LPNHE-2002-
15 juillet 2003

Rapport de Stage de Licence

Olivier ALI

ΦTEM

ENS de Cachan-UPMC

Identification d'électrons non isolés dans Atlas

LPNHE-Paris, Universités Paris 6-7, et IN2P3-CNRS

Table des matières

1 Informations générales

1.1 Le laboratoire

Le Laboratoire de Physique Nucléaire et des Hautes Energies (LPNHE), situé sur le campus inter-universitaire de Paris VI et Paris VII, est un des 18 laboratoires de l'Institut National de Physique Nucléaire et de Physique des Particules (IN2P3) qui constitue le département de physique nucléaire et corpusculaire du CNRS. Cet institut regroupe plus de 2000 personnes impliquées dans une cinquantaine d'expériences nationales et internationales, notamment au travers de partenariats avec le Cern (Laboratoire Européen de Physique des Particules près de Genève), le Fermilab (près de Chicago) et le Slac (près de Stanford).

Le laboratoire regroupe environ 110 personnes dont 26 chercheurs CNRS et 21 enseignants chercheurs, principalement impliqués dans les domaines expérimentaux de la physique des hautes énergies, la cosmologie et la physique des astroparticules, au travers de nombreuses expériences telles que Delphi, Nomad, H1, Babar, Auger, D0, Hess, Harp et Atlas.

1.2 L'expérience Atlas

Le détecteur Atlas (Air Toroidal Lhc Apparatus System) est l'une des deux expériences généralistes destinées à prendre leurs données auprès du collisionneur proton-proton LHC (Large Hardon Collider) du CERN. La collaboration Atlas regroupe près de 1800 chercheurs appartenant à environ 150 laboratoires, répartis entre 34 pays. Le détecteur actuellement en construction est l'un des plus grands et des plus complexes jamais réalisé en physique des particules. Il pèse 7000 tonnes, a un diamètre de 20 mètres, une longueur de 44 mètres et une hauteur de 25 mètres. L'observation directe de nouvelles particules s'effectuera par l'identification de leurs produits de désintégration (électrons, photons, muons, jets ...) et par la mesure de l'énergie déposée dans le détecteur.

Les données récoltées à partir de 2007 pourront ensuite être utilisées dans des analyses relevant de multiples thèmes de la physique, notammment la recherche de bosons de Higgs, particule susceptible d'expliquer l'origine de la masse des particules élémentaires.

1.3 Le groupe Atlas au LPNHE

Ce groupe de travail, dirigé par Philipe Schwemling, est constituté de 9 chercheurs, dont 6 permanents, 1 post-doc et 2 thésards, et d'une équipe de 17 techniciens et ingénieurs en informatique, mécanique et électronique. Il est impliqué dans la réalisation du calorimètre électromagnétique d'Atlas. Il s'agit d'un calorimètre à échantillonnage en plomb et argon liquide. Le groupe est aussi impliqué dans l'analyse de données prises en faisceau test sur les modules de ce même calorimètre, dans le développement de logiciels permettant la reconstruction des événements et la préparation des analyses de physique.

2 Sujet du stage

2.1 Problématique de l'expérience Atlas

La physique des particules s'attache à développer notre connaissance des quarks (les contituants des neutrons et des protons), des leptons (comme l'électron) et des interactions qui les gouvernent. Ces particules sont considérées, à ce jour, comme les constituants élémentaires de la matière. Les avancées, étroitement liées, de la théorie et des observations expérimentales, ont conduit à ce que l'on appelle le "Modèle Standard". Ce modèle décrit la vision que nous avons de l'univers et repose sur 12 constituants élémentaires répartis en trois familles. Le Modèle Standard explique trois des quatres interactions fondamentales, à l'exception de la gravitation. L'interaction électrofaible est le résultat de l'unification des interactions électromagnétique et faible véhiculées respectivement par le photon et les bosons vecteurs $W\pm$ et Z. Les gluons véhiculent l'interaction forte qui est responsable de la cohésion des noyaux. Ce modèle n'a pas été mis en défaut par les observations expérimentales. Cependant, nous savons qu'il ne constitue pas, à lui seul, la théorie ultime - de la physique des particules - puisque, dans sa forme actuelle, certaines questions importantes restent sans réponse, parmi lesquelles pourquoi et comment les particules acquièrent-elles une masse ? et pourquoi sont elles si différentes - de presque 0 à 175 fois la masse du proton ?

Une nouvelle particule - le boson de Higgs - pourrait expliquer ces deux questions et a déjà été recherché activement pendant plus de dix ans au CERN (Laboratoire Européen de Physique des Particules). En 2007 les études sur le Modèle Standard vont connaitre un regain d'activité avec le démarrage de l'accélérateur de particules LHC (Large Hadron Collider). C'est un collisionneur proton-proton avec une énergie dans le centre de masse de 14 TeV et une luminosité de $10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. Les croisement de faisceau sont séparés par 25 ns et on attend 23 interactions par croisement. Il offrira un gain d'un ordre de magnitude en énergie et de deux ordres de magnitudes dans le taux de collision proton-proton, par rapport aux machines actuelles. Il y a ainsi de grands espoirs d'améliorer notre connaissance du Modèle Standard et de découvrir le boson de Higgs, dont la masse est comprise entre 110 GeV (limite expérimentale) et 1 TeV (limite théorique), c'est à dire dans la gamme d'énergie accessible au LHC.

La calorimétrie joue un rôle important dans la détection de particules : les interactions dans la partie active du détecteur transforment l'énergie incidente de chaque particule en une gerbe qui libère des électrons qui sont collectés et induisent un signal mesurable par l'électronique de lecture du détecteur. Le rôle du calorimètre électromagnétique est ainsi de mesurer avec précision l'énergie et la position des électrons, photons et jets. Dans le cas d'Atlas, la mesure la plus précise possible de l'énergie des électrons est nécessaire dans la gamme de quelques dizaines de GeV, c'est pourquoi le développement et l'optimisation d'un algorithme à même d'identifier et d'analyser les traces laissées par ces électrons est un enjeu important. Une partie du groupe de travail, dont Frédéric Derue (mon maître de stage), participe au développement d'un tel algorithme de reconstruction et d'identification des électrons de basse énergie issus de la désintégration du Higgs ($H \rightarrow b\bar{b}$ et $b \rightarrow e$). Au LHC, le taux délectrons isolés est cinq ordres de grandeur plus faible que le taux de jets (pour une même énergie). Ainsi, la capacité d'Atlas à identifier des électrons doit être deux ordres de grandeur meilleur que sur les collisionneurs actuels. De même l'identification des électrons non isolés, à l'intérieur d'un jet, sera aussi extrêmement difficile. Cette identification sera basée sur la combinaison des informations du détecteur interne et du calorimètre électromagnétique.

2.2 L'indentification des électrons non isolés dans Atlas

A l'issue d'une collision des faisceaux de protons, des événements comme la désintégration d'un boson de Higgs H en une paire de quarks $b\bar{b}$. Un des quarks b peut lui-même se désintégrer en un électron. De chaque désintégration est alors issue une grande quantité de traces, essentiellement des pions, mais aussi quelques électrons. Les paramètres de ces traces calculés dans le détecteur interne permettent de localiser la région où un dépôt d'énergie dans le calorimètre électromagnetique doit être recherché. Dans une petite fenêtre on sélectionne la cellule ayant le dépôt d'énergie maximal. Les coordonnées de cette cellule définissent la graine de départ pour la constitution d'un amas, dans lequel on calculera le dépôt d'énergie. Après cette étape, pour chaque trace, on garde en mémoire un certain nombre de variables pertinentes, certaines d'entre elles étant utilisées pour l'identification des électrons. Parmi ces variables certaines caractérisent le développement longitudinal et latéral de la gerbe. Connaissant la forme attendue des distributions des variables pour les électrons et les pions on peut en déduire une "probabilité" pour chaque trace d'être un électron ou un pion.

2.3 Problématique du stage

Le choix de ces variables a été arrêté depuis 2000 lors de l'élaboration de la première version Fortran du code de reconstruction. Un nouveau programme de reconstruction en C++ est en cours d'écriture. Ce stage est donc l'occasion de se pencher à nouveau sur le choix de ces variables ainsi que sur le traitement statistique utilisé pour déduire cette "probabilité" pour chaque trace d'être un électron ou un pion.

Ce stage avait donc trois phases distinctes :

- Pour des raisons pratiques liées à l'écriture du code, les distributions des variables sont découpées d'une manière artificielle en plusieurs intervalles. Un découpage basé sur les particularités physiques des traces constitue la première étape du travail. Éventuellement l'utilisation de certains intervalles devra être discuté.
- Dans un deuxième temps la quantification du pouvoir discriminant de chaque variable. Cette étape n'a jamais été effectuée. Elle doit à la fois permettre de vérifier la pertinence de certaines variables et d'être un point de référence à l'avenir pour un ajout éventuel de nouvelles variables.
- La méthode d'analyse statistique ne prend pas en compte les corrélations éventuelles entre les différentes variables. Il est donc nécessaires de déterminer la matrice de corrélation pour l'ensemble des neuf variables. Enfin on développera une combinaison linéaire des variables, basée sur la méthode de Fisher, qui permet de tenir compte des corrélations entre les variables.

3 Travail personnel réalisé

On utilise des données qui simulent des événements de physique issus des collisions de faisceaux, comme le boson de Higgs H se désintégrant en une paire de quarks $b\bar{b}$ avec un des b se désintégrant lui-même en un électron, ainsi que la réponse du détecteur à ces événements. Ces données contiennent l'ensemble des informations produites par les algorithmes de reconstruction et d'identification. En particulier nous avons accès pour chaque événement à l'ensemble des traces et amas associés, ainsi que les 9 variables discriminantes utilisées par l'algorithme décrit ci-dessus.

Ces données sont analysées en local sur les stations de travail situées à Paris. Pour ce faire, un logiciel, couramment utilisé en physique des particules, développé au CERN, s'appelant Paw est choisi. C'est un utilitaire graphique servant à traiter des données aussi bien expérimentales que simulées. Il gère à la fois son propre langage propre et des routines Fortran, plus pratiques pour certaines tâches.

3.1 Étude des distributions des variables

Cette partie a constitué la mise en route de mon stage, elle m'a donc permis, dans un premier temps, de me familiariser avec les outils graphiques et d'analyse. Pour une valeur donnée d'une variable discriminante, on veut connaître la "densité de probabilité" d'être un électron ou un pion. Les distributions des variables sont traitées comme des fonctions de densité de probabilité. On conserve ces distributions, traitées comme des fonctions de densité de probabilité, par une suite de polynômes (procédure de smoothing de Paw). Les valeurs singulières sont conservées à part. Ainsi, les distributions des variables ont été découpées en 2 à 4 intervalles, soit au total 24 sous-classes pour chaque trace. Bien évidemment ce découpage, s'il permet une écriture relativement simple du code, n'est en rien justifié par des arguments physiques.

J'ai donc écrit et fait tourner des subroutines Fortran au sein du programme d'analyse afin d'estimer la fraction des traces concernées par ces valeurs singulières et de relever les valeurs des autres variables dans ces cas de figure. Si une trace n'est pas associée à un amas dans la première partie du calorimètre, les quatres variables issues de cette partie prennent leur valeur par défaut. Ainsi, on peut réduire de 4 à 1 le nombre de classes pour ce cas de figure. Certaines valeurs singulières sont bel et bien justifiées. J'ai ainsi mis en évidence que les cas $E_1/E = 1$ et $E_3/E = 1$, représentant les fractions d'énergie déposées dans les parties avant et arrière du calorimétre électromagnétique, sont expliqués par la prise en compte du bruit électronique dans la simulation et méritent un traitement spécifique. J'ai par ailleur trouvé un certain nombre de "bugs" dans l'initialisation des variables du programme de reconstruction d'Atlas qui ont amené, après coup, à l'élimination de certaines valeurs singuliéres. Enfin j'ai proposé un nouveau découpage des distributions basé sur l'association ou non d'une trace et d'un amas dans chaque compartiment qui réduit de moitié le nombre de classes utilisées.

3.2 Estimation du pouvoir discriminant des différentes variables

Le but de l'algorithme étant d'identifier les électrons et de rejeter les pions, la deuxième étape de mon travail à consisté à quantifier le pouvoir discriminant de chaque variable, chose qui n'avait jamais été faite. Je me suis inspiré d'une méthode utilisée par des collègues de la collaboration Babar, se basant sur un estimateur statistique appelé séparation défini par $< s^2 >= \frac{1}{2} \int \frac{(g^b(x)-g^s(x))^2}{g^b(x)+g^s(x)} dx$, où x représente la variable discriminante étudiée et $g^b(resp.g^s)$ les distributions des variables normalisées à un, pour les pions et les électrons. Les valeurs obtenues s'échelonnent de 0.6 à 0.04 ; elles ne sont donc pas toutes d'égal intérêt.

La variable s'avérant la plus discriminante provient du détecteur interne, justifiant bien l'emploi de deux détecteurs ayant leur capacité propre à identifier les électrons. La combinaison des variables issues de ces deux détecteurs va ainsi améliorer les performances individuelles de chaque détecteur. La variable possédant la séparation la plus faible est issue du calorimètre et mesure l'asymétrie des gerbes. En faisant tourner alors l'algorithme avec et sans cette variable, les performances en terme de réjection des pions restent inchangées. Cela avait déjà été noté il y a plusieurs années avec l'ancien programme de reconstruction d'Atlas. Ainsi cette constatation reste toujours valable et la variable d'asymétrie sera donc très certainement abandonnée.

J'ai cherché à combiner la variable la plus discriminante avec les autres à tour de rôle afin de déterminer la combinaison optimale. Par itérations successives on peut espérer atteindre une valeur limite pour la séparation, fixant alors le nombre de variables nécéssaires et suffisantes. Cependant, à partir de la combinaison de trois variables, le calcul de la séparation implique l'estimation numérique d'une intégrale à trois dimensions. Du au nombre limité d'événements disponibles, une telle intégrale ne peut être évaluée correctement.

3.3 Prise en compte des corrélations

La méthode utilisée pour combiner les variables (méthode du rapport de vraisemblance) ne tient pas compte les corrélations éventuelles entre les variables, ce qui peut biaiser la pureté du lot d'électrons sélectionné. J'ai évalué, pour la première fois, ces corrélations en construisant une matrice de corrélation 9×9, pour les électrons et les pions. J'ai ainsi confirmé que les variables issues du trajectographe sont très peu corrélées avec celles issues du calorimètre, ce qui est attendu puisque ces deux détecteurs sont indépendants. Les corrélations sont globalement faibles et restent inférieures à 40%. Certaines variables sont corrélées de manière similaire pour les électrons et les pions. Au contraire certaines sont corrélées pour les électrons mais pas pour les pions. J'ai donc cherché à prendre en compte ces corrélations, même si elles restent faibles.

