Tests d'association génétique pour des durées de vie en grappes

Thèse

Martin Leclerc

Doctorat en mathématiques
Philosophiæ doctor (Ph.D.)

Québec, Canada

Résumé

Les outils statistiques développés dans cette thèse par articles visent à détecter de nouvelles associations entre des variants génétiques et des données de survie en grappes. Le développement méthodologique en analyse des durées de vie est aujourd'hui ininterrompu avec la prolifération des tests d'association génétique et, de façon ultime, de la médecine personnalisée qui est centrée sur la prévention de la maladie et la prolongation de la vie.

Dans le premier article, le problème suivant est traité : tester l'égalité de fonctions de survie en présence d'un biais de sélection et de corrélation intra-grappe lorsque l'hypothèse des risques proportionnels n'est pas valide. Le nouveau test est basé sur une statistique de type Cramér-von Mises. La valeur de p est estimée en utilisant une procédure novatrice de bootstrap semi-paramétrique qui implique de générer des observations corrélées selon un devis non-aléatoire. Pour des scénarios de simulations présentant un écart vis-à-vis l'hypothèse nulle avec courbes de survie qui se croisent, la statistique de Cramer-von Mises offre de meilleurs résultats que la statistique de Wald du modèle de Cox à risques proportionnels pondéré. Le nouveau test a été utilisé pour analyser l'association entre un polymorphisme nucléotidique (SNP) candidat et le risque de cancer du sein chez des femmes porteuses d'une mutation sur le gène suppresseur de tumeur BRCA2.

Un test d'association *sequence kernel* (SKAT) pour détecter l'association entre un ensemble de SNPs et des durées de vie en grappes provenant d'études familiales a été développé dans le deuxième article. La statistique de test proposée utilise la matrice de parenté de l'échantillon pour modéliser la corrélation intra-famille résiduelle entre les durées de vie via une copule gaussienne. La procédure de test fait appel à l'imputation multiple pour estimer la contribution des variables réponses de survie censurées à la statistique du score, laquelle est un mélange de distributions du khi-carré. Les résultats de simulations indiquent que le nouveau test du score de type noyau ajusté pour la parenté contrôle de façon adéquate le risque d'erreur de type I. Le nouveau test a été appliqué à un ensemble de SNPs du locus *TERT*.

Le troisième article vise à présenter le progiciel R *gyriq*, lequel implante une version bonifiée du test d'association génétique développé dans le deuxième article. La matrice noyau *identical-by-state* (IBS) pondérée a été ajoutée, les tests d'association génétique actuellement disponibles pour des variables réponses d'âge d'apparition ont été brièvement revus de pair

avec les logiciels les accompagnant, l'implantation du progiciel a été décrite et illustrée par des exemples.

Abstract

The statistical tools developed in this manuscript-based thesis aim at detecting new associations between genetic variants and clustered survival data. Methodological development in lifetime data analysis is today ongoing with the proliferation of genetic association testing and, ultimately, personalized medicine which focuses on preventing disease and prolonging life.

In the first paper, the following problem is considered: testing the equality of survival functions in the presence of selection bias and intracluster correlation when the assumption of proportional hazards does not hold. The new proposed test is based on a Cramér-von Mises type statistic. The p-value is approximated using an innovative semiparametric bootstrap procedure which implies generating correlated observations according to a non-random design. For simulation scenarios of departures from the null hypothesis with crossing survival curves, the Cramer-von Mises statistic clearly outperformed the Wald statistic from the weighted Cox proportional hazards model. The new test was used to analyse the association between a candidate single nucleotide polymorphism (SNP) and breast cancer risk in women carrying a mutation in the BRCA2 tumor suppressor gene.

A sequence kernel association test (SKAT) to detect the association between a set of genetic variants and clustered survival outcomes from family studies is developed in the second manuscript. The proposed statistic uses the kinship matrix of the sample to model the residual intra-family correlation between survival outcomes via a Gaussian copula. The test procedure relies on multiple imputation to estimate the contribution of the censored survival outcomes to the score statistic which is a mixture of chi-square distributions. Simulation results show that the new kinship-adjusted kernel score test controls adequately for the type I error rate. The new test was applied to a set of SNPs from the *TERT* locus.

The third manuscript aims at presenting the R package *gyriq* which implements an enhanced version of the genetic association test developed in the second manuscript. The weighted identical-by-state (IBS) kernel matrix is added, genetic association tests and accompanying software currently available for age-at-onset outcomes are briefly reviewed, the implementation of the package is described, and illustrated through examples.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Mes premiers remerciements vont à mon directeur de thèse Lajmi Lakhal Chaieb pour son engagement à développer mon autonomie de chercheur, pour m'avoir aidé à "penser en dehors de la boîte" et pour son soutien constant. Je remercie mon codirecteur Jacques Simard : pour sa patience, son appui à mes activités de recherche à l'international et son optimisme de tous les instants. Je remercie Antonis Antoniou et son équipe du *Strangeways Research Laboratory* à Cambridge pour leur accueil chaleureux lors de mes deux visites sur place, leur disponibilité et leur sens du partage.

Pendant ce doctorat, j'ai été boursier du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et du Fonds de recherche du Québec - Nature et technologies (FRQNT). J'ai aussi bénéficié de fonds de recherche du Ministère de l'Économie, de l'Innovation et des Exportations du Québec. Je remercie mes concitoyens pour leur investissement envers la recherche universitaire et leur contribution à ma formation.

Les examinateurs externe et internes de cette thèse méritent des remerciements particuliers pour le temps qu'ils ont consacré à lire et commenter le présent document et pour leur présence lors de ma soutenance. D'abord, je veux exprimer ma reconnaissance à Laurent Briollais (Institut de Recherche Lunenfeld-Tanenbaum de l'hôpital Mount Sinai affilié à l'Université de Toronto). Ses commentaires et suggestions, allant du très général au très particulier, ont enrichi mon expérience de doctorant. Je remercie aussi Alexandre Bureau (Département de médecine sociale et préventive) et Thierry Duchesne (Département de mathématiques et de statistique) pour leur regard critique vis-à-vis les aspects de génétique et de durées de vie dans cette thèse incluant la littérature citée de même que pour leurs enseignements à l'hiver et l'automne 2012 respectivement.

À l'hiver 2015, j'ai eu le plaisir d'être chargé d'enseignement pour la première fois. Je remercie Anne-Sophie Charest pour la qualité de son encadrement. Mon travail au Service de consultation statistique du Département en 2014 a été une valeur ajoutée importante à mes études. Je suis reconnaissant envers Hélène Crépeau, Gaétan Daigle et David Émond pour leur professionnalisme et la grande qualité du soutien qu'ils m'ont offert. Mon expérience d'auxiliaire d'enseignement pour deux cours de statistique mathématique enseignés par Claude Bélisle à l'automne 2012 et 2013 a été des plus enrichissantes : je le remercie pour sa confiance et son

implication à faire de moi un meilleur statisticien.

Je remercie mes parents et mes grand-mères : leur amour et leur résilience face à l'adversité ont été une source d'inspiration importante pendant mon doctorat.

Avant-propos

Cette thèse comporte 3 articles présentés aux chapitres 2 à 4. Je suis premier auteur pour chacun d'eux alors que mon directeur de thèse Dr Lajmi Lakhal Chaieb est dernier auteur. Mon codirecteur Dr Jacques Simard est avant-dernier auteur sur chacun des articles. Le Dr Antonis Antoniou figure parmi la liste des auteurs du premier article à titre de collaborateur. Les consortiums ayant rendu disponibles les jeux de données réelles complètent la liste des auteurs pour les deux premiers articles :

— 1er article : *EMBRACE Investigators, GEMO Study Collaborators, INHERIT Investigators*

— 2e article : *Consortium of Investigators of Modifiers of BRCA1/2*

Le 1er article a été publié en ligne par le *Journal of the Royal Statistical Society : Series C (Applied Statistics)* en décembre 2014. Pour cet article, j'ai participé à l'élaboration de la problématique et du développement méthodologique, réalisé les simulations, effectué l'étude du jeu de données réelles et contribué à la discussion. J'ai participé à l'écriture de l'ensemble de l'article de même qu'au processus de révision.

Le 2e article a été publié en ligne par *Genetic Epidemiology* en août 2015. Pour cet article, j'ai proposé la problématique, contribué au développement méthodologique, conçu et réalisé les simulations, effectué l'étude du jeu de données réelles et élaboré la discussion. J'ai écrit la première version de l'article et mené le processus de révision.

Le 3e article est en préparation et sera soumis pour publication au *R Journal*. Pour cet article, j'ai conçu le *package* R, contribué au développement méthodologique et écrit la première version de l'article.

Introduction

Dans cette thèse, une durée de vie T est une variable aléatoire continue à valeurs dans l'intervalle $[0, \infty)$. Un exemple de durée de vie dont il est souvent utile d'en décrire les caractéristiques aléatoires est l'âge au moment du décès d'un individu. Au sens statistique, la notion de durée de vie peut s'appliquer à tout intervalle de temps entre deux événements :

— Temps écoulé jusqu'à une rechute chez un patient ayant subi une greffe de moelle osseuse pour le traitement de la leucémie ;

— Temps nécessaire pour obtenir la floraison d'une plante ;

— Temps écoulé entre deux inséminations successives chez une vache.

Les fonctions les plus souvent utilisées pour décrire le comportement aléatoire de T sont la fonction de survie $S(t)$ et la fonction de risque $\lambda(t)$:

$$S(t) = 1 - F(t) = \mathbb{P}(T > t), \quad t \in [0, \infty)$$

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}, \quad t \in [0, \infty)$$

Il suffit de préciser l'une ou l'autre de ces fonctions pour définir complètement T, en vertu de l'égalité suivante :

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{\frac{1}{\Delta t}\mathbb{P}(t \leq T \leq t + \Delta t)}{\mathbb{P}(T \geq t)} \\ &= \frac{f(t)}{S(t)} \\ &= -\frac{d\{\log[S(t)]\}}{dt} \end{aligned}$$

La fonction $\lambda(t)$ peut être vue comme le risque instantané que l'événement survienne sachant qu'il n'est pas déjà survenu. Dans les exemples qui précèdent, l'événement est le décès, la

rechute, la floraison ou l'insémination. Une autre fonction utile associée à T est le risque cumulé $\Lambda(t)$:

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log[S(t)]$$

Il est possible d'étudier T en utilisant un modèle entièrement paramétrique, par exemple la loi exponentielle : $S(t) = e^{-\mu t}, \mu > 0$. Cependant, les approches semi-paramétriques et non-paramétriques correspondent mieux à plusieurs problèmes rencontrés en pratique.

La censure

Une caractéristique fondamentale empêche le statisticien de traiter la durée de vie au même titre que n'importe quelle variable aléatoire continue : c'est la notion de censure. En pratique, plusieurs circonstances peuvent rendre impossible la mesure du temps écoulé jusqu'à la survenue de l'événement à l'étude chez un individu. Voici quelques exemples :

1. L'étude a pris fin avant que l'événement ne survienne ;
2. Le sujet a été perdu de vue en cours d'étude ;
3. Un événement concurrent est survenu avant l'événement à l'étude.

En ce qui concerne le point 3, il peut s'agir, par exemple, d'une personne décédée dans l'incendie d'un immeuble quand la durée de vie étudiée est l'âge au moment du développement du virus du SIDA. Dans tous les cas, la mesure de la durée de vie s'en trouve censurée. De façon générale, la censure, dénotée C, peut également être vue comme une variable aléatoire continue à valeurs dans l'intervalle $[0, \infty)$. Pour les 3 exemples mentionnés ci-dessus, il s'agit de censure à droite : $C < T$. C'est la forme de censure rencontrée le plus souvent en biostatistique. Les censures à gauche et par intervalle sont également possibles, mais ne sont pas abordées dans cette thèse.

Tout au long de ce document, il est supposé que T et C sont des variables aléatoires indépendantes. Un exemple de durée de vie pour laquelle l'hypothèse d'indépendance est questionnable est le temps écoulé jusqu'à une rechute chez un patient ayant subi une greffe de moelle osseuse pour le traitement de la leucémie. Cette durée de vie peut être censurée par un événement concurrent si le patient décède pendant la période de rémission suivant la greffe. À cet effet, il est probable que les patients à haut risque de rechute soient davantage à risque de décéder en cours de rémission, d'où une dépendance pour T et C.

Soit $\tilde{T} = \min(T, C)$ et $\delta = I(T < C)$, l'indicatrice de censure (vaut 1 si l'événement associé à T a pu être observé). En pratique, les durées de vie observables d'un échantillon de n individus sont représentées par les couples $\{(\tilde{T}_i, \delta_i), i = 1, \cdots, n\}$. Soit D, le nombre d'événements

distincts parmi cet échantillon : $\tilde{T}_1 < \tilde{T}_2 < \cdots < \tilde{T}_D$. Cette notation est plus commode aux fins de définition de l'estimateur de Kaplan-Meier (Kaplan and Meier, 1958) pour $S(t)$:

$$\hat{S}(t) = \begin{cases} 1 & t < \tilde{T}_1 \\ \prod_{\tilde{T}_i \leq t, \delta_i = 1} \left\{ 1 - \frac{d_i}{Y_i} \right\} & t \geq \tilde{T}_1 \end{cases} \tag{1}$$

où Y_i est le nombre d'individus pour qui l'événement survient à $\tilde{T}_i$ ou qui sont encore à risque pour l'événement à $\tilde{T}_i$, et d_i est le nombre d'événements survenant à $\tilde{T}_i$. Il est possible de montrer que $\hat{S}(t)$ est un estimateur sans biais de $S(t)$. L'idée intuitive derrière cet estimateur est la notion de redistribution à droite du risque. Soit e_i le nombre d'individus censurés à $\tilde{T}_i$. Cette notation supplémentaire permet d'écrire :

$$1 - \frac{d_i}{Y_i} = \frac{(Y_i - d_i)(Y_i + e_i)}{Y_i(Y_i + e_i)} = \frac{Y_i + e_i - d_i}{Y_i + e_i} - \frac{d_i e_i / (Y_i + e_i)}{Y_i} \tag{2}$$

L'expression de droite de la dernière égalité en (2) est composée de deux termes. Le deuxième représente le *risque* instantané des individus censurés à $\tilde{T}_i$ redistribué à droite sur les individus encore à risque. Il est à noter que ce terme est nul en l'absence de censure et que le premier terme correspond alors à l'estimateur empirique de $S(t)$.

Les tests d'association génétique

En pratique, il faut la plupart du temps aller au-delà de la description du comportement aléatoire d'une seule durée de vie avec (1). La question de recherche consiste souvent à comparer les durées de vie de plusieurs groupes d'individus. En statistique génétique, cette situation se présente lorsqu'il est question des tests d'association qui visent à identifier les marqueurs de susceptibilité pour une maladie. Il est d'intérêt d'identifier de tels marqueurs à des fins de compréhension des mécanismes biologiques du développement des pathologies. De plus, une stratification du risque en fonction de facteurs génétiques peut mener à de meilleurs programmes de dépistage, à de meilleures décisions en ce qui a trait au type et au moment des interventions et à de meilleurs médicaments. La médecine personnalisée, notamment sur la base du génome, est appelée à connaître de grands développements au cours des prochaines années (Topol, 2014).

Le génome humain est constitué d'environ 3 milliards de paires de bases guanine (G) - cytosine (C) et adenine (A) - thymine (T). Ces paires de bases constituent les brins d'ADN formant les chromosomes contenus dans le noyau d'une cellule. Les chromosomes contiennent des séquences particulières d'ADN appelées gènes qui sont utilisées pour le développement et le fonctionnement d'un individu. À l'intérieur du noyau de la cellule, les chromosomes sont regroupés par paire. La fertilisation fait en sorte que, pour chaque paire, un chromosome provient de la mère alors que l'autre provient du père. Il est d'usage de s'intéresser à un même

emplacement sur chacun des chromosomes d'une paire. Les brins d'ADN étant complémentaires, il ne faut qu'une base (G, C, A ou T) par chromosome pour décrire, dans le cas plus simple, un emplacement : par exemple, CC.

De façon générale, la différence d'ADN entre deux êtres humains est d'environ 0.1% (Kruglyak and Nickerson, 2001). Pour l'exemple ci-dessus, il se peut qu'au sein d'une population la plupart des individus soient CC alors que d'autres sont CT ou TT. Cette forme de variant génétique est appelée polymorphisme nucléotidique (SNP). Chacune des catégories (CC, CT ou TT) formant le SNP est un génotype. De façon générique, la notation AA, Aa et aa est utilisée. Les lettres A et a sont les allèles du SNP. Par convention, l'allèle a est celui le moins fréquent dans la population et porte le nom d'allèle mineur. Pour l'analyse statistique, il est plus commode de convertir l'information génétique en format numérique. Soit donc la variable aléatoire $G \in \{0, 1, 2\}$ égale au nombre de copies de l'allèle mineur pour un individu.

Tester l'association entre un SNP et un trait de survie tel que l'âge au moment d'un premier diagnostic de cancer du sein revient à comparer les fonctions de survie $S_0(t), S_1(t)$ et $S_2(t)$ correspondant aux différentes valeurs de G. Il s'agit ici d'un test d'hypothèses à trois échantillons pour lequel la littérature statistique est abondante en analyse des durées de vie (voir la section 7.3 de Klein and Moeschberger (2003) à ce sujet). Les premiers travaux dans le domaine ne visaient pas spécifiquement à répondre à des questions de recherche en génétique. Cependant, lorsque les durées de vie ne sont pas indépendantes ou ne sont pas identiquement distribuées, des développements méthodologiques particuliers s'avèrent nécessaires.

C'est ce qui se produit pour l'étude des facteurs modifiant le risque de cancer du sein chez les femmes porteuses d'une mutation délétère sur le gène BRCA1 (Miki et al., 1994) ou le gène BRCA2 (Wooster et al., 1995). Ces mutations sont rares : des données à l'échelle de la population suggèrent qu'environ 1 personne sur 800 est porteuse d'une mutation BRCA1 alors qu'environ 1 personne sur 500 est porteuse d'une mutation BRCA2 (Barnes and Antoniou, 2012). Cependant, les femmes qui en sont porteuses sont à haut risque de développer le cancer du sein avant l'âge de 70 ans. Ce risque a été estimé se situant entre 40 et 87 % chez les porteuses d'une mutation BRCA1 et entre 40 et 84 % chez les porteuses d'une mutation BRCA2 (Barnes and Antoniou, 2012).

Pour étudier davantage le risque chez ces sous-populations, des devis de collecte de données standards tels que les études de cohorte ou cas-témoins ne parviendraient à identifier qu'un petit nombre d'individus compte tenu de la rareté des mutations BRCA1 et BRCA2. Afin de pallier à cette difficulté, les données sont recueillies de façon rétrospective selon un devis d'étude cas-famille auprès de familles rencontrées dans des cliniques génétiques. La première personne testée dans une famille est souvent diagnostiquée avec un cancer du sein à un âge relativement jeune. Un tel devis de collecte des données implique donc des probabilités de sélection inégales puisque les jeunes personnes cancéreuses ont une probabilité plus élevée

d'être incluses dans le jeu de données que celles plus âgées et non-cancéreuses. Ainsi, le recrutement n'est pas aléatoire en ce qui a trait à l'âge d'apparition de la maladie et au fait d'être malade ou non, ce qui induit un biais de sélection. De plus, certains individus de l'échantillon appartiennent à une même famille ce qui complique davantage l'analyse puisqu'il faut tenir compte de la dépendance entre les durées de vie. La comparaison des fonctions de survie dans le cadre d'un devis d'étude cas-famille est la problématique qui fait l'objet du premier article de cette thèse. Une statistique de type Cramer-von Mises est utilisée pour tester l'égalité de $S_0(t), S_1(t)$ et $S_2(t)$ et fait intervenir une version pondérée de (1).

Le génotypage est la détermination des variants génétiques qu'un individu possède. Le nombre de SNPs dans le génome humain est estimé à environ 11 millions (Kruglyak and Nickerson, 2001). Avec l'avènement de technologies robustes et moins coûteuses pour le génotypage, il est devenu possible de procéder à des tests d'association sur tout le génome (*GWAS : genome-wide association studies*) pour les maladies humaines complexes. La première GWAS remonte à 2005 ; il s'en fait maintenant plus de 1000 par année et le nombre de SNPs par GWAS se mesure parfois en millions. Voici des exemples de GWAS où l'issue est une durée de vie : (Huang et al., 2009; Pillas et al., 2010; Couch et al., 2013; Gaudet et al., 2013). Bien qu'il soit possible de tester chacun des SNPs un à un, une telle démarche présente des limites (Lin et al., 2011) :

— Le grand nombre de tests effectués et les petites tailles d'effet de chacun des SNPs peuvent entraîner une faible puissance statistique ;

— Souvent, le SNP causal pour la maladie étudiée n'est pas génotypé et la taille d'effet ne peut ainsi être observée que partiellement via un ou plusieurs SNPs associés statistiquement au SNP causal ;

— L'effet d'interaction de deux ou plusieurs SNPs ne peut pas être détecté.

Ces limites ont mené à la mise en place de méthodes permettant de tester simultanément l'association entre un trait de survie et un ensemble de SNPs. Parmi les méthodes publiées jusqu'à maintenant, le modèle à risques proportionnels développé à l'origine par Cox (1972) est une approche fréquemment utilisée pour décrire le comportement aléatoire de T. De façon générale, ce modèle est défini par :

$$\lambda(t|X) = \lambda_0(t) \exp(X\beta) \tag{3}$$

où $\lambda_0(t)$ est la fonction de risque de base, X un vecteur $1 \times p$ de covariables et β un vecteur $p \times 1$ de paramètres. Le modèle est dit semi-paramétrique puisque la portion modélisant l'effet des covariables est paramétrique alors que $\lambda_0(t)$ est traité de façon non-paramétrique. Soient X_1 et X_2 des covariables associées respectivement à deux individus pour lesquels la durée de vie s'exprime au moyen du modèle (3). Le risque relatif entre ces deux individus, donné par

$$\frac{\lambda(t|X_1)}{\lambda(t|X_2)} = \exp\{(X_1 - X_2)\beta\},$$

est constant par rapport au temps, donc proportionnel, d'où cette appellation pour le modèle (3).

Pour $k = 1, \ldots, n$, soit $N_k(t) = I(\tilde{T}_k \leq t, \delta_k = 1)$, le processus qui compte le nombre total d'événements à t ou avant t, et $Y_k(t) = I(\tilde{T}_k > t)$, le processus qui compte le nombre total d'individus pour qui l'événement risque de se produire à t. La vraisemblance partielle introduite par Cox (1972) est utilisée pour l'estimation de β :

$$PL(\beta) = \prod_{l=1}^{n} \prod_{t \geq 0} \left(\frac{Y_l(t) \exp\{X_l\beta\}}{\sum_{k=1}^{n} Y_k(t) \exp\{X_k\beta\}} \right)^{dN_l(t)} \tag{4}$$

Il est possible de donner une appréciation intuitive du ratio dans (4). Le numérateur fait référence au risque instantané que le l-ième individu ait un événement à t^- alors que le dénominateur fait référence au risque instantané d'événement à t^- pour l'ensemble des individus de l'échantillon. La fonction de log-vraisemblance partielle s'écrit :

$$l(\beta) = \sum_{l=1}^{n} \int_0^{\infty} \left[Y_l(t) X_l \beta - \log \left\{ \sum_{k=1}^{n} Y_k(t) \exp\{X_k\beta\} \right\} \right] dN_l(t) \tag{5}$$

Voici l'équation d'estimation pour β issue de la dérivation de (5) :

$$U(\beta) = \sum_{l=1}^{n} \int_0^{\infty} \left\{ X_l - \frac{S^{(1)}(t)}{S^{(0)}(t)} \right\} dN_l(t) = 0,$$

où

$$S^{(1)}(t) = \sum_{k=1}^{n} Y_k(t) X_k \exp(X_k\beta)$$

et

$$S^{(0)}(t) = \sum_{k=1}^{n} Y_k(t) \exp(X_k\beta)$$

La matrice d'information égale au négatif de la dérivée seconde est donnée par :

$$\mathcal{I}(\beta) = \sum_{l=1}^{n} \int_0^{\infty} V(\beta, t) dN_l(t),$$

où

$$V(\beta, t) = \frac{S^{(2)}(t)}{S^{(0)}(t)} - \frac{S^{(1)}(t)^{\bigotimes 2}}{\left\{S^{(0)}(t)\right\}^2},$$

$$S^{(2)}(t) = \sum_{k=1}^{n} Y_k(t) X_k^{\bigotimes 2} \exp(X_k \beta),$$

et $a^{\bigotimes 2} = a'a$.

Soit $\hat{\beta}$ l'estimateur du maximum de la vraisemblance partielle et $\hat{\mathcal{I}} = \mathcal{I}(\hat{\beta})$, la matrice d'information estimée. En génétique, l'hypothèse nulle la plus souvent testée est celle de l'absence d'association entre X et T :

$$H_0 : \beta = 0$$

Voici trois tests courants :

— *Le test du rapport de vraisemblances :* $X^2_{LR} = 2\left\{l(\hat{\beta}) - l(0)\right\}$

— *Le test de Wald :* $X^2_W = \hat{\beta}'\hat{\mathcal{I}}\hat{\beta}$

— *Le test du score :* $X^2_{SC} = U(0)\hat{\mathcal{I}}(0)^{-1}U'(0)$

Sous H_0, la distribution asymptotique de ces trois tests est la loi du khi-carré à p degrés de liberté. Ces tests sont valides si l'ensemble $\{(\tilde{T}_i, \delta_i, X_i), i = 1, \cdots, n\}$ est constitué de triplets indépendants et identiquement distribués, ce qui n'est pas la situation qui prévaut en présence du plan d'échantillonnage cas-famille mentionné précédemment. Lorsque $p = 1$ et que X correspond au génotype d'un seul SNP, le premier article de cette thèse présente une alternative à ces tests quand, en plus, l'hypothèse des risques proportionnels pour le modèle (3) n'est pas raisonnable.

Lorsque X correspond à un ensemble de SNPs, l'approche préconisée dans le deuxième article de cette thèse pour tester l'association génétique est à base de noyau (*kernel-based method*). L'originalité de ces travaux est de rendre possible l'analyse en présence de données familiales en utilisant les matrices de parenté (*kinship matrix*) pour modéliser la dépendance entre individus. Un troisième article a été écrit afin de décrire l'implantation en code R d'une version bonifiée de cette nouvelle méthode et de la rendre facilement accessible aux généticiens.

L'identification de SNPs ou d'ensembles de SNPs associés au risque de cancer du sein au moyen des nouveaux tests d'hypothèses développés dans cette thèse est liée à l'objectif de stratification du risque évoqué précédemment. Pour ce qui est des populations de femmes porteuses d'une mutation BRCA1 ou BRCA2, un commencement de stratification du risque a été décrit récemment par Antoniou et al. (2010). Cette approche fonctionne de la façon suivante. D'abord, l'incidence moyenne du cancer du sein à l'âge t est supposée connue et des estimés $\hat{\beta}$ pour des SNPs significativement associés au risque de cancer du sein sont obtenus à partir d'études publiées. Il est supposé que l'incidence du cancer du sein à l'âge t pour une porteuse de mutation dépend de facteurs modifiant le risque de maladie à travers un modèle à risques proportionnels. Seulement le SNP avec la plus forte association détectée dans chaque région est utilisé. Le modèle suppose l'indépendance entre les SNPs. L'incidence de base du cancer du sein est estimée de façon récursive. Des courbes de survie pour diverses combinaisons de profils de SNPs peuvent alors être obtenues. La Figure 3 de Couch et al. (2013) et la Figure 2 de Gaudet et al. (2013) fournissent des exemples de telles courbes. Cependant, beaucoup de travail reste à faire pour une meilleure gestion clinique personnalisée des porteuses d'une mutation BRCA1 ou BRCA2. Jusqu'à récemment, il était estimé que les SNPs significativement associés au risque de cancer du sein expliquaient environ 3 % de la variabilité génétique du risque de cancer du sein chez les porteuses d'une mutation BRCA1 (Barnes and Antoniou, 2012). Ce pourcentage était d'environ 6 % chez les porteuses d'une mutation BRCA2.

La structure de cette thèse est la suivante. Puisque sa réalisation est par articles, le chapitre 1 détaille certains éléments de théorie complémentaires aux trois articles. Suivent les trois articles proprement dits présentés aux chapitres 2 à 4. Le chapitre 5 expose des résultats de simulations supplémentaires en lien avec le troisième article.

Chapitre 1

Éléments de théorie sur les durées de vie et l'association génétique

Sont détaillés ici quelques éléments de théorie qui ne sont présentés que de façon succincte dans les articles : le test à base de noyau sous le modèle de Cox, les copules, la matrice de parenté et le déséquilibre de liaison.

1.1 Test à base de noyau sous le modèle de Cox

Afin de donner une justification théorique, la formulation générale de Cai et al. (2011) est utilisée, valide pour tout type de noyau. Soit X un vecteur $1 \times p$ de covariables. Il peut s'agir, par exemple, de facteurs environnementaux, de facteurs associés au style de vie ou de composantes principales décrivant l'hétérogénéité parmi une population. Dans toute cette section, G est un vecteur $1 \times m$ correspondant aux génotypes d'un ensemble de m SNPs. En pratique, un échantillon est composé des quadruplets $\{\tilde{T}_i, \delta_i, G_i, X_i, i = 1, \ldots, n\}$. Le modèle à risques proportionnels considéré est

$$\lambda(t_i) = \lambda_0(t_i) e^{\gamma(G_i) + X_i \zeta}$$

où $\zeta = (\zeta_1, \ldots, \zeta_p)'$ est un vecteur $p \times 1$ de coefficients de régression,

$$\gamma(G_i) = \sum_{i'=1}^{n} \alpha_{i'} K(G_i, G_{i'})$$

et $K(\cdot, \cdot)$ est un noyau défini positif, les $\alpha_{i'}$ étant des paramètres inconnus.

Il est à noter qu'aux chapitres 3 et 4, le modèle à risques proportionnels est exprimé de façon alternative en utilisant la classe des modèles linéaires de transformation (*linear transformation*

models) (Cheng et al., 1995) :

$$H(t_i) = -\gamma(G_i) - X_i\zeta + \varepsilon_i \tag{1.1}$$

où $H(\cdot)$ est une fonction monotone croissante inconnue et ε_i est une variable aléatoire continue ayant une distribution quelconque. Si ε_i est distribuée selon la loi des valeurs extrêmes standard et que $H(t) = \log \Lambda_0(t) = \log \int_0^t \lambda_0(q)dq$, alors

$$\mathbb{P}[T_i > t_i] = \mathbb{P}[H(T_i) > \log \Lambda_0(t_i)] = \mathbb{P}[\varepsilon_i > \log \Lambda_0(t_i) + \gamma(G_i) + X_i\zeta] = e^{-\Lambda_0(t_i)e^{\gamma(G_i)+X_i\zeta}},$$

ce qui est la fonction de survie sous le modèle de Cox.

La détection d'une association statistiquement significative entre G et T s'effectue en testant l'hypothèse nulle suivante :

$$H_0 : \gamma(G) = \sum_{i=1}^{n} \alpha_i K(G, G_i) = 0 \tag{1.2}$$

Par définition, soit $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)' \equiv \tau\epsilon$ qui suit une distribution quelconque de moyenne nulle et de matrice de covariance $\tau^2\mathbb{K}^-$ où $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ est un vecteur de variables aléatoires, τ une composante de variance inconnue et $\mathbb{K}^-$ l'inverse généralisé de Moore-Penrose de $\mathbb{K}$, laquelle est une matrice $n \times n$ dont l'élément en position (l, l') est $K_{ll'} \equiv K(G_l, G_{l'})$. Il est à noter que $\gamma(G_l) = \alpha' K_l$ où $K_l = (K_{l1}, \ldots, K_{ln})'$. De plus, $\mathbb{E}[\epsilon] = 0$ et $\mathbb{V}\mathrm{ar}[\epsilon] = \mathbb{K}^-$. Il est possible de montrer que $\gamma(\cdot) = 0$ si et seulement si $\mathbb{V}\mathrm{ar}[\gamma(G)] = 0$ pour tout G. Ainsi, (1.2) est équivalent à tester

$$H_0 : \tau = 0 \tag{1.3}$$

Le test à base de noyau sous le modèle de Cox est en fait un test du score pour (1.3). Il s'agit d'un test de nullité pour l'effet global des m SNPs sur la durée de vie. C'est le caractère aléatoire du vecteur α qui rend le test particulièrement approprié pour un ensemble de SNPs présentant à la fois des effets protecteurs et délétères, en comparaison avec le test à fardeau (*burden test*) standard. La fonction de vraisemblance partielle pour τ avec ζ supposé connu peut s'écrire de la façon suivante :

$$Lik(\tau; \zeta) = \mathbb{E}_\epsilon[L(\tau, \epsilon; \zeta)]$$

où

$$L(\tau, \epsilon; \zeta) = \prod_{l=1}^{n} \prod_{t \geq 0} \left(\frac{Y_l(t) \exp\{\tau \epsilon' K_l + X_l \zeta\}}{\sum_{p=1}^{n} Y_p(t) \exp\{\tau \epsilon' K_p + X_p \zeta\}} \right)^{dN_l(t)} \tag{1.4}$$

L'expression pour $L(\tau, \epsilon; \zeta)$ est analogue à la vraisemblance partielle du modèle de Cox standard. Avant de s'intéresser directement à $Lik(\tau; \zeta)$, il est plus commode de d'abord calculer la dérivée partielle du logarithme de $L(\tau, \epsilon; \zeta)$ par rapport à τ :

$$\frac{\partial \log L}{\partial \tau}(\tau, \epsilon; \zeta) = \frac{\partial \mathcal{L}}{\partial \tau}(\tau, \epsilon; \zeta) = \sum_{l=1}^{n} \int_0^{\infty} \left(\epsilon' K_l - \frac{\sum_{p=1}^{n} \epsilon' K_p Y_p(s) \exp\{\tau \epsilon' K_p + X_p \zeta\}}{\sum_{p=1}^{n} Y_p(s) \exp\{\tau \epsilon' K_p + X_p \zeta\}} \right) dN_l(s)$$

L'évaluation de cette expression à $\tau = 0$ et le calcul de l'espérance vis-à-vis ϵ mène à

$$\mathbb{E}_\epsilon \left[\frac{\partial \mathcal{L}}{\partial \tau}(\tau, \epsilon; \zeta) \Big|_{\tau=0} \right] = \sum_{l=1}^{n} \mathbb{E}_\epsilon \left[\int_0^{\infty} \left(\epsilon' K_l - \frac{\sum_{p=1}^{n} \epsilon' K_p Y_p(s) \exp\{X_p \zeta\}}{\sum_{p=1}^{n} Y_p(s) \exp\{X_p \zeta\}} \right) dN_l(s) \right] = 0$$

Ainsi, le calcul de la dérivée partielle de $\log Lik(\tau; \zeta)$ évaluée à $\tau = 0$ conduit à :

$$\begin{aligned} \frac{\partial \log Lik}{\partial \tau}(\tau; \zeta) \Big|_{\tau=0} &= \frac{\partial \log \mathbb{E}_\epsilon[L(\tau, \epsilon; \zeta)]}{\partial \tau} \Big|_{\tau=0} \\ &= \frac{1}{Lik(0; \zeta)} \, \mathbb{E}_\epsilon \left[\frac{\partial \mathcal{L}}{\partial \tau}(\tau, \epsilon; \zeta) \right] \Big|_{\tau=0} \\ &= \frac{1}{Lik(0; \zeta)} \mathbb{E}_\epsilon \left[\frac{\partial \mathcal{L}}{\partial \tau}(\tau, \epsilon; \zeta) \Big|_{\tau=0} \right] \\ &= 0 \end{aligned} \tag{1.5}$$

Ce résultat est valide sous l'hypothèse que les conditions de régularité sont respectées de façon à pouvoir interchanger la fonction d'évaluation et l'espérance de même que la dérivation et l'espérance. Pour lever l'impasse obtenue, il faut considérer de façon alternative l'hypothèse $H_0 : \tau^2 = 0$, ce qui mène à :

$$\frac{\partial \log Lik}{\partial (\tau^2)} = \frac{\partial \log Lik}{\partial \tau} \frac{\partial \tau}{\partial (\tau^2)} = \frac{\partial \log Lik / \partial \tau}{2\tau} \tag{1.6}$$

où, pour alléger la notation et sans perte de généralité, les arguments τ et ζ sont enlevés. Avec $\tau \to 0$, (1.6) est indéterminée. L'application de la règle de L'Hôpital conduit à :

$$\lim_{\tau \to 0} \frac{\partial \log Lik}{\partial(\tau^2)} = \frac{1}{2} \frac{\partial^2 \log Lik}{\partial \tau^2} \tag{1.7}$$

