau bayésien.

2.3.4 Conclusions

Dans toutes les approches étudiées, les réseaux bayésiens sont vus comme un outil de modélisation du fonctionnement du système. Evidemment, on remarque que c'est un point fort concernant les réseaux bayésiens : les modèles formés possèdent une lisibilité certaine comparés à d'autres approches telles que les réseaux de neurones. De plus, pour des réseaux incluant plusieurs centaines de variables, une représentation sous forme de réseau bayésien orienté objets est très adaptée à une vue décomposée du système entier. Une telle décomposition permettant alors de se focaliser sur des parties cibles à travailler, étudier ou exploiter [159, 160].

Certaines approches entrevues se basent sur un modèle analytique du procédé [1, 125, 157]. Or, dans le cas de la surveillance à partir des données, soit le modèle analytique du

procédé n'est pas disponible, soit l'on veut, pour certaines raisons pratiques, s'en affranchir. Ces approches ne sont donc pas directement transposables au sujet nous intéressant. On peut tout de même tirer quelques conseils concernant l'application de réseaux bayésiens à la surveillance des procédés basée sur les données. Les approches étudiées mettent en lumière quelques points essentiels à étudier, ou tout du moins des points à prendre en compte dans l'approche que nous voulons mettre en place.

- Toutes les approches étudiées (exceptée celle de la décomposition causale [88]) utilisent des nœuds discrets dans leur réseaux. Or, la discrétisation des variables implique une perte d'information. Cependant, les réseaux bayésiens sont capables de prendre en compte des variables continues (sous l'hypothèse de normalité). Il serait donc souhaitable d'utiliser ce type de nœud.
- Nous avons également mis en évidence que dans toutes les approches étudiées, les différents seuils de probabilités permettant de conclure sur l'état du procédé posent problème. Ceci vient principalement du fait que ces approches utilisent des nœuds discrets et qu'il est difficile de fixer des seuils probabilistes sur ce type de nœud. La seule approche étudiée utilisant un seuil statistique exact est celle de Li et al. [88]. Cependant, pour cette approche, le seuil n'est pas directement un seuil de probabilité puisque les décisions impliquées par cette approche ne s'effectuent pas directement dans le réseau bayésien.
- Dans les méthodes se basant sur les données historiques du procédé, nous avons vu que deux types d'approches étaient envisageables : approches se basant sur les données de fautes [34, 89], et approches se basant uniquement sur les données du mode normal de fonctionnement [88, 109]. Il serait intéressant d'exploiter ces deux types d'approches dans un seul et même réseau.
- Dans l'approche de décomposition causale du T^2, proposée par Li et al. [88], les auteurs n'apportent aucune information sur la construction du modèle linéaire gaussien. Au vu des résultats encourageant de leur approche, une étude de ce type de construction serait appréciable.
- Un dernier point permettant de comprendre correctement une approche est l'explication des tables de probabilités conditionnelles. En effet, aucun des travaux étudiés (exceptés ceux de Weber et al. [157]) ne donnent réellement les tables de probabilités conditionnelles du réseau construit. Or, un réseau bayésien ne donnent des résultats cohérents que si ses différentes tables de probabilités sont correctement remplies.

2.4 Conclusion

Ce chapitre nous a permis d'introduire plus précisément les réseaux bayésiens. Nous avons vu que ce type d'outil permettait de prendre en compte des variables discrètes et continues et qu'il modélisait de façon probabiliste l'incertitude des connaissances d'un problème (liens entre les variables, probabilités conditionnelles, etc). Nous avons alors étudié dans la littérature plusieurs approches exploitant les réseaux bayésiens dans le contexte du diagnostic des procédés. Ceci nous a permis de voir qu'aucune approche déjà développée ne permettait réellement de pouvoir envisager par réseaux bayésiens une surveillance de procédé basée sur les données. Cependant, quelques points clés ont été mis en évidence afin d'envisager cette alternative.

Chapitre 3

Réseaux bayésiens pour la surveillance des procédés

Sommaire

3.1 Introduction

L'objectif de ce chapitre est de présenter les différentes contributions que nous apportons au domaine de la surveillance des procédés, dans le cadre de l'utilisation des réseaux bayésiens. Premièrement, nous allons étudier la réalisation d'un principe de détection par

réseaux bayésiens dans la section 3.2. Ensuite, si une faute est détectée, nous allons montrer comment un réseau bayésien va pouvoir discriminer entre plusieurs types de fautes dans la section 3.3. Pour cela, nous introduirons également le domaine de la sélection de variables pertinentes pour la discrimination, ainsi que les contributions apportées dans ce domaine. Cependant, un diagnostic supervisé n'est pas suffisant puisque celui-ci implique l'apport d'exemples de faute. Or, en début de production, l'organe de surveillance doit également pouvoir fournir des indications sur la faute en présence sans jamais l'avoir vue auparavant. Ainsi, dans la section 3.4, nous étudions la réalisation par réseaux bayésiens d'un diagnostic non supervisé (sans exemple de faute) permettant tout de même de donner des indications sur la faute en présence (notamment les variables impliquées). Enfin, dans la section 3.5, nous présentons la structure complète d'un réseau bayésien dédié à la surveillance des procédés. Ce réseau va permettre la détection de faute, ainsi que le diagnostic supervisé et non-supervisée.

3.2 Détection par réseaux bayésiens

Cette section a pour but de démontrer la réalisation de la détection d'une faute dans un procédé multivarié par l'intermédiaire d'un réseau bayésien. La stratégie choisie est de modéliser dans un réseau bayésien les principes des cartes de contrôle multivariées telles que celle du T^2 de Hotelling ou bien celle de la carte MEWMA. Pour cela, nous allons premièrement expliquer que la détection est une tâche de classification monoclasse, puis nous verrons comment réaliser une classification par un réseau bayésien, et enfin, nous étudierons la transposition des cartes de contrôle multivariées en un réseau bayésien. Lors de cette dernière partie, nous démontrerons alors un résultat important : l'équivalence entre un réseau bayésien et une carte de contrôle.

3.2.1 Détection et classification

La détection, comme nous l'avons déjà définie, consiste à déceler la présence de fautes dans le procédé. Elle a pour but de détecter si celui-ci est soumis à l'effet d'une cause spéciale qui impliquera un accroissement de la variabilité à plus ou moins long terme.

Comme nous l'avons déjà évoqué, la détection peut être considérée comme une classification monoclasse ("one-class classification") [137]. Le but de la classification monoclasse est de décrire une classe d'individus, et de pouvoir distinguer si un nouvel individu appartient ou non à cette classe. Dans le cas de la détection, la classe d'intérêt est celle décrite par des exemples d'individus supposés sous contrôle : la classe décrivant le fonctionnement

normal du procédé. Cependant, lorsque l'on parle de classification, il est évident que l'on cherche à faire la distinction entre plusieurs classes. Or, pour traiter le cas de la classification monoclasse, nous n'avons qu'une seule classe à disposition. Nous devons donc créer, au minimum, une seconde classe que nous appellerons classe virtuelle. La première classe représentant le fonctionnement normal du procédé (classe "Sous Contrôle" notée SC), nous dénommerons la seconde classe (virtuelle) : classe "Hors-Contrôle", notée HC. Ainsi, cette seconde classe représente l'ensemble des individus ne pouvant pas appartenir à la classe SC.

Un exemple typique de classification monoclasse est la carte du T^2 de Hotelling. En effet, un nouvel individu est déclaré comme appartenant à la classe SC si la mesure de son T^2 ne dépasse pas une certaine limite de contrôle LC, alors que cet individu est déclaré comme appartenant à la classe HC si son T^2 dépasse la limite de contrôle LC. Dans ce cas, il est clair que la frontière de décision entre les deux classes SC et HC est représentée par la limite de contrôle LC. Dans le cas de cette carte de contrôle, il est possible de représenter, pour un exemple en deux dimensions, la frontière de décision induite par l'application de la limite de contrôle LC. Sur la figure 3.1, on s'aperçoit que cette frontière est une ellipse entourant la classe de fonctionnement normal du procédé (classe SC), et que tout individu à l'extérieur de cette ellipse est considéré comme un individu appartenant à la classe hors-contrôle HC. Bien entendu, cette remarque est également valable pour des dimensions supérieures à 2. En effet, dans le cas à 3 dimensions, la frontière devient une ellipsoïde (forme d'un ballon de rugby), mais dans le cas de dimension supérieure, cette limite ne peut plus être dessinée ou même imaginée par nos sens physique. Cependant, la limite peut toujours être calculée comme le quantile d'une certaine distribution caractérisant la statistique T^2 (voir §1.4.2.2).

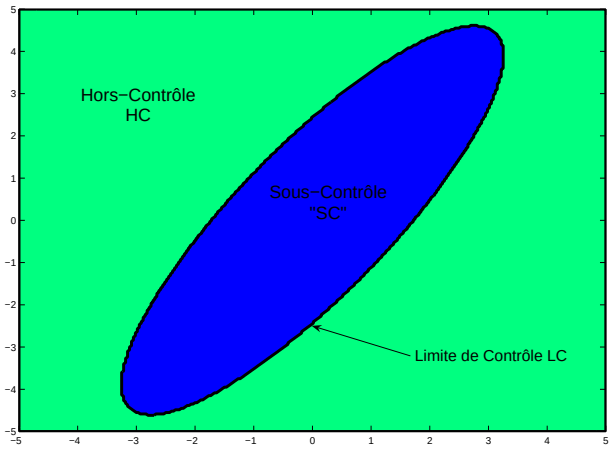


FIG. 3.1 – Frontière de décision de la carte T^2 dans l'espace bivarié

La figure 3.1 montre une chose très importante, la classification engendrée ne peut pas être une classification linéaire. Ainsi, il est impossible d'utiliser une Analyse Discriminante Linéaire (voir §1.5.7.1) pour arriver à une frontière de classification comme celle décrite sur la figure 3.1. Cependant, ce type de frontière est typique d'une Analyse Discriminante Quadratique. De plus, l'Analyse Discriminante Quadratique fait la même hypothèse de normalité des classes que les cartes de contrôle multivariées, hypothèse que d'autres classifieurs ne peuvent pas faire (voir §1.5). Nous nous orientons donc vers ce type de classification, afin de réaliser la classification monoclasse (détection). Pour cela, il nous faut tout d'abord étudier la façon de réaliser une analyse discriminante quadratique dans un réseau bayésien.

3.2.2 Analyse discriminante par réseaux bayésiens

Nous allons pouvons réaliser des analyses discriminantes par réseaux bayésiens. En effet, étant donné que l'inférence dans un réseau bayésien est basée sur la règle de Bayes, et que l'analyse discriminante est également basée sur cette règle de décision, nous pouvons facilement modéliser les fonctions coûts énoncées dans la section 1.5.7.

3.2.2.1 Analyse discriminante complète

La structure pour réaliser des analyses discriminantes classiques (quadratiques, linéaires) sur un système à p variables comprenant k types de fonctionnement peut se modéliser par un nœud continu (variable $\boldsymbol{X}$) normal multivarié de dimension p, relié à un nœud discret (variable C) de dimension k, comme indiqué sur la figure 3.2. Nous avons déjà présenté ce type de réseau, il s'agit d'un réseau bayésien naïf semi-condensé (voir §1.5.9). Ce réseau représente en fait une loi normale multivariée conditionnellement à la classe, tout comme l'est une analyse discriminante.

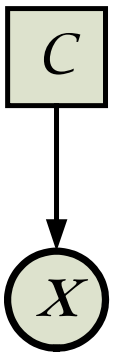


FIG. 3.2 – Analyse discriminante par réseaux bayésiens

Le choix entre les différentes possibilités d'analyse discriminante (quadratique, linéaire, diagonale, sphérique, régularisée, etc) se fait alors sur le choix des k matrices de variance-covariance attribuées à $\boldsymbol{X}$.

On peut également réaliser une analyse discriminante quadratique en faisant apparaître chaque composante X_i de la variable multivariée $\boldsymbol{X}$ puis en reliant toutes ces variables afin de prendre en compte toutes les relations pouvant exister entre les X_i. Un exemple de ceci pour quatre variables est décrit sur la figure 3.3.

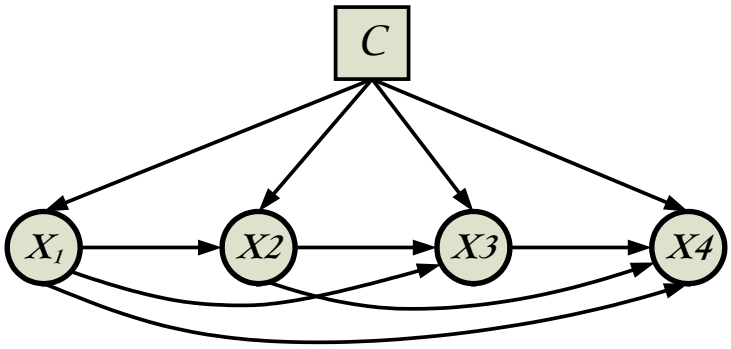


FIG. 3.3 – Analyse discriminante par réseaux bayésiens à variables distinctes

Ce réseau représente en fait une série de $p-1$ régressions linéaires. L'avantage de ce type de réseau est qu'il permet de donner un résultat même si nous n'avons pas les valeurs de chaque variable descriptive. Cependant, l'évaluation de tous les paramètres de régression ainsi que l'inférence demandent plus de calculs que dans le cas précédent.

3.2.2.2 Analyse discriminante à matrice diagonale

Supposons à présent que l'hypothèse de diagonalité des matrices de variance-covariance soit faite. Cela signifie que chaque variable X_i est indépendante des autres variables, excepté la variable de classe C. Dans ce cas, il est toujours possible d'exprimer ceci sous la forme générale vu précédemment (figure 3.2). Il est aussi possible de modéliser une analyse discriminante quadratique diagonale grâce au réseau de la figure 3.4.

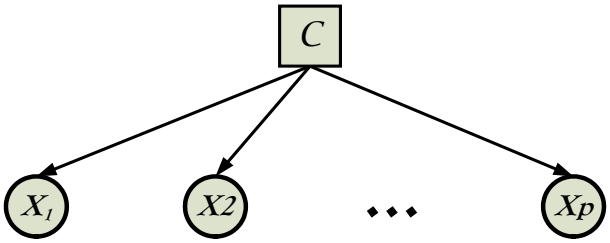


FIG. 3.4 – Analyse discriminante quadratique diagonale

Ce type de réseau se nomme également réseau bayésien naïf, ou classifieur de Bayes, car il fait l'hypothèse naïve que chaque variable X_i est indépendante. Là encore, l'avantage de cette modélisation est le fait de pouvoir obtenir un résultat même si nous n'avons pas les valeurs de chaque X_i.

3.2.2.3 Mélange de gaussiennes

Un réseau bayésien peut également traiter le cas de l'analyse discriminante lorsque les classes ne sont pas gaussiennes. En effet, il est possible d'implémenter l'algorithme EM (Expectation-Maximisation) comme algorithme d'apprentissage (estimation des paramètres des différentes tables de probabilités conditionnelles) d'un réseau bayésien. Ainsi, une analyse discriminante de classes approximées par des mélanges de gaussiennes (voir §1.5.8) se modélise comme sur la figure 3.5.

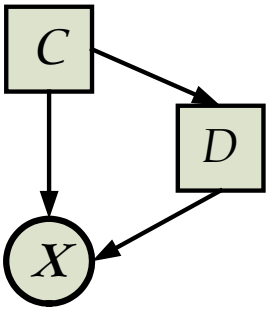


FIG. 3.5 – Mélange de modèles gaussiens par réseaux bayésiens

La dimension du nœud discret D représente le nombre de composantes d pour le mélange, alors que sa table de probabilités conditionnelles va représenter les différents coefficients à attribuer aux gaussiennes définies par $\boldsymbol{X}$, conditionnellement à la classe.

Nous venons de voir que les différentes analyses discriminantes paramétriques sont facilement réalisables dans un réseau bayésien. La section suivante démontre la réalisation des cartes multivariées par réseaux bayésiens, tout en s'appuyant sur la similarité des cartes de contrôle et de l'analyse discriminante.

3.2.3 Cartes multivariées par réseaux bayésiens

Comme nous l'avons vu dans les deux sections précédentes, d'une part il est possible d'assimiler la détection (et notamment le principe des cartes multivariées) à une étape de classification, et d'autre part, il est possible de réaliser une analyse discriminante par réseaux bayésiens. Dans cette section, nous présentons mathématiquement la modélisation des cartes de contrôle multivariées (T^2 et MEWMA) dans un réseau bayésien.

3.2.3.1 Définition de la classe HC

Nous avons vu dans la section 3.2.1 que la détection par carte de contrôle multivariée est assimilable à une classification entre deux classes : la classe de fonctionnement normal (classe SC) et la classe virtuelle représentant tous les individus n'appartenant pas à la première classe (classe HC).

Afin de définir la classe de fonctionnement normal (SC), nous supposons que plusieurs individus supposés sous-contrôle sont disponibles. De même que pour les cartes multivariées, ces individus permettent d'estimer le vecteur des moyennes $\boldsymbol{\mu}$ ainsi que la matrice de variance-covariance $\boldsymbol{\Sigma}$ de notre procédé, lorsque celui-ci est en fonctionnement normal. Concernant la classe virtuelle HC, nous n'avons pas forcement d'individus à disposition. De plus, même si nous en avions, ceux-ci ne suffiraient jamais à couvrir toute la zone de l'espace pour laquelle nous voulons attribuer la classe hors-contrôle HC. En analysant un peu la différence entre les deux classes sur la figure 3.1, on voit qu'il est possible d'assimiler leur centre de classe au même point. Alors la seule chose différenciant les deux classes SC et HC est leur variabilité. En effet, la classe SC possède une variabilité plus faible que la classe HC. Cette remarque nous pousse donc à définir le vecteur des moyennes de la classe HC comme étant le même que celui de la classe SC, mais avec une matrice de variance-covariance exprimant plus de variabilité. Pour cela, nous définissons la matrice de variance-covariance de la classe HC comme étant $c \times \boldsymbol{\Sigma}$ où $\boldsymbol{\Sigma}$ est la matrice de variance-covariance de la classe SC, et c est un coefficient strictement supérieur à 1, permettant ainsi d'augmenter la variabilité de la classe HC par rapport à celle de la classe SC. Nous avons donc les deux classes possédant les paramètres répertoriés dans la table 3.1.

Classe	Distribution
Sous contrôle (SC)	$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Hors-contrôle (HC)	$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, c \times \boldsymbol{\Sigma})$

Tab. 3.1 – Paramètres des classes pour la détection

Cependant, fixer les lois sous-jacentes au fonctionnement du procédé ne suffit pas pour tirer des conclusions sur son état. En effet, la définition de ces paramètres permet l'application d'une analyse discriminante directe ou par réseaux bayésiens. Lorsqu'un nouvel individu $\boldsymbol{x}$ est présenté au classifieur, nous obtenons alors une probabilité $P(SC|\boldsymbol{x})$ que cet individu appartienne à la classe SC et une probabilité $P(HC|\boldsymbol{x})$ que cet individu appartienne à la classe HC (avec $P(SC|\boldsymbol{x}) + P(HC|\boldsymbol{x}) = 1$, puisque seulement deux classes sont présentes). Cependant, comment interpréter ces différentes probabilités, comment prendre une décision correcte en fonction de ces probabilités ? Plusieurs facteurs rentrent en ligne de compte. Comme pour le cas des cartes multivariées, nous devons nous fixer un

certain taux α de fausses alertes. De plus, la valeur du paramètre c définissant la classe HC joue un rôle sur les valeurs des probabilités et donc sur la décision à prendre. Nous allons dans la section suivante nous attacher à adapter les paramètres du réseau bayésien afin d'obtenir la même règle de décision que celle des cartes de contrôle multivariées.

3.2.3.2 Equivalence entre réseaux bayésiens et cartes de contrôle

Après avoir montré l'équivalence qu'il peut y avoir entre un réseau bayésien et une analyse discriminante (voir §3.2.2), nous prouvons ici l'équivalence entre le principe des cartes de contrôle et un réseau bayésien, ou plus précisément, nous définissons les paramètres de celui-ci permettant d'obtenir l'équivalence avec les cartes de contrôle multivariées.

Dans un premier temps, comme pour le cas des cartes de contrôles multivariées et en plus de l'hypothèse de normalité, nous prenons un certain taux α de fausses alarmes. Ainsi, nous nous fixons des probabilités a priori d'être sous-contrôle $P(SC)$ et hors-contrôle $P(HC)$ données par :

$$\begin{aligned} P(SC) &= 1-\alpha \\ P(HC) &= \alpha \end{aligned}$$

Comme pour les cartes de contrôle, nous nous fixons un seuil permettant de prendre une décision : si, pour un individu donné, la probabilité d'être hors-contrôle dépasse α, alors cet individu appartient à la classe HC. Cette règle de décision se traduit donc par : "procédé hors-contrôle si $P(HC|\boldsymbol{x}) > \alpha$", ou de façon équivalente "procédé sous contrôle si $P(SC|\boldsymbol{x}) > 1-\alpha$". Il ne reste qu'un seul paramètre à étudier : le coefficient c permettant l'augmentation de la variabilité de la classe HC. La valeur de ce paramètre permet de créer l'équivalence entre les cartes de contrôle multivariées et le réseau bayésien. L'objectif des développements suivants est donc la définition de c permettant l'équivalence entre la règle de décision fixée pour le réseau bayésien, et la règle de décision des cartes multivariées.

L'objectif est d'obtenir la règle de décision suivante :

$$\boldsymbol{x} \in SC \quad si \quad T^2 < LC \tag{3.1}$$

où T^2 représente la distance de Mahalanobis entre $\boldsymbol{\mu}$ et $\boldsymbol{x}$, à partir de la la règle de décision suivante :

$$\boldsymbol{x} \in SC \quad si \quad P(SC|\boldsymbol{x}) > 1-\alpha \tag{3.2}$$

Développons alors l'inéquation de la seconde règle de décision :

$$\begin{aligned} P(SC|\boldsymbol{x}) &> 1-\alpha \\ P(SC|\boldsymbol{x}) &> (1-\alpha)(P(SC|\boldsymbol{x}) + P(HC|\boldsymbol{x})) \\ P(SC|\boldsymbol{x}) &> (1-\alpha)P(SC|\boldsymbol{x}) + (1-\alpha)P(HC|\boldsymbol{x}) \\ P(SC|\boldsymbol{x}) &> \left(\frac{1-\alpha}{\alpha}\right) P(HC|\boldsymbol{x}) \end{aligned}$$

Or, d'après la loi de Bayes, on a :

$$P(SC|\boldsymbol{x}) = \frac{P(SC)P(\boldsymbol{x}|SC)}{P(\boldsymbol{x})} \tag{3.3}$$

et

$$P(HC|\boldsymbol{x}) = \frac{P(HC)P(\boldsymbol{x}|HC)}{P(\boldsymbol{x})} \tag{3.4}$$

avec $P(SC) = 1-\alpha$ et $P(HC) = \alpha$. Nous obtenons alors :

$$\begin{aligned} \frac{P(SC)P(\boldsymbol{x}|SC)}{P(\boldsymbol{x})} &> \left(\frac{1-\alpha}{\alpha}\right) \frac{P(HC)P(\boldsymbol{x}|HC)}{P(\boldsymbol{x})} \\ \left(\frac{P(SC)}{P(HC)}\right) P(\boldsymbol{x}|SC) &> \left(\frac{1-\alpha}{\alpha}\right) P(\boldsymbol{x}|HC) \\ P(\boldsymbol{x}|SC) &> P(\boldsymbol{x}|HC) \end{aligned} \tag{3.5}$$

Or, dans le cas d'une analyse discriminante à k classes C_i, les probabilités conditionnelles (par rapport aux différentes classes) sont calculées par l'équation 3.6, où ϕ représente la fonction de densité de probabilité de la loi normale multivariée correspondante à la classe.

$$P(\boldsymbol{x}|C_i) = \frac{\phi(\boldsymbol{x}|C_i)}{\sum_{j=1}^{k} P(C_j)\phi(\boldsymbol{x}|C_j)} \tag{3.6}$$

L'inéquation d'équivalence 3.5 s'exprime alors comme :

$$\phi(\boldsymbol{x}|SC) > \phi(\boldsymbol{x}|HC) \tag{3.7}$$

On rappelle que la fonction de densité de la loi normale multivariée de dimension p, de paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$, d'un individu $\boldsymbol{x}$ est donnée par :

$$\phi(\boldsymbol{x}) = \frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \tag{3.8}$$

Si les paramètres de la loi sont $\boldsymbol{\mu}$ et $c \times \boldsymbol{\Sigma}$, alors la fonction de densité devient :

$$\phi(\boldsymbol{x}) = \frac{e^{-\frac{1}{2c}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}c^{p/2}} \tag{3.9}$$

En identifiant l'expression $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ comme étant le T^2 pour l'individu $\boldsymbol{x}$, on écrit les inéquations suivantes :

$$\begin{aligned}
\phi(\boldsymbol{x}|SC) &> \phi(\boldsymbol{x}|HC) \\
\frac{e^{-\frac{T^2}{2}}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} &> \frac{e^{-\frac{T^2}{2c}}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}c^{p/2}} \\
e^{-\frac{T^2}{2}} &> \frac{e^{-\frac{T^2}{2c}}}{c^{p/2}} \\
-\frac{T^2}{2} &> -\frac{T^2}{2c} - \frac{p\ln(c)}{2} \\
T^2 &< \frac{p\ln(c)}{1-\frac{1}{c}}
\end{aligned} \tag{3.10}$$

Or, nous recherchons les valeurs de c permettant d'obtenir une règle de décision équivalente aux cartes de contrôle multivariées, à savoir : $\boldsymbol{x} \in SC \;\; si \;\; T^2 < LC$. Nous obtenons donc l'équation suivante pour c :

$$\frac{p\ln(c)}{1-\frac{1}{c}} = LC \tag{3.11}$$

Cette équation s'exprime sous la forme suivante :

$$1 - c + \frac{pc}{LC}\ln(c) = 0 \tag{3.12}$$

L'équation 3.12 admet deux solutions. L'une des solutions (évidente) est $c = 1$. Mais cette solution n'est pas possible puisque nous avons précédemment défini le paramètre c comme étant strictement supérieur à 1. De plus, attribuer la valeur 1 à c reviendrait à réaliser une analyse discriminante entre deux même classes, ce qui n'aurait pas de sens. L'équation 3.12 admet également une autre racine calculable numériquement. Ainsi, c ne dépendant que de LC et de p, on peut établir des abaques en fonction de la dimension p du système à surveiller et du risque α fixé (en effet, LC est uniquement fonction de p et de α). Nous avons établis ces abaques pour les cartes de contrôle du T^2 de Hotelling ainsi

que pour la carte MEWMA (à différents paramètres λ), pour des valeurs de α de 1% et de 0.5%, et des valeurs de p allant de 1 à 50 variables. Les cartes de contrôle univariés n'étant que des cas particuliers des cartes de contrôle multivariées, la méthode développée ici (ainsi que le calcul du paramètre c) est également valable pour ce type de carte de contrôle. Les différents abaques sont disponibles en annexe A.1. Suivant la carte que l'on veut utiliser, la dimension du système et le risque α choisi, ces abaques donnent la valeur du coefficient c à utiliser afin d'obtenir dans tous les cas une règle de décision comparable à celle de la carte de contrôle choisie.

3.2.3.3 Exemple des cartes T^2 et MEWMA

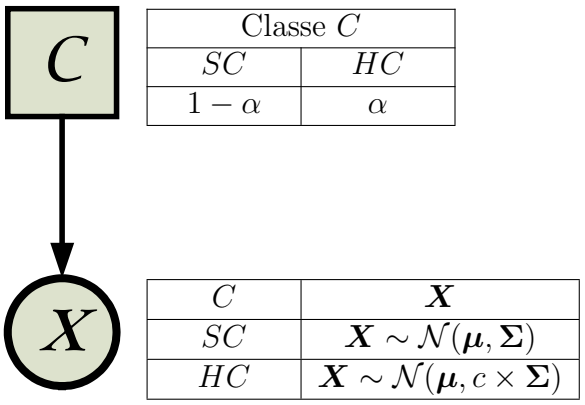
Nous proposons d'illustrer l'approche décrite ci-dessus sur un exemple très simple d'un système à deux dimensions. Nous étudions sur ce système une carte T^2 et sa transposition par un réseau bayésien, ainsi qu'une carte MEWMA (avec $\lambda = 0.1$) et sa transposition par un réseau bayésien, les deux cartes possédant un risque de fausses alertes $\alpha = 1\%$. Lorsque ce système est sous-contrôle, il suit une loi normale multivariée de paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ tels que :

$$\begin{aligned} \boldsymbol{\mu} &= \begin{pmatrix} 5 & 10 \end{pmatrix} \\ \boldsymbol{\Sigma} &= \begin{pmatrix} 1 & 1.2 \\ 1.2 & 2 \end{pmatrix} \end{aligned}$$

Afin de surveiller ce procédé, nous appliquons alors la méthode proposée de détection par réseaux bayésiens. Pour une carte du T^2 de Hotelling, nous obtenons le réseau bayésien de la figure 3.6, où sont également représentées les tables de probabilités conditionnelles de chaque nœud, et où le paramètre c est égal à 95,28.

De la même façon, on peut également surveiller le procédé par une carte MEWMA sous forme du réseau bayésien de la figure 3.7, où le paramètre c est égal à 90.29.

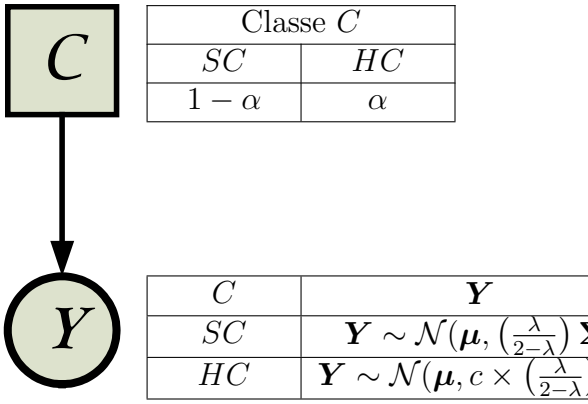
Nous avons simulé ce système sur 30 observations, et nous avons introduit un saut en échelon d'amplitude 0.5 à partir de l'observation 6 sur la première variable. La figure 3.8 présente graphiquement les résultats obtenus. Sur cette figure, les graphiques (a) et (b) représentent respectivement les cartes du T^2 et MEWMA, ils donnent le calcul de la distance statistique de chaque carte de contrôle. Les graphiques (c) et (d) représentent respectivement la modélisation de ces cartes par réseau bayésien. Sur ces deux derniers graphiques est représentés la probabilité de la modalité SC du nœud classe pour la modélisation par réseaux bayésiens de chaque carte. La limite de contrôle est également



Classe C	
SC	HC
$1-\alpha$	α

C	$\boldsymbol{X}$
SC	$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
HC	$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, c \times \boldsymbol{\Sigma})$

FIG. 3.6 – Réseau bayésien similaire à la carte T^2



Classe C	
SC	HC
$1-\alpha$	α

C	$\boldsymbol{Y}$
SC	$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \left(\frac{\lambda}{2-\lambda}\right) \boldsymbol{\Sigma})$
HC	$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, c \times \left(\frac{\lambda}{2-\lambda}\right) \boldsymbol{\Sigma})$

FIG. 3.7 – Réseau bayésien similaire à la carte MEWMA

représentée pour chaque graphique : 9.2 pour la carte T^2, 7 pour la carte MEWMA, et 0.99 (soit $1-\alpha$) pour les modélisations par réseaux bayésiens.

La figure 3.8 permet de voir que les décisions prises à un instant t, entre une carte de contrôle multivariée et son équivalence par réseaux bayésiens, sont les mêmes. De plus, on peut remarquer qu'il existe une certaine ressemblance entre ces signaux, malgré une légère différence du fait de la prise en compte de la borne supérieure et inférieure (à savoir 0 et 1) pour le calcul de $P(SC|\boldsymbol{x})$ [154].

3.2.3.4 Module de détection par réseaux bayésiens

Nous avons précédemment présenté la façon de modéliser une carte de contrôle multivariée (T^2 ou MEWMA) dans un réseau bayésien. Nous allons à présent décrire le module de détection d'un procédé multivarié par réseaux bayésiens. Nous proposons de composer ce module par une carte du T^2 de Hotelling, associée à une carte MEWMA. La carte T^2

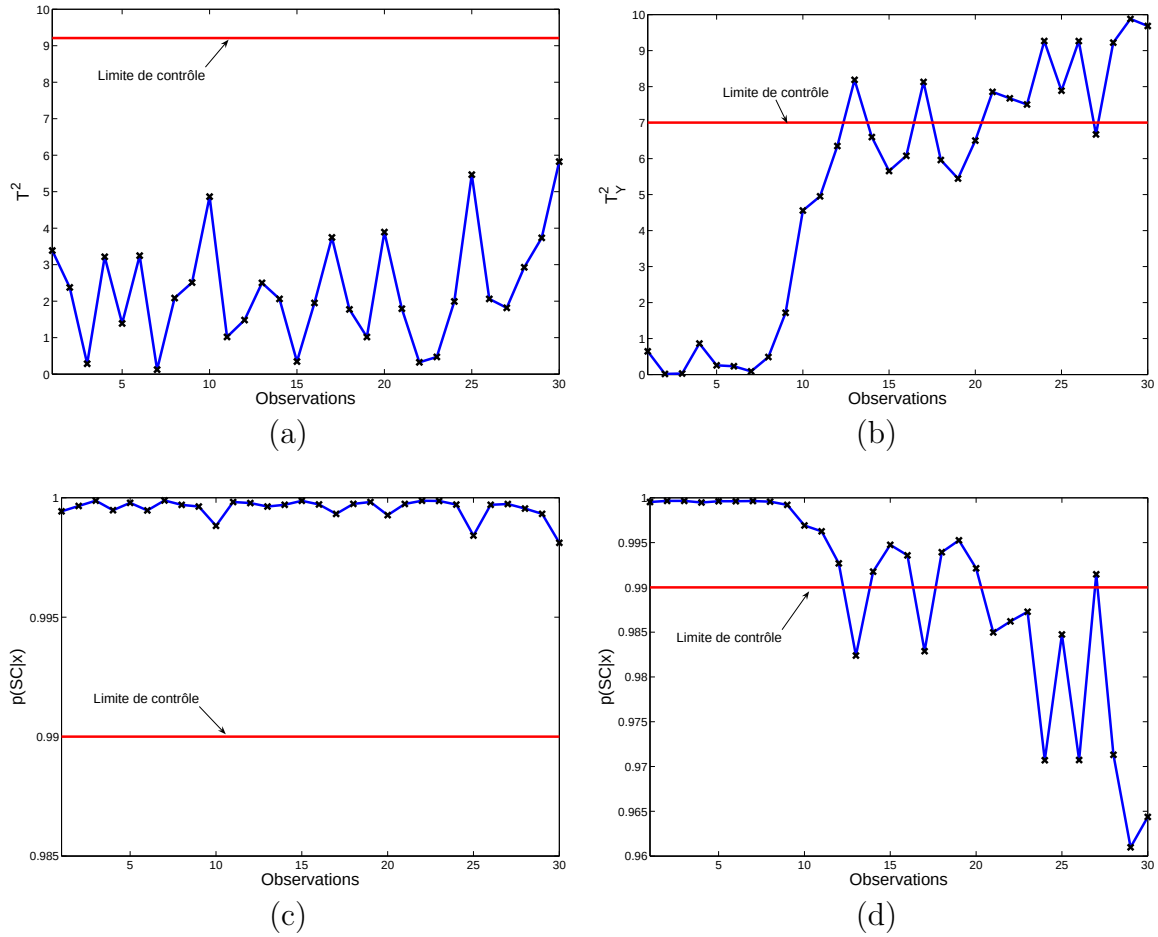


FIG. 3.8 – Résultats des cartes de contrôle T^2 et MEWMA ((a) et (b)), et de leurs équivalences respectives par réseau bayésien ((c) et (d))

détecte très rapidement des sauts de moyenne importants et soudains, alors que la carte MEWMA détecte des sauts d'amplitude plus faible. Nous partons du principe que si l'une ou l'autre des cartes détecte une situation hors-contrôle, alors le procédé est déclaré hors-contrôle. Cette stratégie, bien qu'augmentant le nombre de fausses alarmes, permet de bénéficier des avantages des deux types de cartes. En plus des nœuds nécessaires à la modélisation des deux cartes, nous ajoutons un nœud permettant de statuer sur la présence ou non d'une faute. Nous nommons ce nœud "Détec", et nous le définissons grâce à deux modalités : SC et HC, respectivement pour "sous contrôle" et pour "hors-contrôle". Afin d'exploiter au mieux le réseau, une décision doit être prise sur chaque carte de contrôle. Cette décision ne peut pas être prise directement dans le réseau bayésien. Ainsi, nous réalisons deux inférences successives. La première inférence permet le calcul des différentes probabilités associées à chaque nœud de classe appartenant aux deux cartes. Suivant les probabilités trouvées, soit l'observation est déclarée sous contrôle (SC=1) ou bien hors-

contrôle ($HC = 1$). Les nœuds de classe des cartes de la seconde inférence représentent les résultats des décisions prises (symbole du seuil sur la figure 3.9) entre les deux inférences. Suite à cela, la deuxième inférence permet de statuer sur la présence ou non de faute grâce au nœud "Détec" dont la table de probabilités conditionnelles est donnée dans la table 3.2. Dans cette table, nous nommons la variable de classe de chaque carte multivariée par le nom de la carte représentée. Ainsi le nœud T^2 représente le nœud de classe associé à la modélisation de la carte T^2, alors que le nœud MEWMA représente le nœud de classe associé à la modélisation de la carte MEWMA.

		Détec	
T^2	MEWMA	SC	HC
SC	SC	1	0
	HC	0	1
HC	SC	0	1
	HC	0	1

TAB. 3.2 – Table de probabilités conditionnelles du nœud "Détec"

Ainsi, nous pouvons dresser le schéma (voir figure 3.9) du module de détection de fautes dans un procédé multivarié par réseaux bayésiens. L'originalité de cette méthode est le fait qu'elle peut à la fois détecter des sauts de fortes (carte T^2) et de faibles (carte MEWMA) amplitudes.

Désormais, il est donc possible de pratiquer la détection de faute dans un procédé multivarié, directement par réseaux bayésiens. A présent, nous nous intéressons au diagnostic des fautes détectées.

3.3 Diagnostic supervisé par réseaux bayésiens

Nous avons déjà abordé le problème du diagnostic supervisé, en considérant que celui-ci pouvait être considéré comme une tâche de classification supervisée (voir sections 1.3.2.3 et 1.5.1). Nous avons présenté les principaux classifieurs bayésiens à la section 1.5.9. Nous avons également vu comment réaliser une analyse discriminante (§1.5.7) par réseaux bayésiens (§3.2.2). Nous abordons maintenant deux problèmes importants de la classification supervisée appliquée au diagnostic. Le premier point est l'amélioration des performances de classification par la sélection de variables importantes pour la discrimination. Le second point est l'identification de nouveaux types de fautes pour lesquels aucun exemple n'est disponible. Nous allons donc, dans un premier temps, étudier la sélection de composantes (sélection de variables) importantes pour la classification.