La méthode de Fisher, là aussi utilisée par le groupe Babar du laboratoire, a paru être celle qui correspondait le mieux au problème posé. Elle détermine un hyperplan de l'espace des variables discriminantes sur lequel les projections des distributions électrons/pions de ces variables sont séparées au mieux. Techniquement, il s'agit de construire une matrice de covariance semblable à celle de corrélation. La difficulté principale est son inversion, ce qui a été fait dans un programme Fortran à l'aide d'une bibliothèque de programmes, permettant le calcul des coefficients de Fisher f_μ pour chaque variable. On construit alors une fonction unique $X_{FI} = \sum_{\mu=1}^{N_{var}} f_\mu x_{\mu i}(FI) - f_0$ qui permet de combiner les différentes variables x, tout en tenant compte de leurs corrélations.

Un programme test, générant des distributions gaussiennes, avec différentes corrélations, m'a permis de comprendre l'utilisation de la méthode de Fisher. La méthode de Fisher permet alors de remplacer trois variables corrélées décrivant le développement latéral de la gerbe par une seule. Cette variable est alors peu corrélée avec les variables restantes. Leur combinaison peut ainsi être effectuée par le rapport de vraisemblance usuel.

4 Bilan personnel et conclusion

Ce stage de fin de licence a été mon premier "vrai" contact avec le monde de la recherche, monde dans lequel j'espère évoluer plus tard.

Durant ces 6 semaines, je me suis investi, pour la première fois, dans un projet concret à grande échelle (l'expérience Atlas regroupe environ 1800 chercheurs) aux enjeux importants (les données accumulées par le détecteur pourront confirmer ou infirmer, entre autre, l'existence du boson de Higgs, particule susceptible d'expliquer l'origine de la masse des particules élémentaires). Cet investissement a été d'autant plus intéréssant et enrichissant que la physique des particules est un domaine qui m'intéresse depuis longtemps.

Les connaissances que j'ai pu acquéririrent complètent de surcroît celles que j'avais déjà obtenues, suite à mon TIPE de l'année 2002 sur la TPC du détecteur Delphi. Ces connaissances ont été bien utiles pour comprendre rapidement les enjeux et les motivations du projet Atlas. J'ai également pu mettre en pratique et consolider des enseignements reçus cette année dans le cadre de la licence, notamment en programmation Fortran, notions que j'ai complétées via l'utilisation de l'outil usuel d'analyse de données en physique des particules appelé Paw, et le traitement de texte Lateχ. J'ai également appréhender quelques notions de physique des particules qui me seront surement utiles dès l'année prochaine dans le cadre des enseignements optionnels.

Plus important encore, j'ai découvert ce qu'était la vie d'un groupe de recherche : travail d'équipe, réunions hebdomadaires. De plus l'utilisation d'outils statistiques non usuels dans la collaboration Atlas a mené à des échanges avec des membres d'autres groupes de travail, en particulier avec le groupe Babar. J'ai également profité de l'occasion pour assister à plusieurs exposés aussi bien sur les activités du groupe Atlas que sur les activités des autres groupes du laboratoire comme une soutenance de thèse sur le projet Harp ou les réunions du laboratoire. J'ai aussi essayé de tirer profit de ce stage au travers de toutes les impressions, les conseils et les mises en garde que chacun a bien voulu me donner en ce qui concernait mon orientation future.

Pour conclure je dirais que j'ai eu beaucoup de plaisir à faire ce stage et qu'il a été un moment privilégié de formation et d'apprentissage non seulement d'un point de vue scolaire mais également humain.

5 ANNEXES

5.1 Introduction

La physique des particules s'attache à développer notre connaissance des quarks (les contituants des neutrons et des protons), des leptons (comme l'électron) et des interactions qui les gouvernent. Ces particules sont considérées, à ce jour, comme les constituants élémentaires de la matière. Les avancées, étroitement liées, de la théorie et des observations expérimentales, ont conduit à ce que l'on appelle le "Modèle Standard". Ce modèle décrit la vision que nous avons de l'univers et repose sur 12 constituants élémentaires (cf tableau 1), répartis en trois familles. Le Modèle Standard explique trois des quatres interactions fondamentales, à l'exception de la gravitation. L'interaction électrofaible est le résultat de l'unification des interactions électromagnétique et faible véhiculées respectivement par le photon et les bosons vecteurs $W\pm$ et Z. Les gluons véhiculent l'interaction forte qui est responsable de la cohésion des noyaux. Ce modèle n'a pas été mis en défaut par les observations expérimentales. Cependant, nous savons qu'il ne constitue pas, à lui seul, la théorie ultime - de la physique des particules - puisque, dans sa forme actuelle, certaines questions importantes restent sans réponse :

fermions	leptons		quarks	
$1^{ère}$ famille	électron (e^-)	neutrino électron (ν_e)	haut (u)	bas (d)
$2^{ème}$ famille	muon (μ)	neutrino muon (ν_μ)	charme (c)	étrange (s)
$3^{ème}$ famille	tau (τ)	neutrino tau (ν_τ)	sommet (t)	beauté (b)
bosons	photon (γ)	bosons vecteurs ($Z^0, W^\pm$)	gluon	graviton

TAB. 1 – *Tableau des constituants élémentaires de la matière.*

– Comment pouvons-nous unifier l'interaction électrofaible avec les interactions forte et gravitationelle ? Le nombre élevé de paramètres fondamentaux[1] est également un indice qui tend à faire penser que le Modèle Standard n'est qu'une partie d'un modèle plus vaste dont la SuperSymétrie est l'une des extensions les plus populaires. Elle prédit, entre autres choses, l'existence d'une cohorte de nouvelles particules dites supersymétriques, partenaires des particules formant la matière ordinaire. Cette unification pourrait aussi conduire à la découverte de dimensions cachées bouclées sur elles-même.
- Comment expliquer la disparition de l'antimatière alors qu'à sa naissance l'univers contenait autant de matière que d'antimatière ? Ce mystère pourrait trouver sa solution grâce à l'étude des violations des symétries discrètes CP (C pour conjugaison de charge, P pour parité).
- Pourquoi - et comment - les particules élémentaires ont-elles une masse ? Et pourquoi ces masses sont-elles si différentes (de quelques eV à quelques centaines de GeV) ? Le mécanisme de Higgs pourrait répondre à ces deux questions. Une conséquence remarquable est qu'il prédit l'existence d'une particule scalaire neutre : le boson de Higgs. Cette particule est à ce jour le dernier élément manquant du Modèle Standard.

En 2007 les études sur le Modèle Standard vont connaitre un regain d'activité avec le démarrage de l'accélérateur de particules LHC (Large Hadron Collider). C'est un col-

[1]3 couplages de jauges, 2 paramètres pour le Higgs, les masses des 9 quarks et leptons chargés, 3 angles de mélanges et une phase de violation CP.

lisionneur proton-proton avec une énergie dans le centre de masse de 14 TeV et une luminosité de $10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. Les croisement de faisceau sont séparés par 25 ns et on attend 23 interactions par croisement. Il offrira un gain d'un ordre de magnitude en énergie et de deux ordres de magnitudes dans le taux de collision proton-proton, par rapport aux machines actuelles. Il y a ainsi de grands espoirs d'améliorer notre connaissance du Modèle Standard et de découvrir le boson de Higgs, dont la masse est comprise entre 110 GeV (limite expérimentale) et 1 TeV (limite théorique), c'est à dire dans la gamme d'énergie accessible au LHC. On peut de plus espérer voir apparaitre de la nouvelle physique dite au-delà du Modèle Standard. Ces sujets de physique seront l'enjeu de l'expérience Atlas.

L'expérience Atlas est entrée dans sa phase de construction pour tous ses sous détecteurs de manière à être prête pour les premières collisions en 2007. Les logiciels de reconstruction de la réponse des détecteurs au flot de particules produites par le LHC sont aussi en cours d'écriture. En particulier, le groupe de travail "électron-photon" de Atlas a pour but d'étudier les performances du calorimètre électromagnétique pour les électrons et les photons. Ceci implique la participation au développement d'algorithmes de reconstruction et d'identification, ainsi que l'évaluation et l'optimisation en permanence des performances des détecteurs obtenues avec des électrons et des photons. Dans ce cadre de travail un algorithme d'identification des électrons non isolés a été développé conjointement par le Laboratoire de Physique des Particules et de Hautes Énergies de Paris et par l'Institut de Physique Nucléaire de Cracovie. Un algorithme d'identification des électrons de basse impulsion transverse ($p_T < 20$ GeV[2]), localisés dans les jets produits lors de la fragmentation des quarks issus des collisions proton-proton des faisceaux du LHC, a été récemment développé dans Atlas. Cette identification sera utile dans de nombreux aspects liés à l'expérience Atlas, dont notamment :

- Calibration du calorimètre électromagnétique : La construction d'une échelle d'énergie commune pour le détecteur interne et le calorimètre électromagnétique à argon liquide est d'une importance cruciale pour toutes les analyses qui utiliseront les informations provenant de ces deux détecteurs. Ce sera bien évidemment le cas pour toutes les analyses utilisant des électrons pour la mesure d'une impulsion ou d'une masse invariante. Dans le cas d'Atlas, la mesure la plus précise possible de l'énergie des électrons est nécessaire dans la gamme de quelques dizaines de GeV. Une possibilité d'établir cette échelle d'énergie commune, et de la vérifier avec les mesures faites lors des tests en faisceau, est d'utiliser la désintégration de de bosons Z en paires électrons-positrons. Ces désintégrations sont utilisées pour calibrer le calorimètre électromagnétique dans cette gamme d'énergie. Pour vérifier cette calibration, ainsi que la linéarité du calorimètre, il est important d'obtenir les coefficients de calibration[3] pour des électrons de basse énergie. Ceci peut être fait grâce aux électrons de basse énergie produits par la désintégration de particules J/Ψ (état lié de quarks $c\bar{c}$).
- Etiquetage des jets issus de quarks b : la capacité à identifier des jets provenant de la fragmentation de quarks b joue un rôle clé dans l'analyse de nombreux processus attentus dans l'expérience Atlas. De nombreux de canaux de désintégrations aboutissent à des états finaux comportant des quarks b, par exemple :
  - désintégration du boson de Higgs $H \to b\bar{b}$;

[2]La direction du faisceau définie l'axe z. L'axe $x-y$ définit le plan transverse à la direction du faisceau. L'angle azimuthal ϕ est mesuré autour de l'axe du faisceau et l'angle polaire θ est l'angle par rapport à l'axe du faisceau. La pseudo-rapidité est définie comme $\eta = -\ln\tan(\theta/2)$. L'impulsion transverse p_T est définie dans le plan $x-y$.

[3]rapport entre l'énergie déposée dans le détecteur et l'énergie reconstruite : $E_{vraie} = \alpha \times E_{rec} + \beta$.

- désintégration du quark top ($t \to Wb$ et éventuellement, $t \to WbZ$).

Dans tous les mécanismes ci-dessus les jets originaires des quarks les plus légers et des gluons sont potentiellement sources d'un bruit de fond important. C'est pourquoi l'observation du signal dans les multijets-b requiert généralement une réjection importante des jets sans quarks b et la plus grande efficacité possible pour l'identification des jets contenant des quarks b.

Il a déjà été montré, par les expériences au Tevatron et au LEP, qu'un étiquetage efficace des jets issus de quarks b peut être réalisé de deux façons indépendantes :

– identification de vertex[4] secondaires dus au temps de vie relativement long des hadrons B, impliquant une longueur de vol de l'ordre de quelques millimètres ;
– l'étiquetage des leptons "mous" basé sur l'identification des leptons mous issus des désintégrations semileptoniques des mésons B.

Alors que l'efficacité des algorithmes de reconstruction de vertex est seulement limité par les performances du détecteur, l'efficacité attendue pour la méthode d'étiquetage des leptons "mous" est limitée par la fraction de désintégrations de méson B contenant des leptons ($\approx 17\%$ par famille de leptons).

Il a été montré que la combinaison de ces deux méthodes permet une amélioration de 16% sur la réjection des jets u et 10% sur la réjection des jets c dans Atlas [WM00].

Un algorithme d'identification des électrons de basses impulsions transverses a été récemment implémenté dans le programme de reconstruction Athena [gro] de l'expérience Atlas. Le sujet du stage a pour but d'étudier une éventuelle optimisation de cet algorithme, en particulier dans sa partie d'analyse statistique. En effet, la méthode utilisée jusqu'ici ne prend pas en compte les corrélations entre les variables discriminantes utilisées par l'algorithme. Nous allons étudier une méthode d'analyse multidimensionnelle (dite de Fisher) pour les prendre en compte.

La section 5.2 effectuera un bref rappel des quelques spécificités du détecteur Atlas qui nous serons utiles pour établir l'algorithme d'identification des électrons. La section 5.3 décrira brièvement le lot de données utilisé pour cette étude. La section 5.4 décrit l'algorithme d'identification tel qu'implémenté à l'heure actuelle. Enfin la section 5.5 donnera l'ensemble des résultats obtenus lors de ce stage.

5.2 Le détecteur Atlas

Une description détaillée peut être trouvée dans "Atlas : Detector and Physics Performance Technical Design Report" [col99a]. Une description plus abordable peut être trouvée dans la thèse d'habilitation à diriger des recherches [Sch02] d'où est issue cette description succincte. On distingue sur la figure 1 les systèmes essentiels de détection et de mesure des particules. La forme du détecteur, cylindrique, est caractéristique d'une expérience sur collisionneur avec des couches successives de sous-détecteurs entourant le point d'interaction. En suivant la trajectoire d'une particule, depuis le point d'interaction situé au milieu du détecteur, on voit successivement :

- Le détecteur interne pour la reconstruction des traces chargées et des vertex de production et désintégration ;
- Le calorimètre électromagnétique pour la mesure du dépôt en énergie des particules neutres et chargées formant des gerbes électromagnétiques (comme les photons et les électrons) ;
- Le calorimètre hadronique pour la mesure du dépôt en énergie des particules neutres

[4]un vertex correspond à l'endroit de la désintégration d'une particule en ses produits de désintégration.

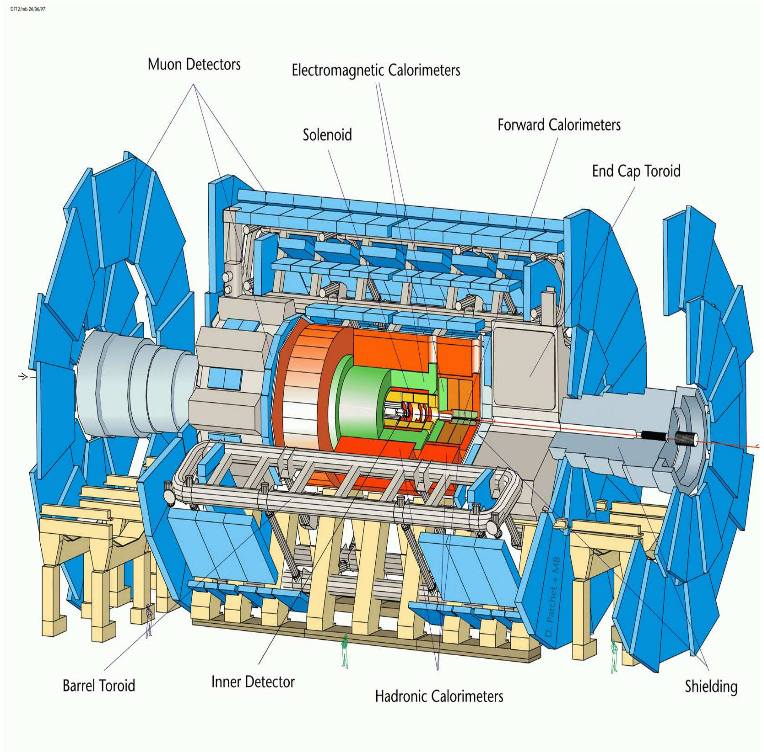

FIG. 1 – *Représentation du détecteur Atlas.*

et chargées formant des gerbes hadroniques (comme les pions, les kaons etc...) ;

- Le spectromètre à muons baignant dans un champ toroïdal.