Le terme de droite de (1.7) correspond à la statistique de test. Dans ce qui suit, la constante 1/2 est omise. À partir du résultat de (1.5), il est possible de montrer, au moyen d'un calcul de dérivées, que

$$\frac{\partial^2 \log Lik}{\partial \tau^2}(\tau; \zeta)\bigg|_{\tau=0} = \frac{\partial^2 \log Lik}{\partial \tau^2}(0; \zeta) = \frac{1}{Lik(0;\zeta)} \frac{\partial^2 Lik}{\partial \tau^2}(0; \zeta)$$

et

$$\frac{\partial^2 L}{\partial \tau^2}(0, \epsilon; \zeta) = L(0, \epsilon; \zeta) \left\{ \left(\frac{\partial \mathcal{L}}{\partial \tau}(0, \epsilon; \zeta) \right)^2 + \frac{\partial^2 \mathcal{L}}{\partial \tau^2}(0, \epsilon; \zeta) \right\}$$

Puisqu'il n'y a aucun terme ϵ dans $L(0, \epsilon; \zeta)$, il appert que $Lik(0; \zeta) = L(0, \epsilon; \zeta)$. Ainsi, la statistique de test est

$$\begin{aligned}
\hat{Q} &= \frac{1}{Lik(0;\zeta)} \frac{\partial^2 Lik}{\partial \tau^2}(0; \zeta) \\
&= \frac{1}{Lik(0;\zeta)} \left. \frac{\partial^2 \mathbb{E}_\epsilon[L(\tau, \epsilon; \zeta)]}{\partial \tau^2} \right|_{\tau=0} \\
&= \frac{1}{Lik(0;\zeta)} \mathbb{E}_\epsilon \left[\frac{\partial^2 L}{\partial \tau^2}(0, \epsilon; \zeta) \right] \\
&= \mathbb{E}_\epsilon \left[\left(\frac{\partial \mathcal{L}}{\partial \tau}(0, \epsilon; \zeta) \right)^2 \right] + \mathbb{E}_\epsilon \left[\frac{\partial^2 \mathcal{L}}{\partial \tau^2}(0, \epsilon; \zeta) \right]
\end{aligned} \tag{1.8}$$

Soit $N(t) = \sum_{k=1}^n N_k(t)$, $S^{(0)}(t) = \sum_{k=1}^n Y_k(t) \exp\{X_k \zeta\}$, $\hat{\Lambda}_0(t) = \int_0^t \frac{dN(s)}{S^{(0)}(s)}$ et $\hat{\Lambda}_k(t) = \exp\{X_k \zeta\} \hat{\Lambda}_0(t)$. Le premier terme à droite de l'égalité dans (1.8) se simplifie de la façon suivante :

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \tau}(0, \epsilon; \zeta) &= \int_0^\infty \sum_{l=1}^n \left(K_l' - \frac{\sum_{p=1}^n K_p' Y_p(s) \exp\{X_p \zeta\}}{\sum_{p=1}^n Y_p(s) \exp\{X_p \zeta\}} \right) \epsilon \, dN_l(s) \\
&= \int_0^\infty \left(\sum_{l=1}^n K_l' dN_l(s) - \frac{\sum_{p=1}^n K_p' Y_p(s) \exp\{X_p \zeta\}}{S^{(0)}(s)} dN(s) \right) \epsilon \\
&= \int_0^\infty \sum_{l=1}^n \left\{ dN_l(s) - Y_l(s) d\hat{\Lambda}_l(s) \right\} K_l' \epsilon \\
&= \sum_{l=1}^n M_l K_l' \epsilon
\end{aligned} \tag{1.9}$$

où

$$M_l = \int_0^\infty \left\{ dN_l(s) - Y_l(s) d\hat{\Lambda}_l(s) \right\}$$

est appelé le résidu martingale. L'équation (1.9) peut aussi s'écrire $\frac{\partial \mathcal{L}}{\partial \tau}(0, \epsilon; \zeta) = M' \mathbb{K} \epsilon$ où $M = (M_1, \ldots, M_n)'$. Ainsi,

$$\begin{aligned} \mathbb{E}_\epsilon \left[\left(\frac{\partial \mathcal{L}}{\partial \tau}(0; \epsilon, \zeta) \right)^2 \right] &= \mathbb{E}_\epsilon \left[\left(M' \mathbb{K} \epsilon \right)^2 \right] \\ &= M' \mathbb{K} \left(\mathbb{V}ar[\epsilon] \right) \mathbb{K} M \\ &= M' \mathbb{K} M \end{aligned} \tag{1.10}$$

Pour ce qui est du deuxième terme à droite de l'égalité dans (1.8), il se simplifie de la façon suivante :

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial \tau^2}(0, \epsilon; \zeta) &= \int_0^\infty \left(\frac{\sum_{p=1}^n \epsilon' K_p Y_p(s) \exp\{X_p \zeta\}}{\sum_{p=1}^n Y_p(s) \exp\{X_p \zeta\}} \right)^2 dN(s) \\ &\quad - \int_0^\infty \frac{\sum_{p=1}^n K_p' \epsilon \epsilon' K_p Y_p(s) \exp\{X_p \zeta\}}{\sum_{p=1}^n Y_p(s) \exp\{X_p \zeta\}} dN(s) \\ &= \int_0^\infty \frac{\left(\sum_{p=1}^n \epsilon' K_p Y_p(s) \exp\{X_p \zeta\} \right)^2}{S^{(0)}(s)} d\hat{\Lambda}_0(s) \\ &\quad - \int_0^\infty \sum_{p=1}^n K_p' \epsilon \epsilon' K_p Y_p(s) \exp\{X_p \zeta\} d\hat{\Lambda}_0(s) \end{aligned}$$

En calculant l'espérance, il est possible d'obtenir après simplification

$$\begin{aligned} \mathbb{E}_\epsilon \left[\frac{\partial^2 \mathcal{L}}{\partial \tau^2}(0, \epsilon; \zeta) \right] &= \sum_{l=1}^n \sum_{l'=1}^n \int_0^\infty \frac{K(G_l, G_{l'}) Y_l(s) Y_{l'}(s) \exp\{X_l \zeta\} \exp\{X_{l'} \zeta\}}{S^{(0)}(s)} d\hat{\Lambda}_0(s) \\ &\quad - \sum_{l=1}^n \int_0^\infty K(G_l, G_l) Y_l(s) \exp\{X_l \zeta\} d\hat{\Lambda}_0(s) \end{aligned}$$

ce qui correspond à $-q$ dans Cai et al. (2011). En combinant cette dernière équation avec (1.8) et (1.10), la statistique de test s'écrit alors $Q = M' \mathbb{K} M - q$. C'est la forme la plus générale des tests à base de noyau dont il est question aux chapitres 3 et 4. Dans cette expression, il est commun d'estimer ζ par son estimateur du maximum de la vraisemblance partielle sous H_0

(Cai et al., 2011). De plus, q n'étant qu'un paramètre de centralité assurant que la statistique du score est de moyenne zéro, il est possible d'omettre ce terme dans la définition du test. Afin d'obtenir une valeur de p pour la statistique du score au moyen du rééchantillonnage, Cai et al. (2011) obtiennent d'abord des expansions asymptotiques de celle-ci s'écrivant sous la forme de processus de martingales. Des réalisations de la statistique du score sous H_0 sont ensuite générées par une approximation de la distribution des processus de martingales faisant intervenir B vecteurs de perturbations aléatoires indépendantes et identiquement distribuées selon la loi normale univariée standard.

1.2 Les copules

De façon générale, une copule est une fonction qui permet de modéliser la dépendance entre deux ou plusieurs variables aléatoires en liant leur distribution de probabilité conjointe à leurs distributions de probabilité marginales. Mathématiquement, une copule de dimension d est une fonction $\mathcal{C}(u_1, \ldots, u_d)$ définie sur $[0,1]^d$, à valeurs dans $[0,1]$ et possédant les propriétés suivantes (Nelsen, 2006, p. 45) :

1. Pour tout vecteur $u \in [0,1]^d$, $\mathcal{C}(u) = 0$ si au moins un élément de u est nul ;
2. Si tous les éléments de u sont égaux à 1 sauf u_k, alors $\mathcal{C}(u) = u_k$;
3. Soient $u = (u_1, \ldots, u_d)', v = (v_1, \ldots, v_d)' \in [0,1]^d$ tels que $u_i \leq v_i$ pour tout $i = 1, \ldots, d$. Alors, l'inégalité du rectangle est respectée pour $\mathcal{C}$:
   $\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} \mathcal{C}(a_{1,i_1}, \ldots, a_{d,i_d}) \geq 0$,où $a_{j,1} = u_j$ et $a_{j,2} = v_j$.

Ces propriétés assurent qu'une copule de dimension d est une fonction de répartition bien définie et que les lois marginales associées sont uniformes. Pour $d > 2$, soient $k \in \{2, \ldots, d-1\}$ et $\{i_1, \ldots, i_k\}$, une sélection de k indices parmi les d indices originaux. Il est possible de montrer que la fonction $\mathcal{C}_k(u_{i_1}, \ldots, u_{i_k})$ définie sur $[0,1]^k$ et à valeurs dans $[0,1]$ est elle-même une copule (Cherubini et al., 2004, p.131). Le théorème suivant met en lumière le lien entre fonction de répartition et copule.

Théorème de Sklar de dimension *d*. *Soit F une fonction de répartition de dimension d dont les fonctions de répartition marginales sont $F_1, \ldots, F_d$. Alors il existe une copule $\mathcal{C}$ de dimension d telle que pour tout $x = (x_1, \ldots, x_d)'$ dans $\mathbb{R}^d$,*

$$F(x_1, \ldots, x_d) = \mathcal{C}\{F_1(x_1), \ldots, F_d(x_d)\} \tag{1.11}$$

Si $F_1, \ldots, F_d$ sont toutes continues, alors $\mathcal{C}$ est unique. À l'inverse, si $\mathcal{C}$ est une copule de dimension d et que $F_1, \ldots, F_d$ sont des fonctions de répartition, alors la fonction F définie

dans l'équation (1.11) est une fonction de répartition de dimension d dont les fonctions de répartition marginales sont $F_1, \ldots, F_d$.

Pour davantage de détails concernant ce théorème et sa démonstration, il est possible de consulter le livre de Nelsen (2006) et les références mentionnées dans cet ouvrage.

Dans le cadre de cette thèse, les copules interviennent pour 2 types de variables aléatoires : une durée de vie T dont la distribution marginale est estimée de façon non-paramétrique (chapitre 2) et une variable ε, par hypothèse distribuée selon la loi des valeurs extrêmes standard (chapitres 3 et 4). Dans chaque cas, les variables issues d'une même famille d'individus sont corrélées et le paramètre d'une copule permet de caractériser la force de la dépendance entre ces variables. La transformation par probabilité (*probability transformation*) est utilisée pour obtenir les distributions marginales : par définition, $S(t)$ et $F_\varepsilon(\varepsilon)$ sont uniformes, F_ε étant ici la fonction de répartition de la loi des valeurs extrêmes standard. Tenant compte du fait que le nombre d'individus par famille n'est pas constant, il est à noter que la dimension d de la copule est une quantité aléatoire tout au long de cette thèse. Voici maintenant trois exemples de copules. Au préalable, soit Γ une matrice $d \times d$ symétrique et définie positive telle que $\text{diag}(\Gamma) = (1, 1, \ldots, 1)'$.

La copule gaussienne

Soit Φ_Γ la fonction de répartition de la loi normale multivariée standard avec matrice de corrélation Γ. La copule gaussienne est définie par :

$$\mathcal{C}_\Gamma(u_1, \ldots, u_d) = \Phi_\Gamma \left\{ \Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d) \right\}$$

où Φ^{-1} est l'inverse de la fonction de répartition de la loi normale univariée standard.

La copule de Student

Soit $t_{\Gamma,\nu}$ la fonction de répartition de la loi de Student multivariée à ν degrés de liberté avec matrice de covariance Γ. La copule de Student est définie par :

$$\mathcal{C}_{\Gamma,\nu}(u_1, \ldots, u_d) = t_{\Gamma,\nu} \left\{ t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d) \right\}$$

où t_ν^{-1} est l'inverse de la fonction de répartition de la loi de Student univariée à ν degrés de liberté.

La copule de Clayton

Cette copule appartient à la famille des copules archimédiennes, soit les copules qui peuvent s'écrire sous la forme :

$$\mathcal{C}(u_1, \ldots, u_d) = \psi^{-1}\left\{\psi(u_1) + \ldots + \psi(u_d)\right\}$$

où la fonction $\psi(u) : [0,1] \to [0,\infty]$ est continue et strictement décroissante, et son inverse ψ^{-1}, complètement monotone sur $[0,\infty]$. Cette fonction est appelée le générateur de la copule. Dans le cas de la copule de Clayton, $\psi(u) = u^{-\alpha} - 1$ et $\psi^{-1}(s) = (s+1)^{-1/\alpha}$, qui est complètement monotone si α, le paramètre d'association, est strictement positif. Ainsi, la copule de Clayton peut s'écrire :

$$\mathcal{C}_\alpha(u_1, ..., u_d) = \left(\sum_{i=1}^{d} u_i^{-\alpha} - d + 1\right)^{-1/\alpha}, \quad \alpha > 0$$

À des fins d'étude de simulation, il est utile de pouvoir générer des variables aléatoires corrélées dont la dépendance est induite via une copule. Dans le cas des copules gaussienne et de Student, il suffit de simuler des observations à partir de lois normale et de Student multivariées, respectivement. Dans le cas de la copule de Clayton, il faut procéder par échantillonnage conditionnel (*conditional sampling*). La clé de cette procédure repose sur le lien qui existe entre probabilité conditionnelle et copule. Soit les variables aléatoires uniformes $U_1, \ldots, U_d$ dont la fonction de répartition conjointe est donnée par une copule quelconque $\mathcal{C}$. Alors (Joe, 2015, p. 29),

$$\mathbb{P}[U_k \leq u_k | U_1 = u_1, \ldots, U_{k-1} = u_{k-1}] = \frac{\partial^{k-1}\mathcal{C}_k(u_1, ..., u_k)}{\partial u_1 \ldots \partial u_{k-1}} \Big/ \frac{\partial^{k-1}\mathcal{C}_{k-1}(u_1, ..., u_{k-1})}{\partial u_1 \ldots \partial u_{k-1}},$$

pourvu que les numérateur et dénominateur existent et que le dénominateur soit différent de zéro.

Dans cette thèse, des copules permettent de modéliser la dépendance entre des variables aléatoires potentiellement censurées à droite. Cette particularité a évidemment un impact sur l'équation de la vraisemblance qui sert à estimer le paramètre d'une copule, que ce soit, à titre d'exemple, le α de la copule de Clayton ou un terme de composante de variance dans la matrice de covariance des copules gaussienne ou de Student. Ainsi, il est plus convivial pour la suite d'utiliser une copule qui lie la fonction de survie conjointe S aux fonctions de survie marginales $S_1, \ldots, S_d$:

$$S(t_1, \ldots, t_d) = \mathcal{C}_\theta\left\{S_1(t_1), \ldots, S_d(t_d)\right\} \tag{1.12}$$

pour tout $t = (t_1, \ldots, t_d)'$ dans $[0, \infty)^d$. Soit un échantillon de m grappes comportant d_i observations par grappe. La fonction de vraisemblance pour l'estimation du paramètre θ dans (1.12) s'écrit alors :

$$L = \prod_{i=1}^{m} \frac{(-1)^{p_i} \partial^{p_i} \mathcal{C}_\theta(u_{i1}, \ldots, u_{id_i})}{\partial u_{i1} \cdot \ldots \cdot \partial u_{ip_i}} \tag{1.13}$$

où $u_{ij} = S(t_{ij})$ pour tout $j \in \{1, \ldots, d_i\}$ et $p_i = \sum_{j=1}^{d_i} \delta_{ij}$ est le nombre d'observations non censurées dans la grappe. Dans (1.13), à des fins de simplification de l'écriture et sans perte de généralité, les p_i premières valeurs de $u_{i1}, \ldots, u_{id_i}$ correspondent aux observations non censurées et les $d_i - p_i$ dernières valeurs, aux observations censurées. Cette équation est analogue à la formule générale de la fonction de vraisemblance dans le cas d'un échantillon d'observations indépendantes censuré à droite :

$$L = \prod_{i=1}^{n} \mathbb{P}[t_i, \delta_i] = \prod_{i=1}^{n} \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}$$

Lorsque d_i est constant et vaut 2, (1.13) devient :

$$L = \prod_{i=1}^{m} \frac{\partial^2 \mathcal{C}_\theta(u_{i1}, u_{i2})^{\delta_{i1}\delta_{i2}}}{\partial u_{i1} \partial u_{i2}} \frac{\partial \mathcal{C}_\theta(u_{i1}, u_{i2})^{\delta_{i1}(1-\delta_{i2})}}{\partial u_{i1}} \frac{\partial \mathcal{C}_\theta(u_{i1}, u_{i2})^{(1-\delta_{i1})\delta_{i2}}}{\partial u_{i2}} \mathcal{C}_\theta(u_{i1}, u_{i2})^{(1-\delta_{i1})(1-\delta_{i2})}$$

1.3 La matrice de parenté

En présence de données d'individus apparentés résultant d'un plan d'échantillonnage cas-famille, il est d'intérêt d'exploiter l'information connue décrivant les relations entre les individus. Une forme d'information privilégiée est le coefficient de parenté (Jacquard, 1970). Pour deux individus A et B, il est défini comme étant la probabilité qu'un gène pris au hasard chez A soit identique à un gène pris au hasard au même locus chez B. En supposant que les ancêtres inconnus des individus de l'échantillon sont tous distincts, non consanguins et non apparentés entre eux, il est possible de calculer de façon théorique le coefficient de parenté. Un exemple simple est celui de deux individus A et B n'ayant qu'un ancêtre commun C (une seule chaîne de parenté). En pareil cas, la probabilité de l'événement *Un gène pris au hasard chez A est identique à un gène pris au hasard au même locus chez B* est donnée par $\left(\frac{1}{2}\right)^{n+p+1}$, où n et p sont respectivement les nombres de degrés par lesquels C est un ancêtre de A et B. Cette probabilité est le produit de trois probabilités :

— La probabilité pour le gène de A de provenir de C : $(1/2)^n$;

— La probabilité pour le gène de B de provenir de C : $(1/2)^p$;

— La probabilité que les gènes de A et B soient tous deux la réplique d'un même gène de C : 1/2.

Par exemple, le coefficient de parenté entre une mère et sa fille se calcule de la façon suivante :

$$\phi_{MF} = \left(\frac{1}{2}\right)^{0+1+1} = \frac{1}{4}$$

En présence de plus d'une chaîne de parenté et toujours sous l'hypothèse que les ancêtres inconnus sont distincts, non consanguins et non apparentés entre eux, il faut faire la somme sur chacune des chaînes i et l'équation devient :

$$\phi_{AB} = \sum_i \left(\frac{1}{2}\right)^{n_i+p_i+1}$$

Voici deux exemples.

Soeur-soeur

$$\phi_{SS} = \left(\frac{1}{2}\right)^{1+1+1} + \left(\frac{1}{2}\right)^{1+1+1} = \frac{1}{4}$$

Tante-nièce

$$\phi_{TN} = \left(\frac{1}{2}\right)^{1+2+1} + \left(\frac{1}{2}\right)^{1+2+1} = \frac{1}{8}$$

La matrice de parenté ϕ est formée des coefficients de parenté de tous les individus pris deux à deux parmi un échantillon. En présence d'un échantillon d'individus non-apparentés et non consanguins, il est à noter que la matrice 2ϕ équivaut alors à la matrice identité.

Avec l'avènement des données GWAS, il est devenu possible d'estimer la matrice de parenté, de façon moyenne, à l'échelle du génome. L'estimation part du principe que deux allèles qui ne proviennent pas d'un ancêtre commun chez deux individus peuvent être vus comme des tirages aléatoires parmi un pool d'allèles. Le coefficient de parenté est alors interprétable comme un coefficient de corrélation pour des variables indiquant si les allèles tirés sont d'un type donné, par exemple A, l'allèle majeur. Soient $x_{ik}, x_{jk} \in \{0, 1, 2\}$, le nombre d'allèles de type A chez les $i^{\text{ième}}$ et $j^{\text{ième}}$ individus d'un échantillon pour un SNP k et p_k, la fréquence de l'allèle A. Alors,

$$\hat{\phi}_{ij} = \frac{1}{L}\sum_{l=1}^{L}\frac{(x_{il}-2p_l)(x_{jl}-2p_l)}{4p_l(1-p_l)}$$

où L est le nombre de SNPs de l'étude GWAS. En pratique, il est possible qu'une portion non observée du pedigree définisse de lointaines parentés entre des individus de l'échantillon (*cryptic relatedness*) ou que l'échantillon soit subdivisé en strates d'individus ayant une ascendance lointaine commune (*population structure*). La matrice de parenté estimée est alors un bon résumé du pedigree à des fins d'ajustement dans le cadre d'une analyse d'association génétique (Astle and Balding, 2009).

1.4 Le déséquilibre de liaison

La méiose est un type de division cellulaire qui mène à la production des gamètes pour la reproduction. Durant ce processus, un phénomène d'enjambement fait en sorte que l'arrangement des bases nucléotidiques sur un chromosome est reconfigué. Ainsi, les chromosomes sont des mosaïques. Si deux loci sont relativement proches, il est davantage probable que leur disposition sur un chromosome demeure intacte sur plusieurs générations. C'est ce que le déséquilibre de liaison (LD) permet de mesurer. De façon générale, le LD entre deux SNPs est inversement proportionnel à la distance qui les sépare et varie en fonction de la région où se trouvent les SNPs sur le génome. Des facteurs autres que l'enjambement, par exemple le taux de mutation et la sélection naturelle, entrent aussi en ligne de compte.

Soit un SNP bi-allélique dont les probabilités des allèles sont p_A et p_a. Soit un deuxième SNP sur le même chromosome dont les probabilités des allèles sont p_B et p_b. Le couple formé par les allèles de ces deux SNPs est appelé haplotype. Soient p_{AB}, p_{Ab}, p_{aB} et p_{ab}, les probabilités des quatre haplotypes possibles. Une variable souvent utilisée pour mesurer le LD est le carré du coefficient de corrélation :

$$r^2 = \frac{(p_{AB}-p_Ap_B)^2}{p_Ap_ap_Bp_b} \tag{1.14}$$

Une population est dite en équilibre de liaison pour ces deux SNPs quand $p_{AB} = p_Ap_B$. Intuitivement, cet équilibre signifie que la probabilité d'observer à la fois les allèles A et B est égale à la probabilité d'observer A et B indépendamment. À l'inverse, une différence entre p_{AB} et p_Ap_B est révélatrice d'une association entre la fréquence d'apparition de l'haplotype AB et celles des allèles A et B. Une formule équivalente à (1.14) est :

$$r^2 = \frac{(p_{AB}p_{ab}-p_{Ab}p_{aB})^2}{p_Ap_ap_Bp_b}$$

Dans cette équation, le numérateur est maximal si $p_{AB}p_{ab} = 0$ ou si $p_{Ab}p_{aB} = 0$, donc si seulement deux haplotypes sont observés. Sans perte de généralité, il peut s'agir par exemple de AB et ab. En pareil cas, $p_A = p_B = p_{AB}$, $p_a = p_b = p_{ab}$ et $r^2 = 1$. Il s'agit de la valeur maximale du r^2.

Avec $r^2 = 1$, la corrélation est parfaite entre les deux SNPs : la connaissance des génotypes de l'un d'eux dans l'échantillon constitue une information complète au sujet de l'autre, les deux SNPs étant alors redondants. Ainsi, si r^2 est élevé, moins de SNPs sont nécessaires pour décrire la variabilité dans une région. En analyse d'association génétique, il est d'intérêt de choisir de façon optimale les SNPs à analyser, l'objectif étant d'aller chercher le maximum d'information pour un nombre de SNPs qui est fixé d'avance en raison du coût du génotypage.

Il est possible de tester l'hypothèse nulle d'équilibre de liaison au moyen de la statistique $2nr^2$ qui est distribuée selon la loi du khi-carré à un degré de liberté sous H_0.

Chapter 2

Analysis of multivariate failure times in the presence of selection bias with application to breast cancer

[Analyse de durées de vie multivariées en présence d'un biais de sélection: application au cas du cancer du sein]

Résumé

Identifier des loci qui modifient le risque de cancer chez les porteurs d'une mutation génétique est un sujet important en oncogénétique. À l'intérieur de ce domaine de recherche, notre intérêt s'est porté vers l'analyse de l'association entre un variant génétique et le risque de cancer chez les femmes porteuses d'une mutation pathogène sur le gène BRCA2. Puisque cette mutation est rare, les données ont été recueillies de façon rétrospective selon un devis d'étude cas-famille faisant appel à des programmes de dépistage génétique. Un tel devis implique un biais de sélection et de la corrélation intra-famille, ce qui complique l'analyse statistique. Dans cet article, une statistique de type Cramer-von Mises est développée afin de tester l'égalité des fonctions de survie spécifiques aux génotypes lorsque le modèle des risques proportionnels n'est pas applicable. Une copule de Clayton est utilisée afin de modéliser la dépendance familiale résiduelle des phénotypes et une procédure de bootstrap semi-paramétrique novatrice est proposée pour approximer la distribution de la statistique de test sous l'hypothèse nulle. Le test proposé est appliqué à des données provenant de porteuses de mutation européennes et nord-américaines et sa performance est évaluée au moyen de simulations.

Abstract

Identifying loci that modify the risk of cancer for genetic mutation carriers is an important topic in oncogenetics. Within this research area, we are concerned with the analysis of the association between a genetic variant (SNP rs13281615) and breast cancer among women with a pathogenic mutation in the BRCA2 gene. As this mutation is rare, data were collected retrospectively according to a case-family design through genetic screening programs. This involves a selection bias and an intrafamilial correlation, which complicates the statistical analysis. We derive a Cramer–von Mises type statistic to test the equality of genotype-specific survival functions when the proportional hazards model does not hold. A Clayton copula is specified to model the residual phenotype familial dependence and an innovative semiparametric bootstrap procedure is proposed to approximate the distribution of the test statistic under the null hypothesis. The proposed test is applied to data from European and North American mutation carriers and its performance is evaluated by simulations. [1]

2.1 Introduction

Breast cancer aggregates in families, such that the disease is approximately twice as common in the first-degree relatives of breast cancer patients as in the general population. Twin and family studies indicate that this familial risk is largely the result of inherited susceptibility due to the combined effects of multiple genetic variants (Lichtenstein et al., 2000). Three classes of breast cancer susceptibility alleles, with different levels of risk and prevalence in the population, have been identified: high, intermediate, and low risk alleles. Alleles in genes such as BRCA1, BRCA2, PTEN and TP53 confer high lifetime risks of the disease but are relatively rare, and explain approximately 20% of the familial risk (Easton, 1999). Mutations in intermediate-risk alleles (CHEK2, ATM, BRIP1, PALB2 and XRCC2) confer risks of 2-4 fold, and explain a further ∼3% of the familial risk. Common low-risk alleles (frequency ≥5%, RR <1.3) explain ∼14% of the inherited susceptibility (Ghoussaini et al., 2013) at the moment.

The risk estimates associated with pathogenic mutations in the BRCA1/2 genes have been found to vary substantially among studies (Antoniou et al., 2003; Begg et al., 2008; Brohet et al., 2014; Milne et al., 2008; Simchoni et al., 2006). There is also variability in age at diagnosis and type of cancer in the index case (Antoniou et al., 2003), even among women who carry the same BRCA mutation (Thorlacius et al., 1997) and among women in the same family (Vogl et al., 2007). Such evidence suggests that genetic or other factors that cluster in families may modify the cancer risks conferred by BRCA1 and BRCA2 mutations.

1. This is the accepted version of the following article: Leclerc, M., EMBRACE Investigators, GEMO Study Collaborators, INHERIT Investigators, Antoniou, A. C., Simard, J. and Lakhal-Chaieb, L. (2015), Analysis of multivariate failure times in the presence of selection bias with application to breast cancer. Journal of the Royal Statistical Society: Series C (Applied Statistics), 64: 525–541. doi: 10.1111/rssc.12091, which has been published in final form at `http://onlinelibrary.wiley.com/doi/10.1111/rssc.12091/abstract`.

Indeed, several factors including common genetic variants, hormonal/lifestyle and imaging variables have recently been shown to modify breast and ovarian cancer risks for BRCA1/2 carriers (Mitchell et al., 2006; Barnes and Antoniou, 2012). Due to the high risk of breast and ovarian cancer conferred by BRCA1/2 mutations, the combined effects of the common alleles and other risk factors can result in large differences in the absolute risk of developing these cancers between extreme genotypes. In this regard, findings from a prospective study support the evidence that common breast cancer susceptibility alleles in combination are predictive of breast cancer risk for BRCA2 carriers (Mavaddat et al., 2013).

Currently, genetic counselling of women with a deleterious mutation in the BRCA1 and BRCA2 breast and ovarian cancer susceptibility genes is based on average cancer risk estimates, which do not take into account the potential modifying effects of genetic and environmental factors. Incorporating the explicit effects of modifiers into a cancer risk prediction model would thus improve our ability to estimate personalized cancer risks for women with BRCA1/2 mutations. Such risk estimates would improve the clinical management of BRCA1/2 mutation carriers, for example in making better decisions concerning the type and timing of interventions and in the delivery of effective screening programs (Milne and Antoniou, 2011).

This work is motivated by the analysis of the association between a genetic variant termed Single-nucleotide polymorphism (SNP) rs13281615 and breast cancer among women with a mutation in BRCA2. The risk of cancer is specified by the distribution of a failure time T corresponding to the age at onset of cancer whereas the SNP under investigation is expressed in terms of the observed genotype $G = g \in \{0, 1, 2\}$ equal to the number of copies of the minor allele. The analysis of the association consists then of testing the equality of the three genotype-specific survival functions of $T : H_0 : S_0(t) = S_1(t) = S_2(t)$ for all $t \geq 0$.

As BRCA2 mutations are rare, standard data collection designs such as cohort and case-control studies will be able to identify only a small number of subjects. To overcome this difficulty, our data were collected retrospectively according to a case-family design through genetic screening programs where members of families with strong cancer history are included if the first person screened in the family has a high probability of carrying a mutation. However, this data collection protocol involves unequal sampling probabilities as young and diseased persons are more likely to be included in the resulting dataset than old and unaffected persons. Therefore, the ascertainment is non-random with respect to the age at onset and the disease status, which induces a selection bias.

Various versions of conditional likelihood functions have been considered in the literature to correct for the bias due to the ascertainment with familial data (Chatterjee et al., 2006; Zhang et al., 2010; Schaid et al., 2010; Barnes et al., 2013). Typically, this strategy requires the knowledge of the relationships between family members.

Without such information, Antoniou et al. (2005) derived a weighted cohort method. This

approach assigns weights to all subjects depending of their age at onset of cancer or last followup/contact, and disease status. These weights, intended to correct for the bias due to the ascertainment, are computed such that the overall estimated weighted incidence of cancer is equal to the true incidence function, assumed to be known. The authors assumed then a proportional hazard model

$$h(t \mid G = g) = h_0(t)e^{\beta g}, \tag{2.1}$$

where h is the hazard function of T, h_0 the baseline hazard and e^{β} the per-allele hazard ratio (HR). The significance of the association between a SNP and the risk of cancer is tested via $H_0 : \beta = 0$ using the Wald statistic $\hat{\beta}/\sqrt{v(\hat{\beta})}$ where $\hat{\beta}$ is obtained by maximizing the weighted partial likelihood and $v(\hat{\beta})$ is given by the robust sandwich estimator as the dependence between the members of the same family is treated as nuisance. Alternatively, one may also consider the model with separate HR

$$h(t \mid G = g) = h_0(t)e^{\beta_1 I(g=1)+\beta_2 I(g=2)} \tag{2.2}$$

and test the significance of the association via $H_0 : \beta_1 = \beta_2 = 0$ using the Wald test. The appropriateness of these proportional hazards models to our data set can be easily checked by plotting the estimated genotype-specific log cumulative hazard functions versus time, which are reported in Figure 2.3. This figure reveals a strong graphical evidence that these models do not hold since there are multiple crossings between the curves. Adding time-dependent variables of the form $g \times \log(t)$ to models (2.1) and (2.2) and performing the global test of proportional hazards based on scaled Schoenfeld residuals as described by Therneau and Grambsch (2000) also indicates departure from proportionality with p-values of 0.0173 and 0.0296 respectively. In this paper, we develop an association test between SNP rs13281615 and the risk of breast cancer among BRCA2 mutation carriers without assuming proportional hazards.

The paper is organized as follows: in Section 2.2, we describe the BRCA2 data set. In Section 2.3, we derive a Cramer–von Mises type test statistic. Section 2.4 is devoted to the approach employed to handle the non-randomness in the data. A semiparametric bootstrap procedure to approximate the p-value of the proposed test is described in Section 2.5. Results from simulation studies and the application to the BRCA2 data are presented in Sections 2.6 and 2.7 respectively. Section 2.8 provides a discussion.

2.2 The BRCA2 data set

The data analyzed come from three studies assessing the association between SNPs and breast cancer risk in BRCA1 and BRCA2 mutation carriers (countries and sample sizes given in

brackets): Epidemiological study of BRCA1 and BRCA2 mutation carriers (EMBRACE) [UK, Ireland; 1,022], Genetic Modifiers of cancer risk in BRCA1/2 mutation carriers (GEMO) [France, USA; 717], Interdisciplinary Health Research International Team on Breast Cancer Susceptibility (INHERIT BRCAs) [Quebec - Canada; 48].

All of the 1,787 subjects are women who carry a mutation in the BRCA2 gene and were recruited through cancer genetics clinics or research studies of high-risk families. The sample consists of 1,289 clusters, 24% having a size greater than one. Among these, the cluster size varies between 2 and 15, for a total of 812 subjects (45% of the sample).

The frequencies of the genotypes AA, AG and GG of rs13281615 are equal to 30%, 51% and 19% respectively, which suggests that the Hardy-Weinberg equilibrium is fulfilled (p-value of the chi-square test is 0.166).

The observations are right-censored if any of the following three events occurs before breast cancer diagnosis: ovarian cancer diagnosis, bilateral prophylactic mastectomy, or loss to follow-up. Censoring rates are 46%, 43%, and 39%, respectively for the genotypes AA, AG, and GG.

2.3 Testing the SNP-breast cancer association

The data consist of n independent clusters, corresponding to families. For $i = 1, \ldots, n$, the i^{th} cluster is composed of d_i triplets of the observed variables $\{(X_{ij}, \delta_{ij}, G_{ij}), j = 1, \ldots, d_i\}$ where $X_{ij} = \min(T_{ij}, C_{ij})$, $\delta_{ij} = I(T_{ij} < C_{ij})$, C_{ij} is the censoring variable, assumed to be independent from the failure time T_{ij}, and G_{ij} the genotype of the individual.