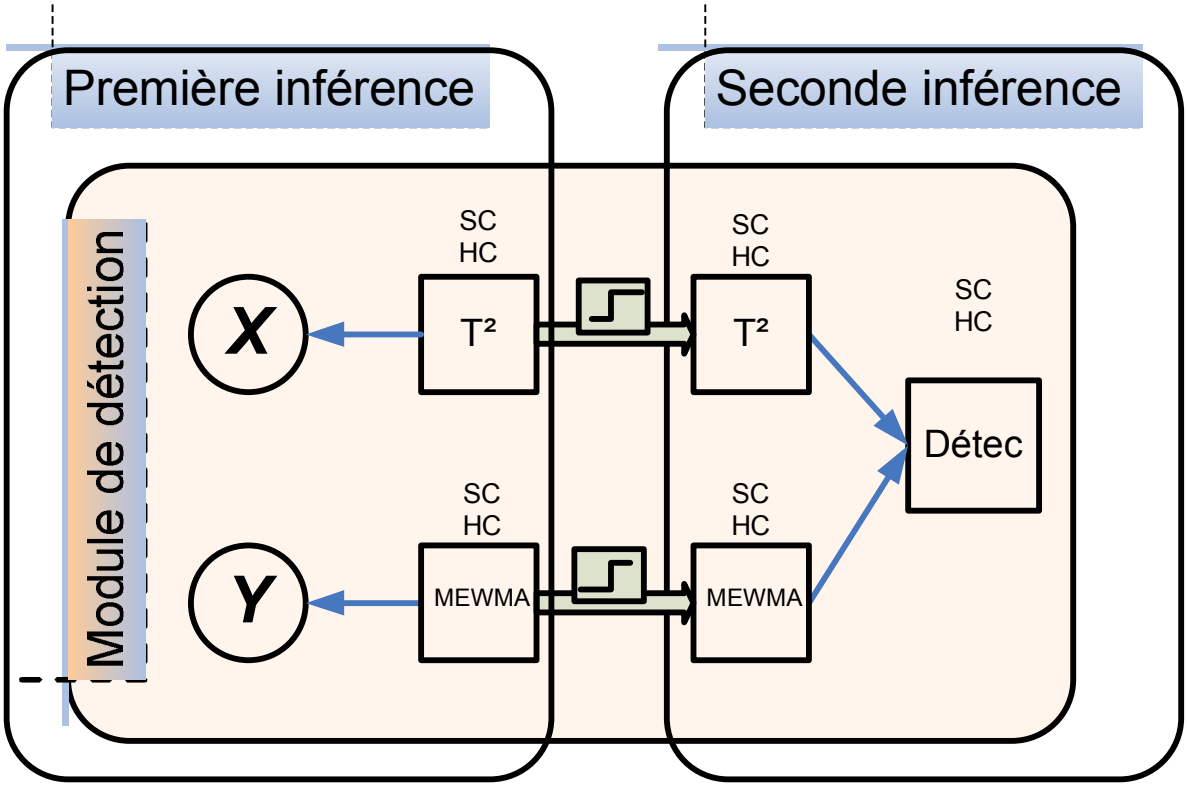


FIG. 3.9 – Module de détection par réseaux bayésiens

3.3.1 Sélection de composantes pour la discrimination

Dans le cadre de la sélection de composantes, beaucoup de méthodes se basent sur des notions provenant de la théorie de l'information [4, 7, 85, 86, 116, 149, 150, 151, 153, 155]. Il est donc nécessaire de présenter cette théorie.

3.3.1.1 Théorie de l'information

La théorie de l'information [26] fournit une mesure quantitative de la notion d'information apportée par un message (ou une observation). Cette notion fut introduite par Claude Shannon en 1948 [133] afin d'étudier les limites du possible en matière de compression de données et de transmission d'informations au moyen de canaux bruités. On trouve de nombreuses applications en télécommunications, en informatique et en statistique notamment. Nous présentons les notions de base de la théorie de l'information, à savoir : l'entropie, l'information mutuelle, ainsi que leurs différentes relations.

Entropie L'entropie est une fonction mathématique correspondant à la quantité d'information contenue ou délivrée par une source d'information [26, 44]. On peut alors considérer une loi de distribution $P(x)$ comme étant une source d'information. Dans le cas où X est une variable discrète à m modalités $(x_1, \ldots, x_m)$, l'entropie (qui peut se considérer comme une mesure de l'imprévisibilité des séquences de X), nommée entropie de Shannon, se calcule à l'aide de l'équation 3.13.

$$H(X) = -\sum_{i=1}^{m} P(x_i) \log P(x_i) \tag{3.13}$$

L'entropie de Shannon est toujours positive. On peut étendre l'équation 3.13 du cas discret vers le cas continu. On obtient alors ce que l'on appelle l'entropie différentielle. On l'appelle différemment car l'entropie différentielle n'est pas le cas limite de l'entropie de Shannon lorsque $m \to \infty$ [26]. Le calcul de l'entropie différentielle d'une variable aléatoire continue X de fonction de densité de probabilité $P(X)$ est donné par l'équation 3.14.

$$h(X) = -\int_{X} P(X) \log P(X) dX \tag{3.14}$$

L'entropie est sans nul doute la notion la plus importante en théorie de l'information. Elle sert notamment au calcul de l'information mutuelle présentée ci-après.

Information mutuelle Supposons à présent que nous possédons deux distributions de variables pouvant être différentes, X et Y. L'information mutuelle peut alors se définir comme étant la mesure de réduction d'incertitude d'une variable au vu de la connaissance de la seconde variable [26]. En d'autres termes, l'information mutuelle de deux variables aléatoires est une quantité mesurant la dépendance statistique de ces variables. L'information mutuelle entre deux variables se calcule comme indiqué par l'équation 3.15 où $P(X, Y)$ représente la densité de probabilité jointe des deux variables.

$$I(X;Y) = \sum_{X} \sum_{Y} P(X,Y) ln \frac{P(X,Y)}{P(X)P(Y)} \tag{3.15}$$

On peut citer plusieurs propriétés de l'information mutuelle :

- $I(X;Y) = 0$ si et seulement si X et Y sont des variables aléatoires indépendantes,
- l'information mutuelle est positive ou nulle : $I(X;Y) \geq 0$,
- l'information mutuelle est symétrique : $I(X;Y) = I(Y;X)$.

Relations Il est possible d'établir certaines relations entre entropie et information mutuelle [44]. Prenons deux distributions de variables aléatoires X et Y. La figure 3.10 montre graphiquement les relations entre les entropies de chaque distribution et l'information mutuelle.

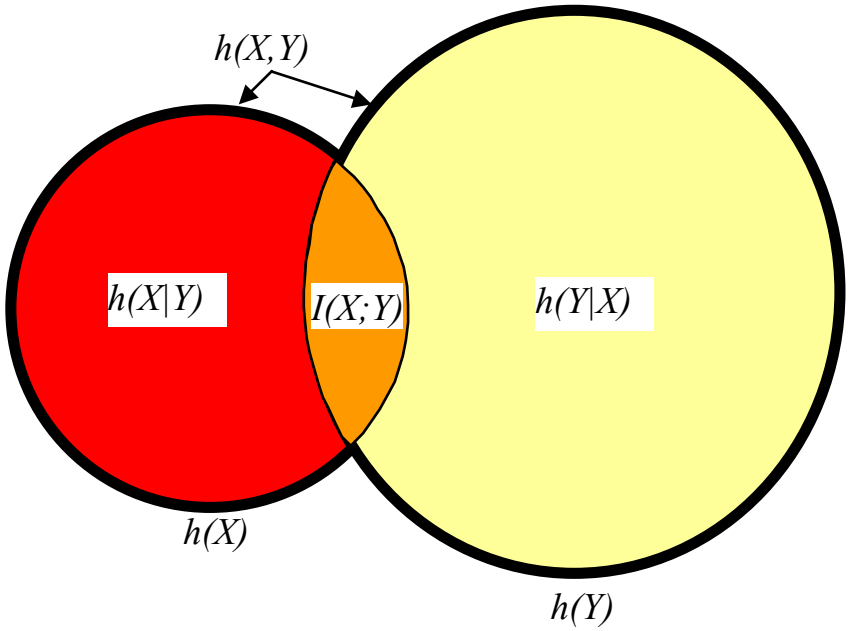


FIG. 3.10 – Représentation graphique des relations sur l'entropie

Il apparaît clairement les relations suivantes :

$$I(X;Y) = h(X) - h(X|Y) \tag{3.16}$$
$$I(X;Y) = h(Y) - h(Y|X) \tag{3.17}$$
$$I(X;Y) = h(X) + h(Y) - h(X,Y) \tag{3.18}$$
$$I(X;X) = h(X) \tag{3.19}$$

où $h(X|Y)$ est l'entropie conditionnelle, alors que $h(X,Y)$ représente l'entropie croisée.

3.3.1.2 Sélection de Composantes et Information Mutuelle

La sélection de composantes représente le choix des variables importantes pour la discrimination des différentes classes du système. Cette sélection permet : d'améliorer les performances de classification, d'accélérer la classification, et enfin de pouvoir mieux

comprendre le fonctionnement sous-jacent du système. Dans le cas de la classification supervisée, l'objectif de la sélection de composantes est d'identifier un groupe de variables donnant de bonnes performances de discrimination entre les différentes classes du système. La solution optimale pour obtenir un classifieur performant est d'estimer le taux de mauvaises classifications de chaque groupe possible du système. Mais, en supposant un système à p variables, le nombre de groupes constructibles possibles est de $\sum_{i=1}^{p} C_n^p$. Par exemple, un système à 20 variables demande l'évaluation de plus d'un million de groupes. Cette méthode peut être efficace pour un système avec peu de variables, mais dans de nombreux cas, cette recherche exhaustive est impossible. Pour répondre à ce problème, certaines approches ont été développées, notamment en utilisant la notion d'information mutuelle [4, 7, 85, 86] décrite précédemment.

Concernant la classification, si l'on calcule l'information mutuelle entre la variable de classe et les descripteurs, on peut alors connaître les descripteurs ou groupes de descripteurs qui sont importants pour la discrimination [116]. Battiti [4], dans le contexte de l'apprentissage neuronal supervisé, propose l'algorithme suivant, nommé MIFS (Mutual Information based Feature Selection), où $\boldsymbol{F}$ et $\boldsymbol{S}$ représentent des ensembles de variables :

1. Initialisation : poser $\boldsymbol{F} \leftarrow$ "p composantes du système" et $\boldsymbol{S} \leftarrow \emptyset$.
2. Calcul de I entre les composantes et la classe C : pour chaque composante $f \in \boldsymbol{F}$, calculer $I(f;C)$.
3. Choix de la première composante : trouver la composante f qui maximise $I(f;C)$; régler $\boldsymbol{F} \leftarrow \boldsymbol{F} \backslash \{f\}$; régler $\boldsymbol{S} \leftarrow \{f\}$.
4. Boucle de sélection : répéter jusqu'à ce que $|\boldsymbol{S}| = k$:
   (a) Calcul de I entre les composantes : pour tous les couples de variables (f, s) avec $f \in \boldsymbol{F}$ et $s \in \boldsymbol{S}$, calculer, si ce n'est pas déjà fait, $I(f;C)$.
   (b) Sélection de la composante suivante : choisir la composante f maximisant $I(f;C) - \beta \sum_{s \in \boldsymbol{S}} I(f;s)$; régler $\boldsymbol{F} \leftarrow \boldsymbol{F} \backslash \{f\}$; régler $\boldsymbol{S} \leftarrow \boldsymbol{S} \cup \{f\}$.
5. $\boldsymbol{S}$ contient les composantes sélectionnées.

Le principal avantage de cette approche est qu'elle considère que les données peuvent suivre n'importe quelle fonction de densité de probabilité. En effet, la méthode approxime les différentes distributions par des histogrammes de données. Un autre avantage de cette méthode est le principe de recherche dans toutes les combinaisons possibles de groupes : il s'agit d'un algorithme forward qui inclut, pas à pas, la variable contribuant à maximiser l'information mutuelle entre la variable de classe et le groupe de variables.

Mais, cette approche possède tout de même quelques inconvénients. Premièrement, la discrétisation des données (afin de construire les différentes fonctions de densité de

probabilité) introduit une perte d'information. Deuxièmement, le nombre de composantes k doit être fixé à l'avance et ne garantie donc pas que le groupe trouvé soit optimal. Enfin, le calcul utilisé pour l'information mutuelle est approximatif. En effet, l'information mutuelle entre des vecteurs de variables est approximée par l'information mutuelle entre les composantes individuelles des vecteurs. De plus, afin de ne pas sélectionner une nouvelle composante possédant trop d'information redondante, Battiti [4] introduit un coefficient β dans le calcul de l'information mutuelle, et le choix de ce coefficient est empirique.

Une autre approche est proposée par Bonnlander et al. [7] dans laquelle les densités de probabilité sont calculées par des noyaux d'Epanechnikov. Cette approche utilise un algorithme de Branch and Bound [36] pour la recherche dans l'espace des différents groupes possibles.

Leray et al. [85] proposent, avec l'utilisation de noyaux d'Epanechnikov pour l'estimation des densités, un algorithme de sélection forward (comme l'approche de Battiti) basé sur l'information mutuelle. Cet algorithme intègre un critère d'arrêt permettant de ne pas fixer initialement le nombre de composantes (contrairement à l'approche de Battiti). Ce critère d'arrêt se base sur la comparaison du taux d'accroissement de l'information mutuelle par rapport à un seuil que Leray et al. [85] fixent à 99%. Cependant, le nombre de composantes sélectionnées dépend de ce seuil, là encore fixé empiriquement. Plus tard, Leray et al. [86] recommandent une procédure différente : premièrement, il faut évaluer les différents groupes grâce à une validation croisée ; deuxièmement, on effectue une comparaison (par un test de Fisher) des différents groupes par rapport au groupe possédant le plus faible taux d'erreur ; enfin, il faut sélectionner, parmi les groupes dont le taux d'erreur est comparable à celui possédant le plus faible taux, le groupe avec la dimension la plus faible (le groupe ayant le moins de composantes).

Le problème commun des différentes approches proposées tient au fait qu'aucune ne soit spécialisée pour un classifieur en particulier. Il est raisonnable de penser que si une méthode de sélection de composantes est adaptée à un classifieur particulier, la combinaison de cette méthode et de ce classifieur pourrait permettre d'augmenter la performance de celui-ci. De plus, dans les approches étudiées, l'information mutuelle n'est pas calculée de manière exacte, mais approximée (notamment du fait de la discrétisation des données). Cependant, en faisant l'hypothèse de normalité, nous montrons que l'information mutuelle entre un groupe de variables et la variable de classe peut être calculée de manière exacte. En effet, dans la section 1.5.9, nous avons présenté certains classifieurs bayésiens et notamment le réseau bayésien semi-naïf condensé, qui suit l'hypothèse de normalité de la variable multivariée. Ainsi, dans le cas de ce classifieur, il est plus intéressant de calculer l'information mutuelle entre le nœud de classe et le nœud multivarié. Pour des

nœuds multivariés de mêmes dimensions, il est possible d'évaluer celui étant le plus informatif. Nous démontrons un nouveau résultat concernant l'information mutuelle entre une variable gaussienne multivariée et une variable multinomiale. Cette information mutuelle peut être calculée comme indiqué dans l'équation 3.20. Dans cette équation, il est supposé que : C est une variable aléatoire multinomiale avec r modalités, ainsi qu'une distribution de probabilités donnée par $P(C = c) = P(c)$; $\boldsymbol{X}$ est une variable aléatoire qui suit une distribution normale multivariée de paramètres $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$; $\boldsymbol{X}$ conditionnellement à $C = c$ suit une distribution normale multivariée de paramètres $\boldsymbol{\mu_c}$ et $\boldsymbol{\Sigma_c}$.

$$I(\boldsymbol{X}; C) = \frac{1}{2}\left[\log(|\boldsymbol{\Sigma}|) - \sum_{c=1}^{r} P(c)\log(|\boldsymbol{\Sigma_c}|)\right] \tag{3.20}$$

En effet, d'après Cover [26], l'entropie h d'une variable $\boldsymbol{X}$ distribuée suivant une loi normale multivariée de dimension p s'écrit :

$$h(\boldsymbol{X}) = -\int_{\boldsymbol{X}} P(\boldsymbol{X})\log(P(\boldsymbol{X}))d\boldsymbol{X} = \frac{1}{2}\log\left((2\pi e)^p|\boldsymbol{\Sigma}|\right) \tag{3.21}$$

De plus, la définition de l'information mutuelle (équation 3.15) donne :

$$\begin{aligned} I(\boldsymbol{X}; C) &= \sum_{c=1}^{r}\int_{\boldsymbol{X}} P(c, \boldsymbol{X})\log\left(\frac{P(c, \boldsymbol{X})}{P(c)P(\boldsymbol{X})}\right)d\boldsymbol{X} \\ &= \sum_{c=1}^{r}\int_{\boldsymbol{X}} P(c)P(\boldsymbol{X}|\, c)\log\left(\frac{P(c)P(\boldsymbol{X}|\, c)}{P(c)P(\boldsymbol{X})}\right)d\boldsymbol{X} \\ &= \sum_{c=1}^{r} P(c)\int_{\boldsymbol{X}} P(\boldsymbol{X}|\, c)\log\left(P(\boldsymbol{X}|\, c)\right)d\boldsymbol{X} - \sum_{c=1}^{r}\int_{\boldsymbol{X}} P(c)P(\boldsymbol{X}|\, c)\log\left(P(\boldsymbol{X})\right)d\boldsymbol{X} \end{aligned} \tag{3.22}$$

Nous pouvons voir que l'intégrale du premier terme représente la définition de l'entropie d'une variable normale multivariée de moyenne $\boldsymbol{\mu_c}$ et de matrice de variance-covariance $\boldsymbol{\Sigma_c}$. Le second terme peut être développé :

$$\begin{aligned}\sum_{c=1}^{r}\int_{\boldsymbol{X}} P(c)P(\boldsymbol{X}|\,c)\log\left(P(\boldsymbol{X})\right)d\boldsymbol{X} &= \int_{\boldsymbol{X}}\sum_{c=1}^{r} P(\boldsymbol{X},c)\log\left(P(\boldsymbol{X})\right)d\boldsymbol{X}\\ &= \int_{\boldsymbol{X}} P(\boldsymbol{X})\log\left(P(\boldsymbol{X})\right)d\boldsymbol{X}\\ &= -\frac{1}{2}\log\left((2\pi e)^p|\boldsymbol{\Sigma}|\right) \qquad (3.23)\end{aligned}$$

Alors,

$$\begin{aligned}I(\boldsymbol{X};C) &= \sum_{c=1}^{r} P(c)\left(-\frac{1}{2}\log\left((2\pi e)^p\,|\boldsymbol{\Sigma_c}|\right)\right)+\frac{1}{2}\log\left((2\pi e)^p\,|\boldsymbol{\Sigma}|\right) \qquad (3.24)\\ &= -\frac{1}{2}\log\left((2\pi e)^p\right)-\frac{1}{2}\sum_{c=1}^{r}P(c)\log\left(|\boldsymbol{\Sigma_c}|\right)+\frac{1}{2}\log\left((2\pi e)^p\right)+\frac{1}{2}\log\left(|\boldsymbol{\Sigma}|\right)\\ &= \frac{1}{2}\left[\log\left(|\boldsymbol{\Sigma}|\right)-\sum_{c=1}^{r}P(c)\log(|\boldsymbol{\Sigma_c}|)\right] \qquad (3.25)\end{aligned}$$

Ainsi, I peut être calculée pour les différents groupes de variables d'un système. Le groupe le plus important pour la tâche de classification est celui possédant une valeur importante pour I. La comparaison des I peut seulement être faite pour des groupes de variables de mêmes dimensions. En effet, en ajoutant des variables au modèle, on accroît l'information de celui-ci. Nous pouvons également donner ce résultat pour le cas particulier univarié. Dans le cas d'une distribution $N(\mu,\sigma^2)$, on a $|\Sigma|=\sigma^2$ et l'équation 3.21 devient :

$$I(X;C)=\frac{1}{2}\left[\log(\sigma^2)-\sum_{c=1}^{r}P(c)\log(\sigma_c^2)\right] \qquad (3.26)$$

Le résultat de cette équation 3.26 a déjà été démontrée par Perez et al. [116] et correspond donc à un cas particulier du nouveau résultat démontré dans l'équation 3.20. L'avantage de l'équation 3.20 est qu'elle permet de prendre en compte l'information portée par les différentes corrélations entre les variables.

Nous exploitons ce nouveau résultat théorique dans un algorithme de sélection de composantes.

3.3.1.3 Algorithme proposée pour la sélection de composantes

La procédure que nous proposons s'effectue en trois étapes : premièrement, rechercher le meilleur groupe (celui maximisant l'information mutuelle) S_k de k variables pour $k = 1$ à p (nous obtenons donc p groupes) ; deuxièmement, pour chaque groupe S_k sélectionné à la première étape, évaluer le taux de mauvaises classifications (moyenne et écart type) par une validation croisée à m parties ; troisièmement, sélectionner le groupe de dimension la plus faible et dont l'erreur moyenne est équivalente à l'erreur moyenne la plus faible. La figure 3.11 représente le schéma de cette procédure.

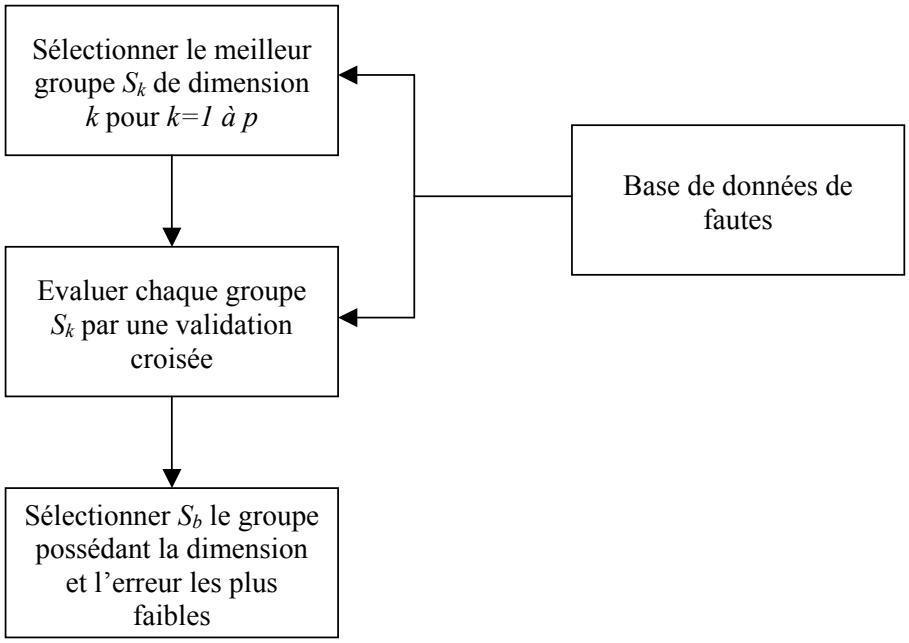


FIG. 3.11 – Procédure de sélection de composantes

Etape 1 : recherche dans l'espace des groupes possibles de variables Pour un système à p variables, le but de cette première étape est de sélectionner le meilleur groupe de variables S_k pour chaque dimension possible k. A la sortie de cette étape, nous obtenons p groupes de variables : le premier groupe possède une variable, le deuxième groupe possède deux variables, ..., le $p^{\text{ième}}$ groupe possède p variables. Pour cela, il est possible d'utiliser un algorithme forward (comme pour l'algorithme de Battiti) ou bien backward. Cet algorithme exploite le nouveau résultat (équation 3.20) que nous avons démontré dans la section 3.3.1.2 : l'information mutuelle entre une variable multivariée normale et une variable multinomiale. Cette procédure se déroule sur p itérations. Dans le cas de l'algorithme forward, à chaque itération, le groupe permettant la maximisation de l'information mutuelle est sélectionné. Ce groupe devient alors le groupe de base pour

l'itération suivante. Le schéma de l'algorithme forward est illustré sur la figure 3.12, alors que la figure 3.13 présente le schéma de l'algorithme backward.

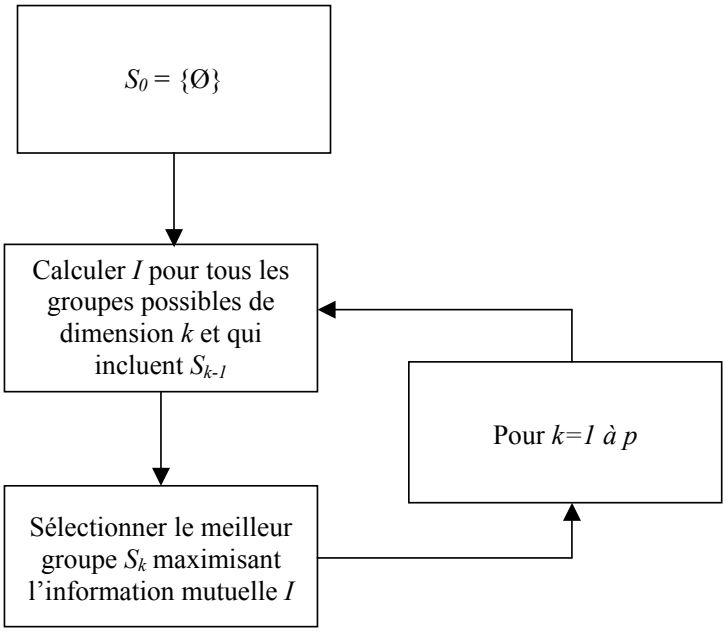


FIG. 3.12 – Algorithme de recherche forward

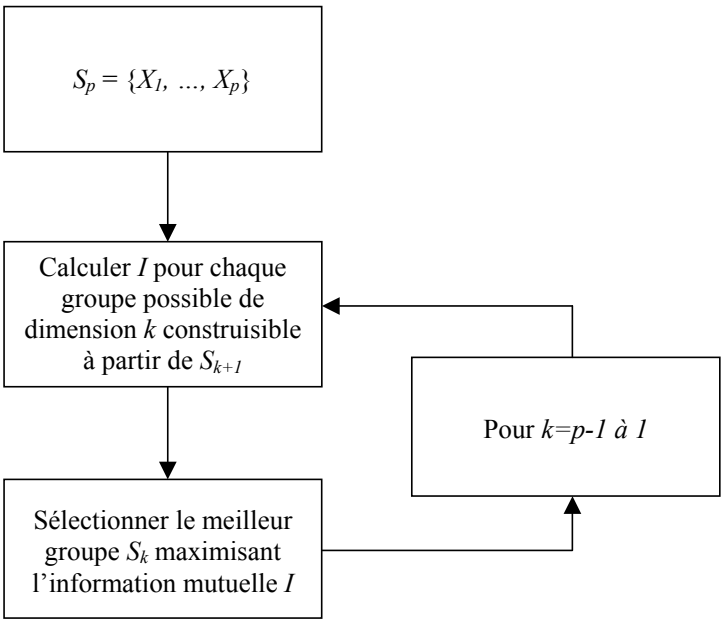


FIG. 3.13 – Algorithme de recherche backward

Afin d'illustrer la technique de recherche, nous avons pris un système à quatre variables soumis à deux types de fautes pour lesquelles nous avons simulées 100 observations chacune. Les paramètres de simulation de ces fautes sont exprimés ci-après :

$$\boldsymbol{\mu}_1 = [1\ 2\ 2\ 1] \qquad \boldsymbol{\mu}_2 = [2\ 1\ 1\ 2]$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.3 & 0.6 & 0.4 \\ 0.3 & 1 & 0.4 & 0.2 \\ 0.6 & 0.4 & 1 & 0.3 \\ 0.4 & 0.2 & 0.3 & 1 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.8 & 0.3 & 0.6 \\ 0.8 & 1 & 0.2 & 0.5 \\ 0.3 & 0.2 & 1 & 0.5 \\ 0.6 & 0.5 & 0.5 & 1 \end{pmatrix}$$

La figure 3.14 représente les différentes étapes de l'algorithme de recherche (algorithme forward) sur cet exemple simple. Au premier pas, l'information mutuelle de chaque variable est calculée. La variable 3 est alors retenue car c'est elle qui maximise l'information mutuelle. Au deuxième pas, tous les groupes possibles de dimension 2, mais comprenant le groupe retenu au pas précédent (soit la variable 3), sont formés et l'information mutuelle de chacun est calculée. Là encore, le groupe maximisant celle-ci est sélectionné, à savoir le groupe {1,3}. Au troisième pas, tous les groupes possibles de dimension 3 comprenant le groupe {1,3} sont formés. Le groupe {1,2,3} est sélectionné car il maximise l'information mutuelle. Enfin, pour le dernier pas, seul un groupe de dimension 4 est faisable (celui comprenant toutes les variables). Puisque ce groupe est unique, le calcul de son information mutuelle n'est normalement pas nécessaire puisqu'il est obligatoirement sélectionné.

Une fois la recherche finie, $\sum_{i=1}^{p} C_{i-1}^{i}$ groupes de variable ont été évalués (pour $p = 20$, 210 évaluations), et p groupes ont été sélectionnés. Il faut choisir, parmi ces p groupes, le plus adapté pour la classification des différentes fautes du procédé. Ceci est l'objectif des étapes 2 et 3.

Etape 2 : Evaluation des groupes sélectionnés L'objectif de cette seconde étape est l'évaluation des différents groupes sélectionnés à l'étape précédente. Cette évaluation ne se fait que sur les données disponibles (données d'apprentissage). Pour cette procédure d'évaluation nous appliquons une technique bien connue : la validation croisée à m parties [25] (voir §1.5.1).

Toujours sur l'exemple donné dans la section précédente, nous pouvons voir sur la table 3.3, la moyenne et l'écart type de l'erreur pour les quatre groupes de variables sélectionnés.

Groupe	S_1	S_2	S_3	S_4
Erreur moyenne	26	15.5	10.5	11
Écart type	12.2	8.9	7.2	7.2

TAB. 3.3 – Résultats de la validation croisée à m parties pour un système à 4 variables

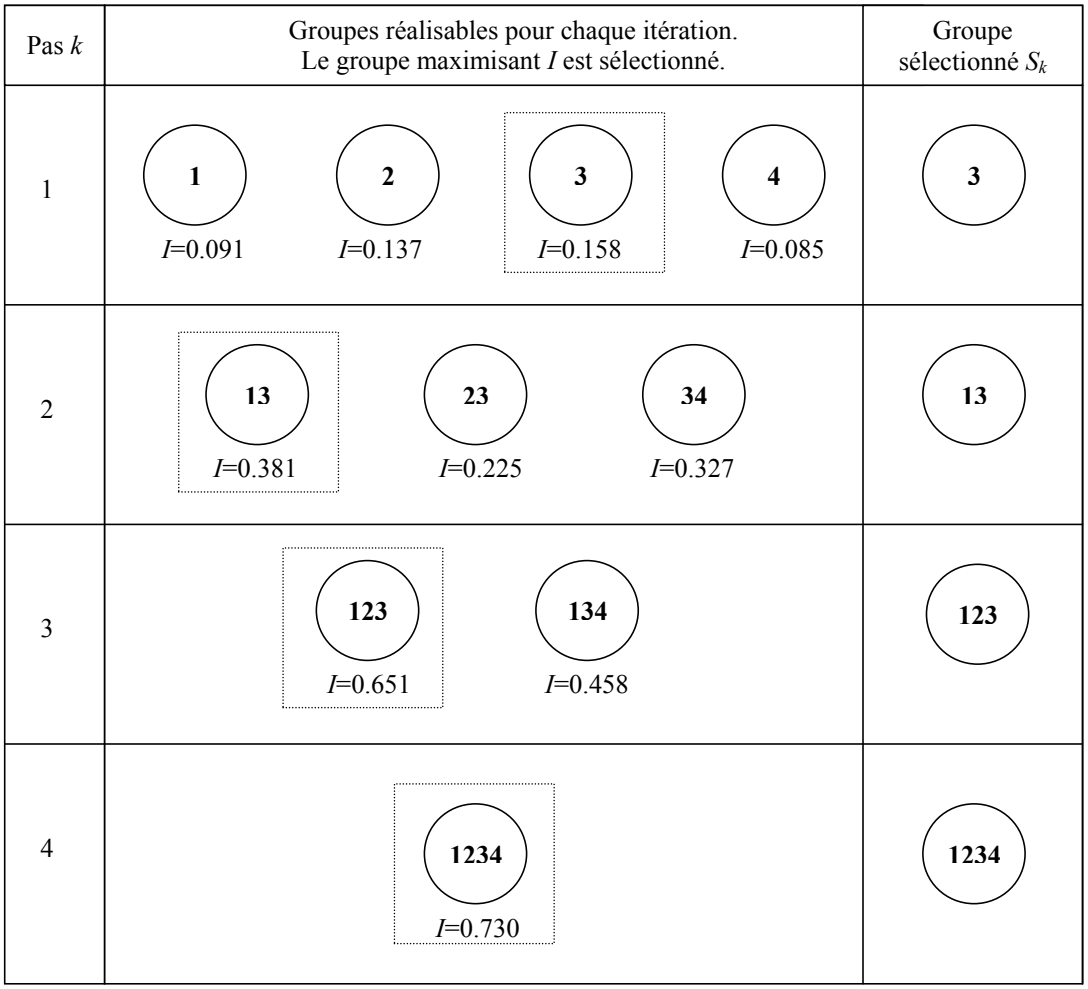


FIG. 3.14 – Exemple de la recherche forward pour un système à 4 variables

Etape 3 : sélection du meilleur groupe de variables Le but de cette troisième étape est de sélectionner le meilleur groupe de variable pour le diagnostic de fautes. Pour cela, nous serions tentés dans un premier temps de sélectionner le groupe possédant la plus faible erreur de classification. Mais, avant cela, représentons l'erreur en fonction du nombre de composantes (voir figure 3.15).

Sur cette figure 3.15, nous pouvons clairement distinguer trois zones : la zone A représente une zone où l'erreur décroît lorsque le nombre de composantes augmente, indiquant alors que le nombre de composantes n'est pas assez important pour obtenir une bonne discrimination ; la zone B est une zone où l'erreur est sensiblement constante, et contenant le groupe ayant la plus faible erreur S_{min} (de dimension N_{min}) ; dans la zone C, plus le nombre de composantes augmente et plus l'erreur augmente également, signifiant alors que trop de composantes apportent du bruit pour la discrimination. Ainsi, si notre but est de sélectionner le groupe de variables donnant l'erreur la plus faible et possédant un

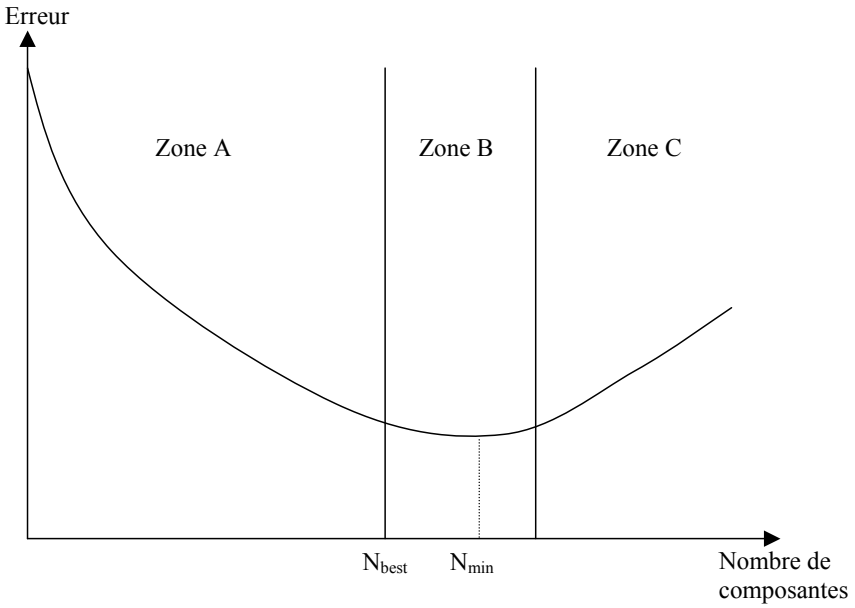


FIG. 3.15 – Erreur moyenne en fonction du nombre de composantes

nombre de composantes le plus faible possible, S_{min} n'est pas le meilleur choix. En effet, le groupe S_{best} possède une erreur de classification équivalente à S_{min}, mais avec un nombre de composantes plus faible N_{best}.

Ainsi, l'idée de cette troisième étape est de sélectionner le groupe S_{min}, puis par une suite de tests d'hypothèse (voir Leray et al. [86]), de trouver S_{best}.

Le test d'hypothèse que nous utilisons ici est un test d'égalité des moyennes de deux distributions (supposées normales) de paramètres μ_1, σ_1^2 et μ_2, σ_2^2, et dont les écart-types sont inconnues et supposés non-égaux. Les écart-types sont supposés inégaux afin de prendre en compte le cas le plus général. Pour le test présenté ici, on utilise toujours une erreur moyenne que l'on sait être la plus faible (erreur moyenne de S_{min}), un test unilatéral est donc employé. Les estimations des paramètres des lois sont respectivement notées $\overline{x}_1, s_1^2$ et $\overline{x}_2, s_2^2$, alors que le nombre d'observations pour chaque distribution est noté n_1 et n_2. α représente le risque de première espèce, classiquement les valeurs de α sont de 1% ou 5%. Ainsi, nous obtenons le test suivant :

$$H0 : \mu_1 = \mu_2 \tag{3.27}$$

$$H1 : \mu_1 > \mu_2 \tag{3.28}$$

On calcule alors la variable Z :

$$Z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.29}$$

Si $Z < Z_\alpha$ (où Z_α est le quantile de la distribution normale centrée et réduite à la valeur α), alors l'hypothèse nulle est vérifiée. Nous faisons remarquer que ce test est correct pour des valeurs de n_1 et n_2 supérieures à 30 observations. Cependant, si n_1 et n_2 sont inférieures à 30 nous pouvons utiliser un test similaire se basant sur une distribution de Student.

Sur l'exemple du système à quatre variables, la table 3.4 donne les résultats des tests d'hypothèse effectués, ainsi que la sélection du meilleur groupe : S_2. Le résultat d'un test d'hypothèse est donné ainsi : "1" signifie que les erreurs sont statistiquement égales, alors que "0" indique que les erreurs ne peuvent pas être considérées comme égales.

