A statistical point of view on fatigue criteria : from supervised classification to positive-unlabeled learning

Olivier Coudray

université
PARIS-SACLAY

A statistical point of view on fatigue criteria: from supervised classification to positive-unlabeled learning

Un point de vue statistique sur les critères de fatigue : de la classification supervisée à l'apprentissage positif-non labellisé

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées
Graduate School : Mathématiques.
Référent : Faculté des sciences d'Orsay

Thèse préparée au sein du **Laboratoire de Mathématiques d'Orsay (LMO) (Université Paris-Saclay, CNRS)**,
sous la direction de **Christine KERIBIN**, Maîtresse de Conférences Hors-Classe, HDR,
le co-encadrement de **Patrick PAMPHILE**, Maître de Conférences,
la co-supervision de **Miguel DINIS**, Ingénieur Expert, Stellantis,
la co-supervision de **Philippe BRISTIEL**, Ingénieur Expert, Stellantis

Thèse soutenue à Paris-Saclay, le 8 décembre 2022, par

Olivier COUDRAY

Composition du jury

Membres du jury avec voix délibérative

Pascal MASSART Professeur, Université Paris-Saclay	Président
Didier CHAUVEAU Professeur, Université d'Orléans	Rapporteur & Examinateur
Adrien SAUMARD Professeur associé, École Nationale de Statistique et Analyse de l'Information (ENSAI), Bruz	Rapporteur & Examinateur
Anne GÉGOUT-PETIT Professeure, Université de Lorraine	Examinatrice

THESE DE DOCTORAT
NNT : 2022UPASM040

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Un point de vue statistique sur les critères de fatigue : de la classification supervisée à l'apprentissage positif-non labellisé

Mots clés: Critère de fatigue, Classification, Bruit d'étiquetage, Apprentissage positif-non labellisé, Bornes de risque.

Résumé: La fiabilité des véhicules est un enjeu majeur pour les constructeurs automobiles. En particulier, la fatigue mécanique est une préoccupation importante du bureau d'études. En effet, la fatigue est un phénomène complexe qui dépend du design de la pièce (géométrie, matériaux utilisés), des procédés de fabrication, et des chargements externes subis par la pièce. Le dimensionnement à la fatigue repose sur une modélisation numérique de la pièce et sur l'application de critères de fatigue déterministes afin d'identifier de potentielles faiblesses sur la conception. Ces critères, bien qu'efficaces sur des géométries simples, ne suffisent pas à prédire correctement les risques d'amorçage sur des composants complexes. Cela entraîne un allongement des temps de développement et une augmentation des coûts liés aux prototypes physiques. Pour y remédier, les constructeurs automobiles recherchent de nouvelles méthodes digitales, pour mieux identifier les zones critiques sur de nouvelles conceptions.

Dans cette thèse, nous construisons une base de données fatigue, à partir d'informations mises à disposition par Stellantis, regroupant des résultats numériques et des comptes rendus d'essais de fatigue. Une analyse non supervisée du jeu de données est réalisée, permettant de mieux comprendre sa structure ainsi que les liens entre les covariables disponibles. Ensuite, l'application de méthodes d'apprentissage supervisé (régression logistique, forêts aléatoires, SVM à noyau...) permet d'estimer des critères de fatigue offrant de meilleures prédictions que le critère mécanique déterministe usuel. Une difficulté de l'analyse provient du fait que l'étiquetage des zones est affecté par un bruit asymétrique, ce qui motive une approche originale fondée sur l'apprentissage positif-non labellisé (PU learning). Cette approche est abordée suivant tous les angles: théorique, méthodologique et appliqué. De nouvelles bornes de risques adaptées à ce cadre spécifique sont démontrées. Une méthodologie est proposée pour l'estimation d'un classifieur PU à partir des données. Enfin, la méthodologie est évaluée sur des jeux de données simulés ainsi que sur les données de fatigue. Les performances obtenues confirment l'intérêt de la méthode et son utilité pour le constructeur automobile.

Title: A statistical point of view on fatigue criteria: from supervised classification to positive-unlabeled learning

Keywords: Fatigue criterion, Classification, Label noise, PU learning, Risk bounds.

Abstract: The reliability of vehicles is a major issue for automotive manufacturers. In particular, mechanical fatigue is an important preoccupation of the design office. Indeed, fatigue is a complex phenomenon that depends on the design of the part (geometry, materials used), the manufacturing and on the external loads it is subjected to. In order to design a safety part against fatigue, the part is numerically modeled and a deterministic fatigue criterion is applied to identify potential weaknesses. If these criteria prove to be effective when evaluated on experimental test data with standardized specimens, they are less effective for rig tests with prototypes. This results in an increase in development costs and duration. In order to remedy this issue, car manufacturers seek new digital tools to better predict the fatigue risks on new design proposals.

In this thesis, we build a fatigue database, based on information provided by Stellantis, gathering numerical results along with fatigue test reports on prototypes. Unsupervised machine learning methods are applied offering a better understanding of the structure of the database and the relations between the available features. Then, the application of supervised machine learning methods (logistic regression, random forests, kernel SVM...) allows to estimate fatigue criteria offering better predictions than the standard fatigue criterion. However, the binary labels in this classification task are affected by a completely asymmetric label noise. This motivates an original approach to fatigue criteria estimation based on Positive-Unlabeled learning (PU learning). This problem is studied from all angles: theory, methodology and application. First, new risk bounds, adapted to this specific framework, are proved. Then, we develop a practical methodology to estimate a PU classifier. Finally, the methodology is evaluated on simulated data and on the fatigue database. The prediction performances confirm the interest of the methodology and its utility for car manufacturers.


université
PARIS-SACLAY
FACULTÉ
DES SCIENCES
D'ORSAY


MATHEMATIQUES
ORSAY


Inria


STELLANTIS


OPENLAB


Fondation mathématique
FMJH
Jacques Hadamard

Contents

Remerciements

Je tiens tout d'abord à adresser mes remerciements à Christine Keribin et Patrick Pamphile qui ont dirigé ma thèse avec brio. Je mesure l'honneur et la chance que j'ai eu de pouvoir apprendre et travailler à vos côtés durant ces trois années. Vos conseils, vos connaissances, vos intuitions et votre expérience m'ont guidé dans cette magnifique aventure qu'est la thèse. Christine, merci de m'avoir transmis ton exigence de rigueur, de clarté et de souci du détail notamment dans la rédaction. Patrick, merci pour ton investissement notamment sur les thématiques de fiabilité et de fatigue mécanique et pour m'avoir parfaitement aiguillé dans cette discipline qui m'était inconnue. Grâce à vous, j'ai la certitude qu'aucune erreur ou imprécision ne passera sous le radar. Vous avez tous les deux su m'épauler aussi bien sur le plan scientifique qu'humain. Merci pour votre soutien durant mes périodes de doute et pour votre disponibilité tout au long de la thèse : que ce soit le mercredi soir pour lire mes comptes rendus envoyés au dernier moment avant le point du jeudi matin, ou encore pendant les vacances d'été pour relire des chapitres de thèse !

Merci également à Gilles Celeux d'avoir participé à l'élaboration du projet de thèse et de nous avoir accompagné lors des premières réunions avec Stellantis (autrefois Groupe PSA). Même si les circonstances t'ont éloigné de l'encadrement, cette thèse n'aurait certainement pas vu le jour sans toi !

Cette thèse est le fruit d'un partenariat industriel avec Stellantis qui m'a permis de vivre une expérience à mi-chemin entre recherche académique et industrie. Je remercie chaleureusement Miguel Dinis et Philippe Bristiel qui ont assuré mon encadrement industriel. Votre complémentarité m'a permis de mener à bien ce projet. Miguel, merci d'avoir veillé au planning et à l'organisation de la thèse et pour tous tes apports sur la partie data qui m'ont aidé à avancer. Philippe, ton expertise en fatigue mécanique m'a été essentielle dans la compréhension des enjeux métiers de la thèse et dans l'analyse des données de fatigue. Je vous remercie tous les deux pour votre disponibilité chaque semaine pour notre rituel (pas toujours à la date et à l'heure prévues mais qu'importe !), et ce malgré vos emplois du temps extrêmement chargés.

Merci à Adrien Saumard et Didier Chauveau d'avoir accepté de rapporter ma thèse. Votre lecture attentive, vos commentaires et remarques m'ont été très précieux. Merci à Anne Gégout-Petit de me faire l'honneur de participer à mon jury de soutenance, trois ans après mon passage à l'IECL qui me semble déjà tellement loin. Je remercie également Pascal Massart d'avoir accepté de faire partie de mon jury. Merci pour ta disponibilité et tes précieux conseils au cours de la thèse qui m'ont remis sur de bons rails à un moment où je ne savais pas quelle direction prendre. Enfin, merci à Laurent Rota d'avoir accepté notre invitation au jury de thèse et d'avoir suivi régulièrement le projet à travers les fameux « points DIM ».

Je ne peux bien sûr pas évoquer Stellantis sans parler du réseau DIM (ex-PSA) qui a joué un rôle central dans le déroulement de cette thèse. Aussi, je tiens à remercier Matteo Luca

Facchinetti pour le soutien et l'aide qu'il a apporté à ce projet. Je remercie également l'ensemble du réseau DIM et notamment (en espérant oublier personne) Benoit Delattre, Ida Raoult, Laurent Rota, Pierre Osmond, Olivier Villars, Romain Hayat pour leur participation active aux échanges. Ces « points réseau DIM » ont considérablement alimenté mes réflexions et m'ont aidé à mieux cerner les enjeux métier de la thèse. Je remercie également Guy Martin Borret qui a aussi soutenu le projet auprès du périmètre Liaison Au Sol. Je remercie le réseau Data de Stellantis et notamment Guillaume Gruel et Mathieu Donain de m'avoir permis d'y présenter mes travaux. Ces présentations ont toujours donné lieu à des échanges fructueux qui ont bénéficié à l'avancée du projet.

Merci à mes collègues du CEMR à Poissy de m'avoir accueilli durant ces trois années. Même si le Covid nous a éloigné pendant presque deux ans, je retiens tous ces moments partagés dans l'open space, au café, au self, et lors de quelques sorties restaurant à Poissy. Les équipes étant en perpétuel remaniement, je me permets de remercier dans le désordre : Fabrice, Alexandre, Céline, Franck, Madjid, David, Olivier (D.), Gauthier, Nora, Thomas (les deux), Guillaume, Bastien, Nicolas.

Merci à la communauté des doctorants de Stellantis, en particulier à Emilien et Enora qui ont travaillé sur des sujets connexes. Je remercie également l'équipe d'animation pour son investissement dans l'accompagnement des doctorants, la vie de la communauté et la valorisation de nos activités auprès du groupe. Merci Jamila, Sandrine, Mathieu, Pascal. Merci à l'équipe Inria Celeste de m'avoir accueilli au LMO. J'ai également une pensée pour mes co-bureaux (et ex-co-bureaux) : Louise, Ruoci puis Lucas et Antoine. J'espère qu'ils trouveront un.e EDPiste pour me remplacer et corriger l'intrus que j'étais !

Je remercie mes amis, avec une pensée particulière pour la troupe du Michigan : Clément, Julien, Luyi. Merci Pierre, Hervé, Maxime (et Clément bien sûr) pour cette superbe semaine de randonnée dans les Vosges qui a été une véritable bouffée d'air au milieu des péripéties que nous ont réservé ces dernières années : Covid, confinements, ... J'en profite pour saluer les clubs de triathlon de Poissy et Sartrouville pour les entraînements, courses et autres événements festifs qui ont été aussi essentiels à mon bien-être.

Enfin, je remercie ma famille de m'avoir permis d'arriver jusqu'ici, en particulier mes parents, ma sœur, mes grands-parents. J'ai une pensée particulièrement émue pour Papé qui aurait certainement été fier de me voir arriver au bout de ces études, et qui, sans le savoir, a largement contribué à mon goût pour les mathématiques. Un immense merci à Ruihua qui a été d'un soutien infaillible au cours de ces trois ans et qui a su sans cesse trouver les mots pour me remotiver dans les moments durs.

Introduction générale (en français)

Au cours de sa durée de vie, la structure d'un véhicule est soumise à diverses contraintes mécaniques résultant des charges externes transférées par les roues et les suspensions à l'ensemble de la voiture. On distingue les charges statiques liées au poids du véhicule et à sa charge utile, et les charges dynamiques induites par les mouvements du véhicule. Ces charges dynamiques sont dues aux conditions de route (nids de poule, dos d'âne...) et aux actions du conducteur (freinage, accélération, virage...).

Après une longue durée d'utilisation, l'accumulation de contraintes combinées à des concentrations de contraintes (dues à la géométrie des composants mécaniques, aux procédés de fabrication...) peuvent provoquer l'amorçage de micro-fissures sur certaines zones du véhicule. Ces micro-fissures peuvent progressivement conduire à l'amorçage d'une macro-fissure qui se propage jusqu'à la rupture complète de la pièce mécanique. Ce phénomène, appelé *fatigue mécanique*, est extrêmement dangereux car il peut entraîner la défaillance brutale d'une pièce mécanique dans des conditions normales d'utilisation sans sollicitation excessive. La fatigue est donc un phénomène dangereux et complexe qui dépend du choix des matériaux, des procédés de fabrication et des contraintes locales lors de l'utilisation du véhicule (Schijve, 2005; Bathias and Pineau, 2010).

L'objectif du dimensionnement à la fatigue est donc de s'assurer que le véhicule répondra aux exigences en termes de performances mais aussi en termes de fiabilité et de durabilité : par exemple, une durabilité d'au moins 10 ans ou 100 000 kilomètres, représentant l'ordre de grandeur de la durée de vie d'une voiture concernée. La fiabilité du véhicule doit être assurée pour une grande variété de modèles et de conditions d'utilisation (conditions de route, sévérités clients...). Ceci est d'autant plus important pour les pièces de sécurité du véhicule où une défaillance peut avoir des conséquences dangereuses pour la sécurité des passagers : arbres d'essieu, système de direction, système de transmission, éléments de suspension, système de freinage...

Ainsi, l'optimisation de la conception des pièces mécaniques est devenue une préoccupation essentielle des constructeurs automobiles. En effet, l'objectif est de construire des véhicules plus légers afin de réduire leur consommation d'énergie tout en assurant leur fiabilité et leur durabilité. Par ailleurs, la conception doit être la plus rapide possible pour réduire les coûts de développement.

Le travail présenté dans cette thèse est le fruit d'une collaboration entre l'*Inria* (équipe *Celeste*) et la société *Stellantis* dans le cadre de l'*Openlab IA* avec le soutien financier de l'*ANRT* pour le contrat CIFRE n°2019/1131.

L'équipe de recherche *Celeste* de l'Inria travaille sur des sujets liés aux statistiques, à l'apprentissage automatique et à l'optimisation en mettant l'accent sur les liens entre théorie, algorithmes et applications.

Stellantis est un constructeur automobile né de la récente fusion entre le groupe italo-

américain *Fiat Chrysler Automobiles (FCA)* et le *Groupe PSA* (français). Plus précisément, ce travail de recherche a été mené au sein d'une équipe de Recherche et Développement de Stellantis travaillant à l'interface entre la modélisation numérique, l'optimisation et la science des données. Le projet est né d'un besoin exprimé par les équipes impliquées dans la conception et la validation des composants de la *Liaison Au Sol (LAS)* des véhicules. Il s'inscrit dans l'axe stratégique *full digital* visant à réduire le nombre d'essais physiques de validation en favorisant la modélisation numérique. Il a été mené en collaboration avec le *réseau DIM* de Stellantis rassemblant des experts en conception et dimensionnement mécanique. Ainsi, ce projet a bénéficié d'échanges fructueux avec divers acteurs de l'entreprise, en particulier des experts en dimensionnement à la fatigue, en modélisation numérique, en procédés de fabrication et en essais de validation.

Le dimensionnement à la fatigue des composants mécaniques chez Stellantis repose sur la méthode contrainte-résistance (cf. Thomas et al., 2005). Cette approche probabiliste prend en compte la diversité des usages clients et la dispersion des résistances, et vise à s'assurer que la résistance à la fatigue est suffisamment élevée compte tenu des choix de conception (matériaux, géométries des composants, procédés de fabrication...).

Le développement d'une pièce mécanique commence par une phase de conception où la géométrie et les matériaux sont définis. Ensuite, un modèle numérique de la pièce permet de simuler les contraintes sur la pièce sous chargement extérieur. Ces résultats numériques sont post-traités et un critère de fatigue est appliqué afin de localiser les faiblesses potentielles de la conception (par exemple le critère de Dang Van, cf. Ballard et al., 1995). Une fois la conception satisfaisante, des essais de fatigue sur prototypes réels sont réalisés pour valider la résistance à la fatigue et vérifier qu'elle répond aux exigences de durabilité (cf. Beaumont et al., 2012).

En raison de la complexité des pièces et du processus de développement, des problèmes peuvent survenir lors de la conception des composants du châssis du véhicule : par exemple, les tests de fatigue ne permettent parfois pas de valider les choix de conception. Par conséquent, la conception doit être corrigée et des tests physiques supplémentaires doivent être effectués. Les essais de fatigue sont particulièrement longs car ils nécessitent de soumettre plusieurs prototypes à des sollicitations cycliques répétitives sur une durée représentative de la durée de vie de la voiture. Plus concrètement, un seul essai peut durer plusieurs semaines et l'ensemble d'une campagne d'essais plusieurs mois. Par conséquent, les itérations entre validation et conception retardent considérablement le développement d'un véhicule. Dans ce contexte, l'objectif de Stellantis est de réduire drastiquement le nombre de tests et de tendre vers une conception entièrement digitale qui ne nécessiterait qu'une seule campagne d'essais de validation.

Les difficultés récurrentes de validation des composants LAS sont dues à de mauvaises corrélations entre les résultats de simulation numérique et les expérimentations physiques. En d'autres termes, le critère de fatigue appliqué aux résultats numériques ne permet pas d'identifier certaines faiblesses de la conception. Ces faiblesses sont alors découvertes lors des essais physiques. Cela nécessite une modification de la conception et une nouvelle campagne d'essais de validation. Pour réduire ces allers-retours entre conception et validation et accélérer le développement d'un véhicule, Stellantis s'intéresse à de nouvelles approches pour améliorer l'identification des zones critiques (zones à risques d'initiation de fissures) à partir du modèle numérique. L'enjeu est d'autant plus important que le passage de la voiture thermique à la voiture électrique nécessite la conception de nouvelles plateformes pour les véhicules. Dans ce contexte, il est crucial de capitaliser sur les expérimentations et les résultats acquis sur les véhicules thermiques pour ne pas repartir de zéro.

Au cours des dernières années, plusieurs projets de thèse ont été menés à Stellantis (ex-Groupe PSA) dans le cadre du dimensionnement à la fatigue, portant sur les protocoles d'essais de fatigue accélérés (cf. Beaumont, 2013), la caractérisation des propriétés de fatigue des soudures

(cf. Florin, 2015), la modélisation numérique des points de soudure électriques (cf. Mainnemare, 2021), et l'étude et la modélisation des contraintes résiduelles dues au procédé de soudage (cf. Tryla, 2022). Deux projets de thèse en cours portent sur l'amélioration de la méthode Contrainte-Résistance avec une description multivariée des sévérités de chargement en service (cf. Baroux et al., 2022) ; et l'analyse des spectres de chargement multiaxiaux et d'amplitude variable pour l'évaluation des dommages dus à la fatigue (cf. Bellec et al., 2022).

L'historique des modèles numériques et des rapports d'essais de validation de fatigue sur les conceptions précédentes représente une source de données riche et importante qui n'a pas été complètement exploitée jusqu'à présent. L'objectif de cette thèse est donc de mener une analyse exploratoire de cette base de données fatigue et de développer de nouveaux outils statistiques pour améliorer la prédiction des phénomènes de fatigue sur de nouveaux modèles numériques. Ces analyses statistiques peuvent offrir plusieurs avantages en complément de l'approche actuelle. Tout d'abord, ils s'appuient sur des données numériques et expérimentales de pièces mécaniques complexes, alors que les prédictions de fatigue classiques sont fondées sur des critères de fatigue calibrés par des essais sur éprouvettes. Les éprouvettes utilisées consistent généralement en des géométries simples qui ne sont pas nécessairement représentatives de la diversité et de la complexité des pièces mécaniques réelles. Ensuite, l'approche proposée peut prendre en compte des descripteurs physiques et géométriques supplémentaires à ceux considérés dans les critères de fatigue traditionnels (triaxialité, gradients de contraintes, caractéristiques descriptives des singularités...). En outre, l'analyse statistique peut guider le choix des variables appropriées pour prédire les risques de d'amorçage sur les pièces mécaniques. Enfin, les modèles statistiques peuvent rendre compte de la dispersion des résultats d'essais de fatigue, ce qui est essentiel lorsque l'on s'intéresse au phénomène de fatigue.

Ces axes offrent un fort potentiel d'amélioration des prédictions du critère de fatigue actuellement mis en œuvre. Un critère de dimensionnement à la fatigue amélioré devrait mieux identifier les zones critiques sur une nouvelle conception numérique et aider à anticiper les problèmes de validation. Cela se traduirait par moins d'itérations entre conception et essais de validation et donc un développement accéléré du véhicule.

D'un point de vue statistique, l'estimation d'un critère de fatigue peut être vue comme une tâche de classification qui pose des problèmes originaux du fait de la nature spécifique des données (conditions expérimentales, essais interrompus). Lors des essais, la présence d'amorçage de fissures assure la criticité d'une zone. Cependant, l'absence de fissure n'est pas une preuve de sécurité : une fissure aurait pu amorcer sous une sévérité plus élevée ou si l'essai avait été prolongé. Ce problème est lié à l'*apprentissage positif-non labellisé* (apprentissage PU), un cadre de classification semi-supervisé où seul un sous-ensemble d'observations positives est étiqueté (cf. Bekker and Davis, 2020). En particulier, l'apprentissage PU sous l'hypothèse *Selected At Random* (SAR, *i.e.* lorsque le sous-ensemble des instances étiquetées est affecté par un biais de sélection) n'a reçu que peu d'attention dans la littérature (Bekker and Davis, 2018b; He et al., 2018; Gong et al., 2021). Une étude théorique de l'apprentissage PU sous l'hypothèse SAR pourrait apporter un nouvel éclairage sur la problématique industrielle et guider le choix de méthodologies appropriées pour la résoudre.

Les paragraphes suivants donnent un bref aperçu de la structure globale de ce manuscrit, organisé en quatre chapitres.

Le chapitre 1 introduit le phénomène de fatigue mécanique et les modèles classiques de la littérature pour représenter et caractériser les risques de fatigue. Une distinction importante est faite entre les modèles de durée de vie en fatigue et les critères de fatigue : les premiers cherchent à modéliser la durée de vie en fatigue d'une pièce mécanique soumise à un chargement répété tandis que les seconds s'attachent à prédire si une pièce satisfait ou non aux exigences de durabilité et consistent donc en des prédictions binaires. Même si les objectifs poursuivis sont différents, les deux types de modèles sont étroitement liés. Le chapitre explique ensuite précisément le rôle

des modèles de fatigue dans le dimensionnement de pièces mécaniques automobiles à la fatigue. En particulier, les rôles de la modélisation numérique et des essais de fatigue sont détaillés, conduisant à la formulation des enjeux industriels et des objectifs de cette thèse : une approche statistique pour améliorer l'identification des défauts de conception sur un modèle numérique et ainsi aider à réduire les itérations entre conception et validation.

Le chapitre 2 présente la base de données de fatigue Stellantis, construite à partir de résultats numériques sur d'anciennes conceptions combinés à des rapports d'essais de fatigue expérimentaux. Constatant le déséquilibre important entre le nombre d'observations avec et sans initiation de fissure, une méthode est proposée pour changer l'unité d'analyse en considérant des zones (groupes d'éléments) au lieu d'éléments individuels. Cela permet de réduire considérablement le déséquilibre. De plus, un ensemble de variables descriptives appropriées est introduit, comprenant des variables standards en fatigue (invariants de contrainte) mais aussi de nouveaux descripteurs spécifiques aux zones (moyennes spatiales d'invariants de contrainte) et aux singularités (soudures et bords de tôles). Une analyse non supervisée de la base de données est effectuée, mettant en évidence des corrélations importantes entre les variables et simultanément une structure entre les individus (zones). Puis, passant à la caractérisation des risques de fatigue, une version probabiliste du critère de fatigue de Dang Van est proposée, permettant d'estimer et de prendre en compte la dispersion du phénomène de fatigue dans un cadre multiaxial. Ce critère est estimé et validé au travers de résultats d'essais de fatigue sur éprouvettes soudées mais montre ses limites lorsqu'il est étendu à des pièces mécaniques complexes de la base de données fatigue. Par conséquent, la construction d'un critère de fatigue est reformulée comme une tâche de classification supervisée, et des techniques d'apprentissage automatique standards sont appliquées. Ces critères de fatigue basés sur la classification peuvent prendre en compte tous les descripteurs disponibles dans la base de données et conduisent à de meilleures performances en prédiction. Il s'agit donc d'une première réponse aux enjeux industriels de la thèse.

Cependant, les techniques de classification standard ignorent un mécanisme de bruit d'étiquette affectant les observations binaires. Si les amorçages de fissures assurent la présence de défauts de conception sur les pièces mécaniques, l'absence de rupture par fatigue n'est pas une preuve de sécurité. Ainsi, certaines zones critiques des pièces mécaniques ne sont pas détectées lors des essais (et donc non étiquetées) du fait par exemple d'une sévérité trop faible ou d'un essai interrompu trop tôt. Ainsi, la construction d'un critère de fatigue consiste en une tâche spéciale de classification semi-supervisée appelée *apprentissage positif-non labellisé* (apprentissage PU).

Le chapitre 3 présente l'apprentissage PU et sa spécificité par rapport à la classification standard. Après une revue bibliographique des approches et méthodologies existantes adaptées à ce cadre de classification, le chapitre se concentre sur une analyse théorique de l'apprentissage PU sous l'hypothèse Selected At Random (SAR). Cette hypothèse stipule que le bruit d'étiquetage affectant l'observation peut dépendre de covariables. De nouvelles bornes de risque supérieures et inférieures sont fournies, soulignant l'impact du bruit d'étiquetage sur les taux de convergence des classifieurs. Ces taux de convergence sont illustrés empiriquement par des expériences numériques.

Le chapitre 4 traite de l'application pratique de l'apprentissage PU pour estimer un critère de fatigue. La construction d'un critère de fatigue est exprimée sous la forme d'une tâche d'apprentissage PU sous l'hypothèse SAR et l'ensemble des variables influençant le bruit d'étiquetage sont identifiées. Nous proposons un modèle paramétrique adapté à l'application fatigue, et l'identifiabilité de ce modèle est discutée. Ensuite, une méthodologie basée sur l'algorithme *Expectation-Maximization* (EM) (cf. Dempster et al., 1977) est développée pour estimer les paramètres du modèle. L'intérêt de cette méthodologie est illustré par des expérimentations numériques sur des données simulées. Enfin, la méthode est appliquée à la base de données fatigue de Stellantis, fournissant un nouveau type de critère de fatigue.

Pour résumer, les principales contributions de cette thèse sont les suivantes :

1. Une présentation originale des modèles de fatigue mettant en évidence les liens entre la

fatigue du point de vue de la fiabilité et de la durabilité (cf. Chapitre 1). Cette connexion est illustrée sur la figure 1.7 et ouvre la voie à des approches fondées sur la classification pour construire de nouveaux critères de fatigue.

2. Une version probabilisée du critère de fatigue de Dang Van utilisé pour l'identification des zones critiques sur une conception (cf. Chapitre 2). La nouveauté de ce critère réside dans sa calibration, permettant d'estimer jointement les paramètres matériau et la dispersion du critère multiaxial.

3. Une analyse exploratoire de la base de données fatigue de Stellantis à l'aide de techniques classiques d'apprentissage supervisé et non supervisé (cf. Chapitre 2). Une définition de zones est proposée pour remédier à l'important déséquilibre du jeu de données et faciliter les analyses statistiques. L'analyse non supervisée permet de mieux comprendre la diversité des conditions de contraintes et des singularités géométriques sur des pièces mécaniques complexes, et comment elles peuvent être caractérisées avec des variables appropriées. Des critères basés sur les données de fatigue peuvent ensuite être estimés à l'aide de méthodes de classification supervisée. L'originalité de ces critères de fatigue est leur capacité à prendre en compte de nombreux descripteurs en plus de ceux traditionnellement considérés, conduisant à de meilleures performances en prédiction.

4. Une nouvelle approche basée sur l'apprentissage PU pour construire un critère de dimensionnement à la fatigue (Chapitres 3 et 4). Comme les essais de fatigue ne permettent d'identifier qu'un sous-échantillon de zones critiques, la méthode de classification doit tenir compte de cette source de bruit d'étiquetage asymétrique dans l'estimation du critère. Les contributions dans ce domaine sont doubles :

   (a) Une étude théorique de l'apprentissage PU sous l'hypothèse Selected At Random (SAR), c'est-à-dire lorsque la probabilité qu'une instance positive demeure non étiquetée dépend de ses covariables. Nous avons prouvé de nouvelles bornes de risque pour l'apprentissage PU, soulignant comment le taux de convergence dépend du bruit d'étiquetage (Chapitre 3).

   (b) Le développement d'une méthodologie pratique pour estimer un classifieur PU adapté au problème de dimensionnement à la fatigue. La méthodologie a été illustrée sur des données simulées et appliquée à la base de données Stellantis (Chapitre 4).

Certaines des contributions ci-dessus font partie de publications soumises au cours de la thèse :

- une version préliminaire de l'analyse non supervisée du chapitre 2 (Coudray et al., 2020b) complétée par l'application de méthodes de classification standard pour l'estimation de critères de fatigue (cf. Coudray et al., 2020a, présentée à la conférence *Lambda-Mu*) ;

- une analyse spécifique des soudures de la base de données de fatigue où des indicateurs supplémentaires sont calculés pour améliorer les prédictions de fatigue (cf. Coudray et al., 2021, présenté à la conférence *SIA Simulation*) ;

- la présentation des résultats théoriques du chapitre 3 (cf. Coudray et al., 2022a, soumis au *Journal of Machine Learning Research*, JMLR). Des simulations numériques illustrant ces résultats théoriques ont été présentées à la "*Conférence pour l'Apprentissage automatique*" (cf. Coudray et al., 2022b).

General introduction

Over its service life, the structure of a vehicle is subjected to various mechanical stresses resulting from external loads transferred by the wheels and suspensions to the whole car. A distinction is made between static loads related to the weight of the vehicle and its live load, and dynamic loads induced by the vehicle's motions. These dynamic loads are due to the road conditions (potholes, humps...) and to the driver's actions (braking, acceleration, turn...).

After a long duration of use, the accumulation of stresses combined with stress concentrations (due to the geometry of mechanical components, manufacturing processes...) can cause the initiation of micro-cracks on certain zones of the vehicle. These micro-cracks can progressively lead to the initiation of a macro-crack which propagates until the complete fracture of the mechanical component. This phenomenon, called *mechanical fatigue*, is extremely dangerous as it can lead to the sudden failure of a mechanical part in normal conditions of use without any excessive load. Fatigue is thus a dangerous and complex phenomenon depending on the choice of materials, the manufacturing processes, and the local stresses during the use of the vehicle (Schijve, 2005; Bathias and Pineau, 2010).

The objective of fatigue design is thus to make reasonably sure that the vehicle will meet the requirements in terms of performance but also in terms of reliability and durability: for instance, a durability of at least 10 years or 100 000 kilometers, representing the order of magnitude of a relevant car's lifetime. The reliability of the vehicle must be ensured for a large variety of models and usage conditions (road conditions, customer severities...). This is even more crucial for safety parts of the vehicle where a failure can have dangerous consequences regarding the security of passengers: axle shafts, steering system, transmission system, suspension components, braking system...

Therefore, optimizing the conception of mechanical parts has become an essential preoccupation of car manufacturers. Indeed, the objective is to construct lighter vehicles in order to reduce their energy consumption while ensuring their reliability and durability. At the same time, the conception should be as fast as possible to reduce the development costs.

The work presented in this thesis is the result of a collaboration between *Inria* (*Celeste* team) and the company *Stellantis* in the framework of the *Openlab AI* with the financial support of the *ANRT* for the CIFRE contract n°2019/1131.

Celeste research team at Inria works on subjects related to statistics, machine learning and optimization with a focus on the relations between theory, algorithms, and applications.

Stellantis is a car manufacturer born from the recent merger between the Italian-American group *Fiat Chrysler Automobiles (FCA)* and the French *Groupe PSA*. More precisely, this research work was conducted within a Research and Development team at Stellantis working at the interface between numerical modeling, optimization, and data science. The project emerged

as a need expressed by teams involved in the design and validation of chassis components of vehicles. It is part of the *full digital* strategic axis aiming at reducing the number of physical validation tests by fostering numerical modeling. It was led in collaboration with the *DIM network* at Stellantis gathering experts in mechanical design and structural integrity. Thus, this project has benefited from fruitful exchanges with diverse actors in the company, including experts in fatigue design, numerical modeling, manufacturing processes, and validation tests.

The fatigue design of mechanical components at Stellantis relies on the Stress-Strength method (cf. Thomas et al., 2005). This probabilistic approach considers the diversity of customer usage and the dispersion of resistances, and aims at ensuring that the fatigue resistance is high enough given the conception choices (materials, geometries of the components, manufacturing processes...).

The development of a mechanical part begins with a conception phase where the geometry and materials are defined. Then, a numerical model of the part allows to simulate the stresses on the part under external loads. These numerical results are post-processed and a fatigue criterion is applied in order to locate potential weaknesses in the conception (for example Dang Van criterion, cf. Ballard et al., 1995). Once the conception is satisfying, fatigue tests on real prototypes are carried out to validate the resistance to fatigue and check that it meets the durability requirements (cf. Beaumont et al., 2012).

Due to a huge and complex framework, project roadblocks may occur when it comes to designing chassis components of the vehicle: for instance, the fatigue tests sometimes fail to validate the design choices. Therefore, the conception needs to be corrected, and additional physical tests must be performed. Fatigue tests are particularly long as they require subjecting multiple prototypes to repetitive cyclic loads over a duration representing the objective lifetime of the car. More concretely, a single test can last several weeks and a whole test campaign several months. Consequently, numerous iterations between validation and conception strongly delay the development of a vehicle. In this context, the objective of Stellantis is to drastically reduce the number of tests and tend towards a *full digital* design that would require only one final validation test.

The recurring difficulties in validating chassis components are due to poor correlations between numerical simulation results and physical experiments. In other words, the fatigue criterion applied to the numerical results fails to identify some weaknesses of the conception. These weaknesses are then discovered during the physical tests. This requires a modification of the conception and another validation test campaign. To reduce design loops and accelerate the development of a vehicle, Stellantis seeks new approaches to improve the identification of critical zones (zones with crack initiation risks) on a numerical model. The issue is all the more critical as the transition from thermal to electric cars necessitates the design of new vehicle platforms. In this context, it is crucial to capitalize on the experiments and results acquired on thermal vehicles in order not to restart from scratch.

Over the past years, multiple thesis research projects were conducted at Stellantis (ex-Groupe PSA) in the context of fatigue design, addressing accelerated fatigue test protocols (cf. Beaumont, 2013), the characterization of fatigue properties of welded joints (cf. Florin, 2015), the numerical modeling of spot welds (cf. Mainnemare, 2021), and the study and modeling of residual stresses due to welding processes (cf. Tryla, 2022). Two ongoing thesis projects focus on improving the Stress-Strength method with a multivariate description of in-service stress severities (cf. Baroux et al., 2022); and the analysis of multiaxial and variable-amplitude load spectra for fatigue damage assessments (cf. Bellec et al., 2022).

The history of numerical models and fatigue validation tests reports about previous designs represents a rich and sizeable source of data that was not completely exploited so far. The

objective of this thesis is thus to conduct an exploratory analysis of this fatigue database and develop new statistical tools to improve the prediction of fatigue phenomena on new numerical models. These statistical analyses can offer several advantages in addition to the current approach. First, they rely on numerical and experimental data from complex mechanical parts, whereas the classical fatigue predictions are based on fatigue criteria calibrated through coupon tests. The coupon specimens used usually consist in simple geometries that are not necessarily representative of the diversity and complexity of real mechanical parts. Second, the proposed approach can account for additional physical and geometric descriptors to those considered in traditional fatigue criteria (triaxiality, stress gradients, descriptive features on singularities...). Third, the statistical analysis can guide the choice of appropriate features to predict fatigue risks on mechanical parts. Finally, statistical models can account for the dispersion of fatigue test results which is essential when dealing with fatigue phenomena.

These axes offer great potential for improving the predictions of the fatigue criterion currently implemented. An improved fatigue design criterion should better identify critical zones on a new numerical conception and help anticipate validation issues. This would result in fewer iterations between conception and validation tests and thus an accelerated development of the vehicle.

From a statistical point of view, the estimation of fatigue criteria can be viewed as a classification task that raises original issues due to the specific nature of the data (experimental conditions, interrupted tests). During tests, the presence of crack initiation asserts the criticality of a zone. However, the absence of crack is not an evidence of safety: a crack could initiate under higher severity or if the test was extended. This issue is connected to *Positive Unlabeled learning* (PU learning), a semi-supervised classification framework where only a subset of positive observations is labeled (cf. Bekker and Davis, 2020). In particular, PU learning under the *Selected At Random* assumption (SAR, *i.e.* when the set of labeled instances is affected by a selection bias) has received only few attention in the literature (Bekker and Davis, 2018b; He et al., 2018; Gong et al., 2021). A theoretical study of PU learning under the SAR assumption could provide a new point of view on the industrial issue and guide the choice of appropriate methodologies to solve it.

The following paragraphs provide a brief overview of the global structure of this manuscript, organized in four chapters.

Chapter 1 introduces the mechanical fatigue phenomenon and the classic models from the literature to represent and characterize fatigue risks. An important distinction is made between fatigue lifetime models and fatigue criteria: the former seeks to model the fatigue lifetime of a mechanical part subjected to repetitive loads while the latter focuses on predicting whether or not a part satisfies the durability requirements and thus consists in binary predictions. Even if the objectives pursued are different, the two types of models are closely connected. The chapter then precisely explains the role of fatigue models in designing automotive mechanical parts against fatigue. In particular, the roles of numerical modeling and fatigue tests are detailed, leading to the formulation of the industrial issues and objectives of this thesis: a statistical data-based approach to improve the identification of design flaws on numerical conceptions and thus to help reduce design loops.

Chapter 2 presents Stellantis fatigue database, constructed upon past results from numerical models combined with experimental fatigue test reports. Noting the severe imbalance between the number of observations with and without fatigue crack initiation, a method is proposed to change the unit of analysis by considering zones (groups of elements) instead of individual elements. This allows to significantly reduce the imbalance. In addition, a set of appropriate descriptive features is introduced, including standard variables in fatigue (stress invariants) but also new descriptors specific to zones (spatial averages of stress invariants) and to singularities (welds and edges). An unsupervised analysis of the database is carried out, highlighting important correlations among features and simultaneously a structure among individuals (zones). Then, moving to the characterization of fatigue risks, a probabilistic version of Dang Van fatigue

criterion is proposed allowing to estimate and account for the dispersion of fatigue phenomenon in a multiaxial framework. This criterion is estimated and validated through fatigue test results on welded coupon specimens but shows its limits when applied to complex mechanical parts of Stellantis fatigue database. Hence, the construction of a fatigue criterion is reformulated as a supervised classification task, and standard machine learning techniques are applied. These classification-based fatigue criteria can account for all the descriptors available in the database and lead to improved prediction performances. Therefore, this is a first answer to the industrial issues of the thesis.

However, standard classification techniques ignore a label noise mechanism affecting the binary observations. If crack initiations assert the presence of design flaws on mechanical parts, the absence of fatigue failure is not evidence of safety. Therefore, some critical zones of mechanical parts may remain undetected at testing (and thus unlabeled) if the severity was not high enough or if the test was interrupted too soon. This means that the construction of a fatigue criterion consists in a special semi-supervised classification task called *Positive-Unlabeled learning* (PU learning).

Chapter 3 introduces PU learning and its specificity compared to the standard classification setting. After a bibliographic review of existing approaches and methodologies adapted to PU learning, the chapter focuses on a theoretical analysis of PU learning under the Selected At Random (SAR) assumption. This assumption states that the label noise affecting the observation can depend on covariates. New upper and lower risk bounds are provided, highlighting the impact of the label noise on the convergence rates of classifiers. These convergence rates are empirically illustrated in numerical experiments.

Chapter 4 deals with the practical application of PU learning to estimate a fatigue criterion. The construction of a fatigue criterion is expressed as a PU learning task under the SAR assumption and the set of variables impacting the label noise are identified. We propose a parametric model adapted to the fatigue application, and the identifiability of this model is discussed. Then, a methodology based on the *Expectation-Maximization* (EM) algorithm (cf. Dempster et al., 1977) is developed to estimate the model's parameters. The interest of this methodology is illustrated through numerical experiments on simulated data. Finally, the method is applied to Stellantis fatigue database, providing a new type of fatigue criterion.

To outline, the main contributions of this work are the following:

1. An original presentation of fatigue models highlighting the links between fatigue under the two points of view of reliability and durability (cf. Chapter 1). This connection is illustrated in Figure 1.7 and paves the way for classification-based approaches to construct new fatigue criteria.

2. A probabilistic version of Dang Van fatigue criterion used in the identification of critical zones on a design proposal (cf. Chapter 2). The novelty of this criterion lies in its calibration, allowing to jointly estimate the material parameters and the dispersion of the multiaxial criterion.

3. An exploratory analysis of Stellantis fatigue database using classical supervised and unsupervised machine learning techniques (cf. Chapter 2). A construction of zones is proposed to remedy the severe imbalance of the data set and facilitate statistical analyses. The unsupervised analysis allows to better understand the diversity of stress conditions and geometric singularities on complex mechanical parts, and how they can be characterized with appropriate descriptive features. Fatigue data-based criteria can then be estimated using supervised classification methods. The originality of these fatigue criteria is their ability to account for many features in addition to those traditionally considered, leading to better predictive performances.

4. A new approach based on PU learning to construct a fatigue design criterion (Chapters 3 and 4). As fatigue tests only allow the identification of a sub-sample of critical zones, the classification method needs to account for this source of asymmetric label noise in the estimation of the criterion. The contributions in this domain are twofold:

   (a) A theoretical study of PU learning under the Selected At Random (SAR) assumption, *i.e.* when the probability for a positive instance to remain unlabeled depends on its covariates. We provided new risk bounds for PU learning, highlighting how the convergence rate depends on the amount of label noise (Chapter 3).

   (b) The development of a practical methodology to estimate a PU classifier adapted to the fatigue design problem. The methodology was illustrated on simulated data and applied to Stellantis database (Chapter 4).

Some of the above contributions are part of publications submitted during the thesis:

- a preliminary version of the unsupervised analysis of Chapter 2 (Coudray et al., 2020b) completed by the application of standard classification methods to the estimation of fatigue criteria (cf. Coudray et al., 2020a, presented at *Lambda-Mu* conference);
- a specific analysis of welds of the fatigue database where additional features are computed to improve the fatigue predictions (cf. Coudray et al., 2021, presented at *SIA Simulation* conference);
- the presentation of the theoretical results of Chapter 3 (cf. Coudray et al., 2022a, submitted at the *Journal of Machine Learning Research*, JMLR). Numerical simulations illustrating these theoretical results were presented at the "*Conférence pour l'Apprentissage automatique*" (cf. Coudray et al., 2022b).

1

Fatigue design of automotive chassis parts

During the vehicle design, the manufacturer needs to ensure that the mechanical components are able to resist the in-service loads they will be subjected to over the car usage. To do so, engineers have to choose appropriate materials and geometries to meet these requirements. Among the different existing failure modes, fatigue is of critical importance. Fatigue affects parts whose mechanical characteristics are modified after a repetition of loadings: it is a wear phenomenon. Fatigue fracture is thus very dangerous because it can happen when a part is subjected to loads below its mechanical resistance. Besides, the fatigue phenomenon is very hard to appraise: a numerical simulation is usually unable to fully characterize the fatigue resistance of a mechanical part. Indeed, numerical models rely on simplifications (*e.g.* geometric simplifications due to meshing, "nominal" material parameters used, poor load representativeness). Therefore, the design against fatigue always requires an experimental validation to ensure that a given part is robust enough.

Section 1.1 introduces the fatigue phenomenon in mechanics and explains why it is difficult to characterize. In Section 1.2, we introduce classic fatigue models in the literature to characterize the fatigue risks. Section 1.3 presents the methodology Stress-Strength method to design and validate complex mechanical parts. Its implementation in the context of fatigue design at Stellantis is also detailed. Finally, in Section 1.4, we present the main issues of this thesis and introduce our approach to solve these challenges.

1.1 - Fatigue of materials

The resistance of a mechanical parts to external loads depends on the material and on the intensity of the force it is subjected to. Multiple failure modes exist (Subsection 1.1.1) including fatigue. Fatigue is a particularly complex failure mode because it is the result of repeated loads and usually happens after a long duration (Subsection 1.1.2). Besides, the fatigue resistance depends on many parameters which makes the characterization of fatigue risks on complex parts very challenging (Subsection 1.1.3).

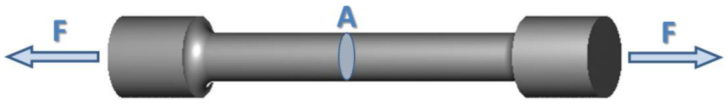

Figure 1.1: Coupon traction test.

1.1.1 Strength of materials

In mechanics of materials, the strength of a given material is usually studied through a traction test (cf. Fortunier, 2001). This test consists in applying a given force F on a sample of the material. The sample used is an elementary geometry called coupon specimen (cf. Fig. 1.1). If A denotes the *area* of the section of the specimen, the *local stress* σ at the center of the specimen is:

$$\sigma = \frac{F}{A} \ .$$

For a given value of F (and thus σ), the induced deformation (*strain* ε) of the specimen is measured:

$$\varepsilon = \frac{l - l_0}{l_0}$$

where l_0 is the nominal length of the specimen and l is the length under load F.

Remark. During the traction test, the area of the section A changes due to *Poisson effect*: an elongation of the specimen results in a decrease in the section area. This phenomenon is not of primary importance here. Usually, we consider that:

$$\sigma = \frac{F}{A_0}$$

where A_0 is the nominal section of the specimen.

The stress as a function of the strain is represented on a stress-strain diagram: the curve characterizes the behaviour of the material (cf. Fig. 1.2). If the stress σ exceeds a certain limit called *ultimate strength* (σ_u), the specimen breaks directly. Below this ultimate strength, the *elastic limit* σ_y defines a boundary between plastic and elastic deformations. In the plastic domain (*i.e.* for σ between σ_y and σ_u), the deformations are not reversible, meaning that once the force F is set back to 0, the specimen does not get back to its initial state. A repetition of such loads can entail a mechanical damage and result in breaking the specimen. For stresses below the elastic limit σ_y, the deformation is said to be elastic: under this regime, the strain can usually be modeled as a linear function of the stress. Besides the deformation is reversible, meaning that once the force F is set back to 0, the specimen gets back to its initial state.

Even under repeated elastic stresses, a specimen can eventually break. The reason is that the deformation is not fully reversible as there can be plastic deformations at the microscopic scale. This creates micro-cracks on the specimen that propagate and merge to form a macro-crack, hence leading to the failure of the specimen (Schijve, 2009; Bathias and Pineau, 2010). This failure mode is called *fatigue*. A material is thus characterized by an additional limit σ_e called *fatigue limit* or *endurance limit*. This limit represents the stress under which the lifetime of the specimen is infinite. This concept remains quite theoretic. For practical purposes, the endurance limit is usually associated to a lifetime sufficiently high considering the application (*e.g.* 10^6 cycles, cf. Subsection 1.2.1).

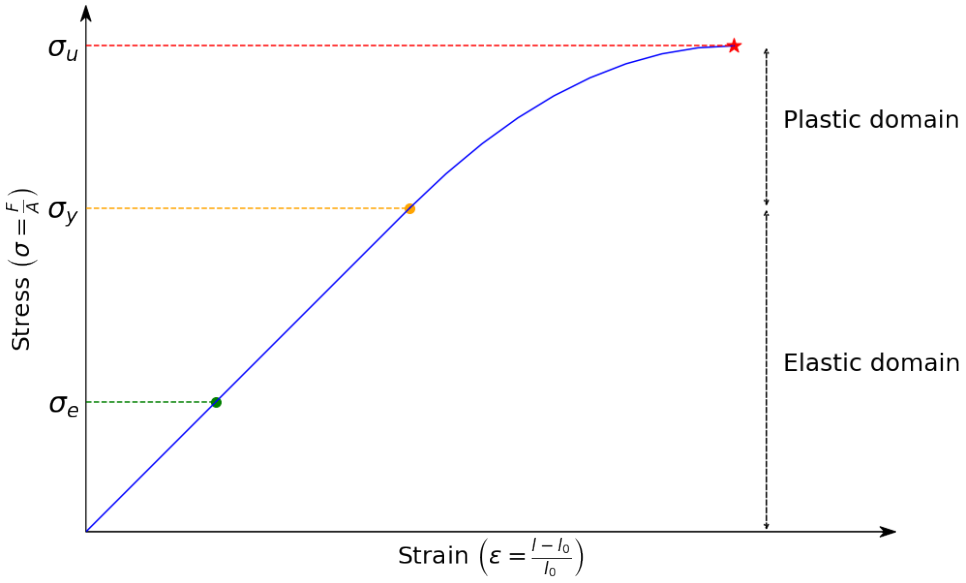

Figure 1.2: Stress-strain diagram: the blue curve characterizes the mechanical properties of a material

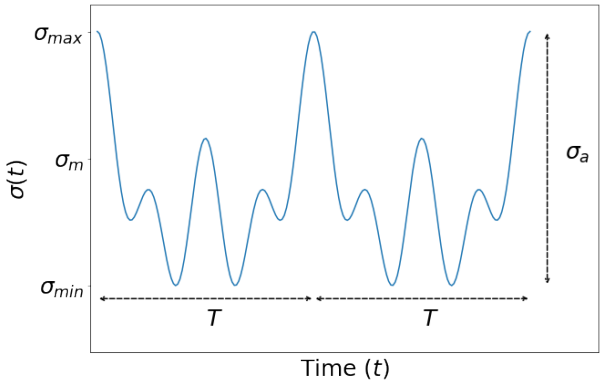

Figure 1.3: Example of cyclic loading.

1.1.2 Definition of fatigue

Fatigue is a mechanical failure mode for a specimen subjected to cyclic loading. This means that the loading is time-dependent and possibly periodic. Considering the illustrative example of Subsection 1.1.1, we now assume that the loading F is a T-periodic function of time $t \mapsto F(t)$. The local stress at the center of the specimen is also time-dependent and T-periodic:

$$\sigma(t) = \frac{F(t)}{A} \ .$$

Usually $(\sigma(t))_{0 \leq t < T}$ is summarized by its maximum and minimum values σ_{max} and σ_{min} which allow to compute the stress mean and amplitude (cf. Fig. 1.3):

$$\sigma_m = \frac{1}{2}\left(\sigma_{max} + \sigma_{min}\right)$$
$$\sigma_a = \left(\sigma_{max} - \sigma_{min}\right)$$

Another useful parameter is the *load ratio* R defined as:

$$R = \frac{\sigma_{min}}{\sigma_{max}} \ .$$

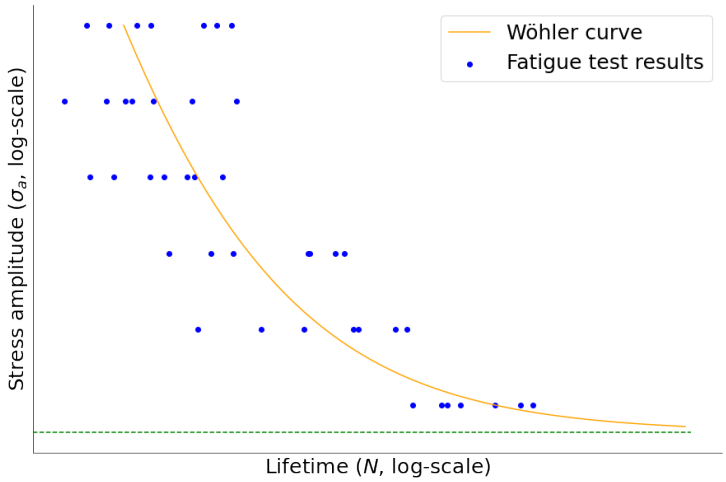

Figure 1.4: Example of S-N diagram with artificial data. The dotted green line represents the endurance limit σ_e.

A particular case is when the mean stress is 0 and thus $R = -1$ (*fully reversed*).

The *lifetime* N of a specimen is defined as the number of load cycles the specimen can endure before breaking. Fatigue is commonly divided into two sub-fields. On the one hand, the *Low Cycle Fatigue* (LCF) applies to situations where the stresses are above the elastic limit, thus inducing non-reversible macroscopic plastic deformations. In this case, the lifetimes considered are usually below 10^4 cycles (Schijve, 2009). On the other hand, *High Cycle Fatigue* (HCF) covers the elastic domain and lifetimes usually superior to 10^4 cycles. In the scope of this thesis, we are interested in HCF as the objective is to ensure the resistance of mechanical parts over the lifetime of the car, evaluated as 10^6 cycles. Under the load intensities considered, there is no macro-plastic deformation as the stresses are below the elastic limit σ_y.

The fatigue lifetime N is usually divided in two phases: first, the initiation of a macro-crack on the specimen (N_i); then, its propagation until the complete failure of the specimen (N_p). In the context of fatigue design in the automotive industry, N_p is usually small compared to N_i. Besides, the objective is to ensure the absence of crack initiation over the vehicle's lifetime; contrary to aeronautics where crack propagation is part of a comprehensive design and maintenance framework. Hence, N_i is the quantity of interest and the lifetime N considered is the number of cycles before observing a visible macroscopic crack.

The fatigue resistance of a material is experimentally characterized by performing tests on a series of coupon specimens subjected to a sinusoidal load with different intensities. Usually the mean stress is fixed and only the amplitude stress σ_a changes. Multiple specimens are tested for different levels of stress: for each test, the experimental lifetime N of the specimen is obtained. The results are represented in an *S-N diagram* (cf. Fig. 1.4) with the lifetime N on the x-axis and the stress amplitude σ_a on the y-axis (usually with logarithmic scale on both axes). The *Wöhler curve* (Schijve, 2009; Bathias and Pineau, 2010) represents the mean fatigue resistance as a function of the lifetime N and characterizes the fatigue properties of the material. Test results usually exhibit an important dispersion, especially in the HCF domain. Subsection 1.2.1 will provide standard models to build Wöhler curves taking into account the randomness of crack initiation.

1.1.3 Parameters influencing the fatigue resistance of metallic materials

Even considering a uniaxial coupon test (elementary and homogeneous geometry) in a controlled environment, the experimental fatigue results are scattered (cf. Fig. 1.4). The reason is that the micro-structure parameters (grain size and shapes, cristallographic structure of the material, inclusions...) have an impact on the fatigue resistance. Hence, even two macroscopically identical specimens have a different microscopic arrangements of grains which leads to different lifetimes under an identical loading. As these parameters are usually impossible to control, it is important to correctly account for the dispersion of fatigue lifetimes.

The fatigue strength of a specimen also depends on macroscopic parameters characterizing the specimen and the stresses it is subjected to.

- The average lifetime of a specimen is a decreasing function of the stress amplitude σ_a. The mean stress σ_m also has an influence on the fatigue lifetime (cf. Dowling, 2004). More generally, when considering complex mechanical parts under more complex loading, the stress is usually neither uniaxial nor univariate. The fatigue properties therefore depend on the triaxiality of the stress.

- The fatigue phenomenon also depends on the geometry of a specimen. First, the distribution of local stresses on a mechanical part depends on its geometry. Second, geometric singularities like holes, corners or notches induce stress concentrations that can accelerate or delay fatigue crack initiations. Third, fatigue depends on the size of the specimen (cf. Sun et al., 2016). Indeed, fatigue is a *weakest link mechanism*: a part is as weak as its weakest element, therefore the probability of having a weaker element increases with the size of the specimen. This is true as far as the stress is uniformly distributed on the specimen, but this rationale is also applicable to more complex cases.

- The material and the manufacturing process have a significant impact on the fatigue resistance. As previously mentioned, the Wöhler curve representing the fatigue properties of a specimen depends on the material it is made of. Besides, the transformations applied to the material prior to or during the assembling of the part (manufacturing process) can modify its fatigue properties. Manufacturing processes (welding, stamping, machining, shot peening, ...) introduce either micro-structural modifications altering the fatigue resistance of the material or residual stresses that need to be considered in addition to the load stresses (cf. Godefroid et al., 2014; Jimenez-Martinez, 2020).

- The environment interactions also have an impact on fatigue phenomenon (*e.g.* corrosion, cf. Schijve, 2009, Chapter 16).

All in all, fatigue is a specific failure mode in mechanics occurring under repeated cyclic loads of moderate intensity (stresses below the elastic limit). Accounting for fatigue risk in the design of automotive parts is essential as the vehicle is subjected to external loads over its lifetime and must resist without any crack initiation on safety parts. The fatigue properties of a material can be characterized experimentally using Wöhler curves. These studies are however limited to simple specimens (coupon specimens) under uniaxial loading conditions and in controlled environments. Indeed, the fatigue lifetime of a specimen depends on many parameters including the stresses, the geometry, the material and the manufacturing processes. Even after accounting for all these parameters, the fatigue phenomenon remains scattered because it also depends on the specific micro-structure of a specimen.

1.2 - Modeling fatigue risks

In this section, we give an overview of standard fatigue models. We can classify these models in two groups serving different purposes. On the one hand, S-N models consist in characterizing the lifetime of a specimen as a function of the stress. On the other hand, fatigue criteria seek to characterize the fatigue endurance of a specimen, *i.e.* the stress conditions under which the fatigue lifetime is infinite. The two types of models are however interconnected. On the one hand, the calibration of multiaxial fatigue criteria usually relies on multiple S-N curves. On the other hand, multiaxial fatigue criteria can be extended to the finite lifetime domain: instead of characterizing the endurance limit of a material, the fatigue criteria characterizes the fatigue limit for a fixed finite lifetime $N_0 < +\infty$.

Subsection 1.2.1 presents standard S-N models to characterize the distribution of the lifetime of a specimen as a function of the stress under uniaxial loading conditions. Under multiaxial loads, the stress can no longer be represented by a univariate feature σ. Subsection 1.2.2 presents the concept of stress tensor and shows how it can be used to characterize multiaxial stress conditions through stress invariants. In Subsection 1.2.3, we present the general principles of multiaxial fatigue criteria. Finally, Subsection 1.2.4 focuses on the definition of Dang Van fatigue criterion and the calibration of its parameters.

1.2.1 S-N fatigue models

A S-N model is a statistical model adjusting the Wöhler curve, *i.e.* the fatigue lifetime of a specimen subjected to a uniaxial cyclic stress. The mean stress σ_m is assumed to be fixed and thus, only the effect of the stress amplitude σ_a is accounted for. The general form of an S-N model is the following:

$$\log(N)\,|_{\sigma_a} = g(\sigma_a) + \varepsilon \ . \tag{1.1}$$

The regression function g usually consists in a parametric model describing the Wöhler curve. The noise ε represents the randomness of fatigue failures. Different choices for each part of the model are provided in Paragraphs a and b. Parameters of a S-N model are estimated on fatigue test results, hence providing an estimate of the Wöhler curve and the dispersion of fatigue test results.

a. Regression models for S-N curves

Different parametric models exist in the literature to model the relation between the mean fatigue lifetime N and the stress amplitude σ_a.

The standard fatigue model called *Basquin model* assumes a linear relation between the logarithm of the lifetime $\log(N)$ and the logarithm of the stress amplitude $\log(\sigma_a)$ (cf. Basquin, 1910):

$$g(\sigma_a) = -b\,\log\left(\sigma_a\right) + c \tag{1.2}$$

where b and c are parameters characterizing the material and the type of test. In particular, the regression coefficient b is called *Basquin slope* and is a crucial fatigue parameter. Basquin model is popular for its simplicity and its efficiency in the domain of HCF with limited endurance (for stresses above the endurance limit σ_e). One major limit of this model is that it does not assume the existence of an endurance limit: according to Basquin model, infinite lifetime is reached when the stress amplitude σ_a tends toward 0.

Other models have been developed to integrate the existence of a fatigue limit. Some consider an additional unknown parameter σ_e representing the endurance limit (Stromeyer, 1914):

$$g(\sigma_a) = -b\,\log\left(\left[\sigma_a - \sigma_e\right]_+\right) + c \tag{1.3}$$

where $[\cdot]_+ = \max(\cdot, 0)$ denotes the positive part. This model is adapted to the description of the Wöhler curve in the HCF domain.

Another model allows a good representation of the Wöhler curve both in HCF and LCF domains (cf. Bastenaire, 1972). It relies on an additional parameter d:

$$g(\sigma_a) = -\log\left([\sigma_a - \sigma_e]_+\right) - \left(\frac{[\sigma_a - \sigma_e]_+}{b}\right)^d + c \ . \tag{1.4}$$

In the context of fatigue design in the automotive industry, Basquin model is generally used as it provides a good description of the fatigue properties in the HCF domain under limited endurance. The fact that it does not account for the endurance limit is not restrictive as we are usually interested in the fatigue limits for a fixed number of cycles N_0 representing the lifetime of the car (usually $N_0 = 10^6$).

b. Modeling the dispersion of S-N test results

Modeling the dispersion of fatigue test results is crucial as the engineers in reliability are usually interested in low order quantiles of the distribution of $\log(N)$ rather than just the median (or mean) Wöhler curve. Different distributions can be used to model the dispersion of the fatigue lifetime N (Schijve, 2005). The most common class of models assume that $\log(N)$ follows a Gaussian distribution: hence ε (from Eq. 1.1) follows a centered Gaussian distribution with unknown variance (*Normal Fatigue Model*).

Another popular choice of distribution for N is the Weibull distribution (*Weibull fatigue model*): in this case the logarithm $\log(N)$ and thus ε follow a Gumbel distribution with unknown location and scale parameters μ and γ. The cumulative distribution function of ε is given by:

$$F(u) = \mathbb{P}(\varepsilon \leq u) = 1 - e^{-\exp\left(\frac{u-\mu}{\gamma}\right)} \ .$$

As explained by Schijve (2009), the validation of these models in practice requires a lot of experimental data points which is rarely the case. The normal fatigue model is the most commonly used.

S-N models are useful models to estimate Wöhler curves and fatigue lifetime dispersion using a series of uniaxial tests on coupon specimens. However, the geometry of mechanical parts are far more complex than coupon specimens and the stress state is usually multiaxial. Indeed, even a uniaxial load can generate multiaxial local stresses on a complex part. In the context of the design of chassis components, the parts of interest are simultaneously subjected to multiple multiaxial loads: longitudinal (acceleration, braking), lateral (turns), vertical (potholes and humps)... Therefore, more complex fatigue models are needed in order to describe the fatigue risks.

1.2.2 Stress tensors and invariants

When studying complex mechanical parts under cyclic loading, the stress state on each point of the part is often multiaxial and thus requires multiple physical indicators to be fully described. In continuum mechanics, the stresses on each point of the structure are represented by stress tensors (cf. Paragraph a). As the components of a stress tensor depend on the basis it is expressed in, it is common to compute stress invariants to characterize and compare the stresses on each location of the structure (cf. Paragraph b).

a. Stress tensor

Let us consider a structure subjected to an external static load. The external load generates a stress field on the structure. The stress on an element of the structure is represented by a stress tensor $\boldsymbol{\sigma}$. Let us first consider a two-dimensional setting. In this case, the stress tensor $\boldsymbol{\sigma}$ is a 2×2 symmetric matrix expressed in a given basis $(\boldsymbol{e_x}, \boldsymbol{e_y})$:

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} .$$

Considering an infinitesimal square around the element, σ_{xx} and σ_{xy} represent the normal and tangential stresses applied on the right edge (cf. Fig. 1.5). Similarly, σ_{yy} and σ_{xy} are the normal and tangential stresses applied on the upper edge. Since the part is at equilibrium, the stresses on the left and lower edges are symmetric (cf. fig. 1.5). There exists an orthonormal basis $(\boldsymbol{e_1}, \boldsymbol{e_2})$ in which the tensor matrix is diagonal. In this particular basis, the stresses applied on each edge of the infinitesimal square around the element are only normal stresses (cf. Fig. 1.6). Their values σ_1 and σ_2 are the eigenvalues of $\boldsymbol{\sigma}$ (*principal stresses*) and $(\boldsymbol{e_1}, \boldsymbol{e_2})$ are the corresponding eigenvectors (*principal directions*).

These definitions generalize to the three-dimensional setting. In this case, a stress tensor is a symmetric 3×3 matrix expressed in a given basis $(\boldsymbol{e_x}, \boldsymbol{e_y}, \boldsymbol{e_z})$:

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} \end{pmatrix} \tag{1.5}$$

The interpretation of the components of $\boldsymbol{\sigma}$ is the following. Considering an infinitesimal box around the element of interest, the vector $(\sigma_{xx}\ \sigma_{xy}\ \sigma_{xz})$ represents the stress applied on the face with normal $\boldsymbol{e_x}$, $(\sigma_{xy}\ \sigma_{yy}\ \sigma_{yz})$ is the stress on the face with normal $\boldsymbol{e_y}$ and $(\sigma_{xz}\ \sigma_{yz}\ \sigma_{zz})$ is the stress on the face with normal $\boldsymbol{e_z}$. More generally, if $\boldsymbol{u}$ is a normed column vector, the vector $\boldsymbol{\sigma} \times \boldsymbol{u}$ is the stress applied on the face with normal $\boldsymbol{u}$.

Similarly, the matrix $\boldsymbol{\sigma}$ can be diagonalized. By convention, the principal stresses σ_1, σ_2 and σ_3 are sorted in decreasing order. The associated principal directions are denoted $\boldsymbol{e_1}$, $\boldsymbol{e_2}$ and $\boldsymbol{e_3}$. The components of the stress tensor depend on the basis the tensor is expressed in. Therefore, comparing the stress tensors of two elements from two different structures is meaningless as the structures can have different orientations and coordinate systems.