In this paper, we correct for the bias due to the ascertainment via the weighting approach of Antoniou et al. (2005). The key assumption behind this approach is that the probability of a mutation carrier to be sampled in the dataset does not depend on the genotype G but depends only on the age at onset of cancer or last followup/contact X and the disease status δ. This assumption is reasonable for the data analysed since BRCA2 mutation carriers are recruited through cancer genetics clinics or research studies of high-risk families and their SNP genotypes are not known in advance. Antoniou et al. (2005) estimated the weights, expressed then as $\Phi(X, \delta)$, by two piecewise constant functions of X, one for each possible value of δ. Technical details of this estimation procedure are given in Appendix A. Individual weights $w_{ij} = \widehat{\Phi}(X_{ij}, \delta_{ij})$ are assigned to all subjects included in the dataset. This weighted cohort approach has been shown to provide unbiased estimates of the association between a risk factor and disease risk (Antoniou et al., 2005). For instance, we follow Miyahara and Wahed

(2010) and define the weighted at-risk and event-counting processes as follows

$$\begin{aligned} N_{ij}^w(t) &= w_{ij} I(X_{ij} \le t; \delta_{ij} = 1), Y_{ij}^w(t) = w_{ij} I(X_{ij} > t), \\ N_g^w(t) &= \sum_{i=1}^{n} \sum_{j=1}^{d_i} I(G_{ij} = g) N_{ij}^w(t), N^w(t) = \sum_{g=0}^{2} N_g^w(t), \\ Y_g^w(t) &= \sum_{i=1}^{n} \sum_{j=1}^{d_i} I(G_{ij} = g) Y_{ij}^w(t), Y^w(t) = \sum_{g=0}^{2} Y_g^w(t). \end{aligned}$$

One may then estimate the genotype-specific survival function S_g, $g = 0, 1, 2$, by the weighted Kaplan–Meier estimator

$$\hat{S}_g(t) = \prod_{X_{ij} \le t, \delta_{ij}=1, G_{ij}=g} \left\{ 1 - \frac{\Delta N_g^w(X_{ij})}{Y_g^w(X_{ij})} \right\},$$

where $\Delta N_g^w(t) = N_g^w(t) - N_g^w(t^-)$ is the magnitude of the jump of the process N_g^w at t. According to Williams (1995), this estimator remains valid even if the failure times of members of the same family are correlated. A natural Cramer–von Mises type statistic to test H_0 : $S_0(t) = S_1(t) = S_2(t)$ for all $0 \le t \le \zeta$ is then given by

$$K = \int_0^{\zeta} \sum_{g=0}^{2} m_g \left\{ \hat{S}_g(t) - \hat{S}_P(t) \right\}^2 d\hat{S}_P(t), \tag{2.3}$$

where $m_g = \sum_{i=1}^{n} \sum_{j=1}^{d_i} I(G_{ij} = g)$, $\hat{S}_P(t) = \prod_{X_{ij} \le t, \delta_{ij}=1} \left\{ 1 - \frac{\Delta N^w(X_{ij})}{Y^w(X_{ij})} \right\}$ and ζ is the largest time with at least one subject at risk for each genotype. The null hypothesis is rejected when K is large. However, the asymptotic distribution of K under H_0 is not easy to derive and one must rely on resampling techniques to approximate the p-value associated with K. Unfortunately, a simple permutation test is not valid for this situation since the genotypes of the members of the same family are correlated. Moreover, permutation tests for clustered data are associated with a significant loss of power (Braun and Feng, 2001) and therefore may not be of great use for genetic studies. In this paper, we propose to approximate the p-value associated with K via an appropriate semiparametric bootstrap procedure.

2.4 Non-random sampling

The main statistical contribution of this paper is to propose a semiparametric bootstrap procedure in the presence of a selection bias, which requires generating correlated observations according to a non-random design with respect to X and δ. The theoretical background behind such sampling is presented in this section.

We begin with the simple case of i.i.d. observations with no censoring and no clusters. Let f be a density function with support on $[0, \infty)$, $S(t) = \int_t^{\infty} f(u)du$ the associated survival function,

$\mathcal{P} : [0, \infty) \to [0, 1]$ an ascertainment sampling probability function, and $w(t)$ a weight function proportional to $1/\mathcal{P}(t)$. Consider a random sample $T_1, \cdots, T_N$ drawn from a distribution with a density function f and construct a new dataset as follows. For $i = 1, \cdots, N$, generate Y_i according to a Bernoulli distribution with a probability of success $\Pr(Y_i = 1) = \mathcal{P}(T_i)$ and include T_i in the new dataset if $Y_i = 1$, as considered by Patil and Rao (1978). The resulting sample of size $n \leq N$, termed a biased or non-random sample in the sequel, follows a distribution with a density function equal to $f_{\mathcal{P}}(t) = \mathcal{P}(t)f(t)/\int_0^\infty \mathcal{P}(u)f(u)du$. Note that $f_{\mathcal{P}}$ is invariant if $\mathcal{P}$ is multiplied by a constant A such that $A \times \mathcal{P}(t) \leq 1$ for all $t \in [0, \infty)$.

Inversely, given a non-random sample $T_1, \ldots, T_n$ following a distribution with a density function $f_{\mathcal{P}}$, one may estimate S by the weighted empirical estimator $\hat{S}(x) = \sum_{i=1}^n I(T_i > x)w(T_i)/\sum_{i=1}^n w(T_i)$ when w is known. Very few attempts have been made to generalize the above theory to clusters of non-random possibly correlated variables. Indeed, the existing literature mainly focuses on the cases where entire clusters are included in the dataset according to a given non-random selection design (Gong et al., 2010). In this paper, we follow the design used to collect our dataset and focus on the selection process of the individuals rather than that of the clusters.

The joint distribution of the genotypes of mutation carriers belonging to the same family follows the Mendelian inheritance law. On the other hand, the dependence structure of ages at onset of cancer given the genotypes can be complex to model. Some studies assume the conditional independence of the phenotypes given the genotypes. However, this ignores the residual correlation possibly due to other genes or shared environmental risk factors and may lead to biased results (Begg, 2002; Kraft and Thomas, 2000). In this paper, we model the residual phenotype dependence via a copula model (Chatterjee et al., 2006; Gong et al., 2010). The conditional joint survival function of ages at onset of cancer of a family of D mutation carriers given their genotypes is then

$$\Pr(T_1 > t_1, \ldots, T_D > t_D \mid G_1 = g_1, \ldots, G_D = g_D) = \mathcal{C}_{\theta,D}\left\{S_{g_1}(t_1), \ldots, S_{g_D}(t_D)\right\},$$

where $\mathcal{C}_\theta$ is a copula model indexed by a parameter θ measuring the residual familial aggregation corresponding to the dependence between the ages at onset within the family that cannot be attributed to the SNP under investigation. In the paper, we focus on a particular copula, namely Clayton's copula given by

$$\mathcal{C}_{\theta,D}(u_1, \cdots, u_D) = \left[\sum_{j=1}^{D} u_j^{-\theta} - D + 1\right]^{-1/\theta}, \theta > 0. \tag{2.4}$$

In Appendix B, we recall some useful properties of this copula and present the procedure of Lee (1993), which generates a set of random variables with a dependence structure given by (2.4). By combining the strategy described above to introduce non-randomness with the

Table 2.1 – Algorithm generating correlated failure times $T_1, \ldots, T_d$ from a non-random sample (no censoring)

- Generate the first variate ($j = 1$)
  - Repeat until SUCCESS = 1
    - Generate $U_1 \sim$ Uniform$[0, 1]$
    - Set $T_1 = S_{g_1}^{-1}(U_1)$
    - SUCCESS $\sim$ Bernoulli$\{\mathcal{P}(T_1)\}$
  - Save U_1 and T_1
- For $j = 2, \ldots, d$
  - Repeat until SUCCESS = 1
    - Generate $V_j \sim$ Uniform$[0, 1]$
    - Solve $\mathcal{H}_{\theta,j}(U_1, \ldots, U_j) = V_j$ to obtain U_j
    - Set $T_j = S_{g_j}^{-1}(U_j)$
    - SUCCESS $\sim$ Bernoulli$\{\mathcal{P}(T_j)\}$
  - Save U_j and T_j

where $\mathcal{H}_{\theta,j}$ is defined in Appendix B and satisfies

$$\mathcal{H}_{\theta,j}\{S_{g_1}(t_1), \ldots, S_{g_j}(t_j)\} =$$

$$\Pr(T_j > t_j \mid T_1 = t_1, \ldots, T_{j-1} = t_{j-1}, G_1 = g_1, \ldots, G_j = g_j).$$

procedure of Lee (1993), we obtain the algorithm of Table 2.1 which generates d correlated observations according to a non-random ascertainment design with respect to $\mathcal{P}(T)$.

2.5 Inference procedures

2.5.1 Estimation of θ

In order to use the semiparametric bootstrap to resample the clustered data, we first need to estimate the dependence parameter θ of the Clayton copula from the original data. This is achieved by adapting the two-stage approach of Shih and Louis (1995). In the first stage, individual genotype-specific survival probabilities $\{\hat{S}_{g_{ij}}(X_{ij}), i = 1, ..., n, j = 1, ..., d_i\}$ are obtained via the weighted Kaplan-Meier estimator introduced in Section 2.3. At the second stage, one may estimate θ by maximizing an appropriate log-likelihood function. In order to correct for the bias due to the ascertainment, the contribution of each cluster to the standard log-likelihood function must be multiplied by a familial weight inversely proportional to the probability that the cluster is included in the dataset. However, a family of size $d \geq 2$ is

included in the dataset if at least two members of the family are included in the dataset, which involves multiple possibilities. Therefore, the computation of the familial weights can be complex and computationally very costly, especially if d is large.

Alternatively, one may consider a pairwise approach (Chen and Yu, 2012). Indeed, the joint distribution of any pair of members of the same family follows $\mathcal{C}_{\theta,2}$. The weight of such a pair is simply the product of the two individual weights since the pair contributes to the log-likelihood function if and only if both members are included in the dataset. Therefore, one may estimate θ by maximizing the pairwise log-likelihood function given by

$$\sum_{i=1}^{n}\sum_{j=1}^{d_i-1}\sum_{k=j+1}^{d_i} w_{ij}w_{ik}\log\left\{L(\theta,\hat{S}_{g_{ij}}(X_{ij}),\hat{S}_{g_{ik}}(X_{ik}))\right\},$$

where

$$\begin{aligned} L(\theta,u_{ij},u_{ik}) &= \left\{\mathcal{C}^{11}_{\theta,2}(u_{ij},u_{ik})\right\}^{\delta_{ij}\delta_{ik}}\times\left\{\mathcal{C}^{10}_{\theta,2}(u_{ij},u_{ik})\right\}^{\delta_{ij}(1-\delta_{ik})}\times \\ &\quad \left\{\mathcal{C}^{01}_{\theta,2}(u_{ij},u_{ik})\right\}^{(1-\delta_{ij})\delta_{ik}}\times\left\{\mathcal{C}_{\theta,2}(u_{ij},u_{ik})\right\}^{(1-\delta_{ij})(1-\delta_{ik})} \end{aligned}$$

and $\mathcal{C}^{lm}_{\theta,2}(u_1,u_2)=\frac{\partial^{l+m}\mathcal{C}_{\theta,2}(u_1,u_2)}{\partial^l u_1\partial^m u_2}, l,m=0,1.$

Discussion of the estimation of the variance of $\hat{\theta}$ is delayed to the end of the next section.

2.5.2 Computation of the p-value

In this section, we derive a semiparametric procedure to approximate the p-value associated with K. The bootstrap procedure involves generating datasets having the same properties as the original dataset under the null hypothesis $H_0 : S_0 = S_1 = S_2$. The data resampling process must then follow these rules:

1. The generated datasets have the same number of clusters n, clusters sizes $d_1,\ldots,d_n$ and genotypes $\{G_{ij}, i=1,\ldots,n, j=1,\ldots,d_i\}$ as the original dataset.
2. The ages at onset of cancer of individuals from the same cluster are generated jointly according to $\mathcal{C}_{\hat{\theta}}$, where $\hat{\theta}$ is estimated as detailed in the previous section.
3. In order to mimic the behavior under the null hypothesis, all ages at onset of cancer are marginally generated according to $\hat{S}_P$.
4. The censoring times of individuals from the same cluster are generated independently from each other and following marginally $\hat{R}_P$, the weighted Kaplan-Meier estimator of the survival function of the censoring variable.
5. The generated datasets are non-random. In order to mimic the ascertainment scheme of the original dataset with respect to X and δ, an estimate of the sampling probability

Table 2.2 – Algorithm generating a bootstrapped dataset

- For $i = 1, \ldots, n$
  - Generate the first variate ($j = 1$)
    - Repeat until SUCCESS = 1
      - Generate $U_{i1}, Z_{i1} \sim U[0,1]$
      - Set $T^*_{i1} = \hat{S}_P^{-1}(U_{i1})$ and $C^*_{i1} = \hat{R}_P^{-1}(Z_{i1})$
      - Set $X^*_{i1} = \min(T^*_{i1}, C^*_{i1})$ and $\delta^*_{i1} = I(T^*_{i1} < C^*_{i1})$
      - SUCCESS $\sim$ Bernoulli$\{\hat{\mathcal{P}}(X^*_{i1}, \delta^*_{i1})\}$
    - Save U_{i1}, X^*_{i1} and δ^*_{i1}
  - If $d_i \geq 2$, then for $j = 2, \ldots, d_i$
    - Repeat until SUCCESS = 1
      - Generate $V_{ij}, Z_{ij} \sim U[0,1]$
      - Solve $\mathcal{H}_{\hat{\theta},j}(U_{i1}, \ldots, U_{ij}) = V_{ij}$ to obtain U_{ij}.
      - Set $T^*_{ij} = \hat{S}_P^{-1}(U_{ij})$ and $C^*_{ij} = \hat{R}_P^{-1}(Z_{ij})$
      - Set $X^*_{ij} = \min(T^*_{ij}, C^*_{ij})$ and $\delta^*_{ij} = I(T^*_{ij} < C^*_{ij})$
      - SUCCESS $\sim$ Bernoulli$\{\hat{\mathcal{P}}(X^*_{ij}, \delta^*_{ij})\}$
    - Save U_{ij}, X^*_{ij} and δ^*_{ij}

function $\mathcal{P}$ is required. This probability function is inversely proportional to the weight function and needs only to be specified up to a multiplicative constant such that all its values fall into $[0,1]$. Therefore, $\mathcal{P}$ can be estimated by $\widehat{\mathcal{P}}(x,\delta) = (1/\widehat{\Phi}(x,\delta))/M$, where $M = \max\{1/\widehat{\Phi}(x,\delta), x \in [0,\infty), \delta = 0,1\}$.

Combining all these facts together with the algorithm described at the end of Section 2.4 yields the semiparametric procedure of Table 2.2 that generates a bootstrapped dataset $\{(X^*_{ij}, \delta^*_{ij}, G_{ij}), i = 1, \ldots, n; j = 1, \ldots, d_i\}$ under $H_0 : S_0 = S_1 = S_2$, from which it is possible to compute the Cramer–von Mises statistic K^* using (2.3). Repeating this B times yields $K^*_1, \ldots, K^*_B$ and the p-value associated with K is approximated by $\sum_{b=1}^{B} I(K^*_b > K)/B$.

Note that substituting $\hat{S}_P$ by $\hat{S}_{g_{ij}}$ in the algorithm described in Table 2.2 enables one to generate a sample having the same properties as the original dataset but not necessarily under $H_0 : S_0 = S_1 = S_2$. From such a sample, it is possible to compute an estimator $\hat{\theta}^*$ for the parameter of the copula. Repeating this B times yields $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ and the variance of $\hat{\theta}$ can be estimated by $\sum_{b=1}^{B}(\hat{\theta}^*_b - \bar{\theta}^*)^2/(B-1)$, where $\bar{\theta}^* = \sum_{b=1}^{B} \hat{\theta}^*_b/B$.

All the procedures described above were implemented using the R software (R Core Team, 2015). An example data set and the R code implementing the semiparametric bootstrap

procedure for the Cramer-von Mises test can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2.6 Simulation studies

Simulations were carried out to evaluate the empirical properties of the proposed inference procedures. The simulation conditions were chosen as close as possible to those of the real dataset described in Section 2.2.

Samples of $n = 2000$ clusters were generated as follows. For each cluster i, the size d_i was generated from a zero-truncated Poisson distribution with a parameter λ equal to either 1.4 or 1.8. We assumed that the members of each cluster are brothers and sisters, in which case the joint distribution of genotypes $\{G_{i1}, \ldots, G_{id_i}\}$ is given by Elandt-Johnson (1971); see Appendix C for details. The minor allele frequency was set to 0.2 and the simulations were performed under the Hardy-Weinberg assumption.

Clustered failure times were generated according to a Clayton copula whose parameter was chosen such that the corresponding Kendall's τ is equal to either 1/5 or 1/3. Given the genotypes, the marginal distribution of the failure times was Weibull with a scale parameter equal to 50. The shape parameter varied according to the genotype: $8 - \gamma, 8$ or $8 + \gamma$ for 0, 1 or 2 copies of the minor allele, respectively with $\gamma \in [0, 4]$. Note that $\gamma = 0$ corresponds to $H_0 : S_0 = S_1 = S_2$ whereas $\gamma > 0$ indicates a departure from the null hypothesis.

On the other hand, the censored variables were simulated, independently from each other, according to a Weibull distribution with a shape and a scale parameter equal to 50 and 8 respectively.

We introduced non-randomness via a piecewise constant sampling probability function $\mathcal{P}(x, \delta)$ equal to $(1, 1, 0.5, 0.5, 0.5)$ on the intervals $[0, 35), [35, 40), [40, 45), [45, 50)$ and $[50, \infty)$ if $\delta = 1$ and $(0.5, 0.5, 0.25, 0.25, 0.25)$ on the same intervals if $\delta = 0$. This leads to an under-representation of the censored subjects. The true incidence rates required for the calculation of the weights were estimated from a simulated population of 100,000 observations for each scenario of simulations.

2.6.1 Results for $\hat{\tau}$

We ran a first set of simulations to evaluate the estimation procedure of the parameter of the copula θ, or equivalently the Kendall's τ, presented in Section 2.5.1. Observations were generated with $\gamma = 3$. Figure 2.1 displays the resulting true genotype-specific survival curves. For each combination of the true values of λ and τ, we generated 1000 samples and for

Table 2.3 – Performance of Kendall's τ estimator (1000 replications)

λ	τ	Mean($\hat{\tau}$)	$SD(\hat{\tau})$	Mean($\widehat{SD}$)	95% Cov.
1.4	.2	.2067	.0383	.0395	94.7
	.33	.3429	.0326	.0330	93.9
1.8	.2	.2073	.0275	.0278	93.2
	.33	.3454	.0232	.0237	92.3

Note: $SD(\hat{\tau})$ = empirical SD and Mean($\widehat{SD}$) = mean estimated SD

Table 2.4 – Rejection rates under H_0 (1000 replications)

λ	τ	SPB	CPT	wtCox PA	wtCox S
1.4	.2	.033	.072	.056	.072
	.33	.047	.060	.053	.062
1.8	.2	.041	.084	.064	.067
	.33	.036	.088	.061	.061

Note:
SPB = Cramer–von Mises - semiparametric bootstrap
CPT = Cramer–von Mises - clustered permutation test
wtCox PA = weighted Cox - per-allele HR
wtCox S = weighted Cox - separate HR

each simulated dataset, we computed $\hat{\theta}$ and $\hat{\tau}$ and obtained an estimate of the standard deviation of $\hat{\tau}$ via the bootstrap procedure with 100 replicates. The empirical mean and standard deviation as well as the mean of the estimated standard deviation and the empirical coverage rate of the 95% confidence interval are reported in Table 2.3. This table shows that the proposed estimator is virtually unbiased, the standard deviations are well estimated and the 95% confidence intervals have reasonable empirical coverage rates.

2.6.2 Results for the Cramer–von Mises statistic

A second set of simulations was performed to compare the proposed test with the existing approaches. Let us consider first the case $\gamma = 0$, corresponding to H_0. For each combination of the true values of λ and τ, we generated 1000 samples. For each simulated dataset, we computed the p-values for (*i*) the proposed Cramer–von Mises type statistic from $B = 1000$ bootstrapped samples (*ii*) the Wald statistic based on the weighted Cox models in (2.1) and (2.2) and (*iii*) the clustered permutation test from $B = 1000$ permuted samples. The percentages of rejection of the null hypothesis at the nominal level of 0.05 are reported in Table 2.4. This table reveals that the proposed test tends to be somewhat conservative whereas the clustered permutation test seems to have an inflated type I error. On the contrary, the rejection rates of the tests based on weighted Cox models are close to the nominal level.

Varying the value of γ within the interval (0, 4] corresponds to scenarios of departures from the null hypothesis with crossing survival curves. For $\lambda = 1.4$ and $\tau = 1/3$, we computed the

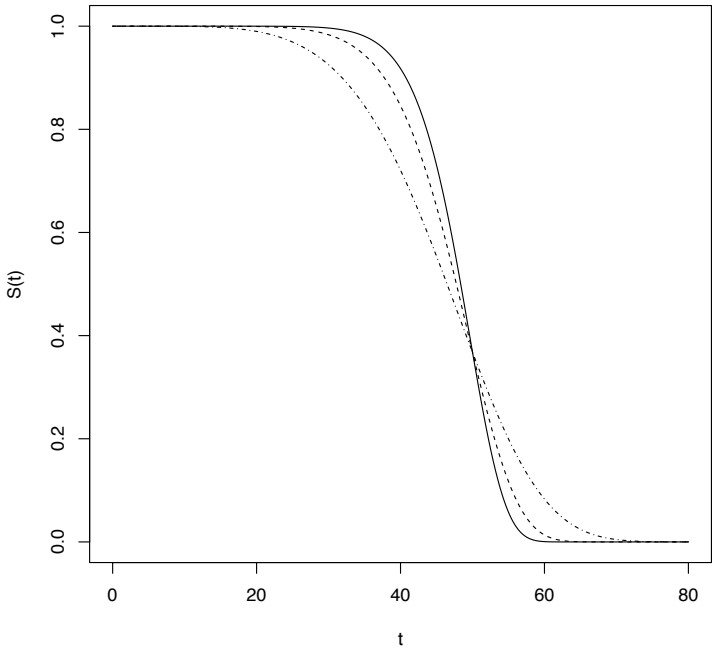

Figure 2.1 – True survival curves used to test the performance of Kendall's τ estimator. The curves differ according to the underlying genotype: 0 [dashed-dotted line], 1 [dotted line] or 2 [solid line] copies of the minor allele.

empirical power of all considered tests for various values of γ. The *LOWESS* smoother of the R software (R Core Team, 2015) was used to draw the curves reported in Figure 2.2 from 21 datapoints. This figure shows that the proposed test clearly outperforms the clustered permutation test and the ones based on the weighted Cox model when the proportional hazards assumption is not fulfilled.

2.7 Results from an application to the BRCA2 data

In the first step of our analysis, the weights are computed based on the true incidence rates of breast cancer for women with a mutation of BRCA2 reported in Table III of Antoniou et al. (2005). The weights obtained for censored subjects range from 1.18 to 2.54 while those for observed subjects vary between 0.20 and 0.67, indicating that affected women are over-represented in the sample. In Figure 2.3, we report the genotype-specific weighted log cumulative hazard functions. These curves are crossing, which suggests that models (2.1) and (2.2) are not appropriate for the data analyzed in the present paper. On the other hand, the weighted Kaplan–Meier estimates reported in Figure 2.4 show that the GG genotype group has the greatest breast cancer risk, especially beyond the age of 60.

Maximizing the pairwise likelihood presented in Section 2.5.1 yields $\hat{\theta} = 0.78$ (s.e. = 0.50), which corresponds to an estimated Kendall's τ equal to 0.28 (s.e. = 0.10). This highlights a residual phenotype dependence which may be explained by other genetic, environmental,

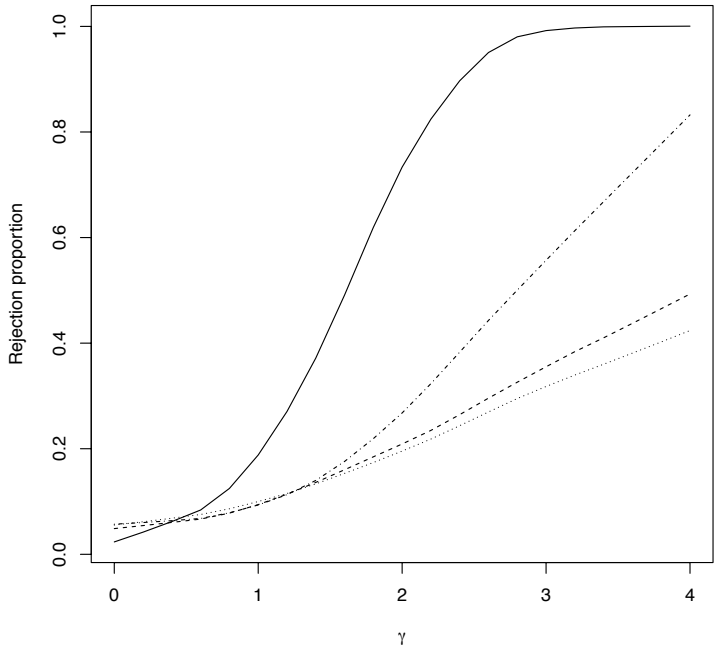

Figure 2.2 – Power comparison between the Cramer–von Mises test with semiparametric bootstrap [solid line], the clustered permutation test [dashed-dotted line] and the Wald test under the weighted Cox model with a common [dashed line] or separate hazard ratios [dotted line]. The γ variable is an indicator of the difference in the shape of the genotype-specific survival curves.

lifestyle, reproductive and hormonal factors. Common modifier alleles identified to date explain a small proportion of the genetic variability in breast or ovarian cancer risk for mutation carriers.

Applying the proposed semiparametric bootstrap procedure with $B = 1000$ bootstrap samples yields a p-value equal to 0.007, which reveals an association between the considered SNP and breast cancer, without assuming a proportional hazards model. On the other hand, the Wald statistics in the weighted Cox models in (2.1) [per-allele HR] and (2.2) [separate HR] lead to p-values of 0.030 and 0.083, respectively. Finally, with the clustered permutation test, we obtain a p-value equal to 0.022.

In order to check the adequacy of the Clayton copula for the data analyzed here, we considered a generalization of the Clayton copula given by

$$C_{\theta,\xi,D}(u_1,\ldots,u_D) = \left\{ \left[\sum_{j=1}^{D} (u_j^{-\theta} - 1)^{\xi} \right]^{1/\xi} + 1 \right\}^{-1/\theta}, \theta > 0, \xi \geq 1, \tag{2.5}$$

which reduces to (2.4) when $\xi = 1$. We fitted (2.5) and obtained $\hat{\xi} = 1.074$. In order to perform the test $H_0 : \xi = 1$ versus $H_1 : \xi > 1$, we approximated the distribution of $\hat{\xi}$ when the true ξ equals 1 by generating $B = 1000$ samples via the modified version of the semiparametric

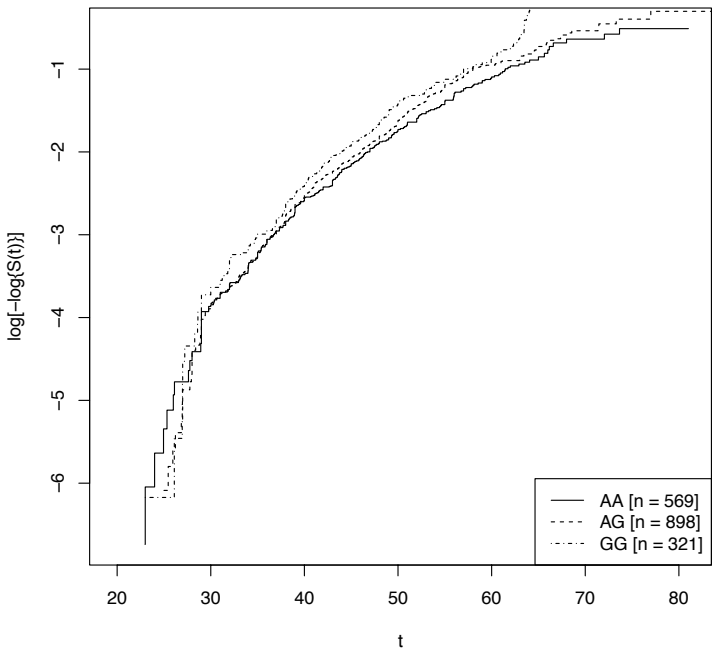

Figure 2.3 – Estimated log cumulative hazard functions for the three genotypes of the SNP rs13281615

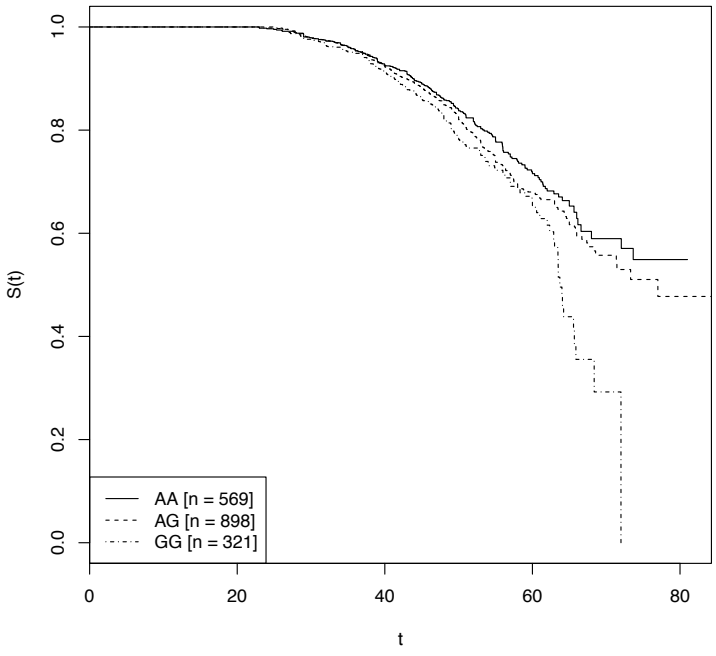

Figure 2.4 – Weighted Kaplan-Meier estimates of the genotype-specific survival functions for the SNP rs13281615

bootstrap procedure that uses $\hat{S}_{g_{ij}}$ instead of $\hat{S}_P$. For each generated sample, we computed $\hat{\xi}^*$ and obtained a p-value equal to $\sum_{b=1}^{B} I(\hat{\xi}_b^* > \hat{\xi})/B = 0.14$. Therefore, ξ is not significantly greater than one, which suggests that the Clayton copula is suitable for this dataset.

2.8 Discussion

This paper presents a semiparametric bootstrap procedure to approximate the p-value of an association test between SNP rs13281615 and breast cancer risk among BRCA2 mutation carriers. The proposed bootstrap procedure involves generating samples of correlated failure times data according to a non-random design. The considered Cramer–von Mises type test statistic presents an interesting alternative when the proportional hazards assumption is questionable. It has specifically been designed to provide greater power to detect crossing hazard rates (Klein and Moeschberger, 2003), which is confirmed by Figure 2.2. It is a more appropriate tool for the situation described in Figure 2.4, in comparison with other statistics such as the log-rank or Gehan test.

The test is based on the Clayton copula, which implicitly assumes that the correlation between the ages at onset of cancer of any pair of members of the same family given their genotypes is constant. The goodness-of-fit test performed at the end of Section 2.7 suggests that this copula is suitable for our dataset. Nevertheless, if there is an evidence against this assumption, one may consider an alternative copula such as the normal copula where the dependence between the ages at onset of cancer of a pair of members of the same family can be expressed in terms of their identical-by-descent sharing. A vector of dependence parameters for the copula could model the different relative types and/or coefficients of the kinship matrix. However, this requires the knowledge of the relationships between family members and may involve complex computations.

Figure 2.4 indicates a high proportion of subjects for which breast cancer diagnosis did not occur up to age 80. In clinical settings, the nonzero proportion at which a survival curve levels off is called the cure fraction, while in genetic studies one may refer to the proportion of nonsusceptibles (Chen and Lu, 2012). Future work may be directed towards the development of methods to test simultaneously the equality of the proportion of nonsusceptibles and the conditional survival curves of the susceptible subjects in the non-random setting of the BRCA1/2 data.

Transition

Dans le chapitre 2, un test d'association génétique entre un seul SNP et une durée de vie a été développé en présence de corrélation intra-famille. Chez les populations de porteuses de mutations sur les gènes BRCA1/2, les SNPs identifiés jusqu'à maintenant expliquent une faible proportion de la variabilité génétique observée pour le risque de cancer du sein. Devant la multitude de tests statistiques effectués, les tailles d'effet plutôt modestes et la disponibilité de variants rares pour l'analyse, l'intérêt se porte maintenant vers le développement d'outils statistiques qui permettent de tester l'association pour un ensemble de SNPs plutôt qu'un seul SNP à la fois. Le prochain chapitre s'attaque à ce problème dans le cas de durées de vie en grappes. La matrice de parenté permet de tenir compte de la corrélation intra-famille.

Chapter 3

SNP set association testing for survival outcomes in the presence of intrafamilial correlation

[Test d'association entre un ensemble de SNPs et des issues de survie en présence de corrélation intra-famille]

Résumé

Dans le cadre de ce travail, un test d'association entre un ensemble de SNPs et des phénotypes censurés est proposé en présence d'un devis d'étude basé sur des familles. Le test proposé est valide pour les variants communs et les variants rares. Un modèle de Cox à risques proportionnels est défini pour la distribution marginale du trait et la dépendance familiale est modélisée au moyen d'une copule gaussienne. Les valeurs censurées sont considérées comme des données partiellement manquantes et une procédure d'imputation multiple est proposée afin de calculer la statistique de test. La valeur de p est ensuite déduite de façon analytique. Les propriétés empiriques de la méthode proposée sont évaluées dans le cas d'échantillons de taille finie et comparées aux approches concurrentes existantes par simulation. L'utilisation du nouveau test est illustrée au moyen d'un jeu de données sur le cancer du sein du *Consortium of Investigators of Modifiers of BRCA1 and BRCA2*.

Abstract

In this work, we propose a SNP-set association test for censored phenotypes in the presence of a family-based design. The proposed test is valid for both common and rare variants. A proportional hazards Cox model is specified for the marginal distribution of the trait and the familial dependence is modelled via a Gaussian copula. Censored values are treated as partially missing data and a multiple imputation procedure is proposed in order to compute the test statistic. The p-value is then deduced analytically. The finite-sample empirical properties of the proposed method are evaluated and compared to existing competitors by simulations and its use is illustrated using a breast cancer data set from the Consortium of Investigators of Modifiers of BRCA1 and BRCA2. [1]

3.1 Introduction

The advent of new genomic sequencing technologies is shifting the biostatistical community to new paradigms with millions of SNPs including low frequency and rare gene variants becoming available for association studies. In recent years, it has been well established that *(i)* the effect sizes for the common variants are typically small to moderate, *(ii)* the low frequency and rare variants explain a portion of missed heritability for complex diseases, and *(iii)* many phenotypes are related to multiple genetic variants through complex dependence structures. Identification of causal variants can help to better understand the biology of disease development and lead to prevention and treatment strategies particularized to each human being via risk prediction models. Consequently, many attempts have been made to develop multimarker association tests that can test jointly multiple common and/or rare variants; see Wang et al. (2013) for a recent review. Such tests are typically more powerful than standard testing strategies based on multiple univariate single-marker analysis with a correction for multiple testing (Han and Pan, 2010; Thomas et al., 2013).

The first developed approaches for testing rare variants are mainly based on collapsing the entire region into a unique score, whose association with the phenotype is then tested (Lee et al., 2014). However, these burden tests experience a significant loss of power when both deleterious and protective variants are present in the tested region. To overcome this drawback, Wu et al. (2011) developed a variance-components test termed the Sequence Kernel Association Test (SKAT), which is based on a mixed model. The kernel function measures the genetic similarity for pairs of subjects via the markers in the region under study. Several strategies have been proposed afterwards to combine the burden tests and the variance-components tests and the test statistics for the rare and the common variants to produce more powerful tests (Ionita-Laza et al., 2013; Derkach et al., 2013).

1. This is the accepted version of the following article: Leclerc, M., The Consortium of Investigators of Modifiers of BRCA1/2, Simard, J. and Lakhal-Chaieb, L. (2015), SNP set association testing for survival outcomes in the presence of intrafamilial correlation. Genetic Epidemiology, 39: 406–414. doi: 10.1002/gepi.21914, which has been published in final form at `http://onlinelibrary.wiley.com/doi/10.1002/gepi.21914/abstract`.

Family-based designs have received a growing interest in genetics as including related individuals allows one to obtain larger samples of better quality and can potentially increase the power of association tests (Kazma and Bailey, 2011). However, failure to appropriately take into account the familial correlation may yield an inflated type 1 error and/or significant loss of power. Consequently, several authors developed region-based association tests in the presence of familial correlation (Oualkacha et al., 2013). Typically, the familial dependence is modelled via a random effect following a multivariate normal distribution with mean zero and covariance matrix proportional to the kinship matrix.

All the approaches cited above were developed for dichotomous and quantitative traits and very little attention has been paid to time-to-event outcomes in the presence of right-censoring. For instance, Cai et al. (2011) proposed a kernel machine score statistic to test the association between a genetic pathway and a survival outcome under a Cox regression framework. This approach was extended by Lin et al. (2011) to the context of GWAS survival studies. In both cases, a perturbation resampling procedure was employed to approximate the distribution of the test statistic under the null hypothesis. Alternatively, Chen et al. (2014) extended SKAT to survival outcomes. The p-value of the resulting test is computed analytically when a linear kernel is specified. However, these few methods for censored traits do not consider family-based designs.