Groupe	S_1	S_2	S_3	S_4
Erreur moyenne	26	15.5	10.5	11
Écart type	12.2	8.9	7.2	7.2
Résultat du test d'hypothèse	0	1	–	–
Meilleur groupe		S_2		

TAB. 3.4 – Séléction du meilleur groupe pour l'exemple du système à 4 variables

Sur la table 3.4, on identifie tout d'abord le groupe possédant l'erreur moyenne la plus faible : S_3. Deux tests d'hypothèse sont alors effectués : $\mu_1 = \mu_3$ et $\mu_2 = \mu_3$. Le test d'hypothèse $\mu_4 = \mu_3$ n'aurait pas de sens puisqu'au mieux l'erreur du groupe S_4 est équivalente à celle de S_3 mais S_4 possède une dimension plus élevée. Le résultat des tests d'hypothèse indique que μ_3 et μ_1 ne peuvent pas être considérées comme statistiquement égales, alors que μ_3 et μ_2 peuvent être considérées comme statistiquement égales. Alors, vu que la dimension de S_2 est plus faible que celle de S_3, et vu que les erreurs moyennes de ces deux groupes peuvent être considérées comme égales, on choisit S_2 comme étant le meilleur groupe.

L'idée de cet algorithme de sélection de variables est de sélectionner à chaque itération le groupe permettant une maximisation de l'information mutuelle. À la fin des p itérations, nous obtenons donc p groupes de variables. Ces groupes sont évalués par une validation croisée. Le meilleur groupe est alors sélectionné grâce à une suite de tests d'hypothèses.

3.3.2 Cas d'un nouveau type de faute

Nous considérons que le système de diagnostic doit être capable de statuer si un nouvel individu hors-contrôle appartient à une des classes de faute existantes, ou bien s'il s'agit d'un nouveau type de faute.

Pour bien comprendre ce problème, impliqué par la classification supervisée, étudions un cas très simple de 3 fautes dans un espace bivarié, illustré sur la figure 3.16 suivante.

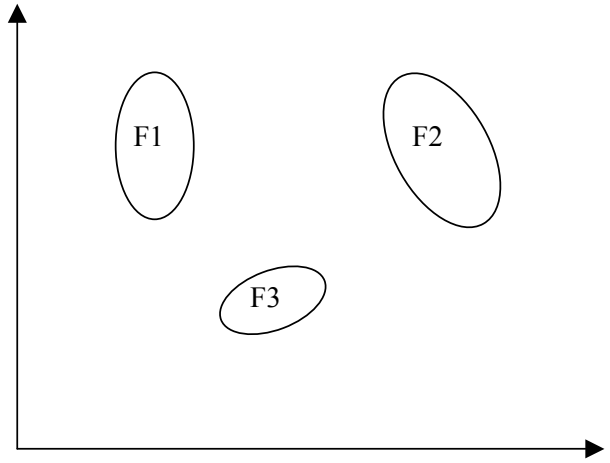


FIG. 3.16 – Trois classes gaussiennes dans l'espace bivarié

Si nous appliquons à présent une analyse discriminante linéaire (ou bien quadratique) afin de séparer ces trois classes dans l'espace bivarié, nous obtenons trois zones de classification, recouvrant tout l'espace, comme indiqué sur la figure 3.17.

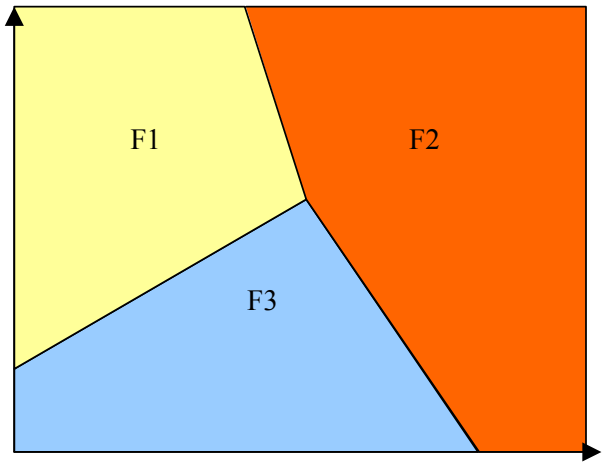


FIG. 3.17 – Zone de classification des trois classes

Le problème de l'apparition d'un nouveau type de faute est alors évident : l'espace est entièrement découpé en trois zones de classification représentant les trois types de fautes connus. Pour répondre à ce problème, nous devons non plus travailler avec 3 classes (représentant les 3 types de fautes connus), mais prendre en compte 4 classes. C'est à dire que nous rajoutons une classe virtuelle NF représentant un type de faute inconnu jusqu'à présent, une nouvelle classe de faute. Ainsi, nous voulons obtenir un espace divisé en quatre zones de classification, comme illustré sur la figure 3.18.

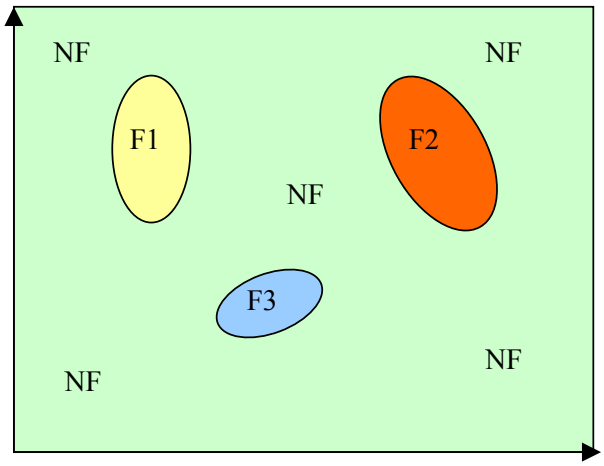


FIG. 3.18 – Zone de classification restreinte des trois classes

Sur cette figure 3.18, on remarque que dès qu'un individu est éloigné des trois classes de faute, on lui attribue la classe virtuelle nouvelle faute NF. C'est ce que l'on appelle un rejet de distance [32, 43] : l'individu appartient à une zone éloignée de celles occupées par l'ensemble d'apprentissage. Or, le rejet de distance d'une classe correspond exactement à une carte de contrôle du T^2 de Hotelling sur la classe d'intérêt. En effet, en ne considérant qu'une seule classe de faute à la fois, le rejet de distance est équivalent à la classification monoclasse présentée à la section 3.2.1 et illustrée par la figure 3.1.

Il est possible de combiner une analyse discriminante avec la notion de rejet de distance, tout cela sous forme d'un réseau bayésien. En effet, d'un côté, nous possédons les probabilités des différents types de fautes connues (grâce à l'analyse discriminante), et de l'autre côté, nous pouvons savoir si tel ou tel type de faute peut être exclu du raisonnement car l'observation est trop loin de la classe (rejet de distance). Le raisonnement basé sur le rejet de distance de chaque type de faute implique une décision à prendre : l'observation est-elle trop loin de la classe de faute F_i ? Pour cela, comme nous l'avons dit, nous utilisons la classification monoclasse. Ceci nous permet de conclure catégoriquement sur l'appartenance possible de l'observation à chaque type de faute, puis de raisonner à

partir de ces conclusions. Malheureusement, l'inférence classique d'un réseau bayésien ne permet pas de faire ceci en une seule fois. Nous allons donc encore une fois (voir §3.2.3.4) enchaîner deux inférences, permettant ainsi de tirer certaines conclusions entre l'application de celles-ci. La figure 3.19 présente le schéma de la classification supervisée par réseaux bayésiens que nous proposons. On peut remarquer que ce module de diagnostic travaille dans un espace réduit $\boldsymbol{X_{sel}}$, tel que défini dans la section 3.3.1 précédente.

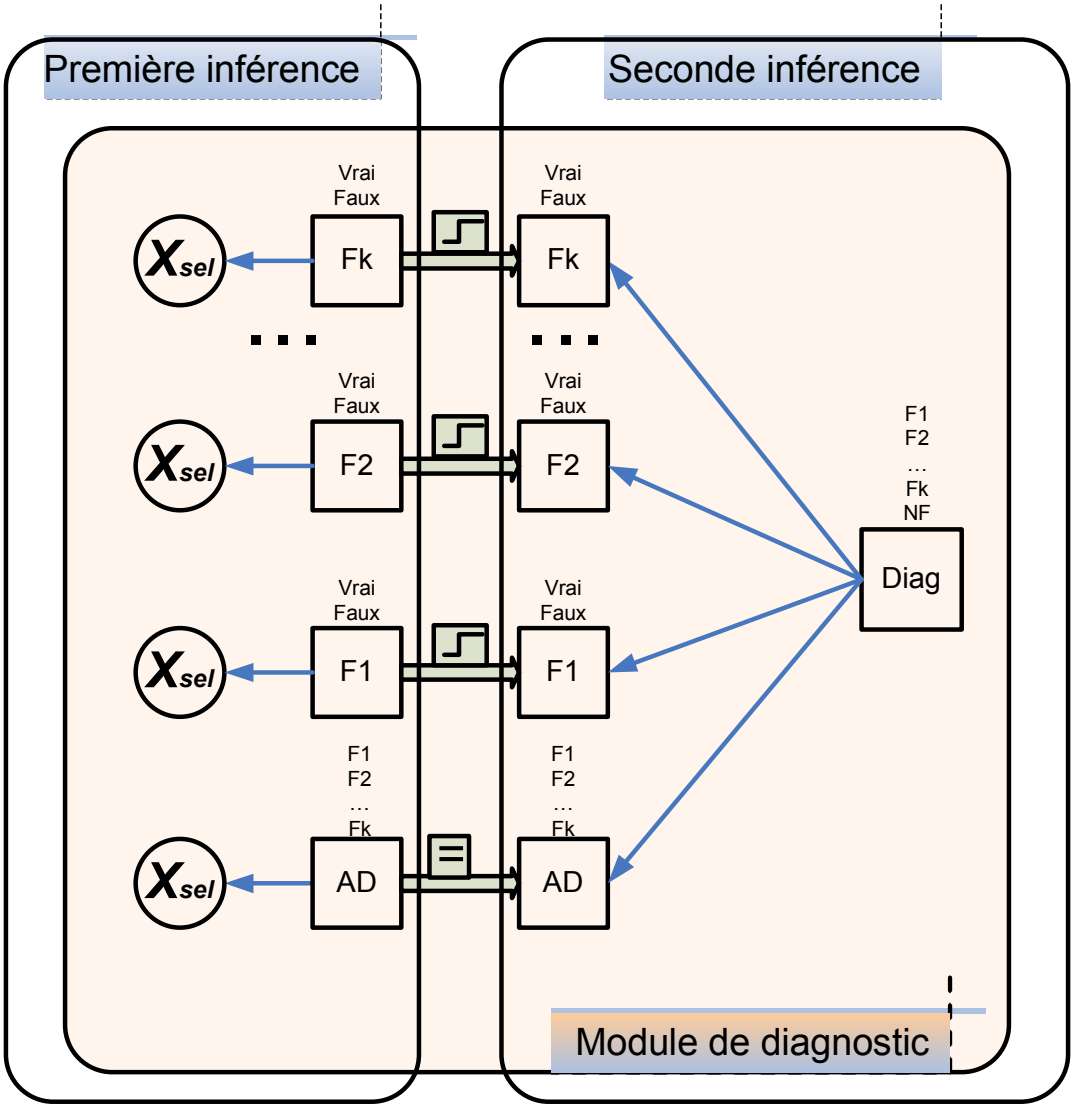
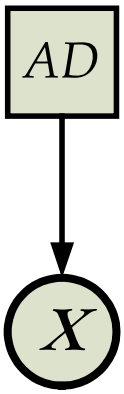


FIG. 3.19 – Module de classification supervisée par réseaux bayésiens

Nous détaillons maintenant les différentes tables de probabilités conditionnelles associées aux différents nœuds du réseau de la figure 3.19, impliqués dans la première inférence. La figure 3.20 présente le réseau bayésien correspondant à l'analyse discriminante basique

(sans prise en compte de nouvelle faute). Pour plus de simplicité, nous fixons les probabilités a priori de chaque classe à $p(F_k) = \frac{1}{k}$. Le nœud $\boldsymbol{X}$ suit les différentes lois de probabilités conditionnellement à la classe de AD, où $\boldsymbol{\mu}_i$ représente le vecteur des moyennes de la faute F_i, et $\boldsymbol{\Sigma}_i$ représente la matrice de variance-covariance de cette faute. Ce réseau permet ainsi d'obtenir des règles de classification similaires à celle de la figure 3.17.

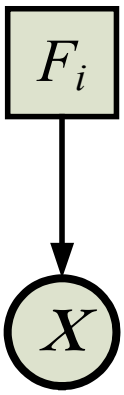


AD		
F_1	...	F_k
$\frac{1}{k}$	...	$\frac{1}{k}$

AD	$\boldsymbol{X}$
F_1	$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
...	...
F_k	$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

FIG. 3.20 – Réseau bayésien correspondant à l'analyse discriminante

Dans la première inférence, nous évaluons également les différentes probabilités que l'observation puisse appartenir à chacune des différentes classes (fautes). Ainsi, pour la faute F_i, nous obtenons le réseau de la figure 3.21, où le nœud de classe est appelé F_i et composé de deux modalités "Vrai" (l'observation appartient à F_i) ou "Faux" (l'observation n'appartient pas à F_i). Dans ce cas, la valeur de α permet de régler la force du rejet. Ainsi, plus α est élevé, plus on rejette le fait que l'observation puisse appartenir à cette classe de faute.



F_i	
Vrai	Faux
$1-\alpha$	α

F_i	$\boldsymbol{X}$
Vrai	$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
Faux	$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_i, c \times \boldsymbol{\Sigma}_i)$

FIG. 3.21 – Réseau bayésien correspondant à l'évaluation de l'appartenance de l'observation la faute F_i

La première inférence nous permet ainsi d'évaluer les différentes probabilités d'appartenance de l'observation à chaque classe de faute. A cet instant, on peut décider du statut des différents nœuds F_i (leurs probabilités dépassent-elles α ?). On les fixe alors à une valeur déterministe : soit l'individu appartient à F_i ($p(F_i = Vrai) = 1$), soit il n'y appartient pas ($p(F_i = Vrai) = 0$). C'est ce passage d'un statut probabiliste des nœuds vers un statut déterministe qui crée la nécessité d'opérer deux inférences. Les nœuds F_i de la seconde inférence représentent le résultat des décisions prises (symbole du seuil sur la figure 3.19) entre les deux inférences. En ce qui concerne le nœud AD de la deuxième inférence, il est tout simplement une copie de son homologue de la première inférence (symbole de l'égalité sur la figure 3.19). Nous détaillons maintenant les tables de probabilités conditionnelles associées aux différents nœuds de la seconde inférence.

Chaque nœud F_i posséde la table de probabilités conditionnelles de la table 3.5. Nous avons fixé les différentes probabilités en appliquant les règles logiques suivantes :

- si $Diag = F_i$, alors nous sommes sûr que l'observation appartient à la faute F_i,
- si $Diag = NF$, alors nous sommes sûr que l'observation n'appartient pas à la faute F_i,
- si $Diag = F_j$ (où $F_j \neq F_i$), alors nous n'apprenons aucune connaissance sur l'appartenance ou non de l'observation à la faute F_i.

	F_i	
Diag	Vrai	Faux
F_1	0.5	0.5
...	...	...
F_i	1	0
...	...	...
F_k	0.5	0.5
NF	0	1

TAB. 3.5 – Table de probabilités conditionnelles des nœuds F_i

La table 3.6 présente la table de probabilités conditionnelles du nœud AD lors de la seconde inférence. On voit que la connaissance d'une faute F_i au niveau du nœud $Diag$ permet de nous fixer la connaissance du nœud AD, ceci étant exprimé par $P(AD = F_i|Diag = F_i) = 1$. A l'inverse, la connaissance sur $Diag$ du nouveau type de faute NF ne nous apporte aucune information sur la discrimination entre les différentes fautes F_i du nœud AD.

En ce qui concerne le nœud $Diag$, il ne possède pas de table de probabilités conditionnelles puisqu'il n'est l'enfant d'aucun nœud. Nous donnons tout de même ses probabilités a priori dans la table 3.7. Les probabilités a priori du nœud $Diag$ sont fixées de telle sorte

	AD		
$Diag$	F_1	...	F_k
F_1	1	...	0
...	...	...	...
F_k	0	...	1
NF	$\frac{1}{k}$	...	$\frac{1}{k}$

TAB. 3.6 – Table de probabilités conditionnelles du nœud AD (deuxième inférence)

qu'aucune modalité de ce nœud ne soit avantagée, elles sont donc toutes égales à $\frac{1}{k+1}$. Cependant, il est possible de privilégier certaines classes de faute si on le souhaite. Pour cela, il suffit d'attribuer des probabilités a priori plus élevées sur les classes que l'on veut privilégier.

$Diag$			
F_1	...	F_k	NF
$\frac{1}{k+1}$	...	$\frac{1}{k+1}$	$\frac{1}{k+1}$

TAB. 3.7 – Table de probabilités a priori du nœud $Diag$

L'intérêt de la seconde inférence est d'ajouter au résultat de l'analyse discriminante de la première inférence, les différents résultats d'appartenance aux classes de fautes obtenus lors de la première inférence et dont les différentes décisions sont prises entre les deux inférences. L'application globale du réseau de la figure 3.19 permet ainsi d'obtenir une règle de classement similaire à celle recherchée (voir figure 3.18). La classification proposée par l'analyse discriminante peut voir ses performances diminuées lorsque peu de données sont disponibles. Dans ce cas, il est toujours possible d'appliquer des méthodes de régularisation comme celles décrites à la section 1.5.7, ou bien d'autres types de classification par réseaux bayésiens proposées à la section 1.5.9 tel qu'un réseau bayésien naïf augmenté par un arbre.

Nous avons proposé ici une méthode permettant la classification d'une faute lorsque des exemples précédents de cette faute sont disponibles. Cette méthode inclut un rejet de distance permettant alors de détecter l'apparition de nouveaux types de fautes dans le procédé. Néanmoins, lorsque cette méthode statue sur la décision NF (nouveau type de faute), il faut pouvoir aider les responsables à identifier au mieux cette nouvelle faute. Ainsi, nous allons proposer une méthode d'identification des variables responsables d'une faute.

3.4 Méthode d'identification MYT par réseaux bayésiens

Cette section a pour but de proposer un réseau bayésien permettant, dans le cas de l'apparition d'un nouveau type de faute, de donner des indications sur les variables impliquées dans cette faute, afin d'aider les opérateurs à identifier physiquement la faute. Pour cela, nous allons nous baser sur les travaux de Li et al. [88], que nous avons présenté au chapitre 2 (§2.3.3.2). Ces auteurs proposent une décomposition MYT (voir §1.4.2.3) basée sur une structure causale d'un réseau bayésien représentant les données, ceci afin d'identifier les variables impliquées dans une faute détectée. Nous étudions quelques points non développés par Li et al. [88], puis nous allons ensuite proposer une amélioration de cette méthode, afin de l'englober entièrement dans un réseau bayésien.

3.4.1 Structure du réseau

3.4.1.1 Algorithmes de recherche de structure

Dans les travaux de Li et al. [88], les auteurs se basent sur une structure causale des données. Cependant, ils ne donnent aucune information concernant la création de cette structure.

La construction de la structure d'un réseau bayésien est un sujet vaste et très étudiée. Plusieurs approches sont possibles, elles peuvent se classer dans les trois catégories [108] :

- les données sont complètes et représentent totalement le problème,
- les données sont incomplètes et/ou il existe des variables latentes,
- peu de données sont disponibles et il faut utiliser la connaissance des experts.

Pour chacun des cas, une recherche exhaustive de la structure du réseau est impossible. En effet, l'espace de recherche est immense puisque Robinson [124] a prouvé que le nombre de structures, noté NS, constructibles à partir de n nœuds est :

$$NS(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} NS(n-i) \tag{3.30}$$

où $NS(0) = NS(1) = 1$. Par exemple : $NS(5) = 29281$ et $NS(10) = 4.2 \times 10^{18}$. Ainsi, beaucoup de chercheurs ont proposé des algorithmes permettant la recherche d'une structure convenable parmi l'espace des structures possibles.

La première approche, proposée à peu près en même temps par deux équipes de recherche [115] et [135], consiste à rechercher les différentes relations causales existant entre

les variables. D'autres approches essaient de quantifier l'adéquation d'un réseau bayésien au problème à résoudre, c'est à dire d'associer un score à chaque réseau bayésien, puis de rechercher la structure permettant de maximiser le score [87]. Enfin, certaines approches sont spécialisées, comme les structures particulières de réseaux bayésiens pour la classification (voir §1.5.9).

Dans cette partie, nous nous intéressons aux approches permettant de définir la structure causale du réseau, en faisant l'hypothèse que les données sont complètes et qu'il n'existe pas de variable latente (suffisance causale). Dans ce contexte, l'algorithme le plus répandu est l'algorithme PC (Peter and Clark) développé par Spirtes et al. [135]. Un algorithme assez similaire a été développé à la même époque par Pearl et Verma, l'algorithme IC (Inductive Causation) [115]. Ces algorithmes partent d'un graphe non orienté complètement relié puis testent alors toutes les indépendances conditionnelles afin de supprimer des arêtes. Suite à cela, ils recherchent toutes les V-structures (trois nœuds reliés en forme de V) et propagent l'orientation des arcs obtenus sur les arêtes adjacentes. Nous présentons plus en détail l'algorithme PC sur la figure 3.22.

3.4.1.2 Test d'indépendance conditionnelle

Comme nous l'avons vu, un algorithme causal implique un test d'indépendance conditionnelle. Dans la littérature, la majorité des applications de l'algorithme PC utilise des variables discrètes. Pour ce type de variable, les tests d'indépendance les plus utilisés sont le test d'indépendance du χ^2 [108], ainsi que le test du rapport de vraisemblance G^2 [135]. Dans le cas qui nous intéresse (travaux de Li et al. [88]), toutes les variables sont continues et possèdent une densité de probabilité gaussienne. Dans ce cas, il est possible d'utiliser la corrélation entre les variables comme indicateur de dépendance. Cependant, la simple corrélation des variables n'est pas adaptée. En effet, la corrélation entre deux variables d'un système à p variables est également fonction des $(p-2)$ autres variables. Une mesure plus intéressante est alors la notion de corrélation partielle. En statistique, la corrélation partielle mesure le degré d'association entre deux variables aléatoires, tout en écartant les effets d'un ensemble d'autres variables aléatoires. Formellement, la corrélation partielle entre deux variables X_i et X_j, étant donné un ensemble $\boldsymbol{X}$ de p variables ($\boldsymbol{X} = \{X_1, X_2, \ldots, X_p\}$), notée $\rho_{X_i,X_j \bullet \boldsymbol{Z}}$ (où $\boldsymbol{Z} = \boldsymbol{X} \backslash \{X_i, X_j\}$), est la corrélation entre les résidus R_{X_i} et R_{X_j} résultant de la régression linéaire respective de X_i avec $\boldsymbol{Z}$, et de X_j avec $\boldsymbol{Z}$.

Plusieurs méthodes permettent le calcul de la corrélation partielle de deux variables [2] : régression linéaire, formule récursive, méthode d'inversion des matrices. La méthode la plus simple à mettre en œuvre est la méthode d'inversion des matrices. Cette méthode

Algorithme PC

- Construction d'un graphe non orienté
  Soit G le graphe reliant complètement tous les nœuds χ
  $i \leftarrow 0$
  Répéter
    Recherche des indépendances conditionnelles d'ordre i
    $\forall \{X_A, X_B\} \in \chi^2$ tels que $X_A - X_B$ et $Card(Adj(G, X_A, X_B)) \geq i$
    $\forall S \subset Adj(G, X_A, X_B)$ tel que $Card(S) = i$
    si $X_A \perp X_B | S$ alors
      suppression de l'arête $X_A - X_B$ dans G
      $Sepset(X_A, X_B) \leftarrow Sepset(X_A, X_B) \cup S$
      $Sepset(X_B, X_A) \leftarrow Sepset(X_B, X_A) \cup S$
    $i \leftarrow i + 1$
  Jusqu'à $Card(Adj(G, X_A, X_B)) < i$, $\forall \{X_A, X_B\} \in \chi^2$
- Recherche des V-structures
  $\forall \{X_A, X_B, X_C\} \in \chi^3$ tels que $\neg\overline{X_A X_B}$ et $X_A - X_C - X_B$,
  si $X_C \notin Sepset(X_A, X_B)$ alors on crée une V-structure :
    $X_A \rightarrow X_C \leftarrow X_B$
- Ajout récursif de $\rightarrow$
  Répéter
  $\forall \{X_A, X_B\} \in \chi^2$,
    si $X_A - X_B$ et $X_A \hookrightarrow X_B$, alors ajout d'une flèche à X_B : $X_A \rightarrow X_B$
    si $\neg\overline{X_A X_B}$, $\forall X_C$ tel que $X_A \rightarrow X_C$ et $X_C - X_B$ alors $X_C \rightarrow X_B$
  Tant qu'il est possible d'orienter des arêtes

Notations	χ	ensemble de tous les nœuds
	$Adj(G, X_A)$	ensemble des nœuds adjacents à X_A dans G
	$Adj(G, X_A, X_B)$	$Adj(G, X_A) \backslash \{X_B\}$
	$X_A - X_B$	il existe une arête entre X_A et X_B
	$X_A \rightarrow X_B$	il existe un arc de X_A vers X_B
	$\overline{X_A X_B}$	X_A et X_B adjacents $X_A - X_B$, $X_A \rightarrow X_B$ ou $X_B \rightarrow X_A$
	$X_A \hookrightarrow X_B$	il existe un chemin dirigé reliant X_A et X_B

FIG. 3.22 – Algorithme PC

possède l'avantage de pouvoir calculer aisément toutes les corrélations partielles de rang n en une seule itération. Pour cela, on doit calculer la matrice de corrélation $\boldsymbol{R}$ du jeu de données comprenant X_i, X_j et $\boldsymbol{Z}$. Ensuite, il faut inverser la matrice $\boldsymbol{R}$: soit $\boldsymbol{P} = \boldsymbol{R}^{-1}$. Le coefficient de corrélation partielle entre X_i et X_j est alors donné par l'équation 3.31, où p_{ij} représente le scalaire de la matrice $\boldsymbol{P}$ pour la ligne i et la colonne j.

$$\rho_{X_i, X_j \bullet \boldsymbol{Z}} = -\frac{p_{ij}}{\sqrt{p_{ii} p_{jj}}} \tag{3.31}$$

Ainsi, nous avons donc à disposition une mesure permettant de quantifier la dépendance conditionnelle des deux variables. Ceci n'est pas suffisant : à partir de quelle valeur ce coefficient de corrélation partielle est-il significatif pour pouvoir conclure de la dépendance ou de l'indépendance conditionnelle de deux variables ? Nous devons effectuer un test d'hypothèse permettant de savoir si l'on peut assimiler une valeur de coefficient de corrélation partielle à une valeur nulle. Dans le cas où les variables impliquées suivent des lois gaussiennes, il est possible d'appliquer le test d'hypothèse suivant [70] :

$$\begin{aligned} H0 &: \rho_{X_i,X_j \bullet \boldsymbol{Z}} = 0 \\ H1 &: \rho_{X_i,X_j \bullet \boldsymbol{Z}} \neq 0 \end{aligned} \tag{3.32}$$

Pour effectuer ce test, on passe par une transformée en z de Fisher, donnée par :

$$z(\rho_{X_i,X_j \bullet \boldsymbol{Z}}) = \frac{1}{2} \log \left(\frac{1 + \rho_{X_i,X_j \bullet \boldsymbol{Z}}}{1 - \rho_{X_i,X_j \bullet \boldsymbol{Z}}} \right) \tag{3.33}$$

Alors, $H0$ est acceptée si la condition suivante est remplie :

$$\left(\sqrt{N - |\boldsymbol{Z}| - 3} \right) z(\rho_{X_i,X_j \bullet \boldsymbol{Z}}) \leq \Phi^{-1}(1 - \alpha/2) \tag{3.34}$$

où Φ représente la fonction de répartition de loi normale centrée réduite, et N représente la taille de l'échantillon qui a permis d'estimer $\rho_{X_i,X_j \bullet \boldsymbol{Z}}$.

Au vu de l'algorithme PC et du test d'indépendance conditionnelle énoncé, la remarque suivante peut être faite : l'algorithme PC ne dépend que du paramètre α, le seuil de confiance du test d'indépendance conditionnelle. Plus α est faible, moins l'algorithme PC forme d'arcs entre les différentes variables, alors que plus α est élevé, et plus la structure finale du réseau posséde d'arcs. Une étude du seuil α a été effectuée par Kalisch et Buhlmann [70]. Les auteurs mettent en avant qu'un α de 0.05 ou de 0.01 permet d'obtenir des performances correctes dans presque toutes les situations. Les auteurs précisent que plus le nombre d'exemples N disponibles pour l'apprentissage est important, et plus la valeur de α devrait se rapprocher de 0.

3.4.2 Paramètres du réseau

3.4.2.1 Calcul des paramètres des nœuds

Nous revenons sur l'utilisation des nœuds continus dans un réseau bayésien. Nous avons vu que pour des raisons de simplicité, seuls les nœuds continus gaussiens sont utilisés. Dans cette section, nous présentons l'apprentissage des paramètres de ce type de nœuds. Nous exposons tout d'abord le cas général où chaque nœud gaussien est de dimension supérieure à 1, puis nous nous intéressons au cas particulier où tous les nœuds gaussiens sont de dimension 1.

L'approche adoptée dans un réseau bayésien est de modéliser la distribution jointe d'un nœud continu et de ses parents continus comme étant une distribution gaussienne multivariée $\boldsymbol{X}$. Afin de calculer les paramètres d'un nœud continu, nous utilisons la régression linéaire. Soit $\boldsymbol{X_1}$, un nœud continu, et $\boldsymbol{X_2}$ ses parents. Ainsi, nous avons :

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \end{pmatrix} ; \boldsymbol{\mu_X} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} ; \boldsymbol{\Sigma_X} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix}$$

Alors, la densité conditionnelle de $\boldsymbol{X_1}$ étant donné $\boldsymbol{X_2}$ est une gaussienne multivariée avec :

$$\boldsymbol{\mu_{X_1|X_2}} = E[\boldsymbol{X_1}|\boldsymbol{X_2} = \boldsymbol{x_2}] = \boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}(\boldsymbol{x_2} - \boldsymbol{\mu_2}) \tag{3.35}$$

et

$$\boldsymbol{\Sigma_{X_1|X_2}} = \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}\boldsymbol{\Sigma_{21}} \tag{3.36}$$

Alors, les paramètres du nœud $\boldsymbol{X_1}$ sont donnés par :

$$\boldsymbol{W_{X_1}} = \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1} \tag{3.37}$$

$$\boldsymbol{\mu_{X_1}} = \boldsymbol{\mu_1} - \boldsymbol{W_{X_1}}\boldsymbol{\mu_2} \tag{3.38}$$

$$\boldsymbol{\Sigma_{X_1}} = \boldsymbol{\Sigma_{11}} - \boldsymbol{W_{X_1}}\boldsymbol{\Sigma_{21}} \tag{3.39}$$

où $\boldsymbol{\mu_{X_1}}$ représente la moyenne de loi normale multivariée associée au nœud $\boldsymbol{X_1}$, $\boldsymbol{\Sigma_{X_1}}$ représente sa matrice de variance-covariance, et $\boldsymbol{W_{X_1}}$ représente sa matrice de régression. Nous illustrons ceci sur un exemple comprenant des nœuds continus gaussiens univariés.

3.4.2.2 Exemple d'apprentissage

Afin de comprendre le principe d'apprentissage des paramètres, prenons l'exemple de trois nœuds X_1, X_2 et X_3, où X_3 dépend de X_1 et X_2. Pour plus de facilité, nous

n'introduisons pas ici de parent discret. Nous disposons d'un jeu de données complet permettant d'apprendre les paramètres du réseau. A partir de ce jeu de données, nous pouvons facilement extraire les valeurs suivantes : μ_1, σ_1^2, μ_2, σ_2^2 et μ_3, σ_3^2, respectivement les moyennes et les variances des variables X_1, X_2 et X_3. Puisque X_1 et X_2 ne possèdent pas de parent, les paramètres estimés sont donc directement les paramètres de ces nœuds dans le réseau. Ceci n'est pas valable pour X_3 puisque ce nœud possède 2 parents (X_1 et X_2). Dans ce cas, nous devons alors rechercher les trois paramètres de ce nœud : μ_{X_3}, $\sigma_{X_3}^2$ et $\boldsymbol{W_{X_3}}$. Pour cela, on utilise la régression linéaire. On se fixe alors $\boldsymbol{Z} = \{X_1, X_2\}$ et donc :

$$\boldsymbol{\mu_Z} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} ; \boldsymbol{\Sigma_Z} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} ; \boldsymbol{\Sigma_{Z3}} = \begin{pmatrix} \sigma_{13} \\ \sigma_{23} \end{pmatrix} ; \boldsymbol{\Sigma_{3Z}} = \boldsymbol{\Sigma_{Z3}}^T.$$

Les trois paramètres du nœud X_3 sont alors donnés par :

- $\boldsymbol{W_{X_3}} = \boldsymbol{\Sigma_{3Z}}\boldsymbol{\Sigma_Z}^{-1}$
- $\mu_{X_3} = \mu_3 - \boldsymbol{W_{X_3}}\boldsymbol{\mu_Z}$
- $\sigma_{X_3}^2 = \sigma_3^2 - \boldsymbol{W_{X_3}}\boldsymbol{\Sigma_{Z3}}$

Si nous faisons l'hypothèse que les données ont été préalablement centrées et réduites, nous obtenons alors :

- $\boldsymbol{W_{X_3}} = \frac{1}{\rho_{12}^2-1}\begin{pmatrix} \rho_{12}\rho_{23} - \rho_{13} & \rho_{12}\rho_{13} - \rho_{23} \end{pmatrix}$
- $\mu_{X_3} = 0$
- $\sigma_{X_3}^2 = \frac{\rho_{12}^2+\rho_{13}^2+\rho_{23}^2-2\rho_{12}\rho_{13}\rho_{23}-1}{\rho_{12}^2-1}$

A présent, nous avons vu comment construire la structure causal d'un réseau bayésien constitué de nœuds gaussiens, et nous avons également étudié le calcul des paramètres de chacun des nœuds. Nous pouvons nous intéresser à l'amélioration de la méthode MYT par réseaux bayésiens, méthode qui nécessitait la compréhension de la construction d'un tel réseau ainsi que l'apprentissage de ces paramètres.

3.4.3 Amélioration de la proposition de Li et al.

La méthode proposée par Li et al. [88] permet, en se basant sur un réseau bayésien à nœuds gaussiens, de connaître les différents termes de la décomposition MYT à calculer. Pour le calcul des différents termes de la décomposition causale du T^2, ainsi que pour les décisions associées (dépassement de limites), les auteurs n'utilisent pas leur réseau de façon optimale. En effet, les auteurs utilisent une carte de contrôle du T^2 extérieurement au réseau, alors que celle-ci peut se modéliser directement à l'intérieur du réseau (voir §3.2.3.3). De même, les auteurs calculent chaque $T^2_{i\bullet PA(X_i)}$ à l'extérieur du réseau, alors qu'il est possible d'effectuer ces calculs dans le réseau.

Nous proposons une extension à la méthode de Li et al. [88] permettant le calcul des différents $T^2_{i\bullet PA(X_i)}$ et des décisions associées à chacun d'entre eux. Le diagnostic par décomposition causale du T^2, tout comme la décomposition MYT, est en fait une surveillance des variables régressées, au moyen de cartes de contrôle univariées. Dans la section 3.2.3, nous avons démontré comment réaliser, par réseaux bayésiens, une carte de contrôle multivariée telle que la carte du T^2 de Hotelling. Or, une carte de contrôle univariée du type carte de contrôle de Shewhart n'est tout simplement qu'un cas particulier d'une carte de contrôle multivariée du type carte du T^2 de Hotelling. En effet, le calcul du T^2 est le suivant :

$$T^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{3.40}$$

Or, dans le cas univarié, $\boldsymbol{x} = x$, $\boldsymbol{\mu} = \mu$ et $\boldsymbol{\Sigma} = \sigma^2$, ainsi l'équation 3.40 devient :

$$T^2 = \frac{(x - \mu)^2}{\sigma^2} \tag{3.41}$$

Dans ce cas univarié, la statistique T^2 suit une loi du χ^2 à un degré de liberté. Or, au vu des démonstrations de la section 3.2.3, ainsi que de leurs transpositions au domaine univarié, il est possible d'envisager une amélioration de la technique développée par Li et al. [88]. Nous proposons ici de suivre directement les différentes valeurs des $T^2_{i\bullet PA(X_i)}$ dans le réseau bayésien. Pour cela, nous rajoutons une variable discrète pour chaque nœud univarié du réseau bayésien. Si nous avons un graphe représentant un système à 3 variables (voir figure 2.20), nous obtenons alors un réseau avec six nœuds : 3 continus (univarié) et 3 discrets (bimodale), comme indiqué sur la figure 3.23.

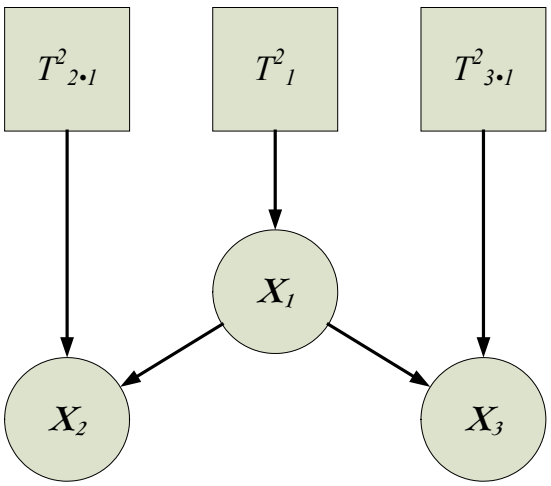
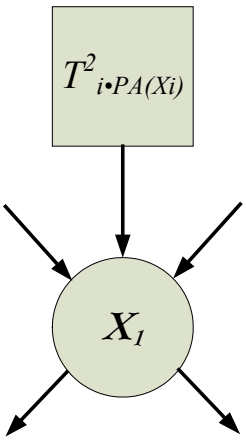


FIG. 3.23 – Exemple de la méthode MYT par réseau bayésien

Nous précisons ici que les variables continues n'ont pas obligatoirement besoin d'avoir été préalablement centrées et réduites. Les nœuds discrets rajoutés à la structure initiale du réseau (celle ne comprenant que les nœuds continus) nous permettent de réaliser directement l'identification des variables incriminées lors d'une situation hors-contrôle. Ces nœuds modélisent une carte de contrôle $T^2_{i\bullet PA(X_i)}$ permettant de conclure sur le statut de chaque variable. Nous rappelons que la modalité SC du nœud discret signifie sous contrôle, alors que la modalité HC signifie hors-contrôle. La figure 3.4.3 détaille la table de probabilités conditionnelles associée à un nœud continu du réseau, ainsi que la table de probabilités a priori de son nœud discret associé.



