ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

REVSTAT

Statistical Journal

Special issue on "Statistical Modelling: Challenges in Health"

Guest Editors:
Lisete Sousa
Valeska Andreozzi
Marília Antunes
Luzia Gonçalves

Volume 9, No.1
March 2011

REVSTAT

Statistical Journal

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E., 2003- . - 30 cm
Semestral. - Continuação de : Revista de Estatística =
ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726

CREDITS

- **PUBLISHER**
  - *Instituto Nacional de Estatística, I.P. (INE, I.P.)*
  Av. António José de Almeida, 2
  1000-043 LISBOA
  PORTUGAL
  Tel.: +351 218 426 100
  Fax: +351 218 426 364
  Web site: http://www.ine.pt
  Customer Support Service
  (National network): 808 201 808
  (Other networks): +351 226 050 748
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE, I.P., by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística, I.P.*
- **EDITION**
  - *300 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*

PRICE

[VAT included]

- Single issue …….. **€ 8**
- Annual subscription (No. 1 Special Issue, No. 2 and No.3)……. **€ 19**
- Annual subscription (No. 2, No. 3) …….. **€ 13**

FOREWORD

This special issue of REVSTAT – *Statistical Journal* contains a selection of invited papers presented at the *Workshop StaM2010 — Statistical Modelling: Challenges in Health* that took place in Parque das Nações — Lisbon, Portugal, from 9 to 12 May 2010.

The workshop was organized under the auspices of the Center of Statistics and Applications, University of Lisbon (CEAUL, `http://www.ceaul.fc.ul.pt`), and the FCT/MCTES research projects: "Statistical Methods in Genetics and Environment" (PTDC/MAT/64353/2006), "Latent Class Models in Tropical Health" (PTDC/SAU-ESA/81240/2006). It brought together researchers interested in advanced statistical applications in Health challenging problems, to promote knowledge and experience exchange and also to encourage cooperation between the participants. A fruitful discussion on the role of statistical modelling in Health with maximum participation in non-parallel sessions included the following areas: health spatial problems; survival analysis; genetics; molecular biology; bioinformatics; latent class models in health.

The six papers in this volume illustrate some of the statistical problems currently of interest in Health. Jow et al. develop a method for estimating the density of BLAST hits across chromosomes on a target genome, which is used to identify genes associated with QTLs from Bovine Hemoglobine Genome by using the Human genome. Anticoagulants are one of the most prescribed groups of drugs. This is the motivation for Henderson et al. to review and develop methods to optimal dynamic treatment regime determination. Teixeira-Pinto and Normand develop likelihood and quasi-likelihood methods to analyse multiple non-commensurate outcomes in the presence of missing data in biomedical researches. Sousa reviews different methods for joint modelling of longitudinal and time to event data, based on the full likelihood of the joint distribution of the two processes. This special issue also includes a review of several nonparametric approaches for non-Markov multi-state survival models. Meira-Machado illustrates his contributions on that topic with well-known real data sets using three-state models. Bailey and Hewson suggest an addition to the multivariate modelling of the geographical distribution of different but potentially related diseases, which incorporates a discrete mixture of latent factors. This is illustrated on data on four carcinomas in some UK geographical areas.

Thanks are due to Maria Antónia Amaral Turkman, Ana Luísa Papoila and Giovani Silva who helped to organize the invited programme. We also express our gratitude to the speakers and the authors of the posters for their valuable contribution to the high scientific standards of the *Workshop Statistical Modelling: Challenges in Health.*

Lisete Sousa
Valeska Andreozzi
Marília Antunes
Luzia Gonçalves

INDEX

Abstracted/indexed in: *Current Index to Statistics, DOAJ, Google Scholar, Journal Citation Reports/Science Edition, Mathematical Reviews, Science Citation Index Expanded®, SCOPUS* and *Zentralblatt für Mathematic.*

REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 1–17

A COMPARATIVE GENOMICS APPROACH TO THE IDENTIFICATION OF QTL CANDIDATE GENES

Authors: Howsun Jow
– School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne U.K.
howsun.jow@ncl.ac.uk

Richard J. Boys
– School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne U.K.
richard.boys@ncl.ac.uk

Darren J. Wilkinson
– School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne U.K.
d.j.wilkinson@ncl.ac.uk

Abstract:

- Despite rapid advances in sequencing technology, many commercially relevant species remain unsequenced, and many that *are* sequenced have very poorly annotated genomes. There is therefore still considerable interest in using comparative approaches to exploit information from well-characterised model organisms in order to better understand related species. This paper develops a statistical method for automating part of a comparative genomics bioinformatic pipeline for the identification of genes and genomic regions in a model organism associated with a QTL region in an unsequenced species. A non-parametric Bayesian statistical model is used for characterising the density of a large number of BLAST hits across a model species genome. The method is illustrated using a test problem demonstrating that markers associated with Bovine hemoglobin can be automatically mapped to a region of the human genome containing human hemoglobin genes. Consequently, by exploiting the (relatively) high quality of genome annotation for model organisms and humans it is possible to quickly identify candidate genes in those well-characterised genomes relevant to the quantitative trait of interest.

Key-Words:

- *Bayesian; non-parametric; density estimation; QTL; BLAST; mapping; comparative genomics.*

AMS Subject Classification:

- 62F15, 62G07, 92D99.

1. INTRODUCTION

The mapping of the genetic component influencing quantitative traits of a species, such as height and weight, can be achieved even in the absence of a complete physical map of a species' genome. This is called quantitative trait loci (QTL) mapping. One method by which QTLs can be mapped utilizes a map of typed genetic markers in order to establish the statistical correlation between a given quantitative trait and a given point, between two markers, on the genetic map ([12]). This allows for the identification of regions which are highly statistically correlated with the quantitative trait and therefore likely to contain a QTL. These regions can then be sequenced and the genes influencing the quantitative trait can be identified.

This method of finding the genes that influence a particular quantitative trait has its drawbacks. For one thing it is dependent on the quality and resolution of the genetic map used to map the QTLs. A low resolution genetic map would lead to a low resolution QTL map in which relatively large regions are identified as being statistically significant and therefore likely to contain a QTL. This in turn requires the sequencing of large portions of the sequence genome. Alternatively the method can be used on high resolution genetic maps. However, this too has problems: constructing high resolution genetic maps is far from a trivial process and can be expensive and labour intensive, especially for traditional linkage maps.

The method described in this paper uses a comparative genomics approach to locate genes which are correlated with the QTL. It works by first identifying statistically significant QTL regions. Then a high resolution map is constructed by integrating available partial maps of the chromosome in which the QTL regions lie into a single map. There are a number of methods available for integrating partial genetic maps ([14, 18, 16, 19, 6, 13, 11]) and in this paper we use a Bayesian approach to map integration developed by Jow *et al.* ([11]).

On obtaining a high resolution integrated map, the markers lying between the QTL flanking markers are identified and a BLAST ([1]) search made of their sequences against the genome of a target species. This gives us a series of "hits" on the target genome, that is, locations where the search sequences match. Using these hits it is possible to estimate the probability density of hits across the target genome using, for example, standard kernel density techniques ([17]) or Bayesian alternatives based on *Dirichlet processes* ([4, 2, 3]). We will use a Bayesian density estimate and then threshold this density to identify regions along the target species which are likely to contain genes performing similar functions to the genes associated with the QTL of the source species.

The rest of this paper is organised as follows. Section 2.1 describes how to construct a Bayesian density estimate from a collection of BLAST hits across

a number of chromosomes. The model is described in detail, with the resulting MCMC algorithm available in the appendix. This section also describes a procedure for determining the location of regions likely to contain genes associated with the QTL. Section 3 validates the MCMC algorithm and implementation on a synthetic example and Section 4 provides a real example of how our method can be used to help identify genes associated with QTLs obtained from the Bovine Hemoglobin genome by using the Human genome.

2. METHODS

In this section we describe how to construct a Bayesian density estimate from a collection of BLAST hits and the procedure for determining the location of intervals likely to contain genes associated with the QTL.

2.1. Bayesian density estimation

Suppose that the target genome consists of C chromosomes with lengths $L_1, ..., L_C$. The data take the form of n BLAST hits describing the location (y) and chromosome (c) on which each hit was made: (y_i, c_i), $i = 1, ..., n$. Let n_c be the number of observed hits on chromosome c. We construct the Bayesian density estimate by modelling these locations as an infinite mixture of normal distributions with unknown means (μ) and variances (σ^2) and with these parameters $\phi = (\mu, \sigma^2)$ resulting from a Dirichlet process with a particular base distribution. Let θ_c denote the probability of a hit occurring on chromosome c. The formulation of the model is slightly complicated by the need to have a continuous density across the C chromosomes. In summary we have, for $i = 1, ..., n$ and $c_i \in \{1, ..., C\}$,

$$\begin{aligned}
\underline{\theta} = (\theta_1, ..., \theta_C)|\alpha &\sim Dir(\alpha \underline{\ell}) \ , \\
Y_i, c_i|\phi_{ic_i}, \theta_{c_i} &\sim N(\mu_{ic_i}, \sigma^2_{ic_i}) \times \text{Bern}(\theta_{c_i}) \ , \\
\phi_{ic_i}|G_{c_i} &\sim G_{c_i} \ , \\
G_{c_i}|\alpha &\sim DP(\alpha, G_{0c_i}) \ , \\
G_{0c_i} &= U(0, L_{c_i}) \times \text{Inv}\,\Gamma(a, L^2_{c_i} b) \ ,
\end{aligned}$$

where $DP(\alpha, G_0)$ denotes a Dirichlet process with base measure G_0 and concentration parameter α, and $\underline{\ell}$ is the normalized form of $\underline{L}$, that is, $\ell_c = L_c / \sum_{j=1}^{C} L_j$. The form of the base distribution has been chosen so that it is independent of the scale used to measure the location of the BLAST hits, for example, Mb or b.

All that remains for a full model specification is to choose the prior distribution for α. In this paper we take a flexible semi-conjugate form with

$$\alpha \sim \Gamma(g, h) \ .$$

Given the data model above and the hit data, it is generally not possible to derive an analytical expression for the probability density at an arbitrary point y on an unknown chromosome k. However, numerical sampling methods can be used approximate this (predictive) density as

$$\pi(y, k|\mathcal{D}) \simeq \frac{1}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} \pi\big(y, k|\phi_{ik}^t, \theta_k^t\big) \ , \tag{2.1}$$

where $(\phi_{ik}^t, \theta_k^t)$, $t = 1, ..., T$, is a sample from the posterior distribution $\pi(\phi_{ik}, \theta_k|\mathcal{D})$ obtained using an appropriate sampling algorithm. In this paper we have used an MCMC algorithm based on one by Escobar and West ([3]); the algorithm is described in the appendix.

2.2. Identification of QTL intervals

On obtaining the probability density of hits across the entire target genome the remaining task is to identify regions with a high probability density. This is done by identifying the highest density regions (HDRs) containing a given percentage of the density; see [10]. For example, a 75% HDR could be found across all the chromosomes. Given that in our model the target genome is one-dimensional, the HDR would be a set of regions across all the chromosomes. These regions can then be searched for genes of interest.

3. SIMULATED DATA

To validate our MCMC algorithm and implementation, we simulated a dataset of 200 "hits" spread over three chromosomes with lengths 100 Mb, 200 Mb and 400 Mb. The distribution of the locations of hits on the different chromosomes were taken to be normal distributions on chromosomes 1 and 3 and a mixture of two normal distributions on chromosome 2; see the dashed lines in Figure 1. Also the probability of a hit being located on a particular chromosome was taken as being proportional to the length of the chromosome, that is, with probability 1/7, 2/7 and 4/7 for chromosomes 1, 2 and 3 respectively.

We specify the base distribution for the cluster variances (σ^2) by taking $a = 2.05$ and $b = 0.000105$, so that $E(\sigma^2) = 10^{-4}L^2$ and $SD(\sigma^2) = \sqrt{2}\,E(\sigma^2)$.

For example, on chromosome 1 this gives $E(\sigma^2) = (1\,\text{Mb})^2$, that is, suggests cluster standard deviations are around 1 Mb. We also input fairly weak prior information for α by taking $g = 4$ and $h = 2$, that is, $E(\alpha) = 2$ and $SD(\alpha) = 1$.

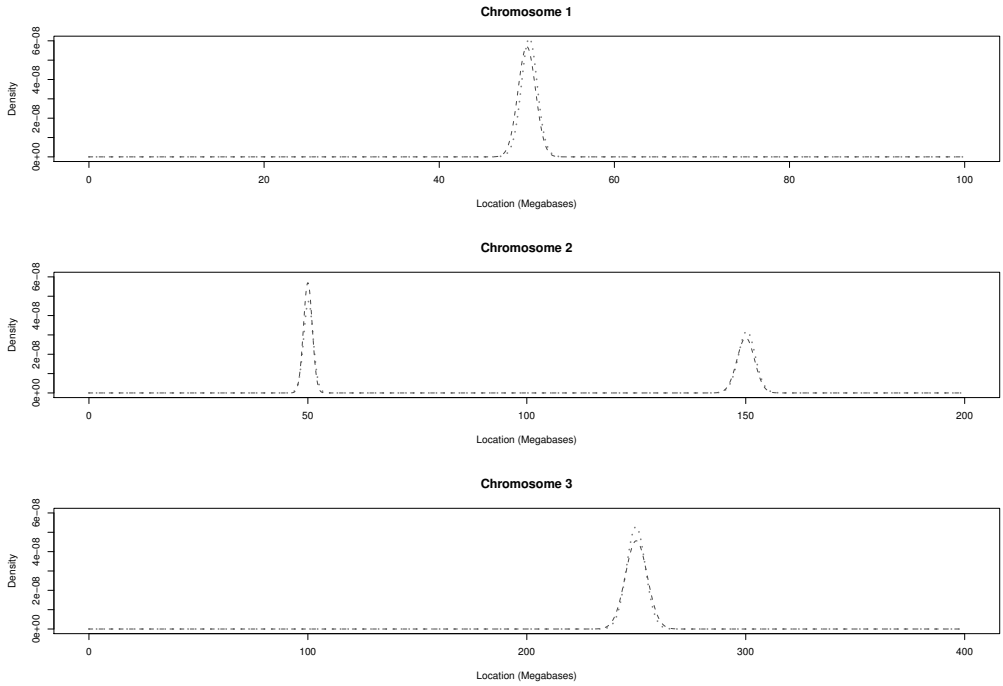

Figure 1: The theoretical distribution of hits along the chromosomes (dashed lines) and the Bayesian density estimate obtained from the simulated data (dotted line).

3.1. Results

The MCMC algorithm outlined in the appendix was applied to the simulated dataset. Convergence was assessed by using informal visual methods and the diagnostics suggested by Gelman and Rubin ([5]) and by Heidelberger and Welch ([7]). We found that a burn-in of $100K$ iterations was required to achieve convergence and we then ran the chain for a further $100K$ iterations, thinning the output by taking every 100^{th} iterate. This gave a posterior sample of size $1K$ observations from which we could calculate the Bayesian density estimate (2.1) across the (simulated) chromosomes. The results are summarized in in Table 1 and Figure 1, and show that there is a reasonably close match between the theoretical and estimated probabilities of a hit being found on a particular chromosome and between the Bayesian density estimate for the location of hits and their generating distribution.

Table 1: Probability of a hit being located on each (simulated) chromosome (to 3 *d.p.*).

Chromosome	Probability
1	0.142
2	0.285
3	0.573

4. BOVINE HEMOGLOBIN MARKER DATA

To illustrate the power of our method, we now show how the Human genome can be used to help identify genes associated with QTLs obtained from the Bovine Hemoglobin genome. The sequences of molecular markers associated with Bovine Hemoglobin genes were taken from the NCBI "GENE" database ([15]) and the markers we use are given in Table 2. For our analysis, we use the same input parameters (a, b, g and h) as in Section 3.

Table 2: Markers associated with Bovine Hemoglobin genes.

Marker name	Associated gene	Gene symbol	Sequence length
REN97351	Hemoglobin Beta	HBB	248
RH69634	Hemoglobin Beta	HBB	141
PMC115301P1	Hemoglobin Beta	HBB	136
GDB:178694	Hemoglobin Beta	HBB	300
HBB	Hemoglobin Gamma	HBG	171
PMC86017P3	Hemoglobin Gamma	HBG	267
PMC21968P1	Hemoglobin Epsilon	HBE	989
Hba-a1	Hemoglobin Alpha	HBA	188
AW312144	Hemoglobin Alpha	HBA	327
CB603723	Hemoglobin Zeta	HBZ	312
BE749596	Hemoglobin Theta 1	HBQ	277
AW428039.1	Hemoglobin Mu	HBM	193

4.1. Results

A BLAST search of these markers was conducted against the reference Human genome (NCBI 36.3 build) using the parameters listed in Table 3, and gave 188 hits distributed across 15 chromosomes. The MCMC algorithm was

then run on these hit data. As with the simulated data, convergence was assessed by using informal visual methods and standard diagnostics tools. Again, we found that a burn-in of $100K$ iterations was required to achieve convergence.

Table 3: BLAST search parameters against the Human genome for the Bovine Hemoglobin markers.

BLAST parameter	argument	value
Expectation Value	−e	0.1
Gap Cost	−G	5
Gap Extension Cost	−E	2
Nucleotide Mismatch Cost	−q	3
Nucleotide Match reward	−r	2

We then ran the chain for a further $100K$ iterations, thinning the output by taking every 100^{th} iterate, to obtain a posterior sample of size $1K$ observations. The results are summarized in Tables 4, 5 and in Figure 2. The posterior probability of a hit being on the target human chromosomes is shown in Table 4.

Table 4: Probability of a hit being located on each chromosome of the Human genome (to 3 *d.p.*).

Chromosome	Probability
1	0.032
3	0.022
5	0.016
6	0.005
7	0.011
9	0.032
11	0.620
12	0.016
13	0.011
14	0.005
15	0.005
16	0.161
17	0.027
19	0.021
20	0.016
2, 4, 8, 10, 18, 21, 22, X, Y	$\simeq 0$

Table 5 contains the 50%, 60% and 75% highest density regions (HDRs) across all chromosomes, calculated using the method of Hyndman ([10]). Figure 2 gives a graphical view of the HDRs for those chromosomes with a hit probability of at least 5%, that is, for chromosomes 11 and 16. The 50% and 60% HDRs

determined over all chromosomes point to genes of interest only on chromosomes 11 and 16. The aim of our method in this example is to identify regions on the human genome which are associated with the Bovine Hemoglobin genome.

Table 5: HDR intervals on the Human genome for the Bovine Hemoglobin markers.

HDR level	Chromosome	Intervals	Number of candidate genes
50%	11	4.59 Mb – 5.77 Mb	95
	16	0.00 Mb – 0.36 Mb	27
60%	11	4.37 Mb – 5.98 Mb	104
	16	0.00 Mb – 0.49 Mb	33
75%	11	3.88 Mb – 6.47 Mb	131
	11	54.52 Mb – 58.11 Mb	121
	16	0.00 Mb – 0.78 Mb	59
	9	124.28 Mb – 124.53 Mb	13
	17	2.51 Mb – 3.72 Mb	35
	20	61.14 Mb – 61.92 Mb	30

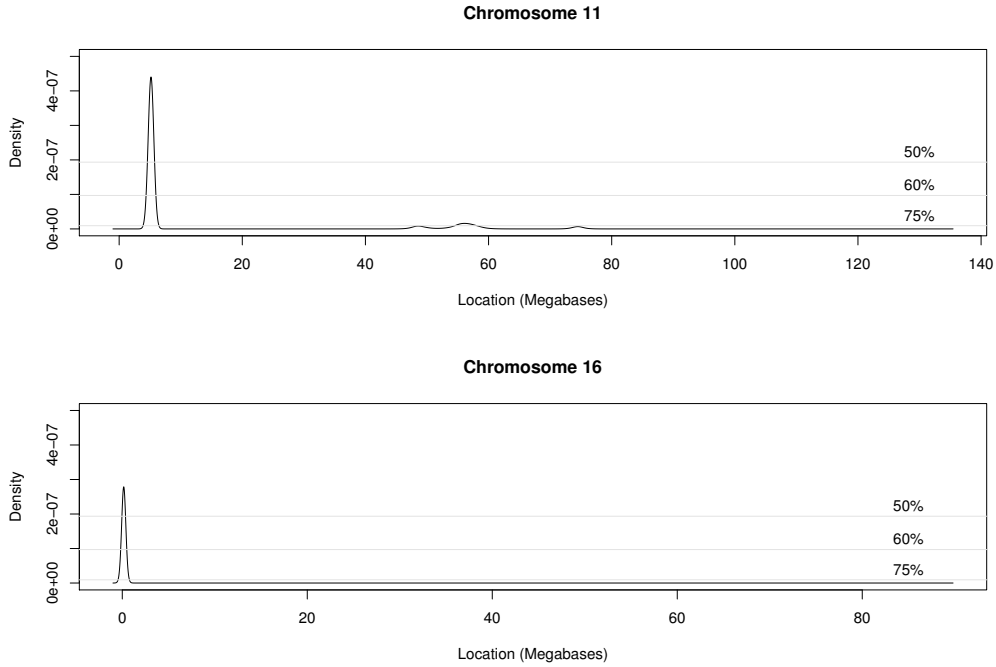

Figure 2: Figure showing the Bayesian density estimate of BLAST hits across chromosomes 11 and 16 of the Human genome.

If we look in detail at the Human genome, its Hemoglobin genes are located in two clusters on chromosomes 11 and 16, with the β-globin cluster spanning an interval of roughly 5.20–5.25 Mb on chromosome 11 and the α-globin cluster span-

ning an interval of roughly 0.14–0.17 Mb on chromosome 16. Thus our method has correctly and reasonably accurately identified the appropriate regions on the Human genome.

If we examine the Human genes found within these 60% HDRs, we find the 33 genes located on chromosome 16 listed in Table 6. These include the five known functional genes and two pseudo-genes of the human α-globin locus.

Table 6: A list of genes in the 60% HDR for chromosome 16. The genes in bold are the 5 known functional genes present in the Human α-globin locus and those in italics are the two known pseudo-genes ([9]).