In this paper, we propose a genomic region-based association test for censored traits in the presence of kinship data. A Cox model is specified for the survival outcome while the familial dependence is modelled via a Gaussian copula with a correlation matrix expressed in terms of the kinship matrix. Gaussian copulas have been regularly employed in genetics since the beginning of this century (Basrak et al., 2004; Li et al., 2006; He et al., 2012). Attractive features of these dependence structures include the ability to assume any correlation matrix to characterize associations between pairs of individuals and the mathematical convenience thanks to their close relationship with the multivariate normal distribution. In this work, censored observations are treated as missing data and a novel multiple imputation procedure is proposed to compute the test statistic. The p-value is then deduced analytically.

In our simulation studies, we assess the empirical performance of the proposed test and compare it to the existing approaches for censored traits in both cases of absence and presence of familial correlation. We also investigate the robustness of the proposed test to the misspecification of the Gaussian copula. In particular, we show that, unlike existing approaches, the proposed method controls very well type 1 error in the presence of familial dependence and is robust to the misspecification of the dependence structure. Finally, we illustrate our approach by analysing time-to-breast cancer phenotype in a large sample of women of European ancestry carrying a deleterious mutation on the BRCA1 gene.

3.2 Methods

3.2.1 Kernel score test for unrelated individuals

We begin by briefly reviewing the association tests developed by Lin et al. (2011) and Chen et al. (2014) in the case of unrelated individuals. Let T_i be the age-at-onset of the disease under investigation and G_i and X_i be row vectors of s genotypes and p non-genetic covariates, for individual i. In practice, the selection of the s SNPs forming the SNP set can be based on genes, biological pathways, LD blocks and recombination hot-spots, or rely on a sliding window method (Lin et al., 2011). The assumed proportional hazards model is

$$\lambda(t_i|G_i, X_i) = \lambda_0(t_i)e^{G_iW\beta+X_i\zeta}, \tag{3.1}$$

where λ and λ_0 are the conditional and the baseline hazard functions, respectively, $W = \text{diag}(w_1, \ldots, w_s)$ is an $s \times s$ diagonal matrix with the weights to be used for the s genotypes, and $\beta = (\beta_1, \ldots, \beta_s)^{'}$ and $\zeta = (\zeta_1, \cdots, \zeta_p)^{'}$ are vectors of regression coefficients. The null hypothesis of the association test is $H_0 : \beta_1 = \cdots = \beta_s = 0$. In the presence of right-censoring, one only observes n quadruplets $\{(Y_i, \delta_i, G_i, X_i), i = 1, \ldots, n\}$ where $Y_i = \min(T_i, C_i)$ is the observed failure time, $\delta_i = I(T_i < C_i)$ is the censoring indicator and C_i is the censoring variable, assumed to be independent from T_i. The test statistics of Lin et al. (2011) and Chen et al. (2014) are both expressed in terms of

$$Q_0 = M^{'}KM, \tag{3.2}$$

where $M = (M_1, \ldots, M_n)^{'}$ is the vector of martingale residuals estimated under the null model, $K = GWWG^{'}$, and G is an $n \times s$ matrix with rows G_i. The $n \times n$ matrix K is the weighted linear kernel whose entries $K_{ij} = \sum_{k=1}^{s} w_k^2 G_{ik}G_{jk}$ capture the similarities in pairs of individuals in the tested region. The association test of Lin et al. (2011) is generalized to any kernel matrix and its p-value is computed using resampling methods. On the other hand, the test of Chen et al. (2014) is limited to the weighted linear kernel and its p-value is computed analytically. These two tests are valid only for unrelated individuals. In what follows, we consider the case of censored outcomes in the presence of intrafamilial correlation.

3.2.2 Kinship-adjusted association model

Under a family-based design, the n individuals of the sample are clustered into families but for simplicity we do not introduce a second subscript. Let ϕ be the $n \times n$ matrix with entries reflecting the proportion of the genome that is IBD between pairs of individuals. For instance, one has $\phi_{ij} = 0$ if the individuals i and j are unrelated.

By Cheng et al. (1995), the proportional hazards model (3.1) can alternatively be written as

$$H(T_i) = -G_i W\beta - X_i\zeta + \varepsilon_i, \tag{3.3}$$

where $H(\cdot)$ is a unknown monotone increasing function and ε_i follows the extreme value distribution with cumulative distribution function (CDF)

$$F(\epsilon) = \mathbb{P}(\varepsilon_i \leq \epsilon) = \exp\{-\exp(-\epsilon)\}, -\infty \leq \epsilon \leq \infty$$

To model the familial correlation, we assume that the joint distribution of $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ follows a Gaussian copula C_Γ where Γ is a $n \times n$ correlation matrix. The diagonal elements of Γ are all equal to 1 and the off-diagonal entries are $\Gamma_{ij} = h^2\phi_{ij}$, where $h^2 \in [0, 1]$ measures the polygenic heritability. It is easy to see that $\Gamma = h^2\phi + (1 - h^2)I_n$, where I_n is the $n \times n$ identity matrix, provided $\phi_{ii} = 1$.

The likelihood ratio test and the Wald test require the estimation of β, which may not be possible in the presence of rare variants (Wu et al., 2011). On the other hand, the score test involves partial derivatives of the log-likelihood function, which is complicated and computationally demanding with censored observations in the presence of a family-based design. In this paper, we consider an alternative approach detailed below.

Let $q_i = \Phi^{-1}[F(\varepsilon_i)]$ be the inverse normal score for individual i. Under the proposed model, $q = (q_1, \ldots, q_n)'$ follows a multivariate normal distribution with mean zero and covariance matrix Γ. Moreover, under $H_0 : \beta = 0$, $q'Kq$ is distributed as a linear combination of independent chi-squared random variables

$$q'Kq \sim \sum_{l=1}^{L} \mu_l \chi^2_{l,1}, \tag{3.4}$$

where $0 < \mu_1 \leq \cdots \leq \mu_L$ are the L positive eigenvalues of $\Gamma^{1/2}K\Gamma^{1/2}$, $\chi^2_{l,1}, l = 1, \ldots, L$ denote independent chi-squared variables with 1 degree of freedom, and $\Gamma^{1/2}$ is the square root of Γ which can be computed by diagonalization (Rao, 1973, Section 3b.4). Motivated by (3.2) and (3.4), we propose the use of the test statistic $r'Kr$, where $r = (r_1, \ldots, r_n)'$ is a vector of appropriately defined residuals, computed under the null hypothesis.

3.2.3 Estimation under the null hypothesis

The functions and parameters involved in our model under the null hypothesis are H, ζ and h^2. These are estimated using a two-stage procedure in the spirit of Othus and Li (2010). In the first stage, estimates $\hat{H}$ and $\hat{\zeta}$ are obtained using the algorithm of Chen et al. (2002) under the working independence assumption. Computational details are given in Appendix D.

In the second stage, we estimate h^2 as follows. We begin by computing

$$\hat{q}_i = \Phi^{-1}\left[F\left(\hat{H}(Y_i) + X_i\hat{\zeta}\right)\right]$$

Actually, $\{\hat{q}_i, i = 1, \ldots, n\}$ is a censored version of $\left\{\Phi^{-1}\left[F\left(\hat{H}(T_i) + X_i\hat{\zeta}\right)\right], i = 1, \ldots, n\right\}$, whose joint distribution is approximately normal with mean zero and covariance matrix Γ. The censoring pattern is preserved as $\Phi^{-1}\left[F\left(\hat{H}(T_i) + X_i\hat{\zeta}\right)\right]$ is a non-decreasing transformation of T_i. The estimate $\hat{h}^2$ is then obtained by maximizing the likelihood function of the right-censored sample $\{(\hat{q}_i, \delta_i), i = 1, \ldots, n\}$. Computational details of the likelihood function involving a multivariate normal distribution in the presence of right-censoring are given in Appendix E.

3.2.4 Test statistic definition and p-value computation

Imputation approaches treating censored observations as missing data have become very popular in recent years (Liu et al., 2011; Lapidus et al., 2014). In this work, we propose to use an imputation procedure to replace the censored $\hat{q}_i$'s by imputed values in order to obtain a completed vector of residuals, whose joint distribution can be easily approximated.

Without loss of generality, we rearrange the indices $\{1, \ldots, n\}$ so that $\delta_1, \ldots, \delta_{n_1} = 1$ and $\delta_{n_1+1} = \cdots = \delta_n = 0$, where $n_1 = \sum_{i=1}^{n} \delta_i$ is the number of uncensored failure times. Consider the partition $\hat{q} = \begin{pmatrix} \hat{q}^{(1)} \\ \hat{q}^{(0)} \end{pmatrix}$ so that $\hat{q}^{(1)}$ and $\hat{q}^{(0)}$ are of lengths n_1 and $n - n_1$, respectively. Similarly, write $\hat{\Gamma} = \hat{h}^2\phi + (1 - \hat{h}^2)I_n = \begin{pmatrix} \hat{\Gamma}^{(11)} & \hat{\Gamma}^{(10)} \\ \hat{\Gamma}^{(01)} & \hat{\Gamma}^{(00)} \end{pmatrix}$. A completed vector of residuals has the form $r = \begin{pmatrix} \hat{q}^{(1)} \\ \tilde{q}^{(0)} \end{pmatrix}$, where the vector of imputed values $\tilde{q}^{(0)}$ is generated from the posterior distribution of the uncensored version of $\hat{q}^{(0)}$ given the observed values $\hat{q}^{(1)}$ with the restriction to be larger than the original censored values, componentwise, i.e. $\tilde{q}^{(0)}$ is a random draw from a truncated multivariate normal distribution with mean $\hat{\Gamma}^{(01)}{\hat{\Gamma}^{(11)}}^{-1}\hat{q}^{(1)}$, covariance matrix $\hat{\Gamma}^{(00)} - \hat{\Gamma}^{(01)}{\hat{\Gamma}^{(11)}}^{-1}\hat{\Gamma}^{(10)}$ and support $[\hat{q}_{n_1+1}, \infty] \times \cdots \times [\hat{q}_n, \infty]$.

Typically, imputation procedures require the generation of multiple completed data sets. Separate analyses are performed on each completed data set and the results are aggregated afterwards using the methodology of Rubin (1987). However, this strategy is mainly useful when the quantity of interest (e.g. parameter estimator or test statistics) follows a normal distribution, which is not the case here since the considered test statistic is a quadratic form following a mixture of chi-squared variables.

In this work, we consider an alternative approach. We generate m completed vectors of residuals and compute their mean, which we denote $r^{(m)}$ in the sequel. In Appendix F, we show that $r^{(m)}$ follows approximately a multivariate normal distribution with mean zero and co-

variance matrix equal to $\rho^{(m)}\hat{\Gamma}$, where $\rho^{(m)}$ is a scale parameter, which reflects the fact that we are using multiple imputed values rather than real observations. This parameter is estimated by its maximum likelihood estimator $\hat{\rho}^{(m)} = r^{(m)'}\hat{\Gamma}^{-1}r^{(m)}/n$. The test statistics is then $Q = r^{(m)'}Kr^{(m)}$. It is distributed under the null hypothesis as $r^{(m)'}Kr^{(m)} \sim \sum_{l=1}^{L} \mu_l \chi^2_{l,1}$, where $0 < \mu_1 \leq \cdots \leq \mu_L$ are the L positive eigenvalues of $\rho^{(m)}\hat{\Gamma}^{1/2}K\hat{\Gamma}^{1/2}$. The p-value is then obtained by employing the Davies approximation (Davies, 1980), which is based on the numerical inversion of the characteristic function of Q and implemented in the R package CompQuadForm (Duchesne and Lafaye De Micheaux, 2010). To summarize, the proposed testing procedure is performed following these steps:

1. Estimate H and ζ by the algorithm of Chen et al. (2002).
2. Estimate h^2 by the algorithm of Othus and Li (2010) and compute $\hat{\Gamma} = \hat{h}^2\phi + \left(1 - \hat{h}^2\right) I_n$.
3. Rearrange the indices $\{1, \ldots, n\}$ and deduce the partitioned expressions of $\hat{q}$ and $\hat{\Gamma}$.
4. Generate m completed vectors of residuals and compute their mean $r^{(m)}$ and the test statistic $Q = r^{(m)'}Kr^{(m)}$.
5. Compute $\hat{\rho}^{(m)} = r^{(m)'}\hat{\Gamma}^{-1}r^{(m)}/n$ and deduce the positive eigenvalues of $\hat{\rho}^{(m)}\hat{\Gamma}^{1/2}K\hat{\Gamma}^{1/2}$.
6. Obtain the p-value by the Davies approximation.

All the procedures described above were implemented by using the R software (R Core Team, 2015). Specific code is available upon request from the first author.

3.2.5 Numerical simulations

Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test. Samples of $n = 600$ individuals from 120 families were generated: 40 families of two parents and one child, 40 families of two parents and two children, and 40 families of three generations (two grand-parents, four parents, and two grandchildren). The coefficients of the block diagonal IBD matrix were fixed at their expected theoretical values. The number of biallelic gene variants was set to $s = 10$. The minor allele frequencies were randomly sampled from Unif(0.001, 0.1). We used the SIMULATE3 program (Terwilliger et al., 1993) to simulate the genotypes of the 600 individuals, assuming a linkage disequilibrium corresponding to a squared correlation coefficient of $r^2 = 0.5$ between consecutive gene variants. Two non-genetic covariates were generated for each individual: $X_{i1} \sim$ Unif(-0.2,0.2) and $X_{i2} \sim$ Bernoulli(0.5). The error terms $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ were generated according to a joint distribution defined by an extreme value marginal distribution and a dependence structure specified as either a Gaussian copula C_Γ, a Student copula $t_{5,\Gamma}$ or a correlated log-normal frailty model with covariance matrix $\frac{1}{2}\Gamma$. The failure times were then computed according to (3.3) with $H(t) = \log(t)$ and $\zeta = (1, 1)'$. The censored variables were either set equal to infinity or simulated, independently from each other, according to a Weibull distribution

with a shape parameter equal to 1.5 and a varying scale parameter controlling the censoring rate. The observed failure times $(\tilde{T}_1, \ldots, \tilde{T}_n)$ and censoring indicators $(\delta_1, \ldots, \delta_n)$ were then deduced. For each simulations scenario, 10,000 samples were generated and for each simulated data set, we estimated h^2 under the null hypothesis $H_0 : \beta = 0$ using the theoretical IBD matrix and computed the p-value of the proposed method with $m = 50$ as well as those of the approaches of Chen et al. (2014) and Lin et al. (2011). The latter test was performed with $B = 1,000$ permutations. In all settings, the weights $w_1, \ldots, w_s$ were defined as the density function of the Beta(1,25) evaluated at the minor allele frequency.

Type 1 Error Rate

We performed two sets of simulations under the null hypothesis by generating observations from (3.3) with $\beta = 0$. The aim of the first set is to investigate the type 1 error rate of the proposed test when the Gaussian copula is correctly specified. Four values of $h^2 : 0, .25, .5$, and .75 and three censoring rates: 0%, 25% and 50% were considered. The second set is intended to compare the performance of the proposed test to the approaches of Lin et al. (2011) and Chen et al. (2014) in terms of control of type 1 error rate and to assess the robustness of the proposed test when the clustering of the survival times is induced via a Student copula or log-normal frailties, or under a scenario of independence. For these simulations, the censoring rate was set to 50% and h^2 fixed at .5.

Power

Following the spirit of SKAT (Wu et al., 2011), observations under the alternative hypothesis were simulated from (3.3) with β generated following a multivariate normal distribution with mean zero and covariance matrix τI_s, for various values of the variance component term τ. For these simulations, the censoring rate was set to 50%, h^2 fixed at .5, m set to 50 and the familial dependence induced via one of the three dependence structures (the Gaussian and Student copulas and the correlated log-normal frailty model). The power was empirically estimated by the percentage of samples with a p-value smaller than the nominal level $\alpha = 5\%$.

3.2.6 Description of the BRCA1 data set

To illustrate the proposed kinship-adjusted association test, we analysed a data set from the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA) which aims to identify genetic factors associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. All of the 13,465 subjects are women of European ancestry who carry a mutation in the BRCA1 gene. These mutation carriers were recruited through cancer genetics clinics or research studies of high-risk families in 25 countries. The sample consists of 9,544 clusters, 24% having a size greater than one. Among these, the family size varies between 2 and 52, for a total of 6,250 subjects (46% of the sample).

The SNP set analysed comprises 111 variants across the TERT locus at 5p15.33 Build 36 (positions 1,280,693 - 1,414,669). SNPs were genotyped on the iCOGS (Collaborative Oncological Gene-environment Study) custom array. iCOGS methodology and quality control procedures are detailed elsewhere (Couch et al., 2013). The MAF of the SNPs ranges from .0134 to .500. The entries of the IBD matrix were estimated using the genotype data of the iCOGS array other than the tested TERT region (Amin et al., 2007; Leutenegger et al., 2003). The phenotype of each individual is defined by age at breast cancer diagnosis or age at last follow-up. The observations are right-censored if any of the following three events occurs before breast cancer diagnosis: ovarian cancer diagnosis, bilateral prophylactic mastectomy, or loss to follow-up. Censoring rate is equal to 49%.

3.3 Results

3.3.1 Type 1 Error Rate

QQ plots of p-values from 10,000 replications using the kinship-adjusted association test when dependence is induced via the Gaussian copula are shown in Figure 3.1. The p-values are close to the reference line of the uniform distribution across all the combinations of censoring rate and heritability parameter. The test tends to be somewhat conservative when $h^2 = 0$. Results of the estimation of h^2 and ζ are presented in Appendix G. Figure 3.2 presents QQ plots for all methods and dependence structures when $h^2 = .5$ and the censoring rate is equal to 50%. When the survival times are independent, the approach of Lin et al. (2011) is conservative. However, performing the simulations with common variants with MAF between .15 and .40 instead of rare or low frequency variants, the p-values were again close to the reference line of the uniform distribution (data not shown). This figure also suggests that both methods of Lin et al. (2011) and Chen et al. (2014) have a significantly inflated type 1 error in the presence of familial dependence whereas the type 1 error of the proposed method matches very well the expected distribution, even when the dependence structure is misspecified. As suggested by an anonymous referee, the empirical type 1 error rate of the proposed kinship-adjusted association test was further validated under other settings corresponding to $s \in \{10, 25, 50\}$ and $r^2 \in \{0, 0.5, 1\}$ and when the genetic variants are associated with the non-genetic covariates. The results, which are presented in Appendix G, confirm that the proposed kinship-adjusted test remains valid under these conditions.

3.3.2 Power

Since both methods of Lin et al. (2011) and Chen et al. (2014) have a significantly inflated type 1 error in the presence of a familial dependence, we only present the power results for the proposed method. In Figure 3.3, we report the empirical power as a function of the variance component τ for each of the three dependence structures at the nominal level $\alpha = 0.05$.

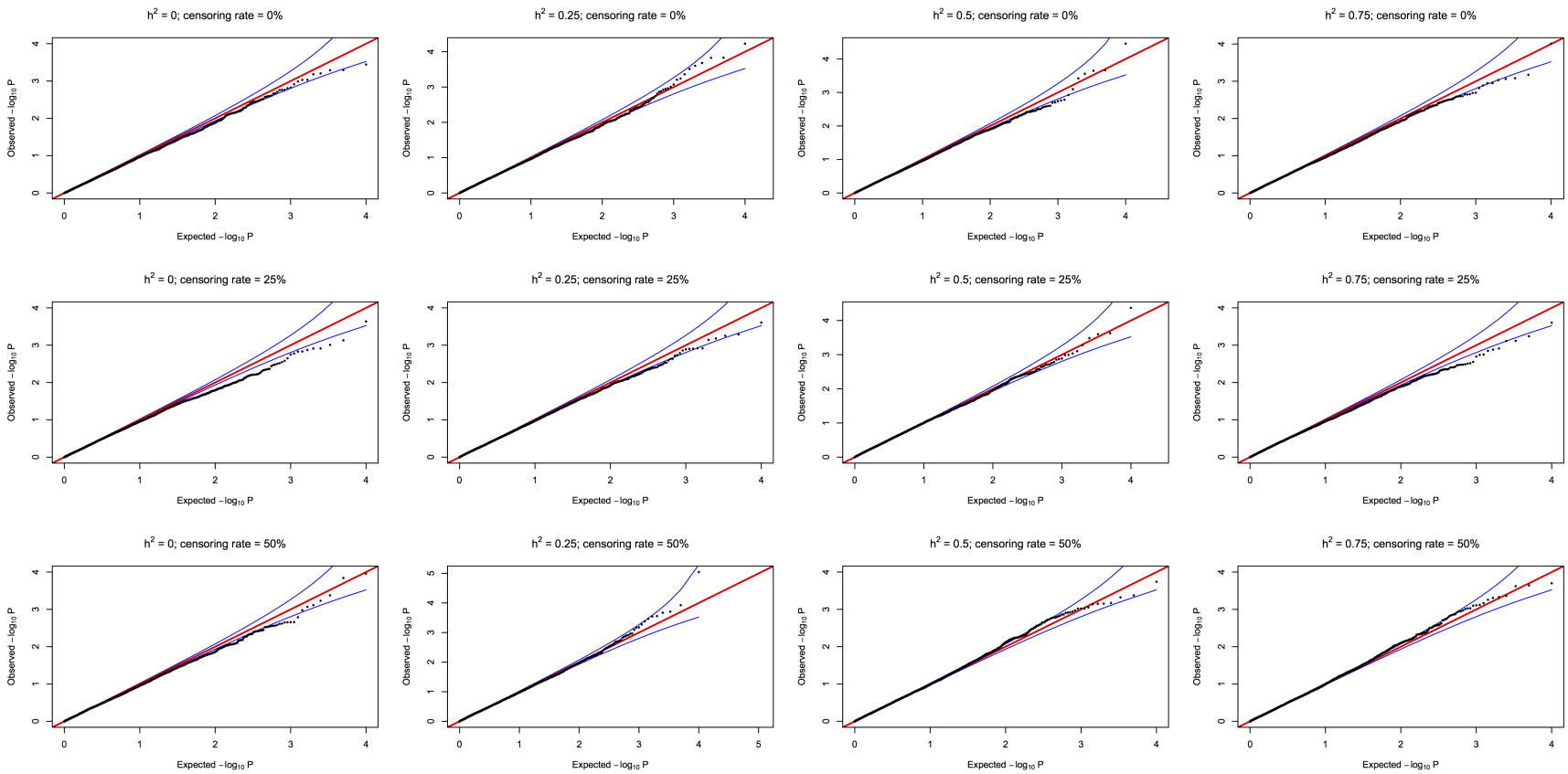

Figure 3.1 – QQ plots of p-values from 10,000 replications under H_0 using the kinship-adjusted association test (reference line of the uniform distribution in red along with the 95 percent confidence interval in blue). The dependence structure between survival times is induced via the Gaussian copula model. The three rows correspond to censoring rates of 0, 25 and 50%, respectively whereas the four columns refer to h^2 equal to 0, 0.25, 0.5 and 0.75 respectively.

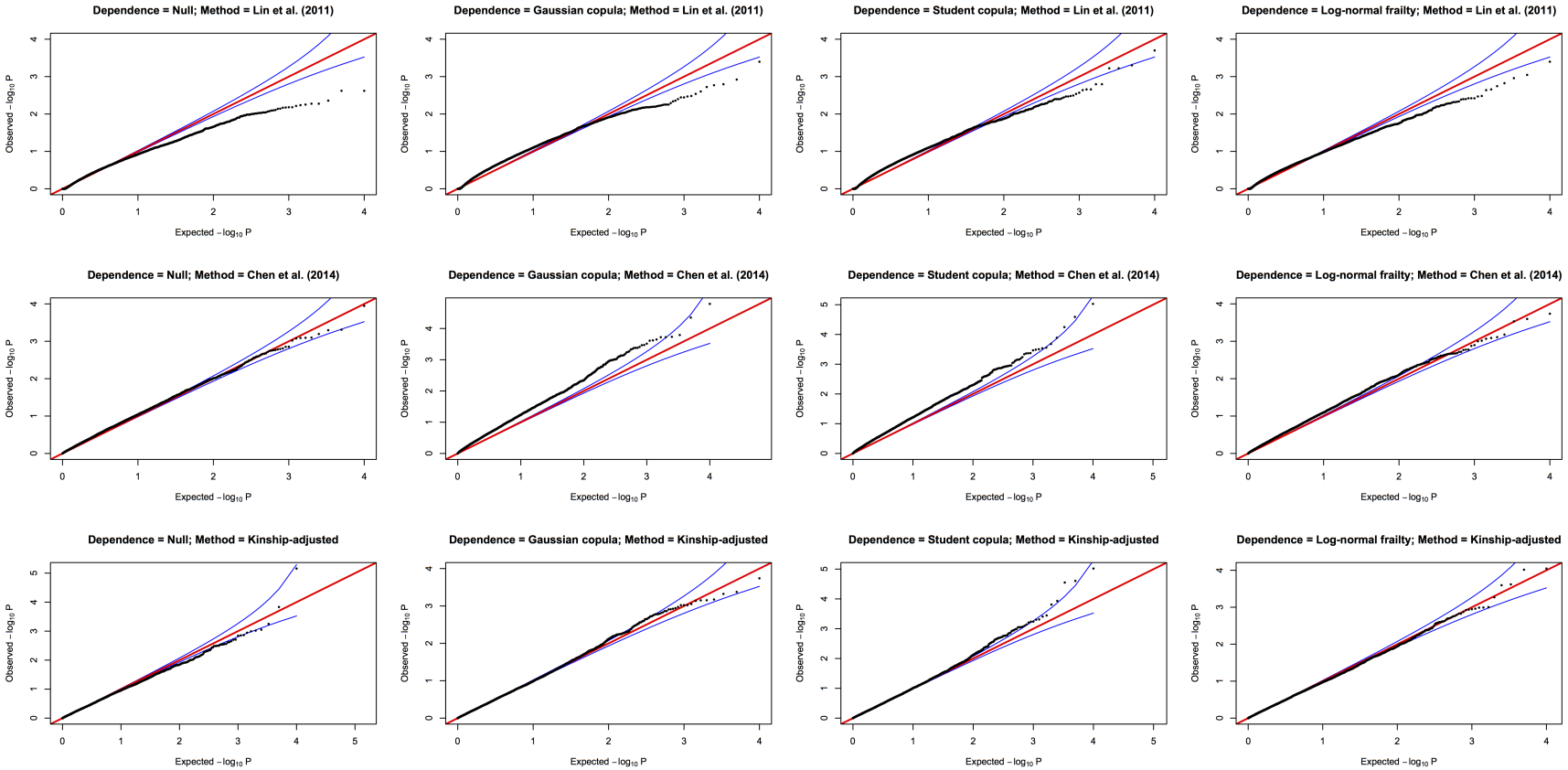

Figure 3.2 – QQ plots of p-values from 10,000 replications under H_0 comparing the kernel machine approach of Lin et al. (2011), the SKAT LRT statistic of Chen et al. (2014) and the proposed kinship-adjusted association test for four types of dependence between survival times. The three rows correspond to the three methods (Lin et al. (2011); Chen et al. (2014) and the proposed kinship-adjusted), whereas the four columns refer to four dependence structures (independence, Gaussian Copula, Student copula and Log-normal frailty terms).

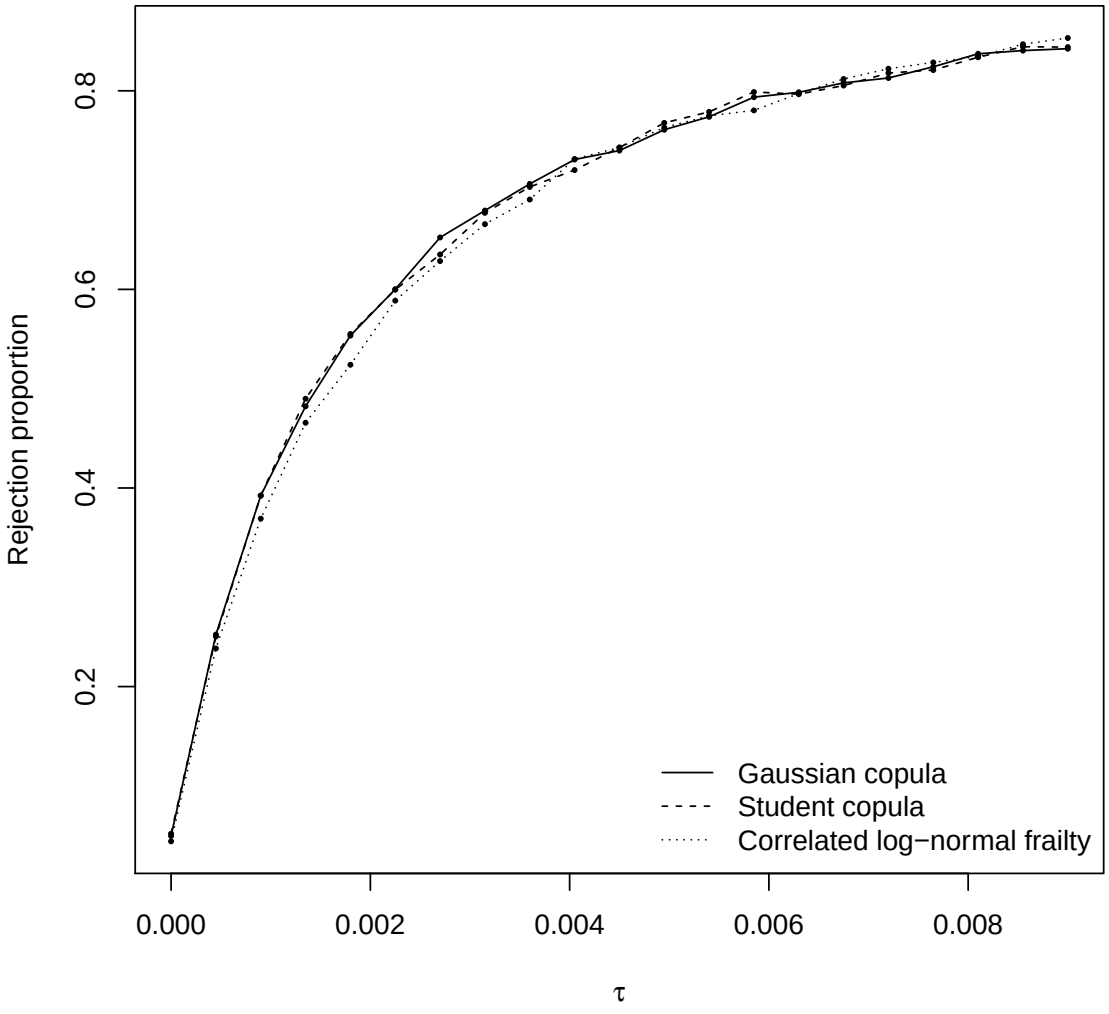

Figure 3.3 – Empirical power of the proposed test when the dependence structure is correct (Gaussian copula) and when the model is misspecified at the nominal level of 5%.

This Figure suggests that the power of the proposed test is robust to the misspecification of the Gaussian copula. Other simulations not shown here show that the powers of the three methods are very close in the absence of familial dependence. As suggested by an anonymous referee, we also investigated the impact of the number of imputations m on the power. The results, which are reported in Appendix G, show that power increases quickly with m and then levels off. These results also suggest that the use of $m = 50$ guarantees a reasonable power in practice.

3.3.3 Application to the BRCA1 data set

The analysis was conducted with the weights set equal to the density function of the Beta(1,25) evaluated at the minor allele frequency. Using the estimated kinship matrix, we obtained $\hat{h}^2 = .566$, which is a measure of the residual familial aggregation corresponding to the dependence between the ages at onset within the family that cannot be attributed to the SNP set under investigation. Such dependence may be explained by other genetic, environmental, lifestyle, reproductive and hormonal factors and has been reported elsewhere [e.g. Leclerc et al. (2015)]. Applying the proposed kinship-adjusted association test with $m = 50$ yields a p-value equal to 2.61×10^{-4}, which indicates evidence of association between the SNPs of the TERT locus and breast cancer risk. Such evidence is weaker for the approaches of Lin et al. (2011) with 1,000 perturbations and Chen et al. (2014) for which we obtain p-values equal to 3.18×10^{-2} and 2.96×10^{-2} respectively. Bojesen et al. (2013) analysed the same data set by considering separate single SNP tests. Using a family-based design test, they detected an association with SNP rs10069690 [HR = 1.16; P = 4.8×10^{-13}, MAF=0.275]. In order to deepen our analysis and to correct for this already established association, we treated this SNP as a covariate and performed association tests for the remaining 110 SNPs of the TERT locus. The estimate of ζ is 0.138, which is in agreement with the results of Bojesen et al. (2013) as $\exp(0.138) = 1.15$. The obtained p-values for the proposed method and the approaches of Lin et al. (2011) and Chen et al. (2014) are 3.85×10^{-3}, 4.13×10^{-1} and 2.86×10^{-2}, respectively. The proposed approach yields a stronger evidence of residual association with the TERT locus, beyond SNP rs10069690. We then used a sliding window method to further examine evidence of association across the SNP set and compare the three methods. Each window includes 25 SNPs, with 15 SNPs each overlapping with the previous and subsequent windows, except for the last window which includes 21 SNPs. Results for the 10 sliding windows with SNP rs10069690 as a covariate (thus 3 windows with 24 SNPs only) are reported in Table 3.1. After adjusting for multiple testing using a stringent procedure such as the Bonferroni correction (experiment-wise significance level of $0.05/10 = 5 \times 10^{-3}$), the only method in Table 3.1 showing evidence of residual association with breast cancer risk beyond SNP rs10069690 is the proposed kinship-adjusted association test. These results highlight the importance of properly modelling the dependence among related individuals.

3.4 Discussion

This paper presents a genomic region association test for survival outcomes, with adjustment for familial relatedness. A multiple imputation procedure that treats censored observations as missing data is proposed to compute the test statistic. The proposed approach is convenient for GWAS as the multiple imputation procedure has to be performed only once and the p-values are computed analytically.

The proposed test is valid for both rare and common variants. However, it may be more

Table 3.1 – Sliding window study results on time-to-breast cancer diagnosis for BRCA1 mutation carriers at the TERT locus with SNP rs10069690 as a covariate

SNPs	Win. size	Start	Stop	Lin et al. (2011)	Chen et al. (2014)	Kinship-adjusted
1-25	25	1280693	1316408	9.11×10^{-2}	7.22×10^{-3}	2.95×10^{-4}
11-35	24	1285775	1333477	2.94×10^{-1}	6.37×10^{-2}	5.72×10^{-4}
21-45	24	1308520	1345983	7.92×10^{-1}	1.39×10^{-1}	9.85×10^{-2}
31-55	24	1329873	1353070	4.16×10^{-1}	5.09×10^{-2}	4.62×10^{-1}
41-65	25	1340290	1368343	4.17×10^{-1}	1.22×10^{-1}	6.34×10^{-1}
51-75	25	1351782	1373957	9.15×10^{-1}	2.91×10^{-1}	7.21×10^{-2}
61-85	25	1366242	1383840	9.56×10^{-1}	3.93×10^{-1}	8.01×10^{-2}
71-95	25	1372353	1399081	8.22×10^{-1}	2.65×10^{-1}	6.28×10^{-2}
81-105	25	1378590	1409450	6.15×10^{-1}	1.85×10^{-1}	9.11×10^{-2}
91-111	21	1390070	1414669	3.49×10^{-1}	1.73×10^{-1}	4.07×10^{-2}

convenient to combine separate test statistics for each type of variants (Ionita-Laza et al., 2013). Investigating procedures to compute and combine these statistics in order to obtain more powerful tests in the setting of censored outcomes with a family-based design is subject of ongoing research.

Our imputation procedure ignores the variability due to the fact that we are using estimates of H, ζ and h^2 instead of the unknown real values. Simulations show that this approximation works very well with the linear kernel. However, this does not seem to be the case with other kernel machines. In such cases, it is of interest to investigate alternative ways to approximate the distribution of the test statistic with relatively small sample sizes (Lee et al., 2012).

Absolute breast cancer risk estimates by SNP profile have been published for the BRCA1 and BRCA2 populations (Couch et al., 2013; Gaudet et al., 2013). However, only the SNP with the strongest evidence of association from each region is used. Future work may be directed towards the inclusion of sets of gene variants such as those considered in this paper for the calculation of individualized breast cancer risk estimates.

Transition

Dans ce chapitre, un test d'association génétique entre un ensemble de SNPs et une durée de vie a été développé en présence de corrélation intra-famille. Cependant, l'analyse est limitée au noyau linéaire pondéré. Le prochain chapitre étend la méthodologie au noyau *identical-by-state* pondéré et présente l'implantation de la nouvelle approche incluant les deux noyaux dans un progiciel R.