$T^2_{i\bullet PA(X_i)}$	
SC	HC
$1-\alpha$	α

$T^2_{i\bullet PA(X_i)}$	X_i
SC	$X_i \sim \mathcal{N}(\mu_{i\bullet PA(X_i)}, \sigma^2_{i\bullet PA(X_i)}$
HC	$X_i \sim \mathcal{N}(\mu_{i\bullet PA(X_i)}, c \times \sigma^2_{i\bullet PA(X_i)})$

FIG. 3.24 – Réseau bayésien similaire à la carte $T^2_{i\bullet PA(X_i)}$

Lorsqu'une faute est détectée dans le procédé, chaque nœud discret (représentant le statut d'une variable régressée) fournit une certaine probabilité que la variable soit sous contrôle. Les variables incriminées dans la faute du procédé sont les variables possédant une probabilité inférieure à $1-\alpha$ (où α représente le risque de fausses alarmes). La personne responsable d'identifier physiquement la faute posséde alors des indications très précieuses puisqu'elle connaît les variables du procédé sur lesquelles la faute a agi.

A la vue de ces extensions, nous pouvons dresser un nouveau module de réseau bayésien permettant, après une prise de décision sur les variables, de déterminer quelles sont celles qui sont incriminées dans une situation hors-contrôle. La figure 3.25 présente ce module.

Comme pour les autres modules présentés précédemment dans ce chapitre, l'incorporation de deux inférences est nécessaire afin d'effectuer la prise de décision suivant les valeurs des différents nœuds discrets.

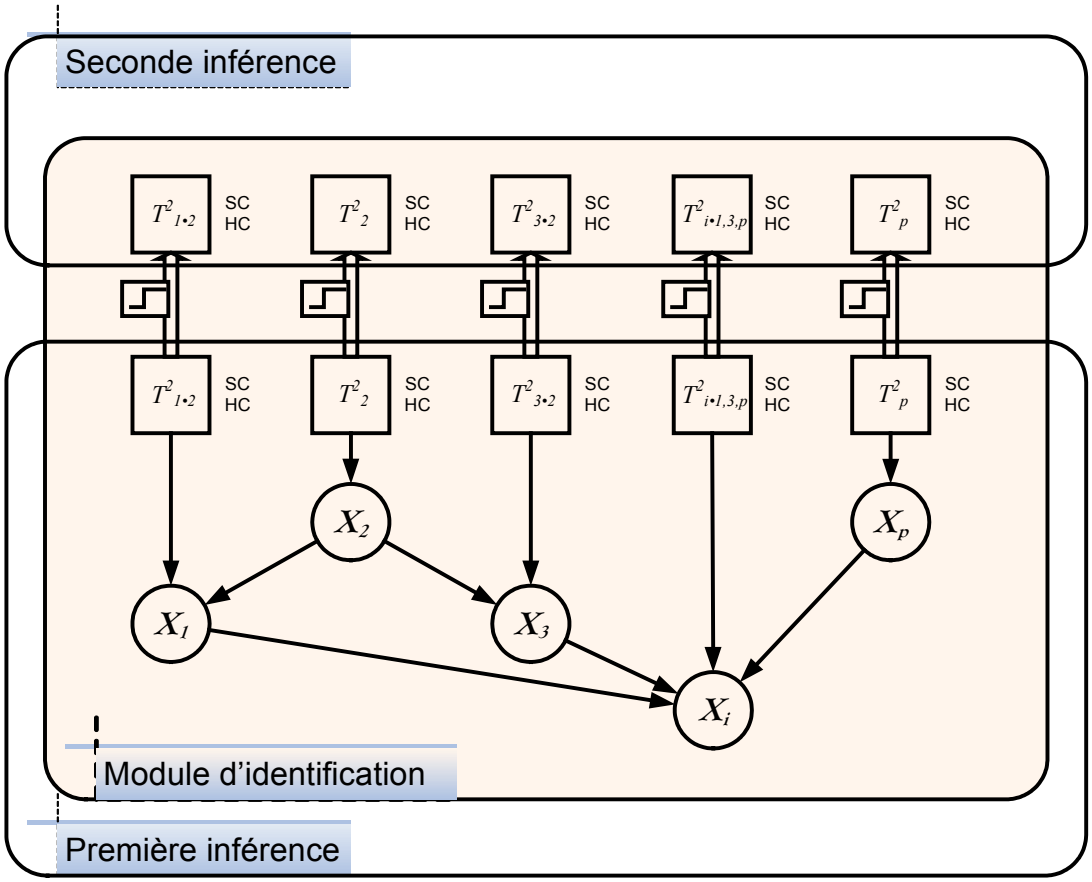


FIG. 3.25 – Module d'identification par réseaux bayésiens

3.5 Surveillance des procédés multivariés par réseaux bayésiens

Dans les sections précédentes, nous avons établis trois points essentiels concernant la surveillance des procédés par réseaux bayésiens : la détection, la classification supervisée (diagnostic), ainsi qu'une méthode d'identification de variables impliquées dans un nouveau type de faute. Dans cette section, nous établissons un schéma global de surveillance des procédés par réseaux bayésiens. Pour cela, nous associons chaque point essentiel de la surveillance à un module.

Nous avons déjà vu que pour chaque module, une phase de décision propre était nécessaire, impliquant alors la réalisation de deux inférences distinctes. Il est possible de ne pas multiplier les inférences : les premières inférences de chaque module peuvent être effectuées en une seule et même inférence, de même que la seconde inférence peut être réalisée en une seule fois pour tous les modules. Lors de cette seconde inférence, nous

n'avons pas développé la prise de décision concernant l'état de notre procédé : il nous faut un nœud permettant de décider dans quel état se trouve notre procédé (est-il sous contrôle, la faute F_k est-elle présente, nouveau type de faute ?). Pour cela, nous utilisons conjointement les informations venant du module de détection, ainsi que du module de classification. Le module d'identification n'est pas pris en compte pour cette tâche car il n'est présent que pour guider les opérateurs du système dans le cas de l'apparition d'un nouveau type de faute. Nous ajoutons un nœud entre les modules de détection et de classification (diagnostic), permettant de prendre une décision concernant l'état du procédé surveillé. Nous nommons ce nœud "Etat". Il comporte plusieurs modalités : SC, le procédé est sous contrôle ; F_1, le procédé est soumis à la faute 1 ; ... ; F_k, le procédé est soumis à la faute k ; NF, le procédé est soumis à un type de faute inconnu. Le nœud "Etat" dépend donc du nœud de détection et du nœud de diagnostic supervisé. Sa table de probabilités conditionnelles est donnée dans la table 3.8.

Détection	Diagnostic	Etat
SC	F_1	SC
	F_2	SC
	⋮	⋮
	F_k	SC
	NF	SC
HC	F_1	F_1
	F_2	F_2
	⋮	⋮
	F_k	F_k
	NF	NF

TAB. 3.8 – Table de probabilités conditionnelles du nœud "Etat"

Le nœud "Etat", comme on le comprend grâce à sa table de probabilités conditionnelles, suit les états des nœuds détection et diagnostic. Si le module de détection ne détecte rien, alors le procédé est déclaré sous contrôle. Par contre, si le module de détection détecte un problème, alors l'état du procédé est une recopie du diagnostic obtenu par le module de diagnostic. Dans le cas où l'état du procédé est NF (apparition d'un nouveau type de faute), le responsable de la surveillance du procédé peut directement visualiser les variables impliquées grâce au module d'identification. On peut désormais dresser le schéma global permettant de surveiller un procédé multivarié grâce à un réseau bayésien sur la figure 3.26.

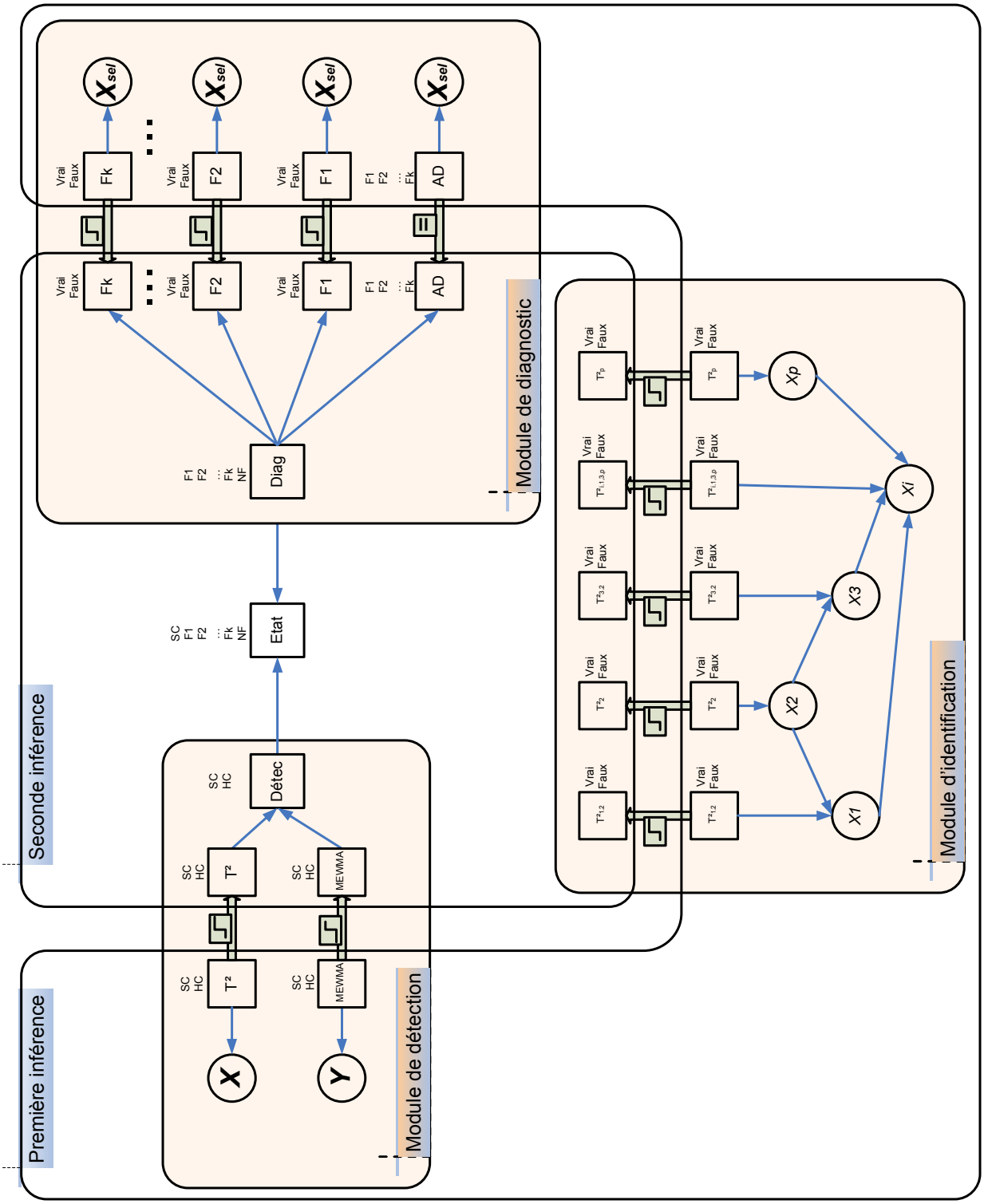


FIG. 3.26 – Réseau bayésien de surveillance des procédés

3.6 Conclusion

Ce chapitre a permis de présenter nos principales contributions apportées au domaine de la surveillance des procédés. Ces contributions portent sur plusieurs points importants. Nous avons tout d'abord pu établir un lien entre cartes de contrôle et analyse discriminante, puis nous avons prouvé mathématiquement l'équivalence qu'il peut y avoir entre une carte de contrôle et un réseau bayésien. Nous avons étudié ensuite la classification supervisée par réseaux bayésiens. Deux contributions originales ont alors été apportées. La première est un nouvel algorithme de sélection de variables importantes pour la discrimination, exploitant un nouveau résultat théorique démontré concernant l'information mutuelle entre une variable gaussienne multivariée et une variable multinomiale. La seconde contribution originale est l'intégration de la notion de rejet de distance directement à l'intérieur du réseau bayésien modélisant une analyse discriminante paramétrique. La troisième partie de ce chapitre a permis de présenter plus en détail le problème de représentation de la régression de variables. En s'appuyant sur les travaux de la première partie de chapitre (à savoir la représentation de cartes de contrôle par réseaux bayésiens) ainsi que sur les travaux de Li et al. [88], nous avons proposé une extension originale aux travaux de ces derniers, permettant l'identification des variables incriminées dans la faute d'un procédé multivarié. Enfin, nous avons présenté la structure complète du réseau bayésien dédié à la surveillance. Ce réseau est principalement composé des trois modules développés (modules de détection, de diagnostic supervisé et de diagnostic non-supervisé). L'assemblage de ces trois modules est effectué grâce au nœud "Etat" traduisant l'état dans lequel se trouve le procédé (SC, F_i, ou bien NF).

Il est désormais légitime de tester si ce réseau est réellement applicable sur un procédé concret. Dans ce but, le prochain chapitre présente l'application de la méthode proposée sur un procédé chimique complexe : le Tennessee Eastman Process (TEP).

Chapitre 4

Application des méthodes proposées sur le TEP

Sommaire

4.1 Introduction

L'application du réseau présenté au chapitre précédent (figure 3.26) demande (comme toute méthode de détection ou de diagnostic basée sur les données) une base de données regroupant des observations de période de fonctionnement normal, ainsi que des observations des différentes fautes déjà connues. Cependant, quelques autres pré-requis sont également nécessaires.

Nous recommandons, avant tout, de travailler sur des données centrées réduites (chaque variable est ramenée à une moyenne nulle et à un écart type unitaire). Bien que cette restriction ne soit pas du tout imposée par la méthode proposée, l'utilisation des données centrées réduites peut notamment permettre le fonctionnement du procédé sur plusieurs points de réglage, et ce, sans changer les paramètres du réseau.

Le second pré-requis est le filtrage des données pour la modélisation de la carte MEWMA. En effet, nous avons fait l'hypothèse que le signal fourni au réseau pour la

carte MEWMA est directement le signal Y. Ce signal doit donc être calculé avant la première inférence du réseau. Une perspective serait une possible intégration du temps (réseau bayésien dynamique) dans le réseau, permettant d'obtenir directement une carte MEWMA à partir des données d'origine X.

Enfin, un troisième pré-requis est l'application de l'algorithme de sélection de variables importantes pour le diagnostic (algorithme proposé à la section 3.3.1). Bien que son application ne soit pas obligatoire (il est possible de diagnostiquer dans l'espace initial), elle semble cependant judicieuse puisqu'elle permet l'amélioration des performances de classification.

De nombreux tests de la méthode proposée ont été effectués sur des données de simulation. Cependant, afin de valider notre méthode sur un exemple réel, nous l'appliquons, dans ce dernier chapitre, sur un procédé chimique complexe : le Tennessee Eastman Process (TEP). La section 4.2 présente ce procédé complexe impliquant 53 variables et 20 types de fautes. Dans la section 4.3, nous étudions, sur ce procédé, les performances du réseau en détection, ainsi qu'en diagnostic supervisé et non-supervisé.

4.2 Présentation du TEP

Le Tennessee Eastman Process (TEP) est un procédé développé par la société Eastman Chemical Company afin de fournir une simulation d'un procédé industriel réel pour le test de méthodes d'asservissements et/ou de surveillance de procédé [40]. Le TEP est basé sur un procédé chimique existant réellement, mais dont certains composants, cinétiques, et conditions opérationnelles ont été modifiés afin d'assurer la confidentialité du procédé réel. Le TEP a été très utilisé par la communauté de la surveillance des procédés afin de comparer certaines méthodes [15, 17, 18, 71]. Ce procédé (voir figure 4.1) est composé de cinq éléments principaux : un réacteur, un compresseur, un décapeur, un séparateur et un condenseur.

Le procédé produit deux composants liquides G et H à partir de quatre gaz réactifs A, C, D et E. Le système implique également un gaz B inerte (non réactif), ainsi qu'un dérivé de production F. Huit composants sont donc impliqués dans le procédé. Les réactions chimiques du procédé sont données par le système d'équation 4.1.

$$\begin{aligned} A(g) + C(g) + D(g) &\rightarrow G(liq) \\ A(g) + C(g) + E(g) &\rightarrow H(liq) \\ A(g) + E(g) &\rightarrow F(liq) \\ 3D(g) &\rightarrow 2F(liq) \end{aligned} \tag{4.1}$$

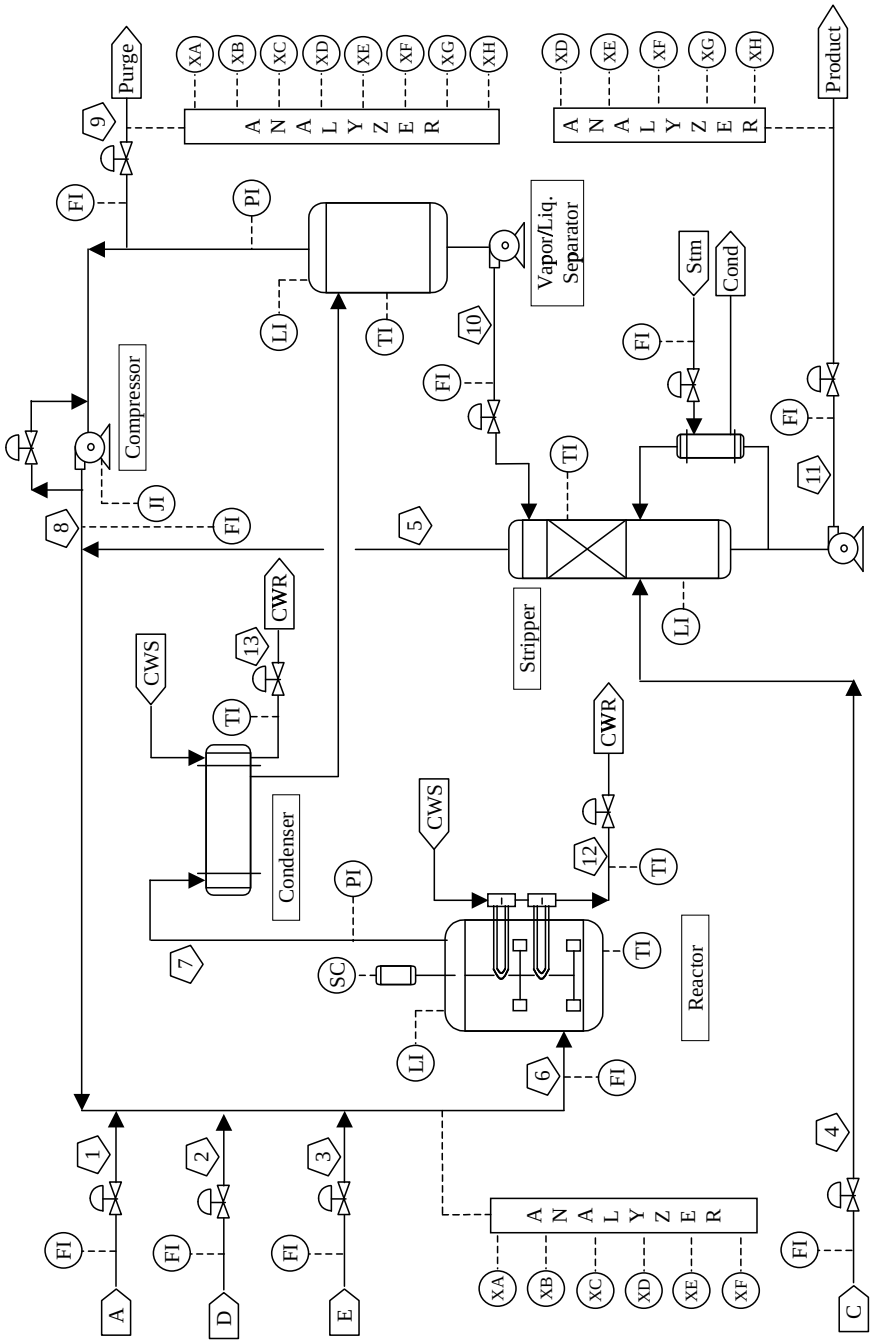


FIG. 4.1 – Schéma du Tennessee Eastman Process

Toutes les réactions sont irréversibles, exothermiques et approximativement de premier ordre en respect des concentrations des différents réactifs. Les taux de réactions suivent une loi d'Arrhenius [80], et la réaction produisant G possède une énergie d'activation élevée, résultant en une forte sensibilité à la température. Les gaz réactifs (A, C, D et E) alimentent le réacteur où ils réagissent et forment, à l'aide d'un catalyseur, les produits

G et H sous forme gazeuse. Un système de refroidissement liquide (par eau) à l'intérieur du réacteur permet l'extraction d'une grande partie de la chaleur produite par celui-ci. Les produits quittent le réacteur, alors que le catalyseur reste dans celui-ci. Le gaz produit est refroidi au moyen d'un condenseur et alimente alors le séparateur liquide-vapeur. La vapeur non condensée dans le séparateur est renvoyée vers le réacteur au moyen d'un compresseur. Le gaz inerte B et le produit dérivé F sont purgés du procédé dans le séparateur. Le flux condensé dans le séparateur est envoyé au décapeur qui a pour but d'éliminer les dernières traces de réactifs. Alors, les produits G et H sont aspirés à l'extérieur du procédé par une unité non représentée sur la figure 4.1.

Ce procédé comporte 53 variables : 12 variables d'asservissement et 41 variables mesurables. Parmi les 41 variables mesurables, 22 sont des variables mesurables en continu (ce sont les valeurs des capteurs du procédé), alors que les autres sont des mesures de compositions telles que des concentrations, et ne sont donc pas disponibles en continu mais échantillonnées. Les 22 variables mesurables en continu sont listées dans la table 4.1 alors que les autres variables mesurables sont visibles dans la table 4.2. Les 12 variables d'asservissement sont données dans la table 4.3.

Variable	Description	Unité
XMES(1)	Débit d'alimentation en A	kscmh
XMES(2)	Débit d'alimentation en D	kg/hr
XMES(3)	Débit d'alimentation en E	kg/hr
XMES(4)	Débit d'alimentation total	kscmh
XMES(5)	Débit de recyclage	kscmh
XMES(6)	Débit d'alimentation du réacteur	kscmh
XMES(7)	Pression du réacteur	kPa
XMES(8)	Niveau du réacteur	%
XMES(9)	Température du réacteur	°C
XMES(10)	Débit de purge	kscmh
XMES(11)	Température du séparateur	°C
XMES(12)	Niveau du séparateur	%
XMES(13)	Pression du séparateur	kPa
XMES(14)	Débit du séparateur	m^3/hr
XMES(15)	Niveau du décapeur	%
XMES(16)	Pression du décapeur	kPa
XMES(17)	Débit du décapeur	m^3/hr
XMES(18)	Température du séparateur	°C
XMES(19)	Débit de gaz au séparateur	kg/hr
XMES(20)	Puissance du compresseur	kW
XMES(21)	Température de ref. liq. en sortie de réacteur	°C
XMES(22)	Température de ref. liq. en sortie de séparateur	°C

TAB. 4.1 – Variables de mesure en continu

Variable	Composant	Période d'échantillonage (en min)	Unité
XMES(23)	A	6	mol%
XMES(24)	B	6	mol%
XMES(25)	C	6	mol%
XMES(26)	D	6	mol%
XMES(27)	E	6	mol%
XMES(28)	F	6	mol%
XMES(29)	A	6	mol%
XMES(30)	B	6	mol%
XMES(31)	C	6	mol%
XMES(32)	D	6	mol%
XMES(33)	E	6	mol%
XMES(34)	F	6	mol%
XMES(35)	G	6	mol%
XMES(36)	H	6	mol%
XMES(37)	D	15	mol%
XMES(38)	E	15	mol%
XMES(39)	F	15	mol%
XMES(40)	G	15	mol%
XMES(41)	H	15	mol%

TAB. 4.2 – Variables de mesure échantillonnées

Variable	Description	Unité
XC(1)	Débit d'alimentation en D	kg/hr
XC(2)	Débit d'alimentation en E	kg/hr
XC(3)	Débit d'alimentation en A	kscmh
XC(4)	Débit d'alimentation en A et C	kscmh
XC(5)	Valve de recyclage du compresseur	%
XC(6)	Valve de purge	%
XC(7)	Débit d'alimentation du séparateur	m^3/hr
XC(8)	Débit d'alimentation du séparateur	m^3/hr
XC(9)	Valve du décapeur	%
XC(10)	Débit du refroidissement liquide au réacteur	m^3/hr
XC(11)	Débit du refroidissement liquide au condenseur	m^3/hr
XC(12)	Vitesse de l'agitateur	tr/min

TAB. 4.3 – Variables de contrôle du TEP

L'intérêt du TEP pour la communauté de l'asservissement est que ce procédé est fortement instable en boucle ouverte. Beaucoup de méthodes ont été proposées afin de l'asservir [94, 99, 122]. Lyman et Georgakis [94] ont fourni plusieurs structures d'asservissement du TEP et ont mis en évidence qu'une seule était réellement performante. Nous avons donc décidé de travailler avec la structure recommandée par ces auteurs. Le schéma du TEP et de son asservissement sont donnés sur la figure 4.2, où seules les variables impliquées dans

l'asservissement sont représentées et où les boucles d'asservissement sont représentées en pointillé.

En ce qui concerne les méthodes de surveillance des procédés, l'intérêt du TEP est qu'il peut être soumis à 20 fautes différentes. Ces fautes sont de diverses natures : saut en échelon de certaines variables internes, augmentation de la variabilité de certaines autres, ou bien faute d'actionneurs tel qu'une vanne bloquée. La description de ces 20 fautes est faite dans le tableau 4.4. On peut observer que les fautes F16 à F20 sont inconnues.

Faute	Description	Type
F1	Ratio d'alimentation A/C	Saut
F2	Composition en B	Saut
F3	Temp. d'alimentation en D	Saut
F4	Temp. d'entrée du ref. liq. au réacteur	Saut
F5	Temp. d'entrée du ref. liq. au condenseur	Saut
F6	Baisse d'alimentation en A	Saut
F7	Perte de pression de l'alimentation en C	Saut
F8	Composition d'alimentation en A, B et C	Variation aléatoire
F9	Temp. d'alimentation en D	Variation aléatoire
F10	Temp. d'alimentation en C	Variation aléatoire
F11	Temp. d'entrée du ref. liq. au réacteur	Variation aléatoire
F12	Temp. d'entrée du ref. liq. au condenseur	Variation aléatoire
F13	Cinétiques des réactions	Dérive lente
F14	Valve du ref. liq. au réacteur	Bloquée
F15	Valve du ref. liq. au condenseur	Bloquée
F16	Inconnue	Inconnue
F17	Inconnue	Inconnue
F18	Inconnue	Inconnue
F19	Inconnue	Inconnue
F20	Inconnue	Inconnue

TAB. 4.4 – Les différentes fautes du TEP

Afin de visualiser le comportement des 52 variables du procédé, nous fournissons, en annexe A.2, le tracé de celles-ci dans le cas du fonctionnement normal. On peut remarquer que toutes ces variables possèdent un certain bruit, et que certaines d'entres-elles suivent une certaine dynamique (variables 18 à 20 par exemple).

Pour mieux comprendre ce que représente une faute, nous prenons le cas de la faute F4. Cette faute est une augmentation de la température du liquide de refroidissement à l'entrée du réacteur. Comme pour les autres fautes, on remarque qu'elle agit sur une variable qui n'est pas pris en compte dans le procédé : aucune variable surveillée ne donne la température du liquide de refroidissement à l'entrée du réacteur. Cependant, cette faute engendre des répercussions sur deux variables incluses dans la surveillance : les variables

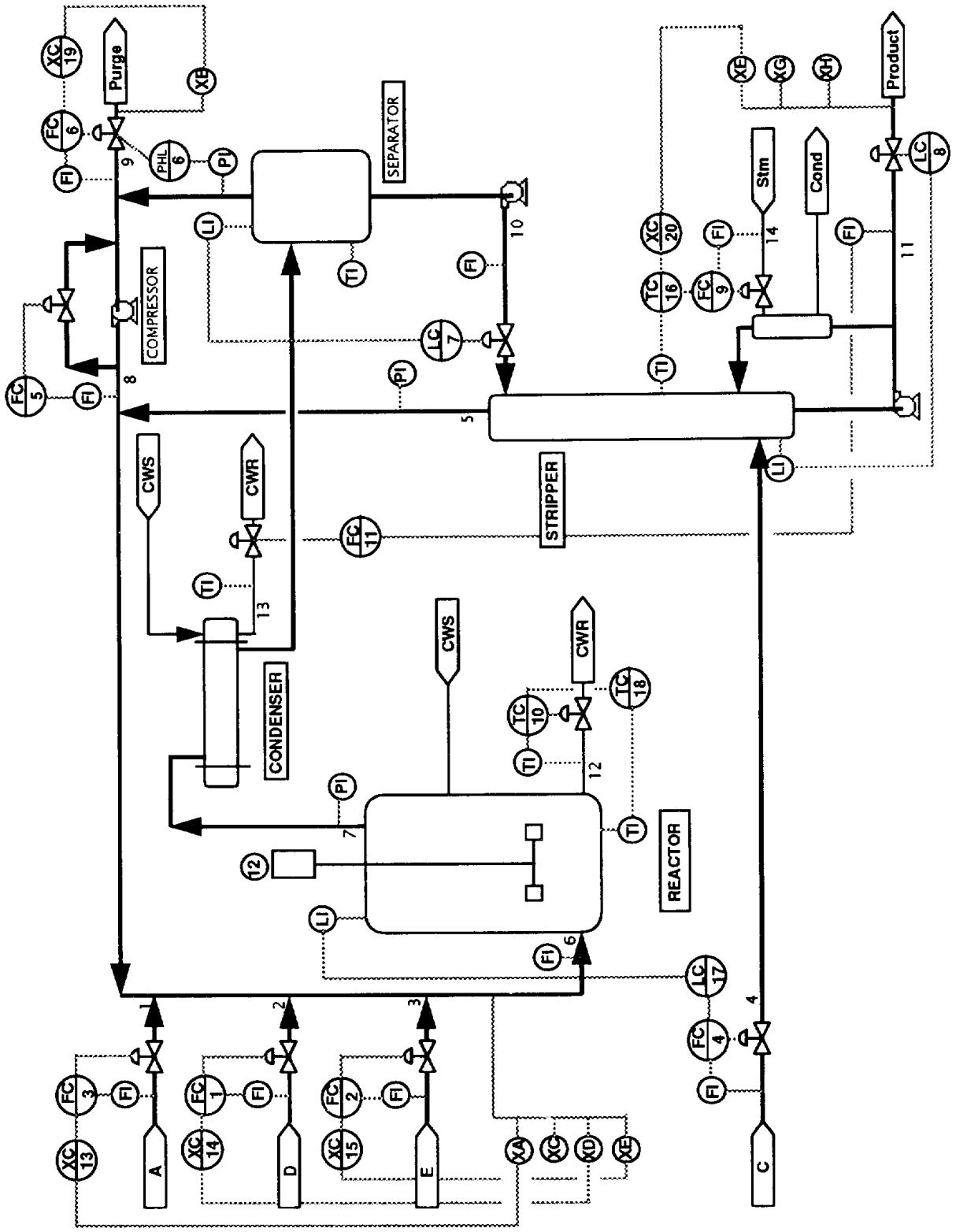


FIG. 4.2 – TEP asservi par Lyman et Georgakis

9 (XMEAS9) et 51 (XC10), respectivement la température du réacteur et le débit de son refroidissement liquide. La figure 4.3 donne la comparaison des variables 9 et 51 pour le cas du fonctionnement normal et pour le cas de la faute F4. Sur les graphiques (c) et (d), la faute F4 est introduite à la 161$^{\text{ième}}$ observation.

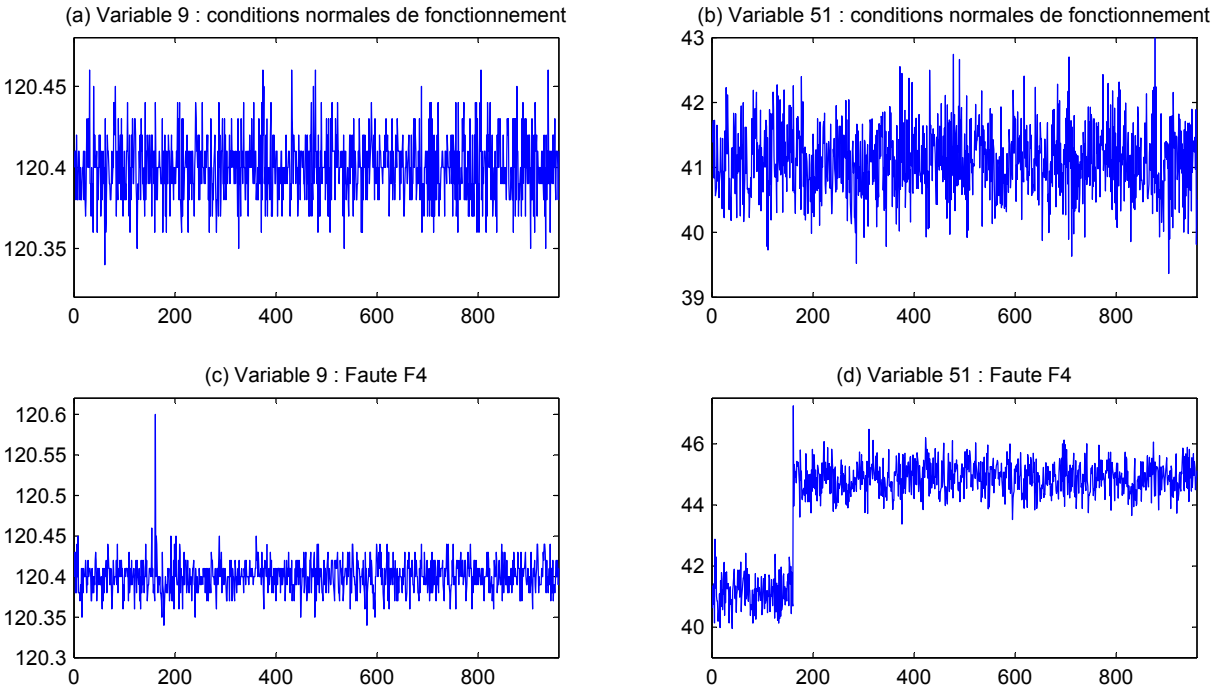


FIG. 4.3 – Comparaison des variables 9 (XMEAS9) et 51 (XC10) pour le fonctionnement normal et pour la faute F4

Il est possible de comprendre le comportement des variables 9 et 51 en réponse à l'introduction de la faute F4 dans le procédé. Dès son apparition (observation 161), la faute F4 (augmentation de la température d'entrée du liquide de refroidissement du réacteur) engendre une augmentation de la température du réacteur. Ceci est bien visible sur le graphique (c) : on observe que la température du réacteur augmente jusqu'à 120.6°C alors qu'elle oscillait normalement autour de 120.4°C. En effet, puisque la température du liquide de refroidissement est plus élevée, l'échange de chaleur entre le liquide et le réacteur est plus faible, engendrant alors une augmentation de la température dans celui-ci. Au vu de cette augmentation de température dans le réacteur, l'asservissement du TEP accroît alors le débit du refroidissement liquide d'environ 41 m^3/hr à environ 45 m^3/hr (visible sur la variable 51, graphique (d)). Puisque le débit augmente, la quantité de chaleur évacuée du réacteur redevient normale, et la température du réacteur retourne à son niveau normal de fonctionnement (aux alentour de 120.4°C).

Sur la figure 4.3, on observe que la faute F4 engendre des conséquences très visibles sur les variables du procédé. Ce type de faute devrait donc être facilement détecté. Cependant, certaines fautes n'entraînent pas de changements si brutaux sur les variables du procédé, rendant la détection moins évidente (exemple de la faute F9 en annexe A.3).

4.3 Surveillance du TEP par réseaux bayésiens

Dans cette section, nous évaluons les performances de la méthode proposée sur un exemple concret, celui du Tennessee Eastman Process. Nous avons repris les données utilisées dans le livre de Chiang et al. [17]. Elles proviennent du TEP couplé à la structure d'asservissement de Lyman et Georgakis [94]. L'intérêt de ces données est qu'elles sont disponibles en ligne à l'adresse suivante http ://brahms.scs.uiuc.edu. Ces données se présentent ainsi (voir table 4.5) : 480 observations d'apprentissage pour chaque type de faute ainsi que pour la période normale, et 800 observations de test pour chaque type de faute ainsi que pour la période normale. Les données d'apprentissage ont été obtenues par simulation de chacune des fautes sur une période de 24 heures, alors que les données de test ont été obtenues sur une durée de 40 heures. La période d'échantillonnage de toutes les variables a été fixée à 3 minutes. Il faut également préciser que les 53 variables n'ont pas été prises en compte puisque la variable XC(12), la vitesse de l'agitateur, reste constante dans n'importe quelle situation (ceci étant dû au système d'asservissement). Ainsi, seul 52 variables sont présentes dans les données utilisées.

Classe	Données d'app.	Données de test
Normale	480	800
Faute 1	480	800
Faute 2	480	800
...	...	...
Faute k	480	800
...	...	...
Faute 20	480	800

Tab. 4.5 – Données utilisées

4.3.1 Détection

Pour tester les performances de notre méthode de surveillance par réseaux bayésiens, nous fixons tout d'abord un taux de fausses alarmes acceptable pour la détection. Tout

comme d'autres chercheurs ayant travaillés sur le TEP [17, 18, 71], nous optons pour un taux de fausses alarmes de 0.01 (1%), soit en moyenne une fausse alarme tous les 100 échantillons. L'échantillonnage étant de 3 minutes, nous obtenons donc une moyenne d'une fausse alarme toutes les 5 heures. Nous avons deux moyens de détection : les modélisations des cartes T^2 et MEWMA. Pour obtenir le taux désiré de fausses alarmes, nous devons donc prendre la moitié du taux global pour chaque carte. Ainsi, chaque carte possède un taux théorique de fausses alarmes de 0.005 (soit 0.5%). La limite de contrôle de la carte T^2 (servant au calcul du coefficient c) est celle décrite dans la table 1.2. Pour la carte MEWMA, nous avons obtenu la limite de contrôle par simulation (comme recommandé par Lowry et al. [93]). Nous avons effectué une vérification séparée des taux de fausses alarmes de chaque carte sur les données de test pour le fonctionnement normal. Pour chaque carte, nous avons obtenu un taux réel de fausses alarmes de 0.625%, soit 5 fausses alarmes pour chaque carte sur les 800 observations de test. Ainsi, au pire des cas, le taux global de fausses alarmes sur les données de test ne peut pas dépasser les 1.3%, ce qui semble acceptable pour un taux théorique de 1%.