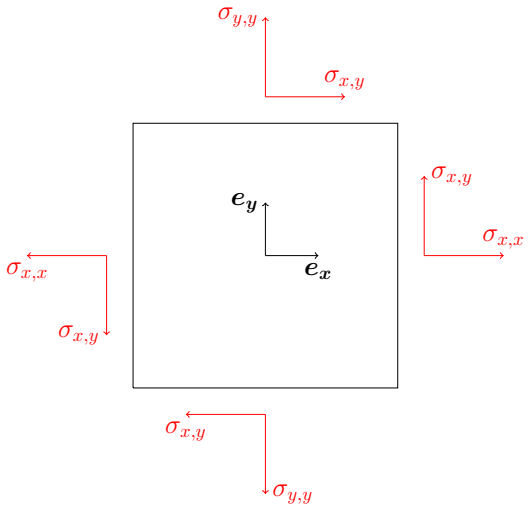

Figure 1.5: Two-dimensional stress tensor in the nominal basis

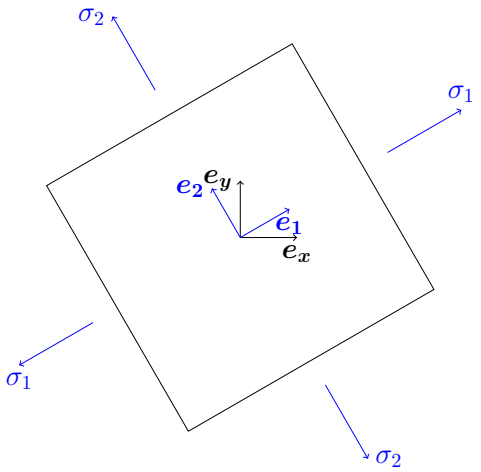

Figure 1.6: Two-dimensional stress tensor in the principal basis

b. Stress invariants

Stress invariants are physical quantities derived from the stress tensor that are independent of the coordinate system the tensor is expressed in. These physical quantities are suitable features to compare the stresses of different elements from different structures.

Hydrostatic stress. The first invariant of the stress tensor $\boldsymbol{\sigma}$ is its trace, denoted I_1:

$$I_1 = \sigma_{xx} + \sigma_{yy} + \sigma_{zz} = \sigma_1 + \sigma_2 + \sigma_3 \ .$$

An important mechanical quantity based on this invariant is the *hydrostatic stress* P equal to the average of the principal stresses:

$$P = \frac{1}{3} I_1 \ .$$

It represents the mean of normal stresses applied on the infinitesimal element. A negative P means that the element is globally subjected to compressive forces which tends to delay the initiation and propagation of cracks. Conversely, a positive P means that the element globally works in traction which facilitate cracks development.

Von Mises stress. A stress tensor $\boldsymbol{\sigma}$ can be decomposed into a spherical part and a deviatoric part:

$$\boldsymbol{\sigma} = P\,\boldsymbol{I} + (\boldsymbol{\sigma} - P\,\boldsymbol{I})$$

where $\boldsymbol{I}$ denotes the identity matrix.

- $P\,\boldsymbol{I}$ is the spherical part.
- $\boldsymbol{D} = \boldsymbol{\sigma} - P\,\boldsymbol{I}$ is the deviatoric part containing the information about shear stresses.

By definition, the trace of $\boldsymbol{D}$ is 0. An important quantity is the second invariant J_2 of the deviatoric part of the stress tensor defined as:

$$J_2 = \frac{1}{2} Tr(\boldsymbol{D}^2)$$

where Tr stands for the trace. This quantity is related to Von Mises stress V:

$$V = \sqrt{3\,J_2} \ .$$

By definition, V is homogeneous to a stress and is always positive.

Tresca shear stress. Tresca invariant τ is another measure of shear stress. It is defined as the amplitude between the maximum and minimum principal stresses and is also positive:

$$\tau = \frac{1}{2}(\sigma_1 - \sigma_3) \ .$$

Stress triaxiality. Stress triaxiality T is an indicator characterizing the type of stress an element is subjected to. It is defined as the ratio between the hydrostatic stress P and Von Mises shear stress V and is thus a quantity without unit:

$$T = \frac{P}{V} \ .$$

The sign of T is identical to the sign of P, therefore positive under traction normal stresses and negative under compressive normal stresses. When $T = 0$, the stress tensor only involves shearing. The triaxiality is equal to $1/3$ under uniaxial traction ($-1/3$ under uniaxial compression) and $2/3$ under equi-biaxial traction. A high triaxiality means that the deviatoric part of the stress tensor is negligible compared to the spherical part.

The stress invariants defined in this paragraph offer a representation of the stress independent from the coordinate system chosen. We will see that these invariants are involved in different multiaxial fatigue criteria. It is important to note that others stress invariants exist even if they are not useful here.

1.2.3 Multiaxial fatigue criteria

During the design of complex structures, engineers want to ensure that every point of the structure is resistant enough to fatigue risks. As the stresses are usually multiaxial, modeling the lifetime of the structure can be very challenging. Besides, engineers only seek a binary answer, *i.e.* whether or not the structure is resistant enough to fatigue crack initiations. Fatigue criteria are then used for this purpose.

Paragraph a clarifies the role and objectives of a fatigue criterion compared to a S-N model. In Paragraph b, we introduce the general formalism of a multiaxial fatigue criterion. Paragraph c presents the different categories of multiaxial fatigue criteria.

a. Objective of a fatigue criterion

A S-N curve represents the mean fatigue lifetime of a specimen (simple, small and homogeneous structure) depending on the stress it is subjected to. If the objective is to characterize the conditions under which the lifetime is infinite, we are no longer interested in the whole S-N curve but only on its asymptote. A fatigue criterion provides a binary answer to the question: is this stress over or below the endurance limit of the material (Nadjitonon, 2010)? The difference between an S-N model and a fatigue criterion is illustrated in Figure 1.7 considering a uniaxial setting. On the left figure, fatigue test data points are represented. After N cycles at stress amplitude σ_a, we denote Y the result of the test: $Y = 0$ (blue points on the graph) if no crack is observed and $Y = 1$ (red stars on the graph) if a crack initiated. Hence the S-N model characterizes the risk (probability of crack initiation) of a specimen with stress amplitude σ_a tested over N cycles to fail during testing. The right figure represents the objective of a fatigue criterion, *i.e.* to predict whether a specimen under stress amplitude σ_a is over ($Z = 1$, orange) or below ($Z = 0$, green) the endurance limit σ_e. An element with stress amplitude over the endurance limit is a *critical element*.

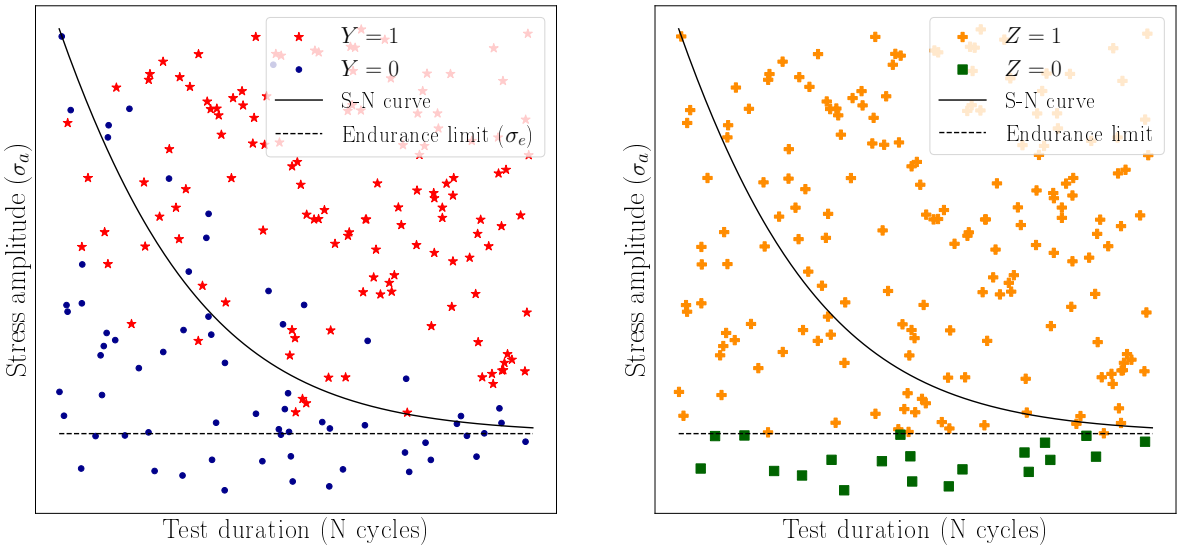

Figure 1.7: S-N curve and fatigue criterion : a uniaxial example (artificial data). In both figures, X-axis represents test duration and Y-axis is the stress amplitude. On the left, red stars represent crack initiations, blue points are non-broken specimens. The figure on the right corresponds to the outputs of the fatigue criterion (orange for critical elements, green for safe elements).

In a uniaxial setting, S-N models and fatigue criteria are connected. Indeed, using for instance Stromeyer or Bastenaire S-N model, one can obtain an estimate of the asymptote of the S-N curve, which defines a uniaxial fatigue criterion.

The concept of fatigue criterion is usually extended beyond the unlimited endurance setting (Fares, 2006). In applications, it is very common to consider fatigue criteria predicting whether a specimen is over or below the fatigue limit for a specified finite lifetime N_0 ($N_0 = 10^6$ in the automotive industry). Again, in a uniaxial setting, standard S-N models (including Basquin model, cf. Subsection 1.2.1) can provide an estimate of this fatigue limit and hence help calibrate the fatigue criterion.

Another important difference between S-N models and fatigue criteria is that the Wöhler curve is used for uniaxial, cyclic and constant-amplitude loads whereas fatigue criteria can be defined for multiaxial (and constant-amplitude) cyclic loads.

b. General formulation of a multiaxial fatigue criterion

A multiaxial fatigue criterion is based on a real-valued function h, called *fatigue function* (cf. Weber, 1999) indicating whether or not the stress cycle on an element of a structure exceeds the endurance limit (infinite lifetime) or the fatigue limit at N_0 cycles (finite lifetime). We recall that the fatigue phenomenon is due to the repetition of a cyclic load on a mechanical part. The response of the structure is characterized by a stress field. In particular, the stress cycle on a given point of the structure is characterized by its stress tensor over a period $(\boldsymbol{\sigma}(t))_{0\leq t<T}$. Hence, the fatigue function h depends on the whole stress cycle $(\boldsymbol{\sigma}(t))_{0\leq t<T}$ and on the material M. By convention, the multiaxial fatigue criterion predicts that the element is:

- safe if $h\left((\boldsymbol{\sigma}(t))_{0\leq t<T}, M\right) < 1$;
- critical if $h\left((\boldsymbol{\sigma}(t))_{0\leq t<T}, M\right) > 1$.

The fatigue limit of the material M is reached when $h\left((\boldsymbol{\sigma}(t))_{0\leq t<T}, M\right) = 1$.

This formulation remains very general: most fatigue functions only account for a limited

number of features extracted from the stress cycle. Usually two types of information are taken into account.

1. Shear stress is the principal cause of crack initiation and thus an important parameter of the fatigue function. Usually, it is represented by the maximum shear stress over the cycle using either Von Mises stress $\max(V(t))$ or Tresca shear stress $\max(\tau(t))$.

2. Hydrostatic stress can accelerate (traction) or delay (compression) crack initiation and is thus also an important parameter. Depending on the criterion, some fatigue functions account for the amplitude, mean and maximum of hydrostatic stress over the cycle.

Moreover, the fatigue function h depends on material parameters that need to be identified empirically. Usually, the calibration of a fatigue criterion requires the knowledge of material fatigue limits in at least two different types of loading (traction and torsion for instance). As explained in Subsection 1.1.3, the fatigue strength not only depends on the material but also on the manufacturing processes (stamping, welding...).

c. Categories of multiaxial fatigue criteria

The literature on fatigue criteria is very rich: the existing criteria can be classified in several ways. Let us consider here the classification proposed by Nadjitonon (2010) with four categories.

1. *Empirical fatigue criteria* are adapted to the characterization of the material fatigue limit for a given type of multiaxial stress. Their main limit is that they are only suited for a single type of loading and thus cannot generalize to different types of loading. An example of such a criterion is the one proposed by Gough & Pollard (Gough et al., 1951).

2. *Critical plane fatigue criteria* consider that a crack is more likely to initiate along a specific plane called *critical plane*. The critical plane is defined as the one maximizing a criterion involving stresses applied on that particular plane. The final fatigue function h depends on invariants of the stress tensor and stresses applied on the critical plane. Dang Van fatigue criterion (Dang Van and Griveau, 1989) commonly used in the automotive industry, belongs to this class. It will be presented in further details in the next Subsection.

3. *Fatigue criteria based on an integral approach* consider the contribution of each plane to the degradation by calculating an integral over all the possible planes of a given function (for instance Fogue criterion, cf. Fogue and Bahuaud, 1985).

4. *Global approach fatigue criteria* are based on stress tensor invariants. For instance, Crossland criterion involves a linear combination between the maximum of the first stress invariant $I_{1,max}$ and the amplitude of the second invariant $J_{2,a}$ (Crossland, 1956). Sines criterion also consider $J_{2,a}$ but in addition to the mean of the first invariant $I_{1,m}$ (Sines and Ohgi, 1981).

1.2.4 Dang Van fatigue criterion

Dang Van criterion is a critical plane fatigue criterion commonly used in the automotive industry (Thomas et al., 2005). As we will frequently refer to this criterion in the rest of the thesis, the purpose of this subsection is to introduce it. Paragraph a explains the construction of Dang Van criterion and the underlying assumptions. In Paragraph b, we show how the corresponding fatigue function is calculated in practice, focusing on proportional and sinusoidal loadings which will be sufficient in the scope of this thesis. Finally, Paragraph c investigates the calibration of the material constants involved in Dang Van criterion.

a. Definition of Dang Van criterion

Dang Van criterion is based on considerations at the microscopic scale of the material. The initiation of fatigue cracks is due the plastic deformations occurring at the microscopic scale in the material. Because of the repetition of stress cycles, micro-cracks progressively appear and can lead to the apparition of a macroscopic crack. Dang Van designed a criterion of non-initiation according to which micro-cracks (and thus fatigue cracks) cannot initiate if the behaviour of the material remains elastic at the microscopic scale (Dang Van and Griveau, 1989). At the microscopic scale, a metallic material consists in a complex arrangement of grains. Dang Van states that no crack initiation can happen if the grains with the worse orientation do not break. The fatigue function associated to Dang Van criterion, denoted h_{DV}, depends on the hydrostatic stress $P(t)$ and on the maximum mesoscopic shear stress[1] taken over all possible planes $\tau_{mes}(t)$. The criterion considers a linear combination of both quantities maximized over the stress cycle:

$$h_{DV}\left((\boldsymbol{\sigma}(t))_{1\leq t<T}, M\right) = \max_{0\leq t<T} \frac{\tau_{mes}(t) + \alpha_M\, P(t)}{\tau_M} \ . \tag{1.6}$$

Here, τ_M and α_M are material parameters that need to be calibrated (cf. Paragraph c).

b. Practical implementation of Dang Van criterion

The application of Dang Van criterion requires the calculation of the hydrostatic stress $P(t)$ and of the maximum mesoscopic shear stress $\tau_{mes}(t)$. The computation of the former is straightforward knowing the stress tensor at every time t ($\boldsymbol{\sigma}(t)$) whereas the second is much more complex to calculate. Dang Van provided a practical methodology to compute $\tau_{mes}(t)$ and thus apply the criterion (Dang Van and Griveau, 1989; Ballard et al., 1995).

First, the deviatoric part $\boldsymbol{D}(t)$ of the stress tensor is computed for every time t. Then, the residual mesoscopic shear $\boldsymbol{s}$ is calculated as the center of the smallest circumscribing hypersphere to the load path in the deviatoric space $(\boldsymbol{D}(t))_{0\leq t<T}$. The mesoscopic stress tensor is defined as the difference between the macroscopic stress tensor and the residual mesoscopic shear: $\boldsymbol{\sigma}_{mes}(t) = \boldsymbol{\sigma}(t) - \boldsymbol{s}$. Finally $\tau_{mes}(t)$ is obtained as Tresca shear stress invariant computed on the mesoscopic stress tensor $\boldsymbol{\sigma}_{mes}(t)$.

The second step involving the computation of the smallest hypersphere's center remains complex in general. In the scope of the thesis, we will be considering sinusoidal and affine stress cycles of the form:

$$\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}_m + \cos\left(\frac{2\pi\, t}{T}\right) \boldsymbol{\sigma}_a \ .$$

$\boldsymbol{\sigma}_m$ and $\boldsymbol{\sigma}_a$ denote the mean and amplitude tensors.

Let us show how the maximum mesoscopic shear stress can be derived in this setting. First, the deviatoric part of tensor $\boldsymbol{\sigma}(t)$ can be expressed as a function of the deviatoric parts of $\boldsymbol{\sigma}_m$

[1] Shear stress at the grain scale of the material (cf. Paragraph b for its calculation in practice).

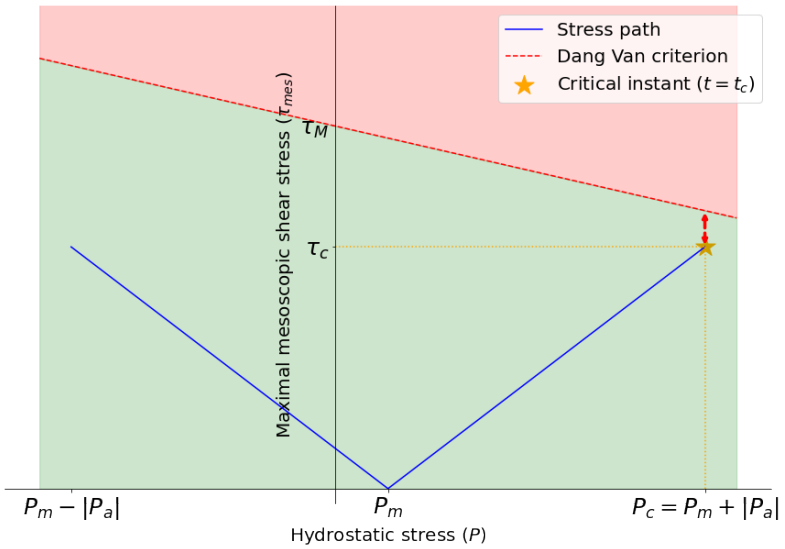

Figure 1.8: Example of stress path $(P(t), \tau_{mes}(t))_{0 \leq t < T}$ in Dang Van plane (blue curve) and representation of the fatigue criterion (red dashed line, with intercept τ_M and slope $-\alpha_M$). The element is safe if the stress path remains in the green zone, critical if it reaches the red zone.

($\boldsymbol{D}_m$) and $\boldsymbol{\sigma}_a$ ($\boldsymbol{D}_a$):

$$\boldsymbol{D}(t) = \boldsymbol{D}_m + \cos\left(\frac{2\pi\, t}{T}\right) \boldsymbol{D}_a \ .$$

Then, in this specific case, the load path in the deviatoric space is a straight line whose center is $\boldsymbol{D}_m$. Hence, the mesoscopic stress tensor can be easily derived:

$$\begin{aligned} \boldsymbol{\sigma}_{mes}(t) &= \boldsymbol{\sigma}(t) - \boldsymbol{D}_m \\ &= P_m \boldsymbol{I} + \cos\left(\frac{2\pi\, t}{T}\right) \boldsymbol{\sigma}_a \end{aligned}$$

where P_m is the mean hydrostatic stress (hydrostatic stress of $\boldsymbol{\sigma}_m$) and $\boldsymbol{I}$ denotes the identity matrix.

Finally, denoting τ_a the Tresca shear stress invariant applied to $\boldsymbol{\sigma}_a$, we can remark that the maximum mesoscopic shear stress $\tau_{mes}(t)$ only depends on τ_a:

$$\tau_{mes}(t) = \left| \cos\left(\frac{2\pi\, t}{T}\right) \right| \times \tau_a \ .$$

The hydrostatic stress $P(t)$ can also be expressed in terms of P_m (mean) and P_a (amplitude):

$$P(t) = P_m + P_a \cos\left(\frac{2\pi\, t}{T}\right) \ .$$

Dang Van criterion can be represented in a *Dang Van diagram* featuring the hydrostatic stress (x-axis) and the maximum mesoscopic shear stress (y-axis). In this plane, Dang Van criterion is a line with intercept τ_M and slope $-\alpha_M$ (cf. Fig. 1.8). A stress path $(P(t), \tau_{mes}(t))_{0 \leq t < T}$ can be visualized in this plane: the stress cycle is safe if it remains below the criterion line. In practice, as the stress cycles we will be analyzing all have this "V" shape, we only need to check the relative position of the top right point (critical instant t_c) to Dang Van line (cf. Fig. 1.8). The values of hydrostatic stress and maximum mesoscopic shear stress at critical instant t_c will be referred as *critical hydrostatic stress* (P_c) and *critical shear stress* (τ_c).

The value of Dang Van fatigue function (cf. Eq. 1.6) can be easily expressed in terms of P_c and τ_c. Engineers usually prefer to compute the *danger coefficient* (CD):

$$CD = h_{DV}\left((\boldsymbol{\sigma}(t))_{1\leq t<T}, M\right) - 1 = \frac{\alpha_M P_c + \tau_c}{\tau_M} - 1 \ .$$

It is a measure of criticality of a stress path: CD is negative for safe instances and positive for critical ones. Besides, one can note that iso-CD lines in Dang Van plane are parallel to Dang Van criterion line.

c. Calibration of Dang Van criterion

The fatigue function h_{DV} characterizing Dang Van criterion depends on two material parameters: α_M and τ_M. Those constants are calculated based on two types of uniaxial fatigue test from which two fatigue limits are estimated. Usually traction and torsion tests are used[2]. For each type of test, multiple coupon specimens are tested for different load intensities in order to estimate the Wöhler curve. Once the two Wöhler curves are estimated, the fatigue limits at lifetime N_0 are calculated: $\sigma_{trac}(N_0)$ and $\sigma_{tors}(N_0)$ are the fatigue limits for the traction and torsion tests. Since each test is characterized by different stress tensors, we can calculate their coordinates in Dang Van plane.

- For the uniaxial traction test, the stress tensor is of the form:

$$\boldsymbol{\sigma} = \cos\left(\frac{2\pi\, t}{T}\right) \begin{pmatrix} \sigma_{trac} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} .$$

  The critical hydrostatic and shear stresses are:

$$P_c = \frac{\sigma_{trac}}{3} \quad \text{and} \quad \tau_c = \frac{\sigma_{trac}}{2} \ .$$

- For the uniaxial torsion test, the stress tensor is:

$$\boldsymbol{\sigma} = \cos\left(\frac{2\pi\, t}{T}\right) \begin{pmatrix} 0 & \sigma_{tors} & 0 \\ \sigma_{tors} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} .$$

  The critical hydrostatic and shear stresses are:

$$P_c = 0 \quad \text{and} \quad \tau_c = \sigma_{tors} \ .$$

Using the fatigue limits $\sigma_{trac}(N_0)$ and $\sigma_{tors}(N_0)$, the two corresponding points are reported in Dang Van plane. The criterion is then defined as the line passing by these two points (cf. Fig. 1.9). In particular, the intercept τ_M and the slope $-\alpha_M$ are:

$$\tau_M = \sigma_{tors}(N_0) \quad \text{and} \quad \alpha_M = 3\left(\frac{\sigma_{tors}(N_0)}{\sigma_{trac}(N_0)} - \frac{1}{2}\right) \ .$$

Dang Van criterion can be defined for any objective lifetime N_0, even infinite: in this case, the criterion will describe the endurance limit. In the context of fatigue design for the automotive industry, we will be considering the standard objective lifetime $N_0 = 10^6$, usually being representative of the service life of a car.

To sum up, multiaxial fatigue criteria (including Dang Van criterion) can be used to assess whether a point of a structure, given its local stress cycle, will be resistant to crack initiation over a fixed lifetime N_0. During the fatigue design of a mechanical part, the objective is to ensure that the whole structure (defined by its geometry, materials, manufacturing processes) is resistant enough to fatigue risks.

[2] A usual roadblock considering steel sheets is that it is difficult to design a coupon torsion test.

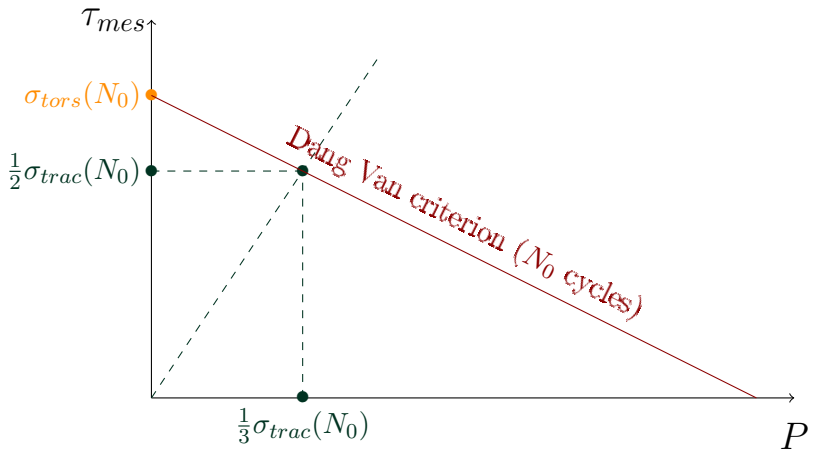

Figure 1.9: Calibration of Dang Van criterion at N_0 cycles using the fatigue limits for uniaxial traction test $\sigma_{trac}(N_0)$ and torsion test $\sigma_{tors}(N_0)$.

1.3 - Fatigue design of complex mechanical parts

During the design of a new vehicle, engineers propose a solution that meets some requirements in terms of functionality, durability and reliability. This is particularly crucial for safety parts of the vehicle for which the reliability requirements are high (cf. Thomas et al., 2005). During the development phase, engineers need to ensure that the components of the vehicle are resistant enough to the in-service loads they may encounter. The fatigue design strategy and the resistance objectives are defined through a Stress-Strength method (cf. Subsection 1.3.1). Then, the validation of the components is carried out using fatigue tests on prototypes (cf. Subsection 1.3.2). As validation fatigue tests are long and expensive, it is important for engineers not to wait for the validation tests and to be able to assess the resistance of the mechanical part through numerical simulations. This pre-validation stage usually relies on the application of a fatigue criterion that can help anticipate potential design flaws on a design proposal, prior to the validation tests (cf. Subsection 1.3.3).

1.3.1 Stress-Strength interference method

The Strength-Stress interference method is a probabilistic approach to the fatigue design of structures (cf. Thomas et al., 2005; Echard et al., 2014). The safety requirements on a structure set a maximum failure probability acceptable p_{max}. The general objective of fatigue design is to ensure that the probability for a random customer to exceed the fatigue limit of a component is lower than p_{max}. The safety requirement can thus be translated into a mathematical formulation:

$$\mathbb{P}\left(S > R\right) \leq p_{max} \ . \tag{1.7}$$

The variable S represents the stress (load) applied to the part during the utilization of the vehicle and R is the resistance of the part, *i.e.* the limit stress value over which the part will fail (crack initiation).

Multiple sources of variability need to be accounted for in S: the motion of the vehicle (acceleration, braking, turns), the road conditions (asphalt, cobblestones, potholes, humps...), the payload of the vehicle and the driving style (smooth, aggressive). The distribution of S is usually modeled by a univariate Gaussian distribution with mean μ_S and variance σ_S^2 (cf. Fig. 1.10, orange curve).

The variability of R is due to the material properties of the part and the manufacturing process: even two macroscopically identical parts will have different resistances. Hence, the

resistance (or strength) is represented by a random variable R following a Gaussian distribution with mean μ_R and variance σ_R^2 (cf. Fig. 1.10, blue curve).

Hence, S represents the intensity of the damages the part is subjected to: the greater S is, the more severe the customer is for the part. The resistance R is the maximum severity the part can endure. Equation 1.7 expresses the durability objective for the car manufacturer: over its service life (*e.g.* 10^5 kilometers), the risk for the part to fail in service should be less than p_{max}. For safety parts, p_{max} is very low, typically 10^{-6}.

Since S and R are assumed independent, the probability $\mathbb{P}(S > R)$ from Equation 1.7 can be expressed in terms of μ_S, σ_S, μ_R and σ_R.

The parameters of the stress distribution (μ_S and σ_S) are estimated using customer usage data. The knowledge of these parameters allows to define an "objective customer" (or "reference customer") as a certain quantile of the distribution of S. At Stellantis (ex-PSA), the objective customer is defined as the quantile of S of a certain level $1 - 1/M$. It represents the severity value that only 1 customer over M will exceed on average. The objective customer F_n is used as a baseline for defining the loading intensity to be simulated on numerical models (cf. Subsection 1.3.3) and for defining the acceptance criterion for prototype validation tests (cf. Subsection 1.3.2). The objective customer F_n is often parameterized as:

$$F_n = \mu_S + \alpha\,\sigma_S \tag{1.8}$$

where the parameter α only depends on the quantile of level $1/M$ of the normal distribution.

Let us now explain how the design requirement of Equation 1.7 is translated into a criterion on the mean resistance μ_R of the structure. Resistance parameters μ_R and σ_R are unknown. Nevertheless, the coefficient of variation $q = \sigma_R/\mu_R$ is assumed to be known through expert knowledge on the materials and manufacturing process (cf. Bergamo et al., 2017). This leaves only one unknown parameter: the mean resistance μ_R. The safety requirement of Equation 1.7 is satisfied if and only if the mean resistance μ_R satisfies a certain validation criterion. This validation criterion is often written in terms of the relative position between μ_R and the objective customer F_n (cf. Fig. 1.10):

$$\mu_R \geq F_n + \beta\sigma_R\ , \tag{1.9}$$

where β depends on μ_S, σ_S, σ_R and on the maximum failure probability p_{max}. Hence, if Equation 1.9 is satisfied, the probability for a random customer S to encounter a mechanical part with a weaker resistance R is less than p_{max}.

1.3.2 Fatigue rig tests for validation

Once the conception of a mechanical part is done, fatigue rig-tests are performed in order to validate its resistance to fatigue. Different series of fatigue tests are carried out for different types of loading. They represent different types of external forces the mechanical part will be subjected to when the car is in service (cornering, longitudinal, transversal, vertical...).

The objective of a fatigue test for a given type of loading is to validate the design by checking that the mean resistance μ_R effectively satisfies the durability requirements (cf. Eq. 1.9). For that purpose, multiple tests are performed and the mean resistance is estimated thanks to the outcomes of the tests (cf. Paragraph a). We will present two different test protocols: Staircase protocol (cf. Paragraph b) and Locati protocol (cf. Paragraph c). As Locati protocol consists in incrementing the severity (loading amplitude) during the fatigue test, the test is no longer performed at constant amplitude. Therefore, an equivalent severity is calculated in order to estimate the resistance. Paragraph d will be dedicated to the introduction of this concept of fatigue equivalent.

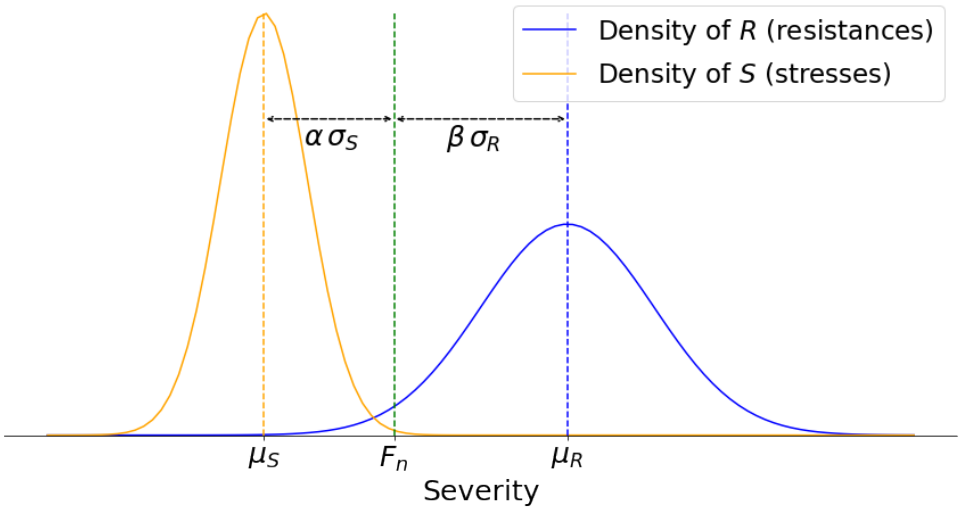

Figure 1.10: Stress-Strength method for fatigue design: the blue curve represents the distribution of R (strength) and the orange curve the distribution of S (stress). The objective customer F_n is an extreme quantile of the stress distribution.

a. General principles on fatigue tests on prototypes

For a given type of loading, a bench is designed in order to reproduce the external forces the part will be subjected to in real conditions (cf. example in Fig. 1.11).

Usually, not less than three identical prototypes are tested. For each prototype, the load cycle is repetitively applied over a certain duration (usually more than 10^6 cycles) or until one or multiple cracks initiate on the part. Testing a single prototype can take several weeks, hence a test campaign may last several months. This is also why companies cannot afford to test much more prototypes.

Depending on the test protocol, the prototype is inspected at the end of the test, and possibly also during the test. This inspection allows to detect potential crack initiations that occurred during testing. As the cracks may be small, penetrant inspection can be used to help detect defects. It consists in using a high contrast liquid on the surface of the mechanical part to help crack's identification (cf. Fig. 1.12).

The final objective is to validate the design by estimating the mean resistance of the mechanical part, *i.e.* the severity for which no crack will appear before 10^6 cycles. For that purpose, only the first crack initiation is important: the number of cycles before the first crack initiation defines the lifetime of the prototype. Nevertheless, it is common to continue the test further (especially for the Locati protocol, cf. Paragraph c) as far as the first crack initiation does not modify significantly the behavior of the part. This way, additional cracks can initiate, informing about others potential weaknesses of the part.

Tests are carried out according to a precise protocol. There are different types of protocols (Beaumont et al., 2012; Beaumont, 2013): Paragraph b presents the Staircase protocol, Paragraph c deals with the Locati protocol which is an accelerated testing protocol also popular in the automotive industry.

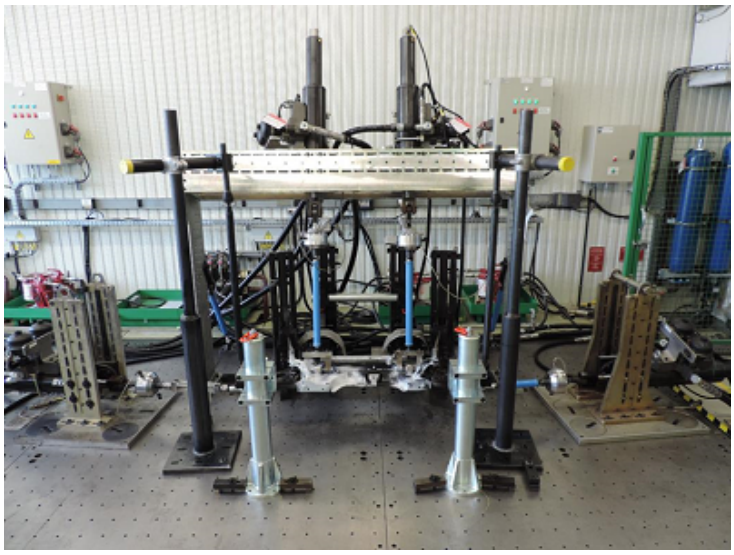

Figure 1.11: Example of bench for testing a cradle model under cornering loading.

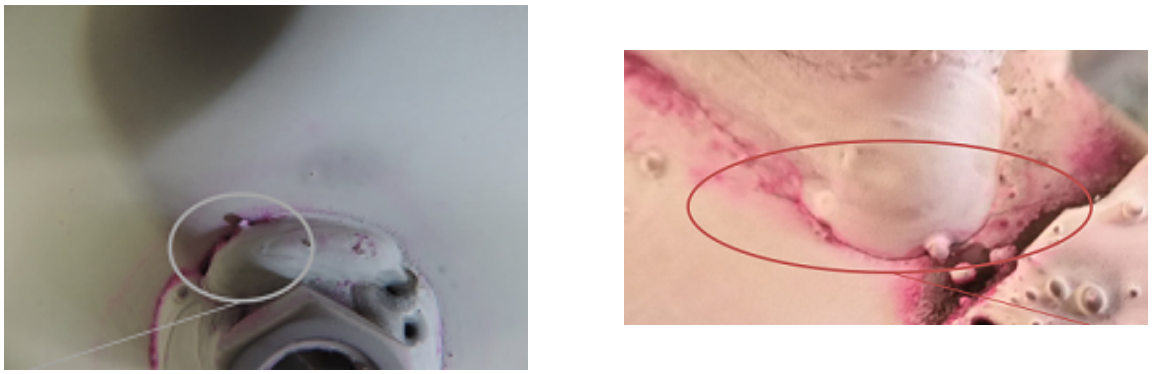

Figure 1.12: Penetrant inspection for fatigue crack detection using a pink liquid to help crack's identification.

b. Staircase fatigue test protocol

The Staircase fatigue test protocol consists in performing constant amplitude tests and adapting the severity of the test (*i.e.* the value of the load amplitude) from the outcome of the previous one (Lin et al., 2001; Zhao and Yang, 2008).

Let f_i denote the severity of test number i. The test is carried out over 10^6 cycles and the outcome of the test is binary: presence or absence of crack initiation. If a crack initiation is observed, then the severity of the following test is decreased: $f_{i+1} = f_i - f_{inc}$. Else, if no crack is observed, the severity is increased: $f_{i+1} = f_i + f_{inc}$ (cf. Fig. 1.13).

The severity of the first test f_1 is usually chosen close to the target mean resistance μ_R that needs to be validated. The increment f_{inc} between the tests is usually set as the standard deviation of the target resistance σ_R.

The mean resistance can be estimated using the outcome of all the tests (Dixon and Mood, 1948).

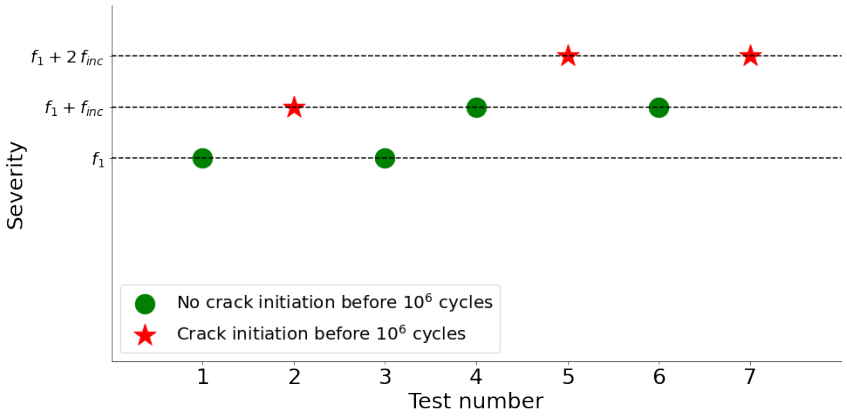

Figure 1.13: Illustration of staircase fatigue test protocol: red stars (green points) represent tests with (without) crack initiation before 10^6 cycles.

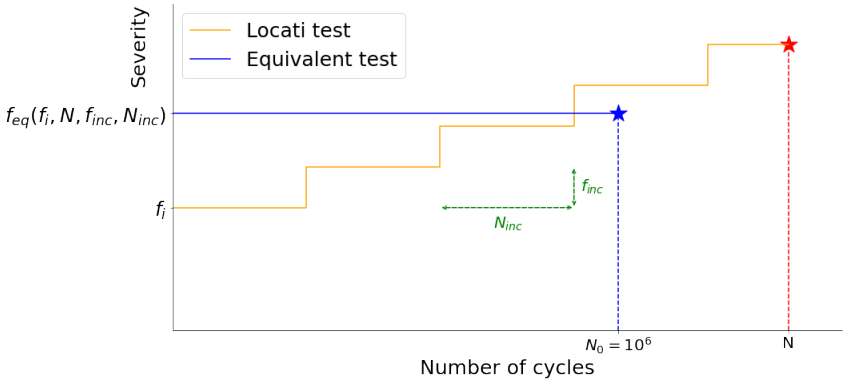

Figure 1.14: Illustration of a Locati test with initial severity f_i. The orange line represents the Locati test: severity (Y-axis) as a function of the number of cycles (X-axis). The blue line represent the equivalent test at constant severity over 10^6 cycles. In terms of fatigue damage, the equivalent test is equivalent to the Locati test.

c. Locati protocol

Locati protocol is an accelerated test life protocol that consists in incrementing the severity during the test (Locati, 1955; Brevet et al., 1978). This way, every prototype is tested until failure. The protocol is characterized by the following parameters:

- the increment in severity during testing f_{inc};
- the number of cycles between two increments N_{inc}.

Let f_i denote the initial severity of test number i. Every N_{inc} cycles, the severity is increased by f_{inc}. Finally, the number of cycles N_i before the first crack initiation is observed (cf. Fig. 1.14). An equivalent severity $f_{eq}(f_i, N_i, f_{inc}, N_{inc})$ is calculated to estimate the resistance of the prototype. The interpretation is the following: the Locati test with parameters $(f_i, N_i, f_{inc}, N_{inc})$ performed is equivalent in terms of fatigue damage to a constant amplitude test at severity $f_{eq}(f_i, N_i, f_{inc}, N_{inc})$ over $N_0 = 10^6$ cycles (cf. Fig. 1.14). The details about the definition of the equivalent severity are given in the next Paragraph.

Once all the k prototypes have been tested, the mean resistance μ_R is estimated as:

$$\widehat{\mu_R} = \frac{1}{k} \sum_{i=1}^{k} f_{eq}(f_i, N_i, f_{inc}, N_{inc}) \ .$$

In the scope of this thesis, the test results were obtained through the Locati protocol.

d. Fatigue equivalent severity of a Locati test

The usual fatigue models (S-N models and fatigue criteria) always rely on the assumption that the stress cycle is repeated over time. During a Locati test though, the severity is incremented gradually, hence this assumption is no longer satisfied. In order to assess the resistance of a mechanical part given the parameters of the Locati test (f_i, N_i, f_{inc}, N_{inc}), engineers calculate the severity of an equivalent test (in terms of fatigue damage) performed over 10^6 cycles with constant severity. The severity of this test is called *equivalent severity* at 10^6 cycles and is denoted $f_{eq}(f_i, N_i, f_{inc}, N_{inc})$.

The calculation of the equivalent severity relies on two ingredients:

1. a cumulative damage law (Miner rule is the simplest and the most frequently used);
2. an S-N curve for the material (in our case, the S-N curve is assumed to follow a Basquin model with a known slope b).

Miner cumulative damage rule (Miner, 1945; Palmgren, 1924) Consider a prototype subjected to n_1 cycles at severity $f^{(1)}$, n_2 cycles at severity $f^{(2)}$,..., and n_p cycles at severity $f^{(p)}$. We assume that a crack initiated at the end of the test.

Now, let us denote N_j the lifetime of the prototype at severity $f^{(j)}$, for j ranging from 1 to p: this means that for a severity $f^{(j)}$, we observe a crack initiation after N_j cycles. The damage undergone by the part is then equal to 1. Miner rule states that the damages are accumulated linearly. Hence, n_j cycles at severity $f^{(j)}$ generate a damage:

$$D_j = \frac{n_j}{N_j} \ . \tag{1.10}$$

Besides the total damage D endured by the part is the sum of the individual damages from Equation 1.10:

$$D = \sum_{j=1}^{p} \frac{n_j}{N_j} \ . \tag{1.11}$$

It is thus assumed that the total damage is independent of the order according to which the different load cycles are applied.

If a crack initiation is observed at the end of the test, the cumulative damage is then equal to 1. Hence, knowing that the part failed at the end of the test, we have the following equality:

$$\sum_{j=1}^{p} \frac{n_j}{N_j} = 1 \ . \tag{1.12}$$

Basquin model. We now assume that the lifetime given the stress amplitude follows a Basquin model with a known Basquin slope b (cf. Subsection 1.2.1). As the stress amplitude is proportional to the severity, the relation of Basquin model can be directly written in terms of the lifetime N and the severity f:

$$\log(N) = a - b\,\log(f)\ . \tag{1.13}$$

If f_{eq} denotes the severity at which the prototype has a lifetime equal to $N_0 = 10^6$ cycles, we have the following relation:

$$\log(N_0) = a - b\log(f_{eq})\ .$$

We can thus replace the unknown parameter a in Equation 1.13, which yields:

$$\log\left(\frac{N_0}{N}\right) = -b\log\left(\frac{f_{eq}}{f}\right)\ .$$

Hence, $1/N$ can be expressed as a function of N_0, f_{eq}, f and b:

$$\frac{1}{N} = \frac{1}{N_0}\left(\frac{f}{f_{eq}}\right)^b\ .$$

Now, considering the cumulative damage given by Equation 1.12, we can re-express $1/N_j$ for every j by using the previous equation:

$$\sum_{j=1}^{p}\frac{n_j}{N_0}\left(\frac{f^{(j)}}{f_{eq}}\right)^b = 1\ . \tag{1.14}$$

Noting that the only unknown variable in Equation 1.14 is f_{eq}, it can be expressed as a function of the remaining variables and parameters:

$$f_{eq} = \left[\sum_{j=1}^{p}\frac{n_j}{N_0}\left(f^{(j)}\right)^b\right]^{\frac{1}{b}}\ . \tag{1.15}$$

The resulting quantity f_{eq} is the equivalent severity in the sense that a crack initiation after n_1 cycles at severity $f^{(1)}$,..., n_p cycles at severity $f^{(p)}$ is equivalent to a crack initiation after N_0 cycles at severity f_{eq}.

Equivalent severity for a Locati test. The Locati test is characterized by the parameters of the increments (N_{inc}, f_{inc}), the initial severity f_0 and the total number of cycles before crack initiation N. Let (q, r) denote the quotient and rest of the euclidean division of N by N_{inc}, the part was subjected to:

- N_{inc} cycles at severity f_0;
- N_{inc} cycles at severity $f_0 + f_{inc}$;
- ...
- N_{inc} cycles at severity $f_0 + (q-1)\,f_{inc}$;
- r cycles at severity $f_0 + q\,f_{inc}$.

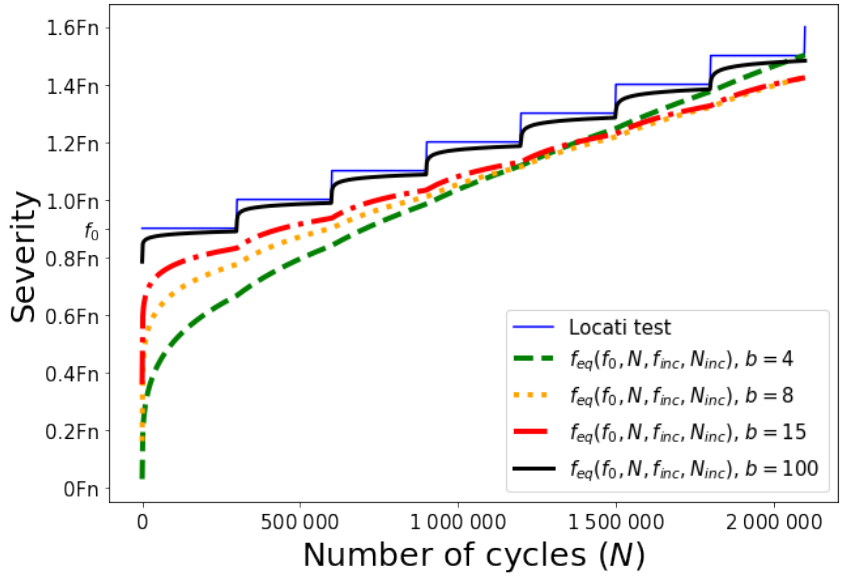

Figure 1.15: Sensitivity analysis of the fatigue equivalent severity to Basquin parameter b. The blue curve represents the Locati test. The other curves represent the fatigue equivalent severity $f_{eq}(f_0, N, f_{inc}, N_{inc})$ as a function of the number of cycles N for a given set of parameters (f_0, N_{inc} and f_{inc}).

Applying the fatigue equivalent formula of Equation 1.15, we obtain the equivalent severity of the Locati test:

$$f_{eq}(f_0, N, f_{inc}, N_{inc}) = \left[\frac{1}{N_0} \sum_{j=0}^{q-1} \left[N_{inc} \, (f_0 + j \, f_{inc})^b \right] + \frac{r}{N_0} \, (f_0 + q \, f_{inc})^b \right]^{\frac{1}{b}} . \tag{1.16}$$

Figure 1.15 represents the equivalent severity $f_{eq}(f_0, N, f_{inc}, N_{inc})$ as a function of the number of cycles N. The other variables are fixed. Multiple colors correspond to different choices of Basquin slope b. A parameter $b = 4$ is standard for welds. Higher values like $b = 15$ are common for metal sheets. The value $b = 8$ represents an intermediate and is often chosen as a default value to compute the equivalent severity associated to a complex mechanical part (*i.e.* containing assembled metal sheets with welds and edges). As far as the number of cycles before failure remains between 800 000 and 2 000 000 cycles, the equivalent severity has similar values for $b = 4$, $b = 8$ and $b = 15$ (the difference is below 10%). Hence, for these values, the equivalent severity is not very sensitive to the parameter b. Finally, the dark solid line ($b = 100$) allows to visualize the behavior of the fatigue equivalent as b gets higher: the contribution of the cycles with lower intensity decreases and the equivalent severity tends to be identical to the maximum severity of the test.

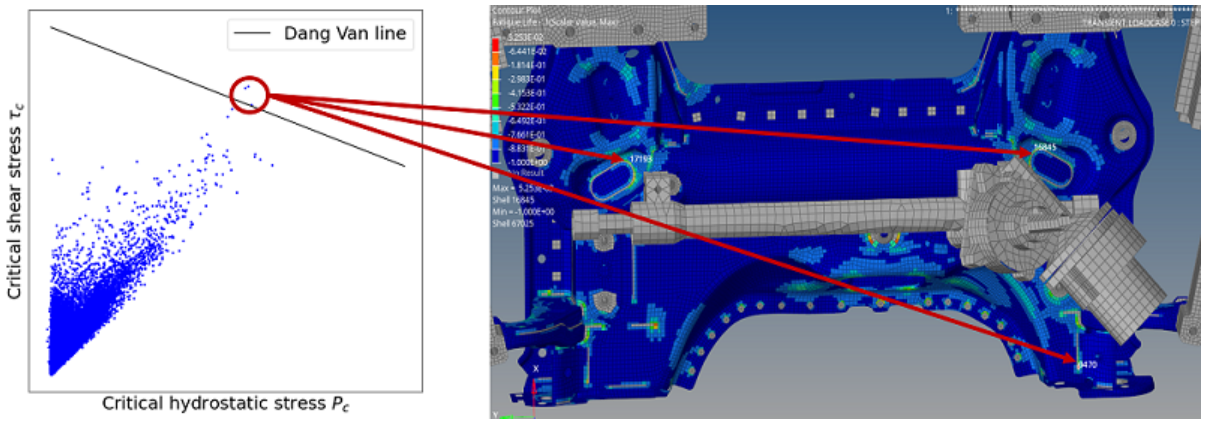

Figure 1.16: Application example of Dang Van criterion on a cradle model under cornering loading (cf. associated test in Fig. 1.11). The left figure represents the position of each element of the model on Dang Van plane. Three points are located over the Dang Van line and thus represent critical points. The right figure represents the FEM where elements are coloured from blue (low CD) to red (high CD). The three critical points are located (red arrows).

1.3.3 Pre-validation of a conception through numerical simulation

When the resistance estimated through fatigue validation tests satisfies the fatigue design requirements, the design proposal is validated. However, if this is not the case, the design needs to be updated. The crack initiations detected during testing show more precisely which zones of the mechanical part need to be strengthened. Once the new conception is ready, a new fatigue test campaign is launched to validate the corrected conception. As the fatigue tests are both long and expensive, engineers are not expected to wait for the fatigue tests to evaluate the resistance of the mechanical part. Indeed, they resort to numerical models to compute the stresses over the part and identify potential crack initiation locations through fatigue criteria.

Hence, prior to testing, the mechanical part is modeled using a Finite Element Model (FEM) and the load cycle is simulated. The FEM allows to calculate the stress cycle on each position of the mechanical part. Then, a fatigue criterion is applied on each position (Ballard et al., 1995). For example, in order to apply Dang Van fatigue criterion, the critical hydrostatic stress P_c and the critical shear stress τ_c are computed for every element of the FEM and represented in a Dang Van diagram. The points located above the line defining the endurance limit are considered critical and thus represent potential weaknesses of the part (cf. Fig. 1.16, left). Alternatively, a danger coefficient CD is calculated on each position on the FEM: locations where CD is positive are critical points of the part. (cf. Fig. 1.16, right) This allows to iterate on the design before launching any test.

1.3.4 Conclusion

The fatigue design of a complex mechanical part is carried out through the Stress-Strength interference method. Knowledge about customer usage allows to define the distribution of severities the part will be exposed to in real situations. In order to satisfy the durability requirements, the fatigue design aims at assessing the resistance of the part. Fatigue validation tests on prototypes are performed in order to estimate the resistance of the part and check that the requirements are met. Prior to that, a pre-validation is performed through FEM and the application of a fatigue criterion on the numerical results. This step is essential in order to identify potential flaws in a design proposal before launching a long and expensive test campaign.

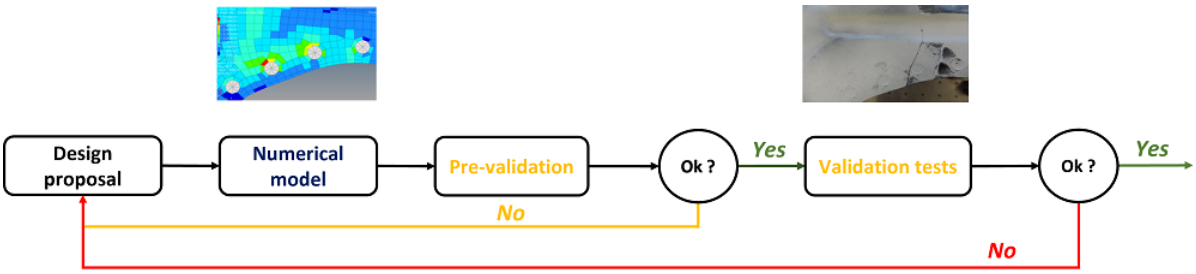

Figure 1.17: Design workflow from design proposal to validation.

1.4 - Issues and objectives

The design of a mechanical component against fatigue includes different stages from the definition of the part to the experimental validation of its functionality and durability. The workflow is represented in Figure 1.17. The first step is the design proposal of the part which consists in choosing the geometry, the materials and the assembly of the part (location of the welded joints...). Then, the mechanical part is modeled numerically using a FEM which allows to simulate the stresses at each location of the part under external loading. A fatigue design criterion is then applied as a pre-validation step, allowing to detect potential design flaws on the design. Once the design is satisfying, fatigue rig tests on prototypes can be launched to validate the resistance of the part.

An ideal workflow would imply only one pass through the design process, and thus only one campaign of validation tests at the end certifying that the durability requirements are met. Unfortunately, this is rarely the case. Numerous roadblocks may occurs during the validation step meaning that the experimental tests fail to validate the design choices. Therefore, the design needs to be corrected, and another test campaign has to be launched to validate the corrected design. These design loops between design proposal and validation strongly delay the development of the part: indeed, the experimental tests are particularly long and expensive.

Numerical computations through FEM offer a significantly faster and cheaper way to assess the behavior of a mechanical part and its resistance to fatigue. Hence, it is preferable to detect design flaws during the pre-validation stage as the iteration between conception and numerical modeling is quick. Unfortunately, the fatigue criterion applied on numerical results fails to identify all the critical elements of the conception. Some weaknesses of the conception are thus discovered only in the validation phase. In other words, the numerical models do not correlate well on experimental fatigue rig tests. The limits of numerical modeling in the characterization of fatigue risks are due to various reasons. First, the fatigue criterion applied on FEM are calibrated on coupon fatigue tests, *i.e.* simple geometries that are not necessarily representative of the diversity and complexity of zones of the designed components. Then, the fatigue predictions given by Dang Van criterion are based on two physical variables (critical hydrostatic stress P_c and critical shear stress τ_c) while the fatigue phenomenon is very complex and depends on many additional parameters (stress concentrations, geometric singularities, manufacturing process...). Finally, the fatigue predictions do not account for various sources of variability: on the one hand, the uncertainties related to the FEM (modeling of welds, mesh size...); on the other hand, the dispersion inherent to fatigue phenomena (cf. Subsection 1.2.1) that can be amplified due to the complexity of the parts and the manufacturing processes.

The objective of car manufacturers is to accelerate the development of vehicles. To do so, the efficiency of the pre-validation stage in anticipating design issues needs to be improved. Hence, Stellantis seeks new design tools to better detect design flaws through numerical simulations.

In this context, the purpose of this thesis is to exploit an alternative source of data consisting

in the history of numerical results and fatigue tests reports on previous designs. The analysis of this database can highlight parameters leading to poor correlations between numerical models and experimental results. More generally, this database can help define additional design tools in order to improve the fatigue predictions through numerical modeling. Hence, the objective of this thesis is to conduct an exploratory analysis of this fatigue database and develop statistical methods allowing to construct new fatigue criteria. These new design tools should help engineers in their conception choices, improve the detection of critical elements in the pre-validation stage, and thus accelerate the whole design workflow.

2

Fatigue database: presentation and first analyses

This chapter focuses on the introduction of Stellantis fatigue database and on first statistical analyses carried out on this database. Section 2.1 details how the fatigue database is built and which features are available. Section 2.2 proposes a methodology to group elements in zones and presents appropriate features to describe a zone. In section 2.3, unsupervised analyses are carried out in order to appraise the variability in the data set and understand correlations between variables. Section 2.4 focuses on an auxiliary data set of welded coupon specimens where we propose a methodology to construct a probabilistic fatigue criterion. In section 2.5, we use classical supervised machine learning methods to estimate statistical fatigue criteria based on the fatigue database. Finally, section 2.6 briefly introduces Positive Unlabeled learning and motivates its use in order to better address the construction of a fatigue criterion.

2.1 - Presentation of the database

In this section, we introduce Stellantis fatigue database relating numerical simulation results to fatigue test results. One of the first steps of conception consists in using a finite element model to represent a mechanical part and simulate its behavior under external loadings. Once the part's design is chosen (shape, size, thickness, materials), validation tests are performed on prototypes in order to check the good resistance against fatigue. Hence, different information are gathered during simulations (Subsection 2.1.1) and tests (Subsection 2.1.2). The fatigue database is constructed by linking data from test reports to numerical results (Subsection 2.1.3).

2.1.1 Simulation results from finite element models

During conception, a mechanical part is modeled by a finite element model (FEM) that consists in a meshing of the part in small elements. Then for a given loading (*i.e.* external forces applied to the part), one can compute the response of the structure to these external forces, described by the distribution of mechanical local stresses on each element of the FEM (cf. Fig. 2.1).

From now on, a case study will consist in a numerical model of a mechanical part along with the simulation results for one type of loading. The database contains a total of 39 case studies with two types of mechanical parts (cradles and cross-members), multiple geometries for each type of part and different types of solicitation (longitudinal, cornering, vertical, transversal).

For each case study, the simulation results give access on each element of the model to different information describing the coordinates of the element, the material, the type of element (sheet, sheet edge, weld) and the local stresses. The mechanical loading of interest is sinusoidal with period $T > 0$. It can be represented as a vector-valued function $(\boldsymbol{F}(t))_{0 \leq t \leq T}$ representing

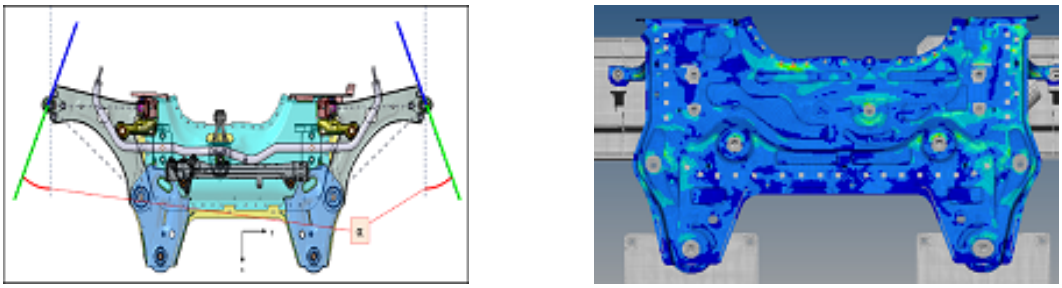

Figure 2.1: On the left, longitudinal loading on a cradle model: instant of maximal loading corresponds to the blue lines, minimum instant to the green lines. On the right, response of the cradle to the longitudinal loading on each element of the structure: the colors represent the value of the first component of the stress tensor at maximum instant (low stress values in blue, higher stress values in light green).

the force applied to the mechanical part:

$$\boldsymbol{F}(t) = \boldsymbol{F}_m + \cos\left(\frac{2\pi t}{T}\right)\boldsymbol{F}_a \ .$$

The simulated loading $\boldsymbol{F}(t)$ corresponds to the *objective customer* F_n (cf. Subsection 1.3.1): the quantities $\boldsymbol{F_m}$ and $\boldsymbol{F_a}$ represent the loading mean and amplitude. For an element e of the finite element model, the simulation results only provide the stress tensors $\boldsymbol{\sigma}_{max}(e)$ and $\boldsymbol{\sigma}_{min}(e)$ representing the local stresses at instants $t = 0$ and $t = T/2$ on the loading cycle. Recall that a stress tensor is a symmetric square matrix (cf. Subsection 1.2.3, Paragraph a). Since the FEM is linear, this information is sufficient to know the stress tensor at any time t. For fatigue applications, we are rather interested by the mean and amplitude tensors that can be calculated as:

$$\boldsymbol{\sigma}_m(e) = \frac{1}{2}\left(\boldsymbol{\sigma}_{max}(e) + \boldsymbol{\sigma}_{min}(e)\right) \text{ and } \boldsymbol{\sigma}_a(e) = \frac{1}{2}\left(\boldsymbol{\sigma}_{max}(e) - \boldsymbol{\sigma}_{min}(e)\right) \ .$$

Besides, the elements are two-dimensional *shell elements* (cf. Cazenave, 2013) which means that they have distinct stress results on the top shell and on the bottom shell. Indeed, there is no interest in looking at any other integration point within the thickness. Hence, each element e has a total of four stress tensors: two mean stress tensors, $\boldsymbol{\sigma}_m^{top}(e)$ and $\boldsymbol{\sigma}_m^{bottom}(e)$; and two amplitude tensors, $\boldsymbol{\sigma}_a^{top}(e)$ and $\boldsymbol{\sigma}_a^{bottom}(e)$.

In particular, Dang Van fatigue criterion introduced in Subsection 1.2.4 uses two mechanical variables from these stress tensors: the critical hydrostatic stress P_c (maximum hydrostatic stress over the loading cycle) and the critical shear stress τ_c (Tresca shear stress calculated over the amplitude stress tensor). We recall that the danger coefficient CD is then defined as:

$$CD = \frac{\alpha\, P_c + \tau_c}{\tau_{mat}} - 1 \ .$$

where α et τ_{mat} are material parameters corresponding to the slope and intercept of Dang Van criterion (cf. Subsection 1.2.4, Paragraph c). As layers top and bottom have different stress tensors, we have two danger coefficient values per element (CD^{top} and CD^{bottom}). The danger coefficient of the element is defined as the maximum of the two. Besides, the corresponding values of P_c and τ_c represent the critical hydrostatic and shear stresses of the element. Other features are accessible through the FEM results, they will be listed in Subsection 2.2.3 and summarized in Table 2.1. The danger coefficient is a measure of *criticality* of an element.

2.1.2 Rig test reports

For each case study, a series of fatigue tests are performed on real prototypes to validate the design proposal. These parts are tested for solicitations identical to those simulated on the numerical model. Usually, between three and seven identical prototypes are tested.

In order to reduce the tests duration, tests are not exactly carried out at the nominal severity "1 F_n" representing the objective customer. The test series follow an accelerated testing protocol in which the severity is incremented gradually (Locati protocol, see Subsection 1.3.2, Paragraph c). Actually, the severity only affects the amplitude of the loading. Hence, under severity αF_n ($\alpha > 0$), the loading $\boldsymbol{F_\alpha}(t)$ is:

$$\boldsymbol{F_\alpha}(t) = \boldsymbol{F}_m + \alpha \cos\left(\frac{2\pi t}{T}\right) \boldsymbol{F}_a \ .$$

By linearity of the numerical model, the amplitude tensors are affected by the same multiplicative constant α while the mean tensors remain the same.

Rig tests reports contain several informations for each prototype tested and each detected crack:

- photo of the crack allowing to identify its location on the part;
- test conditions: initial severity, number of cycles before crack detection and number of cycles before the end of the test.

In the Locati tests analyzed here, the number of cycles between increments along with the value of the increments are fixed: $N_{inc} = 300\,000$ and $f_{inc} = 0.1\,F_n$ (cf. Subsection 1.3.2, Paragraph c).

2.1.3 Including rig test information in the simulation results

As explained above, rig test results provide a set of crack initiations detected on real prototypes. Thanks to the photos of the cracks, it is possible to identify the element or set of elements where a crack initiated and link this information to the numerical models (cf. Fig. 2.2). More particularly, we introduce a binary flag on each element of the numerical model indicating whether or not it is on a crack initiation zone. The initial severity and the total number of cycles of the test are also accessible in the test report and are thus included. This association between numerical models and fatigue tests is the baseline of Stellantis fatigue database.

The process of manually tagging elements on crack zones is not always straightforward. Sometimes, the crack has time to propagate before a photo is taken which makes the identification of the initiation point difficult. This is why, instead of tagging a unique element, we prefer tagging a set of elements in order to be sure that it contains the initiation point of the crack. By doing so, we necessarily introduce labeling errors in the database. This issue will be solved by changing the unit of analysis from elements to groups of elements (zones, cf. Section 2.2).

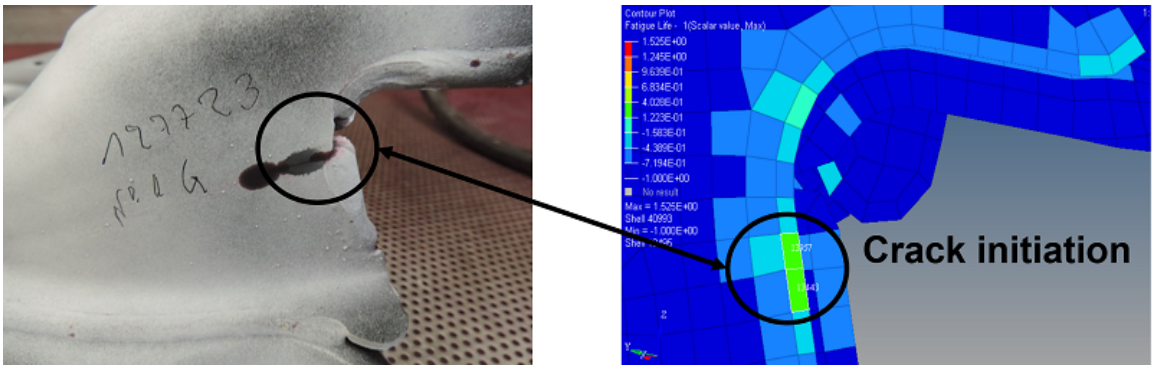

Figure 2.2: Correspondence between numerical models and test results: on the left, crack initiation location from a test report; on the right, the corresponding zone on the FEM. The color scale on the FEM represents the danger coefficient.

