STATISTICAL MODELS AND METHODS FOR COMPLEX CLINICAL DATA

Doctoral Dissertation of:
Stefano Baraldo

Supervisor:
Prof. Anna Maria Paganoni

Tutor:
Prof. Anna Maria Paganoni

The Chair of the Doctoral Program:
Prof. Paolo Biscari

Year 2013 – XXV Cycle

Acknowledgements

I wish to thank all the people named above for their contribution to this work. Above all, I wish to thank my advisor prof. Anna Maria Paganoni, who encouraged my efforts and supported my ideas during the last three years, and taught me so much during my long stay at Politecnico di Milano.

I wish to thank also my fellow PhD students and the MOXstat group, for their direct and indirect contributions to the completion of this work and for the good time spent together. *Man shall not leave by research alone*.

Many friends and loved ones helped me get through these years as a young researcher and a young man. I thank all of them, with a special though for my family. Their love is the only contribution I couldn't do without.

Summary

THIS work presents two distinct applications of statistical modeling and estimation in problems related to health care: Chapter 1 is dedicated to the statistical modeling of hospital admission data and its use for predictive purposes in the analysis of a particular modern care strategy, the home telemonitoring for heart failure; Chapters 2 and 3 instead present various statistical methods for the estimation of diffusion parameters in Magnetic Resonance Imaging (MRI) data, taking into account the distributional properties of the observed signals.

More in detail, Chapter 1 presents briefly the health care practice of home telemonitoring for subjects affected by heart failure, and proposes a framework for analyzing its outcome (completion or early interruption). The prediction of the telemonitoring period's outcome is important for the management of the care program, in particular for deciding patient allocation and for evaluating costs and benefits. The analysis is carried out thanks to the integration of the Public Health Database, which allows to gain information on the clinical history of a subject, and of home telemonitoring survey data. In particular, hospital admissions are modeled as points of a counting process in which the hazard is time-varying and dependent on some covariates. This approach is by itself informative on the characteristics of heart failure patients, but can also be used as a way of summarizing the state of severity of the disease: the time-dependent hazard functions are used as functional covariates, to enrich the predictive model for telemonitoring outcome. A case study is presented, showing the advantages of this approach with respect to other classic models.

Chapter 2 introduces some basics of diffusion estimation in Diffusion-Weighted Imaging, starting with the estimation of Apparent Diffusion Coefficient (ADC) fields in isotropic tissues. The importance of taking into account a reliable model for noise, instead of a Gaussian approximation, is discussed, and some methods for pointwise estimation of Rician distirbution parameters are presented and compared. In this framework, we introduce maximum likelihood estimation of ADC fields coupled with an experimental design for the choice of b-values, aimed at increasing the estimation accuracy in interesting areas, and analyze the results of design optimization on both simulated and real data.

Chapter 3 considers data acquired by MRI scanners with multiple coils, generalizing the estimation framework to noncentral chi-distributed signals and to the Diffusion Tensor Imaging model. The distributional assumptions on the observed signal are exploited to set up a hypothesis test that aims at discriminating the background, pure noise area with the signal area on an MR volume (with a particular focus on brain imaging). Moreover, a likelihood maximization algorithm is presented within the noncentral chi/DTI framework, and its performances are investigated by a simulation study.

Contents

Contents

Counting process modeling for health care data - an application to heart failure telemonitoring

1.1 Introduction

Heart failure is a degenerative disease known worldwide as one of the most frequent causes of hospitalization among the eldest in the population. Since the frequency of crises undergone by a given patient increases along time, a growing employment of health care resources in terms of money, structures and personnel is needed. The necessity of a cost-effective solution for the care of this and other chronic pathologies has led to the experimentation of telemedicine as a possibly convenient strategy ([13], [24] and [47]).

The basic idea of telemonitoring is to keep the patient at home and to instruct her/him about the use of monitoring instruments, which send registered information (ECG, body weight, heart frequency, etc.) to the health institution by a network connection. The physician in charge evaluates received data to properly manage the home care program, for example by modifying drug doses and by scheduling visits.

Telemonitoring databases contain information mostly regarding the telemonitoring period itself, such as duration of the period, number of ECGs transmitted to the hospital, clinical parameters at starting and ending times, the NYHA[1] severity class of the patient's pathology, some features regarding the last hospitalization and others. The telemonitoring outcome, i.e. the conclusion of the planned period (usually 180 days, for what concerns the programs considered in this work) without interruption by adverse events, should be related to the patients' clinical history to get better insight into the effectiveness and applicability of this strategy. To this aim, in this study we consider

---

[1]The New York Heart Association classification divides in four classes the extent of heart failure: 1 is the less severe, 4 the most.

Chapter 1. Counting process modeling for health care data - an application to heart failure telemonitoring

Hospital Dimission Forms (*Schede di Dimissione Ospedaliera*, briefly SDO) extracted from the Italian administrative Public Health Database (PHD), which gathers detailed information about hospitalization periods. The use of information about hospitalizations to study telemonitoring outcome is an innovative approach, since no standard methodology exists to exploit this kind of data. Moreover, data conveyed by administrative databases are used for clinical purposes, to the best of our knowledge, for the first time in Italy and in the field of telemonitoring.

Since heart failure is a pathology that alternates phases of stability to sudden worsenings of the patient's condition, it is not possible to assume a stationary pattern for critical events. When dealing with time dependent observations of localized events, a natural modeling approach, yet new in the field of telemonitoring, is to consider each patient's hospitalizations as points of a non stationary, doubly stochastic counting process. The model we consider is a Cox-type one, a specification of the general class of models introduced in [40] and applied in [27] to the study of intervention effects after cancer relapse. This class of models allows to take into account many aspects that influence hospitalization risk and to compute the realized trajectories of the cumulative hazard process underlying the hospitalizations counting process; these longitudinal data reduce complex characteristics of the patient's clinical history to a single curve that represents each patient's instantaneous risk of hospitalization. Cumulative hazard processes are then studied in the light of functional data analysis techniques (see [44] for a general presentation of the subject) to identify their main features, and used to construct a generalized linear model with functional covariates for predicting telemonitoring outcome.

The methodology proposed in this work can be divided in two distinct phases, basing on the sources of data involved. First, hospital admission historical data from the PHD are used to fit a counting process model, since we reasonably assume that hospitalizations reflect the aggravation of the patient's condition. Hence, data gathered in the PHD are involved in the construction of hazard functions, longitudinal data that reflect the evolution of rehospitalization risk before the application of the treatment (in our case, telemonitoring). The second phase is motivated by the main objective of this work, which is to predict the outcome of a treatment (in our case, the regular conclusion of a telemonitoring period) basing on available data. In this part of the analysis the data collected in the clinical registry are enriched with information gained from the preceding clinical history of the patient, represented by the longitudinal data estimated in the first phase. The two parts of the analysis could be performed also as stand-alone, the first analyzing the risk of hospital readmission for a given time period and class of hospitalization causes (e.g. cardiovascular), the second being a supervised classification method independent from the first. Nonetheless, a problem driven interaction between the two sources of information can bring great improvements in the prediction of the outcome of interest.

The chapter is structured as follows. Section 1.2 describes the theoretical and methodological framework. For what concerns the extraction of pre-telemonitoring longitudinal information, the model for recurrent events is first introduced. Then the smoothing of cumulative hazard functions obtained by realized trajectories of the recurrent event processes is detailed, and the dimensional reduction via functional principal components is described. Generalized linear models with functional covariates are also pre-

sented, as they will be used for the prediction of the telemonitoring outcome. Section 1.3 presents in detail the motivating application, practical issues and results. Finally, section 1.4 contains some concluding remarks and a discussion of possible extensions of this work.

Part of the work described in this Chapter has been published in [7]. The framework for the analysis of telemonitoring clinical registries has been applied in [25] to the Nuove Reti Sanitarie database.

1.2 Theoretical framework

1.2.1 Model for recurrent events

First of all, we introduce the counting process model that describes the patient-specific progression of hospital readmission risk in time.

Let $(\mathcal{F}_t)_{t \in I}$ be a filtration associated to the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $I = [0, \tau]$. We define the counting process $(N(t))_{t \in I}$ adapted to $(\mathcal{F}_t)_{t \in I}$ as follows:

$$N(t) = \sum_{j=0}^{\infty} I\{S_j \leq \min(t, \tau)\}, \tag{1.1}$$

where S_j represents the calendar time of the j-th occurrence of the observed event and τ represents a random censoring time for the process.

Under the standard assumption that N is a submartingale such that the class of random variables $N(T)$, with T an arbitrary stopping time, is uniformly integrable, the Doob-Meyer decomposition theorem states that there exists a unique predictable, non decreasing, *cadlag* (right-continuous with left limits) and integrable compensator (or *cumulative hazard*) process $(\Lambda(t))_{t \in I}$ such that

$$M = N - \Lambda \tag{1.2}$$

is a zero-mean, uniformly integrable martingale (see for example [4]). Hence the distribution of event times is completely characterized by the knowledge of process Λ, on which modeling efforts should then be focused. We will assume that

$$\Lambda(t) = \int_0^t C(s)\lambda(s)ds, \tag{1.3}$$

where $C(s) = I\{s \leq \tau\}$ is the *at-risk process*, and $(\lambda(s))_{s \in I}$ is called *hazard function*, or *intensity process*.

A wide variety of models for the intensity process can be found in counting processes literature, ranging from simple Poisson processes to multiplicative hazard models [17], additive models, frailty and dynamic models (see for instance [4] and [1] for presentations and discussions on various possibilities). Our choice for the target problem is the following Cox-type model: for $i = 1, \ldots, n$, the ith subject has covariate vector $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{iq}(t))^T$ (eventually time dependent) and the intensity is

$$\lambda(t|\boldsymbol{X}_i) = \lambda_0(t)\alpha^{N_i(t^-)}e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(t)}, \tag{1.4}$$

where $\lambda_0(t)$ is an unknown baseline hazard function, α is a real parameter and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ a q-dimensional vector of real coefficients.

We choose to account for unobserved heterogeneity by using the dynamic component $\alpha^{N_i(t^-)}$ instead of a frailty variable, i.e. a multiplicative random effect. Dynamic and frailty modeling can be seen as two related methods for describing subject heterogeneity, but the former is more general and flexible (see [1] for a discussion on these two approaches); however, results obtained in the two cases are compared in section 1.3.2. The dependence of intensity on process state is modeled by the term $\alpha^{N_i(t^-)}$ because of its clear interpretation: values of α higher than 1 indicate that a new event implies a worsening of the patient's condition, increasing future rehospitalization risk, vice versa for α values lower than 1. We also assume the baseline intensity λ_0 to be a function of the total time t, but a wide range of choices is valid within the same framework; see for example the concept of *effective age* introduced in [40].

Adding a censoring variable to account for different observation intervals, the model for cumulative hazard can be written as follows, for patients $i = 1, \ldots, n$:

$$\Lambda_i(t|\boldsymbol{X}_i) = \int_0^t C_i(s)\lambda_0(s)\alpha^{N_i(s^-)} \exp[\boldsymbol{\beta}^T \boldsymbol{X}_i(s)]ds, \qquad (1.5)$$

where $C_i(s) = I\{s \leq \tau_i\}$, i.e. subjects have different, mutually independent censoring times τ_i. Independent censorship as defined in [31] can be assumed for the considered problem, as will be clear from the application described in section 1.3.

1.2.2 Cumulative hazard smoothing and reconstruction

Semiparametric estimation of cumulative hazard, as proposed in [41], produces a step function estimate $\widehat{\Lambda}_0$ of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ that has the following expression: defining t_j as the j-th observed jump time of the aggregated process $N(t) = \sum_{i=1}^n N_i(t)$ and $\tau = \max_{i=1,\ldots,n} \tau_i$, then

$$\widehat{\Lambda}_0(t) = \sum_{t_j \leq t} \frac{1}{\sum_{i=1}^n C_i(t_j)\hat{\alpha}^{N_i(t_j^-)}e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{X}_i(t_j)}}, \qquad t \in (0, \tau],$$

where $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are maximum likelihood estimates of α and $\boldsymbol{\beta}$.

Assuming the true underlying Λ_0 function to be absolutely continuous, we deal with the issue of smoothing its estimate $\widehat{\Lambda}_0$ before moving to the reconstruction of cumulative hazard process realizations for each patient. The function $\Lambda_0(t)$ has two features that we want to be preserved by the smoothing procedure: increasing monotonicity and $\Lambda_0(0) = 0$. A fast method for smoothing functional data while enforcing desired constraints has been proposed in [30] and consists in a minimum absolute deviation estimate of coefficients for a B-spline basis expansion: given a set of observations $\{(x_j, y_j)\}_{j=1,\ldots,m}$ from a function $y = f(x)$ to be smoothed, a set of knots $\{u_0 = 0, u_1, \ldots, u_{K-1}, u_K = \tau\}$ and a fixed polynomial degree d, find $\boldsymbol{a}^* = (a_0^*, \ldots, a_{K+d-1}^*)^T$ such that

$$\boldsymbol{a}^* = \operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^{k+d}} \sum_{j=1}^m \left| y_j - \sum_{k=0}^{K+d-1} a_k B_k^{(d)}(x_i) \right|, \qquad (1.6)$$

$B_0^{(d)}(x), \ldots, B_{K+d-1}^{(d)}(x)$ being the B-spline basis of degree d on the chosen set of knots. In our case $x_1, x_2, x_3, \ldots = 0, t_1, t_2, \ldots$, the global jump times, $f = \widehat{\Lambda}_0$ and

$y_j = \widetilde{\Lambda}_0(t_j)$. If basis functions of polynomial degree $d = 1, 2$ are used, then monotonicity, convexity and pointwise constraints can be written as linear constraints. Since the quantity to be minimized can also be written as a linear objective function, the problem can be solved with linear programming techniques, whose efficiency and reliability are ascertained.

We then reconstruct the realizations of processes $\Lambda_i(t)$ for $i = 1, \ldots, n$ under the chosen model, since we intend to use them as patient-specific functional data. As shown in Appendix A, it is possible to rewrite the model for intensity in a form that allows to plug in $\widetilde{\Lambda}_0$ instead of an estimate of λ_0.

As a qualitative validation of the fitted model, we can compare the averages of counting and cumulative hazard processes: taking conditional expectations in (1.2) we have $\mathbb{E}[\Lambda_i(t)|\mathbf{X}_i(t)] = \mathbb{E}[N_i(t)|\mathbf{X}_i(t)]$ for $i = 1, \ldots, n$. The comparison is not straightforward when curves have different censoring times, in particular if faster growing curves have higher probability of earlier censoring (this is common for risk curves, as frailer patients die earlier). Let $(f(t))_{t \in I}$ be a stochastic process and let τ be a stopping time for this process; then if $\mathbf{f} = (f_1, \ldots, f_n)^T$ is a set of trajectories of the process f, and $\{\tau_1, \ldots, \tau_n\}$ a set of realizations of τ, we can define the pointwise sample mean function as

$$\mu_n[\mathbf{f}](t) = \frac{1}{n(t)} \sum_{i=1}^{n} f_i(t) C_i(t), \qquad \forall t \in [0, \tau], \tag{1.7}$$

where $n(t) = \sum_{i=1}^{n} C_i(t)$, letting $C_i(t) = I(t \leq \tau_i)$ be the censoring process for subject i and $\tau = \max_{i=1,\ldots,n} \tau_i$. In our application this estimator is not monotone and underestimates expected values for large times. As an alternative, we use the following

$$\tilde{\mu}_n[\mathbf{f}](x_k) = \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{C_i(x_j)}{n(x_j)} \Big[f_i(x_j) - f_i(x_{j-1}) \Big], \qquad k = 1, \ldots, m, \tag{1.8}$$

with $\{x_0, \ldots, x_m\}$ an overset of $\{\tau_1, \ldots, \tau_n\}$ and $\tilde{\mu}_n[\mathbf{f}](x_0) = \sum_{i=1}^{n} f_i(x_0) C_i(x_0) = \sum_{i=1}^{n} f_i(x_0)$. Applying this estimator to $\{\widetilde{\Lambda}_i\}_{i=1,\ldots,n}$ and $\{N_i\}_{i=1,\ldots,n}$ enforces monotonicity by definition when all sample curves are monotone. As pointed out in [18], this estimator is unbiased and consistent, and in the case of positively correlated increments it is likely that $\mathbb{V}ar\{\tilde{\mu}_n[\mathbf{f}](t)\} < \mathbb{V}ar\{\mu_n[\mathbf{f}](t)\}$.

Although we specified the counting process model as above, it is important to underline that the methodology presented in the following can be applied to any kind of model that allows for a patient-specific reconstruction of time-dependent hazard.

1.2.3 Functional principal component analysis

From this point, the reconstructed realizations $\tilde{\Lambda}_i(t)$ are considered as functional data, and will be later included as covariates in a generalized linear model for telemonitoring outcome prediction. Since these data are high-dimensional, a common strategy is to perform a suitable dimensional reduction, choosing only relevant components of a proper basis expansion. This approach is similar to the one proposed in [29], where survival functions of cohorts of medflies are regarded as functional data.

Consider the functional ANOVA decomposition of data, as suggested in [37],

$$\widetilde{\Lambda}_i(t) = \mu(t) + D_i(t) + \varepsilon_i(t), \qquad i = 1, \ldots, n \tag{1.9}$$

where $\mu(t)$ is the population mean function, $D_i(t)$ is the residual for subject i and $\varepsilon_i(t)$ the noise term. One of the possibilities for representing $\widetilde{\Lambda}_i(t)$ is to use functional principal component analysis (FPCA), i.e. the Karhunen-Loève decomposition (see for example [20] for some theoretical results and [44] for practical details). At this point we assume that functional data are known on a common domain T, thus allowing to estimate a common Karhunen-Loève basis. Once eigenfunctions $\{\psi_k\}_{k\in\mathbb{N}}$ and eigenvalues $\{\nu_k\}_{k\in\mathbb{N}}$ of the covariance operator for $\widetilde{\Lambda}$ have been found, we express the functional ANOVA decomposition (1.9) in the following form

$$\widetilde{\Lambda}_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t) + \varepsilon_i(t), \qquad i = 1, \ldots, n,$$

where $\xi_{ik} = \int_T D_i(s)\psi_k(s)ds$ is the k-th score for subject i. Eigenfunction-eigenvalue couples $\{(\psi_k, \nu_k)\}_{k\in\mathbb{N}}$ completely explain modes of variation of data, in the sense that eigenfunctions ordered with respect to the associated eigenvalues represent orthonormal directions of decreasing residual variability. It is thus possible to represent data using just the most relevant modes of variation, represented by the first K elements of $\{\psi_k\}_{k\in\mathbb{N}}$. A natural criterion for choosing the number K of components to be included in the truncated representation is the following: fix a threshold c for the proportion of variance they should globally describe, which is based on eigenvalues, and choose the minimum K such that:

$$\frac{\sum_{k=1}^{K} \nu_k}{\sum_{k=1}^{m} \nu_k} \geq c,$$

where m is the number of abscissa values on which functional data are known. We then use the approximation

$$\widetilde{\Lambda}_i^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\psi_k(t) + \varepsilon_i(t), \qquad i = 1, \ldots, n.$$

For the sake of simplicity, from now on we will write $\widetilde{\Lambda}_i(t)$ even when its truncated basis expansion $\widetilde{\Lambda}_i^K(t)$ is used.

1.2.4 Generalized linear models with functional covariates

While the methodological elements presented up to this point regarded the extraction of longitudinal pre-telemonitoring information, the present section involves the analysis of data from clinical registries, eventually enriched with information obtained from estimated cumulative hazard trajectories.

Let us consider a response variable Y such that $Y_i \sim EF(\theta_i, \eta)$, i.e. Y_i for $i \in 1, \ldots, n$ belongs to the exponential family

$$f_{Y_i}(y|\theta_i, \eta) = \exp\left(\frac{y\theta_i - b(\theta_i)}{\eta} + c(y, \eta)\right)$$

$$\mathbb{E}[Y_i] = b'(\theta_i)$$
$$\mathbb{V}ar[Y_i] = \eta b''(\theta_i)$$

with b and c given functions; the *link function* g is s.t. $\mathbb{E}[Y_i] = g^{-1}(\theta_i)$ (i.e. $g^{-1} = b'$), while θ_i is a linear function of covariates related to subject i. This set of covariates is composed by data from the clinical registry, i.e. information recorded at the beginning of the telemonitoring period; we augment this dataset adding the functions $D_i(t)$, $i = 1, \ldots, n$ introduced in section 1.2.3, so that θ_i, $i = 1, \ldots, n$, assumes the (truncated) form

$$\theta_i = \int_T D_i(t)\delta(t)dt + \mathbf{w}_i^T \boldsymbol{\gamma} \approx \int_T \delta(t) \sum_{k=1}^{K} \zeta_{ik}\psi_k(t)dt + \mathbf{w}_i^T \boldsymbol{\gamma},$$

where $\delta : T \mapsto \mathbb{R}$ is a functional parameter, $\boldsymbol{\gamma} \in \mathbb{R}^p$ is a vector of parameters to be estimated and $\mathbf{w}_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$, is a vector of covariates available from the clinical survey. Notice that we used the K most relevant principal components to represent $D_i(t)$. If also $\delta(\cdot)$ is represented using the principal components basis, i.e. $\delta(t) = \sum_{j=1}^{K} \delta_j\psi_j(t)$, for the orthonormality of $\{\psi_k\}_{k\in\mathbb{N}}$ we obtain

$$\theta_i = \sum_{k=1}^{K} \zeta_{ik}\delta_k + \mathbf{w}_i^T \boldsymbol{\gamma}, \quad i = 1, \ldots, n.$$

In the end, the model is reduced to a classical generalized linear model, in which the unknowns are represented by the parameter vector $(\delta_1, \ldots, \delta_K, \gamma_1, \ldots, \gamma_p)$. In our application we will consider a logistic regression model, where the response variable is $Y_i \sim B(p_i)$ for $i \in 1, \ldots, n$ and $\theta_i = \log(p_i/(1 - p_i))$.

1.3 Application to telemonitoring data analysis and results

In Lombardia region an experimentation of heart failure telemonitoring started in 2003, involving 34 health care institutions (see [14] for an overview of program and protocols). Four studies (*Criteria*, *Piano Urbano*, *Nuove Reti Sanitarie* and *Telemaco*) were devoted to collect, under prior informed consent, information about telemonitoring periods, then gathered in a comprehensive database. Each record of this database refers to a single telemonitoring period, and contains anagraphical data of the involved patient, number of transmitted electrocardiograms, diagnosis and disease etiology at the last hospital admission and other relevant clinical quantities.

The enrollment protocol adopted during the period 2004–2008 includes adult citizens of Lombardia with a NYHA class (describing the severity of heart disease) of III or IV who experienced at least one hospitalization for heart failure during the 6 months before the beginning of telemonitoring. The telemonitoring period is planned for a 180-day duration, with possible re-enrollment under particular conditions. The period may be interrupted, by protocol, if a hospitalization lasting more than 8 days occurs or because of surgical intervention; we considered these two cases and decease as "negative" conclusions of the treatment, while other types of "external" events that forced interruption, such as a change of dwelling or the decision by the patient herself/himself to stop the therapy, have been considered as drop-outs.

Since data regarding telemonitoring periods have a limited scope, we requested an interrogation of the regional administrative PHD, to obtain hospital discharge data (SDOs) stored during the five years of interest. Each one of these records contains

extensive information about a single hospitalization, such as date, duration, drugs received, DRG (Diagnosis Related Group, a classification code for patients discharged from hospital, based on the type of resources used during the stay) and other data of clinical interest. This information is used to estimate pre-telemonitoring hazard functions, so the subset of records regarding just pre-telemonitoring hospital admissions for each patient has been extracted from the full database. Each one of the available subjects is identified by a unique code, derived from an anonymizing procedure applied to her/his identity number, and used to retrieve from the SDO database her/his hospitalization history in the period 2004–2008. The linkage and matching of information between these two databases resulted in the constitution of an initial sample of 1081 patients, anyway a further reduction was operated in the second part of the analysis, to include only subjects whose telemonitoring period started at least in 2006; in this way, a 2-year time window before telemonitoring is available for all of them, ensuring better background for predictive tasks.

The risk of hospital readmission is obviously zero during each hospital stay, which typically lasts some days (mean duration = 17.18 ± 28.03 days). We deal with this issue by removing the hospital stay period from the process time count and by counting just once consecutive hospitalization periods that were eventually registered as multiple in the database because of a patient's transfer to a different structure or for any other reason. The time variable is expressed in days passed from 1st January 2004.

The following analyses have been carried out using the statistical software R [43]. For hazard estimation packages `gcmrec` [26] and `frailtypack` [28] have been used, while package `cobs` [38] has been used for constrained smoothing.

1.3.1 Hazard estimation and dimensional reduction

The first step of the analysis consists in the estimation of model (1.5) for cumulative hazard functions, using the procedure explained in section 1.2.2.

The beginning of telemonitoring is introduced as a censoring time τ_i, $i = 1, \ldots, n$, for the hospitalization counting processes, assuming that this event does not influence preceding hospitalizations; this assumption seems reasonable, on the basis of the enrollment protocol.