Dans ce qui suit, nous nous intéresserons uniquement aux sous-détecteurs qui nous concernent directement, à savoir le détecteur interne et le calorimètre électromagnétique.

5.2.1 le détecteur interne

Le détecteur interne est lui-même composé de plusieurs sous-détecteurs. Les parties les plus proches du point d'interaction, cruciales pour la résolution en impulsion et la reconstruction des vertex, sont constituées de trois couches de pixels de silicium (à des rayon de 4, 11 et 14 cm du point d'interaction) et de détecteur à micropistes (à des rayons de 30, 37,3, 44,7 et 52 cm du point d'interaction). L'agencement des couches est fait de telle manière que chaque trace chargée voit au moins quatre couches de détecteur à micropistes et trois couches de pixels. Les couches les plus externes (pour les rayons compris entre 56 et 107 cm) ont été réalisées à l'aide d'un détecteur à tubes à pailles (TRT, Transition Radiation Tracker) en raison du coût élevé de la technologie des détecteurs à silicium. Chaque trace chargée est échantillonnée sur 36 points par ce détecteur. De plus, et surtout en ce qui concerne ce stage, l'emploi de xénon en tant que milieu actif permet de détecter les photons de rayonnement de transition[5] engendrés par par le passage des traces dans le radiateur situé entre les pailles, ce qui permet d'identifier les électrons. Cette particularité sera primordiale dans la mise au point de l'algorithme d'identification des électrons. La figure 2 à gauche montre la simulation d'un événement $H \to b\bar{b}$ tel qu'il sera vu dans le détecteur interne. On reconnaît aisément les trois couches de silicium entourant le vertex, les quatres couches supplémentaires situées à un rayon plus grand.

[5]photons X émis lors du passage d'électrons dans les radiateurs situés entre les pailles.

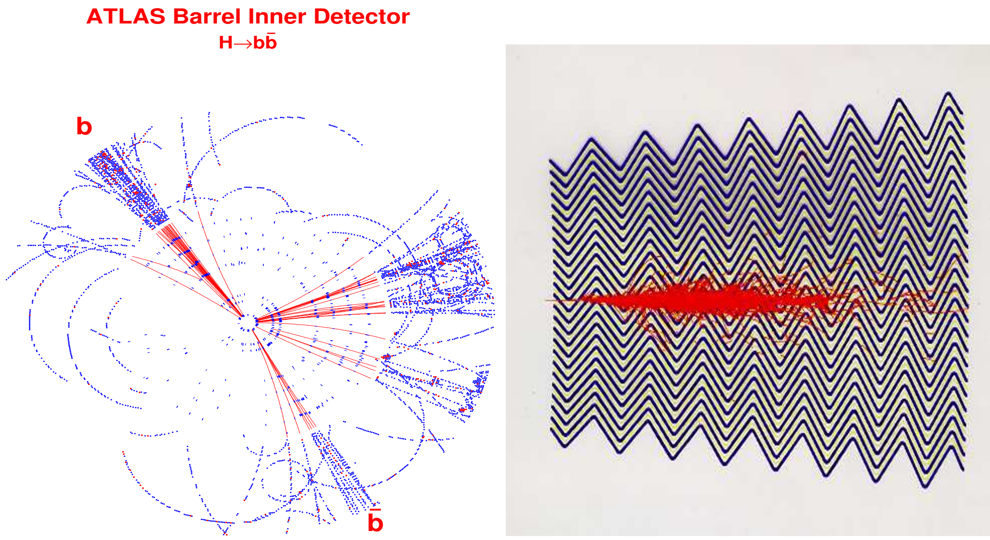

Fig. 2 – *Simulation d'un événement $H \to b\bar{b}$ dans le détecteur interne (à gauche). Simulation du développement d'une gerbe électromagnétique dans le calorimètre électromagnétique à (droite).*

Enfin chaque trace est largement échantillonnée par le grand nombre de coups dans le détecteur TRT à la périphérie du détecteur interne.

Accompagnant le détecteur interne, on trouve un solénoïde formé d'une bobine supraconductrice de 2.44 m de diamètre interne, fonctionnant à une température de 4.5 K et parcourue par un courant de 7600 Ampères. Son rôle est de créer un champ magnétique de 2 T dans lequel baigne le détecteur interne, de manière à courber les trajectoires des particules chargées pour la mesure de leur impulsion.

5.2.2 Le calorimètre électromagnétique

Il s'agit du détecteur sur lequel repose l'essentiel du travail du groupe Atlas du LPHNE. C'est un calorimètre à échantillonnage, l'absorbeur est constitué de plaques de plomb pliées en forme d'accordéon et le milieu actif est constitué d'argon liquide. Les interactions dans les aborbeurs (en plomb) transforment l'énergie incidente de chaque particule en une gerbe (cf. Fig 2 à droite). Dans l'argon elles libèrent des électrons qui sont collectés et induisent un signal mesurable par l'électronique de lecture du détecteur. Le rôle du calorimètre électromagnétique est ainsi de mesurer avec précision l'énergie et la position des électrons, photons et jets.

Le calorimètre électromagnétique se compose d'une partie centrale et de deux parties "bouchons" avant et arrière. Les dimensions de la partie centrale sont données dans le tableau 2. Les parties avant et arrières couvrent quant à elles le domaine $1.45 < |\eta| < 3.20$. Comme le calorimètre électromagnétique est situé derrière la bobine, il est précédé d'un pré-échantillonneur dont le rôle est d'estimer la fraction d'énergie perdue dans la bobine et la paroi du cryostat entourant le calorimètre, permettant de maintenir l'argon liquide à une température de 89.3 K. Le pré-échantillonneur est constitué d'électrodes de cuivre et d'époxy baignant dans l'Argon liquide. Son mode de fonctionnement est très semblable à celui du calorimètre, la différence essentielle étant qu'il ne comporte pas d'absorbeur.
Le calorimètre électromagnétique est divisé en trois compartiments (cf. Fig 3), utilisés par l'algorithme d'identification des électrons :

rayon interne	125 cm
rayon externe	225 cm
longueur	640 cm
domaine couvert	$\|\eta\| < 1.45$

TAB. 2 – *dimensions du calorimètre électromagnétique*

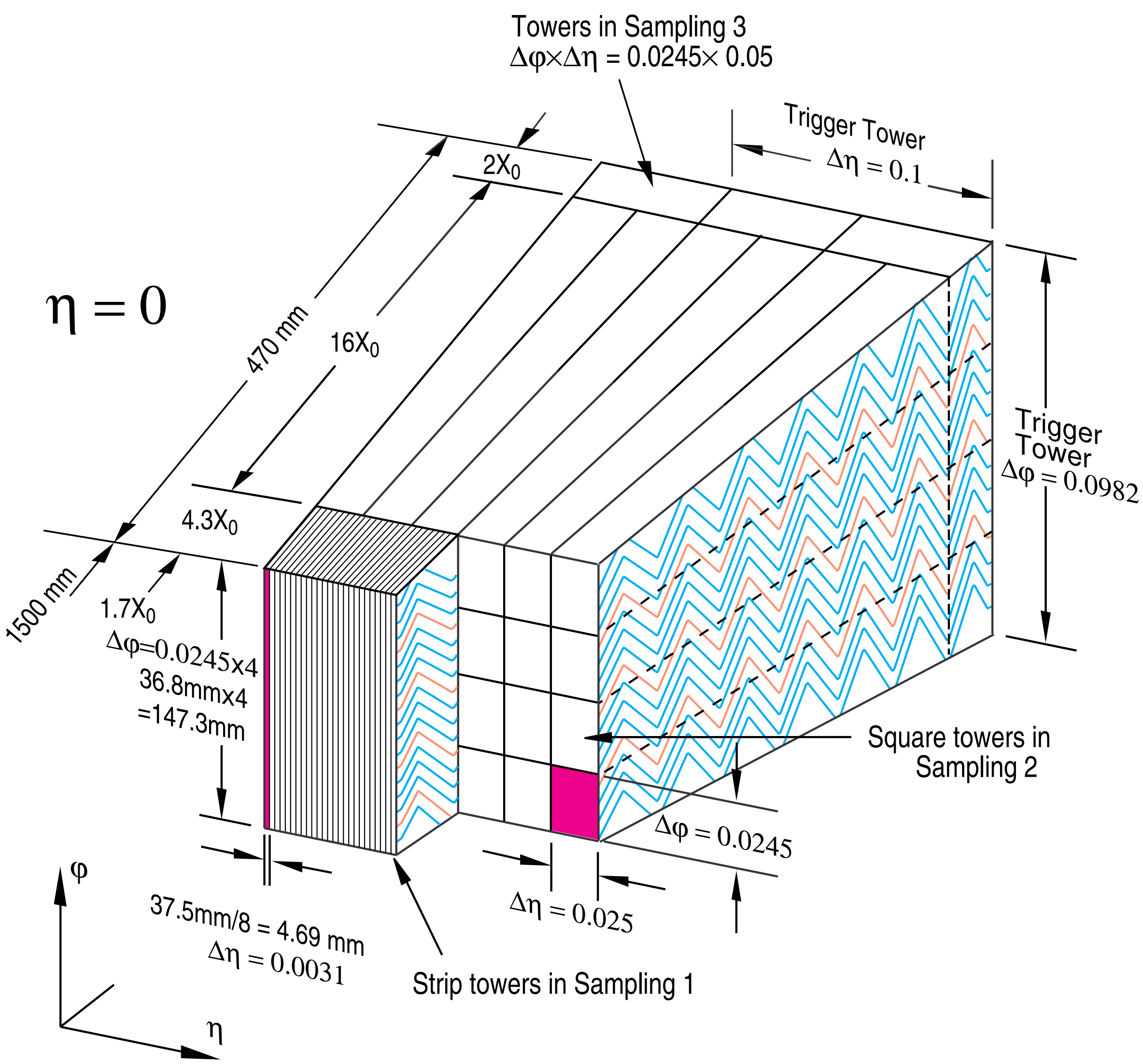

FIG. 3 – *La géométrie projective, la structure en accordéon, la structure en trois compartiments et la granularité du calorimètre électromagnétique.*

- le premier compartiment possède une granularité très fine (environ dix fois plus fine en pseudo-rapidité que les autres compartiments) de manière à mesurer très précisément l'angle et la position du développement de la gerbe, utile pour la mesure de la masse invariante du Higgs dans sa désintégration en deux photons ;
- le deuxième compartiment, le plus grand, va contenir l'essentiel de la gerbe et va donc permettre la mesure de l'essentiel du dépôt en énergie ;
- le troisième compartiment va permettre la mesure des gerbes de très grandes énergies qui pourraient ne pas être entièrement contenues dans les deux premiers comparti-

ments. De plus, ce compartiment sera essentiel pour la séparation entre les gerbes issues d'électrons de basse énergie qui vont déposer peu dénergie dans ce compartiment et les hadrons qui eux vont y déposer une part substantielle de leur énergie.

5.3 Simulation et reconstruction dans Atlas

Pour la majorité des utilisateurs, analysant des événements issus de collisions protons-protons, l'expérience Atlas peut être subdivisée en trois sous-systèmes : le détecteur, pris comme une source de données, le système de déclenchement permettant de sélectionner rapidement les événements intéressants et des les enregistrer, et le logiciel produisant des outils de reconstruction, d'identification et d'analyse pour extraire la physique des signaux ainsi collectés.

Des programmes classiquement utilisés en physique des particules permettent de simuler la physique issue des collisions de faisceaux et de simuler la réponse du détecteur à ces événements. On peut ainsi optimiser certaines parties du détecteur avant la phase de construction, évaluer les performances du détecteur et mettre au point des algorithmes de sélection spécifiques à certaines particules. Voyons maintenant brièvement ces étapes, toujours dans l'optique de la mise au point d'un algorithme d'identification des électrons.

5.3.1 Simulation de la physique

D'une collision de deux faisceaux de protons à 7 TeV le générateur Monte Carlo Pythia 5.7 [Sjo93] simule la création des particules et des processus qui s'en suivent. Après les cascades de désintégrations, une liste de particules stables est produite. Des événements WH contenant un boson W et un boson de Higgs H avec des désintégrations $W \to \mu\nu$ et $H \to b\bar{b}$ pour une masse du boson de Higgs $m_H = 100$ GeV ainsi que des événements $B_d \to J/\Psi K_S$, avec une désintégration du J/Ψ en une paire électron-positron, sont engendrés. C'est sur ce lot d'événements simulés que nous travaillerons par la suite.

5.3.2 Simulation du détecteur

Pour le détecteur Atlas on utilise une description complète du détecteur permettant de simuler la réponse de l'appareillage au passage des particules. L'environnement Geant 3.21 [ea86] est utilisé pour la description de la géométrie ainsi que la simulation du détecteur, incluant le suivi des traces associées à chaque particule et la réponse de l'électronique des éléments actifs du détecteur.

5.3.3 Reconstruction des événements

Le logiciel de reconstruction de l'expérience Atlas est appelé Athena et est basé sur un langage orienté objet (C++). Ce logiciel met en relation de nombreux algorithmes visant à reconstruire, pour chaque détecteur, des particules à partir des informations issues de la simulation de l'appareillage. On aura ainsi accès à des traces - pour les particules chargées et à des cellules dans le calorimètre contenant les dépôts d'énergie.

Reconstruction dans le détecteur interne Un algorithme appelé xKalman permet la reconstruction des traces dans le détecteur interne. La procédure de recouvrement du Bremsstrahlung[6] est appliquée. Celle-ci est cruciale puisque les électrons qui nous

[6] rayonnement de freinage correspondant à l'émission de photons par des particules chargées dont la trajectoire est courbée.

intéressent ici ont une faible impulsion transverse.