2.2 - From elements to groups of elements: definition of zones

The previous section introduced Stellantis fatigue database and explained its construction. In this section, we motivate and present a methodology to change the unit of analysis on the fatigue database: an observation in the database will no longer be an element but a zone, *i.e.* a group of elements (Subsections 2.2.1 and 2.2.2). Zones will be described by appropriate features defined in Subsection 2.2.3.

2.2.1 Grouping elements by zones: motivations

The analysis of the fatigue database raises multiple difficulties. First, the data set is extremely imbalanced. Indeed, each numerical model contains about 10^5 elements and usually no more than 10 crack initiations. Second, the stress field remains low on the majority of the mechanical parts. Figure 2.3 illustrates this fact, showing that for a majority of elements, the stress is weak. Hence we only have a few potentially dangerous locations on the part. Third, the stress field over the mechanical part is continuous. Therefore stress values on close elements are very correlated and provide similar information. Figure 2.3 provides examples of such groups of elements. Finally, as explained in Subsection 2.1.3, it is not always straightforward to locate precisely the element responsible for the crack initiation. It appears that the large majority of cracks initiate and propagate near singularities (edges, holes, corners, welds). Hence, in order to model the risk of crack initiation, it may be relevant to account for features describing the whole zone and not just a single element. All these considerations led us to reduce the number of observations by grouping elements from FEM.

2.2.2 Method for grouping elements

We now describe the method used to build groups of elements. It is based on two main principles. The first one is that the analysis should focus only on relevant zones which means that we will only consider zones with a sufficient level of stress. We rely on the danger coefficient CD (from Dang Van criterion, cf. Subsection 1.2.3) to perform this selection: only elements with a danger coefficient greater or equal to -0.8 are selected. This threshold value is empirically chosen: at the same time small enough so that every tagged element (with detected crack initiation) is selected, and sufficiently high to limit the number of selected points. The second one is that, as fatigue is a local phenomenon, any statistical criterion defined should remain local; hence we limit the radius of each zone to 25 millimeters. This size allows to account for singularities located near critical points. At the same time, it provides flexibility on the location of the crack:

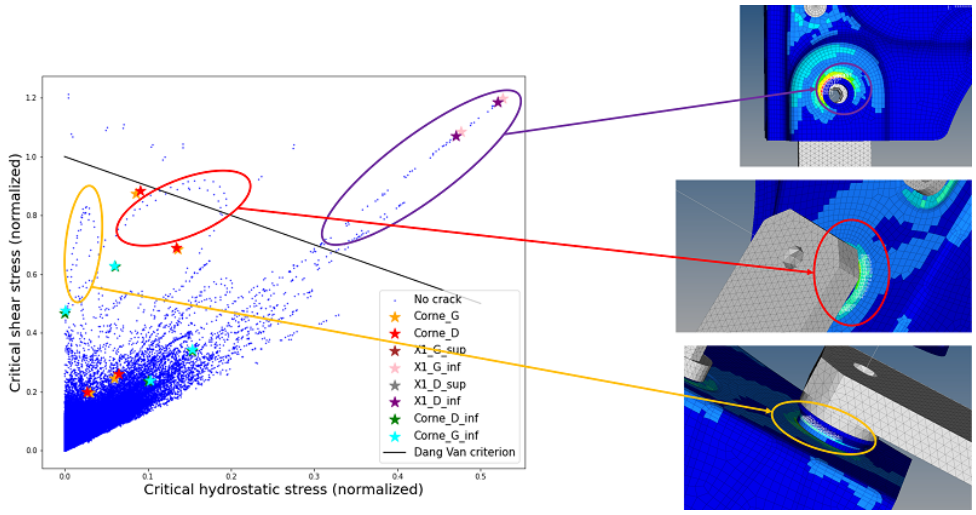

Figure 2.3: Simulation results on a cradle case study: each point represents an element of the model, stars highlight cracks detected during rig tests. Results are represented in Dang Van diagram: X-axis features hydrostatic stress, Y-axis represents critical shear stress. Both quantities are normalized by the material fatigue parameter corresponding to the element. Ellipses represents some groups of neighboring FEM elements.

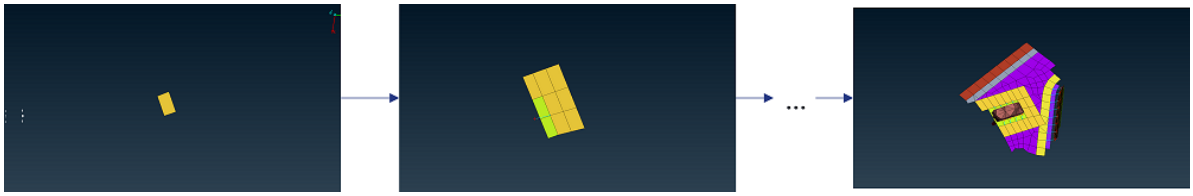

Figure 2.4: Construction of a zone: starting from a selected element with danger coefficient higher than -0.8 (left figure), the zone is iteratively extended by adding neighbors until all elements in a radius of 25 millimeters have been selected (center and right figures).

in other words, if the crack initiation was not precisely located on the numerical model, the zone should cover it. Hence, for each selected element, a zone is built around it by performing a Breadth-First Search (Cormen et al., 2022, Chap. 22) over the FEM graph to find elements at distance less than 25 millimeters. This way, only elements connected to the center through the FEM and at distance less than 25 millimeters are part of the zone. The algorithm is illustrated on an example (cf. Fig. 2.4).

This pre-processing leads to a reduced data set. For instance, on the case study of Fig. 2.3, the number of observations is drastically reduced: from about 83 000 elements to 357 zones. Meanwhile, the number of descriptive features per individual increases as we are now considering groups of elements which gives access to additional information: indeed, instead of having only covariates describing a single element, we now have a set of covariates describing each element of a zone.

2.2.3 Features to describe zones

The unit of analysis has now changed: an individual in the fatigue database does not consist in a single element anymore but in a group of elements representing a zone. Different zones may contain different numbers of elements. We therefore need appropriate features to describe zones.

a. Most critical element of a zone

Since each element of a zone can be described by a set of features and since we want to identify whether or not the zone is critical, a classical idea in fatigue is to consider that a zone is as weak as its weakest element (weakest link theory, cf. Wormsen and Härkegård, 2004). It is then natural to represent a zone by the features of its most critical element. Dang Van danger coefficient provides a way to compare the elements of the zone and the one with the maximum danger coefficient can be considered as the location most likely to initiate. Hence, for each zone, we will consider the descriptive covariates of its most critical element according to Dang Van danger coefficient. As the local stresses on this element are characterized by four stress tensors (cf. Subsection 1.2.3, Paragraph a), we extract the following tensor invariants from each: Von Mises stress (V), Tresca shear stress (τ), hydrostatic stress (P) and stress triaxiality (T). Recall that these so-called tensor invariants do not depend on the basis the tensor is expressed in (cf. Subsection 1.2.3, Paragraph b). Finally, for each feature obtained, we compute the mean and amplitude between top and bottom layers. The mean gives a synthesis of the feature evaluated on the element while the amplitude provides a notion of gradient over the thickness of the sheet.

Example Let us consider the features derived from hydrostatic stress invariant P on an a zone with critical element e. First, the hydrostatic stress is computed on the four stress tensors (mean and amplitude tensors on top and bottom layers): $\boldsymbol{\sigma}_m^{top}(e)$, $\boldsymbol{\sigma}_m^{bottom}(e)$, $\boldsymbol{\sigma}_a^{top}(e)$, $\boldsymbol{\sigma}_a^{bottom}(e)$. This gives access to the mean and amplitude hydrostatic stresses on top and bottom layers: $P_m^{top}(e)$, $P_m^{bottom}(e)$, $P_a^{top}(e)$, $P_a^{bottom}(e)$. Finally, the means and amplitudes over the layers are computed giving the final features:

$$
\begin{aligned}
P_m^m(e) &= \frac{1}{2}\left(P_m^{top}(e) + P_m^{bottom}(e)\right) \\
P_a^m(e) &= \frac{1}{2}\left(P_a^{top}(e) + P_a^{bottom}(e)\right) \\
P_m^a(e) &= \frac{1}{2}\left|P_m^{top}(e) - P_m^{bottom}(e)\right| \\
P_a^a(e) &= \frac{1}{2}\left|P_a^{top}(e) - P_a^{bottom}(e)\right|
\end{aligned}
$$

The absolute value is considered in the amplitude in order to keep the formula symmetric: indeed, an identical element with top and bottom layers exchanged represents the same risk of crack initiation and thus should have the same feature values.

This naming convention for variables will be used throughout the chapter: blue subscript m or a stands for mean or amplitude over the cycle load while red superscript m or a represents the mean or amplitude between top and bottom layers.

Note We decided not to include the two variables P_c and τ_c involved in Dang Van criterion. Indeed, this information is already contained in the considered features: P_c is the maximum hydrostatic stress (sum of the mean and amplitude over the cycle load) and τ_c is Tresca shear stress calculated over the amplitude stress tensor. Both values are calculated on the most critical shell (top or bottom).

The stresses on the most critical element are described by 16 features. In addition, we consider the thickness of the element H (geometric information) and the material fatigue parameter τ_{mat}. This amounts to a total of 18 features (cf. Table 2.1).

b. Features averaged over the zone

The above process for defining features can result in a loss of valuable information about the zone. Hence, in addition to considering the features of the most critical element, we can also compute spatial averages of physical values over the zone or over subparts of the zones. As we have access to the coordinates of the nodes delimiting the elements of the mesh, the spatial averages are easily accessible.

The interest of such features is illustrated on Figure 2.5. In this example, we compare two zones centered around a welded joint. Both are characterized by the same material fatigue parameter at the critical point of the zone (weld). However, the material parameter τ_{mat} averaged over each zone is different. This is due to the fact that the first zone contains other welds in its neighborhood while the second does not. Hence, adding spatial average features offers a richer basis for comparison between the two zones highlighting a geometric difference: while both are centered around welds, the first contains more singularities in its neighborhood.

In the scope of this study, only averages on the whole zone will be considered. Hence, in addition to the features describing the most critical element of a zone, we consider the same set of features averaged over the zone (18 features, cf. Table 2.1).

c. Features specific to welds and edges

More than 90% of the crack instances in the database are located near singularities, more specifically near edges and welds. In order to enhance the characterization of failures, it is interesting to better describe those specific singularities. In particular, the relative orientation of the stresses and the singularity may be relevant for the identification of critical zones.

For each element tagged as weld or edge, we identify a local coordinate system attached to the element: an X-axis parallel to the edge or weld, a Y-axis orthogonal to the singularity (cf. Fig. 2.6). By expressing the stress tensor in this local coordinate system, we obtain stresses oriented with respect to the singularity which better characterize the stress state around the singularity: longitudinal traction, transverse traction and shear stress. These physical quantities are implied in the growth of micro-cracks (Lemaignan, 2012). These features are computed for each of the four stress tensors on the element and with similar operations as in Paragraph 2.2.3.a in order to calculate the mean and amplitude features between top and bottom layers (12 features in total).

These features are computed element-wise and only on weld and edge elements. When considering the whole zone, we only retain the features describing the most critical edge and the most critical weld (24 features). If a zone does not contain any edge or weld, the values along these features are set to 0 by default.

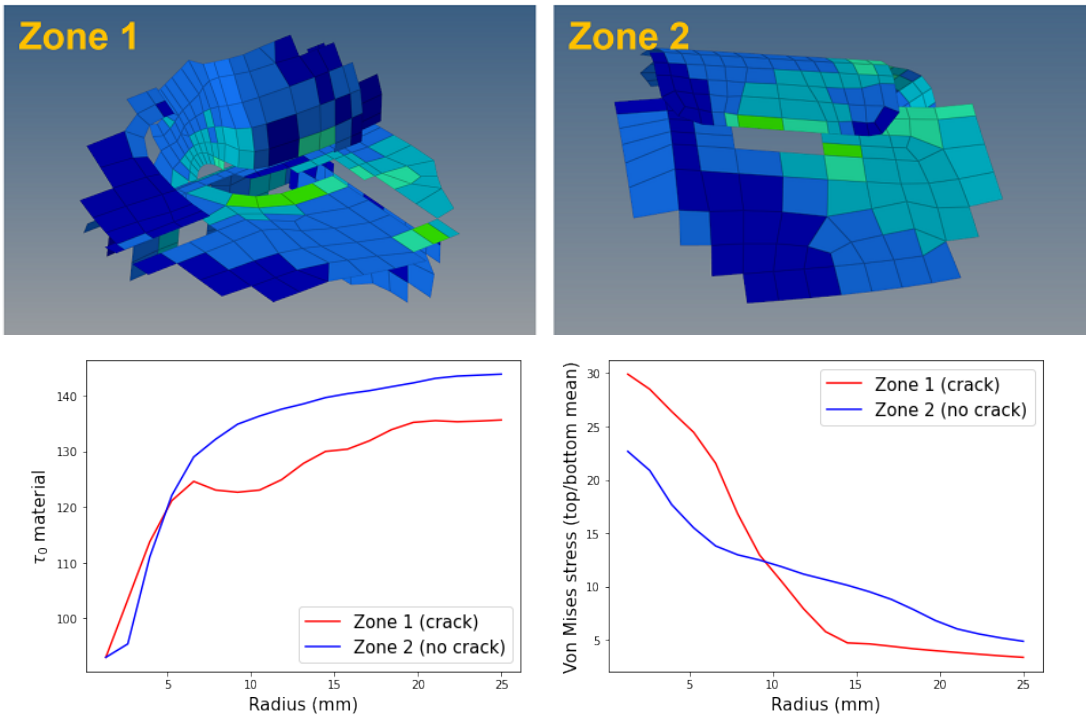

Figure 2.5: Illustration of spatial average features on two zones: material fatigue parameter τ_{mat} (lower left), Von Mises stress (lower right) averaged spatially for two zones. The zones are represented above. The curves represent the spatial averages of these quantities: X-axis denotes the radius over which the spatial mean is calculated. For radius r ranging from 0 to 25, spatial averages are considered over elements located at distance lower than r from the center of the zone. For $r = 0$, only the center is considered while for $r = 25$, the spatial average over the whole zone is considered.

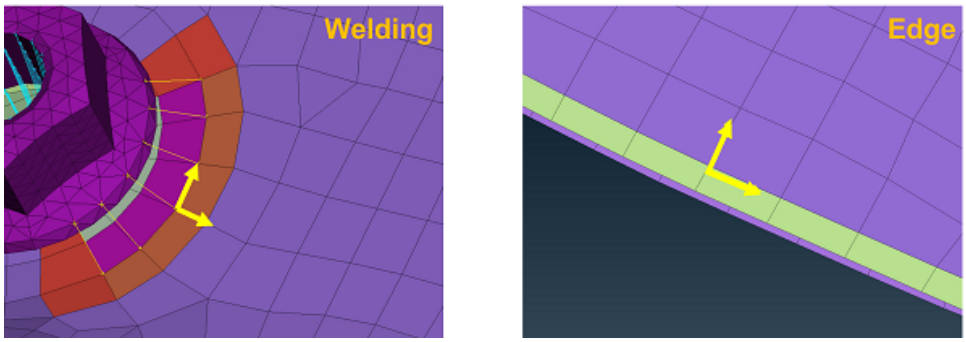

Figure 2.6: Definition of a local coordinate system attached to weld (left) and edge (right) elements

d. Wrapping up

All in all, the fatigue database contains 19 367 zones among which 291 are tagged as crack initiations. Each zone of the database is described by a set of 60 features:

- features describing the most critical element of the zone (18 features);
- features averaged over the zone (18 features);
- features specific to singularities (24 features).

The detailed list of features along with the notations are summarized in Table 2.1. In the next section, we will investigate about potential relations between the defined features through unsupervised analyses.

Table 2.1: List of features and notations. For stress features, blue subscripts a (m) refer to the amplitude (mean) stress tensors over the load cycle while red superscripts a (m) indicate that amplitude (mean) between top and bottom shell is considered. The material and thickness variables are independent from the loading and the shell.

		Mean over cycle load $(\cdot_m)$		**Amplitude over cycle load** $(\cdot_a)$	
	Notation	**Shell mean** $(\cdot^m)$	**Shell amplitude** $(\cdot^a)$	**Shell mean** $(\cdot^m)$	**Shell amplitude** $(\cdot^a)$
Most critical element:					
Von Mises stress	V	V_m^m	V_m^a	V_a^m	V_a^a
Tresca shear stress	τ	τ_m^m	τ_m^a	τ_a^m	τ_a^a
Hydrostatic stress	P	P_m^m	P_m^a	P_a^m	P_a^a
Triaxiality	T	T_m^m	T_m^a	T_a^m	T_a^a
Material parameter	τ_{mat}	τ_{mat}			
Thickness	H	H			
Spatial averages:					
Von Mises stress	$\overline{V}$	$\overline{V_m^m}$	$\overline{V_m^a}$	$\overline{V_a^m}$	$\overline{V_a^a}$
Tresca shear stress	$\overline{\tau}$	$\overline{\tau_m^m}$	$\overline{\tau_m^a}$	$\overline{\tau_a^m}$	$\overline{\tau_a^a}$
Hydrostatic stress	$\overline{P}$	$\overline{P_m^m}$	$\overline{P_m^a}$	$\overline{\tau_a^m}$	$\overline{\tau_a^a}$
Triaxiality	$\overline{T}$	$\overline{T_m^m}$	$\overline{T_m^a}$	$\overline{T_a^m}$	$\overline{T_a^a}$
Material parameter	$\overline{\tau_{mat}}$	$\overline{\tau_{mat}}$			
Thickness	$\overline{H}$	$\overline{H}$			
Weld features					
Longitudinal traction	WL	WL_m^m	WL_m^a	WL_a^m	WL_a^a
Transversal traction	WT	WT_m^m	WT_m^a	WT_a^m	WT_a^a
Shear stress	WS	WS_m^m	WS_m^a	WS_a^m	WS_a^a
Edge features					
Longitudinal traction	EL	EL_m^m	EL_m^a	EL_a^m	EL_a^a
Transversal traction	ET	ET_m^m	ET_m^a	ET_a^m	ET_a^a
Shear stress	ES	ES_m^m	ES_m^a	ES_a^m	ES_a^a

2.3 - Unsupervised analysis

While the mechanical Dang Van fatigue criterion relies on two features to predict whether a zone is critical or not, we defined 60 features describing zones and wish to use them to better identify and characterize critical zones. In this section, we carry out a multivariate analysis of the fatigue data set. Among the 60 available features, many are very correlated: hence we use a dimensional reduction technique (Principal Component Analysis) in order to identify potential interesting directions of observation (Subsection 2.3.1). Then, as some variables report similar information, we will identify groups of variables through feature clustering (Subsection 2.3.2). Finally, noting that covariates can be associated with types of zones, we apply a co-clustering technique which leads to different groups of features and provides simultaneously groups of zones, resulting in useful insights on the data set (Subsection 2.3.3).

2.3.1 Principal Component Analysis

In this subsection, we carry out a Principal Component Analysis (PCA) on the fatigue data set. PCA is a linear dimensionality reduction technique that consists in changing the basis of representation of the data in order to highlight the directions explaining most of the variance (cf. Escofier and Pagès, 1998, Chap. 1).

The fatigue data set is represented as a matrix $\boldsymbol{X}$ of size (n, p) where $n = 19\ 367$ is the number of individuals (*i.e.* zones) and $p = 60$ is the number of features. Columns of $\boldsymbol{X}$ represent variables and rows represent individuals. The vector of observed labels $\boldsymbol{Y}$ (crack initiation or not) is not part of the analysis. It will be only added as an illustrative variable. Prior to the analysis, the features of matrix $\boldsymbol{X}$ are centered and scaled. We keep the same notation $\boldsymbol{X}$ for the standardized matrix. Then the correlation matrix can be calculated as:

$$\boldsymbol{C} = \frac{1}{n}\boldsymbol{X}^T\,\boldsymbol{X}$$

where $\boldsymbol{X}^T$ denotes the transpose of $\boldsymbol{X}$. The correlation matrix $\boldsymbol{C}$ is a symmetric matrix with entries in $[-1, 1]$ and with 1 on the diagonal. Entry $\boldsymbol{C}_{j_1,j_2}$ represents the correlation between features j_1 and j_2.

The correlation matrix is represented on Figure 2.7. It already highlights some couples of highly correlated variables and some structure. For instance, it appears that Von Mises stress (V_q^r) and Tresca shear stress (τ_q^r) are very correlated no matter the quantity considered ($q = a$ or m, $r = a$ or m). This is perfectly understandable from a mechanical point of view as both physical variables convey very similar information. Similarly, it appears that the stresses specific to welds (WL, WT, WS) are correlated.

We then perform the PCA. Let $\lambda_1 \geq ... \geq \lambda_p$ denote the ordered eigenvalues of $\boldsymbol{C}$ (inertia of the principal axes) and $\boldsymbol{u}_1, ..., \boldsymbol{u}_p$ the corresponding eigenvectors (principal vectors). The profile of the eigenvalues (Fig. 2.8) shows that the first ten principal axes represent almost 80% of the total inertia.

We now analyze the projection of the data set along the first principal axes. For $1 \leq k \leq p$, the k^{th} principal component $\boldsymbol{F}_k$ is:

$$\boldsymbol{F}_k = \boldsymbol{X}\,\boldsymbol{u}_k\ .$$

Besides, the contributions of each variable $\boldsymbol{X}_{.,j}$ (j^{th} column of $\boldsymbol{X}$) to the k^{th} principal component can be calculated through the squared cosines:

$$R_{j,k} = \frac{1}{\lambda_k\, n^2}\langle \boldsymbol{X}_{.,j}^T,\, \boldsymbol{F}_k\rangle^2\ .$$

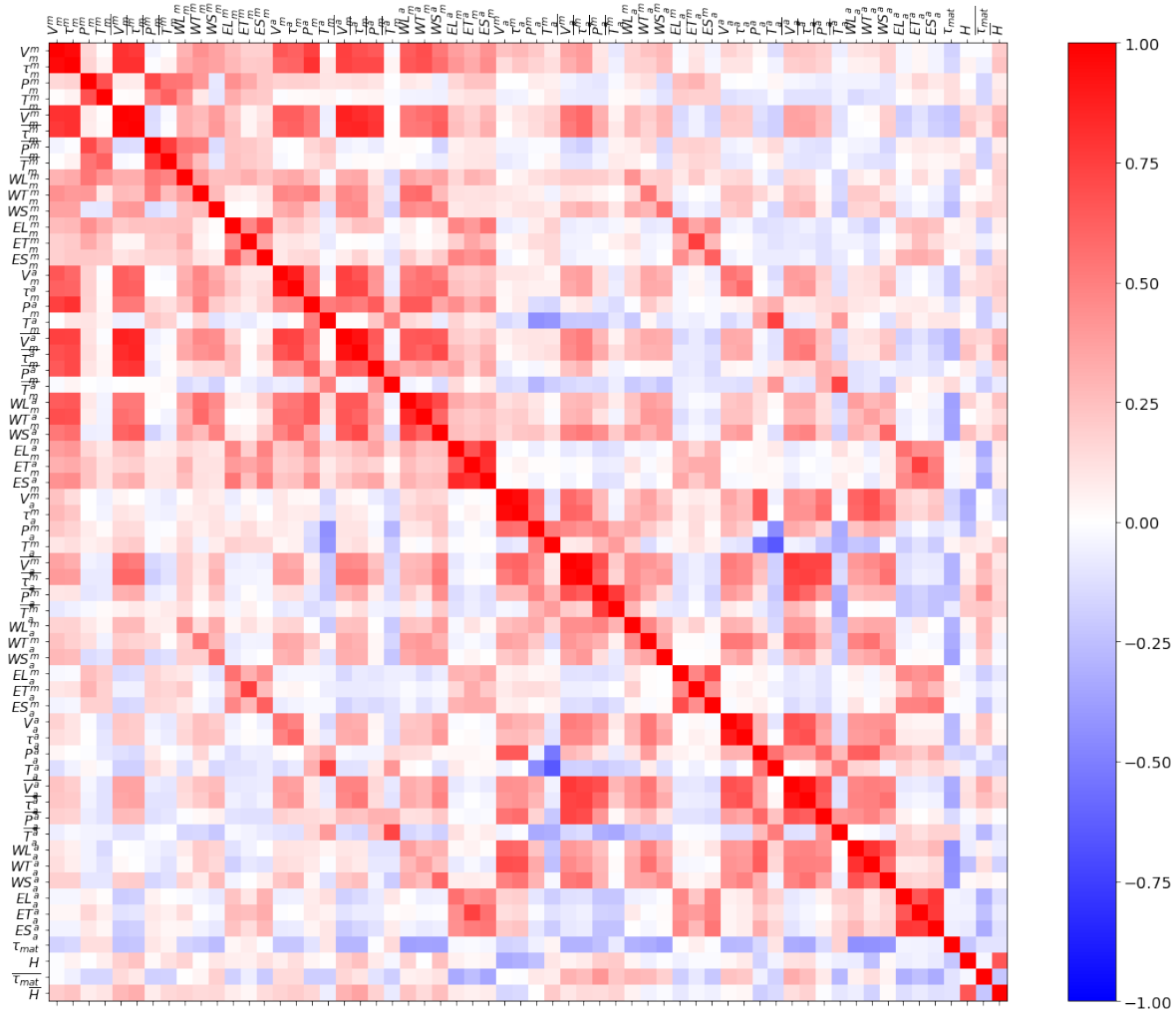

Figure 2.7: Correlation matrix

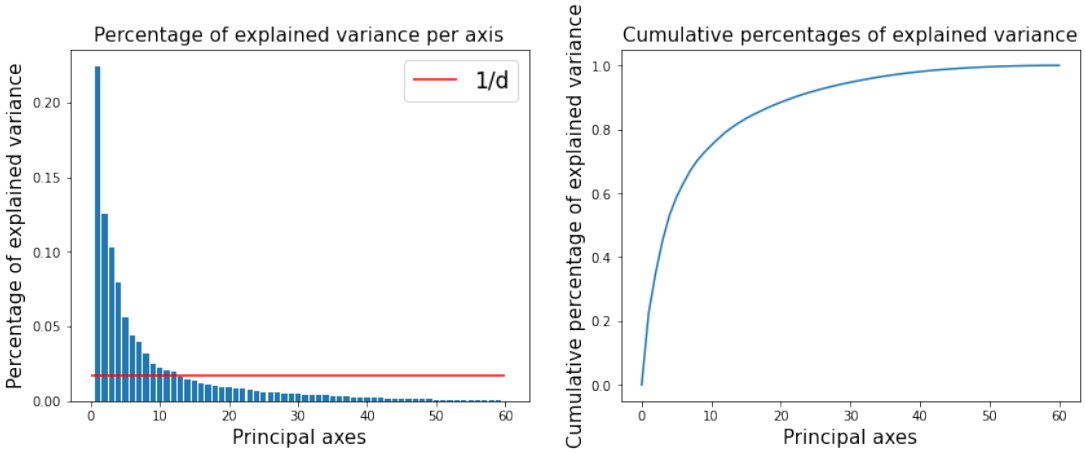

Figure 2.8: Decreasing profile of the singular values in the PCA decomposition (left) and cumulative percentage of explained variance (right).

In the above equation, $\langle v, w \rangle$ denotes the dot product between vectors v and w in $\mathbb{R}^n$:

$$\langle v, w \rangle = \sum_{i=1}^{n} v_i \, w_i \ .$$

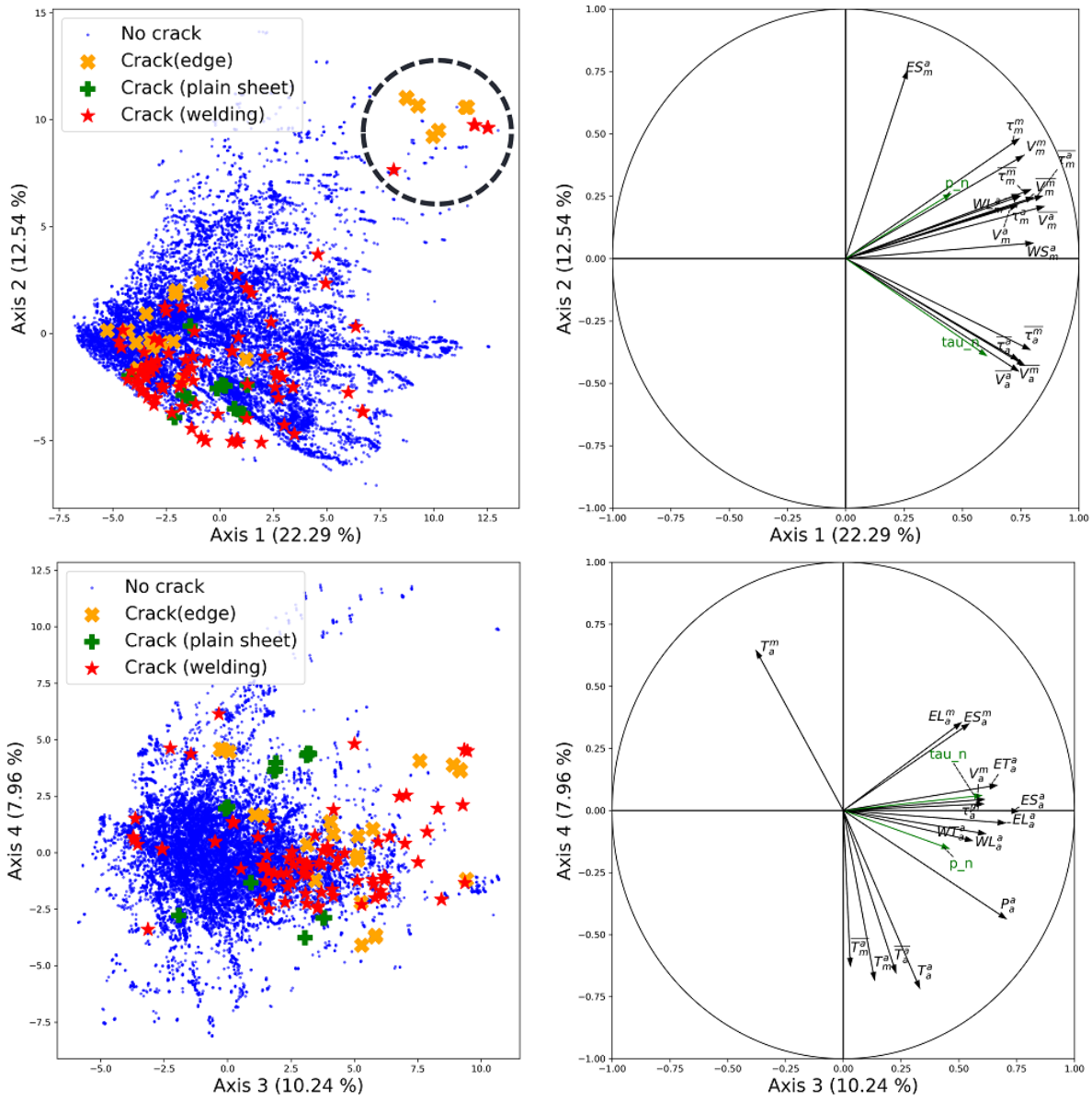

Figure 2.9: Projection of the observations (left) and of the variables (right) on the first principle plane (top figures) and second principal plane (bottom figures). Crack initiations are highlighted: red stars when the critical element of the zone is a weld, orange cross when it is an edge and green "plus" sign when it is a plain metal sheet.

It is then possible to represent the principal components along with their correlation circles: the first two principal planes are represented in Figure 2.9. Due to the great number of variables, only the 15 variables contributing the most to the principal plane are represented. Zones with crack initiations are highlighted along with the type of singularity: red stars when the critical element of the zone is a weld, orange cross when it is an edge and green "plus" sign when it is a plain metal sheet.

The first principal plane (cf. Fig. 2.9) is characterized mainly by two sets of variables: on the one hand, variables related to mean stresses over the loading cycle (on the upper right quadrant of the correlation circle) and spatial averages of amplitude stresses over the zone (lower right quadrant). In particular, a group of crack initiations is very well characterized on this plane (cf. dashed circle on Fig. 2.9): it appears that the critical zones of this group are characterized by important mean stress values. In fact these critical zones belong to cross-member models under

vertical loading. While cradle models are not affected by mean stress values, it is not the case for cross-members that support the weight of the car. The effect of this mean stress is even more visible when the loading is vertical. More generally, the data points in the upper part of the first principal plane belong (for the majority) to cross-member models. The fact that the remaining crack initiations are not well characterized in this first plane means that the spatial average stresses alone (well represented on this first plane) can help identify critical zones but are not sufficient.

The second principal plane (cf. Fig. 2.9) leads to a better characterization of the majority of failure points. Features most correlated to the third principal component are stress invariants commonly used in fatigue criteria. In particular, it is worth noting that Dang Van criterion variables (P_c and τ_c) are correlated to this axis. Besides, variables specific to welds and edges significantly contribute to this third principal axis. The fourth principal axis is also interesting as it is characterized by variables related to the triaxiality and thus accounts for the type of local stresses on the zone.

We can analyze the type of singularity of crack initiations displayed in the PCA results (cf. Fig. 2.9). The figure confirms that the majority of crack initiations is related to welds and only a minority is located on plain metal sheet. Nevertheless, the type of singularity for crack initiation zones is not well characterized on the first two planes.

The first four principal axes explain 53% of the total inertia. The following axes with lower inertia are not represented: they do not provide useful insights and are harder to interpret from a mechanical point of view. All in all, this multivariate analysis shows that there are important correlations among the 60 available features. Besides, not all of them help the characterization of critical zones. Indeed, we saw for instance that the spatial average stresses characterizing the first axis do not really allow the identification of critical zones.

2.3.2 Feature clustering

The correlation matrix and the PCA performed in subsection 2.3.1 already highlight interesting correlations among features. It seems that a lot of features provide very similar information. In order to investigate further these correlations, we use hierarchical clustering. The objective is to identify groups of highly correlated variables.

Hierarchical clustering (Hastie et al., 2009, Chap. 14) is an unsupervised classification technique that does not require the specification of the number of clusters. The algorithm relies on a measure of dissimilarity between groups of observations to iteratively build a tree. In order to cluster individuals (rows), Euclidean distance can be used. In order to cluster standardized variables (columns) we resort to the cosine similarity (correlation) as a similarity measure. We thus have a natural metric to compare pairs of variables.

In fact, hierarchical clustering algorithm needs a *linkage strategy* to compare groups of variables. At each stage, the two groups closest to each other are grouped together. When the distance metric is euclidean, *Ward* linkage is the usual choice. Other strategies (compatible with the cosine similarity metric) exist to define the similarity between groups of variables: *average*, *single* and *complete* linkage measure this similarity as the average, maximum and minimum similarity of pairs of each groups. We choose the average linkage strategy which is a compromise between the two others. In our case though, each linkage strategy yield to quite similar results.

The results of hierarchical feature clustering on the fatigue data set are represented in Figure 2.10 as a dendogram representing the hierarchical structure between the variables. This representation first confirms some preliminary remarks on the high correlation between Von Mises stress (V) and Tresca shear stress (τ). Besides, it is interesting to note that features specific to edges (EL, ET, ES) are grouped together. Finally, the geometric and material information (τ_{mat}, $\overline{\tau_{mat}}$, H, $\overline{H}$) appear close to each other.

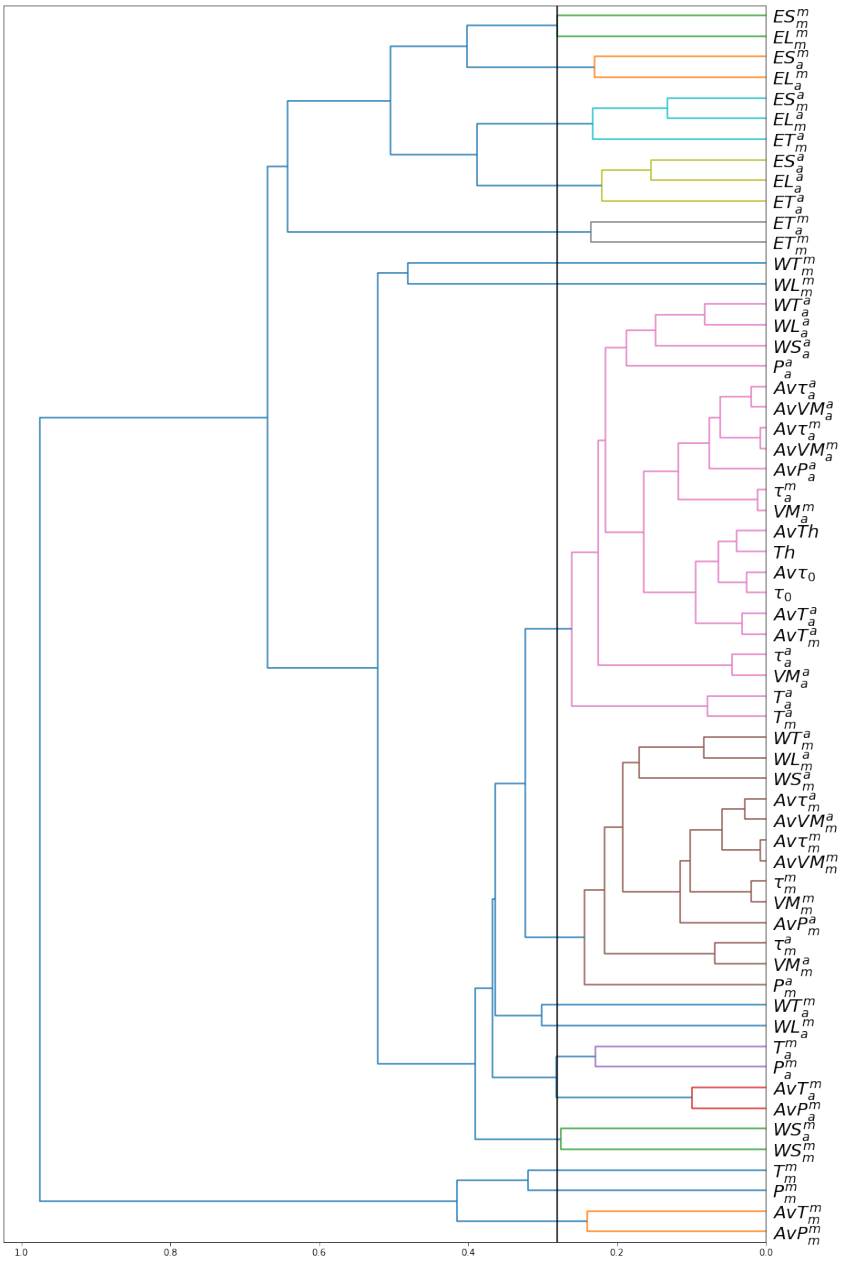

Figure 2.10: Dendogram for feature hierarchical clustering on the fatigue data set using average linkage strategy. Selection of the number of clusters was not performed. However, a threshold indicates the groups obtained by setting the number of clusters to 17 (number of clusters considered later for co-clustering). This will allow to compare it to the partition estimated through co-clustering in Subsection 2.3.3 for a same number of clusters.

While feature clustering study the structure on columns of the data matrix $\boldsymbol{X}$, we also have an heterogeneity between the zones (individuals). In particular, crack initiations happen on different types of singularities (welds, edges, both) with different geometries, hence each type of zone will trigger different physical indicators. It is thus interesting to account simultaneously for the structure on individuals when trying to identify groups of variables. This can be achieved using co-clustering techniques.

2.3.3 Co-clustering

In this subsection, we carry on the unsupervised analysis on the fatigue data set. The objective is to study simultaneously the structure of rows and columns of the data set. We use a Gaussian Latent Block Model to estimate this structure.

a. Gaussian Latent Block Model

Latent Block Model is a model-based approach for co-clustering (Govaert and Nadif, 2013; Keribin et al., 2017). For a specified number of row clusters (K) and column clusters (L), the entries of the matrix $\boldsymbol{X}$ are modeled as a mixture model. More formally, let matrix $\boldsymbol{Z}$ of size (n, K) denote the row cluster memberships and matrix $\boldsymbol{W}$ of size (p, L) the column cluster memberships: $Z_{i,k} = 1$ if individual i belongs to cluster k (else $Z_{i,k} = 0$) and $W_{j,l} = 1$ if variable j belongs to cluster l (else $W_{j,l} = 0$). The model assumes that the variables $(\boldsymbol{Z}_i)_{1\leq n}$ and $(\boldsymbol{W}_j)_{1\leq p}$ are independent, distributed respectively according to multinomial distributions $\mathcal{M}(1, \boldsymbol{\pi})$ and $\mathcal{M}(1, \boldsymbol{\rho})$ where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ and $\boldsymbol{\rho} = (\rho_1, ..., \rho_L)$ satisfying:

$$\sum_{k=1}^{K} \pi_k = \sum_{l=1}^{L} \rho_l = 1 \ .$$

The conditional distributions of the entries $\boldsymbol{X}_{i,j}$ of matrix $\boldsymbol{X}$ given the cluster memberships $(\boldsymbol{Z}, \boldsymbol{W})$ are assumed independent and belong to a same family of parametric probability densities $(f_\theta)_{\theta\in\Theta}$ whose parameter only depends on the row and column cluster memberships:

$$(\boldsymbol{X}_{i,j} \,|\, Z_{i,k}W_{j,l} = 1) \sim f_{\theta_{k,l}} \ .$$

When studying continuous data, a natural choice is the Gaussian family of distributions (Lomet, 2012). In our case, the variance parameter σ^2 is shared by each cluster.

The unknown parameters $\boldsymbol{\theta} = (\theta_{k,l})_{1\leq k\leq K, 1\leq l\leq L}$, σ, $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ are estimated by maximizing the likelihood $\mathcal{L}(\boldsymbol{\theta}, \sigma, \boldsymbol{\pi}, \boldsymbol{\rho}; \boldsymbol{x})$:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma, \boldsymbol{\pi}, \boldsymbol{\rho}; \boldsymbol{x}) = \sum_{\boldsymbol{z},\boldsymbol{w}} \prod_{i,j,k,l} \pi_k^{z_{i,k}} \rho_j^{w_{j,l}} \left(f_{\theta_{k,l}}(\boldsymbol{x}_{i,j})\right)^{z_{i,k}w_{j,l}} \ .$$

The direct maximization of the likelihood is intractable. An approximation of the maximum likelihood estimate can be calculated using a variational approximation (Block Expectation Maximization algorithm, BEM, Govaert and Nadif, 2008). Besides, as stated at the beginning of the paragraph, the number of row and column clusters are both specified. The best model (*i.e.* the best number of clusters) can be chosen using an ICL criterion (Integrated Complete Likelihood, cf. Biernacki et al. 2000 for the general principle and Lomet et al. 2012 for the Gaussian LBM case).

b. Application to the fatigue data set

We use the R package *Blockmodels* (Leger, 2016) in order to perform the estimation for Gaussian LBM along with the model selection. To take into account the exponentially growth of the execution time with the number of rows and columns, the analysis is carried out on a sub-sample of size $n = 2\,000$ of the fatigue data set. The sub-sample contains every crack initiation zones and a random sub-sample of zones without crack initiation.

Figure 2.11 represents the ICL as a function of the total number of clusters (row and column clusters). The ICL reaches its maximum for almost 97 clusters (74 row clusters and 23 column clusters) which is large considering the number of individuals ($n = 2000$) and variables ($p = 60$). We will study the results for a more reasonable number of clusters even if it does not maximize

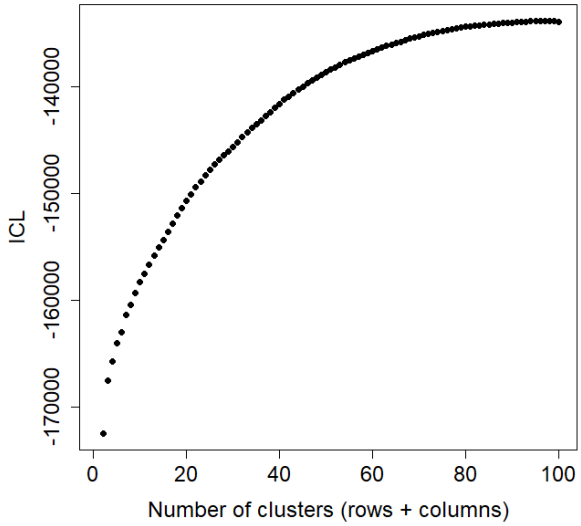

Figure 2.11: ICL as a function of the total number of clusters (row clusters and column clusters).

the ICL: $K = 23$ row clusters and $L = 17$ column clusters. The relative gap between the ICL for this solution and the optimal one is about 5%.

Using the estimated classes, we represent the re-ordered matrix $\boldsymbol{X}$ (cf. Fig. 2.12). The resulting clusters have very heterogeneous sizes: in particular, some row clusters are small (less than 30 individuals). Detailed statistics on the row clusters are reported in Table 2.2: number of individuals per cluster and percentage of crack initiations. The composition of column clusters is presented in Table 2.3. The clusters of variables are quite different from those obtained through feature clustering (cf. Fig. 2.10). In particular, the clusters are more balanced in size and most groups of features obtained are easily interpretable. For instance, it is clear that column clusters 2, 6, 7 and 17 contain features specific to edges and column clusters 10, 12 and 14 are specific to welds (see Table 2.3).

The analysis of co-clusters is insightful as we can relate some clusters of individuals to groups of variables characterizing them. The composition of row and column clusters is detailed in Tables 2.2 and 2.3, where groups of individuals containing more than 10% of crack initiations are highlighted. Using Figure 2.12, we can interpret these clusters:

- Row cluster 1 contains 60% of crack initiations (but only 7% of the total number of crack initiations) and is mainly characterized by feature cluster 6 representing amplitude stresses on edges.
- Row clusters 4 and 19 are well characterized by feature clusters 14 and 15 containing variables specific to welds and standard fatigue criterion features, even though this characterization is less visible for cluster 19 than for cluster 4. Cluster 19 contains 42.5% of the total number of crack initiations.
- Row cluster 11 contains zones with high mean stresses represented by variables of cluster 4.
- Finally, row cluster 17 contains high values for features of cluster 15 representing the shear stress amplitude. This group contains high stress zones on plain metal sheet, without nearby singularities.

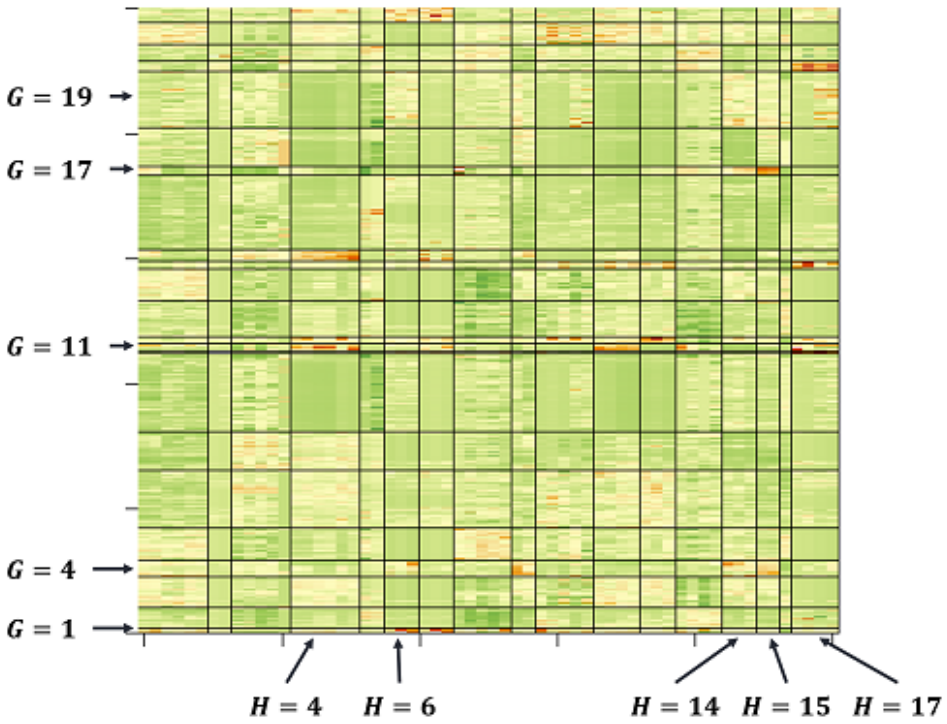

Figure 2.12: Co-clustering results: matrix $\boldsymbol{X}$ re-ordered according to the estimated class memberships (green for low values, red for high values). Row clusters are ordered from bottom to top and column clusters from left to right. Compositions of the clusters are provided in Tables 2.2 and 2.3.

Table 2.2: Descriptive statistics on row clusters. Highlighted clusters are localized on Fig. 2.12.

Cluster	Number of individuals	Number of cracks	Percentage of cracks (%)
1	**15**	**9**	**60.0**
2	70	4	5.7
3	97	4	4.1
4	**56**	**12**	**21.4**
5	104	3	2.9
6	187	4	2.1
7	123	0	0.0
8	256	22	8.6
9	8	0	0.0
10	23	0	0.0
11	**23**	**3**	**13.0**
12	118	7	5.9
13	101	2	2.0
14	25	1	4.0
15	38	0	0.0
16	240	2	0.8
17	**31**	**8**	**25.8**
18	126	1	0.8
19	**181**	**41**	**22.7**
20	35	3	8.6
21	52	0	0.0
22	73	0	0.0
23	44	0	0.0
TOTAL	**2026**	**126**	**6.2**

The co-clustering analysis carried out in this subsection is interesting as an exploratory analysis of the fatigue data set. It allows to better understand the structure of the data set. However, this unsupervised analysis cannot yield a satisfying characterization of the critical

Table 2.3: Co-clustering results: list of features for each column cluster. Highlighted clusters are localized on Fig. 2.12.

Cluster number	List of features
1	$\overline{V_a^m}$, $\overline{\tau_a^m}$, V_a^a, τ_a^a, $\overline{V_a^a}$, $\overline{\tau_a^a}$
2	ET_m^m, ET_a^m
3	T_m^a, $\overline{T_m^a}$, T_a^a, $\overline{T_a^a}$, τ_{mat}
4	$\boldsymbol{V_m^m}$, $\boldsymbol{\tau_m^m}$, $\boldsymbol{\overline{V_m^m}}$, $\boldsymbol{\overline{\tau_m^m}}$, $\boldsymbol{P_m^a}$
5	H, $\overline{H}$
6	$\boldsymbol{EL_a^a}$, $\boldsymbol{ET_a^a}$, $\boldsymbol{ES_a^a}$
7	EL_m^a, ET_m^a, ES_m^a
8	P_a^m, T_a^m, $\overline{P_a^m}$, $\overline{T_a^m}$, WL_a^m
9	P_a^a, $\overline{P_a^a}$
10	WL_m^m, WT_m^m, WS_m^m, WT_a^m, WS_a^m
11	V_m^a, τ_m^a, $\overline{V_m^a}$, $\overline{\tau_m^a}$
12	WL_m^a, WT_m^a, WS_m^a
13	P_m^m, T_m^m, $\overline{P_m^m}$, $\overline{T_m^m}$
14	$\boldsymbol{WL_a^a}$, $\boldsymbol{WT_a^a}$, $\boldsymbol{WS_a^a}$
15	$\boldsymbol{V_a^m}$, $\boldsymbol{\tau_a^m}$
16	$\overline{\tau_{mat}}$
17	$\boldsymbol{EL_m^m}$, $\boldsymbol{ES_m^m}$, $\boldsymbol{EL_a^m}$, $\boldsymbol{ES_a^m}$

zones.

2.3.4 Conclusion

The unsupervised analyses carried on in this section provide useful insights on the fatigue data set. The PCA allows to better understand the distribution of the data by highlighting interesting directions of observations. In particular, some variables carry similar information as those involved in Dang Van criterion. However, other features like triaxiality, edge and weld specific features seem to help the characterization of critical zones. The clustering of features helps to identify groups of correlated variables. Looking at the same time at the structure among individuals and variables allow some interesting interpretations on the fatigue data set. All in all, this exploratory analysis shows that there is heterogeneity among zones and among features: in particular, different types of crack initiations are characterized by different groups of variables.

2.4 - Probabilistic fatigue criterion using welded coupon specimen

In fatigue design, engineers resort to fatigue criteria to identify critical zones on a FEM (cf. Subsection 1.3.3). One of the limits of the fatigue criteria used in fatigue design is that the variability inherent to the tests is poorly addressed. Hence, in this section, we propose a methodology to construct a probabilistic fatigue criterion. The underlying model extends a traditional uniaxial fatigue S-N model to a multiaxial setting. More importantly, the method provides an estimation of the variability essential to account for the randomness of crack initiation.

The methodology is illustrated on Fayard welded coupon specimens (cf. Fayard, 1996). This auxialiary data set contains elementary geometries of welded specimens, *i.e.* small-scale structures containing welded joints. The Dang Van fatigue criterion used at Stellantis for welds is based on this data set. Subsection 2.4.1 introduces Fayard coupon specimens and the corresponding data set and presents the methodology used to estimate the deterministic Dang Van criterion on welds. Then, in Subsection 2.4.2, we construct a probabilistic Dang Van fatigue criterion. Finally, Subsection 2.4.3 presents the results obtained on Fayard coupon specimens.

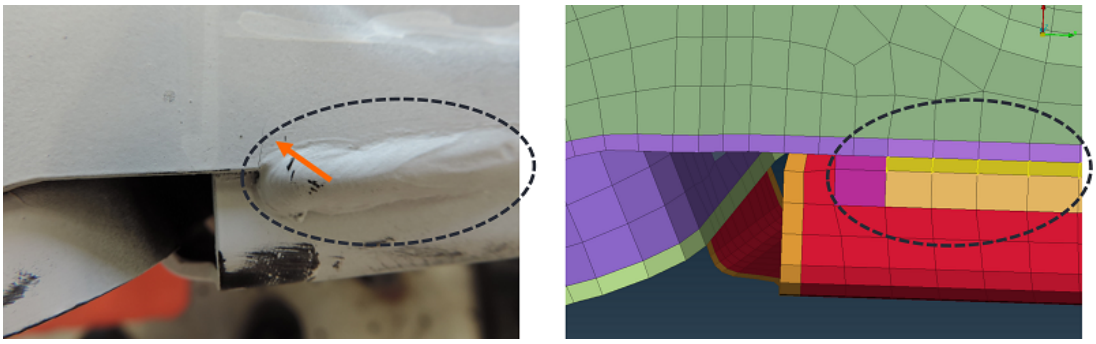

Figure 2.13: Fayard FE structural model for welds: picture of the real weld on the left, its FE modeling on the right. This example is taken from Stellantis fatigue database.

2.4.1 Fayard welded specimens and estimation of a deterministic fatigue criterion

Welded joints are zones of particular interest for the design of automotive mechanical parts as the majority of crack initiations are observed on this type of singularity. During the welding process, under the effect of high temperature, the material warps locally and the metallic microstructure is altered. After cooling, residual stresses and defects appear locally, potentially affecting the material resistance. Hence, the fatigue material properties of these zones are very hard to characterize because an important number of parameters have to be accounted for: geometry of the weld, stress concentration, residual stresses linked to the process effects.

Fayard (1996) introduced a FEM model specific for welds allowing to compute stresses at the welds toes (cf. Paragraph a). In parallel to numerical modeling, multiple fatigue tests on elementary geometries of welds were carried out (cf. Paragraph b), allowing to calibrate an appropriate Dang Van fatigue criterion (cf. Paragraph c).

a. Fayard structural modeling for welds

Fayard (1996) developed a *structural* modeling of welds for complex mechanical parts. In FEM, welds are meshed using this specific methodology (Fig. 2.13). Actually, the set of elements used to model the welds do not represent the exact geometry of the weld (which would be very complex) but allow to account for the structural effects induced by the welding. More particularly, the geometric stresses calculated at the weld toe elements characterize the risk of crack initiation.

b. Fayard welded coupon specimens

Fayard (1996) then proposed a fatigue criterion adapted to welds: it uses the same formalism as Dang Van criterion but requires the estimation of the corresponding fatigue parameters (slope and intercept). For that purpose, Fayard used welded coupon specimens, *i.e.* different elementary geometries of welded components under different types of loading (cf. Fig. 2.14). Each geometry is meshed using the structural modeling for welds and the FEM allows to compute the stress tensors on each element. In parallel, real specimens are tested under similar loading conditions. In this section only, we consider this data set which is built upon the same principles as Stellantis fatigue database.

For each geometry and each loading conditions, we obtain the stress tensors calculated on each element of the structure. Test results provide the severity of the test (force or moment F applied to the structure) and the number of cycles before crack initiation N for each specimen tested. This data set is in many ways less complex than Stellantis fatigue database. First, for each model, there is only one critical element and this element is known. We therefore only consider

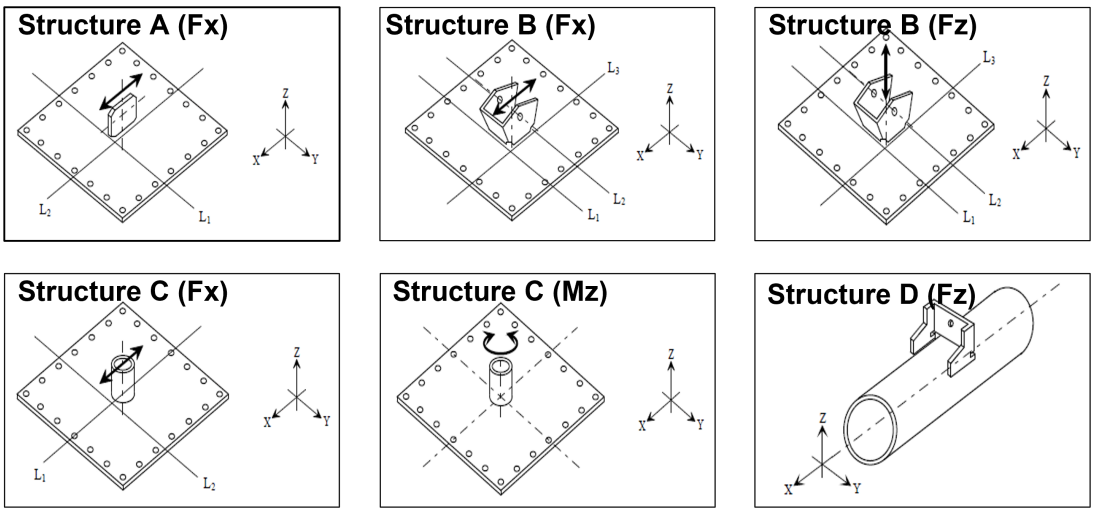

Figure 2.14: Fayard coupon specimens: geometries and loadings

the stresses on the critical element. More particularly, only the variables P_c and τ_c involved in Dang Van criterion are considered. Second, every test is performed until crack initiation, there is no censorship. Finally, unlike the Locati protocol used on prototypes, the testing procedure is standard: the loading amplitude and mean remain constant along the test. Hence, the observed lifetime N and the loading F are sufficient to describe the testing conditions.

c. Fayard methodology for the calibration of a fatigue criterion

For each geometry and loading condition, results can be represented in a Wöhler diagram (Fig. 2.15). Instead of a traditional S-N diagram where the Y-axis represents the stress (cf. Subsection 1.1.2), here the Y-axis features the load F. The principle, however, remains identical. Different loading ratios are considered (see Subsection 1.1.2 for the definition of the loading ratio R). For each geometry, loading type and ratio, a Basquin S-N model is used to represent the fatigue lifetime N as a function of the loading F (cf. Fig. 2.16, left for the case of Structure A under fully reversed loading, $R = -1$).

Each one of the 11 estimated S-N curves leads to an estimate of the fatigue limit F_{lim} at $N = 10^6$ cycles. The FE models are then used to calculate the corresponding values of critical hydrostatic stress P_c and shear stress τ_c for this loading $F = F_{lim}$. This is illustrated in Figure 2.16 for the case of welded structure A under fully reversed loading ($R = -1$).

For each structure, loading type and ratio, the same methodology is applied in order to obtain the fatigue limits and critical stresses through FEM. The 11 points obtained are than reported on a Dang Van diagram, and a linear regression is performed to estimate the final criterion (cf. Fig. 2.17).

The model used by Fayard contains a lot of parameters compared to the number of observations ($n = 144$): 22 parameters for the estimation of S-N curves (2 for each), 2 parameters (slope and intercept) in the final linear regression. In particular, the estimation of Basquin slope in S-N models is impossible in situations where all the experiments are carried out for the same loading level (for instance, for structure A with load ratio $R = 0$). In that case, Fayard assumed the slope to be identical to the same tests with another load ratio R for which it can be properly estimated. Besides, the estimation is performed in two distinct stages: first the estimation of the S-N curves, then the estimation of the fatigue criterion using the estimated fatigue lifetimes at first stage. In particular, the uncertainties on the fatigue lives estimations are not accounted for in the second stage. As a result, we do not have a proper estimate of the variance of the estimated fatigue criterion. Our objective is to construct a probabilistic fatigue criterion that

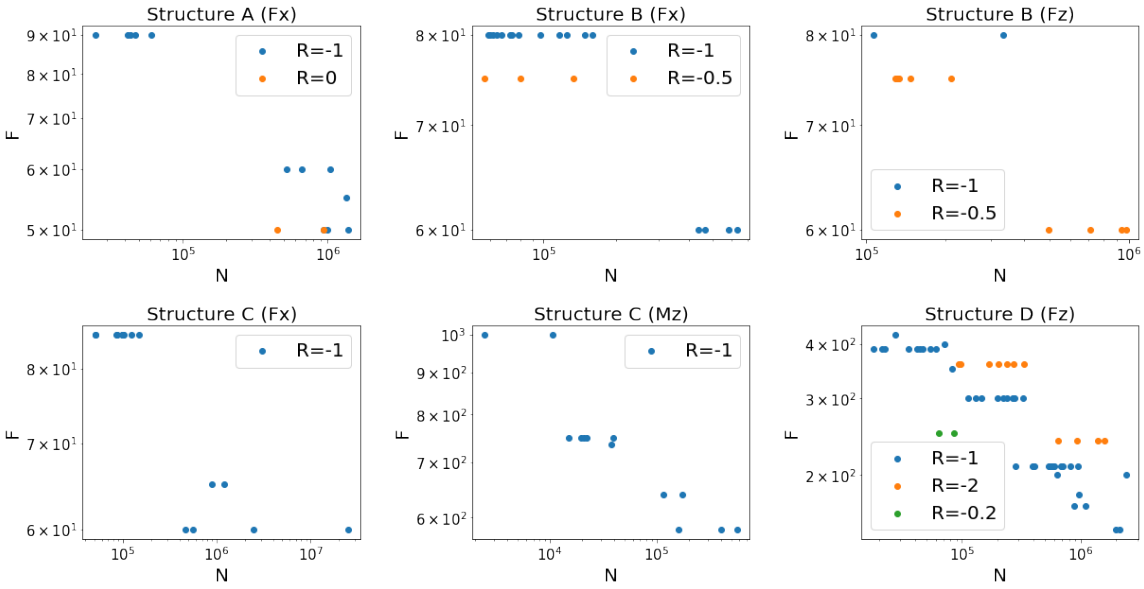

Figure 2.15: Experimental results on Fayard coupon specimens in S-N diagrams. The different colors indicate different load ratios.

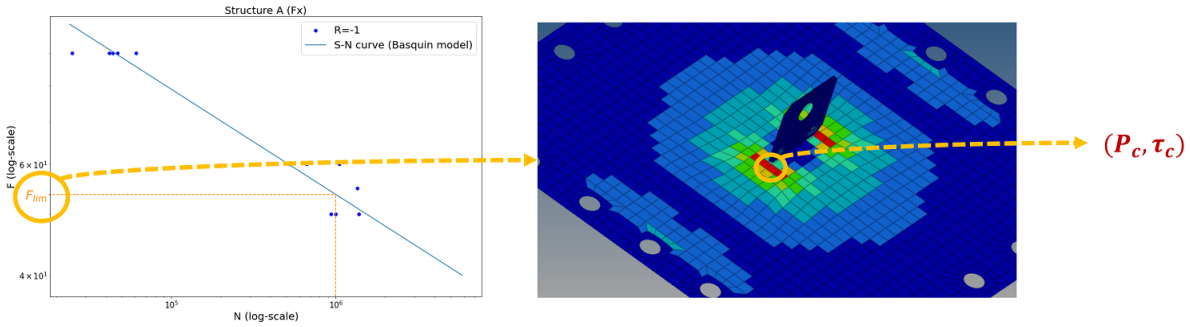

Figure 2.16: Identification of critical stresses (P_c, τ_c) for welded structure A under Fx loading. The fatigue limit F_{lim} is identified on the S-N curve (left plot). Then, the FEM (right figure) allows to calculate the stresses at the critical point (weld toe) under loading F_{lim}.

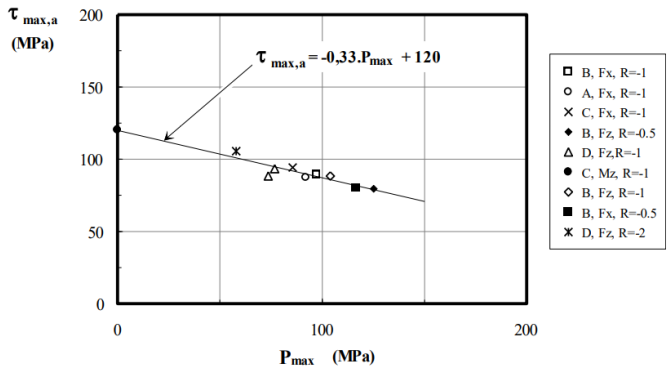

Figure 2.17: Criterion identified by Fayard, figure from Fayard (1996)

estimates this variability.

2.4.2 Construction of a probabilistic Dang Van criterion

We now seek to construct a probabilistic fatigue criterion based on Dang Van variables: critical hydrostatic and shear stresses (P_c, τ_c). In other words, given some point with coordinates (P_c, τ_c) in Dang Van plane, the criterion should output a probability for this point to be critical. This problem of accounting for the scattering of fatigue results was addressed in the literature. In uni-axial loading settings, S-N models account for this variability. Castillo and Fernández-Canteli (2009) and Fouchereau (2014) introduced probabilistic modeling of S-N curves. In the multiaxial setting, Sghaier et al. (2007) proposed a methodology to construct a probabilistic Crossland fatigue criterion: as the material parameters are identified using two types of loadings (bending and torsion), the variability of the estimates of the two fatigue limits are modeled and propagated in the final criterion. Other works consider an extension of S-N curves beyond the uni-axial setting by replacing the uni-axial stress by a multiaxial damage parameter (cf. Susmel and Lazzarin, 2002; Correia et al., 2017). In the context of Dang Van fatigue criterion, Roux et al. (2014) used a linear combination of the critical hydrostatic stress P_c and shear stress τ_c as a multiaxial damage parameter. Our approach also consider such a linear combination in a generalized S-N model. The originality is that the regression coefficient involved in this linear combination (slope of Dang Van criterion) is unknown and thus jointly estimated with the other fatigue parameters.