Ensembl Gene ID	Gene name
ENSG00000220481	Z84812.3
ENSG00000181404	WASH4P
ENSG00000219509	Z84723.2
ENSG00000185203	Z84723.1
ENSG00000161980	POLR3K
ENSG00000161981	C16orf33
ENSG00000007384	RHBDF1
ENSG00000103152	MPG
ENSG00000103148	C16orf35
ENSG00000130656	**HBZ**
ENSG00000206178	*Z84721.1*
ENSG00000206177	**HBM**
ENSG00000218072	*Z84721.4*
ENSG00000188536	**HBA2**
ENSG00000206172	**HBA1**
ENSG00000207243	Y_RNA
ENSG00000086506	**HBQ1**
ENSG00000007392	LUC7L
ENSG00000206168	Z69890.1
ENSG00000167930	ITFG3
ENSG00000215289	AC004754.1
ENSG00000076344	RGS11
ENSG00000206156	ARHGDIG
ENSG00000185615	PDIA2
ENSG00000103126	AXIN1
ENSG00000086504	MRPL28
ENSG00000129925	TMEM8
ENSG00000216963	Z97634.3
ENSG00000103200	NME4
ENSG00000103202	DECR2
ENSG00000090565	RAB11FIP3
ENSG00000201034	Y_RNA
ENSG00000217816	RP1-196A12.1

Additionally, C16orf35 is known to be involved in the regulation of α-globin. The corresponding list for chromosome 11 contains 104 genes and includes the five

known functional Hemoglobin genes in the human β-globin locus (HBE1, HBG1, HBG2, HBD and HBB) and one known hemoglobin pseudogene (HBBP1). The annotations available for the remaining genes in these lists show no known direct link to Hemoglobin or its regulation.

5. CONCLUSIONS

In this paper we have developed a method for estimating the density of BLAST hits across chromosomes on a target genome. This estimate can then be used to determine highest density regions (HDRs) on the target genome for genes associated with the QTL of interest.

The method has been shown to work well on both simulated data and real data. In this latter case this involved obtaining BLAST hits for a number of Bovine Hemoglobin markers (given in Table 2) against the Human genome. We were able to construct the density estimate of BLAST hits across the Human genome and thereby determine the highest density regions. The regions obtained were found to contain the Human α-globin and β-globin loci ([8]).

Currently our method uses a fairly superficial treatment of BLAST hits and does not, for example, distinguish between poor BLAST hits and good ones. Future work might involve exploring how to incorporate properly weighted BLAST hits so that the better hits contribute more to the density estimate and this might lead to more accurate HDRs. Also, because the chromosomes have finite length, strictly the density across the chromosomes should have finite support. This could be achieved, for example, by replacing the Gaussian distribution for the location of clusters by (a mixture of) truncated Gaussian distributions. Unfortunately, such a modification does lead to analytical intractability in the calculations underpinning the Bayesian density estimate, though research into using such distributions is also a possible area of future work.

APPENDIX

The MCMC algorithm is a Gibbs sampler for the cluster parameters $\phi_{ic_i} = \left(\mu_{ic_i}, \sigma^2_{ic_i}\right)$, $i = 1, ..., n$, and the parameters $(\alpha, \underline{\theta})$. In the following sections, we derive the posterior conditional distributions for these parameters.

A. The cluster parameters

Here we derive the posterior conditional distributions for the $\phi_{ic_i}=(\mu_{ic_i},\sigma^2_{ic_i})$, $i=1,...,n$. Letting $\phi'_{ic_i}=\{\phi_{jc_i}\colon j\neq i\}$, the conditional prior density for $\phi_{ic_i}|\phi'_{ic_i}$ is

$$\pi\big(\phi_{ic_i}|\phi'_{ic_i}\big) \;=\; \frac{\alpha}{\alpha+n_{c_i}-1}\; g_{0c_i}(\phi_{ic_i}) + \sum_{j\neq i}\frac{1}{\alpha+n_{c_i}-1}\;\delta_{\phi_{jc_i}}(\phi_{ic_i})$$

where g_{0c_i} is the probability density corresponding to the distribution G_{0c_i}, n_{c_i} is the number of observed hits on chromosome c_i and $\delta_y(x)$ is Dirac's delta function ($\delta_y(x)=0$ if $x\neq y$ and $\int\delta_y(x)\,dx=1$).

Multiplying this by the likelihood $\pi(y_i,c_i|\phi_{ic_i},\theta_{c_i})$, we get the conditional posterior density

$$\pi\big(\phi_{ic_i}|\phi'_{ic_i},y_i,\theta_{c_i}\big) \;=\; q_{i0}\,g_{ic_i}(\phi_{ic_i}) + \sum_{j\neq i} q_{ij}\,\delta_{\phi_{jc_i}}(\phi_{ic_i})$$

where

$$\begin{aligned}
q_{ij} \;&=\; \kappa\,\pi(y_i,c_i|\phi_{jc_i},\theta_{c_i})\;,\\
q_{i0} \;&=\; \kappa\,\alpha\int\pi(y_i,c_i|\phi_{ic_i},\theta_{c_i})\,g_{0c_i}(\phi_{ic_i})\,d\phi_{ic_i}\;,\\
g_{ic_i}(\phi_{ic_i}) \;&=\; \pi(y_i,c_i|\phi_{ic_i})\,g_{0c_i}(\phi_{ic_i})\Big/\int\pi(y_i,c_i|\phi_{ic_i})\,g_{0c_i}(\phi_{ic_i})\,d\phi_{ic_i}
\end{aligned}$$

and κ is a normalizing constant such that

$$q_{i0}+\sum_{j\neq i}q_{ij} \;=1\;.$$

We can derive closed form expressions for the densities g_{ic_i} and the q_{ij} by using the base distribution for the Dirichlet process G_{c_i}, $G_{0c_i}=U(0,L_{c_i})\times \mathrm{Inv}\,\Gamma(a,L^2_{c_i}b)$, as follows. Let $\phi(\cdot|a,b^2)$ denote the $N(a,b^2)$ density, $\psi_a(\cdot|b,c)$ the $St(a,b,c)$ density and $\Psi_a(\cdot)$ the t_a distribution function. Note that if $X\sim t_a$ then $b+\sqrt{c}\,X\sim St(a,b,c)$. Also, to simplify notation, we write $\tau=\sigma^2$. Then

$$q_{ij} \;=\; \kappa\,\pi(y_i,c_i|\phi_{ic_i},\theta_{c_i}) \;=\; \kappa\,\theta_{c_i}\phi(y_i|\mu_{ic_i},\tau_{ic_i})$$

and

$$\begin{aligned} q_{i0} &= \kappa\,\alpha \int \pi(y_i, c_i|\phi_{ic_i}, \theta_{c_i})\, g_{0c_i}(\phi_i)\, d\phi_{ic_i} \\ &= \kappa\,\alpha \int_0^{L_{c_i}} \int_0^{\infty} \theta_{c_i}\, \phi(y_i|\mu_{ic_i}, \tau_{ic_i}) \times \frac{1}{L_{c_i}} \times \frac{(L_{c_i}^2 b)^a\, \tau_{ic_i}^{-a-1}\, e^{-L_{c_i}^2 b/\tau_{ic_i}}}{\Gamma(a)}\, d\tau_{ic_i}\, d\mu_{ic_i} \\ &= \frac{\kappa\,\alpha\,\theta_{c_i}}{L_{c_i}} \int_0^{L_{c_i}} \psi_{2a}(\mu|y_i, L_{c_i}^2 b/a)\, d\mu \\ &= \frac{\kappa\,\alpha\,\theta_{c_i}}{L_{c_i}} \left\{ \Psi_{2a}\!\left(\frac{1 - y_i/L_{c_i}}{\sqrt{b/a}}\right) - \Psi_{2a}\!\left(-\frac{y_i/L_{c_i}}{\sqrt{b/a}}\right) \right\}. \end{aligned}$$

Also, for $0 \le \mu_{ic_i} \le L_{c_i},\ \tau_{ic_i} > 0$

$$\begin{aligned} g_{ic_i}(\phi_{ic_i}) &= \frac{\pi(y_i, c_i|\phi_i, \theta_{c_i})\, g_0(\phi_{ic_i})}{\displaystyle\int \pi(y_i, c_i|\phi_{ic_i}, \theta_{c_i})\, g_0(\phi_{ic_i})\, d\phi_{ic_i}} \\ &= \frac{\phi(y_i|\mu_{ic_i}, \tau_{ic_i}) \times (L_{c_i}^2 b)^a\, \tau_{ic_i}^{-a-1}\, e^{-L_{c_i}^2 b/\tau_{ic_i}} \big/ \Gamma(a)}{\Psi_{2a}\!\left(\frac{1-y_i/L_{c_i}}{\sqrt{b/a}}\right) - \Psi_{2a}\!\left(-\frac{y_i/L_{c_i}}{\sqrt{b/a}}\right)} \\ &= \frac{(L_{c_i}^2 b)^a\, \tau_{ic_i}^{-a-3/2}}{\sqrt{2\pi}\ \Gamma(a) \left\{ \Psi_{2a}\!\left(\frac{1-y_i/L_{c_i}}{\sqrt{b/a}}\right) - \Psi_{2a}\!\left(-\frac{y_i/L_{c_i}}{\sqrt{b/a}}\right) \right\}} \\ &\quad \times\ \exp\left\{ -\left(L_{c_i}^2 b + \frac{(y_i - \mu_{ic_i})^2}{2} \right) \Big/ \tau_{ic_i} \right\}. \end{aligned}$$

For simulation purposes, it is useful to note that

$$g_{ic_i}(\phi_{ic_i}) \;=\; \pi(\mu_{ic_i})\, \pi(\sigma_{ic_i}^2|\mu_{ic_i})$$

where

$$\mu_{ic_i} \sim St\big(2a, y_i, L_{c_i}^2 b/a\big)\ , \qquad 0 \le \mu_{ic_i} \le L_{c_i}$$

and

$$\sigma_{ic_i}^2|\mu_{ic_i} \sim \operatorname{Inv}\Gamma\left(a + \frac{1}{2},\ L_{c_i}^2 b + \frac{(y_i - \mu_{ic_i})^2}{2} \right).$$

B. The remaining parameters

Here we derive the posterior conditional distributions for α and $\underline{\theta}$. The procedure is a generalisation of that used by Escobar and West ([3]).

Suppose that chromosome c has n_c hits arranged in k_c clusters (c=1,2,...,C). Then the probability function for the number of clusters on chromosome c is

$$\pi(k_c|n_c,\alpha,\underline{\theta}) \propto \begin{cases} \alpha^{k_c}\dfrac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}, & k_c=1,2,...,n_c\,, \quad \text{if } n_c>0\,, \\ 1, & k_c=0\,, \quad \text{if } n_c=0\,. \end{cases}$$

Let $\underline{k}=(k_1,k_2,...,k_C)$. As we have independent Dirichlet processes for each chromosome, $k_c|n_c,\alpha$ are independent for $c=1,2,...,C$ and so

$$\pi(\underline{k}|\underline{n},\alpha,\underline{\theta}) \propto \prod_{c=1}^{C}\alpha^{k_c}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\,,$$

for $k_c=1,2,...,n_c$ if $n_c>0$ or $k_c=0$ if $n_c=0$ ($c=1,2,...,C$). This can be simplified slightly by letting $A=\{c\colon n_c>0\}$ with size $|A|$, and renumbering the chromosomes so that $n_c>0$ for $c=1,2,...,|A|$ and $n_c=0$ for $c=|A|+1,|A|+2,...,C$, giving

$$\pi(\underline{k}|\underline{n},\alpha,\underline{\theta}) \propto \prod_{c=1}^{|A|}\alpha^{k_c}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\,.$$

The probability function for the number of hits on each chromosome has a multinomial distribution, with

$$\pi(\underline{n}|\alpha,\underline{\theta}) \propto \prod_{c=1}^{C}\theta_c^{n_c}\,,$$

and so the likelihood function for $(\alpha,\underline{\theta})$ is

$$\pi(\underline{k},\underline{n}|\alpha,\underline{\theta}) = \pi(\underline{k}|\underline{n},\alpha,\underline{\theta})\,\pi(\underline{n}|\alpha,\underline{\theta}) \propto \prod_{c=1}^{|A|}\alpha^{k_c}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\times\prod_{c=1}^{C}\theta_c^{n_c}\,.$$

Thus, if we take a gamma $\Gamma(g,h)$ prior distribution for α, the joint posterior density is

$$\begin{aligned}\pi(\alpha,\underline{\theta}|\underline{k},\underline{n}) &\propto \pi(\underline{k},\underline{n}|\alpha,\underline{\theta})\,\pi(\underline{\theta}|\alpha)\,\pi(\alpha) \\ &\propto \prod_{c=1}^{|A|}\alpha^{k_c}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\times\prod_{c=1}^{C}\theta_c^{n_c+\alpha\ell_c-1}\times\alpha^{g-1}e^{-h\alpha} \\ &\propto \prod_{c=1}^{|A|}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\times\prod_{c=1}^{C}\theta_c^{n_c+\alpha\ell_c-1}\times\alpha^{g+(\bar{k}-1)|A|-1}e^{-h\alpha}\,,\end{aligned}$$

where $\bar{k}=\sum_{c=1}^{|A|}k_c/|A|$ be the mean cluster size over chromosomes with hits. Therefore the (conditional) posterior density for $\underline{\theta}$ is

$$\pi(\underline{\theta}|\alpha,\underline{k},\underline{n}) \prod_{c=1}^{C}\theta_c^{n_c+\alpha\ell_c-1}\,,$$

that is, a $Dir(\underline{n}+\alpha\underline{\ell})$ distribution. Also the (conditional) posterior density for α is

$$\pi(\alpha|\underline{\theta},\underline{k},\underline{n}) \;\propto\; \alpha^{G-1}e^{-H\alpha}\prod_{c=1}^{|A|}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\;,$$

where $G = g + (\bar{k}-1)\,|A|$ and $H = h - \sum_{c=1}^{C}\ell_c\log\theta_c$. Using the identity

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)} \;=\; \frac{(\alpha+n_c)\,B(\alpha+1,n_c)}{\alpha\,\Gamma(n_c)}\;,$$

where $B(\cdot,\cdot)$ is the Beta function, we obtain

$$\pi(\alpha|\underline{\theta},\underline{k},\underline{n}) \;\propto\; \alpha^{G-1}e^{-H\alpha}\prod_{c=1}^{|A|}(\alpha+n_c)\,B(\alpha+1,n_c)\;. \tag{B.1}$$

As the Beta function has integral representation

$$B(\alpha+1,n_c) \;=\; \int_0^1 x_c^{\alpha}(1-x_c)^{n_c-1}\,dx_c$$

it is clear that

$$\pi(\alpha,\underline{\eta}|\underline{\theta},\underline{k},\underline{n}) \;\propto\; \alpha^{G-1}e^{-H\alpha}\prod_{c=1}^{|A|}(\alpha+n_c)\,\eta_c^{\alpha}(1-\eta_c)^{n_c-1}\;,$$

where $\underline{\eta} = (\eta_1,\eta_2,...,\eta_{|A|})'$ are beta distributed auxiliary variables, has distribution (B.1) when marginalised over $\underline{\eta}$. Therefore, letting $\bar{\eta}_g = \big(\prod_{c=1}^{|A|}\eta_c\big)^{1/|A|}$ be the geometric mean of the components of $\underline{\eta}$, we have

$$\pi(\alpha|\underline{\eta},\underline{\theta},\underline{k},\underline{n}) \;\propto\; \alpha^{G-1}\exp\big\{-(H-|A|\log\bar{\eta}_g)\,\alpha\big\}\prod_{c=1}^{|A|}(\alpha+n_c)\;. \tag{B.2}$$

Now

$$\prod_{c=1}^{|A|}(\alpha+n_c) \;=\; e_0(\underline{n})\,\alpha^{|A|} + e_1(\underline{n})\,\alpha^{|A|-1} + e_2(\underline{n})\,\alpha^{|A|-2} + \cdots + e_{|A|}(\underline{n})$$

where

$$e_0(\underline{n}) = 1\,,\quad e_1(\underline{n}) = \sum_{i=1}^{|A|} n_i\,,\quad e_2(\underline{n}) = \sum_{1=i<j}^{|A|} n_i\,n_j\,,\quad ...,\quad e_{|A|}(\underline{n}) = \prod_{i=1}^{|A|} n_i\;.$$

Here the $e_k(\underline{n})$ are elementary symmetric polynomials which may be calculated efficiently by using the Newton–Girard formula

$$k\,e_k(\underline{n}) = \sum_{i=1}^{k}(-1)^{i-1}e_{k-i}(\underline{n})\,S_k(\underline{n}) \qquad \text{where}\;\; S_k(\underline{n}) = \sum_{i=1}^{|A|} n_i^k\;.$$

Substituting this power series expansion into (B.2) gives

$$\pi(\alpha|\underline{\eta},\underline{\theta},\underline{k},\underline{n}) \;\propto\; \sum_{i=0}^{|A|} e_i(\underline{n})\,\alpha^{G+|A|-i-1}\exp\{-(H-|A|\log\bar{\eta}_g)\,\alpha\}\;,$$

which is a mixture of Gamma distributions, that is,

$$\alpha|\underline{\eta},\,\underline{\theta},\,\underline{k},\,\underline{n} \;\sim\; \sum_{i=0}^{|A|} p_i\,\Gamma\big(\alpha;\,G+|A|-i,\,H-|A|\log\bar{\eta}_g\big)$$

with mixture proportions

$$p_i \;=\; \frac{e_i(\underline{n})\,\Gamma(G+|A|-i)}{\sum_{j=0}^{|A|} e_j(\underline{n})\,\Gamma(G+|A|-j)\,(H-|A|\log\bar{\eta}_g)^{j-i}}\;, \qquad i=0,1,...,|A|\;.$$

Finally, for $c=1,2,...,|A|$,

$$\eta_c|\alpha,\,\underline{k},\,\underline{n} \;\sim\; Beta(\alpha+1,n_c), \quad \text{independently}\,.$$

ACKNOWLEDGMENTS

This work was conducted as part of the ComparaGRID project and funded by the UK Biotechnology and Biological Sciences Research Council grant number BBS/B/17158.

REFERENCES

[1] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool, *Journal of Molecular Biology*, **215**(3), 403–410.

[2] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics*, **2**(6), 1152–1174.

[3] Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Society*, **90**(430), 577–588.

[4] Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **1**(2), 209–230.

[5] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**(4), 457–511.

[6] Givry, S.; Bouchez, M.; Chabrier, F.; Milan, S. and Schiex, T. (2005). Cartha-GENE: multipopulation integrated genetic and radiation hybrid mapping, *Bioinformatics*, **21**(8), 1703–1704.

[7] Heidelberger, P. and Welch, P. (1982). Simulation run length control in the presence of an initial transient, *Operations Research*, **31**(6), 1109–1144.

[8] Higgs, D.R.; Vickers, M.A. and Wilkie, A.O. (1989). A review of the molecular genetics of the human alpha-globin gene cluster, *Blood*, **73**(5), 1081–1104.

[9] Hubbard, T.J.P.; Aken, B.L.; Beal1, K.; Ballester1, B.; Caccamo, M.; Chen, Y.; Clarke, L.; Coates, G.; Cunningham, F.; Cutts, T.; Down, T.; Dyer, S.C.; Fitzgerald, S.; Fernandez-Banet, J.; Graf, S.; Haider, S.; Hammond, M.; Herrero, J.; Holland, R.; Howe, K.; Howe, K.; Johnson, N.; Kahari, A.; Keefe, D.; Kokocinski, F.; Kulesha, E.; Lawson, D.; Longden, I.; Melsopp, C.; Megy, K.; Meidl, P.; Overduin, B.; Parker, A.; Prlic, A.; Rice, S.; Rios, D.; Schuster, M.; Sealy, I.; Severin, J.; Slater, G.; Smedley, D.; Spudich, G.; Trevanion, S.; Vilella, A.; Vogel, J.; White, S.; Wood, M.; Cox, T.; Curwen, V.; Durbin, R.; Fernandez-Suarez, X.M.; Flicek, P.; Kasprzyk, A.; Proctor, G.; Searle, S.; Smith, J.; Ureta-Vidal, A. and Birney, E. (2007). Ensembl 2007, *Nucleic Acids Res.*, **35** (Database issue), 610–617.

[10] Hyndman, R.J. (1996). Computing and graphing highest density regions, *The American Statistician*, **50**(2), 120–126.

[11] Jow, H.; Bhattacharjee, M.; Boys, R.J. and Wilkinson, D.J. (2010). The integration of genetic maps using Bayesian inference, *Journal of Computational Biology*, **17**, 825–840.

[12] Lander, E.R. and Botstein, D. (1989). Mapping mendialian factors underlying quantitative traits using RFLP linkage maps, *Genetics*, **121**, 185–199.

[13] Liao, W.; Collins, A.; Hobbs, M.; Khatkar, M.S.; Luo, J. and Nicholas, F.W. (2007). A comparative location database (CompLDB): map integration within and between species, *Mammalian Genome*, **18**(5), 287–299.

[14] Morton, N.E.; Collins, A.; Lawrence, S. and Shields, D.C. (1992). Algorithms for a location database, *Annals of Human Genetics*, **56**, 223–232.

[15] Pruitt, K.D.; Tatusova, T. and Maglott, D.R. (2005). Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Research*, **33**, 501–504.

[16] Schiex, T. and Gaspin, C. (1997). *Carthagene: Constructing and joining maximum likelihood genetic maps*. In "Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology", pp. 258–267.

[17] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

[18] Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: Joinmap, *The Plant Journal*, **3**(5), 739–744.

[19] Stassen, H.H. and Scharfetter, C. (2000). Integration of genetic maps by polynomial transformations, *American Journal of Medical Genetics (Neuropsychiatric Genetics)*, **96**(1), 108–113.

REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 19–36

OPTIMAL DYNAMIC TREATMENT METHODS

Authors: Robin Henderson
– Mathematics & Statistics, Newcastle University, UK
Robin.Henderson@newcastle.ac.uk

Phil Ansell
– Mathematics & Statistics, Newcastle University, UK
P.S.Ansell@newcastle.ac.uk

Deyadeen Alshibani
– Mathematics & Statistics, Newcastle University, UK
Deyadeen.Alshinani@newcastle.ac.uk

Abstract:

- This paper reviews and develops methods for implementing in practice recent ideas in the field of optimal dynamic treatment allocation. Given longitudinal sequences of observational data on health status and treatment selection for a cohort of patients, the aim is to determine a regime, or decision rule, which can be used to select treatment in order to optimise some final response or outcome. The approach to this problem that has been taken in the causal inference literature is shown to be extendable to problems in the field of stochastic optimisation. New diagnostic techniques to aid in model assessment are developed, and an application in anticoagulation is presented.

Key-Words:

- *anticoagulation; bandit problems; causal inference; diagnostics; regret functions; wild bootstrap.*

AMS Subject Classification:

- 62J07, 62L07.

1. INTRODUCTION

Individualised medicine, which is one of the growing areas in health research, presents a number of statistical challenges. Without the luxury of major clinical trials, can we find methods to tailor treatment to a patient's individual circumstances, especially for those with chronic conditions? In this paper we give an overview of a selection of methods for *optimal dynamic treatment regime determination* from observational data [1], [3], [6], [11]–[13]. Our interest in the area is motivated by a collaboration in which an algorithm to determine decision rules for anticoagulation doseage is required. Anticoagulants are used to maintain blood clotting speed and reduce risk of thrombosis. They are one of the most prescribed groups of drugs in the world, being used for both treatment and prophylaxis for conditions like deep venous thrombosis, stroke, atrial fibrillation, acute myocardial infarction, prosthetic heart valves and many more. A difficulty is that there is no standard dose: the amount required varies not just between patients but also over time within patients, in response to lifestyle and dietary changes, in particular the amount of vitamin K within the body. Given a patient's current and previous values of blood clotting time, and their history of anticoagulation, can we find decision rules to provide the optimal current dose?