Following medical advice, subject age is included as covariate $X_i(s)$ in (1.5), providing the following model for patients $i = 1, \ldots, n$

$$\Lambda_i(t|\boldsymbol{X}_i) = \int_0^t C_i(s)\lambda_0(s)\alpha^{N_i(s^-)} \exp[\beta X_i(s)]ds,$$

with $C_i(s) = I\{s \leq \tau_i\}$. Including age at $t = 0$ as a time-independent covariate would yield similar results in the following, anyway we include it as a continuous variable, coherently with the fact that hospital admissions occur for each patients at different distances from the beginning of observation time. Moreover, this choice is coherent with the assumption that the influence of a covariate on hazard does not depend on the particular time instant when we start observing the subject.

Estimated baseline cumulative hazard $\widehat{\Lambda}_0(t)$ is represented in Figure 1.1 (dashed line), while parameter estimates are shown in Table 1.1. We notice that parameter α, describing the effect of a new event on the risk of future hospitalizations, is significantly higher than 1, according to a one-sided hypothesis test; this means that a new

event represents an increase in rehospitalization risk. Parameter β, related to the age covariate, is surprisingly negative, indicating that the risk of rehospitalization is slightly lower for older patients; this could be explained by the fact that in the old population considered (the mean age is 67.82 ± 11.19) subjects who survived up to a higher age are the less frail ones.

	estimate	std. dev.	p-value
α	1.21	0.00887	$< 2 \cdot 10^{-16}*$
β	-0.00336	0.00172	0.051

Table 1.1: *Results of hazard parameters estimation. p-value $*$ refers to a test with null hypothesis $\alpha \leq 1$. Both tests are carried out with a normal approximation for maximum likelihood estimators.*

Since the nonparametric estimator used for $\Lambda_0(t)$ produces a step function, we perform a smoothing of this estimate with the method proposed in section 1.2.2; for the B-Spline basis, we choose polynomial order 2 and 20 equally spaced knots. A comparison between the nonparametric estimate and the B-spline smoothed estimate is shown in Figure 1.1. In this picture we also notice that the cumulative baseline hazard function $\widehat{\Lambda}_0(t)$ has a convex behavior, describing a gradual increase of instantaneous risk due to the disease, common to the whole population.

Once $\widetilde{\Lambda}_0$ has been computed, we can reconstruct individual cumulative hazard processes by the plug-in estimator proposed in Appendix A, obtaining the functional dataset shown in Figure 1.2a. To verify that the condition $\mathbb{E}[\Lambda_i(t)|X_i(t)] = \mathbb{E}[N_i(t)|X_i(t)]$ holds, it is possible to visualize the mean functions of point processes and of cumulative hazard processes, computed using the cumulative mean increment estimator $\tilde{\mu}_n$ suggested in section 1.2.2. To address the problem of computing this conditional expectation, we can split the sample in classes $A_{c_1}, A_{c_2}, \ldots$ of similar initial age and average on these classes to approximate $\mathbb{E}[\Lambda_i(t)|X_i(t)]$ and $\mathbb{E}[N_i(t)|X_i(t)]$. For example, the martingale residual trajectories and their average for subjects belonging to the age class $A_{60} = \{i : a_i \in (55, 65]\}$ are shown in Figure 1.2b; we can see that residuals $\widehat{M}_i(t) = N_i(t) - \widetilde{\Lambda}_i(t)$, $i \in A_{60}$, seem to have the expected behavior. Figure 1.3 shows a comparison between average curves computed using pointwise estimator μ_n 1.7 and estimator $\tilde{\mu}_n$ 1.7 respectively; in the left panel we notice that average curves estimated with μ_n are non monotone and heavily biased because of right censoring, while average curves estimated with $\tilde{\mu}_n$, depicted in the right panel, suffer from censoring only at the very right end of the domain.

A dimensional reduction of the functional dataset of cumulative hazard functions is operated via principal component analysis, so that FPC scores can be used as indicators of pre-telemonitoring risk status and be added to other variables retrieved from the telemonitoring survey in the subsequent logistic regression model. To avoid the problem of censoring, as previously mentioned, we choose patients for which at least 2 years of clinical history before telemonitoring are available in our records, and restrict the time window for our analyses to exactly the 2 years preceding telemonitoring. Doing so, we obtain a dataset of $n = 747$ curves, evaluated on a grid of length $m = 730$ (hazard functions were computed on a uniform time grid with daily spacing).

Before proceeding to principal component analysis, curves are centered by subtracting their mean function $\mu_n(t)$ (which coincides with estimator $\tilde{\mu}_n(t)$ on the chosen
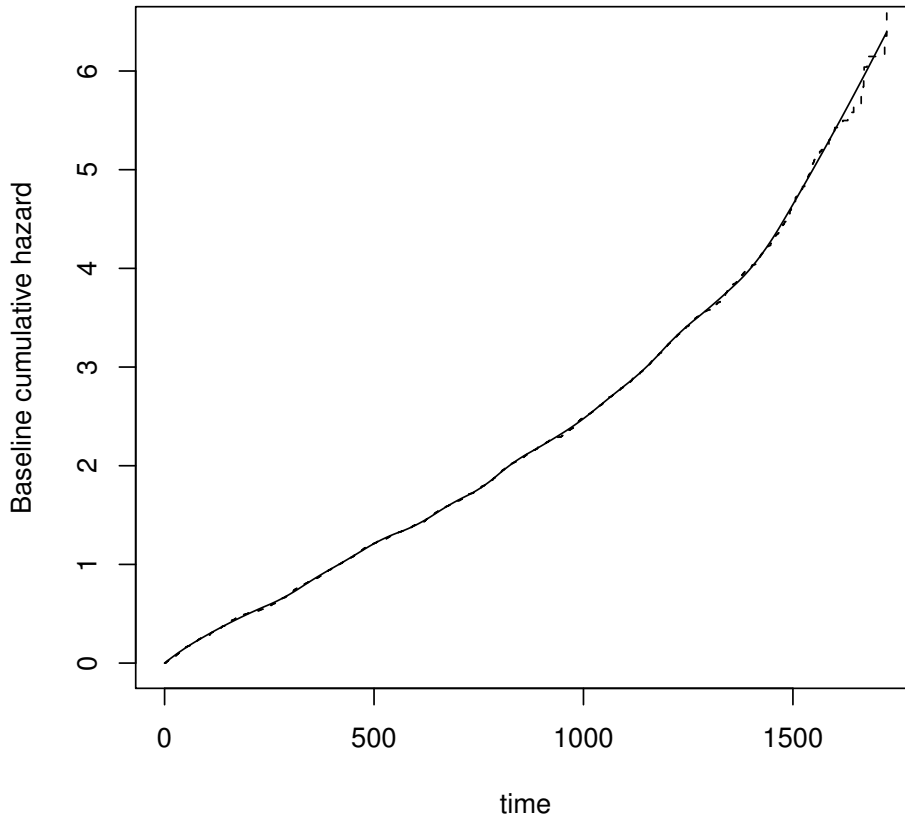
Figure 1.1: *Results of baseline cumulative hazard function estimation:* $\widehat{\Lambda}_0$ *(dashed line) and its smoothed version* $\widetilde{\Lambda}_0$ *(solid line).*

subset of data); moreover, the noise term $\varepsilon_i(t)$ is discarded, since curves have already been estimated with smoothness.

We shall now select the components to be considered in the subsequent analysis. A simple and effective criterion consists in choosing the first K components, such that their associated eigenvalues explain a proportion of variance $c > 95\%$. This criterion leads to the choice of the first $K = 2$ components, as detailed in Table 1.2.

	ν_1	ν_2
value	777.04	45.64
% variance	94.08	5.53
cum. % variance	94.08	99.60

Table 1.2: *First $K = 2$ eigenvalues obtained with FPCA.*

Figure 1.4 shows the two relevant functional principal components in the top panels and $\mu_n(t) \pm \nu_k \phi_k(t)$, $k = 1, 2$ in the bottom panels. The first component is monotone increasing, and is highly dominant in the description of data curves, while the second

(a) *Reconstructed realizations of cumulative hazard processes*

(b) *Trajectories of residuals* $\widehat{M_i}(t) = N_i(t) - \widetilde{\Lambda}_i(t)$, $i \in A_{60}$, *and their average (thick line).*
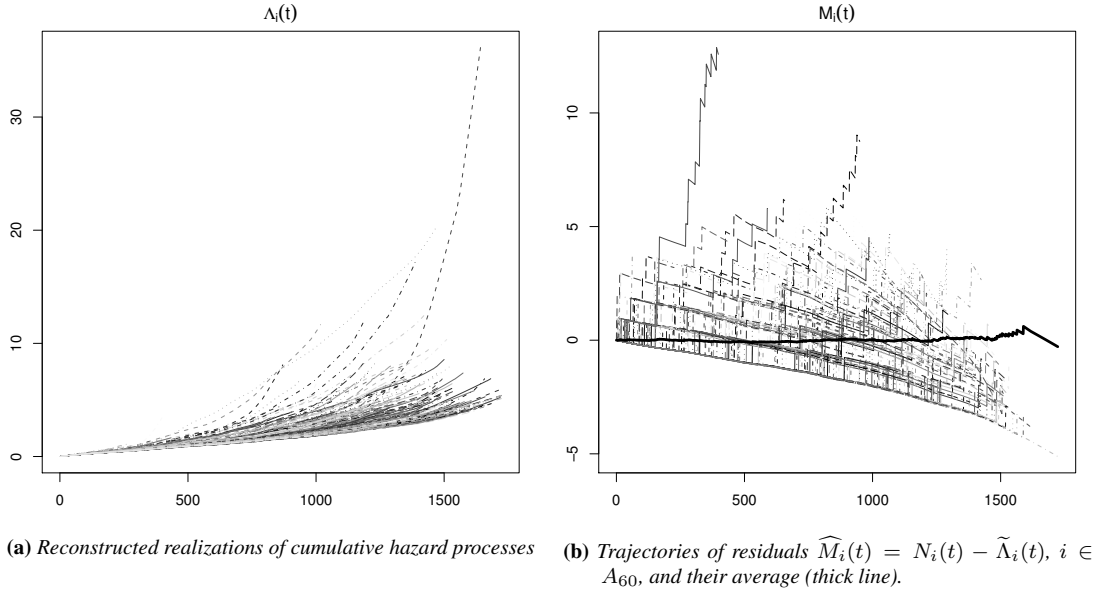
Figure 1.2: *Estimated trajectories and martingale residuals.*

one is decreasing, characterizing curves that do not grow very fast also on a long time period.

1.3.2 Generalized linear model estimation

The final goal of this work is the prediction of telemonitoring outcome, defined as a binary variable with value 1 if telemonitoring has regular conclusion and 0 if the period is terminated by an adverse event, i.e. hospitalization, surgical intervention or decease. The scores of principal components 1 and 2 of individual hazard functions are considered as covariates that summarize the features of the patients' pre-telemonitoring clinical history, and they are used to predict telemonitoring outcome together with a subset of the variables present in the telemonitoring database. In particular, clinicians suggested to take into account diagnosis and etiology of the last hospitalization before telemonitoring and sex. Variable diagnosis has 3 levels (*congestive*, *left* or *unspecified* heart failure), while etiology has 5 levels (*hypertensive*, *hyschaemic*, *primary*, *valve*, *other*). Following the notation introduced in section 1.2.4, the probability of normal outcome p_i takes the following form

$$p_i = \frac{exp\left(\sum_{k=1}^{2} \zeta_{ik}\delta_k + \mathbf{w}_i^T \boldsymbol{\gamma} \right)}{1 + exp\left(\sum_{k=1}^{2} \zeta_{ik}\delta_k + \mathbf{w}_i^T \boldsymbol{\gamma} \right)}, \quad \text{for } i = 1, \dots, n,$$

where $p_i = \mathbb{E}(Y_i|\mathbf{w}_i, \boldsymbol{\zeta}_{i1}, \boldsymbol{\zeta}_{i2})$ and $\mathbf{w}_i$ is composed by dummy variables representing cathegorical covariates sex, diagnosis and etiology.

The model output is reported in Table 1.3. Scores 1 and 2 are both significant, and their signs are coherent with a possible interpretation: principal component 1 is
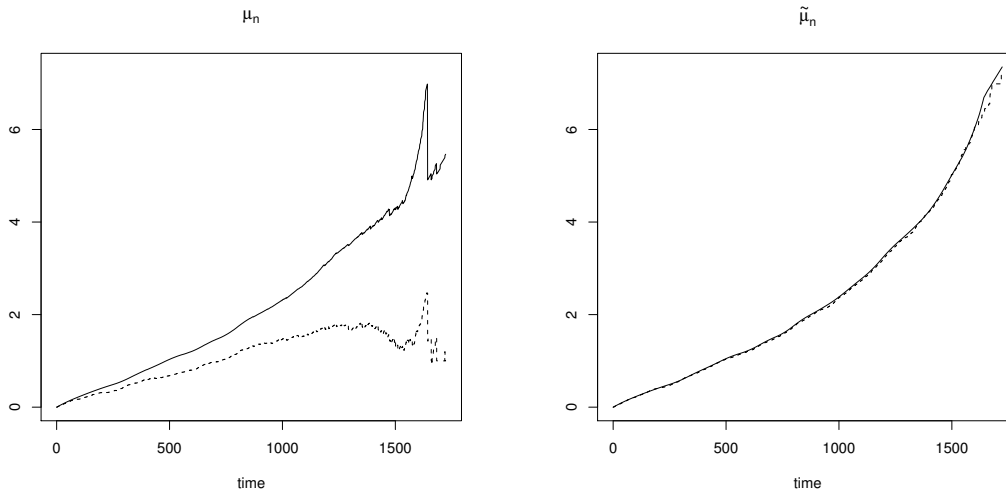
(a) *Mean curves obtained with the pointwise estimator μ_n.*   **(b)** *Mean curves obtained with estimator $\tilde{\mu}_n$.*

Figure 1.3: *Average curves of counting process data (dashed lines) and of reconstructed cumulative hazard functions (solid lines) for age class A_{60}.*

an increasing function, so a larger score, which represents a steeper cumulative hazard process, implies a lower probability of regular conclusion; component 2, instead, is decreasing, and its estimated coefficient has opposite sign, indicating that patients who have lower cumulative hazard for long times have a higher probability of normal conclusion of the telemonitoring period. Also, we can notice a slight dependence on etiology; in particular, valvular etiology seems to increase the probability of early conclusion of telemonitoring caused by an adverse event. Instead, there is no significant difference in the probability of adverse events among either men and women, nor among subjects with different types of diagnoses.

Parameter	Estimate	Std. Error	p-value
γ_0 (Intercept)	14.8108	437.0832	0.9730
γ_1 (Sex)	0.1557	0.2167	0.4726
γ_2 (Etiology - Hypertensive)	0.0669	0.4469	0.8810
γ_3 (Etiology - Hyschaemic)	-0.0187	0.2506	0.9405
γ_4 (Etiology - Primary)	-0.0819	0.3199	0.7981
γ_5 (Etiology - Valve)	-0.8867	0.4673	0.0790
γ_6 (Etiology - Other)	-0.6599	0.3593	0.1248
γ_7 (Diagnosis - Congestive)	-13.8204	437.0832	0.9748
γ_8 (Diagnosis - Left)	-13.1587	437.0832	0.9760
γ_9 (Diagnosis - Unspecified)	-13.6343	437.0833	0.9751
δ_1 (FPCA score 1)	-0.0144	0.0039	0.0003
δ_2 (FPCA score 2)	0.0567	0.020490	0.0056

Table 1.3: *Estimates, standard errors and p-values for the parameters of the logistic regression with FPCA scores.*

We shall now compare the approach proposed in this work with two other methods: a frailty model for hazard functions, in which estimated frailties are used as predictors
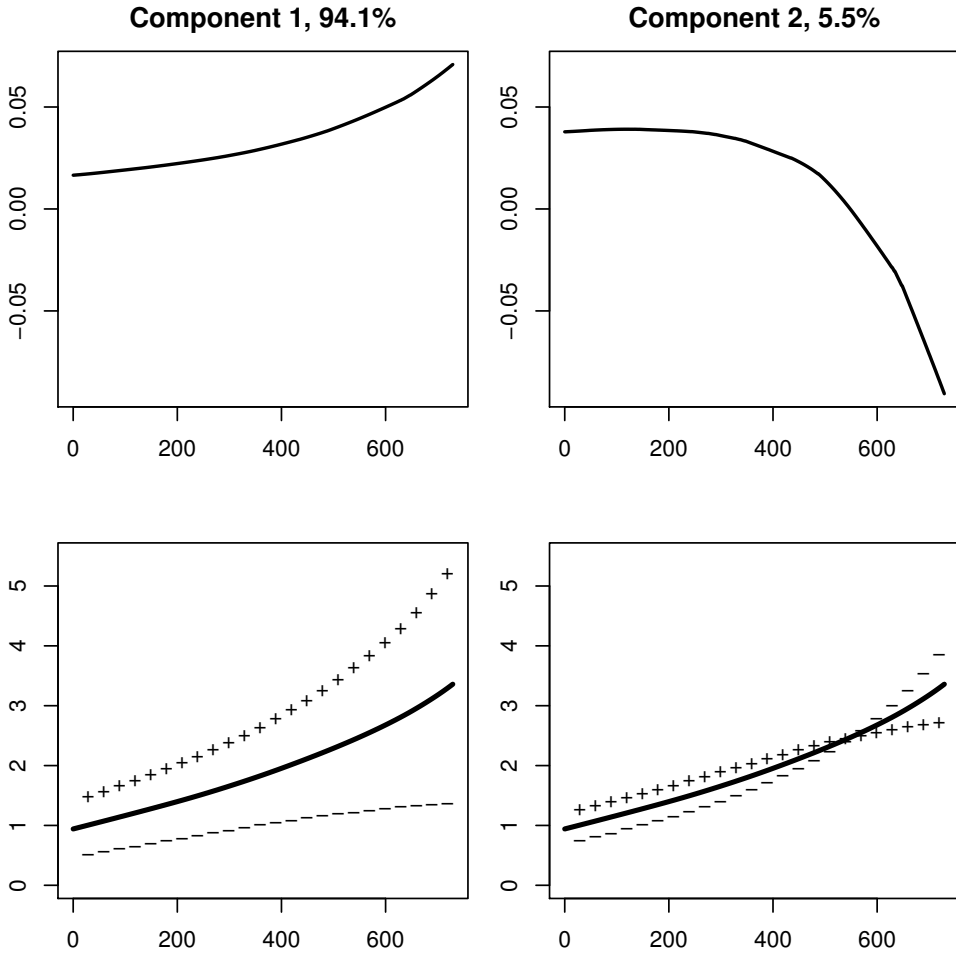
Figure 1.4: *In the upper panels, first $K = 2$ eigenfunctions obtained with FPCA; in the lower panels, representation of $\mu_n(t)$ (solid line) and $\mu_n(t) \pm \nu_k \phi_k(t)$ ('+' or '−' respectively), $k = 1, 2$.*

for the logistic regression, and a functional k-means classification, using instead cluster labels as added covariates.

The shared gamma frailty model is formulated as

$$\lambda(t|\boldsymbol{X}_i, Z_i) = Z_i \lambda_0(t) \alpha^{N_i(t^-)} e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(t)},$$

where $Z_1, \ldots, Z_n$ are i.i.d. from a Gamma$(1/\theta, 1/\theta)$ distribution; notice that θ represents the frailty variance, so that in the limit $\theta \to 0$ the variable is constant and equal to 1. Since we are interested in individual hazard processes, a full likelihood approach, such as the one proposed in [45], is not viable, because random effects are integrated out and only population parameters are estimated. Anyway, this approach is useful to easily assess the significance of random effects, while a computationally heavier EM algorithm can be used to estimate also individual z_is. We used the R package `frailtypack` to check the significance of the parameters in the frailty model, while the estimates of the z_is have been obtained with package `gcmrec`, using the EM algorithm presented in [41]. When the multiplicative frailty component is considered, $\theta \simeq 0.68$ with a significant difference from 0 (a Wald test gives a p-value near 0),

while age and process state $N_i(t)$ in this case are not significant (p-values $> 10\%$). On this basis, we fit the model without covariates using the EM algorithm implemented in package `gcmrec`, and use the estimated frailty realizations $z_1, \ldots, z_n$ as predictors in a logistic regression model; the output of this model is presented in Table 1.4.

Parameter	Estimate	Std. Error	p-value
γ_0 (Intercept)	15.2024	436.2230	0.9722
γ_1 (Sex)	0.1513	0.2160	0.4833
γ_2 (Etiology - Hypertensive)	0.15180	0.4460	0.7336
γ_3 (Etiology - Hyschaemic)	0.1492	0.2503	0.5511
γ_4 (Etiology - Primary)	-0.0452	0.3084	0.8834
γ_5 (Etiology - Valve)	-0.9230	0.4657	0.0474
γ_6 (Etiology - Other)	-0.4346	0.3624	0.2305
γ_7 (Diagnosis - Congestive)	-13.7154	436.2230	0.9749
γ_8 (Diagnosis - Left)	-13.0917	436.2230	0.9761
γ_9 (Diagnosis - Unspecified)	-13.688	436.2230	0.9750
Z (Frailty)	-0.6865	0.1769	0.0001

Table 1.4: *Estimates, standard errors and p-values for the parameters of the logistic regression with estimated frailties.*

The k-means clustering for functional data is presented in [50]. Analogously to k-means clustering for vector data, the method iteratively assigns each one of the observed functions $f_1, \ldots, f_n \in \mathcal{H}$ to a group $k \in 1, \ldots, \mathcal{K}$, choosing the one having the nearest *centroid* with respect to a proper distance d on $\mathcal{H}$, then recomputes centroids basing on updated group assignments. The main types of centroids used in practice are the pointwise mean and the medoid, i.e. the member of the dataset which minimizes the sum of squared distances in the group; the latter choice is usually more convenient when the averaging would produce a function that falls outside the class which the data come from. For our analysis we chose to detect groups basing on the Sobolev H^1 distance

$$d(f_i, f_j)^2 = \int_T (f_i(t) - f_j(t))^2 dt + \int_T (f_i'(t) - f_j'(t))^2 dt,$$

since the rate of hazard increase is also important in characterizing a subject. Moreover, all the curves considered are 0 at the time origin, monotone increasing and have the same support, so it is possible to use the pointwise mean as a centroid, since it will preserve these features. A scree plot representing within-groups variability normalized with total variability versus the number of clusters (Figure 1.5) shows that a quantized description of hazard does not seem to be proper for this dataset, since the described variability decreases very gradually when increasing the number of clusters.

As a preliminary analysis, we can explore the correlation between outcome labels and k-means labels, for $K = 2$ (Table 1.5). The table does not show very strong grouping, nonetheless group 2 includes a clearly higher percentage of positive outcomes w.r.t. group 1, as confirmed by an exact Fisher test (p-value 0.008538).

To compare directly with the preceding models, here we present the logistic regression output only for the case $K = 2$ (Table 1.6), introducing the group labels resulting from the k-means classification as a binary variable to enrich the telemonitoring dataset, while performance indices are presented in Table for $K = 2, 3, 4, 5$. In all cases, only
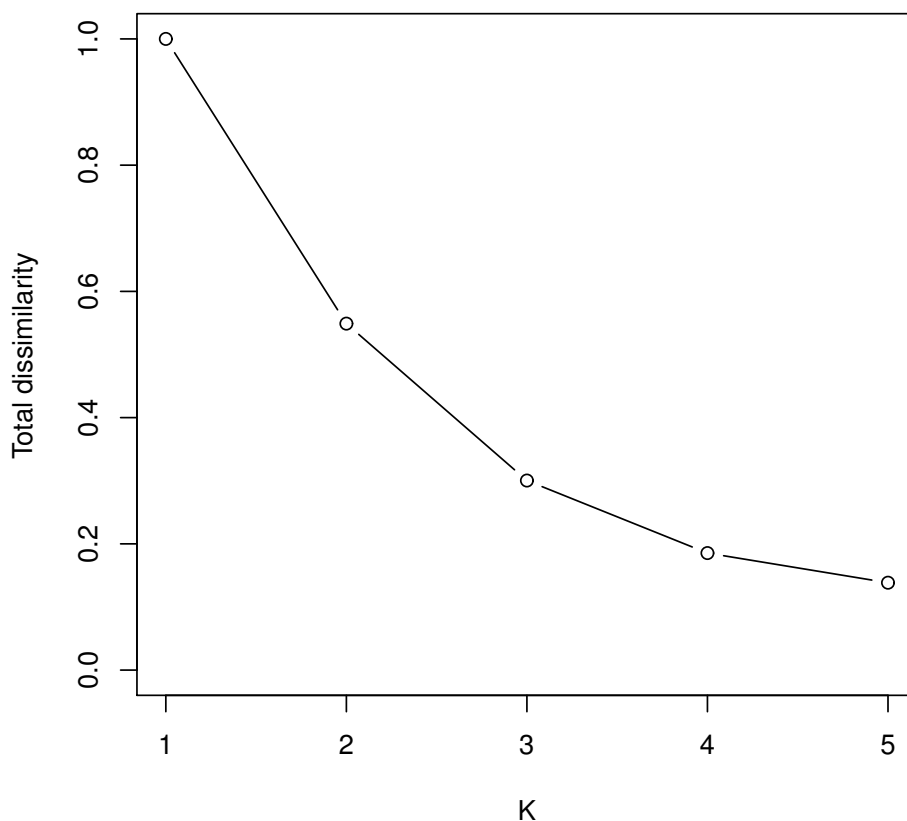
Figure 1.5: *Scree plot of within-groups variability. On the vertical axis, within-groups variability normalized with total variability. On the horizontal axis, corresponding number of clusters.*

one of the $K - 1$ dummy variables introduced in the logistic regression model to represent group labels is significant, with the p-value increasing as K grows.