In the uniaxial setting, Basquin model can be used to relate the loading amplitude to the number of cycles to failure: $F \times N^b$ is assumed constant where F is the loading, N the lifetime and b Basquin slope. However, the different welded geometries are tested under different loadings that are not comparable: for structure A, the loading is measured as a force whereas the structure C is subjected to a moment. Hence, we cannot have a general model relating the loading F to the lifetime N covering every geometries, loading types and ratio. Instead, we directly rely on the local stresses at the critical spots of the specimens. More particularly, Basquin model states that $N \times S^b$ is constant where b is Basquin slope and S is the stress amplitude. Following Susmel and Lazzarin (2002) and Correia et al. (2017), the uniaxial stress S can be replaced by a multiaxial invariant. As Dang Van criterion suggests that the risk of crack initiation depends on a linear combination of critical hydrostatic stress (P_c) and critical shear stress (τ_c), S can be replaced by $\alpha\, P_c + \tau_c$, where α is the unknown slope of Dang Van criterion. Then, by assuming that the variations of $\log(N)$ are Gaussian, we have the following model:

$$\log(N)\,|_{P_c,\tau_c} = a - b\,\log(\alpha\, P_c + \tau_c) + \sigma\,\varepsilon\ . \tag{2.1}$$

In the above equation, $\theta = (a, b, \alpha, \sigma)$ is the unknown parameter and ε is a standardized Gaussian noise. The components of θ represent physical fatigue parameters:

- a is related to the intercept of Dang Van criterion;
- b is Basquin slope, an important fatigue parameter;
- α is the slope of Dang Van criterion;
- σ represents the variability of the fatigue lifetime and will be related to the variability of the fatigue criterion.

Remark The load ratio R does not appear in the model of Equation 2.1 because it is already accounted for in the local stresses (P_c, τ_c). Indeed, the numerical FEM simulation is performed for each structure and each load type and ratio: in particular, different load ratio leads to different stresses (P_c, τ_c).

Table 2.4: Maximum likelihood estimates with the 95% asymptotic confidence intervals.

	inf 95%	mean	sup 95%
a	34.37	37.21	40.06
b	4.39	4.94	5.48
α	0.26	0.35	0.44
σ	0.66	0.75	0.83

The model in Equation 2.1 directly leads to a probabilistic fatigue criterion. An element with stresses (P_c, τ_c) is considered critical if its lifetime is below 10^6 cycles. This happens with probability:

$$p\left(P_c, \tau_c\right) = \mathbb{P}\left(N \leq 10^6 \,|\, P_c, \tau_c\right) = \Phi\left(\frac{\log\left(10^6\right) - a + b\,\log\left(\alpha\, P_c + \tau_c\right)}{\sigma}\right) \tag{2.2}$$

where Φ denotes the cumulative distribution function of the standardized normal distribution. The probability $p\left(P_c, \tau_c\right)$ is a function of the parameter θ. Hence, by plugging in an estimate of the unknown parameter $\widehat{\theta} = (\widehat{a}, \widehat{b}, \widehat{\alpha}, \widehat{\sigma})$ in Eq. 2.2, we obtain an estimate of the probability for a zone with stresses (P_c, τ_c) to be critical:

$$\widehat{p}\left(P_c, \tau_c\right) = \Phi\left(\frac{\log\left(10^6\right) - \widehat{a} + \widehat{b}\,\log\left(\widehat{\alpha}\, P_c + \tau_c\right)}{\widehat{\sigma}}\right) .$$

2.4.3 Results on Fayard coupon specimens

We now use the fatigue results on Fayard coupon specimens to estimate the parameter θ characterizing the probabilistic fatigue criterion of Equation 2.1. The data set consists in $n = 144$ independent observations of lifetimes $(N_i)_{1 \leq i \leq n}$ for given values of critical hydrostatic and shear stresses $(P_{c,i}, \tau_{c,i})_{1 \leq i \leq n}$. The likelihood $\ell(\theta)$ can be expressed as:

$$\ell(\theta) = -\frac{n}{2}\log(2\,\pi\,\sigma^2) - \sum_{i=1}^{n} \frac{\left[\log(N_i) - a + b\log(\alpha P_{c,i} + \tau_{c,i})\right]^2}{2\,\sigma^2} . \tag{2.3}$$

An estimator $\widehat{\theta}$ of the parameter is obtained through maximum likelihood. Since the maximization of the log-likelihood in Eq. 2.3 cannot be solved analytically, a numerical approximation is found using Newton's method. The estimates are presented in Table 2.4 along with the 95% asymptotic confidence intervals. The estimate of α is very close to the slope identified by Fayard. In addition, the value of Basquin fatigue parameter b is perfectly standard for welds (cf. Bergamo et al., 2017). Finally, we provide an estimate of the variability through σ.

The fit can be visualized in a "S-N like" diagram by representing the lifetime on the x-axis and $\widehat{\alpha}\, p_c + \tau_c$ on the y-axis (cf. Fig. 2.18). The probabilistic Dang Van criterion is represented in Fig. 2.19. This criterion is no longer represented as a line in Dang Van plane but as a probability field. Coupon specimens test results are also represented in Dang Van diagram of Fig. 2.19 indicating those that failed before 10^6 cycles and those that failed after. These observations agree with the estimated criterion. In addition, the deterministic criterion identified by Fayard is represented. It is very close to the probabilistic criterion at the level of probability 0.5 (white part of the diagram in Fig. 2.19). The added value of our methodology is that it provides confidence intervals on the estimated parameters. Besides, the variability of the criterion, represented by σ, is also estimated.

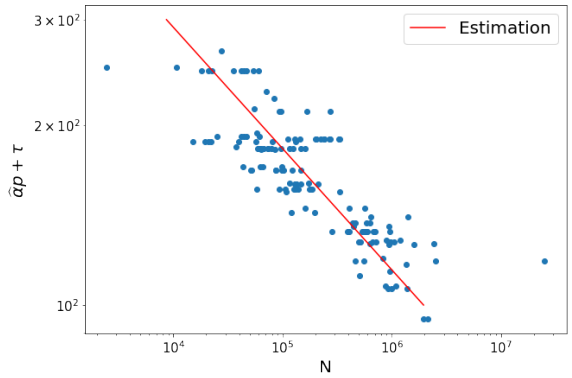

Figure 2.18: Estimated regression line in an S-N like diagram (log scale on both axes): y axis represents the linear combination of P_c et τ_C with the estimated slope $\widehat{\alpha}$.

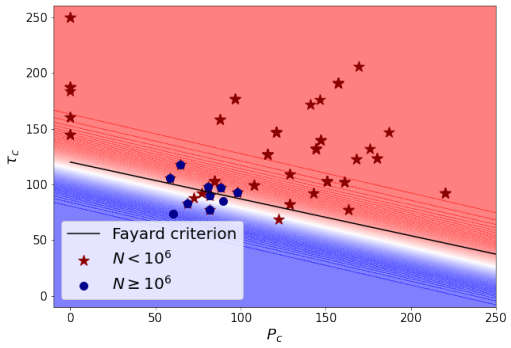

Figure 2.19: Illustration of probabilistic Dang Van criterion: blue (red) represents low (high) probability of failure before 10^6 cycles. Blue points (red stars) represent tests with lifetime superior (inferior) to 10^6 cycles.

2.4.4 Conclusion

In this section, we introduced a methodology to construct a multiaxial probabilistic fatigue criterion based on an S-N regression model. Using the model on the lifetime N of the zone, one can estimate the probability for N to be lower than 10^6 and thus for the zone to be critical. This criterion only involves two variables (P_c, τ_c). If this criterion is well adapted to the prediction of crack initiations on simple geometries like Fayard coupon specimens, it unfortunately generalizes poorly to the fatigue database which contains welds with more complex geometries. Figure 2.20 represents the numerical results from Stellantis fatigue database (only welded zones) on top of the probabilistic fatigue criterion. It clearly shows that multiple critical welds are poorly identified by the criterion.

In order to identify a fatigue criterion better adapted to the complexity of mechanical parts, we need to estimate this criterion on the fatigue database directly. Besides, as there are very different types of zones, it may be beneficial to consider additional features to improve the predictions.

2.5 - Supervised classification methods for fatigue crack predictions

In this section, we seek to estimate a fatigue criterion directly using the fatigue database. This problem can be viewed as a supervised machine learning problem where the goal is to find a binary classification rule able to discriminate crack initiation zones from the others (Subsection 2.5.1). Different classification methods are considered, listed in Subsection 2.5.2. Subsection 2.5.3 defines appropriate performance metrics for classification under class imbalance. Finally, the results on the fatigue database are presented in Subsection 2.5.4.

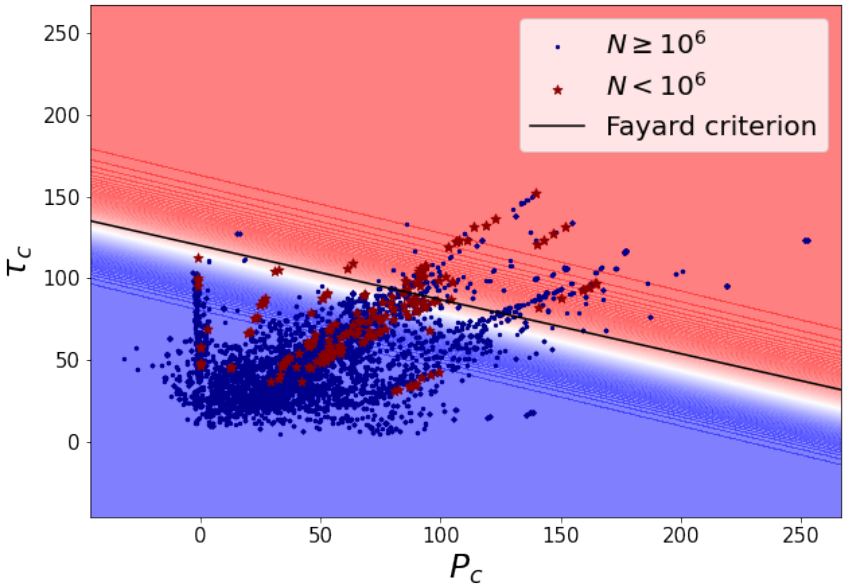

Figure 2.20: Probabilistic Dang Van criterion and welded zones in the fatigue database. Each point represents a zone of Stellantis fatigue database where the critical element is a weld. The position of the points depend on the critical hydrostatic and shear stresses calculated through FEM: red stars (blue points) represent zones with lifetime below (over) 10^6 cycles. The background color represents the probabilistic criterion estimated in Subsection 2.4.3.

2.5.1 Fatigue crack prediction as a supervised classification task

The fatigue database consists in a set of $n = 19\,367$ zones $(X_i, Y_i)_{1 \leq i \leq n}$ where $X_i \in \mathbb{R}^p$ is a covariate vector of size $p = 60$ describing the geometry and the stresses on the zone, and Y_i is the binary output indicating whether $(Y = 1)$ or not $(Y = 0)$ a crack initiated on the zone.

The objective is, for a new zone with covariates $x \in \mathbb{R}^p$, to predict whether it can fail $(Y = 1)$ or not $(Y = 0)$ before 10^6 cycles under the customer objective load. On the one hand, we do not want the criterion to fail in identifying zones that could break: indeed, in that case, the conception would not pass the validation tests and thus require an additional iteration. On the other hand, the criterion should not be too strict in order to avoid useless reinforcements. As Y is a binary target, this problem is a *binary classification task*. This statistical criterion will be estimated using the fatigue data set.

This approach offers several advantages compared to traditional fatigue criteria. First, it relies on fatigue data from complex geometries and loadings to define the fatigue criterion: hence, we can expect this statistical criterion to be better suited for the fatigue design of complex components. Instead, traditional fatigue criteria are usually calibrated on simple specimens (coupon tests) and thus tend to generalize poorly to complex geometries (cf. Section 2.4 for welds). Second, as explained in the Section 2.3, the fatigue database contains 60 descriptive features for each zone which is more informative than the two features used in Dang Van probabilistic criterion of Section 2.4. Hence, the classification method can account for this additional information. Finally, the probabilistic interpretation of the criterion will remain: indeed, most classification methods provide an estimate of the probability for a new instance to be positive $(Y = 1)$. However, contrary to the probabilistic Dang Van criterion, these purely

statistical criteria do not have any mechanical founding.

In the application of binary classification methods to fatigue criterion estimation, we are faced with multiple challenges. First, despite the pre-processing of Section 2.2 to gather elements by zones, the fatigue data set remains imbalanced: only 1.5% of the observations are positive ($Y = 1$). This imbalance issue will be taken into account in the training phase and in the evaluation of performances (cf. 2.5.3) by using appropriate metrics. Second, if the number of features ($p = 60$) is low compared to the size of the data set ($n = 19\,367$), it is of the same order as the number of observations from the positive class (only $n_+ = 291$ crack initiations). Finally, the fatigue data set is affected by an asymmetric label noise: while a crack initiation asserts the criticality of a zone, some unbroken zones may be critical due to the duration of the test (the test may have been stopped before crack initiation) and the randomness of fatigue crack initiation. This issue will be ignored in this section: it will be the subject of further developments in Chapters 3 and 4.

2.5.2 Supervised classification methods

In this subsection, we give a brief overview of the classification methods we will use. Let us first describe the general setting for supervised classification.

The objective of supervised binary classification is to find a binary function $g^* : \mathbb{R}^p \longrightarrow \{0, 1\}$ minimizing the prediction errors:

$$g^* \in \underset{g\,:\,\mathbb{R}^p \longrightarrow \{0,1\}}{\operatorname{Arginf}} \; \mathbb{P}\left(g(X) \neq Y\right) \ .$$

The optimal classifier is known as Bayes classifier and can be expressed in terms of $\mathbb{P}$:

$$g^*(x) = \mathbb{1}_{\mathbb{P}(Y=1\,|\,X=x) \geq \frac{1}{2}} \ .$$

Therefore, a classifier can be obtained by estimating the function $h(x) = \mathbb{P}(Y = 1|X = x)$ representing the conditional probability of Y given x, and setting a threshold t so that the predicted class is 1 if $h(x) \geq t$, else 0 (the natural threshold being $t = 1/2$).

As the distribution of the data is unknown, we rely on a training sample $(X_i, Y_i)_{1 \leq i \leq n}$ to built an estimator $\widehat{g}$ of the classification rule. For that purpose, there exists multiple classification methods. Each rely on two main ingredients.

1. A model $\mathcal{G}$, *i.e.* a set of classifiers assumed to contain the optimal classifier (or at least close to the optimal one). This set can be restricted to linear classifiers (logistic regression, linear Support Vector Machine), piece-wise constant functions (tree-based methods) or more complex sets of functions (Support Vector Machine with Gaussian kernel).

2. A loss function $\ell\left(g(x), y\right)$ measuring the quality of a class prediction $g(x)$ (or prediction probability $h(x)$) given the true class y (logistic loss for logistic regression, hinge loss for Support Vector Machine...).

The estimator $\widehat{g}$ is built by minimizing the risk $R(g) = \mathbb{E}\left[\ell\left(g(X), Y\right)\right]$ estimated by the empirical mean:

$$\widehat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(g(X_i), Y_i\right)$$

over the set of classifiers $\mathcal{G}$. Different methods will be considered in this section:

- Logistic Regression with Lasso regularization (LR);
- Linear and Quadratic Discriminant Analysis (LDA, QDA);
- Support Vector Machine (SVM) with linear and Gaussian kernels;
- Random Forests (RF).

a. Logistic Regression with Lasso regularization (LR)

Logistic Regression (cf. Hastie et al., 2009, Chapter 4) is a linear classification model. Observations $(Y_i)_{1\leq i\leq n}$ are assumed independent and Y_i is assumed to follow a Bernoulli distributions with parameter $p(X_i)$. The probability p is a function of a linear regression on x with coefficients β and intercept β_0. A natural choice is the logistic function which leads to the following function p:

$$p(x) = \frac{1}{1+e^{-\beta_0-\beta^T x}},$$

where $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ are unknown parameters.

The criterion minimized is the opposite of the log-likelihood, also known as *logistic loss*. As the dimension is large, a L^1 penalization is added in order to force the estimated vector $\widehat{\beta}$ to be sparse. Hence, the parameters are estimated by solving the following optimization problem:

$$(\beta_0,\beta) \in \underset{\beta_0,\beta}{\text{Arginf}}\, \lambda\,||\beta||_1 + \sum_{i=1}^{n}\left[Y_i\,\log\left(1+e^{-\beta_0-\beta^T X_i}\right) + (1-Y_i)\log\left(1+e^{\beta_0+\beta^T X_i}\right)\right],$$

where $||\cdot||_1$ denotes the L^1 norm and λ is an hyper-parameter which can be selected using cross-validation (cf. Hastie et al., 2009, Chap. 7).

b. Linear and Quadratic Discriminant Analysis (LDA, QDA)

In QDA, the conditional distributions of X given $Y=0$ and $Y=1$ are assumed to be Gaussian with parameters (μ_0,Σ_0) and (μ_1,Σ_1) (cf. Hastie et al., 2009, Chapter 4). The class prior $\pi = \mathbb{P}(Y=1)$ is also unknown. The parameter $\theta = (\pi,\mu_0,\Sigma_0,\mu_1,\Sigma_1)$ is estimated through maximum likelihood:

$$\widehat{\theta} \in \underset{\theta}{\text{Argsup}} \sum_{i=1}^{n} Y_i \log\left(\pi\, f_{\mu_1,\Sigma_1}(X_i)\right) + (1-Y_i)\,\log\left((1-\pi)\, f_{\mu_0,\Sigma_0}(X_i)\right),$$

where f_{μ_1,Σ_1} (f_{μ_0,Σ_0}) is the density of the Gaussian distribution with mean vector μ_1 (μ_0) and covariance matrix Σ_1 (Σ_0). The posterior probability $\widehat{h}(x)$ is computed using Bayes theorem:

$$\widehat{h}(x) = \frac{\widehat{\pi}\, f_{\widehat{\mu_1},\widehat{\Sigma_1}}(x)}{\widehat{\pi}\, f_{\widehat{\mu_1},\widehat{\Sigma_1}}(x) + (1-\widehat{\pi})\, f_{\widehat{\mu_0},\widehat{\Sigma_0}}(x)}\ .$$

The corresponding classifier $g(x)$ uses the threshold $t=1/2$ to output the binary prediction:

$$g(x) = \mathbb{1}_{\widehat{h}(x)\geq\frac{1}{2}}\ .$$

The decision function for QDA is quadratic. Compared to QDA, LDA further assumes that both distributions share the same covariance ($\Sigma_0=\Sigma_1$) which leads to a linear decision function.

c. Support Vector Machine (SVM)

In this paragraph only, we will assume that instances of the negative class are encoded as $Y = -1$ (instead of $Y = 0$) while instances of the positive class are still represented as $Y = 1$. Linear SVM consists in finding a linear classification rule by minimizing a hinge loss with a penalization term proportional to the L^2 norm of the parameter (Bishop and Nasrabadi, 2006, Chap. 7):

$$\widehat{g} \in \underset{\beta_0,\beta}{\text{Arginf}} \sum_{i=1}^{n} \left[1 - Y_i \left(\beta_0 + \beta^T X_i\right)\right]_+ + \lambda ||\beta||^2,$$

where λ is an hyper-parameter, $||\cdot||$ denotes the L^2 norm and $[\cdot]_+ = \max(0, \cdot)$ is the positive part.

The above optimization problem is equivalent to a constrained optimization problem whose dual only involves the dot products of pairs of elements $(X_i^T X_j)_{1 \leq i,j \leq n}$ (cf Bishop and Nasrabadi, 2006, Chap. 7 for the detailed formulation). It is thus common to extend SVM to non-linear boundaries by mapping the covariate vectors $(X_i)_{1 \leq i \leq n}$ to a higher dimensional space $(\Phi(X_i))_{1 \leq i \leq n}$. A linear decision function in this higher dimensional space $\mathbb{F}$ then results in a non-linear boundary in the original feature space $\mathbb{R}^p$. In practice, neither the mapping Φ nor the higher dimensional space $\mathbb{F}$ need to be explicit. One only needs to specify the dot products of elements in the new feature space $\langle \cdot, \cdot \rangle$ which can be represented by a kernel function K:

$$K(X_i, X_j) = \langle \Phi(X_i), \Phi(X_j) \rangle \ .$$

This is the *kernel trick*. Of course, choosing $K(X_i, X_j) = X_i^T X_j$ boils down to linear SVM. Another popular choice of kernel is the Radial Basis Function (or Gaussian kernel) of the form:

$$K(X_i, X_j) = e^{-\frac{||X_i - X_j||^2}{2\sigma^2}} \ .$$

d. Random Forest (RF)

Random Forest is a classification method that consists in estimating an ensemble of classification trees and aggregating the predictions (Breiman, 2001). A classification tree is a partition of the feature space $\mathbb{R}^p$ in regions $R_1, ..., R_m$ where each region R_i can be mapped to a class c_i representing the class obtaining the majority of votes in the leaf. The classifier g is then of the form:

$$g(x) = \sum_{i=1}^{m} c_i \mathbb{1}_{x \in R_i} \ .$$

A classification tree is estimated by finding iteratively binary splits in the feature space. Each split is performed over a unique feature. Both the feature and the splitting point are chosen as a minimizer of a criterion: standard choices are Gini index and cross-entropy. The size of the tree is controlled by additional hyper-parameters like the maximum depth of the tree or the minimum number of samples per leaf.

Random forest considers a set of B classification trees $(\widehat{g}_k)_{1 \leq k \leq B}$. The trees are different.

- Each of them is trained on a random bootstrap re-sample of the original training set.
- During the estimation of the trees, the binary splits are chosen among a sub-sample of the original features.

Finally, the final classifier $\widehat{g}$ is given by the majority of votes among the B individual trees of the forest. In addition, for a given instance x, the proportion of trees predicting 1 can be interpreted

as a prediction probability:

$$\widehat{h}(x) = \frac{1}{B}\sum_{k=1}^{K} \widehat{g}_k(x) \ .$$

Of course, the classifier $\widehat{g}$ returns 1 if $\widehat{h}(x) \geq 1/2$, else 0.

2.5.3 Performance metrics for classification

Prior to training, the data set is randomly split between train and test sub-samples. Once a classification model is estimated on the training set, one needs to evaluate its performances on the test set. Multiple metrics exist to measure the performances of a classifier. In the context of fatigue applications, we need appropriate evaluation metrics for imbalanced classes.

Let us consider a trained classifier and $(x_i, y_i)_{1\leq i\leq n}$ a test data set (unseen during training). We usually have access to prediction probabilities $\widehat{h}(x_i)$ and not directly labels. In order to provide the predicted labels for covariate vectors $(x_i)_{1\leq i\leq n}$, we can specify a threshold t not necessarily equal to 1/2 which defines the decision rule:

$$\widehat{y}_i = \mathbb{1}_{\widehat{h}(x_i)\geq t} \ .$$

There are two types of performance metrics for classification: those that evaluate the quality of the predicted labels and those that directly rely on the prediction probabilities.

a. Performance metrics on binary predictions

The most common metric for evaluating the quality of the predicted labels is the accuracy that measures the proportion of well classified instances:

$$accuracy = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\widehat{y}_i = y_i} \ .$$

However, the accuracy is a poor choice of metric in imbalanced settings. For instance, in the context of our fatigue application, there are only 1.5% of positive instances. This means that a naive classifier always predicting 0 will get an accuracy of 98.5%. However such a classifier cannot characterize any crack initiation.

In imbalanced setting, it is thus crucial to look at other metrics. Let us first introduce the confusion matrix which gathers the number of truly and wrongly predicted instances for each class: *True Positives* (TP), *False Positive* (FP), *True Negatives* (TN), *False Negatives* (FN). The formulas are given below (Eq. 2.4) and the matrix is presented in Fig. 2.21 with an example using the probabilistic Dang Van criterion of Section 2.4.

$$\begin{aligned} TP &= \sum_{i=1}^{n} \mathbb{1}_{\widehat{y}_i=1, y_i=1} \qquad FN = \sum_{i=1}^{n} \mathbb{1}_{\widehat{y}_i=0, y_i=1} \\ FP &= \sum_{i=1}^{n} \mathbb{1}_{\widehat{y}_i=1, y_i=0} \qquad TN = \sum_{i=1}^{n} \mathbb{1}_{\widehat{y}_i=0, y_i=0} \ . \end{aligned} \tag{2.4}$$

From the four above quantities, we can define appropriate metrics focusing on the performances of a classifier on the minority class (positive class).

- The *recall* (also called *True Positive Rate*, or *sensitivity*) measures the proportion of well classified positive instances:

$$Recall = \frac{TP}{TP + FN} \ .$$

		Predicted labels $\widehat{Y}$		Total
		Crack $\widehat{Y} = 1$	No crack $\widehat{Y} = 0$	
True labels Y	Crack $Y = 1$	$TP = 165$	$FN = 126$	Number of cracks (291)
	No crack $Y = 0$	$FP = 3\,837$	$TN = 15\,239$	Number of non-cracks (19 076)
Total		Number of predicted cracks (4 002)	Number of predicted non-cracks (15 365)	Number of points (19 367)

Figure 2.21: Confusion matrix for Dang Van probabilistic criterion with standard threshold $t = 0.5$. The matrix is evaluated on the whole fatigue data set.

In the context of fatigue, the recall represents the proportion of cracks detected by the criterion. We expect it to be as high as possible.

- The *precision* is the proportion of true positive instances among the positive predictions:

$$Precision = \frac{TP}{TP + FP} .$$

In fatigue, it characterizes the proportion of true cracks among the positive zones identified by the criterion.

- The *False Positive Rate* (FPR) measures the proportion of prediction errors on the negative instances:

$$FPR = \frac{FP}{FP + TN} .$$

- As we usually seek a compromise between *Precision* and *Recall*, a popular metric is the *F1 score*, defined as the harmonic mean of the two:

$$F1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} .$$

Table 2.5: Performances of Dang Van probabilistic criterion on the fatigue database for different thresholds.

	$t = 0.01$	$t = 0.5$	$t = 0.99$
Recall	0.85	0.57	0.22
Precision	0.03	0.04	0.07
FPR	0.44	0.20	0.05
F1 score	0.06	0.08	0.10

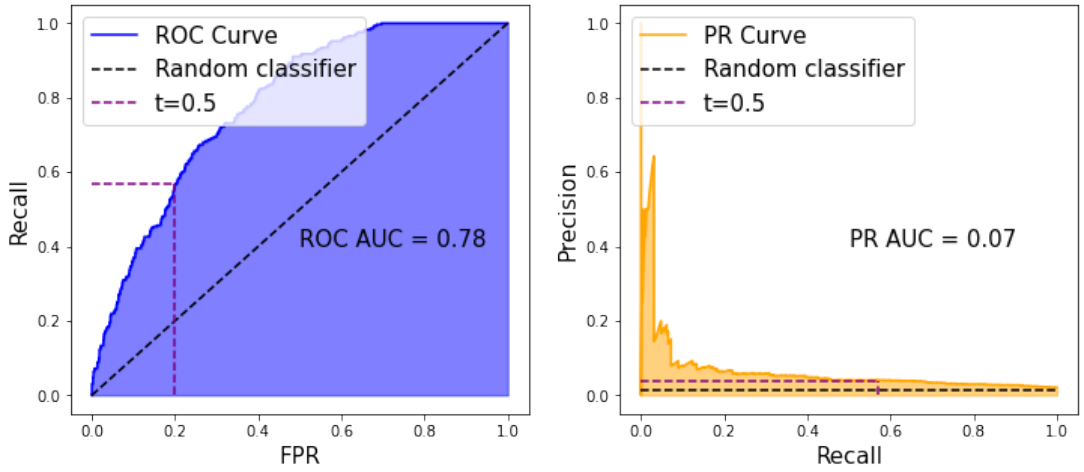

Figure 2.22: ROC (on the left) and PR (on the right) curves for probabilistic Dang Van criterion. Black dashed lines represent the performances of a random classifier. Purple dashed lines represent the performances obtained by setting the threshold to $t = 0.5$.

b. Performance metrics on prediction probabilities

As explained in the introduction of this subsection, the performance metrics defined in the previous paragraph depend on the threshold t used to provide the predicted labels $(\widehat{y}_i)_{1\leq i\leq n}$ given the predicted probabilities $(\widehat{g}(x_i))_{1\leq i\leq n}$. For instance, Table 2.5 presents the binary performance metrics of probabilistic Dang Van criterion for different choices of thresholds. Increasing the threshold tends to improve the Precision and reduce the FPR but at the same time reduces the Recall.

In order to assess the performances of a classifier without having to specify the threshold t, we study the Receiver Operating Characteristic (ROC) curve and the Precision Recall curve (PR).

- ROC curve represents the recall (Y-axis) as a function of the FPR (X-axis) for every choice of threshold.
- PR curve similarly represents the precision (Y-axis) as a function of the recall (X-axis).

ROC and PR area under the curve (AUC) represent the area under ROC and PR curve and can be used as global performance metrics. Assuming there exists a perfect classifier separating the two classes, its ROC and PR AUC will be equal to 1. For a random classifier, ROC AUC will be 0.5 and PR AUC will be equal to the proportion of positive instances. ROC and PR curves for probabilistic Dang Van criterion are illustrated in Fig. 2.22: in the context of fatigue applications, Dang Van criterion leads to small precision scores, lower than 10%, for a recall higher than 20%.

ROC and PR curves lead to similar metrics as they both integrate the performances of a classifier over all the possible thresholds. Nevertheless, they can lead to different conclusions.

The crucial difference is that while both metrics consider the recall, ROC relies on the FPR while PR computes the precision. Saito and Rehmsmeier (2015) argue that, for highly imbalanced data sets, PR curves may be more informative than ROC curves. One drawback though of the PR AUC metric compared to ROC AUC is that the baseline performance (performance of a random classifier) depends on the proportion of positive instances, and thus changes from one data set to another. This is why, we will consider both ROC and PR AUC when evaluating classification performances.

Remark This is not the only solution to deal with imbalanced classes in classification. An alternative solution would be to down-sample the majority class. We do not consider this solution as the imbalance is quite severe. Down-sampling the negative class would result in eliminating a lot of zones which would reduce significantly the size of the data set.

2.5.4 Application to the fatigue database

In this subsection, we apply the supervised classification methods listed in Subsection 2.5.2 to the fatigue data set. We will thus obtain several fatigue criteria calibrated through machine learning techniques. Their performances will be evaluated and compared to the Dang Van fatigue criterion.

Paragraph a describes the procedure to estimate the classifiers, select hyper-parameters and evaluate the performances. This procedure will be illustrated in Paragraph b by considering a LR classifier on the fatigue data set with all features ($p = 60$). Finally, Paragraph c presents the classification results for the different methods.

a. Description of the procedure

The fatigue data set is randomly split in two sub-samples of equal sizes: training and test sets. The classification rule is estimated on the training set and its performances are evaluated on the test set.

Some classification methods require the specification of hyper-parameters:

- multiplicative coefficient on the L^1 penalty in LR (Lasso regularization);
- multiplicative coefficient on the L^2 penalty in SVM (linear and Gaussian kernels);
- maximum depth of the trees in RF.

These hyper-parameters are selected during the training phase using K-fold cross-validation with $K = 5$ (cf. Hastie et al., 2009, Chap. 7). The training set is divided in K sub-samples. For k ranging from 1 to K, a classifier is estimated on the training set except the k^{th} fold. Then the classifier performances on the k^{th} fold are evaluated. The mean performances over the K folds is then calculated. This experiment is performed for multiple values of the hyper-parameter: the value achieving the best performance is selected.

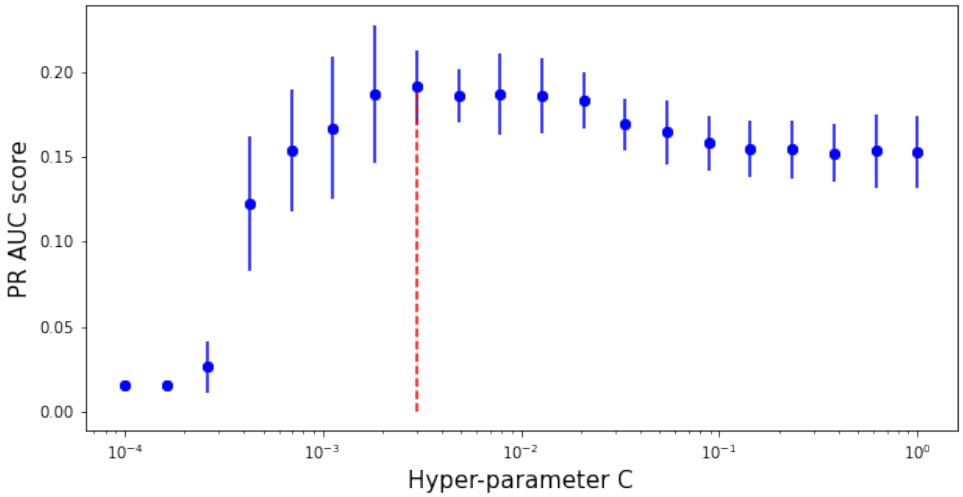

Figure 2.23: Cross-validation PR AUC scores for different values of hyper-parameter C. Error bars represent the standard deviation over the 5-fold performance estimates. Red dashed line represent the hyper-parameter value resulting in the maximum PR AUC.

Remarks:

1. The random partition in K folds is stratified to ensure that the proportion of positive instances in each fold is approximately the same (Stratified K-Folds cross-validation). In particular, this ensures that each fold contains samples of the minority class.

2. The choice of the number of folds, $K = 5$, is standard and often recommended in the literature (cf. Hastie et al., 2009, Chap. 7).

Once the hyper-parameter is selected, a final classifier is estimated over the whole training set using the optimal value for the hyper-parameter. Finally, the performances are evaluated on the test set.

We will see that the performance results change depending on the initial random split between train and test sets. The reason is that there are few positive instances in the data set which leads to a high variance in the estimation of the performance results. In order to assess the performances with more consistency, the whole procedure is repeated $B = 100$ times. Hence, for each method, we will be looking at the distribution of the performances over the B repetitions.

b. Illustration of the procedure

We illustrate here the general procedure described above on an example: the classification model is a LR (with Lasso regularization). The hyper-parameter $C = 1/\lambda$ is selected through $5-fold$ cross-validation. The mean and standard deviation over the PR AUC scores for different choices of hyper-parameter C are presented in Fig. 2.23. For low values of C (high penalty), the performances are close to 0: the penalization is too strong compared to the logistic loss. Conversely, as hyper-parameter becomes high (low penalty), the penalization is small compared to the loss function: the model tends to perform as if there was no variable selection, which also results in poor performances. From these results, we can identify the optimal hyper-parameter $C_{opt} \simeq 2 \times 10^{-2}$.

The final model is trained on the whole training set using the optimal hyper-parameter C_{opt}. The classifier is then evaluated on the test set. In this illustration example, the whole experiment is only performed once. In the next subsection, it will be repeated $B = 100$ times

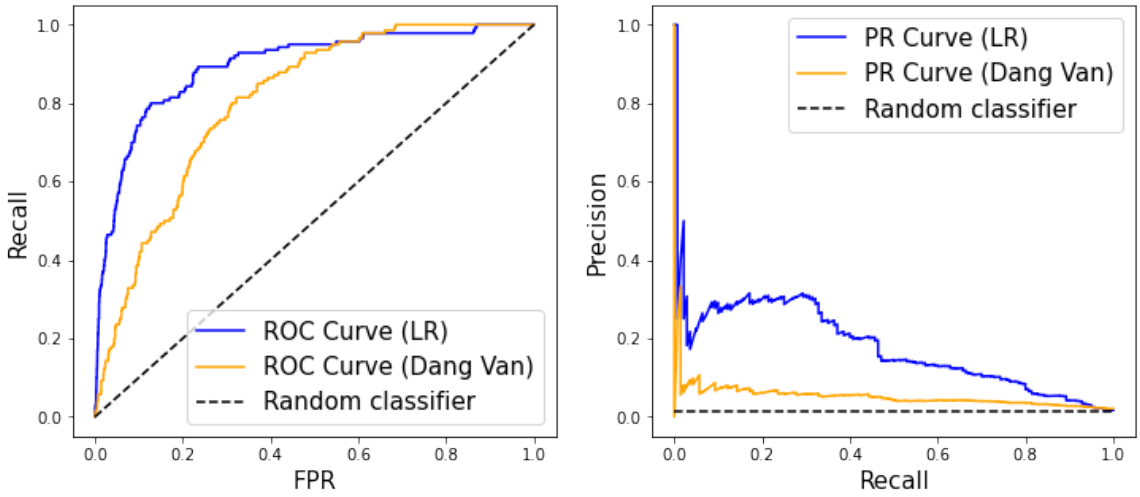

Figure 2.24: ROC (left) and PR (right) curves for trained LR and Dang Van criterion. Dashed lines represent the performances of a random classifier.

in order to account for the variability on the performance estimates. Figure 2.24 represents the ROC and PR curves for the estimated classifier. The performances of Dang Van criterion are also represented highlighting the substantial gain in performances achieved by the machine learning model. Indeed, ROC AUC score on the test set is equal to 0.90 (0.81 for Dang Van criterion) and PR AUC score is 0.18 (0.05 for Dang Van classifier). We will see that this gain is consistent when repeating the same procedure on multiple train-test splits (cf. Subsection c).

c. All classification results

We now extend the comparison of the results for all the supervised classification methods presented in Subsection 2.5.2 on our two scenarii of interest:

1. a first scenario where only the two variables involved in Dang Van criterion are considered ($p = 2$);
2. a second scenario where all the variables (as defined in Subsection 2.2.3) are considered ($p = 60$).

This allows us to compare the two scenarii and thus assess the potential gain in performances when using all the information contained in the fatigue data set to help identify critical zones. As presented in Paragraph a, for each scenario and each method, the estimation and evaluation procedure is repeated $B = 100$ times in order to study the distribution of performance scores. Classification performances of Dang Van criterion are also evaluated on the test set (for each experiment). Contrary to the supervised classification methods, Dang Van criterion is not estimated on the training data set, as it is a priori defined.

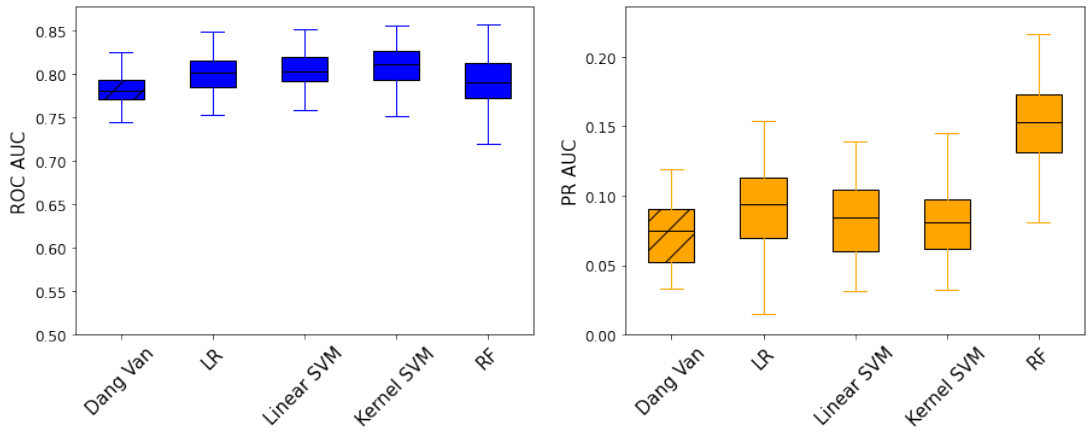

Figure 2.25: First scenario ($p = 2$): distribution of ROC AUC (left) and PR AUC (right) performances for Dang Van criterion (hatched boxplots) and supervised classification methods.

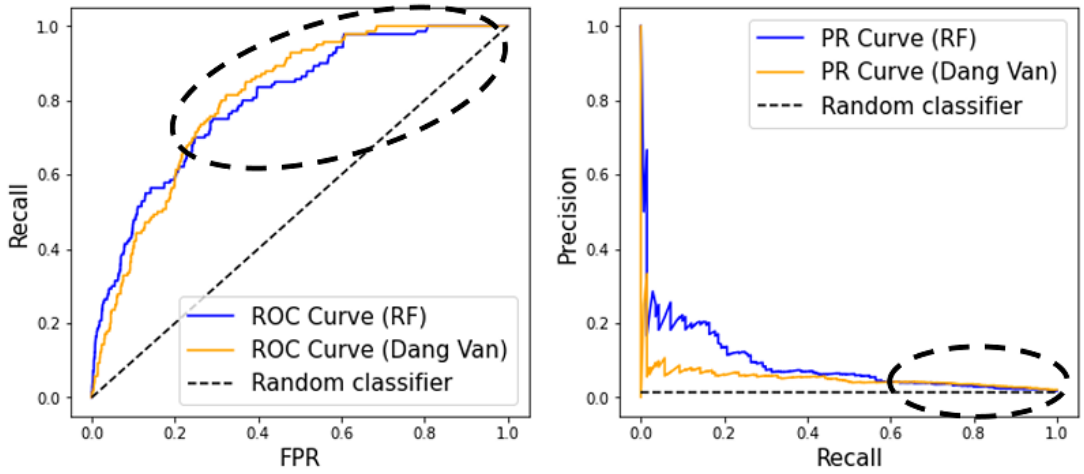

Figure 2.26: First scenario ($p = 2$): ROC (left) and PR (right) curves for Dang Van criterion (orange) and Random Forest (blue). Dashed ellipses highlight the part of the curves for which recall is over 0.6.

First scenario The results on the first scenario are presented in Figure 2.25. The ROC AUC performances of classification methods are not significantly different from Dang Van criterion. In terms of PR AUC, the results are more heterogeneous. While, LR and Linear SVM have a similar behavior and provide performances comparable to Dang Van criterion, Kernel SVM and RF are different. In particular, the distribution of PR AUC scores for RF is significantly higher than Dang Van criterion which seems surprising at first glance because ROC AUC scores are similar. In order to help understand this phenomenon, Figure 2.26 represents the ROC and PR curves for Dang Van criterion and RF ($p = 2$ variables). The major difference between the two PR curves occurs for small values of recall. This means that the Random Forest can identify the most critical zones with higher precision than Dang Van criterion. However, the ROC curve shows that Dang Van is better than RF for a recall higher than 0.6. This difference is not visible in the PR plot because both methods yield similar small precision scores (cf. dashed ellipses in Fig. 2.26).

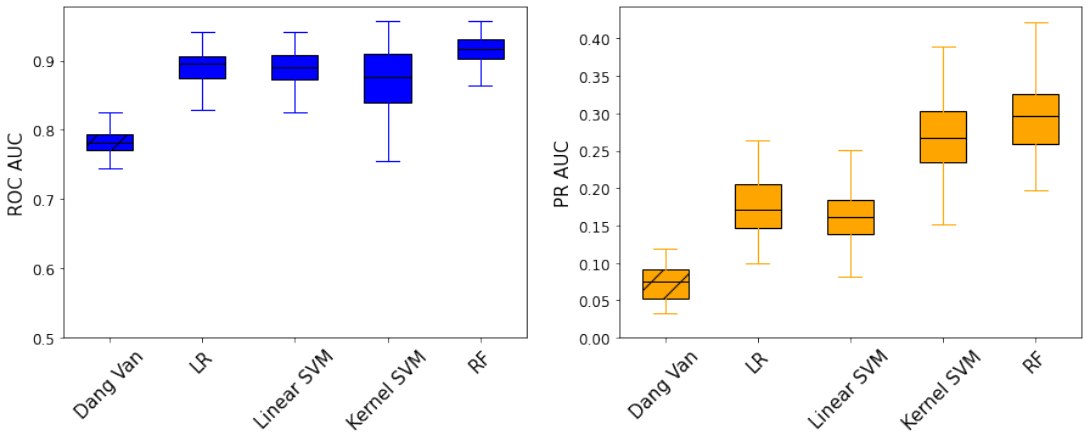

Figure 2.27: Second scenario ($p = 60$): distribution of ROC AUC (left) and PR AUC (right) performances for Dang Van criterion (hatched boxplots) and supervised classification methods.

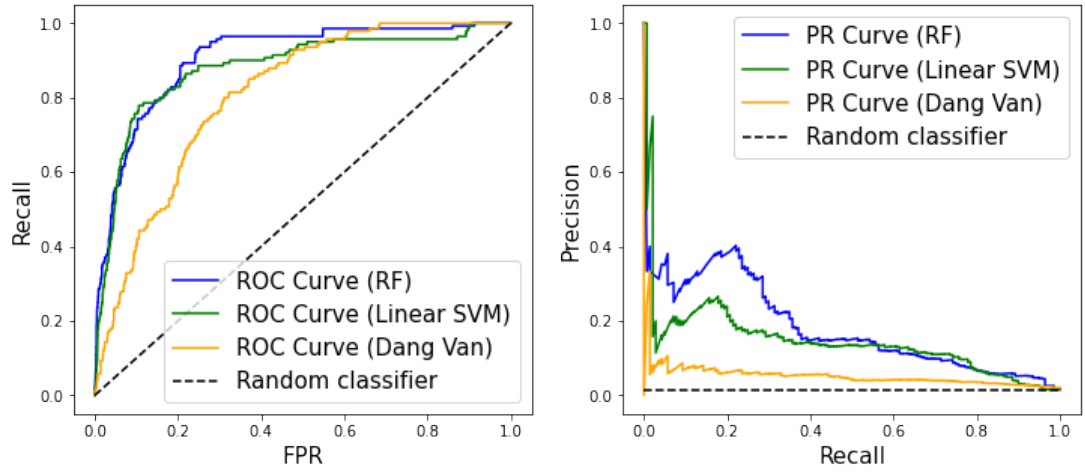

Figure 2.28: Second scenario ($p = 60$): ROC (left) and PR (right) curves for Dang Van criterion (orange), Random Forest (blue) and Linear SVM (green).

Second scenario Performance results on the second scenario ($p = 60$ variables) are represented in Figure 2.27. This time, the gain in performances of supervised classification methods compared to Dang Van criterion is significant, both in terms of ROC AUC and PR AUC. Linear models (LR and Linear SVM) exhibit similar behaviors. Kernel SVM results are more scattered than the other methods (especially for ROC AUC performances). Random Forest achieves the best mean performances both in terms of ROC AUC and PR AUC. As for the first scenario, we can analyze further the differences of PR AUC scores among the methods. Figure 2.28 presents a comparison between the ROC and PR curves obtained for Dang Van criterion, a linear classifier (Linear SVM) and a non-linear one (RF). Again, the main differences between the classifiers on the PR curve are concentrated on its first part, *i.e.* for a recall lower than 0.4.

In this section, supervised machine learning algorithms were implemented in order to estimate fatigue criteria based on the fatigue data set. The results confirm the limits of Dang Van variables (P_c, τ_c) to characterize critical zones on complex specimens: indeed, linear and non-linear classification methods do not achieve a significant improvement over Dang Van fatigue criterion in this case. When all the covariates are considered though, it is clear that the classification methods provide better performances: for instance, Logistic Regression with Lasso regularization yields a mean improvement of 0.11 in terms of ROC AUC and 0.10 in terms of PR AUC.

From a more practical point of view, the Logistic Regression can identify 50% of true crack initiations with a precision of approximately 14% (only 4% for Dang Van criterion). If the recall is now set to 90% (as the objective is to identify most of the crack initiations), the precision drops to 4% (2% for Dang Van criterion). Hence, despite the additional variables, some critical zones are still poorly characterized.

2.6 - Conclusion

In this chapter, we introduced Stellantis fatigue database built upon numerical simulation results and reports from fatigue tests on prototypes. The database contains zones of numerical models characterized by several features (describing the geometry and the stresses) along with the test results indicating whether or not a crack initiated on each zone. We introduced a notion of zone, allowing to reduced significantly the severe imbalance between positive and negative instances, better appraise the local distribution of stresses, and account for nearby singularities through appropriate features. A multivariate analysis allowed to better understand the variance in the data set and the correlation structure among variables. Besides, we identified and characterized simultaneously a structure among individuals and variables through co-clustering. As the objective of fatigue design is to identify critical zones on a numerical model, we analyzed the Dang Van fatigue criterion for welds and proposed a probabilistic version of it accounting for the randomness of crack initiation. This criterion is based on different structures of standardized welded specimens. Although Dang Van criterion works well on standardized specimens, it generalizes poorly to the more complex zones contained in the fatigue database. Considering the estimation of a fatigue criterion as a supervised classification task, we estimated and compared classic classifiers. Contrary to Dang Van criterion (which relies on critical hydrostatic and shear stresses), these criteria account for all the variables available in the fatigue data set ($p = 60$), which leads to better prediction performances. This means that the classification-based criteria can better identify critical zones on a design, which would help to reduce iterations between conception and validation.

Still, the prediction performances of fatigue criteria estimated through supervised classification are far from being perfect. In particular, some crack initiations are still poorly characterized by the available features. The FEM is an important source of uncertainty because it does not account for several phenomena influencing the resistance against fatigue: complex geometric effects (stress concentrations), manufacturing processes generating residual stresses (arc welding, stamping)... Hence, having access to these variables could help the characterization of critical zones and thus improve the performances of classification-based fatigue criteria. Unfortunately, the current FEM do not provide that information. Even if the improvement of FEM is an active topic in the automotive industry, it is beyond the scope of this thesis.

While evaluating supervised classification methods on the fatigue data set, we could notice that there are many false positives (non-broken zones predicted as positive) which results in the precision being low: 14% precision for a 50% recall (Logistic Regression). This is due to the fact that not every critical zone breaks during testing. Hence, only a subset of critical zones effectively

broke during testing. However, the objective of a fatigue criterion should not be the prediction of crack initiations but rather the prediction of the criticality of a zone. As only some critical zones are labeled, the construction of a fatigue criterion can be viewed as a classification task under a completely asymmetric label noise. This task, known as Positive-Unlabeled learning (PU learning) is an important contribution of this thesis and will be the subject of the next two chapters: Chapter 3 introduces PU learning and states new theoretical risk bounds for classification in this particular setting, Chapter 4 focuses on the application of PU learning to the fatigue data set.

3

Theoretical risk bounds for Positive-Unlabeled Learning under the Selected At Random assumption

The estimation of a fatigue criterion using Stellantis fatigue database is a classification task (cf. Chapter 2): the goal is to predict whether or not a zone of a mechanical part is critical, *i.e.* whether or not a crack can initiate before 10^6 cycles. However, this classification is not fully supervised: only a subset of positive instances (critical zones) are actually observed and thus labeled positive (crack initiations). Besides an unlabeled zone (without crack initiation) can either be critical (because the severity was not high enough to observe the crack) or safe. This task is known as Positive-Unlabeled learning (PU learning). The challenge is then to find the correct classifier despite this lack of information.

In this chapter, we are interested in establishing risk bounds for PU learning under the general Selected At Random assumption, *i.e.* when the probability for a positive instance to be labeled depends on its covariates. In addition, we quantify the impact of label noise on PU learning compared to the standard classification setting. Finally, we provide a lower bound on the minimax risk proving that the upper bound is almost optimal.

Section 3.1 introduces the traditional classification setting and recall risk bounds results. In section 3.2, we introduce the PU learning setting. Section 3.3 presents an overview of existing approaches on PU learning. Section 3.4 focuses on the bias issue with labeled-unlabeled classification and motivates the use of an unbiased empirical risk. In 3.5, we present the main results of this chapter: a general upper bound on the excess risk for PU learning under covariate-dependent label noise and a lower bound on the minimax risk. Section 3.6 illustrates the theoretical results through numerical experiments. The proofs and technical lemma can be found in Section 3.7.

3.1 - Traditional classification setting

In this section, we introduce the standard classification setting (Subsection 3.1.1) and recall risk bounds results (Subsection 3.1.2). This will be the opportunity to introduce general notations used throughout the chapter.

3.1.1 General setting

Let $(X_1, Z_1), ..., (X_n, Z_n)$ be independent couples of random variables in $\mathbb{R}^d \times \{0,1\}$ identically distributed according to some unknown distribution denoted $\mathbb{P}$. For each i, X_i is a *covariate* vector with marginal distribution $\mathbb{P}_X$ and Z_i is the *class*, either *negative* ($Z_i = 0$) or *positive* ($Z_i = 1$). Let $\pi = \mathbb{P}(Z = 1)$ denote the class prior. Using $\mathbb{P}_0$ ($\mathbb{P}_1$) the conditional distribution of X given that the class is negative, $Z = 0$ (positive, $Z = 1$), we write the convenient decomposition:

$$\mathbb{P}_X = (1 - \pi)\mathbb{P}_0 + \pi\mathbb{P}_1 \ . \tag{3.1}$$

In classification, the goal is to find a classifier, *i.e.* a binary function $g : \mathbb{R}^d \to \{0,1\}$, minimizing some risk function R. In this chapter, R will denote the misclassification risk:

$$R(g) = \mathbb{P}(g(X) \neq Z) \ .$$

Given the regression function $\eta(x) = \mathbb{P}(Z = 1 | X = x)$, the minimizer of misclassification risk is Bayes classifier g^* that depends explicitely on $\mathbb{P}$:

$$g^*(x) = \mathbb{1}_{\eta(x) \geq \frac{1}{2}}.$$

In order to assess how close a given classifier g is to the optimal one g^*, we are interested in the excess risk $\ell(g, g^*)$:

$$\ell(g, g^*) = R(g) - R(g^*) .$$

Since $\mathbb{P}$ is unknown, neither g^* nor the risk function R can be computed. We rely instead on the training sample $(X_1, Z_1), ..., (X_n, Z_n)$ to build a classifier $\widehat{g}$. Let $r(g, (X, Z)) = \mathbb{1}_{g(X) \neq Z}$ the misclassification error for one observation, the true risk R can be estimated by the empirical mean:

$$\widehat{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} r(g, (X_i, Z_i)) \ .$$

An empirical classifier $\widehat{g}$ is then identified as a minimizer of the empirical risk over a predefined class of classifiers $\mathcal{G}$.

$$\widehat{g} \in \underset{g \in \mathcal{G}}{\text{Argmin}} \ \widehat{R}_n(g) .$$

This procedure is known as Empirical Risk Minimization. Let $g^{\mathcal{G}}$ be the minimizer of the true risk R over $\mathcal{G}$. The excess risk of the classifier $\widehat{g}$ can be decomposed as follows:

$$\ell(\widehat{g}, g^*) = \left(R\left(g^{\mathcal{G}}\right) - R(g^*)\right) + \left(R(\widehat{g}) - R\left(g^{\mathcal{G}}\right)\right)$$

where the first term is the approximation error depending on $\mathcal{G}$ and the second one is the statistical error. Since we are only interested in assessing the statistical error, we assume that Bayes classifier g^* belongs to $\mathcal{G}$, hence the first term vanishes. It is important to note that $\ell(\widehat{g}, g^*)$ depends on $\mathbb{P}$ (through the risk R) and on the training sample $(X_1, Z_1), ..., (X_n, Z_n)$.

3.1.2 Risk bounds in the standard classification

In order to assess the convergence rate of the excess risk $\ell(\widehat{g}, g^*)$ in a non-asymptotic framework, we need an upper bound on $\mathbb{E}[\ell(\widehat{g}, g^*)]$. The expectation is taken with respect to the distribution of the training sample $\mathbb{P}^{\otimes n}$. Moreover the upper bound needs to be uniform over a set of distributions $\mathbb{P}$. We introduce $\mathcal{P}(\mathcal{G})$ a set of probability distributions on $\mathbb{R}^d \times \{0,1\}$ such that g^* belongs to $\mathcal{G}$. In this case, Lugosi (2002) proved that for some absolute constant $C_1 > 0$:

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{G})} \mathbb{E}[\ell(\widehat{g}, g^*)] \leq C_1 \sqrt{\frac{V}{n}} , \tag{3.2}$$

where V is the *Vapnik-Chervonenkis dimension* of $\mathcal{G}$ (VC dimension, see Vapnik, 1999, Chapter 3). We recall that the VC dimension is the maximum integer V such that there exists V points $x_1, ..., x_V$ in $\mathbb{R}^d$ *shattered* by $\mathcal{G}$, namely classified in every way possible by elements of $\mathcal{G}$. In other words:

$$V = \sup_{v \in \mathbb{N}^*} \left\{ v \ s.t. \ \exists x_1, ..., x_v \in \mathbb{R}^d, \ |\{(g(x_1), ..., g(x_v)), \ g \in \mathcal{G}\}| = 2^v \right\} .$$

The VC dimension V measures the complexity of class $\mathcal{G}$ and has to be finite for Equation 3.2 to be meaningful, which we assume for the rest of the chapter.

The upper bound in Equation 3.2 remains true regardless of the form of the regression function η. Actually, η is closely linked to the *label noise*: when $\eta(x)$ is close to $1/2$, the observed class can be positive or negative with probability close to $1/2$, which makes the classification of x more difficult. Hence, the closer η is to $1/2$, the noisier the observed class is. Massart and Nédélec (2006) showed that whenever $\eta(x)$ is uniformly and symmetrically bounded away from $1/2$ by a quantity $h > \sqrt{V/n}$, the upper bound on the risk excess can be improved. Let $\mathcal{P}(\mathcal{G}, h)$ denote the subset of probability distributions in $\mathcal{P}(\mathcal{G})$ such that for every $x \in \mathbb{R}^d$, $|2\eta(x) - 1| \geq h$. Massart and Nédélec (2006) showed that there exists an absolute constant $C_2 > 0$ such that:

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{G}, h)} \mathbb{E}[\ell(\widehat{g}, g^*)] \leq C_2 \frac{V}{nh} \left(1 + \log\left(\frac{nh^2}{V}\right)\right) . \tag{3.3}$$

Hence, as h gets higher, the label noise gets smaller, and the convergence rate can be improved up to V/n, letting aside the logarithm. However, when h is smaller than $\sqrt{V/n}$, Equation 3.2 remains better. Equation 3.3 provides a fine control on the excess risk depending on the difficulty of the classification task, accounted through h.

A lower bound was obtained by Lugosi (2002), extended by Massart and Nédélec (2006), allowing to prove the optimality of the convergence rates. In fact, the optimality of the refined bound of Equation 3.3 is up to the logarithmic term.

3.2 - PU learning context

In the standard classification setting, the classes $(Z_i)_{1 \leq i \leq n}$ are observed. This is no longer the case in PU learning where only an incomplete set of positive data is available, the remaining is unlabeled. For each i, the observed label Y_i is 1 if the class Z_i is positive and *selected* (*i.e.* labeled). Otherwise, the label Y_i is 0 (unlabeled). The objective of PU learning is to use the incomplete information $(X_1, Y_1), ...(X_n, Y_n)$ to build a classifier able to predict the class Z given a new instance with covariates X.

PU learning is motivated by various applications listed in Subsection 3.2.1. It can arise in different settings: one-sample setting or two-sample setting (cf. Subsection 3.2.2). The label noise in PU learning is usually represented by a propensity function which can be assumed constant or covariate-dependent (cf. Subsection 3.2.3)

3.2.1 PU learning applications

PU learning is motivated by various kinds of applications.

- *Reliability* : this work is primarily motivated by an application in mechanical design against fatigue. The idea is to use past simulation data combined with test results to build a classifier able to predict critical zones on new numerical models. In this scenario, experimental tests only provide a fraction of critical zones that effectively broke during testing. However, the absence of crack initiation is not an evidence of safety: maybe other zones could have initiated a crack if the tests had been extended. Hence all the other zones need to be considered unlabeled.

- *Health*: PU learning has been widely applied to *gene disease detection*. The objective is to identify genes related to various human diseases. Each gene is described by different biological information used as features. However, only a fraction of genes are known to be related to a disease. Hence, the task of identifying disease-related genes among the remaining unlabeled examples is a PU Learning task (Yang et al., 2012, 2014; Nikdelfaz and Jalili, 2018). Another class of applications is about automatic diagnosis. Chen et al. (2020) uses MRI (Magnetic Resonance Imaging) data to predict early-stage Alzheimer disease. When traditional classification methods considers non-diagnosed patients as negative examples, PU learning is able to account for the fact that some non-diagnosed patients can be positive (but not yet diagnosed).

- *Text classification*: PU learning naturally arises in text classification problems where the task is to identify texts related to a certain topic. Usually, the training set is made of only a few known positive instances (texts related to the topic of interest). Other documents are added to the training set, but as the labeling processes is long, they are not labeled. This scenario motivated the first PU learning methods (Liu et al., 2002, 2003).

- *Spam review detection*: in many commercial websites, customers are free to leave a comment regarding the product or service they purchased. However, some comments turn out to be fake reviews that could wrongly influence potential buyers. Spam review detection aims at identifying fake reviews. Since labeling is complicated, only a sample of fake reviews are labeled and PU learning techniques are used to build a classifier (Fusilier et al., 2015; Li et al., 2014; He et al., 2020).

- *Anomaly detection* is a general topic that often belongs to the category of PU learning tasks. In this kind of applications, the training set contains a sample of identified anomalies, however the rest of the training set may still contain anomalies mixed with "normal" examples. In this field, PU learning has been applied to predict newborn defects (Jiang et al., 2018), quality flaws on web pages (Ferretti et al., 2014), intrusion detection in cybersecurity (Luo et al., 2018),...

3.2.2 PU learning settings

Missing labels in PU learning can arise from different settings.

- In the *two-sample setting*, the positive and unlabeled instances are sampled separately and are therefore not identically distributed. A first sample $(X_i)_{1\leq i\leq n_L}$ only contains labeled instances $(Y = 1)$ distributed according to the conditional distribution of X given $Z = 1$. A second sample $(X_i)_{n_L+1\leq i\leq n}$ is distributed according to the marginal distribution of X and remains unlabeled $(Y = 0)$. The quantity n_L denotes the number of labeled instances. It is a *case-control* situation.

- In the *one-sample setting*, all the instances $(X_i)_{1\leq i\leq n}$ are i.i.d and some positive instances are labeled. We will focus on this setting in the rest of the thesis.

There is also an important distinction between *transductive* and *inductive* PU learning.

- In a *transductive* PU learning task, the goal is to use the set of positive and unlabeled instances to predict the class of the unlabeled instances of the set.

- In an *inductive* PU learning task, the goal is to use the same training set to predict the class of a new instance. In the scope of the thesis, we will be interested in this setting.

3.2.3 Propensity function and assumptions

In PU learning, the true classes are affected by a class-dependent (thus asymmetric) label noise. The probability for a positive instance to be labeled is generally called the *propensity* (Bekker and Davis, 2020) and it may depend on the covariates:

$$e(x) = \mathbb{P}\left(Y = 1|Z = 1, X = x\right).$$

On the other hand, negative instances are never labeled:

$$\mathbb{P}\left(Y = 1|Z = 0, X = x\right) = 0\ .$$

The regression function associated with Y, $\tilde{\eta}(x) = \mathbb{P}\left(Y = 1|X = x\right)$ depends on this additional label noise:

$$\tilde{\eta}(x) = e(x)\,\eta(x)\ . \tag{3.4}$$

This concept of completely asymmetric label noise was first pointed out by Elkan and Noto (2008). It is now common to define two general types of assumptions: Selected Completely At Random (SCAR) and Selected At Random (SAR).

SCAR: PU learning without selection bias. The propensity $e(x) = e_m$ does not depend on the covariates x. This applies in situations where every positive instance has an equal probability to be selected (labeled). In this case, the conditional distributions of X given $Z = 1$ $(\mathbb{P}_1)$ and given $Y = 1$ $(\tilde{\mathbb{P}}_1)$ are the same. In other words, labeled instances are a representative sub-sample of positive instances.

SAR: PU learning with selection bias. The probability for a positive instance to be labeled depends on its covariates. Hence, labeled instances are a biased sample of positive instances. For example, in mechanical design, a specimen subjected to higher stress is more likely to break, which results in a higher probability of a crack being detected. This is clearly a situation where the SCAR assumption does not hold. This is why, we will focus on the SAR assumption.

3.3 - State of the art on PU learning methodologies

In this section, we present an overview of existing methods to address PU learning tasks. The difficulty lies in the fact that unlabeled data may contain positive instances. Besides, the set of labeled instances may not be a representative sample of positive instances: the propensity $e(x) = \mathbb{P}(Y = 1 \,|\, Z = 1, X = x)$ may depend on x (SAR assumption). The majority of existing methods only work under the SCAR assumption ($e(x) = e > 0$). Nevertheless, over the past few years, new methods emerged to address specifically the more general SAR assumption (when $e(\cdot)$ is not constant).

3.3.1 Non-traditional classifiers

A first way to address PU learning tasks is to simply ignore the label noise. Non-traditional classification techniques consist in considering unlabeled data as negative and labeled data as positive. Then, a classifier is estimated based on the noisy labels $(x_i, y_i)_{1 \leq i \leq n}$ and this learned classifier is used to predict the class Z of a new data point X. From a theoretical point of view (cf. Cannings et al., 2020), a non-traditional classifier remains consistent as far as the amount of label noise is limited (see Subsection 3.4.1).

Besides, a non-traditional classifier under the SCAR assumption ($e(x) = e$) can still rank correctly data points in terms of their probability of being positive. This can be helpful even if it cannot yield a proper estimate of the output probability. This consideration lies on the following remark (cf. Elkan and Noto, 2008):

$$\begin{aligned}\mathbb{P}(Y = 1 \,|\, X = x) &= \mathbb{P}(Y = 1 \,|\, Z = 1,\ X = x) \times \mathbb{P}(Z = 1 \,|\, X = x) \\ &= e \times \mathbb{P}(Z = 1 \,|\, X = x)\ .\end{aligned}$$

Hence $\mathbb{P}(Z = 1 \,|\, X = x)$ is proportional to $\mathbb{P}(Y = 1 \,|\, X = x)$. If one can get a good estimate of $\mathbb{P}(Y = 1 \,|\, X = x)$, then, under the SCAR assumption, it can provide a correct ranking on the predictions. Additionally, if one can get a good estimate of e, the target probability can be estimated. Elkan and Noto (2008) use this key principle to construct a PU classifier upon a non-traditional classifier. The authors rely on a validation set to provide an estimate of the constant propensity e.

3.3.2 Two-step methods

The difficulty of PU learning lies in the absence of tagged negative data. An important class of methods addresses this issue by using heuristics to identify reliable negative instances among the unlabeled data. The PU classifier is then obtained by training a standard classifier using the labeled positive instances and the reliable negative data. These methods are called two-step methods. Different techniques can be used for both steps.