Nous nous sommes intéressés à deux critères de détection : la fiabilité (voir [71]) et l'instant de la première détection. La fiabilité consiste à obtenir le nombre d'alertes obtenues sur la période de test (soit 40 heures) et à diviser ce nombre par le nombre total d'échantillon de la période de test. L'instant de première détection représente le numéro d'échantillon de la première alerte.

Nous précisons que la détection effectuée prend bien en compte les 52 variables disponibles. En effet, certains auteurs [15, 71] ont fait une sélection des variables importantes pour la surveillance de certaines fautes. Cependant, face à un procédé réel, le nombre de faute n'est jamais connu à l'avance, et la réduction du nombre de variables pour la surveillance peut s'avérer inadéquate lorsqu'un nouveau type de faute (non détectable sur les variables sélectionnées) intervient.

Les résultats obtenus pour la détection sont présentés dans la table 4.6, où l'indice de classe 0 représente la période normale de fonctionnement (sans faute) et où les indices de classe de 1 à 20 représentent respectivement les fautes de F1 à F20.

La table 4.6 nous permet de conclure sur quelques points. Concernant la classe de fonctionnement normal, on peut voir que le taux de fausses alarmes (1.13%) est bien respecté puisque la fiabilité de cette classe correspond globalement au taux théorique fixé précédemment de 1% (soit une alerte tous les 100 échantillons).

Concernant les différentes fautes, on s'aperçoit très vite d'une grande disparité entre elles. En effet, certaines fautes sont détectées dans tous les cas : les fautes F4, F5, F6, F7 et F14. D'autres fautes ne sont que peu détectées : les fautes F3, F9 et F15. Avant de

Classe	Instant de la première détection	Fiabilité (en %)	Fiabilité privée d'inertie (en %)
0	97	1.13	1.28
1	3	99.75	100
2	13	98.5	100
3	34	35	36.51
4	1	100	100
5	1	100	100
6	1	100	100
7	1	100	100
8	18	97.75	99.87
9	7	15.88	15.99
10	18	97	99.11
11	7	90.88	91.56
12	2	99.88	100
13	37	95.5	100
14	1	100	100
15	146	30.5	37.25
16	9	99	100
17	20	97.5	99.87
18	57	92.38	99.33
19	2	96.5	96.62
20	65	91.88	99.86

TAB. 4.6 – Performances en détection

tirer d'autres conclusions sur ces données, nous devons faire une remarque importante : certaines fautes possèdent une inertie. En effet, la faute F13 par exemple, semble être détectée assez correctement (fiabilité de 95,50%) malgré que les 36 premiers échantillons de cette faute ne soient pas détectés. Or, en ne prenant pas en compte les 36 premiers échantillons, la fiabilité de cette faute passe alors à 100%. Nous étudions les fautes en ne prenant pas en compte cette éventuelle inertie. Pour cela, il est nécessaire de recalculer la fiabilité à partir du premier échantillon détecté, que nous nommons fiabilité privée d'inertie. La dernière colonne de la table 4.6 donne pour chaque classe sa fiabilité privée d'inertie. Cette table nous permet de conclure sur la détection performante de plusieurs fautes : 15 fautes (F1, F2, F4, F5, F6, F7, F8, F10, F12, F13, F14, F16, F17, F18 et F20) possèdent une fiabilité privée d'inertie supérieure à 99% (dont 10 à 100%). Parmi les autres fautes, deux (F11 et F19) sont relativement bien détectées (respectivement 91.56% et 96.62%).

Les fautes F3, F9 et F15 sont difficiles à détecter (<40%). Chiang et al. [18], en utilisant les méthodes de détection par ACP (voir §1.4.3) sur les mêmes données du

TEP, sont arrivés à la même conclusion. En réalité, ces fautes sont presque inobservables lorsque l'on voit les données des différentes variables. Nous donnons en annexe A.3 le comportement des 52 variables du procédé dans le cas de la faute F9, ainsi que dans le cas du fonctionnement normal. On s'aperçoit très bien que pour chaque variable, il n'y a pas de différences visibles (saut, rampe, augmentation de la variabilité, etc) entre le fonctionnement normal et la faute F9. Il faut préciser que la modélisation de la carte T^2 n'est pas efficace sur ce type de faute. Ces fautes sont majoritairement détectées par la modélisation de la carte MEWMA car elles impliquent des sauts de moyenne très faibles (non détectables sur les graphes donnés en annexe).

Afin de visualiser graphiquement les résultats du réseau pour chacune des fautes ainsi que pour la période de fonctionnement normal, les figures 4.4 et 4.5 représentent temporellement la modalité "Sous Contrôle" (SC) des nœuds T^2 et MEWMA. Nous précisons qu'une observation est déclarée hors-contrôle dès que la probabilité de la modalité est inférieure à 0.995 (puisque le taux théorique de fausses alarmes a été fixé à 0.005 pour cette carte). Sur cette figure 4.4, nous retrouvons les conclusions tirées précédemment. En effet, un signal presque toujours à 1 signifie que le procédé est sous contrôle, ou bien qu'il ne détecte rien (en présence de faute). A l'inverse, un signal toujours près de 0 (ou du moins inférieur à 0.995) signifie qu'une faute est présente dans le procédé. Ainsi, on remarque que pour la modélisation de la carte T^2, les fautes F3, F9 et F15 possèdent une probabilité presque toujours autour de 1, comme dans le cas du fonctionnement normal. On comprend bien que ces fautes ne sont pas détectées. A l'inverse des fautes telles que F6 ou F7 ont un signal presque constant à 0. Cela signifie donc qu'une faute est présente, et qu'elle possède une amplitude importante par rapport au fonctionnement normal, contrairement à la faute F4 (par exemple) dont l'amplitude est plus minime puisque quelques pics remontent vers la valeur supérieure pour la carte T^2. Enfin, il est également possible de remarquer des oscillations, notamment pour les fautes F11 et F19, mettant en évidence des oscillations (avec des passages dans la zone de fonctionnement normal) dans les signaux des différentes variables du procédés.

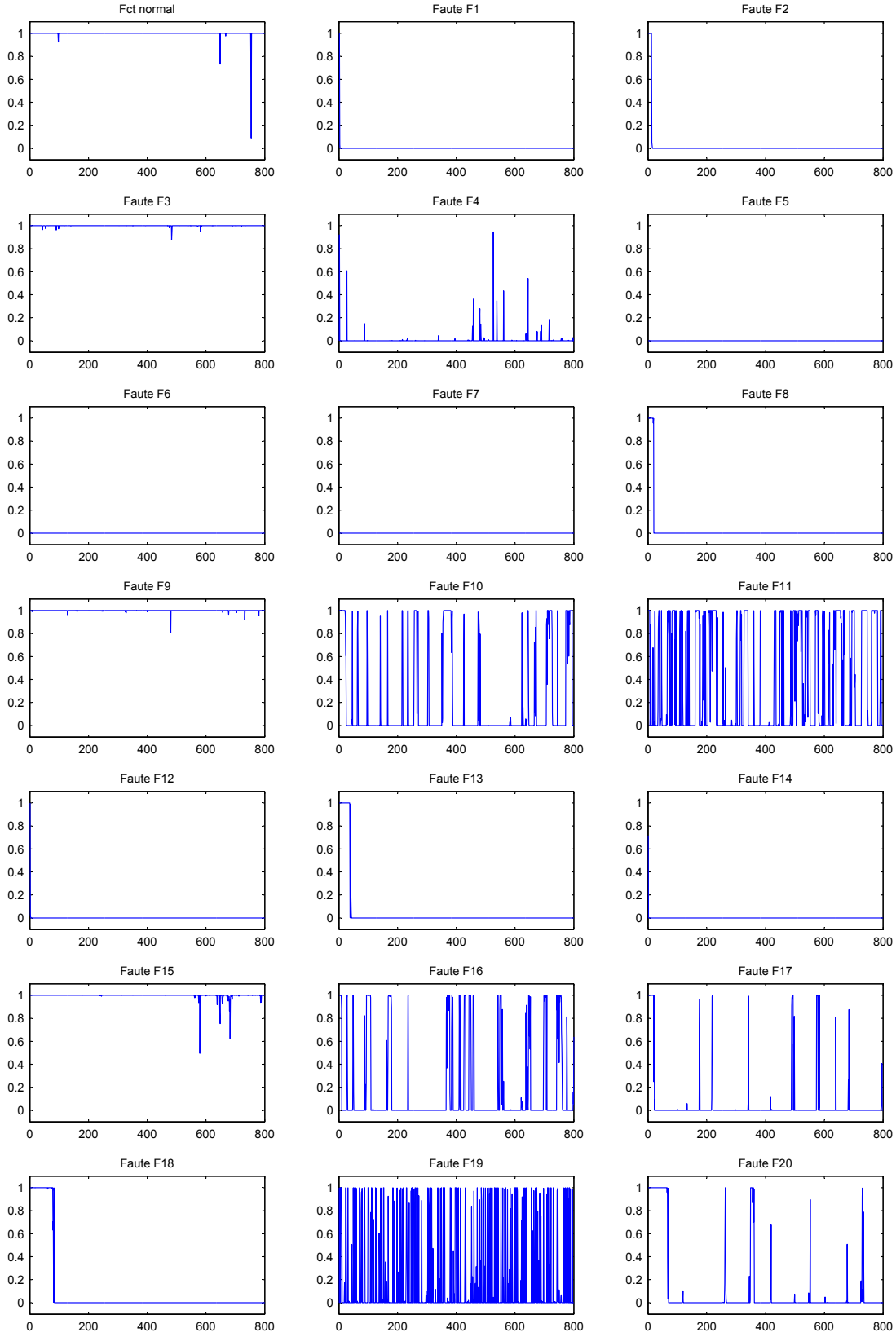


FIG. 4.4 – Modalité SC du nœud T^2

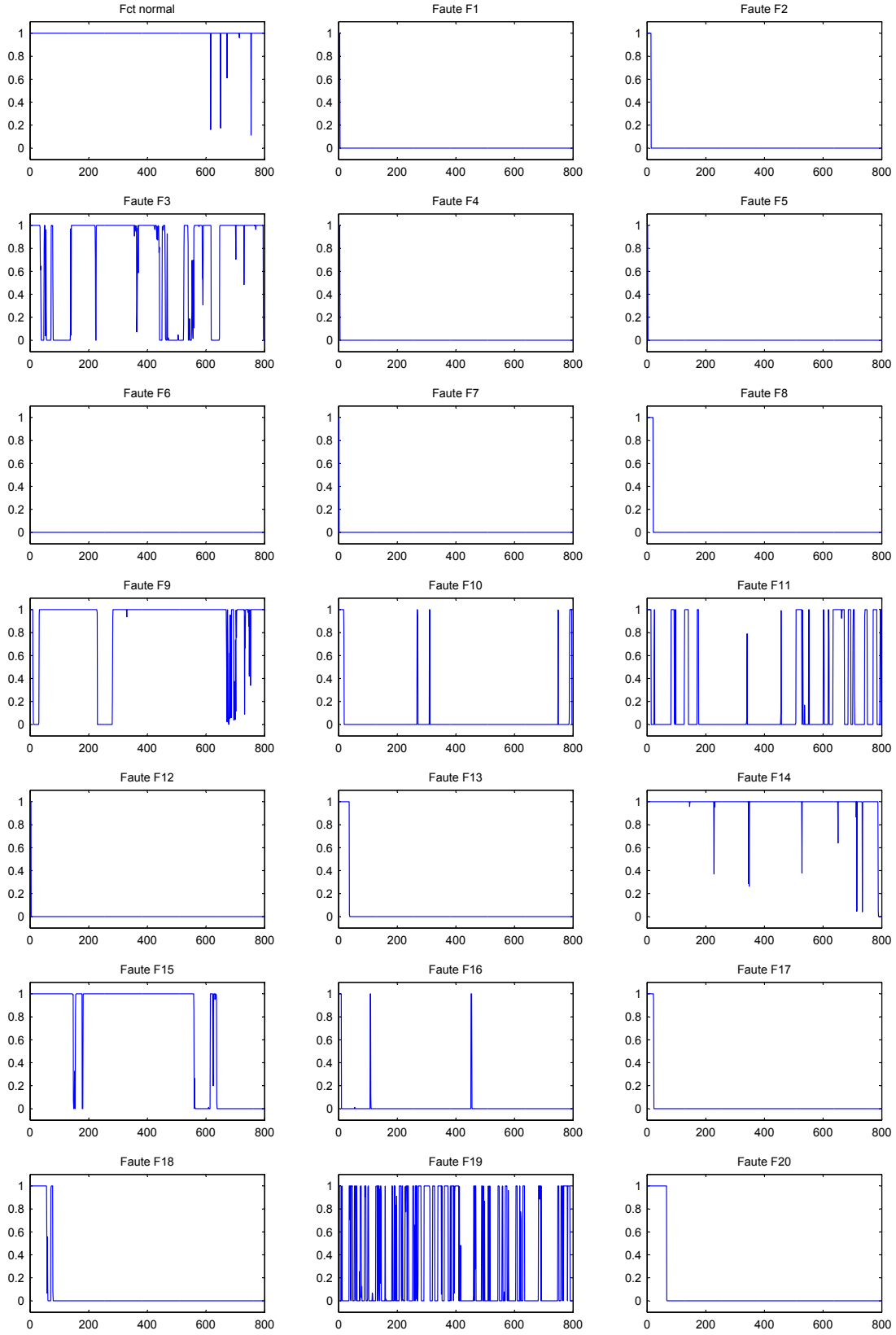


FIG. 4.5 – Modalité SC du nœud MEWMA

La comparaison des figures 4.4 et 4.5 pour les mêmes fautes permet de bien rendre compte de l'intérêt de la modélisation des deux cartes. En effet, pour certaines fautes (F3, F9, F10, F11, F15 et F16), dont l'amplitude est assez faible, la carte MEWMA détecte mieux que la carte T^2. Par contre, pour la faute F14, la carte MEWMA détecte seulement quelques échantillons alors que la carte T^2 les détecte tous (signifiant qu'il s'agit de saut d'amplitude importante). Ceci est dû au phénomène d'inertie lié à la carte MEWMA. Pour expliquer cela, nous donnons sur la figure 4.6 le signal de la variable 21 sur quelques échantillons. Les 20 premières valeurs sont issues du fonctionnement normal et la faute est introduite à partir de l'échantillon numéro 21. Sur cette figure, on observe que l'amplitude de la faute est grande, raison pour laquelle la carte T^2 détecte. Cependant, échantillon par échantillon, le signal est en dents de scie. La carte MEWMA subit alors son principal défaut : son inertie. Dans le calcul du signal Y de cette carte, une valeur élevée d'échantillon est compensée par la valeur faible de l'échantillon précédent, rendant le signal filtré Y relativement constant. La détection de saut est alors rendu impossible (par la carte MEWMA) sur ce type de faute. Cette faute met en évidence l'intérêt de l'utilisation, dans le même outil, des deux cartes T^2 et MEWMA : nous profitons des avantages de chaque carte.

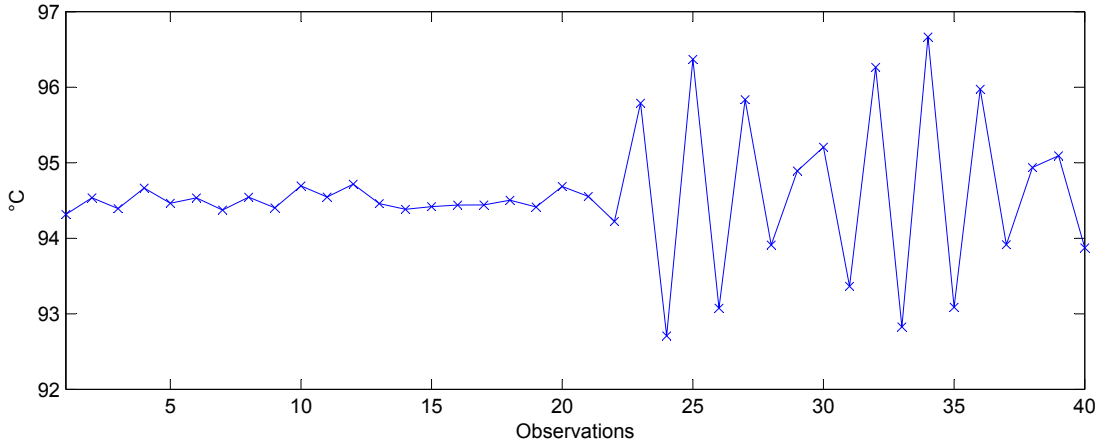


FIG. 4.6 – Signal de la variable 21 pour la faute F14

4.3.2 Diagnostic supervisé

Afin de tester les performances en diagnostic supervisé de la méthode proposée, nous étudions tout d'abord les taux de classification de notre méthode, sans prendre en compte de rejet de distance. Nous prenons en compte toutes les fautes, sans sélection de composantes et en supposant que chaque observation est détectée. Les bases de données d'apprentissage et de test sont celles décrites dans la table 4.5. Ainsi, nous avons 800 observations de chaque faute à classer, soit 16 000 observations.

Les résultats de cette classification sont présentés sous forme de trois matrices de confusion (matrice d'occurrences, matrice de précision et matrice de fiabilité). La matrice d'occurrences (table 4.7) donne pour chaque colonne testée (représentant 800 observations de la faute F_i) les différents classements du classifieur. Ainsi, la trace de cette matrice représente le nombre de bonnes classifications. Nous présentons également la matrice de précision (table 4.8). Cette matrice est construite en divisant chaque cellule de la matrice d'occurrences par la somme de la colonne (ici 800 observations pour chaque colonne), et elle est exprimée en pourcentage. Enfin, nous présentons la matrice de fiabilité (table 4.9), construite de la même manière que la matrice de précision, mais en prenant la somme de la ligne, et non plus de la colonne.

Sur la matrice de précision, on peut observer que certaines fautes sont très bien reconnues par le classifieur (fautes F1, F2, F5, F6, F7, F8, F12, F14 et F19), avec un taux supérieur à 95%. D'autres fautes sont très mal reconnues (F3, F9 et F15) avec un taux de reconnaissance inférieur à 25%. Enfin, les 8 fautes restantes sont moyennement reconnues puisque leurs taux de classification se situe entre 75% et 90%. Il est cependant intéressant de remarquer que pour les fautes F3, F9 et F15, les confusions portent principallement entre ces mêmes fautes : la faute F3 va être classée comme F9 ou F15 par exemple. Le taux de reconnaissance global (moyenne de la diagonale) atteint la valeur de 79.71%. Cela signifie qu'en moyenne, si l'on présente cinq observations au classifieur, il trouve la bonne classe, et donc le bon diagnostic, pour quatre de ces observations.

Un autre critère important est la fiabilité du classifieur. Il s'agit du nombre de fois que le classifieur a bien classé une faute, sur le nombre de fois qu'il a déclaré cette faute. Prenons par exemple le cas de la faute F18, la précision du classifieur pour cette faute est de $\frac{548}{800} = 68.5\%$. Sa fiabilité vaut $\frac{548}{562} = 97.51\%$, montrant alors que si le classifieur attribue la classe F18 à une observation, alors on est presque certain (à 97.51%) que ceci est le bon diagnostic. La table 4.9 donne les résultats de fiabilité pour chacune des fautes. Nous pouvons notamment observer que l'on ne peut pas accorder de confiance au diagnostic des fautes F3, F9 et F15 puisque leur taux de fiabilité est d'environ 25%. Cependant, le taux de fiabilité globale (moyenne de la diagonale) est tout de même de 79.55%.

!

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20
F1	780	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
F2	0	785	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	176	0	0	0	0	2	201	8	18	0	16	0	118	15	1	38	6	17
F4	0	0	0	659	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0
F5	0	0	0	0	784	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
F6	0	0	0	0	0	800	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	800	0	0	0	0	0	0	0	0	0	0	0	0	0
F8	18	5	1	0	0	0	0	776	0	4	0	1	109	0	1	26	0	0	0	0
F9	0	8	171	0	0	0	0	11	181	25	24	1	4	0	233	15	7	13	9	34
F10	0	0	48	0	0	0	0	0	40	695	9	0	0	0	64	48	0	3	5	6
F11	0	0	43	141	0	0	0	3	42	5	604	0	2	1	43	2	30	3	2	3
F12	0	0	0	0	16	0	0	4	0	6	0	786	41	0	4	10	0	168	0	23
F13	0	0	0	0	0	0	0	2	0	3	0	1	609	0	3	4	0	0	0	3
F14	0	0	17	0	0	0	0	0	10	3	28	0	0	790	20	4	71	0	0	1
F15	0	1	215	0	0	0	0	0	221	12	34	0	11	0	188	6	9	9	2	7
F16	0	1	85	0	0	0	0	0	39	35	5	0	2	0	82	645	1	10	4	3
F17	1	0	7	0	0	0	0	0	6	3	42	0	0	9	1	3	680	0	1	2
F18	0	0	0	0	0	0	0	0	0	0	0	8	5	0	0	0	0	548	1	0
F19	1	0	32	0	0	0	0	0	54	1	7	0	1	0	38	16	1	1	769	2
F20	0	0	5	0	0	0	0	0	6	0	2	0	0	0	5	6	0	7	1	699

TAB. 4.7 – Matrice d'occurrences

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**	**F19**	**F20**
F1	97.5	0	0	0	0	0	0	0.25	0	0	0	0	0	0	0	0	0	0	0	0
F2	0	98.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	22	0	0	0	0	0.25	25.12	1	2.25	0	2	0	14.75	1.88	0.13	4.75	0.75	2.13
F4	0	0	0	82.38	0	0	0	0	0	0	3.38	0	0	0	0	0	0	0	0	0
F5	0	0	0	0	98	0	0	0	0	0	0	0.38	0	0	0	0	0	0	0	0
F6	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
F8	2.25	0.63	0.13	0	0	0	0	97	0	0.5	0	0.13	13.63	0	0.13	3.25	0	0	0	0
F9	0	1	21.38	0	0	0	0	1.38	22.63	3.13	3	0.13	0.5	0	29.13	1.88	0.88	1.63	1.13	4.25
F10	0	0	6	0	0	0	0	0	5	86.88	1.13	0	0	0	8	6	0	0.38	0.63	0.75
F11	0	0	5.38	17.63	0	0	0	0.38	5.25	0.63	75.5	0	0.25	0.13	5.38	0.25	3.75	0.38	0.25	0.38
F12	0	0	0	0	2	0	0	0.5	0	0.75	0	98.25	5.13	0	0.5	1.25	0	21	0	2.88
F13	0	0	0	0	0	0	0	0.25	0	0.38	0	0.13	76.13	0	0.38	0.5	0	0	0	0.38
F14	0	0	2.13	0	0	0	0	0	1.25	0.38	3.5	0	0	98.75	2.5	0.5	8.88	0	0	0.13
F15	0	0.13	26.88	0	0	0	0	0	27.63	1.5	4.25	0	1.38	0	23.5	0.75	1.13	1.13	0.25	0.88
F16	0	0.13	10.63	0	0	0	0	0	4.88	4.38	0.63	0	0.25	0	10.25	80.63	0.13	1.25	0.5	0.38
F17	0.13	0	0.88	0	0	0	0	0	0.75	0.38	5.25	0	0	1.13	0.13	0.38	85	0	0.13	0.25
F18	0	0	0	0	0	0	0	0	0	0	0	1	0.63	0	0	0	0	68.5	0.13	0
F19	0.13	0	4	0	0	0	0	0	6.75	0.13	0.88	0	0.13	0	4.75	2	0.13	0.13	96.13	0.25
F20	0	0	0.63	0	0	0	0	0	0.75	0	0.25	0	0	0	0.63	0.75	0	0.88	0.13	87.38

TAB. 4.8 – Matrice de précision exprimée en %

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20
F1	99.74	0	0	0	0	0	0	0.21	0	0	0	0	0	0	0	0	0	0	0	0
F2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	28.57	0	0	0	0	0.21	27.31	0.87	1.95	0	2.56	0	16.5	1.64	0.13	6.76	0.65	2.33
F4	0	0	0	96.06	0	0	0	0	0	0	2.92	0	0	0	0	0	0	0	0	0
F5	0	0	0	0	99.62	0	0	0	0	0	0	0.28	0	0	0	0	0	0	0	0
F6	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
F8	2.3	0.64	0.16	0	0	0	0	82.47	0	0.44	0	0.09	17.44	0	0.14	2.85	0	0	0	0
F9	0	1.02	27.76	0	0	0	0	1.17	24.59	2.72	2.6	0.09	0.64	0	32.59	1.64	0.93	2.31	0.98	4.65
F10	0	0	7.79	0	0	0	0	0	5.43	75.71	0.97	0	0	0	8.95	5.26	0	0.53	0.54	0.82
F11	0	0	6.98	20.55	0	0	0	0.32	5.71	0.54	65.37	0	0.32	0.11	6.01	0.22	3.97	0.53	0.22	0.41
F12	0	0	0	0	2.03	0	0	0.43	0	0.65	0	74.29	6.56	0	0.56	1.1	0	29.89	0	3.15
F13	0	0	0	0	0	0	0	0.21	0	0.33	0	0.09	97.44	0	0.42	0.44	0	0	0	0.41
F14	0	0	2.76	0	0	0	0	0	1.36	0.33	3.03	0	0	83.69	2.8	0.44	9.4	0	0	0.14
F15	0	0.13	34.9	0	0	0	0	0	30.03	1.31	3.68	0	1.76	0	26.29	0.66	1.19	1.6	0.22	0.96
F16	0	0.13	13.8	0	0	0	0	0	5.3	3.81	0.54	0	0.32	0	11.47	70.72	0.13	1.78	0.43	0.41
F17	0.13	0	1.14	0	0	0	0	0	0.82	0.33	4.55	0	0	0.95	0.14	0.33	90.07	0	0.11	0.27
F18	0	0	0	0	0	0	0	0	0	0	0	0.76	0.8	0	0	0	0	97.51	0.11	0
F19	0.13	0	5.19	0	0	0	0	0	7.34	0.11	0.76	0	0.16	0	5.31	1.75	0.13	0.18	83.32	0.27
F20	0	0	0.81	0	0	0	0	0	0.82	0	0.22	0	0	0	0.7	0.66	0	1.25	0.11	95.62

TAB. 4.9 – Matrice de fiabilité exprimée en %

4.3.2.1 Diagnostic sur les 15 premières fautes avec sélection de composantes

Bien que les performances de ce classifieur soient acceptables pour ce type de procédé et vu le nombre de classe, il n'est pas réellement adapté au diagnostic de système puisqu'il est incapable de reconnaître un nouveau type de faute. Nous allons à présent intégrer au classifieur la notion de rejet de distance, et la sélection de variables importantes. Afin d'évaluer la prise en compte de nouveaux types de faute, nous ne prenons désormais que les 15 premières fautes (F1 à F15) comme ensemble d'apprentissage. Les 5 dernières fautes (F16 à F20) sont inconnues pour le classifieur. Celui-ci doit donc les classer en tant que nouveau type de faute (classe NF).

Afin d'étudier l'influence de la sélection de variables sur les performances de classification, nous évaluons tout d'abord le classifieur (sur les fautes F1 à F15) dans l'espace décrit par les 52 variables. Les résultats détaillés de cette évaluation sont donnés par les différentes matrices de confusion (occurrences, précision et fiabilité) en annexe A.4. La table 4.10 présentée ci-dessous, donne les résultats principaux de cette classification (diagonales des matrices de précision et de fiabilité). Il est possible d'améliorer ces résultats en appliquant une sélection des variables importantes. Dans ce but, nous avons appliqué la méthode proposée à la section 3.3.1. L'algorithme proposé a permis de sélectionner les 27 variables suivantes (en ordre de la plus informative vers la moins informative) : 51, 10, 1, 44, 9, 16, 46, 13, 50, 45, 21, 18, 34, 19, 11, 38, 20, 47, 31, 42, 4, 30, 43, 5, 35, 52 et 17. Les diagonales des matrices de précision et de fiabilité obtenues en appliquant cette sélection de variables sont également présentées dans la table 4.10, alors que les détails de chaque matrice de confusion sont donnés en annexe A.5.

Dans la table 4.10, on remarque que la sélection de variables permet une augmentation de la précision d'environ 0.77%, soit près de 100 observations mieux classées. Cependant, l'augmentation de la fiabilité est encore plus forte que celle de la précision puisque nous passons d'un taux de 79.32% à 80.7%, soit une augmentation de 1.38%. La sélection de variables est donc bénéfique au classifieur puisqu'elle permet d'augmenter à la fois sa précision et sa fiabilité. Le bénéfice reste assez minime vu le nombre de classes (15 fautes) et le nombre de variables initiales (52 variables). Pour des applications où le nombre de classes est bien plus faible que le nombre de variables, l'amélioration obtenue par la sélection de variables est impressionnante. Par exemple, en ne prenant en compte que les fautes F4, F9 et F11, le taux d'erreur diminue d'environ 19% (avec les 52 variables) à environ 5% (avec trois variables sélectionnées : 51, 9 et 21) [152].

	Précision (en %)		Fiabilité (en %)	
Faute	52 variables	27 variables	52 variables	27 variables
F1	97.5	98.13	99.74	99.75
F2	98.13	98.25	100	100
F3	25	28.75	33.84	33.09
F4	82.38	91.88	95.92	95.95
F5	98	98.38	99.62	99.49
F6	100	100	100	100
F7	100	100	100	100
F8	97	97	84.62	83.35
F9	25	24.88	27.1	27.07
F10	90.75	90.38	75.78	79.54
F11	80.75	81.88	66.74	81.27
F12	99.25	98.75	91.37	91.65
F13	76.38	72	98.55	98.8
F14	99.88	99.88	88.38	94.44
F15	27.88	29.38	28.09	26.02
Moyenne	79.86	80.63	79.32	80.7

TAB. 4.10 – Précision et fiabilité du classifieur sur les 15 premières fautes

4.3.2.2 Prise en compte du rejet de distance

Nous allons maintenant ajouter au classifieur (toujours avec sélection de variables) le rejet de distance, pour ainsi pouvoir diagnostiquer l'apparition de nouveaux types de fautes. Pour cela, le taux de rejet de distance a été fixé à 0.001 sur chaque type de faute. Nous précisons que ce taux peut être choisi différemment pour chaque type de faute. Dans un premier temps, nous étudions les performances en ne testant que les fautes F1 à F15, afin de voir quelles répercussions engendre la prise en compte du rejet de distance sur les fautes que le classifieur est sensé connaître. La matrice de confusion des différentes occurrences est donnée sur la table 4.11. On remarque que, comme auparavant, les fautes F3, F9 et F15 sont confuses (l'une est prise pour l'autre et vice versa). Un nouvel élément à analyser est le fait que deux fautes (F6 et F13) sont très souvent classées en tant que nouvelle faute NF. Sur les 800 observations de test de la faute F6, 798 sont classées comme NF. Or, auparavant (sans rejet de distance), cette faute était bien classée dans 100% des cas. La même chose se passe (en plus atténuée) pour la faute F13. En réalité, toute méthode de classification classique avec rejet de distance devra subir cet inconvénient.

La faute F6 provoque un phénomène de saturation du procédé (voir [17]). La faute F6 est un défaut d'alimentation en gaz A. Dès qu'il n'y a plus de gaz A dans le réacteur, les réactions chimiques ne se font plus, malgré que le réacteur soit toujours alimenté en

gaz D et E. Cette alimentation en D et E fait augmenter la pression du réacteur jusqu'à un point de sécurité de 2950kPa. Cette saturation entraîne un blocage de beaucoup de variables (toujours à la même valeur pendant des heures), réduisant leurs variabilités. Or, le blocage des variables ne se situe pas toujours exactement au même point, engendrant une moyenne différente d'une faute F6 à une autre faute F6. Comme la variabilité est faible et que la moyenne est différente, la faute F6 fait l'objet d'un rejet de distance. Pour mieux comprendre, nous donnons l'exemple des variables 3 et 11. La figure 4.7 présente ces deux variables dans le cas des données d'apprentissage et de test (en pointillé) de la faute F6. Nous donnons des graphiques similaires pour les 52 variables en annexe A.6.

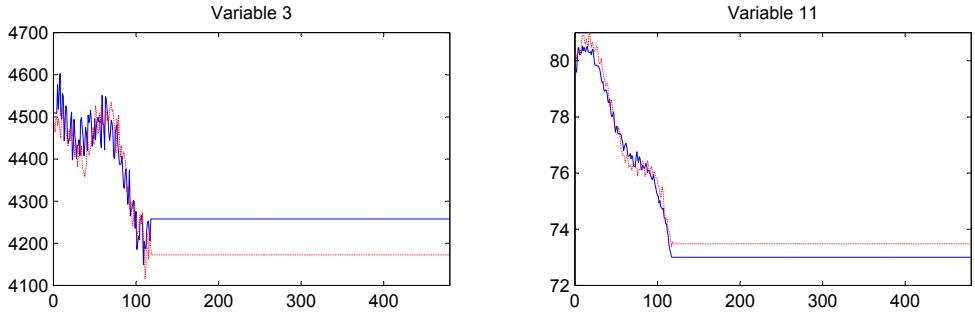


FIG. 4.7 – Variable 3 et 11 dans le cas de la faute F6 pour les données d'apprentissage (en bleu) et de test (en rouge pointillé)

En pratique, le diagnostic de cette faute ne sera pas trop compliquée. Les valeurs des différentes variables, bien que différentes d'une faute F6 à une autre faute F6, restent bloquées à des valeurs très grandes mais finalement assez proches comparées aux valeurs prises par les variables pour les autres types de fautes. Un simple aperçu des données, suite à la classification en nouvelle faute NF, permet de reconnaître la faute F6. Nous avons testé cinq nouvelles fautes (non connues par le classifieur) : les fautes F16 à F20. La matrice de confusion des différentes occurrences est donnée sur la table 4.12. Le classifieur reconnaît une nouvelle faute NF dans 2081 observations sur les 4000 (5×800) testées, soit un taux de reconnaissance de nouvelles fautes d'environ 52%.

Nous avons établi, au vu de ses performances, que la méthode proposée permet un diagnostic correct de nombreuses fautes du procédé, et qu'elle est aussi capable de reconnaître des nouveaux types de faute. Nous avons observé un léger inconvénient du rejet de distance. Celui-ci a complètement ignoré la faute F6 alors qu'elle était parfaitement reconnue avant (de même pour F13). De manière récurrente, nous avons également observé que le diagnostic des fautes F3, F9 et F15 est difficile. Lorsque le classifieur attribue la classe nouvelle faute NF à une observation, la méthode que nous proposons permet d'aider le responsable du procédé à diagnostiquer cette faute grâce à un diagnostic non-supervisé.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
F1	779	0	0	0	0	0	0	1	0	0	0	0	0	0	0
F2	0	778	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	229	0	0	0	0	2	241	16	13	0	20	0	173
F4	0	0	0	731	0	0	0	0	0	0	30	0	0	0	0
F5	0	0	0	0	782	0	0	0	0	0	0	4	0	0	0
F6	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	774	0	0	0	0	0	0	0	0
F8	15	3	1	0	0	0	0	696	0	3	0	2	53	0	2
F9	0	8	189	0	0	0	0	12	197	25	30	2	7	0	259
F10	0	0	69	0	0	0	0	1	19	704	9	0	2	0	88
F11	0	0	21	63	1	0	0	2	21	5	640	0	3	1	35
F12	0	0	0	0	15	0	0	2	0	17	4	697	15	0	3
F13	0	0	0	0	0	0	0	3	0	2	2	2	157	0	0
F14	1	0	11	1	0	0	0	0	5	2	21	0	0	799	6
F15	1	3	280	0	0	0	0	0	317	17	38	0	10	0	234
NF	4	8	0	5	2	798	26	81	0	9	13	93	533	0	0

TAB. 4.11 – Matrice d'occurrence sans test de nouvelles fautes

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	NF
F1	779	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
F2	0	778	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	229	0	0	0	0	2	241	16	13	0	20	0	173	162
F4	0	0	0	731	0	0	0	0	0	0	30	0	0	0	0	0
F5	0	0	0	0	782	0	0	0	0	0	0	4	0	0	0	2
F6	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	774	0	0	0	0	0	0	0	0	0
F8	15	3	1	0	0	0	0	696	0	3	0	2	53	0	2	420
F9	0	8	189	0	0	0	0	12	197	25	30	2	7	0	259	105
F10	0	0	69	0	0	0	0	1	19	704	9	0	2	0	88	179
F11	0	0	21	63	1	0	0	2	21	5	640	0	3	1	35	106
F12	0	0	0	0	15	0	0	2	0	17	4	697	15	0	3	462
F13	0	0	0	0	0	0	0	3	0	2	2	2	157	0	0	73
F14	1	0	11	1	0	0	0	0	5	2	21	0	0	799	6	324
F15	1	3	280	0	0	0	0	0	317	17	38	0	10	0	234	86
NF	4	8	0	5	2	798	26	81	0	9	13	93	533	0	0	2081

TAB. 4.12 – Matrice d'occurrence avec test de nouvelles fautes

4.3.3 Diagnostic non-supervisé

Dans le cas d'une nouvelle faute, il est important de pouvoir fournir des indications au responsable du procédé, afin que celui-ci puisse diagnostiquer la faute et y remédier. La méthode de surveillance par réseaux bayésiens permet une identification des variables responsables d'une situation hors-contrôle.

Pour illustrer cette méthode sur le TEP, nous nous proposons d'étudier l'observation 130 des données test de la faute F4. Cette observation détectée a été classée NF (nouveau type de faute) par le module de diagnostic supervisé. On cherche donc à essayer de caractériser cette faute.

La première action effectuée est de construire le réseau conditionnel gaussien du module de diagnostic supervisé (voir §3.4). Pour cela, l'algorithme PC (§3.4.1.1) est utilisé pour la construction de la structure du réseau. Pour l'algorithme PC, nous avons employé le test d'indépendance conditionnelle énoncé à la section 3.4.1.2 avec un seuil α de 0.05, comme préconisé par Kalish et Buhlmann [70]. La figure 4.8 présente la structure de réseau construit. On ajoute alors tous les nœuds de contrôle de paramètres, c'est à dire tous les nœuds $T^2_{i\bullet PA(X_i)}$, avec un taux de fausses alertes de 0.005. Suite à cela, les paramètres de chaque nœud sont calculés. Les données utilisées pour la construction et l'apprentissage du réseau sont les données d'apprentissage de la période de fonctionnement normal (soit 480 observations sans faute).