Table 1.7 shows a comparison between the logistic regression fits with FPC scores, estimated frailty realizations and k-means clustering labels. Notice that the model with FPC scores has lower AIC with respect to the other models, while the lower Brier score denotes a better fit to data. Moreover, a lower leave-one-out cross-validation error, computed using a mean square error cost function, denotes a higher predictive power for the first model.

The fact that by using FPC scores related to dominant principal components we potentially introduce many covariates for the logistic regression, while the frailty approach produces only one predictor, should be considered as an additional feature of the FPC approach. Estimated frailty realizations represent just an amplification factor for population hazard, while the functional principal component analysis is capable of catching arbitrarily complex behaviors of intensity in time. In this case, the second score describes a late stabilization of the patient's health, a phenomenon that is not captured by the frailty variable. It is also worth noticing that the estimation process for

	Group 1	Group 2
$y = 0$	64	84
$y = 1$	171	349

Table 1.5: *Contingency table of telemonitoring outcome ($y = 1$ regular, $y = 0$ early interruption) versus k-means group labels.*

Parameter	Estimate	Std. Error	p-value
γ_0 (Intercept)	15.55561	438.40026	0.9717
γ_1 (Sex)	0.10531	0.21415	0.6229
γ_2 (Etiology - Hypertensive)	0.19180	0.44757	0.6683
γ_3 (Etiology - Hyschaemic)	-0.08477	0.24442	0.7287
γ_4 (Etiology - Primary)	-0.20171	0.30471	0.5080
γ_5 (Etiology - Valve)	-0.81283	0.46640	0.0814
γ_6 (Etiology - Other)	-0.68874	0.35602	0.0530
γ_7 (Diagnosis - Congestive)	-13.93055	438.40016	0.9747
γ_8 (Diagnosis - Left)	-13.21030	438.40017	0.9760
γ_9 (Diagnosis - Unspecified)	-13.69457	438.40026	0.9751
L (Group label)	-0.45040	0.19599	0.0216

Table 1.6: *Estimates, standard errors and p-values for the parameters of the logistic regression with k-means clustering labels.*

the shared gamma frailty model is computationally much heavier with respect to the estimation process for the first model, despite the FPCA step. The k-means clustering with pointwise means as centroids is computationally cheap, but increases in complexity when the number of clusters grows; moreover, the reduction of the patient's risk status to a discrete variable with few level implies a great loss of information, which is reflected by the poorer performance of the method. It should be noticed that among models with k-means labels as covariates the best performance is achieved at $K = 4$, which evidently has the best tradeoff between number of groups and groups separation.

1.4 Concluding remarks

In this work a novel approach to the analysis of telemonitoring data has been proposed, aimed at getting deeper insight on the patient's health status using data from clinical registries and PHD. The presented methodology, involving database integration, counting process modeling of critical events and generalized linear models, can be applied to the study of many different pathologies, thanks to its capability of dealing with complex data.

The counting process model is a natural way of representing the occurrence of hospitalizations in time, and enables us to include a large piece of information contained in the PHD to describe the clinical history of a patient. The model used is very general and allows to describe complex dynamics in an easily interpretable form.

Although it can seem contradictory to define functional data as "synthetic", in the considered application the effect of complex and heterogeneous variables is more easily and synthetically described via a single process, expressing each patient's instantaneous risk of hospitalization. The obtained trajectories are thus studied as functional

Model	AIC	Brier score	CV error
FPCA	688.6547	0.1614153	0.1675431
Frailty	691.8656	0.1630643	0.1688114
k-means, $K = 2$	701.8175	0.1657858	0.1714261
k-means, $K = 3$	701.4385	0.164899	0.1713723
k-means, $K = 4$	697.9778	0.1634312	0.1695719
k-means, $K = 5$	699.6043	0.1634444	0.1701560

Table 1.7: *Comparison between the logistic regression model with FPCA scores, estimated frailties ($\theta = 0.68$) and k-means labels: AIC, Brier score and cross-validation error (MSE).*

covariates in the framework of generalized linear models with functional covariates, which offer a powerful tool to analyze dependencies and to perform classification and prediction in a wide range of applications, also in such complex practical contexts as the considered one. Using the FPCA it is possible to perform dimensional reduction of functional data, allowing to use well established methods for GLM estimation and inference in a multivariate setting and borrowing strength from both techniques.

The application of the proposed methodology is a novelty in the study of home telemonitoring, representing the first example of use of information from the PHD to reconstruct the patient's clinical history in a synthetic way. The obtained results could have an impact on the planning of this care strategy and provide support to the decision of allocating a patient to telemonitoring, basing on her/his probability of concluding it regularly. Further development of this framework in cooperation with medical staff could include the selection and use of various different time dependent and independent variables to study telemonitoring effectiveness on quality of life, mortality and costs, and could lead to the definition of a useful tool for health care assessment and treatment planning.

CHAPTER 2

Diffusion weighted imaging: ADC estimation and design optimization in the Rician framework

2.1 Introduction

Diffusion magnetic resonance (MR) is as an important tool in clinical research, as it allows to characterize some key properties of biological tissues. The signal acquired by the MR scanner is a proxy observation of the diffusion strength in the considered area, and allows to draw important conclusions about the morphology and histology of the examined tissue. For example, in brain imaging the the local directions of maximum diffusion allow to detect fiber orientation. When tumor areas are analyzed using this technique, it can be observed that the diffusion tensor, estimated from the MR magnitude signal, has reduced values in lesions with respect to surrounding physiological tissues, allowing to identify pathological areas or necrosis.

When the region of interest can be considered as isotropic, the *Apparent Diffusion Coefficient* (ADC) is sufficient to characterize the diffusion properties of the tissue, and it is usually estimated as the exponential decay rate of the signal with respect to the *b-value*, one of the main MR acquisition parameters. The assumption of isotropy is common and reasonable in various cases, like breast and prostate cancer (see for example [53] and [46]). The generalization to anisotropic diffusion will be treated in Chapter 3, presenting a more general framework both for the diffusion model and for the random signal distribution.

In many practical situations it may not be possible to collect more than few measures at different b-values, thus limiting the accuracy of the estimation. A reduction in the total number of measures necessary to achieve a certain accuracy is convenient in term of costs, and allows to keep the patient involved in the MR procedure for a shorter amount of time, as the experience may be unpleasant for patients suffering for the

constrained posture or the closed environment. In this chapter, different frequentist and Bayesian approaches for the estimation of the ADC are compared, underlining their statistical properties and computational issues, and an effective experimental design for data acquisition is proposed, with the aim of maximizing accuracy within a limited amount of acquisitions.

2.2 Rice distribution: definition and general properties

The magnitude/intensity signal available from diffusion MR images is the result of a Fourier-transformed frequency signal. The Rice distribution, which is our focus for this chapter, arises from taking the modulus of the transformed complex signal, so first of all it is important to understand some details of this probability distribution.

A *Rice* or (or *Rician*) distributed random variable, denoted $M \sim \text{Rice}(\nu, \sigma)$, has probability density function

$$f_M(m) = \frac{m}{\sigma^2} e^{-\frac{m^2 + \nu^2}{2\sigma^2}} I_0 \left(\frac{m\nu}{\sigma^2} \right) \mathbb{I}_{(0, +\infty)}(m), \tag{2.1}$$

where I_0 is the zero$^{\text{th}}$-order modified Bessel function of the first kind. This special function admits the following power series representation (see [2]):

$$I_\alpha(x) = \sum_{m=0}^{\infty} \frac{1}{m! \Gamma(m + \alpha + 1)} \left(\frac{x}{2} \right)^{2m + \alpha}.$$

Expected value and variance of the Rice distribution are

$$\mathbb{E}[M] = \sigma \sqrt{\frac{\pi}{2}} L_{1/2} \left(-\frac{\nu^2}{2\sigma^2} \right),$$

$$\mathbb{V}\text{ar}(M) = 2\sigma^2 + \nu^2 - \frac{\pi \sigma^2}{2} L_{1/2}^2 \left(-\frac{\nu^2}{2\sigma^2} \right),$$

where

$$L_{1/2}(x) = e^{\frac{x}{2}} \left[(1 - x) I_0 \left(\frac{-x}{2} \right) - x I_1 \left(\frac{-x}{2} \right) \right],$$

a Laguerre polinomial. Notice that these two central moments, as all the higher order ones, do not have a simple relation with the distribution parameters, and no location-scale formulation of the density seems to be obtainable. Nonetheless, as will be clear in the following, ν can be easily interpreted as a location parameter, while σ^2 as a dispersion parameter, although their interplay in determining location and dispersion of the distribution is not trivial. Remarkably, the second raw moment of the distribution is $\mathbb{E}[M^2] = \nu^2 + 2\sigma^2$. The Rice distribution is unimodal and has support on the whole positive real axis; as a result, this random variable is right-skewed and its asymmetry depends on the Signal-to-Noise Ratio (SNR) ν/σ. In particular, high SNR values make the law of this random variable similar to a Gaussian r.v. with mean ν and variance σ^2, while the special case $\nu = 0$ gives a Rayleigh random variable. Figures 2.1a and 2.1b display different density functions at varying ν and SNR values.

A Rician variable M of parameters ν and σ^2 can be obtained as the Euclidean distance from the origin of a point generated by a spherical bivariate Gaussian random,
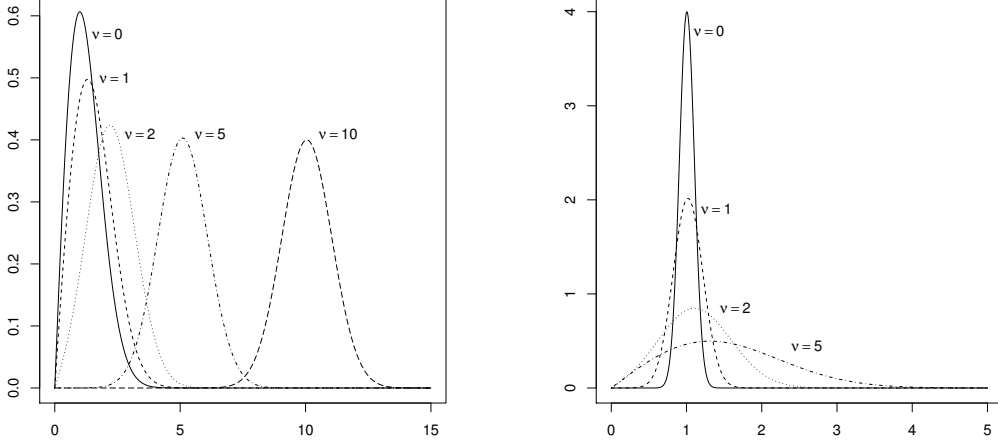
(a) *Varying ν, $\sigma = 1$.*

(b) *Varying SNR (σ), $\nu = 1$.*

Figure 2.1: *Rician probability density functions for different ν and SNR values.*

variable, i.e. $(\mathbf{Z}) = (X, Y) \sim N_2(\boldsymbol{\mu}, \sigma^2 I)$, with $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and being I the 2×2 identity matrix. In the case of diffusion MRI data, (X, Y) represents the vector having as component X the real part and as component Y the imaginary part of the Fourier-transformed frequency signal. The noise variables that corrupt both the components of this signal can be reasonably considered as Gaussian and uncorrelated, thus providing a covariance matrix multiple of the identity. In the following we shall see then how the Rice distribution arises from this bivariate random variable. Operating a change of variables to polar coordinates $(X, Y) \to (M, \Psi)$ and reparametrizing $(\mu_1, \mu_2) = (\nu cos(\varphi), \nu sin(\varphi))$, the Gaussian density reads

$$f(m, \psi | \nu, \varphi, \sigma^2) = \frac{m}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\{m^2 + \nu^2 - 2m\nu\cos(\psi - \varphi)\}} \mathbb{I}(m)_{(0, +\infty)} \mathbb{I}(\psi)_{[0, 2\pi)}.$$

Marginalizing with respect to ψ and using the integral representation

$$I_0(x) = \int_0^\pi e^{x\cos\theta} d\theta$$

we obtain exactly the p.d.f. (2.1).

Notice that the distribution of $M = \sqrt{X^2 + Y^2}$ is independent of φ. Moreover, by the characterization as modulus of a bivariate spherical normal distribution it is easy to verify the following property: $M \sim \text{Rice}(\nu, \sigma^2)$ implies $M/\eta \sim \text{Rice}(\nu/\eta, \sigma^2/\eta^2)$.

A Rician r.v. can be obtained also as $M = \sigma\sqrt{X}$, where X is a mixture of χ^2_{2P+2} distributions with $P \sim Poisson(\nu^2/2\sigma^2)$. The relevance of this fact will be evident in Section 2.5.

2.3   Maximum likelihood estimation for the Rician parameters

The log-likelihood function for a sample of magnitudes M_i, $i = 1, \ldots, n$, reads

$$l(\nu, \sigma^2 | \mathbf{m}) = -n \frac{\nu^2}{2\sigma^2} - \frac{\sum_{i=1}^n m_i^2}{2\sigma^2} + \sum_{i=1}^n \log \left(\frac{m_i}{\sigma^2} I_0 \left(\frac{m_i \nu}{\sigma^2} \right) \right).$$

The ML estimators for ν and σ^2 (the angle parameter φ disappears) obtained by the maximum likelihood criterion are the ones presented below; their computation will be presented in a more general framework in Section 2.4.

$$\hat{\nu}_M = \frac{1}{n} \sum_{i=1}^n m_i \frac{I_1 \left(\frac{m_i \hat{\nu}}{\hat{\sigma}^2} \right)}{I_0 \left(\frac{m_i \hat{\nu}}{\hat{\sigma}^2} \right)}, \tag{2.2}$$

$$\hat{\sigma}_M^2 = \frac{\sum_{i=1}^n m_i^2}{2n} - \frac{\hat{\nu}^2}{2}. \tag{2.3}$$

Notice that the first equation is implicit in $\hat{\nu}$ and that both equations involve both the parameters, so the MLE must be solved by a non linear optimization algorithm. The marginalization removes explicit information on observation angles, thus causing an important loss of information. In particular, it is not possible to obtain an explicit form for the maximum likelihood estimator, since the Rician distribution is not an exponential family (see for example [36]). This comes as a consequence of the Fisher-Darmois-Koopman-Pitman theorem [8], which states that, in parametric families of distributions with parameter-free support and suitable regularity assumptions, the fact of being an exponential family is equivalent to the existence of a vector of sufficient statistics with dimension not increasing with sample size. The MLE is always a function of the sufficient statistics, and in our case the Rician class of distributions is not an exponential family, as proved in Appendix B.3. As a consequence, it is not possible to give an explicit form for the MLE in the Rician (and in the more general noncentral chi) case.

2.4   Isotropic Stejskal-Tanner model and Rice exponential regression

Diffusion MR works thanks to the relation between the physiological and magnetic properties of biological tissues. The description of how the MR scan works and the modeling of the relation between magnetic pulse sequence and intensity response is behind the scope of this work (see [33] for some reviews of various topics in diffusion MR); in this chapter we concentrate on an isotropic framework, suitable for example for the detection and analysis of breast and prostate cancer, and consider a special case of the popular Stejskal-Tanner model (see for example [51], [9])

$$\nu_i = \nu_0 \exp(-\alpha b_i), \qquad i = 1, \ldots, n, \tag{2.4}$$

where α is the *Apparent Diffusion Coefficient* (ADC), the isotropic diffusion parameter (its common unit of measure in biological tissues is mm^2/s), b_i is the *b-value* at acquisition i, a combination of parameters of the MR scan, ν_i is the intensity signal and ν_0 is simply the signal at $b = 0$. In this section we first introduce the likelihood of Rice-distributed data with location parameter ν following a general exponential model,

then we concentrate on estimation in the special case of the isotropic model (2.4), presenting a simulation study of the proposed estimators. In Chapter 3 the full anisotropic Stejskal-Tanner model is considered, anyway even this case falls in the more general class of exponential models presented in the following.

We name *Rice exponential regression* an estimation problem where $\{(M_i, \mathbf{x}_i)\}_{i=1,\ldots,n}$ are observed and $M_i|\mathbf{x}_i \sim Rice(\nu_i, \sigma)$, with

$$\nu_i = \exp(\boldsymbol{\theta}^T \mathbf{x}_i), \qquad i = 1, \ldots, n \tag{2.5}$$

and where $\boldsymbol{\theta}$ is a q-dimensional vector of unknown parameters. The likelihood for such a random sample of independent observations is

$$L(\boldsymbol{\theta}, \sigma|\mathbf{m}, \mathbf{x}) = \prod_{i=1}^{n} \frac{m_i}{\sigma^2} \exp\left(-\frac{m_i^2 + e^{2\boldsymbol{\theta}^T \mathbf{x}_i}}{2\sigma^2}\right) I_0\left(\frac{m_i e^{\boldsymbol{\theta}^T \mathbf{x}_i}}{\sigma^2}\right).$$

The log-likelihood is then

$$\log L(\boldsymbol{\theta}, \sigma|\mathbf{m}, X) = \sum_{i=1}^{n} \log(m_i) - n \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} m_i^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} e^{2\boldsymbol{\theta}^T \mathbf{x}_i}$$
$$+ \sum_{i=1}^{n} \log\left[I_0\left(\frac{m_i e^{\boldsymbol{\theta}^T \mathbf{x}_i}}{\sigma^2}\right)\right].$$

From now on we consider the parameter σ as known, so the whole vector of parameters for the Rice exponential regression likelihood is just $\boldsymbol{\theta}$.

In the case of known σ^2, to maximize the log-likelihood we have to solve the score equations

$$\frac{\partial}{\partial \theta_j} L(\boldsymbol{\theta}|\sigma, \mathbf{m}, X) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} e^{2\boldsymbol{\theta}^T \mathbf{x}_i} 2x_{ij} + \sum_{i=1}^{n} \frac{I_1\left(\frac{m_i e^{\boldsymbol{\theta}^T \mathbf{x}_i}}{\sigma^2}\right)}{I_0\left(\frac{m_i e^{\boldsymbol{\theta}^T \mathbf{x}_i}}{\sigma^2}\right)} \frac{m_i e^{\boldsymbol{\theta}^T \mathbf{x}_i}}{\sigma^2} x_{ij}$$

for $j = 1, \ldots, q$; this expression is obtained using the fact that $I_0'(x) = I_1(x)$.

2.5 Bayesian modeling

A Bayesian approach to Rician exponential regression may also be adopted. In this case the dispersion parameter σ^2 could also be considered as unknown; notice that if the noise parameter is uniform over an MR image, the estimate of σ^2 can borrow information from all the pixels, so introducing this unknown parameter should not raise identifiability issues.

To specify the Bayesian formulation, first of all we have to remind that, if M is a Rician variable, M^2/σ^2 can be sampled by a mixture of $\chi^2(2P+2)$ distributions, with $P \sim \mathcal{P}(\nu^2/2\sigma^2)$. Using this fact as it is would require to know the value of σ^2, but we can express the distribution of $R = M^2$ in a different way to overcome this problem. First of all, R has density

$$f_R(r|\nu, \sigma^2) = \frac{e^{-\frac{r+\nu^2}{2\sigma^2}}}{2\sigma^2} I_0\left(\frac{\sqrt{r}\nu}{\sigma^2}\right) \mathbb{I}(r)_{(0,+\infty)} \tag{2.6}$$

(a noncentral chi-squared multiplied by σ^2). We adopt the following reparametrization: $\eta = \nu^2/2\sigma^2$ and $\tau = 1/2\sigma^2$. Expanding the modified Bessel function I_0, we obtain:

$$f_R(r|\eta, \tau) = \tau e^{-r\tau} e^{-\eta} \sum_{k=0}^{\infty} \frac{1}{k!k!}(r\eta\tau)^k \mathbb{I}(r)_{(0,+\infty)}$$

$$= \sum_{k=0}^{\infty} \frac{\tau^{k+1} r^k e^{-r\tau}}{k!} \mathbb{I}(r)_{(0,+\infty)} \cdot \frac{e^{-\eta}\eta^k}{k!},$$

which is a mixture of $\mathrm{Gamma}(K + 1, \tau)$ distributions with $K \sim \mathcal{P}(\eta)$.

Plugging in the Stejskal-Tanner model, for observations $i = 1, \ldots, n$ on pixels $j = 1, \ldots, n_p$, with $\nu_{ij} = \nu_{0j} e^{-\alpha_j b_i}$ we have the following likelihood

$$L(\boldsymbol{\nu}_0, \boldsymbol{\alpha}, \sigma^2|\mathbf{m}, \mathbf{b}) = \prod_{i=1}^{n} \prod_{j=1}^{n_p} \frac{m_{ij}}{\sigma^2} \exp\left(-\frac{m_{ij}^2 + \nu_{0j}^2 e^{-2\alpha_j b_i}}{2\sigma^2}\right) I_0\left(\frac{m_{ij}\nu_{0j}e^{-\alpha_j b_i}}{\sigma^2}\right),$$

where $\mathbf{m} = \{m_{ij}\}_{i=1,\ldots n; j=1,\ldots,n_p}$ and $b = (b_1, \ldots, b_n)^T$. With the reparametrization, setting $\eta_{ij} = \eta_{0j} e^{-2\alpha_j b_i}$, we have

$$L(\boldsymbol{\eta}_0, \boldsymbol{\alpha}, \tau|\mathbf{r}, \mathbf{b}) = \prod_{i=1}^{n} \prod_{j=1}^{n_p} \sum_{k_{ij}=0}^{\infty} \frac{\tau^{k_{ij}+1} r_{ij}^{k_{ij}} e^{-r_{ij}\tau}}{k_{ij}!} \cdot \frac{e^{-\eta_{0j} e^{-2\alpha_j b_i}} (\eta_{0j} e^{-2\alpha_j b_i})^{k_{ij}}}{k_{ij}!}.$$

This formulation allows to easily implement the Bayesian model for example on BUGS [23], using random number generators for classical distributions. The extension to a general Rice exponential regression model is straightforward.

2.6 Comparison of pointwise estimation methods

In this Section we present different methods for the estimation of α, the unknown parameter of interest. We consider a sample of signal intensities on a single pixel $M_i \sim \mathrm{Rice}(\nu_0 e^{-\alpha b_i}, \sigma^2)$, $i = 1, \ldots, n$, and their respective realizations $\mathbf{m} = m_1, \ldots, m_n$ at b-values $\mathbf{b} = b_1, \ldots, b_n$.

The dispersion parameter σ^2 is usually measured over regions where almost pure noise is observed, and used as a known parameter in the subsequent estimates. This estimate of σ^2 is considered as reliable, since it can be based on a very large number of pixels, so we will consider the case of known dispersion parameter.

We consider nonlinear least squares, maximum likelihood and three bayesian point estimators. In the case of a simple $\mathrm{Rice}(\nu, \sigma^2)$ random variable an iterative method of moments estimator has been proposed in [32], but this technique has no straightforward extension to the case of covariate-dependent ν, while moment equations would be difficult to invert in the considered case. Moreover, under the model assumptions presented in Section 2.2 a decoupling of noise and signal in the fashion of Signal Detection Theory could not be pursued (see [35] for a brief presentation and a Bayesian implementation of the SDT classical scheme).

The different estimation methods for the couple (ν_0, α) presented here will be tested under different Signal-to-Noise Ratios (SNRs) ν/σ in Section 2.6.4 .

2.6.1 Nonlinear regression

A standard approach for the estimation of ν_0 and α is to solve a nonlinear least squares problem, which is equivalent to approximating $M_i \approx \nu_0 exp(-\alpha b_i) + \varepsilon_i$ for $i = 1, \ldots, n$, where ε_i are iid, zero mean, gaussian noise terms. The estimators $\hat{\nu}_0^{LS}$ and $\hat{\alpha}^{LS}$ are defined as

$$(\hat{\nu}_0^{LS}, \hat{\alpha}^{LS}) = \underset{(\nu_0, \alpha)}{\operatorname{argmin}} \sum_{i=1}^{n} (m_i - \nu_0 e^{-\alpha b_i})^2,$$

for $\nu_0, \alpha > 0$, which is equivalent to the solution of the following equations

$$\begin{cases} \nu_0 \sum_{i=1}^{n} e^{-2\alpha b_i} = \sum_{i=1}^{n} m_i e^{-\alpha b_i}, \\ \nu_0 \sum_{i=1}^{n} b_i e^{-2\alpha b_i} = \sum_{i=1}^{n} m_i b_i e^{-\alpha b_i}. \end{cases} \tag{2.7}$$

The approximation to a nonlinear regression model is inconsistent with the phenomenon under study, most evidently for the fact that in this case the noise term is symmetric and it can assume real values. This inconsistency is negligible for high SNR values, since a Rice(ν, σ^2) distribution in this case approaches a $\mathcal{N}(\nu_0, \sigma^2)$, but becomes important with medium and low SNRs. In [51], the behavior of the Rice distribution with fixing $\sigma = 1$ and varying ν is examined, observing that normality can be considered a good approximation at about $\nu/\sigma > 2.64$, but the sample variance approaches σ^2 only for SNR values greater than 5.19. Even for pixels with high SNRs at $b = 0$, for large b-values the real signal could reach the same order of magnitude of noise, depending on the unknown value of α, and this could lead to very biased estimates. However, the least squares approach is computationally simpler and quicker to carry out, since it can be seen from (2.7) that ν_0 can be expressed as a function of α, thus requiring just a one-dimensional optimization to compute the estimates.