The first step consists in identifying reliable negative examples among the unlabeled ones. Liu et al. (2002) suggests to contaminate the unlabeled set with some labeled instances (*spies*) and then to learn a non-traditional classifier. The reliable negative examples are then identified as those for which the output probability is below the output probability of spies. This, however, relies on the assumption that the set of spies is representative of the positive unlabeled instances, which boils down to assuming SCAR assumption. Some strategies rely on different metrics to identify reliable negatives as those that are far enough from labeled instances. Most of them use a non-traditional classifier and use its output probability as a criterion to select reliable negative instances, for example *Naive Bayes* classifier (cf. Liu et al., 2002)), *Rochhio* classification (cf. Li and Liu, 2003) or *1-DNF* that consists in identifying strong positive features (cf. Hwanjo Yu et al., 2004)). Most of these techniques were specifically designed for text classification as it was one of the first application to be addressed by PU learning.

The second step consists in applying a standard classification technique using labeled and reliable negative data. A common choice is to use a SVM classifier (Li and Liu, 2003; Li et al., 2010).

Usually both steps are iterated, meaning that the classifier estimated in the second step is used to update the set of reliable negative data. Then a new classifier is estimated using the updated reliable negative instances. Bekker and Davis (2020) provides an exhaustive list of existing methods for both steps.

Up to now, most applications of PU learning use two-step strategies: text classification Liu et al. (2002, 2003); Li and Liu (2003); Ferretti et al. (2014), fake reviews detection Li et al. (2014); He et al. (2020) and disease genes identification Yang et al. (2012); Nikdelfaz and Jalili (2018).

3.3.3 Neyman-Pearson classification

Under the SCAR assumption, the conditional probability of X given $Y = 1$ is the same as the conditional probability of X given $Z = 1$. Hence, Blanchard et al. (2010) made two fundamental remarks. First, for a given classifier $g : \mathbb{R}^d \to \{0, 1\}$, the risk of predicting 0 instead of 1 is:

$$R_1(g) = \mathbb{P}\left(g(X) \neq 1 \,|\, Z = 1\right) = \mathbb{P}\left(g(X) \neq 1 \,|\, Y = 1\right), \tag{3.5}$$

which can be estimated using PU data.

Besides, even if the second type risk $R_0(g) = \mathbb{P}\left(g(X) \neq 0 \,|\, Z = 0\right)$ cannot be directly estimated using PU data, one can compute an estimate of the unlabeled risk $R_U(g) = \mathbb{P}\left(g(X) \neq 0 \,|\, Y = 0\right)$. Blanchard et al. (2010) provide theoretical guarantees showing that the following optimization problem:

$$\underset{g \in \mathcal{G},\, R_1(g) \leq \alpha}{\text{Arginf}} \quad R_0(g) \tag{3.6}$$

is equivalent to

$$\underset{g \in \mathcal{G},\, R_1(g) \leq \alpha}{\text{Arginf}} \quad R_U(g)\ . \tag{3.7}$$

From a hypothesis testing point of view, the second problem consists in finding an optimal α-level test for testing whether the distribution of a new instance X is $\mathbb{P}\left(\cdot \,|\, Z = 1\right)$ (null hypothesis) or $\mathbb{P}\left(\cdot \,|\, Y = 0\right)$. When both distributions have a density, the test statistics is given by the ratio of the densities (Neyman-Pearson lemma).

The method proposed by Blanchard et al. (2010) consists in using the training set to estimate both $\mathbb{P}_1$ and $\mathbb{P}_U$ densities, and then use their ratio to predict the class of new data. Of course, the ratio of densities do not provide the output probabilities but a criterion giving a ranking on the output probabilities. We can then use a validation set in order to calibrate a correct threshold for predictions.

3.3.4 PU learning as cost-sensitive learning

PU learning tasks under SCAR or SAR assumption can be re-written as cost-sensitive learning. The idea is to assign specific weights to labeled and unlabeled instances in order to correct the bias of non-traditional classifiers. This approach is introduced in this subsection and will be studied in further details in Sections 3.4 and 3.5 for $0-1$ loss, leading to theoretical guarantees (consistency, risk bounds). Numerical experiments will also be performed using a logistic loss function (cf. Section 3.6).

The cost-sensitive approach to PU learning is not restrained to $0-1$ and logistic loss functions. In fact, any binary classification loss function is suitable. Let l be a loss function, *i.e.* a positive function defined on $\mathbb{R} \times \{0, 1\}$. Plessis et al. (2014, 2015); Bekker et al. (2020) showed that

minimizing the expected risk $L(h) = \mathbb{E}\left[l\left(h(X), Z\right)\right]$ for $h : \mathbb{R}^d \to \mathbb{R}$ in a predefined set of functions $\mathcal{H}$, is equivalent to minimizing:

$$L_{PU}(h) = \pi \left(\mathbb{E}\left[l\left(h(X),1\right) - l\left(h(X),0\right) \mid Z=1\right]\right) + \mathbb{E}\left[l\left(h(X),0\right)\right] \tag{3.8}$$

$$= \mathbb{E}\left[\frac{\mathbb{1}_{Y=1}}{e(X)}\left(l\left(h(X),1\right) - l\left(h(X),0\right)\right)\right] + \mathbb{E}\left[l\left(h(X),0\right)\right] \tag{3.9}$$

where $\pi = \mathbb{P}\left(Z=1\right)$ denotes the class prior. Each of the expectations in Eq. 3.9 can be estimated using PU data if the propensity $e(\cdot)$ is known for labeled observations. A PU classifier is then estimated as a minimizer of the empirical risk:

$$\widehat{h_{PU}} \in \underset{h\in\mathcal{H}}{\text{Arginf}} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}_{Y_i=1}}{e(X_i)}\left(l\left(h(X_i),1\right) - l\left(h(X_i),0\right)\right)\right] + \frac{1}{n}\sum_{i=1}^{n} l\left(h(X_i),0\right) \ . \tag{3.10}$$

Under SCAR assumption, Plessis et al. (2014, 2015) suggest to minimize an empirical risk that does not depend on the constant propensity e, but on the class prior $\pi = \mathbb{P}\left(Z=1\right)$. The estimator rely on an empirical estimate of Eq. 3.8. In this case, the classifier is identified as:

$$\widehat{h_{PU}} \in \underset{h\in\mathcal{H}}{\text{Arginf}} \frac{\pi}{n_L}\sum_{i=1}^{n}\left[Y_i\ \left(l\left(h(X_i),1\right) - l\left(h(X_i),0\right)\right)\right] + \frac{1}{n}\sum_{i=1}^{n} l\left(h(X_i),0\right) \ .$$

Multiple choices of loss functions were discussed. A crucial remark is that even if the loss function l is convex, the PU learning optimization problem is not necessarily convex because the difference of two convex functions $\left(l\left(\cdot,1\right) - l\left(\cdot,0\right)\right)$ is not necessarily convex. In fact Plessis et al. (2015) show that for the PU problem to remain convex, the function $l\left(\cdot,1\right) - l\left(\cdot,0\right)$ must be linear, which is the case for logistic loss, but not for hinge loss (used in Support Vector Machine).

Besides, the unbiased risk estimate methodology was extended to more complicated models including deep learning architectures by Kiryo et al. (2017). When the number of parameters increases, the cost-sensitive PU learning tends to overfit rapidly. They explained this trend by the fact that the empirical risk in PU learning can take negative values whereas the true risk is always positive. Then, an over-parameterized model tends to reach a minimal value that is negative and this results in severe overfitting. To address this issue, the authors introduced a non-negative risk estimator under SCAR assumption:

$$\widehat{h_{nnPU}} \in \underset{h\in\mathcal{H}}{\text{Arginf}} \frac{\pi}{n_L}\sum_{i=1}^{n} Y_i\left(l\left(h(X_i),1\right)\right) + \max\left\{0, \frac{1}{n}\sum_{i=1}^{n} l\left(h(X_i),0\right) - \frac{\pi}{n_L}\sum_{i=1}^{n} Y_i\, l\left(h(X_i),0\right)\right\} \ .$$

It can be extended to SAR assumption following Bekker et al. (2020), by minimizing the following non-negative empirical risk:

$$\widehat{h_{nnPU}} \in \underset{h\in\mathcal{H}}{\text{Arginf}} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}_{Y_i=1}}{e(X_i)}\left(l\left(h(X_i),1\right)\right)\right] + \max\left\{0, \frac{1}{n}\sum_{i=1}^{n} l\left(h(X_i),0\right) - \frac{\mathbb{1}_{Y_i=1}}{e(X_i)} l\left(h(X_i),0\right)\right\} \ .$$

The term with the maximum operates as a regularization by preventing the estimated negative risk (the risk of predicting 1 when Z is 0) from taking negative values.

As previously mentioned, the weights used in the empirical risk depend either on the class prior (under SCAR assumption), or on the propensity (under SAR assumption). Hence, for this method to be applied, one needs to know this additional information. In practice, this is rarely the case. For instance, under the SCAR assumption, π is usually estimated which is another challenge: class prior estimation for PU learning has been an important research topic over the last decade (cf. Blanchard et al., 2010; Plessis et al., 2016; Jain et al., 2016; Bekker and Davis, 2018a; Garg et al., 2021). Under the more general SAR assumption, the propensity is required, at least for labeled instances. A solution to overcome this issue is to estimate the propensity, which however results in an even more difficult task than class prior estimation. The joint estimation of the classifier and propensity has been recently addressed by several authors (cf. Bekker and Davis, 2018b; Gong et al., 2021).

3.3.5 Ensemble methods

Ensemble techniques were suggested in order to address PU learning. Mordelet and Vert (2014) introduced a bagging-SVM classifier that consists in fitting several SVM classifiers with the labeled data and bootstrap samples of unlabeled data. The final classifier averages the predictions of all the classifiers. Claesen et al. (2015) resort to a similar procedure but also re-sample labeled instances. One of the interest of bagging procedures is to reduce the instability of the estimated classifier.

Yang et al. (2014) used an ensemble of different PU classifier for gene disease identification (Ensemble PU learning).

3.3.6 Deep Generative Modeling

Over the past few years, several authors have worked on deep learning methods adapted to PU learning under both SCAR and SAR assumptions, relying on recent advances in Deep Generative Models. Na et al. (2020) present a method based on variational autoencoders. Hou et al. (2017); Chiaroni et al. (2018) use GANs (Generative Adversarial Networks) in order to obtain a generative model for the distribution of negative instances (and also for positive instances in Hou et al. (2017)). The authors applied these methods to image PU classification tasks.

3.3.7 Conclusion

In the application of PU learning to the definition of a fatigue criterion, assuming SCAR assumption would be highly restrictive. This is why we will focus on the SAR assumption. For that purpose, we will consider approaches based on unbiased empirical risk minimization (cf. Section 3.4) and provide theoretical risk bounds (cf. Section 3.5). Chapter 4 will be dedicated to the application of PU learning to fatigue design.

3.4 - Unbiased risk estimators for PU learning

In this section, we focus on the definition of loss functions that enable learning in PU learning setting. After explaining why labeled-unlabeled classifiers are limited (Subsection 3.4.1), we will introduce an unbiased empirical risk for PU learning under the SCAR assumption (Subsection 3.4.2), which generalizes to the SAR assumption (Subsection 3.4.3).

3.4.1 Bias issue with labeled-unlabeled classification

A natural idea to address a PU learning problem is to consider labeled instances as positive and every unlabeled instance as negative. Standard classification methods then allow to identify a classifier $\widehat{g}_{NT}$. In the literature, such a classifier is called a *non-traditional classifier* (Elkan and Noto, 2008) because it is meant to give good predictions on Y instead of Z. As the number of training examples increases, we can then expect $\widehat{g}_{NT}$ to get closer to Bayes classifier $\tilde{g}^*$ for the classification of Y given X which is not what we are looking for. Indeed, $\tilde{g}^*$ is *a priori* different from g^* as the regression function $\tilde{\eta}(x) = \mathbb{P}(Y = 1|X = x)$ is different from $\eta(x)$ (cf. Equation 3.4).

Nevertheless, in specific situations, the non-traditional classifier is robust to PU learning label noise. Cannings et al. (2020) showed for example that $\tilde{g}^* = g^*$ if:

$$e(x) \geq \frac{1}{2\eta(x)}, \text{ for all } x \in \mathbb{R}^d \text{ such that } \eta(x) \geq \frac{1}{2} \,. \tag{3.11}$$

In fact, this is part of a more general result from Cannings et al. (2020) that encompasses binary classification with asymmetric and instance-dependent label noise. Under the conditions from Equation 3.11, any consistent non-traditional classifier is a consistent traditional classifier. In other words, as the training sample size increases, $\widehat{g}_{NT}$ gets closer to $\tilde{g}^*$ which is identical to g^*.

This condition requires every positive instance ($\eta(x) > \frac{1}{2}$) difficult to classify ($\eta(x)$ close to $\frac{1}{2}$) to have propensity close enough to 1. Instances easier to classify ($\eta(x)$ close to 1) can undergo label noise without harming the consistency. However, the label noise cannot exceed $\frac{1}{2}$ or, in other words, the propensity can never be smaller than $\frac{1}{2}$.

This condition is thus restrictive in the context of PU learning under the SAR assumption for two main reasons. On the one hand, in many realistic situations, the propensity (*i.e.* the probability for a positive instance to be labeled) is correlated to the difficulty of classifying the observation. A positive instance difficult to classify tends to have low propensity, which clearly violates the condition given in Equation 3.11. On the other hand, we cannot expect the propensity to be greater than $\frac{1}{2}$. In *text classification* or *spam review detection*, as the process of labeling is both difficult and time-consuming, only a small fraction of positive instances gets labeled, which suggests a propensity lower than $\frac{1}{2}$.

Before dealing with convergence rates, it is crucial to have methods for building consistent classifiers under more general conditions than Equation 3.11.

3.4.2 Unbiased empirical risk minimization under the SCAR assumption

In this subsection, we assume that the SCAR assumption is satisfied, which means that the propensity is constant:

$$e(x) = e_m > 0 \ .$$

In order to compensate for label noise due to PU Learning under the SCAR assumption, Plessis et al. (2014) showed in the case-control setting that a consistent classifier can be found by minimizing an unbiased version of the risk. Using the convenient decomposition of $\mathbb{P}_X$ distribution (Equation 3.1), the misclassification risk can be rewritten only with $\mathbb{P}_X$ and $\mathbb{P}_1$.

$$\begin{aligned} R(g) &= \pi \, \mathbb{P}_1(g(X) \neq 1) + (1-\pi)\, \mathbb{P}_0(g(X) \neq 0) \\ &= \pi \left(\mathbb{P}_1(g(X) \neq 1) - \mathbb{P}_1(g(X) \neq 0)\right) + \mathbb{P}_X(g(X) \neq 0) \, . \end{aligned} \tag{3.12}$$

Therefore, as labeled instances are a representative sub-sample of positive instances, a consistent classifier can be found by minimizing the following risk:

$$\widehat{R}_n^{SCAR}(g) = \frac{\pi}{N_L} \sum_{i=1}^{n} \mathbb{1}_{Y_i=1} \left[\mathbb{1}_{g(X_i)\neq 1} - \mathbb{1}_{g(X_i)\neq 0}\right] + \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{g(X_i)\neq 0}$$

where $N_L = \sum_{i=1}^{n} \mathbb{1}_{Y_i=1}$ is the number of labeled instances. In fact, Plessis et al. (2014) considered the case-control setting where the number of labeled instances N_L is fixed which is slightly different from our setting. One of the main properties of $\widehat{R}_n^{SCAR}(g)$ is that it is an unbiased estimate of the true risk, as we have:

$$\mathbb{E}\left[\widehat{R}_n^{SCAR}(g)\right] = \mathbb{P}(g(X) \neq Z) \ .$$

The proof of Plessis et al. (2014) extends to the one-sample-setting where N_L is random:

$$\mathbb{E}\left[\widehat{R}_n^{SCAR}(g)\right] = \sum_{i=1}^{n} \mathbb{E}\left[\frac{\pi}{N_L} \mathbb{1}_{Y_i=1} \mathbb{E}\left[\mathbb{1}_{g(X_i)\neq 1} - \mathbb{1}_{g(X_i)\neq 0} \,|\, Y_i\right]\right] + \mathbb{P}_X(g(X) \neq 0)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\pi}{N_L} \mathbb{1}_{Y_i=1} \left[\mathbb{P}\left(g(X_i) \neq 1 \,|\, Y_i\right) - \mathbb{P}\left(Y_i = 1, g(X_i) \neq 0 \,|\, Y_i\right)\right]\right] \\
&+ \mathbb{P}_X\left(g(X) \neq 0\right) \\
&= \pi \sum_{i=1}^{n} \mathbb{E}\left[\frac{\mathbb{1}_{Y_i=1}}{N_L} \left(\mathbb{P}_1\left(g(X) \neq 1\right) - \mathbb{P}_1\left(g(X) \neq 0\right)\right)\right] + \mathbb{P}_X\left(g(X) \neq 0\right) \qquad (3.13a) \\
&= \pi \left[\mathbb{P}_1\left(g(X) \neq 1\right) - \mathbb{P}_1\left(g(X) \neq 0\right)\right] + \mathbb{P}_X\left(g(X) \neq 0\right) \ . \qquad (3.13b)
\end{aligned}
$$

Equation 3.13a results from the fact that under the SCAR assumption the conditional distribution of X given $Y = 1$ is the same as the conditional distribution of X given $Z = 1$ ($\mathbb{P}_1$). Finally, Equation 3.13b matches the decomposition of Equation 3.12, ending the proof.

Computing the risk $\widehat{R}_n^{SCAR}$ requires π to be known. Alternatively, another empirical risk can be written:

$$
\widehat{R'}_n^{SCAR}(g) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbb{1}_{Y_i=1}}{e_m} \left(\mathbb{1}_{g(X_i)\neq 1} - \mathbb{1}_{g(X_i)\neq 0}\right) + \mathbb{1}_{g(X_i)\neq 0}\right] \ . \qquad (3.14)
$$

This risk remains unbiased and consistent but requires the knowledge of the constant propensity e_m instead of the class prior π. The unbiasedness of $\widehat{R'}_n^{SCAR}$ will be proved in Subsection 3.4.3 as a special case of the more general SAR setting.

3.4.3 Extension to PU learning under the SAR assumption

For now, PU learning under the SAR assumption is a difficult problem and there are only a few results in the literature (cf. Bekker et al., 2020; He et al., 2018; Gong et al., 2021). We recall that empirical risk minimization under the SCAR assumption requires extra knowledge on the model (class prior or propensity). In practice, these parameters are usually estimated (cf. Blanchard et al., 2010; Plessis et al., 2016; Jain et al., 2016; Bekker and Davis, 2018a; Garg et al., 2021). In order to provide a consistent empirical risk in the SAR setting, additional assumptions are needed to avoid identifiability issues. In the literature, different settings have been studied. He et al. (2018) assume that the propensity $e(x)$ is an increasing function of $\eta(x)$. Bekker and Davis (2018b) and Gong et al. (2021) suggest a *parametric* model on the propensity. Bekker et al. (2020) studied the case where the propensity is known for labeled instances which enables an empirical risk minimization approach similar to Plessis et al. (2014).

In this chapter, following Bekker et al. (2020), we will focus on PU learning under the SAR assumption where the propensity is known for labeled instances. We argue that this setting is sufficient to derive interesting risk bounds and assess the difficulty of PU learning tasks. However restrictive this assumption may seem, we insist that only the propensity for labeled instances is needed; therefore an exhaustive knowledge of the propensity is not required. In practice, the propensity can be estimated using prior knowledge on the labeling mechanism (when available) or by defining a parametric model on the propensity (Bekker and Davis, 2018b; Gong et al., 2021).

Under this assumption, Bekker et al. (2020) generalized the empirical risk in Equation 3.14 to obtain an unbiased empirical risk for PU learning under the SAR assumption. More particularly, they define the following loss function:

$$
\begin{aligned}
r_{SAR}\left(g, (X, Y)\right) &= \frac{\mathbb{1}_{Y=1}}{e(X)} \left(\mathbb{1}_{g(X)\neq 1} - \mathbb{1}_{g(X)\neq 0}\right) + \mathbb{1}_{g(X)\neq 0} \\
&= \frac{\mathbb{1}_{Y=1}}{e(X)} \left(2\,\mathbb{1}_{g(X)\neq 1} - 1\right) + \mathbb{1}_{g(X)\neq 0} \ .
\end{aligned}
$$

The empirical risk is then the empirical mean:

$$
\widehat{R}_n^{SAR}(g) = \frac{1}{n} \sum_{i=1}^{n} r_{SAR}\left(g, (X_i, Y_i)\right) \ . \qquad (3.15)
$$

This time, the labeled instances are weighted by the inverse of their propensity. Clearly, $\widehat{R'}_n^{SCAR}$ in Equation 3.14 is a special case of $\widehat{R}_n^{SAR}$ under the SCAR assumption ($e(x) = e_m$).

Bekker et al. (2020) studied maximum deviations between this latter empirical risk $\widehat{R}_n^{SAR}$ and the empirical risk for standard classification $\widehat{R}_n$. They then used concentration inequalities to derive an upper bound with high probability on the deviations between the two quantities. As we are interested in studying the deviations between $\widehat{R}_n^{SAR}$ and the true risk R directly, we compute the total expectation of $\mathbb{E}\left[\widehat{R}_n^{SAR}(g)\right] = \mathbb{E}\left[r_{SAR}\left(g,(X,Y)\right)\right]$ shedding light on the fact that for any g, $\widehat{R}_n^{SAR}(g)$ is an unbiased estimate of the true risk $R(g)$.

$$
\begin{aligned}
\mathbb{E}\left[r_{SAR}\left(g,(X,Y)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[r_{SAR}\left(g,(X,Y)\right)|X\right]\right] \\
&= \mathbb{E}\left[\frac{1}{e(X)}\left(\mathbb{1}_{g(X)\neq 1} - \mathbb{1}_{g(X)\neq 0}\right)\mathbb{P}\left(Y=1|X\right)\right] + \mathbb{P}_X\left(g(X)\neq 0\right) \\
&= \mathbb{E}\left[\frac{1}{e(X)}\left(\mathbb{1}_{g(X)\neq 1} - \mathbb{1}_{g(X)\neq 0}\right)\eta(X)e(X)\right] + \mathbb{P}_X\left(g(X)\neq 0\right) \\
&= \mathbb{E}\left[\left(\mathbb{1}_{g(X)\neq 1} - \mathbb{1}_{g(X)\neq 0}\right)\mathbb{1}_{Z=1}\right] + \mathbb{P}_X\left(g(X)\neq 0\right) \\
&= \pi\left(\mathbb{P}_1\left(g(X)\neq 1\right) - \mathbb{P}_1\left(g(X)\neq 0\right)\right) + \mathbb{P}_X\left(g(X)\neq 0\right) \\
&= R(g)\ .
\end{aligned}
$$

where the last line comes from Equation 3.12. Then, $\widehat{R}_n^{SAR}$ is indeed unbiased:

$$
\mathbb{E}\left[\widehat{R}_n^{SAR}(g)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[r_{SAR}\left(g,(X_i,Y_i)\right)\right] = \mathbb{P}\left(g(X)\neq Z\right) = R(g)\ . \tag{3.16}
$$

3.5 - Upper and lower risk bounds for PU learning under the SAR assumption

We are now in a position to state our results. We first present an upper bound on the excess risk for PU learning under the SAR assumption. We then show that the rate achieved is almost optimal by providing a lower bound on the minimax risk. Both bounds explicitly quantify the impact of label noise due to PU learning.

3.5.1 An upper bound for PU learning excess risk under the SAR assumption

We recall that, in PU learning, the true classes $(Z_i)_{1\leq i\leq n}$ are no longer available for training. A classifier is then built as a minimizer of the unbiased empirical risk introduced in Equation 3.15:

$$
\widehat{g}_{PU} \in \operatorname*{Argmin}_{g\in\mathcal{G}} \widehat{R}_n^{SAR}(g)\ .
$$

We recall that the risk $\widehat{R}_n^{SAR}$ is unbiased (Equation 3.16) and we will denote $\overline{R_n^{SAR}}$ the centered empirical risk:

$$
\overline{R_n^{SAR}}(g) = \widehat{R}_n^{SAR}(g) - \mathbb{P}\left(g(X)\neq Z\right)\ .
$$

Bekker et al. (2020) study the deviations between $\widehat{R}_n^{SAR}\left(\widehat{g}_{PU}\right)$ and $\widehat{R}_n\left(\widehat{g}_{PU}\right)$ and provide an upper bound in the case where $\mathcal{G}$ is a *finite* family of classifiers. Besides, the influence of $e(\cdot)$ on the upper bound is not discussed. Our objective here is to provide a uniform upper bound on $\ell(\widehat{g}_{PU}, g^*)$ and explicitly show its dependence in $e(\cdot)$. In our setting, $\mathcal{G}$ is an *infinite* set of functions. Its complexity is controlled by its VC dimension $V < +\infty$. Following Massart and Nédélec (2006), we consider the following separability assumption which is key to work with the possibly uncountable class $\mathcal{G}$:

(A_1) There exists a countable subset $\mathcal{G}'$ dense in $\mathcal{G}$ in the sense that for each $g \in \mathcal{G}$, there exists a sequence $(g_k)_{k\geq 0}$ such that, for every $(x, y) \in \mathbb{R}^d \times \{0, 1\}$:

$$r_{SAR}\left(g_k, (x, y)\right) \underset{k\to+\infty}{\longrightarrow} r_{SAR}\left(g, (x, y)\right) \ .$$

In addition, we want our upper bound on the excess risk to explicitly account for the difficulty of the classification task. Then, as $|2\eta(x) - 1|$ quantify the difficulty of classifying x, we introduce the following assumption (Massart and Nédélec, 2006):

(A_2) $\quad \exists\, h > 0,\ \forall x \in \mathbb{R}^d,\ |2\eta(x) - 1| \geq h \ .$

Assumption (A_2) will be referred to as *Massart noise* assumption in the rest of the chapter.

We are now able to state our upper bound for PU learning under the SAR assumption.

Theorem 3.5.1: Upper risk bound for PU learning under the SAR assumption

Let $\widehat{g}_{PU}$ be a minimizer of the unbiased empirical risk for PU learning under the SAR assumption:

$$\widehat{g}_{PU} \in \underset{g\in\mathcal{G}}{\operatorname{Argmin}}\, \widehat{R}_n^{SAR}(g) \ .$$

Suppose that separability (A_1) and Massart noise (A_2) assumptions hold, and that the propensity $e(\cdot)$ is greater than $e_m > 0$. Then, we have the following upper bound on the excess risk:

$$\mathbb{E}\left[\ell\left(\widehat{g}_{PU}, g^*\right)\right] \leq \kappa_1 \left[\frac{V}{n\, e_m\, h}\left(1 + \log\left(\frac{n\, h^2}{V} \vee 1\right)\right) \wedge \sqrt{\frac{V}{n\, e_m}}\right] \tag{3.17}$$

where $\kappa_1 > 0$ is an absolute constant.

Remarks: The upper bound in Equation 3.17 is uniform on the set of probability distributions for which $g^* \in \mathcal{G}$ and Massart noise condition (A_2) is satisfied with constant h ($\mathcal{P}(\mathcal{G}, h)$). This can be re-written as follows:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{G},h)} \mathbb{E}\left[\ell\left(\widehat{g}_{PU}, g^*\right)\right] \leq \kappa_1 \left[\frac{V}{n\, e_m\, h}\left(1 + \log\left(\frac{n\, h^2}{V} \vee 1\right)\right) \wedge \sqrt{\frac{V}{n\, e_m}}\right] \tag{3.18}$$

The assumption $e(x) \geq e_m$ is an additional assumption on the label noise. As the biased regression function is $\tilde{\eta}(x) = \eta(x)\, e(x)$ (cf. Equation 3.4), this assumption together with assumption (A_2) control the difficulty of the PU learning task.

In Equation 3.17, the convergence rate is of order $\mathcal{O}(\frac{V}{n\, h\, e_m})$ (if we let aside the logarithmic term) when h is higher than $\sqrt{V/n\, e_m}$. When h becomes smaller than $\sqrt{V/n\, e_m}$, the rate is of order $\mathcal{O}(\sqrt{V/n\, e_m})$. These two regimes are analogous to standard classification risk bounds as recalled in Subsection 3.1.2. In particular, when $e_m = 1$, all positive examples are labeled and we are then in a standard classification setting ($Y = Z$). In this case, the upper bound exactly matches the known upper bound rates in the standard classification setting (Equation 3.3 and Equation 3.2). Conversely, as e_m gets lower, the upper bound increases. This means without surprise that PU learning deteriorates the generalization bound: Theorem 3.5.1 quantifies this effect through the coefficients $1/e_m$ and $1/\sqrt{e_m}$.

Let N_L be the number of labeled instances in the training set. Under the SCAR assumption ($e(x) = e_m$), $n\, e_m$ from Equation 3.17 is linked to the expectation of the number of labeled

instances in the training set:

$$\mathbb{E}[N_L] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{Y_i=1}\right] = n\,\mathbb{P}(Y=1) = n\,\pi\,e_m$$

where $\pi = \mathbb{P}(Z=1)$ is the class prior. This illustrates a natural intuition on PU learning: the upper bound on the excess risk is related to the number of fully labeled examples. Hence, good prediction performances cannot be expected if the number of labeled examples among the positives is too low, or equivalently if the propensity is too low.

The detailed proof of Theorem 3.5.1 can be found in Subsection 3.7.1. It consists in establishing controls on the variance of increments of $r_{SAR}(\cdot)$ and uniform bounds on the empirical process $\left(\overline{R_n^{SAR}}(g)\right)_{g\in\mathcal{G}}$. A general risk bound result for empirical risk minimizers is then applied.

So far, we have provided an upper bound on generalization risk for unbiased empirical risk minimization in PU learning under the SAR assumption. There is however no proof that this rate is optimal. In other words, is there another procedure that can learn a classifier $\widehat{g}$ that outperforms $\widehat{g}_{PU}$? A lower bound will help to answer this question.

3.5.2 A lower bound on the minimax risk

In order to assess the optimality of the upper bound (Equation 3.17), we analyze and provide a lower bound on the minimax risk.

The minimax risk is the risk of the classification procedure that performs best in the worst case. For any given estimate $\widehat{g}$, we recall that its generalization risk is measured as $\mathbb{E}[\ell(\widehat{g}, g^*)]$. The minimax risk is denoted $\mathcal{R}(\mathcal{G}, h)$ and is defined as follows:

$$\mathcal{R}(\mathcal{G}, h) = \inf_{\widehat{g}\in\mathcal{G}} \left[\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{G},h)} \mathbb{E}[\ell(\widehat{g}, g^*)]\right]$$

where the infimum is taken over the set of functions $\widehat{g}$ of $(X_i, Y_i)_{1\leq i\leq n}$ such that $\widehat{g}$ belongs to $\mathcal{G}$.

The bound in Equation 3.18 is an obvious upper bound on the minimax risk. Theorem 3.5.2 establishes a lower bound on the minimax risk for PU learning under the SCAR assumption. Proposition 3.5.1 extends it to the SAR assumption.

Theorem 3.5.2: Lower bound on the minimax risk under the SCAR assumption

Suppose that $V \geq 2$ and $n\,e_m \geq V$. Let $h' = \sqrt{\frac{V}{n\,e_m}}$.
Assuming $e(x) = e_m$, $\forall x \in \mathbb{R}^d$, there exists an absolute constant $\kappa_2 > 0$ such that:

(C_1) if $h \geq h'$:

$$\mathcal{R}(\mathcal{G}, h) \geq \kappa_2 \frac{V-1}{h\,n\,e_m}\ ; \tag{3.19}$$

(C_2) if $h \leq h'$:

$$\mathcal{R}(\mathcal{G}, h) \geq \kappa_2 \sqrt{\frac{V-1}{n\,e_m}}\ . \tag{3.20}$$

Remarks

The lower bounds in Theorem 3.5.2 explicitly depend on V, n, h and e_m. The cases (C_1) and (C_2) highlight a trade-off between the expected number of fully labeled instances (proportional

to $n\,e_m$), the complexity of the model V and the noise condition (A_2) represented by h. The restriction of these results to the standard classification setting ($e_m = 1$) exactly matches existing results (see Massart and Nédélec, 2006). Theorem 3.5.2 moreover provides the influence of propensity e_m in PU learning framework under the SCAR assumption. As for the upper bound (cf. Theorem 3.5.1), the lower bound (Equation 3.19) is affected the same way with a degradation of order $1/e_m$ over the minimax rate when Massart noise condition (A_2) is satisfied with h high enough, in case (C_1). In this case, the lower bound rate almost matches the upper bound up to a logarithmic factor. In the second case (C_2), the lower bound (Equation 3.20) is of order $\sqrt{V/n\,e_m}$ which exactly matches the rate of the upper bound in this regime. In this sense, $\widehat{g}_{PU}$ obtained through unbiased empirical risk minimization is almost optimal as it almost achieves the minimax convergence rates.

The detailed proof of Theorem 3.5.2 can be found in the Paragraph a of Subsection 3.7.2. It makes use of similar arguments as for minimax lower bounds in the standard classification setting. First, the expression of the minimax risk is simplified by choosing a specific set of probabilities satisfying the noise conditions. Then Assouad lemma (Yu, 1997) is applied to provide a lower bound on this simplified expression, where the singularity of PU learning mainly interferes.

To extend the result to the SAR assumption, we need an extra condition:

(A_3) $\forall \varepsilon > 0$, $\exists\,(x_1, ..., x_V) \in \left(\mathbb{R}^d\right)^V$ shattered by $\mathcal{G}$ and such that:

$$\sup_{i \in \{1,...,V\}} e(x_i) \leq e_m + \varepsilon \ .$$

This assumption is technical. It is used in the first step of the proof of the minimax lower bound as it allows us to choose a convenient family of discrete probability distributions satisfying the noise assumptions. Assumption (A_3) is fulfilled in natural situations, for example, when $e(\cdot)$ is continuous and $\mathcal{G}$ is the set of linear classifiers in $\mathbb{R}^d$.

Proposition 3.5.1: Lower bound on minimax risk under the SAR assumption

Theorem 3.5.2 extends to the SAR assumption if the propensity $e(\cdot)$ greater than $e_m > 0$ and if assumption (A_3) is satisfied.

The proof of the above proposition can be found in the Paragraph b of Subsection 3.7.2. The same remarks as for Theorem 3.5.2 remain valid under the SAR assumption when assumption (A_3) is satisfied. In particular, in regimes (C_1) and (C_2), the minimax rate still matches the upper bound rate Equation 3.17 up to the logarithmic factor.

3.6 - Numerical experiments

In this section, we now study numerically the performances of the PU estimator minimizing the empirical risk of Equation 3.15. Subsection 3.6.1 describes the simulation setting. Then, we show that using the PU learning empirical risk enables to estimate the right classifier when the naive non-traditional approach fails to do so (Subsection 3.6.2). Then, the convergence rates are studied empirically, emphasizing how they are affected by both the sample size and the propensity (Subsection 3.6.3). Finally, Subsection 3.6.4 extends the study beyond the scope of the theoretical results by considering a convex loss function instead of the $0-1$ loss in the PU learning empirical risk.

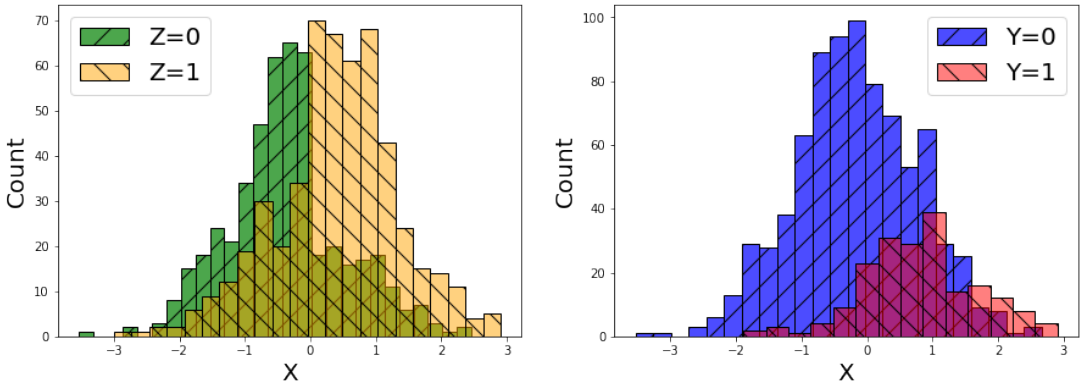

Figure 3.1: Example of simulation with $n = 1000$, $h = 0.5$ and logistic propensity (with $e_m = 0.05$). On the left, the histograms of the positive and negative instances (true labels); on the right, the histograms of labeled and unlabeled instances (noisy labels). We can note a significant overlap between the distributions in both figures.

3.6.1 Simulation setting

We consider examples of PU learning tasks in one dimension ($d = 1$), as the minimization of the $0-1$ loss function remains tractable. Of course, a natural solution to overcome this difficulty is to resort to convex surrogate loss, which will be discussed in Subsection 3.6.4. For every i, the covariate X_i is drawn i.i.d. according to the standard normal distribution. The distribution of Z given X is chosen to satisfy Assumption (A_2) with $h > 0$. We simplify it by choosing $\mathbb{P}(Z = 1|X = x)$ equal to either $\frac{1+h}{2}$ (when $X \geq 0$) or $\frac{1-h}{2}$ (when $X < 0$). For each i:

$$Z_i \sim \mathcal{B}\left(\frac{1+h}{2}\mathbb{1}_{X_i \geq 0} + \frac{1-h}{2}\mathbb{1}_{X_i<0}\right) ,$$

where $\mathcal{B}$ denotes the Bernoulli distribution and h is a constant in $(0, 1)$.

Under this setting, the Bayes classifier f^* is known explicitly:

$$f^*(x) = \mathbb{1}_{x \geq 0} .$$

In order to generate the labels $(Y_i)_{1 \leq i \leq n}$, we define two models of propensity:

1. constant propensity (SCAR assumption):

$$e(x) = e_m, \text{ with } e_m > 0 \tag{3.21}$$

2. logistic propensity (SAR assumption):

$$e(x) = \max\left(e_m, \frac{1}{1+e^{x-1}}\right), \text{ with } e_m > 0 . \tag{3.22}$$

This propensity mimics a selection bias on the positive instances. It is lower bounded by $e_m > 0$ and thus respects the assumptions of Theorem 3.5.1.

An example of simulation is shown in Figure 3.1. In these simulations, the objective is to use only the observations $(X_i, Y_i)_{1 \leq i \leq n}$ (cf. Fig. 3.1, right) to estimate the classifier.

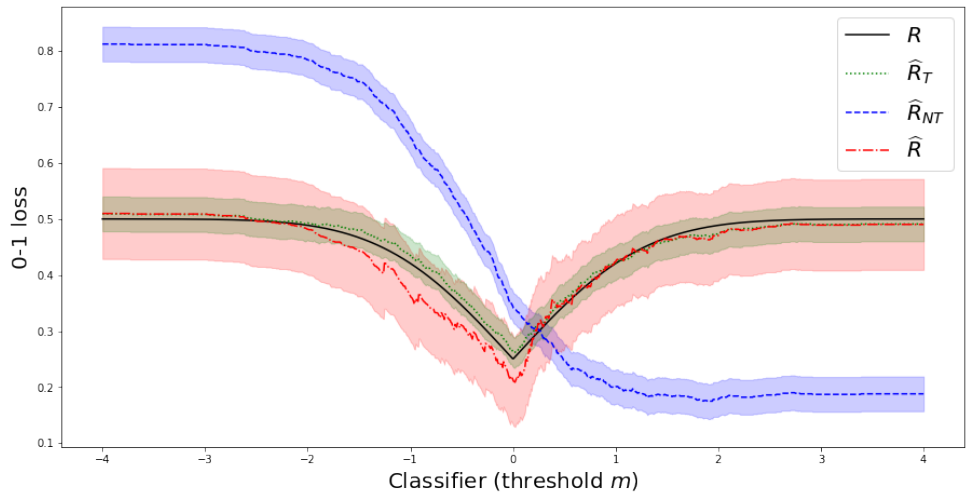

Figure 3.2: Comparison between the different empirical risk functions and R evaluated on the simulated data depicted on Fig. 3.1: $\widehat{R}_T$ in green (cf. Eq. 3.23), $\widehat{R}_{NT}$ in blue (cf. Eq. 3.24) and $\widehat{R}$ in red (cf. Eq. 3.25). Abscissa represents the threshold m corresponding to the classifier $x \mapsto \mathbb{1}_{x \geq m}$. Estimated curves are represented within the 95% confidence intervals.

3.6.2 PU learning empirical risks

In this simulation setting, searching a linear classifier is equivalent to identifying a threshold $m \in \mathbb{R}$ for the classification. Hence, we consider the following hypothesis space $\mathcal{G} = \{x \mapsto \mathbb{1}_{x \geq m},\ m \in \mathbb{R}\}$.

We recall that different empirical risks exist to approximate the true risk $R(f)$:

1. the traditional approach in standard binary classification using the proportion of missclassified training instances:

$$\widehat{R}_T(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{g(X_i) \neq Z_i} \tag{3.23}$$

   which is inapplicable in PU learning context since the true classes are unobserved;

2. the non-traditional approach uses an analogous empirical risk by ignoring the label noise due to PU learning:

$$\widehat{R}_{NT}(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{g(X_i) \neq Y_i}\ ; \tag{3.24}$$

3. the unbiased empirical risk that accounts for the propensity:

$$\widehat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbb{1}_{Y_i=1}}{e(X_i)} \left(2\mathbb{1}_{g(X_i) \neq 1} - 1 \right) + \mathbb{1}_{g(X_i) \neq 0} \right]. \tag{3.25}$$

These three empirical risks are compared to the true risk in Fig. 3.2. Despite a higher variance, $\widehat{R}$ correctly estimates R and can at least identify its minimum. Instead, $\widehat{R}_{NT}$ is clearly a biased estimate of R and fails to identify the right classifier.

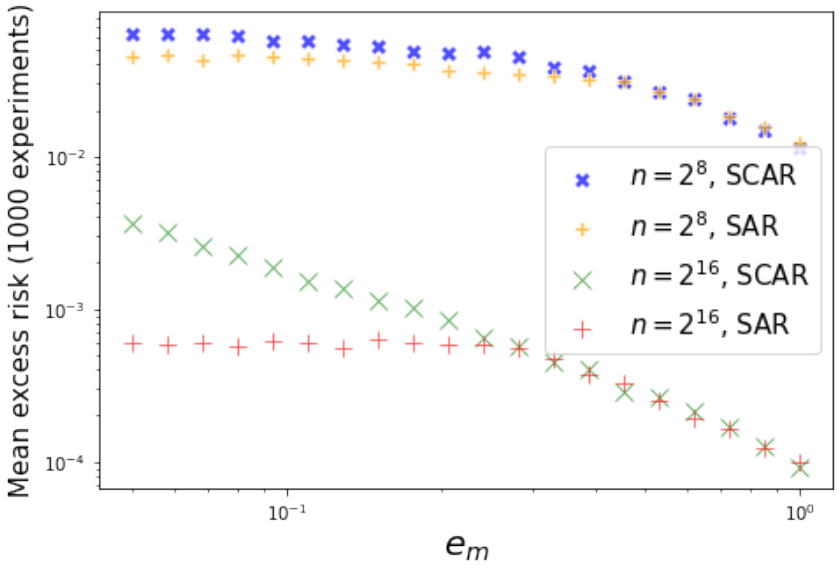

Figure 3.3: Mean excess risk as a function of e_m for n fixed ($n = 2^8$ and $n = 2^{16}$), log scale on both axes.

3.6.3 Convergence rates

We now illustrate numerically the rates of convergence of PU learning empirical risk minimizers when the number of observations n and the minimum propensity e_m change. To do so, we repeat B times the following steps:

1. simulate a training set of size n with propensity $e(\cdot)$ (chosen among the models described in Eq. 3.21 and 3.22)

2. estimate a classifier $\widehat{g}$ as a minimizer of PU learning empirical risk.

3. evaluate the excess risk $\ell(\widehat{g}, g^*) = R(\widehat{g}) - R(g^*)$

We then estimate the mean excess risk by the empirical average over the B runs. Multiple experiments were realized with n ranging from 100 to 30,000 and e_m ranging from 0.05 to 1. Massart noise parameter is fixed for these experiments: $h = 0.25$. This value for parameter h was chosen low enough to allow both convergence regimes (cf. Eq. 3.17) to be observable.

The results for both propensity models are presented in Fig. 3.3, 3.4 and 3.6, each on logarithmic scale. In Fig. 3.3, we clearly see that the mean excess risk decreases when e_m increases but the decrease happens faster when n is high for both SCAR and SAR situations. On the other hand, we can notice that the SCAR selection bias ($e(x) = e_m$) jeopardizes the classification more than the SAR selection bias ($e(x) \geq e_m$). In addition, a small value of e_m in the SAR propensity model does not alter much the propensity function, we could even choose $e_m = 0$. This means that, in practice, we can allow the propensity to take arbitrary small values as far as this occurs with small probability. When e_m is greater, the SAR propensity behaves almost like the SCAR propensity (cf. Eq. 3.22).

Fig. 3.4 shows that the mean excess risk effectively depends on the term $n \times e_m$ which is closely related to the expected number of labeled instances. We find the two convergence speeds: fast when $n \times e_m$ is higher than 200, slow when $n \times e_m$ is lower. The behaviour of the mean excess risk under SAR assumption confirms the observations of Fig. 3.3: when n is fixed, the performances remain almost identical when e_m vanishes and follow SCAR propensity when e_m tends to 1.

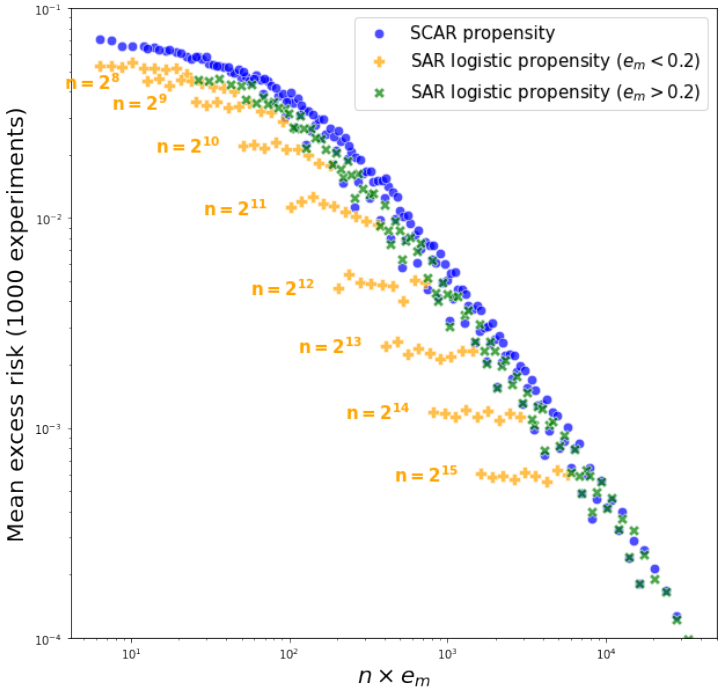

Figure 3.4: Mean excess risk for both propensity models (SCAR and logistic SAR) for different values of n and e_m, log scale on both axes. Experiments under logistic SAR model are split in two: in orange those for which $e_m \leq 0.2$, in green the rest. Aligned orange points correspond to equal values of n (cf. annotations).

The results of mean excess risk as a function of $n \times e_m$ under SCAR assumption remain a bit scattered even if the general trend is well captured. Representing $n \times \frac{e_m}{2-e_m}$ in abscissa seems to better explain the observed results (cf. Fig. 3.6). In fact, looking closer at the theoretical result, the proof shows that the risk upper bound depends on e_m through the term $\frac{2-e_m}{e_m}$ which was then upper bounded by $\frac{2}{e_m}$ in the final result (cf. Subsection 3.7.1). A linear regression is performed at the logarithmic scale on the results under SCAR assumption for $n\,e_m \geq 200$. The estimated slope is close to -1 ($[-1.012, -0.993]$, 95% confidence interval). Hence, this allows to identify the value of the exponent on $\left(\frac{2-e_m}{n\,e_m}\right)$ and asserts the decrease in $\mathcal{O}\left(\frac{2-e_m}{n\,e_m}\right)$ of the excess risk.

3.6.4 Using a tractable loss function

The theoretical results of Section 3.5 are based on a procedure that consists in minimizing an empirical risk based on $0-1$ loss. If this framework is convenient to study theoretical properties of PU learning, it is not directly useful for applications because the minimization of $0-1$ loss requires solving difficult combinatorial optimization problems. It is thus natural to resort to convex loss functions instead. The use of convex loss functions adapted to PU learning was discussed in Plessis et al. (2014) under SCAR assumption. Bekker et al. (2020) present a natural extension to SAR assumption.

In this section, we investigate the use of a continuous and convex loss function which is of course more suitable for applications.

Coming back to our simulation example ($d = 1$), we change the estimation of the classifier. Replacing the $0-1$ loss function by a logistic loss yields the following optimization problem:

$$\widehat{g} \in \underset{g \in \mathcal{G}}{\operatorname{Argmin}}\, \widehat{R}_C(g),$$

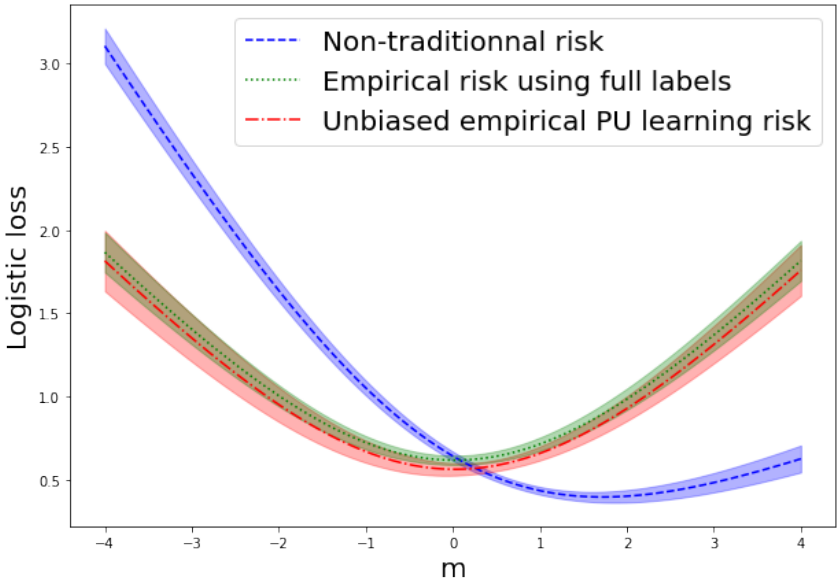

Figure 3.5: Comparison between the different loss functions: in green the traditional logistic loss function using the true classes $(Z_i)_{1\leq i\leq n}$, in blue the non-traditional logistic loss function ignoring the label noise, in red the logistic loss adapted to PU learning (cf. Eq. 3.26). Around the curves are represented the 95% confidence intervals.

where $\mathcal{G} = \{x \mapsto x - m,\ m \in \mathbb{R}\}$ and where

$$\widehat{R}_C(g) = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\mathbb{1}_{Y_i=1}}{e(X_i)}g(X_i) + \log\left(1+e^{g(X_i)}\right)\right]. \tag{3.26}$$

The corresponding classifier is $\widehat{f}(x) = \mathbb{1}_{\widehat{g}(x)\geq 0}$. This time the loss function (as a function of $m \in \mathbb{R}$) is continuous, convex and one can check that it remains an unbiased estimate of the logistic risk:

$$\mathbb{E}\left[\widehat{R}_C(g)\right] = \mathbb{E}\left[-Zg(X) + \log\left(1+e^{g(X)}\right)\right]. \tag{3.27}$$

As in Subsection 3.6.2, we can check that the PU learning empirical risk provides a good estimate of the true one contrary to the non-traditional risk, *i.e.* ignoring the label noise (cf. Fig. 3.5).

We perform similar experiments as in Subsection 3.6.3, using now the logistic loss function to estimate the classifier. We study the mean excess risk under both propensity models (cf. Fig. 3.7). The numerical results confirm, at least for the SCAR propensity model, that the mean excess risk depends on $\frac{n\,e_m}{2-e_m}$. Unless, this time, the estimated slope is around $-\frac{1}{2}$ ($[-0.508, -0.495]$, 95% confidence interval) which suggests a decrease of the mean excess risk at the parametric rate $\mathcal{O}\left(\sqrt{\frac{2-e_m}{n\,e_m}}\right)$. This is not surprising as the logistic regression is a parametric model optimized through maximum of likelihood, it is then normal to retrieve a parametric rate of convergence on the mean excess risk.

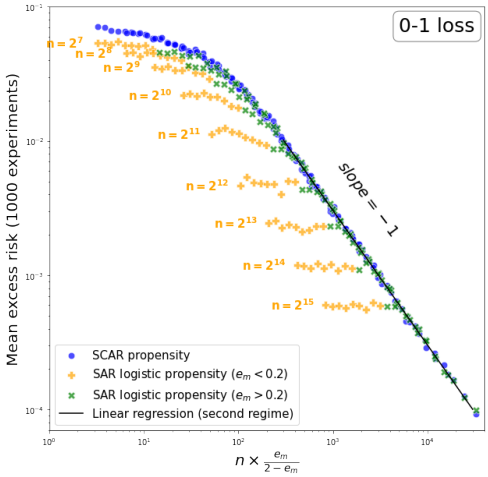

Figure 3.6: Estimated mean excess risk for both propensity models (SCAR and logistic SAR), log scale on both axes. Contrary to Fig. 3.4, abscissa corresponds to $n \times \frac{e_m}{2-e_m}$.

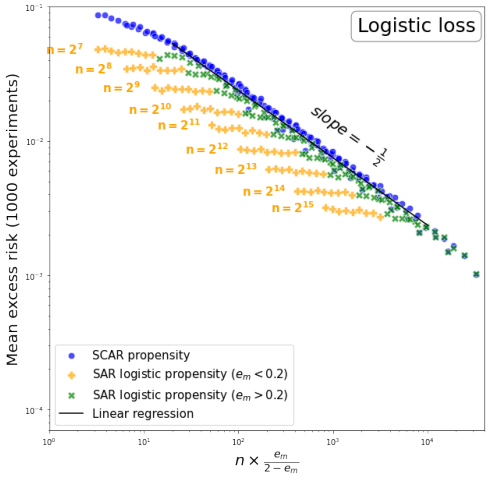

Figure 3.7: Mean excess risk as a function of $\frac{2-e_m}{n\,e_m}$. A linear regression on the experiments under the SCAR assumption allows to estimate a slope close to $-\frac{1}{2}$ which asserts a convergence rate in $\sqrt{n\,e_m}/\sqrt{2-e_m}$.

3.6.5 Conclusion

In this section, we provided a numerical study of convergence rates for PU learning under the SAR assumption. Our simulations highlight two convergence rates depending on the value of $n\,e_m$ as stated by Theorem 3.5.1: slow in $\sqrt{n\,e_m}$ or fast in $1/n\,e_m$. Besides, we extended our experiments to a tractable loss function that is suitable for applications. In this case, we observed a parametric convergence rate on the mean excess risk.

3.7 - Proofs and technical lemmas

This section contains the proofs of the theoretical results presented in Section 3.5. Subsection 3.7.1 presents the proof of Theorem 3.5.1 and Subsection 3.7.2 the proofs of Theorem 3.5.2 and Proposition 3.5.1. Finally, Subsection 3.7.3 recall useful definitions and properties concerning the universal entropy metrics and Subsection 3.7.4 provides the proof of some technical lemmas.

3.7.1 Proof of Theorem 3.5.1

The proof is organized as follows. We first state a general upper bound result for empirical risk minimizers adapted to the case where the loss function takes values in an arbitrary interval $[a, b]$ with $a < b$ (cf. Paragraph a). Then, we show that the PU learning loss function satisfies the assumptions of this general result (cf. Paragraph b). Finally, we deduce the upper bound as the solution of a fixed point equation (cf. Paragraph c).

a. General risk upper bound on empirical risk minimizers

We begin by stating a general upper bound theorem for empirical risk minimizers.

Theorem 3.7.1: General upper bound for empirical risk minimizers

Let r be an unbiased loss function with values in $[a,b]$, $\widehat{R}_n$ the empirical risk, $\overline{R_n}$ the centered empirical risk. Let g^* denote the Bayes classifier and let $\widehat{g}$ be a minimizer of the empirical risk over a class $\mathcal{G}$ for which we assume separability condition (A_1). Let ℓ denote the excess risk. We assume that:

(B_1) there exists a positive and symmetric function d such that for any couple of classifiers (g, g'):

$$Var\left[r\left(g', (X,Y)\right) - r\left(g, (X,Y)\right)\right] \leq d^2\left(g', g\right);$$

(B_2) there exists a non-decreasing function w continuous on $\mathbb{R}_+$, such that $x \mapsto \frac{w(x)}{x}$ is non-increasing on $\mathbb{R}_+^*$, with $w(\sqrt{b-a}) \geq b-a$ and ensuring for any classifier g:

$$d(g^*, g) \leq w\left(\sqrt{\ell(g^*, g)}\right);$$

(B_3) there exists a non-decreasing function Φ continuous on $\mathbb{R}_+$, such that $x \mapsto \frac{\Phi(x)}{x}$ is non-increasing with $\Phi(b-a) \geq b-a$ and ensuring:

$$\forall h \in \mathcal{G}', \sqrt{n}\,\mathbb{E}\left[\sup_{g \in \mathcal{G}', d(g,h) \leq \sigma} \overline{R_n}(h) - \overline{R_n}(g)\right] \leq \Phi(\sigma).$$

for every positive σ such that $\Phi(\sigma) \leq \sqrt{n}\frac{\sigma^2}{b-a}$, where $\mathcal{G}'$ comes from separability condition (A_1).

Then there exists an absolute constant $\kappa > 0$ such that:

$$\mathbb{E}\left[\ell(g^*, \widehat{g})\right] \leq \kappa\, \varepsilon_*^2, \tag{3.28}$$

where ε_* is the unique positive solution of the following equation:

$$\sqrt{n}\,\varepsilon_*^2 = \Phi\left(w\left(\varepsilon_*\right)\right). \tag{3.29}$$

Proof. The above result follows from the application of Massart and Nédélec's theorem (2006, Theorem 2) using the re-scaled risk $\tilde{r} = \frac{r-a}{b-a}$ and the functions $\tilde{d}(g, g') = \frac{d(g,g')}{b-a}$, $\tilde{w}(x) = \frac{1}{b-a}\, w\left(x\sqrt{b-a}\right)$ and $\tilde{\Phi}(x) = \frac{1}{b-a}\Phi\left((b-a)\,x\right)$. This leads to the upper bound in Equation 3.28 solution of Equation 3.29. □

It is worth noting that now, contrary to Massart and Nédélec's original result, (B_2) and (B_3) explicitly involve the length of the interval $[a,b]$. This will be accounted for in our proof.

b. Verification of assumptions of Theorem 3.7.1 in the PU learning setting

We first recall the definition and the main property of PU learning loss function as defined in Subsection 3.4.3. We then exhibit three functions d, w, Φ fulfilling conditions (B_1), (B_2) and (B_3). Hence we show that the general upper bound result (*i.e.* Theorem 3.7.1) can be applied in PU learning context.

In the context of PU learning under the SAR assumption, we recall that the loss function r_{SAR} is defined as follows:

$$r_{SAR}\left(g,(X,Y)\right) = \frac{\mathbb{1}_{Y=1}}{e(X)}\left(2\,\mathbb{1}_{g(X)\neq 1}-1\right)+\mathbb{1}_{g(X)\neq 0}$$

where $e(x) = \mathbb{P}\left(Y=1 \,|\, Z=1, X=x\right)$ is the propensity assumed to be known for labeled observations. Knowing that the propensity is greater than $e_m > 0$, the loss function is then at values in $\left[1-\frac{1}{e_m}, \frac{1}{e_m}\right]$, an interval of length:

$$C_e = \frac{2}{e_m} - 1 \; . \tag{3.30}$$

We have seen that this empirical risk is an unbiased estimate of the true risk (cf. Equation 3.16):

$$\mathbb{E}\left[r_{SAR}\left(g,(X,Y)\right)\right] = \mathbb{P}\left(g(X)\neq Z\right) \; .$$

In order to apply the general upper bound theorem (Theorem 3.7.1) to the PU learning risk minimizer, we need to identify three functions d, w, Φ satisfying conditions (B_1), (B_2) and (B_3). These functions are crucial since the upper bound is the solution of a fixed point equation involving them. The choice of functions d, w and Φ will be a consequence of Propositions 3.7.1, 3.7.2 and 3.7.3.

Proposition 3.7.1

For any pair of classifiers (g, g'):

$$Var\left[r_{SAR}\left(g',(X,Y)\right) - r_{SAR}\left(g,(X,Y)\right)\right] \leq 2\,C_e\,\mathbb{E}\left[|g(X)-g'(X)|^2\right] \; ,$$

where C_e is given by Equation 3.30.

Remark A direct consequence of the above proposition is that the function d defined as:

$$d(g,g') = \sqrt{2C_e}\sqrt{\mathbb{E}\left[|g(X)-g'(X)|^2\right]} \tag{3.31}$$

satisfies condition (B_1).

Proof. We first provide an upper bound on the variance of increments of r_{SAR}:

$$\begin{aligned} Var\left[r_{SAR}\left(g\right) - r_{SAR}\left(g'\right)\right] &\leq \mathbb{E}\left[\left(r_{SAR}\left(g\right) - r_{SAR}\left(g'\right)\right)^2\right] \\ &= \mathbb{E}\left[\left(g(X)-g'(X)\right)^2\left(1-\frac{2\mathbf{1}_{Y=1}}{e(X)}\right)^2\right] \\ &= \mathbb{E}\left[\left(g(X)-g'(X)\right)^2\mathbb{E}\left[\left(1-\frac{2\mathbf{1}_{Y=1}}{e(X)}\right)^2 |X\right]\right] \end{aligned}$$

$$= \mathbb{E}\left[\left(g(X) - g'(X)\right)^2 \left(1 + 4\eta(X)\frac{1 - e(X)}{e(X)}\right)\right] \tag{3.32a}$$

$$\leq \left(1 + 4\frac{1 - e_m}{e_m}\right) \mathbb{E}\left[\left(g(X) - g'(X)\right)^2\right] \tag{3.32b}$$

$$\leq 2\, C_e\, \mathbb{E}\left[\left(g(X) - g'(X)\right)^2\right] \ .$$

We then use the fact that $\mathbb{E}\left[\mathbb{1}_{Y=1}|X\right] = \eta(X)\, e(X)$ to get Equation 3.32a. And Equation 3.32b results from the fact that $\eta(X)$ is less than 1 and $e(X)$ is greater than e_m. □

Proposition 3.7.2

For any classifier g:

$$d(g, g^*) \leq \sqrt{\frac{2\, C_e}{h}} \sqrt{\ell(g, g^*)} \ .$$

for d defined in Equation 3.31.

Remark As a consequence, the function w defined as:

$$w(x) = \sqrt{\frac{2\, C_e}{h}}\, x \ . \tag{3.33}$$

satisfies Assumption (B_2): w is continuous on $\mathbb{R}_+$, non-decreasing, such that $x \mapsto \frac{w(x)}{x}$ is non-increasing and $w\left(\sqrt{C_e}\right) \geq C_e$, and such that:

$$d(g^*, g) \leq w\left(\sqrt{\ell\left(g^*, g\right)}\right) \ .$$

Let $h' = \sqrt{V/n\, e_m}$. Then, the function

$$w_0(x) = \sqrt{2\, C_e} \vee x\, \sqrt{2\, C_e / h'} \tag{3.34}$$

also satisfies assumption (B_2).

Proof. The excess risk can be expressed in terms of $\eta(X)$ as follows:

$$\begin{aligned} \ell(g, g^*) &= \mathbb{P}\left(g(X) \neq Z\right) - \mathbb{P}\left(g^*(X) \neq Z\right) \\ &= \mathbb{E}\left[|g(X) - g^*(X)|^2 |2\,\eta(X) - 1|\right] \ . \end{aligned} \tag{3.35}$$

Then, using the noise assumption (A_2) and the definition of d (cf. Equation 3.31), we have the following lower bound on the excess risk:

$$\begin{aligned} \ell(g, g^*) &= \mathbb{E}\left[\left(g(X) - g^*(X)\right)^2 |2\,\eta(X) - 1|\right] \\ &\geq h\, \mathbb{E}\left[\left(g(X) - g^*(X)\right)^2\right] \\ &= \frac{h}{2\, C_e} d^2(g, g^*) \ . \end{aligned}$$

Taking the square root on both sides finishes the proof. □

The next proposition states the existence of Φ fulfilling (B_3). We recall that the subset $\mathcal{G}' \subset \mathcal{G}$ is given by the separability assumption (A_1) and that the constant C_e is defined in Equation 3.30.

Proposition 3.7.3

Assume $\mathcal{G}$ has finite VC dimension V and $\mathcal{G}'$ is given by separability assumption (A_1). There exists an absolute constant $K \geq 1$ such that the function Φ defined as

$$\Phi(\sigma) = K\sigma\sqrt{V\left[1 + \log\left(\frac{C_e}{\sigma} \vee 1\right)\right]} \tag{3.36}$$

satisfies:

$$\sqrt{n}\,\mathbb{E}\left[\sup_{g\in\mathcal{G}', d(g,h)\leq\sigma} \overline{R_n^{SAR}}(g_0) - \overline{R_n^{SAR}}(g)\right] \leq \Phi(\sigma)$$

for all $g_0 \in \mathcal{G}'$ and for every σ such that $\Phi(\sigma) \leq \sqrt{n}\,\frac{\sigma^2}{C_e}$.

Proof. We consider a fixed $g_0 \in \mathcal{G}'$ along the proof and use the notation:

$$W = \sup_{g\in\mathcal{G}', d(g,g_0)\leq\sigma} \overline{R_n^{SAR}}(g_0) - \overline{R_n^{SAR}}(g)\ .$$

The main steps of the proof are: (i) rewrite W as the supremum of an empirical process over a class of functions; (ii) split the expression of W in two terms depending on the sign of $(g_0(x) - g(x))$ (W^+ and W^-) that will be processed similarly and independently; (iii) provide an upper bound on $\mathbb{E}[W^+]$ using a symmetrization principle (cf. Bousquet et al., 2003); (iv) apply a chaining inequality and Haussler bound (Bousquet et al., 2003; Massart and Nédélec, 2006); (v) a few calculations finish the proof. This proof uses the notion of entropy metrics: the definition and some useful properties are recalled in Subsection 3.7.3.