Three classes of methods for a general version of this problem are summarised in Section 2. We consider model formulation and estimation, and illustrate through simulations. In Section 3 we draw attention to links between recent optimal dynamic treatment methods and the longstanding stochastic scheduling research in the operational research literature. In Section 4 we propose a suite of diagnostic tests for model adequacy based on wild bootstrap residuals. In Section 5 we describe an application of the methods to the warfarin anticoagulation application which motivated our interest.

2. REGRETS, BLIPS AND REGRESSION

2.1. Modelling approaches

We assume there are K decision times, for example clinic visits. At each decision time a *state* variable is recorded, $S_1, S_2, ..., S_K$. This might be the health of a patient and can be multivariate or scalar. A decision on the *action* to be taken is then made, such as treatment allocation, leading to an action sequence $A_1, A_2, ..., A_K$. The objective is to maximise some final value Y, which may not be revealed until all K decisions have been taken, or which may accrue with

time, as in the warfarin example in Section 5. Panel (a) of Figure 1 illustrates the sequence that is followed. Throughout we will assume independence between subjects and will take the standard assumption of no unmeasured confounders: all non-random elements influences action choices are captured in the observed data. We omit further technical detail on the conditions needed for valid inference.

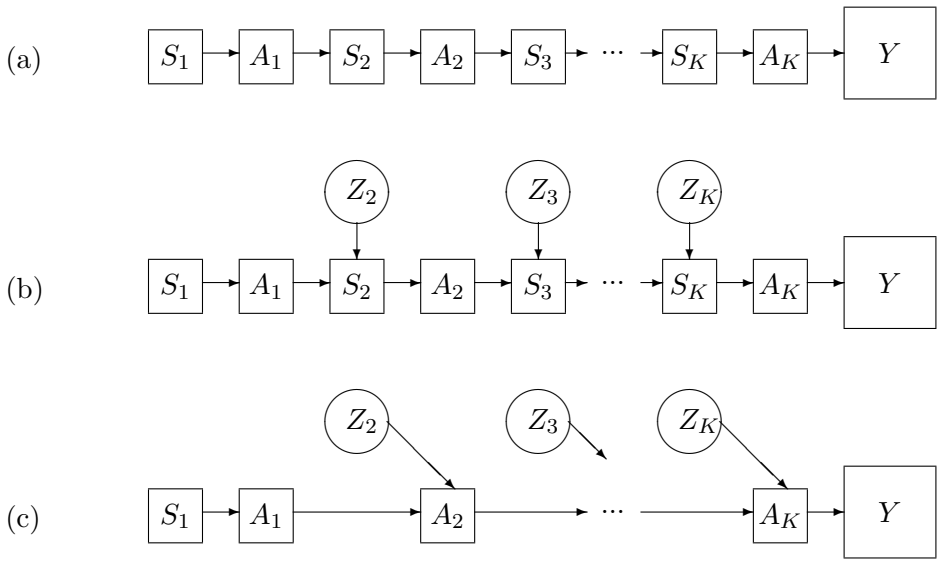

Figure 1: State, action and outcome sequence: (a) the general scenario; (b) inclusion of exogenous variables; (c) orthogonalisation.

Define $\bar{S}_j = (S_1, ..., S_j)$ and $\bar{A}_j = (A_1, ..., A_j)$ to indicate the history of states and actions respectively, up to and including time j. The information available just before action j is selected is $\mathcal{F}_j = (\bar{S}_j, \bar{A}_{j-1})$ and the aim is to obtain decision *rules* $d_j(\mathcal{F}_j)$ which will maximise the expected value of Y given the information to hand. We will use $\underline{d}_j^{\text{ref}}$ to denote a known standard or reference regime, with the underscore being read as meaning all times from j to K. Similarly $\underline{d}_j^{\text{opt}}$ is the *optimal* regime, which is unknown and is the target for analysis.

Robins [14] proposed a structural nested mean model [8] approach to the problem, based on *blip functions*, which can be defined as

$$\gamma_j(a_j|\mathcal{F}_j) \;=\; E(Y \mid \mathcal{F}_j, a_j, \underline{d}_{j+1}^{\text{opt}}) - E(Y \mid \mathcal{F}_j, d_j^{\text{ref}}, \underline{d}_{j+1}^{\text{opt}}) \;.$$

Here γ_j is a function of the possible actions a_j which are available at time j, given the history $\mathcal{F}_j$ of states and actions up to that point. The blip contrasts two expectations. The first is of the final response Y given that a_j is selected at

time j and under the possibly counterfactual assumption that the optimal reference regime will be followed from $j+1$ onward. The second expectation is similar except action a_j is replaced by the reference regime d_j^{ref} at time j. Robins chose the name because in each expectation the past F_j is the same, the future policy is the same, and the only difference or "blip" is between a_j and d_j^{ref} at time j.

Under Robins' approach a parametric form $\gamma_j(a_j|\mathcal{F}_j;\theta)$ is assumed for the blip function. For example we might take

$$\gamma_j(a_j|\mathcal{F}_j;\theta) \;=\; \theta_1\Big(a_j - (\theta_2 + \theta_3 S_j + \theta_4 S_{j-1})\Big)\, I(a_j \neq d_j^{\text{ref}})$$

where $I(\cdot)$ is an indicator function introduced to ensure the blip is zero if the reference action is selected. Otherwise the effect of the action a_j is assumed to depend on current and previous states S_j and S_{j-1} respectively. This is a strong assumption, but an advantage of the approach is that once parameter θ is estimated it is straightforward to determine the causal effect of actions.

Murphy [13] prefers to work with *regret functions*

$$\mu_j(a_j|\mathcal{F}_j) \;=\; E(Y \mid \mathcal{F}_j, \underline{d}_j^{\text{opt}}) - E(Y \mid \mathcal{F}_j, a_j, \underline{d}_{j+1}^{\text{opt}}) \;.$$

These are of similar form to blip functions except they contrast the effect at time j of action a_j with the as-yet-unknown optimal rule. Thus the first expectation assumes the optimal decision is taken from j onward, whereas in the second expectation action a_j is chosen at j and then the optimal policy followed from $j+1$ onward. Regrets are non-negative since the objective is to maximise Y. They give a direct measure of the effect of choosing a sub-optimal action at time j.

Again a parametric form is assumed, for example

$$\mu_j(a_j|\mathcal{F}_j;\psi) \;=\; \psi_1\Big(a_j - (\psi_2 + \psi_3 S_j + \psi_4 S_{j-1})\Big)^2 \;.$$

This guarantees the non-negativity of the regrets and assumes the optimal action — that which has zero regret — is a linear combination of S_j and S_{j-1}. Once ψ is known the optimal action is therefore immediately obtained.

The parametric forms assumed for blips or regrets cannot be checked, since they are models for differences in counterfactuals. An alternative approach introduced independently by Almirall and colleagues [1] and Henderson and colleagues [6] attempts to incorporate parametrised regrets into a model for the actual response Y. The authors note first that the final response Y is determined by three groups of factors:

1. The initial conditions.

2. The actions selected.

3. Chance development over time.

It is straightforward to introduce initial conditions as a function of S_1 into a model for Y. The effect of actions can be modelled by regrets as above. To model chance development over time, [1] and [6] envisage a sequence of exogenous variables $Z_1, Z_2, ..., Z_K$ which influence states and Y over and above the effects of the chosen actions, as summarised in panel (b) of Figure 1. If the $\{Z_j\}$ are observed then we can model final response as

$$E[Y|\bar{S}_K, \bar{A}_K] \;=\; \beta_1(S_1) + \sum_{j=2}^{K} \beta_j^T(\bar{S}_{j-1}, \bar{A}_{j-1})\, Z_j - \sum_{j=1}^{K} \mu_j(A_j|\mathcal{F}_j)\ , \tag{2.1}$$

where β_1 is an appropriate function to capture the effect of initial conditions, and $\beta_2, \beta_3, ..., \beta_K$ are coefficients which measure the effect of the exogenous variables. In principle these can depend on the complete history of states and actions: in practice dimensionality can be managed by allowing them to depend only on recent history. The regrets μ measure the effects of actions and complete the three components. Since the $\{Z_j\}$ are unknown, Henderson *et al.* propose they be estimated by residuals from models for S_j on previous states and actions $(\bar{S}_{j-1}, \bar{A}_{j-1})$. If linear models are used then the residuals are orthogonal to the covariates. Thus, we can separate the effect of exogenous variables from the effect of earlier decisions, as displayed in panel (c) of Figure 1. See [6] for further information.

2.2. Estimation

Moodie *et al.* [11] provide a very clear description of the estimation procedures proposed by Robins and Murphy. We provide only a brief outline here. The blips of Robins [14] can be obtained by first obtaining constructed variables which estimate at each j the response under the optimal policy:

$$H_j(\theta) \;=\; Y + \sum_{k \geq j} \Big\{ \gamma_j(d_j^{\text{opt}}|\mathcal{F}_j; \theta) - \gamma_j(A_j|\mathcal{F}_j; \theta) \Big\}\ .$$

A user-specified vector $V_j(A_j)$ of length $\dim(\theta)$ is then specified. By construction $H_j(\theta)$ is independent of $V_j(A_j)$ and so

$$0 \;= \sum_j H_j(\theta) \Big\{ V_j(A_j) - E\big[V_j(A_j)|\bar{S}_j, \bar{A}_{j-1}\big] \Big\}$$

is an unbiased estimating equation.

Murphy [13] takes a different approach. She defines a sum of squares involving two versions of the parameter vector ψ, say ψ and ψ^*, together with a

stabilising constant c:

$$f_n(\psi, \psi^*, c) \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\Bigg(Y^i + c + \sum_{l=1, l\neq j}^{K}\mu_l(\bar{S}_l^i, \bar{A}_l^i; \psi) \\ + \mu_j(\bar{S}_j^i, \bar{A}_j^i; \psi^*) \;-\; \hat{E}_{A_j}\Big(\mu_j(\bar{S}_j^i, \bar{A}_j^i; \psi^*)|\bar{S}_j^i, \bar{A}_{j-1}^i\Big)\Bigg)^2 .$$

Murphy shows that consistent estimation is possible through an iterative procedure to find $(\psi, \hat{c})$ such that

$$f_n(\psi, \psi, \hat{c}) \;\leq\; f_n(\psi, \psi^*, c)$$

for all (ψ^*, c). Note that this is not the same as minimising f.

The estimation methods of Robins and Murphy are at best computationally challenging. By contrast, the approach of Almirall *et al.* [1] and Henderson *et al.* [6] is based on a model for the observed response (2.1) which means standard methods are available. Henderson *et al.* propose ordinary least squares between observed and expected responses, which is valid without any distributional assumption for responses. More efficient procedures may be possible if further assumptions are made.

2.3. Illustration

We will illustrate using a simple two-timepoint example with Normal states and binary actions as also used by Moodie *et al.* [11]. Data were generated as $S_1 \sim \mathrm{N}(450, 150^2)$, $A_1 \sim \mathrm{Bern}(0.5)$, $S_2 \sim \mathrm{N}(1.25\,S_1, 60^2)$ and $A_2 \sim \mathrm{Bern}(0.5)$. Blip functions were parametrised, leading to regrets

$$\mu_1(a_1|S_1; \psi) \;=\; \begin{cases} I(a_1 = 0)\,(\psi_{10} + \psi_{11}S_1)\,, & \psi_{10} + \psi_{11}S_1 > 0\ , \\ -I(a_1 = 1)\,(\psi_{10} + \psi_{11}S_1)\,, & \psi_{10} + \psi_{11}S_1 < 0\ , \end{cases}$$

$$\mu_2(a_1|\bar{S}_2, A_1; \psi) \;=\; \begin{cases} I(a_1 = 0)\,(\psi_{20} + \psi_{21}S_1)\,, & \psi_{20} + \psi_{21}S_2 > 0\ , \\ -I(a_1 = 1)\,(\psi_{20} + \psi_{21}S_2)\,, & \psi_{20} + \psi_{21}S_2 < 0\ , \end{cases}$$

and then response $Y \sim \mathrm{N}\Big(400 + 1.6\,S_1 - \mu_1(A_1|S_1; \psi) - \mu_2(S_1|\bar{S}_2, A_1; \psi), 60^2\Big)$.

Table 1 compares G-estimation as used by Moodie *et al.* with the regret-regression method proposed by [6]. For the latter we used ordinary least squares to fit the correctly specified model

$$E[Y|\bar{S}_2, \bar{A}_2] \;=\; \beta_0 + \beta_1 S_1 - \mu_1(A_1|S_1; \psi) - \mu_2(S_1|\bar{S}_2, A_1; \psi)\ .$$

The `nlm` routine in R was used for parameter estimation. In all simulations the algorithm converged very quickly. Both methods produce apparently unbiased estimators, as they should, with smaller standard errors under the regret-regression method.

Table 1: Summary of simulation results based on Moodie *et al.* scenario. One thousand repetitions at sample size $n = 500$.

True ψ	G-estimation*		Regret-regression	
	Mean	SE	Mean	SE
250.0	250.01	17.20	250.20	11.39
−1.0	−1.00	0.04	−1.00	0.03
720.0	720.30	24.05	719.85	10.82
−2.0	−2.00	0.04	−2.00	0.02

* These results are taken from Moodie *et al.* (2007), who used the doubly robust form of G-estimation: their equation (2), which is the most efficient of the methods they considered.

Table 2 investigates how estimated parameters translate into decision regime performance. One thousand repetitions at sample size $n = 500$ were generated. After each repetition a further 10 000 observations were generated using each of four different decision rules: the gold standard of always choosing the optimal decision; equally likely randomised decisions; and following the estimated decision rules obtained from the first stage data by G-estimation of the regret functions and by the regret-regression procedure. Column $\bar{Y}$ gives the mean achieved response for each procedure, and column "Err" gives the overall percentage of times a suboptimal decision was made, pooled over both decision times. Columns 'Cut 1' and 'Cut 2' summarise the estimated cutpoints at each decision time, with the true values given in the gold standard row. Again we see that both G-estimation and regret-regression perform well, with again less variability when regret-regression is used.

Table 2: Further summary of simulation results based on Moodie *et al.* scenario. See text for explanation.

	$\bar{Y}$	SE	Err	Cut 1	SE	Cut 2	SE
Gold	1120.1	2.4	0.0	250.0		360.0	
Random	780.0	3.5	50.0				
Regrets (G-est.)	1119.6	2.8	0.6	249.9	9.9	359.5	12.7
Regret-regression	1120.0	2.5	0.3	250.5	6.3	359.9	2.6

3. REGRET-REGRESSION FOR A TWO-ARM BANDIT PROBLEM

The methods summarised above were developed with the aim of causal inference from observational data. In this section we argue that they can also be applied to problems from the stochastic optimisation literature. We will illustrate using the classic two-arm bandit problem.

At time j the state value S_j is a 2-vector (M_{0j}, M_{1j}), where $M_{0j} \in \{0, 1\}$ is the value of *arm zero* and $M_{1j} \in \{0, 1\}$ of *arm one*. The action A_j is to choose one of the arms. Response Y is then incremented by a reward which depends on the current value of the chosen arm. In our example the rewards are 6 or 4 for the two values of arm zero, and 8 or 3 for the two values of arm one. If arm zero is selected then M_{0j} is updated for time $j + 1$ according to a Markov chain but M_{1j} remains at its previous value. The opposite happens if arm one is selected: M_{1j} is updated but M_{0j} is unchanged. In our example the transition matrices are

$$P_0 = \begin{pmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{pmatrix} \qquad \text{and} \qquad P_1 = \begin{pmatrix} 0.4 & 0.6 \\ 0.5 & 0.5 \end{pmatrix}.$$

This is a special case of the so-called multi-armed bandit problem. A single resource is available to process a collection of competing projects (arms) over an infinite horizon. At each decision time $j = 0, 1, ...$, a decision must be taken as to which arm will be selected for processing. If arm k is chosen at time j then a discounted reward of

$$\lambda^j R_k(M_{kj})$$

is gained, where $\lambda \in [0, 1)$ is a discount rate, $R_k(\cdot)$ is a reward function and M_{kj} is the value of a Markov chain modelling the evolution of arm k at time j. After a unit of time dedicated to project k, it changes state according to a Markov law of motion P_k. The states of the other arms remain unchanged.

The objective is to find a policy for allocating arms for processing that maximises the total expected discounted reward over an infinite horizon. In principle, for particular problems the use of dynamic programming and the application of Bellman's principle of optimality [2] would allow these classical problems to be solved. However, as the size of the problem increases, the computational difficulties become intractable. Additionally, no insight into the structure of the optimal policy is obtained. An alternative method of solution, based around *forwards induction*, was introduced by Gittins and Jones [5]. They defined a dynamic allocation index (DAI) as

$$G_k(x_k) = \sup_{\tau > 0} \frac{E\left[\sum_{t=0}^{\tau-1} \lambda^t R(M_t) | M_0 = x_k\right]}{1 - E\left[\lambda^\tau\right]},$$

where the bandit is initially in state x_k and τ is a positively valued stopping time defined on the process. The Gittins Index policy is the one that selects the arm

with the current largest DAI. Such policies, since Whittle [17], are now referred to as *Gittins Index policies*. There are a number of methods for calculating the Gittins index including direct calculation, calibration methods, linear programming and special purpose algorithms; see [4] for more details. The "largest to smallest algorithm" [15] was implemented for the illustration here.

For the special two-arm two-value case described at the opening of this section, the Gittins Index policy under almost no discounting ($\lambda = 0.9999$) is to choose arms 1,0,1,1 for states $(0,0), (0,1), (1,0)$ and $(1,1)$ respectively. Note that a play-the-winner rule which optimises current reward would be 1,0,1,0 in the same order. The difference is at state (1,1) where the rewards on offer are (4,3). The Gittins policy of choosing arm one acknowledges future expectation — the possibility that the arm one reward value could change from 3 to 8 — whereas the play-the-winner rule is myopic and takes the higher immediate reward of 4 on offer from arm zero.

The Gittins policy is derived under an assumption that the process continues indefinitely and the optimal policy is stationary. We can use the regret-regression method to examine optimal dynamic policies for fixed length horizons. We simply simulate the process with actions chosen randomly and then fit a linear model incorporating regrets and residuals from dummy variables to describe the values. After each action the model includes residuals associated with eight dummy variables: one for each state/action combination. We choose optimal actions by working from the final timepoint and changing the action to ensure regrets are positive, starting with a working guess at which actions are optimal. Since linear models are used, this is a trivial task even when large samples are used to smooth out the noise generated by the Markov chains.

Table 3 illustrates for $K = 5$, showing the optimal action for each state S_j $(j = 1, 2, ..., 5)$ for this problem, along with the regrets for choosing a suboptimal action. It is interesting to compare the optimal choices with the Gittins policy. In states (0,0) and (1,0) they are the same: choose action $A_j = 1$ and hence take reward 8 units. State (1,1) has Gittins and optimal actions the same at $A_j = 1$ until time $j = 5$ at which final time the higher short-term reward under action $A_j = 0$ should be taken. State (0,1) also has a change in optimal action near the end, but this time at the penultimate decision stage. When in this state at earlier times, the optimal dynamic policy is to choose action $A_j = 1$ whereas the stationary Gittins policy is to choose $A_j = 0$.

For reference we give the mean reward under four decision regimes:

Regime	Mean Y
Random, prob 0.5	25.2
Play-the-winner	26.0
Gittins	27.1
Optimal dynamic	27.8

Table 3: Optimal actions and regrets for two-arm bandit problem with horizon $K = 5$. The first reward is obtained if $A_j = 0$, the second if $A_j = 1$. See text for other parameter values.

State S_j	Rewards	Optimal action				
		$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
(0,0)	6 or 8	1	1	1	1	1
(0,1)	6 or 3	1	1	1	0	0
(1,0)	4 or 8	1	1	1	1	1
(1,1)	4 or 3	1	1	1	1	0

State S_j	Rewards	Regret				
		$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
(0,0)	6 or 8	0.32	0.32	0.32	0.80	2.00
(0,1)	6 or 3	0.32	0.32	0.32	0.40	3.00
(1,0)	4 or 8	0.88	0.88	0.88	1.60	4.00
(1,1)	4 or 3	0.88	0.88	0.88	0.40	1.00

4. DIAGNOSTICS

We return to the general problem of Section 2 and focus on the regret-regression approach based on (2.1). An advantage of this approach is that we model the actual responses and hence can obtain residuals between observed and fitted values. Plots of residuals against covariates, fitted values, selected actions or estimated regrets can be used for diagnostic assessment and model comparisons. However, we have made no assumptions on response Y other than independence and our model (2.1) for the mean. In particular we have not assumed homogeneity of variance, which implies that whilst there should be no trends in the means of plots of residuals there may well be systematic patterns in the scatter, even for a correctly specified model. Further, standard bootstrap methods can be problematic when observations are independent but not identically distributed.

We propose to test for trend in residual plots using the wild bootstrap or conditional multiplier method [7], [10]. Suppose we have variables $\{D_i\}$ $(i = 1, 2, ..., n)$ which are independent with zero mean and finite but not necessarily equal variance. Suppose further that $T_0 = n^{-1/2} \sum_{i=1}^{n} D_i$ converges in distribution to some variable D. Let $\{\xi_i\}$ $(i = 1, 2, ..., n)$ be independent and identically distributed with zero mean and unit variance. Then $T_1 = n^{-1/2} \sum_{i=1}^{n} \xi_i D_i$ also converges in distribution to D. The wild bootstrap resampling method is to generate N independent copies of $\{\xi_i\}$ and use the resulting N copies of T_1 as an empirical estimator of the distribution of T_0. Note that all original variables D_i contribute exactly once to each T_1: there is no omission or duplication as in the standard bootstrap.

A complication is that residuals $R_i = Y_i - E[Y_i|\bar{S}_{iK}, \bar{A}_{iK}]$ are not independent. Our proposal is to base a test statistic on a contrast: $T_0 = n^{-1/2}\sum_{i=1}^{n} c_i R_i$ where $\sum_{i=1}^{n} c_i = 0$. We then obtain N resamples of $T_1 = n^{-1/2}\sum_{i=1}^{n} \xi_i c_i R_i$ and we compare the observed T_0 with the empirical distribution of T_1 to obtain a test of trend. The detail is as follows:

1. Order the residuals against a chosen covariate (or the fitted value).

2. Divide the residuals into six equally sized groups 1 to 6 corresponding to lowest sixth to highest sixth covariate values (with minor adjustments below if the six groups cannot be equal).