2.6.2 Maximum likelihood

The maximum likelihood approach allows to take into account the asymmetry of the signal distribution, always providing admissible values of the parameters. The objective function is the log-likelihood

$$l(\nu_0, \alpha | \mathbf{m}, \mathbf{b}, \sigma^2) = \log L(\nu_0, \alpha | \mathbf{m}, \mathbf{b}, \sigma^2) = \sum_{i=1}^{n} \log f_{M_i}(m_i | \nu_0 e^{-\alpha b_i}, \sigma^2)$$

$$\propto -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \nu_0^2 e^{-2\alpha b_i} + \sum_{i=1}^{n} \log \left[I_0 \left(\frac{m_i \nu_0 e^{-\alpha b_i}}{\sigma^2} \right) \right],$$

where f_{M_i} is the Rician density (2.1), for $i = 1, \ldots, n$. The ML estimator is then

$$(\hat{\nu}_0^{ML}, \hat{\alpha}^{ML}) = \underset{(\nu_0, \alpha)}{\operatorname{argmax}} \, l(\nu_0, \alpha | \mathbf{m}, \mathbf{b}, \sigma^2),$$

for $\nu_0, \alpha > 0$.

Looking for stationary points of l and using the fact that $I_0'(x) = I_1(x)$, we obtain

the following score equations

$$\begin{cases} \nu_0 \sum_{i=1}^n e^{-2\alpha b_i} = \sum_{i=1}^n \frac{I_1(\frac{m_i \nu_0 e^{-\alpha b_i}}{\sigma^2})m_i}{I_0(\frac{m_i \nu_0 e^{-\alpha b_i}}{\sigma^2})}e^{-\alpha b_i}, \\ \nu_0 \sum_{i=1}^n b_i e^{-2\alpha b_i} = \sum_{i=1}^n \frac{I_1(\frac{m_i \nu_0 e^{-\alpha b_i}}{\sigma^2})m_i}{I_0(\frac{m_i \nu_0 e^{-\alpha b_i}}{\sigma^2})}b_i e^{-\alpha b_i}. \end{cases}$$

Notice that these equations differ from (2.7) only for the Bessel functions ratio I_1/I_0, which multiplies observations m_i. In particular, this factor decreases the values of observations, since $0 < I_1(x)/I_0(x) < 1$ for $x > 0$, and increases asymptotically to 1 for large SNRs, so that the score equations tend to (2.7).

As shown in [48], the maximum likelihood estimator for ν obtained from an iid sample $M_1, \ldots, M_n \sim \text{Rice}(\nu, \sigma^2)$ and known σ^2 becomes exactly 0 when the moment estimator for $\mathbb{E}[M^2] = \nu^2 + 2\sigma^2$ becomes inadmissible, i.e. when $\sum_{i=1}^n M_i^2/n - 2\sigma^2 \leq 0$, even if the real value of ν is larger than 0. The case of Rice exponential regression suffers of a similar problem in a non trivial way, and would require σ^2 to be estimated with the other parameters to keep parameter values coherent with the model. This issue is addressed in this Chapter 3, where a method for the joint estimation of all the parameters is proposed.

2.6.3 Bayesian approaches

We consider also three different estimators based on a Bayesian posterior distribution: its mean, median and mode. To allow an easy implementation using BUGS code, we introduce a slightly different formulation of the model.

If $M \sim \text{Rice}(\nu, \sigma^2)$, then $R = M^2/\sigma^2$ has noncentral χ^2 distribution with 2 degrees of freedom and noncentrality parameter $\lambda = \nu^2/(2\sigma^2)$. Be now $R_1, \ldots, R_n$ the random sample considered, with $R_i = M_i^2/\sigma^2$ and $M_i \sim \text{Rice}(\nu_0 e^{-\alpha b_i}, \sigma^2)$ for $i = 1, \ldots, n$, and let $\mathbf{r} = (r_1, \ldots, r_n)$ be the observations from this sample. Let $\pi(\nu_0)$ and $\pi(\alpha)$ be the prior distributions of the two unknown parameters respectively, while the density of each R_i will be denoted as $f_{R_i}(r_i)$, with parameter $\lambda_i = \nu_0^2 e^{-2\alpha b_i}/2\sigma^2$. The joint posterior distribution of ν_0 and α is then

$$p(\nu_0, \alpha | \mathbf{r}, \mathbf{b}, \sigma^2) \propto \prod_{i=1}^n f_{R_i}(r_i | \lambda_i)\pi(\nu_0)\pi(\alpha)$$

As anticipated in Section 2.2, a noncentral χ^2 distribution of noncentrality λ can be sampled as a mixture of $\chi^2(2P + 2)$ with $P \sim \mathcal{P}(\lambda)$. This allows an easy BUGS implementation of these estimators.

2.6.4 Simulation study

We compared 5 estimators for α - least squares (LS), maximum likelihood (ML) and posterior mean (PMe), median (PMd) and mode (PMo) - in terms of mean and mean square error. For the two frequentist approaches, ranges for the possible parameter values have been chosen, considering $\nu_0 \in [0.1, 10]$ and $\alpha \in [0.1, 5]$, while the fixed parameter σ^2 has been taken always equal to 1. For the Bayesian point estimators we chose uninformative, uniform priors, with the same support as the ranges chosen for

LS and ML. The first two estimators have been computed with R 2.12.2 (see [43]), using built-in optimization functions: `optimize` for the one-dimensional minimization required in LS and `optim`, using the L-BFGS-B method, for the likelihood maximization, with startup values $(\nu_{0\text{start}}, \alpha_{\text{start}}) = (1, 1)$. Bayesian posterior distributions have been computed using a Gibbs sampler implemented in JAGS (see [42]). In particular, the following model code (valid for any program supporting BUGS-type language) was used:

```
model{
  for(i in 1:n){
    lambda[i]<-(nu0*nu0)*exp(-2*alpha*b[i])/(2*sigma*sigma)
    p[i]~dpois(lambda[i])
    k[i]<-2*p[i]+2
    M[i]~dchisqr(k[i])
  }
  alpha~dunif(0.1,5)
  nu0~dunif(0.1,10)
}
```

As it can be seen from the model code, uniform prior distributions have been chosen, with supports equal to the search ranges for LS and ML. 10000 Gibbs sampling iterations have been run for each different sample, with a thinning of 10, and standard diagnostics revealed a good behavior of the generated chains.

We chose b-values in a typical range for diffusion MR machine settings, i.e. from 0 to $1000\text{s}/\text{mm}^2$, on equally spaced grids of $n = 5, 10, 15, 20, 25, 30$ points. Different simulations have been run with parameter values $\nu_0 = 2, 4, 8$, which represent a low, an intermediate and a high SNR, and $\alpha = 0.7, 1, 3$, typical low, intermediate and high physiological values of ADC.

It must be reported that the ML estimator, in cases of low SNR, reached the boundaries of the optimization region in various simulations. In the combination $n = 5$, $\nu_0 = 2$, $\alpha = 3$ only 45% of the simulations gave ML estimates that converged to a value inside the predefined ranges of parameters search, while in the other cases this number oscillated around 70% when $\alpha = 1$ or 100% when $\alpha = 0.7$. These degenerate results have been removed for the computation of bias and variance.

Figure 2.2 displays the decaying exponential curves we aim to estimate in the 9 different combinations of ν_0 and α, along with a horizontal line at level σ, to represent the order of magnitude of noise with respect to the signal. The quality of estimates depends both on the SNR at $b = 0$ and on the ADC, as will be clear from simulations.

Figure 2.3 shows the behavior of bias for the estimators of α with different sample sizes n. For what concerns the frequentist estimators (LS and ML), there is no uniform ordering through the considered values of n when the signal decays slowly ($\alpha = 0.7$), but in the other cases, when noise is stronger along the curve, the maximum likelihood estimate is always less biased than the least squares one; notice also that the least squares estimates do not seem to have a decreasing bias when n increases among the considered values. Concerning the 3 Bayesian estimators, no striking differences arise among them, while with respect to the frequentist estimators in many cases they have comparable or higher bias, with the exception of the "worst case" $\nu_0 = 2$, $\alpha = 3$, where they are uniformly more accurate.

From what concerns variance, analyzed in Figure 2.4, the LS estimator shows almost always the best performance, excepted for low sample sizes when $\alpha = 3$. The other estimators have similar performances and behaviors at different sample sizes n, with ML
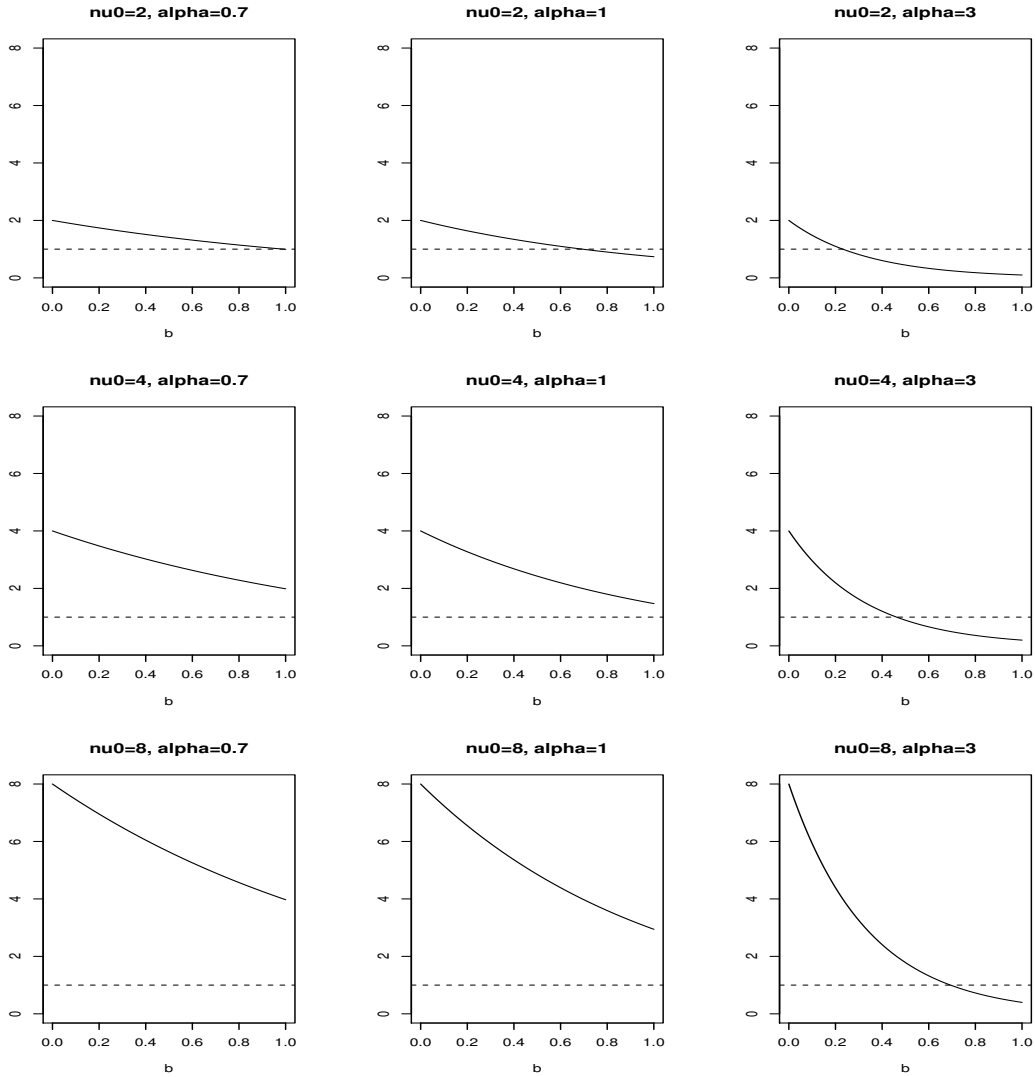
Figure 2.2: *Stejskal-Tanner model in simulation parameter combinations. b-values are expressed in* $1000s/mm^2$.

and PMe having strikingly higher variance in some noisy cases. As expected, variance notably decreases for all estimators at increasing n in most combinations of parameters, but with very low SNR ($\nu_0 = 2$) the only one showing empirical convergence of variance to 0 is LS.

An overall index of estimator performance can be evaluated by the mean square error (MSE). Since the MSE is the sum of square bias and variance, the orders of magnitude of these two characteristics assume an important role. As it can be seen from Figure 2.5, the LS estimator has the lowest MSE when $\alpha = 0.7, 1$, but exhibits the worst performances in the critical cases of high ADC, where Bayesian estimators seem to work better.

Results for ν_0 are not detailed here, but it is worth mentioning that, since it is necessary to estimate the two parameters jointly, the precisions and accuracies of their estimators are mutually influenced. Anyway, estimators for ν_0 show a more classical
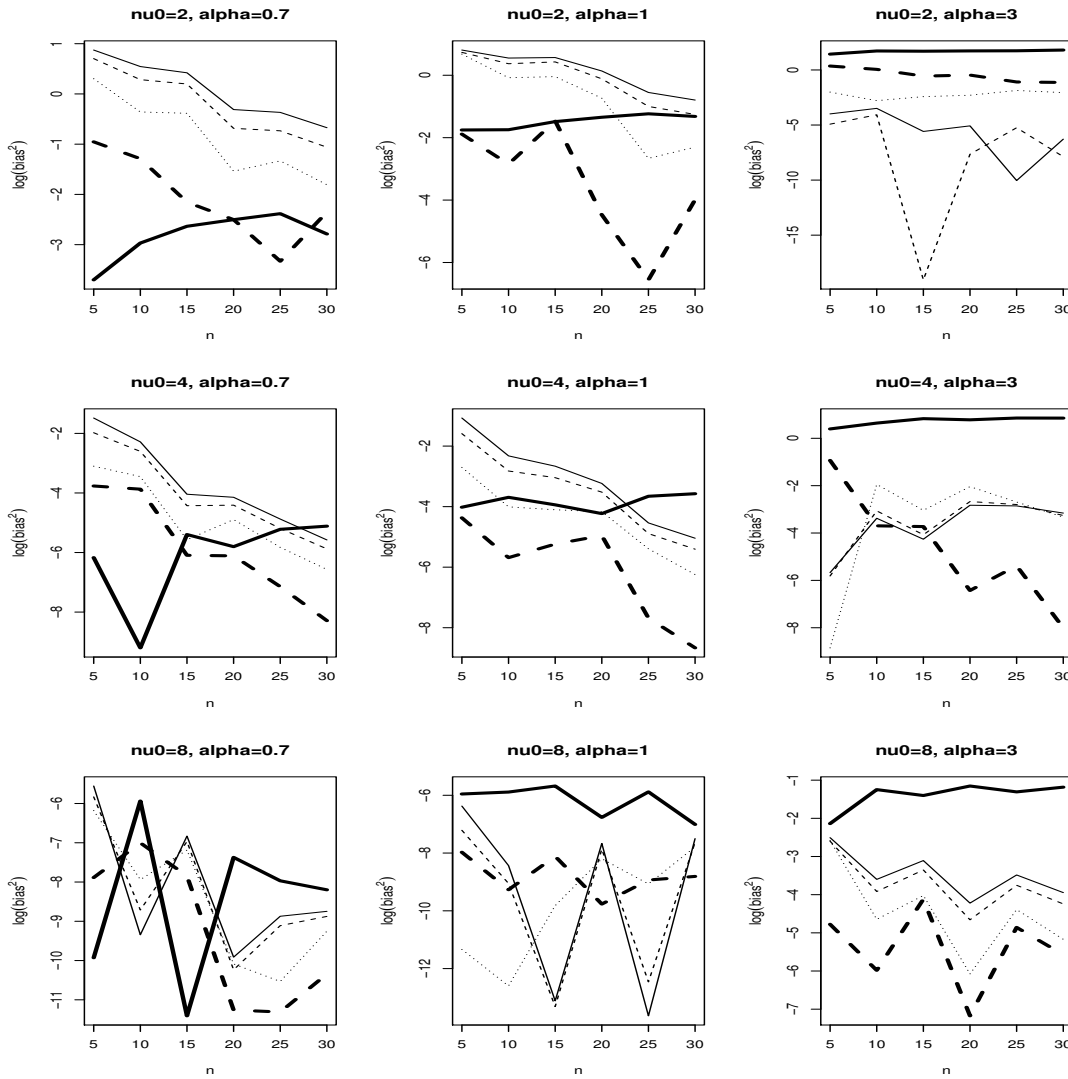
Figure 2.3: *Bias of estimators for α. Bold lines: solid=LS, dashed=ML; slim lines: solid=PMe, dashed=PMd, dotted=PMo.*

behavior: the LS estimator is in all cases less accurate but more precise (high bias and low variance), and the consistency of all estimators is evident when increasing n. The summary plots for the MSE of the estimators for ν_0 can be seen in Figure 2.6.

2.7 Experimental design

Experimental design optimization is another strategy for achieving higher accuracy in estimation. In this section we couple maximum likelihood estimation of the Rician model with a criterion of accuracy that derives naturally from the framework of the likelihood principle: the Fisher Information Matrix (briefly FIM). Since the inverse of the FIM $\mathcal{I}^{-1}(\theta|\sigma^2, \mathbf{b})$ for large samples approaches the covariance matrix of a vector of ML estimators $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ [12], samplings can be designed so that a proper functional of this matrix is optimized.
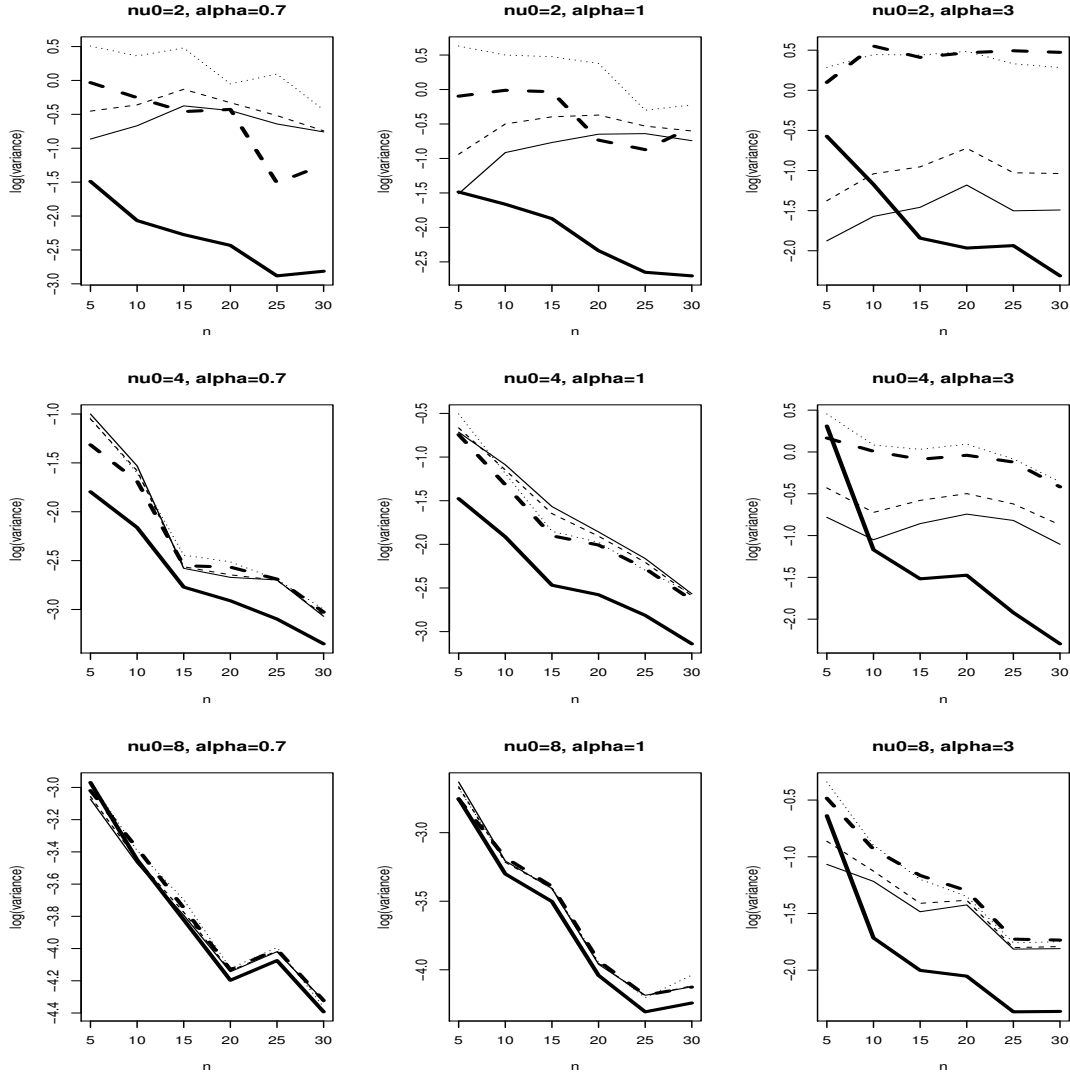
Figure 2.4: *Variance of estimators for α. Bold lines: solid=LS, dashed=ML; slim lines: solid=PMe, dashed=PMd, dotted=PMo.*

The FIM is the matrix of elements

$$[\mathcal{I}(\boldsymbol{\theta})]_{jk} = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_j}\log L(\boldsymbol{\theta}|\sigma,\mathbf{m},\mathbf{x})\right)\left(\frac{\partial}{\partial\theta_k}\log L(\boldsymbol{\theta}|\sigma,\mathbf{m},\mathbf{x})\right)\right], \text{ for } i,j = 1,\dots,q,$$

where L represents the likelihood function. Alternatively, if some regularity conditions are satisfied, this is also equal to

$$[\mathcal{I}(\boldsymbol{\theta})]_{jk} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log L(\boldsymbol{\theta}|\sigma,\mathbf{M},\mathbf{x})\right].$$

Using the likelihood of the Rice exponential regression model we obtain the following
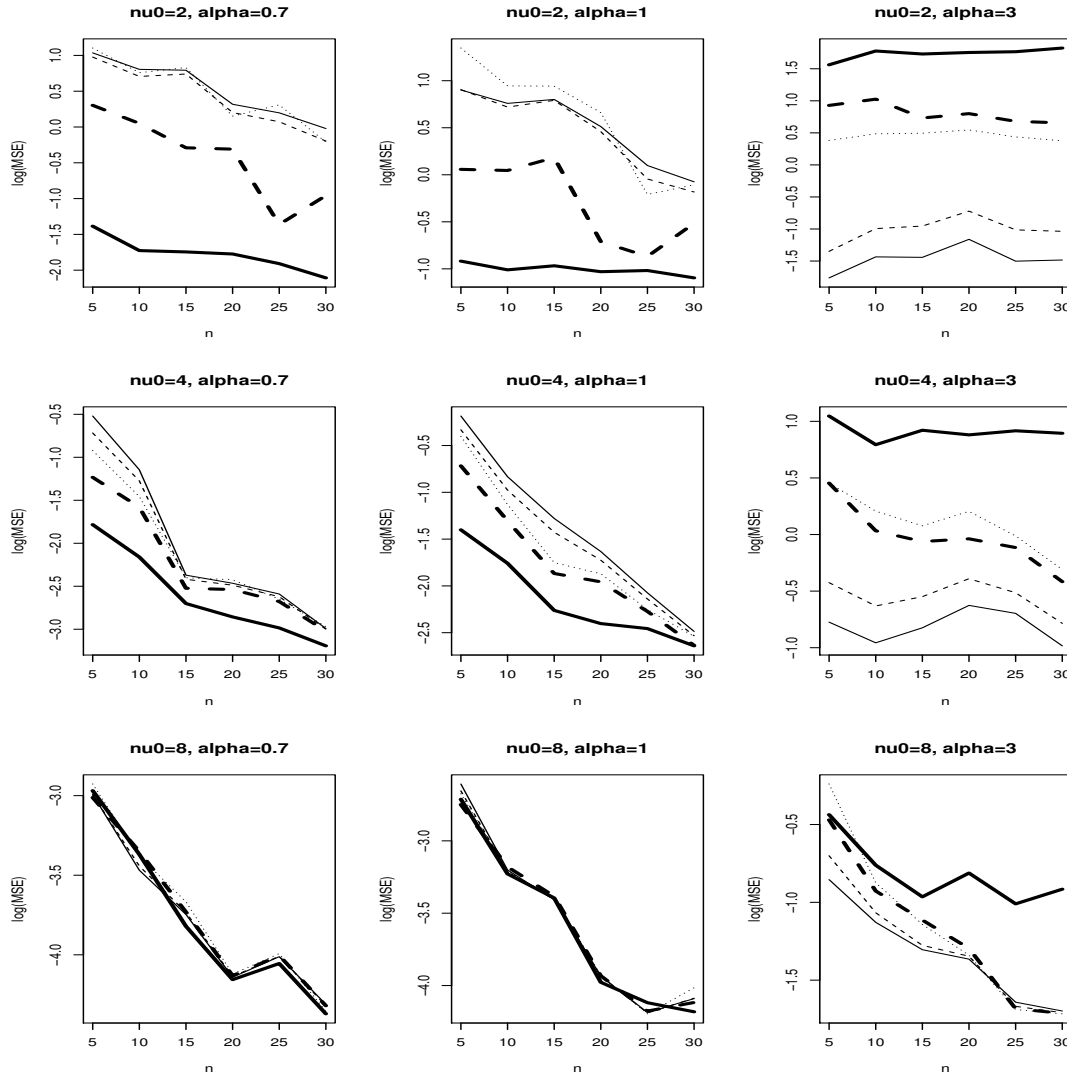
Figure 2.5: *MSE of estimators for α. Bold lines: solid=LS, dashed=ML; slim lines: solid=PMe, dashed=PMd, dotted=PMo.*

expression for the FIM element (j, k) (see Appendix B.1):

$$[\mathcal{I}(\boldsymbol{\theta})]_{jk} = -\frac{1}{\sigma^4} \sum_{i=1}^{n} e^{2\boldsymbol{\theta}^T \mathbf{x}_i} x_{ij} x_{ik} \left\{ e^{2\boldsymbol{\theta}^T \mathbf{x}_i} - \mathbb{E}\left[M_i^2 \frac{I_1^2\left(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}_i M_i}}{\sigma^2}\right)}{I_0^2\left(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}_i M_i}}{\sigma^2}\right)} \right] \right\}.$$

Let us remind the notation $\nu_i = \exp(\boldsymbol{\theta}^T \mathbf{x}_i)$. Equation (2.7) requires to integrate on $\mathbb{R}^+$ the function

$$g(m|\nu_i, \sigma) = m^2 \frac{I_1^2\left(\frac{\nu_i m}{\sigma^2}\right)}{I_0^2\left(\frac{\nu_i m}{\sigma^2}\right)} f_M(m|\nu_i, \sigma). \tag{2.8}$$

An approach for carrying out this task is described in Appendix B.2.