Chapter 4

gyriq: An R package for testing the association of sets of genetic variants with a survival trait in the presence of familial clustering

[gyriq : un progiciel R pour tester l'association d'ensembles de variants génétiques avec un trait de survie en présence de regroupements par famille]

Résumé

Des procédures de tests d'hypothèses entre des ensembles de variants génétiques et des issues d'âge d'apparition ont été développées récemment pour des études de puces d'ADN, de pharmacogénétique et d'analyse d'association sur le génome entier (GWAS) conduites selon des schémas d'échantillonnage standards. Davantage de puissance statistique est espérée de la part de devis basés sur la famille, notamment lorsque la co-ségrégation avec une maladie résulte en une occurrence plus fréquente de variants causaux parmi les membres d'une famille. Nous avons développé le progiciel R *gyriq* afin d'effectuer des tests d'association génétique pour des durées de vie en grappes avec ajustement pour la parenté. Ce progiciel prolonge le développement méthodologique présenté par Leclerc et al. (2015) avec l'ajout de la matrice noyau *identical-by-state* (IBS) pondérée. Nous passons en revue de façon brève les tests d'association génétique et les logiciels les accompagnant actuellement disponibles pour des issues d'âge d'apparition, décrivons comment le nouveau progiciel est implanté et illustrons son utilisation au moyen d'exemples.

Abstract

Hypothesis testing procedures between sets of genetic variants and age-at-onset outcomes have been developed recently for microarray, pharmacogenetics and genome-wide association studies (GWAS) conducted under standard sampling schemes. Higher statistical power is expected from family-based designs, especially when co-segregation with a disease result in more frequent occurrence of causal variants among family members. We developed the R package *gyriq* to perform kinship-adjusted genetic association testing for clustered lifetime data. This package extends the methodological development presented in Leclerc et al. (2015) with the addition of the weighted identical-by-state (IBS) kernel matrix. We briefly review the genetic association tests and accompanying software currently available for age-at-onset outcomes, describe how the new package is implemented, and illustrate its use through examples.

4.1 Introduction

Comparison of affected cases with unaffected controls via logistic regression is quite common in genetic association testing. However, in addition to disease status, genetic variants may be associated with the age at the onset of the disease, especially when the disease prevalence is high. Moreover, a subject with a censored age at onset may observe the event of interest afterwards and switch from the control to the case group. The age at the onset carries more information about the etiology of a disease than the disease status (Lin, 2014). Furthermore, simulation studies have shown that analyses performed using a survival trait instead of a simple binary endpoint can be more powerful (Hsu, 2003). The development of statistical methods especially focused on detecting genetic associations with survival outcomes has been ongoing with the proliferation of genotypic data in the last decade. Motivation for such development comes, for example, from microarray studies (Goeman et al., 2005), GWAS survival studies (Lin et al., 2011), or pharmacogenetics studies (Tzeng et al., 2014).

The Cox proportional hazards model (Cox, 1972) is widely used in genetic epidemiology to describe the simultaneous effects of age, genes, and environmental factors on disease risk in the presence of censored age-at-onset traits (Thomas, 2004). Typically, each genetic variant under investigation is a biallelic single nucleotide polymorphism (SNP) with three levels known as genotypes. Large-scale association studies where the genetic association of the survival trait with each SNP is tested in turn is referred to as the single-SNP approach in the literature. It has several drawbacks, notably the lack of statistical power which may result from the large number of tests being conducted and the small effect sizes for each of the SNPs. Alternative approaches have been developed, such as aggregating the SNPs of a given region into a unique score and variance-components tests. An advantage of the latter over the former is the ability to deal with the presence of both deleterious and protective variants in the tested region. Recently, several papers proposed variance-components tests especially focused on association

testing for survival outcomes (Goeman et al., 2005; Cai et al., 2011; Lin et al., 2011; Chen et al., 2014; Tzeng et al., 2014) with some of them implemented into user-friendly software.

All the variance-components tests mentioned above have been designed under a standard cohort setting where individuals are unrelated. However, many genetic studies especially those involving case-family designs, sample multiple relatives from the same family. Higher statistical power is expected from family-based designs with pedigrees enriched with multiple copies of disease-related variants. Results from Barnes et al. (2013) suggest that family-based approaches have greater power to detect associations than standard case-control analysis that ignores pedigree structure. Co-segregation with a disease can result in more frequent occurrence of causal variants among family members. From this perspective, Leclerc et al. (2015) introduced a kinship-adjusted SNP-set association test for survival outcomes which relies on a Gaussian copula to model the familial dependence. Up to now, this new method has not been implemented in a publicly available software. The aim of this paper is to fill this gap by introducing the R package *gyriq* (Leclerc and Chaieb, 2016) to the community. It is freely available from the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/web/packages/gyriq/index.html`. In Section 4.2, we present an overview of the genetic association tests and accompanying software currently developed for survival outcomes. Describing techniques especially designed for clustered survival data is the aim of Section 4.3, which includes the statistical approach behind the package *gyriq*. Section 4.4 presents extensions to arbitrary kernels. Section 4.5 explains how the test is implemented in terms of R code. In Section 4.6, we illustrate how kinship-adjusted SNP-set association testing is performed using the new package.

4.2 Genetic association tests for independent observations

Let T_i be the survival trait under investigation, and $G_i = (G_{i1}, \cdots, G_{is})$ and $X_i = (X_{i1}, \cdots, X_{ip})$ be row vectors of s genotypes and p non-genetic covariates, for individual i. In practice, the selection of the s genetic variants can be based on genes, biological pathways, linkage disequilibrium (LD) blocks and recombination hot-spots, or rely on a sliding window method (Lin et al., 2011). Typically, each $G_{ij} \in \{0, 1, 2\}$ is the genotype of a biallelic SNP corresponding to the number of copies of the minor allele. We consider the following proportional hazards model

$$\lambda(t_i|G_i, X_i) = \lambda_0(t_i)e^{G_iW\beta + X_i\zeta}, \tag{4.1}$$

where λ is the hazard function of T_i, λ_0 is the baseline hazard, $W = \text{diag}(w_1, \ldots, w_s)$ is a $s \times s$ diagonal matrix of weights to be used for the s genotypes, and $\beta = (\beta_1, \cdots, \beta_s)'$ and $\zeta = (\zeta_1, \ldots, \zeta_p)'$ are vectors of regression coefficients. One choice of weights is $\sqrt{w_j} = \text{Beta}(\text{MAF}_j; 1, 25)$ where MAF_j is the minor allele frequency of the j^{th} SNP in the set. Such

weighting gives more weight to the rare variants relative to variants with MAF 1%-5% and was suggested by Wu et al. (2011). It might be used to prevent the effects of rare variants to be smoothed over (Lin et al., 2011). The null hypothesis of the association test is

$$H_0 : \beta = 0. \tag{4.2}$$

In the presence of right-censoring, one only observes n quadruplets $\{Y_i, \delta_i, G_i, X_i, i = 1, \ldots, n\}$ where $Y_i = \min(T_i, C_i)$ is the observed survival trait, $\delta_i = I(T_i < C_i)$ is the censoring indicator and C_i is the censoring variable, assumed to be independent from T_i.

4.2.1 Single-SNP tests

Testing each genetic variant in turn, that is model (4.1) with $s = 1$, is the standard method for large-scale association studies. Computation of the Wald, likelihood ratio (LR), and score statistics is straightforward using the function `coxph` of the R package *survival* (Therneau, 2015b). Recent developments around the single-SNP tests address the issues of multiple testing and CPU time needed to run the analysis when the number of SNPs is large (up to millions). Resampling-based multiple hypothesis testing procedures for the Cox model are implemented in the R package *multtest* (Pollard et al., 2005) available from Bioconductor. Mendolia et al. (2014) proposed a Cox-Snell score test which is superior to the classical Wald and score tests but inferior to Lin's test (Lin, 2005) in terms of computational efficiency.

4.2.2 Aggregation tests of many SNPs

A natural choice is the standard burden test which is based on the following model

$$\lambda(t_i|G_i, X_i) = \lambda_0(t_i)e^{\kappa \sum_{j=1}^{s} w_j G_{ij} + X_i \zeta},$$

where $\sum_{j=1}^{s} w_j G_{ij}$ is the burden score and κ is a regression coefficient. Again, the test $H_0 : \kappa = 0$ can be performed using the Wald, LR, or score statistics. To prevent the effects of the SNPs from cancelling each other out, Adewale et al. (2008) proposed the following aggregation statistic to test the association between a set of s genes corresponding to a pathway and overall survival to breast cancer

$$\Psi = \sum_{i=1}^{s} \left(\frac{\beta_i}{se(\beta_i)} \right)^2,$$

where β_i is the parameter associated with the i^{th} gene under the single-SNP approach, and $se(\beta_i)$ its corresponding standard error. Other approaches were discussed by Crijns et al. (2009) and Lee et al. (2011).

4.2.3 Variance-components tests

Under this class of tests, β is treated as a random effect in model (4.1), with mean 0 and variance τI_s. Testing $H_0 : \tau = 0$ is then equivalent to (4.2). Let $R_i = \hat{\Lambda}_0(t_i)e^{X_i\hat{\zeta}}$, where $\hat{\Lambda}_0(t_i)$ is the Breslow's estimator of the baseline cumulative hazard function under the null model, G be the $n \times s$ matrix with rows G_i, and $M = (M_1, \cdots, M_n)'$ with $M_i = \delta_i - R_i$. The M_i's are known as the martingale residuals under H_0. The variance-components test statistic is based on the marginal likelihood with respect to β. Using a second-order Taylor expansion around $\beta = 0$ and derivating the marginal log-likelihood with respect to τ leads to

$$Q_0 = \frac{1}{2}M'KM - \mathrm{tr}(\tilde{R}K), \tag{4.3}$$

where $\tilde{R} = \mathrm{diag}(R_1, \ldots, R_n)$ and $K = GWWG'$. The score test proposed by Goeman et al. (2005) is equal to twice the first term in (4.3) and has been implemented in the R package *globaltest* (Goeman and Oosting, 2015) available from Bioconductor. Chen et al. (2014) proposed an alternative version based on signed square-root LR statistics which has better small-sample performance. Cai et al. (2011) generalized (4.3) to any kernel matrix. Following this paper, Lin et al. (2011) introduced the identical-by-state (IBS) kernel in the context of GWAS survival studies for modeling SNP-SNP interactions. The procedures are available in the R package *coxKM* (Lin and Zhou, 2015). Tzeng et al. (2014) proposed a similar method, only the resampling scheme being different as it relies on the generation of a large set of independent and identically distributed chi-squared random variables. Marceau et al. (2015) extended this approach by adding interaction terms comprising the kernel matrix and non genetic covariates. R functions of this extension are available from the authors' website. Other methods include the accelerated failure time model of Sinnott and Cai (2013), approaches based on boosting and permutation (Lee et al., 2011), supervised component analysis (Chen et al., 2008), and Bayesian statistics (Tachmazidou et al., 2010, 2008).

4.3 Presence of familial clustering

4.3.1 Single-SNP tests

Ignoring the clustering of failure times in the Cox model may lead to serious underestimation of the variance of the parameter estimates and consequently inflated type I error rates (Lin, 1994). The first solutions to this problem were the marginal (Wei et al., 1989) and shared frailty (Vaupel et al., 1979) approaches. Treating dependence between clustered observations as nuisance, the marginal approach allows robust estimation of the variance of the log-hazard ratio via the well-known sandwich estimator. In contrast, the shared frailty explicitly models the dependence between failure times. Both approaches are implemented in the R packages *survival* and *coxme* (Therneau, 2015a) respectively. However, in the context of genetic as-

sociation studies, these approaches would only make use of the identifier of the correlated observations, namely family membership. Often more is known about the relationships between individuals. The kinship matrix ϕ, which reflects the proportion of the genome that is identical-by-descent between pairs of individuals, may be available or can be estimated from the genotypic data (Astle and Balding, 2009, and references herein). An extension of the shared frailty approach which allows modelling of the dependence induced by ϕ is the correlated frailty model (Wienke, 2010):

$$\lambda(t_i|G_i, X_i) = \lambda_0(t_i)e^{G_iW\beta+X_i\zeta+a_i},$$

where $a = (a_1, \ldots, a_n)'$ follows a multivariate normal distribution with mean 0 and variance equal to $2\sigma_a^2\phi$, where σ_a^2 is a polygenic variance component. This model is implemented in R *coxme*. Parameter estimation is based on the Laplace approximation. Diao and Lin (2006) considered a correlated frailty model under the broader class of semiparametric linear transformation models with random effects. Adaptive Gaussian quadrature approximation (Pinheiro and Bates, 1995) is used for parameter estimation in the special case of the Cox model. Tests on β are performed using the LR statistic. Recently, Karyadi et al. (2015) used the correlated frailty model for single-SNP testing in survival GWAS.

4.3.2 Variance-components test

First note that model (4.1) can be alternatively written as a linear transformation model (Cheng et al., 1995)

$$H(t_i) = -G_iW\beta - X_i\zeta + \varepsilon_i,$$

where $H(\cdot)$ is an unknown monotone increasing function and ε_i follows the extreme value distribution. The kinship-adjusted kernel score test of Leclerc et al. (2015) models the familial dependence between individuals using a Gaussian copula. Under this setting, the joint cumulative distribution function (CDF) of the ε_i's is

$$\mathbb{P}(\varepsilon_1 \leq \epsilon_1, \ldots, \varepsilon_n \leq \epsilon_n) = \mathcal{C}_\Gamma\{F(\epsilon_1), \ldots, F(\epsilon_n)\},$$

where

$$\mathcal{C}_\Gamma\{u_1, \ldots, u_n\} = \Phi_n\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)\}$$

is the Gaussian copula, Γ is a correlation matrix, $F(\epsilon) = 1-e^{-\exp(\epsilon)}$ is the CDF of the extreme value distribution, and Φ and Φ_n are respectively the standard univariate and multivariate

normal CDFs. The diagonal elements of Γ are all equal to 1 and the off-diagonal entries are $\Gamma_{ij} = h^2\phi_{ij}$, where $h^2 \in [0,1]$ is a parameter measuring the polygenic heritability. Without loss of generality, we arrange the indices $\{1,\ldots,n\}$ so that $\delta_1 = \ldots = \delta_{n_1} = 1$ and $\delta_{n_1+1} = \ldots = \delta_n = 0$, where $n_1 = \sum_{i=1}^n \delta_i$ is the number of observed failure times. The testing procedure of Leclerc et al. (2015) can be summarized in five steps:

1. The parameters H, ζ and h^2 are estimated under the null model using a two-stage procedure (Othus and Li, 2010) and estimates $\hat{q}_i = \Phi^{-1}\left[F\left(\hat{H}(Y_i) + X_i\hat{\zeta}\right)\right]$ are computed.
2. Consider the partition $\hat{q} = \begin{pmatrix}\hat{q}^{(1)}\\ \hat{q}^{(0)}\end{pmatrix}$ so that $q^{(1)}$ and $q^{(0)}$ are of lengths n_1 and $n-n_1$, respectively. Let m be the number of required imputations. For $j = 1,\ldots,m$, imputed values $\tilde{q}^{(0)}$ are generated from the posterior distribution of the uncensored version of $\hat{q}^{(0)}$ given the observed values $\hat{q}^{(1)}$ with the restriction to be larger than the original censored values, componentwise. Then, we set $r_j = \begin{pmatrix}\hat{q}^{(1)}\\ \tilde{q}^{(0)}\end{pmatrix}$.
3. The completed vector of residuals $r^{(m)}$ is estimated as the mean of the r_j's and its approximate multivariate normal distribution is deduced. Full detail is given in Leclerc et al. (2015).
4. The test statistic is computed: $Q = r^{(m)'}Kr^{(m)}$. The Davies' (Davies, 1980) method is used to approximate the distribution of Q which is a linear combination of independent chi-squared random variables. The p value associated with Q is then obtained.

4.4 Extensions to arbitrary kernels

We now consider the model

$$H(t_i) = -\gamma(G_i) - X_i\zeta + \varepsilon_i,$$

where

$$\gamma(G_i) = \sum_{i'=1}^{n} \beta_{i'} K(G_i, G_{i'}),$$

$K(\cdot,\cdot)$ is a positive definite kernel, and $\beta = (\beta_1,\ldots,\beta_n)'$ is a random vector of regression coefficients. Examples of kernels include the weighted linear (LIN) and IBS kernels respectively given by

$$K_{\text{LIN}}(G_i, G_j) = \sum_{k=1}^{s} w_s G_{ik} G_{jk}$$

$$K_{\text{IBS}}(G_i, G_j) = \frac{\sum_{k=1}^{s} w_s \text{IBS}(G_{ik}, G_{jk})}{\sum_{k=1}^{s} 2w_s}$$

where $\text{IBS}(G_{ik}, G_{jk}) = 2 - |G_{ik} - G_{jk}| \in (0, 1 \text{ or } 2)$ is the number of alleles shared IBS by individuals i and j at SNP k in the SNP-set. The weighted IBS kernel is an interesting alternative to the weighted LIN kernel because it does not assume linearity or interactions of a particular order and is invariant to the type of genotype encoding (Wu et al., 2011). Let $\mathbb{K}$ be the $n \times n$ matrix whose $(i, j)^{\text{th}}$ element is $K(G_i, G_j)$. In Leclerc et al. (2015), we showed that $r^{(m)} \sim N(0, \rho^{(m)}\hat{\Gamma})$ where $\rho^{(m)}$ is a scale parameter and $\hat{\Gamma} = \hat{h}^2\phi + (1 - \hat{h}^2)I_n$. Let $z^{(m)} = \hat{\Gamma}^{-1/2} r^{(m)} / \sqrt{\rho^{(m)}}$. Elements of $z^{(m)}$ are independent. To test $H_0 : \beta = 0$, we consider the following statistic

$$\tilde{Q} = z^{(m)'} \mathbb{K} z^{(m)}$$

A standard permutation procedure can be used to obtain the corresponding p value. Consider $B = 10,000$ permutations of the elements of $z^{(m)}$, denoted as $\left\{ z_{b*}^{(m)}, b = 1, \ldots, 10,000 \right\}$, and let $\tilde{Q}_{b*} = z_{b*}^{(m)'} \mathbb{K} z_{b*}^{(m)}$. The p value is given by

$$\frac{1}{B} \sum_{b=1}^{B} I(\tilde{Q}_{b*} > \tilde{Q}) \tag{4.4}$$

In the context of a GWAS, multiple imputation (steps 1-3 in Section 4.3.2) and the B permutations of the completed vector of residuals have to be performed only once. To further reduce the computational burden in the presence of tens of thousands of SNP-sets and $n >> 1000$, we propose to first compute a p value for all the SNP-sets using (4.4) with $B = 1000$. Then, the matching moments approach (Lee et al., 2012) with $B = 10,000$ can be used to obtain a better approximation of the smallest p values. This method works as follows. Let

$$\hat{\mu} = \frac{1}{B} \sum_{b=1}^{B} \tilde{Q}_{b*}, \qquad \hat{\sigma}^2 = \frac{1}{B} \sum_{b=1}^{B} \left(\tilde{Q}_{b*} - \hat{\mu} \right)^2, \qquad \text{and} \qquad \hat{\mu}_4 = \frac{1}{B} \sum_{b=1}^{B} \left(\tilde{Q}_{b*} - \hat{\mu} \right)^4$$

The p value is

$$1 - F_d \left[\frac{\left(\tilde{Q} - \hat{\mu} \right) \sqrt{2d}}{\hat{\sigma}} + d \right]$$

where F_d is the CDF of the chi-square distribution with $d = 12/\hat{\kappa}$ degrees of freedom, and $\hat{\kappa} = \mu^4 / \hat{\sigma}^4 - 3$. Results from simulations studies presented in Chapter 5 indicate that the

matching moments approach under both LIN and IBS kernels controls very well type I error in the presence of familial dependence with $B = 10,000$ and is robust to the misspecification of the Gaussian copula.

4.5 Package *gyriq*: Kinship-adjusted survival SNP-set analysis

The function `genComplResid` implements the steps 1-3 of the proposed test. The function takes as arguments: the survival times (`U`), the censoring indicator (`Delta`), the kinship matrix (`Phi`), a vector called `blkID` with entries identifying correlated groups of observations, the number `m` of imputations (default = 50), and an optional matrix `X` corresponding to covariates. No missing data is allowed for all these arguments. Simulation studies reported in Leclerc et al. (2015) suggest that the default number of imputations guarantees a reasonable power in practice. The function will stop if one of the covariates has no variation. If no covariates are present, the matrix of covariates can be left as `NULL`.

Calculation of the completed vector of residuals requires proper estimation of the correlation matrix of the Gaussian copula. To this end, functions `dmvnorm` and `pmvnorm` are imported from the R package *mvtnorm* (Genz et al., 2015) to compute the corresponding log-likelihood function. Consequently, the number of censored individuals in each group identified by `blkID` cannot exceed 1000 which is the maximum dimension in the function `pmvnorm`. This should not be a problem because correlated groups identified by `blkID` most often correspond to families or blocks of the kinship matrix. However, larger groups such as regions of residence might also be used, for example to take into account population stratification or cryptic relatedness.

The function `genComplResid` produces a list consisting of:

- `compResid`: the completed vector of residuals
- `herit`: the estimate of h^2
- `covPar`: the estimate of ζ (if applicable)

Computation of the test statistic along with its corresponding p value is performed by the function `testGyriq` which has the following arguments:

- the completed vector of residuals
- a matrix (`G`) containing the set of SNPs
- a vector (`w`) of weights for the SNPs
- a string (`ker`) defining the type of kernel matrix

— an optional integer (`asv`) corresponding to the number of approximate eigenvalues to be estimated if the implicitly-restarted Lanczos bidiagonalization implemented in the R package *irlba* (Baglama and Reichel, 2015) is to be used for the spectral decomposition of the kernel matrix as a way to reduce the computional burden for big matrices (Davies' approximation only)

— a string (`method`) defining the procedure used to obtain the p value of the test

— an optional matrix (`starResid`) of permuted residuals to be used along with the matching moments or standard permutation approach

— two optional vectors (`bsw` and `tsw`) defining the lower and upper bounds of the sliding windows considered for the SNP-set

— an optional vector (`pos`) corresponding to the SNP positions

— a logical parameter (`sf`) indicating whether or not cluster computing is to be used via the R package *snowfall* (Knaus, 2015) in order to reduce wall-clock time when the number of sliding windows is still relatively modest (see the reference manual of the package *gyriq* for details)

— a string (`fileOut`) defining the name and path of the file where the results of each sliding window are printed

No missing data are allowed for `G` and `w`. The SNP genotypes should be equal to the number of copies of the minor allele (0, 1 or 2). The approximation of the p value following Davies' approximation is performed using the function `davies` imported from the R package *CompQuadForm* (Duchesne and Lafaye De Micheaux, 2010). If lower and upper bounds of sliding windows are not provided, the test is performed once on the whole SNP-set and the function produces a list consisting of:

— `score`: the score statistic of the test

— `pVal`: the p value

Otherwise, the test is applied to each sliding window and the output of the function is a data frame where each row represents a window tested. In addition to the observed value of the test statistic along with its corresponding p value, information is given regarding the definition of the sliding window, and the positions of the SNPs. If the calculation of the p value has failed, an error message is displayed. When the number of sliding windows is quite large, such as in a GWAS, the best option is to parallelize the test outside the scope of R.

4.6 Example

`simGyriq` is a dataset of phenotypic, genotypic and kinship data available in the package *gyriq*. It was simulated under conditions described in Leclerc et al. (2015). A sample of $n = 600$

individuals from 120 families was generated. The number of biallelic SNPs in matrix `G` was set to $s = 50$. Two covariates following the Bernoulli(0.5) and Uniform(-0.2, 0.2) distributions were generated. h^2 and ζ were set equal to 0.5 and $(1, 1)'$, respectively. The dataset also includes simulated positions (`pos`) for the 50 SNPs, and the lower (`bsw`) and upper (`tsw`) bounds of 4 sliding windows. Each window includes 10 SNPs, overlapping with the previous and subsequent windows. Note that through an appropriate choice of the lower and upper bounds it is also possible to define sliding windows with varying numbers of SNPs, for example, as a function of distance in kilobases between SNPs. Steps 1-3 of the testing procedure are performed by

```
> set.seed(1)
> library("gyriq")
> data("simGyriq")
> for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])
> cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
> str(cr)
```

```
List of 3
 $ compResid: num [1:600] -0.945 0.141 -0.148 0.132 -0.713 ...
 $ herit    : num 0.54
 $ covPar   : Named num [1:2] 1.399 0.869
  ..- attr(*, "names")= chr [1:2] "X1" "X2"
```

Estimates $\hat{h}^2$ and $\hat{\zeta}$ are relatively close to the true values. In Leclerc et al. (2015), we show by simulation that $\hat{h}^2$ and $\hat{\zeta}$ are virtually unbiased. The next chunk of code shows how to perform the analysis on each sliding window using the LIN kernel. The p value is computed using Davies' approximation.

```
> testGyriq(cr$compResid, G, w, ker="LIN", method="davies", bsw=bsw, tsw=tsw,
       pos=pos)
```

```
[1] "Processing sliding window 1"
[1] "Processing sliding window 2"
[1] "Processing sliding window 3"
[1] "Processing sliding window 4"
  FirstSNP LastSNP winSize   Start    Stop   Score  P_value Message
1        1      10      10 1254496 1273626 7710630 0.226984      OK
2        6      15      10 1266860 1282141 2648930 0.613296      OK
```

```
3       11      20      10 1274414 1291020  444691 0.883496      OK
4       16      25      10 1282346 1321322  951039 0.830385      OK
```

The vector of size $B \times n$ of permuted row indices ($B = 1,000$) needed for the standard permutation approach is included in the dataset `simGyriq`. Here is how to use the IBS kernel along with this approach.

```
> theResid <- cr$compResid[indResid]
> starResid <- matrix(theResid, nrow=1000, byrow=TRUE)
> testGyriq(cr$compResid, G, w, ker="IBS", method="rspOrd",
       starResid=starResid, bsw=bsw, tsw=tsw, pos=pos)
```

```
[1] "Processing sliding window 1"
[1] "Processing sliding window 2"
[1] "Processing sliding window 3"
[1] "Processing sliding window 4"
  FirstSNP LastSNP winSize   Start    Stop   Score  P_value Message
1        1      10      10 1254496 1273626 106.731   0.881      OK
2        6      15      10 1266860 1282141 124.812   0.639      OK
3       11      20      10 1274414 1291020 121.738   0.671      OK
4       16      25      10 1282346 1321322 116.944   0.709      OK
```

4.7 Discussion

In this paper, we introduced the R package *gyriq* for association testing between a set of SNPs and a survival trait in the presence of intrafamilial correlation. This package implements the statistical methodology developed in Leclerc et al. (2015). It has been extended with the implementation of the weighted IBS kernel and two methods for the computation of the p value: matching moments and standard permutation procedures. We also presented available software for single-SNP association testing in the presence of survival outcomes, with adjustment for familial relatedness.

Simulation studies (Lin et al., 2011; Chen et al., 2014) have shown under which scenarios kernel-based association testing outperforms single-SNP analysis and vice versa. Combining the test statistics or p values from the single-SNP, aggregation and variance-components approaches through omnibus tests in the presence of intrafamilial correlation require further investigation.

Risk prediction models are being developed for survival outcomes where only the SNP with the strongest evidence of association from a given region of the genome is included in the model

(Antoniou et al., 2010). Further investigation is needed to see how the inference procedures implemented in the R package *gyriq* could be combined with approaches for estimation and prediction under the kernel Cox regression model (Li and Luan, 2003) in order to improve predictive power in risk prediction models.

Chapitre 5

Résultats de simulations pour les noyaux linéaire et *identical-by-state* pondérés

À travers l'implantation de la méthodologie du chapitre 3 dans le progiciel R *gyriq* au chapitre 4, le test d'association génétique entre un ensemble de SNPs et un trait de survie a été généralisé par l'ajout du noyau *identical-by-state* pondéré. Le présent chapitre fournit les résultats de simulations qui illustrent que l'approximation de la valeur de p du test avec l'approche de permutation basée sur le couplage des moments est valide avec ce nouveau noyau et le noyau linéaire pondéré. Les scénarios de simulations utilisés sont similaires à ceux décrits au chapitre 3.

Les figures 5.1 et 5.2 montrent, pour chacun des noyaux, les graphiques quantile-quantile des valeurs p issues du test d'association appliqué dans le cadre de 10 000 répétitions sous H_0 avec dépendance familiale induite au moyen de la copule gaussienne. Les valeurs p sont très près de la ligne de référence de la distribution uniforme pour toutes les combinaisons du taux de censure et du paramètre d'héritabilité.

Les figures 5.3 et 5.4 présentent, pour chacun des noyaux, les graphiques quantile-quantile sous H_0 pour les quatre types de structure de dépendance quand $h^2 = .5$ et que le taux de censure est égal à 50 %. Les valeurs p du test sont près de la distribution attendue, même lorsque la structure de dépendance est mal spécifiée.

Pour davantage d'information concernant l'utilisation du progiciel R *gyriq*, le manuel de référence se trouve à l'annexe H.

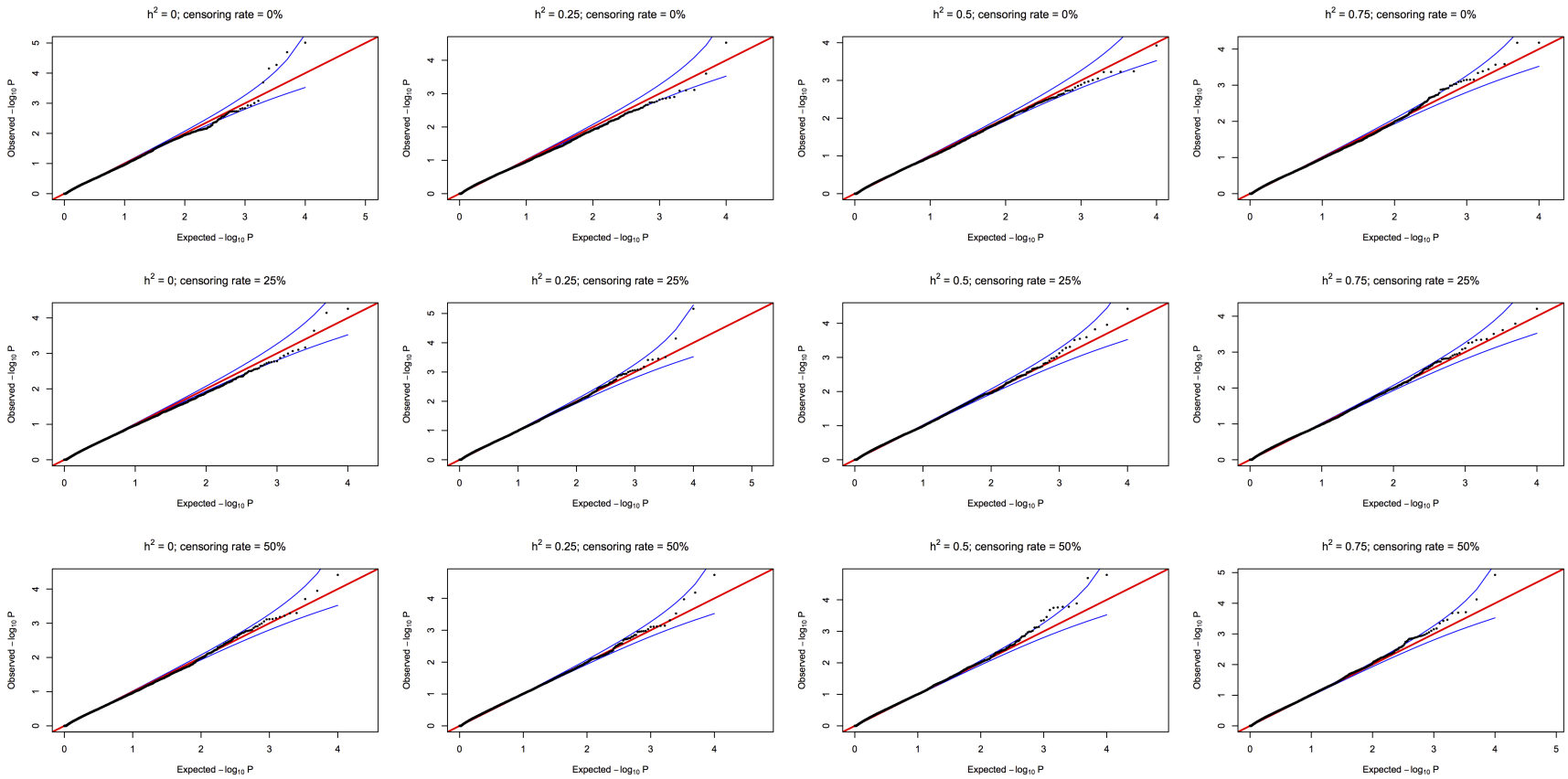

FIGURE 5.1 – Graphiques quantile-quantile des valeurs p de 10 000 répétitions produites sous H_0 en utilisant le test d'association du progiciel *gyriq* avec noyau **IBS** pondéré et approximation de la valeur de p via l'approche de permutation basée sur le couplage des moments (la ligne de référence de la distribution uniforme est en rouge et l'intervalle de confiance à 95 % est en bleu). La structure de dépendance entre les traits de survie est induite via un modèle de copule gaussienne. Les trois rangées correspondent à des taux de censure de 0, 25 et 50 % respectivement alors que les quatre colonnes font référence à un paramètre d'héritabilité h^2 égal à 0, 0.25, 0.5 et 0.75 respectivement.

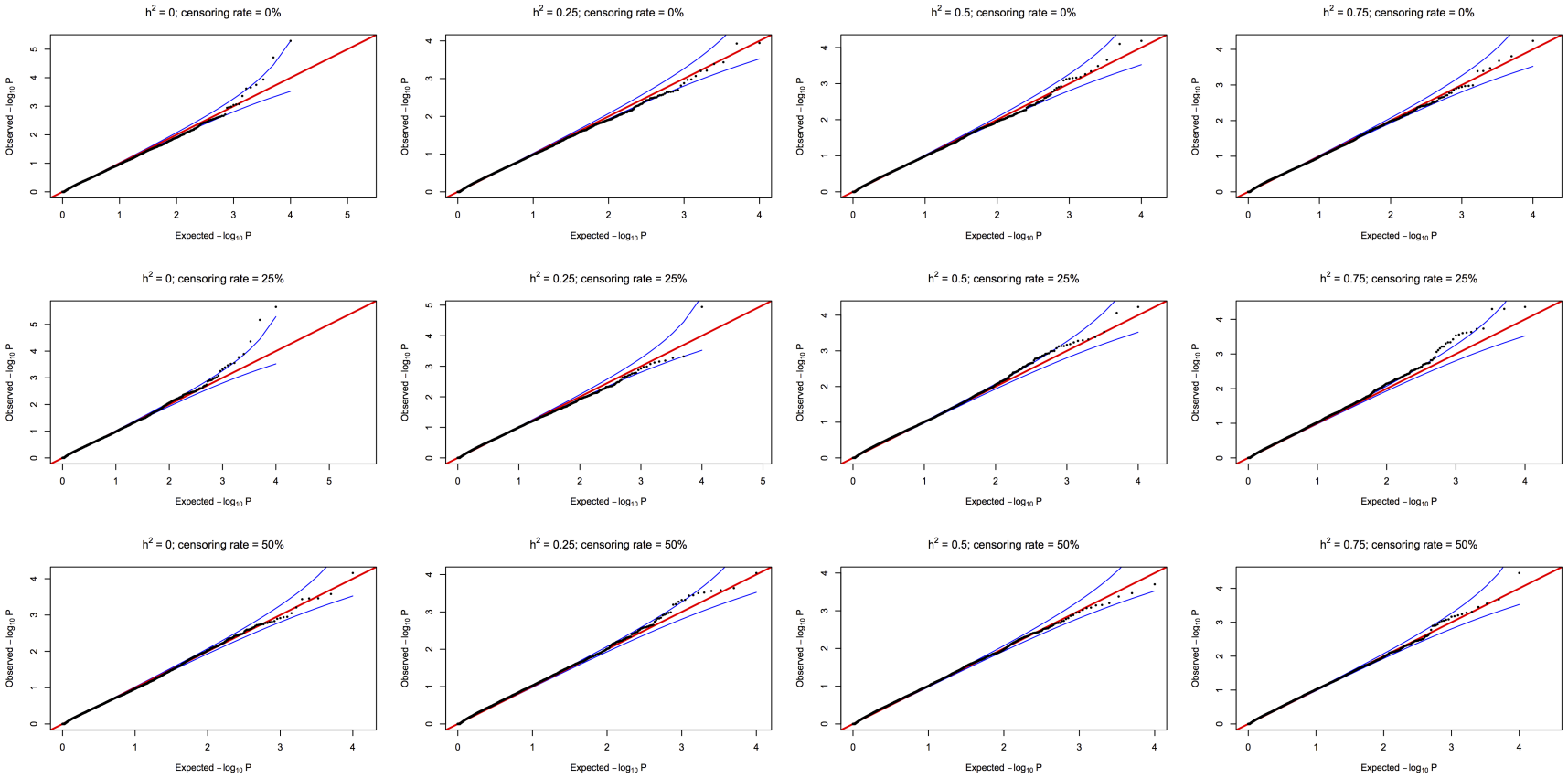

FIGURE 5.2 – Graphiques quantile-quantile des valeurs p de 10 000 répétitions produites sous H_0 en utilisant le test d'association du progiciel *gyriq* avec noyau **linéaire** pondéré et approximation de la valeur de p via l'approche de permutation basée sur le couplage des moments (la ligne de référence de la distribution uniforme est en rouge et l'intervalle de confiance à 95 % est en bleu). La structure de dépendance entre les traits de survie est induite via un modèle de copule gaussienne. Les trois rangées correspondent à des taux de censure de 0, 25 et 50 % respectivement alors que les quatre colonnes font référence à un paramètre d'héritabilité h^2 égal à 0, 0.25, 0.5 et 0.75 respectivement.