(i) We start by rewriting the expression inside the supremum in W:

$$\begin{aligned}
\overline{R_n^{SAR}}(g_0) - \overline{R_n^{SAR}}(g) &= \widehat{R}_n^{SAR}(g_0) - \widehat{R}_n^{SAR}(g) - \mathbb{E}\left[\widehat{R}_n^{SAR}(g_0) - \widehat{R}_n^{SAR}(g)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(r_{SAR}\left(g_0, (X_i, Y_i)\right) - r_{SAR}\left(g, (X_i, Y_i)\right)\right) - \mathbb{E}\left[\widehat{R}_n^{SAR}(g_0) - \widehat{R}_n^{SAR}(g)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(g_0(X_i) - g(X_i)\right)\left(\frac{2\,\mathbb{1}_{Y_i=1}}{e(X_i)} - 1\right) - \mathbb{E}\left[\left(g(X) - g_0(X)\right)\left(\frac{2\,\mathbb{1}_{Y=1}}{e(X)} - 1\right)\right] \\
&= (\mathbb{P}_n - \mathbb{P})\,(f_g),
\end{aligned}$$

where $\mathbb{P}_n f_g$ and $\mathbb{P} f_g$ denote the empirical mean and the expectation of the function f_g:

$$f_g : (x, y) \mapsto \left(g_0(x) - g(x)\right)\left(\frac{2\,\mathbb{1}_{y=1}}{e(x)} - 1\right).$$

Hence, denoting $\mathcal{F}(\sigma) = \{f_g : g \in \mathcal{G}',\ d(g_0, g) \leq \sigma\}$, we can write W as the supremum of the empirical process $(\mathbb{P}_n - \mathbb{P})(\cdot)$ over the set of functions $\mathcal{F}(\sigma)$:

$$W = \sup_{f\in\mathcal{F}(\sigma)} (\mathbb{P}_n - \mathbb{P})(f).$$

(ii) For any $g \in \mathcal{G}'$, we can decompose f_g depending on the sign of $(g_0(x) - g(x))$:

$$f_g(x,s) = \left(\frac{2\,\mathbb{1}_{s=1}}{e(x)} - 1\right)\mathbb{1}_{g(x)>g_0(x)} - \left(\frac{2\,\mathbb{1}_{s=1}}{e(x)} - 1\right)\mathbb{1}_{g_0(x)>g(x)}\ .$$

Then, introducing the following classes of functions

$$\mathcal{F}^+(\sigma) = \left\{ f : \mathbb{R}^d \times \{0,1\} \to \mathbb{R},\ \exists\, g \in \mathcal{G}',\ f(x,s) = \left[\frac{2\mathbb{1}_{s=1}}{e(X)} - 1\right]\mathbb{1}_{g(x)>g_0(x)}\ ,\ d(g_0,g) \le \sigma \right\}$$

$$\mathcal{F}^-(\sigma) = \left\{ f : \mathbb{R}^d \times \{0,1\} \to \mathbb{R},\ \exists\, g \in \mathcal{G}',\ f(x,s) = \left[\frac{2\mathbb{1}_{s=1}}{e(X)} - 1\right]\mathbb{1}_{g(x)<g_0(x)}\ ,\ d(g_0,g) \le \sigma \right\}$$

and the corresponding suprema

$$W^+ = \sup_{f \in \mathcal{F}^+(\sigma)} (\mathbb{P}_n - \mathbb{P})(f)$$
$$W^- = \sup_{f \in \mathcal{F}^-(\sigma)} (\mathbb{P} - \mathbb{P}_n)(f),$$

we decompose $\mathbb{E}[W]$ as follows:

$$\mathbb{E}[W] \le \mathbb{E}\left[W^+\right] + \mathbb{E}\left[W^-\right].$$

We now process both terms separately focusing on W^+ (the proof for the other term is almost identical).

(iii) We first apply a symmetrization principle to provide an upper bound on $\mathbb{E}[W^+]$ depending on a Rademacher average (cf. Bousquet et al., 2003):

$$\mathbb{E}\left[W^+\right] \le \frac{2}{n}\,\mathbb{E}\left[\sup_{f \in \mathcal{F}^+(\sigma)} \sum_{i=1}^{n} \varepsilon_i f(X_i, Y_i)\right]$$

where $(\varepsilon_i)_{1 \le i \le n}$ are i.i.d. Rademacher variables (*i.e.* $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}$).

(iv) Let $\delta^2 = \sup_{f \in \mathcal{F}^+(\sigma)} \mathbb{P}_n\left(f^2\right) \vee \sigma^2$. We apply a chaining inequality (lemma A.2, Massart and Nédélec 2006) which gives us the following inequality:

$$\mathbb{E}\left[W^+\right] \le \frac{6}{\sqrt{n}}\mathbb{E}\left[\delta \sum_{j=0}^{+\infty} 2^{-j}\sqrt{H\left(2^{-j-1}\delta, \mathcal{F}_+(\sigma)\right)}\right] \tag{3.37}$$

where H is the universal entropy metric (cf. Subsection 3.7.3).

Let $\mathcal{A}_+ = \left\{\mathbb{1}_{g(x)>g_0(x)}, g \in \mathcal{G}'\right\}$, which can be consider as a set of classifiers and has VC dimension V at most. Using the fact that $H(\cdot, \mathcal{F}_+(\sigma))$ is non-increasing (cf. Proposition 3.7.4), we have $\forall j \ge 0$:

$$H\left(2^{-j-1}\delta, \mathcal{F}_+(\sigma)\right) \le H\left(2^{-j-1}\sigma, \mathcal{F}_+(\sigma)\right)\ .$$

Applying Proposition 3.7.5, we obtain the following upper bound on the entropy of $\mathcal{F}_+(\sigma)$ in terms of the entropy of $\mathcal{A}_+$:

$$H\left(2^{-j-1}\delta, \mathcal{F}_+(\sigma)\right) \le H\left(2^{-j-1}\frac{\sigma}{C_e},\ \mathcal{A}_+(\sigma)\right)\ .$$

We are then in a position to apply Haussler bound (Proposition 3.7.6), to get an upper bound on the entropy in terms of the VC dimension of $\mathcal{A}_+$ which is no more than V:

$$H\left(2^{-j-1}\delta, \mathcal{F}_+(\sigma)\right) \le \kappa V\left(1 + \log\left(2^{j+1}\frac{C_e}{\sigma} \vee 1\right)\right) \tag{3.38}$$

for some absolute constant $\kappa > 1$.

(v) Injecting Equation 3.38 in Equation 3.37, we get:

$$\begin{aligned}\mathbb{E}\left[W^+\right] &\leq 6\sqrt{\frac{\kappa\, V}{n}}\left[\sum_{j=0}^{+\infty} 2^{-j}\sqrt{1+\log\left(2^{j+1}\frac{C_e}{\sigma}\vee 1\right)}\right]\mathbb{E}\left[\delta\right] \\ &\leq C(\sigma)\sqrt{\frac{V}{n}}\mathbb{E}\left[\delta\right] \qquad (3.39a) \\ &\leq C(\sigma)\sqrt{\frac{V}{n}}\sqrt{\mathbb{E}\left[\delta^2\right]} \qquad (3.39b)\end{aligned}$$

where $C(\sigma) = 12\ (1+\log(2))\ \sqrt{\kappa}\sqrt{1+\log\left(\frac{C_e}{\sigma}\vee 1\right)}$. Equation 3.39a is a consequence of technical Lemma 3.7.1 in Subsection 3.7.4, Equation 3.39b follows from Cauchy-Schwartz inequality.

Now, we provide an upper bound on $\mathbb{E}\left[\delta^2\right]$ in terms of $\mathbb{E}\left[W^+\right]$:

$$\begin{aligned}\mathbb{E}\left[\delta^2\right] &\leq \sigma^2 + \mathbb{E}\left[\sup_{f\in\mathcal{F}_+(\sigma)}\mathbb{P}_n\left(f^2\right)\right] \\ &\leq \sigma^2 + C_e\,\mathbb{E}\left[\sup_{f\in\mathcal{F}_+(\sigma)}\mathbb{P}_n\left(f\right)\right] \\ &\leq \sigma^2 + C_e\,\mathbb{E}\left[\sup_{f\in\mathcal{F}_+(\sigma)}\left(\mathbb{P}_n-\mathbb{P}\right)\left(f\right)\right] + C_e\sup_{f\in\mathcal{F}_+(\sigma)}\mathbb{P}(f) \qquad (3.40)\end{aligned}$$

Let $f\in\mathcal{F}_+(\sigma)$ and define $g\in\mathcal{G}'$ such that $f(x,s) = \left[\frac{2\mathbb{1}_{s=1}}{e(x)}-1\right]\mathbb{1}_{g_0(x)>g(x)}$ (and $d(g_0,g)\leq\sigma$). We have:

$$\begin{aligned}\mathbb{P}(f) &= \mathbb{E}\left[\mathbb{E}\left[\frac{2\mathbb{1}_{Y=1}}{e(X)}-1|X\right]\mathbb{1}_{g_0(X)>g(X)}\right] \\ &= \mathbb{E}\left[\left(2\eta(X)-1\right)\mathbb{1}_{g_0(X)>g(X)}\right] \\ &\leq \mathbb{E}\left[|g_0(X)-g(X)|^2\right] \\ &= \frac{d^2(g_0,g)}{2\,C_e} \\ &\leq \frac{\sigma^2}{2\,C_e}\end{aligned}$$

using Equation 3.31 and the definition of $\mathcal{F}_+(\sigma)$. We can note that the above upper bound does not depend on $f\in\mathcal{F}_+(\sigma)$. Hence, we can use it in Equation 3.40 to obtain:

$$\mathbb{E}\left[\delta^2\right] \leq C_e\,\mathbb{E}\left[W^+\right] + \frac{3}{2}\sigma^2$$

Hence, coming back to $\mathbb{E}\left[W^+\right]$:

$$\mathbb{E}\left[W^+\right] \leq C(\sigma)\sqrt{\frac{V}{n}}\sqrt{C_e\,\mathbb{E}\left[W^+\right]+\frac{3}{2}\sigma^2}\ .$$

Taking the square on both sides and solving the second-order inequation in $\mathbb{E}\left[W^+\right]$ yields:

$$\mathbb{E}\left[W^+\right] \leq \frac{1}{2}C(\sigma)\sqrt{\frac{V}{n}}\left(C(\sigma)\,C_e\sqrt{\frac{V}{n}}+\sqrt{\frac{C(\sigma)^2\,C_e^2\,V}{n}+6\sigma^2}\right)\ .$$

Therefore, whenever $\sigma \geq C(\sigma)\,C_e\sqrt{\frac{V}{n}}$:

$$\sqrt{n}\,\mathbb{E}\left[W^+\right] \leq 2\,\sigma\,C(\sigma)\sqrt{V}\ .$$

We can prove a similar upper bound on $\mathbb{E}[W^-]$. If we define $\Phi(\sigma) = 4\,\sigma\,C(\sigma)\sqrt{V}$, for all σ such that $\Phi(\sigma) \leq \sqrt{n}\,\frac{\sigma^2}{C_e}$ (condition of Proposition 3.7.3):

$$\sigma \geq C(\sigma)\,C_e\sqrt{\frac{V}{n}}\ .$$

Hence, we have the desired upper bound on $\mathbb{E}[W]$:

$$\sqrt{n}\,\mathbb{E}[W] \leq \Phi(\sigma)\ .$$

Besides, the constant $K = 4\,C(\sigma)$ is greater than 1. □

c. Upper bounds on the risk

In the previous paragraph, we checked that Theorem 3.7.1 can be applied to PU learning under the SAR assumption. Hence, the upper bound on risk excess ε_*^2 is the unique solution to the fixed point equation:

$$\sqrt{n}\,\varepsilon_*^2 = \Phi\left(w(\varepsilon_*)\right) \tag{3.41}$$

where w is given in Equation 3.33 (or w_0 in Equation 3.34) and Φ in Equation 3.36.

$$w(x) = \sqrt{\frac{2\,C_e}{h}}\,x\ ,$$

$$w_0(x) = \sqrt{2\,C_e} \vee x\sqrt{\frac{2\,C_e}{h'}}\ ,$$

$$\Phi(\sigma) = K\sigma\sqrt{V\left[1 + \log\left(\frac{C_e}{\sigma} \vee 1\right)\right]}\ .$$

We cannot explicitly solve this equation, but we can provide an upper bound on the solution which is enough to complete the proof of Theorem 3.5.1. The choice of w as Equation 3.33 or Equation 3.34 leads to two different upper bounds that together complete the proof of Theorem 3.5.1.

First case Using the known definitions of w in Equation 3.33 and Φ in Equation 3.36, Equation 3.41 can be rewritten as:

$$\sqrt{n}\,\varepsilon_*^2 = K\,\varepsilon_*\sqrt{\frac{2\,C_e}{h}}\sqrt{V\left[1 + \log\left(\frac{\sqrt{C_e\,h}}{\sqrt{2}\varepsilon_*} \vee 1\right)\right]}$$

Because the logarithmic term is always non-negative and $K \geq 1$, we get:

$$\varepsilon_* \geq \sqrt{\frac{2\,C_e\,V}{n\,h}}\ .$$

Using this on the logarithmic term, we obtain the following upper bound on ε_*:

$$\begin{aligned}\varepsilon_* &\leq K\sqrt{\frac{2\,C_e\,V}{n\,h}}\sqrt{1 + \log\left(\frac{\sqrt{n}\,h}{2\sqrt{V}} \vee 1\right)} \\ &\leq K\sqrt{\frac{2\,C_e\,V}{n\,h}}\sqrt{1 + \log\left(\frac{n\,h^2}{V} \vee 1\right)}\end{aligned}$$

Noting that $C_e \leq \frac{2}{e_m}$, we get the desired result:

$$\varepsilon_*^2 \leq 4\,K^2\,\frac{V}{n\,h\,e_m}\left[1 + \log\left(\frac{n\,h^2}{V} \vee 1\right)\right]\ .$$

■

Second case We now consider Equation 3.41 where w is given by Equation 3.34. We can note that the logarithmic term is necessarily 0. If we assume that the solution ε_* of Equation 3.41 satisfies $\varepsilon_* \geq \sqrt{h'}$, then $w(x) = \varepsilon_* \sqrt{\frac{2\,C_e}{h'}}$. We obtain:

$$\varepsilon_*^2 \leq 4\,K^2 \sqrt{\frac{V}{n\,e_m}}\ .$$

Else, $\varepsilon_* \leq \sqrt{h'}$ which implies that

$$\varepsilon_*^2 \leq h' = \sqrt{\frac{V}{n\,e_m}}\ .$$

Both bounds provide the same convergence rate.

Paragraphs c and c together complete the proof of Theorem 3.5.1.

■

3.7.2 Proof of minimax lower bounds

We remind the reader that the minimax risk is defined as:

$$\mathcal{R}\left(\mathcal{G}, h\right) = \inf_{\widehat{g} \in \mathcal{G}} \left[\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{G}, h)} \mathbb{E}\left[\ell\left(\widehat{g}, g^*\right)\right] \right]\ .$$

The lower bound on the minimax risk is proved in Paragraph a for the SCAR assumption (cf. Theorem 3.5.2) and in Paragraph b for the SAR assumption (cf. Proposition 3.5.1).

a. Under the SCAR assumption (proof of Theorem 3.5.2)

The proof consists in exhibiting a finite subset of probability distributions on which the excess risk is worst. It is organised as follows: (i) we provide a lower bound on the minimax risk expression by restricting ourselves to this subset of distributions; (ii) we use Massart noise condition and simplify the remaining expression; (iii) the application of Assouad lemma finishes the proof.

(i) We start by introducing a family of probability distributions which plainly exploits the noise condition (A_2). Let $x_1, ..., x_V$ be V points of $\mathbb{R}^d$ shattered by $\mathcal{G}$. This is possible because the VC dimension of $\mathcal{G}$ is V. For some parameter $p < \frac{1}{V-1}$, we define a discrete probability distribution on $\{x_1, ..., x_V\} \subset \mathbb{R}^d$ verifying:

$$\mathbb{P}\left(X = x_i\right) = p \quad \forall\, i \leq V-1 \text{ and } \mathbb{P}\left(X = x_V\right) = 1 - p\ (V-1)\ .$$

For some binary vector $b \in \{0,1\}^{V-1}$, we consider $\mathbb{P}_b$ the probability distribution such that:

$$\forall\, 1 \leq i \leq V-1,\ \mathbb{P}_b\left(Z = 1 \,|\, X = x_i\right) = \frac{1}{2}\left[1 + (2\,b_i - 1)\ h\right]$$

for $h > 0$. We can consider by default that each point in $\mathbb{R}^d \setminus \{x_1, ..., x_{V-1}\}$ has class 0 almost surely. This has no incidence on the rest of the proof. Moreover:

$$\mathbb{P}_b\left(Y = 1 \,|\, X = x_i, Z = y\right) = y\, e(x_i)$$

following the definition of propensity.

Hence, $(\mathbb{P}_b)_{b\in\{0,1\}^{V-1}}$ defines a family of distributions on (X,Y) that satisfies Massart noise condition (A_2) at its limit: the regression function $|2\eta(x_i)-1|$ equals h for every $i \in \{1,...V-1\}$. Furthermore, for every $b \in \{0,1\}^{V-1}$, the Bayes classifier g_b^* is known:

$$\forall\, 1 \leq i \leq V-1,\ g_b^*(x_i) = b_i\ .$$

As $(x_1,...,x_V)$ is shattered by $\mathcal{G}$, g_b^* necessarily belongs to $\mathcal{G}$.
Hence, $(\mathbb{P}_b)_{b\in\{0,1\}^{V-1}} \subset \mathcal{P}(\mathcal{G},h)$ and therefore:

$$\mathcal{R}(\mathcal{G},h) \geq \inf_{\widehat{g}\in\mathcal{G}} \left[\sup_{b\in\{0,1\}^{V-1}} \mathbb{E}_b\left[\ell\left(\widehat{g},g_b^*\right)\right] \right]$$

where $\mathbb{E}_b$ denotes the expectation according to $\mathbb{P}_b$ distribution.

(ii) Let $\widehat{g}$ be a classifier, function of the training sample $(X_i,Y_i)_{1\leq i\leq n}$. We use the following decomposition of ℓ (cf. Equation 3.35):

$$\ell(\widehat{g},g_b^*) = \mathbb{E}\left[|2\eta(X)-1||\widehat{g}(X)-g_b^*(X)|\right]\ .$$

Combined with Massart noise condition (A_2), this yields:

$$\mathcal{R}(\mathcal{G},h) \geq h \inf_{\widehat{g}\in\mathcal{G}} \left[\sup_{b\in\{0,1\}^{V-1}} \mathbb{E}_b\left[|\widehat{g}(X)-g_b^*(X)|\right] \right]$$

For every $\widehat{g}$, we define $\widehat{b}$ such that:

$$\widehat{b} = \operatorname*{Argmin}_{b\in\{0,1\}^{V-1}} \mathbb{E}_X\left[|g_b^*(X)-\widehat{g}(X)|\right]$$

where the expectation is taken with respect to the marginal distribution of X and conditionally to the training sample. Hence, $\widehat{b}$ is a function of the training sample $(X_i,Y_i)_{1\leq i\leq n}$. By triangular inequality and then by definition of $\widehat{b}$:

$$|g_{\widehat{b}}^*(X)-g_b^*(X)| \leq |g_{\widehat{b}}^*(X)-\widehat{g}(X)| + |\widehat{g}(X)-g_b^*(X)| \leq 2\,|\widehat{g}(X)-g_b^*(X)|\ .$$

Hence:

$$\begin{aligned}
\mathcal{R}(\mathcal{G},h) &\geq \frac{h}{2} \inf_{\widehat{g}\in\mathcal{G}} \left[\sup_{b\in\{0,1\}^{V-1}} \mathbb{E}_b\left[|g_{\widehat{b}}^*(X)-g_b^*(X)|\right] \right] \\
&= \frac{h}{2} \inf_{\widehat{b}\in\{0,1\}^{V-1}} \left[\sup_{b\in\{0,1\}^{V-1}} \mathbb{E}_b\left[|g_{\widehat{b}}^*(X)-g_b^*(X)|\right] \right] \\
&= \frac{p\,h}{2} \inf_{\widehat{b}\in\{0,1\}^{V-1}} \left[\sup_{b\in\{0,1\}^{V-1}} \mathbb{E}_b\left[\sum_{i=1}^{V-1} \mathbb{1}_{b_i\neq\widehat{b}_i}\right] \right]
\end{aligned}$$

where the last line is obtained by developing the expectation according to the marginal distribution of X which is discrete.

(iii) With this simplified expression, we apply Assouad lemma (cf. Yu, 1997) which provides the following general lower bound:

$$\inf_{\widehat{b}\in\{0,1\}^{V-1}}\left[\sup_{b\in\{0,1\}^{V-1}}\mathbb{E}_b\left[\sum_{i=1}^{V-1}\mathbb{1}_{b_i\neq\widehat{b}_i}\right]\right]\geq\frac{V-1}{2}\left(1-\sqrt{\gamma\,n}\right)$$

where γ is an upper bound on the square Hellinger distance between probability distributions $\mathbb{P}_b$ and $\mathbb{P}'_b$ on (X,Y) when b and b' only differ on one coordinate. Using technical Lemma 3.7.2 in Subsection 3.7.4, we have the following upper bound on the square Hellinger distance $\mathcal{H}^2\left(\mathbb{P}_b,\mathbb{P}_{b'}\right)$:

$$\mathcal{H}^2\left(\mathbb{P}_b,\mathbb{P}_{b'}\right)\leq 2\,p\,e_m\,h^2\ . \tag{3.42}$$

Applying Assouad lemma together with Equation 3.42, we get the following inequality:

$$\mathcal{R}\left(\mathcal{G},h\right)\geq\frac{p\,h}{4}\left(V-1\right)\left(1-\sqrt{2\,p\,e_m\,h^2\,n}\right)\ .$$

In case (C_1), we choose $p=\frac{2}{9\,e_m\,h^2\,n}$ that is lower than $\frac{1}{V-1}$, we obtain the desired lower bound on $\mathcal{R}\left(\mathcal{G},h\right)$:

$$\mathcal{R}\left(\mathcal{G},h\right)\geq\frac{V-1}{54\,e_m\,h\,n}\ .$$

Else, in case (C_2), we choose $p=\frac{2}{9\,e_m\,h'^2\,n}$ where we recall that $h'=\sqrt{\frac{V}{n\,e_m}}$. As $h\leq h'$:

$$\mathcal{R}\left(\mathcal{G},h\right)\geq\mathcal{R}\left(\mathcal{G},h'\right)\geq\frac{V-1}{54\,e_m\,h'\,n}\geq\frac{1}{54\sqrt{2}}\sqrt{\frac{V-1}{n\,e_m}}\ .$$

■

b. Proof of Proposition 3.5.1

This proof relies on the same tools as SCAR assumption case. We alter (i) by choosing $x_1,...,x_V$ satisfying assumption (A_3) for $\varepsilon>0$. (ii) remains unchanged. In (iii), the upper bound in Equation 3.42 has to be replaced by $2\,p\,h^2\,(e_m+\varepsilon)$. This yields the following lower bounds:

1. in case (C_1):
$$\mathcal{R}\left(\mathcal{G},h\right)\geq\frac{V-1}{54\,\left(e_m+\varepsilon\right)\,h\,n}\ ;$$
2. in case (C_2):
$$\mathcal{R}\left(\mathcal{G},h\right)\geq\frac{1}{54\sqrt{2}}\sqrt{\frac{V-1}{\left(e_m+\varepsilon\right)\,h\,n}}\ .$$

It remains to note that these lower bounds are valid for any $\varepsilon>0$ to complete the proof.

■

3.7.3 Universal entropy metric and related properties

In this subsection, we recall some definitions and properties concerning the universal entropy metric. These properties are used for the proof of Proposition 3.7.3 in Subsection 3.7.1.

Let us consider $(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. random variables with values in $\mathbb{R}^d \times \{0,1\}$ and $\mathcal{F}$ a set of functions on $\mathbb{R}^d \times \{0,1\}$.

Definition 3.7.1: Universal entropy metric, cf. Massart and Nédélec (2006)

Let $\varepsilon > 0$ and $\mathbb{Q}$ be a probability measure. Define $h(\mathcal{F}, \varepsilon, \mathbb{Q})$ as the logarithm of the largest number N of functions $f_1, ..., f_N$ separated by a distance ε, namely $\mathbb{E}_{\mathbb{Q}}\left[(f_i(X,Y) - f_j(X,Y))^2\right] > \varepsilon^2$, $\forall i \neq j$. Then the universal entropy metric $H(\mathcal{F}, \varepsilon)$ is defined as:

$$H(\mathcal{F}, \varepsilon) = \sup_{\mathbb{Q}} h(\mathcal{F}, \varepsilon, \mathbb{Q}) \ .$$

Proposition 3.7.4

For a fixed $\mathcal{F}$, $H(\mathcal{F}, \cdot)$ is a decreasing function.

Proposition 3.7.5

Let ψ be a function defined on $\mathbb{R}^d \times \{0,1\}$ and $\mathcal{F}$ be a family of functions such that:

$$\mathcal{F} = \{(x,s) \mapsto \psi(x,s)\, g(x,s) \ , \ g \in \mathcal{G}\}$$

where $\mathcal{G}$ is another family of functions on $\mathbb{R}^d \times \{0,1\}$. Then:

$$\forall \varepsilon > 0, \ H(\mathcal{F}, \varepsilon) \leq H\left(\mathcal{G}, \frac{\varepsilon}{||\psi||_\infty}\right) \ .$$

Proof. Let $\mathbb{Q}$ be a probability distribution and N such that $h\left(\mathcal{G}, \frac{\varepsilon}{||\psi||_\infty}, \mathbb{Q}\right) < \log(N)$. Then, for any set of functions $g_1, ..., g_N$, there is $i \neq j$ such that $\mathbb{E}_{\mathbb{Q}}\left[(g_i(X,Y) - g_j(X,Y))^2\right] \leq \left(\frac{\varepsilon}{||\psi||_\infty}\right)^2$. This implies that $\mathbb{E}_{\mathbb{Q}}\left[(\psi(X,Y)\ [g_i(X,Y) - g_j(X,Y)])^2\right] \leq \varepsilon^2$ and then that $h(\mathcal{F}, \varepsilon, \mathbb{Q}) < \log(N)$. Then, we have that $h(\mathcal{F}, \varepsilon, \mathbb{Q}) \leq h\left(\mathcal{G}, \frac{\varepsilon}{||\psi||_\infty}, \mathbb{Q}\right)$. Considering the supremum over the probability distributions $\mathbb{Q}$, we obtain the desired result. □

Finally, we recall Haussler bound which provides an upper bound on the universal entropy metric of a set of classifiers in terms of its VC dimension.

Proposition 3.7.6: Haussler bound, cf. Bousquet et al. (2003)

Assuming that $\mathcal{F}$ is a set of indicator functions with finite Vapnik dimension V. Then, $\forall \varepsilon > 0$:

$$H(\mathcal{F}, \varepsilon) \leq \kappa V \left(1 + \log\left(\varepsilon^{-1} \vee 1\right)\right)$$

where $\kappa \geq 1$ is an absolute constant.

3.7.4 Technical lemmas

Lemma 3.7.1

Let $C_e > 1$ and $\sigma > 0$. Then:

$$\sum_{j=0}^{+\infty} 2^{-j} \sqrt{1 + \log\left(2^{j+1} \frac{C_e}{\sigma} \vee 1\right)} \leq 2\,(1 + \log(2)) \sqrt{1 + \log\left(\frac{C_e}{\sigma} \vee 1\right)}$$

Proof.

$$\begin{aligned}
\sum_{j=0}^{+\infty} 2^{-j} \sqrt{1 + \log\left(2^{j+1} \frac{C_e}{\sigma} \vee 1\right)} &\leq \sum_{j=0}^{+\infty} 2^{-j} \sqrt{1 + (j+1)\log(2) + \log\left(\frac{C_e}{\sigma} \vee 1\right)} \\
&\leq \sum_{j=0}^{+\infty} 2^{-j} \sqrt{1 + (j+1)\log(2)} \sqrt{1 + \log\left(\frac{C_e}{\sigma} \vee 1\right)} \\
&\leq \sum_{j=0}^{+\infty} 2^{-j} \left(1 + (j+1)\frac{\log(2)}{2}\right) \sqrt{1 + \log\left(\frac{C_e}{\sigma} \vee 1\right)} \\
&= 2\,(1 + \log(2)) \sqrt{1 + \log\left(\frac{C_e}{\sigma} \vee 1\right)}
\end{aligned}$$

□

Lemma 3.7.2

Let $x_1, ..., x_V$ be vectors of $\mathbb{R}^d$. Let e be a function on R^d with values in $(0, 1]$. Let $p \leq \frac{1}{V-1}$ and consider $(\mathbb{P}_b)_{b \in \{0,1\}^{V-1}}$ the family of probability distributions on $\{x_1, ..., x_V\} \times \{0, 1\}$ defined in Paragraph i of Subsection 3.7.2. If b and b' are binary vectors of $\{0, 1\}^{V-1}$ which only differ at coordinate i, then:

$$\mathcal{H}\left(\mathbb{P}_b, \mathbb{P}_{b'}\right) \leq 2\, p\, e(x_i)\, h^2 \ .$$

Proof. Recall that b and b' only differ at coordinate i, hence:

$$\begin{aligned}
\mathcal{H}^2\left(\mathbb{P}_b, \mathbb{P}_{b'}\right) &= \frac{1}{2} \sum_{j=1}^{V} \left(\sqrt{\mathbb{P}_b\left(X = x_j,\ Y = 1\right)} - \sqrt{\mathbb{P}_{b'}\left(X = x_j,\ Y = 1\right)}\right)^2 \\
&+ \frac{1}{2} \sum_{j=1}^{V} \left(\sqrt{\mathbb{P}_b\left(X = x_j,\ Y = 0\right)} - \sqrt{\mathbb{P}_{b'}\left(X = x_j,\ Y = 0\right)}\right)^2 \\
&= \frac{1}{2} \left(\sqrt{\mathbb{P}_b\left(X = x_i,\ Y = 1\right)} - \sqrt{\mathbb{P}_{b'}\left(X = x_i,\ Y = 1\right)}\right)^2 \\
&+ \frac{1}{2} \left(\sqrt{\mathbb{P}_b\left(X = x_i,\ Y = 0\right)} - \sqrt{\mathbb{P}_{b'}\left(X = x_i,\ Y = 0\right)}\right)^2 \ .
\end{aligned}$$

Let us now calculate the probabilities using the definition of $\mathbb{P}_b$:

$$\begin{aligned}
\mathbb{P}_b\left(X = x_i, S = 1\right) &= p\, \frac{e(x_i)}{2} \left[1 + (2b_i - 1)\ h\right] \ , \\
\mathbb{P}_b\left(X = x_i, S = 0\right) &= p \left(1 - \frac{e(x_i)}{2} \left[1 + (2b_i - 1)\ h\right]\right) \ .
\end{aligned}$$

Noting that either $(b_i, b_i') = (0,1)$ or $(b_i, b_i') = (1,0)$, we have in both cases:

$$\left(\sqrt{\mathbb{P}_b\left(X = x_i,\ Y = 1\right)} - \sqrt{\mathbb{P}_{b'}\left(X = x_i,\ Y = 1\right)}\right)^2 = p\, e(x_i)\ \left[1 - \sqrt{1 - h^2}\right]\ ,$$

and

$$\begin{aligned}
&\left(\sqrt{\mathbb{P}_b\left(X = x_i,\ Y = 0\right)} - \sqrt{\mathbb{P}_{b'}\left(X = x_i,\ Y = 0\right)}\right)^2 \\
&= p\left[2 - e(x_i) - 2\sqrt{1 - \frac{e(x_i)}{2}\left(1 + h\right)}\sqrt{1 - \frac{e(x_i)}{2}\left(1 - h\right)}\right]\ .
\end{aligned}$$

We then sum the two results together:

$$\begin{aligned}
\mathcal{H}^2\left(\mathbb{P}_b, \mathbb{P}_{b'}\right) &= \frac{p}{2}\left[2 - e(x_i)\sqrt{1 - h^2} - 2\sqrt{1 - e(x_i) + \frac{e(x_i)^2}{4}\left(1 - h^2\right)}\right] \\
&= p\left[1 - \frac{e(x_i)}{2}\sqrt{1 - h^2} - \sqrt{\left(1 - \frac{e(x_i)}{2}\sqrt{1 - h^2}\right)^2 - e(x_i)\left(1 - \sqrt{1 - h^2}\right)}\right] \\
&= p\left[1 - \frac{e(x_i)}{2}\sqrt{1 - h^2}\right]\left[1 - \sqrt{1 - \frac{e(x_i)\left(1 - \sqrt{1 - h^2}\right)}{\left[1 - \frac{e(x_i)}{2}\sqrt{1 - h^2}\right]^2}}\right] \\
&\leq \frac{p\ e(x_i)\left(1 - \sqrt{1 - h^2}\right)}{1 - \frac{e(x_i)}{2}\sqrt{1 - h^2}} \\
&\leq 2\, p\, e(x_i)\, h^2
\end{aligned}$$

In the above calculation, we applied the inequality $1 - \sqrt{1 - h^2} \leq h^2$ for $h^2 \in [0,1]$. □

4

Fatigue criterion construction through Positive-Unlabeled Learning

As seen in Chapter 2, a fatigue criterion is used to predict critical zones of a mechanical part (*i.e.* zones that may result in crack initiation for customers) given simulation results obtained from a numerical model. Fatigue rig tests are also carried out on prototypes. After the test, the part is inspected and zones with crack initiations are identified as critical. Every critical zone does not result in failure because fatigue crack initiation is a random event and depends on different parameters like the severity of the test or the observability of the crack. The construction of a fatigue criterion can then be viewed as a PU learning problem. The objective of this chapter is to define and calibrate a fatigue criterion through PU classification. The definition of the PU classification model involves a classifier and the propensity (cf. Chapter 3). In fatigue applications, the classifier is the fatigue criterion we want to estimate and the propensity represents the risk for a critical zone to fail during testing: it can then be modeled using fatigue lifetime models (S-N curves).

Section 4.1 explains how the construction of a fatigue criterion can be viewed as a PU learning task and specifies how the propensity can be modeled. In Section 4.2, we indicate which PU learning methodologies can be used for fatigue applications and provide parametric models for the classifier and the propensity suited for fatigue applications. In section 4.3 we discuss the identifiability of the proposed models. Section 4.4 presents SAR-EM, a methodology introduced by Bekker and Davis (2018b) that consists in jointly estimating the classifier and the propensity. We show how this general methodology applies to our parametric models. In section 4.5, we illustrate the interest of the methodology through numerical experiments on simulated data. Finally, in section 4.6 we apply the method to Stellantis data set and analyze the results.

4.1 - Fatigue criterion under the point of view of PU learning

From a statistical point of view, a fatigue criterion is a binary classifier that predicts the criticality of a zone. Since the tests cannot assert the safety of a zone, only the zones with observed crack initiations are labeled. This can be viewed as a label noise and more particularly as a PU label noise. The estimation of the fatigue criterion is then a PU learning task. In order to estimate the classifier, we need to account for the propensity, which is, by definition, the probability of crack initiation for a critical zone. It depends on the features describing the zone (local stresses, material, geometry, test conditions), hence the SCAR assumption does not hold, we are then under the SAR assumption. Besides, this propensity can be modeled using $S-N$ fatigue models.

4.1.1 Fatigue criterion and PU classification

The fatigue data set consists in a set of n individuals $(x_i, y_i)_{1\leq i\leq n}$ representing zones of a tested prototype.

- The covariate vector $x_i \in \mathbb{R}^d$ can be divided in two sub-vectors $(\tilde{x}_i, t_i) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ ($d_1 + d_2 = d$):
  - $\tilde{x}_i$ contains data from the numerical simulation results on the zone (material, stresses) when the part is subjected to a nominal loading representing the client objective (nominal severity, see Section 1.3).
  - t_i represents the testing conditions: initial severity, number of cycles before ending the test. We recall that the tests follow an accelerated test protocol in order to reduce their duration (fatigue tests with Locati method, see Section 1.3).
- The response y_i is a binary label indicating whether or not a crack initiated on the zone during the test ($y_i = 1$ or $y_i = 0$). We say that an instance is labeled if $y_i = 1$ and unlabeled if $y_i = 0$.

For a mechanical part to be valid, every zone must be set below the endurance limit. In other words, there should not be any crack initiation over the car lifetime. Therefore, for each zone with covariates $x_i = (\tilde{x}_i, t_i)$, we seek to predict a binary class Z_i indicating whether the zone may fail over the car lifetime ($Z_i = 1$, critical) or not ($Z_i = 0$, safe). In fact, the fatigue criterion should only depend on $\tilde{x}_i$ and not on t_i because we are only interested in predicting the criticality of the zone for the nominal severity. Our objective is then to estimate the classification rule η where:

$$\eta(x) = \mathbb{P}\left(Z = 1 \,|\, X = x\right) = \mathbb{P}\left(Z = 1 \,|\, \tilde{X} = \tilde{x}\right) \ . \tag{4.1}$$

We remark that the true classes $(Z_i)_{1\leq i\leq n}$ are not fully observed. The observed labels only provide limited information about the true classes. An observed crack asserts the criticality of a zone:

$$\mathbb{P}\left(Z = 1 \,|\, X = x, Y = 1\right) = 1 \ . \tag{4.2}$$

However, not every critical zone will fail during testing. In other words, the probability for a positive instance to be labeled is not necessarily 1:

$$\mathbb{P}\left(Y = 1 \,|\, X = x, Z = 1\right) = \ e(x) \in (0, 1]. \tag{4.3}$$

A particular case is when the propensity is equal to 1, which corresponds to the standard classification setting ($Z = Y$ almost surely).

In the estimation of a classifier η given a set of training observations $(x_i,\ y_i)_{1\leq i\leq n}$, the above quantity (Eq. 4.3) is the propensity and represents the probability for a critical zone to fail under specific testing conditions.

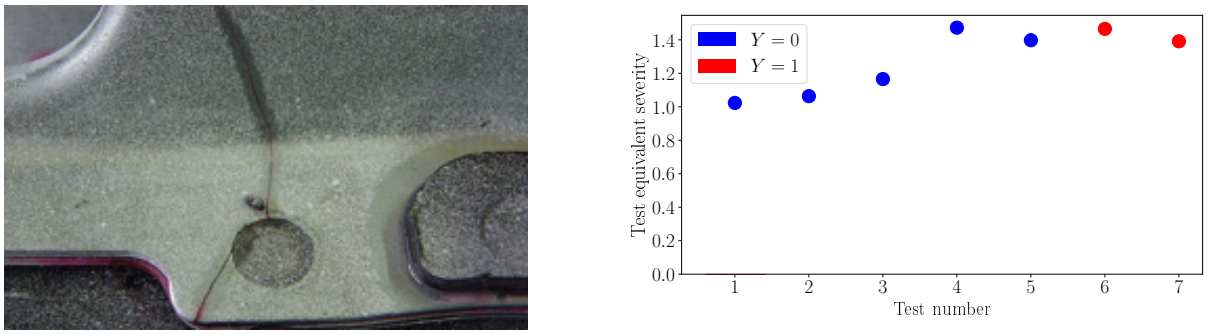

Figure 4.1: Example of critical zone on a cradle model under longitudinal sollicitations. The picture on the left represents the zone on the sixth specimen tested. The figure on the right represents the severity (multiplicative coefficient of the client objective F_n) of the seven tests performed. Only two (the red ones) resulted into a crack initiation.

4.1.2 Modeling the propensity in fatigue

We now consider the calibration of a fatigue criterion as a PU learning task: the goal is to estimate η (see Eq. 4.2) given a set of observations $(x_i, y_i)_{1 \leq i \leq n}$ affected by PU learning label noise. The probability for an instance to be labeled is:

$$\mathbb{P}(Y = 1 \,|\, X = x) = \mathbb{P}(Z = 1 \,|\, X = x) \times \mathbb{P}(Y = 1 \,|\, Z = 1, X = x) = \eta(\tilde{x}) \times e(\tilde{x}, t) \ . \quad (4.4)$$

Even if we are only interested in estimating η, the propensity e plays a crucial role and will need to be characterized to solve the classification problem.

The propensity represents the probability for a critical zone to fail during testing: it can be viewed as a selection bias, *i.e.* the probability for a positive instance with covariates x to be labeled. The phenomenon is observable considering the multiplicity of the tests performed for a same design. For instance, Fig. 4.1 (left) represents a crack initiation detected on a cradle part which asserts the criticality of the zone. Among the seven identical prototypes tested, only two resulted in a crack initiation at this specific location (see Fig. 4.1, right). This also means that if only the first five prototypes had been tested, the critical zone would not have been labeled. It is then likely that several critical zones remain unlabeled. This clearly illustrates the PU learning label noise affecting the observations. It is important to note that this label noise is completely asymmetric: we only have false negatives (unlabeled positive instances) but no false positive (labeled negative instance). Moreover, Fig. 4.1 illustrates the randomness of crack initiation: although tests on prototypes 4 and 6 have a similar severity, only the latter resulted in crack initiation.

The testing conditions t can influence the propensity in several ways. A higher severity can accelerate the initiation of a crack in a critical zone. Furthermore, increasing the duration of the test will leave more time for a crack to initiate and propagate enough to be observable. Hence, the severity and the number of cycles both have an influence on the propensity. Usually, we rely on a single variable to represent the testing conditions: the equivalent severity that depends on the initial severity of the test and the total number of cycles (cf. Section 1.3). Fig. 4.1 already illustrates the effect of equivalent severity on propensity as we clearly see that the critical zone broke for two of the most severe tests. We can confirm this statement looking at the severity for every known critical zone of the database, *i.e.* those that broke at least for one test among the repetitions (cf. Fig. 4.2, left). Even if the two histograms seem close, a rough estimate of propensity for each bin (Fig. 4.2, right), asserts the increasing trend of propensity when the equivalent severity increases.

The propensity also depends on the covariates $\tilde{x}$ (or at least some of them). Indeed, a higher local stress and lower material resistance will accelerate the crack initiation in a critical zone.

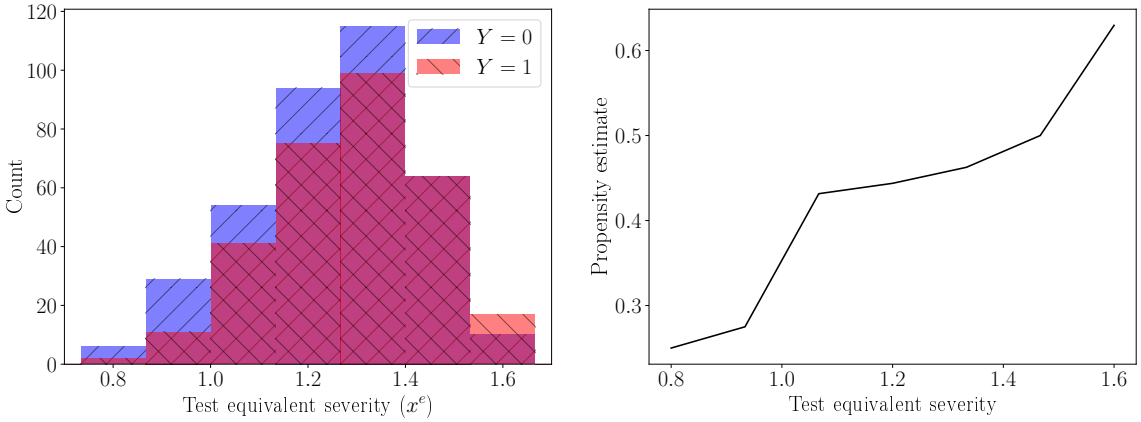

Figure 4.2: Influence of test severity on propensity. On the left, blue (red) histogram represents the empirical distribution of test severity for unbroken (broken) critical zones. Each bin leads to an estimate of propensity (as the frequency of crack initiations among the total number of observations in the bin) represented on the right.

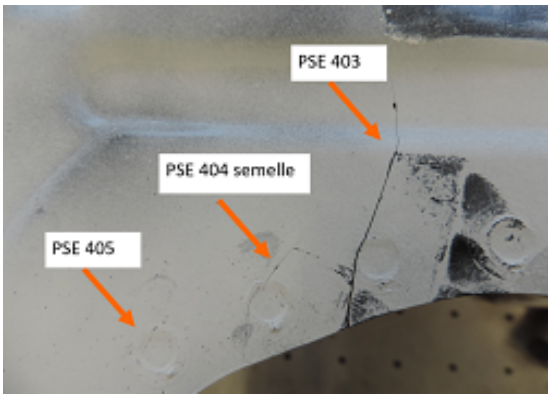

Figure 4.3: Multiple cracks detected in close locations during a test on a cradle part under longitudinal solicitations. The crack labeled "PSE 405" initiated after the two others during testing.

It also depends on the size of the crack and more generally on its observability. A crack in a hidden location on the mechanical part is less likely to be detected. Likewise, the size of the crack and the effort to detect crack initiations on the part have an influence on the label noise. In some testing experiments, penetrant inspection is used to help detect crack initiations (cf. Section 1.3). This makes the detection of cracks easier and thus increases the propensity.

Finally, let us recall that multiple cracks can initiate on a same part during testing. Sometimes, different cracks initiate on close locations (*e.g.* Fig. 4.3). Hence, there can be a dependence effect facilitating the initiation of cracks around an already broken zone or making it harder. Unfortunately, these parameters are not easily accessible and thus cannot be properly accounted for.

Hence we will stick to a propensity only depending on the available information, *i.e.* the covariates $x = (\tilde{x}, t)$.

Since the propensity represents a probability of crack initiation for a critical zone (a zone with finite lifetime), S-N models can provide useful ways of modeling it.

Let us first consider a one-dimensional setting where t is a scalar representing the total number of cycles and where the tests are performed at constant severity. A regression model on the lifetime N of an individual given the local stresses $\tilde{x}$ can yield to a model on the propensity

because the probability for a crack to be observed is, by definition, given by the cumulative distribution function of the lifetime N:

$$\begin{aligned} e(x) &= e(\tilde{x}, t) \\ &= \mathbb{P}\left(N \leq t \,|\, \tilde{X} = \tilde{x}\right) . \end{aligned}$$

In order to reduce the test duration, acceleration strategies are used. We recall that it consists in increasing gradually the severity during the test to accelerate potential crack initiations. For further details on the test protocol, refer to Section 1.3. The test conditions are described by two features: $t = (f, n)$ where f is the initial severity of the fatigue test with Locati method and n is the duration of the test (number of cycles). We compute $f_{eq}(t)$ that represents the equivalent severity at $n_0 = 10^6$. Section 1.3 explains how this equivalent severity is calculated. A crack initiation under testing conditions t is equivalent to a crack initiation under a constant-severity $f_{eq}(t)$ test before n_0 cycles. Then, we can also rely on a $S-N$ model to define the propensity:

$$e(x) = \mathbb{P}\left(N \leq n_0 \,|\, \tilde{X} = \tilde{x}_{eq}\right) , \tag{4.5}$$

where $\tilde{x}_{eq}$ denotes the local stresses under loading severity $f_{eq}(t)$. Hence $\tilde{x}_{eq} \in \mathbb{R}^{d_1}$.

4.1.3 Conclusion

We have seen that the estimation of a fatigue criterion is a PU learning task. In order to estimate the criterion η depending on covariates $\tilde{x}$ describing the local stresses at the nominal severity, we have to account for the test conditions t that may increase the propensity $e(\tilde{x}, t)$, *i.e.* the probability of observing a crack initiation during testing. Since the propensity represents a probability of crack initiation for critical zones, we can resort to classical S-N models to model it.

4.2 - PU learning: methods and models

We showed that estimating a fatigue criterion from simulation and rig test results is a PU classification task. Besides, the propensity depends on the covariates: SCAR assumption does not hold. In Chapter 3, we approached this question from the theoretical point of view and gave theoritical risk bounds under the general SAR assumption, extending the particular SCAR case. We now need a practical method to estimate the classifier. In this section, we first analyze which PU learning methods are suited for fatigue applications. Then, we define a parametric PU learning model by specifying the parametric models both on the classifier and the propensity.

4.2.1 Methods

In Section 3.3, we presented the main categories of methodologies to address PU learning tasks. Some methods like *Semi-Supervised Novelty Detection* and bagging strategies strongly rely on the SCAR assumption. Thus, they are not suited for fatigue applications because this assumption does not hold. Two-step methods do not explicitly assume the SCAR assumption, but they rely on heuristics to identify reliable negative instances. They are thus difficult to apply in our context. Finally, methods based on the minimization of an unbiased empirical risk can handle the SAR assumption setting. However, they require the knowledge of the propensity scores for labeled observations which are unavailable in practice. In order to overcome this difficulty, the methodology SAR-EM introduced by Bekker and Davis (2018b) consists in jointly estimating the classifier and the propensity. This approach is well suited for the fatigue application since

we already know how to model the propensity (cf. Subsection 4.1.2). In the next Subsection, we provide further details on the models chosen for both the classifier and the propensity.

In numerical experiments and application (Sections 4.5 and 4.6), this PU methodology will be compared to a non-standard approach to estimate the classifier. The non-standard approach is in fact the method used in Section 2.5: a classifier is estimated ignoring the PU learning label noise. Hence Y is considered as the target to predict.

4.2.2 Models

Our objective is to estimate the classification rule $\eta(x) = \mathbb{P}(Z = 1 \,|\, X = x)$. For that purpose, we need to provide a model on η. However, since we only observe $(x_i, y_i)_{1 \leq i \leq n}$, we also provide a model on $e(x) = \mathbb{P}(Y = 1 \,|\, Z = 1, X = x)$. We choose parametric models for both. Hence, from now on, a PU learning model will consist in a couple of parametric models described by parameters (θ, ϕ) where θ characterizes the classification rule and ϕ the propensity. The conditional distribution of Y given $X = x$ is denoted $\mathbb{P}_{\theta,\phi}$:

$$\mathbb{P}_{\theta,\phi}(Y = 1 \,|\, X = x) = \eta_\theta(x) \times e_\phi(x) \ .$$

Besides, we recall that the classification rule only depends on a subset of variables $\tilde{x}$ whereas the propensity may depend on all the features $x = (\tilde{x}, t)$ so that:

$$\mathbb{P}_{\theta,\phi}(Y = 1 \,|\, X = (\tilde{x}, t)) = \eta_\theta(\tilde{x}) \times e_\phi(\tilde{x}, t) \ .$$

We now provide explicit parametric models for the classification rule and the propensity.

a. Classification models

There are two ways of modeling the classifier $\eta(\tilde{x})$. A first solution is to directly provide a model on the conditional probability of Z given $\tilde{X}$. A second solution is to model the conditional probabilities of $\tilde{X}$ given $Z = 1$ and $Z = 0$, and then use Bayes theorem to retrieve the probability distribution of Z given $\tilde{X}$.

Linear logistic regression: In the first case, we use a linear logistic regression to model the class probability $\eta_\theta(\tilde{x})$:

$$\eta_\theta(\tilde{x}) = \frac{1}{1 + e^{-\alpha_0 - \alpha^T \tilde{x}}} \tag{4.6}$$

where $\theta = (\alpha_0, \alpha) \in \mathbb{R} \times \mathbb{R}^{d_1}$.

Linear Discriminant Analysis In the second case, we resort to a Linear Discriminant Analysis model. Hence, we assume that the conditional distributions of $\tilde{X}$ given $Z = 1$ and $Z = 0$ are Gaussian with parameters (μ_1, Σ) and (μ_0, Σ). The mean vectors μ_1 and μ_0 are both in $\mathbb{R}^{d_1}$. The shared covariance Σ is a symmetric positive definite matrix in $\mathbb{R}^{d_1 \times d_1}$. The class prior is $\pi = \mathbb{P}(Z = 1)$. The posterior class probability can be obtained using Bayes rule:

$$\eta_\theta(\tilde{x}) = \frac{\pi \, f_{\mu_1,\Sigma}(\tilde{x})}{\pi \, f_{\mu_1,\Sigma}(\tilde{x}) + (1 - \pi) f_{\mu_0,\Sigma}(\tilde{x})} \tag{4.7}$$

where $\theta = (\pi, \mu_0, \mu_1, \Sigma)$ is the set of parameters and $f_{\mu_1,\Sigma}$ ($f_{\mu_0,\Sigma}$) is the density of $\tilde{X}|Z = 1$ ($\tilde{X}|Z = 0$).

b. Propensity models

In Subsection 4.1.2 we gave a list of factors influencing the propensity and showed that it could be modeled using S-N fatigue models. Given a feature vector $(\tilde{x}, t)$, we can calculate the test equivalent severity $f_{eq}(t)$ and then the local stresses $\tilde{x}_{eq} \in \mathbb{R}^{d_1}$ at the equivalent severity: $\tilde{x}_{eq} \in \mathbb{R}^{d_1'}$ represents a transformed feature vector depending on $\tilde{x}$ and t. Then, the probability of crack initiation is modeled as:

$$e_\phi(x) = F\left(\log\left(\left[\beta^T \tilde{x}_{eq}\right]_+\right)\right) \tag{4.8}$$

where F belongs to a parametric family of cumulative distribution functions and $\beta \in \mathbb{R}^{d_1'}$.

Different choices are possible for the parametric family F belongs to: we will consider the *log-normal fatigue model* and the *Weibull fatigue model*.

In the log-normal fatigue model, the fatigue lifetime follows a log-normal distribution meaning that the logarithm of fatigue lifetime follows a normal distribution, hence F belongs to the family $(F_\sigma)_{\sigma \in \mathbb{R}_+^*}$ where F_σ denotes the cumulative distribution function of a centered normal distribution with variance σ^2. The propensity e_ϕ is then represented by the set of parameters $\phi = (\beta, \sigma)$.

In the Weibull fatigue model, the fatigue lifetime follows a Weibull distribution, hence its logarithm follows a Gumbel distribution: F belongs to $\left\{F_{a,b},\ a, b \in \mathbb{R} \times \mathbb{R}_+^*\right\}$ where $F_{a,b}$ denotes the cumulative distribution function of a Gumbel distribution with parameters (a, b):

$$F_{a,b}(u) = 1 - e^{-\exp\left(\frac{u-a}{b}\right)}\ .$$

The propensity e_ϕ is then represented by the set of parameters $\phi = (\beta, a, b)$.

The two above models are derived from classic S-N models in the literature and are thus specific to fatigue applications (cf. Castillo and Fernández-Canteli, 2009). In addition, we also consider a general statistical model consisting in a logistic regression based on the transformed feature vector $\tilde{x}_{eq}$ (*Logistic propensity*):

$$e_\phi(x) = \frac{1}{1 + e^{-\beta_0 - \beta^T \tilde{x}_{eq}}}\ .$$

The set of parameters is $\phi = (\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^{d_1'}$.

c. Summary

A PU learning model is the combination of a classification model on $\eta(\cdot)$ and a propensity model on $e(\cdot)$. Tables 4.1 and 4.2 summarize the list of parametric models considered for the classifier and the propensity. From now on, a PU learning model will be represented by the parameter (θ, ϕ) where θ characterizes the classifier and ϕ the propensity. No matter the propensity model, we will talk about:

- *PU Logistic Regression* (PU-LR) when the classifier is a linear logistic regression model;
- *PU Discriminant Analysis* (PU-DA) when the classifier is a Linear Discriminant Analysis.

Table 4.1: List of parametric models used for the classifier η

Model name	**Formula**	**Parameters**
Linear Logistic Regression	$\eta_\theta(\tilde{x}) = \frac{1}{1+e^{-\alpha_0-\alpha^T\tilde{x}}}$	$\theta = (\alpha_0, \alpha) \in \mathbb{R} \times \mathbb{R}^{d_1}$
Linear Discriminant Analysis	$\eta_\theta(\tilde{x}) = \frac{\pi f_{\mu_1,\Sigma}(x)}{\pi f_{\mu_1,\Sigma}(x)+(1-\pi)f_{\mu_0,\Sigma}(x)}$	$\theta = (\pi, \mu_0, \mu_1, \Sigma)$ $\in [0,1] \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1 \times d_1}$

Table 4.2: List of parametric models used for the propensity e

Model name	**Formula**	**Parameters**
Normal fatigue model	$e_\phi(x) = F_\sigma\left(\log\left(\left[\beta^T\tilde{x}_{eq}\right]_+\right)\right)$ $F_\sigma(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\,\pi\,\sigma^2}} e^{-\frac{s^2}{2\,\sigma^2}}\,\mathrm{d}s$	$\phi = (\beta, \sigma) \in \mathbb{R}^{d_1'} \times \mathbb{R}$
Weibull fatigue model	$e_\phi(x) = F_{a,b}\left(\log\left(\left[\beta^T\tilde{x}_{eq}\right]_+\right)\right)$ $F_{a,b}(u) = 1 - e^{-a\,\exp(b\,u)}$	$\phi = (\beta, a, b) \in \mathbb{R}^{d_1'} \times \mathbb{R}_+^* \times \mathbb{R}_+^*$
Logistic propensity	$e_\phi(x) = \frac{1}{1+e^{-\beta_0-\beta^T\tilde{x}_{eq}}}$	$\phi = (\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^{d_1'}$

4.3 - Identifiability

In Subsection 4.2.2, we defined a parametric PU learning model consisting in a model on the classifier η_θ and another on the propensity $e(\phi)$. Hence, the estimation is performed on the set of parameters (θ, ϕ). Before studying the estimation of the parameters (θ, ϕ), we need to ensure that the PU learning model is identifiable. A PU model $(\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi\in\Phi}$ is identifiable if and only if the parameters (θ, ϕ) uniquely characterize the distribution $\mathbb{P}_{\theta,\phi}$:

$$\mathbb{P}_{\theta,\phi} = \mathbb{P}_{\theta',\phi'} \implies (\theta, \phi) = (\theta', \phi') \ .$$

We remark that, if we drop the parametric assumptions on η and e, the decomposition

$$\mathbb{P}(Y = 1 \,|\, X = x) = \eta(x) \times e(x) \ .$$

is not unique. In general, the classifier and the propensity are clearly not identifiable. In this section, we provide sufficient conditions ensuring the identifiability of the PU learning parametric model. First, we assume that the parametric model on the propensity is identifiable:

$$(E) \quad \left(\forall \tilde{x}, t,\ e_\phi(\tilde{x}, t) = e_{\phi'}(\tilde{x}, t)\right) \implies \left(\phi = \phi'\right) \ .$$

This assumption is necessary for PU-LR and PU-DA settings. Indeed, if the model on the propensity is not identifiable, then the PU cannot be identifiable. The propensity models defined in Subsection 4.2.2 satisfy condition (E).

Subsections 4.3.1 and 4.3.2 provide sufficient conditions for the identifiability in the PU logistic regression setting and the PU discriminant analysis setting.

4.3.1 Identifiability in PU-LR setting

In this paragraph, we consider the PU-LR setting, meaning that the model on the classifier is a logistic regression. In this setting $\mathbb{P}_{\theta,\phi}$ is the conditional distribution of Y given X:

$$\mathbb{P}_{\theta,\phi}(Y = 1 \,|\, X = x) = \eta_\theta(\tilde{x}) \times e_\phi(\tilde{x}, t) \ .$$

Proposition 4.3.1 provide sufficient conditions for the PU learning model to be identifiable.

Proposition 4.3.1: Identifiability for PU-LR

Consider a PU learning model $(\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi\in\Phi}$ where the classification model is a logistic regression. Assume that the propensity model satisfies (E) and that the logistic regression classification model is identifiable. The model $(\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi\in\Phi}$ is identifiable if:

(D_1) $\forall \phi \in \Phi,\ \forall \tilde{x},\ \sup_{t\in\mathbb{R}^{d_2}} e_\phi(\tilde{x}, t) = 1$.

Before proving the proposition, let us make a few comments on the conditions.

Remarks:

1. The logistic regression model is identifiable, at least if the covariates are not linearly dependent. Condition (D_1) is naturally adapted to fatigue applications. Indeed, coavariate vectors $\tilde{x}_{eq}$ represent local stresses and contain strictly positive values. Besides, increasing the stress results in a higher probability to break (higher propensity): hence we will assume that the regression parameter β in the propensity also contains positive values. Therefore, when the equivalent severity $f_{eq}(t)$ tends to $+\infty$, local stresses also tend to $+\infty$, and the propensity tends to 1. Thus, condition (D_1) is satisfied.
2. We will see that the proof for Proposition 4.3.1 can be extended beyond the logistic regression model. In fact, we only need the classification model $(\eta_\theta)_{\theta\in\Theta}$ to be identifiable. Hence, Proposition 4.3.1 remains true if we replace the logistic regression model by any identifiable model on the conditional distribution of Z given X.

Let us now prove Proposition 4.3.1.

Proof. Suppose there exists (θ, ϕ) and (θ', ϕ') such that $\mathbb{P}_{\theta,\phi} = \mathbb{P}_{\theta',\phi'}$. We then have:

$$\forall\, \tilde{x}, t,\ \eta_\theta(\tilde{x})\, e_\phi(\tilde{x}, t) = \eta_{\theta'}(\tilde{x})\, e_{\phi'}(\tilde{x}, t) \ . \tag{4.9}$$

Taking the supremum over t in Eq. 4.9 yields:

$$\forall \tilde{x},\ \eta_\theta(\tilde{x}) = \eta_{\theta'}(\tilde{x})$$

because the model satisfies condition (D_1). Using the identifiability of the classification model, $\theta = \theta'$. Now, coming back to Eq. 4.9, we have:

$$\forall \tilde{x}, t,\ e_\phi(\tilde{x}, t) = e_{\phi'}(\tilde{x}, t)$$

which implies $\phi = \phi'$ because the model on the propensity is identifiable (E). □

4.3.2 Identifiability in PU-DA setting

We now consider the PU-DA setting, the classification model is thus a Linear Discriminant Analysis. In this setting, $\mathbb{P}_{\theta,\phi}$ is the distribution of (X, Y):

$$\mathbb{P}_{\theta,\phi}(Y = y, X = x) = [\pi f_{\mu_1,\Sigma}(\tilde{x}) e_\phi(\tilde{x}, t)]^y [\pi f_{\mu_1,\Sigma}(\tilde{x})(1 - e_\phi(\tilde{x}, t)) + (1 - \pi) f_{\mu_0,\Sigma}(\tilde{x})]^{1-y} . \tag{4.10}$$

Proposition 4.3.2 provides sufficient conditions for identifiability under this setting.

Proposition 4.3.2: Identifiability for PU-DA

Consider a PU learning model $(\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi\in\Phi}$ where the classification model is a Linear Discriminant Analysis. Assume that the propensity model satisfies (E). The model $(\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi\in\Phi}$ is identifiable if:

(D_2) $\forall\,\theta, \forall\,\phi, \exists\, x,$ such that $\pi f_{\mu_1,\Sigma}(\tilde{x}) e_\phi(\tilde{x}, t) > (1 - \pi) f_{\mu_0,\Sigma}(\tilde{x})$.

Let us make a few comments on this result.

Remarks

1. Condition (D_2) is a compatibility condition that requires the density of labeled positive instances to be higher than the density of negative instances for some $\tilde{x} \in \mathbb{R}^{d_1}$. If condition (D_2) is not fulfilled, there are at most two sets of parameters representing the distribution of (X, Y). Below, we provide such an example.

2. Condition (D_1) implies (D_2). Indeed, since $f_{\mu_0,\Sigma}$ and $f_{\mu_1,\Sigma}$ are densities of Gaussian distributions with same covariance matrix, there exists $\tilde{x}$ such that $\pi\, f_{\mu_1,\Sigma}(\tilde{x}) > (1 - \pi)\, f_{\mu_1,\Sigma}(\tilde{x})$. Then condition (D_1) allows us to choose t such that condition (D_2) is fulfilled with $x = (\tilde{x}, t)$. This means that conditions (D_2) is less restrictive than (D_1). However, we can note that the assumptions on the classifier are stronger in the case of Linear Discriminant Analysis since the conditional distributions of $\tilde{X}$ given $Z = 1$ and $Z = 0$ are assumed to be Gaussian. This is not the case for logistic regression.

3. As we will see, the proof of Proposition 4.3.2 can be extended to other classification models similar to Linear Discriminant Analysis. As far as the mixture model on the marginal distribution of $\tilde{X}$ is identifiable, the same proof remains valid. For instance, a Quadratic Discriminant Analysis would also lead to an identifiable PU model.

Example of non-identifiable propensity when (D_2) is not fulfilled: In this example, we consider a one-dimensional setting ($d_1 = 1$) and drop the dependency on t so that both η and e both only depend on $x = \tilde{x} \in \mathbb{R}$. The set of parameters $\theta = (\pi, \mu_0, \mu_1, \Sigma)$ for the Linear Discriminant Analysis model η_θ is chosen as follows:

$$\pi = 0.4,\ \mu_0 = -1.5,\ \mu_1 = 1.5,\ \Sigma = 2 .$$

Let us denote $f(\cdot, Y = 1)$ and $f(\cdot, Y = 0)$ the densities of labeled and unlabeled instances. In this example, condition (D_2) is not satisfied. And the model is not identifiable, indeed we can exhibit two sets of parameters $\theta = (\pi, \mu_0, \mu_1, \Sigma)$ and $\theta' = (1 - \pi, \mu_1, \mu_0, \Sigma)$ along with propensity functions e and e' leading to the same distribution on $(\tilde{X}, Y)$ (cf. Fig. 4.4). However, in this example, if condition (D_2) were fulfilled, only one of the two solutions would have been admissible and the model would be identifiable. This shows that condition (D_2) is necessary.

Let us now prove Proposition 4.3.2.

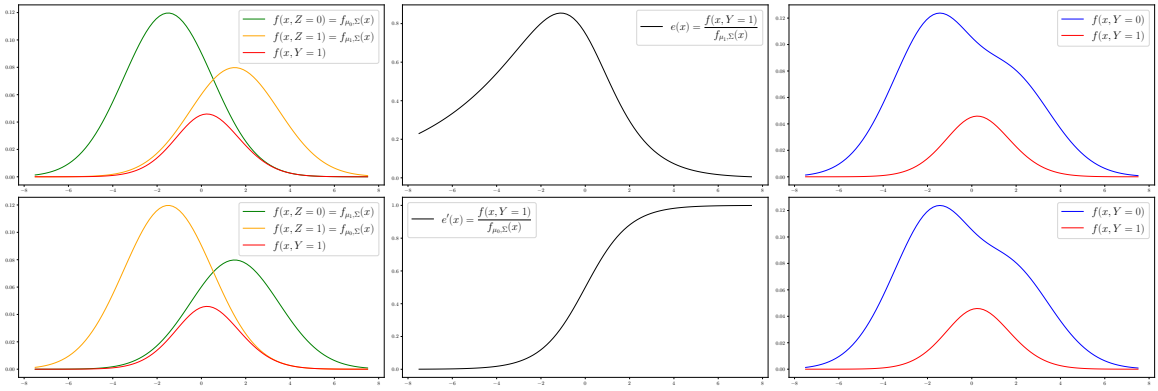

Figure 4.4: Non-identifiable example in PU-DA when condition D_2 is not fulfilled: the figure on the left represent the densities of positive and negative instances along with the density of labeled instances. Note that the densities $f(x, Z = 0)$ and $f(x, Z = 1)$ were inverted. In the middle the corresponding propensity functions are represented. These two situations lead to the same distributions of labeled and unlabeled instances (figures on the right). Hence the distribution on (X, Y) can arise either from parameter θ with propensity e (upper-left figures) or from parameter θ' with propensity e' (lower-left figures).