The parameter σ^2 is commonly estimated separately before the estimation of other parameters, considering areas of tissue where the measured signal is believed to be almost pure noise. It is possible also to estimate σ^2 jointly with the other parameters:
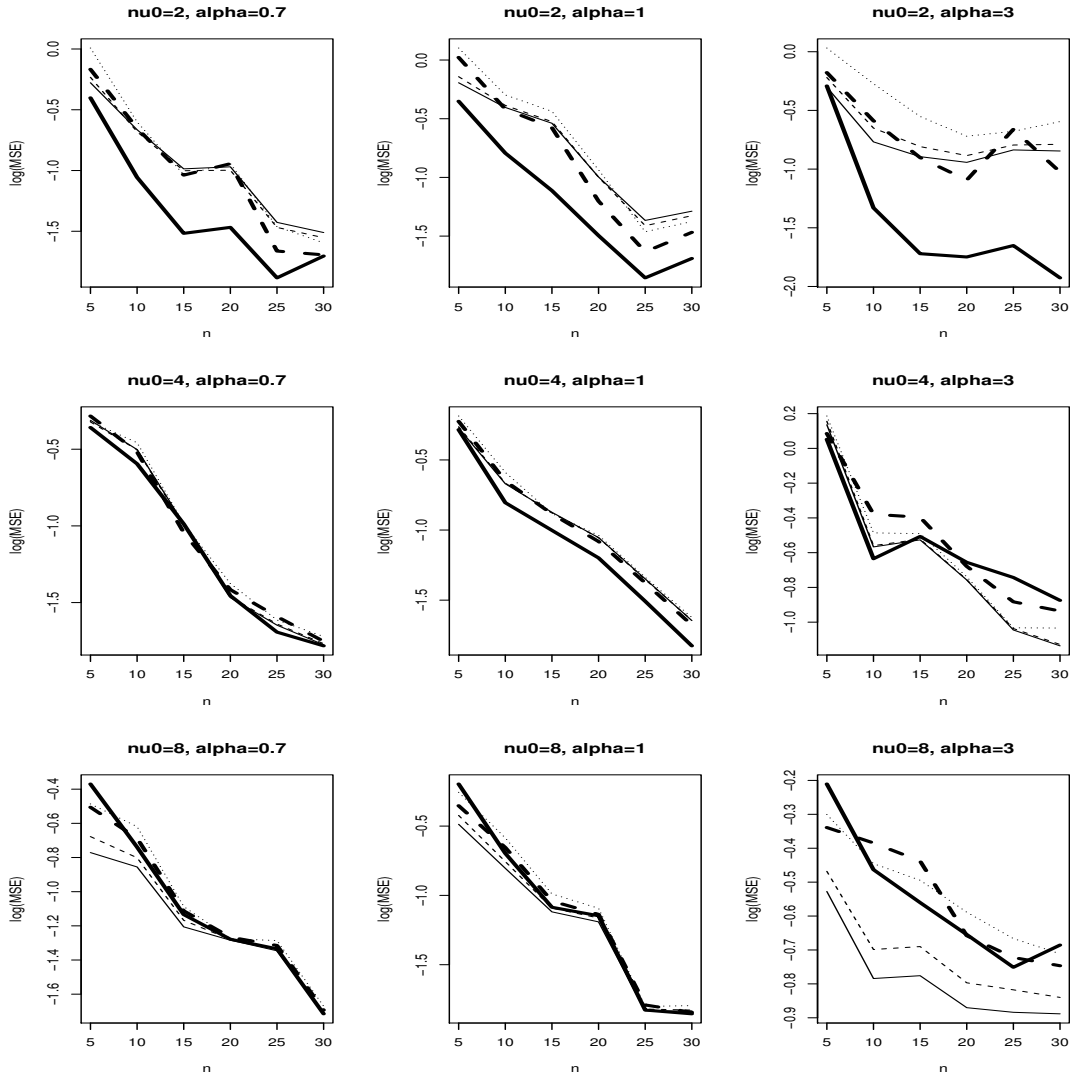
Figure 2.6: *MSE of estimators for ν_0. Bold lines: solid=LS, dashed=ML; slim lines: solid=PMe, dashed=PMd, dotted=PMo.*

considering a joint likelihood for the whole MR image, if σ^2 can be considered approximately uniform on the MR volume then it can also be estimated using an alternated maximization approach: fixing the vectors ν_0 and α of parameter values on different pixels, σ^2 is estimated by ML, then σ^2 is fixed to its updated estimate and $\hat\nu_0$ and $\hat\alpha$ are obtained separately on each pixel. If the number of acquisitions is large enough, it is possible also to estimate pixelwise the dispersion parameter, using the full likelihood. This method will not be applied in this chapter, but it will be deepened in Chapter 3.

2.7.1 Adaptive algorithm

We propose an experimental design performed by choosing a convenient b-value for a new measurement, basing on estimates obtained in previous measurements. A convenient functional of the FIM is iteratively optimized with respect to $\mathbf{b} = (b_1, \ldots, b_n)$ for

a fixed sample size n. The pointwise likelihood estimation is carried out following the framework described in Section 2.6, i.e. using the likelihood function

$$L(\nu_0, \alpha, \sigma^2|\mathbf{m}, \mathbf{b}) = \prod_{i=1}^{n} \frac{m_i}{\sigma^2} e^{-\frac{m_i^2 + \nu_0^2 e^{-2\alpha b_i}}{2\sigma^2}} I_0\left(\frac{m_i \nu_0 e^{-\alpha b_i}}{\sigma^2}\right) I_{(0,+\infty)}(m_i). \qquad (2.9)$$

The optimal values of ν_0 and α for equation (2.9) cannot be obtained explicitly, so a numerical optimization method is required to obtain estimates $\hat{\nu}_0$ and $\hat{\alpha}$. In this work we use the *L-BFGS* method (see [39]) with interval constraints, implemented in the R function `optim` [43].

The FIM depends on the values of the unknown parameters ν_0 and α in a non trivial way, so we perform the optimization fixing their values and assuming that they are near to the real ones. We propose an adaptive greedy optimization, that alternates the following steps:

1. Compute estimates $\hat{\nu}_0^{(h)}$ and $\hat{\alpha}^{(h)}$ of ν_0 and α using measurements performed on $b_1^*, \ldots, b_h^*$.

2. Fixing $(b_1, \ldots, b_h) = (b_1^*, \ldots, b_h^*)$ and $(\nu, \alpha) = (\hat{\nu}_0^{(h)}, \hat{\alpha}^{(h)})$, find

$$b_{h+1}^* = \arg\min_b F\left[\mathcal{I}(\hat{\nu}^{(h)}, \hat{\alpha}^{(h)}|\sigma^2, b_1^*, \ldots, b_h^*, b)\right],$$

being F a suitable functional of the FIM, then perform a new measurement m_{h+1} at point b_{h+1}^* and increment h by 1.

Various choices are possible for $F\left[\mathcal{I}(\hat{\nu}_0, \hat{\alpha}|\sigma^2, b_1^*, \ldots, b_n^*)\right]$. Some examples are $\mathcal{I}_{22}^{-1}$, the asymptotic variance of the ML estimator $\hat{\alpha}$, or the determinant and the trace of $\mathcal{I}^{-1}$, which take into account the variability of all the estimates.

The method described above refers to a single pixel of a MR image, but in applications it is necessary to improve some index of global precision on the whole image. This can be achieved, for example, by minimizing the sum of variances of ADC estimators on all pixels of the image.

The experimental design can be validated using measures performed on a dummy with known diffusion properties, collected on a fine grid of b-values. A subset of the available b-values and related measures can be used as fixed design observations, while the adaptive design can be approximated by choosing adaptively the available b-value nearest to the computed optimum. Estimates obtained in both cases can then be compared with documented physical values, at different sample sizes. Simulation studies have underlined the potential efficacy of the adaptive approach based on the sum of asymptotic variances with respect to a fixed design [6].

2.7.2 Simulation study

We tested our algorithm simulating signals from an ADC field over a 16×16 grid, which represents the considered area of tissue. The pattern chosen for the ADC distribution on the grid is formed by a central part, where the ADC α is fixed at $1.2 \times 10^{-3} mm^2/s$, an intermediate region where $\alpha = 1 \times 10^{-3} mm^2/s$ and an external region where $\alpha = 0.7 \times 10^{-3} mm^2/s$, as we can see in Figure 2.7. $\nu_0 = 1$ and $SNR_0 = \nu_0/\sigma = 7$ for all the cells of the grid.
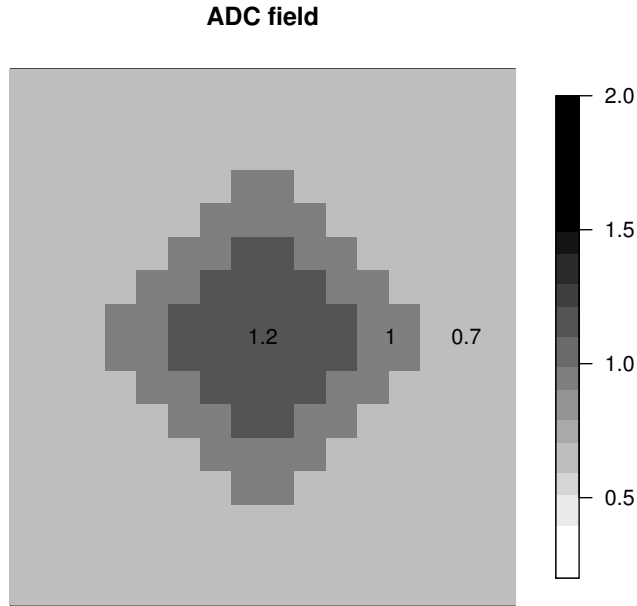
Figure 2.7: *Simulation ADC pattern. Reported values are ADC$\times 10^3$, in expressed in mm^2/s.*

Two cases are compared: ADC estimates performed using signals sampled with an equally spaced grid of $n = 4, 6, 10$ and 20 b-values ranging from 0 to 1000 s/mm^2, and ADC estimates performed using signals sampled with n b-values obtained from the adaptive design algorithm. The estimated ADC field with the two methods is represented in Figure 2.8.

Comparing the adaptive method (right panels) with the regular grid method (left panels), from Figure 2.8 we notice a slightly more defined mass in the central part and less noise over the whole image in the adaptive case. White pixels correspond to ADC estimates that hit the search domain $(0, 5)$. The difference is more evident in a comparison of mean squared errors on the whole image, represented in Figure 2.9: with all sample sizes, the MSE obtained by the adaptive grid is about half the MSE obtained with the fixed design.

2.7.3 Validation with phantom data

To test the adaptive design of experiment, various estimates were performed using MRI images of a container of distilled water immersed into a larger container of ice. This *phantom* is similar to the one used in [15], here taken as a reference for what concerns the ADC in water at almost 0°C. To simulate an adaptive experimental design, acquisitions at many different b-values were performed, so that estimates obtained using observations obtained on a fixed, regularly-spaced grid could be compared with estimates performed on observations obtained with b-values selected by the algorithm. Due to machine constraints, 16 b-values were used: $0, 50, 100, \ldots, 450, 500$ and
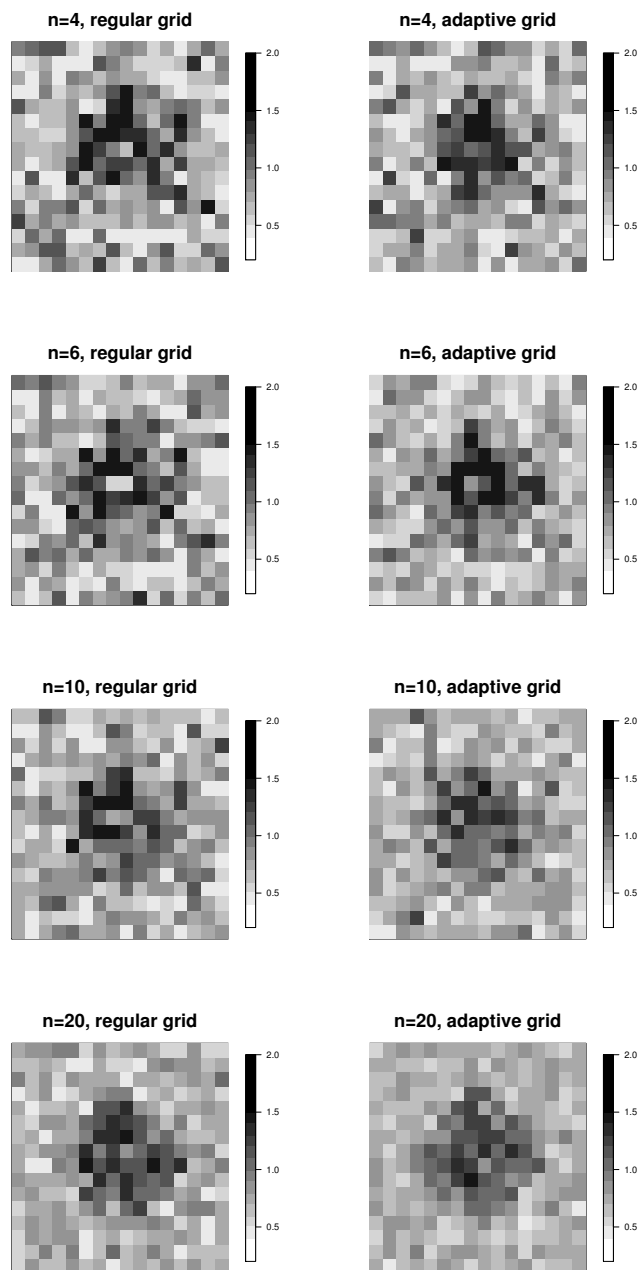
Figure 2.8: *ADC field estimates obtained using a uniform grid of b-values (left panels) and an adaptive sequence (right panels) with $n = 4, 6, 10, 20$ observations.*

$600, 700, \dots, 1000$. The method presented in Section 2.7.1 could be applied by choosing the best next b-value among the available ones, basing on the sum of a given target functional over the pixels of a region of interest (ROI). The chosen functional was the determinant of the FIM on a pixel, to be maximized (remind that the inverse FIM is related to the asymptotic MLEs covariance matrix).

The ROI was chosen by hand so that it contained only pixels of distilled water, avoiding areas disturbed by artifacts or where the discrimination between water and
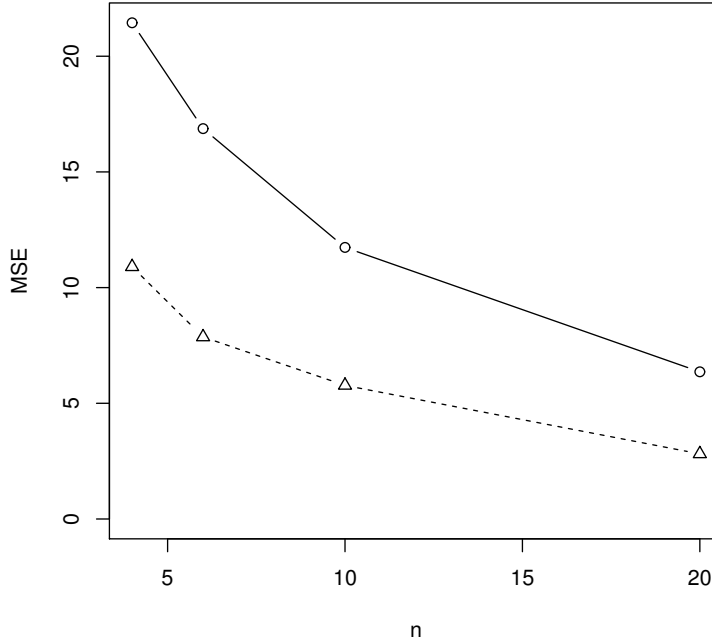
Figure 2.9: *Global mean square errors of fixed (circles and solid line) and adaptive (triangles and dashed line) design estimators.*

the wall of the test tube was unclear. Figure 2.10 shows the slices from the $b = 0$ acquisition that contain the ROI (in red). Estimates were performed only on the ROI, a subset of 2624 pixels of the original $132 \times 132 \times 25$ volume.

The noise parameter σ^2 was considered as uniform on the whole volume, and estimated in advance with the following method: pixels with observation at $b = 0$ lower than 50 were considered as background ones (set $\mathcal{N}$), and used to compute the estimate $\hat{\sigma^2} = \sum_{j \in \mathcal{N}} m_j^2 / (2|\mathcal{N}|)$. Several other methods for estimating the background noise dispersion parameter are present in literature, see [3] for a review of other possibilities. Our choice seemed compatible with the very high SNR observed in signal pixels (an example is displayed in Figure 2.11): we obtained $\hat{\sigma^2} = 11.754$, while the minimum magnitude value in the ROI at $b = 1000$ is 269. This fact, together with the relatively low ADC of distilled water that keeps the magnitude signal well away from the noise floor, would give estimates of high quality even with classical methods; anyway, testing on a substance with well documented properties allows to validate the performance of the method. Moreover, a small but constant advantage of the adaptive design in terms of estimates accuracy is evident from Table 2.1.

Observations at $b = 0$ were always included as starting points, as it is very common to perform an acquisition with the highest possible SNR. The mean values across pixels of the ROI is compatible with the literature, while the empirical standard deviation of the estimates is almost always slightly lower with the adaptive method. The only exceptions are given by regular grids including $b = 1000$ observations: when the most

36
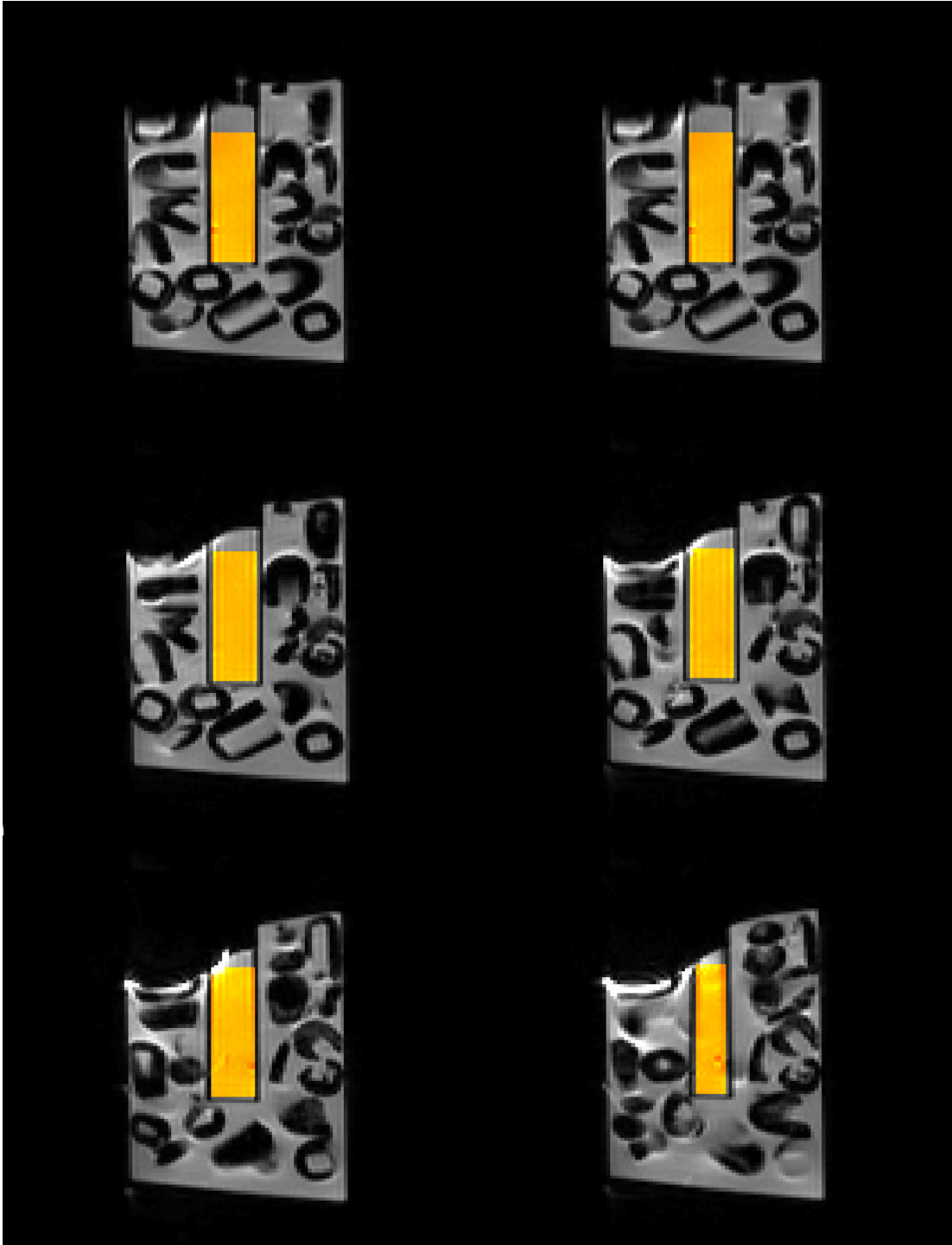
Figure 2.10: *Slices 11 to 16 of the MR acquisition at $b = 0$. The ROI is highlighted in yellow.*

"useful" points ($b = 0$ and $b = 1000$) in terms of accuracy are taken into account, the adaptation can fail in predicting the best solution at the next few steps. In general, since the SNR observed in these experiment is very high, the adaptive method chooses extremal points as the ones that maximize accuracy, as expected. This was not the case in the simulation study, in which the noise dispersion was large enough to provide non

Figure 2.11: *Plot of b-values vs observed intensities in a pixel of the ROI.*

trivial choices for the best points to select.

2.8 Concluding remarks

In the first part of this chapter we presented different methods for estimating pixelwise the ADC from diffusion MR signals, f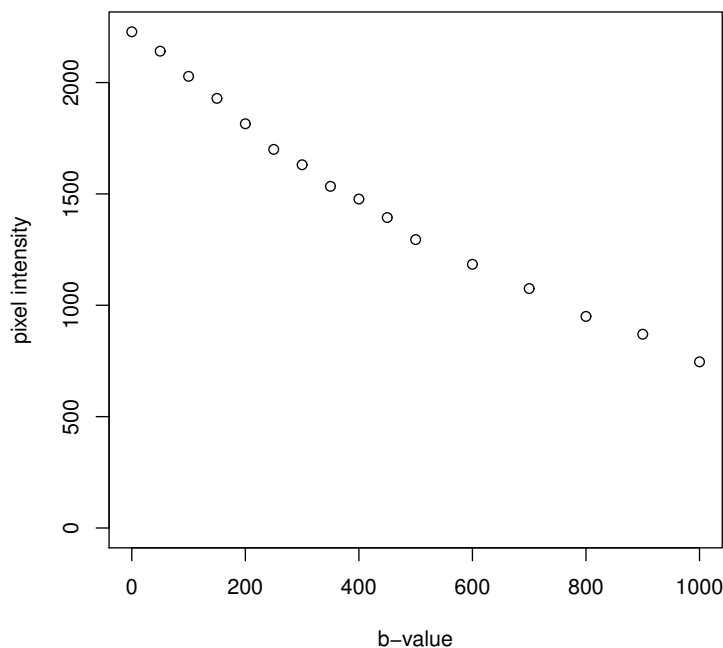ollowing the Rice noise model and the isotropic Stejskal-Tanner equation for magnitude decay. The presented estimators exhibit different features that should be taken into account when approaching real data.