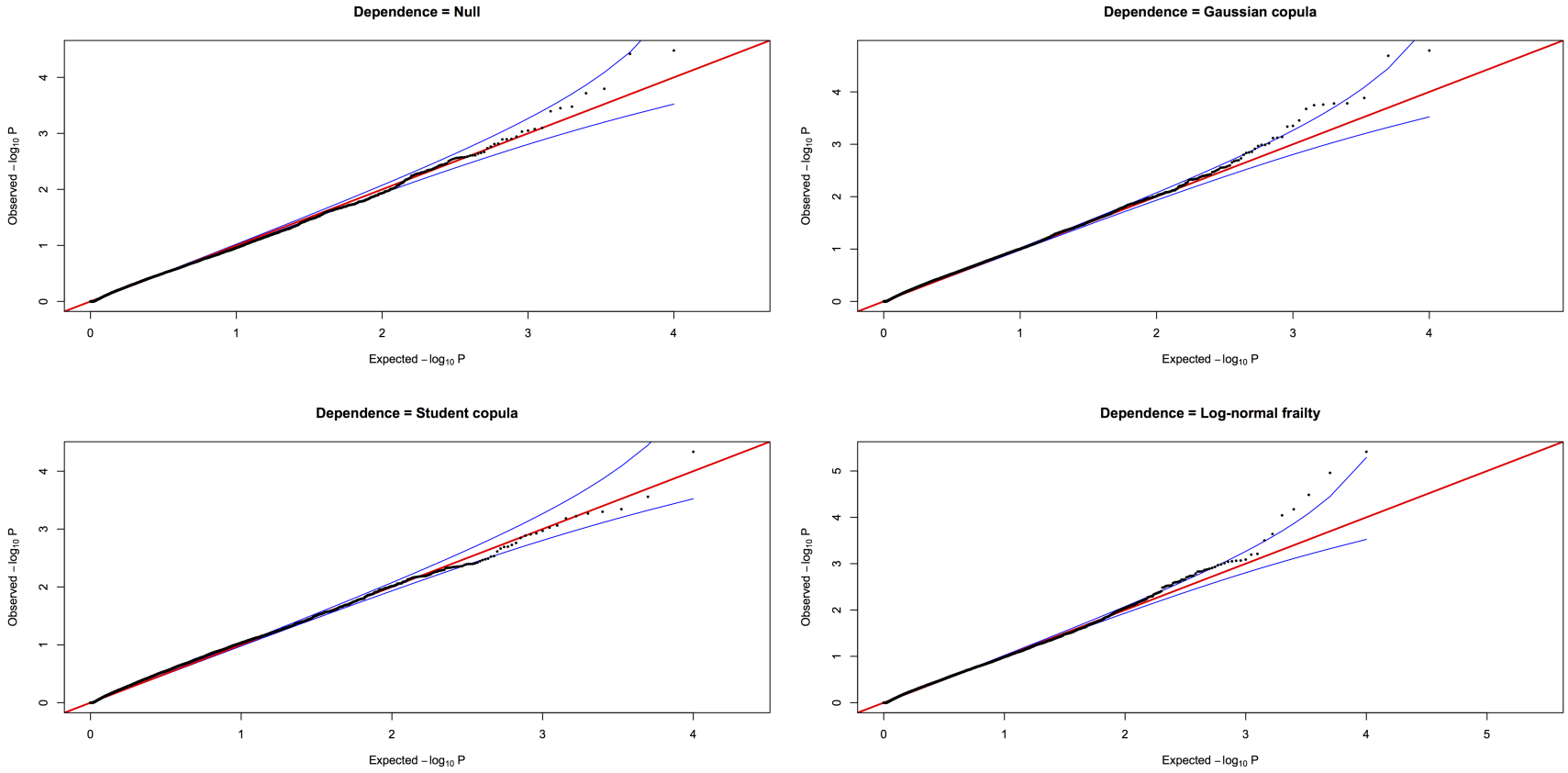

FIGURE 5.3 – Graphiques quantile-quantile des valeurs p de 10 000 répétitions produites sous H_0 en utilisant le test d'association du progiciel *gyriq* avec noyau **IBS** pondéré et approximation de la valeur de p via l'approche de permutation basée sur le couplage des moments pour quatre types de structure de dépendance entre les temps de survie (indépendance, copule gaussienne, copule de Student et termes de *frailty* distribués selon la loi log-normale).

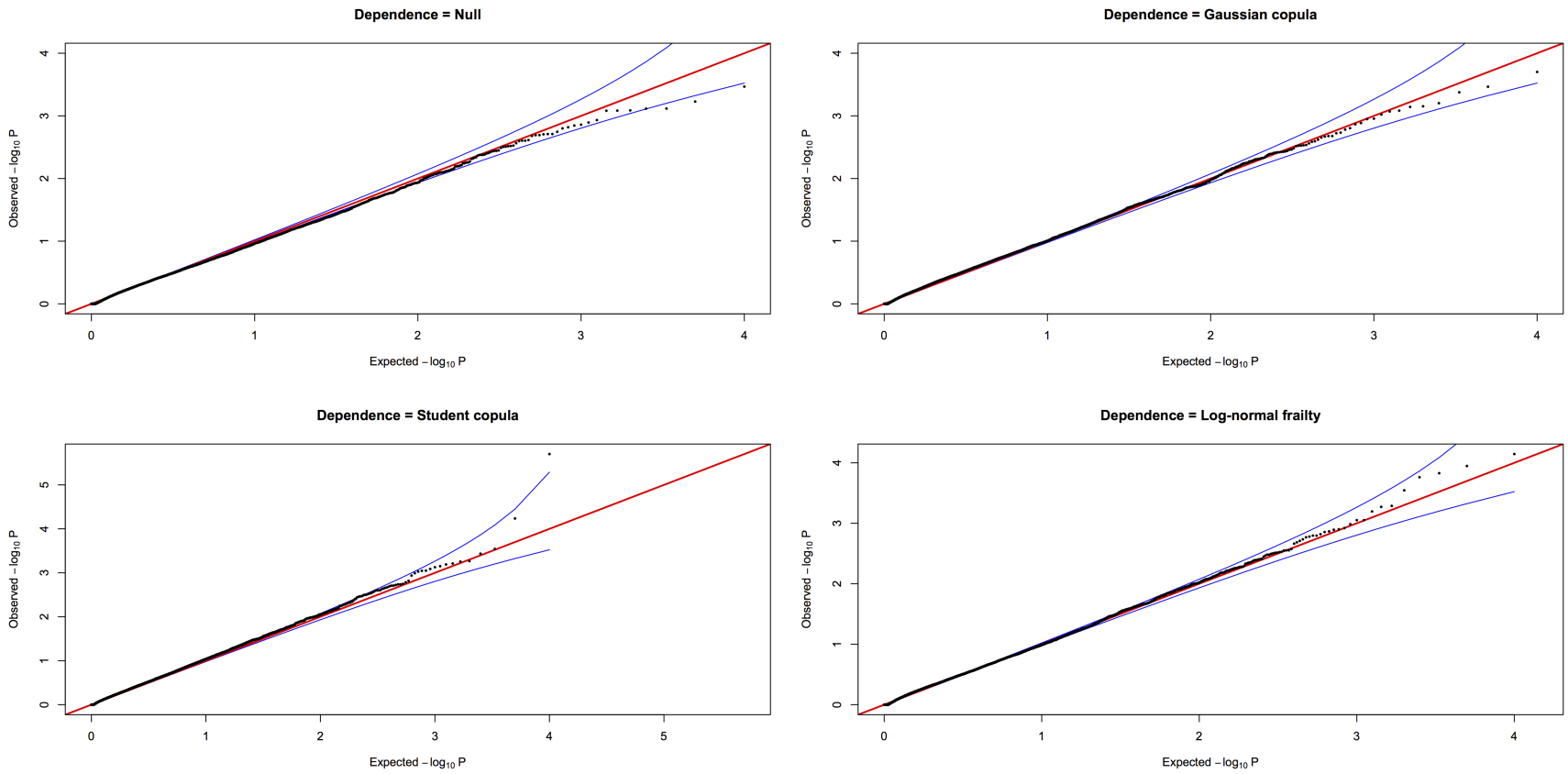

FIGURE 5.4 – Graphiques quantile-quantile des valeurs p de 10 000 répétitions produites sous H_0 en utilisant le test d'association du progiciel *gyriq* avec noyau **linéaire** pondéré et approximation de la valeur de p via l'approche de permutation basée sur le couplage des moments pour quatre types de structure de dépendance entre les temps de survie (indépendance, copule gaussienne, copule de Student et termes de *frailty* distribués selon la loi log-normale).

Conclusion

Dans cette thèse, il a été question de l'analyse des durées de vie dans le contexte spécifique des tests d'association génétique. Davantage de puissance statistique est espérée lorsqu'un trait de survie est utilisé comme variable réponse plutôt qu'une variable binaire. Cependant, les caractéristiques du plan d'échantillonnage cas-famille rend nécessaire le développement d'outils statistiques plus avancés et c'est ce qui constitue le coeur de cette thèse. Un premier test a d'abord été élaboré pour tester l'association SNP-cancer lorsque l'hypothèse des risques proportionnels du modèle de Cox n'est pas satisfaite. Ensuite, une approche a été développée pour détecter l'association entre un ensemble de SNPs et le risque de cancer du sein. Dans ce dernier cas, un modèle de copule utilisant les matrices de parenté et une approche innovante d'imputation multiple des observations censurées assure un contrôle approprié de l'erreur de type I du test.

Les articles 2 et 3 de cette thèse rendent possible l'identification d'ensembles de SNPs associés au risque de maladie. Il s'agit d'une étape intermédiaire du pipeline de la recherche qui mène vers la médecine personnalisée. Une question qui intéresse les généticiens est l'identification de variants génétiques qui ne sont pas seulement associés, mais qui causent la maladie (Edwards et al., 2013). En présence d'un ou plusieurs ensembles de SNPs présentant une association statistiquement significative, il serait d'intérêt de développer une méthode statistique qui permette d'identifier de façon plus efficace et efficiente les meilleurs SNPs candidats pour la relation de cause à effet.

Cette thèse s'est concentrée sur le développement d'un test d'hypothèses entre un ensemble de SNPs et des durées de vie en grappes et ce, sans fournir d'estimé pour la taille de l'effet potentiellement détecté. Des travaux ont été réalisés en lien avec la quantification d'un tel effet sous l'approche à noyau (Li and Luan, 2003; Evers and Messow, 2008; Van Belle et al., 2011). Selon la méthode originale de Li and Luan (2003), le négatif de la vraisemblance partielle sous le modèle de Cox est vu comme une fonction de perte avec paramètre de réglage (*tuning parameter*) estimé par validation croisée. Des travaux d'adaptation pourraient être effectués afin de rendre cette approche utilisable en présence d'un biais de sélection et de corrélation intra-famille.

La notion de risques concurrents fait habituellement référence aux données de survie pour

lesquelles deux ou plusieurs événements mutuellement exclusifs peuvent survenir (Andersen et al., 2012; Haller et al., 2013). Une telle situation se présente dans le contexte des données BRCA1/2 lorsque l'objectif est d'identifier des SNPs qui modifient le risque de développer, en tant que premier événement, le cancer du sein ou de l'ovaire. Dans cette thèse, l'âge d'apparition du cancer de l'ovaire, considéré comme une censure, est supposé indépendant de l'âge d'apparition du cancer du sein. Il serait pertinent d'examiner la sensibilité des tests d'hypothèses développés pour différents scénarios de dépendance entre ces deux événements concurrents. Une approche possible est d'utiliser une copule pour modéliser l'association entre les durées de vie des deux types d'événements (cancer du sein et cancer de l'ovaire) et de procéder aux tests statistiques pour différents niveaux d'association fixés d'avance (Huang and Zhang, 2008). Dans le cadre d'un plan d'échantillonnage cas-famille et pour un test portant sur un ensemble de SNPs, cette approche nécessiterait des développements particuliers.

D'autre part, pour l'analyse de données réelles dans cette thèse, les individus apparentés sont tous porteurs d'une mutation pathogène sur le gène BRCA1 ou BRCA2. Sachant que le risque de cancer du sein est associé aux antécédents familiaux, il serait approprié de développer de nouvelles méthodes statistiques qui prennent en compte toute l'information génotypique et phénotypique disponible auprès des membres de la famille, qu'ils soient porteurs de mutation ou non. Davantage de puissance statistique peut en résulter et une nouvelle estimation du risque pour les SNPs déjà associés à la maladie permettrait de mesurer l'ampleur du biais potentiel.

Cette thèse s'est concentrée sur l'identification de SNPs ou d'ensembles de SNPs significativement associés à la durée de vie étudiée. Dans le cadre de modèles de prédiction du risque, l'avènement de données massives sur les variants génétiques rend nécessaire le développement de nouvelles approches statistiques pour l'optimisation du pouvoir prédictif des modèles, approches qui devraient aussi inclure les SNPs ou ensembles de SNPs qui ne franchissent pas nécessairement le seuil pour être jugés significatifs. Ces variants peuvent contenir une quantité substantielle d'information pertinente pour la prédiction. La revue de littérature de Sinnott and Cai (2014) offre un point de départ intéressant en ce qui concerne les méthodes existantes en analyse des durées pour la prédiction du risque avec données de haute dimension.

Effets des variants rares, estimés d'héritabilité incorrects, variants non détectés avec petites tailles d'effet, etc. : la quête pour la variabilité génétique inexpliquée est ouverte sur plusieurs fronts.

Appendix A

Weighted cohort method

Antoniou et al. (2005) proposed to estimate the weights by two functions of x, $\widehat{\Phi}(x,1)$ and $\widehat{\Phi}(x,0)$, both piecewise constant over the L intervals $[A_0, A_1), [A_1, A_2), \ldots, [A_{L-1}, A_L)$, where $0 = A_0 < A_1 < \ldots < A_L$ is an arbitrary partition such that $A_L > \max_{ij}\{X_{ij}\}$. Let $a_l = \widehat{\Phi}(x,1)$ and $b_l = \widehat{\Phi}(x,0)$ for $x \in [A_{l-1}, A_l)$, $l = 1, \cdots, L$. The weighted cohort method estimates $\{(a_l, b_l), l = 1, \ldots, L\}$ as follows. For each time interval $[A_{l-1}, A_l)$ define

- $t_l = A_l - A_{l-1}$: the length of the interval;
- $r_l = \sum_{i,j} \delta_{ij} I(A_{l-1} \leq X_{ij} < A_l)$: the number of failure events;
- $p_l = \sum_{i,j} (X_{ij} - A_{l-1}) \delta_{ij} I(A_{l-1} \leq X_{ij} < A_l)$: the at-risk time accumulated by the r_l failure events;
- $s_l = \sum_{i,j} (1 - \delta_{ij}) I(A_{l-1} \leq X_{ij} < A_l)$: the number of censoring events;
- $q_l = \sum_{i,j} (X_{ij} - A_{l-1})(1 - \delta_{ij}) I(A_{l-1} \leq X_{ij} < A_l)$: the at-risk time accumulated by the s_l censoring events.

Let μ_l be the true incidence rate for the time interval l (assumed to be known). The sampling weights for the failure and censoring events, a_l and b_l respectively, satisfy

$$\mu_l = \frac{a_l r_l}{a_l p_l + b_l q_l + t_l \sum_{m>l} (b_m s_m + a_m r_m)}$$

The numerator in the term on the right-hand side is an estimate of the number of failure events within the time interval while the denominator represents the at-risk time accumulated within the interval. Contraints are added to get equal sample sizes for weighted and unweighted data

$$a_l r_l + b_l s_l = r_l + s_l$$

Solving for a_l and b_l we obtain

$$a_l = \frac{\mu_l \left\{ q_l(r_l + s_l) + t_l s_l \sum_{m>l}(r_m + s_m) \right\}}{r_l s_l + \mu_l(q_l r_l - p_l s_l)}$$

$$b_l = \frac{1}{s_l}(r_l + s_l - a_l r_l)$$

Appendix B

Copula models

Let $(T_1, \ldots, T_D)^T$ be a D-dimensional random variable with a dependence structure given by (2.4) and marginal survival functions equal to $\{S_{g_1}, \ldots, S_{g_D}\}$. Under such a model, any subset of $(T_1, \ldots, T_D)$ of size $d \leq D$ follows $C_{\theta,d}$. Furthermore, for $j = 2, \ldots, d$, one has

$$\Pr(T_j > t_j \mid T_1 = t_1, \ldots, T_{j-1} = t_{j-1}, G_1 = g_1, \ldots, G_j = g_j) = \mathcal{H}_{\theta,j}\{S_{g_1}(t_1), \ldots, S_{g_j}(t_j)\},$$

where

$$\mathcal{H}_{\theta,j}(u_1, \ldots, u_j) = \frac{\frac{\partial^{j-1}}{\partial u_1 \ldots, \partial u_{j-1}} \mathcal{C}_{\theta,j}(u_1, \ldots, u_j)}{\frac{\partial^{j-1}}{\partial u_1 \ldots, \partial u_{j-1}} \mathcal{C}_{\theta,j-1}(u_1, \ldots, u_{j-1})}.$$

The algorithm of Lee (1993) generates such a subset $\{T_1, \ldots, T_d\}$ and works as follows

- Generate the first variate ($j = 1$)
  - Generate $U_1 \sim$ Uniform$[0, 1]$
  - Set $T_1 = S_{g_1}^{-1}(U_1)$
- For $j = 2, \ldots, d$
  - Generate $V_j \sim$ Uniform$[0, 1]$
  - Solve $\mathcal{H}_{\theta,j}(U_1, \ldots, U_j) = V_j$ to obtain U_j
  - Set $T_j = S_{g_j}^{-1}(U_j)$

It is worth mentioning that $\mathcal{H}_{\theta,j}(U_1, \ldots, U_j) = V_j$ has an explicit solution given by $U_j = \{b(a-1) + a\}^{-1/\theta}$ where $a = V_j^{-\theta/\{\theta(j-1)+1\}}$ and $b = \sum_{l=1}^{j-1} U_l^{-\theta} - (j-1)$.

Appendix C

Joint distribution of $\{G_1, \ldots, G_d\}$

The joint distribution of the genotypes $\{G_1, \ldots, G_d\}$ of d siblings is clearly discrete. Table C.1, deduced from Elandt-Johnson (1971), gives the seven possible genotypic types of $\{G_1, \ldots, G_d\}$ along with their respective probabilities when $d \geq 3$. For instance, if $d = 3$, the sixth line of Table C.1 states that

$$\begin{aligned} & P(G_1 = aa, G_2 = AA, G_3 = AA) = P(G_1 = AA, G_2 = aa, G_3 = AA) \\ = \quad & P(G_1 = AA, G_2 = AA, G_3 = aa) = p^2 \times (1-p)^2 \times (1/4)^{d-1}. \end{aligned}$$

where $p = \mathbb{P}(a)$ is the minor allele frequency. When $d = 2$, the situation is even simpler as there are only 6 possible genotypic types. Finally, when $d = 1$, the probabilities $P(G = aa)$, $P(G = Aa)$ and $P(G = AA)$ are equal to p^2, $2p(1-p)$ and $(1-p)^2$, respectively.

Table C.1 – Joint probabilities for genotypic types of a genotype vector of d siblings, $d \geq 3$

Genotypic type	Probability $\times 4^{d-1}$	Combinations
$d \times AA$	$p^2(1+\alpha p)^2$	1
$d \times aa$	$q^2(1+\alpha q)^2$	1
$d \times Aa$	$2pq(1+\alpha p+\alpha q+\alpha(\alpha+1)pq)$	1
$r \times AA, (d-r) \times Aa$	$2p^2q(1+\alpha p+(2^{d-1-r}-1)q)$	$\binom{d}{r}$
$r \times aa, (d-r) \times Aa$	$2q^2p(1+\alpha q+(2^{d-1-r}-1)p)$	$\binom{d}{r}$
$r \times AA, (d-r) \times aa$	p^2q^2	$\binom{d}{r}$
$r_1 \times AA, r_2 \times Aa,$ $(d-r_1-r_2) \times aa$	$2^{r_2}p^2q^2$	$\binom{d}{r_1, r_2, d-r_1-r_2}$

Note: $p = \mathbb{P}(a)$, $q = 1-p$ and $\alpha = 2^{d-1}-1$

Appendix D

Estimation procedure for H and ζ

The algorithm of Chen et al. (2002) estimates the parameters H and ζ under the null hypothesis $H_0 : \beta = 0$ based on the data $\{(Y_i, \delta_i, X_i), i = 1, \cdots, n\}$. This procedure works as follows:

— Compute $\hat{\zeta}$ as the standard maximum partial likelihood estimator of the Cox model $\lambda(t_i|X_i) = \lambda_0(t_i)e^{X_i\zeta}$.

— Estimate H by a step function with jumps at the observed failure times, i.e., Y_i with $\delta_i = 1$. Let $t_1 < \ldots < t_{n_1}$ be the n_1 ordered observed failure times. The step function $\hat{H}$ is obtained recursively as follows:

  — Compute $\hat{H}(t_1)$ by solving $\sum_{i=1}^{n} I(Y_i \geq t_1)e^{X_i\hat{\zeta}+H(t_1)} = 1$.

  — For $k = 2, \ldots, n_1$, one has $\hat{H}(t_k) = \hat{H}(t_{k-1}) + \frac{1}{\sum_{i=1}^{n} I(Y_i \geq t_k)e^{X_i\hat{\zeta}+\hat{H}(t_{k-1})}}$

Appendix E

Likelihood function of the right-censored sample $\{(\hat{q}_i, \delta_i), i = 1, \ldots, n\}$

Without loss of generality, we rearrange the indices $\{1, \ldots, n\}$ so that $\delta_1 = \cdots = \delta_{n_1} = 1$ and $\delta_{n_1+1} = \ldots = \delta_n = 0$, where $n_1 = \sum_{i=1}^{n} \delta_i$ is the number of uncensored failure times. Consider the partition $\hat{q} = \begin{pmatrix} \hat{q}^{(1)} \\ \hat{q}^{(0)} \end{pmatrix}$ so that $\hat{q}^{(1)}$ and $\hat{q}^{(0)}$ are of lengths n_1 and $n - n_1$, respectively. Similarly, write

$$\Gamma = h^2\phi + (1 - h^2)I_n = \begin{pmatrix} \Gamma^{(11)} & \Gamma^{(10)} \\ \Gamma^{(01)} & \Gamma^{(00)} \end{pmatrix}$$

The likelihood function is then $L\left(\hat{q}^{(1)}, \hat{q}^{(0)}\right) = L\left(\hat{q}^{(1)}\right) \times L\left(\hat{q}^{(0)}|\hat{q}^{(1)}\right)$, where $L\left(\hat{q}^{(1)}\right)$ is the density function of a n_1-variate normal distribution with mean 0 and covariance matrix $\Gamma^{(11)}$ evaluated at $\hat{q}^{(1)}$ and $L\left(\hat{q}^{(0)}|\hat{q}^{(1)}\right)$ is the joint survival function of a $(n-n_1)$-variate normal distribution with mean $\Gamma^{(01)}{\Gamma^{(11)}}^{-1}\hat{q}^{(1)}$ and covariance matrix $\Gamma^{(00)} - \Gamma^{(01)}{\Gamma^{(11)}}^{-1}\Gamma^{(10)}$ evaluated at $\hat{q}^{(0)}$. The estimate $\hat{h}^2$ is then obtained by maximizing $L\left(\hat{q}^{(1)}, \hat{q}^{(0)}\right)$.

Appendix F

Proof of the joint distribution of the completed vector of residuals r

Let A be a random variable that denotes an arbitrary element of a completed vector of residuals r. Actually, A corresponds to either an observed or an imputed value, therefore it is distributed according to a mixture of two distributions. We begin by showing that this mixture corresponds to a standard normal distribution. Let Q be a random variable following a standard normal distribution and C a censoring variable, independent from Q with density f. Generate A as follows.

- Generate Q and C and denote by q and c the obtained values.
  - If $q \leq c$, then set $A = q$.
  - If $q > c$, then generate Z according to a standard normal distribution truncated to $[c, \infty]$ and set $A = Z$.

For an arbitrary real number a, one has

$$\begin{aligned} P(A \leq a) &= P(A \leq a, Q < C) + P(A \leq a, Q > C) \\ &= P\{Q < \min(a, C)\} + P(A \leq a, Q > C) \\ &= D + E \end{aligned}$$

with

$$\begin{aligned} D &= P\{Q < \min(a, C)\} \\ &= \int_{c=-\infty}^{\infty} P\{Q < \min(a, c)\} f(c) dc \\ &= \int_{c=-\infty}^{a} \Phi(c) f(c) dc + \Phi(a) P(C > a) \end{aligned}$$

where Φ is the CDF of the standard normal distribution and

$$\begin{aligned}
E &= P(A \leq a, Q > C) \\
&= \int_{c=-\infty}^{\infty} P(Z \leq a, Q > c | C = c) f(c) dc \\
&= \int_{c=-\infty}^{\infty} P(Z \leq a | C = c) \{1 - \Phi(c)\} f(c) dc \\
&= \int_{c=-\infty}^{a} \frac{\Phi(a) - \Phi(c)}{1 - \Phi(c)} \{1 - \Phi(c)\} f(c) dc \\
&= \Phi(a) P(C < a) - \int_{c=-\infty}^{a} \Phi(c) f(c) dc
\end{aligned}$$

and therefore $P(A \leq a) = D + E = \Phi(a)$, which implies that each of the elements of the completed vector r marginally follows a standard normal distribution.

Consider now two independent pairs of random variables: (Q_1, Q_2) following a bivariate normal distribution with mean μ and covariance matrix Σ and (C_1, C_2) with a joint density function $f_C(c_1, c_2)$. Define (R_1, R_2) as follows:

— Generate two independent pairs (Q_1, Q_2) and (C_1, C_2). Denote these (q_1, q_2) and (c_1, c_2).

— If $q_1 \leq c_1$ and $q_2 \leq c_2$, then set $R_1 = q_1$ and $R_2 = q_2$.

— If $q_1 \leq c_1$ and $q_2 > c_2$ then set $R_1 = q_1$ and generate R_2 from the conditional normal distribution of Q_2 given $Q_1 = q_1$ restricted to $[c_2, \infty]$.

— If $q_1 > c_1$ and $q_2 \leq c_2$ then generate R_1 from the conditional normal distribution of Q_1 given $Q_2 = q_2$ restricted to $[c_1, \infty]$ and set $R_2 = q_2$.

— If $q_1 > c_1$ and $q_2 > c_2$, then generate (R_1, R_2) from a bivariate normal distribution with mean μ and covariance matrix Σ restricted to $[c_1, \infty] \times [c_2, \infty]$.

Note that the definition of (R_1, R_2) corresponds to the imputation process with $m = 1$ and $n = 2$. Let us show that (R_1, R_2) has the same joint distribution as (Q_1, Q_2).

Let (r_1, r_2) be a pair of real numbers. From the law of total probabilities, one has

$$\begin{aligned}
P(R_1 \leq r_1, R_2 \leq r_2) &= P(R_1 \leq r_1, R_2 \leq r_2, Q_1 < C_1, Q_2 < C_2) \\
&\quad + P(R_1 \leq r_1, R_2 \leq r_2, Q_1 < C_1, Q_2 > C_2) \\
&\quad + P(R_1 \leq r_1, R_2 \leq r_2, Q_1 > C_1, Q_2 < C_2) \\
&\quad + P(R_1 \leq r_1, R_2 \leq r_2, Q_1 > C_1, Q_2 > C_2) \\
&= A_1 + A_2 + A_3 + A_4.
\end{aligned}$$

By conditioning on $C_1 = c_1; C_2 = c_2$, we obtain

$$\begin{aligned} A_1 &= P\{Q_1 \le \min(r_1, C_1), Q_2 \le \min(r_2, C_2)\} \\ &= \int_{c_1=-\infty}^{r_1}\int_{c_2=-\infty}^{r_2} P(Q_1 < c_1, Q_2 < c_2) f(c_1, c_2) dc_1 dc_2 \\ &\quad + \int_{c_1=-\infty}^{r_1}\int_{c_2=r_2}^{\infty} P(Q_1 < c_1, Q_2 < r_2) f(c_1, c_2) dc_1 dc_2 \\ &\quad + \int_{c_1=r_1}^{\infty}\int_{c_2=-\infty}^{r_2} P(Q_1 < r_1, Q_2 < c_2) f(c_1, c_2) dc_1 dc_2 \\ &\quad + \int_{c_1=r_1}^{\infty}\int_{c_2=r_2}^{\infty} P(Q_1 < r_1, Q_2 < r_2) f(c_1, c_2) dc_1 dc_2. \end{aligned}$$

On the other hand, one has

$$\begin{aligned} A_2 &= P\{Q_1 \le \min(r_1, C_1), R_2 \le r_2, Q_2 > C_2\} \\ &= \int_{c_1=-\infty}^{r_1}\int_{c_2=-\infty}^{r_2} P(R_2 \le r_2 | Q_2 > c_2, Q_1 \le c_1) P(Q_1 \le c_1, Q_2 > c_2) f(c_1, c_2) dc_1 dc_2 \\ &\quad + \int_{c_1=r_1}^{\infty}\int_{c_2=-\infty}^{r_2} P(R_2 \le r_2 | Q_2 > c_2, Q_1 \le r_1) P(Q_1 \le r_1, Q_2 > c_2) f(c_1, c_2) dc_1 dc_2 \\ &= \int_{c_1=-\infty}^{r_1}\int_{c_2=-\infty}^{r_2} P(Q_1 \le c_1, c_2 \le Q_2 \le r_2) f(c_1, c_2) dc_1 dc_2 \\ &\quad + \int_{c_1=r_1}^{\infty}\int_{c_2=-\infty}^{r_2} P(Q_1 \le r_1, c_2 \le Q_2 \le r_2) f(c_1, c_2) dc_1 dc_2. \end{aligned}$$

Similarly, one has

$$\begin{aligned} A_3 &= \int_{c_1=-\infty}^{r_1}\int_{c_2=-\infty}^{r_2} P(c_1 \le Q_1 \le r_1, Q_2 \le c_2) f(c_1, c_2) dc_1 dc_2 \\ &\quad + \int_{c_1=-\infty}^{r_1}\int_{c_2=r_2}^{\infty} P(c_1 \le Q_1 \le r_1, Q_2 \le r_2) f(c_1, c_2) dc_1 dc_2. \end{aligned}$$

Finally,

$$A_4 = \int_{c_1=-\infty}^{r_1}\int_{c_2=-\infty}^{r_2} P(c_1 \le Q_1 \le r_1, c_2 \le Q_2 \le r_2) f(c_1, c_2) dc_1 dc_2.$$

By summing up these pieces, one obtains $P(R_1 \le r_1, R_2 \le r_2) = P(Q_1 \le r_1, Q_2 \le r_2)$, for all (r_1, r_2).

The extension of this result to an arbitrary n is straightforward. Therefore, a vector of residuals $r^{(1)}$ obtained by a single imputation ($m = 1$) follows a multivariate normal distribution with mean zero and covariance matrix Γ. Moreover, the distribution of $\Gamma^{-1/2} r^{(1)}$ is a multivariate normal with mean zero and covariance matrix I_n, the identity of size n.

A vector of residuals $r^{(m)}$ obtained from an imputation with an arbitrary m is the mean of m vectors of residuals obtained from an single imputation, i.e.

$$r^{(m)} = \frac{1}{m} \sum_{j=1}^{m} r_j.$$

Therefore, one has

$$\Gamma^{-1/2} r^{(m)} = \frac{1}{m} \sum_{j=1}^{m} \Gamma^{-1/2} r_j.$$

For $j \in \{1, \ldots, m\}, \Gamma^{-1/2} r_j$ is distributed according to a multivariate normal distribution with mean zero and covariance matrix I_n, and hence, has independent components. Therefore, the components of $\Gamma^{-1/2} r^{(m)}$ are also independent. Moreover, these components are identically distributed and thus the covariance matrix of $\Gamma^{-1/2} r^{(m)}$ has the form ρI_n, where ρ is the marginal variance of any element of $\Gamma^{-1/2} r^{(m)}$. Therefore, $\Gamma^{-1/2} r^{(m)}$ follows approximately a multivariate normal distribution with mean zero and covariance matrix ρI_n, which can be expressed as $r^{(m)} \sim \mathcal{N}(0, \rho\Gamma)$.

Appendix G

Additional simulation results

As suggested by anonymous referees, we performed several sets of additional simulations in order to further investigate the empirical properties of the proposed association test in various situations. The conditions and results of these simulations are presented below.

G.1 Impact of the number m of imputations on power

In the first set of simulations, we investigated the impact of the number m of imputations on power. We considered scenarios corresponding to censoring rates of 25 and 50%, h^2 equal to 0, 0.25, 0.5 and 0.75, respectively. For each combination of these parameters, 10,000 replications were generated under the alternative hypothesis using a Gaussian copula model and we computed the empirical power using different values of m. The coefficients $\beta = (\beta_1, \cdots, \beta_s)'$ were generated following a multivariate normal distribution with mean zero and variance-covariance matrix equal to τI_s, where τ satisfies

$$\begin{aligned} h^2_{QTL} &= \frac{\text{Var}(G_i W \beta)}{\text{Var}(G_i W \beta) + \text{Var}(\epsilon_i)} \\ &= \frac{\tau E[G_i W W G_i']}{\tau E[G_i W W G_i'] + \pi^2/6} = 1\%. \end{aligned}$$

The results are reported in Figure G.1, which shows that the power increases quickly with m and then levels off in all considered scenarios. These results also suggest that the use of $m = 50$ guarantees a reasonable power in practice.

G.2 Type I error rates for various LDs and tested SNP-set sizes s

In this simulations set, we investigated the empirical properties of the proposed method under various LDs and SNP-set sizes s. We considered scenarios corresponding to $r^2 = 0, 0.5$ and 1

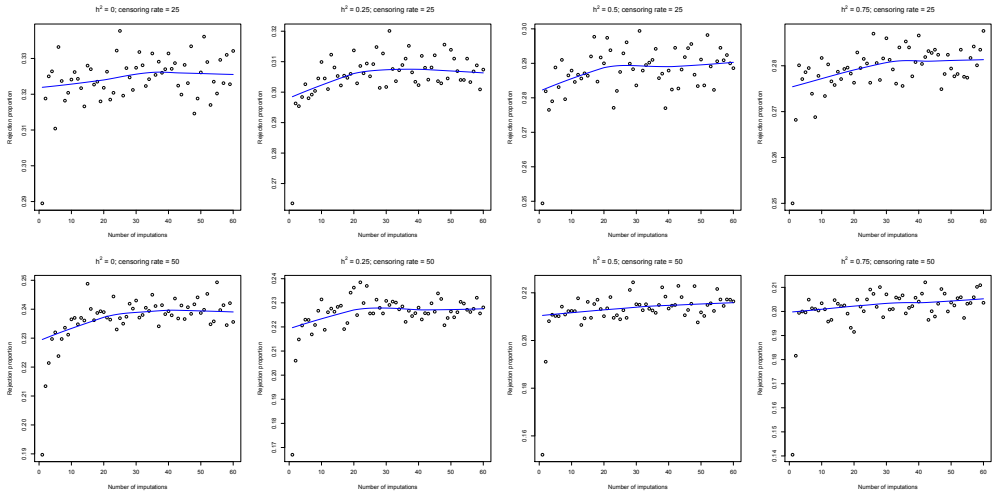

Figure G.1 – Power of the kinship-adjusted association test as a function of the number m of imputations used to estimate the residuals for the censored survival times. Each rejection rate was computed from 10,000 replications at the 5% significance level. The dependence structure between survival times is induced via the Gaussian copula model. The two rows correspond to censoring rates of 25 and 50%, respectively whereas the four columns refer to h^2 equal to 0, 0.25, 0.5 and 0.75 respectively. The blue curve is obtained by the LOWESS smoother of the R software.

and $s = 10, 25$ and 50. For each combination of these parameters, we generated 10,000 replications under the null hypothesis. The dependence structure between survival times was induced via the Gaussian copula model with the heritability parameter h^2 fixed at 0.5 and the censoring rate at 50%. For each simulated dataset, we computed the p-value of the proposed association test with $m = 50$ and the QQ-plots of these p-values are reported in Figure G.2. This figure clearly shows that the proposed method remains valid under these settings since the estimated quantiles of the QQ-plots are very close to their expected counterparts.