Reconstruction dans le calorimètre électromagnetique Les paramètres de la trace calculés dans le détecteur interne permettent de localiser la région où un dépôt d'énergie dans le calorimètre électromagnétique doit être recherché. Les coordonnées de ce point doivent être corrigées de manière à prendre en compte le champ magnétique appliqué dans le détecteur interne. Dans une petite fenêtre en $\Delta\eta \times \Delta\phi$ on sélectionne la cellule ayant le dépôt d'énergie maximal. Les coordonnées de cette cellule définissent la graine de départ pour la constitution d'un amas, dans lequel on calculera le dépôt d'énergie. La même procédure est appliquée pour chacun des compartiments du calorimètre. Après cette étape, pour chaque trace, on garde en mémoire un certain nombre de variables pertinentes, certaines d'entre elles étant utilisées pour l'identification des électrons.

A la suite de ces étapes un algorithme d'identification est appelé. Sa description fait l'objet de la section suivante. Après cette étape, des fichiers de données "reconstruites" sont ainsi produits et mis à disposition de la communauté Atlas pour la mise au point et l'optimisation d'autres algorithmes ou pour l'étude des performances de ces mêmes algorithmes.

5.4 Description de l'algorithme d'identification

On se contentera ici de décrire brièvement le principe de l'algorithme d'identification des électrons. Cet algorithme est implémenté dans le logiciel de reconstruction d'Atlas [DF03].

5.4.1 Variables d'identification

De manière à identifier des traces issues d'électrons tout en rejetant les traces issues de pions[7], un ensemble de variables discriminantes, utilisant les informations du détecteur interne et du calorimètre électromagnétique, est définie. Ces variables peuvent être classées de la manière suivante.

- Variables n'utilisant que le détecteur interne :
  - N_{TR} - nombre de coups (TR hits) dans le TRT issus du rayonnement de transition ;
  - A_0 - le paramètre d'impact transverse de la trace au niveau du vertex de la collision.
- Variables utilisant les informations du calorimètre électromagnétique :
  - $E_1(3 \times 1)/E_{core}$ - fraction d'énergie déposée dans le premier compartiment longitudinal du calorimètre électromagnétique. L'énergie $E_1(3 \times 1)$ est calculée à partir des trois pistes en η, centrés autour de la piste de plus grande énergie. $E(core)$, l'énergie de cœur est définie comme la somme des énergies définies dans les fenêtres :
    - 3×3 dans le pré-échantilloneur ;
    - 15×3 dans le premier compartiment ;
    - 5×5 dans le compartiment du milieu ;
    - 3×5 dans le compartiment arrière.
  - $E_3(3{\times}3)/E_{core}$ - fraction d'énergie déposée dans le trosième compartiment longitudinal du calorimètre électromagnétique. L'énergie $E_3(3 \times 3)$ dans le compartiment

[7] On cherche à séparer les traces issus délectrons de signal, provenant de désintégrations de quarks b ou de J/Ψ, des autres traces, à savoir des hadrons (essentiellement des pions), et des électrons de bruit de fond (provenant de photons de conversion ou de désintégration Dalitz.)

arrière est la somme des énergies déposées dans une fenêtre $\Delta\eta \times \Delta\phi = 3 \times 3$ autour de la cellule la plus énergétique;
- $ISO = 1 - E_{3\times3}/E_{3\times7}$ - isolation de la gerbe électromagnétique dans le calorimètre électromagnétique. Plus précisément, l'amas 3×3 est construit en sommant les énergies déposées dans chaque compartiment dans les fenêtres :
  - 1×1 dans le pré-échantilloneur;
  - 3×1 dans le premier compartiment;
  - 3×3 dans le compartiment du milieu;
  - 3×3 dans le compartiment arrière.

  L'amas 3×7 est construit en sommant les énergies déposées dans chaque compartiment dans les fenêtres :
  - 3×3 dans le pré-échantilloneur;
  - 15×3 dans les pistes;
  - 3×7 dans le compartiment du milieu;
  - 3×7 dans le compartiment arrière.
- $W = \sqrt{\sum_{i=imax-1}^{i=imax+1} E_i \times (i - i_{max})^2 / \sum_{i=imax-1}^{i=imax+1} E_i}$, largeur de la gerbe dans 3 pistes, autour de la piste la plus énergétique ($imax$). E_i est l'énergie déposée dans la cellule i dans la direction η pour une $\phi = \phi_{max}$ constant.
- $ASY = \left| \frac{\sum_{i=imax+1}^{i=imax+3} E_i - \sum_{i=imax-3}^{i=imax-1} E_i}{\sum_{i=imax+1}^{i=imax+3} E_i + \sum_{i=imax-3}^{i=imax-1} E_i} \right|$, l'asymétrie du développement latéral de la gerbe dans trois pistes, autour de la piste le plus énergétique.

- Variables combinant les informations des deux détecteurs :
  - $E_T(core)/p_T$, le rapport de l'énergie transverse déposée dans le calorimètre dans une cellule 3×3, sur l'impulsion transverse de la trace;
  - $POS = \frac{\sum_{i=imax-7}^{i=imax+7} E_i \times (i - imax)}{\sum_{i=imax-7}^{i=imax+7} E_i}$, la différence entre la position extrpolée de la trace dans le calorimètre et la position de la gerbe électromagnétique, mesurée en unité de pistes. $imax$ est ici la cellule d'impact attendue pour la trace extrapolée.

La Figure 4 montre la forme des distributions obtenues pour des traces d'électrons et de pions issus d'événements $H \to b\bar{b}$. On notera d'ores et déjà le grand pouvoir séparateur de la variable N_{TR}, avec, comme attendu, une distribution piquée à des faibles valeurs pour les pions, et à des grandes valeurs pour les électrons. On notera aussi le grand pouvoir discriminant de la variable E_T/p_T, piquée à 1 pour les électrons comme il se doit.

5.4.2 Construction d'une fonction discriminante

Pour séparer un signal d'un bruit de fond, la méthode la plus simple consiste à appliquer des coupures sur les variables discriminantes. Cette méthode n'est cependant pas optimale puisque d'une part elle coupe du signal, d'autre part elle n'utilise pas des informations telles la forme des distributions ou la possible corrélation entre les variables. La méthode utilisée ici repose sur un rapport de vraisemblance et est construite à partir de la distribution des variables discriminantes obtenues pour les deux classes d'événements à distringuer (électrons et pions). Idéalement, pour N variables discriminantes, le rapport de vraissemblance s'écrit :

$$X_{PA} = \frac{\mathcal{L}^1(x_1, ...x_N)}{\mathcal{L}^1(x_1, ...x_N) + \mathcal{L}^2(x_1, ...x_N)}, \tag{1}$$

où $\mathcal{L}^1 = g^1(x_1, ...x_N)$ est la distribution de densité à N dimensions (ou vraisemblance) pour la classe 1 (électrons), et $\mathcal{L}^2 = g^2(x_1, ...x_N)$ pour la classe 2 (pions).

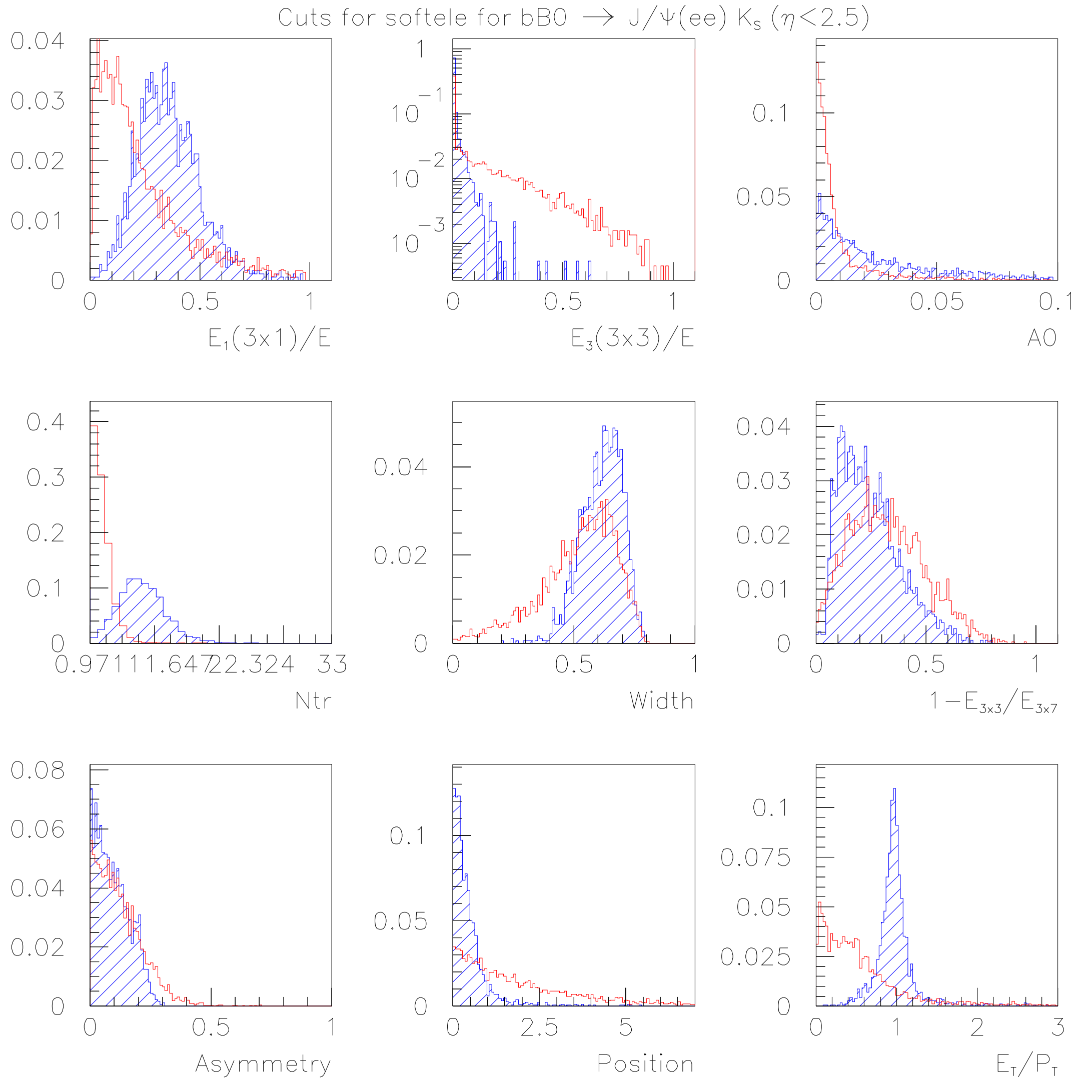

FIG. 4 – *Fonctions de densité de probabilité des variables utilisées pour l'identification des électrons dans des événements $H \to b\bar{b}$. L'histogramme hachuré correspond aux électrons (signal), l'autre correspond aux traces de pions (bruit de fond).*

Les distributions des variables continues sont traitées en tant que fonctions de densité de probabilité et paramétrées par une suite de polynômes par le programme Paw (opération dite de "smoothing"). La variable discrète N_{TR} est elle conservée par une table. L'approximation suivante est effectuée :

$$X_{PA} = \frac{\prod_i g_i^1(x_i)}{\prod_i g_i^1(x_i) + \prod_i g_i^2(x_i)}, \tag{2}$$

où $g_i^1(x_i)$ (resp. $g_i^2(x_i)$) est la distribution de densité à une dimension de la variable x_i pour les événements de la classe 1 (resp. 2). Cette approximation consiste à négliger les corrélations entre les variables.

Plusieurs commentaires peuvent être effectués :

- par construction, X_{PA} tend vers 1 pour la classe 1 (électrons) et vers 0 pour les événements de la classe 2 (pions), cf. Fig 5 ;
- si une variable discriminante x_i n'est pas utile ($g_i^1(x_i) \sim g_i^2(x_i)$) sur toute la gamme considérée pour x_i, elle ne dilue pas l'information provenant des autres variables (ce qu'on pourra noter dans la Fig 6 à droite) ;
- s'il n'y a pas de corrélation entre les variables, la variable combinée X_{PA} est optimale ;
- dans le cas où il y a des corrélations entre les variables discriminantes, de l'information est perdue. De plus l'utilisation de X_{PA} comme variable combinée peut introduire un biais dans les résultats, qui est de plus difficile à déterminer. On peut ainsi noter sur la figure 5 qu'un grand nombre de traces de pions ont une grande valeur de X_{PA}. Ces traces étant largement plus nombreuses que les traces d'électrons (au moins dix fois plus nombreuses), on voit que les corrélations résiduelles entre les variables peuvent limiter la pureté du lot de traces d'électrons sélectionné. Enfin l'ajout d'un trop grand nombre de variables corrélées va finir par diluer le pouvoir discriminant de cette méthode.

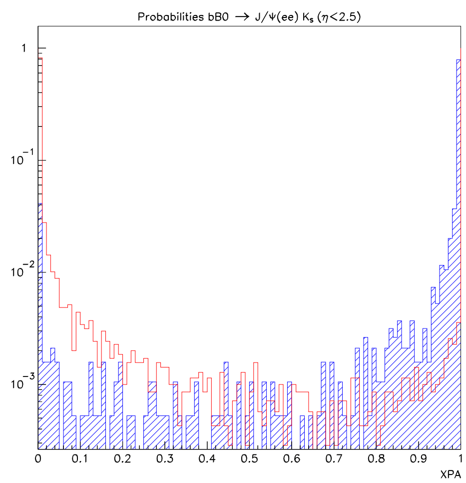

FIG. 5 – *Distribution obtenue pour X_{PA}. L'histogramme du signal (hachuré) pique vers 1, l'histogramme du bruit de fond (trait plein) pique vers 0.*

5.4.3 Procédure d'identification des électrons

L'identification d'un candidat comme trace d'électron est basée sur la valeur de la fonction discriminante associée à chaque trace. Les traces pour lesquelles X_{PA} est en-dessous d'une valeur donnée seront rejetées comme candidats. Pour quantifier les performances de l'algorithme on définit :

- Une efficacité d'identification des électrons :

$$\epsilon^e = \frac{N_e^t}{N_e}, \tag{3}$$

  où N_e^t est le nombre de traces d'électrons ayant X_{PA} au-dessus du seuil, et N_e le nombre total de traces d'électrons.

- Une réjection pour les pions :

$$R^{\pi} = \frac{N_{\pi}}{N_{\pi}^{t}}, \tag{4}$$

où N_{π}^{t} est le nombre de traces de pions ayant X_{PA} au-dessus du seuil, et N_{π} le nombre total de traces de pions.