The least squares approach is the fastest and has low variance, but becomes less accurate when the conditional signal distribution at different b-values is more distant from normality.

The maximum likelihood estimator is slightly slower, requiring a nonlinear maximization on 2 variables instead of 1, and has the lowest bias in many cases, but, as pointed out before, it has a certain probability of providing invalid estimates with samples from noisy signals. This problem disappears when estimating jointly diffusion parameters and noise dispersion, as the MLE in this case never collapses or diverges. In Chapter 3 an implementation of this kind of estimation is proposed, showing some bias in the estimate of σ^2 but good convergence properties.

Bayesian estimators are the most expensive in terms of computational costs, and may require further tuning for improving their performances; they are the best in terms of mean square error at the high ADCs here tested, and offer the advantage of providing the whole posterior distribution for inferential purposes, while inferential tools regarding LS and ML should rely, at present time, on normal approximations, which may not

38

fixed grid	adaptive grid	mean (fixed)	sd (fixed)	mean (adaptive)	sd (adaptive)	sd difference
0, 500, **1000**	50	1.0844	0.0365	1.0844	0.0366	-0.0000
0, 450, **900**	1000	1.0807	0.0390	1.0806	0.0389	0.0001
0, 400, **800**	1000	1.0782	0.0444	1.0782	0.0443	0.0001
0, 350, **700**	1000	1.0758	0.0502	1.0757	0.0500	0.0002
0, 300, **600**	1000	1.0763	0.0582	1.0763	0.0580	0.0003
0, 250, **500**	1000	1.0786	0.0655	1.0787	0.0651	0.0004
0, 200, **400**	1000	1.0800	0.0715	1.0801	0.0710	0.0005
0, 150, **300**	1000	1.0747	0.0801	1.0749	0.0793	0.0007
0, 100, **200**	1000	1.0587	0.0987	1.0590	0.0976	0.0011
0, 50, **100**	1000	1.0124	0.1593	1.0134	0.1576	0.0017
0, 300, 700, **1000**	50, 900	1.0844	0.0365	1.0844	0.0366	-0.0000
0, 300, 600, **900**	1000, 50	1.0807	0.0391	1.0807	0.0390	0.0000
0, 300, 500, **800**	1000, 50	1.0783	0.0444	1.0783	0.0443	0.0001
0, 200, 500, **700**	1000, 50	1.0758	0.0502	1.0758	0.0501	0.0002
0, 300, 700, **1000**	1000, 50	1.0758	0.0501	1.0758	0.0501	0.0001
0, 200, 400, **600**	1000, 50	1.0764	0.0583	1.0764	0.0580	0.0002
0, 300, 600, **900**	1000, 50	1.0764	0.0581	1.0764	0.0580	0.0001
0, 150, 350, **500**	1000, 50	1.0786	0.0655	1.0787	0.0652	0.0004
0, 200, 500, **700**	1000, 50	1.0786	0.0654	1.0787	0.0652	0.0002
0, 150, 250, **400**	1000, 900	1.0800	0.0715	1.0802	0.0708	0.0007
0, 200, 400, **600**	1000, 900	1.0800	0.0714	1.0802	0.0708	0.0006
0, 150, 200, **300**	1000, 900	1.0747	0.0801	1.0750	0.0791	0.0010
0, 150, 300, **450**	1000, 900	1.0747	0.0800	1.0750	0.0791	0.0009
0, 300, 600, **900**	1000, 900	1.0750	0.0792	1.0750	0.0791	0.0001
0, 50, 150, **200**	1000, 900	1.0587	0.0987	1.0593	0.0974	0.0014
0, 100, 300, **200**	1000, 900	1.0587	0.0986	1.0593	0.0974	0.0013
0, 200, 400, **600**	1000, 900	1.0592	0.0980	1.0593	0.0974	0.0006
0, 50, 100, **150**	1000, 900	1.0125	0.1592	1.0139	0.1570	0.0022
0, 100, 200, **300**	1000, 900	1.0132	0.1585	1.0139	0.1570	0.0015

Table 2.1: *Comparison between fixed and adaptive design on the ROI. The first two columns present the b-values of the fixed grid, marking in bold the two shared with the adaptive design, and the b-values chosen by the proposed method. Other columns present sample mean and standard deviation over the ROI for the two methods. The last column reports the difference* (sd fixed) − (sd adaptive).

be reliable with low sample sizes and SNRs. A possible way of improving the speed of Bayesian computation could be to design an integrated nested Laplace approximations approach for the Rician exponential regression problem, thus cutting out simulation times for markov chain sampling.

The last part of the chapter deals with improving estimation accuracy by optimizing the experimental design. The proposed method shows good performances in simulation, when compared to a fixed design, and promising results with real data, although the improvement is of small entity because of the build of the phantom available for testing. The application of this approach could lead to an improvement in the quality of diffusion MRI scans while avoiding the extension of acquisition times. The adaptive design should be further validated by testing its performance on different kinds of tissues, with stronger diffusion properties. Moreover, with biological data many possible choices for the single-pixel functional and the multi-pixel combination function could be tested, adopting a more goal-oriented approach to estimation improvement.

Diffusion weighted imaging: noncentral chi estimation in DTI

In recent years, many efforts have been done to account for the non-gaussianity of the intensity signal in diffusion MRI. In particular, various estimation methods for the Diffusion Tensor Imaging (DTI) model [10] have been extended to the Rician noise model (see for example [34], [21] and [5], which arises from taking the modulus of a single complex signal, assumed to be gaussian in both its real and imaginary component. The Rician model is effective in accounting for the signal skewness at low Signal-to-Noise Ratios (SNRs), but the assumptions for its validity are violated in presence of multi-coil scanners (or multi-channel coils) [3]. If the signals received by the N coils are combined by sum of squares, the intensity signal follows a noncentral chi distribution [16]. In this chapter we present a framework for accounting for the distribution of a multi-coil signal under the DTI model. First, we propose a hypothesis test-based method for discriminating background from signal voxels basing on the noise distribution, with the possibility of accounting locally for the apparent number of coils. Moreover, we propose a method for estimating the diffusion parameters of the DTI model in noncentral chi-distributed signals. The effectiveness of the noise voxels discrimination test is demonstrated on real DWI brain data, while a simulation study shows the properties of the likelihood method in single-voxel estimation, compared to other classical methods.

3.1 The DTI model for noncentral chi-distributed data

If the signals registered by multiple receiver coils can be approximately considered as independent, the sum of two squared Gaussian variables described in Section 2.2 gets generalized to a sum of $2N$ terms. Diffusion-weighted MR signal magnitudes become then *scaled noncentral chi-distributed* random variables [16], denoted $M_{ij} \sim$

$\mathrm{NC}\chi_{2N}(\nu_{ij}^2/\sigma_j^2)$, i.e. for any acquisition $i = 1, \ldots, n$ the signal intensity M_{ij} measured on any voxel $j = 1, \ldots, n_v$ of an observed volume has probability density function defined as

$$f_{M_{ij}}(m_{ij}) = \frac{\nu_{ij}}{\sigma^2} \left(\frac{m_{ij}}{\nu_{ij}}\right)^N e^{-\frac{m_{ij}^2 + \nu_{ij}^2}{2\sigma^2}} I_{N-1}\left(\frac{m_{ij}\nu_{ij}}{\sigma^2}\right),$$

where ν_i is the unknown "true" signal at voxel j and m_{ij} its noisy observation, N is the number of coils of the MR scanner, σ^2 is the noise parameter (the noise variance in the k-space) and I_{N-1} is the modified Bessel function of the first kind of degree $N-1$. In the DTI framework, the signal ν_{ij} depends on the local diffusion tensor D_j, the *b-value* b_i and the applied magnetic gradient $\mathbf{g}_i$ by the Stejskal-Tanner equation [9]

$$\nu_{ij} = \nu_{0j} e^{-b_i \mathbf{g}_i^T D_j \mathbf{g}_i}. \tag{3.1}$$

The methods proposed in this chapter make use of startup estimates, which can be obtained through a low accuracy but fast method. The following section describes a natural, simple way of obtaining a preliminary estimate of the diffusion tensor.

3.2 Preliminary DTI estimation

Considering the full Stejskal-Tanner model, it is possible to take the logarithm on both sides, and to approximate the problem to a regression on the log scale. Equation (3.1) can be conveniently rewritten as

$$\nu_{ij} = e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}, \tag{3.2}$$

with

$$\mathbf{x}_i = -b_i \cdot (1, g_{i1}^2, g_{i2}^2, g_{i1}^2, 2g_{i1}g_{i2}, 2g_{i1}g_{i3}, 2g_{i2}g_{i3}), \quad i = 1, \ldots, n$$
$$\boldsymbol{\beta}_j = (\log(\nu_{0j}), d_{11}, d_{22}, d_{33}, d_{12}, d_{13}, d_{33})^T, \quad j = 1, \ldots, n_v$$

being d_{kl} elements of D_j. Taking the logarithm on both sides of the equation, least squares estimates can be obtained by solving the regression model

$$\log(m_{ij}) = \mathbf{x}_i^T \boldsymbol{\beta}_j + \epsilon_{ij}, \quad i = 1, \ldots, n, \tag{3.3}$$

where ϵ_{ij} is a residual, zero-mean error. In analogy with classical linear regression notation, the elements of $\boldsymbol{\beta}_j$ will be indexed as $\beta_{0j}, \ldots, \beta_{6j}$, due to the clear meaning of β_{0j} as intercept in the *log-linear model* (3.3).

3.3 Individuation of pure noise voxels

The first method we propose aims at discriminating background from tissue voxels, basing on the noise distribution with null signal. This strategy is not intended for performing segmentation: the main purpose is to achieve an accurate quantification of background noise, by computing an accurate estimate of σ_G^2, the global noise parameter for the background region. This is done by basing the estimate on a set $\mathcal{N}$ of pure noise voxels.

The first step consists in obtaining a startup subset of $\mathcal{N}$, individuating pixels that can be most easily identified as background ones. See Section 3.3.1 for some ideas on how to find a safe startup set.

A refinement of this initial set $\mathcal{N}$ is then performed exploiting the distribution of M_{ij} under the null hypothesis that $\nu_j(b_i, \mathbf{g}_i) = 0$ for any b-value of b and applied magnetic gradient $\mathbf{g}$. In this case, M_{ij} follows a (central) chi distribution with $2N$ degrees of freedom for each $i = 1, \ldots, n$, implying that

$$\widehat{\sigma_j^2} = \frac{\sum_{i=1}^{n} M_{ij}^2}{2Nn} \sim \frac{\sigma^2}{2Nn} \chi^2(2Nn).$$

Consequently, the distribution of $\widehat{\sigma^2}_{\mathcal{N}}$, a global estimator of σ^2 obtained by pooling on the set of pure noise voxels $\mathcal{N}$, is the following

$$\widehat{\sigma_{\mathcal{N}}^2} = \frac{\sum_{i=1}^{n} \sum_{j \in \mathcal{N}} M_{ij}^2}{2Nnm} \sim \frac{\sigma^2}{2Nnm} \chi^2(2Nnm),$$

where $m = |\mathcal{N}|$. This allows to define, for each voxel $j \notin \mathcal{N}$, the test statistic:

$$T_j = \frac{\widehat{\sigma_j^2}}{\widehat{\sigma^2}_{\mathcal{N}}} \sim \frac{\chi^2(2Nn)/2Nn}{\chi^2(2Nnm)/2Nnm} = F(2Nn, 2Nnm),$$

where $F(2Nn, 2Nnm)$ is a Fisher distribution with $2Nn$, $2Nnm$ degrees of freedom. It is then finally possible to establish whether a pixel $j \in \mathcal{N}$ containins signal or not by performing a multiple hypothesis test using the valid p-values:

$$p_A(T_j) = [1 - F_{F(2Nn,2Nnm)}(T_j)] \cdot A(j, n_v - m),$$

where $F_{F(2Nn,2Nnm)}$ is the c.d.f. of a Fisher random variable with $2Nn$, $2Nnm$ degrees of freedom and $A(j, n_v)$ is an adjustment factor for test multiplicity. A good strategy for correcting p-values is for example the one from Benjamini and Hochberg [11], that allows to compare the adjusted p-values to α, a desired False Discovery Rate (FDR) level to ensure. Voxels for which the null hypothesis is rejected are considered as pure background voxels.

The test described above can be easily adjusted in case of spatially varying number of coils: it suffices to replace the denominators $\widehat{\sigma_j^2}$ and $\widehat{\sigma^2}_{\mathcal{N}}$ and the associated degrees of freedom of the Fisher distribution with $2N_j n$ and $2n \sum_{j \in \mathcal{N}} N_j$, respectively, where N_j is the apparent number of coils at voxel j.

3.3.1 Startup of set $\mathcal{N}$

Initializing the set $\mathcal{N}$ is a task that requires some attention, since introducing heavy selection bias could have a negative impact on the final performance of the hypothesis test above. On one hand, simply selecting voxels with a very low sample mean value of the observations could lead to sampling just very low quantiles of the noise distribution, thus not allowing the algorithm to move much from this set. On the other hand, choosing arbitrarily a threshold for some voxelwise statistic could lead also to consider as pure noise pixels some signal pixels; since the algorithm can only enlarge the set of pure noise pixels, this would brake up the method as well. Here we propose a couple of more elaborate criteria, although still fast to carry out.

Quantile method. If at least two different b-values have been used in the acquisitions, an initial set of pure noise voxels can be selected by the following criterion: let the observations $m_{1j}, \ldots m_{nj}$ at voxel j be sorted by increasing b-values. The median (quantile of order 0.5) of observations $m_{2j}, \ldots, m_{nj}$ is computed, and compared to m_{1j}. If the observation that should theoretically be the highest is not highly ranked amongst the sample, a decaying exponential model w.r.t the b-value is unlikely to hold, indicating that the pixel should be regarded as one of pure noise. To improve the discrimination of pure noise voxels from signal points with a high noise level, this criterion is applied again, progressively halving the order of the quantile for comparison. This is stopped when the estimated global variance computed on selected pure noise voxels has a variation smaller than a fixed, roughly tuned tolerance.

Intercept test method. An initial set of pure noise voxels could also be selected by performing a multiple hypothesis test on coefficients β_{0j}. We test the null hypothesis H_{0j}: $\beta_{0j} < \delta$ *in voxel j*, where δ is a threshold "small" enough to be considered non arbitrary and to ensure that the voxel contains just pure noise and no signal. Without loss of generality, fixing $\delta = 0$ (which implies $\nu_{0j} < 1$) would give the p-values

$$p_A(\hat{\beta}_{0j}/s_j) = [1 - F_{t(n-7)}(\hat{\beta}_{0j}/s_j)] \cdot A(j, n_v),$$

where $F_{t(n-7)}$ is the c.d.f. of a Student's T random variable with $n - 7$ degrees of freedom (n minus the number of regression parameters) and $A(j, n_v)$ is an adjustment factor for test multiplicity.

3.3.2 Test on brain data

We tested the background discrimination method described above on brain imaging data acquired on a healthy subject by a Siemens Magnetom Verio syngo MR B17 scanner with a 32 channels coil. Here we present the results on 4 slices (20,25,30 and 35) from the original $128 \times 128 \times 60$ volume. Figure 3.1 shows on the left column the initialization obtained with the quantile method, while on the right column the final result is shown. Figure 3.2 displays the same comparison, presenting instead the initialization by intercept test on the left.

Comparing the two figures, it is evident that the quantile method selects a smaller set of startup points with respect to the intercept test methods, nonetheless the two initializations yield two similar results, as can be seen on right columns. The small difference in the final set of pure noise voxels gives slightly different results: in both cases the algorithm stops after 2 iterations, but in the first case (quantile method) $\widehat{\sigma^2} = 0.531189$ (based on 680410 voxels), while in the second $\widehat{\sigma^2} = 0.273694$ (660654 voxels). From the figures we notice that the quantile method startup leads the test to select also some pixels inside the brain region, thus providing a higher value for the noise dispersion parameter estimate. In general, on this kind of data, initializing with the quantile method leads to the selection of more voxels as pure noise, which on one side provides a larger sample for estimation, but on the other side chooses many "unsafe" voxels. Notice that in both cases, as anticipated before, the result is not a segmentation of the brain, which is not the aim of this method, but rather the definition of an area which does not contain signal-generating tissues. The signal area is larger than the actual brain circumference, since surrounding tissues (skin, muscles) can exhibit light diffusion properties in the

MR scan; this tissues are correctly identified by the algorithm, with both the startups, as signal areas.

3.4 Maximum likelihood estimation

Let us consider a sample of observations M_i, $i = 1, \ldots, n$, following a $NC\chi_{2N}(\nu_i^2/\sigma^2)$, with $\nu_i = \nu(\boldsymbol{\beta}; \mathbf{x})$ dependent on a vector of parameters $\boldsymbol{\beta}$ and a vector of covariates $\mathbf{x}$. The log-likelihood for a sample of size n following this model is

$$\log L(\boldsymbol{\beta}, \sigma^2 | \mathbf{m}, \mathbf{x}) = \sum_{i=1}^{n} (2N-1)\log(m_i) - N\log(\sigma^2) - \frac{m_i^2 + \nu_i^2}{2\sigma^2} + \log\left(\frac{I_{N-1}\left(\frac{m_i\nu_i}{\sigma^2}\right)}{\left(\frac{m_i\nu_i}{\sigma^2}\right)^{N-1}}\right). \tag{3.4}$$

Since we will always consider N as a known parameter, notice that the term $(2N-1)\log(m_i)$ does not depend on any of the unknowns, so it can be simply removed from the likelihood for maximization purposes. The particular grouping of terms chosen in (3.4) is useful because it allows to exploit the following property of modified Bessel functions of the first kind: $d(x^{-k}I_k(x))/dx = x^{-k}I_{k+1}(x)$. Using this fact, the partial derivatives of $\log L$ w.r.t. parameters $\beta_0, \ldots, \beta_q$ assume the form

$$\frac{\partial}{\partial \beta_j} \log L(\boldsymbol{\beta}, \sigma^2 | \mathbf{m}, \mathbf{x}) = \sum_{i=1}^{n} \left[-\frac{\nu_i}{\sigma^2} + \frac{I_N\left(\frac{m_i\nu_i}{\sigma^2}\right)}{I_{N-1}\left(\frac{m_i\nu_i}{\sigma^2}\right)} \frac{m_i}{\sigma^2} \right] \frac{\partial \nu_i}{\partial \beta_j}. \tag{3.5}$$

The Bessel functions ratio I_N/I_{N-1} requires particular attention, since evaluating modified Bessel functions of the first kind individually may lead to arithmetic overflow or underflow. We adopted the modified Bessel functions ratio approximation proposed in [22], in particular we use Perron's continuous fraction representation, which gave accurate results by keeping just the first 10 terms of the representation. The method was tested in R by comparing it to the direct evaluation of the ratio $I_N(x)/I_{N-1}(x)$, in the range where it was applicable, at $N = 32$ (over/underflow problems are more evident at high N). Results were concordant up to the 7th decimal digit when approaching the upper limit $x = 709$, and Perron's method keeps providing valid numbers (lower than 1) up to $x = 10^8$, and 1 for greater values. The same precision is observable up to the lower limit of direct evaluation, about 10^{-8}, with Perron's method still providing valid numbers for lower values of the argument.

The Bessel function logarithm in equation (3.4) could also have overflow or underflow problems, so asymptotic expansions for large (> 600) and small values ($< \sqrt{(N+1)/100}$) were used, the former truncated to the second term (see [2]) and the latter taken as $x/(2N!)$. Approximations yielded satisfactory results also in this case.

If the diffusion model admits tractable expressions for the partial derivatives with respect to the diffusion parameters, the optimization can be performed taking advantage of the gradient of the target functional to be optimized. In particular, when considering the exponential model (3.2) we have

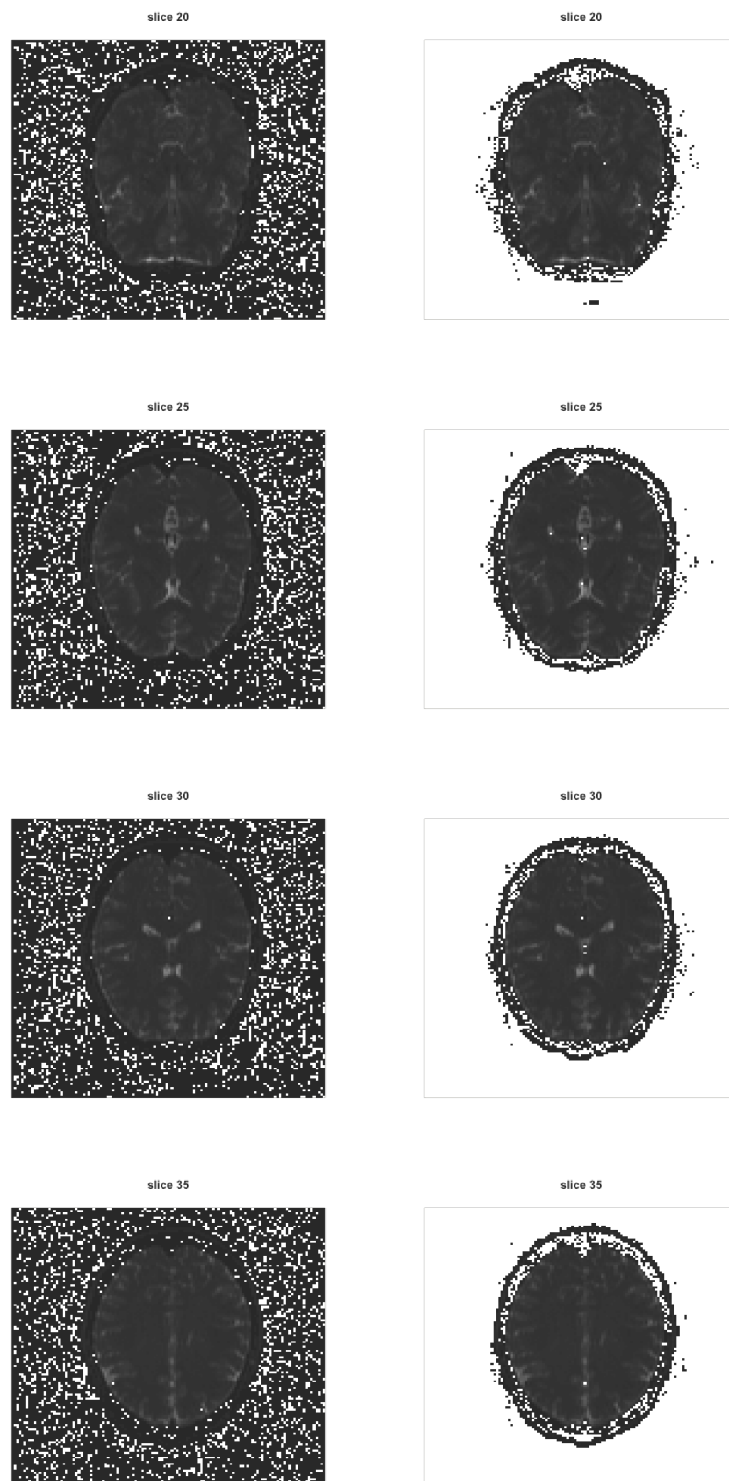$$\frac{\partial \nu_i}{\partial \beta_j} = \nu_i x_{ij},$$