Suite à l'apprentissage de la structure et des paramètres du réseau, nous présentons l'observation pour laquelle nous voulons obtenir un diagnostic non-supervisé. On observe alors les probabilités de tous les nœuds $T^2_{i\bullet PA(X_i)}$. La probabilité de SC (sous contrôle) de chacun de ces nœuds est supérieur à 0.995 (supérieur au seuil d'alerte) excepté pour deux nœuds. Il s'agit des variables 9 et 51 (avec des probabilités respectives d'environ 30% et 0%). Ces deux variables sont donc impliquées dans la situation de hors-contrôle détectée et diagnostiquée comme nouvelle faute. La variable 9 représente la température du réacteur, alors que la variable 51 représente le débit du refroidissement liquide au réacteur. On comprend tout de suite qu'un problème de refroidissement liquide est intervenu dans le procédé, ce qui est bien le cas de la faut eF4.

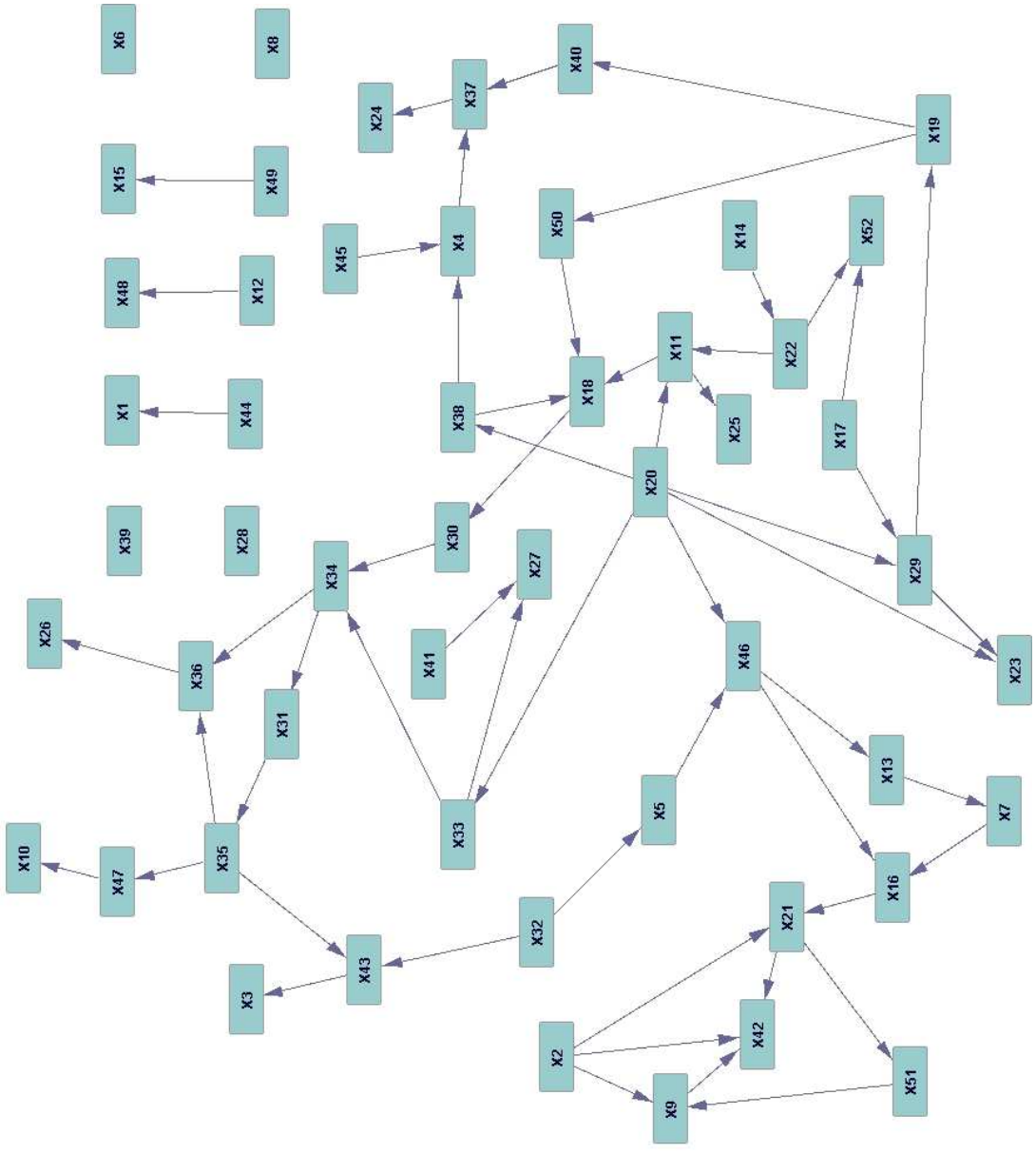


FIG. 4.8 – Réseau issu de l'algorithme PC

4.4 Conclusion

Ce chapitre a permis d'illustrer notre approche de surveillance des procédés par réseaux bayésiens. Pour cela, nous avons choisi un procédé multivarié très utilisé dans la littérature : le Tennessee Eastman Process (TEP). Suite à une présentation détaillée du TEP (procédé à 52 variables et 20 types de fautes), nous avons étudié les réponses du réseau que ce soit en terme de détection ou de diagnostic. Nous avons pu remarquer que pour une majorité des fautes, la détection était rapide et fiable, mais que pour certaines (les fautes F3, F9 et F15), les moyens de détection mis en place sont insuffisants. Nous avons également pu apprécier les performances de classification supervisée de notre méthode. Nous avons alors pu constater que l'algorithme de sélection de variables était bénéfique, et que l'inclusion du rejet de distance pouvait engendrer quelques problèmes (cas des fautes F6 et F13). Là encore, les fautes F3, F9 et F15 posent problème, elles ne sont pas bien diagnostiquées. Ceci est tout de même logique puisque si elles ne sont pas bien détectées, c'est que leur distributions respectives ressemblent beaucoup à celle du fonctionnement normal, et que donc elles se ressemblent également entre-elles. Il est donc logique d'avoir des difficultés à les discriminer si elles se ressemblent. Enfin, nous avons également pu apprécier, sur un exemple simple, l'apport d'information auquel contribue le module d'identification (diagnostic non-supervisé).

Conclusion générale

Dans le contexte économique actuel, la performance des entreprises doit être toujours croissante. Celles-ci doivent produire toujours mieux, à moindre coût et dans des conditions de sécurité de plus en plus sévères. De plus, les procédés sont de plus en plus complexes et de plus en plus informatisés. Ainsi, il est de moins en moins évident ou intuitif de savoir si tout se passe bien dans un procédé. Dans ce but, la surveillance des procédés permet la détection et le diagnostic d'anomalies (de fautes). Ainsi, plus une faute est rapidement détectée, et correctement diagnostiquée, et plus la production du procédé sera conforme aux exigences requises, dans les conditions de sécurité requises. Nous avons ainsi proposé une surveillance des procédés multivariés par réseaux bayésiens permettant d'inclure dans un seul et même outil : un organe de détection et un organe de diagnostic supervisé et non-supervisé.

Dans le premier chapitre, nous avons pu voir rapidement les principales approches possibles (méthodes à base de modèles analytiques, méthodes à base de connaissances, méthodes basées sur les données) pour la surveillance des procédés. Nous avons également présenté les différentes étapes permettant la maîtrise d'un procédé (détection, diagnostic, reconfiguration). En ciblant notre étude sur les méthodes à base de données historiques du procédé, nous avons alors passé en revue quelques méthodes de détection et de diagnostic non-supervisées (se basant uniquement sur des échantillons de période de fonctionnement normal du procédé), mais également quelques méthodes supervisées (se basant sur des échantillons de période de faute dans le procédé). Nous avons alors, sur la base de plusieurs critères, effectué un choix parmi les classifieurs présentés dans les méthodes supervisées. Ce choix s'est porté sur les réseaux bayésiens. Les réseaux bayésiens ont notamment comme qualité le fait de pouvoir manipuler des variables discrètes et continues pouvant être corrélées, de fournir des réponses rapidement et ce pour des temps d'apprentissages raisonnables. De plus, ils permettent facilement la prise en compte de plusieurs classes de fonctionnement et les algorithmes sont capables de gérer des données manquantes. Enfin, les réseaux bayésiens sont à la base de beaucoup d'extensions comme les réseaux bayésiens dynamiques (prise en compte du temps), les réseaux bayésiens orientés objet (permettant

une représentation plus simple d'importants systèmes) ou bien les diagrammes d'influence (prise en compte de décisions et d'utilités associées aux variables du réseau).

Suite au choix de ce classifieur, nous avons présenté les réseaux bayésiens de manière plus approfondie dans le second chapitre. Après la définition formelle d'un réseau bayésien, nous avons présenté le principe de fonctionnement et les calculs associés sur un exemple très simple. Nous avons alors détaillé un peu plus le fonctionnement d'un réseau en étudiant les différentes relations possibles entre les nœuds (discrets et continus). Nous nous sommes alors également intéressés aux extensions des réseaux bayésiens. Ainsi, nous avons brièvement présenté ce que sont les réseaux bayésiens dynamiques, les réseaux bayésiens orientés objet, ainsi que les diagrammes d'influence. Suite à cette présentation des réseaux bayésiens, nous avons établi un état de l'art des méthodes de diagnostic utilisant cet outil. Ainsi, nous avons pu voir plusieurs approches intéressantes, et nous avons pu souligner des erreurs à éviter et des idées intéressantes à exploiter.

Dans le troisième chapitre, nous nous sommes intéressés à la réalisation, par réseaux bayésiens, des différentes phases de la surveillance des procédés : détection, diagnostic supervisé et diagnostic non-supervisé. Dans un premier temps, nous avons proposé de réaliser la détection de faute dans un réseau bayésien. Dans ce but, nous avons prouvé l'équivalence entre une analyse discriminante (modélisée par réseaux bayésiens) et une carte de contrôle multivariée. Ainsi, nous avons rendu possible l'application des cartes de contrôle multivariées de T^2 de Hotelling et MEWMA (l'application étant également possible pour les cartes univariées) dans un réseau bayésien. Suite à cela, nous avons proposé un module de diagnostic supervisé. En plus de la fonction classique de classification pour le diagnostic, ce module permet un rejet de distance capable de diagnostiquer l'apparition de nouveaux types de fautes. De plus, nous avons démontré un nouveau résultat concernant l'information mutuelle. Ce résultat a permis d'établir un algorithme de sélection de variables permettant une augmentation des performances du classifieur supervisé. Concernant le diagnostic non-supervisé, nous avons proposé l'amélioration d'une méthode exploitant déjà les réseaux bayésiens. Enfin, nous avons exposé de manière plus générale le principe de surveillance des procédés par réseaux bayésiens. Ainsi, nous avons présenté un réseau permettant de prendre en compte les trois modules développés (modules de détection, de diagnostic supervisé et non supervisé).

Le dernier chapitre a permis d'illustrer la méthode proposée sur un exemple concret de procédé multivarié : le procédé Tennessee Eastman. Nous avons alors pu voir le comportement du réseau face à ce procédé complexe (52 variables) et soumis à 20 types de fautes différentes.

Les perspectives de ces travaux sont nombreuses. Concernant la détection, la modélisation des cartes de contrôle multivariées fait intervenir la multiplication de chaque terme de la matrice de variance-covariance par un coefficient c. Il serait envisageable d'étudier les performances en détection lorsque c n'est plus le même pour chaque terme, privilégiant ainsi la détection sur des régions particulières de l'espace multivarié. De même, pour la carte MEWMA, le filtrage des données est supposé être réalisé avant l'entrée des données dans le réseau. Il serait envisageable d'étudier la possibilité d'une carte de contrôle MEWMA modélisée par un réseau bayésien dynamique, permettant ainsi le filtrage des données directement dans le réseau. Un autre aspect encourageant est le fait qu'il est possible de modéliser une Analyse en Composantes Principales (ACP) dans un réseau bayésien. Il serait alors intéressant d'étudier la transposition des différentes techniques exploitant l'ACP, dans un réseau bayésien.

Une autre perspective évidente tient au fait que toutes les données continues sont pour le moment modélisées comme étant des variables gaussiennes. Bien entendu, en pratique, ceci n'est pas toujours vrai. Or, il est possible de réaliser des modèles à mélange de gaussiennes par réseaux bayésiens. Bien que pour le diagnostic supervisé, l'application est directe, concernant le rejet de distance et la détection, une étude serait demandée. De même, il serait également intéressant d'adapter l'algorithme de sélection de variables aux mélanges de gaussiennes. Cependant, à notre connaissance, l'entropie d'un mélange de gaussiennes ne possède pas de formule analytique et l'algorithme est donc pour le moment difficilement transposable.

Enfin, les perspectives les plus directes et les plus intéressantes tiennent dans l'application des extensions des réseaux bayésiens. En effet, il serait possible de modéliser la méthode proposé par réseaux bayésiens orientés objet, afin de pouvoir représenter plus facilement les différents modules, ainsi que l'intérieur de chacun d'eux (surtout les modules de diagnostic supervisé et non-supervisé). De même, l'extension de la méthode proposée aux diagrammes d'influence permettrait au responsable du procédé d'opter pour des décisions en prenant en compte des aspects de coûts toujours très importants.

Pour conclure, nous estimons que les réseaux bayésiens sont un outil prometteur dans le domaine de la surveillance des procédés complexes, une nouvelle voie malheureusement très peu exploitée jusqu'à présent et qui mériterait d'être davantage approfondie dans l'avenir.

Annexes

A.1 Abaques du coefficient c

La table A.1 présente les valeurs du coefficient c permettant de modéliser une carte MEWMA, pour des taux de fausses alertes α de 1% ou 0.5%, pour un nombre de variables p de 1 à 50, et pour un λ de 0.05 à 1 (équivalent à la carte T^2).

λ	α \ p	1	2	5	10	20	50
1	1%	754.54	95.28	17.14	7.46	4.17	2.48
	0.5%	2634.5	194.63	24.91	9.53	4.92	2.75
0.5	1%	727.41	93.33	16.91	7.37	4.1	2.47
	0.5%	2517.67	184.47	24.67	9.53	4.92	2.75
0.2	1%	698.18	91.26	16.69	7.32	4.06	2.46
	0.5%	2466.7	127.9	23.86	9.43	4.89	2.73
0.1	1%	650.81	90.29	16.57	7.3	4.01	2.45
	0.5%	2228.61	111.49	23.42	9.37	4.81	2.72
0.05	1%	553.46	89.2	16.25	7.22	3.96	2.42
	0.5%	1518.64	107.94	22.93	9.1	4.69	2.68

Tab. A.1 – Coefficient c pour la carte MEWMA

A.2 Variables du TEP en fonctionnement normal

Les figures A.1 à A.5 représentent les 52 variables du TEP lors du fonctionnement normal. Les variables mesurées en continu sont les variables de 1 à 22, les variables échantillonnées des concentrations des différents composés chimiques sont les variables de 23 à 41, et les 11 variables d'asservissement sont les variables de 42 à 52.

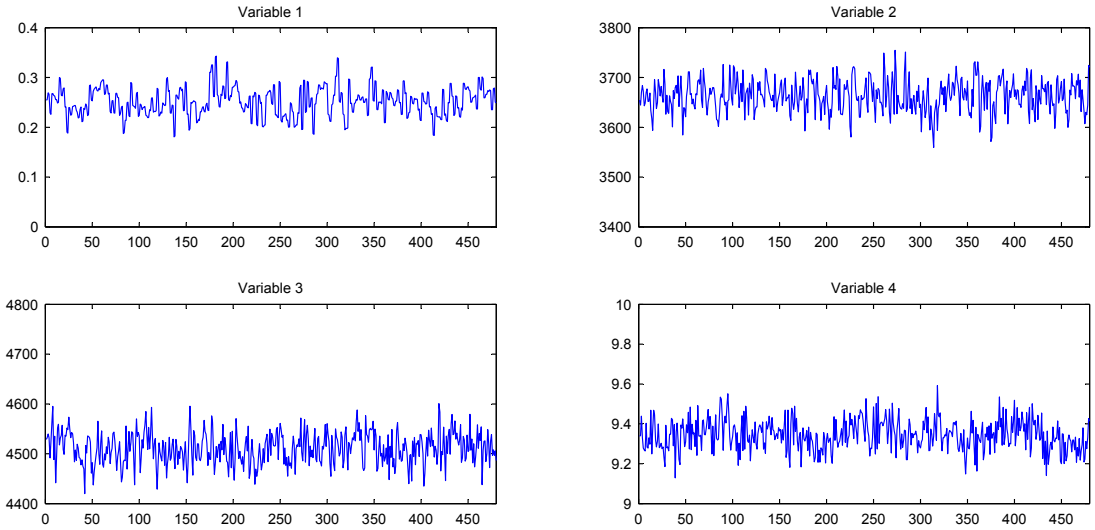


FIG. A.1 – Variables 1 à 4 en fonctionnement normal

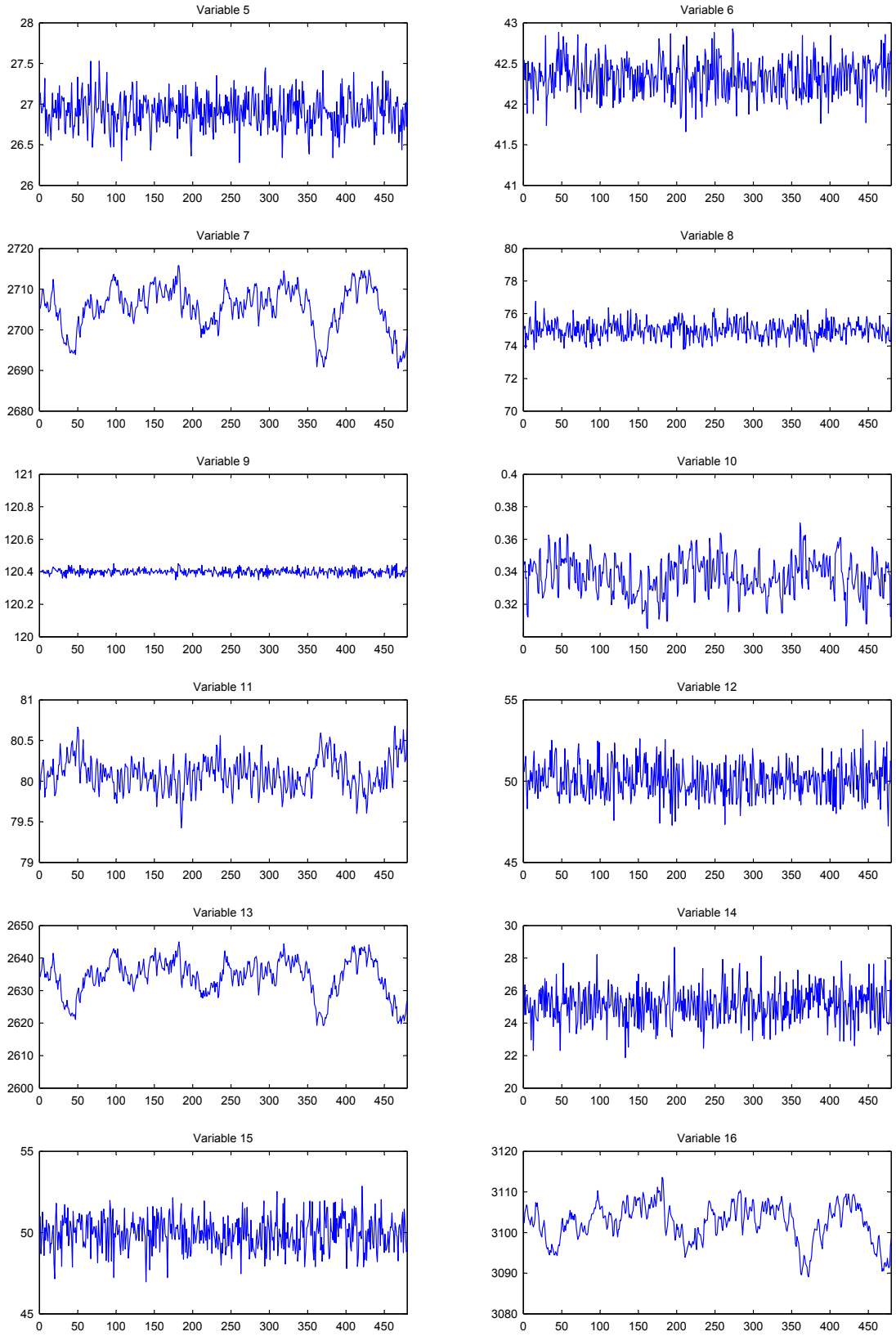


FIG. A.2 – Variables 5 à 16 en fonctionnement normal

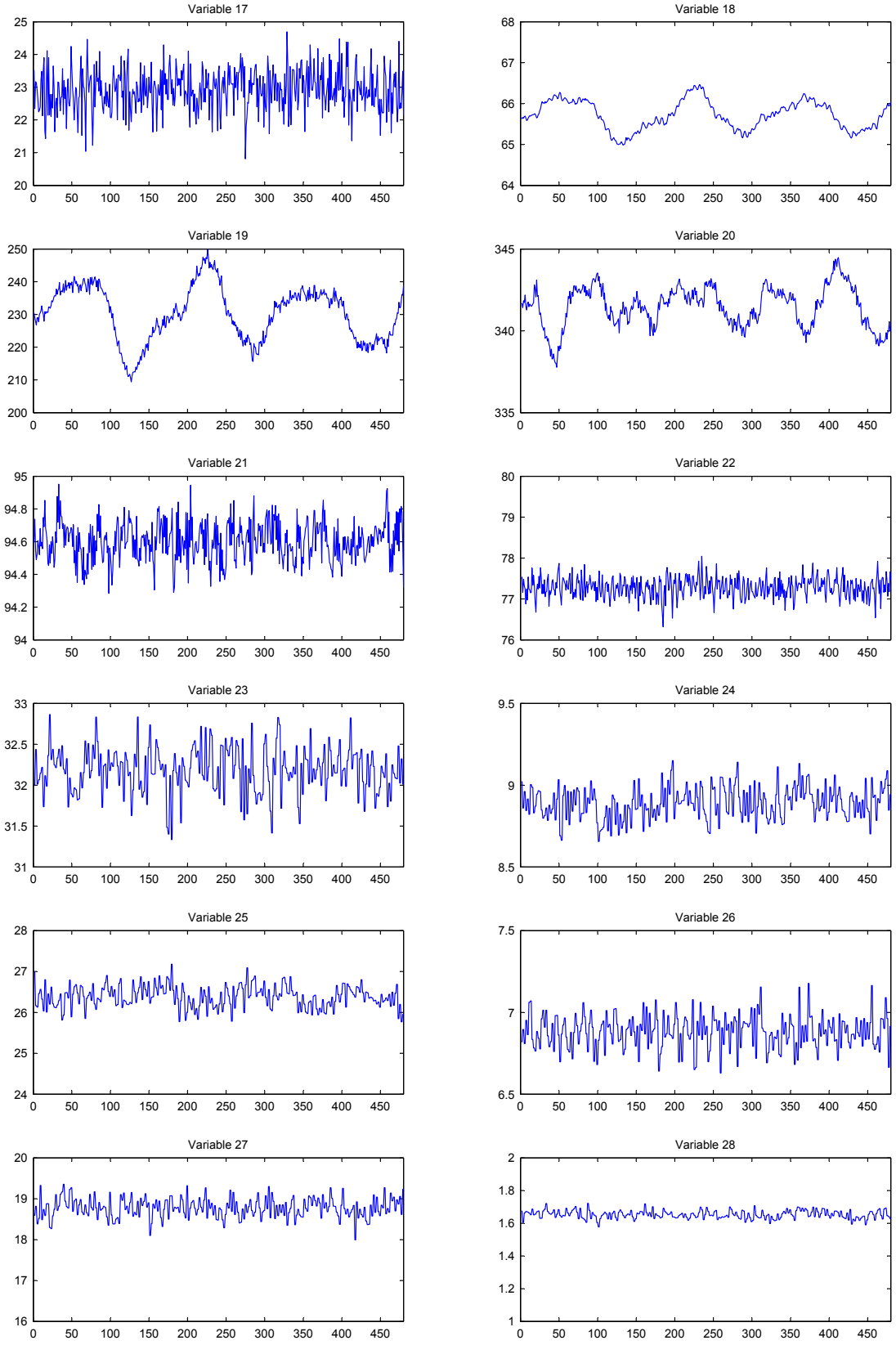


FIG. A.3 – Variables 17 à 28 en fonctionnement normal

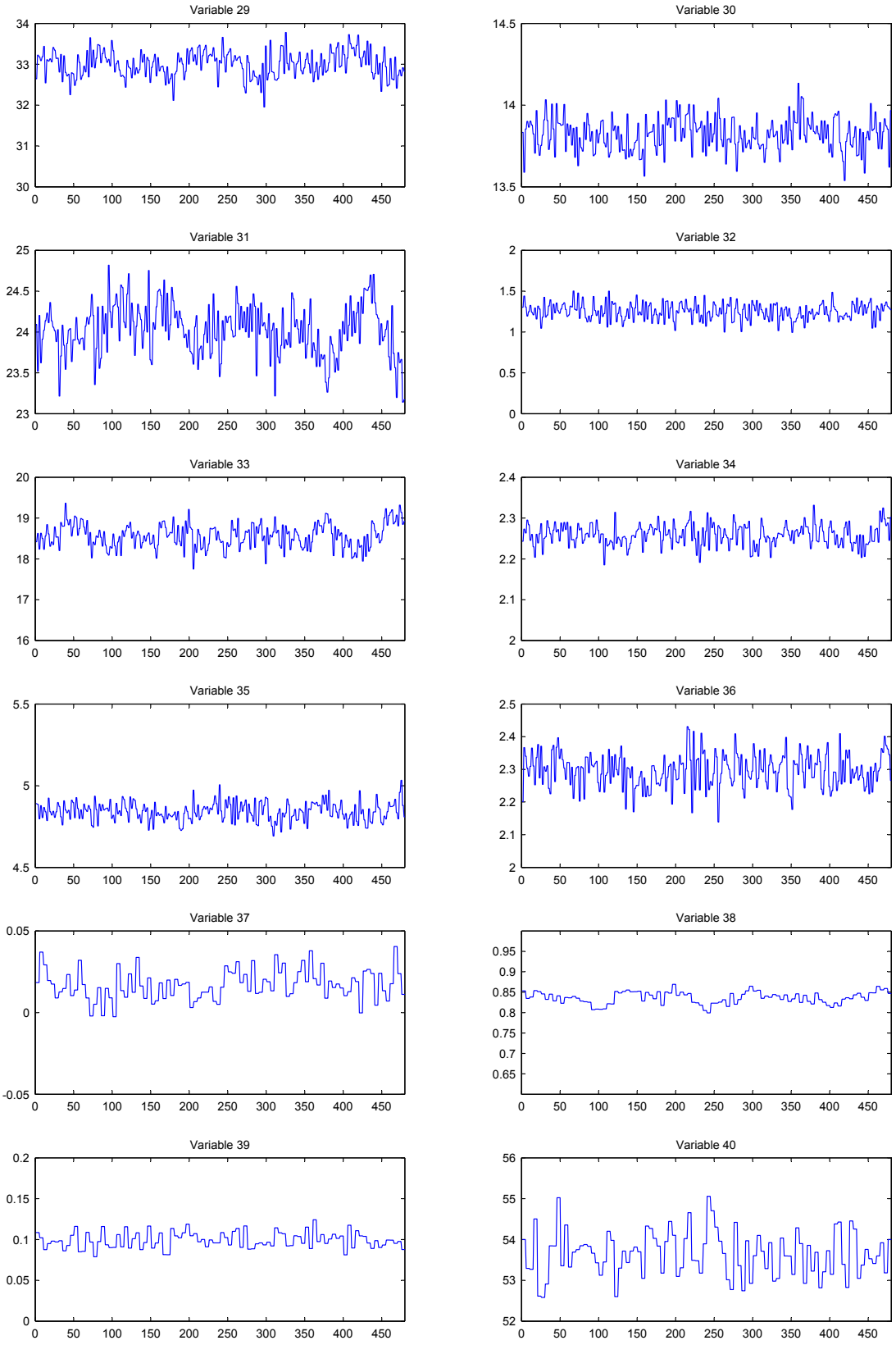


FIG. A.4 – Variables 29 à 40 en fonctionnement normal

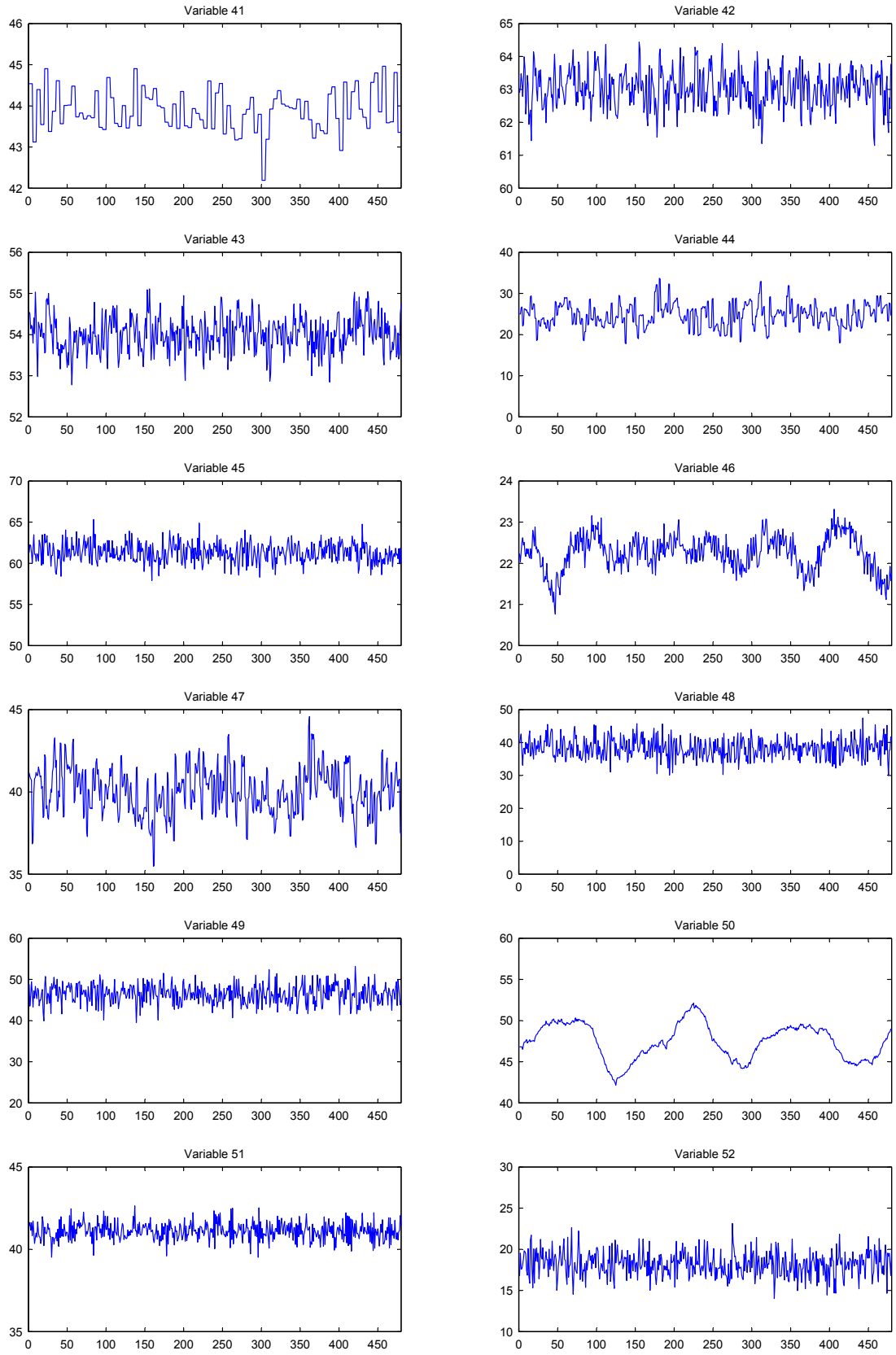


FIG. A.5 – Variables 41 à 52 en fonctionnement normal

A.3 Variables du TEP pour la faute F9

Les figures A.6 à A.13 représentent les 52 variables du TEP lors du fonctionnement normal (F0), et lors de la faute F9.