3. Select a contrast set from the following:

Test		Contrast coefficients c					
		1	2	3	4	5	6
T_1	Trend	1	1	1	−1	−1	−1
T_2	Curvature	1	1	−2	−2	1	1
T_3	Lower tail	1	−1	0	0	0	0
T_4	Upper tail	0	0	0	0	1	−1

4. Compute T_0 with the chosen contrasts for the six groups. Compute N wild bootstrap versions as described above using standard Normal $\{\xi_i\}$ and obtain an empirical p-value as the proportion of resampled test statistics which are more extreme than T_0.

Simulation results (not shown) indicate that all of the tests have the correct size for correctly specified models and that none uniformly dominates for power. We propose that all four be adopted in practice and in addition we recommend a fifth test based on the extremum of the cumulative residuals:

$$T_5 = \max_j \left\{ \sum_{i=1}^{j} R_i \right\}.$$

5. APPLICATION

Rosthøj *et al.* [16] and Henderson *et al.* [6] describe analyses of data on warfarin treatment of patients on long term anticoagulation. There are 303 patients with 14 clinic visits each. At each visit the International Normalised Ratio (INR) of blood clotting time was recorded, along with the *change* in prescribed dose of anticoagulant. If INR is too high then patients have risk of severe bleeding, whereas if INR is too low then there is risk of thrombosis. The aim therefore is to adjust dose to maintain as closely as possible INR within a target range, which can depend on underlying condition of the patient.

The response variable Y used in previous analysis is the overall percentage time the patient spent in range (PTR), which was to be maximised. The state variable S_j used by [16] and [6] is a standardised version of the INR, defined to be zero if the patient has INR in range, and otherwise the scaled distance to the nearest target boundary, with scaling by the population standard deviation. This ensures comparability between patients with different conditions and target intervals. The action variable A_j is the change in dose, in mg Warfarin. The first four visits are considered as a stabilisation period, and since there is no information after the final visit this is not used for the analyses to come. Thus $K = 9$ and the data for analysis consist of states $S_1, S_2, ..., S_9$ and actions $A_1, A_2, ..., A_9$ for the 303 patients. Henderson *et al.* [6] also worked with a discretised state S_j^* given by

$$S_j^* = \begin{cases} 1\,, & S_j \le -0.3 & \text{(very low)}\,, \\ 2\,, & -0.3 < S_j < 0 & \text{(low)}\,, \\ 3\,, & S_j = 0 & \text{(in range)}\,, \\ 4\,, & 0 < S_j < 0.55 & \text{(high)}\,, \\ 5\,, & S_j \ge 0.55 & \text{(very high)}\,. \end{cases}$$

Rosthøj *et al.* used the methods of [13] and were able to fit only one very simple regret model:

$$\mu_j(a_j|\mathcal{F}_j) = \begin{cases} I(a_j \neq 0)\left(5.84 + 1.59\, a_j^2\right), & S_j = 0\,, \\ 0.24\big(a_j + 2.01\, S_j\big)^2\,, & S_j \neq 0\,. \end{cases} \tag{5.1}$$

Here the optimal decision by construction is to leave dose unchanged if INR is within range, and is otherwise to change in proportion to state. For high states the dose should be increased so as to reduce clotting time, and the opposite for low states. The regret for a suboptimal decision increases quadratically as the dose change moves away from optimal. The model is overly simple and not claimed to be realistic, but Rosthøj *et al.* were unable to obtain convergence of either the G-estimation or iterative methods (see Section 2) for more realistic models.

Henderson *et al.* used the regret-regression approach based on (2.1) and had no difficulty in fitting more realistic models. Their final selection assumed that the regret function depended on the current discretised state S_j^* and the previous standardised state S_{j-1}. For category s of S_j^* the model is:

$$\mu_j(a_j|\mathcal{F}_j, S_j^* = s; \psi) = \psi_{s1}\, f(A_j - \psi_{s2} - \psi_{s3}\, S_{j-1})\,, \tag{5.2}$$

where $f(u) = u$ if $u \ge 0$ and $f(u) = u^2$ otherwise. Parameter estimates and bootstrap standard errors from 100 resamples are given in the upper part of Table 4.

Table 4: Parameter estimates and bootstrap standard errors for anticoagulation model example. Upper section: $\lambda = 1$ and analysis as [6]. Lower section: $\lambda = 0.3$.

$S_j^* = s$	ψ_{s1}	SE	ψ_{s2}	SE	ψ_{s3}	SE
−2	0.67	0.32	2.15	0.27	−1.11	0.38
−1	0.38	0.11	2.74	0.18	−1.57	0.67
0	0.97	0.36	−0.14	0.32	−1.12	0.74
1	2.38	0.27	−2.33	0.26	−0.98	0.27
2	2.83	0.79	−3.00	0.44	0.25	0.21
1	0.28	0.17	1.86	0.33	−1.05	0.58
2	0.12	0.11	3.00	0.43	−1.54	0.81
3	0.23	0.27	−0.10	0.17	−1.21	0.75
4	1.24	0.37	−1.57	0.42	−0.27	0.43
5	1.47	0.60	−1.98	0.69	0.39	0.39

To illustrate our diagnostic test suggestion, we will consider residuals from the two fitted models plotted against the regret following the first considered visit time. Figure 2 shows that the residuals from model (5.1) are more variable than those from model (5.2), with perhaps more evidence of trend in the early and later segments. To investigate, we applied the five wild bootstrap tests of Section 4. p-values from 200 wild bootstrap samples are given in Table 5. They confirm the early and late trends for model (5.1) are significant and the model is not fully adequate, but there are no significant trends for model (5.2).

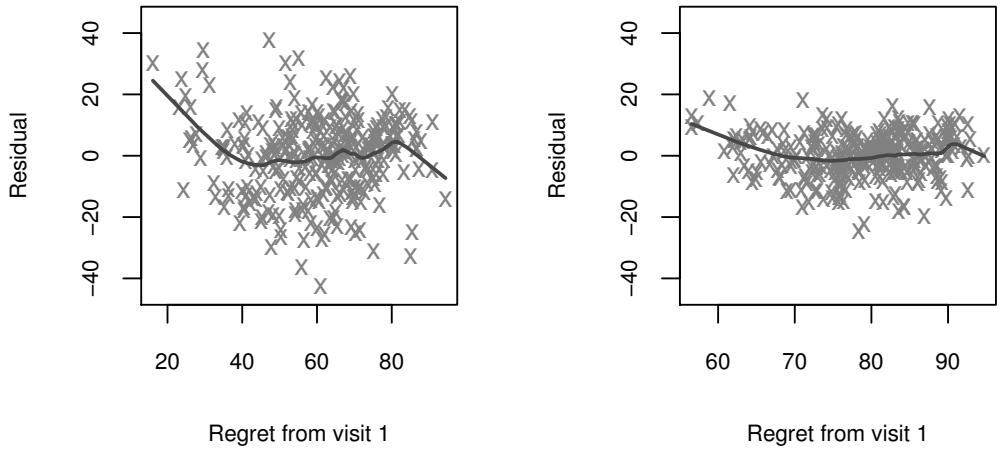

Figure 2: Warfarin residuals against regret at time 1. Left plot: model (5.1); right plot: model (5.2). The solid line is a smooth through the data.

Table 5: Wild bootstrap p-values for residuals in Figure 2.

Model	Test				
	T_1	T_2	T_3	T_4	T_5
(5.1)	0.830	0.000	0.593	0.018	0.055
(5.2)	0.753	0.611	0.136	0.870	0.816

We summarise now a new analysis of the warfarin data with a revised response variable. As well as a decision on the dose to be taken, at each clinic visit there is also a recommendation as to the timing of the next visit. Generally overly frequent visits are discouraged. Letting $N(\tau)$ be the number of visits in follow-up time τ we propose a new response

$$Y = Y(\tau) = \lambda\, PTR(\tau) + (1-\lambda)\frac{\tau}{N(\tau)}\ , \qquad 0 \le \lambda \le 1\ , \tag{5.3}$$

which weights together percentage time in range and average time between visits. Overly frequent visits thus reduce the response. For the warfarin data $N(\tau)$ is fixed at nine visits of interest but the time τ taken varies considerably between patients.

We will use model (5.2) for analysis. Choosing $\lambda = 1$ gives the previous results. To explore, we also analysed for a variety of other values for λ. To illustrate, the lower part of Table 4 gives parameter estimates at $\lambda = 0.3$ together with bootstrap standard errors obtained from 100 resamples. The general trend against s is the same as for $\lambda = 1$ but since the response is on a different scale it is hard to make a direct comparison. Instead, in Figure 3 we show the estimated optimal actions at a variety of combinations of current and previous state. The crosses indicate the values obtained when $\lambda = 1$ and the other points indicate values at a sequence of decreasing λ. As expected, increase in dose is indicated when INR is low, and decrease when INR is high, with previous INR moderating the action. Generally there is little effect of λ except at high INR, where the recommendation would be to reduce dose by a smaller amount if timing of visits is of interest. The rationale is that large dose changes are usually followed by quick return visits to monitor the effect. If this is to be discouraged then more modest changes are recommended. There lack of effect of λ at the low values of INR reflects the asymmetry in risk: very low values of INR need immediate strong action.

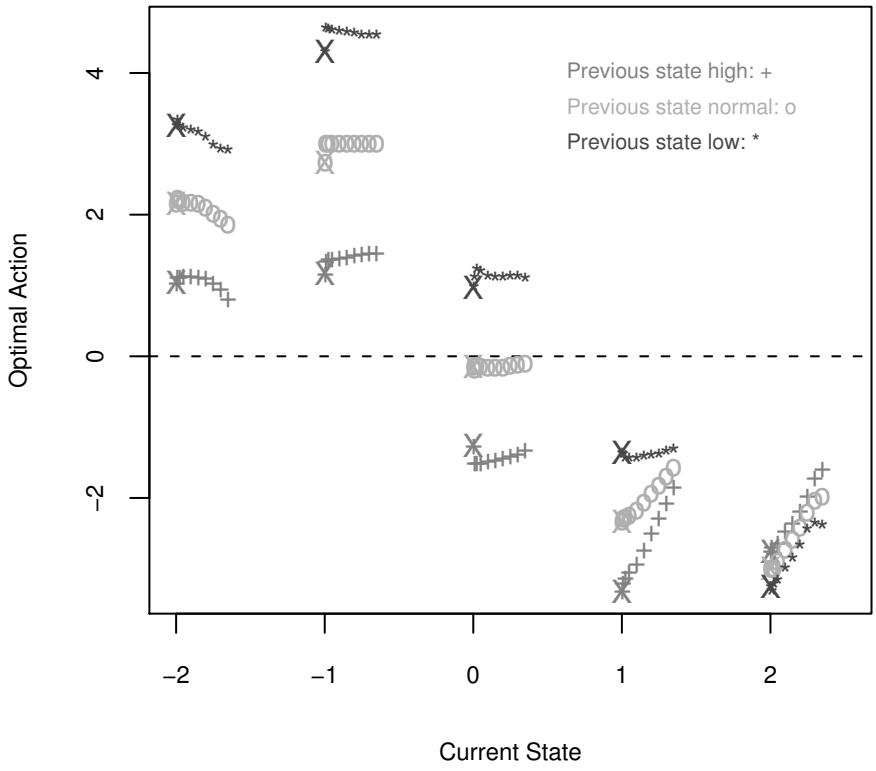

Figure 3: Effect of changing λ in response 5.3. The crosses mark optimal actions when $\lambda = 1$. The other points show how the optimal action changes as λ varies through $\{0.995, 0.99, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3\}$, moved to the right for display purposes.

6. DISCUSSION

We have presented an overview of the structural nested mean model approach to optimal dynamic treatment regime determination, with focus on blip models [14], regret models [13], and regret-regression models [1], [6]. Although there has been growing discussion in the literature on causal inference for dynamic treatment regimes, the area is still very much underdeveloped and there are few genuine applications in realistic problems. One issue is the computational challenge faced for reasonable sized data sets. Another is the assumption of balanced data, in the sense of common clinic or visit times. Methods which allow irregular timing of visits are needed. In this case the definition of regrets and blips is problematic. The counting process approach may be fruitful [9] but much further research is needed. Nonetheless we see great promise in this type of approach.

ACKNOWLEDGMENTS

We are grateful to the organisers of the meeting *Statistical Modelling: Challenges in Health, 2010*, for the invitation to present this work.

REFERENCES

[1] Almirall, D.; Ten Have, T. and Murphy, S.A. (2010). Structural nested mean models for assessing time-varying effect moderation, *Biometrics*, to appear.

[2] Bellman, R.E. (1957). *Dynamic Programming*, Princeton University Press, Princeton.

[3] Chakraborty, and Murphy, S.A. (2009). Inference for nonregular parameters in optimal dynamic treatment regimes, *Statistical Methods in Medical Research*, **19**, 317–343.

[4] Gittins, J.C. (1989). *Multi-Armed Bandit Indices*, Wiley, Chichester.

[5] Gittins, J.C. and Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments, *Progress in Statistics, European Meeting of Statisticians*, **1**, 241–266.

[6] Henderson, R.; Ansell, P. and Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes, *Biometrics*, to appear.

[7] Lin, D.Y.; Fleming, T.R. and Wei, L.J. (1994). Confidence bands for survival curves under the proportional hazards model, *Biometrika*, **81**, 73–81.

[8] Lok, J.; Gill, R.; van der Vaart, A. and Robins, J.M. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models, *Statistica Neerlandica*, **58**, 1–25.

[9] Lok, J. (2008). Statistical modeling of causal effects in continuous time, *Annals of Statistics*, **36**, 1464–1507.

[10] Martinussen, T. and Scheike, T.H. (2006). *Dynamic Regression Models for Survival Data*, Spring-Verlag, New York.

[11] Moodie, E.; Richardson, T.S. and Stephens, D. (2007). Demystifying optimal dynamic treatment regimes, *Biometrics*, **63**, 447–455.

[12] Moodie, E.E.M. and Richardson, T.S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null, *Scandinavian Journal of Statistics*, **37**, 126–146.

[13] Murphy, S. (2003). Optimal dynamic treatment regimes (with discussion), *Journal of the Royal Statistical Society Series B*, **65**, 331–366.

[14] Robins, J.M. (2004). Optimal structured nested models for optimal sequential decisions. In "Proceedings of the Second Seattle Symposium on Biostatistics" (D.Y. Lin and P.J. Heagerty, Eds.), Springer, New York, 189–326.

[15] Robinson, D.R. (1982). Algorithms for evaluating the dynamic allocation index, *Operations Research Letters*, **1**, 72–74.

[16] Rosthøj, S.; Fullwood, C.; Henderson, R. and Stewart, S. (2006). Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach, *Statistics in Medicine*, **25**, 4197–4215.

[17] Whittle, P. (1980). Multi-armed bandits and the Gittins index, *Journal of the Royal Statistical Society Series B*, **42**, 143–149.

REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 37–55

MISSING DATA IN REGRESSION MODELS FOR NON-COMMENSURATE MULTIPLE OUTCOMES

Authors: Armando Teixeira-Pinto
– Serviço de Bioestatística e Informática Médica,
CINTESIS, Faculdade de Medicina, Universidade do Porto,
Portugal
tpinto@post.harvard.edu

Sharon-Lise Normand
– Department of Health Care Policy, Harvard Medical School,
and Department of Biostatistics, Harvard School of Public Health,
Boston, USA
normand@hcp.med.harvard.edu

Abstract:

- Biomedical research often involves the measurement of multiple outcomes in different scales (continuous, binary and ordinal). A common approach for the analysis of such data is to ignore the potential correlation among the outcomes and model each outcome separately. This can lead not only to loss of efficiency but also to biased estimates in the presence of missing data. We address the problem of missing data in the context of multiple non-commensurate outcomes. The consequences of missing data when using likelihood and quasi-likelihood methods are described, and an extension of these methods to the situation of missing observations in the outcomes is proposed. Two real data examples illustrate the methodology.

Key-Words:

- *mixed outcomes; multivariate; latent variable; non-commensurate; missing data; maximum likelihood; direct maximization; weighted generalized estimating equations.*

AMS Subject Classification:

- 62J12, 62H99.

1. INTRODUCTION

Many biomedical studies involve measurements of multiple outcomes on each subject. When the outcomes are commensurate, i.e., are measured on the same scale and are measuring the same underlying variable, classical tools of multivariate statistics can be used. However, multivariate methods to analyze outcomes measured on different scales or measuring different underlying variables, i.e., non-commensurate outcomes, are less common and rarely used in data analysis. A common solution used in the presence of non-commensurate outcomes is to analyze each outcome separately, ignoring the potential correlation between the outcomes. There are several disadvantages of this approach. First, there might be a loss of efficiency by ignoring the extra information contained in the correlation between the outcomes. Second, with separate analysis it is harder to answer intrinsic multivariate questions such as the existence of a covariate effect on the underlying outcome. Third, if some outcomes are missing for some individuals, different samples of individuals will be included in the analysis of the effect of exposure on different outcomes. Finally, the situation of missing data may also produce biased results if the missing data depends on the other outcomes.

The main difficulty of modeling non-commensurate variables is that there is no obvious multivariate distribution. Mainly, three approaches to model non-commensurate outcomes have been described in the literature. The first has its roots in the general location model ([10]) and has been extended to accommodate covariates ([2]) and clustered data ([6], [12]). The key idea is to factorize the likelihood as the product of marginal and conditional distributions, and model each term of the product. However, this approach does not generalize easily when the number of outcomes is increased. The second approach uses latent variables to induce the correlation between the outcomes and assumes that conditional on these latent quantities, the outcomes are independent ([14], [17], [5]). The third approach extends the framework of generalized estimating equations (GEE) to multivariate discrete and continuous outcomes ([11], [20], [19]). The main advantages of the GEE over likelihood methods is the lack of assumptions regarding the distribution of the data and its robustness to misspecification of the correlation between the outcomes. Naturally this will lead to less efficient but more robust estimates (see Teixeira-Pinto and Normand ([19]) for a summary of these and other approaches).

With the measurement of multiple outcomes there is a higher risk of missing data. Few authors have addressed the problem of missing data in the setting of non-commensurate outcomes. Fitzmaurice and Laird ([7]) proposed the use of the EM-algorithm ([3]) to fit the extension of the general location model in the presence of missing data. Shafer ([15]) described likelihood-based data

augmentation approaches to missing data assuming a general location model. In Little and Rubin's ([9]) nomenclature, the missing data is defined as missing at random (MAR) if it only depends on the observed data. If the missing data does not depend on the observed or unobserved data, the missing data is designated as missing completely at random (MCAR). In contrast, if the the missing data depends on unobserved data, the missing mechanism is said to be missing not at random (MNAR). The GEE gives consistent estimates in the presence of missing data only if the data are MCAR. This would also apply to the GEE extension proposed by several authors ([11], [20] and, [19]). However, Robins *et al.* ([13]) extended the common GEE methodology to situations of MAR by weighting each observation by its inverse probability of being observed.

In this paper we describe the properties of the latent variable model under missing data and extend the weighted GEE (WGEE) to multiple non-commensurate outcomes for MAR data. A study investigating the association between participation in a managed behavioral health care carve-out and quality of health care measured using bivariate mixed outcomes ([4]), and a study evaluating health-related quality of life after discharge from an intensive care unit using the Euroqol-5d instrument([8]), illustrate our methods.

2. LATENT VARIABLE MODEL FOR MULTIPLE CONTINUOUS AND BINARY OUTCOMES

Let $(y_{1i}, ..., y_{qi})$ represent a multivariate outcome for the i^{th}-individual $(i = 1, ..., n)$. We will use the symbol $\cdot$ in the subscript of $y_{k\cdot}$ to designate all the observations for outcome k or $y_{\cdot i}$ to indicate all the outcomes for the individual i. Let $\boldsymbol{x}_{ji}$ represent a vector of covariates for the i^{th}-individual associated with the j^{th}-outcome. We allow each outcome to be associated with its own set of covariates. Let R_{ji} be an indicator variable with value 1 if y_{ji} is observed and 0 otherwise. The superscript 'obs' is used to denote *observed* data. We assume throughout that the covariates are fixed and completely observed, and thus will be suppressed when writing the conditional distributions.

2.1. Latent variable model with outcome data MAR

One approach to model non-commensurate outcomes in a multivariate framework is to introduce latent variables, $\boldsymbol{u}_i = (u_{1i}, ..., u_{pi};\, p < q)$, to induce the correlation between the outcomes. Conditional on the latent variables $\boldsymbol{u}$ the outcomes are assumed to be independent ([5]). We assume that one of the outcomes, y_{1i}, has some missing observations and that these observations are MAR,

i.e., $P\big(R_{1i}=1 \mid y_{\cdot i}, \boldsymbol{x}_{ji}\big)$ depends on the observed data, for example, $y_{2i}, ..., y_{qi}$ and $\boldsymbol{x}_{1i}$. Let θ be the vector of parameters associated with the distribution of $y_{\cdot i} \,|\, \boldsymbol{x}_{ji}$. The log-likelihood for the observed data is given by

$$\begin{aligned} \log L\big(\theta; y_{1\cdot}^{\text{obs}}, ..., y_{q\cdot}, R_{1\cdot}, \boldsymbol{x}_{ji}\big) \;\propto\; & \log \prod_{i=1}^{n} \Big(f(y_{\cdot i} \,|\, \boldsymbol{x}_{ji})\, P\big(R_{1i}=1 \,|\, y_{\cdot i}, \boldsymbol{x}_{ji}\big) \Big)^{R_{1i}} \\ & \times \left(\int f(y_{\cdot i} \,|\, \boldsymbol{x}_{ji})\, P\big(R_{1i}=0 \,|\, y_{\cdot i}, \boldsymbol{x}_{ji}\big)\, \partial y_{1i} \right)^{(1-R_{1i})} . \end{aligned} \tag{2.1}$$

With some algebraic manipulation and using the fact that R_{1i} does not depend on y_{1i} we can re-write (2.1) as

$$\begin{aligned} = & \sum_{i=1}^{n} \Big(R_{1i} \log f(y_{\cdot i} \,|\, \boldsymbol{x}_{ji}) + (1-R_{1i}) \log f(y_{2i}, ..., y_{qi} \,|\, \boldsymbol{x}_{ji}) \Big) \\ & + \sum_{i=1}^{n} \Bigg(R_{1i} \log \Big(P\big(R_{1i}=1 \,|\, y_{2i}, ..., y_{qi}, \boldsymbol{x}_{ji}\big) \Big) \\ & + (1-R_{1i}) \log \Big(P\big(R_{1i}=0 \,|\, y_{2i}, ..., y_{qi}, \boldsymbol{x}_{ji}\big) \Big) \Bigg). \end{aligned} \tag{2.2}$$

The terms in the log-likelihood involving the missingness mechanism $P\big(R_{1i} \,|\, y_{2i}, ..., y_{qi}, \boldsymbol{x}_{ji}\big)$ will not involve the parameters θ associated with the distribution of $y_{\cdot i} \,|\, \boldsymbol{x}_{ji}$. These terms will not contribute for the estimation of θ and for this reason they can be ignored. Therefore, the log-likelihood can be written as the sum of terms associated with the distribution for complete observations and terms associated with the distribution for incomplete observations. Thus, the presence of missing data does not add extra difficulty to the maximization of the likelihood. In this case we say that the likelihood can be directly maximized because it does not require a more complex method, such as the EM-algorithm nor multiple imputation, to compute the maximum likelihood estimates.