45

Figure 3.1: *Startup pure noise pixels provided by the quantile method, on the left, and final set of pure noise pixels, on the right. Pure noise pixels are in white, superimposed to slices 20,25,30 and 35 at $b = 0$.*
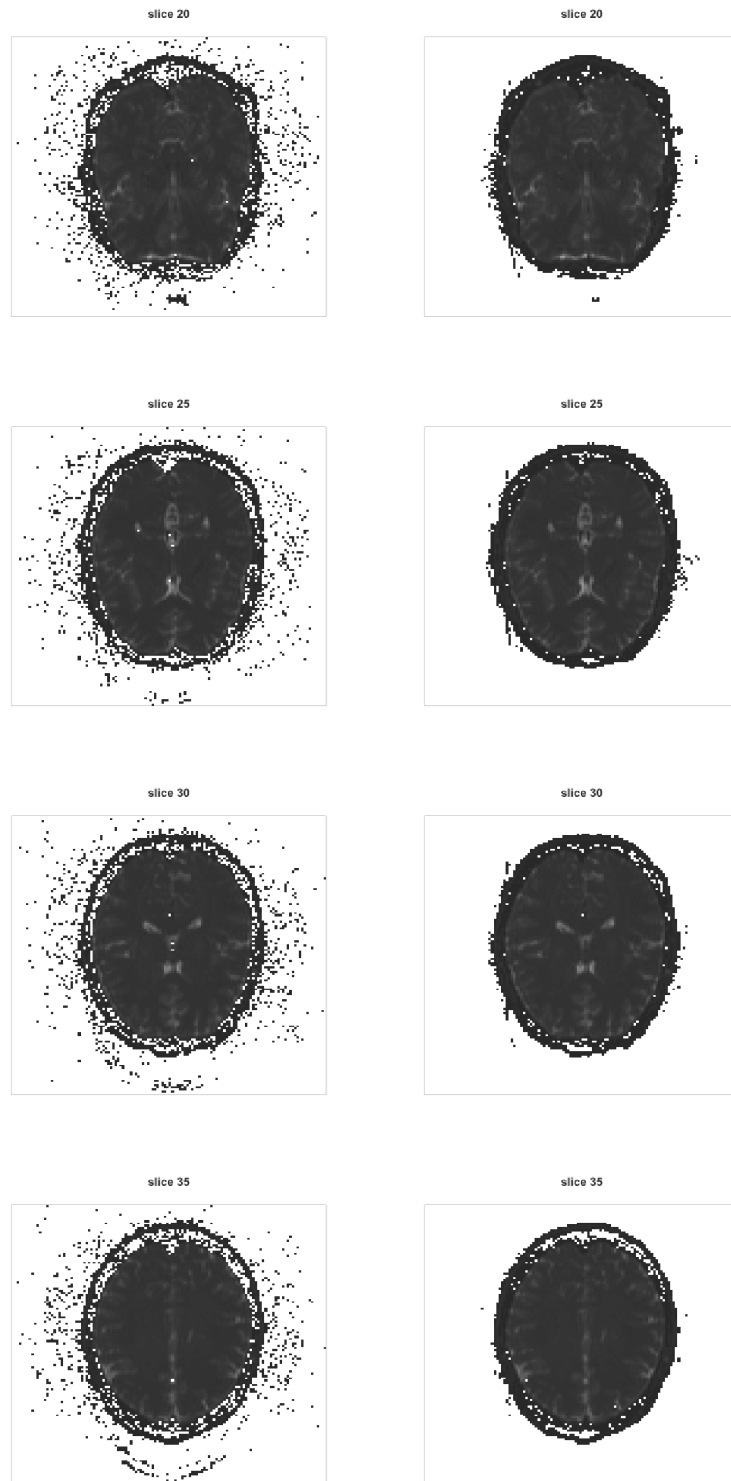
Figure 3.2: *Startup pure noise pixels with provided by the intercept test method, on the left, and final set of pure noise pixels, on the right. Pure noise pixels are in white, superimposed to slices 20,25,30 and 35 at b = 0.*

47

where x_{ij} is the j-th element of vector $\mathbf{x}$. Then we can rewrite 3.5 as

$$\frac{\partial}{\partial \beta_j} \log L(\boldsymbol{\beta}, \sigma^2 | \mathbf{m}, \mathbf{x}) = \sum_{i=1}^{n} \left[-\frac{\nu_i^2}{\sigma^2} + \frac{I_N\left(\frac{m_i \nu_i}{\sigma^2}\right)}{I_{N-1}\left(\frac{m_i \nu_i}{\sigma^2}\right)} \frac{m_i \nu_i}{\sigma^2} \right] x_{ij}. \qquad (3.6)$$

The partial derivative w.r.t. σ^2 is useful too, since we now intend to estimate this parameter locally:

$$\frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \sigma^2 | \mathbf{m}, \mathbf{x}) = -\frac{nN}{\sigma^2} + \sum_{i=1}^{n} \left[\frac{m_i^2 + \nu_i^2}{2(\sigma^2)^2} - \frac{I_N\left(\frac{m_i \nu_i}{\sigma^2}\right)}{I_{N-1}\left(\frac{m_i \nu_i}{\sigma^2}\right)} \cdot \frac{m_i \nu_i}{(\sigma^2)^2} \right].$$

Parameter values are initialized with the diffusion parameters estimate described in section 3.2 and with the local dispersion parameter estimate

$$\widehat{\sigma_j^2} = \sum_{i=1}^{n} \frac{(m_{ij} - \nu_{ij})^2}{nN}.$$

We propose an optimization scheme that alternates the optimization of diffusion parameters and recomputation of σ^2.

During the $\boldsymbol{\beta}$ estimation step, the cost function to maximize in each voxel is the likelihood function (the footer j is omitted)

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\frac{\nu_i^2}{2\sigma^2} + \log \left(\frac{I_{N-1}\left(\frac{m_i \nu_i}{\sigma^2}\right)}{\left(\frac{m_i \nu_i}{\sigma^2}\right)^{N-1}} \right), \qquad (3.7)$$

obtained removing from (3.4) terms depending only on the observations or on the local noise parameter σ^2. Partial derivatives (3.6) can be used when considering the DTI, for example optimizing by a BFGS method with box constraints.

The estimate of σ^2 is computed separately, basing on its score equation (the partial derivative of the log-likelihood w.r.t. σ^2 equated to 0). Since it is an implicit equation, it is updated using the last estimated value of the parameter itself:

$$\sigma_{h+1}^2 = \sum_{i=1}^{n} \left[\frac{m_i^2 + \nu_i^2}{2nN} - \frac{I_N\left(\frac{m_i \nu_i}{\sigma_h^2}\right)}{I_{N-1}\left(\frac{m_i \nu_i}{\sigma_h^2}\right)} \cdot \frac{m_i \nu_i}{nN} \right].$$

It is reasonable to assume, and verified in practical situations, that the noise level in regions of interest is higher than in the background, due for example to partial volume effects and model misspecification. For this reason, the values that the local estimator of σ^2 can assume are truncated to the left at $\widehat{\sigma_N^2}$, the background noise parameter estimate.

The two steps above are repeated up to the stabilization of the parameters or when a maximum, preset number of iterations is reached.

3.5   Simulation study

We tested the proposed likelihood maximization algorithm with simulated, single-voxel data. A number of coils $N = 8$ was chosen to show the differences in estimation quality

in presence of model misspecification, i.e. when using the log-linear model (3.3) or the Rice exponential regression model, and when using the correct scaled noncentral chi model.

Simulation studies were carried out under 4 different parameter settings, which combine 2 different noise dispersion parameters and 2 diffusion matrices. ν_0 has been fixed to 1000 in all the experiments, so the SNRs at fixed b-value depend only on σ. The values of σ used in the simulations are $\sigma_1 = 10$ and $\sigma_2 = 100$, while the diffusion matrices are

$$D_1 = 10^{-3} \begin{bmatrix} 0.7 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 0.7 \end{bmatrix}, \qquad D_2 = 10^{-3} R^T \begin{bmatrix} 0.7 & 0 & 0 \\ 0 & 2.5 & 0 \\ 0 & 0 & 0.7 \end{bmatrix} R,$$

where R is a (counter-clockwise) rotation matrix of angle $\pi/4$ around the z axis. Notice that matrix D_2 has a stronger anisotropy, moreover it will have non-zero extra diagonal terms d_{12}, d_{21}, do to the rotation around the z axis.

For samples of size $n = 11, 31, 51, 71, 91$ we present the Root Mean Square Errors (RMSE). Individual simulations are organized in the following way: the $b = 0$ observation is always included (the gradient orientation does not matter in this case), while the remaining $10, 30, 50, 70, 90$ observations are performed at $b = 1000$ with angles sampled uniformly on the unit sphere.

The log-linear model estimate is used also for the initialization of the Rician and noncentral chi likelihood methods, based on the algorithm described in Section 3.4 with $N = 1$ for the Rician case and $N = 8$ for the noncentral chi case.

Figures 3.3 and 3.4 show the RMSEs for the diffusion parameters in D_1 at $\sigma = 10$ and $\sigma = 100$ respectively. In the first case, the one with high SNR, we notice that the three methods have almost the same performance in all cases, anyway the noncentral chi estimation has always the lowest RMSE. In the case with lower SNR, on the contrary, the noncentral chi estimation has remarkably poorer performances at low sample sizes, while it becomes more accurate for $n = 51, 71, 91$. A further analysis shows that the variance of this estimator is higher with respect to the others considered, while the bias is much lower, particularly for what concerns β_1, β_2 and β_3. The advantage of using the correct ML approach is then found mostly when considering sample sizes that decrease variance enough, since among possibly biased estimators it is not the one with uniformly lowest variance.

When considering the analyses performed with $D = D_2$, we see again a better behavior of the noncentral chi estimator at $\sigma = 10$ (Figure 3.5), now with a remarkable difference also on parameter $\beta_5 = d_{12}$. Figure 3.6 shows again that the noncentral chi estimator has poorer performances at low angular resolutions, due to high variance with low SNRs, but gains accuracy when increasing the sample size, while the other methods have almost constant performances even at large n.

The results obtained in this simulation study are somewhat coherent with the likelihood framework: imposing the correct model reduces bias in estimation, nonetheless the MLE is not necessarily the one with the smallest variance, particularly for small sample sizes and higher SNRs.
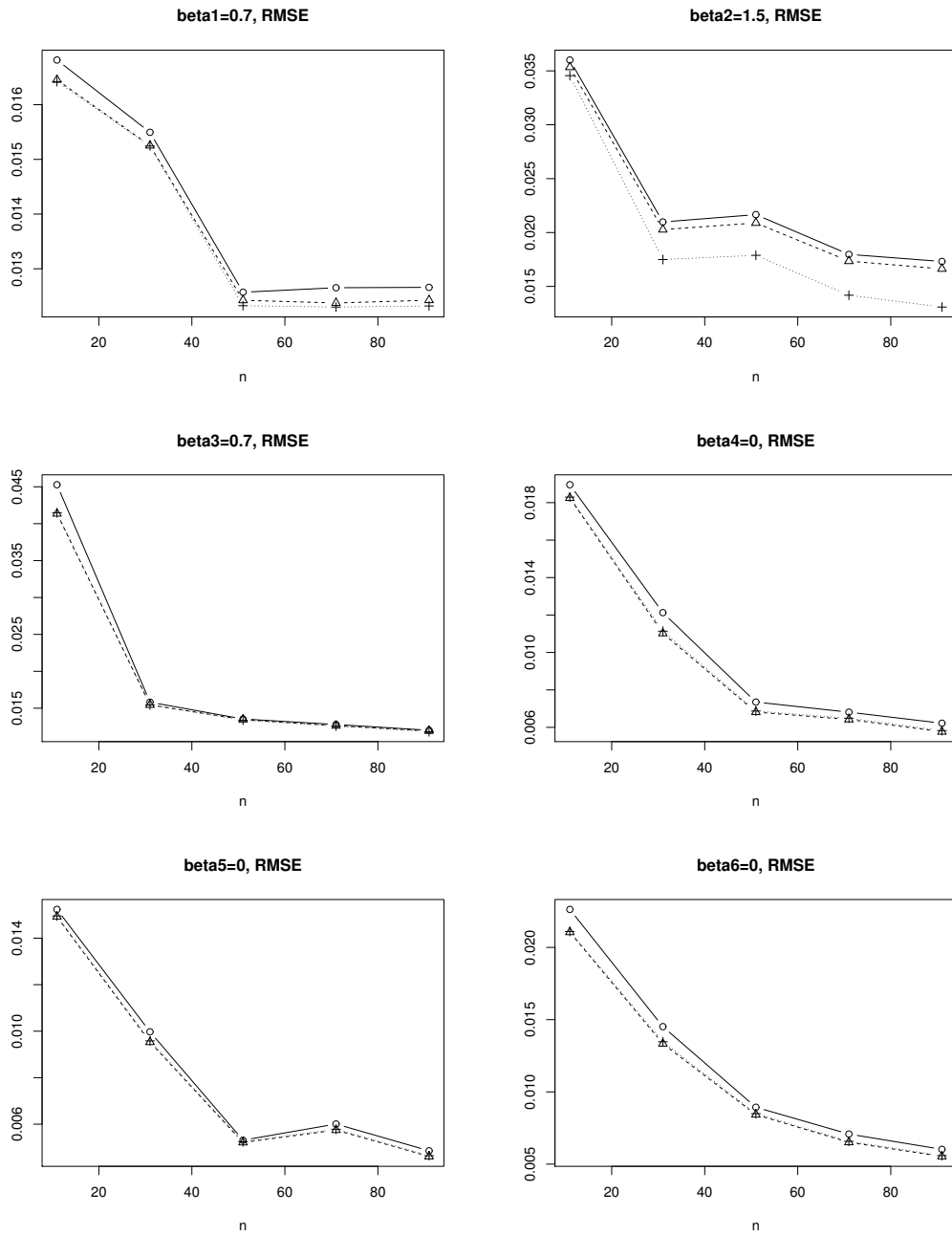
Figure 3.3: *Root Mean Square Error for the diffusion parameters* $\beta_2, \ldots, \beta_7$, *in the case* $\mathbf{D} = \mathbf{D_1}$, $\sigma = 10$.
Circles and solid line: log linear. Triangles and dashed line: Rice. Crosses and dotted line: noncentral chi.

Figure 3.4: *Root Mean Square Error for the diffusion parameters $\beta_2, \ldots, \beta_7$, in the case $\mathbf{D} = \mathbf{D_1}$, $\sigma = 100$.*
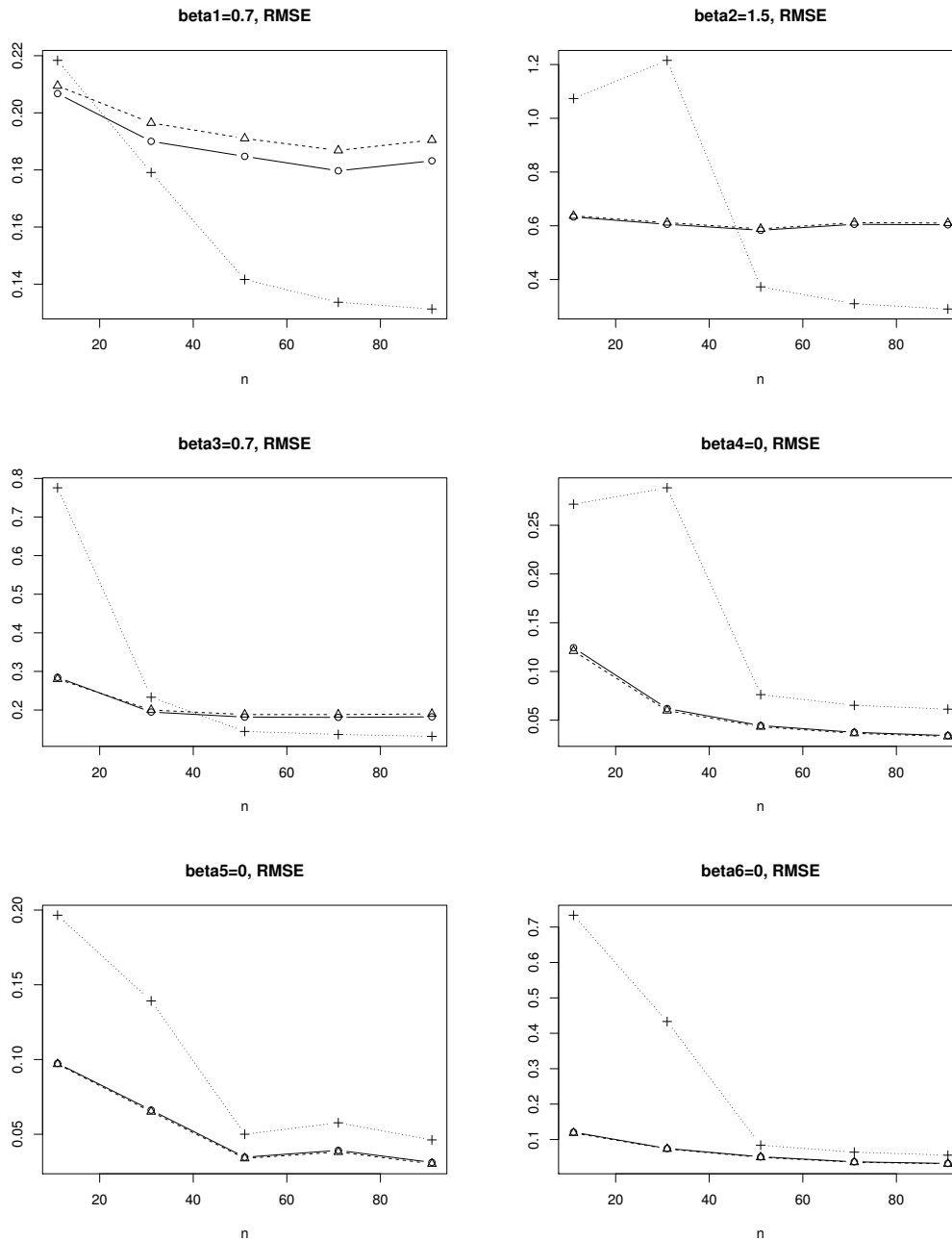*Circles and solid line: log linear. Triangles and dashed line: Rice. Crosses and dotted line: noncentral chi.*

51

Figure 3.5: *Root Mean Square Error for the diffusion parameters* $\beta_2, \ldots, \beta_7$, *in the case* $\mathbf{D} = \mathbf{D_2}$, $\sigma = 10$.
Circles and solid line: log linear. Triangles and dashed line: Rice. Crosses and dotted line: noncentral chi.
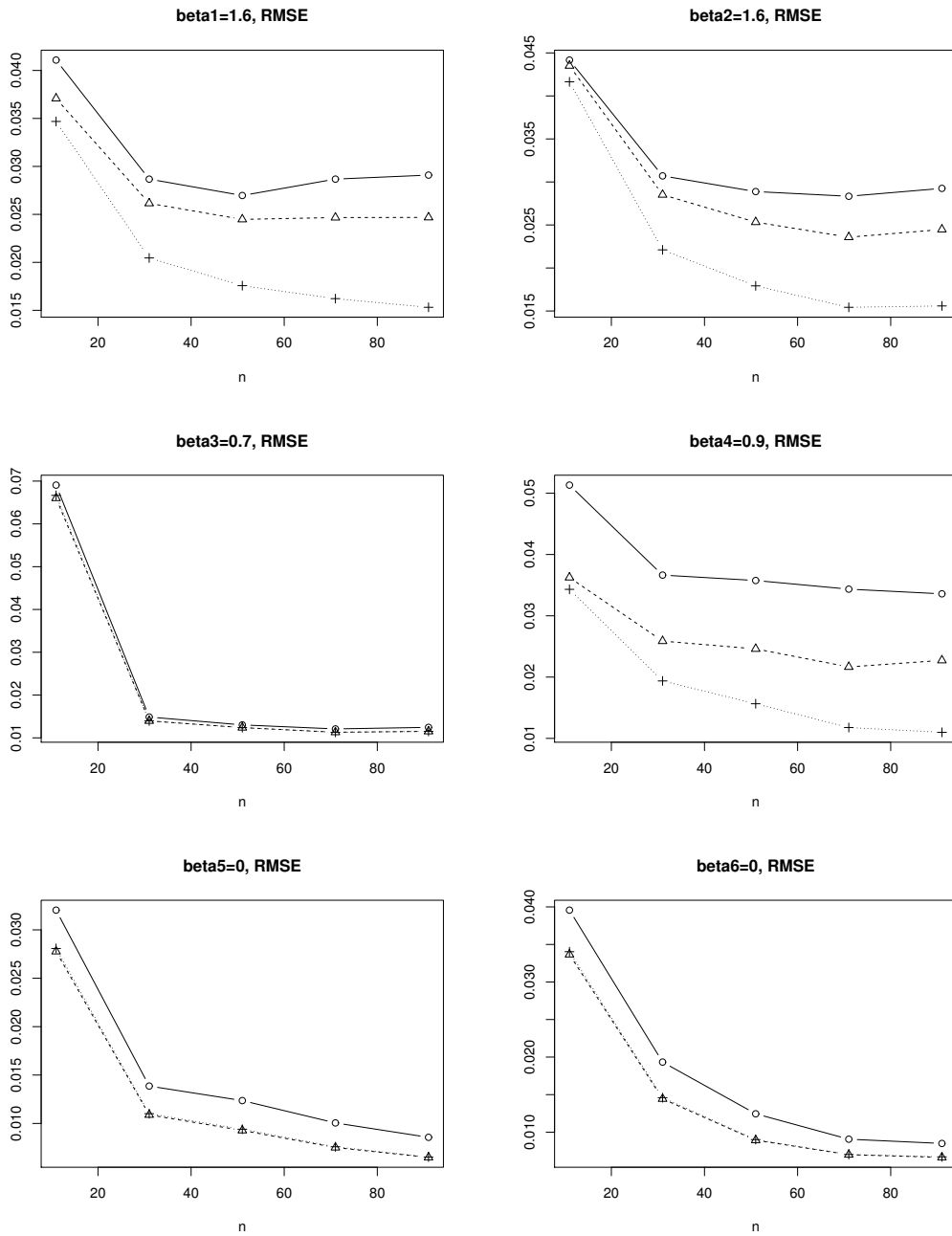
Figure 3.6: *Root Mean Square Error for the diffusion parameters $\beta_2, \ldots, \beta_7$, in the case $\mathbf{D} = \mathbf{D_2}$, $\sigma = 100$.*
Circles and solid line: log linear. Triangles and dashed line: Rice. Crosses and dotted line: noncentral chi.
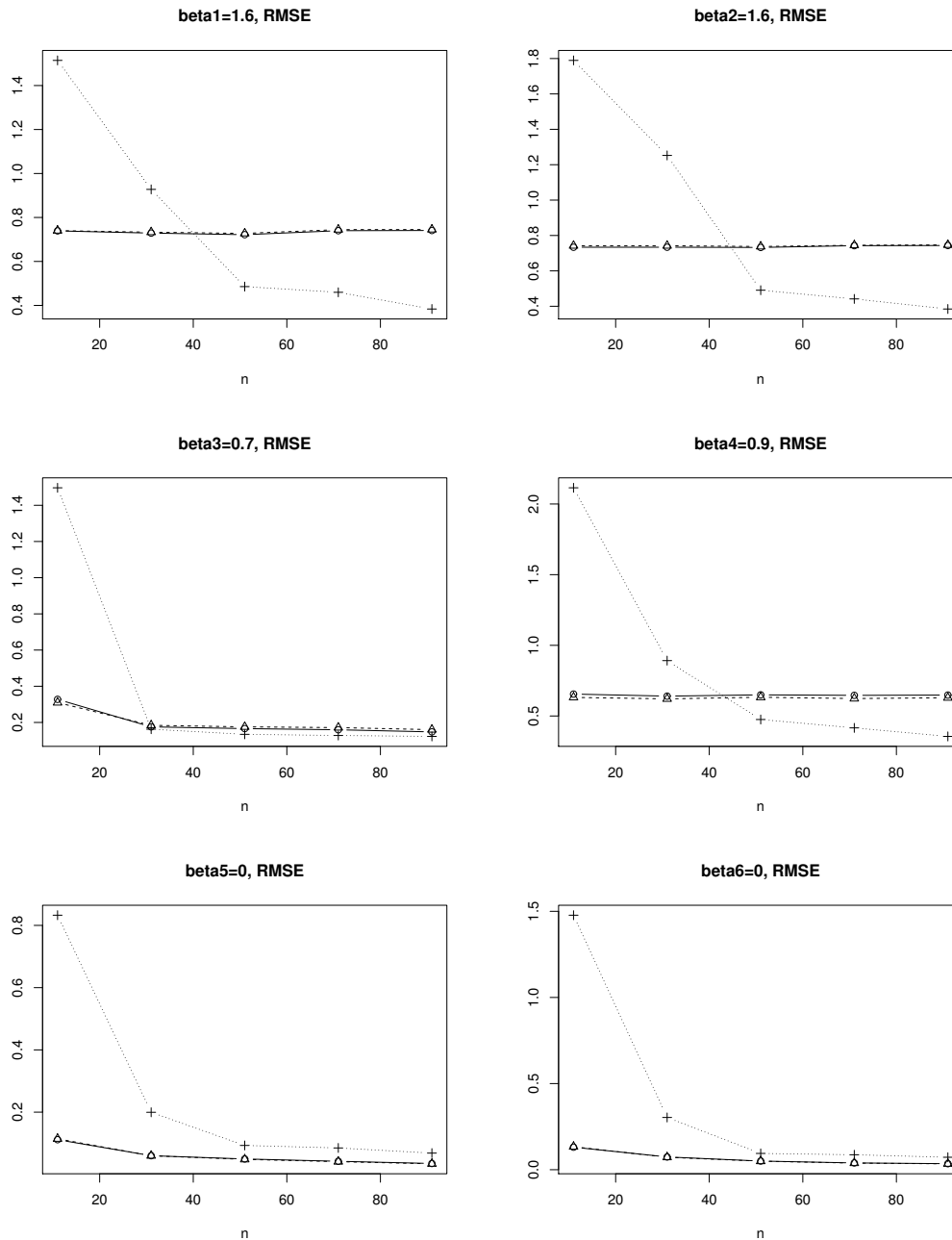
3.6 Final remarks and perspectives

Considering a proper model specification, in our case for what concerns the statistical distribution of the DWI signal, allows to set up more accurate estimation tools and to give a deeper interpretation to the observed random variables.