En faisant varier la coupure sur X_{PA} on peut ainsi construire la courbe de réjection des pions en fonction de l'efficacité d'identification des électrons. Une telle courbe pour des traces issues d'événements $H \rightarrow b\bar{b}$ est présentée Fig 6. Les performances ainsi obtenues

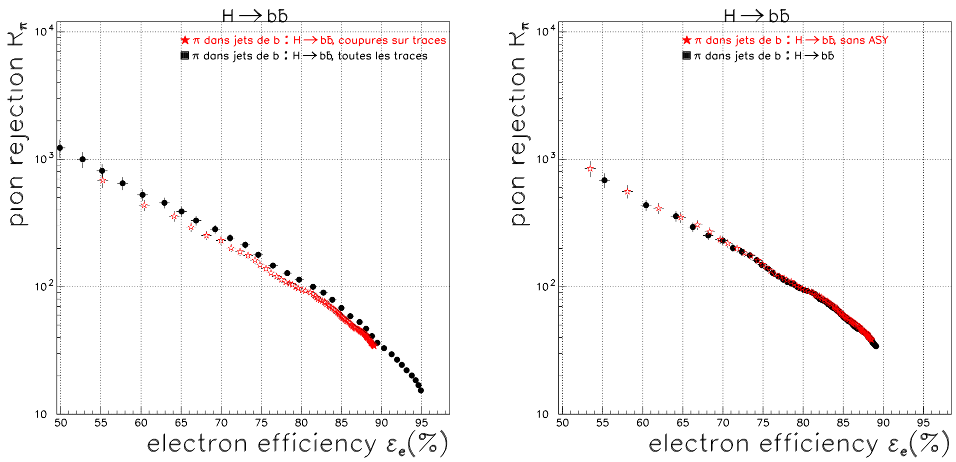

FIG. 6 – *Réjection des traces de pions en fonction de l'efficacité d'identification des électrons pour du $H \rightarrow b\bar{b}$. Les courbes représentent les résultats obtenus sur l'ensemble des traces et sur le lot de traces retenues pour la suite. A droite la même courbe mais avec et sans utilisation de la variable ASY.*

sont similaires à celles obtenues dans [A.00].

5.5 Travail personnel

On présente ici le travail personnel effectué lors de ce stage. Celui-ci s'est effectué sur un ensemble de données ayant passé les étapes de génération (cf. section 5.3.1), simulation (cf. section 5.3.2) et reconstruction (cf. section 5.3.3). Ces données contiennent l'ensemble des informations produites par les algorithmes de reconstruction et d'identification. En particulier nous avons accès pour chaque événement à l'ensemble des traces et amas associés issus de l'algorithme décrit ci-dessus.

Ces données sont analysées en local sur les stations de travail situées à Paris. Pour ce faire, un logiciel, couramment utilisé en physique des particules, développé au CERN, s'appelant Paw est choisi. C'est un utilitaire graphique servant à traiter des données aussi bien expérimentales que simulées. Il gère à la fois son propre langage propre et des routines Fortran, plus pratiques pour certaines tâches.

Une première étape, décrite dans la section 5.5.1, a permis à la fois de s'initier à Paw et d'améliorer notre connaissance de la forme des distributions des variables discriminantes. La section 5.5.2 montre le développement d'une variable quantifiant le pouvoir discriminant de chaque variable. Enfin la section 5.5.3 qui constitue l'essentiel du stage vise à la prise en compte des corrélations entre les variables discriminantes.

5.5.1 Découpage des PDFs (Probability Density Functions)

Pour une valeur donnée d'une variable discriminante, on veut connaitre la "densité de probabilité" d'être un électron ou un pion. Les distributions des variables sont traitées comme des fonctions de densité de probabilité. On conserve ces distributions, traitées comme des fonctions de densité de probabilité, par une suite de polynômes (procédure de smoothing de Paw). Les valeurs singulières sont conservées à part. Ainsi, les distributions des variables ont été découpées en quelques intervalles. Pour les variables E_1/E et E_3/E quatres intervalles sont utilisés :

- l'intervalle égal à -9999 signifiant que l'extrapolation de la trace n'aboutit pas à une cellule dans laquelle il y a bien un dépôt d'énergie ;
- l'intervalle égal à 0 signifiant qu'il n'y a pas de dépot d'énergie ;
- l'intervalle égal à 1 signifiant que toute l'énergie est déposée dans la partie testée du calorimètre ;
- pour tous les intervalles entre 0 et 1 on garde la paramétrisation de la distribution.

Un tel découpage est illustré sur la figure 7 pour la variable E_1/E. Pour les variables POS, ASY et W, deux intervalles sont utilisés :

- l'intervalle égal à -9999 signifiant que l'extrapolation de la trace n'aboutit pas à une cellule chaude ;
- tous les canaux entre 0 et 1 : on garde la paramétrisation de la distribution.

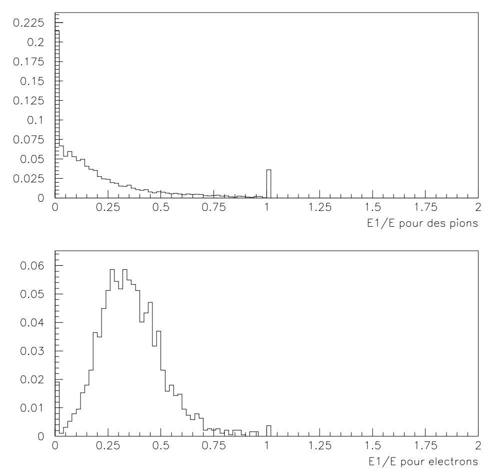

FIG. 7 – *Histogrammes représentant les distributions de la variable E_1/E pour les hypothèses électron et pions*

Pour la variable ISO trois intervalles sont utilisés :

- l'intervalle égal à -9999 signifiant que l'extrapolation de la trace n'aboutit pas à une cellule chaude ;
- l'intervalle égal à 0 signifiant qu'il n'y a pas de dépot d'énergie ;
- pour tous les intervalles entre 0 et 1 on garde la paramétrisation de la distribution.

Pour la variable E_T/p_T, trois intervalles sont utilisés :

- l'intervalle égal à 0 signifiant que l'extrapolation de la trace n'aboutit pas à une cellule chaude dans aucun des compartiments du calorimètre ;
- tous les canaux entre 0 et 3 : on garde la paramétrisation de la distribution ;
- les traces ayant $E_T/p_T > 3$ sont traitées à part.

Pour la variable n_{TR}, deux intervalles sont utilisés :
- l'intervalle entre 0 et 33 pour l'ensemble des traces ;
- si $n_{TR} > 33$ on considère que la trace est un électron.

Pour la variable A_0, deux intervalles sont utilisés :
- l'intervalle entre 0 et 0.1 pour leqeul on garde la paramétrisation ;
- si $A_0 > 1$ on considère que la trace n'est pas un électron.

Soit au total 24 sous-classes pour chaque trace. Bien évidemment ce découpage, s'il permet une écriture relativement simple du code, n'est en rien justifié par des arguments physiques. Il serait préférable de le simplifier, dans la mesure du possible.

En regardant le tableau 4 donnant les fractions des traces ayant une valeur singulière de E_3/E, il nous a semblé étrange que près de 50% des électrons et 40% des pions n'aient pas de dépôt d'énergie dans le troisième compartiment. Nous avons donc recensé les valeurs singulières (principalement les canaux -9999 et 0) pour chaque variable et regardé dans chaque cas les valeurs prises par les autres variables. Les tableaux 3 et 4 illustrent les cas des variables E_1/E et E_3/E et donnent la fraction f des traces, issues de pions et d'électrons, prenant des valeurs singulières. L'erreur statistique correspondante est donnée par

$$\sigma(f) = \sqrt{f(1-f)/N_{tot}}, \tag{5}$$

où N_{tot} est le nombre total de trace. Ainsi, le pic noté sur la figure 7 pour les pions correspondant à la fraction de traces ayant $E_1/E = 0$ (près de 16% d'après le tableau 3) nous a paru suspect. En effet, celui-ci semble être lié à la présence de bruit électronique dans les cellules du calorimètre. De plus, les traces ayant $E_1/E = 0$ avaient aussi des valeurs de $E_1(3 \times 1)$ et $E_1(15 \times 3)$ (cf. section 5.4.1 pour la définition de ces variables) égales à -9999, signe de l'absence de dépôt d'énergie autour de l'extrapolation dans le premier compartiment du calorimètre. Il est apparu que le code C++ implémenté dans le programme de reconstruction était mal protégé pour l'initialisation de certaines variables. Les corrections alors apportées ont permis de résoudre ce problème. Le tableau 3 donne la fraction des traces ayant des valeurs singulières pour E_1/E avant et après correction. Après correction, on voit que l'intervalle 0 n'est plus peuplé. La table 5 résume les

	avant correction		après correction	
	électrons	pions	électrons	pions
-9999	0.4 ± 0.1	3.1 ± 0.4	1.7 ± 0.2	19.7 ± 0.5
0	1.7 ± 0.3	16.5 ± 0.8	0	0

TAB. 3 – *Fractions d'électrons et de pions dans les 2 canaux singuliers pour la variable E_1/E. Dans le cas d'un évènement $J/\Psi \to e^+e^-$ avec bruit électronique.*

	avant correction		après correction	
	électrons	pions	électrons	pions
-9999	0.2 ± 0.1	3.2 ± 0.4	41.9 ± 0.4	31.9 ± 0.3
0	57.8 ± 0.6	40.1 ± 0.3	16.1 ± 0.5	11.4 ± 0.2

TAB. 4 – *Fractions d'électrons et de pions dans les 2 canaux singuliers pour la variable E_3/E. Dans le cas d'un évènement $J/\Psi \to e^+e^-$ avec bruit électronique.*

fraction de traces se retrouvant dans le découpage actuel des distributions des variables discriminantes. On notera que :

- le cas $n_{TR} > 33$ utilisé comme protection contre des valeurs aberrantes semble inutile ;
- les résultats pour cas $A_0 > 0.1$ masquent en fait une coupure sur la qualité des traces appliquée en amont de l'algorithme et qui impose justement un $A_0 < 0.1$. Cet intervalle est en fait utilisé uniquement comme protection pour le code C++.
- de E_T/p_T on déduit que seules 0.1% des traces d'électrons n'ont pas d'amas associé dans le calorimètre, contre 2% des traces de pions.
- Seules 1.5% des traces d'électrons n'ont pas d'amas associé dans la partie avant du calorimètre ($E_1/E = -9999$), contre près de 20% des traces de pions. La valeur singulière $E_1/E = 1$, impliquant que $E_2(3\times3) = E_3(3\times3) = 0$, est essentiellement due au bruit électronique et que l'on impose dans le code de reconstruction que si l'énergie d'un amas est trouvée négative (ce qui peut être le cas avec du bruit électronique), alors on impose que cette énergie soit nulle strictement.
  près de 40% des traces d'électrons et 30% des traces de pions ne déposent pas d'énergie dans le trosième compartiment du calorimètre ($E_3/E = -9999$).
  POS, W, ASY et E_1/E prennent leur valeur par défaut désormais dans la même proportion.

variables	intervalles	fraction d'électrons (%)	fraction de pions (%)
n_{TR}	$0 < ... < 33$	100	100
	> 33	0	0
A_0	$0 < ... < 0.1$	100	100
	> 0.1	0	0
E_T/p_T	0	0.1 ± 0.1	2.0 ± 0.1
	$0 < ... < 3$	99.6 ± 0.1	96.2 ± 0.1
	> 3	0.3 ± 0.1	1.8 ± 0.1
E_1/E	-9999	1.5 ± 0.2	19.7 ± 0.2
	0	0	0
	$0 < ... < 1$	98.0 ± 0.2	76.4 ± 0.2
	1	0.5 ± 0.1	3.6 ± 0.1
E_3/E	-9999	41.9 ± 0.4	31.9 ± 0.3
	0	16.1 ± 0.5	11.4 ± 0.2
	$0 < ... < 1$	41.9 ± 0.5	55.1 ± 0.2
	1	0.03 ± 0.1	1.6 ± 0.2
W	-9999	1.9 ± 0.2	25.0 ± 0.3
	$0 < ... < 1$	98.1 ± 0.2	75.0 ± 0.3
ASY	-9999	1.9 ± 0.2	25.0 ± 0.3
	$0 < ... < 1$	98.1 ± 0.2	75.0 ± 0.3
POS	-9999	1.9 ± 0.2	25.0 ± 0.3
	$0 < ... < 1$	98.1 ± 0.2	75.0 ± 0.3
ISO	-9999	0.5 ± 0.3	14.4 ± 0.3
	0	0	0
	$0 < ... < 1$	99.5 ± 0.3	85.6 ± 0.2

TAB. 5 – *Fractions d'électrons et de pions dans les différents intervalles.*

La figure 8 montre le découpage des distributions des variables discriminantes qui est maintenant proposé. Seules 14 sous classes restent - ce qui est encore beaucoup, mais surtout le découpage commence à être basé sur des explications physiques et non techniques.

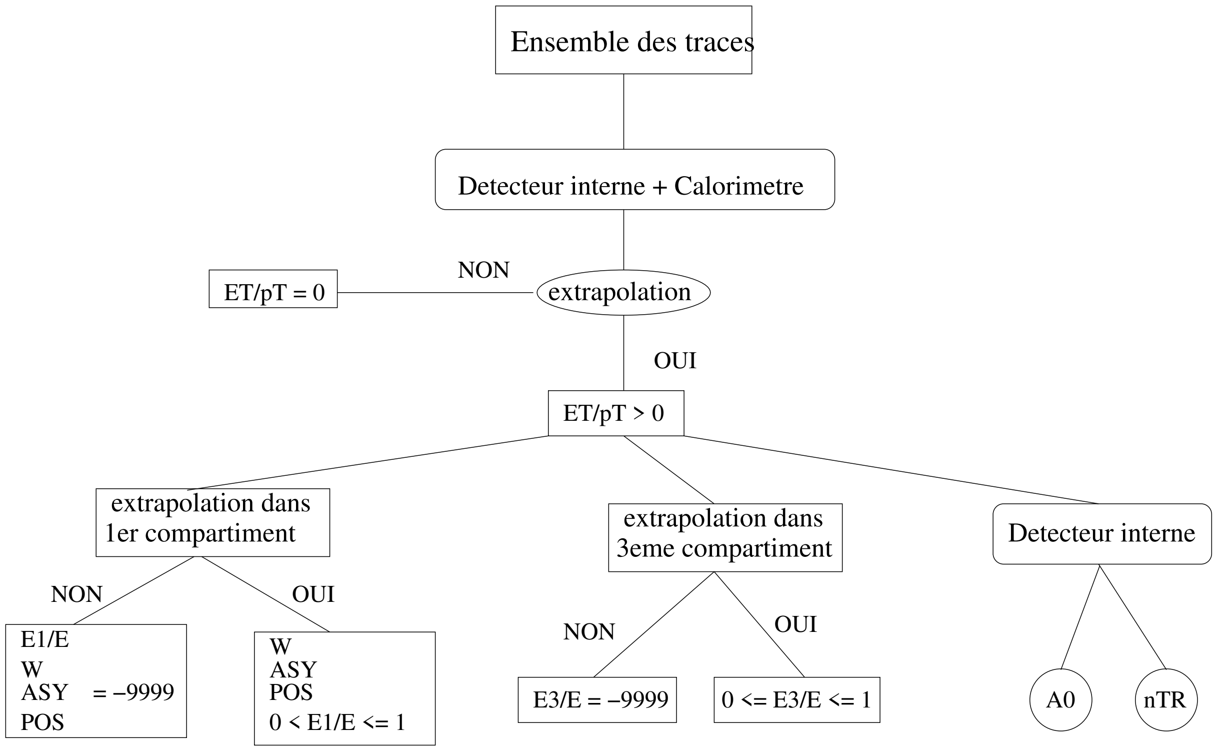

FIG. 8 – *Diagramme donnant la nouvelle décomposition des distributions des variables discriminantes.*

La suite de l'étude nécessitant l'estimation de la moyenne et de la dispersion de chaque distribution pour chaque variable, ou bien de son intégrale, on travaillera sur les traces telles que : $0 < E_T/p_T < 3$, $0 < E_1/E < 1$, $0 < E_3/E < 1$.