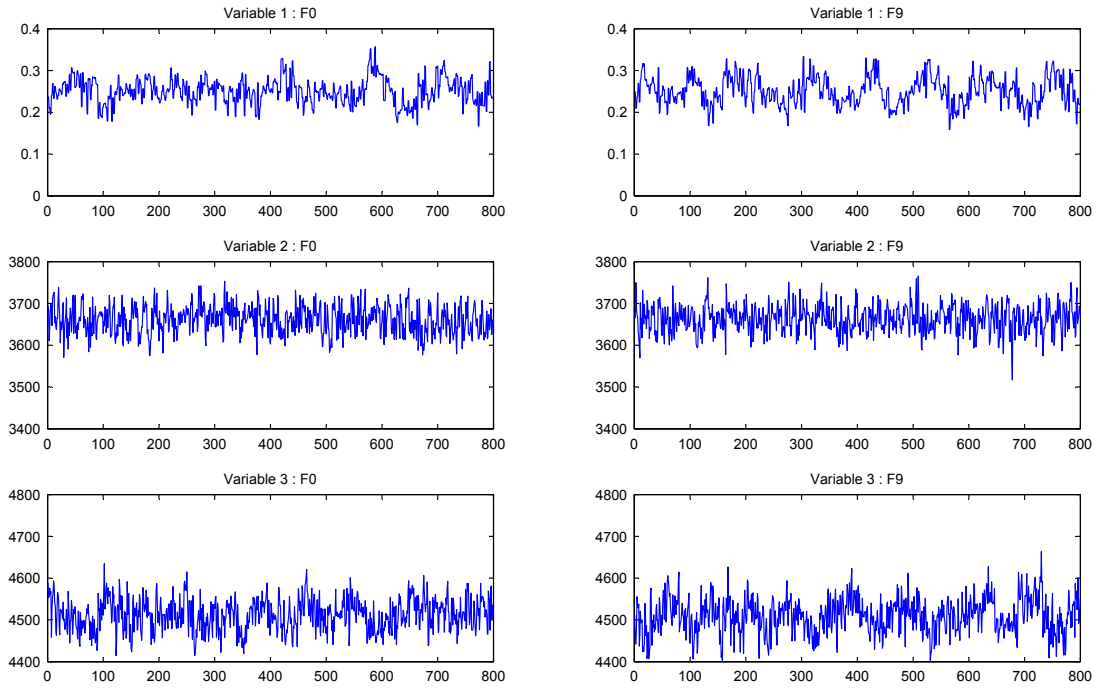


FIG. A.6 – Variables 1 à 3 en fonctionnement normal (F0) et pour la faute F9

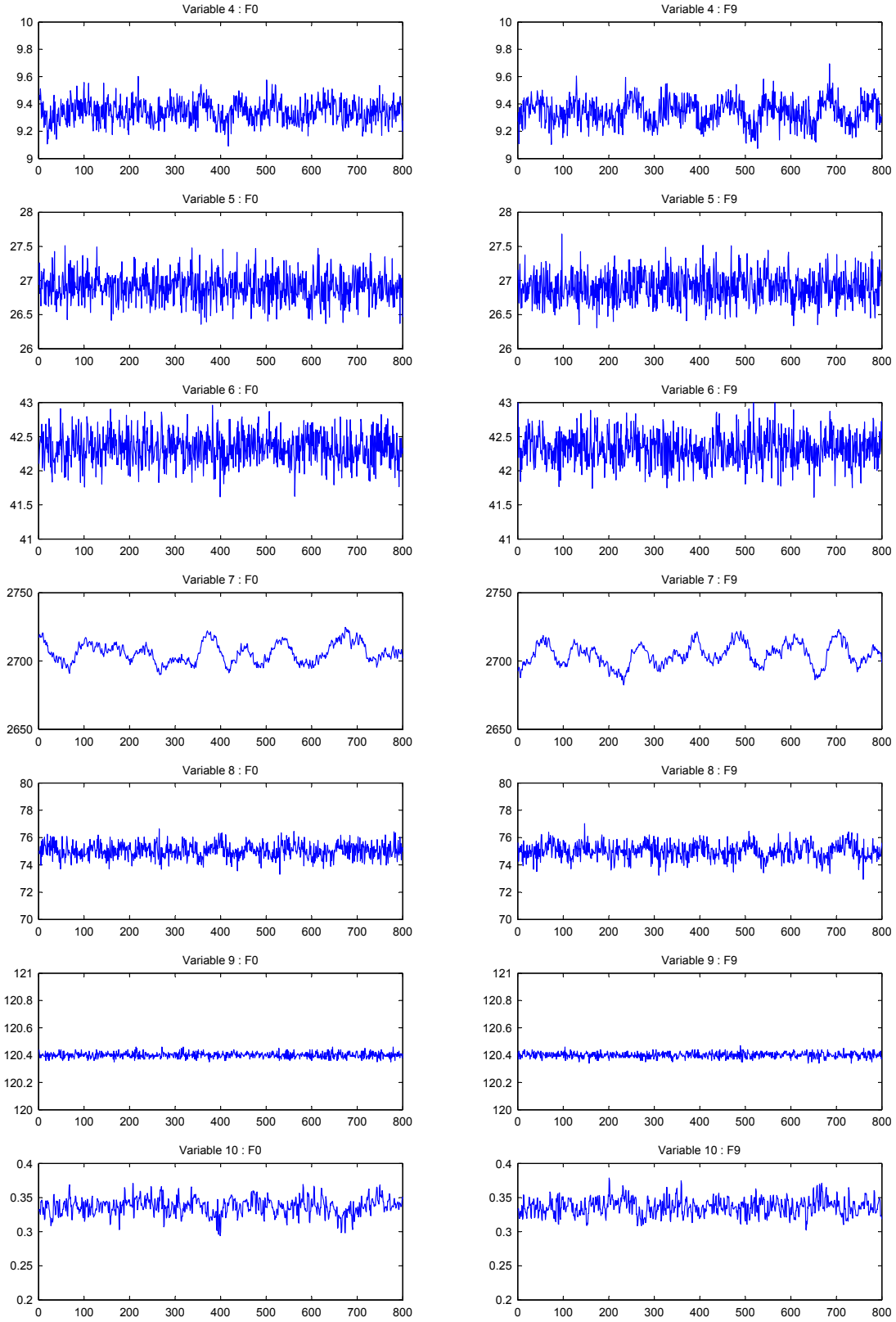


FIG. A.7 – Variables 4 à 10 en fonctionnement normal (F0) et pour la faute F9

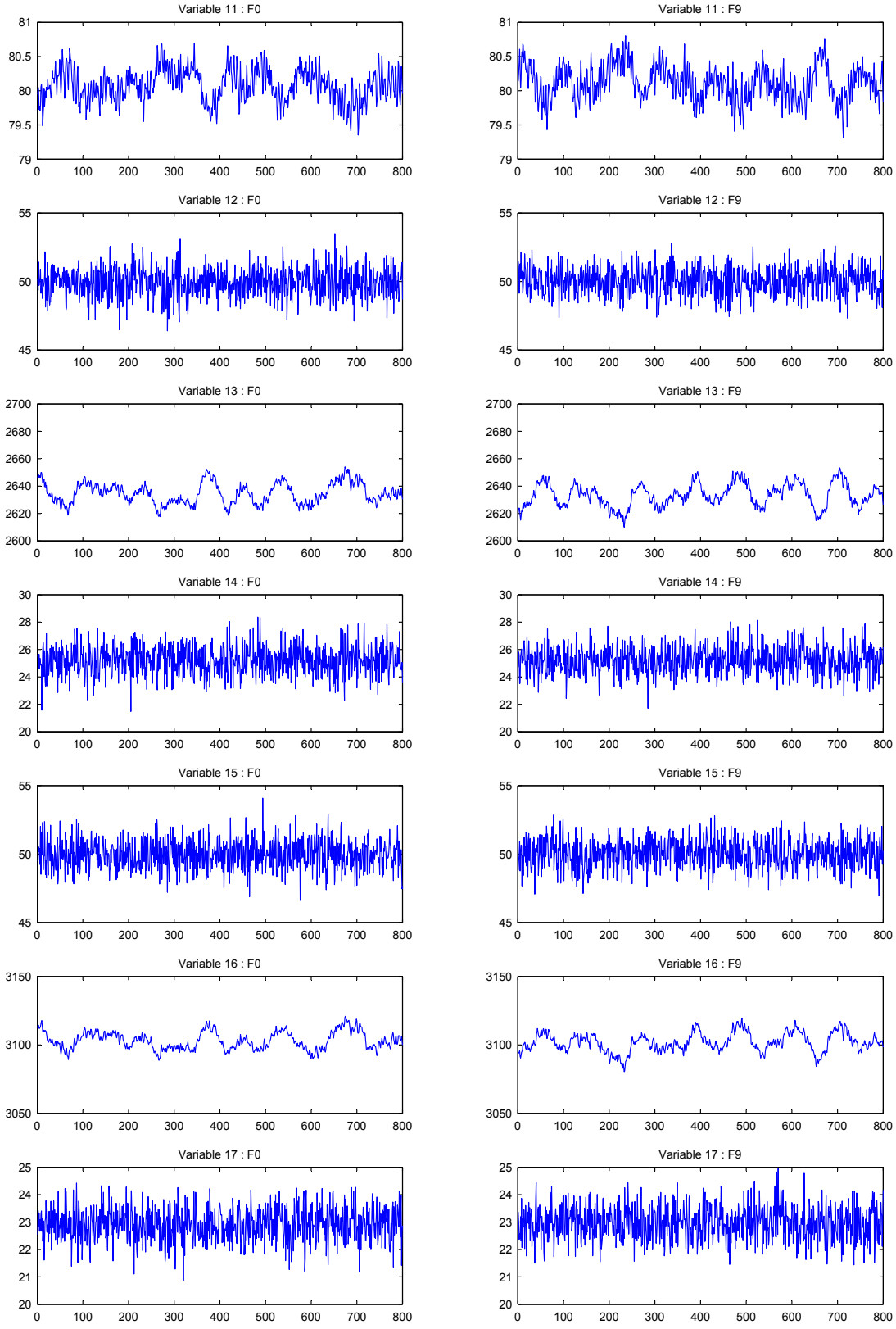


FIG. A.8 – Variables 11 à 17 en fonctionnement normal (F0) et pour la faute F9

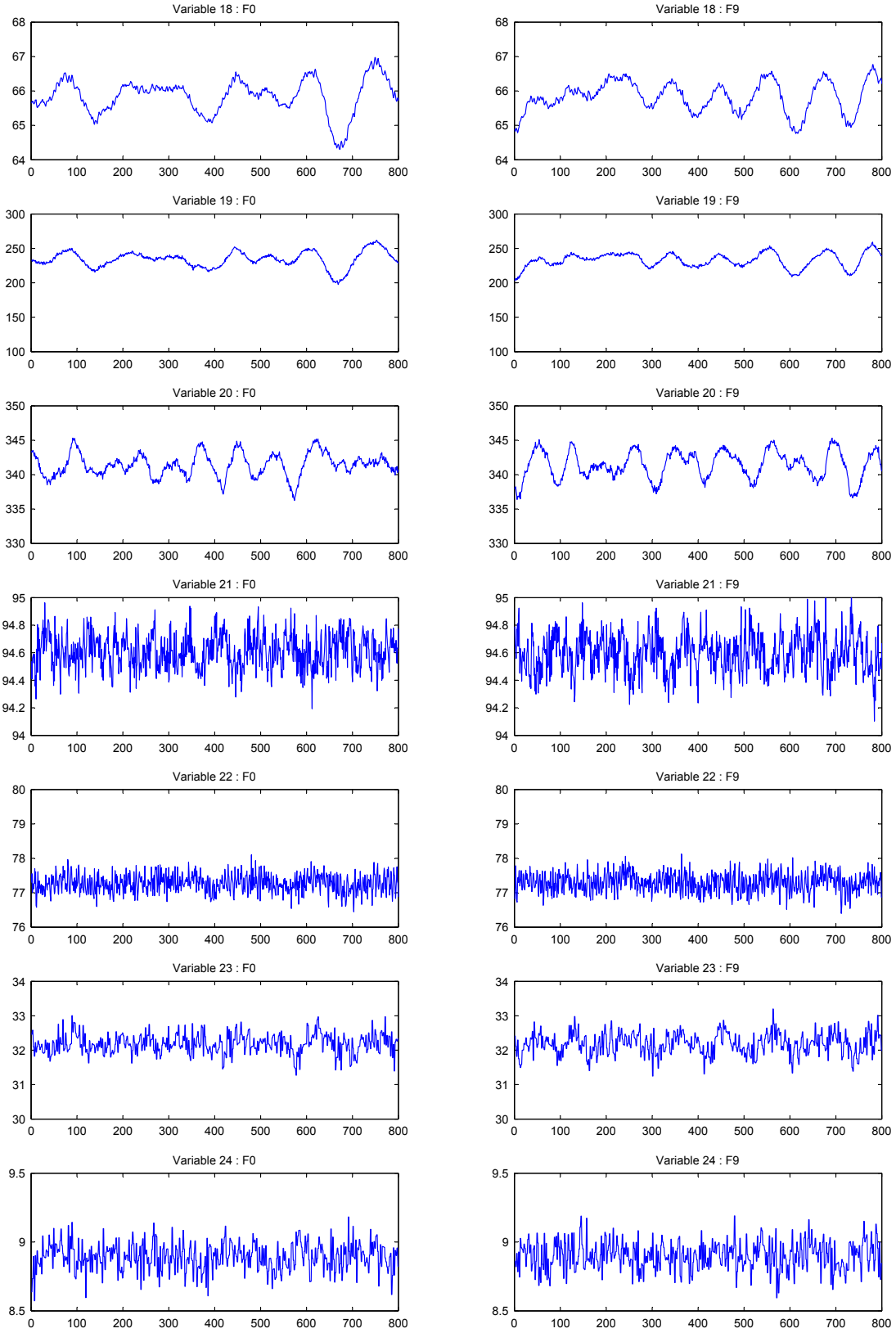


FIG. A.9 – Variables 18 à 24 en fonctionnement normal (F0) et pour la faute F9

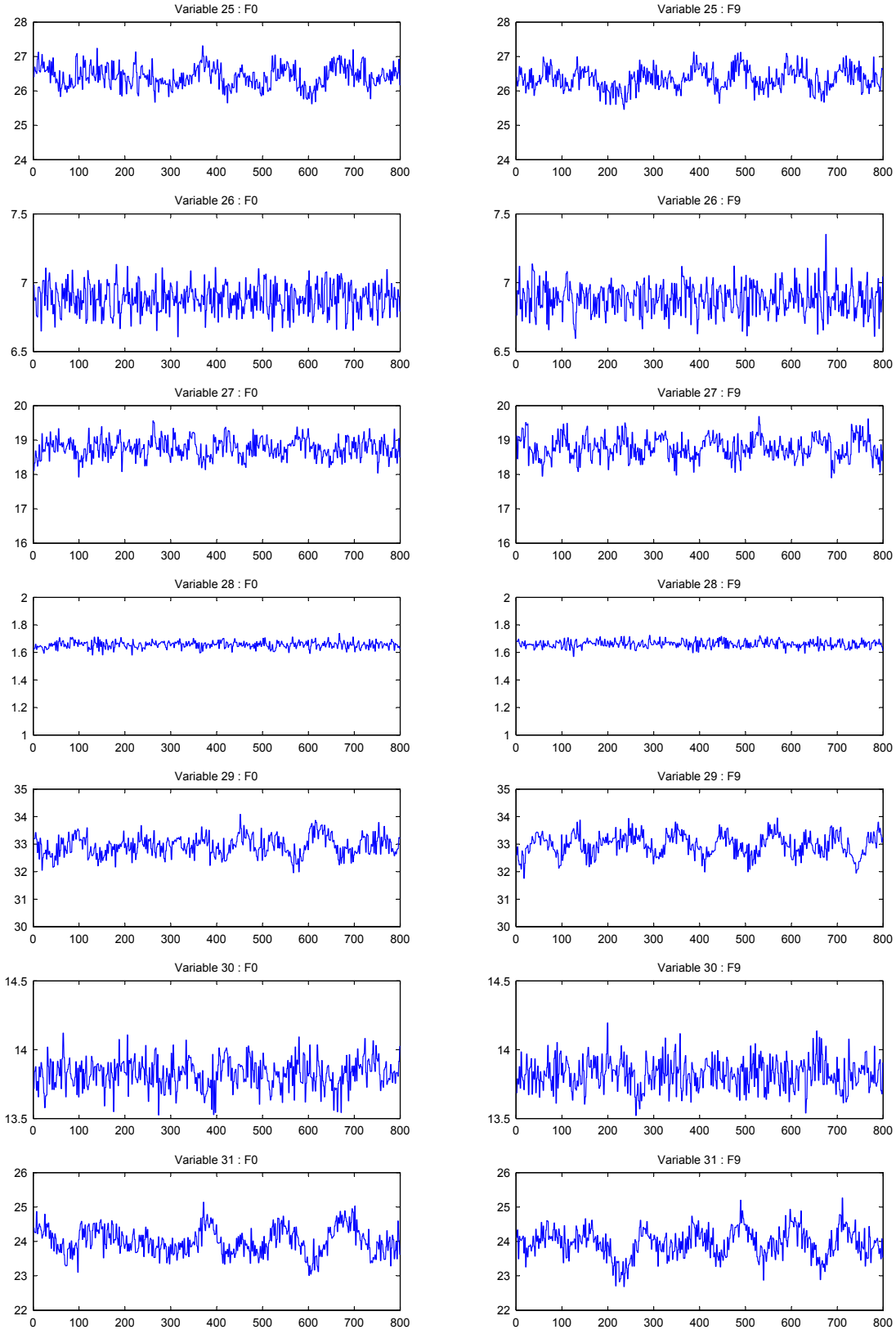


FIG. A.10 – Variables 25 à 31 en fonctionnement normal (F0) et pour la faute F9

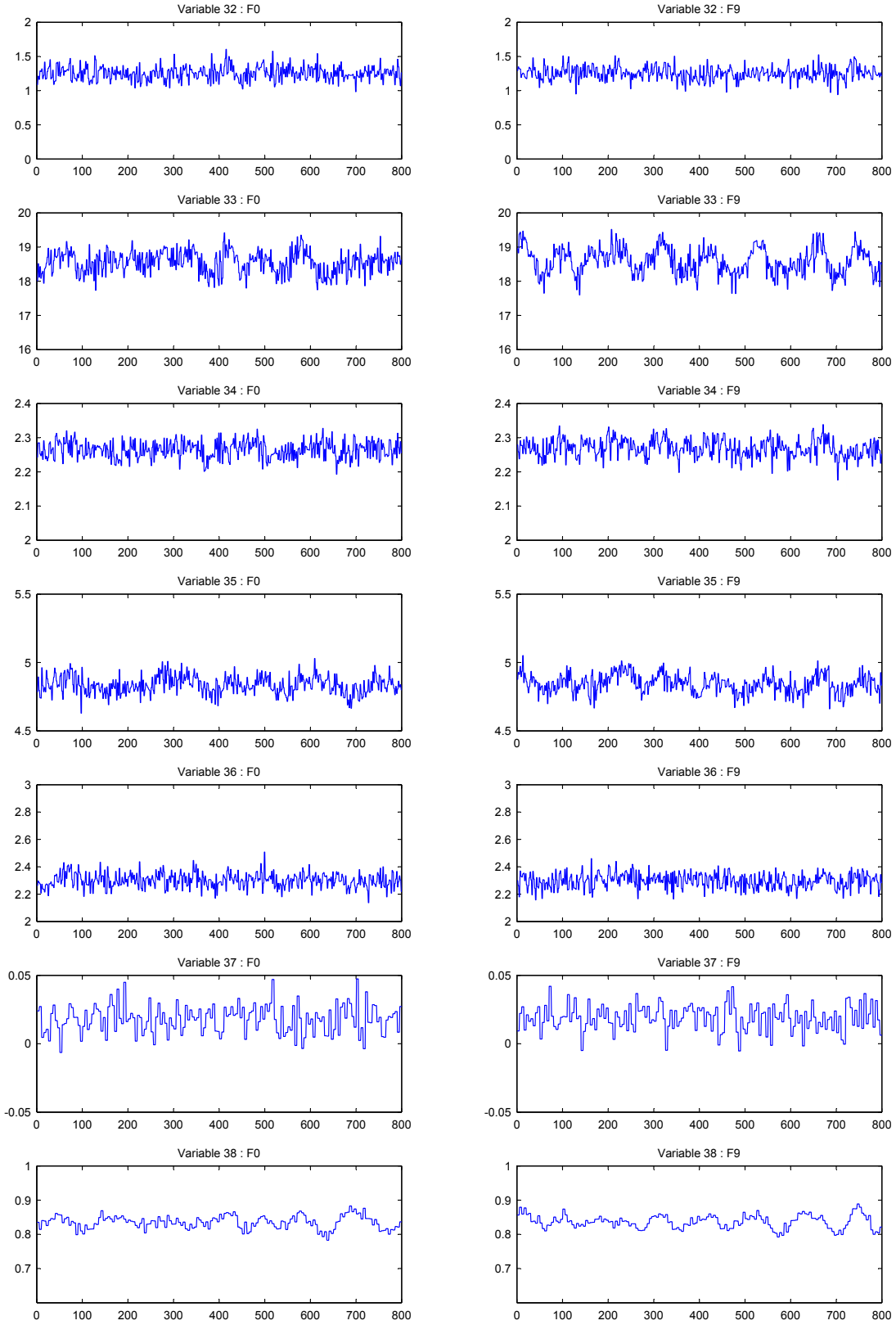


FIG. A.11 – Variables 32 à 38 en fonctionnement normal (F0) et pour la faute F9

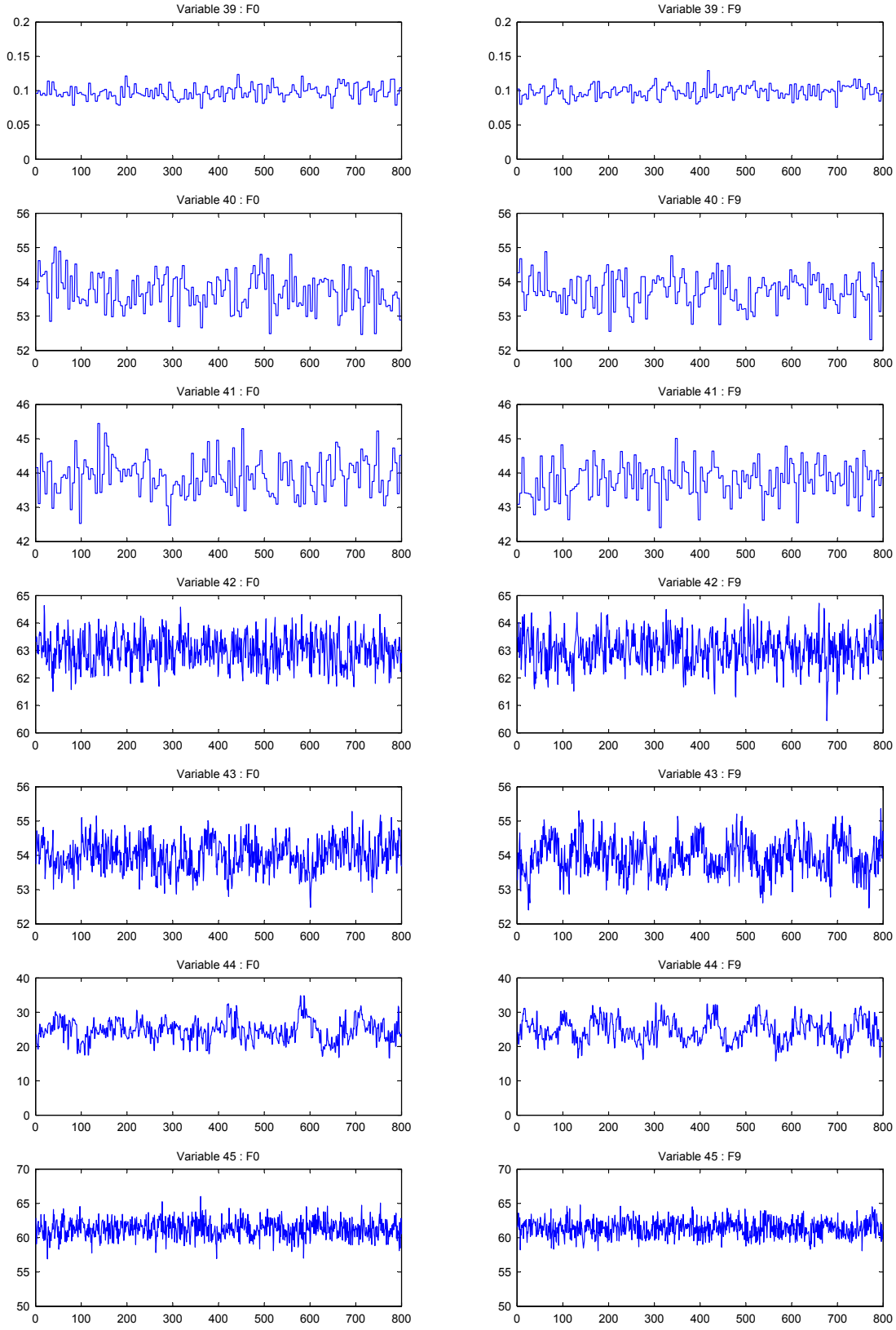


FIG. A.12 – Variables 39 à 45 en fonctionnement normal (F0) et pour la faute F9

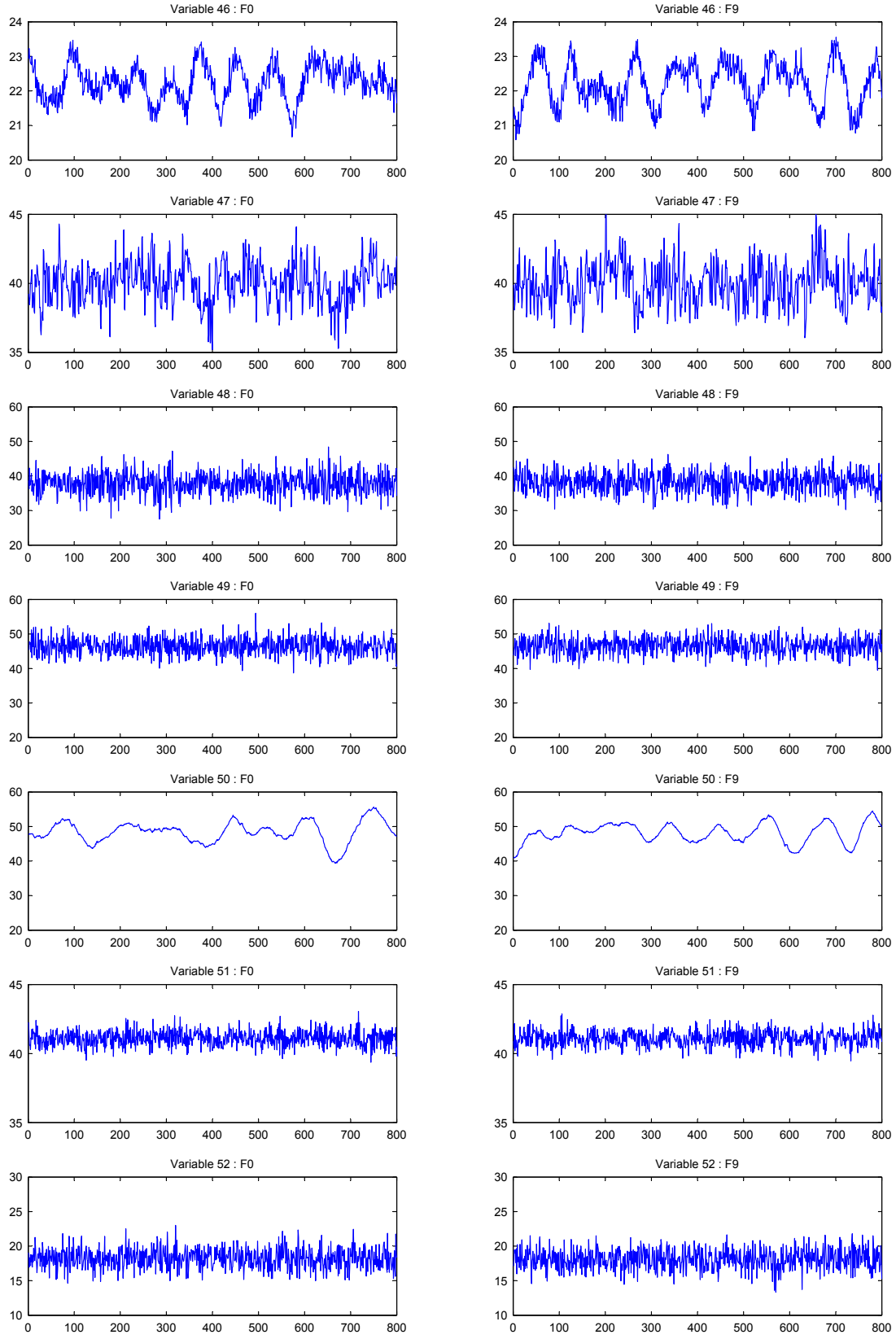


FIG. A.13 – Variables 46 à 52 en fonctionnement normal (F0) et pour la faute F9

A.4 Matrices de confusion sur les 15 premières fautes

Ces résultats concernent la classification des observations de test des fautes F1 à F15, sur le classifieur n'intégrant ni de sélection de composantes, ni de rejet de distance. La table A.2 donne la matrice des occurrences. Les tables A.3 et A.4 donnent respectivement les matrices de précision et de fiabilité.

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**
F1	780	0	0	0	0	0	0	2	0	0	0	0	0	0	0
F2	0	785	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	200	0	0	0	0	2	218	9	19	0	17	0	126
F4	0	0	0	659	0	0	0	0	0	0	28	0	0	0	0
F5	0	0	0	0	784	0	0	0	0	0	0	3	0	0	0
F6	0	0	0	0	0	800	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	800	0	0	0	0	0	0	0	0
F8	18	5	1	0	0	0	0	776	0	4	0	1	111	0	1
F9	0	8	189	0	0	0	0	11	200	30	29	1	4	0	266
F10	0	1	76	0	0	0	0	0	54	726	11	0	0	0	90
F11	0	0	58	141	0	0	0	3	55	6	646	0	2	1	56
F12	0	0	0	0	16	0	0	4	0	6	0	794	42	0	7
F13	0	0	0	0	0	0	0	2	0	3	0	1	611	0	3
F14	1	0	29	0	0	0	0	0	14	3	30	0	0	799	28
F15	1	1	247	0	0	0	0	0	259	13	37	0	13	0	223

Tab. A.2 – Matrice d'occurrences

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**
F1	97.5	0	0	0	0	0	0	0.25	0	0	0	0	0	0	0
F2	0	98.13	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	25	0	0	0	0	0.25	27.25	1.13	2.38	0	2.13	0	15.75
F4	0	0	0	82.38	0	0	0	0	0	0	3.5	0	0	0	0
F5	0	0	0	0	98	0	0	0	0	0	0	0.38	0	0	0
F6	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
F8	2.25	0.63	0.13	0	0	0	0	97	0	0.5	0	0.13	13.88	0	0.13
F9	0	1	23.63	0	0	0	0	1.38	25	3.75	3.63	0.13	0.5	0	33.25
F10	0	0.13	9.5	0	0	0	0	0	6.75	90.75	1.38	0	0	0	11.25
F11	0	0	7.25	17.63	0	0	0	0.38	6.88	0.75	80.75	0	0.25	0.13	7
F12	0	0	0	0	2	0	0	0.5	0	0.75	0	99.25	5.25	0	0.88
F13	0	0	0	0	0	0	0	0.25	0	0.38	0	0.13	76.38	0	0.38
F14	0.13	0	3.63	0	0	0	0	0	1.75	0.38	3.75	0	0	99.88	3.5
F15	0.13	0.13	30.88	0	0	0	0	0	32.38	1.63	4.63	0	1.63	0	27.88

TAB. A.3 – Matrice de précision exprimée en %

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**
F1	99.74	0	0	0	0	0	0	0.22	0	0	0	0	0	0	0
F2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	33.84	0	0	0	0	0.22	29.54	0.94	1.96	0	2.74	0	15.87
F4	0	0	0	95.92	0	0	0	0	0	0	2.89	0	0	0	0
F5	0	0	0	0	99.62	0	0	0	0	0	0	0.35	0	0	0
F6	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
F8	2.3	0.64	0.17	0	0	0	0	84.62	0	0.42	0	0.12	17.9	0	0.13
F9	0	1.02	31.98	0	0	0	0	1.2	27.1	3.13	3	0.12	0.65	0	33.5
F10	0	0.13	12.86	0	0	0	0	0	7.32	75.78	1.14	0	0	0	11.34
F11	0	0	9.81	20.52	0	0	0	0.33	7.45	0.63	66.74	0	0.32	0.11	7.05
F12	0	0	0	0	2.03	0	0	0.44	0	0.63	0	91.37	6.77	0	0.88
F13	0	0	0	0	0	0	0	0.22	0	0.31	0	0.12	98.55	0	0.38
F14	0.13	0	4.91	0	0	0	0	0	1.9	0.31	3.1	0	0	88.38	3.53
F15	0.13	0.13	41.79	0	0	0	0	0	35.09	1.36	3.82	0	2.1	0	28.09

Tab. A.4 – Matrice de fiabilité exprimée en %

A.5 Matrices de confusion sur les 15 premières fautes avec sélection des variables importantes

Ces résultats concernent la classification des observations de test des fautes F1 à F15, sur le classifieur n'intégrant pas de rejet de distance, mais intégrant une sélection des variables importante pour la classification. L'algorithme proposé à la section 3.3.1 a permis de sélectionner les 27 variables suivantes (en ordre de la plus informative vers la moins informative) : 51, 10, 1, 44, 9, 16, 46, 13, 50, 45, 21, 18, 34, 19, 11, 38, 20, 47, 31, 42, 4, 30, 43, 5, 35, 52 et 17. La table A.5 donne la matrice des occurrences. Les tables A.6 et A.7 donnent respectivement les matrices de précision et de fiabilité.

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**
F1	785	0	0	0	0	0	0	2	0	0	0	0	0	0	0
F2	0	786	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	230	0	0	0	0	2	241	16	13	0	20	0	173
F4	0	0	0	735	0	0	0	0	0	0	31	0	0	0	0
F5	0	0	0	0	787	0	0	0	0	0	0	4	0	0	0
F6	0	0	0	0	0	800	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	800	0	0	0	0	0	0	0	0
F8	13	3	0	0	0	0	0	776	0	3	0	2	133	0	1
F9	0	8	190	0	0	0	0	13	199	25	30	2	7	0	261
F10	0	0	68	0	0	0	0	0	18	723	10	0	2	0	88
F11	0	0	21	64	1	0	0	2	20	5	655	0	3	1	34
F12	0	0	0	0	12	0	0	2	0	7	0	790	49	0	2
F13	0	0	0	0	0	0	0	3	0	2	0	2	576	0	0
F14	1	0	11	1	0	0	0	0	4	2	22	0	0	799	6
F15	1	3	280	0	0	0	0	0	318	17	39	0	10	0	235

Tab. A.5 – Matrice d'occurrences

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**
F1	98.13	0	0	0	0	0	0	0.25	0	0	0	0	0	0	0
F2	0	98.25	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	28.75	0	0	0	0	0.25	30.13	2	1.63	0	2.5	0	21.63
F4	0	0	0	91.88	0	0	0	0	0	0	3.88	0	0	0	0
F5	0	0	0	0	98.38	0	0	0	0	0	0	0.5	0	0	0
F6	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
F8	1.63	0.38	0	0	0	0	0	97	0	0.38	0	0.25	16.63	0	0.13
F9	0	1	23.75	0	0	0	0	1.63	24.88	3.13	3.75	0.25	0.88	0	32.63
F10	0	0	8.5	0	0	0	0	0	2.25	90.38	1.25	0	0.25	0	11
F11	0	0	2.63	8	0.13	0	0	0.25	2.5	0.63	81.88	0	0.38	0.13	4.25
F12	0	0	0	0	1.5	0	0	0.25	0	0.88	0	98.75	6.13	0	0.25
F13	0	0	0	0	0	0	0	0.38	0	0.25	0	0.25	72	0	0
F14	0.13	0	1.38	0.13	0	0	0	0	0.5	0.25	2.75	0	0	99.88	0.75
F15	0.13	0.38	35	0	0	0	0	0	39.75	2.13	4.88	0	1.25	0	29.38

TAB. A.6 – Matrice de précision exprimée en %

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**
F1	99.75	0	0	0	0	0	0	0.21	0	0	0	0	0	0	0
F2	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
F3	0	0	33.09	0	0	0	0	0.21	32.79	1.76	1.61	0	3.43	0	19.16
F4	0	0	0	95.95	0	0	0	0	0	0	3.85	0	0	0	0
F5	0	0	0	0	99.49	0	0	0	0	0	0	0.46	0	0	0
F6	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
F8	1.65	0.38	0	0	0	0	0	83.35	0	0.33	0	0.23	22.81	0	0.11
F9	0	1.02	27.34	0	0	0	0	1.4	27.07	2.75	3.72	0.23	1.2	0	28.9
F10	0	0	9.78	0	0	0	0	0	2.45	79.54	1.24	0	0.34	0	9.75
F11	0	0	3.02	8.36	0.13	0	0	0.21	2.72	0.55	81.27	0	0.51	0.12	3.77
F12	0	0	0	0	1.52	0	0	0.21	0	0.77	0	91.65	8.4	0	0.22
F13	0	0	0	0	0	0	0	0.32	0	0.22	0	0.23	98.8	0	0
F14	0.13	0	1.58	0.13	0	0	0	0	0.54	0.22	2.73	0	0	94.44	0.66
F15	0.13	0.38	40.29	0	0	0	0	0	43.27	1.87	4.84	0	1.72	0	26.02

Tab. A.7 – Matrice de fiabilité exprimée en %

A.6 Variables du TEP pour la faute F6

Les figures A.14 à A.18 représentent les 52 variables du TEP lors de la faute F6. Les traits continus représentent les données d'apprentissage, alors que les traits pointillés représentent les 480 premières observations des données de test.