As seen in Section 3.3, accounting for the correct noise model allows to derive a rigorous hypothesis test for discriminating signal from pure noise voxels. The proposed method can be further extended to filling in or deleting spots of isolated pixels, either by search-and-correct or by introducing hierarchical testing methods, which borrows strength from partial results in neighboring areas.

The maximum likelihood approach allows to estimate parameters in DTI while accounting in the correct way for signal skewness. The estimation algorithm proposed in this work overcomes some technical difficulties that kept noncentral chi estimation away from being proposed in literature and used in clinical estimation. In particular, the formulation proposed here reduces the noncentral chi case to a straightforward generalization of the likelihood equations of the Rician case. Other added difficulties, mainly related to dealing with Bessel functions, are addressed here by introducing consistent approximations. The simulation study anyway highlights a "caveat", which is that the MLE may not be the best estimator in terms of variance, in presence of relatively low SNRs and and low angular resolution designs. Moreover, the proposed algorithm should be integrated in a framework of spatial estimation with real data, by introducing proper regularization methods for tensor data. Some successful work in this field has been proposed for example in [21] for Rice-distributed data, and an analogous approach could straightforwardly be applied to the single-voxel approach proposed here. Another important issue that is not treated in this work is ensuring the positive-definiteness of estimated tensors. Different parametrizations of the DTI model may be easily used in the noncentral chi framework, starting back from equations (3.4) and (3.5). Future developments of this work should then integrate spatially-regularized tensor estimation, while considering eventually more complex diffusion models, like multi-fiber DTI or the DDI model proposed in [49].

As anticipated in Section 3.3, the noncentral chi model could also account for a spatially varying apparent number of coils, if considering the parameter N as free. Different parallel acquisition methods and different kinds of signal combination methods can modify the nature of this signal [52], which often results as a noncentral chi variable with a reduced, spatially varying number of degrees of freedom [19]. Not only this could be considered in the noise discrimination test, but also the likelihood estimation could be extended to consider the number of coils/channels not known a priori. Some successful preliminary tests have already been performed, showing some potential for this further extension.

Appendix

A.1 Plug-in estimator of individual cumulative hazard functions

Letting $(t_1^{(i)}, \ldots, t_{N_i(\tau_i)}^{(i)})$ be the jump times for patient i and setting $t_0^{(i)} = 0$ and $t_{N_i(\tau_i)+1}^{(i)} = \tau_i$, we have

$$
\begin{aligned}
\Lambda_i(t) &= \int_0^t \lambda_0(s) \alpha^{N_i(s^-)} e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)} ds \\
&= \sum_{k=0}^{N_i(t^-)} e^{k \log \alpha} \int_{t_k^{(i)}}^{t_{k+1}^{(i)} \wedge t} \lambda_0(s) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)} ds, \quad i = 1, \ldots, n,
\end{aligned}
\tag{A.1}
$$

for $t \in [0, \tau_i)$. If the covariate vector $\boldsymbol{X}_i^T$ is differentiable on each interval $[t_k^{(i)}, t_{k+1}^{(i)})$ for $i = 1, \ldots, n$ and $k = 1, \ldots, N_i(\tau_i)$, it is possible to express each Λ_i as a function of Λ_0, so that an estimate of λ_0 will not be required: defining

$$
P_{\boldsymbol{X}_i}(t) = \Lambda_0(t) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)} - \int_0^t \Lambda_0(s) \boldsymbol{\beta}^T \boldsymbol{X}_i'(s) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i(s)} ds,
\tag{A.2}
$$

where

$$
\boldsymbol{X}_i'(s) = \sum_{k=0}^{N_i(s^-)} \left(\frac{dX_{i1}(s)}{ds}, \ldots, \frac{dX_{iq}(s)}{ds} \right)^T I_{[t_k^{(i)}, t_{k+1}^{(i)})}(s),
$$

and integrating by parts (A.1) leads to the expression

$$
\Lambda_i(t) = \sum_{k=0}^{N_i(t^-)} e^{k \log \alpha} \left[P_{\boldsymbol{X}_i}(t_{k+1}^{(i)} \wedge t) - P_{\boldsymbol{X}_i}(t_k^{(i)}) \right],
\tag{A.3}
$$

55

which does not involve λ_0.

In the case study considered in section 1.3 we reconstruct individual cumulative hazard processes by letting $\mathbf{X}_i(t) = X_i(t)$, representing the age of patient i. We can express age as a variable explicitly dependent on time (in days) writing $\boldsymbol{X}_i(t) = a_i + t/365$, where a_i represents the age of patient i at the beginning of the observation period (1st January 2004). It is then possible to rewrite the subject-specific cumulative hazard functions in the form given by (A.3), with

$$P_{\boldsymbol{X}}(t) = e^{\beta a_i} \left[\Lambda_0(t) e^{\frac{\beta t}{365}} - \frac{\beta}{365} \int_0^t \Lambda_0(s) e^{\frac{\beta s}{365}} ds \right].$$

Plugging $(X_i(t), N_i(t), \widetilde{\Lambda}_0, \tau_i, \hat{\alpha}, \hat{\beta})$ for $i = 1, \ldots, n$ in (A.3) yields the reconstruction of cumulative hazard processes for all the considered patients.

Appendix

B.1 Fisher information matrix for the Rice exponential regression model

Starting from the alternative definition

$$[\mathcal{I}(\boldsymbol{\theta})]_{jk} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log L(\boldsymbol{\theta}|\sigma, \mathbf{M}, \mathbf{x})\right],$$

defining $y_i = M_i e^{\boldsymbol{\theta}^T\mathbf{x}_i}/\sigma^2$ and using the fact that $I_1'(x) = I_0(x) - I_1(x)/x$, we find the following expression for the second order derivatives of the log-likelihood:

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log L(\boldsymbol{\theta}|\sigma, \mathbf{M}, \mathbf{x}_i) =$$

$$-\frac{2}{\sigma^2}\sum_{i=1}^n x_{ij}x_{ik}e^{2\boldsymbol{\theta}^T\mathbf{x}_i} + \sum_{i=1}^n \frac{I_1(y_i)}{I_0(y_i)}y_i x_{ij}x_{ik} + \sum_{i=1}^n y_i x_{ij}\frac{\partial}{\partial\theta_k}\frac{I_1(y_i)}{I_0(y_i)} =$$

$$-\frac{2}{\sigma^2}\sum_{i=1}^n x_{ij}x_{ik}e^{2\boldsymbol{\theta}^T\mathbf{x}_i} + \sum_{i=1}^n \frac{I_1(y_i)}{I_0(y_i)}y_i x_{ij}x_{ik} + \sum_{i=1}^n y_i x_{ij}x_{ik}\left[y_i\left(1 - \frac{I_1^2(y_i)}{I_0^2(y_i)}\right) - \frac{I_1(y_i)}{I_0(y_i)}\right] =$$

$$-\frac{2}{\sigma^2}\sum_{i=1}^n x_{ij}x_{ik}e^{2\boldsymbol{\theta}^T\mathbf{x}_i} + \sum_{i=1}^n y_i^2 x_{ij}x_{ik}\left[1 - \frac{I_1^2(y_i)}{I_0^2(y_i)}\right] =$$

$$-\frac{2}{\sigma^2}\sum_{i=1}^n x_{ij}x_{ik}e^{2\boldsymbol{\theta}^T\mathbf{x}_i} + \frac{1}{\sigma^4}\sum_{i=1}^n e^{2\boldsymbol{\theta}^T\mathbf{x}_i}M_i^2 x_{ij}x_{ik}\left[1 - \frac{I_1^2(y_i)}{I_0^2(y_i)}\right].$$

Taking conditional expectation and inverting the sign of this expression, we obtain the FIM element (j, k)

$$[\mathcal{I}(\boldsymbol{\theta})]_{jk} = +\frac{2}{\sigma^2} \sum_{i=1}^{n} x_{ij} x_{ik} e^{2\boldsymbol{\theta}^T \mathbf{x}_i} - \frac{1}{\sigma^4} \sum_{i=1}^{n} e^{2\boldsymbol{\theta}^T \mathbf{x}_i} x_{ij} x_{ik} \left\{ 2\sigma^2 + e^{2\boldsymbol{\theta}^T \mathbf{x}_i} - \mathbb{E}\left[M_i^2 \frac{I_1^2(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}_i M_i}}{\sigma^2})}{I_0^2(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}_i M_i}}{\sigma^2})} \right] \right\}$$

$$= -\frac{1}{\sigma^4} \sum_{i=1}^{n} e^{2\boldsymbol{\theta}^T \mathbf{x}_i} x_{ij} x_{ik} \left\{ e^{2\boldsymbol{\theta}^T \mathbf{x}_i} - \mathbb{E}\left[M_i^2 \frac{I_1^2(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}_i M_i}}{\sigma^2})}{I_0^2(\frac{e^{\boldsymbol{\theta}^T \mathbf{x}_i M_i}}{\sigma^2})} \right] \right\}, \tag{B.1}$$

which was obtained using the fact that $\mathbb{E}[M_i^2] = 2\sigma^2 + \nu_i^2$ when $M_i \sim \text{Rice}(\nu_i, \sigma^2)$.

B.2   Numerical integration of $g(m|\nu_i, \sigma)$

This paragraph deals with the definition of a suitable interval of integration for the function $g(m|\nu_i, \sigma)$. This function is dominated, for large values, by the "quadratic" exponential decay of $\exp(-m^2/(2\sigma^2))$; in fact, modified Bessel functions $I_h(x)$ of any order h are asymptotic, for large values of x, to $e^x/\sqrt{2\pi x}$. Using this approximation in equation (2.8) we get

$$g(m|\nu_i, \sigma) \sim m^2 \frac{e^{\frac{\nu_i m}{\sigma^2}}}{\sqrt{2\pi \frac{\nu_i m}{\sigma^2}}} \frac{m}{\sigma^2} e^{-\frac{m^2 + \nu^2}{2\sigma^2}} \quad \text{for "large" } m.$$

We want to define an upper integration bound $\bar{m}$ (the lower bound is taken as 0) such that $g(m|\nu_i, \sigma) < \epsilon$ on $(\bar{m}, +\infty)$, being ϵ a very small value (we'll see that it is possible to choose ϵ as the machine epsilon).

Taking the logarithm on both sides of the inequality $g(m|\nu_i, \sigma^2) < \epsilon$ and using the approximation for Bessel functions, we get

$$-\frac{\nu_i^2}{2\sigma^2} - \log(\sigma) - \frac{1}{2}\log(2\pi\nu_i) + \frac{5}{2}\log(m) - \frac{m^2}{2\sigma^2} + \frac{\nu_i m}{\sigma^2} < \log(\epsilon).$$

To simplify the notation a bit, we define the redundant parameter $R_i = \nu_i/\sigma$ and perform the change of variable $s = m\nu_i/\sigma^2$, obtaining

$$-\frac{R_i^2}{2} - \log(\sigma) - \frac{1}{2}\log(2\pi\nu_i) + \frac{5}{2}\log\left(\frac{s\sigma^2}{\nu_i}\right) - \frac{s^2\sigma^4}{2\nu_i^2\sigma^2} + s < \log(\epsilon)$$

$$\frac{R_i^2}{2} + 4\log(\sigma) + \frac{1}{2}\log(2\pi) + 3\log(\nu_i) - \frac{5}{2}\log(s) + \frac{s^2}{2R_i^2} - s > \log\left(\frac{1}{\epsilon}\right).$$

To get an explicit formula for the upper bound $\bar{m}$ we can bound inferiorly the term $-5\log(s)/2$ with $-5(s-1)/2$, so that if the resulting quadratic function exceeds

$\log(1/\epsilon)$ the original one will exceed it too. This way we get

$$\frac{R_i^2}{2} + 4\log(\sigma) + \frac{1}{2}\log(2\pi) - 3\log(\nu_i) - \frac{5}{2}(s-1) + \frac{s^2}{2R_i^2} - s - \log\left(\frac{1}{\epsilon}\right) > 0$$

$$\frac{R_i^2}{2} + 4\log(\sigma) + \frac{1}{2}\log(2\pi) - 3\log(\nu_i) + \frac{5}{2} - \frac{7}{2}s + \frac{s^2}{2R_i^2} - \log\left(\frac{1}{\epsilon}\right) > 0$$

$$\frac{s^2}{2R_i^2} - \frac{7}{2}s + \left\{\frac{5}{2} - \log\left(\frac{1}{\epsilon}\right) + \frac{R_i^2}{2} - 3\log(R_i) + \log(\sigma) + \frac{1}{2}\log(2\pi)\right\} > 0.$$

$$\text{(B.2)}$$

Defining as $c_i(\epsilon)$ the part of equation (B.2) that does not depend on s, it's clear that we just have to find the largest root of a quadratic polynomial in s; so the bound should be

$$\bar{m} = \frac{\sigma^2}{\nu}\left\{\frac{7R_i^2}{2} + R^2\sqrt{\frac{49}{4} - \frac{2c_i(\epsilon)}{R^2}}\right\}.$$

As anticipated, ϵ can be chosen, for example, as the 64-bit machine epsilon: in `R` we get a `.Machine$double.eps` equal to 2^{-52}, but in the final expression we have a dependence on ϵ of the form $\log\left(\frac{1}{\epsilon}\right) = \log(2^{52}) = 52\log(2) \simeq 36$, which is an "easy" number!

B.3 Non-exponentiality of the noncentral chi family

A possible approach to prove that the noncentral chi family of distributions is not an exponential one is to show that the class of functions $\{\ln(I_0(\eta x))\}_{\eta \in \mathbb{R}^+}$ is linearly independent. To do this, we show that the difference between two elements with parameters $\eta_1, \eta_2 > 0$, $\eta_1 \neq \eta_2$ does not remain within the same family. If the contrary was true, then we would have

$$\ln(I_0(\eta_3 x)) = \ln(I_0(\eta_1 x)) - \ln(I_0(\eta_2 x)),$$

which would mean, exponentiating both sides of the equation, that $I_0(\eta_1 x) = I_0(\eta_2 x)I_0(\eta_3 x)$. This is an equation between to power series, that must have equal coefficients for each order to yield the same result. Comparing the coefficients of the terms of order 2 and 4, we get the system of equations

$$\begin{cases} \frac{\eta_1^2}{4} = \frac{\eta_2^2 + \eta_3^2}{4} \\ \frac{\eta_1^4}{64} = \frac{\eta_2^4 + \eta_3^4 + 4\eta_2^2\eta_3^2}{64}, \end{cases}$$

which admit a solution only if $\eta_2 = 0$ or $\eta_3 = 0$. The considered family is then linearly independent.

Bibliography

[1] Odd O Aalen, Ø rnulf Borgan, and Håkon K Gjessing. *Survival and event history analysis: a process point of view*. Springer-Verlag New York, 2006.

[2] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publications, 1964.

[3] Santiago Aja-Fernández, Antonio Tristán-Vega, and Carlos Alberola-López. Noise estimation in single- and multiple-coil magnetic resonance data based on statistical models. *Magnetic resonance imaging*, 27(10):1397–409, 2009.

[4] Per Kragh Andersen, Ø rnulf Borgan, Richard D Gill, and Niels Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag New York, 1993.

[5] Jesper L R Andersson. Maximum a posteriori estimation of diffusion tensor parameters using a Rician noise model: Why, how and but. *NeuroImage*, 42(4):1340–1356, 2008.

[6] Stefano Baraldo, Francesca Ieva, Luca Mainardi, and Anna Maria Paganoni. Adaptive Design of Experiment for the Estimation of Apparent Diffusion Coefficients in { MRI}. In *Proceedings of the 7th Conference on Statistical Computation and Complex Systems (SCo 2011), Padova, September 19-21*, 2011.

[7] Stefano Baraldo, Francesca Ieva, Anna Maria Paganoni, and Valeria Vitelli. Outcome Prediction for Heart Failure Telemonitoring Via Generalized Linear Models with Functional Covariates. *to appear on the Scandinavian Journal of Statistics*, 2012.

[8] Edward W Barankin and Ashok P Maitra. Generalization of the Fisher-Darmois-Koopman-Pitman Theorem on Sufficient Statistics. *Sankhyā: The Indian Journal of Statistics*, 25(3):217–244, 1963.

[9] Peter J Basser. Inferring Microstructural Features and the Physiological State of Tissues from Diffusion-Weighted Images. *NMR in Biomedicine*, 8:333–334, 1995.

[10] Peter J. Basser, James Mattiello, and Denis LeBihan. MR diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–67, 1994.

[11] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[12] P J Bickel and K A Doksum. *Mathematical Statistics*. Prentice Hall, 2006.

[13] Soccorso Capomolla, Giandomenico Pinna, Maria Teresa La Rovere, Roberto Maestri, Monica Ceresa, Marina Ferrari, Oreste Febo, Angelo Caporotondi, Giampaolo Guazzotti, Francesca Lenta, Sonia Baldin, Andrea Mortara, and Franco Cobelli. Heart Failure Case Disease Management Program: a Pilot Study of Home Telemonitoring Versus Usual Care. *European Heart Journal Supplements*, 6 (Supplem:F91–F98, 2004.

[14] CEFRIEL. `http://ftp.cefriel.it/nrs/`, 2010.

[15] Thomas L Chenevert, Craig J Galbán, Marko K Ivancevic, Susan E Rohrer, Frank J Londy, Thomas C Kwee, Charles R Meyer, Timothy D Johnson, Alnawaz Rehemtulla, and Brian D Ross. Diffusion coefficient measurement using a temperature-controlled fluid for quality control in multicenter studies. *Journal of magnetic resonance imaging*, 34:983–987, 2011.

Bibliography

[16] Chris D. Constantinides, Ergin Atalar, and Elliot R. McVeigh. Signal-to-noise measurements in magnitude images from nmr phased arrays. *Magnetic Resonance in Medicine*, 38(5):852–857, 1997.

[17] David R Cox. Regression Models and Life Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[18] John I Crowell. Nonparametric estimation of a process mean from censored data. *Statistics & Probability Letters*, 15:253–257, 1992.

[19] Olaf Dietrich, José G. Raya, Scott B. Reeder, Michael Ingrisch, Maximilian F. Reiser, and Stefan O. Schoenberg. Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics. *Magnetic resonance imaging*, 26(6):754–762, 2008.

[20] Frederic Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis*. Springer-Verlag New York, 2006.

[21] Pierre Fillard, Xavier Pennec, Vincent Arsigny, and Nicholas Ayache. Clinical DT-MRI estimation, smoothing, and fiber tracking with log-Euclidean metrics. *IEEE transactions on medical imaging*, 26(11):1472–1482, 2007.

[22] W. Gautschi and J. Slavik. On the computation of modified Bessel function ratios. *Mathematics of Computation*, 32(143):865–875, 1978.

[23] W R Gilks, W R Thomas, and D J Spiegelhalter. A language and program for complex Bayesian modelling. *The Statistician*, 43(1):169–177, 1983.

[24] A Giordano, S Scalvini, E Zanelli, U Corrá, G L Longobardi, V A Ricci, P Baiardi, and F Glisenti. Multicenter randomised trial on home-based telemanagement to prevent hospital readmission of patients with chronic heart failure. *International Journal of Cardiology*, 131(2):192–199, 2008.

[25] Amerigo Giordano, Simonetta Scalvini, Anna Maria Paganoni, Stefano Baraldo, Maria Frigerio, Claudia Vittori, Gabriella Borghi, Maurizio Marzegalli, and Ornella Agostoni. Home-based telesurveillance programme in patients with heart failure: effects on clinical status and implications for one-year prognosis. The experience of Lombardy region with "Nuove Reti Sanitarie". *To appear on Telemedicine and e-Health*, 2013.

[26] J R González, E H Slate, and Edsel A Peña. *gcmrec: General Class of Models for Recurrent Event Data*, 2009.

[27] Juan R González, Edsel A Peña, and Elizabeth H Slate. Modelling intervention effects after cancer relapses. *Statist. Med.*, 24:3959–3975, 2005.

[28] Juan R. González, Virginie Rondeau, and Yassin Mazroui. *frailtypack: Frailty models using maximum penalized likelihood estimation*, 2010. Avaliable at: `http://cran.r-project.org/web/packages/frailtypack/index.html`.

[29] Guozhong He, Hans-Georg Müller, Jane-Ling Wang, and Yang Wenjing. Functional linear regression via canonical analysis. *Bernoulli*, 16(3):705–729, 2010.

[30] Xuming He and Pin Ng. COBS: Qualitatively Constrained Smoothing via Linear Programming. *Computational Statistics*, 14:315–337, 1999.

[31] John D Kalbfleisch and Ross L Prentice. *The Statistical Analysis of Failure Data*. John Wiley & Sons, New York, 1980.

[32] Cheng Guan Koay and Peter J Basser. Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. *Journal of Magnetic Resonance*, 179:317–322, 2006.

[33] Landini Luigi, Vincenzo Positano, and Maria Filomena Santarelli, editors. *Advanced Image Processing in Magnetic Resonance Imaging*. CRC Press, 2005.

[34] B Landman, P.-L. Bazin, and J Prince. Diffusion Tensor Estimation by Maximizing Rician Likelihood. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007.

[35] Michael Lee. BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods*, 40(3):450–456, 2008.

[36] Carl N Morris. Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, 11(2):515–529, 1983.

[37] Hans-Georg Müller and Ulrich Stadtmüller. Generalized Functional Linear Models. *The Annals of Statistics*, 33(2):774–805, 2005.

[38] Pin T. Ng and Martin Maechler. *cobs: COBS – constrained B-splines (sparse matrix based)*, 2009. Available at: `http://cran.r-project.org/web/packages/cobs/index.html`.

[39] George Nocedal and Stephen J Wright. *Numerical Optimization*. Springer-Verlag New York, 1999.

[40] Edsel A Peña and M Hollander. Models for recurrent events in reliability and survival analysis. In T Mazzuchi R. Soyer and N Singpurwalla, editors, *Mathematical Reliability: An Expository Perspective*, chapter 6, pages 493–514. Kluwer Academic Publishers, Dordrecht, 2004.

[41] Edsel A Peña, E H Slate, and J R González. Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference*, 137(6):1727–1747, 2007.

[42] Martyn Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria, 2003.

[43] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009.

[44] James O Ramsay and Bernard W Silverman. *Functional Data Analysis*. Springer Science+Business Media, 2005.

[45] Virginie Rondeau, Daniel Commenges, and Pierre Joly. Maximum Penalized Likelihood Estimation in a Gamma-Frailty Model. *Lifetime Data Analysis*, 9:139–153, 2003.

[46] Chiho Sato, Shinji Naganawa, Tatsuya Nakamura, Hisashi Kumada, Shunichi Miura, Osamu Takizawa, and Takeo Ishigaki. Differentiation of noncancerous tissue and cancer lesions by apparent diffusion coefficient values in transition and peripheral zones of the prostate. *Journal of magnetic resonance imaging : JMRI*, 21(3):258–62, March 2005.

[47] Simonetta Scalvini, Emanuela Zanelli, Maurizio Volterrani, Giovanna Martinelli, Doriana Baratti, Osvaldo Buscaya, Paola Baiardi, Fulvio Glisenti, and Amerigo Giordano. A Pilot Study of Nurse-Led, Home-Based Tele cardiology for Patients with Chronic Heart Failure. *Journal of Telemedicine and Telecare*, 10:113–117, 2004.

[48] Jan Sijbers, Arnold J den Dekker, Paul Scheunders, and Dirk Van Dyck. Maximum Likelihood estimation of Rician distribution parameters. *IEEE Transactions on Medical Imaging*, 17:357–361, 1998.

[49] Aymeric Stamm, Patrick Pérez, and Christian Barillot. Diffusion Directions Imaging (DDI). Rapport de recherche, Computational Medicine and Neurosciences July, INRIA, 2011.

[50] Thaddeus Tarpey and Kimberly K J Kinateder. Clustering Functional Data. *Journal of Classification*, 20:93–114, 2003.

[51] Simon Walker-Samuel, Matthew Orton, Lesley D McPhail, and Simon P Robinson. Robust estimation of the apparent diffusion coefficient ({ADC}) in heterogeneous solid tumors. *Magnetic Resonance in Medicine*, 62(2):420–429, 2009.

[52] David O. Walsh, Arthur F. Gmitro, and Michael W. Marcellin. Adaptive reconstruction of phased array MR imagery. *Magnetic Resonance in Medicine*, 43(5):682–690, 2000.

[53] Reiko MD Woodhams, Keiji MD Matsunaga, Keiichi MD Iwabuchi, Shinichi MD Kan, Hirofumi RT Hata, Masaru MD Kuranami, Masahiko MD Watanabe, and Kazushige MD Hayakawa. Diffusion-Weighted Imaging of Malignant Breast Tumors: The Usefulness of Apparent Diffusion Coefficient (ADC) Value and ADC Map for the Detection of Malignant Breast Tumors and Evaluation of Cancer Extension. *Journal of Computer Assisted Tomography*, 29(5):644–649, 2005.