5.5.2 Estimation du pouvoir séparateur des variables

Afin d'estimer le pouvoir séparateur de chacune des variables discriminantes nous utilisons un estimateur appellé "séparation". L'utilisation de cette variable est courante dans une expérience comme Babar (étude de la violation des symétries CP). Nous nous inspirons ici de ces méthodes ([col99b], [Ver99]) très utile pour trouver et comparer différents jeux de variables discriminantes. La séparation est définie comme :

$$< s^2 >= \frac{1}{2}\int \frac{(g^b - g^s)^2}{g^b + g^s} dx, \tag{6}$$

avec x la variable discriminante pour laquelle on calcule la séparation et g^b et g^s sont des distributions normalisées à un ($\int g dx = 1$) respectivement pour le bruit de fond et le signal, les mêmes qu'utilisées dans l'équation 2 et dont on voit les formes dans la figure 4. La séparation $< s^2 >$ tend vers 0 si les distributions sont peu discriminantes et vers un si elles le sont.

Nous avons ainsi pu classer les variables par ordre croissant de pouvoir séparateur, comme le montre le tableau 6. On notera que la variable n_{TR} est la plus discriminante, ainsi que la variable E_T/p_T, c'est à dire les variables utilisant les informations du détecteur interne. D'autre part on notera que la variable d'asymétrie de la gerbe électromagnétique montre une séparation très faible. Ainsi, si on enlève la variable ASY de l'algorithme, les performances de l'algorithme restent similaires (cf. figure 6 à droite). Cette particularité avait déjà été notée puisque cette variable utilisée dans les premières versions de l'algorithme développé en 1998 [JS98], ne l'était plus dans [A.00]. Cette particularité est toujours valide avec la nouvelle implémentation, avec de nouvelles données. De plus on quantifie cette absence de pouvoir séparateur, ce qui n'avait jamais été fait. La variable ASY sera très certainement abandonnée.

Variables	n_{TR}	E_T/p_T	POS	E_1/E	E_3/E	W	A_0	ISO	ASY
séparation	0.613	0.565	0.346	0.279	0.243	0.161	0.159	0.076	0.042

TAB. 6 – *Séparation $< s^2 >$ pour chacune des variables. Les valeurs sont rangées par ordre décroissant.*

Dans un second temps, on combine la variable donnant la séparation la plus grande avec les $N-1$ restantes. On retient alors la combinaison optimale. La même procédure est alors appliquée à partir de cette combinaison et des $N-2$ variables restantes. Ainsi, la séparation tend à augmenter avec le nombre de variables ,si celles-ci ne sont pas trop corrélées. Le tableau 7 donne la séparation obtenue en conbinant la variable N_{TR} avec les 8 autres restantes. La combinaison optimale est alors $N_{TR} \oplus E_T/p_T$. La procédure est alors réitérée pour ce couple de variables. Les résultats sont contenus dans le tableau 8. A ce stade, la meilleur combinaison est $n_{TR} \oplus E_T/p_T \oplus POS$.

Cependant, les distributions des 8 variables[8] sont étudiées dans des histogrammes de 100 canaux. Ainsi, pour deux variables, on sépare chaque trace dans près de 100^2 intervalles. Ayant près de 25000 traces, la plupart des intervalles sont remplis[9]. Une possible optimisation du calcul en réduisant le nombre d'intervalles de chaque histogramme ou bien en utilisant non pas des histogrammes mais une fonction (comme une suite de polynômes) décrivant ce même histogramme n'a pas été entrepris.

Pour trois variables, les traces sont réparties dans près d'un million de cases, rendant difficile une intégration numérique (tout du moins sa précision). Ainsi, si les résultats semblent corrects à première vue, la séparation augmentant avec le nombre de variables, les combinaisons optimales correspondent à ce que l'on prévoyait, on ne peut cependant pas avoir une absolue confiance dans la valeur des séparations obtenues pour trois variables. Dès lors, les calculs pour quatre variables et plus semblent inutiles.

Variables	E_T/p_T	POS	E_1/E	E_3/E	A_0	W	ISO	ASY
séparation	0.788	0.705	0.686	0.676	0.656	0.654	0.632	0.623

TAB. 7 – *Séparation $< s^2 >$ pour chacun des variables combinées avec N_{TR}. Les valeurs sont rangées par ordre décroissant.*

5.5.3 Prise en compte des corrélations entre les variables

On a noté que la fonction de discrimination X_{PA} ne prenait pas en compte les corrélations entre les variables. De ceci résulte les grandes queues de distribution que l'on peut noter dans la figure 5.

[8] toutes sauf n_{TR} qui est découpée en 34 intervalles

[9] Ceci supposerai ten fait qu'il n'y ait pas de corrélation entre les variables et que l'histogramme à deux dimensions ainsi obtenu soit uniformément remplit.

Variables	POS	E_1/E	E_3/E	W	A_0	ISO	ASY
séparation	0.845	0.834	0.830	0.815	0.812	0.800	0.796

TAB. 8 – *Séparation $< s^2 >$ pour chacun des variables combinées avec $N_{TR} \oplus E_T/p_T$. Les valeurs sont rangées par ordre décroissant.*

Estimation de la matrice de corrélations La figure 9 montre des diagrammes à deux dimensions donnant une idée des corrélations entre certaines variables parmi les plus corrélées. Dans un premier temps nopus allons quantifier les corrélations entre les

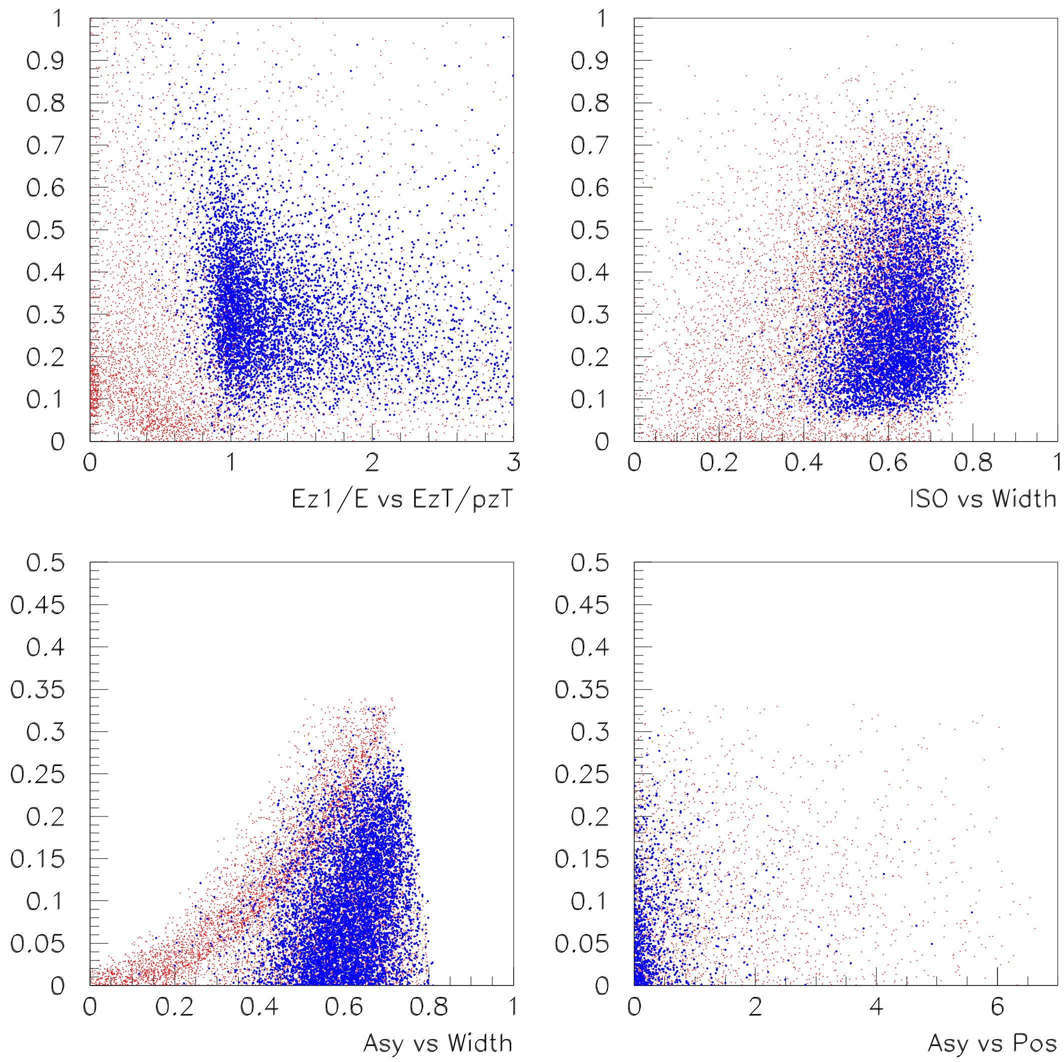

FIG. 9 – *Diagrammes montrant les distributions de E_1/E en fonction de E_T/p_T (en haut à gauche), ISO vs W (en haut à droite), ASY vs W (en bas à gauche) et ASY vs POS (en bas à droite). Les gros points représentent les valeurs obtenues pour les électrons, les petites points représentent les valeurs obtenues pour les pions.*

différentes variables. Pour ce faire nous allons construire la matrice de corrélation 9×9, pour le signal et le bruit de fond, dont chaque élément est donné par :

$$\rho_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{\mu i}-\bar{x}_\mu)(x_{\nu i}-\bar{x}_\nu)}{\sigma(x_\mu)\sigma(x_\nu)}, \tag{7}$$

avec $\bar{x}$ la moyenne de la distribution de la variable x, $\sigma(x)$ sa dispersion. La table 9 donne les valeurs de $\bar{x}$ et $\sigma(x)$ pour chaque variable pour les électrons et les pions. Les tables

	$\bar{x}^1$	$\bar{x}^2$	$\sigma(x)^1$	$\sigma(x)^2$
E_1/E	0.32	0.24	0.14	0.22
$E_3/3$	0.02	0.13	0.05	0.17
A_0	0.02	0.01	0.02	0.02
n_{TR}	8.14	2.11	3.79	1.97
W	0.61	0.50	0.08	0.18
ISO	0.28	0.31	0.14	0.19
ASY	0.10	0.11	0.07	0.08
POS	0.42	1.34	0.46	1.44
E_T/p_T	1.30	0.71	0.46	0.65

TAB. 9 – *Valeur moyenne et dispersion des distributions de chaque variable pour les électrons (1) et les pions (2).*

	E_1/E	E_3/E	A_0	n_{TR}	W	ISO	ASY	POS	E_T/p_T
E_1/E	1.000	−0.194	0.098	−0.100	0.141	0.281	−0.007	−0.012	−0.215
E_3/E	−0.194	1.000	0.022	0.021	−0.037	0.126	−0.019	0.057	0.321
A_0	0.098	0.022	1.000	−0.023	0.025	0.134	−0.015	0.044	0.038
n_{TR}	−0.100	0.021	−0.023	1.000	−0.068	−0.096	−0.046	−0.079	0.118
W	0.141	−0.037	0.025	−0.068	1.000	0.216	0.316	0.128	−0.037
ISO	0.281	0.126	0.134	−0.096	0.216	1.000	−0.032	0.264	0.334
ASY	−0.007	−0.019	−0.015	−0.046	0.316	−0.032	1.000	0.176	−0.040
POS	−0.012	0.057	0.044	−0.079	0.128	0.264	0.176	1.000	0.125
E_T/p_T	−0.215	0.321	0.038	0.118	−0.037	0.334	−0.040	0.125	1.000

TAB. 10 – *Matrice de corrélation pour le signal (électrons) dans les événements $H \to b\bar{b}$.*

10 et 11 donnent les matrices de corrélation pour le signal et le bruit de fond. On notera que :

- les variables n_{TR} et A_0 issues des informations du seul détecteur interne sont, comme attendu, très peu corrélées avec les variables issues des informations du calorimètre ;
- d'une manière générale les coefficients de corrélation sont faibles, ne dépassant pas $\sim 40\%$;
- des variables comme E_1/E et E_3/E sont assez fortement corrélées mais de manière similaire pour le signal et le bruit de fond ;
- des couples de variables comme (E_1/E,E_T/p_T) ou (E_3/E,E_T/p_T) présentent une corrélation importante pour le signal mais pas pour le bruit de fond.

C'est cette dernière remarque qui est peut être la plus importante, dans l'espoir qu'utiliser les corrélations entre les variables va apporter une meilleur séparation entre les fonctions de discrimination signal/bruit de fond et ainsi d'améliorer les performances de l'algorithme.

	E_1/E	E_3/E	A_0	n_{TR}	W	ISO	ASY	POS	E_T/p_T
E_1/E	1.000	−0.328	0.040	−0.057	0.263	0.375	0.052	0.170	0.037
E_3/E	−0.328	1.000	0.010	0.047	−0.173	−0.136	−0.046	−0.144	−0.071
A_0	0.040	0.010	1.000	0.007	0.079	0.143	0.004	0.041	0.141
n_{TR}	−0.057	0.047	0.007	1.000	0.013	−0.015	−0.011	−0.021	0.064
W	0.263	−0.173	0.079	0.013	1.000	0.293	0.416	0.203	0.286
ISO	0.375	−0.136	0.143	−0.015	0.293	1.000	0.030	0.248	0.323
ASY	0.052	−0.046	0.004	−0.011	0.416	0.030	1.000	0.050	0.011
POS	0.170	−0.144	0.041	−0.021	0.203	0.248	0.050	1.000	0.184
E_T/p_T	0.037	−0.071	0.141	0.064	0.286	0.323	0.011	0.184	1.000

TAB. 11 – *Matrice de corrélation pour le bruit de fond (pions) dans les événements* $H \to b\bar{b}$.

Estimation de la matrice de Fisher Des méthodes d'analyse prenant en compte les corrélations entre les variables sont couramment utilisées dans des expériences comme Babar (étude de la violation des symétries CP). Nous nous inspirons ici de ces méthodes ([col99b], [Ver99], [Piv03]).

Dans les méthodes d'analyses multidimensionnelles, l'ensemble initial de N variables discriminantes est combinée linéairement de manière à déduire la meilleure séparation entre les différentes classes d'événements (signal et bruit de fond). Le but de ces méthodes est de trouver un axe (un hyper plan à $N-1$ dimensions), dans l'espace à N dimensions des variables discriminantes, tel que les deux classes d'événements soient séparées au mieux.