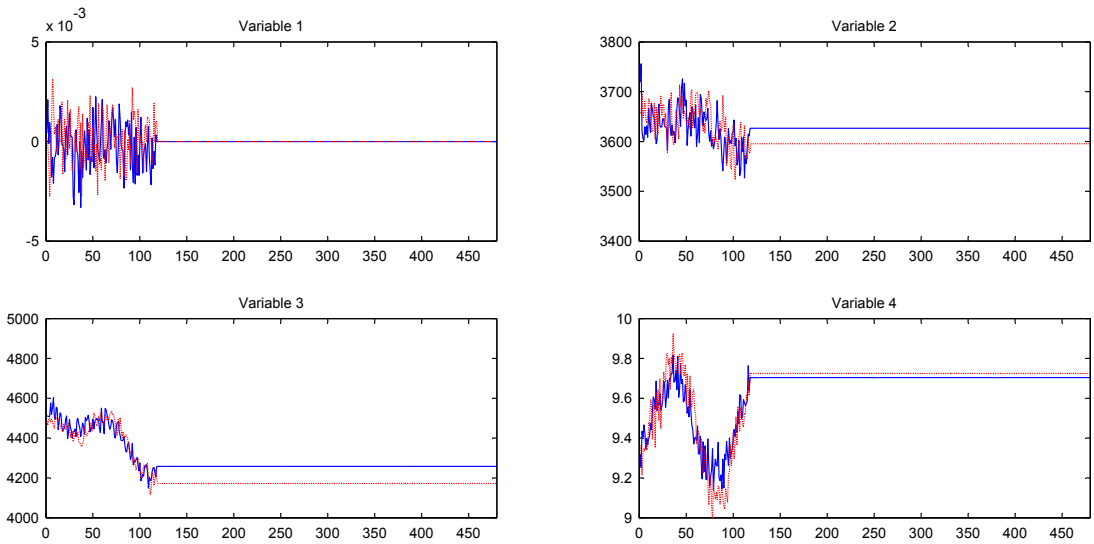


FIG. A.14 – Variables 1 à 4 pour la faute F6

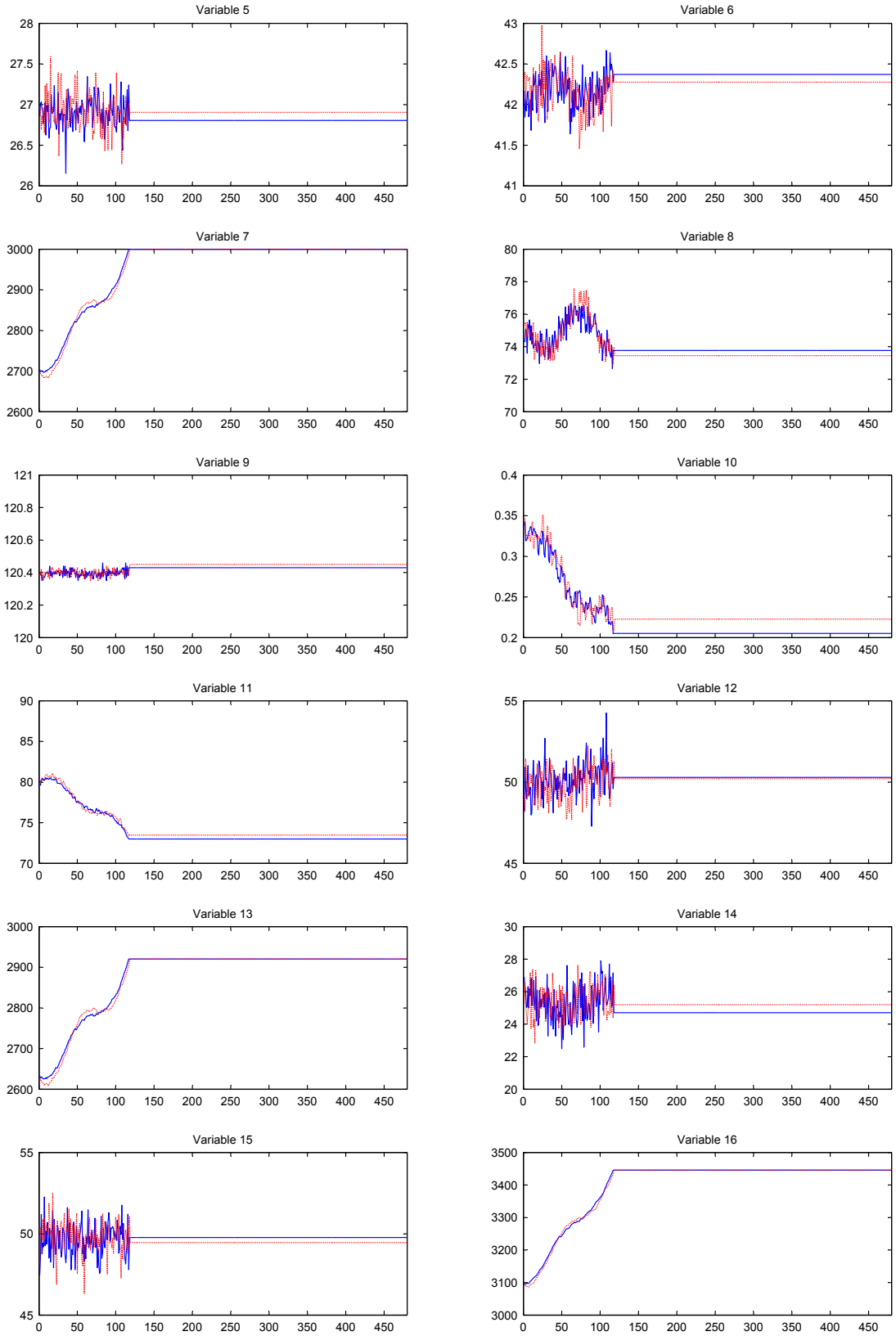


FIG. A.15 – Variables 5 à 16 pour la faute F6

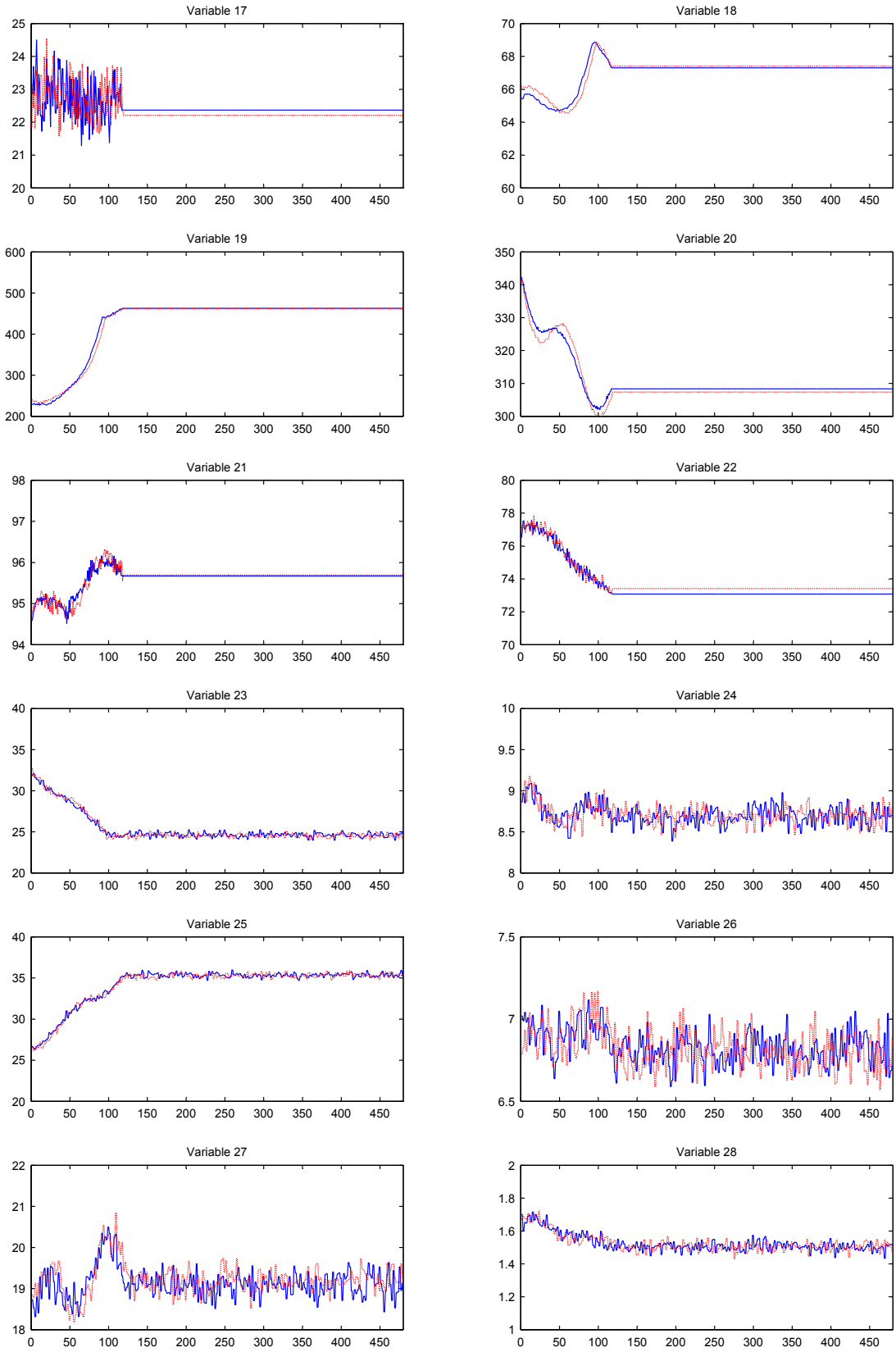


FIG. A.16 – Variables 17 à 28 pour la faute F6

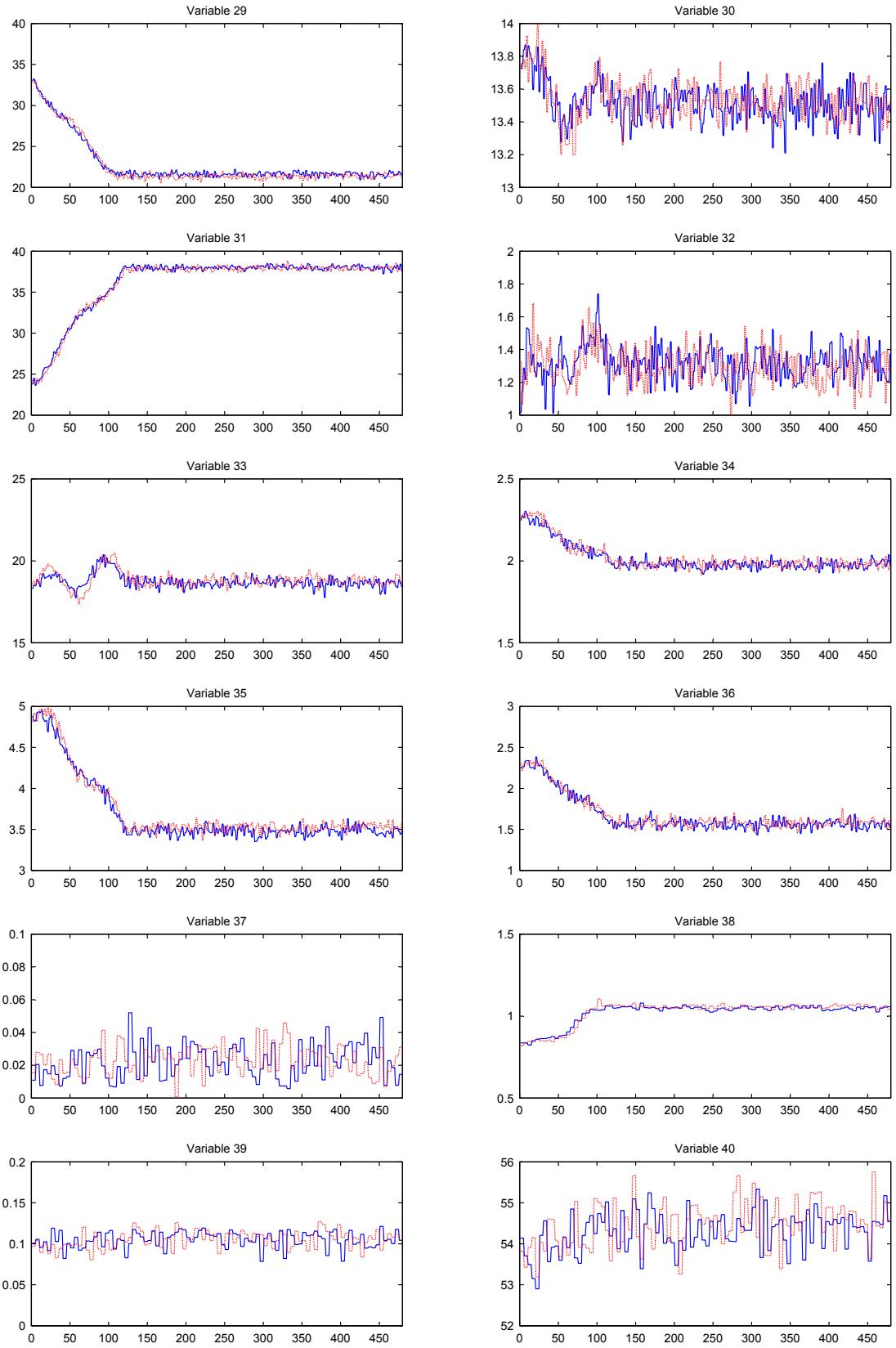


FIG. A.17 – Variables 29 à 40 pour la faute F6

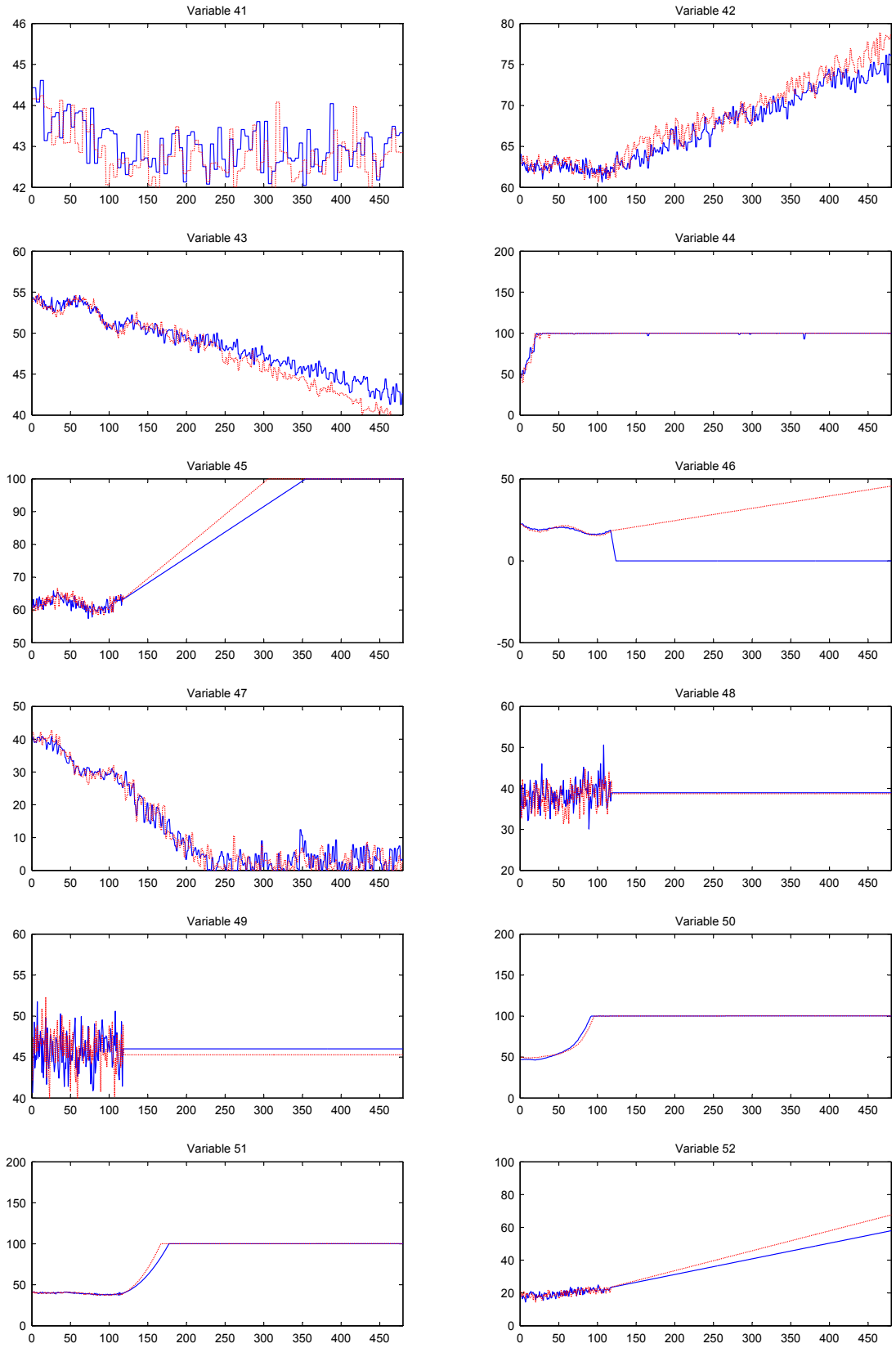


FIG. A.18 – Variables 41 à 52 pour la faute F6

Bibliographie

[1] H.B. Aradhye. Sensor fault detection, isolation, and accommodation using neural networks, fuzzy logic, and bayesian belief networks. Mémoire de Master, University of New Mexico, Albuquerque NM, 1997.

[2] Kunihiro Baba, Ritei Shibata, et Masaaki Sibuya. Partial correlation and conditionnal correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4) :657–664, 2004.

[3] C. G. Bai, Q. P. Hu, M. Xie, et S. H. Ng. Software failure prediction based on a markov bayesian network model. *Journal of Systems and Software*, 74(3) :275–282, February 2005.

[4] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5 :537–550, 1994.

[5] Bayesia. Logiciel bayesialab - http ://www.bayesia.com/.

[6] David Bellot. *Fusion de données avec des réseaux bayésiens pour la modélisation des systèmes dynamiques et son application en télémédecine*. Thèse de Doctorat, Nancy, 2002.

[7] B.V. Bonnlander et A.S. Weigend. Selecting input variables using mutual information and nonparamteric density estimation. Dans *Proceedings of the 1994 International Symposium on Artificial Neural Networks*, pages 42–50, Tainan, Taiwan, 1994.

[8] Bernhard E. Boser, Isabelle M. Guyon, et Vladimir N. Vapnik. Training algorithm for optimal margin classifiers. Dans *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Univ of California, Berkeley, United States, 1992.

[9] Hichem Boudali et Joanne B. Dugan. A temporal bayesian network reliability framework. Dans *Mathematical Method in Reliability*, 2004.

[10] N. Boudaoud et Z. Cherfi. Maîtrise statistique des processus multivariés : Avantages

et limites des différentes approches sur les cartes de contrôle multivariées. *Journal Européen des Systèmes Automatisés*, 34(2-3) :379–390, 2000.

[11] Leo Breiman, Jerome Friedman, Charles J. Stone, et R.A. Olshen. *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, 1993.

[12] R.B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1 :245–276, 1966.

[13] Eugene Charniak. Bayesian networks without tears. *AI Magazine*, 12(4) :50–63, 1991.

[14] Jean-Noël Chatain. *Diagnostic par système expert*. Traité des nouvelles technologies. Série Diagnostic et maintenance. Hermes Sciences Publications, 1993.

[15] Gang Chen et Thomas J. McAvoy. Predictive on-line monitoring of continuous processes. *Journal of Process Control*, 8(5-6) :409–420, 1998.

[16] Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, et Weiru Liu. Learning Bayesian networks from data : An information-theory based approach. *Artificial Intelligence*, 137(1–2) :43–90, 2002.

[17] Leo H. Chiang, Evan L. Russell, et Richard D. Braatz. *Fault detection and diagnosis in industrial systems*. New York : Springer-Verlag, 2001.

[18] L.H. Chiang, M.E. Kotanchek, et A.K. Kordon. Fault diagnosis based on fisher discriminant analysis and support vector machines. *Computers and Chemical Engineering*, 28(8) :1389–1401, 2004.

[19] C. Chow et C. Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3) :462–467, 1968.

[20] Meng Koon Chua et Douglas C. Montgomery. Investigation and characterization of a control scheme for multivariate quality control. *Quality and Reliability Engineering International*, 8 :37–44, 1992.

[21] B.R. Cobb, R. Rumi, et A. Salmeron. Modeling conditional distributions of continuous variables in bayesian networks. Dans *Lecture Notes in Computer Science*, volume 3646 LNCS, pages 36–45, 2005.

[22] B. Conrard, J.-M. Thiriet, et M. Robert. Distributed system design based on dependability evaluation : A case study on a pilot thermal process. *Reliability Engineering and System Safety*, 88(1) :109–119, 2005.

[23] Antoine Cornuéjols, Laurent Miclet, et Yves Kodratoff. *Apprentissage artificiel : concepts et algorithmes*. Eyrolles, 2002.

[24] Franck Corset. *Aide à l'optimisation de maintenance à partir de réseaux bayésiens et fiabilité dans un contexte doublement censuré.* Thèse de Doctorat, Université Joseph Fourier - Grenoble I, 2003.

[25] T. M. Cover. *Learning in pattern recognition.* Methodologies of Pattern Recognition, NY, s. watanabe (ed.) edition, 1969.

[26] Thomas M. Cover et Joy A. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.

[27] T.M. Cover et P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13 :21–27, 1967.

[28] R.B. Crosier. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30(3) :291–303, 1988.

[29] Y. Le Cun. Learning scheme for asymmetric threshold networks. *Cognitiva 85, Paris, France*, page 599, 1985.

[30] Belur V. Dasarathy. *Nearest Neighbor : Pattern Classification Techniques.* Ieee Computer Society, 1991.

[31] A. P. Dempster, N. M. Laird, et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39 :1–38, 1977.

[32] T. Denoeux, M. Masson, et B. Dubuisson. Advanced pattern recognition techniques for system monitoring and diagnosis : A survey. *Journal Europeen des Systemes Automatises*, 31(9-10) :1509–1539, 1997.

[33] Murat Deviren et Khalid Daoudi. Apprentissage de structures de réseaux bayésiens dynamiques pour la reconnaissance de la parole. Dans *XXIVèmes Journées d'étude sur la parole*, 2002.

[34] S. Dey et J.A. Stori. A bayesian network approach to root cause diagnosis of process variations. *International Journal of Machine Tools and Manufacture*, 45(1) :75–91, 2005.

[35] Necip Doganaksoy, Frederick .W. Faltin, et William T. Tucker. Identification of out of control quality characteristics in a multivariate manufacturing environment. *Communications in Statistics - Theory and Methods*, 20(9) :2775–2790, 1991.

[36] A. Doig et A. H. Land. An automatic method for solving discrete programming problems. *Econometrica*, 28 :497, 1960.

[37] Pedro Domingos et Michael J. Pazzani. Beyond independence : Conditions for the optimality of the simple bayesian classifier. Dans *International Conference on Machine Learning*, 1996.

[38] D. Dong et T.J. Mcavoy. Nonlinear principal component analysis - based on principal curves and neural networks. *Computers and Chemical Engineering*, 20(1) :65–78, 1996.

[39] J. Dougherty, R. Kohavi., et M. Sahami. Supervised and unsupervised discretization of continuous features. Dans *Proceedings of the 12th International Conference on Machine Learning*, 1995.

[40] J.J. Downs et E.F. Vogel. Plant-wide industrial process control problem. *Computers and Chemical Engineering*, 17(3) :245–255, 1993.

[41] Gérard Dreyfus, Jean-Marc Martinez, Mannuel Samuelides, Mirta Gordon, Fouad Badran, Sylvie Thiria, et Laurent Hérault. *Réseaux de neurones : Méthodologie et applications*. Eyrolles, 2ème édition, 2004.

[42] Bernard Dubuisson. *Diagnostic et reconnaissance des formes*. Traité des nouvelles technologies. Série Diagnostic et maintenance. Hermès, 1990.

[43] Bernard Dubuisson. *Diagnostic, intelligence artificielle et reconnaissance des formes*. Traité IC2 information. Série productique. Hermès sciences publications, 2001.

[44] R. O. Duda, P. E. Hart, et D. G. Stork. *Pattern Classification 2nd edition*. Wiley, 2001.

[45] Tapio Elomaa et Juho Rousu. On decision boundaries of naive bayes in continuous domains. Dans *Lecture Notes in Artificial Intelligence*, 2003.

[46] Brigitte Escofier et Jérôme Pages. *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation, 3ème édition*. Dunod, 1998.

[47] Jean Faucher. *Pratique de l'AMDEC*. Dunod, 2004.

[48] P.M. Frank. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy. a survey and some new results. *Automatica*, 26(3) :459–474, 1990.

[49] J. H. Friedman. Regularized discriminant analysis. *J. Amer. Statist. Assoc*, 84(405) :165–175, 1989.

[50] N. Friedman, D. Geiger, et M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3) :131–163, 1997.

[51] P. Geladi et B. Kowalski. Partial least-squares regression : A tutorial. *Analytica Chimica Acta*, 185 :1–17, 1986.

[52] J. Gertler. Fault detection and isolation using parity relations. *Control Engineering Practice*, 5(5) :653–661, 1997.

[53] J. Gertler, W. Li, Y. Huang, et T. McAvoy. Isolation enhanced principal component analysis. *AIChE Journal*, 45(2) :323–334, 1999.

[54] Yaniv Gurwicz et Boaz Lerner. Rapid spline-based kernel density estimation for bayesian networks. Dans *ICPR '04 : Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3*, pages 700–703, Washington, DC, USA, 2004. IEEE Computer Society.

[55] T.J. Harris. Assessment of control loop performance. *Canadian Journal of Chemical Engineering*, 69 :48–57, 1991.

[56] Douglas M. Hawkins. Multivariate quality control based on regression-adjusted variables. *Technometrics*, 33(1) :61–75, 1991.

[57] Douglas M. Hawkins. Regression adjustment for variables in multivariate quality control. *Journal of Quality Technology*, 25(3) :170–182, 1993.

[58] Joseph P. Hoffbeck et David A. Landgrebe. Covariance matrix estimation and classification with limited training data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7) :763–767, 1996.

[59] A. Hoskuldsson. Pls regression methods. *Journal of Chemometrics*, 2 :211–228, 1988.

[60] Harold Hotelling. Multivariate quality control. *Techniques of Statistical Analysis*, :111–184, 1947.

[61] C.-N. Hsu, H.-J. Huang, et T.-T. Wong. Implications of the dirichlet assumption for discretization of continuous variables in naive bayesian classifiers. *Machine Learning*, 53(3) :235–263, 2003.

[62] R. Isermann. Fault diagnosis of machines via parameter estimation and knowledge processing - tutorial paper. *Automatica*, 29(4) :815–835, 1993.

[63] J. E. Jackson. Quality control for several related variables. *Technometrics*, 1 :359–377, 1959.

[64] J. Edward Jackson. *A User's Guide to Principal Components*. Wiley, 2003.

[65] J.Edward Jackson et Govind S. Mudholkar. Control procedures for residuals associated with principal component analysis. *Technometrics*, 21(3) :341–349, 1979.

[66] Luan Jaupi. *Contrôle de la qualité : MSP, analyse des performances et contrôle de réception*. Dunod, 2002.

[67] S. Jong. Simpls : An alternative pproach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18 :251–263, 1993.

[68] Stéphanie Jonquières. *Application des réseaux Bayésiens à la reconnaissance active d'objets 3D : contribution à la saisie d'objets*. Thèse de Doctorat, INP de Toulouse, 2000.

[69] M. I Jordan. *Learning in Graphical Models*. Kluwer Academic Publishers, Dordecht, The Netherlands, 1998.

[70] M. Kalisch et P. Buhlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8 :613–636, 2007.

[71] M. Kano, K. Nagao, S. Hasebe, I. Hashimoto, H. Ohno, R. Strauss, et B.R. Bakshi. Comparison of multivariate statistical process monitoring methods with applications to the eastman challenge problem. *Computers and Chemical Engineering*, 26(2) :161–174, 2002.

[72] E. Keogh et M. Pazzani. Learning augmented bayesian classifiers : A comparison of distribution-based and classification-based approaches. Dans *Proceedings of the Seventh International Workshop on Artificial Intelligence*, pages 225–230, 1999.

[73] D. Koller et A. Pfeffer. Object-oriented bayesian networks. Dans *Proceedings of the UAI-97*, pages 302–313, 1997.

[74] Igor Kononenko. Semi-naive bayesian classifier. Dans *EWSL-91 : Proceedings of the European working session on learning on Machine learning*, pages 206–219, 1991.

[75] T. Kourti et J.F. MacGregor. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28(1) :3–21, 1995.

[76] Theodora Kourti et John F. MacGregor. Multivariate spc methods for process and product monitoring. *Journal of Quality Technology*, 28(4) :409–428, 1996.

[77] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2) :233–243, 1991.

[78] W. Ku, R.H. Storer, et C. Georgakis. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 30(1) :179–196, 1995.

[79] Vincent Labatut. *Réseaux causaux probabilistes à grande échelle : un nouveau formalisme pour la modélisation du traitement de l'information cérébrale*. Thèse de Doctorat, Université Paul Sabatier de Toulouse, 2003.

[80] K.J. Laidler. *The World of Physical Chemistry*. Oxford University Press, 1993.

[81] Pat Langley, Wayne Iba, et Kevin Thompson. An analysis of bayesian classifiers. Dans *National Conference on Artificial Intelligence*, 1992.

[82] Pat Langley et Stephanie Sage. Induction of selective bayesian classifiers. Dans *In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 1994.

[83] Ludovic Lebart, Alain Morineau, et Marie Piron. *Statistique exploratoire multidimensionnelle*. DUNOD, 2000.

[84] Daniel Lepadatu. *Optimisation des procédés de mise en forme par approche plan d'expériences*. Thèse de Doctorat, Université d'Angers, 2006.

[85] P. Leray et P. Gallinari. Feature selection with neural networks. *Behaviormetrika (special issue on Analysis of Knowledge Representation in Neural Network Models)*, 26(1) :145–166, 1999.

[86] P. Leray et P. Gallinari. De l'utilisation d'obd pour la sélection de variables dans les perceptrons multicouches. *Revue d'Intelligence Artificielle*, 15(3-4) :373–391, 2001.

[87] Philippe Leray. Réseaux bayésiens : apprentissage et modélisation de systèmes complexes. Dans *Soutenance Habilitation à Diriger les Recherches*, 2006.

[88] Jing Li, Jionghua Jin, et Jianjun Shi1. Causation-based t2 decomposition for multivariate process monitoring and diagnosis. Dans *Industrial Engineering Research Conference*, 2006.

[89] Jing Li et Jianjun Shi. Knowledge discovery from observational data for process control using causal bayesian networks. *IIE Transactions*, 39 :681–690, 2007.

[90] Nikolaos Limnios. *Arbres de défaillances - 2ème édition*. Lavoisier, 2005.

[91] J.S. Liu, F. Liang, et W.H. Wong. The use of multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95 :121–134, 2000.

[92] Cynthia A. Lowry et Douglas C. Montgomery. Review of multivariate control charts. *IIE Transactions (Institute of Industrial Engineers)*, 27(6) :800–810, 1995.

[93] Cynthia A. Lowry, William H. Woodall, Charles W. Champ, et Steven E. Rigdon. A multivariate exponentially weighted moving average control chart. *Technometrics*, 34(1) :46–53, 1992.

[94] P.R. Lyman et C. Georgakis. Plant-wide control of the tennessee eastman problem. *Computers and Chemical Engineering*, 19(3) :321–331, 1995.

[95] J.F. MacGregor et T. Kourti. Statistical process control of multivariate processes. *Control Engineering Practice*, 3(3) :403–414, 1995.

[96] Michael G. Madden. The performance of bayesian network classifiers constructed using different techniques. Dans *Proceedings of European Conference on Machine Learning, Workshop on Probabilistic Graphical Models for Classification*, September 2003.

[97] R.L. Mason, N.D. Tracy, et J.C. Young. A practical approach for interpreting multivariate t^2 control chart signals. *Journal of Quality Technology*, 29(4) :396–406, 1997.

[98] Robert L. Mason, Nola D. Tracy, et John C. Young. Decomposition of T^2 for multivariate control chart interpretation. *Journal of Quality Technology*, 27(2) :99–108, 1995.

[99] T.J. McAvoy et N. Ye. Base control for the tennessee eastman problem. *Computers and Chemical Engineering*, 18(5) :383–413, 1994.

[100] G. McLachlan et K. Basford. *Mixture Models : Inference and Applications to Clustering*. Marcel Dekker, 1988.

[101] N. Mehranbod, M. Soroush, et C. Panjapornpon. A method of sensor fault detection and identification. *Journal of Process Control*, 15(3) :321–339, 2005.

[102] N. Mehranbod, M. Soroush, M. Piovoso, et B.A. Ogunnaike. Probabilistic model for sensor fault detection and identification. *AIChE Journal*, 49(7) :1787–1802, 2003.

[103] W.E. Molnau, D.C. Montgomery, et G.C. Runger. Statistically constrained economic design of the multivariate exponentially weighted moving average control chart. *Quality and Reliability Engineering International*, 17(1) :39–49, 2001.

[104] Douglas C. Montgomery. *Introduction to Statistical Quality Control, Third Edition*. John Wiley and Sons, 1997.

[105] Djamel Mouss, Hayet Mouss, et K. Mouss. Acquisition des connaissances et diagnostic des défaillances d'un procédé industriel. *Phoebus*, 35 :61–74, 2005.

[106] Kevin Patrick Murphy. Dynamic bayesian networks - http ://www.ai.mit.edu/ murphyk/Thesis/thesis.html.

[107] Kevin Patrick Murphy. *Dynamic Bayesian Networks : Representation, Inference and Learning*. Thèse de Doctorat, U.C. Berkeley, 2002.

[108] Patrick Naim, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, et Anna Becker. *Réseaux bayésiens - 2ème édition*. Eyrolles, 2004.

[109] T.D. Nielsen et F.V. Jensen. On-line alert systems for production plants : A conflict based approach. *International Journal of Approximate Reasoning*, 45(2) :255–270, 2007.

[110] Paul Nomikos et John F. MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, 40(8) :1361–1373, 1994.

[111] E. S. Page. Continuous inspection schemes. *Biometrika*, 41 :100–115, 1954.

[112] R.J. Patton et J. Chen. Observer-based fault detection and isolation : Robustness and applications. *Control Engineering Practice*, 5(5) :671–682, 1997.

[113] M. Pazzani. Searching for dependencies in bayesian classifiers. *Learning from Data Artificial Intelligence and Statistics*, 5 :239–248, 1997.

[114] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann Publishers, 1988.

[115] Judea Pearl et Tom S. Verma. A theory of inferred causation. Dans James F. Allen, Richard Fikes, et Erik Sandewall, editors, *KR'91 : Principles of Knowledge Representation and Reasoning*, pages 441–452, San Mateo, California, 1991. Morgan Kaufmann.

[116] Aritz Perez, Pedro Larranaga, et Inaki Inza. Supervised classification with conditional gaussian networks : Increasing the structure complexity from naive bayes. *International Journal of Approximate Reasoning*, 43 :1–25, 2006.

[117] F. Pernkopf. Bayesian network classifiers versus selective k-nn classifier. *Pattern Recognition*, 38(1) :1–10, 2005.

[118] M. Perzyk, R. Biernacki, et A. Kochanski. Modeling of manufacturing processes by learning systems : The naive bayesian classifier versus artificial neural networks. *Journal of Materials Processing Technology*, 164-165 :1430–1435, 2005.

[119] J.J. Pignatiello et G.C. Runger. Comparisons of multivariate cusum charts. *Journal of Quality Technology*, 22(3) :173–186, 1990.

[120] Maurice Pillet. *Appliquer la maîtrise statistique des procédés MSP/SPC 3ème édition*. Les Editions d'Organisation, 2001.

[121] S.S. Prabhu et G.C. Runger. Designing a multivariate ewma control chart. *Journal of Quality Technology*, 29(1) :8–15, 1997.

[122] N.L. Ricker. Decentralized control of the tennessee eastman challenge process. *Journal of Process Control*, 6(4) :205–221, 1996.

[123] S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3) :239–250, Août 1959.

[124] R. W. Robinson. Counting unlabeled acyclic digraphs. Dans C. H. C. Little, editor, *Combinatorial Mathematics V*, volume 622 of *Lecture Notes in Mathematics*, pages 28–43, Berlin, 1977. Springer.

[125] Carlos Rojas-Guzman et Mark A. Kramer. Comparison of belief networks and rule-based expert systems for fault diagnosis of chemical processes. *Engineering Applications of Artificial Intelligence*, 6(3) :191–202, June 1993.

[126] F. Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65 :386–407, 1958.

[127] Alain Ruegg. *Probabilités et statistique 3ème édition*. Presses Polytechniques Romandes, 1989.

[128] D.E. Rumelhart, G.E. Hinton, et R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088) :533–536, 1986.

[129] G.C. Runger et M.C. Testik. Multivariate extensions to cumulative sum control charts. *Quality and Reliability Engineering International*, 20(6) :587–606, 2004.

[130] Mehran Sahami. Learning limited dependence bayesian classifiers. Dans *Second International Conference on Knowledge Discovery in Databases*, 1996.

[131] M. Sanchez, G. Sentoni, S. Schbib, S. Tonelli, et J. Romagnoli. Gross measurements error detection/identification for an industrial ethylene reactor. *Computers and Chemical Engineering*, 20(SUPPL.2) :–, 1996.

[132] Bruno Scibilia. *Développement et améliorations des méthodes d'optimisation des procédés par les plans d'expériences*. Thèse de Doctorat, Université d'Angers, 2000.

[133] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27 :379–423, 623–656, 1948.

[134] Walter A. Shewhart. *Economic control of quality of manufactured product*. New York : D. Van Nostrand Co., 1931.

[135] Peter Spirtes, Clark Glymour, et Richard Scheines. *Causation, prediction, and search*. Springer-Verlag, 1993.

[136] Joe H. Sullivan et William H. Woodall. A comparison of multivariate control charts for individual observations. *Journal of Quality Technology*, 28(4) :398–408, October 1996.

[137] David M. J. Tax et Robert P. W. Duin. Combining one-class classifiers. *Lecture Notes in Computer Science*, 2096 :299–308, 2001.

[138] M.C. Testik et C.M. Borror. Design strategies for the multivariate exponentially weighted moving average control chart. *Quality and Reliability Engineering International*, 20(6) :571–577, 2004.

[139] M.C. Testik, G.C. Runger, et C.M. Borror. Robustness properties of multivariate ewma control charts. *Quality and Reliability Engineering International*, 19(1) :31–38, 2003.

[140] C.E. Thomaz, D.F. Gillies, et R.Q. Feitosa. A new covariance estimate for bayesian classifiers in biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(2) :214–223, 2004.

[141] Teodor Tiplica. *Contribution à la Maîtrise Statistique des Processus Industriels Multivariés*. Thèse de Doctorat, ISTIA, 2002.

[142] Teodor Tiplica, Abdessamad Kobi, et Alain Barreau. Utilisation de méthode t2 hotelling dans le contrôle statistique des processus multivariés. Dans *CCF*, 2000.

[143] Teodor Tiplica, Abdessamad Kobi, et Alain Barreau. Optimisation et maîtrise des processus multivariés. la méthode fnad. *Journal Européen des Systèmes Automatisés*, 37(4) :477–500, 2003.

[144] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[145] S. Vempala et G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4) :841–860, 2004.

[146] V. Venkatasubramanian, R. Rengaswamy, et S.N. Kavuri. A review of process fault detection and diagnosis part ii : Qualitative models and search strategies. *Computers and Chemical Engineering*, 27(3) :313–326, 2003.

[147] V. Venkatasubramanian, R. Rengaswamy, S.N. Kavuri, et K. Yin. A review of process fault detection and diagnosis part iii : Process history based methods. *Computers and Chemical Engineering*, 27(3) :327–346, 2003.

[148] V. Venkatasubramanian, R. Rengaswamy, K. Yin, et S.N. Kavuri. A review of process fault detection and diagnosis part i : Quantitative model-based methods. *Computers and Chemical Engineering*, 27(3) :293–311, 2003.

[149] Sylvain Verron, Teodor Tiplica, et Abdessamad Kobi. Bayesian networks and mutual information for fault diagnosis of industrial systems. Dans *Workshop on Advanced Control and Diagnosis*, 2006.

[150] Sylvain Verron, Teodor Tiplica, et Abdessamad Kobi. Fault diagnosis with bayesian networks : Application to the tennessee eastman process. Dans *IEEE International Conference on Industrial Technology*, Mumbai, India, 2006.

[151] Sylvain Verron, Teodor Tiplica, et Abdessamad Kobi. A new procedure based on mutual information for fault diagnosis of industrial systems. Dans *Workshop on Advanced Control and Diagnosis*, 2006.

[152] Sylvain Verron, Teodor Tiplica, et Abdessamad Kobi. Fault detection and identification with a new feature selection based on mutual information. *Journal of Process Control*, In Press, Corrected Proof :–, 2007.

[153] Sylvain Verron, Teodor Tiplica, et Abdessamad Kobi. Fault diagnosis of industrial systems with bayesian networks and mutual information. Dans *European Control Conference*, 2007.

[154] Sylvain Verron, Teodor Tiplica, et Abdessamad Kobi. Multivariate control charts with a bayesian network. Dans *4th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, 2007.

[155] Sylvain Verron, Teodor Tiplica, et Abdessamad Kobi. Procedure based on mutual information and bayesian networks for fault diagnosis of industrial systems. Dans *American Control Conference*, 2007.

[156] Kajiro Watanabe, Atsushi Komori, et Tsuyoshi Kiyama. Diagnosis of instrument fault. *Conference Record - IEEE Instrumentation and Measurement Technology Conference*, 1 :386–389, 1994.

[157] P. Weber, D. Theilliol, C. Aubrun, et A. Evsukoff. Increasing effectiveness of model-based fault diagnosis : A dynamic bayesian network design for decision making. Dans *6th IFAC Symposium on Fault Detection, Supervision and Safety of technical processes*, 2006.

[158] Philippe Weber et Lionel Jouffe. Reliability modelling with dynamic bayesian network. Dans *SafeProcess 2003, 5th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, Washington D.C.*, 2003.

[159] Philippe Weber et Marie-Christine Suhner. Modélisation de processus industriels par réseaux bayésiens orientés objet (rboo). *Revue d'Intelligence Artificielle*, 18 :299–326, 2004.

[160] G. Weidl, A.L. Madsen, et S. Israelson. Applications of object-oriented bayesian networks for condition monitoring, root cause analysis and decision support on operation of complex continuous processes. *Computers and Chemical Engineering*, 29(9) :1996–2009, 2005.

[161] Barry M. Wise, N. Lawrence Ricker, et David J. Veltkamp. Upset and sensor failure detection in multivariate processes. Technical report, Eigenvector Research, Manson, Washington, 1989.

[162] B.M. Wise et N.B. Gallagher. The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, 6(6) :329–348, 1996.

[163] S. Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20 :397–405, 1978.

[164] Ying Yang et Geoff I. Webb. A comparative study of discretization methods for naive-bayes classifiers. Dans *Proceedings of PKAW 2002*, 2002.

[165] Gilles Zwingelstein. *Diagnostic des défaillances, théorie et pratique pour les systèmes industriels*. Ed. HERMES, 1995.

Résumé

Cette thèse porte sur la surveillance (détection et diagnostic) des procédés multivariés par réseaux bayésiens. Ceci permet l'unification dans le même outil, un réseau bayésien, de plusieurs méthodes dédiées à la surveillance des procédés, telles que les cartes de contrôles multivariées, l'analyse discriminante ou bien la méthode MYT. Le premier chapitre expose les différents points clés de la surveillance des procédés, en étudiant les diverses approches permettant de réaliser celle-ci. Des méthodes de surveillance supervisées et non-supervisées sont présentées et une étude de différents classifieurs pour la surveillance est effectuée. Le choix d'un classifieur se porte alors sur les réseaux bayésiens. Le second chapitre est l'objet d'une présentation plus approfondie des réseaux bayésiens et des extensions possibles et intéressantes de ce genre d'outil dans le contexte de la surveillance des procédés. Puis, un état de l'art des méthodes de surveillance ou de diagnostic basées sur les réseaux bayésiens est étudié. Le troisième chapitre expose les contributions apportées au domaine de la surveillance des procédés par réseaux bayésiens. Les contributions apportées se répartissent en trois parties : détection, diagnostic supervisé et diagnostic non-supervisé. En s'appuyant sur ces contributions, la structure complète d'un réseau bayésien dédié à la surveillance des procédés est proposée. Le dernier chapitre présente une application de la méthode proposée sur un exemple classique : le procédé Tennessee Eastman. Les performances du réseau en terme de détection et de diagnostic sont évaluées. Finalement, les conclusions et perspectives de l'approche proposée sont émises.

Mots-clés: Surveillance des procédés, réseaux bayésiens, détection, diagnostic, supervisé, non-supervisé, analyse discriminante, sélection de composantes.

Abstract

This thesis is about the multivariate process monitoring (detection and diagnosis) with bayesian networks. It allows to unify in a same tool (a bayesian network) some monitoring dedicated methods like multivariate control charts, discriminant analysis and the MYT method. The first chapter gives some essential points of the process monitoring, with a study of different approaches. Some supervised and non-supervised monitoring methods are presented and a study of different classifiers for monitoring purpose is made. A classifier is then chosen : bayesian networks. The second chapter gives a more precise presentation of bayesian networks and their possible extensions in the context of process monitoring. After that, a state of the art of diagnosis and monitoring methods with bayesian networks is studied. The third chapter explains the contributions given to the topic of process monitoring with bayesian networks. These contributions are in three groups : detection, supervised diagnosis and non-supervised diagnosis. Based on these contributions, a complete structure of a bayesian network dedicated to process monitoring is given. The last chapter presents an application of the proposed method on a benchmark problem : the Tennesse Eastman Process. Efficiency of the network is evaluated for detection and for supervised and non-supervised diagnosis. Finally, conclusions and outlooks of the proposed approach are given.

Keywords: Process monitoring, bayesian networks, detection, diagnosis, supervised, non-supervised, discriminant analysis, feature selection.