G.3 Estimation of ζ and h^2 under the null hypothesis

In Table G.1 below, we report the results of the estimation of h^2 and ζ for the simulations conducted under the null hypothesis with conditions reported in Section 3.2.5. This table clearly shows that the estimation of ζ and h^2 under the null hypothesis is very accurate.

G.4 Type I error rates when the genetic variants are associated with the non-genetic covariates

Finally, in this simulations set, we investigated the validity of the proposed method when the genetic variants are associated with the non-genetic covariates. We considered three scenarios: in the first two, the genetic variants are associated with a discrete and a continuous

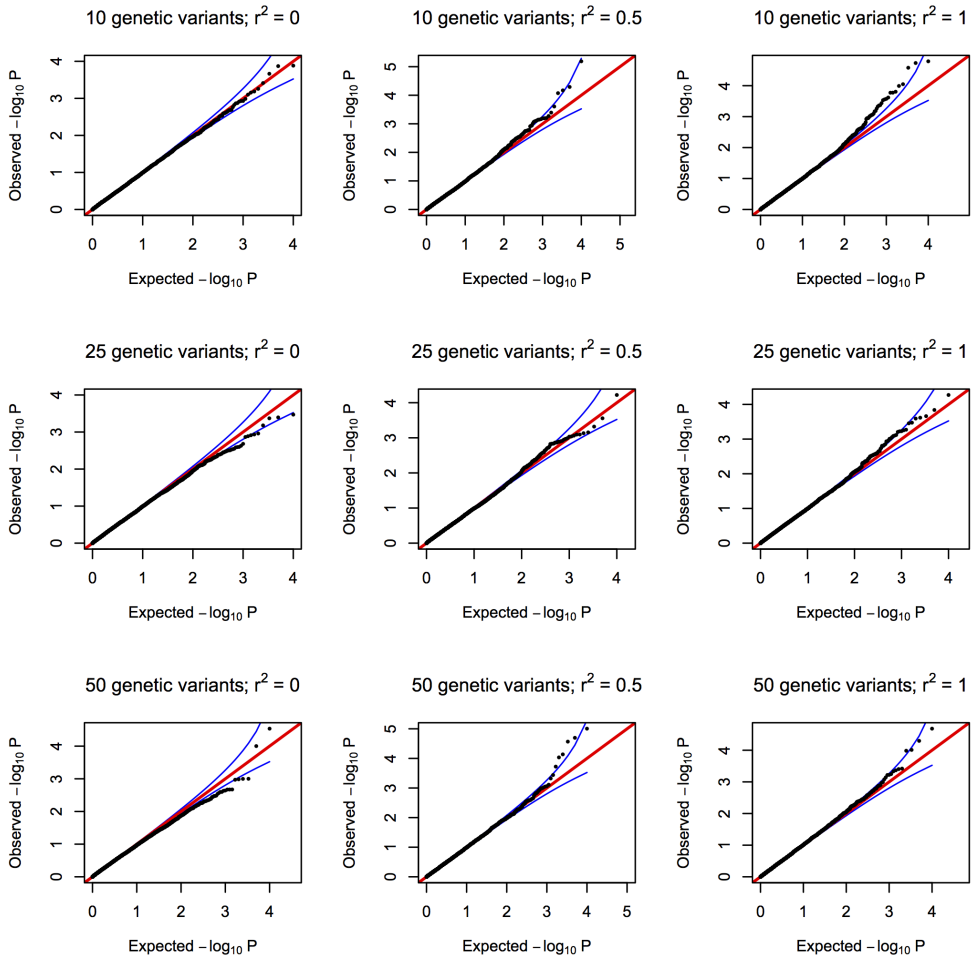

Figure G.2 – QQ plots of p-values from 10,000 replications under H_0 using the kinship-adjusted association test for various numbers of genetic variants and various values of linkage disequilibrium (r^2) between consecutive variants. The dependence structure between survival times is induced via the Gaussian copula model with heritability parameter h^2 fixed at 0.5. The censoring rate is set equal to 50%. The rows correspond to $s = 10$, 25 and 50 variants, respectively whereas the columns refer to $r^2 = 0$, 0.5 and 1, respectively.

Table G.1 – Mean estimates of the heritability h^2 and non genetic covariate (ζ_1 and ζ_2) parameters under H_0 (10,000 replications). ζ_1 refers to the Uniform covariate whereas ζ_2 refers to the Bernoulli covariate. True values for (ζ_1, ζ_2) are $(1, 1)$.

		Heritability parameter h^2			
Estimate	Censoring rate	0	.25	.50	.75
h^2	0	0.03	0.25	0.50	0.75
	25	0.03	0.25	0.50	0.75
	50	0.04	0.25	0.50	0.75
ζ_1	0	1.00	1.00	1.00	1.01
	25	1.01	1.00	1.01	1.00
	50	1.00	1.00	1.01	1.01
ζ_2	0	1.00	1.00	1.01	1.01
	25	1.00	1.00	1.00	1.00
	50	1.00	1.01	1.00	1.00

non-genetic covariate, respectively. In the last scenario, the association between the genetic variants and the non-genetic covariates is induced through a linkage disequilibrium. For each of these settings, 10,000 replications were generated under the null hypothesis. The dependence structure between survival times is induced via the Gaussian copula model with an heritability parameter h^2 fixed at 0.5 and a censoring rate equal to 50%. For each simulated dataset, we computed the p-value using the proposed association test with $m = 50$ imputations. The QQ-plots of these p-values are reported in Figures G.3, G.4, and G.5 respectively.

These figures clearly show that the proposed association test remains valid when the genetic variants are associated with the non-genetic covariates. These results were expected since we correct for the non-genetic covariates.

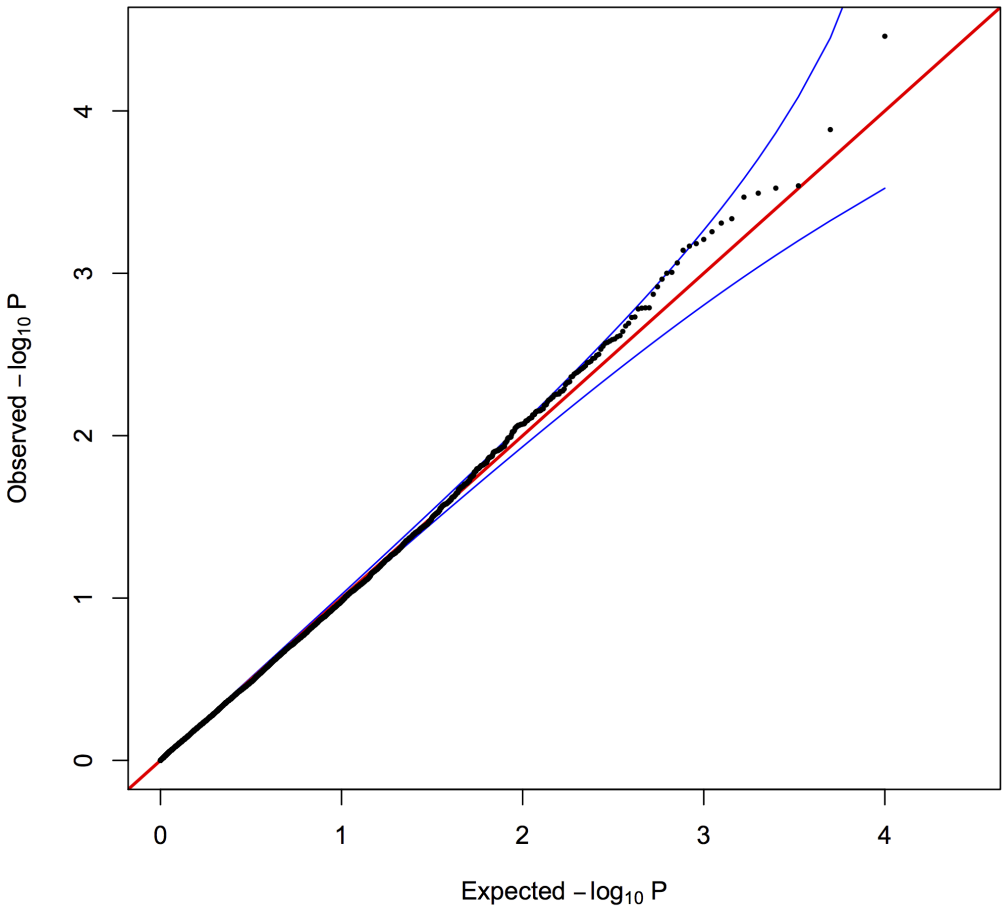

Figure G.3 – QQ plot of p-values from 10,000 replications under H_0 using the kinship-adjusted association test where $X_{i2} \sim \text{Bernoulli}(p_i)$ with $p_i = \frac{\exp\{\kappa(G_i)\}}{1+\exp\{\kappa(G_i)\}}$ and $\kappa(G_i) = 5G_{i1} - 4G_{i2} + 3G_{i3} - 2G_{i4} + G_{i5} - G_{i6} + 2G_{i7} - 3G_{i8} + 4G_{i9} - 5G_{i10}$. The dependence structure between survival times is induced via the Gaussian copula model with an heritability parameter h^2 fixed at 0.5. The censoring rate is set equal to 50%.

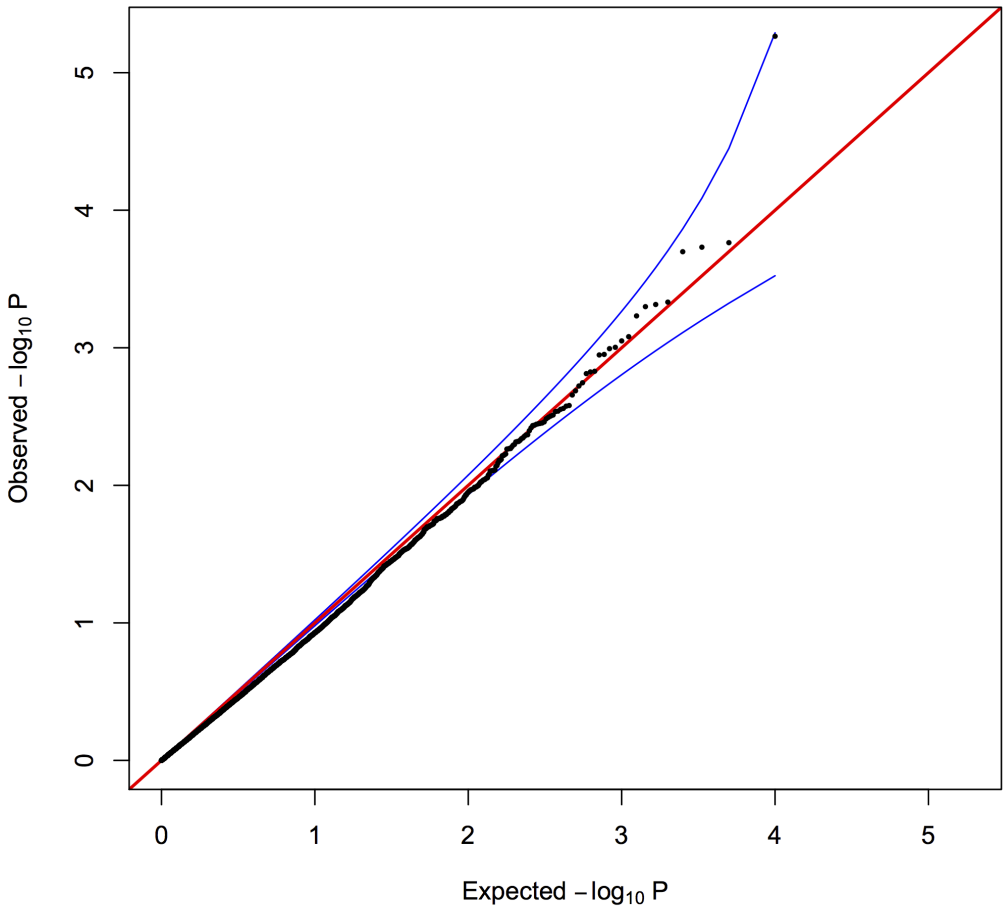

Figure G.4 – QQ plot of p-values from 10,000 replications under H_0 using the kinship-adjusted association test where $X_{i1} \sim \text{Unif}(\gamma_{i1}, \xi_{i1})$ with $\gamma_{i1} = -0.2 - |\kappa(G_i)|/5$ and $\xi_{i1} = 0.2 + |\kappa(G_i)|/20$. The dependence structure between survival times is induced via the Gaussian copula model with an heritability parameter h^2 fixed at 0.5. The censoring rate is set equal to 50%.

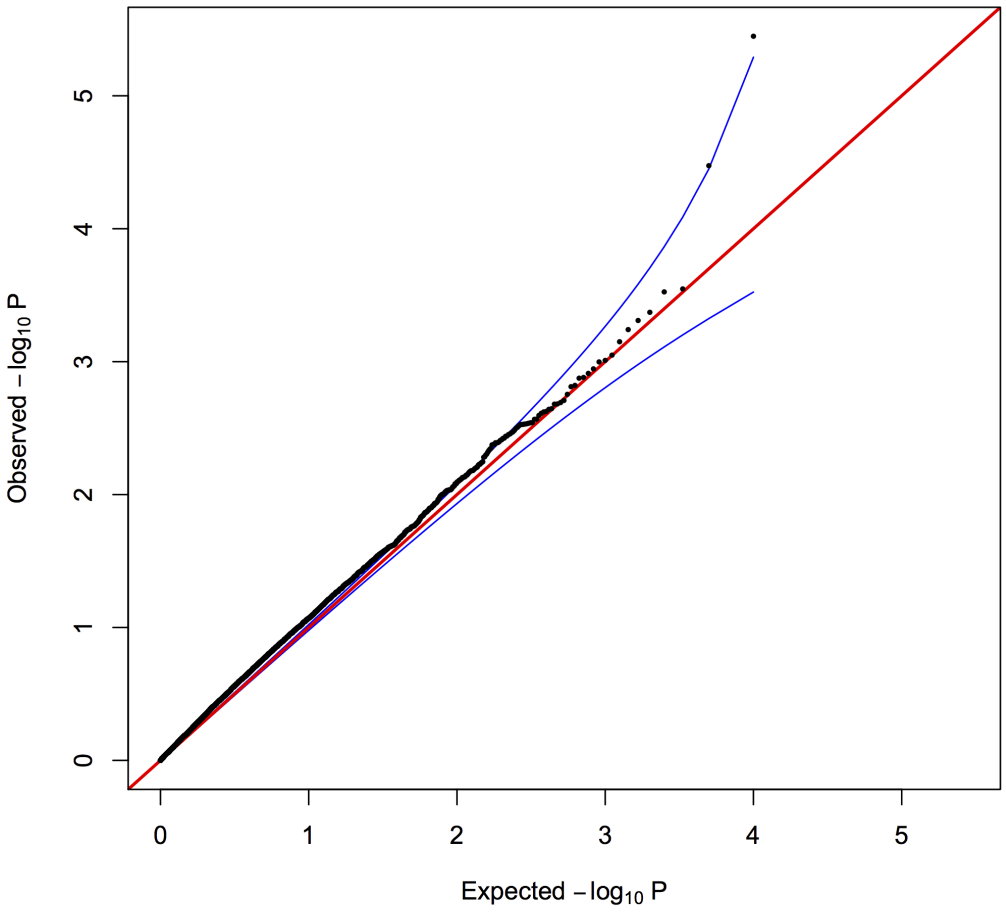

Figure G.5 – QQ plot of p-values from 10,000 replications under H_0 using the kinship-adjusted association test where $X_{i2} = G_{i,11}$ with $G_{i,11}$ as the eleventh simulated genotype with minor allele frequency equal to 0.30 and a squared correlation coefficient of $r^2 = 0.5$ with $G_{i,5}$ and $G_{i,6}$. The dependence structure between survival times is induced via the Gaussian copula model with an heritability parameter h^2 fixed at 0.5. The censoring rate is set equal to 50%.

Appendix H

Reference manual of the R package *gyriq*

Package 'gyriq'

January 7, 2016

Type Package

Title Kinship-Adjusted Survival SNP-Set Analysis

Version 1.0.2

Date 2016-01-06

Author Martin Leclerc and Lajmi Lakhal Chaieb

Maintainer Martin Leclerc <martin.leclerc.5@ulaval.ca>

Description
SNP-set association testing for censored phenotypes in the presence of intrafamilial correlation.

Imports CompQuadForm, irlba, mvtnorm, survival

Suggests snowfall

License GPL (>= 2)

NeedsCompilation yes

Depends R (>= 2.10)

Repository CRAN

Date/Publication 2016-01-07 13:55:34

R topics documented:

`gyriq-package` *Kinship-Adjusted Survival SNP-Set Analysis*

Description

SNP-set association testing for censored phenotypes in the presence of intrafamilial correlation

Details

Package:	gyriq
Type:	Package
Version:	1.0.2
Date:	2016-01-06
License:	GPL (>= 2)

This variance-components test between a set of SNPs and a survival trait is valid for both common and rare variants. A proportional hazards Cox model (written as a transformation model with censored data; Cheng et al., 1995) is specified for the marginal distribution of the survival trait. The familial dependence is modelled via a Gaussian copula with a correlation matrix expressed in terms of the kinship matrix. The statistical procedure has been described in full detail by Leclerc et al. (2015).

Censored values are treated as partially missing data and a multiple imputation procedure is employed to estimate vectors of residuals. These residuals and the SNPs in the genomic region under study are used to compute measures of phenotypic and genotypic similarity between pairs of subjects. The contribution to the score statistic is maximal when these measures are both high which corresponds to departure from the null hypothesis of no association between the set of SNPs and the survival outcome. The selection of the SNPs forming the SNP set can be based on biological information such as linkage disequilibrium (LD) blocks or rely on a sliding window method.

The procedure is convenient for GWAS as the multiple imputation procedure for the estimation of a completed vector of residuals has to be performed only once using the function genComplResid. A sliding window approach can then be used to examine the evidence of association across the SNP set. In each run, the p-value is computed with the function testGyriq.

Author(s)

Martin Leclerc <martin.leclerc.5@ulaval.ca> and Lajmi Lakhal Chaieb <lakhal@mat.ulaval.ca>

References

Cheng SC, Wei LJ, Ying Z. 1995. Analysis of transformation models with censored data. Biometrika 82:835-845.

Leclerc M, The Consortium of Investigators of Modifiers of BRCA1/2, Simard J, Lakhal-Chaieb L. 2015. SNP set association testing for survival outcomes in the presence of intrafamilial correlation. Genetic Epidemiology 39:406-414.

Lin X, Zhou Q. 2015. coxKM: Cox kernel machine SNP-set association test. R package version 0.3, URL http://www.hsph.harvard.edu/xlin/software.html#coxkm.

Lin X, Cai T, Wu M, Zhou Q, Liu G, Christiani D, Lin X. 2011. Survival kernel machine SNP-set analysis for genome-wide association studies. Genetic Epidemiology 35:620-631.

Cai T, Tonini G, Lin X. 2011. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67:975-986.

Examples

```
data(simGyriq)
for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])

cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
testGyriq(cr$compResid, G, w, ker="LIN", asv=NULL, method="davies",
starResid=NULL, bsw, tsw, pos)
```

`genComplResid` *genComplResid*

Description

Generates a completed vector of residuals

Usage

```
genComplResid(U, Delta, Phi, blkID, m = 50, X = NULL)
```

Arguments

`U`	a nx1 vector containing the survival times. `U = min(C, T)` where `C` is the censoring time, and `T` the failure time
`Delta`	a nx1 vector containing the censoring indicator
`Phi`	a nxn kinship matrix
`blkID`	a nx1 vector with entries identifying correlated groups of observations. The number of censored individuals in each group cannot exceed 1000 (see **Details**)
`m`	default=50. Number of imputations used to generate the completed vector of residuals
`X`	a nxp **matrix** of p covariates. Each row represents a different individual, and each column represents a different numeric covariate. If no covariates are present, `X` can be left as `NULL`

Details

This function involves three steps. The first two are similar in spirit to the two-stage procedure of Othus and Li (2010).

1. The vector of covariate parameters and the monotone increasing function of the transformation model with censored data (Cheng et al., 1995) are estimated under the working independence assumption following the algorithm of Chen et al. (2002) and used to compute raw residuals;
2. The polygenic heritability parameter is estimated which is a measure of the dependence between the survival traits of correlated groups that cannot be attributed to the SNP set under investigation. This estimate is used to deduce the approximate covariance matrix of the raw residuals.
3. An imputation procedure is employed to replace the censored raw residuals by the mean of multiple imputed values generated from the posterior distribution of the uncensored version with the restriction to be larger than the original censored values, componentwise. The completed vector of residuals is then deduced and standardized. A scale parameter is used to reflect the fact that we are using multiple imputed values rather than real observations.

Warning: Correlated groups identified by the vector `blkID` most often corresponds to families or blocks of the block-diagonal kinship matrix **Phi**. Larger groups such as regions of residence can be considered, for example to take into account population stratification or cryptic relatedness. However, the number of censored individuals in each group cannot exceed 1000 as the test makes use of the distribution function of the multivariate normal distribution for which the maximum dimension is 1000 in the function **pmvnorm** of the package **mvtnorm**.

Simulation studies reported in Leclerc et al. (2015) suggest that the use of `m = 50` imputations guarantees a reasonable power in practice.

Warning: No missing data is allowed for `U`, `Delta`, `Phi`, `blkID`, and `X`.

Value

The function produces a list consisting of:

`compResid`	the completed vector of residuals
`herit`	the estimate of the polygenic heritability parameter
`covPar`	the estimate of the vector of covariate parameters (if applicable)

Author(s)

Martin Leclerc <martin.leclerc.5@ulaval.ca> and Lajmi Lakhal Chaieb <lakhal@mat.ulaval.ca>

References

Chen K, Jin Z, Ying Z. 2002. Semiparametric analysis of transformation models with censored data. Biometrika 89:659-668.

Cheng SC, Wei LJ, Ying Z. 1995. Analysis of transformation models with censored data. Biometrika 82:835-845.

Leclerc M, The Consortium of Investigators of Modifiers of BRCA1/2, Simard J, Lakhal-Chaieb L. 2015. SNP set association testing for survival outcomes in the presence of intrafamilial correlation. Genetic Epidemiology 39:406-414.

Othus M, Li Y. 2010. A gaussian copula model for multivariate survival data. Stat Biosci 2:154-179.

Examples

```
data(simGyriq)
for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])

cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
```

`simGyriq` *Simulated SNP-set*

Description

Simulated dataset of phenotypic, genotypic and kinship data.

Format

A list containing the following elements:

U 600x1 vector containing the survival times. `U = min(C, T)` where `C` is the censoring time, and `T` the failure time

Delta 600x1 vector containing the censoring indicator

Phi 600x600 kinship matrix

blkID 600x1 vector with entries identifying correlated groups of observations

X 600x2 matrix of 2 covariates

G 600x50 matrix containing the set of 50 SNPs

w 50x1 vector of weights for the 50 SNPs

bsw 4x1 vector containing the lower bounds of the 4 sliding windows considered for the SNP-set

tsw 4x1 vector containing the upper bounds of the 4 sliding windows considered for the SNP-set

pos 50x1 vector of SNP positions (used for the output only)

indResid 10,000*600x1 vector of permuted row indices

Details

This dataset was generated under conditions described in Leclerc et al. (2015).

Samples of n = 600 individuals from 120 families were generated: 40 families of two parents and one child, 40 families of two parents and two children, and 40 families of three generations (two grand-parents, four parents, and two grandchildren). The coefficients of the block diagonal kinship matrix were fixed at their expected theoretical values. The number of biallelic SNPs was set to s = 50. The minor allele frequencies were randomly sampled from Unif(0.001, 0.1). The genotypes of the 50 SNPs were simulated assuming a linkage disequilibrium corresponding to a squared correlation coefficient of r^2 = 0.5 between consecutive SNPs.

The two covariates follow Bernoulli(0.5) and Uniform(-0.2, 0.2) distributions respectively. The polygenic heritability parameter was fixed at 0.5. Each covariate parameter was set equal to 1 and

the monotone increasing function of the transformation model with censored data (Cheng et al., 1995) was fixed at H(t) = log(t) in order to generate the survival traits. The censoring rate was equal to 50%. The weight of each SNP was defined as the density function of the Beta (1, 25) evaluated at the corresponding minor allele frequency.

The dataset includes simulated positions for the 50 SNPs, and the lower and upper bounds of 4 sliding windows. Each window includes 10 SNPs, overlapping with the previous and subsequent windows. A vector of size B*n of permuted row indices is also included, where B=10,000. This is to be used to compute the p-value of the test following the standard or matching moments permutation approach.

References

Cheng SC, Wei LJ, Ying Z. 1995. Analysis of transformation models with censored data. Biometrika 82:835-845.

Leclerc M, The Consortium of Investigators of Modifiers of BRCA1/2, Simard J, Lakhal-Chaieb L. 2015. SNP set association testing for survival outcomes in the presence of intrafamilial correlation. Genetic Epidemiology 39:406-414.

Examples

```
data(simGyriq)
for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])

cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
testGyriq(cr$compResid, G, w, ker="LIN", asv=NULL, method="davies",
starResid=NULL, bsw, tsw, pos)
```

`testGyriq` *testGyriq*

Description

Calculates the p-value of the kinship-adjusted SNP-set association test for censored traits

Usage

```
testGyriq(compResid, G, w, ker = "LIN", asv = NULL, method = "davies",
  starResid = NULL, bsw = NULL, tsw = NULL, pos = NULL, sf = FALSE,
  fileOut = "outGyriq.out")
```

Arguments

`compResid`	a nx1 vector containing the completed residuals
`G`	a nxs matrix containing the set of SNPs. Each row represents a different individual and each column represents a separate SNP. The SNP genotypes should be equal to the number of copies of the minor allele (0, 1 or 2).
`w`	a sx1 vector of weights for the s SNPs

`ker` (default="LIN") Type of kernel matrix: weighted linear ("LIN") or weighted identical-by-state ("IBS")

`asv` (default=NULL) Number of approximate eigenvalues to be estimated for the kernel matrix using the implicitly-restarted Lanczos bidiagonalization implemented in the package **irlba** (Baglama and Reichel, 2005). If the spectral decomposition of the matrix is to be conducted using the R base function **eigen**, `asv` can be left as `NULL`. This argument has no effect if `method` is not equal to "davies".

`method` (default="davies") Procedure used to obtain the p-value of the test. "davies" represents the approximation of Davies (1980), "rspMom" represents the permutation approach based on matching moments described in Lee et al. (2012), and "rspOrd" represents the standard permutation procedure.

`starResid` (default=NULL) a Bxn **matrix** of permuted residuals used to obtain the p-value of the test following a permutation procedure (method based on matching moments or standard permutation method). Each row represents a different permutation sample, and each column represents a different individual. This argument has no effect if `method` is not equal to "rspOrd" or "rspMom".

`bsw` (default=NULL) a vx1 vector containing the lower bounds of the v sliding windows considered for the SNP-set, taking values between 1 and s

`tsw` (default=NULL) a vx1 vector containing the upper bounds of the v sliding windows considered for the SNP-set, taking values between 1 and s

`pos` (default=NULL) a sx1 vector of SNP positions

`sf` (default=FALSE) logical: indicates whether or not cluster computing is used via the package **snowfall** in order to reduce wall-clock time. Initialisation and loading of the package **gyriq** on all nodes including master must be called beforehand using the functions `sfInit` and `sfLibrary` respectively. See the reference manual of `snowfall` for details. When cluster computing is used, the p-value for each sliding window is computed on a separate node.

`fileOut` (default="outGyriq.out") a string containing the name and path of the output file where the results are printed (used only if lower and upper bounds of sliding windows are also given as input; the file is appended for each sliding window in order to reduce resource wastage)

Details

If the lower and upper bounds of sliding windows are not provided, the test is performed once on the whole SNP-set `G`. Otherwise, the score statistic and the p-value are computed for each window sequentially.

In each run, the score statistic, which has a quadratic form following a mixture of chi-squared variables, is calculated from the completed vector of residuals and a kernel matrix. The p-value is obtained using a permutation approach based on matching moments described in Lee et al. (2012), a standard permutation procedure or the Davies approximation (Davies, 1980) implemented in the package **CompQuadForm** (Duchesne and Lafaye De Micheaux, 2010).

Warning: No missing data is allowed for `compResid`, `G`, `w` and `starResid`.

Value

If the lower and upper bounds of sliding windows are not provided, the function produces a list consisting of:

`score`	the score statistic of the test
`pVal`	the p-value

Otherwise, the function produces a data frame where each row represents a sliding window tested. For each window, the following information is provided:

- `FirstSNP`: Rank of the SNP corresponding to the lower bound of the sliding window in the SNP-set
- `LastSNP`: Rank of the SNP corresponding to the upper bound of the sliding window in the SNP-set
- `winSize`: Number of SNPs in the sliding window
- `Start`: Position of the SNP corresponding to the lower bound of the sliding window
- `Stop`: Position of the SNP corresponding to the upper bound of the sliding window
- `Score`: Score statistic of the association test
- `P-value`: P-value of the association test
- `Message`: If the calculation of the p-value failed, the corresponding error message is given. Otherwise, "OK" is displayed.

Author(s)

Martin Leclerc <martin.leclerc.5@ulaval.ca> and Lajmi Lakhal Chaieb <lakhal@mat.ulaval.ca>

References

Baglama J, Reichel L. 2005. Augmented implicitly restarted Lanczos bidiagonalization methods. SIAM J Sci Comput 27:19-42.

Davies RB. 1980. The distribution of a linear combination of χ^2 random variables. J R Stat Soc Ser C 29:323-333.

Lee S, Emond MJ, Bamshad MJ et al. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 91:224-237.

Duchesne P, Lafaye De Micheaux P. 2010. Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. Comput Stat Data Anal 54:858-862.

Lin X, Zhou Q. 2015. coxKM: Cox kernel machine SNP-set association test. R package version 0.3, URL http://www.hsph.harvard.edu/xlin/software.html#coxkm.

Lin X, Cai T, Wu M, Zhou Q, Liu G, Christiani D, Lin X. 2011. Survival kernel machine SNP-set analysis for genome-wide association studies. Genetic Epidemiology 35:620-631.

Cai T, Tonini G, Lin X. 2011. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67:975-986.

Examples

```
data(simGyriq)
for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])

cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
testGyriq(cr$compResid, G, w, ker="LIN", asv=NULL, method="davies",
starResid=NULL, bsw, tsw, pos)
```

Index

Bibliographie

[1] Adewale, A. J., I. Dinu, J. D. Potter, Q. Liu, and Y. Yasui (2008). Pathway analysis of microarray data via regression. *Journal of Computational Biology 15*(3), 269–277.

[2] Amin, N., C. M. van Duijn, and Y. S. Aulchenko (2007). A genomic background based method for association analysis in related individuals. *PLoS One 2*(12), e1274.

[3] Andersen, P. K., R. B. Geskus, T. de Witte, and H. Putter (2012). Competing risks in epidemiology : possibilities and pitfalls. *International Journal of Epidemiology 41*(3), 861–870.

[4] Antoniou, A., P. D. P. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, N. Loman, H. Olsson, O. Johannsson, Å. Borg, B. Pasini, P. Radice, S. Manoukian, D. M. Eccles, N. Tang, E. Olah, H. Anton-Culver, E. Warner, J. Lubinski, J. Gronwald, B. Gorski, H. Tulinius, S. Thorlacius, H. Eerola, H. Nevanlinna, K. Syrjäkoski, O.-P. Kallioniemi, D. Thompson, C. Evans, J. Peto, F. Lalloo, D. G. Evans, and D. F. Easton (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history : a combined analysis of 22 studies. *American Journal of Human Genetics 72*(5), 1117–1130.

[5] Antoniou, A. C., J. Beesley, L. McGuffog, O. M. Sinilnikova, S. Healey, S. L. Neuhausen, et al. (2010). Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers : implications for risk prediction. *Cancer Research 70*(23), 9742–9754.

[6] Antoniou, A. C., D. E. Goldgar, N. Andrieu, J. Chang-Claude, R. Brohet, M. A. Rookus, and D. F. Easton (2005). A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes. *Genetic Epidemiology 29*(1), 1–11.

[7] Astle, W. and D. J. Balding (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science 24*(4), 451–471.

[8] Baglama, J. and L. Reichel (2015). *irlba : Fast Truncated SVD, PCA and Symmetric Eigendecomposition for Large Dense and Sparse Matrices*. R package version 2.0.0.

[9] Barnes, D. R. and A. C. Antoniou (2012). Unravelling modifiers of breast and ovarian cancer risk for BRCA1 and BRCA2 mutation carriers : update on genetic modifiers. *Journal of Internal Medicine 271*(4), 331–343.

[10] Barnes, D. R., D. Barrowdale, J. Beesley, X. Chen, kConFab Investigators, Australian Ovarian Cancer Study Group, P. A. James, J. L. Hopper, D. Goldgar, G. Chenevix-Trench, A. C. Antoniou, and G. Mitchell (2013). Estimating single nucleotide polymorphism associations using pedigree data : applications to breast cancer. *British Journal of Cancer 108*(12), 2610–2622.

[11] Basrak, B., C. A. J. Klaassen, M. Beekman, N. G. Martin, and D. I. Boomsma (2004). Copulas in QTL mapping. *Behavior Genetics 34*(2), 161–171.

[12] Begg, C. B. (2002). On the use of familial aggregation in population-based case probands for calculating penetrance. *Journal of the National Cancer Institute 94*(16), 1221–1226.

[13] Begg, C. B., R. W. Haile, Å. Borg, K. E. Malone, P. Concannon, D. C. Thomas, B. Langholz, L. Bernstein, J. H. Olsen, C. F. Lynch, H. Anton-Culver, M. Capanu, X. Liang, A. J. Hummer, C. Sima, and J. L. Bernstein (2008). Variation of breast cancer risk among BRCA1/2 carriers. *Journal of the American Medical Association 299*(2), 194–201.