Consider the case of a binary outcome, $y_{1\cdot}$, and a continuous outcome, $y_{2\cdot}$, where some entries of $y_{1\cdot}$ are missing. In this case $q=2$ and $p=1$. We assume the following model for the outcomes:

$$\begin{aligned} \text{probit}\big(E(y_{1i} \,|\, \boldsymbol{x}_{1i}, u_i)\big) \;&=\; \boldsymbol{\beta}_1^{*\mathrm{T}} \boldsymbol{x}_{1i} + u_i \;, \\ y_{2i} \,|\, \boldsymbol{x}_{1i}, u_i \;&=\; \boldsymbol{\beta}_2^{\mathrm{T}} \boldsymbol{x}_{2i} + \sigma_2\, u_i + \epsilon_{2i} \;, \end{aligned} \tag{2.3}$$

where $\epsilon_{2i} \sim N(0, \sigma_2^2)$ and u_i is a latent variable with $u_i \sim N(0, \sigma_u^2)$. The latent variable u_i in the model induces the correlation between the outcomes and the parameter σ_2 that multiplies the latent variable is introduced to standardize the different scales of the two outcomes. For more details see Teixeira-Pinto and Normand ([19]).

The log-likelihood for the observed data can be written as

$$\log L\big(\theta; y_{1\cdot}^{\text{obs}}, y_{2\cdot}, R_{1\cdot}, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}\big) \;\propto \tag{2.4}$$

$$\propto \sum_{i=1}^{n}\Big(R_{1i}\log f(y_{1i}, y_{2i}\,|\,\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}) + (1-R_{1i})\log f(y_{2i}\,|\,\boldsymbol{x}_{2i})\Big)$$

$$= \sum_{i=1}^{n}\Big(R_{1i}\log\int f(y_{1i}\,|\,\boldsymbol{x}_{1i}, u_i)\, f(y_{2i}\,|\,\boldsymbol{x}_{2i}, u_i)\, f(u_i)\,\partial u_i \;+\; (1-R_{1i})\log f(y_{2i}\,|\,\boldsymbol{x}_{2i})\Big).$$

Depending on the link functions used for each outcome it might be possible to have a closed-form representation for the marginal distribution of each outcome. Using the identity link for the continuous outcome and the probit link for the binary as in (2.3), the model for the marginal means of each outcome can be written as

$$\begin{aligned}\text{probit}\big(P(y_{1i}=1\,|\,\boldsymbol{x}_{1i})\big) \;&=\; \text{probit}\Big(\int P(y_{1i}=1\,|\,\boldsymbol{x}_{1i}, u_i)\, f(u_i)\, du_i\Big) \;=\; \frac{\boldsymbol{\beta}_1^{*\mathrm{T}}\boldsymbol{x}_{1i}}{\sqrt{1+\sigma_u^2}}\;,\\ y_{2i}\,|\,\boldsymbol{x}_{2i} &= \boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{x}_{2i} + \epsilon_{2i}^{*}\,, \qquad \text{where}\;\; \epsilon_{2i}^{*} \sim N\big(0, \sigma_2^2(1+\sigma_u^2)\big)\,.\end{aligned} \tag{2.5}$$

If instead we choose a logit link for the binary outcome in equation (2.3), the model for the marginal mean does not have a closed-form representation.

3. WEIGHTED GENERALIZED ESTIMATING EQUATIONS FOR NON-COMMENSURATE OUTCOMES

3.1. WGEE with data MAR

Suppose we are in the same setting as in the previous section with a binary and a continuous outcome to motivate the WGEE. We adapt the WGEE proposed by Robins *et al.* ([13]) to the situation of multiple non-commensurate outcomes.

The generalization to multiple outcomes is relatively straightforward but some remarks will be made.

Let $y_{\cdot i} = (y_{1i}, y_{2i})^{\mathrm{T}}$ be a vector of a binary and a continuous outcome with the following marginal model for the outcomes:

$$\mu_{ji} \;=\; g_j^{-1}(\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_{ji})\;, \tag{3.1}$$

where $\mu_{ji} = E(y_{ji}\,|\,\boldsymbol{x}_{ji})$, $j = (1, 2)$, g_j is the probit link for $j = 1$ and the identity link for $j = 2$. If both outcomes are completely observed, the estimating equation

is

$$\sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1} (y_{\cdot i} - \mu_{\cdot i}) \;=\; 0 \tag{3.2}$$

and has a solution that is a consistent and asymptotically normal estimator for $\boldsymbol{\beta}_j$ ([20], [19]) with variance $\Gamma^{-1}\,\Omega\,\Gamma^{-1}$, where $D_i = \left(\frac{\partial \mu_{\cdot i}}{\partial \boldsymbol{\beta}}\right)_j$, V_i is a 'working' covariance matrix for y_{1i} and y_{2i}, $\Gamma = E(D_i^{\mathrm{T}} V^{-1} D_i)$ and $\Omega = E\big(D_i^{\mathrm{T}} V^{-1} (y_{\cdot i} - \mu_{\cdot i}) \cdot \cdot (y_{\cdot i} - \mu_{\cdot i})^{\mathrm{T}} V^{-1} D_i\big)$. Typically, D_i is a block-diagonal matrix because the equations for each outcome do not share the regression parameters. The solution for the estimating equation is a consistent estimator of $\boldsymbol{\beta}$ even if V_i is misspecified. In the case of missing data, this result holds if the data are MCAR but not for MAR.

Suppose that some observations of y_{1i} are missing and the missing mechanism depends on y_{2i} and x_{ji}. If the variables y_{2i} and x_{ji} are always observed then y_{1i} is MAR. In this case $E(y_{1i}^{\mathrm{obs}} \,|\, \boldsymbol{x}_{ji}) \neq \mu_{1i}$ because

$$\begin{aligned} E(y_{1i}^{\mathrm{obs}} \,|\, \boldsymbol{x}_{ji}) \;&=\; E(R_{1i}\, y_{1i} \,|\, \boldsymbol{x}_{ji}) \;=\; E\big(E(R_{1i}\, y_{1i} \,|\, y_{1i}, y_{2i}, \boldsymbol{x}_{ji})\big) \\ &=\; E\big(y_{1i}\, E(R_{1i} \,|\, y_{2i}, y_{1i}, \boldsymbol{x}_{ji})\big)\,. \end{aligned} \tag{3.3}$$

R_{1i} does not depend on y_{1i} because the data are MAR, and $E\big(y_{1i}\, E(R_{1i} \,|\, y_{2i}, y_{1i}, \boldsymbol{x}_{ji})\big)$ simplifies to $E\big(y_{1i}\, P(R_{1i} = 1 \,|\, y_{2i}, \boldsymbol{x}_{ji})\big)$. Therefore, this expectation is not equal to μ_{1i} so the solution for the equation (3.2) is no longer a consistent estimate of $\boldsymbol{\beta}_1$. However, if we weight y_{1i} by its inverse probability of being observed $\pi_{1i} = P(R_{1i} \,|\, y_{2i}, \boldsymbol{x}_{ji})$, we have:

$$E\left(\frac{R_{1i}}{\pi_{1i}}\,(y_{1i} - \mu_{1i}) \mid \boldsymbol{x}_{1i}\right) \;=\; E\left(E\left(\frac{R_{1i}}{\pi_{1i}}\,(y_{1i} - \mu_{1i}) \mid y_{1i}, y_{2i}, \boldsymbol{x}_{ji}\right) \,\middle|\, \boldsymbol{x}_{1i}\right)$$

and, because $E(R_{1i} \,|\, y_{2i}, \boldsymbol{x}_{ji}) = \pi_{1i}$,

$$=\; E\big(y_{1i} - \mu_{1i} \,|\, \boldsymbol{x}_{1i}\big) \;=\; 0\;.$$

This motivates the following weighted estimating equation:

$$\sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1} \Delta_i (y_{\cdot i} - \mu_{\cdot i}) \;=\; 0 \tag{3.4}$$

and

$$\Delta_i \;=\; \begin{pmatrix} R_{1i}\, \pi_{1i}^{-1} & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.5}$$

The estimating equation (3.4) has a solution $\hat{\boldsymbol{\beta}}$ which is a consistent estimate of $\boldsymbol{\beta}$ and has an asymptotic normal distribution with a consistent estimator

of its variance given by $\hat{\Gamma}^{-1}\left(\sum_{i=1}^{n}\hat{C}_i\,\hat{C}_i^{\mathrm{T}}\right)\Gamma^{-1^{\mathrm{T}}}$ where $\hat{\Gamma}=\sum_{i=1}^{n}(D_i^{\mathrm{T}}V^{-1}\Delta_i D_i)$, $\hat{C}_i=D_i^{\mathrm{T}}V^{-1}(y_{\cdot i}-\mu_{\cdot i})-\left(\sum_{i=1}^{n}D_i^{\mathrm{T}}V^{-1}(y_{\cdot i}-\mu_{\cdot i})\,S_i^{\mathrm{T}}\right)\left(\sum_{i=1}^{n}S_i\,S_i^{\mathrm{T}}\right)S_i$ and S_i is the score component for the i^{th}-individual from the model for π_{1i}.

The last entry in the matrix Δ_i is 1 because only $y_{1\cdot}$ is missing for some subjects and $y_{2\cdot}$ is always observed. The weights π_{ji} are unknown and have to be estimated. We can use, for example, a logistic regression to estimate $\pi_{1i}=P(R_{1i}=1\,|\,y_{2i},\boldsymbol{x}_{ji})$ as in (3.6) and plug in the estimates in equation (3.4).

$$\text{logit}(\pi_{1i}) \;=\; \zeta_0+\zeta_1\,y_{2i}+\zeta_2\,\boldsymbol{x}_{ji}\;. \tag{3.6}$$

The assumption of MAR implies that if R_{ji} depends on the other outcomes, then only one outcome can have missing observations. However, if there are missing observations in y_{2i} or in one of the covariates involved in the model (3.6), we no longer have a case of MAR and we are not able to estimate all the weights π_{1i}.

3.2. Estimation of the Covariance Parameters

Although we are mainly interested in the estimation of the parameters $\boldsymbol{\beta}_j$, consistent estimators for the parameters in $V_i=\begin{pmatrix}\sigma_1^2 & \rho\,\sigma_2\,\sigma_1\\ \rho\,\sigma_2\,\sigma_1 & \sigma_2^2\end{pmatrix}$ are needed in equation (3.4). One way of obtaining these estimators is to add estimating equations for these parameters. Because we are not concerned about estimating σ_1, σ_2 and ρ efficiently, we can use the following unbiased equations based on the method of moments:

$$\sum\frac{R_{1i}}{\pi_{1i}}\left(\sigma_1-\sqrt{\frac{\sum(y_{1i}-\mu_{1i})^2}{n}}\right)\;=\;0\;, \tag{3.7}$$

$$\sum\left(\sigma_2-\sqrt{\frac{\sum(y_{2i}-\mu_{2i})^2}{n}}\right)\;=\;0\;, \tag{3.8}$$

$$\sum\frac{R_{1i}}{\pi_{1i}}\left(\rho-\frac{\sum(y_{1i}-\mu_{1i})\,(y_{2i}-\mu_{2i})}{\sigma_2\,\sqrt{n\sum(y_{1i}-\mu_{1i})^2}}\right)\;=\;0\;.$$

Equations (3.4) and (3.7) can be solved jointly to obtain estimates for all the parameters. If instead of missing observations in y_{1i} we had missing observations in y_{2i}, then the terms in equation for σ_2 would also require to be weighted in order to obtain an unbiased estimator for σ_2.

This entire approach can be applied to more than two outcomes. However, the assumption of MAR implies that missingness mechanism has to depend only in completely observed outcomes. If this is not the case the data are MNAR.

4. SIMULATION STUDY

Data were generated using the model

$$(4.1)\quad (y_{1i}^*, y_{2i}) \,|\, (x_i, z_{1i}, z_{2i}) \;\sim\; MVN\left(\begin{pmatrix} .5 + 2\,x_i + 2\,z_{1i} \\ 5 + 10\,x_i + 2\,z_{2i} \end{pmatrix}, \begin{pmatrix} 1 & 6\times 1\times .8 \\ & 36 \end{pmatrix}\right),$$

with x_i generated from a Bernoulli(.5), z_{1i} generated from a Uniform$(-1, 0)$ and z_{2i} from $N(1, 4)$. Then, y_{1i}^* was categorized in the following way:

$$(4.2)\qquad y_{1i} = \begin{cases} 0\,, & \text{if } y_{1i}^* \le 0\ , \\ 1\,, & \text{if } y_{1i}^* > 0\ . \end{cases}$$

By using a probit link to model y_{1i} as $\text{probit}\big(P(y_{1i} = 1 \,|\, x_i, z_i)\big) = \alpha_1 + \beta_1 x_i + \gamma_1 z_i$, we have $P(y_{1i} = 1 \,|\, x_i, z_i) = P(y_{1i}^* > 0 \,|\, x_i, z_i) = \Phi\big(\frac{.5 + 2\,x_i + 2\,z_i}{\sigma_1}\big)$. By construction $\sigma_1 = 1$ thus, the true parameters for the probit regression maintain the same value as in (4.1), i.e., $\alpha_1 = .5$, $\beta_1 = 2$ and $\gamma_1 = 2$.

We generated 1000 datasets with 400 bivariate observations each. Some observations for the continuous outcome were deleted according to the model $\text{logit}\big(P(R_{2i} = 1 \,|\, y_{1i}, x_i)\big) = .5 - 3.5\,y_{1i} - x_i$. The parameters were chosen to obtain approximately 25% of missing observations (about 40% of missing y_{2i} when $x_i = 0$ and 5% when $x_i = 1$).

4.1. Univariate analysis

We fit separate regressions for each outcome, ignoring the missingness mechanism and the correlation between the outcomes. We used a probit regression for the binary outcome (4.3) and a linear regression for the continuous (4.4):

$$(4.3)\qquad \text{probit}\big(E(y_{1i} \,|\, x_i, z_{1i})\big) \;=\; \alpha_1 + \beta_1 x_i + \gamma_1 z_{1i}\ ,$$

$$(4.4)\qquad E(y_{2i} \,|\, x_i, z_{2i}) \;=\; \alpha_2 + \beta_2 x_i + \gamma_2 z_{2i}\ .$$

4.2. Latent variable model

We fit the latent variable model,

$$(4.5)\qquad \text{probit}\big(E(y_{1i} \,|\, x_i, z_{1i}, u_i)\big) \;=\; \alpha_1^* + \beta_1^* x_i + \gamma_1^* z_{1i} + u_i\ ,$$

$$(4.6)\qquad E(y_{2i} \,|\, x_i, z_{2i}, u_i) \;=\; \alpha_2 + \beta_2 x_i + \gamma_2 z_{2i} + \sigma_2 u_i\ .$$

It can shown that the above model is the correct model for the data generation process. To obtain marginal effects of the covariates as in the other models we have to average over the latent variable u_i. In this case the marginal effects can be obtained by dividing the parameters by $\sqrt{1+\sigma_u^2}$, for example, the marginal effect of x on y_1 is $\beta_1 = \frac{\beta_1^*}{\sqrt{1+\sigma_u^2}}$. We used PROC NLMIXED from SAS to fit the latent variable model. The initial parameters were obtained by fitting separate regressions for each outcome (univariate analysis). The initial value for the correlation parameter was set to be 0.5.

4.3. Weighted generalized estimating equations

We assumed the following model for the means of the outcomes:

$$\text{probit}\big(E(y_{1i}\,|\,x_i, z_{1i})\big) \;=\; \text{probit}(\mu_{1i}) \;=\; \alpha_1 + \beta_1 x_i + \gamma_1 z_{1i}\ , \tag{4.7}$$

$$E(y_{2i}\,|\,x_i, z_{2i}) \;=\; \mu_{2i} \;=\; \alpha_1 + \beta_1 x_i + \gamma_1 z_{2i}\ . \tag{4.8}$$

We solved the WGEE:

$$\sum_{i=1}^{n} \begin{pmatrix} -\phi(A_i) & 0 \\ -x_i\,\phi(A_i) & 0 \\ -z_{1i}\,\phi(A_i) & 0 \\ 0 & 1 \\ 0 & x_i \\ 0 & z_{2i} \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_2\,\sigma_1 \\ \rho\,\sigma_2\,\sigma_1 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & \frac{R_{2i}}{\hat{\pi}_{2i}} \end{pmatrix} \begin{pmatrix} y_{1i}-\mu_{1i} \\ y_{2i}-\mu_{2i} \end{pmatrix} = 0 \tag{4.9}$$

with $A_i = \alpha_1 + \beta_1 x_i + \gamma_1 z_{1i}$ and $\sigma_1 = \sqrt{\Phi(A_i)\,(1-\Phi(A_i))}$. The weights $\hat{\pi}_{2i}$ were estimated using the logistic regression

$$\text{logit}\big(R_{2i}=1\,|\,y_{1i}, x_i\big) \;=\; \text{logit}(\pi_{2i}) \;=\; \zeta_0 + \zeta_1\,y_{1i} + \zeta_2\,x_i\ . \tag{4.10}$$

Two additional equations were added to the system of equations (4.9) to obtain estimates of the unknown parameters σ_2 and ρ:

$$\begin{aligned} &\sum \frac{R_{2i}}{\pi_{2i}} \left(\sigma_2 - \sqrt{\frac{\sum (y_{2i}-\mu_{2i})^2}{n}} \right) = 0\ , \\ &\sum \frac{R_{2i}}{\pi_{2i}} \left(\rho - \frac{\sum (y_{1i}-\mu_{1i})\,(y_{2i}-\mu_{2i})}{\sigma_2 \sqrt{n \sum (y_{1i}-\mu_{1i})^2}} \right) = 0\ . \end{aligned} \tag{4.11}$$

The WGEE were solved using a program developed in SAS with PROC IML.

4.4. Results

The results of the simulations are summarized in Tables 1 and 2. Overall, the latent variable model performed better than the univariate approach and the WGEE. The estimates of the parameters associated with the continuous outcome, $\hat{\alpha}_2$ and $\hat{\beta}_2$, were biased for the univariate model, and the mean square errors (MSE) were about 4 and 6 times higher than the corresponding MSE estimates obtained from the latent variable model. The remaining estimates for the univariate approach were not biased but they had slightly higher standard errors than the latent model. This is explained by the fact that the latent variable model uses the additional information of the correlation between the outcomes as described by Teixeira-Pinto and Normand ([19]).

Table 1: Estimates and standard errors averaged over the results of 1000 simulated datasets with sample size equal to 400. About 25% data were deleted for the continuous outcome using a model for the missingness mechanism that depends on the binary outcome.

Estimates (true value)	Univariate		Latent		WGEE	
	Mean	(SE)	Mean	(SE)	Mean	(SE)
Binary outcome						
$\hat{\alpha}_1$ $(\alpha_1 = .5)$	0.521	(0.167)	0.521	(0.148)	0.519	(0.159)
$\hat{\beta}_1$ $(\beta_1 = 2)$	2.025	(0.181)	2.025	(0.172)	2.019	(0.181)
$\hat{\gamma}_1$ $(\gamma_1 = 2)$	2.045	(0.305)	2.044	(0.257)	2.035	(0.288)
Continuous outcome						
$\hat{\alpha}_2$ $(\alpha_2 = 5)$	6.523	(0.581)	5.009	(0.556)	5.033	(0.601)
$\hat{\beta}_2$ $(\beta_2 = 10)$	8.737	(0.702)	9.980	(0.685)	9.944	(0.737)
$\hat{\gamma}_2$ $(\gamma_2 = 2)$	2.001	(0.170)	1.999	(0.145)	1.999	(0.171)

Table 2: Mean square error (MSE) and relative bias (estimate/true value) averaged over the results of 1000 simulated datasets with sample size equal to 400. About 25% data were deleted for the contin uous outcome using a model for the missingness mechanism that depends on the binary outcome.

Estimates	Mean square error			Relative bias		
	Univ.	Latent	WGEE	Univ.	Latent	WGEE
Binary outcome						
$\hat{\alpha}_1$	0.030	0.024	0.027	1.042	1.042	1.037
$\hat{\beta}_1$	0.033	0.031	0.038	1.013	1.013	1.009
$\hat{\gamma}_1$	0.097	0.071	0.089	1.023	1.022	1.018
Continuous outcome						
$\hat{\alpha}_2$	2.669	0.317	0.371	1.305	1.002	1.007
$\hat{\beta}_2$	2.108	0.494	0.857	0.874	0.998	0.994
$\hat{\gamma}_2$	0.028	0.021	0.031	1.001	1.000	1.000

The WGEE estimates had very similar bias to the latent variable model, although the MSEs for all estimates were higher in the WGEE due to higher variances for the estimates. This loss of efficiency is expected when compared to a full likelihood method such as the latent variable model. The sandwich estimator for the variance of the estimates is robust to the misspecification of the correlation between the outcomes and for this reason is more conservative.

5. EXAMPLES

5.1. Example 1: Managed Care and Quality of Care for Schizophrenia

Dickey *et al.* ([4]) conducted a prospective observational study of 420 adults with schizophrenia who sought care for a psychiatric crisis. The main objective of the study was to compare care for patients who were and were not enrolled in managed care because advocates for those with mental illness worried that patients who had their care managed may have worse care than those who did not. Two outcomes, one binary (whether the patient was prescribed an atypical anti-psychotic medication) and one continuous (self-reported quality of interpersonal interactions between patient and clinician) were measured for the 197 patients who had their care managed and the 223 patients whose care was not managed. The self-reported quality of interpersonal interactions between patient and clinician was missing for 26 patients (6%). The information regarding the prescription of an atypical anti-psychotic was available for all the subjects. There was a significant difference in the proportion of patients who were prescribed an atypical anti-psychotic medication between the group without data on the quality of interpersonal interactions between patient and clinician (50%) and the group of patients with data on this outcome (71%) ($\chi^2_2 = 5$, p-value $= 0.03$). This result suggests that the data are MAR. There was no statistical significant association between the missing indicator and the sociodemographic characteristics using a significance level of 0.05.

We used separate regression models for each outcome (the univariate approach) ignoring the correlation between the outcomes and the missing data (equations 4.3 and 4.4). We fit the latent variable model (4.5) and the WGEE (4.9). For the latent variable model we computed the marginal effects estimates of managed care on the outcomes by dividing the regression coefficients by $\sqrt{1+\sigma^2_u}$ as described in section 2.1. The weights for the WGEE were obtained from a logistic model for the probability of missing observation in the self-reported quality of interpersonal interactions between patient and clinician outcome using the prescription of an atypical anti-psychotic and managed care status as covariates. The estimates for the weights were given by the inverse of the estimated probabilities from the logistic model, $\text{logit}(\hat{\pi}_{2i}) = \text{logit}\big(\hat{P}(R_{2i} \,|\, y_{1i}, x_i)\big) = 2.23 + 0.88\, y_{1i} - 0.11\, x_i$.