Proof. Assume that the model satisfies (D_2) and that there exists (θ, ϕ) and (θ', ϕ') such that $\mathbb{P}_{\theta,\phi} = \mathbb{P}_{\theta',\phi'}$. Recall that in this setting, $\mathbb{P}_{\theta,\phi}$ and $\mathbb{P}_{\theta',\phi'}$ represent probability distributions on (X, Y). The equality $\mathbb{P}_{\theta,\phi} = \mathbb{P}_{\theta',\phi'}$ implies that the marginal distributions over X are equal. Hence, after integrating Equation 4.10 over y, we have:

$$\forall \tilde{x},\ \pi\, f_{\mu_1,\Sigma}(\tilde{x}) + (1-\pi)\, f_{\mu_0,\Sigma}(\tilde{x}) = \pi'\, f_{\mu_1',\Sigma'}(\tilde{x}) + (1-\pi')\, f_{\mu_0',\Sigma'}(\tilde{x})\ .$$

Since the mixture of Gaussian densities is identifiable up to a permutation of the components, we necessarily have either

$$(\pi, \mu_0, \mu_1, \Sigma) = \left(\pi', \mu_0', \mu_1', \Sigma'\right), \tag{4.11}$$

or

$$(\pi, \mu_0, \mu_1, \Sigma) = \left(1-\pi', \mu_1', \mu_0', \Sigma'\right). \tag{4.12}$$

Using that $\mathbb{P}_{\theta,\phi}(Y = 1, X = x) = \mathbb{P}_{\theta',\phi'}(Y = 1, X = x)$ for all x yields either

$$\forall \tilde{x}, t,\ e_{\phi'}(\tilde{x}, t) = e_{\phi}(\tilde{x}, t) \text{ if Eq. 4.11 is satisfied} \tag{4.13}$$

or

$$\forall \tilde{x}, t,\ e_{\phi'}(\tilde{x}, t) = \frac{\pi f_{\mu_1,\Sigma}(\tilde{x}) e_{\phi}(\tilde{x}, t)}{(1-\pi) f_{\mu_0,\Sigma}(\tilde{x})} \text{ if Eq. 4.12 is satisfied.} \tag{4.14}$$

However, if Eq. 4.14 were true, condition (D_2) would imply the existence of x such that $e_{\phi'}(x) > 1$ which is impossible. Hence, only Eq. 4.11 and 4.13 are valid. Together with the fact that the model on the propensity e is identifiable, we have:

$$(\theta, \phi) = (\theta', \phi')\ .$$

□

4.3.3 Conclusion

We have provided sufficient conditions ensuring the identifiability of the parametric PU learning model under PU Logistic Regression and PU Discriminant Analysis settings. The identifiability of the propensity model (E) is always fulfilled for the models defined in Section 4.2. Conditions (D_1) and (D_2) are also fulfilled in our fatigue application.

4.4 - Joint estimation of classification rule and propensity

We have defined a parametric model $(\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi\in\Phi}$, and conditions (Propositions 4.3.1 and 4.3.2) to satisfy its identifiability. In this section, we deal with the estimation of the parameters on training observations through maximum likelihood. Since we are in the presence of missing data, we will use EM algorithm to maximize the likelihood.

In Subsection 4.4.1, we explain why the EM algorithm is well suited for optimizing the log-likelihood for PU learning models. Then, in Subsection 4.4.2, we recall the general principle of the EM algorithm. Subsections 4.4.3 and 4.4.4 motivate the use of the EM algorithm in PU-LR and PU-DA settings and provide the detailed steps of the EM algorithm for maximum likelihood estimation.

In PU-LR setting, the methodology is identical to the one used by Bekker and Davis (2018b). In PU-DA setting though, we no longer perform the estimation conditionally to the covariates which leads to a different objective function and thus a different algorithm.

4.4.1 Maximum likelihood with the EM algorithm

We consider a set of n independent observations $(\tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Y}) = (\tilde{X}_i, T_i, Y_i)_{1\leq i\leq n}$. Let us denote $\ell(\theta, \phi \,|\, \tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Y})$ the log-likelihood. The objective is to find the parameters $(\widehat{\theta}, \widehat{\phi})$ maximizing the log-likelihood:

$$\left(\widehat{\theta}, \widehat{\phi}\right) = \underset{\theta,\phi}{\text{Argmax}}\ \ell\Big(\theta, \phi \,|\, \tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Y}\Big)\ . \tag{4.15}$$

The maximization of Eq. 4.15 is difficult to solve directly. Besides we have latent variables $(Z_i)_{1\leq i\leq n}$ and we will show that under both PU-LR and PU-DA settings, the maximization of the complete log-likelihood $\ell(\theta, \phi \,|\, \tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{Y})$ results in a much simpler optimization problem. Hence, Expectation Maximization (EM) algorithm is well suited for maximizing the log-likelihood.

4.4.2 EM algorithm

The EM algorithm, introduced by Dempster et al. (1977), enables the calculation of maximum of likelihood estimates for models with latent variables. The algorithm consists in iterating through an expectation step (E step) and a maximization step (M step) (cf. Algorithm 1). The likelihood increases at each iteration and the algorithm stops when it reaches a local maximum. As the likelihood is not necessarily convex, there can be multiple local maxima. Therefore, depending on the initialization, the algorithm does not necessarily converge to the global maximum. Initialization of the EM algorithm for PU learning will be discussed in Subsection 4.4.5.

Algorithm 1 General EM algorithm for PU learning

Initialization: start with initial parameters $\left(\theta^{(0)}, \phi^{(0)}\right)$

Iterate until convergence:

E step Given the parameters $\left(\theta^{(c)}, \phi^{(c)}\right)$ obtained at step c, compute the conditional expectation of the complete log-likelihood:

$$Q^{(c)}\left(\theta, \phi\right) = \mathbb{E}\left[\ell\big(\theta, \phi \,|\tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{Y}\big) \,\Big|\, \tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Y}, \widehat{\theta}^{(c)}, \widehat{\phi}^{(c)}\right]$$

M step Maximize the conditional expectation over (θ, ϕ):

$$\left(\widehat{\theta}^{(c+1)}, \widehat{\phi}^{(c+1)}\right) = \underset{\theta,\phi}{\text{Argmax}}\ Q^{(c)}\left(\theta, \phi\right)$$

4.4.3 Estimation for PU logistic regression

In the PU-LR setting, the estimation is done conditionally to the covariates. The log-likelihood is:

$$\ell\Big(\theta, \phi \,|\, \tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Y}\Big) = \sum_{i=1}^{n} \Big[Y_i \log\Big(\eta_\theta(\tilde{X}_i)\, e_\phi(\tilde{X}_i, T_i)\Big) + (1 - Y_i) \log\Big(1 - \eta_\theta(\tilde{X}_i)\, e_\phi(\tilde{X}_i, T_i)\Big)\Big] \ . \tag{4.16}$$

Even if the form of the log-likelihood looks similar to a logistic regression, it cannot be written as a generalized linear model. Besides, we cannot separate the effects of θ and ϕ. The complete log-likelihood, however, separates the effects of both models and can be optimized efficiently. It is given by:

$$\begin{aligned}\ell\left(\theta, \phi \,|\, \tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{Y}\right) = &\sum_{i=1}^{n} \Big[Z_i \log\Big(\eta_\theta(\tilde{X}_i)\Big) + (1 - Z_i) \log\Big(1 - \eta_\theta(\tilde{X}_i)\Big)\Big] \\ &+ \sum_{i=1}^{n} Z_i \Big[Y_i \log\Big(e_\phi(\tilde{X}_i, T_i)\Big) + (1 - Y_i) \log\Big(1 - e_\phi(\tilde{X}_i, T_i)\Big)\Big]\end{aligned} \tag{4.17}$$

Hence, the EM algorithm is well suited for optimizing the log-likelihood in PU-LR setting. This methodology was used by Bekker and Davis (2018b). Let us rewrite the expectation and maximization steps in this setting, with a propensity given by one of the models of Table 4.2.

a. Expectation

Let us assume that the parameter after iteration c in the EM algorithm is $\left(\theta^{(c)}, \phi^{(c)}\right)$. We want to calculate $Q^{(c)}\left(\theta, \phi\right)$. Thanks to the linearity of the expectation, we only need to compute the posterior probabilities $\gamma_i^{(c)}$:

$$\gamma_i^{(c)} = \mathbb{E}\left[Z_i \,\middle|\, \tilde{X}_i, T_i, Y_i, \theta^{(c)}, \phi^{(c)}\right] = \mathbb{P}_{\theta^{(c)}, \phi^{(c)}}\left(Z_i = 1 \,\middle|\, \tilde{X}_i, T_i, Y_i\right) \ .$$

We recall that, in PU learning, a labeled instance $(Y_i = 1)$ is necessarily positive, hence:

$$\mathbb{P}_{\theta^{(c)}, \phi^{(c)}}\left(Z_i = 1 \,\middle|\, \tilde{X}_i, T_i, Y_i = 1\right) = 1 \ .$$

The posterior probability for unlabeled instances can be computed using Bayes theorem:

$$\mathbb{P}_{\theta^{(c)}, \phi^{(c)}}\left(Z_i = 1 \,\middle|\, \tilde{X}_i, T_i, Y_i = 0\right) = \mathbb{P}_{\theta^{(c)}, \phi^{(c)}}\left(Z_i = 1 \,|\, \tilde{X}_i, T_i, Y_i = 0\right)$$

$$= \frac{\mathbb{P}_{\theta^{(c)},\phi^{(c)}}\left(Z_i = 1, Y_i = 0 \,|\, \tilde{X}_i, T_i\right)}{\mathbb{P}_{\theta^{(c)},\phi^{(c)}}\left(Y_i = 0 \,|\, \tilde{X}_i, T_i\right)}$$
$$= \frac{\eta_{\theta^{(c)}}(\tilde{X}_i)\,\left(1 - e_{\phi^{(c)}}(\tilde{X}_i, T_i)\right)}{1 - \eta_{\theta^{(c)}}(\tilde{X}_i)\, e_{\phi^{(c)}}(\tilde{X}_i, T_i)} \,.$$

The conditional expectation of the log-likelihood is then:

$$Q^{(c)}(\theta, \phi) = \sum_{i=1}^{n} \left[\gamma_i^{(c)} \log\left(\eta_\theta(\tilde{X}_i)\right) + (1 - \gamma_i^{(c)}) \log\left(1 - \eta_\theta(\tilde{X}_i)\right)\right]$$
$$+ \sum_{i=1}^{n} \gamma_i^{(c)} \left[Y_i \log\left(e_\phi(\tilde{X}_i, T_i)\right) + (1 - Y_i) \log\left(1 - e_\phi(\tilde{X}_i, T_i)\right)\right] \,.$$

b. Maximization

The conditional expectation of the log-likelihood can be separated in two terms, one involving only θ, the other involving only ϕ. Hence, the maximization step consists in solving two maximization problems:

$$\theta^{(c+1)} \in \underset{\theta}{\text{Argmax}} \sum_{i=1}^{n} \left[\gamma_i^{(c)} \log\left(\eta_\theta(\tilde{X}_i)\right) + (1 - \gamma_i^{(c)}) \log\left(1 - \eta_\theta(\tilde{X}_i)\right)\right]$$
$$\phi^{(c+1)} \in \underset{\phi}{\text{Argmax}} \sum_{i=1}^{n} \gamma_i^{(c)} \left[Y_i \log\left(e_\phi(\tilde{X}_i, T_i)\right) + (1 - Y_i) \log\left(1 - e_\phi(\tilde{X}_i, T_i)\right)\right] \tag{4.18}$$

The first maximization problem is a weighted logistic regression where each observation is considered as positive with weight $\gamma_i^{(c)}$ and negative with weight $1-\gamma_i^{(c)}$. The second is a weighted logistic regression based on the observed labels $(Y_i)_{1\leq i\leq n}$ and with a link function depending on the propensity model: log-normal for log-normal fatigue model, Gumbel for Weibull fatigue model and logistic for the logistic propensity.

4.4.4 Estimation for PU-DA

In PU-DA setting, we model the joint distribution of (X, Y). The log-likelihood is then:

$$\ell\Big(\theta, \phi \,|\, \tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Y}\Big) = \sum_{i=1}^{n} \left[Y_i \log\left(\pi\, f_{\mu_1,\Sigma}(\tilde{X}_i)\, e_\phi(\tilde{X}_i, T_i)\right)\right]$$
$$+ \sum_{i=1}^{n} \left[(1 - Y_i) \log\left(\pi\, f_{\mu_1,\Sigma}(\tilde{X}_i)(1 - e_\phi(\tilde{X}_i, T_i)) + (1 - \pi)\, f_{\mu_0,\Sigma}(\tilde{X}_i)\right)\right] \,. \tag{4.19}$$

The direct maximization of 4.19 is difficult because of the sum in the logarithm of the second term in Eq. 4.19. Besides, we have latent classes $(Z_i)_{1\leq i\leq n}$ and, as for PU-LR, the complete log-likelihood $\ell\Big(\theta, \phi \,|\, \tilde{\boldsymbol{X}}, \boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{Y}\Big)$ results in a much simpler optimization problem.

a. Expectation

The expectation step is similar to the PU-LR setting. The same formula remains valid for calculating the posterior probabilities. For Linear Discriminant Analysis, $\eta_\theta(\tilde{x})$ is obtained as:

$$\eta_\theta(\tilde{x}) = \frac{\pi f_{\mu_1,\Sigma}(\tilde{x})}{\pi f_{\mu_1,\Sigma}(\tilde{x}) + (1-\pi) f_{\mu_0,\Sigma}(\tilde{x})} \ .$$

The conditional expectation of the log-likelihood is:

$$\begin{aligned} Q^{(c)}(\theta,\phi) = &\sum_{i=1}^{n} \left[\gamma_i^{(c)} \left(\log(\pi) + \log\left(f_{\mu_1,\Sigma}(\tilde{x}_i)\right)\right) + (1-\gamma_i^{(c)}) \left(\log(1-\pi) + \log\left(f_{\mu_0,\Sigma}(\tilde{x}_i)\right)\right)\right] \\ &+ \sum_{i=1}^{n} \gamma_i^{(c)} \left[y_i \log\left(e_\phi(\tilde{x}_i, t_i)\right) + (1-y_i) \log\left(1 - e_\phi(\tilde{x}_i, t_i)\right)\right] \end{aligned} \tag{4.20}$$

b. Maximization

As for the previous model, the maximization step consist in solving two separate maximization problems. The maximization involving the propensity model is strictly identical to Eq. 4.18. The other maximization consists in performing a weighted Linear Discriminant Analysis:

$$\begin{aligned} \theta^{(c+1)} \in \underset{\theta=(\pi,\mu_0,\mu_1,\Sigma)}{\text{Argmax}} &\sum_{i=1}^{n} \left[\gamma_i^{(c)} \log\left(f_{\mu_1,\Sigma}(\tilde{x}_i)\right) + (1-\gamma_i^{(c)}) \left(\log\left(f_{\mu_0,\Sigma}(\tilde{x}_i)\right)\right)\right] \\ &+ \sum_{i=1}^{n} \left[\gamma_i^{(c)} \log(\pi) + (1-\gamma_i^{(c)}) \log(1-\pi)\right] \end{aligned}$$

which can be solved explicitly:

$$\begin{aligned} \pi^{(c+1)} &= \frac{1}{n} \sum_{i=1}^{n} \gamma_i^{(c)} \\ \mu_0^{(c+1)} &= \frac{\sum_{i=1}^{n} (1-\gamma_i^{(c)}) \tilde{X}_i}{\sum_{i=1}^{n} (1-\gamma_i^{(c)})} \\ \mu_1^{(c+1)} &= \frac{\sum_{i=1}^{n} \gamma_i^{(c)} \tilde{X}_i}{\sum_{i=1}^{n} \gamma_i^{(c)}} \\ \Sigma^{(c+1)} &= \frac{1}{n} \sum_{i=1}^{n} \left[\gamma_i^{(c)} \left(\tilde{X}_i - \mu_1^{(c+1)}\right)^T \left(\tilde{X}_i - \mu_1^{(c+1)}\right) + \left(1-\gamma_i^{(c)}\right) \left(\tilde{X}_i - \mu_0^{(c+1)}\right)^T \left(\tilde{X}_i - \mu_0^{(c+1)}\right)\right] \end{aligned}$$

In fact, the updating of θ is similar to the maximization step in EM algorithm when estimating the parameters of a Gaussian mixture model. The difference is that, here, some of the posterior probabilities are exactly equal to 1 (every label instance is positive). In this sense, PU Discriminant Analysis is an intermediate between fully supervised Linear Discriminant Analysis and unsupervised Gaussian mixture models. In the first case, the coefficients $(\gamma_i)_{1\leq i\leq n}$ would be replaced by the observed classes $(Z_i)_{1\leq i\leq n}$; in the second, each posterior probability would be in $(0,1)$. PU Discriminant analysis clearly operates in a semi-supervised setting as the posterior probabilities are 1 for labeled instances and lie in $(0,1)$ for unlabeled ones.

4.4.5 Initialization and stopping criterion

As mentioned earlier, EM algorithm does not necessarily converge to the global optimum depending on the initialization. In order to overcome this issue, we use the *Small EM* strategy (cf. Biernacki et al. (2003)). We perform multiple runs of the EM algorithm with random intializations over a few iterations. Then we keep the model that reaches the highest likelihood and continue its optimization until convergence. In practice, we stop the algorithm when the likelihood does not increase more than a predefined tolerance threshold.

4.4.6 Conclusion

Since we have missing data in the data set, the joint estimation of the PU classifier and the propensity can be performed through EM algorithm. It involves basic calculations in the expectation step and weighted classification problems in the maximization step (cf. Algorithm 2).

The maximization step can be performed independently on the classification model and on the propensity model. The algorithm is thus very flexible as we can easily change either the classification model or the propensity without having to worry about the interactions between the two.

The iterative procedure is similar to two-step methods (cf. Subsection 3.3.2). The weights computed in the expectation step represent the probability (given the current parameters) for the instances to be positive. To some extent, this is a way to identify reliable negative instances. In the maximization step, the classifier is updated based on the weights calculated in the expectation step, which is also very similar to step 2 in two-step methods. The key difference is that the identification of reliable negative instances no longer relies on heuristics, but is part of EM algorithm used to maximize the likelihood.

4.5 - Numerical experiments on simulated data

In this section, we illustrate the methodology for estimating PU learning models introduced in Section 4.4. We first rely on simulated data in order to validate the methodology and assess its performances. Multiple experiments are carried out highlighting the interest of PU learning approach compared to a non-traditional approach ignoring the PU label noise.
Subsection 4.5.1 explains how the artificial data sets are generated. In Subsection 4.5.2, we analyze the convergence of the EM algorithm and illustrate the estimation results on a few examples. Subsection 4.5.3 provides a detailed analysis of the classification performances obtained for different series of experiments: the impact of sample size, separability between the classes and misspecification are studied.

4.5.1 Simulation setting

We seek to simulate positive unlabeled data under the SAR assumption in order to test the estimation method described in Section 4.4. In this subsection we describe how the artificial data sets are generated. The simulation is done in three steps: first generate the covariates vectors and their corresponding classes $(\tilde{X}_i, Z_i)_{1\leq i\leq n}$, then simulate the observed labels $(Y_i)_{1\leq i\leq n}$ by applying a selection bias on the positive instances, finally drop the information about the classes $(Z_i)_{1\leq i\leq n}$ in order to keep only the PU data. In the two-dimensional examples considered, the classification and the propensity will each depend on one variable. This way, we can check whether or not the PU learning classifier is able to identify which variable is related to the classification and which one influences the propensity.

Algorithm 2 Detailed EM algorithm for PU learning

Initialization: start with initial parameters $\left(\theta^{(0)}, \phi^{(0)}\right)$
Iterate until convergence:

E step Given the parameters $\left(\theta^{(c)}, \phi^{(c)}\right)$ obtained at step c, compute the posterior probabilities $(\gamma_i^c)_{1\leq i\leq n}$:

$$\gamma_i^c = \begin{cases} 1 & \text{if } Y_i = 1 \\ \frac{\eta_{\theta^{(c)}}(\tilde{X}_i)\left(1-e_{\phi^{(c)}}(\tilde{X}_i,T_i)\right)}{1-\eta_{\theta^{(c)}}(\tilde{X}_i)\, e_{\phi^{(c)}}(\tilde{X}_i,T_i)} & \text{if } Y_i = 0 \end{cases}$$

M step Update parameters (θ, ϕ):

Classifier ($\boldsymbol{\theta}$)

1. PU Logistic Regression setting: solve the following weighted logistic regression problem.

$$\theta^{(c+1)} \in \underset{\theta}{\text{Argmax}} \sum_{i=1}^{n} \left[\gamma_i^{(c)} \log\left(\eta_\theta(\tilde{X}_i)\right) + (1-\gamma_i^{(c)}) \log\left(1-\eta_\theta(\tilde{X}_i)\right)\right] .$$

2. PU Discriminant Analysis setting:

$$\begin{aligned}
\pi^{(c+1)} &= \frac{1}{n}\sum_{i=1}^{n} \gamma_i^{(c)} \\
\mu_0^{(c+1)} &= \frac{\sum_{i=1}^{n}(1-\gamma_i^{(c)})\tilde{X}_i}{\sum_{i=1}^{n}(1-\gamma_i^{(c)})} \\
\mu_1^{(c+1)} &= \frac{\sum_{i=1}^{n}\gamma_i^{(c)}\tilde{X}_i}{\sum_{i=1}^{n}\gamma_i^{(c)}} \\
\Sigma^{(c+1)} &= \frac{1}{n}\sum_{i=1}^{n}\Big[\gamma_i^{(c)}\left(\tilde{X}_i-\mu_1^{(c+1)}\right)^T\left(\tilde{X}_i-\mu_1^{(c+1)}\right) \\
&\qquad + \left(1-\gamma_i^{(c)}\right)\left(\tilde{X}_i-\mu_0^{(c+1)}\right)^T\left(\tilde{X}_i-\mu_0^{(c+1)}\right)\Big]
\end{aligned}$$

Propensity ($\boldsymbol{\phi}$) Solve the following weighted logistic regression problem (link function depending on the model on e):

$$\phi^{(c+1)} = \underset{\phi}{\text{Argmax}} \sum_{i=1}^{n} \gamma_i^{(c)} \left[Y_i \log\left(e_\phi(\tilde{X}_i,T_i)\right) + (1-Y_i)\log\left(1-e_\phi(\tilde{X}_i,T_i)\right)\right] .$$

a. Generation of the covariate vectors and the classes

Without loss of generality, we consider 2-dimensional examples meaning that each covariate vector $\tilde{X}_i$ is in $\mathbb{R}^2$. Two simulation settings are then considered: a first setting where the classes are drawn according to a logistic regression model; a second where the assumptions of Linear Discriminant Analysis are satisfied.

- **PU-LR simulation setting:** this represents a setting where the classifier satisfies the assumptions of PU-LR. In this setting the vectors $(\tilde{X}_i)_{1\leq i\leq n}$ is an i.i.d. sample following a multivariate Gaussian distribution with fixed mean vector μ and covariance matrix C_1:

$$\mu = (2,2) \text{ and } C_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$$

  Then, the classes Z_i are drawn independently according to Bernoulli distribution with parameter $\eta_\theta(\tilde{X}_i)$ where η_θ depends on $\tilde{X}_i$ through a logistic regression model with parameter $\theta = (\alpha_0, \alpha) \in \mathbb{R} \times \mathbb{R}^2$:

$$\eta_\theta(x) = \frac{1}{1 + e^{-\alpha_0 - \alpha^T x}} \ .$$

  In the examples, choosing the parameters of the form $\alpha_0 = -2\,\rho$ and $\alpha = (\rho, 0)$ with a strictly positive ρ allows to control the overlap between the two classes according to ρ: a small value of ρ will lead to a poor separation while a high value will lead to a better separability.

- **PU-DA simulation setting:** it satisfies the assumptions of PU-DA setting. In this setting, we first draw an i.i.d. sample of size n of Bernoulli variables (Z_i) with parameter $p = 0.5$. Then, for each i, $\tilde{X}_i$ is generated according to a multivariate normal distribution with mean vector μ_{Z_i} and covariate vector C_2 (only the mean depends on the class). In our examples:

$$\mu_0 = (2-\nu, 1)\ ,\ \mu_1 = (2+\nu, 1) \text{ and } C_2 = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

  Similarly to ρ in PU-LR simulation setting, the parameter $\nu > 0$ controls how well the classes are separated: a small value of ν will induce a significant overlap between the classes while a higher ν will lead to a better separability.

Examples of simulations are represented in Fig. 4.5. We can note that, in both settings, the optimal classifiers are identical and only involve the first variable.

b. Generation of the labels

In order to mimic fatigue applications, we simulate a severity F_i for every instance representing the equivalent severity instance i was tested at. Recall that the severity depends only on the testing conditions and affects the probability of crack initiation on critical zones. In our experiments, the F_i are drawn independently and uniformly between f_{min} and f_{max}. We compute $\tilde{X}_{i,eq} = F_i \times \tilde{X}_i$ which represents the transformed covariate vector in the test conditions. Then each positive instance with equivalent covariates $\tilde{X}_{i,eq}$ is labeled with probability $e_\phi(\tilde{X}_{i,eq})$ where e_ϕ is a propensity functions chosen among the models of Table 4.2 with a given parameter ϕ. From this step, the data can be represented either with the covariates $\tilde{X}_i$ or with the transformed covariates $\tilde{X}_{i,eq}$.

Examples based on the simulations of Fig. 4.5 are represented in Fig. 4.6. The propensity model used for simulation is a logistic regression propensity (cf. Table 4.2).

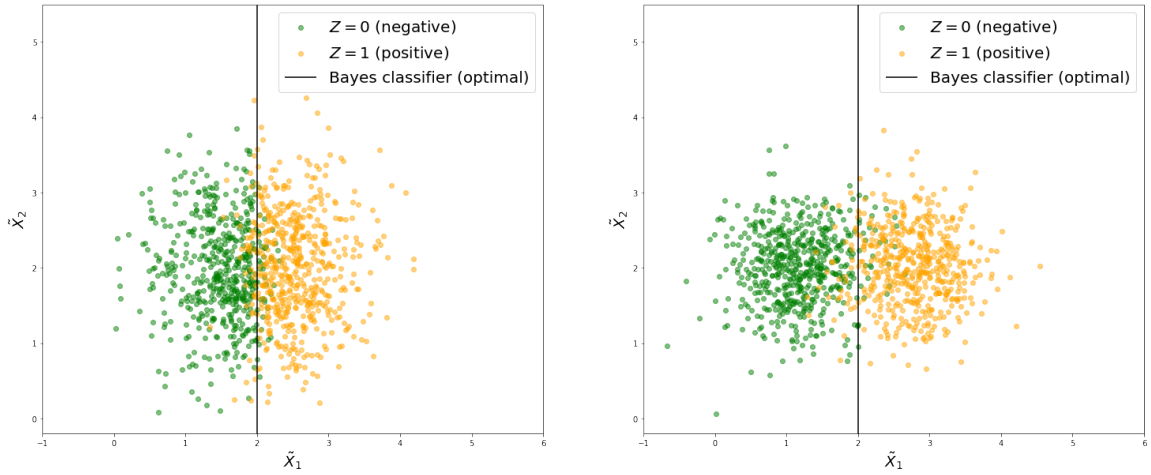

Figure 4.5: Simulated data $(\tilde{X}_i, Z_i)_{1\leq i\leq n}$ where $(\tilde{X}_i)_{1\leq i\leq n}$ are the covariate vectors and $(Z_i)_{1\leq i\leq n}$ are the classes ($n = 1000$): PU-LR simulation setting on the left ($\rho = 5$), PU-DA simulation setting on the right ($\nu = 0.8$). Note that the optimal classifier (Bayes classifier) only involves the first variable.

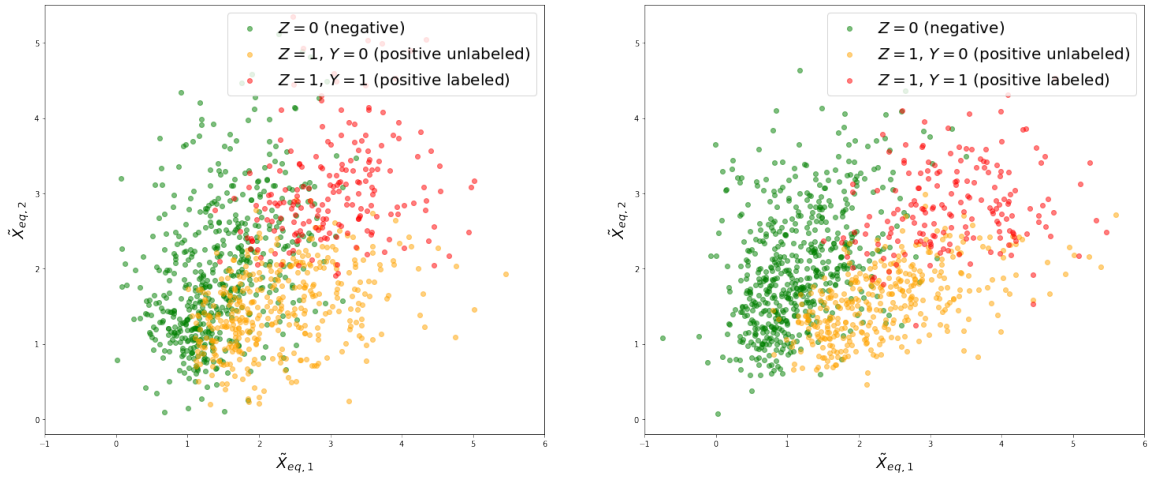

Figure 4.6: Simulated data with transformed covariate vectors $(\tilde{X}_{i,eq})_{1\leq i\leq n}$, classes $(Z_i)_{1\leq i\leq n}$ and labels $(Y_i)_{1\leq i\leq n}$ ($n = 1000$): PU-LR simulation setting on the left ($\rho = 5$), PU-DA simulation setting on the right ($\mu = 0.8$). The propensity function used in both simulations is a logistic regression propensity with parameters: $\beta_0 = -14$ and $\beta = (\beta_1, \beta_2) = (0, 6)$.

c. PU data set

Once the labels are simulated, we drop the information about the classes $(Z_i)_{1\leq i\leq n}$: only the labels are available for training. When testing the performances of the estimated classifiers, we will keep the knowledge of the classes Z in the test data sets which gives us access to the ground truth. This will allow assessing the performances of the estimated classifiers. Unfortunately, for real applications, we only have PU data even for testing: this means that we will not be able to evaluate our methods in the same way.

Examples of simulated PU data sets are represented in Fig. 4.7. They follow from Fig. 4.5 and 4.6. These examples were chosen so that the theoretical classifier only involves the first variable. At the same time, the selection bias depends on the second variable. This explains why most labeled instances appear in the top right of the diagrams.

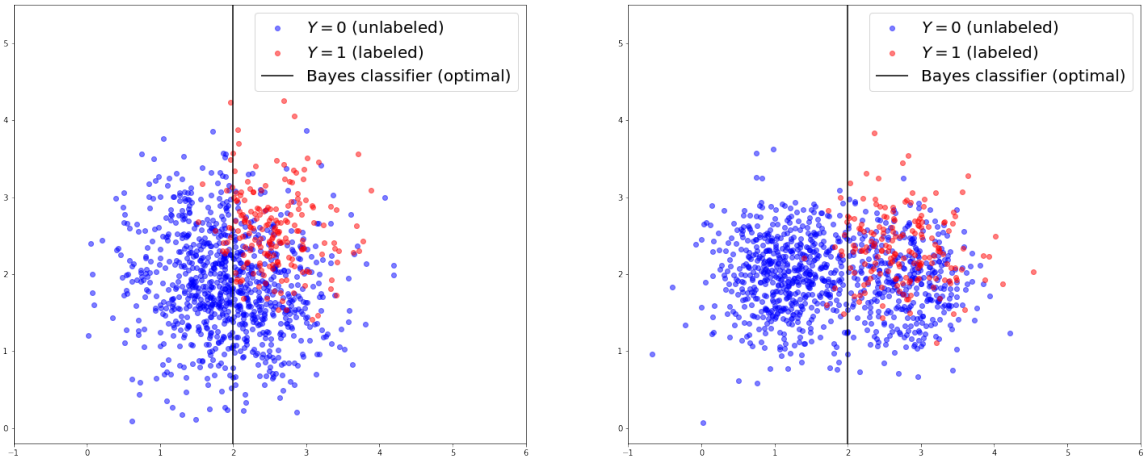

Figure 4.7: Simulated PU data sets ($n = 1000$): PU-LR simulation setting on the left ($\rho = 5$), PU-DA simulation setting on the right ($\mu = 0.8$). The propensity function used in both simulations is the same as in Fig. 4.6.

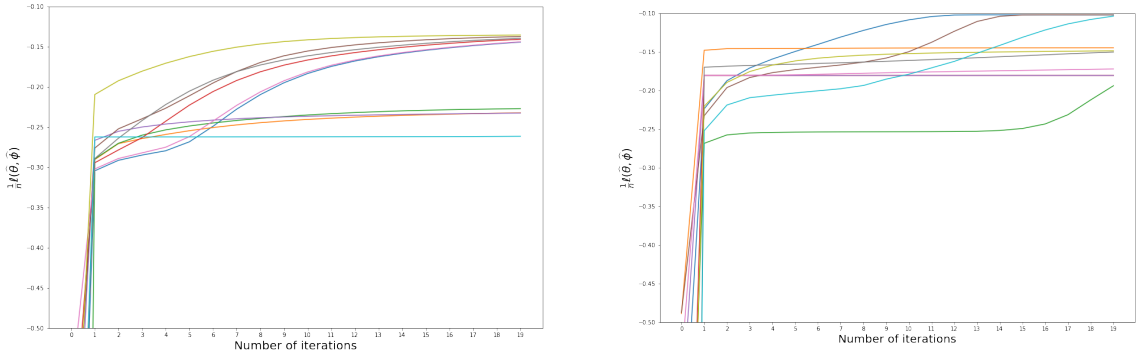

Figure 4.8: Convergence of the EM algorithm in the PU-LR setting on the left, PU-DA setting on the right. Each line represents the log-likelihood values over a run of the EM algorithm with a random initialization (Small EM). Some trajectories converge toward local maxima.

4.5.2 Parameter estimation

We now want to illustrate the methodology presented in Section 4.4 using these simulated data sets. We will check that the EM algorithm correctly converges and that classifier and the propensity are correctly estimated.

In this subsection, we consider the two simulated data sets of Subsection 4.5.1. In each case, we apply the initialization strategy described in Subsection 4.4.5. We perform multiple runs of the EM algorithm over 20 iterations. For some runs, the log-likelihood remains stuck in a local maxima (cf. Fig. 4.8). For others, the log-likelihood seem to converge to the same value which appears to be the global maximum. We therefore keep the run that reaches the highest likelihood after the 20 iterations and optimize it until convergence. We remark that the majority of trajectories have already converged after 20 iterations, which shows that the convergence of EM algorithm is quite fast. This illustrates the interest of the *Small EM* initialization strategy of the EM algorithm for PU learning.

We then check that the parameters of the PU learning models are correctly estimated. For this purpose, we study the empirical distribution of the maximum likelihood estimate $(\widehat{\theta}, \widehat{\phi})$ over $B = 200$ replicated data sets and verify that the estimation is coherent with the theoretical values (cf. Fig. 4.9 and 4.10). The estimation is satisfactory as the histograms are centered around the

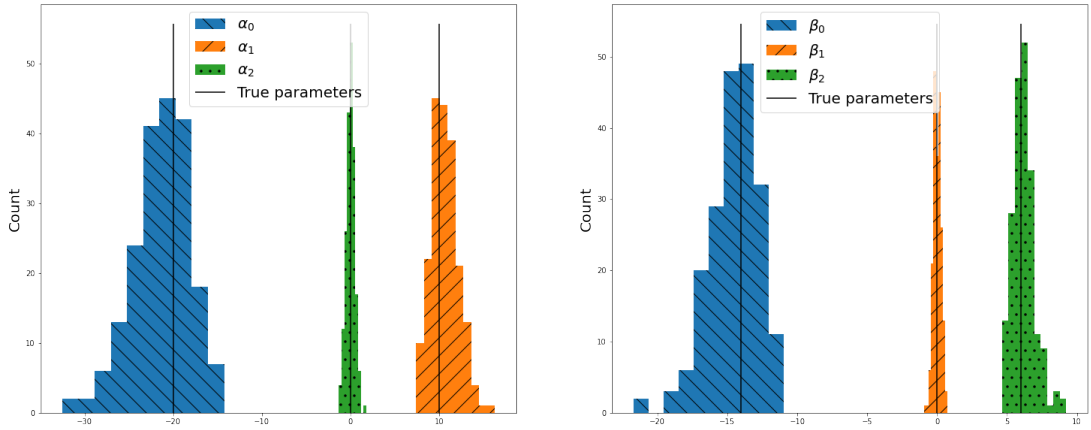

Figure 4.9: Empirical distribution of maximum likelihood estimate in PU-LR setting ($B = 200$ experiments on data sets of size $n = 1000$): on the left, the parameters of the classifier $\widehat{\theta} = (\widehat{\alpha_0}, \widehat{\alpha})$; on the right, the parameters of the logistic propensity $\widehat{\phi} = (\widehat{\beta}_0, \widehat{\beta})$.

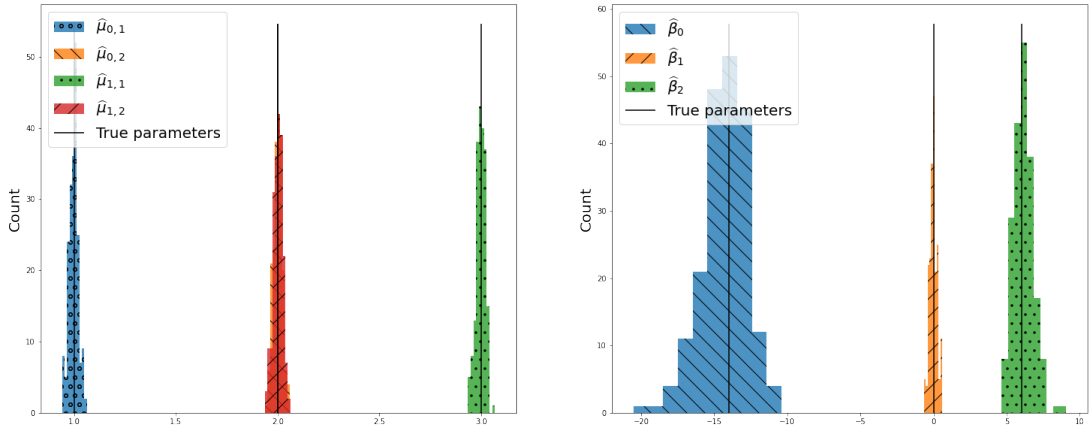

Figure 4.10: Empirical distribution of maximum likelihood estimate in PU-DA setting ($B = 200$ experiments on data sets of size $n = 1000$): on the left, some of the parameters of the classifier $\widehat{\mu}_0 = (\widehat{\mu}_{0,1}, \widehat{\mu}_{0,2})$ and $\widehat{\mu}_1 = (\widehat{\mu}_{1,1}, \widehat{\mu}_{1,2})$; on the right, the parameters of the logistic propensity $\widehat{\phi} = (\widehat{\beta}_0, \widehat{\beta})$. The histograms on $\widehat{\mu}_{0,2}$ and $\widehat{\mu}_{1,2}$ are almost confounded, which is logical because the theoretical parameters are identical.

theoretical values. This therefore illustrates the methodology of Section 4.4.

Finally, we assess the performances of the estimated classifiers. For that purpose, we use test data set simulated in the same conditions as the training set but keeping the information of the classes Z. The performance metric used is the area under Receiver Operating Characteristic curve (ROC AUC, cf. 2.5.3). It will be used throughout this subsection. We insist on the fact that the performances are calculated on the predictions of the classes Z and not of the labels Y. The performances are compared to a non-traditional classification method: Logistic Regression or Linear Discriminant Analysis on Y given $\tilde{X}$ depending on the setting. We also report the performances of the theoretical Bayes classifier which represents the optimal (cf. Table 4.3). In both settings, the performance of the PU learning classifier is close to the optimal one. Conversely, the non-traditional classifiers are significantly outperformed by PU learning.

These first experiments illustrate the interest of the methodology introduced in Section 4.4. The EM algorithm for PU learning leads to a correct estimation of the parameters of the model and yields to almost optimal classification performances.

Setting	Bayes classifier (optimal)	Non-traditional classification	PU learning
PU-LR	0.98	0.85	0.98
PU-DA	0.99	0.90	0.98

Table 4.3: Classification performances (ROC AUC). Comparison of the PU learning approach with a non-traditional classification and the optimal performances given by Bayes classifier. The table presents the estimates of the mean performances. Standard deviations are below 0.01.

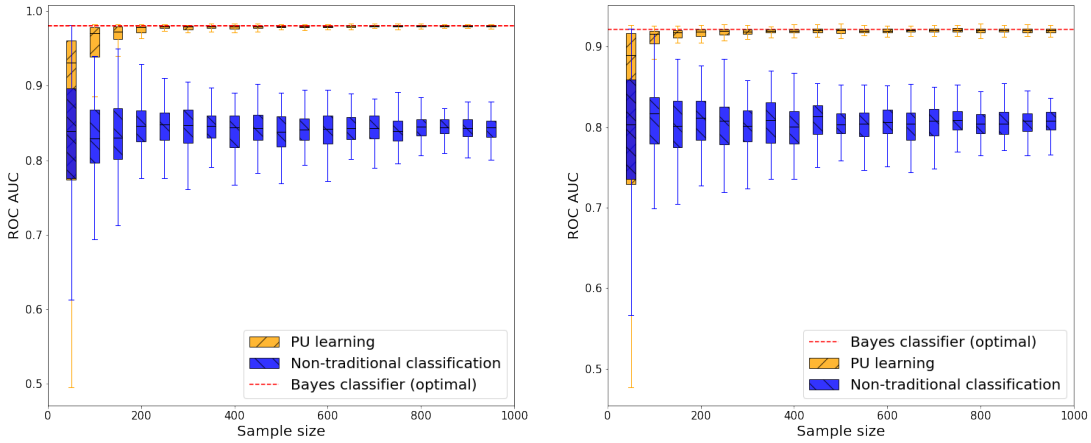

Figure 4.11: Classification performances on simulated data depending on the training sample size: PU-LR simulation setting on the left, PU-DA simulation setting on the right.

4.5.3 Classification performances on multiple experiments

We now carry out multiple series of experiments to study the sensitivity of PU classifiers to different simulation parameters. The objective is to degrade the simulation setting (small sample size, poor separability, misspecification) and to see how the performances of PU learning are affected.

a. Sample size

In this first case of experiments, we keep the same simulation settings and parameters as in Subsection 4.5.2 and only alter the size of the training set. Sample sizes range from 50 to 1000. For each sample size, $B = 100$ identical experiments are repeated in order to estimate the mean performances along with their variability. The results show that the performances tend to the optimal ones (fully supervised classification) when the sample size increases whereas the non-traditional classification performances do not increase (cf. Fig. 4.11). We remark that for very small sample sizes though ($n < 100$), PU learning performances are highly scattered compared to non-standard classification. In such situations, we may prefer the stability of non-standard classification, despite its bias.

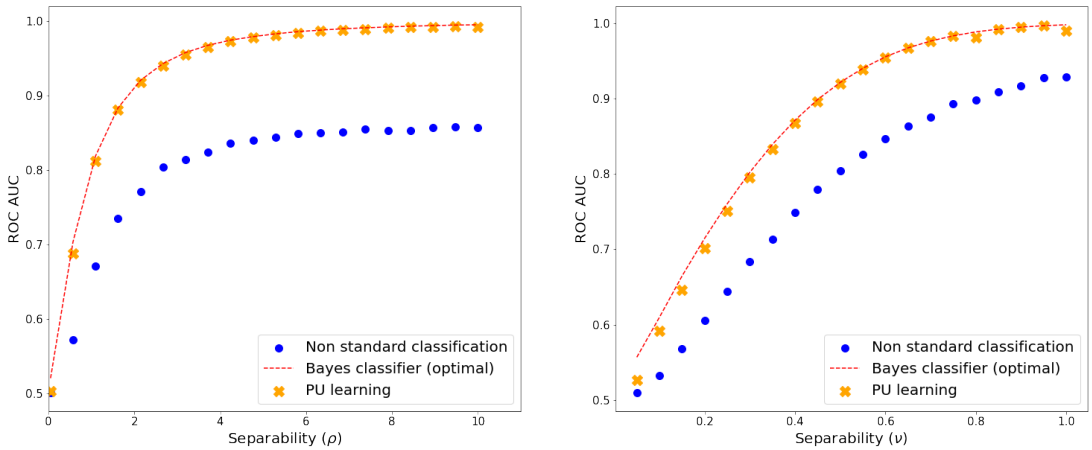

Figure 4.12: Classification performances on simulated data depending on the separability: ρ parameter in PU-LR simulation setting (left), ν parameter on the PU-DA simulation setting (right).

b. Separability

The sample size n is now set to 1000. In this paragraph, we are interested in the performances of PU learning when the separability is poor which of the course makes the classification problem more difficult. In the PU-LR simulation setting, we study the performances of PU learning as a function of ρ. Recall that it controls the overlap between the two classes. In the PU-DA setting we will change the distance between the means of the Gaussian distributions of the classes. We will therefore play with parameter ν (cf. Subsection 4.5.1).

Results are presented in Fig. 4.12. The separability has a clear impact on the performances in both settings which is logical: less separable classes lead to a more difficult classification problem. Therefore the maximum achievable performance (*i.e.* performance for the optimal classifier) also decreases. However, we observe that the performances of PU learning remain close to the optimal classifier. Again, the performances for non-traditional classification are significantly below which illustrates the interest of PU learning.

c. Misspecification

In this paragraph, we study the behaviour of PU learning classification performances when the PU learning parametric model used does not correspond to the simulation setting. More particularly, we want to know how a PU-DA model will perform on a data set not respecting the Gaussian assumptions. Hence, experiments are performed using a PU-LR model on PU-DA simulated data and vice versa. Similar experiments are carried out regarding the propensity models. In summary, four simulation examples are studied:

1. PU-LR simulation labeled with logistic propensity;
2. PU-LR simulation labeled with normal fatigue propensity;
3. PU-DA simulation labeled with logistic propensity;
4. PU-DA simulation labeled with normal fatigue propensity.

For each simulation, we use the four different PU models to estimate the classifier and compare the performances (cf. Table 4.4). It seems that the classification model does not have much effect on the classification performances. Hence, using a PU-LR model or a PU-DA model on PU-DA or PU-LR simulated data yield approximately the same performances. This is not the

Table 4.4: Classification performances depending on the PU learning model used (for classification and propensity). Each experiment is repeated 20 times. The table reports the mean performances along with the standard deviations ($mean \pm std$). Normal propensity refers to the normal fatigue model on propensity (cf. Table 4.2).

Classification model	PU-LR		PU-DA	
Propensity model / **Simulation example**	Logistic	Normal	Logistic	Normal
Ex 1: PU-LR / logistic propensity	$\mathbf{0.80 \pm 0.01}$	0.76 ± 0.06	0.79 ± 0.02	0.75 ± 0.09
Ex 2: PU-LR / normal propensity	0.79 ± 0.04	$\mathbf{0.79 \pm 0.01}$	0.77 ± 0.08	0.77 ± 0.08
Ex 3: PU-DA / logistic propensity	0.80 ± 0.01	0.75 ± 0.07	$\mathbf{0.79 \pm 0.02}$	0.75 ± 0.09
Ex 4: PU-DA / normal propensity	0.78 ± 0.07	0.79 ± 0.02	0.77 ± 0.05	$\mathbf{0.80 \pm 0.01}$

case, however, for the propensity model as we see on Ex 1 and Ex 3 that the models with logistic propensity perform better and are more stable.

4.5.4 Conclusion

In this section, we illustrated the benefits of the methodology introduced in Section 4.4 on simulated data. The advantage of using simulated data is that we can obtain the classes (ground truth) which allow us to properly estimate the performances of PU classifiers. Multiple experiments were performed to study the classification performances of PU learning in different conditions illustrating the interest of the methodology.

4.6 - Application of PU learning to the estimation of a fatigue design criterion

We now want to apply PU learning classification methods in order to identify a fatigue criterion for the design of mechanical parts. Recall that the fatigue criterion denoted η is a classifier designed to predict whether a zone is critical or not. Dang Van mechanical fatigue criterion rely on two features: maximum hydrostatic pressure and critical shear stress. After describing the estimation and evaluation procedure, we apply PU learning to the fatigue data set in a two-dimensional setting only using the two features involved in Dang Van criterion. We highlight some instability issues that can be solved by simplifying the model on the propensity. Finally, we move on to the application of PU learning using additional features which leads to better prediction results.

4.6.1 Estimating and evaluating fatigue criteria

In this subsection, we describe the experimental setting for estimating and testing PU learning fatigue criteria.

The data set is split in two sub-samples with equal sizes: one will be used for training, the other one for testing. In the training phase, the parameters of the PU learning model are estimated on the training data. The model used is PU-LR with a logistic propensity. In parallel, an analogous non-traditional classifier is estimated. This analogous classifier is of the same type as the classification model in the PU learning model: Logistic Regression. The major difference, is that the non-traditional classifier ignores the propensity.

Once the estimation is done, we evaluate the performances of the models on the test set. As previously mentioned, the test data is itself PU data, therefore it does not provide the ground truth on the classes. Hence, we cannot properly assess the performances of the estimated classifier

$\widehat{\eta}$. We have two alternative ways to evaluate the PU learning model.

1. For a covariate vector $(\tilde{x}, t)$, the probability of crack initiation $(Y = 1)$ is modeled as a product $\eta(\tilde{x}) \times e(\tilde{x}, t)$. Using the estimated classifier $\widehat{\eta}$ and propensity $\widehat{e}$, it is thus possible to estimate this probability $\widehat{p}(\tilde{x}, t)$:

$$\widehat{p} = \widehat{\eta}(\tilde{x}) \times \widehat{e}(\tilde{x}, t) \ .$$

   Thus, comparing these posterior probabilities to the labels on the test set provides a first set of performance indicators.

2. Even if the true classes $(Z_i)_{1 \leq i \leq n}$ are not available, we have access at least to a sub-sample of positive instances. Indeed, we already know that labeled instances ($Y = 1$, crack initiations) are critical ($Z = 1$). Besides, we have access to multiple test outputs for each zone (usually 3 to 7). In particular, a zone with at least one crack initiation among the tests performed is critical. Hence, we know that these instances share the same class ($Z = 1$), even those that did not result in crack initiation. It is worth noting that the corresponding individuals in the data set are not strictly identical as the test severities are different. Therefore, we have access to the knowledge of an extended subset of positive instances (critical zones). We denote $\tilde{Z}$ the variable indicating whether a zone initiated at least once ($\tilde{Z} = 1$) or never ($\tilde{Z} = 0$). These "approximate classes" $\tilde{Z}_i$ can be compared to the classification predictions $\widehat{\eta}(x_i)$, allowing to assess the performances of the PU classifier. We insist though that these evaluations may be biased since $\tilde{Z}_i$ does not provide the ground truth on the classes.

Performances are computed on three models: the PU learning model of interest, its non-traditional counterpart and Dang Van fatigue criterion. As in Section 2.5, the performances are evaluated in terms of ROC and PR AUC (cf. Subsection 2.5.3).

As noted in Section 2.5, performance assessment is particularly difficult as the variance on the performance estimation is important. In order to evaluate our models with more consistency, we repeat several times the procedure described above. For each repetition, the train-test split is randomly chosen. This gives us access to the distribution of the performances and will allow us to compare the models.

4.6.2 PU learning: 2D criterion using Dang Van variables

We apply the PU learning methodology defined in Section 4.4 to the fatigue data set in a two-dimensional setting. The two variables used here are the standard variables involved in Dang Van fatigue criterion: critical hydrostatic stress P_c and critical shear stress τ_c.

Looking at the quantitative performances over $B = 100$ repetitions of the estimation-evaluation procedure, we can observe that the PU learning method performs similarly as the non-traditional classification method when evaluating the method on the prediction of labels Y (cf. Fig. 4.13, top). However, we notice that the performances measured on the class predictions are extremely variable for PU learning and that the non-traditional approach is preferable in this case (cf. Fig. 4.13, bottom).

The estimation results for one experiment are represented in Figure 4.14. The left diagram features the estimated classification rule $\eta_{\widehat{\theta}}$ of interest (fatigue criterion) as a function of the covariates $\tilde{x} = (P_c, \tau_c)$ representing the stresses at the nominal severity. The diagram on the right focuses on the estimated propensity function $e_{\widehat{\phi}}$ as a function of the transformed feature vector $\tilde{x}_{eq} = (P_{c,eq}, \tau_{c,eq})$ representing the stresses at the equivalent severity of the test. We recall that $\tilde{x}_{eq}$ depends on t. The estimated probability of being labeled is given by the product $\eta_{\widehat{\theta}}(\tilde{x}) \times e_{\widehat{\phi}}(\tilde{x}_{eq})$.

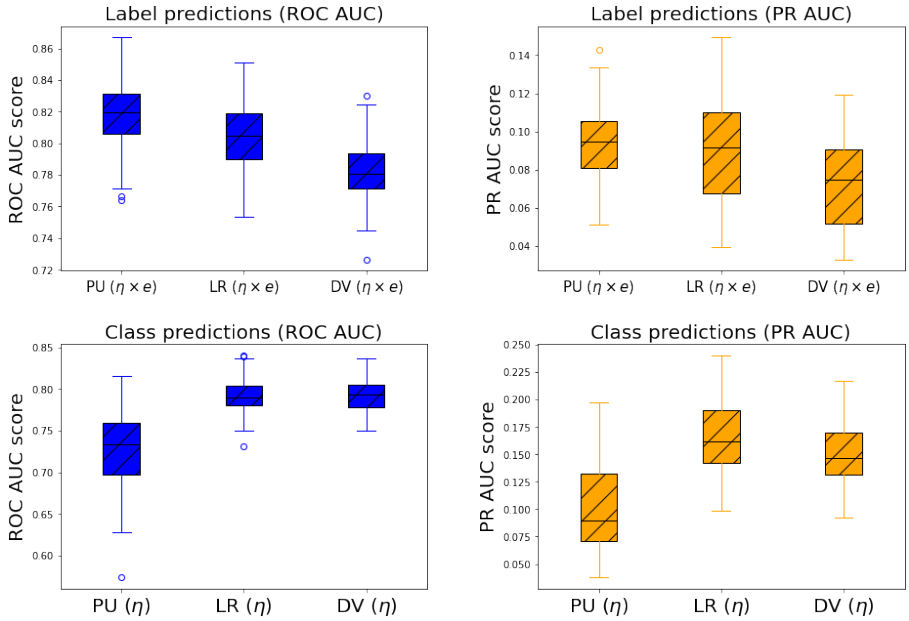

Figure 4.13: Quantitative performances of models with two variables: prediction performances on the labels (top, $\eta \times e$) and on the classes (bottom, η). The metrics used are ROC AUC (left, blue) and PR AUC (right, orange). Each boxplot represents the distribution of the prediction performances. Each series of three boxplots corresponds to (from left to right): PU-LR, non-traditional Logistic Regression and mechanical Dang Van criterion.

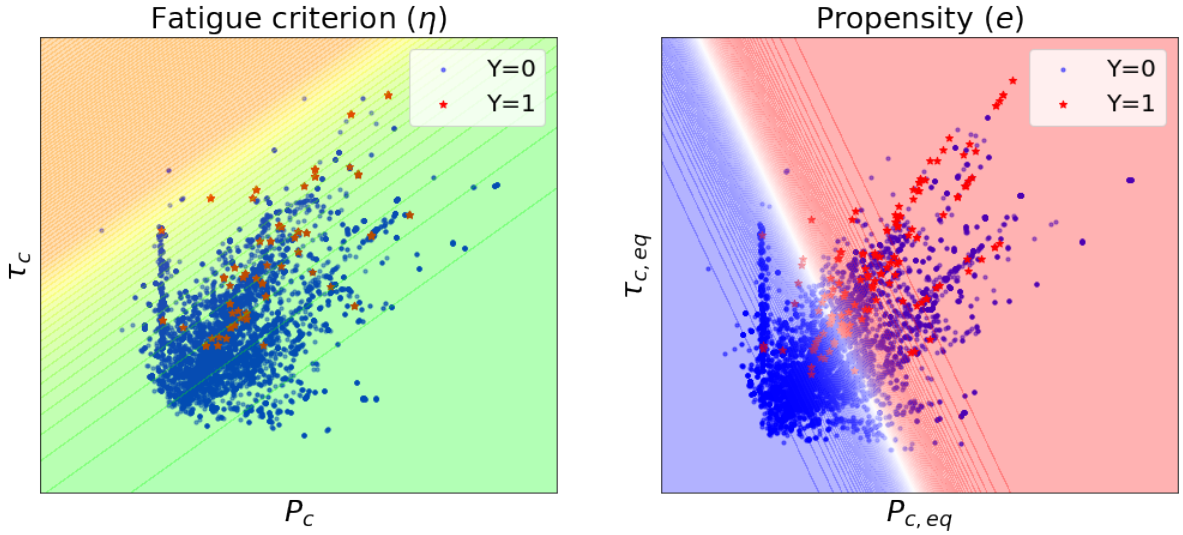

Figure 4.14: Example of PU learning classification results with two variables: estimated fatigue criterion on the left, estimated propensity on the right. On the left, the background color represents the criticality: from green (low values of η) to orange (high values of η). On the right, the background color represents the propensity: red for high values, blue for low values.

In the results of Figure 4.14, the estimated classifier is quite different from the Dang Van criterion commonly used in fatigue design of automotive parts as the linear criterion has a positive slope whereas Dang Van criterion has a negative slope. According to this PU criterion, the probability for a zone to be critical is an increasing function of τ_c and a decreasing function of P_c. The estimated propensity shows that the probability of crack initiation for critical instances is an increasing function of $P_{c,eq}$ and $\tau_{c,eq}$ calculated at the test equivalent severity, which is

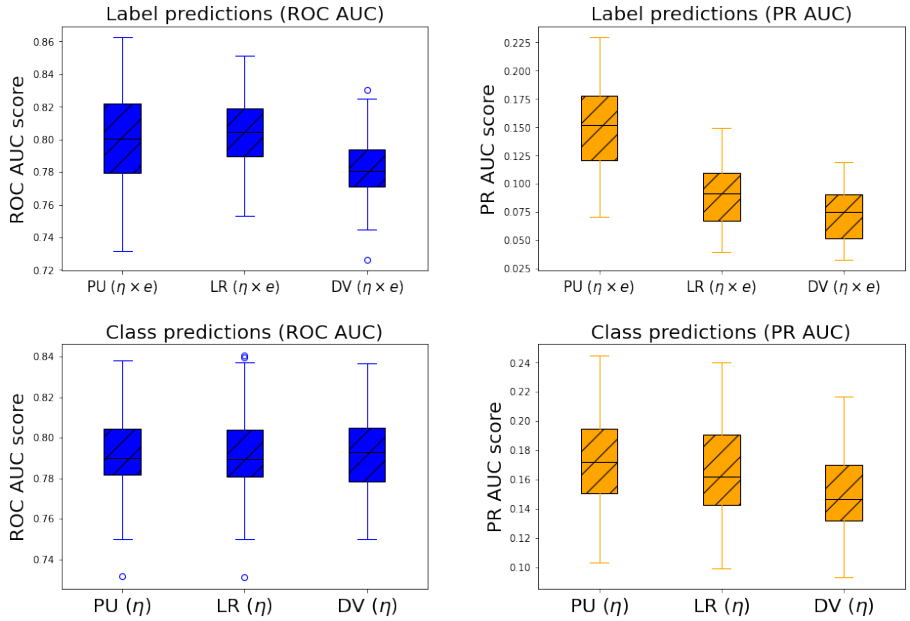

Figure 4.15: Quantitative performances of models with the simplified propensity (cf. Eq. 4.21): prediction performances on the labels (top, $\eta \times e$) and on the classes (bottom, η). The metrics used are ROC AUC (left) and PR AUC (right). Each boxplot represents the distribution of the prediction performances. Each series of three boxplots corresponds to (from left to right): PU-LR, non-traditional Logistic Regression and mechanical Dang Van criterion.

natural. This example allows to understand the quantitative performances described above. Indeed, while the product $\eta_{\widehat{\theta}}(\tilde{x}) \times e_{\widehat{\phi}}(\tilde{x}_{eq})$ provides decent performances in the predictions of the labels (Y), the classifier alone $\eta_{\widehat{\theta}}$ estimates poorly the risk for a zone to be critical (Z).

This unwanted behaviour of PU classifier is mainly explained by the high correlation between the feature vector $\tilde{x}$ used in the classification model η and the transformed vector $\tilde{x}_{eq}$ involved in the propensity e. Besides $\tilde{x}$ and $\tilde{x}_{eq}$ both have a similar influence on the crack initiation probability: a higher nominal stress results in a higher probability for a zone to be critical and, at the same time, in a higher probability for this zone to crack. In some sense, the estimated propensity in Figure 4.14 absorbed part of the information that should be contained in the classifier η. It is thus crucial to clearly separate the effects accounted for in the classification model and in the propensity. We achieve this separation by simplifying the propensity model so that it only accounts for the testing conditions t through the equivalent severity $f_{eq}(t)$. This way, the probability for an instance to be labeled is modeled as:

$$\mathbb{P}_{\theta,\phi}\left(Y = 1 \,|\, X = (\tilde{x}, t)\right) = \eta_\theta(\tilde{x}) \times e_\phi\left(f_{eq}(t)\right) \ . \tag{4.21}$$

In addition, this model remains perfectly logical from a physical point of view. The classifier η is the fatigue criterion predicting the criticality of a zone given the stresses $\tilde{x}$. On the other hand, the propensity e characterizes the probability of crack initiation for critical zone: considering two zones with equal criticality, then the probability of crack initiation depends on the test conditions t, through the equivalent severity $f_{eq}(t)$.

This simpler PU learning model satisfies the conditions for identifiability presented in 4.3 and the estimation methodology presented in Section 4.4 can be applied identically.

The results show that the performances are far more stable (cf. Fig. 4.15). The performances evaluated on label predictions (Fig. 4.15, upper diagram) show that the PU learning models seem to better predict the risk of crack initiation compared to Dang Van criterion and the non traditional classifier (higher prediction performances in terms of PR AUC). This effect is logical in the sense that the PU learning explicitly account for the effect of test conditions on

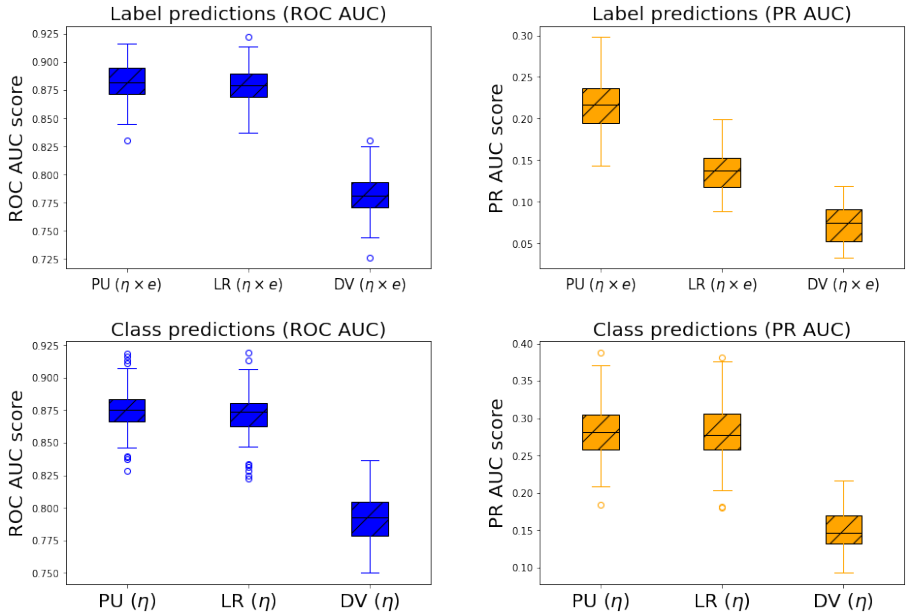

Figure 4.16: Quantitative performances (ROC AUC) of models with five variables: prediction performances on the labels (top, $\eta \times e$) and on the classes (bottom, η). The metrics used are ROC AUC (left) and PR AUC (right). Each boxplot represents the distribution of the prediction performances. Each series of three boxplots corresponds to (from left to right): PU-LR, non-traditional Logistic Regression and mechanical Dang Van criterion.

crack initiation. When evaluating the predictions of the classifier alone, the performances of PU learning are quite similar to non-traditional classification.

4.6.3 PU learning with additional variables

We now consider additional variables in order to better characterize critical zones. In this subsection, the covariate vector $\tilde{x}$ contains five variables:

- the critical shear stress τ_a^m;
- the mean triaxiality T_m^m;
- the mean material parameter $\overline{\tau_{mat}}$ over the zone;
- the maximum longitudinal stress over the edges elements of the zone EL_a^a;
- the maximum transversal stress over the welds of the zone WL_a^a.

These features not only characterize the stress on the zone but also account for the singularities (edges and welds). For further information on the features, refer to Chapter 2.

We carry out the same experiments as in Subsection 4.6.2 and compare the prediction performances of PU learning, non-traditional classification and Dang Van criterion. We recall that Dang Van criterion still rely on two variables whereas both PU learning and non-traditional classification use the variables listed above. The results are presented in Figure 4.16.

We notice that both statistical methods get higher performances compared to Dang Van criterion. In the prediction of labels, PU learning seems to outperform non-traditional classification (higher PR AUC scores).

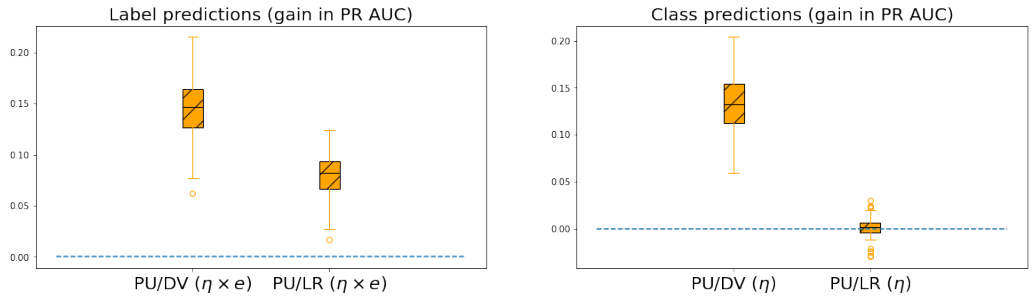

Figure 4.17: Performance comparisons (PR AUC): label predictions on the left, class predictions on the right. For each figure, the boxplot on the left represents the difference of performances between PU learning and Dang Van criterion, the second compares PU learning and non-traditional classification.