[14] Bojesen, S. E., K. A. Pooley, S. E. Johnatty, J. Beesley, K. Michailidou, J. P. Tyrer, S. L. Edwards, H. A. Pickett, H. C. Shen, C. E. Smart, K. M. Hillman, P. L. Mai, K. Lawrenson, M. D. Stutz, Y. Lu, R. Karevan, N. Woods, R. L. Johnston, J. D. French, X. Chen, M. Weischer, S. F. Nielsen, M. J. Maranian, M. Ghoussaini, S. Ahmed, C. Baynes, M. K. Bolla, Q. Wang, J. Dennis, L. McGuffog, D. Barrowdale, a. Lee, S. Healey, M. Lush, D. C. Tessier, D. Vincent, F. Bacot, I. Vergote, S. Lambrechts, E. Despierre, H. A. Risch, A. Gonzalez-Neira, M. A. Rossing, G. Pita, J. A. Doherty, N. Alvarez, M. C. Larson, B. L. Fridley, N. Schoof, J. Chang-Claude, M. S. Cicek, J. Peto, K. R. Kalli, A. Broeks, S. M. Armasu, M. K. Schmidt, L. M. Braaf, B. Winterhoff, H. Nevanlinna, G. E. Konecny, D. Lambrechts, L. Rogmann, P. Guenel, A. Teoman, R. L. Milne, J. J. Garcia, A. Cox, V. Shridhar, B. Burwinkel, F. Marme, R. Hein, E. J. Sawyer, C. A. Haiman, S. Wang-Gohrke, I. L. andrulis, K. B. Moysich, J. L. Hopper, K. Odunsi, A. Lindblom, G. G. Giles, H. Brenner, J. Simard, G. Lurie, P. A. Fasching, M. E. Carney, P. Radice, L. R. Wilkens, A. Swerdlow, M. T. Goodman, H. Brauch, M. Garcia-Closas, P. Hillemanns, R. Winqvist, M. Durst, P. Devilee, I. Runnebaum, A. Jakubowska, J. Lubinski, A. Mannermaa, R. Butzow, N. V. Bogdanova, T. Dork, L. M. Pelttari, W. Zheng, A. Leminen, H. Anton-Culver, C. H. Bunker, V. Kristensen, R. B. Ness, K. Muir, R. Edwards, A. Meindl, F. Heitz, K. Matsuo, A. du Bois, A. H. Wu, P. Harter, S.-H. Teo, I. Schwaab, X.-O. Shu, W. Blot, S. Hosono, D. Kang, T. Nakanishi, M. Hartman, Y. Yatabe, U. Hamann, B. Y. Karlan, S. Sangrajrang, S. K. Kjaer, V. Gaborieau, A. Jensen, D. Eccles, E. Hogdall, C.-Y. Shen, J. Brown, Y. L. Woo,

M. Shah, M. A. N. Azmi, R. Luben, S. Z. Omar, K. Czene, R. A. Vierkant, B. G. Nordestgaard, H. Flyger, C. Vachon, J. E. Olson, X. Wang, D. A. Levine, A. Rudolph, R. P. Weber, D. Flesch-Janys, E. Iversen, S. Nickels, J. M. Schildkraut, I. D. S. Silva, D. W. Cramer, L. Gibson, K. L. Terry, O. Fletcher, A. F. Vitonis, C. E. van der Schoot, E. M. Poole, F. B. L. Hogervorst, S. S. Tworoger, J. Liu, E. V. Bandera, J. Li, S. H. Olson, K. Humphreys, I. Orlow, C. Blomqvist, L. Rodriguez-Rodriguez, K. Aittomaki, H. B. Salvesen, T. A. Muranen, E. Wik, B. Brouwers, C. Krakstad, E. Wauters, M. K. Halle, H. Wildiers, L. A. Kiemeney, C. Mulot, K. K. Aben, P. Laurent-Puig, A. M. Altena, T. Truong, L. F. A. G. Massuger, J. Benitez, T. Pejovic, J. I. A. Perez, M. Hoatlin, M. P. Zamora, L. S. Cook, S. P. Balasubramanian, L. E. Kelemen, a. Schneeweiss, N. D. Le, C. Sohn, A. Brooks-Wilson, I. Tomlinson, M. J. Kerin, N. Miller, C. Cybulski, B. E. Henderson, J. Menkiszak, F. Schumacher, N. Wentzensen, L. Le Marchand, H. P. Yang, A. M. Mulligan, G. Glendon, S. A. Engelholm, J. A. Knight, C. K. Hogdall, C. Apicella, M. Gore, H. Tsimiklis, H. Song, M. C. Southey, A. Jager, A. M. W. den Ouweland, R. Brown, J. W. M. Martens, J. M. Flanagan, M. Kriege, J. Paul, S. Margolin, N. Siddiqui, G. Severi, A. S. Whittemore, L. Baglietto, V. McGuire, C. Stegmaier, W. Sieh, H. Muller, V. Arndt, F. Labreche, Y.-T. Gao, M. S. Goldberg, G. Yang, M. Dumont, J. R. McLaughlin, A. Hartmann, A. B. Ekici, M. W. Beckmann, C. M. Phelan, M. P. Lux, J. Permuth-Wey, B. Peissel, T. A. Sellers, F. Ficarazzi, M. Barile, A. Ziogas, A. Ashworth, A. Gentry-Maharaj, M. Jones, S. J. Ramus, N. Orr, U. Menon, C. L. Pearce, T. Bruning, M. C. Pike, Y.-D. Ko, J. Lissowska, J. Figueroa, J. Kupryjanczyk, S. J. Chanock, A. Dansonka-Mieszkowska, A. Jukkola-Vuorinen, I. K. Rzepecka, K. Pylkas, M. Bidzinski, S. Kauppila, A. Hollestelle, C. Seynaeve, R. A. E. M. Tollenaar, K. Durda, K. Jaworska, J. M. Hartikainen, V.-M. Kosma, V. Kataja, N. N. Antonenkova, J. Long, M. Shrubsole, S. Deming-Halverson, A. Lophatananon, P. Siriwanarangsan, S. Stewart-Brown, N. Ditsch, P. Lichtner, R. K. Schmutzler, H. Ito, H. Iwata, K. Tajima, C.-C. Tseng, D. O. Stram, D. van den Berg, C. H. Yip, M. K. Ikram, Y.-C. Teh, H. Cai, W. Lu, L. B. Signorello, Q. Cai, D.-Y. Noh, K.-Y. Yoo, H. Miao, P. T.-C. Iau, Y. Y. Teo, J. McKay, C. Shapiro, F. Ademuyiwa, G. Fountzilas, C.-N. Hsiung, J.-C. Yu, M.-F. Hou, C. S. Healey, C. Luccarini, S. Peock, D. Stoppa-Lyonnet, P. Peterlongo, T. R. Rebbeck, M. Piedmonte, C. F. Singer, E. Friedman, M. Thomassen, K. Offit, T. V. O. Hansen, S. L. Neuhausen, C. I. Szabo, I. Blanco, J. Garber, S. A. Narod, J. N. Weitzel, M. Montagna, E. Olah, a. K. Godwin, D. Yannoukakos, D. E. Goldgar, T. Caldes, E. N. Imyanitov, L. Tihomirova, B. K. Arun, I. Campbell, A. R. Mensenkamp, C. J. van Asperen, K. E. P. van Roozendaal, H. Meijers-Heijboer, J. M. Collee, J. C. Oosterwijk, M. J. Hooning, M. A. Rookus, R. B. van der Luijt, T. A. M. Os, D. G. Evans, D. Frost, E. Fineberg, J. Barwell, L. Walker, M. J. Kennedy, R. Platte, R. Davidson, S. D. Ellis, T. Cole, B. Bressac-de Paillerets, B. Buecher, F. Damiola, L. Faivre, M. Frenay, O. M. Sinilnikova, O. Caron, S. Giraud, S. Mazoyer, V. Bonadona, V. Caux-Moncoutier, A. Toloczko-Grabarek, J. Gronwald, T. Byrski, A. B. Spurdle, B. Bonanni, D. Zaffaroni, G. Giannini, L. Bernard, R. Dolcetti,

S. Manoukian, N. Arnold, C. Engel, H. Deissler, K. Rhiem, D. Niederacher, H. Plendl, C. Sutter, B. Wappenschmidt, A. Borg, B. Melin, J. Rantala, M. Soller, K. L. Nathanson, S. M. Domchek, G. C. Rodriguez, R. Salani, D. G. Kaulich, M.-K. Tea, S. S. Paluch, Y. Laitman, A.-B. Skytte, T. A. Kruse, U. B. Jensen, M. Robson, A.-M. Gerdes, B. Ejlertsen, L. Foretova, S. A. Savage, J. Lester, P. Soucy, K. B. Kuchenbaecker, C. Olswold, J. M. Cunningham, S. Slager, V. S. Pankratz, E. Dicks, S. R. Lakhani, F. J. Couch, P. Hall, A. N. A. Monteiro, S. A. Gayther, P. D. P. Pharoah, R. R. Reddel, E. L. Goode, M. H. Greene, D. F. Easton, a. Berchuck, A. C. Antoniou, G. Chenevix-Trench, and A. M. Dunning (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature Genetics 45*(4), 371–384.

[15] Braun, T. M. and Z. Feng (2001). Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association 96*(456), 1424–1432.

[16] Brohet, R. M., M. E. Velthuizen, F. B. L. Hogervorst, H. E. J. Meijers-Heijboer, C. Seynaeve, M. J. Collée, S. Verhoef, M. G. E. M. Ausems, N. Hoogerbrugge, C. J. van Asperen, E. Gómez García, F. Menko, J. C. Oosterwijk, P. Devilee, L. J. van't Veer, F. E. van Leeuwen, D. F. Easton, M. A. Rookus, A. C. Antoniou, and HEBON Resource (2014). Breast and ovarian cancer risks in a large series of clinically ascertained families with a high proportion of BRCA1 and BRCA2 Dutch founder mutations. *Journal of Medical Genetics 51*, 98–107.

[17] Cai, T., G. Tonini, and X. Lin (2011). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics 67*(3), 975–986.

[18] Chatterjee, N., Z. Kalaylioglu, J. H. Shih, and M. H. Gail (2006). Case-control and case-only designs with genotype and family history data : estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics 62*(1), 36–48.

[19] Chen, C.-M. and T.-F. C. Lu (2012). Marginal analysis of multivariate failure time data with a surviving fraction based on semiparametric transformation cure models. *Computational Statistics & Data Analysis 56*(3), 645–655.

[20] Chen, C.-M. and C.-Y. Yu (2012). A two-stage estimation in the Clayton-Oakes model with marginal linear transformation models for multivariate failure time data. *Lifetime Data Analysis 18*(1), 94–115.

[21] Chen, H., T. Lumley, J. Brody, N. L. Heard-Costa, C. S. Fox, L. A. Cupples, and J. Dupuis (2014). Sequence kernel association test for survival traits. *Genetic Epidemiology 38*(3), 191–197.

[22] Chen, K., Z. Jin, and Z. Ying (2002). Semiparametric analysis of transformation models with censored data. *Biometrika 89*(3), 659–668.

[23] Chen, X., L. Wang, J. D. Smith, and B. Zhang (2008). Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics 24*(21), 2474–2481.

[24] Cheng, S. C., L. J. Wei, and Z. Ying (1995). Analysis of transformation models with censored data. *Biometrika 82*(4), 835–845.

[25] Cherubini, U., E. Luciano, and W. Vecchiato (2004). *Copula Methods in Finance*. Chichester : John Wiley & Sons.

[26] Couch, F. J., X. Wang, L. McGuffog, A. Lee, C. Olswold, K. B. Kuchenbaecker, P. Soucy, Z. Fredericksen, D. Barrowdale, J. Dennis, M. M. Gaudet, E. Dicks, M. Kosel, S. Healey, O. M. Sinilnikova, A. Lee, F. Bacot, D. Vincent, F. B. L. Hogervorst, S. Peock, D. Stoppa-Lyonnet, A. Jakubowska, kConFab Investigators, P. Radice, R. K. Schmutzler, SWE-BRCA, S. M. Domchek, M. Piedmonte, C. F. Singer, E. Friedman, M. Thomassen, Ontario Cancer Genetics Network, T. V. O. Hansen, S. L. Neuhausen, C. I. Szabo, I. Blanco, M. H. Greene, B. Y. Karlan, J. Garber, C. M. Phelan, J. N. Weitzel, M. Montagna, E. Olah, I. L. andrulis, a. K. Godwin, D. Yannoukakos, D. E. Goldgar, T. Caldes, H. Nevanlinna, A. Osorio, M. B. Terry, M. B. Daly, E. J. van Rensburg, U. Hamann, S. J. Ramus, A. Ewart Toland, M. A. Caligo, O. I. Olopade, N. Tung, K. Claes, M. S. Beattie, M. C. Southey, E. N. Imyanitov, M. Tischkowitz, R. Janavicius, E. M. John, A. Kwong, O. Diez, J. Balmaña, R. B. Barkardottir, B. K. Arun, G. Rennert, S.-H. Teo, P. A. Ganz, I. Campbell, A. H. van der Hout, C. H. M. van Deurzen, C. Seynaeve, E. B. Gómez Garcia, F. E. van Leeuwen, H. E. J. Meijers-Heijboer, J. J. P. Gille, M. G. E. M. Ausems, M. J. Blok, M. J. L. Ligtenberg, M. A. Rookus, P. Devilee, S. Verhoef, T. A. M. van Os, J. T. Wijnen, HEBON, EMBRACE, D. Frost, S. Ellis, E. Fineberg, R. Platte, D. G. Evans, L. Izatt, R. A. Eeles, J. Adlard, D. M. Eccles, J. Cook, C. Brewer, F. Douglas, S. Hodgson, P. J. Morrison, L. E. Side, A. Donaldson, C. Houghton, M. T. Rogers, H. Dorkins, J. Eason, H. Gregory, E. McCann, A. Murray, A. Calender, A. Hardouin, P. Berthet, C. Delnatte, C. Nogues, C. Lasset, C. Houdayer, D. Leroux, E. Rouleau, F. Prieur, F. Damiola, H. Sobol, I. Coupier, L. Venat-Bouvet, L. Castera, M. Gauthier-Villars, M. Léoné, P. Pujol, S. Mazoyer, Y.-J. Bignon, GEMO Study Collaborators, E. Złowocka-Perłowska, J. Gronwald, J. Lubinski, K. Durda, K. Jaworska, T. Huzarski, A. B. Spurdle, A. Viel, B. Peissel, B. Bonanni, G. Melloni, L. Ottini, L. Papi, L. Varesco, M. G. Tibiletti, P. Peterlongo, S. Volorio, S. Manoukian, V. Pensotti, N. Arnold, C. Engel, H. Deissler, D. Gadzicki, a. Gehrig, K. Kast, K. Rhiem, A. Meindl, D. Niederacher, N. Ditsch, H. Plendl, S. Preisler-Adams, S. Engert, C. Sutter, R. Varon-Mateeva, B. Wappenschmidt, B. H. F. Weber, B. Arver, M. Stenmark-Askmalm, N. Loman, R. Rosenquist, Z. Einbeigi, K. L. Nathanson, T. R. Rebbeck, S. V. Blank, D. E. Cohn, G. C. Rodriguez, L. Small, M. Friedlander, V. L. Bae-Jump, A. Fink-Retter, C. Rappaport, D. Gschwantler-Kaulich, G. Pfeiler, M.-K. Tea, N. M. Lindor, B. Kaufman, S. Shimon Paluch, Y. Laitman, A.-B. Skytte, A.-M. Gerdes, I. S. Pedersen, S. T. Moel-

ler, T. A. Kruse, U. B. Jensen, J. Vijai, K. Sarrel, M. Robson, N. Kauff, A. M. Mulligan, G. Glendon, H. Ozcelik, B. Ejlertsen, F. C. Nielsen, L. Jønson, M. K. andersen, Y. C. Ding, L. Steele, L. Foretova, A. Teulé, C. Lazaro, J. Brunet, M. A. Pujana, P. L. Mai, J. T. Loud, C. Walsh, J. Lester, S. Orsulic, S. A. Narod, J. Herzog, S. R. Sand, S. Tognazzo, S. Agata, T. Vaszko, J. Weaver, A. V. Stavropoulou, S. S. Buys, A. Romero, M. de la Hoya, K. Aittomäki, T. A. Muranen, M. Duran, W. K. Chung, A. Lasa, C. M. Dorfling, A. Miron, BCFR, J. Benitez, L. Senter, D. Huo, S. B. Chan, A. P. Sokolenko, J. Chiquette, L. Tihomirova, T. M. Friebel, B. A. Agnarsson, K. H. Lu, F. Lejbkowicz, P. A. James, P. Hall, A. M. Dunning, D. Tessier, J. Cunningham, S. L. Slager, C. Wang, S. Hart, K. Stevens, J. Simard, T. Pastinen, V. S. Pankratz, K. Offit, D. F. Easton, G. Chenevix-Trench, and A. C. Antoniou on behalf of CIMBA (2013). Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genetics 9*(3), e1003212.

[27] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society : Series B (Methodological) 34*(2), 187–220.

[28] Crijns, A. P. G., R. S. N. Fehrmann, S. de Jong, F. Gerbens, G. J. Meersma, H. G. Klip, H. Hollema, R. M. W. Hofstra, G. J. te Meerman, E. G. E. de Vries, and A. G. J. van der Zee (2009). Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS Medicine 6*(2), e1000024.

[29] Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society : Series C (Applied Statistics) 29*(3), 323–333.

[30] Derkach, A., J. F. Lawless, and L. Sun (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology 37*(1), 110–121.

[31] Diao, G. and D. Y. Lin (2006). Semiparametric variance-component models for linkage and association analyses of censored trait data. *Genetic Epidemiology 30*(7), 570–581.

[32] Duchesne, P. and P. Lafaye De Micheaux (2010). Computing the distribution of quadratic forms : further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics & Data Analysis 54*(4), 858–862.

[33] Easton, D. F. (1999). How many more breast cancer predisposition genes are there ? *Breast Cancer Research 1*(1), 14–17.

[34] Edwards, S. L., J. Beesley, J. D. French, and A. M. Dunning (2013). Beyond GWASs : Illuminating the dark road from association to function. *American Journal of Human Genetics 93*(5), 779–797.

[35] Elandt-Johnson, R. C. (1971). Joint genotype distributions of s children and a parent, and of s siblings. Multiple alleles. *American Journal of Human Genetics 23*(5), 442–461.

[36] Evers, L. and C.-M. Messow (2008). Sparse kernel methods for high-dimensional survival data. *Bioinformatics 24*(14), 1632–1638.

[37] Gaudet, M. M., K. B. Kuchenbaecker, J. Vijai, R. J. Klein, T. Kirchhoff, L. McGuffog, D. Barrowdale, A. M. Dunning, A. Lee, J. Dennis, S. Healey, E. Dicks, P. Soucy, O. M. Sinilnikova, V. S. Pankratz, X. Wang, R. C. Eldridge, D. C. Tessier, D. Vincent, F. Bacot, F. B. L. Hogervorst, S. Peock, D. Stoppa-Lyonnet, P. Peterlongo, R. K. Schmutzler, K. L. Nathanson, M. Piedmonte, C. F. Singer, M. Thomassen, T. v. O. Hansen, S. L. Neuhausen, I. Blanco, M. H. Greene, J. Garber, J. N. Weitzel, I. L. andrulis, D. E. Goldgar, E. D'andrea, T. Caldes, H. Nevanlinna, A. Osorio, E. J. van Rensburg, A. Arason, G. Rennert, A. M. W. van den Ouweland, A. H. van der Hout, C. M. Kets, C. M. Aalfs, J. T. Wijnen, M. G. E. M. Ausems, D. Frost, S. Ellis, E. Fineberg, R. Platte, D. G. Evans, C. Jacobs, J. Adlard, M. Tischkowitz, M. E. Porteous, F. Damiola, L. Golmard, L. Barjhoux, M. Longy, M. Belotti, S. F. Ferrer, S. Mazoyer, A. B. Spurdle, S. Manoukian, M. Barile, M. Genuardi, N. Arnold, A. Meindl, C. Sutter, B. Wappenschmidt, S. M. Domchek, G. Pfeiler, E. Friedman, U. B. Jensen, M. Robson, S. Shah, C. Lazaro, P. L. Mai, J. Benitez, M. C. Southey, M. K. Schmidt, P. A. Fasching, J. Peto, M. K. Humphreys, Q. Wang, K. Michailidou, E. J. Sawyer, B. Burwinkel, P. Guénel, S. E. Bojesen, R. L. Milne, H. Brenner, M. Lochmann, K. Aittomäki, T. Dörk, S. Margolin, A. Mannermaa, D. Lambrechts, J. Chang-Claude, P. Radice, G. G. Giles, C. A. Haiman, R. Winqvist, P. Devillee, M. García-Closas, N. Schoof, M. J. Hooning, A. Cox, P. D. P. Pharoah, A. Jakubowska, N. Orr, A. González-Neira, G. Pita, M. R. Alonso, P. Hall, F. J. Couch, J. Simard, D. Altshuler, D. F. Easton, G. Chenevix-Trench, A. C. Antoniou, K. Offit, K. Investigators, Ontario Cancer Genetics Network, HEBON, EMBRACE, GEMO Study Collaborators, and The GENICA Network (2013). Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk. *PLoS Genetics 9*(3), e1003173.

[38] Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2015). *mvtnorm : Multivariate Normal and t Distributions*. R package version 1.0-3.

[39] Ghoussaini, M., P. D. P. Pharoah, and D. F. Easton (2013). Inherited genetic susceptibility to breast cancer. *American Journal of Pathology 183*(4), 1038–1051.

[40] Goeman, J. J. and J. Oosting (2015). *globaltest R package version 5.25.0*.

[41] Goeman, J. J., J. Oosting, A.-M. Cleton-Jansen, J. K. Anninga, and H. C. van Houwelingen (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics 21*(9), 1950–1957.

[42] Gong, G., N. Hannon, and A. S. Whittemore (2010). Estimating gene penetrance from family data. *Genetic Epidemiology 34*(4), 373–381.

[43] Haller, B., G. Schmidt, and K. Ulm (2013). Applying competing risks regression models : an overview. *Lifetime Data Analysis 19*(1), 33–58.

[44] Han, F. and W. Pan (2010). Powerful multi-marker association tests : unifying genomic distance-based regression and logistic regression. *Genetic Epidemiology 34*(7), 680–688.

[45] He, J., H. Li, A. C. Edmondson, D. J. Rader, and M. Li (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. *Biostatistics 13*(3), 497–508.

[46] Hsu, L. (2003). Genetic association tests with age at onset. *Genetic Epidemiology 24*(2), 118–127.

[47] Huang, X. and N. Zhang (2008). Regression survival analysis with an assumed copula for dependent censoring : A sensitivity analysis approach. *Biometrics 64*(4), 1090–1099.

[48] Huang, Y.-T., R. S. Heist, L. R. Chirieac, X. Lin, V. Skaug, S. Zienolddiny, A. Haugen, M. C. Wu, Z. Wang, L. Su, K. Asomaning, and D. C. Christiani (2009). Genome-wide analysis of survival in early-stage non–small-cell lung cancer. *Journal of Clinical Oncology 27*(16), 2660–2667.

[49] Ionita-Laza, I., S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin (2013). Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics 92*(6), 841–853.

[50] Jacquard, A. (1970). *Structures Génétiques des Populations*. Paris : Institut national d'études démographiques.

[51] Joe, H. (2015). *Dependence Modeling with Copulas*. Boca Raton : CRC Press.

[52] Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

[53] Karyadi, D. M., S. Zhao, Q. He, L. McIntosh, J. L. Wright, E. A. Ostrander, Z. Feng, and J. L. Stanford (2015). Confirmation of genetic variants associated with lethal prostate cancer in a cohort of men from hereditary prostate cancer families. *International Journal of Cancer 136*(9), 2166–2171.

[54] Kazma, R. and J. N. Bailey (2011). Population-based and family-based designs to analyze rare variants in complex diseases. *Genetic Epidemiology 35*(Suppl 1), S41–S47.

[55] Klein, J. P. and M. L. Moeschberger (2003). *Survival Analysis : Techniques for Censored and Truncated Data* (2nd ed.). New York : Springer.

[56] Knaus, J. (2015). *snowfall : Easier Cluster Computing (Based on snow)*. R package version 1.84-6.1.

[57] Kraft, P. and D. C. Thomas (2000). Bias and efficiency in family-based gene-characterization studies : conditional, prospective, retrospective, and joint likelihoods. *American Journal of Human Genetics 66*(3), 1119–1131.

[58] Kruglyak, L. and D. A. Nickerson (2001). Variation is the spice of life. *Nature Genetics 27*(3), 234–236.

[59] Lapidus, N., S. Chevret, and M. Resche-Rigon (2014). Assessing assay agreement estimation for multiple left-censored data : a multiple imputation approach. *Statistics in Medicine 33*(30), 5298–5309.

[60] Leclerc, M. and L. L. Chaieb (2016). *gyriq : Kinship-Adjusted Survival SNP-Set Analysis*. R package version 1.0.2.

[61] Leclerc, M., EMBRACE Investigators, GEMO Study Collaborators, INHERIT Investigators, A. C. Antoniou, J. Simard, and L. Lakhal-Chaieb (2015). Analysis of multivariate failure times in the presence of selection bias with application to breast cancer. *Journal of the Royal Statistical Society : Series C (Applied Statistics) 64*(3), 525–541.

[62] Leclerc, M., The Consortium of Investigators of Modifiers of BRCA1/2, J. Simard, and L. Lakhal-Chaieb (2015). SNP set association testing for survival outcomes in the presence of intrafamilial correlation. *Genetic Epidemiology 39*(6), 406–414.

[63] Lee, A. J. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *American Statistician 47*(3), 209–215.

[64] Lee, S., G. R. Abecasis, M. Boehnke, and X. Lin (2014). Rare-variant association analysis : study designs and statistical tests. *American Journal of Human Genetics 95*(1), 5–23.

[65] Lee, S., M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. Christiani, M. Wurfel, and X. Lin (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics 91*(2), 224–237.

[66] Lee, S., J. Kim, and S. Lee (2011). A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC Bioinformatics 12*, 377.

[67] Leutenegger, A.-L., B. Prum, E. Génin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E. A. Thompson (2003). Estimation of the inbreeding coefficient through use of genomic data. *American Journal of Human Genetics 73*(3), 516–523.

[68] Li, H. and Y. Luan (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. In *Pacific Symposium on Biocomputing*, Singapore, pp. 65–76. World Scientific.

[69] Li, M., M. Boehnke, G. R. Abecasis, and P. X.-K. Song (2006). Quantitative trait linkage analysis using Gaussian copulas. *Genetics 173*(4), 2317–2327.

[70] Lichtenstein, P., N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki (2000). Environmental and heritable factors in the causation of cancer – analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine 343*(2), 78–85.

[71] Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data : the marginal approach. *Statistics in Medicine 13*(21), 2233–2247.

[72] Lin, D. Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics 21*(6), 781–787.

[73] Lin, D. Y. (2014). Survival analysis with incomplete genetic data. *Lifetime Data Analysis 20*(1), 16–22.

[74] Lin, X., T. Cai, M. C. Wu, Q. Zhou, G. Liu, D. C. Christiani, and X. Lin (2011). Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology 35*(7), 620–631.

[75] Lin, X. and Q. Zhou (2015). *coxKM : Cox Kernel Machine SNP-Set Association Test.* R package version 0.3.

[76] Liu, L. X., S. Murray, and A. Tsodikov (2011). Multiple imputation based on restricted mean model for censored data. *Statistics in Medicine 30*(12), 1339–1350.

[77] Marceau, R., W. Lu, S. Holloway, M. M. Sale, B. B. Worrall, S. R. Williams, F.-C. Hsu, and J.-Y. Tzeng (2015). A fast multiple-kernel method with applications to detect gene-environment interaction. *Genetic Epidemiology 39*(6), 456–468.

[78] Mavaddat, N., S. Peock, D. Frost, S. Ellis, R. Platte, E. Fineberg, D. G. Evans, L. Izatt, R. A. Eeles, J. Adlard, R. Davidson, D. Eccles, T. Cole, J. Cook, C. Brewer, M. Tischkowitz, F. Douglas, S. Hodgson, L. Walker, M. E. Porteous, P. J. Morrison, L. E. Side, M. J. Kennedy, C. Houghton, A. Donaldson, M. T. Rogers, H. Dorkins, Z. Miedzybrodzka, H. Gregory, J. Eason, J. Barwell, E. McCann, A. Murray, A. C. Antoniou, and D. F. Easton on behalf of EMBRACE (2013). Cancer risks for BRCA1 and BRCA2 mutation carriers : results from prospective analysis of EMBRACE. *Journal of the National Cancer Institute 105*(11), 812–822.

[79] Mendolia, F., J. P. Klein, E. W. Petersdorf, M. Malkki, and T. Wang (2014). Comparison of statistics in association tests of genetic markers for survival outcomes. *Statistics in Medicine 33*(5), 828–844.

[80] Miki, Y., J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding, R. Bell, J. Rosenthal, C. Hussey, T. Tran, M. McClure, C. Frye, T. Hattier, R. Phelps, A. Haugen-Strano, H. Katcher, K. Yakumo, Z. Gholami, D. Shaffer, S. Stone, S. Bayer, C. Wray, R. Bogden, P. Dayananth, J. Ward, P. Tonin, S. Narod, P. K. Bristow, F. H. Norris, L. Helvering, P. Morrison, P. Rosteck, M. Lai, J. C. Barrett, C. Lewis, S. Neuhausen, L. Cannon-Albright, D. Goldgar, R. Wiseman, A. Kamb, and M. H. Skolnick (1994). A strong candidate for the breast and ovarian-cancer susceptibility gene BRCA1. *Science 266*(5182), 66–71.

[81] Milne, R. L. and A. C. Antoniou (2011). Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers. *Annals of Oncology 22*(suppl 1), i11–i17.

[82] Milne, R. L., A. Osorio, T. R. Cajal, A. Vega, G. Llort, M. de la Hoya, O. Díez, M. C. Alonso, C. Lazaro, I. Blanco, A. Sánchez-de-Abajo, T. Caldés, A. Blanco, B. Graña, M. Durán, E. Velasco, I. Chirivella, E. E. Cardeñosa, M.-I. Tejada, E. Beristain, M.-D. Miramar, M.-T. Calvo, E. Martínez, C. Guillén, R. Salazar, C. San Román, A. C. Antoniou, M. Urioste, and J. Benítez (2008). The average cumulative risks of breast and ovarian cancer for carriers of mutations in BRCA1 and BRCA2 attending genetic counseling units in Spain. *Clinical Cancer Research 14*(9), 2861–2869.

[83] Mitchell, G., A. C. Antoniou, R. Warren, S. Peock, J. Brown, R. Davies, J. Mattison, M. Cook, I. Warsi, D. G. Evans, D. Eccles, F. Douglas, J. Paterson, S. Hodgson, L. Izatt, T. Cole, L. Burgess, EMBRACE collaborators, R. Eeles, and D. F. Easton (2006). Mammographic density and breast cancer risk in BRCA1 and BRCA2 mutation carriers. *Cancer Research 66*(3), 1866–1872.

[84] Miyahara, S. and A. S. Wahed (2010). Weighted Kaplan-Meier estimators for two-stage treatment regimes. *Statistics in Medicine 29*(25), 2581–2591.

[85] Nelsen, R. B. (2006). *An Introduction to Copulas* (2e ed.). New York : Springer.

[86] Othus, M. and Y. Li (2010). A Gaussian copula model for multivariate survival data. *Statistics in Biosciences 2*(2), 154–179.

[87] Oualkacha, K., Z. Dastani, R. Li, P. E. Cingolani, T. D. Spector, C. J. Hammond, J. B. Richards, A. Ciampi, and C. M. T. Greenwood (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic Epidemiology 37*(4), 366–376.

[88] Patil, G. P. and C. R. Rao (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics 34*(2), 179–189.

[89] Pillas, D., C. J. Hoggart, D. M. Evans, P. F. O'Reilly, K. Sipilä, R. Lähdesmäki, I. Y. Millwood, M. Kaakinen, G. Netuveli, D. Blane, P. Charoen, U. Sovio, A. Pouta, N. Freimer,

A.-L. Hartikainen, J. Laitinen, S. Vaara, B. Glaser, P. Crawford, N. J. Timpson, S. M. Ring, G. Deng, W. Zhang, M. I. McCarthy, P. Deloukas, L. Peltonen, P. Elliott, L. J. M. Coin, G. D. Smith, and M.-R. Jarvelin (2010). Genome-wide association study reveals multiple loci associated with primary tooth development during infancy. *PLoS Genetics 6*(2), e1000856.

[90] Pinheiro, J. C. and D. M. Bates (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics 4*(1), 12–35.

[91] Pollard, K. S., S. Dudoit, and M. J. van der Laan (2005). Multiple testing procedures : R multtest package and applications to genomics. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 249–271. New York : Springer.

[92] R Core Team (2015). *R : A Language and Environment for Statistical Computing*. Vienna : R Foundation for Statistical Computing.

[93] Rao, C. R. (1973). *Linear Inference and its Applications* (2nd ed.). New York : Wiley.

[94] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : Wiley.

[95] Schaid, D. J., S. K. McDonnell, S. M. Riska, E. E. Carlson, and S. N. Thibodeau (2010). Estimation of genotype relative risks from pedigree data by retrospective likelihoods. *Genetic epidemiology 34*(4), 287–298.

[96] Shih, J. H. and T. A. Louis (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics 51*(4), 1384–1399.

[97] Simchoni, S., E. Friedman, B. Kaufman, R. Gershoni-Baruch, A. Orr-Urtreger, I. Kedar-Barnes, R. Shiri-Sverdlov, E. Dagan, S. Tsabari, M. Shohat, R. Catane, M.-C. King, A. Lahad, and E. Levy-Lahad (2006). Familial clustering of site-specific cancer risks associated with BRCA1 and BRCA2 mutations in the Ashkenazi Jewish population. *Proceedings of the National Academy of Sciences of the United States of America 103*(10), 3770–3774.

[98] Sinnott, J. A. and T. Cai (2013). Omnibus risk assessment via accelerated failure time kernel machine modeling. *Biometrics 69*(4), 861–873.

[99] Sinnott, J. A. and T. Cai (2014). High-dimensional regression models. In *Handbook of Survival Analysis*, pp. 93–112. Boca Raton : CRC Press.

[100] Tachmazidou, I., T. Andrew, C. J. Verzilli, M. R. Johnson, and M. De Iorio (2008). Bayesian survival analysis in genetic association studies. *Bioinformatics 24*(18), 2030–2036.

[101] Tachmazidou, I., M. R. Johnson, and M. De Iorio (2010). Bayesian variable selection for survival regression in genetics. *Genetic Epidemiology 34*(7), 689–701.

[102] Terwilliger, J. D., M. Speer, and J. Ott (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genetic Epidemiology 10*(4), 217–224.

[103] Therneau, T. M. (2015a). *coxme : Mixed Effects Cox Models*. R package version 2.2-5.

[104] Therneau, T. M. (2015b). *A Package for Survival Analysis in S*. version 2.38.

[105] Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data : Extending the Cox Model*. New York : Springer.

[106] Thomas, D. C. (2004). *Statistical Methods in Genetic Epidemiology*. New York : Oxford University Press.

[107] Thomas, D. C., Z. Yang, and F. Yang (2013). Two-phase and family-based designs for next-generation sequencing studies. *Frontiers in Genetics 4*(276).

[108] Thorlacius, S., S. Sigurdsson, H. Bjarnadottir, G. Olafsdottir, J. G. Jonasson, L. Tryggvadottir, H. Tulinius, and J. E. Eyfjörd (1997). Study of a single BRCA2 mutation with high carrier frequency in a small population. *American Journal of Human Genetics 60*(5), 1079–1084.

[109] Topol, E. J. (2014). Individualized medicine from prewomb to tomb. *Cell 157*(1), 241–253.

[110] Tzeng, J.-Y., W. Lu, and F.-C. Hsu (2014). Gene-level pharmacogenetic analysis on survival outcomes using gene-trait similarity regression. *Annals of Applied Statistics 8*(2), 1232–1255.

[111] Van Belle, V., K. Pelckmans, S. Van Huffel, and J. A. K. Suykens (2011). Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics 27*(1), 87–94.

[112] Vaupel, J. W., K. G. Manton, and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*(3), 439–454.

[113] Vogl, F. D., M. D. Badzioch, L. Steele, S. L. Neuhausen, and D. E. Goldgar (2007). Risks of cancer due to a single BRCA1 mutation in an extended Utah kindred. *Familial Cancer 6*(1), 63–71.

[114] Wang, X., N. J. Morris, X. Zhu, and R. C. Elston (2013). A variance component based multi-marker association test using family and unrelated data. *BMC Genetics 14*(17).

[115] Wei, L. J., D. Y. Lin, and L. Weissfeld (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association 84*(408), 1065–1073.

[116] Wienke, A. (2010). *Frailty Models in Survival Analysis*. Boca Raton : Chapman & Hall/CRC.

[117] Williams, R. L. (1995). Product-limit survival functions with correlated survival times. *Lifetime Data Analysis 1*(2), 171–186.

[118] Wooster, R., G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, G. Micklem, R. Barfoot, R. Hamoudi, S. Patel, C. Rice, P. Biggs, Y. Hashim, A. Smith, F. Connor, A. Arason, J. Gudmundsson, D. Ficenec, D. Kelsell, D. Ford, P. Tonin, D. T. Bishop, N. K. Spurr, B. A. J. Ponder, R. Eeles, J. Peto, P. Devilee, C. Cornelisse, H. Lynch, S. Narod, G. Lenoir, V. Egilsson, R. B. Barkadottir, D. F. Easton, D. R. Bentley, P. A. Futreal, A. Ashworth, and M. R. Stratton (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature 378*(6559), 789–792.

[119] Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics 89*(1), 82–93.

[120] Zhang, H., S. Olschwang, and K. Yu (2010). Statistical inference on the penetrances of rare genetic mutations based on a case-family design. *Biostatistics 11*(3), 519–532.