The mean (SD) age of patients was 40 (8.5) and 41 (7.9) in the managed care and not managed care group, respectively. Other sociodemographic characteristics of the patients are described in Table 3. No significant differences were observed for the two outcomes analyzed regarding the sociodemographic characteristics. Seventy one percent of the patients in the managed care group received atypical anti-psychotic medication versus 68% in the not managed care group. The mean (SD) self-reported quality of interpersonal interactions between patient and clinician was 3.20 (0.67) for the managed care group and 3.21 (0.65) for the not managed group.

Table 3: Sociodemographic characteristics of 420 patients with schizophrenia.

Sociodemographic characteristics	Type of care		p-value
	Managed ($n = 197$)	Not Managed ($n = 223$)	
Age			
< 35 years	24	21	
35–44 years	46	44	0.338
45–54 years	21	29	
55–64 years	8	6	
Male sex	47	66	< 0.001
Race or Ethnicity			
White	51	66	
African American	31	22	0.005
Other	18	12	
Never married	64	68	0.364
High school education or less	74	59	0.002
Homeless	15	9	0.069
English speaking	90	93	0.277

The effect estimates of managed care on the outcomes were identical and not statistically significant at the 0.05 level for all the models (Table 4). This suggests no difference in the quality of care between the managed and not managed care groups. For the outcome with some missing observations, patient/clinician relationship outcome, the estimated effect of managed care was the same for the latent variable model and the WGEE ($\hat{\beta}_2 = -0.019$). The effect estimate for the univariate approach was slightly smaller ($\hat{\beta}_2 = -0.017$). Although this result is consistent with the simulation study, it is hard to argue that such a small difference is a consequence of ignoring the MAR mechanism rather than random variation. The WGEE provided identical standard errors of the estimators to the other two approaches. This can be explained by the low correlation between the outcomes (0.059 as estimated by the WGEE).

Table 4: Managed care effect on the two outcomes related to quality of care: "patient/clinician relationship" and "prescription of anti-psychotic medication". Data on 420 patients with schizophrenia but only 394 patients had information regarding patient/clinician's relationship.

Estimated effects	Model		
	Univariate	Latent	WGEE
	β (Std. Error) p-value	β (Std. Error) p-value	β (Std. Error) p-value
Binary: Prescription of anti-psychotic ($n = 420$)			
Intercept	0.549 (0.089) ≤ 0.001	0.548 (0.089) ≤ 0.001	0.549 (0.088) ≤ 0.001
Managed care	-0.081 (0.129) 0.530	-0.079 (0.129) 0.538	-0.081 (0.128) 0.527
Continuous: Patient/clinician relationship ($n = 394$)			
Intercept	3.213 (0.045) ≤ 0.001	3.213 (0.045) ≤ 0.001	3.213 (0.045) ≤ 0.001
Managed Care	-0.017 (0.066) 0.799	-0.019 (0.066) 0.775	-0.019 (0.067) 0.771
$\hat{\sigma}_2$	0.656	0.630	0.656
$\hat{\sigma}_u$	—	0.286	—
$\hat{\rho}$	—	—	0.059

5.2. Example 2: Quality of life after discharge from Intensive Care

Granja *et al.* ([8]) evaluated the health-related quality of life (HRQOL) of adult patients discharged from an intensive care unit (ICU) located in Portugal. The 485 patients who agreed to participate in the study were asked to complete the Euroqol 5D (EQ-5D) instrument to evaluate their HRQOL ([1]), 6 months after discharge from ICU. This instrument includes two main sections. The first contains five questions that measure functional dimensions of HRQOL (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) and it is summarized by a general score designated as the EQ-5D index. The EQ-5D index varies from 0 to 100, where 100 indicates no disability in the 5 dimensions. The second part of the instrument is a visual analogue scale (VAS) in which patients mark their perception of their health state in a 0 to 100 scale (100 – best imaginable state, 0 – worst imaginable state). For the analysis the VAS scale was dichotomized using the middle point of its scale (less or equal to 50 and more than 50).

In this example we will focus on the impact of patient's severity when admitted to the ICU (measured by the Apache II score) on the HRQOL after discharge (measured by the EQ-5D index and dichotomized VAS). Some studies reported that most of the patients who survive ICU do not show significant decrease in physical ability but they report psychological problems ([18], [16]). This finding suggests that the effect of the severity of the episode that led to ICU admission may be different for functional HRQOL and for patient's perception of their HRQOL. If this is the case, both aspects of HRQOL should be reported in HRQOL studies.

The effect of patient's severity at ICU admission on HRQOL should be adjusted to age and previous health state (non-chronic disease, chronic disease with no disability and chronic disease with disability). All the patients completed the first part of the questionnaire involved in the calculation of the EQ-5D index, but only 366 completed the VAS question.

Table 5 summarizes some demographic and clinical information from the 485 patients. The mean (SD) age of the 485 patients was 55.2 (17.4) years old. Twenty eight percent (28%) of the patients reported that they had no chronic disease prior to admission to ICU and 21% reported they had chronic disease that caused some kind of disability. The remaining 51% indicated that they had chronic disease with no disability before admission to ICU. The mean (SD) Apache II score at admission was 13.0 (6.8). For the 366 patients who completed the VAS scale, 64% reported a value above 50. The mean (SD) for the EQ-5D index was 74.2 (17.4). The group of patients that completed both parts of the questionnaire had significantly higher EQ-5D index than those who did not completed the VAS question (77.9 vs. 52.6, p-value < 0.001).

Table 5: Demographic and clinical characteristics of 485 patients that participated in the study of HRQOL after ICU admission.

Demographic and clinical characteristics	($n = 485$)
Age (mean (SD))	55.2 (17.4)
Male sex (n (%))	275 (57)
Apache II score (mean (SD))	13.0 (6.8)
Previous health state (n (%))	
non-chronic disease	138 (28)
chronic disease with no disability	245 (51)
chronic disease with disability	102 (21)
ICU length of stay in days (median (IQR))	2 (1–6)

Similarly to example 1, we run separate models for each outcome (a linear regression for the EQ-5D index and a probit regression for the dichotomized VAS) and we fit the latent model and WGEE using the same link functions as

the univariate models. The effect of previous health state on both measures of HRQOL was linear for the three categories, so it entered the model as an interval variable with no need to create dummy variables for the categories. The weights for the WGEE were obtained from a logistic model for the probability of missing observation in the VAS question using the EQ-5D index, Apache II score, age and the previous health state as covariates. The estimates for the weights were given by the inverse of the estimated probabilities from the logistic model, $\text{logit}(\hat{\pi}_{1i}) = \text{logit}\big(\hat{P}(R_{1i} \,|\, y_{2i}, x_{1i}, x_{2i}, x_{3i})\big) = 0.85 + 0.04\, y_{1i} - 0.03\, x_{1i} - 0.04\, x_{2i} - 0.17\, x_{3i}$.

The results are summarized in Table 6. The HRQOL is associated with patient's age and the health state previous to admission. The severity at admission measured by Apache II is not associated with the functional aspect of HRQOL (p-value $= 0.999$). These results were consistent in all approaches.

Table 6: Effect of severity at admission to ICU (Apache II), adjusted to age and previous health state, on health-related quality of life measured (D-VAS and EQ-5D index), 6 months after discharge from an ICU. A total of 485 patients entered the study but only 366 completed the question regarding D-VAS.

Estimated effects	Model		
	Univariate	Latent	WGEE
	β (Std. Error) p-value	β (Std. Error) p-value	β (Std. Error) p-value
Binary: D-VAS ($n = 366$)			
Intercept	−2.069 (0.290) < 0.001	2.018 (0.280) < 0.001	2.027 (0.280) < 0.001
Age	−0.011 (0.004) 0.014	−0.009 (0.004) 0.029	−0.012 (0.004) 0.014
Previous health state	−0.460 (0.111) < 0.001	−0.494 (0.106) < 0.001	−0.442 (0.111) < 0.001
Apache II	−0.018 (0.011) 0.093	−0.028 (0.011) 0.009	−0.025 (0.012) 0.040
Continuous: EQ-5D ($n = 485$)			
Intercept	100.3 (3.902) < 0.001	100.3 (3.886) < 0.001	103.3 (3.489) < 0.001
Age	−0.244 (0.061) < 0.001	−0.244 (0.061) < 0.001	−0.244 (0.055) < 0.001
Previous health state	−8.116 (1.540) < 0.001	−8.116 (1.533) < 0.001	−8.115 (1.458) < 0.001
Apache II	≈ 0 (0.157) 0.999	≈ 0 (0.157) 0.999	≈ 0 (0.163) 0.999
$\hat{\sigma}_2$	21.94	14.86	21.94
$\hat{\sigma}_u$	—	1.086	—
$\hat{\rho}$	—	—	0.532

The major difference between the univariate and the multivariate methods is the result for the effect of Apache II on the dichotomized VAS. The estimate in the latent model and WGEE is higher than that in the univariate approach and it becomes statistically significant at the 0.05 level. This may indicate that the patient's perception about his or her own HRQOL is affected by the degree of severity of the episode leading to ICU admission. This fact would not be identified in the univariate analysis.

6. CONCLUSION

We developed likelihood and quasi-likelihood methods to analyze multiple non-commensurate outcomes in the presence of missing data. Although this type of data is common in biomedical studies, the usual approach is to analyze each outcome separately ignoring the correlation among the outcomes. This can lead to loss of efficiency and biased estimates in the case of MAR. The WGEE has the advantage of being robust to the misspecification of the correlation between the outcomes and MAR while the latent variable model is a full likelihood approach which typically gives more efficient estimates but assumes that the mean and covariance models are correctly specified. Another alternative to WGEE is to use the multiple imputation methodology. We could assume a model to impute values for the missing observations and repeat the process to create several complete datasets. Then we could solve a regular GEE for each dataset and obtain the estimates of the regression parameters as the mean over the estimates obtained in each complete dataset.

We have shown both in simulations and in real data analysis that the estimation of associations can be biased in situations of MAR in the outcomes. The bias can be substantially reduced by jointly model the outcomes in a multivariate framework.

ACKNOWLEDGMENTS

This work was supported by Grant R01-MH54693 (Teixeira-Pinto and Normand) and R01-MH61434 (Normand), both from the National Institute of Mental Health and PTDC/SAU-ESA/100841/2008 from Fundação para a Ciência e Tecnologia. The schizophrenia managed care data were generously provided through the efforts of Barbara Dickey, Ph.D., Harvard Medical School, Boston, MA. The health-related quality of life data was generously provided by Cristina Granja, MD, Ph.D., Hospital Pedro Hispano, Porto, Portugal.

REFERENCES

[1] Brooks, R. and the EuroQol Group (1996). EuroQol: the current state of play, *Health Policy*, **37**, 53–72.

[2] Cox, D.R. and Wermuth, N. (1992). Response models for binary and quantitative variables, *Biometrika*, **79**, 441–461.

[3] Dempster, A.; Laird, N. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society, Series B*, **39**, 1–38.

[4] Dickey, B.; Normand, S.-L.T.; Hermann, R.C.; Eisen, S.V.; Cortes, D.E.; Cleary, P.D. and Ware, N. (2003). Guideline recommendations for treatment of schizophrenia: the impact of managed care, *Arch. Gen. Psychiatry*, **60**, 340–8.

[5] Dunson, D.B. (2000). Bayesian latent variable models for clustered mixed outcomes, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **62**, 355–366.

[6] Fitzmaurice, G.M. and Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering, *Journal of the American Statistical Association*, **90**, 845–852.

[7] Fitzmaurice, G.M. and Laird, N.M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values, *Biometrics*, **53**, 110–122.

[8] Granja, C.; Teixeira-Pinto, A. and Costa-Pereira, A. (2002). Quality of life after intensive care: evaluation with EQ-5D questionnaire, *Intensive Care Medicine*, **28**, 898–907.

[9] Little, R.J. and Rubin, D. (2002). *Statistical Analysis with Missing Data*, John Wiley and Sons, Inc., Hoboken, New Jersey, U.S.A.

[10] Olkin, I. and Tate, R. (1961). Multivariate correlation models with mixed discrete and continuous variables, *The Annals of Mathematical Statistics*, **32**, 448–465.

[11] Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics*, **47**, 825–839.

[12] Regan, M.M. and Catalano, P.J. (1999). Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology, *Biometrics*, **55**, 760–768.

[13] Robins, J.; Rotnitzky, A. and Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, **90**, 106–121.

[14] Sammel, M.D.; Ryan, L.M. and Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes, *Journal of the Royal Statistical Society, Series B: Methodological*, **59**, 667–678.

[15] Schafer, J. (1997). *Analysis of Incomplete Multivariate Data (Chapter 9)*, Chapman and Hall / CRC Monographs on Statistics and Applied Probability, 72, New York, NY.

[16] Schelling, G.; Stoll, C.; Haller, M.; Briegel, J.; Manert, W.; Hummel, T.; Lenhart, A.; Heyduck, M.; Polasek, J.; Meier, M.; Preuss, U.; Bullinger, M.; Schffel, W. and Peter, K. (1998). Health-related quality of life and posttraumatic stress disorder in survivors of the acute respiratory distress syndrome, *Critical Care Medicine*, **26**, 634–635.

[17] Shi, J. and Lee, S. (2000). Latent variable models with mixed continuous and polytomous data, *Journal of the Royal Statistical Association, Series B*, **62**, 77–87.

[18] Sukantarat, K.; Greer, S.; Brett, S. and Williamson, R. (2007). Physical and psychological sequelae of critical illness, *British Journal of Health Psychology*, **12**, 65–74.

[19] Teixeira-Pinto, A. and Normand, S.-L.T. (2009). Correlated bivariate continuous and binary outcomes: issues and applications, *Statistics in Medicine*, **28**, 1753–73.

[20] Zhao, L.P.; Prentice, R.L. and Self, S.G. (1992). Multivariate mean parameter estimation by using a partly exponential model, *Journal of the Royal Statistical Society, Series B*, **54**, 805–811.

REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 57–81

A REVIEW ON JOINT MODELLING OF LONGITUDINAL MEASUREMENTS AND TIME-TO-EVENT

Authors: Inês Sousa
– Departamento de Matemática e Aplicações,
Universidade do Minho, Portugal
isousa@math.uminho.pt

Abstract:

- In longitudinal studies subjects are measured for one or more response variable, over time. Although the underlying evolution of such response variables is continuous in time, in practice the measurements are observed at discrete time points. In longitudinal clinical trials it is also common to observe relevant events, generating time-to-event data. If both types of data are available, we might be interested in the association between the two processes, longitudinal and time-to-event. Commonly, when death is considered the event, the observation sequence of longitudinal measurements is terminated by the event process. When the two observed processes are related, the analysis of the data set should be suited to the specific objectives. We distinguish three situations: if the interest is to analyse the longitudinal outcome response variable with drop-out at the time-to-event; to analyse time-to-event, whilst exploiting correlation with a noisy version of a time-varying risk factor; or to analyse the relationship between the two processes. Joint models assume a full distribution for the joint distribution of longitudinal and time-to-event processes, which includes a description of the relation between the two processes.

Key-Words:

- *longitudinal; time-to-event; survival; Gaussian; correlation structure.*

AMS Subject Classification:

- 62F12.

1. INTRODUCTION

Longitudinal studies are characterised by observation of repeated measurements on a number of subjects at a series of time points. In this work we will only consider continuous response variables. It is of interest, particularly in longitudinal clinical trials, to test significant differences between the underlying processes of the same response variables for different treatment groups.

Time to event data are a set of times on individuals, induced by multiple or single events. We will for this work only consider single events. In clinical trials patients are usually assigned to different treatment groups, or in different age or gender groups. Therefore, the aim of time-to-event analysis is to identify differences in the time-to-event distributions of different groups.

In medical studies it is common to have data on repeated measurements jointly with time-to-event. The interest on data analysis is sometimes on the analysis of time-to-event, allowing for correlation with a time dependent variable, or on the analysis of longitudinal outcome with potentially informative missing data. Individual longitudinal and survival models might be considered. However, the notion of joint modelling is motivated in a setting of dependent longitudinal and time-to-event data.

If the interest of inference is on the association between the response variable and the survival mechanism, the two processes have to be modelled jointly, including parameters that represent their correlation. The proposal goes to the so called joint models for longitudinal and time-to-event. These models are based on a joint distribution for the two processes, longitudinal and failure time.

2. LONGITUDINAL DATA ANALYSIS

A longitudinal data set is characterised by repeated measurements of one or more response variables on a number of subjects at a series of time points. We first introduce linear models for repeated measurements with focus on general linear mixed effects models. For the analysis of repeated measurements it is common to assume independence between subjects, to have the replication across subjects for the analysis of time trajectory. However, this assumption is not adequate for measurements within the same subject, as measurements in time from a same person tend to be correlated. Moreover, measurements from different subjects and within a same individual are also subject to measurement error.

2.1. Notation

In the context of repeated measurements of a response variable, we let Y_{ij} be a response variable measured on subject $i = 1, ..., n$ at time point t_{ij}, with $j = 1, ..., m_i$. We include a set of p explanatory variables given by the vector $\boldsymbol{x}_{ij}$ with dimension p, which can be time dependent or only measured at baseline. The full set of repeated measurements for subject i is represented by the vector $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{im_i})$, with mean $\mathrm{E}[\boldsymbol{Y}_i] = \boldsymbol{\mu}_i$, and variance covariance matrix $\mathrm{Var}(\boldsymbol{Y}_i) = V_i$ of dimension $(m_i \times m_i)$, where each element (j, k) of this matrix is the covariance $\mathrm{Cov}(Y_{ij}, Y_{ik}) = v_{ijk}$, and $\mathrm{Var}(Y_{ij}) = v_{ij}$, for $j = k$.

The most common model-based approach for longitudinal repeated measurements assumes independence between subjects i, where each measurement is a realisation of a Gaussian random variable. The linear model is based on the regression of explanatory variables:

$$Y_{ij} \;=\; \mu_i(t_{ij}) + \epsilon_{ij} \;. \tag{2.1}$$

Different models for longitudinal data differ on the correlation structure for the errors ϵ_{ij}. For the entire data set of $N = \sum_{i=1}^{n} m_i$ longitudinal measurements, we use the notation $\boldsymbol{Y} = (\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n)$ as the random variable of all measurements for all subjects, with the linear model for longitudinal measurements as

$$\boldsymbol{Y} \sim \; MVN\big(\boldsymbol{X\beta}, \boldsymbol{V}(\psi)\big) \,,$$

where $\boldsymbol{X}$ is the $(N \times p)$ design matrix of explanatory variables. The matrix $\boldsymbol{V}$, with dimension $(N \times N)$ and parameters ψ, is a block diagonal matrix, because we assume independence between subjects, with each diagonal matrix $\boldsymbol{V}_i$ representing the variance covariance matrix for subject i.

2.2. General linear mixed models

We will be using linear longitudinal models as defined previously, with ideas from [1] and [2]. The general idea of linear mixed effects models is to assume a structure for the ϵ_{ij}'s as in (2.1), separating pure measurement error from variability between and within individuals. The general linear mixed effects model is defined as

$$Y_{ij} \;=\; \mu_i(t_{ij}) + \Omega_i(t_{ij}) + Z_{ij} \;, \tag{2.2}$$

where $\Omega_i(t_{ij})$ is an unobserved random process, and Z_{ij} are independent realisations of a zero-mean Gaussian random variable with variance τ^2, representing

pure measurement error. Diggle *et al.* [2] propose to decompose the unobserved random process $\Omega_i(t_{ij})$ into two components in an additive way,

$$\Omega_i(t_{ij}) \;=\; \boldsymbol{d}'_{ij}\,\boldsymbol{U}_i + W_i(t_{ij}) \;,$$

where $\boldsymbol{U}_i$ are n independent realisations of a r-dimension multivariate Gaussian random variable with mean zero and variance covariance matrix G, and $\boldsymbol{d}_{ij}$ are r-dimension vectors of explanatory variables for the random process $\boldsymbol{U}_i$. The $W_i(t_{ij})$ are n independent realisations of a stationary Gaussian process with mean zero, variance σ^2 and correlation function $\rho(u)$, with u being time lag. The processes $\boldsymbol{U}_i$ and $W_i(t_{ij})$ are in [2] terminology random effects and serial correlation components, interpreted as the variability between and within individuals, respectively.

Notice that decomposing ϵ_{ij} in the previous additive way implies that

$$\mathrm{Var}(\epsilon_{ij}) \;=\; D_i G D'_i + \sigma^2 H_i + \tau^2 I_i \;,$$

where H_i is a matrix with (j,k) element $h_{ijk} = \rho(|t_{ik} - t_{ij}|)$.

For estimation of model parameters we will use likelihood-based methods. The full likelihood is easily available for the entire data set.

2.3. Modelling missing process in longitudinal analysis

In this section we will be referring to balanced longitudinal study designs. This meaning that the study specifies that all subjects are observed at the same equally spaced time points, the same number of times. It is common that not all subjects provide the complete set of measurements for the study, originating missing values. Therefore, we review longitudinal models that cope with potentially informative missing data. In particular, we consider longitudinal models where the event is "drop-out of the study". This will lead us to distinguish different reasons for missing values, and how they can be associated with the repeated measurement processes.

Missing values in longitudinal studies occur in two different ways. They can be missing at intermittent times in the sequence, which means that other measurements are observed following missing values; for example, when a patient does not feel well for the visit, or just forgets the appointment. The other type of missing values appear when all other values after this are also missing, and the patient is said to have dropped-out of the study (measurement sequence terminates prematurely [3]). There might be several reasons for a patient to drop-out a study such as death, feeling the treatment is not helpful to them or just moving house.

The main concern of longitudinal analysis with missing data arises when there is an association between the longitudinal profile and the missing process. For example, if a patient drops-out the study because he/she believes that the treatment is not being effective, the missing values should not be dissociated from the measurement process. Therefore, it is necessary to distinguish between different reasons for missing values, to be possible to conjecture on possible association. Little and Rubin [4] classified the nature of missing data mechanism as:

> **MCAR** — Missing Completely At Random: when the probability of missing does not depend on either the observed or unobserved measurements. For example, when a patient forgets to attend the appointment.
>
> **MAR** — Missing At Random: when the probability of missing depends on the observed data, but not on the unobserved measurements. Conditional on the observed measurements, missing process and data are independent. For example, the patients leaves the study on doctors advice based on previous observed longitudinal measurements.
>
> **MNAR** — Missing Not At Random: when the probability of missing depends on observed and unobserved data. For example, when a patient leaves the study because he/she feels ill on the day of their appointment, and the illness is related with all the longitudinal profile, including those measurements that would have been observed if they would have kept on going to the appointments.