On introduit les notations suivantes : $x^c_{\mu i}$, avec $i = 1, ...N$ est l'indice de l'événement, $\mu = 1, ...p$ est la composante du vecteur de variables discriminantes et $c = (1, ...q)$ est la classe à laquelle appartient l'événement (dans notre cas $q = 2$).

La projection du point i sur un axe $\vec{u} = (u_1, ...u_p)$ est une combinaison linéaire des p composantes du vecteur $\vec{x}$ de variables discriminantes : $\vec{u}_i = \sum_{\mu=1}^{p} u_\mu (x_{\mu i} - \bar{x}_\mu)$ On introduit la valeur moyenne $\vec{x}$ de la composante μ du vecteur et la valeur moyenne de la composante μ pour une classe donnée :

$$\bar{x}_\mu = \frac{1}{N} \sum_{i=1}^{N} x_{\mu i} \quad et \quad \bar{x}^c_\mu = \frac{1}{N_c} \sum_{i=1}^{N_c} x^c_{\mu i}, \tag{8}$$

où N_c est le nombre d'événements de la classe c.

Pour trouver l'axe de projection, on doit calculer la matrice de covariance du lot d'événements (matrice $p \times p$) :

$$T_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} (x_{\mu i} - \bar{x}_\mu)(x_{\nu i} - \bar{x}_\nu) \ . \tag{9}$$

En introduisant les deux classes d'événements, la matrice peut s'écrire (théorème de Huygens) :

$$T_{\mu\nu} = \frac{1}{N} \sum_{c=1}^{q} \sum_{i=1}^{N_c} (x_{\mu i} - \bar{x}^c_\mu) + \sum_{c=1}^{q} \frac{N_c}{N} (\bar{x}^c_\mu - \bar{x}_\mu)(\bar{x}^c_\nu - \bar{x}_\nu), \tag{10}$$

c'est à dire, en terme de matrices :

$$T = W + B. \tag{11}$$

La matrice W (pour "within") représente la dispersion à l'intérieur des classes, tandis que la matrice B ("between") représente la dispersion des classes par rapport au lot total d'événements.

Le critère de Fisher pour déterminer l'axe $\vec{u}$ consiste à maximiser le rapport $\vec{u}^T B\vec{u}/\vec{u}^T W\vec{u}$.

Dans le cas à deux classes (électrons et pions), on introduit :

$$\bar{x}_\mu = \frac{1}{N_1 + N_2}\sum_{c=1}^{2}\sum_{i=1}^{N_c} x_{\mu i}^c \quad et \quad \bar{x}_\mu^c = \frac{1}{N_c}\sum_{i=1}^{N_c} x_{\mu i}^c. \tag{12}$$

Les éléments de matrice sont donnés par :

$$W_{\mu\nu} = \frac{1}{N_1 + N_2}\sum_{c=1}^{2}\sum_{i=1}^{N_c}(x_{\mu i} - \bar{x}_\mu^c)(x_{\nu i} - \bar{x}_\nu^c) \quad et \quad B_{\mu\nu} = \frac{1}{N_1 + N_2}\sum_{c=1}^{2} N_c(\bar{x}_\mu^c - \bar{x}_\mu)(\bar{x}_\nu^c - \bar{x}_\nu).$$

Les coefficients de Fisher f_i pour chaque variable discriminante sont donnés par :

$$f_\mu = \frac{\sqrt{N_1 N_2}}{N_1 + N_2}\sum_{i=1}^{N_{var}} W_{\mu\nu}^{-1} \times (\bar{x}_{\mu i}^1 - \bar{x}_{\mu i}^2). \tag{13}$$

Le paramètre

$$f_0 = \sum_{\mu=1}^{N_{var}} f_\mu \times (\bar{x}_{\mu i}^1 + \bar{x}_{\mu i}^2)/2, \tag{14}$$

permet de centrer à zéro la variable de Fisher pour l'ensemble des deux lots.

La variable de Fisher X_{FI} pour chaque trace est donnée par :

$$X_{FI} \;=\; \sum_{\mu=1}^{N_{var}} f_\mu x_{\mu i}(FI) - f_0. \tag{15}$$

Validation de la méthode La méthode a été testée avec un "toy Monte-Carlo"[10] permettant de générer des distributions gaussiennes avec différents niveaux de corrélation. Nous avons engendré 10000 événements "signal" avec deux variables $X - Y$ selon des gausiennes $G_X(-1, 1)$ (de moyenne -1 et de largeur 1) et $G_Y(0.5, 2)$ et 10000 événements "bruit de fond" avec deux gausiennes $G_X(-5, 3)$ et $G_Y(2, 2.5)$. Deux cas ont été considérés : pas de corrélation entre $X - Y$ et un cas avec $\rho_{XY} = 0.9$ pour le signal et $\rho_{XY} = -0.7$ pour le bruit de fond. La figure 10 montre les distributions ainsi obtenues. La figure 11 montre les distributions ainsi obtenues pour X_{PA} et X_{FI} pour les deux lots d'événements (avec et sans corrélation). En particulier on peut remarquer qu'une fraction non négligeable de traces "bruit de fond" ont une valeur de X_{PA} proche de 1 et seront donc pris pour du signal. Cette fraction est encore plus important lorsqu'il y a du bruit de fond. Ceci peut introduire un biais important dans la pureté du lot de signal si le rapport signal/bruit est faible. La distribution de X_{FI} est elle bien distribuée autour de 0, avec des valeurs positives pour le signal et des valeurs négatives pour le bruit de fond. La figure 12 montre les performances ainsi obtenus par le rapport de vraisemblance X_{PA} et par la méthode de Fisher. On voit qu'en l'absence de corrélation le Fisher est inapproprié. En effet, on voit bien sur la figure 10 qu'alors il est très difficile de trouver un axe de projection qui

[10]courtoisement donné par J. Ocariz de la collaboration BaBar.

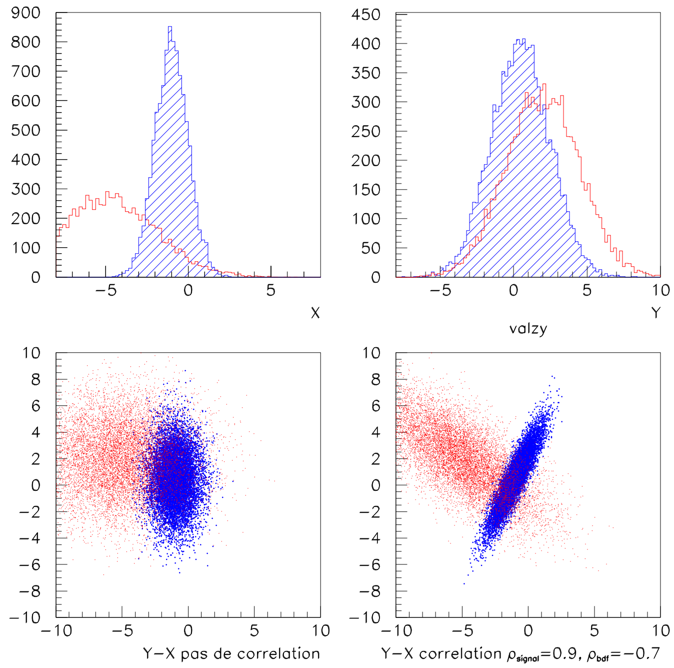

FIG. 10 – *Distributions gaussiennes issues de "toy Monte-Carlo" : nous avons engendré 10000 événements "signal" avec deux variables $X - Y$ selon des gausiennes $G_X(-1, 1)$ (de moyenne -1 et de largeur 1) et $G_Y(0.5, 2)$ et 10000 événements "bruit de fond" avec deux gausiennes $G_X(-5, 3)$ et $G_Y(2, 2.5)$. Deux cas ont été considérés : pas de corrélation entre $X - Y$ et un cas avec $\rho_{XY} = 0.9$ pour le signal et $\rho_{XY} = -0.7$ pour le bruit de fond.*

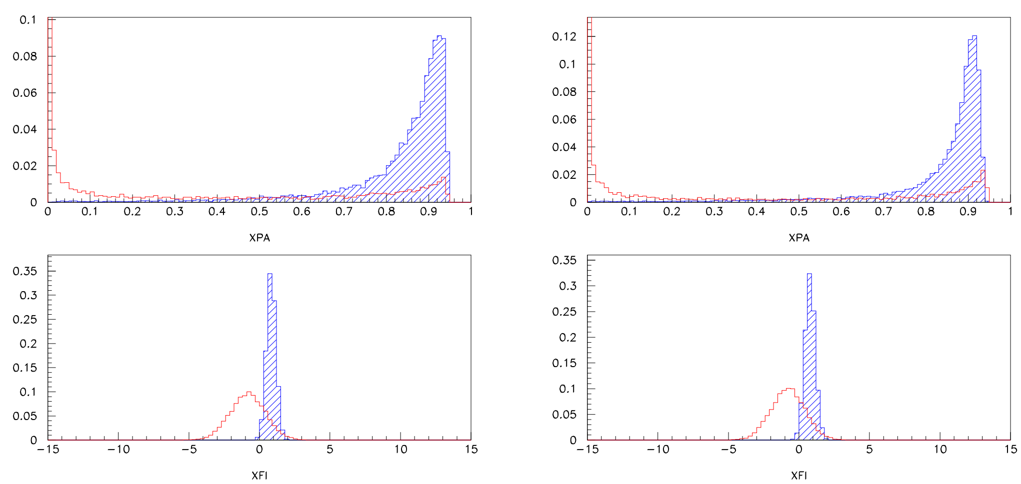

FIG. 11 – *Fonction discriminantes X_{PA} et X_{FI} obtenues dans le cas de la combinaison de deux gausiennes non corrélées (à gauche) et avec une grande corrélation (à droite).*

permette de séparer correctement à la fois les distributions signal/bruit de fond de X (bien séparée en une dimension) et la distribution de Y. Au contraire avec une grande corrélation les performances obtenues par les deux méthodes sont identiques. Par contre

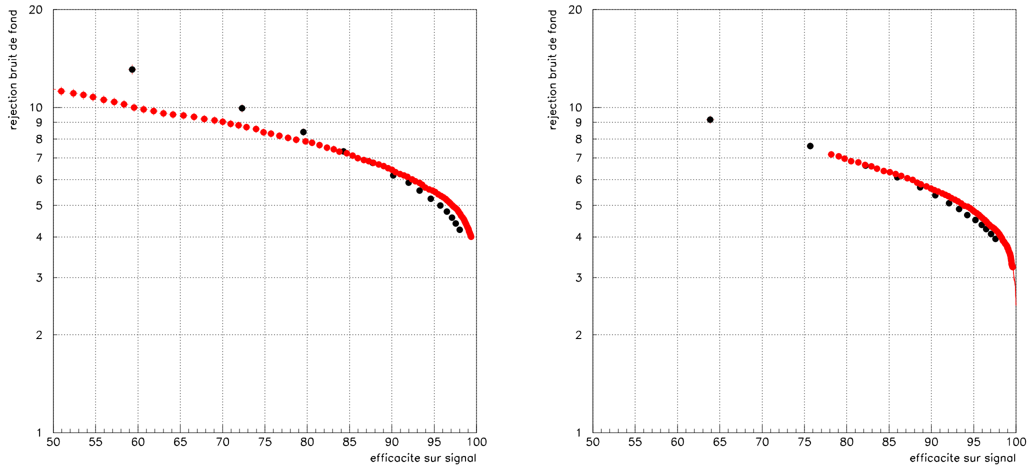

FIG. 12 – *Réjection du bruit de fond en fonction de l'efficacité d'identification du signal pour la combinaison de deux gausiennes non corrélées (à gauche) et avec une grande corrélation (à droite).*

la méthode de Fisher permet de remplacer deux variables très corrélées par une seule, ce qui peut être très profitable.

Résultats obtenus Effectuons quelques remarques générales sur la matrice W obtenue pour les événement $H \to b\bar{b}$ pour commencer :

- les valeurs N_1 et N_2 doivent être similaires. Ayant presque dix fois plus de traces de pions que d'électrons nous n'avons ainsi gardé qu'une trace de pions sur dix ;
- Vu la précision statistique avec laquelle nous travaillons (près de 8000 traces pour chaque catégorie d'événement), la matrice W est représentée avec trois chiffres significatifs après la virgule ;
- la variables A_0 étant peu corrélées avec les autres, il s'en suit que les éléments correspondants de la matrice inverse peuvent prendre des valeurs énormes. Il faudra donc exclure cette variable pour laquelle la méthode de Fisher ne peut s'appliquer.

	E_1/E	E_3/E	A_0	n_{TR}	W	ISO	ASY	POS	E_T/p_T
E_1/E	0.033	−0.007	0.000	−0.039	0.006	0.010	0.000	0.026	−0.004
E_3/E	−0.007	0.016	0.000	0.010	−0.003	−0.002	0.000	−0.017	0.000
A_0	0.000	0.000	0.000	−0.001	0.000	0.000	0.000	0.001	0.001
n_{TR}	−0.039	0.010	−0.001	9.270	−0.009	−0.028	−0.007	−0.100	0.146
W	0.006	−0.003	0.000	−0.009	0.019	0.006	0.004	0.028	0.016
ISO	0.010	−0.002	0.000	−0.028	0.006	0.027	0.000	0.042	0.031
ASY	0.000	0.000	0.000	−0.007	0.004	0.000	0.005	0.006	0.000
POS	0.026	−0.017	0.001	−0.100	0.028	0.042	0.006	1.127	0.099
E_T/p_T	−0.004	0.000	0.001	0.146	0.016	0.031	0.000	0.099	0.321

TAB. 12 – *Matrice de covariance W obtenue pour des événements $H \to b\bar{b}$.*

$$
\left(\begin{array}{cccccccccc}
 & E_1/E & E_3/E & A_0 & n_{TR} & W & ISO & ASY & POS & E_T/p_T \\
E_1/E & 39.63 & 13.68 & -10.69 & 0.06 & -7.94 & -14.47 & 4.01 & -0.18 & 2.35 \\
E_3/E & 13.68 & 68.81 & -13.99 & -0.01 & 5.81 & -2.92 & -1.41 & 0.70 & 0.06 \\
A_0 & -10.69 & -13.99 & 2761.74 & 0.20 & -4.94 & -32.86 & 8.96 & 0.08 & -4.79 \\
n_{TR} & 0.06 & -0.01 & 0.20 & 0.10 & 0.00 & 0.14 & 0.11 & 0.00 & -0.06 \\
W & -7.94 & 5.81 & -4.94 & 0.00 & 72.17 & -8.49 & -49.38 & -0.71 & -2.62 \\
ISO & -14.47 & -2.92 & -32.86 & 0.14 & -8.49 & 50.85 & 7.10 & -1.01 & -4.26 \\
ASY & 4.01 & -1.41 & 8.96 & 0.11 & -49.38 & 7.10 & 222.43 & -0.42 & 2.05 \\
POS & -0.18 & 0.70 & 0.08 & 0.00 & -0.71 & -1.01 & -0.42 & 0.97 & -0.17 \\
E_T/p_T & 2.35 & 0.06 & -4.79 & -0.06 & -2.62 & -4.26 & 2.05 & -0.17 & 3.77
\end{array}\right)
$$

TAB. 13 – *Matrice inverse de covariance W^{-1} obtenue pour des événements $H \to b\bar{b}$.*

Les tableaux 12 et 13 donnent les matrices W et W^{-1} ainsi obtenues. On pourra vérifier que (dans notre cas $N_1 = N_2$) :

$$W_{\mu\nu} = \frac{N_1}{N_1 + N_2} \left((\sigma^1_{\mu\nu})^2 + (\sigma^2_{\mu\nu})^2 \right). \tag{16}$$

Les combinaisons linéaires ainsi obtenues sont :

$$
\begin{aligned}
X_{FI}(ASY, W) &= 1.6 + 3.4W - 3.3ASY \\
X_{FI}(ASY, POS) &= -0.35 - 0.50ASY - 0.40POS \\
X_{FI}(ISO, W) &= 2.8 + 3.1W - 1.1ISO \\
X_{FI}(W, POS) &= 1.32 + 1.84 - 0.40POS \\
X_{FI}(E_1/E, E_T/p_T) &= -1.27 + 1.30E_1/E + 0.90E_T/p_T
\end{aligned} \tag{17}
$$

La figure 13 montre la comparaison entre les performances de l'algorithme en ne combinant que deux variables, soit par la méthode de vraisemblance soit par la méthode de Fisher On voit que les performances sont seulement similaires pour les couples (ASY, W), (ASY, POS). Par contre pour le couple (E_1/E, E_T/p_T) on obtient des performances bien pires avec le Fisher. On peut en effet noter que s'il y a bel et bien une corrélation de 30% pour le signal, les projections à une dimension des distributions de signal et du bruit de fond sont déjà bien séparées (ce que nous avions quantifier avec la variable $< s^2 >$). Ainsi, la seule valeur du coefficient de corrélation ne semble pas être un argument en soi, la valeur de la séparation rentrant aussi en compte.