In order to compare more efficiently PU learning to non-traditional classification, let us look at the differences of performance achieved experiment by experiment. For j between 1 and 100, denote s_j^{PU} and s_j^{NT} the scores of PU learning and non-traditional classification for the j^{th} experiment. Now, instead of looking at the empirical distributions $(s_j^{PU})_{1\leq j\leq 100}$ and $(s_j^{NT})_{1\leq j\leq 100}$ separately, we rather consider $(s_j^{PU} - s_j^{NT})_{1\leq j\leq 100}$. This is legitimate as for each j, s_j^{PU} and s_j^{NT} are calculated over the same test set for models estimated over the same training sets. A similar comparison is provided between PU learning and Dang Van criterion. Results are presented in Figure 4.17. Looking at the performances on the label and class predictions, PU learning outperforms Dang Van criterion for every experiment with a significant median gain of approximately 0.14 over PR AUC. On label predictions, PU learning provides a median gain of approximately 0.08 over non-traditional classification. Finally, when comparing PU learning to non-traditional classification, it appears that both methods provide similar performances for class predictions. This can seem surprising: indeed, as the goal of PU learning is to better estimate the classifier by accounting for PU label noise, we should expect better performances on class predictions. However, as explained earlier, the performances of the PU classifier may remain biased as we do not know all the critical zones in the test set. Hence, the results only tell that PU learning classifier is as good as its non-traditional counterpart in characterizing already labeled critical zones.

4.6.4 Conclusion

In this section, we applied PU learning to the identification of a fatigue criterion. The results confirm those of Chapter 2 in the sense that additional features significantly improve the identification of critical zones. PU learning seems to provide slightly better results than non-traditional classification in the prediction of crack initiations. The results on class predictions are similar but remain biased as we do not have access to the true classes to properly evaluate the methods. More importantly, PU learning is able to account for the presence critical zones among the unlabeled ones. Finally, the fatigue criterion is far from being perfect as some critical zones are still poorly characterized through this statistical criterion. This is due to the fact that we are still lacking important variables to characterize critical zones (effects of manufacturing processes, residual stresses...).

Conclusion et perspectives (en français)

Conclusion générale

L'objectif de cette thèse était le développement de critères de fatigue via l'utilisation de méthodes statistiques pour mieux identifier les défauts de conception sur un modèle numérique de pièce mécanique. Cette amélioration de la détection et de la caractérisation des zones dites *critiques* permet d'éviter des itérations longues et coûteuses entre conception et essais de validation. Par conséquent, le coût et la durée de développement peuvent être réduits. Les résultats présentés dans cette thèse s'appuient sur une base de données fatigue rassemblant un historique de données de modèles numériques et d'essais de validation correspondants. Alors que la plupart des modèles de fatigue sont calibrés sur des géométries élémentaires (essais sur éprouvettes), nous avons étudié comment cette nouvelle source de données pouvait être exploitée pour améliorer la prédiction des risques de fatigue.

Nous avons introduit les principaux modèles de fatigue existants en mettant en évidence une différence entre les modèles de durée de vie en fatigue et les critères de fatigue. Les premiers cherchent à modéliser la durée de vie d'une pièce mécanique compte tenu des contraintes auxquelles elle est soumise : d'un point de vue statistique, il s'agit d'une tâche de régression. Les seconds visent à indiquer si le cycle de contrainte appliqué à la pièce mécanique est supérieur ou non à la limite de fatigue : il s'agit d'une tâche de classification. L'objectif du dimensionnement à la fatigue étant de garantir que la pièce satisfait aux exigences de durabilité, nous nous sommes intéressés aux critères de fatigue.
Nous avons ensuite effectué une analyse exploratoire de la base de données fatigue de Stellantis. Un pré-traitement des données a permis de regrouper les observations par zones avec un avantage triple. Tout d'abord, le déséquilibre entre zones cassées et non cassées a été considérablement réduit. Deuxièmement, cela a permis de considérer les observations (zones) comme indépendantes alors que deux éléments proches (sur un modèle à éléments finis) sont fortement corrélés. Troisièmement, cette notion de zone est plus robuste aux erreurs de localisation des amorces de fissures. Un ensemble de variables appropriées a été introduit pour décrire les zones : invariants de contraintes sur l'élément le plus critique de la zone, moyenne spatiale de grandeurs physiques sur la zone, informations géométriques et caractéristiques propres aux singularités (soudures et bords de tôles). La richesse de la base de données fatigue a été étudiée dans le cadre d'une analyse non supervisée permettant d'identifier différents types de zones caractérisées par différents groupes de variables. Même si cette analyse préliminaire n'aide pas à l'identification des zones critiques, elle a fourni des informations utiles sur la structure des données.
Nous avons ensuite abordé l'estimation de critères de fatigue multiaxiale. Une version probabilisée

du critère de Dang Van a été introduite, permettant d'estimer jointement les paramètres matériau et de dispersion. Ce critère a été estimé sur un jeu de données auxiliaire rassemblant des résultats numériques et des données d'essais de fatigue sur des structures élémentaires avec soudures (éprouvettes). Cependant, ce critère probabiliste montre ses limites lorsqu'il est appliqué à l'identification de zones critiques sur des pièces mécaniques réelles et complexes. Par conséquent, nous nous sommes directement appuyés sur la base de données Stellantis pour estimer des critères de fatigue à l'aide de méthodes de classification supervisée. Ces critères statistiques de fatigue offrent des performances en prédiction meilleures que le critère de Dang Van. La principale valeur ajoutée de ces critères par rapport aux méthodes standards est leur capacité à prendre en compte un plus grand nombre de descripteurs, ce qui explique ce gain de performance significatif.

Les méthodes de classification supervisée ne traitent que les amorces de fissures comme des observations positives. Cependant, une zone sans amorce de fissure détectée peut en réalité être critique. En effet, du fait de la sévérité de l'essai, de sa durée et du caractère aléatoire de l'amorçage de fissure, une zone critique peut ne pas être détectée lors d'un essai de fatigue et donc rester non étiquetée. A l'inverse, une amorce de fissure est une zone critique avérée. La prise en compte de ce bruit d'étiquette asymétrique dans la classification est l'objectif de l'apprentissage PU (*Positive Unlabeled*). Dans un certain sens, l'apprentissage PU peut être considéré comme un mécanisme de censure dans un cadre de classification. Alors que la censure affecte généralement l'estimation des modèles de régression (modèles de durée de vie en fiabilité par exemple), l'apprentissage PU traite de l'estimation des classifieurs lorsqu'une partie des osbervations de la classe positive est censurée. Ce problème a été étudié sous les angles théorique, méthodologique et pratique.

D'un point de vue théorique, nous avons étudié l'apprentissage PU sous l'hypothèse SAR (*Selected At Random*), c'est-à-dire lorsque la probabilité qu'une observation de classe positive soit étiquetée (*i.e.* la propension) dépend des covariables. Nous avons démontré des bornes de risque supérieures et inférieures en soulignant comment le taux de convergence dépend de la propension (*i.e.* la quantité de bruit d'étiquetage). Les résultats ont été illustrés par des expériences numériques qui permettent de retrouver les taux de convergence théoriques.
Globalement, les résultats théoriques permettent de comprendre la difficulté de l'estimation d'un classifieur dans le cadre de l'apprentissage PU. Une faible propension se traduit par moins d'observations positives étiquetées et donc une tâche plus difficile. Inversement, lorsque la propension tend vers 1, les performances de l'apprentissage PU tendent vers celles de la classification supervisée. Le rôle de la propension est donc crucial. Dans les applications de fatigue, l'augmentation de la sévérité (ou de la durée) de l'essai augmente la probabilité qu'une zone critique amorce et soit ainsi étiquetée. Par conséquent, la propension est augmentée. Ainsi, l'augmentation de la sévérité des essais de fatigue peut faciliter l'estimation des critères de fatigue par apprentissage PU. Cependant, en pratique, la sévérité de l'essai ne peut pas être augmentée indéfiniment : cela exposerait la pièce mécanique à d'autres modes de défaillance qui ne font pas partie du dimensionnement à la fatigue (plasticité par exemple). In fine, les résultats théoriques renforcent un principe déjà bien établi dans la validation de la résistance à la fatigue: les essais de fatigue doivent être réalisés pour une sévérité proche de la résistance de la pièce. Si la sévérité est trop faible, il est peu probable que le composant casse à l'essai et la résistance à la fatigue risque d'être impossible à estimer.

Nous avons enfin développé une méthodologie pratique pour estimer un classifieur PU adapté à la problématique de fatigue mécanique. Cette méthodologie s'appuie sur des modèles paramétriques à la fois sur le critère de fatigue (classifieur) et sur la propension. La propension rend compte des conditions d'essais par le calcul d'une sévérité équivalente. L'estimation des paramètres est réalisée par maximum de vraisemblance à l'aide de l'algorithme EM. L'intérêt de la méthodologie

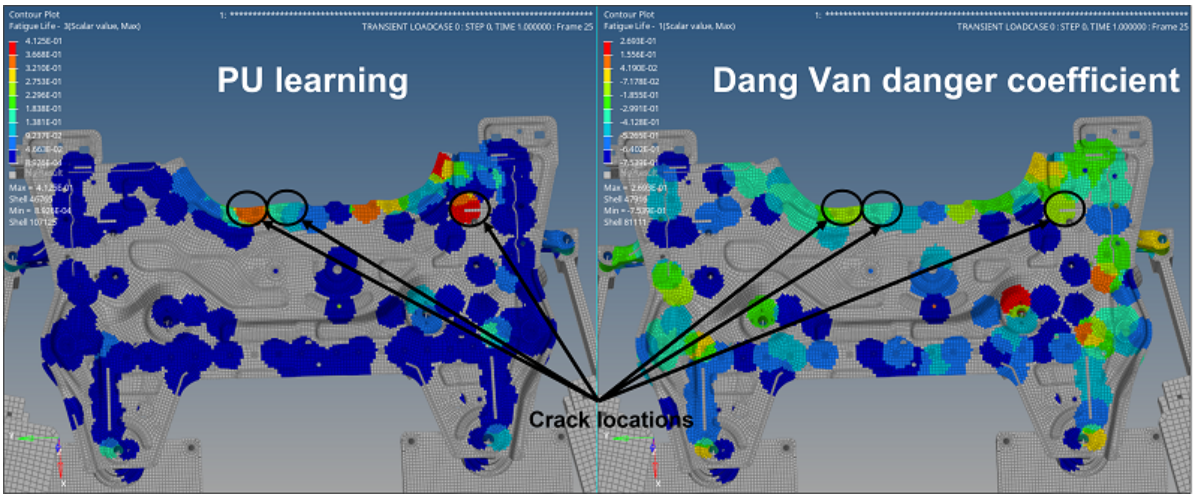

Figure A: Prédictions pour le classifieur PU (gauche) et le critère de Dang Van (droite) sur un modèle à éléments finis (berceau sous chargement longitudinal) : le rouge correspond aux zones critiques, le bleu aux zones sûres. Les zones grises n'ont pas été sélectionnées durant la phase de pré-traitement (construction des zones, cf. sous-section 2.2.2) : elles doivent être considérées comme bleues. Les zones d'amorçage sur cette face de la pièce sont repérées par des cercles noirs.

est illustré sur des données simulées. Ensuite, l'application de la méthode au jeu de données de fatigue fournit des résultats intéressants. Les performances du classifieur PU sont comparées à la classification standard et au critère de Dang Van. Par rapport aux classifieurs standards, l'apprentissage PU modélise mieux le risque d'amorçage de fissure. Lors de l'évaluation des modèles sur les classes prédites, les classifieurs standards et PU fournissent des performances similaires. Cependant, les mesures d'évaluation peuvent être biaisées en raison de la présence de bruit d'étiquette PU y compris dans les données de test. Comme pour les méthodes supervisées classiques, les performances des classifieurs PU dans l'identification des zones critiques sont significativement supérieures à celles du critère Dang Van traditionnel. Enfin, l'intérêt principal de l'apprentissage PU réside dans la modélisation qui sépare les effets du critère de fatigue (classifieur PU) et des conditions d'essai (propension) qui ont une influence sur les étiquettes observées.

Dans le cadre de cette thèse, nous avons développé de nouveaux outils statistiques pour l'identification de zones critiques à partir de modèles numériques : des classifieurs obtenus par des méthodes classiques de classification supervisée et des classifieurs PU qui tiennent compte du bruit d'étiquette spécifique aux essais de fatigue. Une fois les classifieurs estimés, ils peuvent être facilement déployés dans un logiciel d'analyse par éléments finis pour fournir des prédictions sur différentes zones d'une conception. Comme pour le coefficient de danger du critère de Dang Van, il est simple d'évaluer la probabilité que chaque zone soit critique.

Le déploiement de la méthodologie est représenté sur la Figure A. Sur cet exemple de modèle de berceau sous chargement longitudinal, les prédictions du critère de Dang Van et celles du classifieur PU sont comparées. Rappelons que le coefficient de danger d'une zone est défini comme le coefficient de danger maximal parmi les éléments de la zone. Il est tout d'abord important de noter que le critère identifié par apprentissage PU ne met en évidence que quelques zones critiques, la majorité reste bleue (sûre). A l'inverse, le coefficient de danger de Dang Van est élevé (valeurs proches du seuil 0 ou au-dessus) pour un grand nombre de zones. Par conséquent, suivre les prédictions du critère de Dang Van nécessiterait de nombreux renforcements de la pièce, la plupart de ces renforcements étant probablement inutiles. Au lieu de cela, l'apprentissage PU génère moins de faux positifs et est plus à même de guider les équipes de conception afin qu'elles

se concentrent sur les zones les plus critiques (celles qui ont le plus de chance de casser à l'essai). De plus, les zones connues d'amorçage de fissures sont mieux identifiées par le critère PU que par le coefficient de danger de Dang Van. Globalement, ces observations confirment le gain substantiel en performances obtenu par les critères statistiques de fatigue par rapport au critère classique de Dang Van.

Perspectives

Ces travaux ont ouvert la voie au développement et au déploiement de critères statistiques de fatigue facilitant le dimensionnement à la fatigue des pièces de sécurité des véhicules. Plusieurs directions de recherche restent ouvertes et pourraient conduire à des améliorations significatives de la méthodologie et de son efficacité.

Extension de l'approche à d'autres pièces

Dans le cadre de cette thèse, seuls des modèles de berceaux et de traverses ont été considérés. Ces pièces de sécurité des véhicules concentrent la majorité des problèmes de dimensionnement du fait de leur complexité (géométrie, grand nombre de soudures...). Cependant, le dimensionnement à la fatigue ne se limite pas à ces deux types de pièces mécaniques. Par conséquent, la méthodologie développée dans cette thèse peut être étendue au-delà de ces deux types de pièces mécaniques et même au-delà du périmètre de la Liaison Au Sol (LAS) : par exemple, des applications au système de propulsion (packs batteries) pourraient être envisagées. Un défi supplémentaire serait de mélanger des modèles à éléments finis volumiques (3D) avec des modélisations coques (2D) dans la base de données fatigue et dans l'estimation des critères de fatigue. Pour l'instant, seuls des modèles coques sont considérés. Enfin, une autre perspective serait d'étudier l'applicabilité de cette méthodologie à la fatigue vibratoire.

Amélioration de la base de données fatigue

Initialement composée de quelques études de cas, la base de données a ensuite été enrichies au cours de la thèse, atteignant environ 40 études de cas et 300 observations de fissures. Cependant, le potentiel de données disponibles est bien plus conséquent. L'augmentation de la taille de la base de données permettrait d'accroître la diversité des géométries et des chargements considérés et ainsi de renforcer les critères statistiques estimés.

D'un point de vue pratique, le développement de la base de données fait face à deux principaux défis. Tout d'abord, cette base s'appuie sur deux sources de données différentes (résultats de calculs par éléments finis et rapports d'essais de fatigue). Les calculs numériques et les essais de fatigue n'étant pas réalisés par les mêmes équipes, il n'est pas toujours aisé de rassembler *a posteriori* un résultat de calcul par éléments finis avec le rapport d'essai de fatigue correspondant. Par ailleurs, la construction de la base de données fatigue nécessite d'étiqueter manuellement les zones d'amorçage de fissure sur le modèle numérique. Ce processus d'étiquetage manuel est parfois compliqué car nous n'avons accès qu'aux photos des amorces de fissures (via le rapport d'essai de fatigue) et il n'est pas toujours simple de localiser ces zones critiques sur le modèle numérique.

À l'avenir, il serait certainement plus facile de faire l'étiquetage lors des essais de fatigue. En effet, les équipes en charge des essais de validation pourraient mieux localiser les zones d'initiation de fissures car elles ont accès au prototype et n'ont pas à se fier aux photos. De plus, comme les calculs par éléments finis sont effectués avant les essais, le modèle numérique pourrait être transmis à l'équipe en charge de la conduite des essais de fatigue afin qu'elle puisse reporter les résultats des essais sur le modèle numérique. Cela permettrait d'alimenter la base de données de

façon continue et efficace.

Un autre axe d'amélioration concernant la base de données est la précision des calculs par éléments finis. Dans ce travail (et dans le dimensionnement en fatigue en général), les critères de fatigue appliqués reposent sur des grandeurs physiques calculées à l'aide du modèle numérique. Les résultats des calculs par éléments finis restent incertains, en particulier sur certaines zones à géométrie complexe (trous, encoches, soudures...). Une grande partie des difficultés des modèles de fatigue à prévoir les zones critiques est due aux limites de la modélisation par éléments finis qui ne tient pas compte de plusieurs phénomènes (procédés de fabrication, contraintes résiduelles). L'amélioration de ces modèles étant un sujet actif dans l'industrie automobile, nous pouvons nous attendre à ce que les modèles à éléments finis gagnent en précision dans un futur proche. En conséquence, la base de données fatigue pourrait bénéficier de jeux de données de meilleure qualité, ce qui améliorerait également la caractérisation des phénomènes de fatigue par l'utilisation des méthodologies développées dans cette thèse.

Meilleure prise en compte de la géométrie dans les critères de fatigue

Les critères de fatigue développés dans cette thèse s'appuient sur plusieurs variables dont des descripteurs géométriques de zones (épaisseur, contraintes exprimées dans un repère local pour les singularités...). Ces caractéristiques géométriques ont été choisies au début de l'étude (cf. Chapitre 2). À l'avenir, il serait intéressant d'étudier comment ces choix peuvent être justifiés statistiquement et d'explorer d'autres méthodes qui pourraient mieux rendre compte de la géométrie.

Rappelons tout d'abord que la construction des zones repose sur certains paramètres choisis empiriquement dans la section 2.2 : un critère de sélection pour limiter le nombre de zones et un rayon délimitant la taille d'une zone. Il serait intéressant d'étudier comment les modèles et performances estimés dépendent de ces hyper-paramètres. Cela aiderait à concevoir une méthodologie pour sélectionner ces hyper-paramètres.

De plus, parmi les 60 variables introduites dans la sous-section 2.2.3, certaines caractérisent l'élément le plus critique de la zone. Cet élément est identifié comme celui ayant le coefficient de danger de Dang Van maximal sur la zone, considéré comme l'élément ayant le plus de chance d'amorcer à l'essai. Cette hypothèse pourrait être assouplie en laissant le modèle d'apprentissage traiter l'incertitude concernant l'emplacement précis de l'initiation de la fissure. En particulier, le *Multiple Instance Learning* (MIL, cf. Sabato and Tishby, 2012; Herrera et al., 2016) pourrait fournir une solution à ce problème. Le MIL est une méthode de classification supervisée où seuls des sacs (groupes) d'observations sont étiquetés. Dans le cadre binaire, un groupe d'observations est étiqueté positif si au moins une observation est positive ; sinon, il est étiqueté négatif. Par conséquent, le MIL semble bien adapté à notre tâche, car nous n'avons que des zones étiquetées (groupes d'éléments) et ne savons pas avec certitude quel(s) élément(s) a causé l'initiation de la fissure.

Enfin, les critères de fatigue estimés dans le cadre de cette thèse ne tiennent compte que d'un nombre limité de caractéristiques géométriques. Comme nous l'avons vu, certaines zones critiques sont encore mal identifiées par les critères en raison des limites et des incertitudes du modèle à éléments finis. Une meilleure prise en compte de la géométrie pourrait aider à identifier les zones critiques. Même si les effets liés aux procédés de fabrication et aux contraintes résiduelles ne peuvent pas être calculés, deux zones avec une géométrie et des matériaux similaires auront une résistance à la fatigue similaire. Par conséquent, considérer l'ensemble de la géométrie de la zone (forme, courbure, ...) dans le critère pourrait aider à améliorer les prédictions. Une perspective intéressante serait de considérer la zone comme une image tridimensionnelle et d'utiliser des méthodes d'apprentissage profond adaptées pour extraire des caractéristiques permettant de

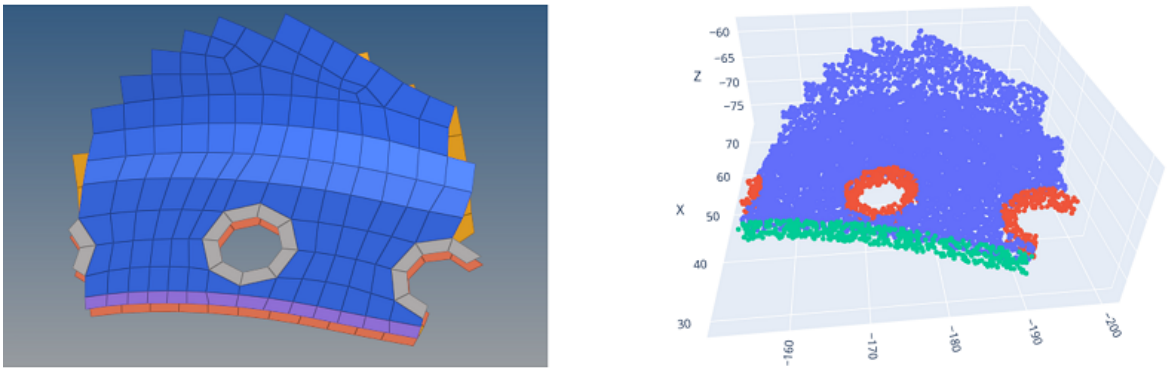

Figure B: Exemple de zone sur un modèle à éléments finis (gauche) et de sa conversion en nuage de points (droite).

mieux prédire le risque d'initiation de fissures. Bien entendu, une image tridimensionnelle (voxels) n'est probablement pas le format le plus approprié, les pièces mécaniques étant des structures planes. Une solution serait de convertir une zone en nuage de points en échantillonnant des points sur la zone (cf. Fig. B) et de considérer ce format pour les données d'entrée. Une autre solution serait de considérer directement les données brutes (maillage de la zone) : une zone peut ainsi être considérée comme un graphe avec des sommets (éléments) et des arêtes (liens entre éléments). Au cours des dernières années, différentes architectures de réseaux de neurones ont été développées pour répondre à ce type de tâches. Par exemple, *PointNet* est une architecture de réseau de neurones conçue pour résoudre des tâches d'apprentissage supervisé en considérant les nuages de points comme données d'entrée (cf. Qi et al., 2017). De même, *MeshCNN* est conçu pour des prédictions à partir de données de maillage (Hanocka et al., 2019). Ces architectures font partie d'un cadre émergent dans l'apprentissage profond appelé *Geometric Deep Learning* qui vise à développer des architectures de réseaux de neurones pour résoudre des tâches impliquant des structures de données spécifiques telles que des nuages de points, des graphes, des maillages et des variétés (Bronstein et al., 2017). Ces méthodes pourraient aider à prédire les fissures de fatigue et pourraient également servir à d'autres applications impliquant l'analyse de modèles à éléments finis.

Perspectives pour l'apprentissage PU

L'une des principales contributions de cette thèse est l'étude de l'apprentissage PU sous l'hypothèse SAR et le développement d'une méthodologie pratique pour estimer des critères de fatigue. L'apprentissage PU demeure un sujet de recherche actif et certains problèmes restent ouverts.

Une extension naturelle des résultats théoriques démontrés dans le chapitre 3 serait d'étudier l'apprentissage PU pour des fonctions de perte convexes. En effet, ce travail se concentre sur un risque empirique fondé sur une fonction de perte $0-1$ qui est intéressante pour étudier les propriétés théoriques de l'apprentissage PU. Cependant, en pratique, l'optimisation d'un tel risque est intractable numériquement. Une solution classique pour surmonter cette difficulté est de considérer des fonctions de perte convexes pour lesquelles l'optimisation est tractable. Il serait donc intéressant d'étendre les garanties théoriques à ces fonctions de perte. Ce problème a déjà été étudié dans le cadre de la classification binaire standard (cf. Bartlett et al., 2006; Blanchard et al., 2008). En outre, Plessis et al. (2014) ont prouvé des bornes de risque sous l'hypothèse SCAR en utilisant des fonctions de perte convexes, mettant en évidence un taux de convergence paramétrique. Dans la thèse, des expériences numériques ont été réalisées sous l'hypothèse SAR

en utilisant une perte logistique, qui présente également un taux de convergence paramétrique affecté par la propension (cf. Section 3.6).

Les bornes de risque du chapitre 3 dépendent de la propension, plus particulièrement de son minimum qui est supposé strictement positif. Cette hypothèse sur la propension reste forte et il serait intéressant d'étudier dans quelle mesure elle pourrait être relâchée. Par exemple, on pourrait autoriser la propension à prendre des valeurs arbitrairement proches de 0 dans la mesure où cela ne se produirait qu'avec une faible probabilité. Les simulations de la section 3.6 traitent le cas d'une propension logistique écrêtée afin de rester au-dessus d'une valeur minimale $e_m > 0$. Cependant, lorsque e_m tend vers 0, l'excès de risque se stabilise, ce qui signifie que nous avons toujours un taux de convergence similaire même si la propension peut prendre des valeurs arbitrairement petites.

La méthodologie d'apprentissage PU développée dans ce manuscrit soulève également d'autres défis. D'abord, la sélection de variables reste un problème crucial, surtout si le nombre de variables est du même ordre de grandeur que le nombre d'observations étiquetées. Ce problème est même double car on pourrait avoir besoin de sélectionner des variables à la fois dans le modèle de classification et dans la propension.

Ensuite, une autre direction intéressante serait de considérer d'autres modèles de classification comme classifieur PU (SVM, forêts aléatoires...) et d'adapter la méthodologie à ces modèles. En particulier, nous avons vu que les forêts aléatoires obtenait de meilleurs résultats que la régression logistique sur le jeu de données de fatigue en classification standard (cf. Section 3.1). Par conséquent, nous pouvons nous attendre à de meilleures performances de classification en tirant parti de ces méthodes dans le cadre de l'apprentissage PU.

Enfin, une approche alternative à l'apprentissage PU est la classification de Neyman-Pearson. Blanchard et al. (2010) ont développé une méthodologie d'apprentissage PU dans ce cadre sous l'hypothèse SCAR. L'intérêt de la classification de Neyman-Pearson est qu'elle vise à minimiser le risque de seconde espèce (taux de faux positifs) sous contrôle du risque de première espèce (taux de faux négatifs). Ceci est particulièrement intéressant pour l'application au dimensionnement à la fatigue car on n'accorde pas la même importance aux deux types de risques : le risque de première espèce est plus important (proportion de zones critiques non identifiées) que celui de seconde espèce (proportion de zones sûres prédites comme critiques). Par conséquent, il serait intéressant d'étudier l'extension potentielle de la classification de Neyman-Pearson à l'apprentissage PU sous l'hypothèse SAR.

Conclusion and perspectives

General conclusion

The objective of this thesis was the development of statistical fatigue criteria to better identify weaknesses in numerical models of mechanical parts. This improvement in the detection and characterization of so-called *critical zones* can avoid lengthy and expensive iterations between conception and validation tests. Consequently, the development cost and duration can be reduced. The results presented in this thesis rely on a fatigue database gathering data from previous numerical designs along with corresponding validation tests. While most fatigue models are calibrated on small-scale and elementary components (coupon tests), we investigated how this new source of data could be leveraged to improve the prediction of fatigue risks.

We introduced existing fatigue models highlighting a difference between fatigue lifetime models and fatigue criteria. The former seeks to model the lifetime of a mechanical part given the stresses it is subjected to: from a statistical point of view, this is a regression task. The latter aims at indicating whether or not the stress cycle applied to the mechanical part is above the fatigue limit: this is a classification task. As the objective of fatigue design is to guarantee that the part satisfies the durability requirements, we are interested in fatigue criteria.
We then carried out an exploratory analysis of Stellantis fatigue data set. A pre-processing of the data allowed to group observations by zones with a triple benefit. First, the imbalance between broken and unbroken zones was drastically reduced. Second, it allowed to consider the observations (zones) as independent whereas two nearby elements are strongly correlated. Third, it is more robust to errors in the localization of crack initiations. A set of appropriate features was introduced to describe zones: stress invariants on the most critical element of the zone, spatial average of physical quantities, geometric information and features specific to singularities (welds and edges). The richness of the fatigue database was demonstrated by conducting an unsupervised analysis helping to identify different types of zones characterized by different groups of variables. Even if this preliminary analysis did not help the identification of critical zones, it provided useful insights on the data set.
We then addressed the estimation of multiaxial fatigue criteria. A probabilistic Dang Van criterion was introduced, allowing to jointly estimate the material and dispersion parameters. This criterion was estimated on an auxiliary data set gathering numerical results and fatigue test data on welded elementary structures. However, this probabilistic criterion poorly generalizes to the identification of critical zones on real-scale and complex mechanical parts. Therefore, we directly relied on Stellantis data set to estimate fatigue criteria using supervised classification methods. These statistical fatigue criteria provide improved prediction performances compared

to Dang Van criterion. The main added value of these criteria compared to the standard ones is their ability to account for a larger number of features, which explains this significant gain in performance.

The supervised classification methods treat only the crack initiations as positive instances. However, a zone without detected crack initiation can still be critical. Indeed, due to the severity of the test, its duration and the randomness of crack initiation, a critical zone may remain undetected during a fatigue test and thus unlabeled. Conversely, a crack initiation is an asserted critical zone. Accounting for this asymmetric label noise in classification is the objective of PU learning. In some sense, PU learning can be seen as a censorship mechanism in a classification framework. While censorship usually affects the estimation of regression models (lifetime models in reliability for instance), PU learning deals with the estimation of classifiers under censorship among class observations. This problem was studied theoretically and practically.

From a theoretical point of view, we studied PU learning under the SAR assumption, meaning that the probability for a positive instance to be labeled (*i.e.* the propensity) depends on the covariates. We proved upper and lower risk bounds emphasizing how the convergence rate depends on the propensity (*i.e.* the amount of label noise). The results were illustrated through numerical experiments that support the theoretical convergence rates.
Overall, the theoretical results help understanding the difficulty of the estimation of a PU learning classifier. A low propensity results in fewer positive instances labeled and thus a more difficult task. Conversely, as the propensity tends to 1, the performances of PU learning tend to those of fully supervised classification. The role of propensity is thus crucial. In fatigue applications, augmenting the severity (or duration) of the test increases the probability for a critical zone to initiate and thus be labeled. Hence, the propensity is increased. Therefore, increasing the severity of fatigue tests can facilitate the estimation of PU learning fatigue criteria. However, in practice, the test severity cannot be augmented indefinitely: this would expose the mechanical part to other failure modes that are not part of fatigue design (plasticity). All in all, the theoretical results strengthen an already well established principle in the validation of resistance in fatigue design: the fatigue tests should be performed at a severity close to the resistance of the part. If the severity is too low, the component is unlikely to fail and the fatigue resistance might be impossible to estimate.

We finally developed a practical methodology to estimate a PU learning classifier adapted to the context of fatigue. This methodology relies on parametric models both on the fatigue criterion (classifier) and the propensity. The propensity accounts for the testing conditions through the calculation of an equivalent severity. The estimation of the parameters is carried out through maximum likelihood using the EM algorithm. The interest of the methodology is illustrated on simulated data. Then, the application of the method to the fatigue data set provides interesting results. The performances of PU learning are compared to standard classification and to Dang Van criterion. Compared to standard classifiers, PU learning better models the risk of crack initiation. When evaluating the models on the predicted classes, both non-traditional and PU classifier provide similar performances. However, the evaluation metrics may be biased due to the presence of PU label noise in the testing data. As for classical supervised methods, the performances of PU learning classifiers in the identification of critical zones are significantly higher than those of the traditional Dang Van criterion. Finally, the main interest of PU learning lies in the model that separates the effects of the fatigue criterion (PU classifier) and the testing conditions (propensity) that impact the observed labels.

In the scope of this thesis, we have developed new statistical tools for the identification of critical zones in numerical models: classifiers obtained through classical supervised classification

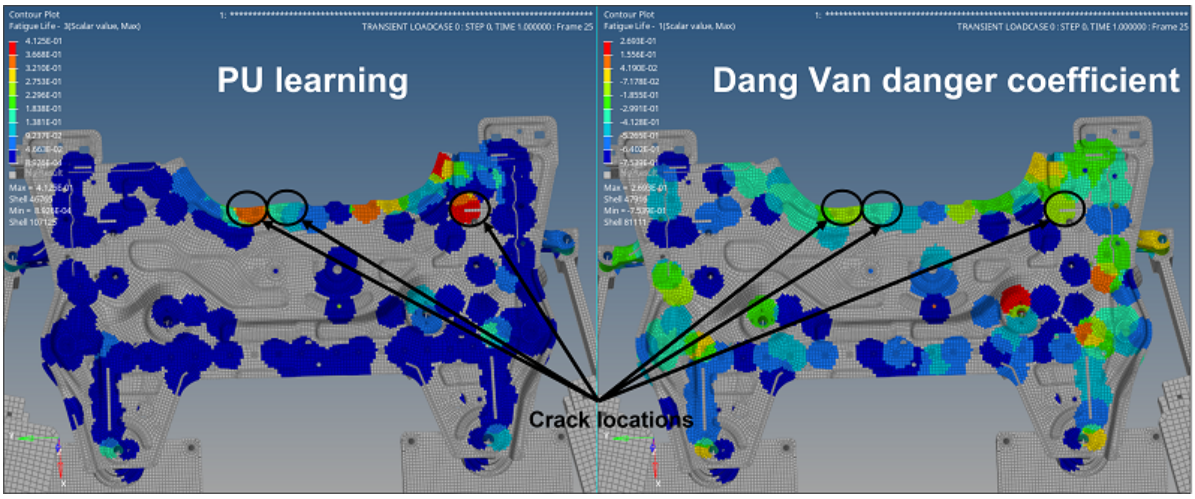

Figure A: PU learning predictions (left) and Dang Van criterion (right) on a FEM (cradle model under longitudinal loading): dangerous zones in red, safe zones in blue. The grey zones on the FEM were not selected during the pre-processing (construction of zones, cf. Subsection 2.2.2): they should be considered as blue. The crack initiation locations on this side of the part are highlighted as dark circles.

methods and PU classifiers that account for label noise specific to fatigue tests. Once the classifiers are estimated, they can be easily deployed in Finite Element software to provide predictions on different zones of a conception. As for the danger coefficient of Dang Van criterion, it is straightforward to evaluate the probability for each zone to be critical.

The deployment of the methodology is represented in Figure A. On this example of cradle model under longitudinal loading, the predictions of Dang Van criterion and those of PU learning are compared. We recall that the danger coefficient of a zone is defined as the maximum danger coefficient among the elements of the zone. It is first important to note that PU learning criterion highlights only a few critical zones, the majority remains blue (safe). Conversely, Dang Van danger coefficient is high (values close to the threshold 0 or above) for a large number of zone. Hence, following the predictions of Dang Van criterion would require many reinforcements on the part: most of these reinforcement may be unnecessary. Instead, PU learning has less false alarms and is more likely to guide the conception teams into focusing on the most critical locations (those that are more likely to fail during testing). Moreover, the known crack initiation locations are better identified by PU learning criterion than through Dang Van danger coefficient. Overall, these observations confirm the substantial gain in performances achieved by statistical fatigue criteria compared to the classical Dang Van criterion.

Perspectives

This work paved the way for the development and deployment of statistical fatigue criteria improving the numerical design of safety parts of vehicles. Several research directions remain open and could lead to significant improvements on the methodology and on its efficiency.

Extending the approach to other mechanical parts

In the scope of this thesis, only cradle and cross-member models were considered. These safety parts of the vehicles concentrate the majority of design issues due to their complexity (complex geometry, great number of welds...). However, fatigue design is not limited to these two types of mechanical parts. Therefore, the methodology developed in this thesis can be extended beyond these two types of mechanical parts and even beyond the scope of chassis components: for instance, applications to propulsion system (battery packs) could be considered. An additional challenge would be to mix volumetric FEM with shell models in the fatigue database and in the estimation of fatigue criteria. For now, only shell models are considered. Finally, another perspective would be to study the applicability of this methodology to vibration fatigue.

Improving the fatigue database

Initially composed of a few case studies, the database was consequently developed during the thesis, reaching about 40 case studies and 300 crack instances. However, the potential of available data is far more consequent. Increasing the size of the database would allow to enhance the diversity of geometries and loadings considered and thus strengthen the estimated statistical criteria.

From a practical point of view, the development of the database faces two main challenges. First, it relies on two different sources of data (FEM results and fatigue test reports). As the numerical calculations and the fatigue tests are not performed by the same teams, it is not always easy to gather *a posteriori* a FEM result with the corresponding fatigue test report. Besides, the construction of the fatigue database requires to manually label the crack initiation zones on the numerical model. This manual labeling process is sometimes complicated as we only have access to photos of crack initiations (through the fatigue test report) and it is not always straightforward to identify these critical locations on the numerical model.

In the future, it would be easier to do the labeling during the fatigue tests. Indeed, the teams in charge of validation tests could better locate crack initiation zones as they can see the prototype and do not have to rely on photos. Besides, as the FEM calculations are performed before testing, the numerical model could be transferred to the team in charge of conducting the fatigue tests so that they can report the tests outcome on the FEM. This would allow to develop continuously and efficiently the database.

Another axis of improvement regarding the database is the accuracy of FEM calculations. In this work (and in fatigue design in general), the fatigue criteria applied rely on physical quantities calculated through numerical models. The FEM results remain uncertain, especially on locations with complex geometry (holes, notches, welds...). A great part of the difficulties of fatigue models in predicting critical zones is due to the limits of FEM that do not account for several phenomena (manufacturing processes, residual stresses). As the improvement of FEM is an active topic in the automotive industry, we can expect the FEM to gain in accuracy in the near future. As a consequence, the fatigue database could benefit from higher quality data sets, which would also improve the characterization of fatigue phenomena using the methodologies developed in this thesis.

Better account for geometry in fatigue criteria

The fatigue criteria developed in this thesis rely on multiple features including geometric descriptors of zones (thickness, stresses expressed in a local coordinate system for singularities...). These geometric features were chosen at the beginning of the study (cf. Chapter 2). As future perspective, it would be interesting to investigate how these choices can be justified statistically and to explore other methods that could better account for the geometry.

First, we recall that the construction of zones relies on a few parameters empirically chosen in Section 2.2: a selection criteria to limit the number of zones and a radius delimiting the size of the zone. It would be interesting to study how the estimated models and performances depend on these hyper-parameters. This would help designing a methodology to select those hyper-parameters.

Second, among the 60 features introduced in Subsection 2.2.3, some of them characterize the most critical element of the zone. This element is identified as the one with the maximum Dang Van fatigue coefficient over the zone, considered as the element most likely to cause the crack initiation. This assumption could be relaxed by letting the machine learning model deal with the uncertainty concerning the precise location of crack initiation. In particular, *Multiple Instance Learning* (MIL, cf. Sabato and Tishby, 2012; Herrera et al., 2016) might provide a solution to this issue. MIL is a supervised classification method where only bags (groups) of observations are labeled. In the binary setting, a group of observations is labeled positive if at least one observation is positive; else, it is labeled negative. Hence, MIL seems to be well suited for the task as we only have labeled zones (groups of elements) and we do not know for sure which element(s) caused the crack initiation.

Finally, the fatigue criteria estimated in the scope of this thesis only account for a limited number of geometric features. As we have seen, some critical zones are still poorly identified by the criteria due to the limits and uncertainties in FEM. Accounting better for the geometry may help the identification of critical zones. Even if manufacturing process effects and residual stresses cannot be calculated, two zones with similar geometry and materials will have similar fatigue resistance. Therefore, considering the whole geometry of the zone (shape, curvature, ...) in the criterion can help improve the predictions. An interesting perspective would be to consider the zone as a three-dimensional image and use adapted deep learning methods to extract high level features and predict the risk of crack initiation. Of course, a three-dimensional image (voxels) may not be the most appropriate format: as the mechanical parts are planar structures, such a representation will raise sparsity issues. Alternatively, a zone can be easily converted into a point cloud by sampling points on the zone (cf. Fig. B). Another solution is to consider directly the raw data entry (meshing of the zone): a zone can thus be considered as a graph with vertices (elements) and edges (links between elements). Over the past few years, different neural network architectures have been developed to address this kind of tasks. For instance, *PointNet* is a neural network architecture designed to solve tasks considering point clouds as entry data (cf. Qi et al., 2017). Similarly, *MeshCNN* is designed for predictions based on meshing data (Hanocka et al., 2019). These architectures are part of an emerging framework in deep learning called *Geometric Deep Learning* that aims at developing neural network architectures for solving tasks involving specific data structures like point clouds, graphs, meshes and manifolds (Bronstein et al., 2017). These methods could help predict fatigue cracks and could also serve for other applications involving the analysis of FEM.

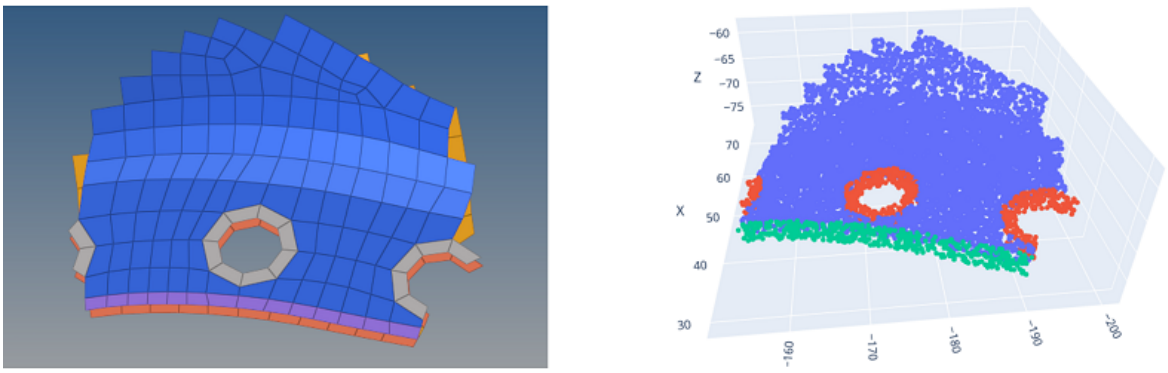

Figure B: Example of zone on FEM (left) and its conversion as a point cloud (right).

Perspectives on PU learning

One of the main contributions of this thesis is the study of PU learning under the SAR assumption and the development of a practical methodology to estimate fatigue criteria. PU learning remains an active research topic and some problems remain open.

A natural extension to the theoretical results proved in Chapter 3 would be to study PU learning with surrogate loss functions. Indeed, this work focus on an empirical risk based on a $0-1$ loss function which is interesting to study theoretical properties of PU learning. However, in practice, the optimization of such a risk is numerically intractable. A standard solution to overcome this difficulty is to consider convex loss functions for which the optimization is numerically tractable. It would thus be interesting to extend the theoretical guarantees to these loss functions. This problem was already studied for standard binary classification setting (cf. Bartlett et al., 2006; Blanchard et al., 2008). Besides, Plessis et al. (2014) proved risk bounds under the SCAR assumption using surrogate loss functions, highlighting a parametric convergence rate. In the thesis, numerical experiments were carried out under the SAR assumption using a logistic loss, which also exhibits a parametric convergence rate affected by the propensity (cf. Section 3.6).

The risk bounds of Chapter 3 depend on the propensity through its minimum which is assumed to be strictly positive. This assumption on the propensity remains strong and it would be interesting to investigate if it could be relaxed. For instance, one could only assume that the propensity can take values arbitrarily close to 0 as far as this occurs with small probability. The simulations of Section 3.6 cover the case of a logistic propensity clipped to remain above a minimum value $e_m > 0$. However, as e_m tends to 0 the excess risk stabilizes meaning that we still have a similar convergence rate even if the propensity can take arbitrarily small values.

The PU learning methodology developed in this manuscript also raises additional challenges. First, variable selection remains a crucial problem especially if the number of features is of the same order as the number of labeled instances. This challenge is even double as there may be variable selection at the same time in the classification model and in the propensity.

Then, another interesting direction would be to consider other classification models as PU classifier (SVM, Random Forest...) and adapt the methodology to these models. In particular, we have seen that Random Forest achieved better results than Logistic Regression on the fatigue data set in non-traditional classification (cf. Section 3.1). Therefore, we can expect better classification performances by leveraging these methods in the PU learning framework.

Finally, an alternative approach to PU learning is Neyman-Pearson classification. Blanchard et al. (2010) developed a PU learning methodology based on this framework under the SCAR

assumption. The interest of Neyman-Pearson classification is that it aims at minimizing the second type risk (False Positive Rate) given a control on the first type risk (False Negative Rate). This is particularly interesting in fatigue design application as we do not give the same importance to both types of risks: the first type risk is more important (proportion of critical zones that are not identified) than the second type (proportion of safe zones predicted as critical). Therefore, it would be interesting to study the potential extension of Neyman-Pearson classification to PU learning under the SAR assumption.

Bibliography

Ballard, P., Van, K. D., Deperrois, A., and Papadopoulos, Y. (1995). High cycle fatigue and a finite element analysis. *Fatigue & Fracture of Engineering Materials & Structures*, 18(3):397–411.

Baroux, E., Delattre, B., Constantinescu, A., Pamphile, P., and Raoult, I. (2022). Analysis of real-life multi-input loading histories for the reliable design of vehicle chassis. *Procedia Structural Integrity*, 38:497–506.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Basquin, O. (1910). The exponential law of endurance tests. In *Proc Am Soc Test Mater*, volume 10, pages 625–630.

Bastenaire, F. (1972). New method for the statistical evaluation of constant stress amplitude fatigue-test results. In *Probabilistic aspects of fatigue*. ASTM International.

Bathias, C. and Pineau, A. (2010). *Fatigue of materials and structures*. Wiley Online Library.

Beaumont, P. (2013). *Optimisation des plans d'essais accélérés Application à la tenue en fatigue de pièces métalliques de liaison au sol*. PhD thesis, Université d'Angers.

Beaumont, P., Guérin, F., Lantieri, P., Facchinetti, M. L., and Borret, G. M. (2012). Accelerated fatigue test for automotive chassis parts design: An overview. In *2012 Proceedings Annual Reliability and Maintainability Symposium*, pages 1–6. IEEE.

Bekker, J. and Davis, J. (2018a). Estimating the class prior in positive and unlabeled data through decision tree induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Bekker, J. and Davis, J. (2018b). Learning from positive and unlabeled data under the selected at random assumption. In *Proceedings of The Learning with Imbalanced domains: Theory and Application Workshop @ ECML 2018*, volume 94, pages 8–22. Journal of Machine Learning Research.

Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4):719–760.

Bekker, J., Robberechts, P., and Davis, J. (2020). Beyond the Selected Completely at Random Assumption for Learning from Positive and Unlabeled Data. In Brefeld, U., Fromont,

E., Hotho, A., Knobbe, A., Maathuis, M., and Robardet, C., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 11907, pages 71–85. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Bellec, E., Facchinetti, M., Doudard, C., Calloch, S., and Moyne, S. (2022). Multiaxial variable amplitude loading for automotive parts fatigue life assessment: A loading classification-based approach proposal. *Procedia Structural Integrity*, 38:202–211.

Bergamo, S., Schimmerling, P., Triboulet, F., Wilson, P., Facchinetti, M. L., Monin, M., F., L., and Weber, B. (2017). Préconisations pour les caractéristiques statistiques de résistance en fatigue. *SIA*.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Blanchard, G., Bousquet, O., and Massart, P. (2008). Statistical performance of support vector machines. *The Annals of Statistics*, 36(2):489–531.

Blanchard, G., Lee, G., and Scott, C. (2010). Semi-Supervised Novelty Detection. *Journal of Machine Learning Research*, 11(99):2973–3009.

Bousquet, O., Boucheron, S., and Lugosi, G. (2003). Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brevet, P., François, D., Gourmelon, J.-P., and Raharinaivo, A. (1978). *FATIGUE DES OUVRAGES D'ART METALLIQUES SOUDES-RAPPORT INTRODUCTIF A UN PROGRAMME DE RECHERCHE*. RAPP RECH LPC.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

Cannings, T. I., Fan, Y., and Samworth, R. J. (2020). Classification with imperfect training labels. *Biometrika*, 107(2):311–330.

Castillo, E. and Fernández-Canteli, A. (2009). *A unified statistical methodology for modeling fatigue damage*. Springer Science & Business Media.

Cazenave, M. (2013). *Méthode des éléments finis-2e éd.: Approche pratique en mécanique des structures*. Dunod.

Chen, X., Chen, W., Chen, T., Yuan, Y., Gong, C., Chen, K., and Wang, Z. (2020). Self-PU: Self boosted and calibrated positive-unlabeled training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1510–1519.

Chiaroni, F., Rahal, M.-C., Hueber, N., and Dufaux, F. (2018). Learning with A Generative Adversarial Network From a Positive Unlabeled Dataset for Image Classification. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1368–1372. ISSN: 2381-8549.

Claesen, M., De Smet, F., Suykens, J. A., and De Moor, B. (2015). A robust ensemble approach to learn from positive and unlabeled data using svm base models. *Neurocomputing*, 160:73–84.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2022). *Introduction to algorithms*. MIT press.

Correia, J., Apetre, N., Arcari, A., De Jesus, A., Muñiz-Calvente, M., Calçada, R., Berto, F., and Fernández-Canteli, A. (2017). Generalized probabilistic model allowing for various fatigue damage variables. *International Journal of Fatigue*, 100:187–194.

Coudray, O., Bristiel, P., Dinis, M., Keribin, C., and Pamphile, P. (2020a). Caractérisation de zones critiques pour le dimensionnement en fatigue d'une pièce mécanique. In *E-congrès 2020 Lambda λμ22-22e Congrès de Maîtrise des Risques et Sûreté de Fonctionnement λμ22*.

Coudray, O., Bristiel, P., Dinis, M., Keribin, C., and Pamphile, P. (2021). Fatigue data-based design: statistical methods for the identification of critical zones. In *SIA Simulation Numérique*.

Coudray, O., Keribin, C., Massart, P., and Pamphile, P. (2022a). Risk bounds for pu learning under selected at random assumption. *arXiv preprint arXiv:2201.06277*.

Coudray, O., Keribin, C., and Pamphile, P. (2022b). Convergence rates for positive-unlabeled learning under selected at random assumption: sensitivity analysis with respect to propensity. In *Conférence sur l'Apprentissage automatique*.

Coudray, O., Keribin, C., Pamphile, P., Dinis, M., and Bristiel, P. (2020b). Caractérisation de zones critiques pour le dimensionnement en fatigue d'une pièce mécanique. In *SFdS2020-52èmes Journées de Statistiques de la Société Française de Statistique*.

Crossland, B. (1956). Effect of large hydrostatic pressures on the torsional fatigue strength of an alloy steel. In *Proc. Int. Conf. on Fatigue of Metals*, volume 138, pages 12–12. Institution of Mechanical Engineers London.

Dang Van, K. and Griveau, B. (1989). On a new multiaxial fatigue limit criterion- theory and application. *Biaxial and multiaxial fatigue(A 90-16739 05-39). London, Mechanical Engineering Publications, Ltd., 1989,*, pages 479–496.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dixon, W. J. and Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association*, 43(241):109–126.

Dowling, N. E. (2004). Mean stress effects in stress-life and strain-life fatigue. *SAE Technical Paper*, 32(12):1004–1019.

Echard, B., Gayton, N., and Bignonnet, A. (2014). A reliability analysis method for fatigue design. *International journal of fatigue*, 59:292–300.

Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 213, Las Vegas, Nevada, USA. ACM Press.

Escofier, B. and Pagès, J. (1998). *Analyses factorielles simples et multiples*. Dunod, Paris.

Fares, Y. (2006). *Dimensionnement en fatigue des assemblages boulonnés à l'aide de critères de fatigue multiaxiale*. PhD thesis, Institut National des Sciences Appliquées de Toulouse.

Fayard, J.-L. (1996). *Dimensionnement à la fatigue polycyclique de structures soudées*. PhD thesis, Ecole Polytechnique.

Ferretti, E., Errecalde, M. L., Anderka, M., and Stein, B. (2014). On the Use of Reliable-Negatives Selection Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia. In *2014 25th International Workshop on Database and Expert Systems Applications*, pages 211–215. ISSN: 2378-3915.

Florin, P. (2015). *Caractérisation rapide des propriétés à la fatigue à grand nombre de cycle des assemblages métalliques soudés de type automobile: vers une nouvelle approche basée sur des mesures thermométriques*. PhD thesis, Brest.

Fogue, M. and Bahuaud, J. (1985). Fatigue multiaxiale a durée de vie illimitée. In *7eme Congres Franais de Mécanique, Bordeaux*, pages 30–31.

Fortunier, R. (2001). Comportement mécanique des matériaux. *cours, ENS des Mines de Saint-Etienne*, page 214.

Fouchereau, R. (2014). *Modélisation probabiliste des courbes SN*. PhD thesis, Paris 11.

Fusilier, D. H., Montes-y-Gómez, M., Rosso, P., and Cabrera, R. G. (2015). Detecting positive and negative deceptive opinions using pu-learning. *Information processing and management*, 51(4):433–443.

Garg, S., Wu, Y., Smola, A. J., Balakrishnan, S., and Lipton, Z. (2021). Mixture Proportion Estimation and PU Learning:A Modern Approach. In *Advances in Neural Information Processing Systems*, volume 34, pages 8532–8544. Curran Associates, Inc.

Godefroid, L. B., Faria, G. L. d., Cândido, L. C., and Araujo, S. (2014). Fatigue failure of a welded automotive component. *Procedia materials science*, 3:1902–1907.

Gong, C., Wang, Q., Liu, T., Han, B., You, J. J., Yang, J., and Tao, D. (2021). Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gough, H. J., Pollard, H., and Clenshaw, W. (1951). Some experiments on the resistance of metals to fatigue under combined stresses. *Reports and Memoranda*.

Govaert, G. and Nadif, M. (2008). Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245.

Govaert, G. and Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons.

Hanocka, R., Hertz, A., Fish, N., Giryes, R., Fleishman, S., and Cohen-Or, D. (2019). Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

He, D., Pan, M., Hong, K., Cheng, Y., Chan, S., Liu, X., and Guizani, N. (2020). Fake Review Detection Based on PU Learning and Behavior Density. *IEEE Network*.

He, F., Liu, T., Webb, G. I., and Tao, D. (2018). Instance-dependent pu learning by bayesian optimal relabeling. *arXiv:1808.02180*.

Herrera, F., Ventura, S., Bello, R., Cornelis, C., Zafra, A., Sánchez-Tarragó, D., and Vluymans, S. (2016). Multiple instance learning. In *Multiple instance learning*, pages 17–33. Springer.

Hou, M., Chaib-Draa, B., Li, C., and Zhao, Q. (2017). Generative adversarial positive-unlabelled learning. *arXiv preprint arXiv:1711.08054*.

Hwanjo Yu, Jiawei Han, and Chang, K. (2004). PEBL: web page classification without negative examples. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):70–81.

Jain, S., White, M., and Radivojac, P. (2016). Estimating the class prior and posterior from noisy positives and unlabeled data. *Advances in neural information processing systems*, 29:2693–2701.

Jiang, Y., Haihong, E., Song, M., and Zhang, K. (2018). Research and Application of Newborn Defects Prediction Based on Spark and PU-learning. *5th IEEE International Conference on Cloud Computing and Intelligence Systems*, pages 657–663.

Jimenez-Martinez, M. (2020). Manufacturing effects on fatigue strength. *Engineering Failure Analysis*, 108:104339.

Keribin, C., Celeux, G., and Robert, V. (2017). The latent block model: a useful model for high dimensional data. In *ISI 2017-61st world statistics congress*, pages 1–6.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. (2017). Positive-Unlabeled Learning with Non-Negative Risk Estimator. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Leger, J.-B. (2016). Blockmodels: A r-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates. *arXiv preprint arXiv:1602.07587*.

Lemaignan, C. (2012). *La rupture des matériaux*. EDP sciences.

Li, H., Chen, Z., Liu, B., Wei, X., and Shao, J. (2014). Spotting Fake Reviews via Collective Positive-Unlabeled Learning. In *2014 IEEE International Conference on Data Mining*, pages 899–904, Shenzhen, China. IEEE.

Li, X. and Liu, B. (2003). Learning to classify texts using positive and unlabeled data. In *IJCAI*, volume 3, pages 587–592.

Li, X.-L., Liu, B., and Ng, S. K. (2010). Negative training data can be harmful to text classification. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 218–228.

Lin, S.-K., Lee, Y.-L., and Lu, M.-W. (2001). Evaluation of the staircase and the accelerated test methods for fatigue limit distributions. *International journal of fatigue*, 23(1):75–83.

Liu, B., Dai, Y., Li, X., Lee, W., and Yu, P. (2003). Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pages 179–186.

Liu, B., Lee, W. S., Yu, P. S., and Li, X. (2002). Partially supervised classification of text documents. In *ICML*, volume 2, pages 387–394. Sydney, NSW.

Locati, L. (1955). Le prove di fatica come ausilio alla progettazione ed alla produzione. *La Metallurgia Italiana*, 9:301.

Lomet, A. (2012). *Sélection de modèle pour la classification croisée de données continues*. PhD thesis, Compiègne.

Lomet, A., Govaert, G., et al. (2012). Model selection in block clustering by the integrated classification likelihood. In *20th International Conference on Computational Statistics (COMPSTAT 2012)*, pages 519–530.

Lugosi, G. (2002). Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer.

Luo, Y., Cheng, S., Liu, C., and Jiang, F. (2018). PU Learning in Payload-based Web Anomaly Detection. *Third International Conference on Security of Smart Cities, Industrial Control System and Communications*, pages 1–5.

Mainnemare, F. (2021). *Modélisation du point de soudure électrique pour la tenue en service des structures automobiles*. PhD thesis, université Paris-Saclay.

Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *Annals of Statistics*, 34(5).

Miner, M. A. (1945). Cumulative damage in fatigue. *Journal of Applied Mechanics*.

Mordelet, F. and Vert, J.-P. (2014). A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209.

Na, B., Kim, H., Song, K., Joo, W., Kim, Y.-Y., and Moon, I.-C. (2020). Deep generative positive-unlabeled learning under selection bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1155–1164.

Nadjitonon, N. (2010). *Contribution à la modélisation de l'endommagement par fatigue*. PhD thesis, Université Blaise Pascal-Clermont-Ferrand II.

Nikdelfaz, O. and Jalili, S. (2018). Disease genes prediction by HMM based PU-learning using gene expression profiles. *J. Biomed. Inf.*, 81:102–111.

Palmgren, A. (1924). Die lebensdauer von kugellargern. *Zeitshrift des Vereines Duetsher Ingenieure*, 68(4):339.

Plessis, M., Niu, G., and Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. *Advances in Neural Information Processing Systems*, 1:703–711.

Plessis, M., Niu, G., and Sugiyama, M. (2016). Class-prior Estimation for Learning from Positive and Unlabeled Data. In *Asian Conference on Machine Learning*, pages 221–236. PMLR. ISSN: 1938-7228.

Plessis, M. D., Niu, G., and Sugiyama, M. (2015). Convex Formulation for Learning from Positive and Unlabeled Data. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1386–1394. PMLR. ISSN: 1938-7228.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660.

Roux, C., Lorang, X., Maitournam, H., and Nguyen-Tajan, M. (2014). Fatigue design of railway wheels: a probabilistic approach. *Fatigue & Fracture of Engineering Materials & Structures*, 37(10):1136–1145.

Sabato, S. and Tishby, N. (2012). Multi-instance learning with any hypothesis class. *The Journal of Machine Learning Research*, 13(1):2999–3039.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

Schijve, J. (2005). Statistical distribution functions and fatigue of structures. *international Journal of Fatigue*, 27(9):1031–1039.

Schijve, J. (2009). *Fatigue of structures and materials*. Springer.

Sghaier, R. B., Bouraoui, C., Fathallah, R., Hassine, T., and Dogui, A. (2007). Probabilistic high cycle fatigue behaviour prediction based on global approach criteria. *International journal of fatigue*, 29(2):209–221.

Sines, G. and Ohgi, G. (1981). Fatigue criteria under combined stresses or strains. *Journal of Engineering Materials and Technology*.

Stromeyer, C. (1914). The determination of fatigue limits under alternating stress conditions. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 90(620):411–425.

Sun, C., Zhang, X., Liu, X., and Hong, Y. (2016). Effects of specimen size on fatigue life of metallic materials in high-cycle and very-high-cycle fatigue regimes. *Fatigue & Fracture of Engineering Materials & Structures*, 39(6):770–779.

Susmel, L. and Lazzarin, P. (2002). A bi-parametric wöhler curve for high cycle multiaxial fatigue assessment. *Fatigue & Fracture of Engineering Materials & Structures*, 25(1):63–78.

Thomas, J.-j., Nguyen-Tajan, T., and Burry, P. (2005). Structural durability in automotive design. *Mat.-wiss. u. Werkstofftech*, 36(11).

Tryla, H. (2022). *Évolution des contraintes résiduelles des assemblages soudés sous sollicitations mécaniques-Étude et modélisation*. PhD thesis, HESAM Université.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Weber, B. (1999). *Fatigue multiaxiale des structures industrielles sous chargement quelconque*. PhD thesis, Lyon, INSA.

Wormsen, A. and Härkegård, G. (2004). A statistical investigation of fatigue behaviour according to weibull's weakest link theory. *ECF15*.

Yang, P., Li, X., Chua, H.-N., Kwoh, C.-K., and Ng, S.-K. (2014). Ensemble Positive Unlabeled Learning for Disease Gene Identification. *PLOS ONE*, 9(5):e97079. Publisher: Public Library of Science.

Yang, P., Li, X.-L., Mei, J.-P., Kwoh, C.-K., and Ng, S.-K. (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647.

Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer.

Zhao, Y. and Yang, B. (2008). Probabilistic measurements of the fatigue limit data from a small sampling up-and-down test method. *International Journal of Fatigue*, 30(12):2094–2103.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Un point de vue statistique sur les critères de fatigue : de la classification supervisée à l'apprentissage positif-non labellisé

Mots clés: Critère de fatigue, Classification, Bruit d'étiquetage, Apprentissage positif-non labellisé, Bornes de risque.

Résumé: La fiabilité des véhicules est un enjeu majeur pour les constructeurs automobiles. En particulier, la fatigue mécanique est une préoccupation importante du bureau d'études. En effet, la fatigue est un phénomène complexe qui dépend du design de la pièce (géométrie, matériaux utilisés), des procédés de fabrication, et des chargements externes subis par la pièce. Le dimensionnement à la fatigue repose sur une modélisation numérique de la pièce et sur l'application de critères de fatigue déterministes afin d'identifier de potentielles faiblesses sur la conception. Ces critères, bien qu'efficaces sur des géométries simples, ne suffisent pas à prédire correctement les risques d'amorçage sur des composants complexes. Cela entraîne un allongement des temps de développement et une augmentation des coûts liés aux prototypes physiques. Pour y remédier, les constructeurs automobiles recherchent de nouvelles méthodes digitales, pour mieux identifier les zones critiques sur de nouvelles conceptions.

Dans cette thèse, nous construisons une base de données fatigue, à partir d'informations mises à disposition par Stellantis, regroupant des résultats numériques et des comptes rendus d'essais de fatigue. Une analyse non supervisée du jeu de données est réalisée, permettant de mieux comprendre sa structure ainsi que les liens entre les covariables disponibles. Ensuite, l'application de méthodes d'apprentissage supervisé (régression logistique, forêts aléatoires, SVM à noyau...) permet d'estimer des critères de fatigue offrant de meilleures prédictions que le critère mécanique déterministe usuel. Une difficulté de l'analyse provient du fait que l'étiquetage des zones est affecté par un bruit asymétrique, ce qui motive une approche originale fondée sur l'apprentissage positif-non labellisé (PU learning). Cette approche est abordée suivant tous les angles: théorique, méthodologique et appliqué. De nouvelles bornes de risques adaptées à ce cadre spécifique sont démontrées. Une méthodologie est proposée pour l'estimation d'un classifieur PU à partir des données. Enfin, la méthodologie est évaluée sur des jeux de données simulés ainsi que sur les données de fatigue. Les performances obtenues confirment l'intérêt de la méthode et son utilité pour le constructeur automobile.

Title: A statistical point of view on fatigue criteria: from supervised classification to positive-unlabeled learning

Keywords: Fatigue criterion, Classification, Label noise, PU learning, Risk bounds.

Abstract: The reliability of vehicles is a major issue for automotive manufacturers. In particular, mechanical fatigue is an important preoccupation of the design office. Indeed, fatigue is a complex phenomenon that depends on the design of the part (geometry, materials used), the manufacturing and on the external loads it is subjected to. In order to design a safety part against fatigue, the part is numerically modeled and a deterministic fatigue criterion is applied to identify potential weaknesses. If these criteria prove to be effective when evaluated on experimental test data with standardized specimens, they are less effective for rig tests with prototypes. This results in an increase in development costs and duration. In order to remedy this issue, car manufacturers seek new digital tools to better predict the fatigue risks on new design proposals.

In this thesis, we build a fatigue database, based on information provided by Stellantis, gathering numerical results along with fatigue test reports on prototypes. Unsupervised machine learning methods are applied offering a better understanding of the structure of the database and the relations between the available features. Then, the application of supervised machine learning methods (logistic regression, random forests, kernel SVM...) allows to estimate fatigue criteria offering better predictions than the standard fatigue criterion. However, the binary labels in this classification task are affected by a completely asymmetric label noise. This motivates an original approach to fatigue criteria estimation based on Positive-Unlabeled learning (PU learning). This problem is studied from all angles: theory, methodology and application. First, new risk bounds, adapted to this specific framework, are proved. Then, we develop a practical methodology to estimate a PU classifier. Finally, the methodology is evaluated on simulated data and on the fatigue database. The prediction performances confirm the interest of the methodology and its utility for car manufacturers.