In a setting of time-to-event, it is reasonable to consider missing values as events, and the design times at which the missing values occur as the set of possible event times. The events associated with intermittent missingness are multiple events in a same subject. However, it is commonly assumed that this type of missing data is missing completely at random, because other measurements are observed after in time. Hence, intermittent missing values are treated as ignorable and inferences can me made using likelihood based methods.

The drop-out missing value originates a single event, identified as the time that terminates the longitudinal sequence. It is usual in clinical trials to record the cause of the patient's drop-out. This information helps to identify the nature of the missing data.

Let $\boldsymbol{Y}$ be the random variable associate with the complete data vector for a single subject, that can be decompose as $\boldsymbol{Y} = (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$ with observed and missing measurements, respectively. Also, $\boldsymbol{D}$ be the missing data indicator (0/1) for the same subject, for observed and missing measurements, respectively. The model for the complete data requires the specification of the joint distribution $[\boldsymbol{Y}, \boldsymbol{D}]$, where $[\cdot]$ represents the density distribution. Using this notation, Little [5] contrasts different models for the drop-out mechanism that come in parallel with

the nature of the missing data as before. These are:

Covariate-Dependent Drop-out — when the drop-out mechanism does not depend on any longitudinal values, but is allowed to depend on the covariates:

$$[\boldsymbol{D}|\boldsymbol{Y}] = [\boldsymbol{D}] \quad \equiv \quad \textbf{MCAR} \; .$$

Missing-at-Random Drop-out — when the drop-out mechanism depends only on observed data:

$$[\boldsymbol{D}|\boldsymbol{Y}] = [\boldsymbol{D}|\boldsymbol{Y}_{\text{obs}}] \quad \equiv \quad \textbf{MAR} \; .$$

Nonignorable Outcome-Based Drop-out — when the drop-out may depend on missing components of $\boldsymbol{Y}$:

$$[\boldsymbol{D}|\boldsymbol{Y}] = [\boldsymbol{D}|\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}] \quad \equiv \quad \textbf{MNAR} \; .$$

It is proved that likelihood-based inferences on the model parameters are unbiased when ignoring the missing values of the data [6], if the data is believed to be MCAR. The standard procedure for testing for MCAR is to compare the empirical distributions of complete observed variables for respondents and non-respondents subjects, using t-tests [7].

If the likelihood function can be factorised into two independent parts, one corresponding to the response parameters and the other corresponding to missing parameters, the missing process is considered to be at least MAR with respect to the response process. Under the MAR assumption Rubin [6] shows that if the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, on the distributions $f(\boldsymbol{y}|\boldsymbol{\theta})$ and $f(\boldsymbol{d}|\boldsymbol{y}, \boldsymbol{\psi})$, are distinct, then likelihood inference is possible, by integrating out the density of $\boldsymbol{y}_{\text{mis}}$. If the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ do not have common components it is possible to factorise

$$f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\psi}) \; = \int f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\theta}) \, f(\boldsymbol{d}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \boldsymbol{\psi}) \; d\boldsymbol{y}_{\text{mis}} \; ,$$

and under the MAR assumption $f(\boldsymbol{d}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \boldsymbol{\psi}) = f(\boldsymbol{d}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\psi})$, so

$$f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\psi}) \; = \; f(\boldsymbol{d}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\psi}) \, f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta}) \; .$$

Therefore, maximisation of the likelihood for model parameters, requires the maximisation of two independent terms, that do not share common parameters. However a theorem proved by Molenberghs *et al.* [8] implies that MAR is untestable without additional assumptions no matter how much data are available. Also, Molenberghs and colleagues [9] derive the bias on parameter estimates when data is MCAR and MAR and simple methods like last observation carried forward and complete case analysis are used, and show that likelihood-base methods provide consistent estimators.

Missing values which are MCAR or MAR are known in the literature as ignorable, because longitudinal analysis can be performed ignoring them. However, missing values originated by MNAR are said to be informative or non-ignorable [10].

3. TIME-TO-EVENT DATA ANALYSIS

Time-to-event data is generated by observing several subjects until a single or multiple event occur, and the data is the waiting time. For example, in a medical context a single time-to-event is the time to recurrence of a health condition, time of response to a treatment or time to death from a certain cause. To determine time-to-event correctly, it is necessary to choose an appropriate time origin, which has to be easily identified and common for all patients. Usually, time from randomisation, time from diagnosis or time from beginning of medication is chosen. From now we will refer to failure time with the same meaning as time-to-event.

The special difficulty with time-to-event data, is that the event will not occur for some subjects during the follow-up period of the study. The only information available for these patients is a maximum time, up to which it is known not to have observed the event. For example, in a clinical trial where failure time is time to death, not all patients will die during the study. For these patients we observe a right-censored time, which in the maximum is the follow-up time of the study. The set of failure and censored times we call survival data, or sometimes observed lifetime. Therefore, the analysis of time-to-event data is also commonly called survival data analysis.

The observed censored times can represent subjects still alive when the study is finished, or subjects who drop-out of the study. We consider drop-out time, the time when a subject drops out of the study, and we use it analogously to time-to-event.

3.1. Notation

Let the random variable F denote the time-to-event and let $f_1, ..., f_n$ be a random sample from F on $i = 1, ..., n$ subjects. However, the event is not always observed and every subject i has associated a censored time, coming from a random variable C, where $c_1, ..., c_n$ is the random sample from C for the same subjects. Therefore, the observed survival data is the realisation $s_i = \min\{f_i, c_i\}$, $i = 1, ..., n$, of a random variable S.

A common assumption in survival analysis is non-informative censoring, meaning that random variables C and F are independent. Therefore, if $F \leq C$, $S = F$ and a failure time is observed, if $C < F$, $S = C$ and a censoring time is observed. The observed data are realisations (s_i, δ_i), where s_i is defined as before, and δ_i is a subject indicator (1/0), for failure or censored time, respectively.

To describe the distribution of failure time f, it is more appropriate to use the survival and hazard functions. The survival function $S(t)$ is defined as the probability of failure time being beyond some point t, $S(t) = \mathrm{P}(F > t)$. The hazard function is the probability of failure time occur in the next short period of time, given that failure time did not occur up to that time and all the past history,

$$\lambda(t) \;=\; \lim_{\Delta t \to 0^+} \frac{\mathrm{P}\big(t \leq F < t + \Delta t \,|\, F \geq t\big)}{\Delta t} \;,$$

and is defined as the instantaneous death rate for an individual surviving to time t. It is possible to combine the two definitions and get the relation

$$\lambda(t) \;=\; \frac{\mathtt{f}(t)}{S(t)} \;, \tag{3.1}$$

where $\mathtt{f}(t)$ is the density function of F.

For observed survival data (s_i, δ_i) on subjects $i = 1, ..., n$, the likelihood function of model parameter is the product of probabilities given the observed data, for all subjects i. Usually the censoring mechanism is ignored [11] and the likelihood of interest is

$$L(\theta; \boldsymbol{s}, \boldsymbol{\delta}) \;\propto\; \prod_i \mathtt{f}(s_i)^{\delta_i} \times S(s_i)^{1-\delta_i} \;, \tag{3.2}$$

where each failure time contributes with the density function and each censored time contributes with the survival function.

3.2. Survival models

When modelling survival data, the most common non-parametric method is the product-limit estimator [12], sometimes called the Kaplan–Meier estimator. Consider the ordered subset of $k \leq n$ unique observed failure times from the observed survival times, $s_{(1)} < ... < s_{(k)}$. Let d_i be the number of failures which occur at t_i and n_i the number of individuals who are at risk just before time t_i, making up the risk set $R(t_i)$, say. Notice that n_i represents the number of subjects that survive at least until time t_i. Therefore, d_i/n_i is an estimate of the probability of failure at time t_i, conditional on surviving up to t_i. The product-limit

estimator is then defined as

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < s_{(1)} \ , \\ \prod_{s_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) & \text{if } t \geq s_{(1)} \ . \end{cases} \tag{3.3}$$

The most common way to model survival data is through the hazard function, including a set of q explanatory variables $\boldsymbol{W}$ measured at baseline to predict failure time. The most widely use semi-parametric model is the so called Cox proportional hazards model [13],

$$\lambda(t|\boldsymbol{W}) = \lambda_0(t) \exp(\boldsymbol{W}'\boldsymbol{\alpha}) \ , \tag{3.4}$$

where $\boldsymbol{\alpha}$ is a $(q \times 1)$ vector of parameters to estimate, and $\lambda_0(t)$ is the unknown hazard function at the baseline variables $\boldsymbol{W} = \boldsymbol{0}$. This is a semi-parametric model, because the baseline covariates are modelled parametrically whereas the baseline hazard function is modelled non-parametrically with no specific form. The function $\lambda_0(t)$ is considered a nuisance parameter in the Cox proportional hazards analysis. Therefore, when writing the likelihood function for this model as in (3.2), it is not possible to estimate simultaneously the baseline hazard function and parameters of interest $\boldsymbol{\alpha}$. Consequently, Cox [14] suggests an estimation method based on conditional probabilities at the set of failure times, which is based on maximising the partial likelihood. The main advantage of the partial likelihood is that it does not depend on the baseline proportional hazard function $\lambda_0(\cdot)$, and only parameters of explanatory variables are estimated. For complete details on parameter estimates in the partial likelihood and score function vector see [15].

3.3. Time dependent covariates in time-to-event analysis

When the interest is on inference for the model parameters of a time-to-event process, we allow for survival data analysis, which deals with censored event times. The most popular model is the Cox proportional hazards model, where the hazard of an individual with some covariates is proportional to a baseline function of time [13], as discussed before. This model allows for fixed covariates that do not change over time [16], and parameters are estimated by maximising the partial-likelihood [14]. However, it is often the case that time-dependent covariates are available and these also want to be included in the survival model.

The Cox proportional hazards model can be extended to incorporate the observed time-dependent covariates [17], with the partial likelihood evaluated at

each event time in the form

$$\prod_{t_i:\ \text{event time}} \frac{\exp\{\boldsymbol{W}_i'(t_i)\,\boldsymbol{\alpha}\}}{\sum_{j\in R(t_i)} \exp\left\{\boldsymbol{W}_j'(t_i)\,\boldsymbol{\alpha}\right\}}\ , \tag{3.5}$$

where $R(t_i)$ represents the set of subjects at risk at event time t_i. This model is described in [16] Chapter 8, and the efficiency of the parameter estimates using partial likelihood is compared with those obtained from a fully parametric model. Moreover, Hougaard [11] in section 2.4.4 argues that time dependent covariates have to be predicted, which means the trajectory of the covariate has to be known at every time points.

There are also ways for handling with missing time dependent covariates, as in longitudinal models mention in the previous section. Lin and Ying [18] estimate from the subjects with complete measurements, the conditional expectation of missing covariates at all time points. Thus, in the parameter estimating equations this is subtracted to the observed covariate. They claim this method is generally more efficient than using only subjects with complete data. However, it is stated that the validity of the method *"depends critically on the MCAR assumption"*.

Paik and Tsai [19] suggest a very similar estimator, with the advantage that is consistent under the missingness mechanism. But also in this work, the authors conclude that when the missing probability depends on unobserved values of the covariate, their estimator is biased.

If we consider longitudinal measurements as time dependent covariates, it is ignored that these are measured with error, and the observe measurements are a noisy version of true process. A drawback of the previous methods is that they do not account for measurement error in the repeated measurements. Prentice [20] shows that regression coefficients on the partial likelihood are asymptotically biased when it accommodates covariates measured with error, and he suggests a modified partial likelihood, using conditional expectations on the relative observed hazard. Altman and DeStavola [21] review the different problems of including time dependent covariates measured with error in survival data analysis. Following this, models for the joint distribution of time-to-event and longitudinal response variables have been proposed, included in the so called area of joint analysis of longitudinal and time-to-event outcomes.

4. JOINT MODELLING

In the context of joint modelling it is necessary to establish a clear framework to distinguish terminology from longitudinal and time-to-event processes.

Two processes are considered, the longitudinal $\boldsymbol{Y}$ and time-to-event $\boldsymbol{F}$ processes with possible association, which we are interested in. Another common issue in any data set is the missing process generating missing data. Therefore, the missing data can be missing of longitudinal measurements or missing to observe the event. When the event is not observed the missing process is called censoring $\boldsymbol{C}$, and missing of longitudinal measurements is called missing data $\boldsymbol{D}$. The censoring process $\boldsymbol{C}$ is usually assumed non-informative, in the sense that is considered independent of the time-to-event and longitudinal processes.

The missing of longitudinal measurements can be intermittent or terminating the sequence of longitudinal measurements, as discussed in section 2. In the case of intermittent missingness we assume these MCAR and it is known that these can be ignored in the likelihood function. If a missing longitudinal measurement terminates the sequence of longitudinal measurements, we call it a drop-out time from a drop-out process $\boldsymbol{D}$, as corresponds to a subject drop-out of the study. Moreover, the drop-out process cannot be ignored in most of the cases and it is considered to be MNAR.

We then have four processes, longitudinal process $\boldsymbol{Y}$, drop-out process $\boldsymbol{D}$, time-to-event process $\boldsymbol{F}$ and censoring process $\boldsymbol{C}$, and assumptions on possible associations on these processes is necessary. For example, we might assume that the event of interest is drop-out time, and so processes $\boldsymbol{D}$ and $\boldsymbol{F}$ are the same. This is an assumption of many clinical trials, as there is no record of an actual time-to-event, and the time of the last observation is considered the failure time. Furthermore, the missing longitudinal measurements caused by drop-out time are allowed to be associated with the event time, that is MNAR.

Different associations are possible between the four processes. We will consider the situation that time-to-event is available in the data set and the event generates missing data. Therefore, the time-to-event process completely determines the drop-out process. Figure 1 represents graphically this situation.

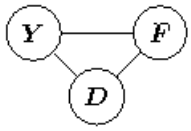

Figure 1: Graphical representation of possible associations between longitudinal, time-to-event and drop-out processes.

Another situation would be the longitudinal process associated with both processes time-to-event and drop-out, but these not associated with each other.

4.1. Model based joint models

We review full likelihood methods for exact estimation of model parameters in the joint distribution of repeated measurements and time-to-event. These are what we will call *joint models*, which model the joint distribution $[\boldsymbol{Y}, \boldsymbol{F}]$, for $\boldsymbol{Y}$ and $\boldsymbol{F}$ the random variables of repeated measurements and failure time, respectively. Inference on model parameters is done through the decomposition of the full likelihood. Nevertheless, it is not clear a joint distribution for the two random variables. Therefore, in joint models the joint distribution is factorised using Bayes rule.

The two different factorisations of the joint distribution generates different model strategies that contrast model interpretations, and consequently their suitability for individual problems. These are pattern-mixture and selection models [4, 22] that factorise:

pattern-mixture models	selection models
$[\boldsymbol{Y}, \boldsymbol{F}] = [\boldsymbol{F}]\,[\boldsymbol{Y}\vert\boldsymbol{F}]$	$[\boldsymbol{Y}, \boldsymbol{F}] = [\boldsymbol{Y}]\,[\boldsymbol{F}\vert\boldsymbol{Y}]$.

The parameters involved in each of the model components have different interpretations, in one model they are the parameters of the conditional distribution in the other they are the parameters of the marginal distribution. Depending on the context, the parameters of interest for inference will also be different. Notice that, if event is drop-out, $\boldsymbol{F} = \boldsymbol{D}$ in the terminology of section 2, and if the missing process is MCAR the two model strategies are equivalent, as the two processes are independent.

The model strategy depends mostly on the nature of the statistical problem and the scientific questions to be answered. Although mathematically the models describe exactly the same joint distribution, they have different statistical interpretations. Selection models are mainly used when inference is on time-to-event model parameters, improving the inference by allowing for correlation in the longitudinal measurements. In opposition, when primary interest is on the longitudinal trajectory, which might be associated with an event pattern, the pattern-mixture models are more commonly used. Therefore, the two different approaches lead to different understanding and inferences of the model parameters, together with different views on how to store the data.

Pattern-mixture models stratify regression models by missing pattern cohort, then model the marginal distribution of the response as a mixture of distributions over the patterns. These models are useful as an exploratory tool to check on longitudinal profile differences between drop-out groups. Selection models assume a model for the complete longitudinal data and then multiply by the

probability of observing the event given the complete data, though the observed data does not match the complete data.

Selection models can be seen as an alternative to pattern-mixture models for data with many complex missing patterns. The terminology of these models is clear for pattern-mixture models, they model a mixture of conditional distributions each for each missing pattern data. For selection models they model the selection of drop-outs condition on the measurement history.

The models above can be extended to incorporate random effects, in this case they are called random pattern-mixture models and random selection models. The individual unobserved random effects in the selection models are included in the marginal longitudinal model, whereas in the pattern-mixture models these come in the marginal distribution of the event times. Therefore, when jointly modelling repeated measurements $\boldsymbol{Y}$, event times $\boldsymbol{F}$ and random effects $\boldsymbol{U}$, the joint distributions are:

random pattern-mixture models	random selection models
$[\boldsymbol{Y}, \boldsymbol{F}, \boldsymbol{U}] = [\boldsymbol{U}]\ [\boldsymbol{F}\|\boldsymbol{U}]\ [\boldsymbol{Y}\|\boldsymbol{F}]$	$[\boldsymbol{Y}, \boldsymbol{F}, \boldsymbol{U}] = [\boldsymbol{U}]\ [\boldsymbol{Y}\|\boldsymbol{U}]\ [\boldsymbol{F}\|\boldsymbol{Y}]\ .$

Diggle [23] defines one different class of joint models, these as *random effects models*. Random effects joint models assume that both repeated measurements and event time depend on a unobserved random effect, these specified through a certain bivariate distribution. The random effects joint model is described by assuming conditional independence between $\boldsymbol{Y}$ and $\boldsymbol{F}$ given the random effects $\boldsymbol{U} = (U_1, U_2),$ as

random effects model

$$[\boldsymbol{Y}, \boldsymbol{F}, \boldsymbol{U}] = [\boldsymbol{U}]\ [\boldsymbol{Y}|U_1]\ [\boldsymbol{F}|U_2]\ .$$

In random effects joint models the association between longitudinal measurements and time-to-event is completely determine by the correlation structure between the two random effects U_1 and U_2. The three different strategies to model the joint distribution, can be distinguish visually by diagrams presented by Diggle [23] and shown here in Figure 2.

The diagrams in Figure 2 represent conditional independence graphs for the three random variables. The absence of an edge indicates conditional independence between the two vertices of the edge, given the third vertice involved in the graph. In Figure 2(a) it is represented the saturated model, where all the associations are possible. Figure 2(b) represents selection models, where longitudinal measurements are influenced by their individual random effects, and it is the realisation of the measurement process that will influence the event, and not the random effect. On the contrary, in pattern-mixture models in Figure 2(c)

the individual random effects will determine the time of event, which after being predefined develops the individual longitudinal profile with some error. Regarding random effects joint models, Figure 2(d) suggests that both processes are a joint response to an unobserved individual specific process, and conditional on the responses being independent of each other.

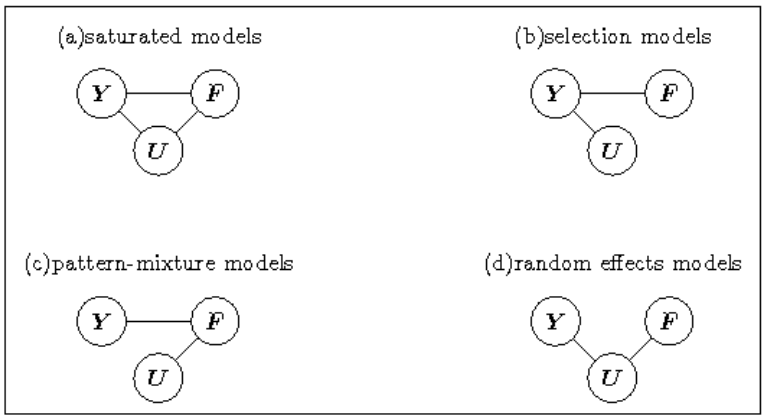

Figure 2: Graphical representation of saturated, selection, pattern-mixture and random effects models as in [23].

Little [5] produces a detailed review on selection and pattern-mixture models, when different missing mechanisms are present, in parallel with examples of data sets. Hogan and Laird [24] give a good comparison between pattern-mixture and selections models, and refers to random effect pattern-mixture and random effect selection models, but not with the same definition as we introduce here. We refer to this below, when giving examples of random effects models.

4.1.1. Pattern-mixture models

In mixture models it is necessary to specify a model for the marginal distribution of the event times $[\boldsymbol{F}]$ and the conditional distribution of $[\boldsymbol{Y}|\boldsymbol{F}]$. For the former, standard distributions would be the multinomial, or through modelling the hazard function with a Cox model, additive model or accelerated life model. The latter distribution is not always established, as the sample space of drop-out patterns can be integrated out of the conditional distribution, and inferences are made directly on the marginal distribution of $[\boldsymbol{Y}]$. The main goal of pattern-mixture models is to adjust the inference about $\boldsymbol{Y}$ for the effects of drop-out, with the convenience of not having to specify the event time marginal distribution.

One of the first pattern-mixture models was proposed by Wu and Bailey [25] whose aim was to compare rates of change of a continuous variable under informative missingness, for k different treatment groups. A conditional linear random effects model is proposed for the continuous variable, where the random effects $\boldsymbol{U}_i = (U_{1i}, U_{2i})$ are conditional on the event time. Especially the random slope is a polynomial function of some degree, on the event time,

$$(U_{2i}|\boldsymbol{F}_i = f_i) \;=\; \sum_{l=0}^{L} \gamma_{lk}\, s_i^l \;.$$

Two alternative estimation methods based on simple linear regression are proposed for the expected values of the random slope in the k group, namely linear minimum variance unbiased estimator and linear minimum mean squared errors estimator. These are compared by simulation studies with three other estimates and found to be more efficient. Testing for MNAR in this model, corresponds to test for the alternative hypothesis $H_A\colon \gamma_l \neq 0$, for all l, when the null hypothesis is $H_0\colon \gamma_l = 0$, for all l. Wu and Bailey [26] consider the particular case of $l = 1$.

Little [22] proposes a pattern-mixture model, where the drop-out patterns are considered realisations of a multinomial model, such that model parameters are the probabilities of having each of the drop-out patterns. For each drop-out pattern, the conditional distribution of the longitudinal measurements are assumed to be multivariate Gaussian. Usually, pattern-mixture models have a large number of parameters due to possible high number of drop-out patterns, which can cause identifiability problems. Therefore, it is mentioned the need to have at least one more longitudinal complete case than number of response variables to obtain consistent estimates for each pattern.