Nous avons essayé de combiner les variables qui semblaient présenter le plus d'intérêt vis à vis de la méthode de Fisher. En combinant ASY, W et POS on obtient :

$$X_{FI}(ASY, W, POS) = 1.5 + 4.1W - 3.3ISO - 0.5ASY, \tag{18}$$

dont la distribution est donnée sur la figure 14. Les performances obtenues avec cette combinaison sont du même ordre que celles obtenus avec le rapport de vraisemblance obtenu à partir de ces mêmes variables (cf. Fig 14 à droite). Ainsi de trois variables corrélées nous pouvons extraire une variable unique.

Il n'a pas été possible d'aller plus en avant dans cette étude lors de ce stage. Néanmoins, de ces études préliminaires nous pouvons dresser le protocole suivant qui devra être suivi à l'avenir :

- prendre les N variables discriminantes ;

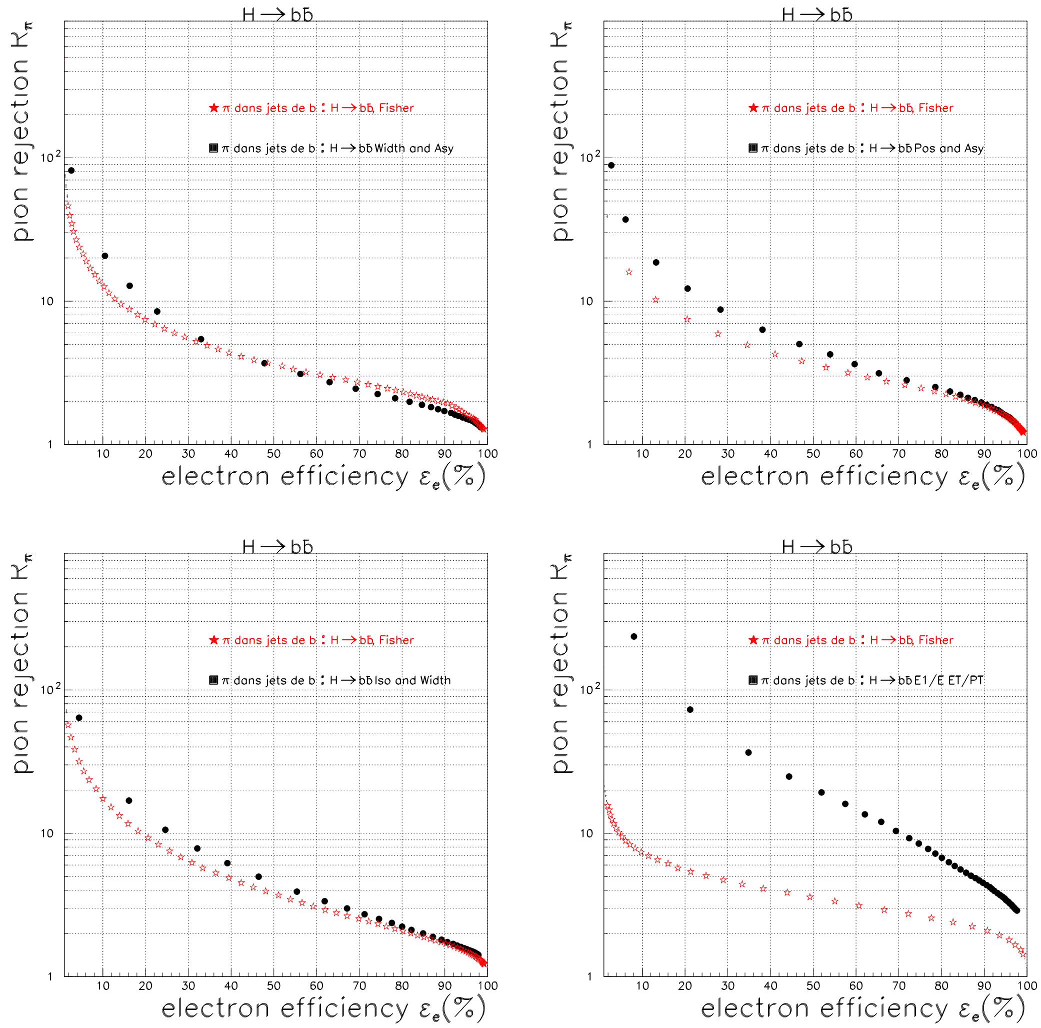

FIG. 13 – *Réjection des traces de pions en fonction de l'efficacité d'identification des électrons pour du $H \to b\bar{b}$. On présente les résultats obtenus par le rapport de vraisemblance X_{PA} ou par un Fisher X_{FI}. Les résultats sont présentés pour la combinaison (ASY, W) en haut à gauche, (ASY, POS) en haut à droite, (ISO, W) en bas à gauche et (E_1/E, E_T/p_T) en bas à droite.*

- estimer leur séparation $< s^2 >$;
- voir lesquelles possèdent de fortes corrélations ;
- combiner les variables présentant un intérêt[11] à l'aide de la méthode de Fisher ;
- construire un rapport de vraisemblance combinant les variables non corrélées et la variable X_{FI} ainsi obtenue ;

[11] Cette partie nécessitera certainement l'essai fastidieux des différentes combinaisons de variables

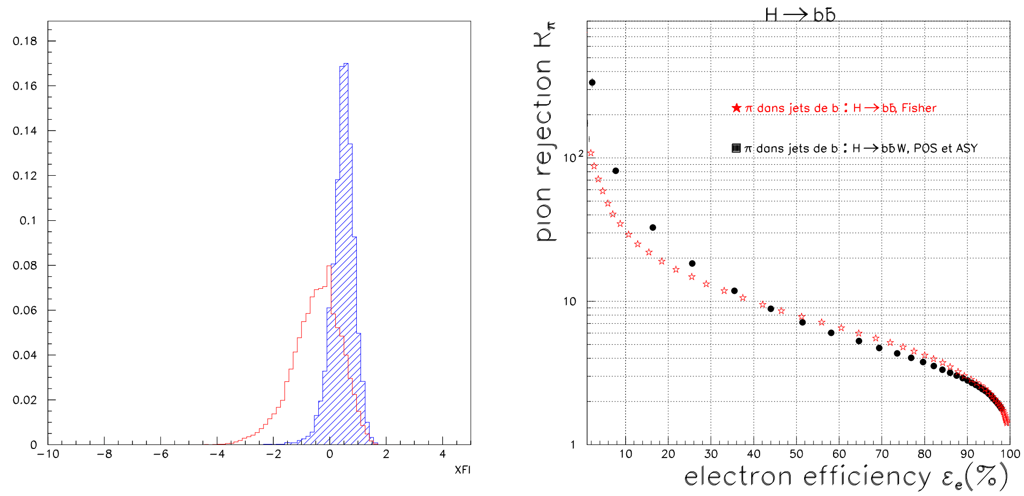

FIG. 14 – *Distribution de X_{FI} (à gauche) et réjection des traces de pions en fonction de l'efficacité d'identification des électrons pour du $H \to b\bar{b}$. On présente les résultats obtenus par le rapport de vraisemblance X_{PA} ou par un Fisher X_{FI}. Les résultats sont présentés pour la combinaison POS, ASY et W.*

6 Conclusion

Un algorithme pour l'identification des électrons de basses impulsions transverses, de fait non isolés dans des jets, a été récemment intégré au logiciel de reconstruction d'Atlas. Cet algorithme combine des informations issues du détecteur interne (impulsion transverse, nombre de coups dûs au rayonnement de transition dans le TRT ...) et du calorimètre électromagnétique (forme de la gerbe électromagnétique). Toutes ces variables discriminantes (un ensemble de neuf variables) sont combinées en une seule variable via la méthode du rapport de vraisemblance utilisant notre connaissance a-priori des formes des distributions attendues pour des électrons et des pions.

Dans un premier temps, lors de ce stage, nous avons étudié la manière dont ces distributions étaient découpées pour leur utilisation. De ces neuf variables discriminantes, un ensemble de 24 sous-classes permettaient de caractériser une trace. Outre le fait d'être non satisfaisante, cette classification n'était pas basée sur des arguments de classification physique mais sur des arguments purement techniques : "la variable prend (ou pas) sa valeur par défaut", plutôt que "il existe une correspondance entre une trace dans le détecteur interne et un amas dans le calorimètre électromagnétique". Des erreurs dans l'initialisation des variables dans le code de reconstruction ont été trouvées et une nouvelle classification, réduisant à 14 le nombre de sous-classes est présenté ici. Cette classification peut sans doute être encore simplifiée.

Le choix des variables discriminantes a été arrêté il y a plusieurs années et une phase de réoptimisation de l'algorithme va être à nouveau entreprise. Il est important de développer des outils permettant de quantifier le caractère discriminant (entre les hypothèses signal=électron et bruit de fond=pions) pour chaque variable. Ainsi avant d'être incorporée de nouvelles variables pourront ainsi être comparées à la brochette de variables actuelle. Une variable appelée "séparation" permet de quantifier ce pouvoir séparateur. Ainsi les variables les plus puissantes se trouvent être celles contenant les informations du détecteur interne (comme n_{TR}) et celles combinant les informations des deux sous-détecteurs (comme E_T/p_T). La variable dite d'asymétrie de la gerbe électromagnétique présente une séparation très faible et sa non-utilisation ne déteriore pas les performances de l'algorithme. Cette variable sera sans doute abandonnée par la suite.

La méthode de combinaison des variables discriminantes en une seule ne prend pas en compte les corrélations entre les variables. La méthode de Fisher permet de trouver un axe de projection dans un espace à N dimensions où l'ensemble des N variables discriminantes va se projeter avec la plus grande séparation possible. Cette méthode prend en compte les corrélations entre les variables. Tout d'abord les matrices de corrélations entre les variables, pour les électrons et les pions, ont été calculées pour la première fois. On retrouve bien évidemement le fait que les variables du détecteur interne et du calorimètre sont peu corrélées mais de plus que les coefficients de corrélations restent assez faibles (toujours moins que 30%).

Pour les variables les plus corrélées on montre que, si les performances ne sont pas améliorées par rapport à la méthode du rapport de vraissemblance, on ainsi la possibilité de remplacer un lot de deux variables corrélées par une seule. Le choix des variables sur lesquelles appliquer un Fisher n'est cependant pas aisé, puisque dans certains cas (comme pour le couple $E_1/E, E_T/p_T$), malgré une corrélation assez forte (30% pour les électrons), les distributions à une dimension sont déjà bien séparées et il n'y a pas d'axe de projection à deux dimensions privilégié. Ainsi la méthode de Fisher, utiliser "bêtement" peut même induire des pertes de performances.

L'utilisation de la méthode de Fisher est ainsi moins aisée qu'attendu naïvement. Elle peut cependant permettre, dans l'état actuel de notre compréhension, de remplacer trois des neufs variables (ASY, POS et W) par une seule tout en gardant des performances similaires. De plus amples études sont nécessaires pour mettre au point une procédure visant non pas à appliquer un rapprot de vraisemblance combinant les neuf variables, mais un rapport de vraisemblance combinant des variables non corrélées, les corrélations étant inclus dans des variables de Fisher, combinant de manière linéaire des variables qui, elles, sont corrélées.

Cette apparente difficulté est cependant le reflet d'un choix pertinent des variables discriminantes qui sont d'ores et déjà peu corrélées.

Références

[A.00] Kaczmarska A. Tagging b-jets using low-p_t electrons (full rapidity range), March 2000. ATLAS-PHYS-2000-012.

[col99a] ATLAS collaboration. Tdr atlas detector and physics performance, vol. 1. *CERN/LHCC 99-14, ATLAS-TDR-14*, 1999.

[col99b] Babar collaboration. The babar physics book. *SLAC-R-504*, 1999.

[DF03] Kaczmarska A. Derue F. Soft electron identification algorithm within athena, 2003. en préparation.

[ea86] R. Brun et al. Geant 3.21, 1986. CERN/DD/EE/84-1.

[gro] Athena group. Athena. http ://atlas.web.cern.ch/Atlas/GROUPS/SOFTWARE/OO/domains/Reconstruction/index.html.

[JS98] Wolter M. Jagielski S., Kaczmarska A. Tagging b-jets using low-p_t electrons, November 1998. ATLAS-PHYS-98-129.

[Piv03] M. Pivk. Etude de la violation de cp dans la désintégration b0-¿h+h- auprès du détecteur babar à slac, 2003. These de doctorat, Université Paris VII.

[Sch02] P. Schwemling. Les électrodes du calorimètre électromagnétique du détecteur atlas, 2002. These de doctorat, Université Paris VII.

[Sjo93] T. Sjostrand. Pythia 5.7 and jetset 7.4 physics and manual, Les Électrodes du Calorimètre Électromagnétique du détecteur Atlas - 1993. CERN/TH/7112/93.

[Ver99] S. Versille. La violation de cp dans babar : étiquetage des mésons b et étude du canal b-¿3 pi, 1999. These de doctorat, Université Paris XI.

[WM00] Kaczmarska A. Wolter M. Combining b-tagging methods using a neural network approach, November 2000. ATLAS-INDET-2000-023.