For a saturated model, where each multivariate distribution has distinct parameters, not all the parameters are possible to estimate. For example, the model parameters of missing patterns with no observed measurements, will not come out in the likelihood. The work of Little proposes parameter restrictions to the conditional models, which would reflect a certain missing process. The same approach is extended to categorical response variables, where the multinomial distribution is defined through a contingency table. However, it is noticed that the methodology can be inefficient because it requires a reasonable large number of complete cases.

The model proposed by Little does not allow for observed censored survival times, because it specifically models the probability of observing an event. Nevertheless, this can be extended to different multinomial distributions for each subsets of subjects with failure and censored times. This is a reasonable model when we want to assume independence between censoring and failure processes.

The work by Hogan and Laird [27] is motivated by limitations of the previous models that assume fully observed events, and parametric models for the drop-out process. Accordingly, the distribution $[\boldsymbol{F}]$ is specified non-parametrically by estimating multinomial probabilities with incomplete data, and the conditional distribution $[\boldsymbol{Y}|\boldsymbol{F}]$ is assumed a linear regression with individual random effects $\boldsymbol{b}_i$.

An advantage of using a Gaussian model for the conditional distribution is that the unconditional distribution is a mixture of conditional normal distributions. When the event times are completely observed the maximum likelihood estimates are easily obtained by maximising the likelihood.

The work of Li and Schluchter [28] examines different conditional models for the random effects $\boldsymbol{b}_i|\boldsymbol{F}_i$. Firstly, they consider a conditional quadratic and linear models, where random effects follow quadratic or linear regression curves on survival times. Secondly, a pattern-mixture model is described in the general form of the mixed effects model, using a single parameter for each missing pattern. These models differ on the design matrix $\boldsymbol{d}_i$, and many other models can be defined depending on the observed survival time allowing for censoring. For example, for a non-parametric model it is possible to use a piecewise linear or spline model.

In the opinion of Hogan and Laird [24] pattern-mixture models appeared to approximate selection models, as these are difficult to fit, have problems with identifiability and sensitivity to parametric assumptions. In addition, they considered pattern-mixture models not very appealing, because they mainly focus on the stratification of the sample by time of drop-out. However, they consider their main advantage over selection models, to be able to integrate out the cumulative distribution of $\boldsymbol{F}$. Therefore, it is possible to make inference on the marginal longitudinal parameters, without specifying a model for the drop-out process. When specifying a model for the drop-out, non-parametric estimators are usually use, like Kaplan–Meier as in [27].

All pattern-mixture models presented here make the assumption that cannot be verified of $f_{Y_{\text{obs}}|F} = f_{Y|F}$, which is not equivalent to $f_{Y_{\text{obs}}} = f_Y$ [27]. Other authors discuss more carefully about identification problems of pattern-mixture models, generated by unverifiable assumptions between the distribution of the complete data and only the observed measurements.

Thijs and co-authors [29] look at sensitivity analysis for pattern-mixture models, and propose three different strategies to fit pattern-mixture models under identified restrictions. The model strategies allow extrapolation beyond the time of drop-out, and inference on the distribution of the unobserved outcomes given the observed ones is possible. When restrictions are made it is plausible to perform a sensitivity analysis as the model assumptions are well identified. Birmingham *et al.* [30] present three class of restrictions that identify marginal distributions of the outcome, and are comparable to restrictions in selection models.

4.1.2. Selection models

In selection models the marginal distribution of longitudinal measurements is modelled, whereas the model for event time is conditional on the response variable. The most common approach for the distribution of repeated measurements is a linear mixed effects model, usually only random intercept and slope. Generally, choices for the conditional time-to-event distribution are logistic linear regression, probit regression, where probabilities are modelled as function of some longitudinal measurements. However, in some works the Cox proportional hazards model and accelerated life model are proposed.

One of the earliest proposals on selection models, as defined here, is [31] where the event time depends directly on the repeated measurements. The probability of drop-out at any time t_k is a parametric logistic linear model, with regression parameters on all the history of the observed measurements $(y_1, ..., y_{k-1})$ and on the unobserved measurement at drop-out time y_k, that is,

$$\text{logit}\big\{P(\boldsymbol{F} = k|\boldsymbol{y})\big\} \;=\; \beta_0 + \beta_1 y_k + \sum_{j=2}^{k} \beta_j \, y_{k+1-j} \;.$$

If $\beta_1 = \beta_2 = ... = \beta_k = 0$ the missing process is completely at random, and if only $\beta_1 = 0$ and all other different from 0 the missing values are missing at random.

This same model is used in [23], and likelihood ratio tests are performed on the β parameters to test for random drop-out and informative drop-out. A drawback of this model is the restriction on a monotone drop-out. This model does not deal with censoring times, but the model can be extended to accommodate for that.

Scharfstein and colleagues [32] refer to a general logistic regression model, in the context of a selection model, for the probability of drop-out given the complete vector of measurements,

$$\text{logit}\big\{P(\text{drop-out}|\boldsymbol{Y})\big\} \;=\; \beta_0 + q(\boldsymbol{Y}) \;,$$

where $q(\cdot)$ can be any function. For the particular application, they consider the class of functions $\mathcal{Q} = \{\alpha \log(\boldsymbol{Y}) \colon \alpha \in \mathbb{R}\}$, where α is a selection bias parameter. That is, it is possible to test for the value of α to be zero to understand the missing process, as before for the values of β.

In this paper the advantage of having a methodology that depends on a general function $q(\cdot)$ is discussed. The flexibility of function $q(\cdot)$ quantifies the influence of the response on the probability to drop-out, which allows a straightforward sensitivity analysis, where different assumptions can be tested.

Most of the joint models described here model repeated measurements data with the popular parametric linear mixed effects model. However, in many applications the data may not fit well by linear models, or it is of interest to model the response non-parametrically. Brown *et al.* [33] propose a cubic B-spline to model the longitudinal data, so that there is no parametric assumption on the trajectory of subject's longitudinal profile. This approach is relevant when inference on the effect of the longitudinal measures on the time-to-event is of interest, but not on the longitudinal process or its trajectory over time. Therefore, this approach allows a much more flexible modelling of the longitudinal data. Bayesian methods are used for the estimation, and the B-spline is extended to accommodates estimation on multiple response variables.

4.1.3. Random effects models

There are many models called selections models, that we include in our classification of random effects models. Although, the conditional distribution of time-to-event is modelled, this is conditional on a latent process, and the longitudinal and time-to-event processes are assumed independent conditional on the latent process. Therefore, we include these models in the class of random effects joint models. These models are also called shared parameter models, because the longitudinal response and missing mechanisms are modelled by sharing random effects.

In random effects joint models we assume both event time and longitudinal process dependent on a underlying disease or illness progression, defined by a random effect, rather than to the actual outcome. Moreover, the two processes are independent conditional on the unobserved random effects. For example, in [24] the joint distribution is defined as

$$[\boldsymbol{Y}, \boldsymbol{F}] = \int_{\boldsymbol{U}} [\boldsymbol{U}]\ [\boldsymbol{Y}|\boldsymbol{U}]\ [\boldsymbol{F}|\boldsymbol{U}]\ d\boldsymbol{U}\,.$$

Wu and Carroll [34] propose a random effects model called an "informative right censoring" model. In this model it is assumed that the repeated measurements follow a linear mixed effect model, with subject specific random effects, in particular random intercept and random slope. They further assume a general density function $M(t)$ for the failure process, conditional on the subject specific random effects. In particular they use a probit drop-out model for $M(t)$, and estimation of model parameters is obtained by maximising a pseudo-likelihood.

Testing for non-informative missingness under this model, is equivalent to test for the regression parameters that relate the conditional probability with the random effects to be zero. A test statistic is proposed for testing for non-informative missing process.

The model propose by Schluchter [35] is also a random effects model, as they use a trivariate Gaussian distribution to model the joint distribution of logarithm transformation of time-to-event, and the random intercept and random slope. These are random effects in a linear mixed effects model for the longitudinal measurements, and it is assumed that the event time is associated with an underlying process that is unobserved. The parameter estimates are acquired with an EM algorithm on the complete log-likelihood of the parameters given the observed data.

Some of the advantages of this model enumerated by the author are, the allowance for unbalanced data due to staggered entries or unequally-time visits, it is possible to use all the data available and possibility to apply likelihood ratio tests on the model parameters. In particular, on the correlation parameters of the trivariate distribution, which represent the association between random effects and event time. However, there is the computational disadvantage of this model, that may require large amounts of data to obtain convergence in the EM algorithm. We think this models simplifies the association structure, by only having two cross-correlation parameters, between the event time and two random effects, initial value and slope.

DeGruttola and Tu [36] propose to extend the two random effects joint model to include a general structure for the random effects. The conditional distribution of any transformation of failure times is modelled as a linear mixed effects model, and longitudinal and time-to-event processes share the random effects. Thus, this model assumes that both processes are measurements with error of the same unobserved latent process that represent health deterioration. The estimation of the model parameters is by an EM algorithm.

In all random effects models mentioned before, the survival time is modelled parametrically. Another very popular approach to the conditional distribution of the event times is by semiparametric survival models, such as the Cox proportional hazards model. Faucett and Thomas [37] propose one of the first random effects models with proportional hazards for the event time. They consider the joint analysis of longitudinal measurements and survival time as the joint distribution of two models, covariate tracking model and disease risk model. The former models the longitudinal response as a linear mixed model with subject specific random effects, intercept and slope, as in a linear growth curve model,

$$Y_{ij} \;=\; U_{1i} + U_{2i}\, t_{ij} + \epsilon_{ij} \;.$$

The latter allows a Cox proportional hazards model for the disease risk, with the same random effects as in the longitudinal model, assuming that these describe the true latent process

$$\lambda(t|\boldsymbol{U}_i) \;=\; \lambda_0(t)\, \exp\big\{\beta(U_{1i} + U_{2i}t)\big\}\,.$$

Wulfsohn and Tsiatis [38] consider the same model as Faucett and Thomas, as an alternative to the two-stage model. In this proposal, they notice that the normality of the random effects is on the overall subjects, and constant over time, which does not imply normality on the random effects of the subjects at risk at a certain time point.

In the two stage model the random effects are estimated in the first stage, by fitting a longitudinal model, and these are input to the Cox proportional hazards model in a second stage. In the model propose by Wulfsohn and Tsiatis [38] the parameters are estimated using all the information available at each time point, by maximising the full likelihood of the joint distribution. Although, the models by Faucett and Thomas and by Wulfsohn and Tsiatis are the same, they use different approaches for the parameter estimation. Faucett and Thomas follow a MCMC approach, with a Gibbs sampling whereas Wulfsohn and Tsiatis use an EM algorithm for the estimation.

Henderson and colleagues [39] propose an extension to the previous model, including a Gaussian stochastic process to each longitudinal response linear model

$$Y_{ij} \;=\; \mu(t_{ij}) + \Omega_{1i}(t_{ij}) + \epsilon_{ij}$$

and event time hazard model

$$\lambda(t|\boldsymbol{\Omega}) \;=\; \lambda_0(t)\,\exp\big\{\alpha(t) + \Omega_{2i}(t)\big\}\,.$$

The stochastic processes are components of a bivariate Gaussian process $\boldsymbol{\Omega}(t) = \{\Omega_1(t), \Omega_2(t)\}$. This is an extension of the previous model as each Gaussian stochastic process is assumed as in [40],

$$\Omega_{1i}(t) \;=\; \boldsymbol{d}_i \boldsymbol{U}_i + W_i(t)\;,$$

where $\boldsymbol{U}_i$ are the associated Gaussian random effects and $W_i(t)$ is a stationary Gaussian process that introduces serial autocorrelation. The last component is not considered in any of the previous models. It is then assumed that both processes are independent given $\boldsymbol{\Omega}(t)$. Therefore, the association between the two processes is interpreted by the correlation between the two latent variables. Moreover, in the absence of association between the two processes, the analysis becomes as two independent longitudinal and survival analyses.

Guo *et al.* [41] propose a model which they call a random pattern-mixture model, that incorporates aspects from both selection and mixture models. This model considers random subject-specific effects on the conditional longitudinal response, as in most of the cases, and a random pattern specific effects $\boldsymbol{V}$. The model implies the factorisation

$$[\boldsymbol{Y}, \boldsymbol{F}, \boldsymbol{U}] \;=\; \int_{\boldsymbol{V}} [\boldsymbol{V}]\;[\boldsymbol{F}|\boldsymbol{V}]\;[\boldsymbol{U}|\boldsymbol{V}]\;[\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{V}]\;d\boldsymbol{V}\,.$$

5. DISCUSSION

The approach for the analysis of repeated measurements and time-to-event data, depends on research interests and on the assumptions we are ready to make on the available data. We have seen how these models can incorporate extra information. However, the assumption on association between processes need to be tested and we have seen that, for example MAR is not testable without additional assumptions. If the primary interest is on the time-to-event process, repeated measurements are used as time-dependent covariates in a Cox proportional hazards model. Conversely, if the interest is to make inference on the longitudinal profile, the missing pattern has to be considered.

We reviewed different methods for joint modelling of longitudinal and time-to-event data, based on the full likelihood of the joint distribution of the two processes. Different factorisations of the joint distribution lead to different model interpretations, namely pattern-mixture and selection models. We argue the approach of the analysis depends on scientific questions that need to be answered, and on the nature of association between processes. Cox [42] describes four different types of relation between a longitudinal process and failure times, not only in medical context and shows the implication of these on appropriate analysis in each case.

The model specification of selection models is more intuitive, usually with linear mixed effects model for the marginal distribution of the repeated measurements and a proportional hazards model for the conditional time-to-event distribution. However, these models usually involve intensive computational methods, as numerical integration and convergence difficulties.

The most common model for the longitudinal response variable, in pattern-mixture and selection models, is the general linear mixed effects model. Though, we notice the model proposals mainly differ in the random effects to use. Tsiatis and Davidian [43] give an interesting discussion on the philosophical issues of which of the fixed effects, random effects and stochastic processes should be included to model the longitudinal measurements. Their arguments are mainly related with biological processes that are involved in the specific data sets.

In particular we are not aware of pattern-mixture models that include a stochastic process in the conditional distribution of the longitudinal measurements. This could be related with model restrictions to include a stochastic processes on a conditional distribution which already has a time dependent process.

In this work, the focus is on the informative missingness of longitudinal measurements, due to an event. However, subjects do not always experience the event, and a censoring time is the only information available. The censoring

mechanism is always assumed non-informative and independent of time-to-event and longitudinal processes. In more complex models the censoring mechanism can be considered informative with an associated distribution, which would imply different models.

ACKNOWLEDGMENTS

This work has been supported by the grant number SFRH/BD/10266/2001 from FCT-Portugal, that supported the PhD of the author.

REFERENCES

[1] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, **38**, 963–974.

[2] Diggle, P.J.; Heagerty, P.J.; Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, Oxford University Press, Oxford (second edition).

[3] Diggle, P.J. (1989). Testing for random dropouts in repeated measurement data, *Biometrics*, **45**, 1255–1258.

[4] Little, R.J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, John Wiley, New York (second edition).

[5] Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association*, **90**(431), 1112–1121.

[6] Rubin, D.B. (1976). Inference and missing data, *Biometrika*, **63**(3), 581–592.

[7] Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data, *Journal of the American Statistical Association*, **77**(378), 270–280.

[8] Molenberghs, G.; Michiels, B.; Kenward, M.G. and Diggle, P.J. (1998). Monotone missing data and pattern-mixture models, *Statistica Neerlandica*, **52**(2), 153–161.

[9] Molenberghs, G.; Thijs, H.; Jansen, I.; Beunckens, C.; Kenward, M.G.; Mallinckrodt, C. and Carroll, R.J. (2004). Analyzing incomplete longitudinal clinical trial data, *Biostatistics*, **5**(3), 445–464.

[10] Hogan, J.W. and Laird, N.M. (1996). Intention-to-treat Analyses for incomplete repeated measures data, *Biometrics*, **52**, 1002–1017.

[11] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Statistics for Biology and Health, Springer, London.

[12] Kaplan, E.L. and Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Society*, **53**, 457–481.

[13] Cox, D.R. (1972). Regression models and life tables (with Discussion), *Journal of the Royal Statistical Society – series B Methodological*, **34**(2), 187–220.

[14] Cox, D.R. (1975). Partial likelihood, *Biometrika*, **72**(2), 269–276.

[15] Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis – Techniques for Censored and Truncated Data*, Springer, New York.

[16] Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*, Monographs on Statistics and Applied Probability, 21, Chapman and Hall, London.

[17] Therneau, T.M. and Grambsch, P.M. (2000). *Modeling Survival Data – Extending the Cox Model*, Statistics for Biology and Health, Springer, London.

[18] Lin, D.Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements, *Journal of the American Statistical Association*, **88**(424), 1341–1349.

[19] Paik, M.C. and Tsai, W.-Y. (1997). On using the Cox proportional hazards model with missing covariates, *Biometrika*, **84**(3), 579–593.

[20] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika*, **69**(2), 331–342.

[21] Altman, D.G. and DeStavola, B.L. (1994). Practical problems in fitting a proportional hazards models to data with updated measurements of the covariates, *Statistics in Medicine*, **13**, 301–341.

[22] Little, R.J.A. (1993). Pattern-Mixture models for multivariate incomplete data, *Journal of the American Statistical Association*, **88**(421), 125–134.

[23] Diggle, P.J. (1998). *Dealing with missing values in longitudinal studies*. In "Recent Advances in the Statistical Analysis of Medical Data" (B.S. Everitt and G. Dunn, Eds.), Arnold, London, 203–228.

[24] Hogan, J.W. and Laird, N.M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data, *Statistics in Medicine*, **16**, 259–272.

[25] Wu, M.C. and Bailey, K. (1989). Estimation and comparison of changes in the presence of informative right censoring:conditional linear model, *Biometrics*, **45**, 939–955.

[26] Wu, M.C. and Bailey, K. (1988). Analysing changes in the presence of informative right censoring cause by death and withdrawal, *Statistics in Medicine*, **7**, 337–346.

[27] Hogan, J.W. and Laird, N.M. (1997). Mixture models for the joint distribution of the repeated measurements and event times, *Statistics in Medicine*, **16**, 239–257.

[28] Li, J. and Schluchter, M.D. (2004). Conditional mixed models adjusting for non-ignorable drop-out with administrative censoring in longitudinal studies, *Statistics in Medicine*, **23**, 3489–3503.

[29] Thijs, J.; Molenberghs, G.; Michiels, B.; Verbeke, G. and Curran, D. (2002). Strategies to fit pattern-mixture models, *Biostatistics*, **3**(2), 245–265.

[30] Birmingham, J.; Rotnitzky, A. and Fitzmaurice, G.M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns, *Journal of the Royal Statistical Society – series B Methodological*, **65**(1), 275–297.

[31] Diggle, P.J. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion), *Applied Statistics*, **43**(1), 49–93.

[32] Scharfstein, D.O.; Daniels, M. and Robins, J.M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes, *Biostatistics*, **4**(4), 495–512.

[33] Brown, E.R.; Ibrahim, J.G. and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival, *Biometrics*, **61**, 64–73.

[34] Wu, M.C. and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics*, **44**, 175–188.

[35] Schluchter, M.D. (1992). Methods for the analysis of informatively censored longitudinal data, *Statistics in Medicine*, **11**, 1861–1870.

[36] DeGruttola, V. and Tu, X.M. (1994). Modelling progression of CD4-Lymphocyte count and its relationship to survival time, *Biometrics*, **50**, 1003–1014.

[37] Faucett, C.L. and Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach, *Statistics in Medicine*, **15**, 1663–1685.

[38] Wulfsohn, M.S. and Tsiatis, A.A. (1997). A Joint model for survival and longitudinal data measured with error, *Biometrics*, **53**, 330–339.

[39] Henderson, R.; Diggle, P.J. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data, *Biostatistics*, **1**(4), 465–480.

[40] Diggle, P.J. (1988). An approach to the analysis of repeated measurements, *Biometrics*, **44**, 959–971.

[41] Guo, W.; Ratcliffe, S.J. and Have, T.T. (2004). A random pattern-mixture model for longitudinal data with dropouts, *Journal of the American Statistical Association*, **99**(468), 929–937.

[42] Cox, D.R. (1999). Some remarks on failure-times, surrogate markers, degradation, wear, and the quality of life, *Lifetime Data Analysis*, **5**, 307–314.

[43] Tsiatis, A.A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview, *Statistica Sinica*, **14**, 809–834.

REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 83–98

INFERENCE FOR NON-MARKOV MULTI-STATE MODELS: AN OVERVIEW

Authors: Luís Meira-Machado
– Department of Mathematics and Applications,
University of Minho, Portugal
lmachado@math.uminho.pt

Abstract:

- In longitudinal studies of disease, patients can experience several events across a follow-up period. Analysis of such studies can be successfully performed by multi-state models. This paper considers nonparametric and semiparametric estimation of important targets in multi-state modeling, such as the transition probabilities and bivariate distribution function (for sequentially ordered events). These estimators are shown to be consistent even for data which is non-Markov. We illustrate the methods on two data sets.

Key-Words:

- *bivariate censoring; Markov property; multi-state models; Kaplan–Meier; presmoothing; transition probabilities.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

In many cancer studies, the main outcome under assessment is the time to death. However, other types of events can be observed during the follow-up period. For example, in colon cancer studies more than one event is often observed such as "local recurrence", "distant metastasis" and "dead". The occurrence of these intermediate events often affect patient's prognosis and can be modeled using a Cox proportional hazards model with a time-dependent covariate. Alternatively, a natural way to model such data is by using a multi-state model with states based on the values of these categorical-valued time-dependent covariates.

A multi-state model is a model for a stochastic process which occupies one of a set of discrete states at any time. These models are well adapted for modeling complex event histories (Andersen *et al.* [1]; Hougaard [2]; Meira-Machado *et al.* [3]). The use of such models is very useful for describing event history data offering a better understanding of the process of the illness, and leading to a better knowledge of the evolution of the disease over time. Issues of interest include the estimation of progression rates, assessing the effects of individual risk factors, survival rates or prognostic forecasting.

The complexity of a multi-state model greatly depends on the number of states defined and by the transitions allowed among these states. The simplest form of multi-state model is the "two-state model", or mortality model, for survival analysis (with only two states, "Alive" and "Dead", and a single transition). Splitting the "Alive" state from the simple mortality model for survival data into two transient states, we therefore obtain the simplest progressive three-state model, illustrated in Figure 1. Graphically, multi-state models may be illustrated using diagrams with rectangular boxes to represent possible states and with arrows between the states representing the allowed transitions. States can be transient or absorbing. A state is said to be an absorbing state if no transitions can emerge from the state (e.g. death). Irreversible illness-death models are often used to model disease processes in medical cancer studies.

Figure 1: Progressive three-state model.

In these models, individuals may pass from the initial state (e.g. disease-free; state 1), to the intermediate event or disease state (e.g. recurrence; state 2) and then to the absorbing state (e.g. dead; state 3). Individuals are at risk of death

in each transient state (states 1 and 2). Figure 2 shows the schematic diagram of transitions involved in the illness-death model.

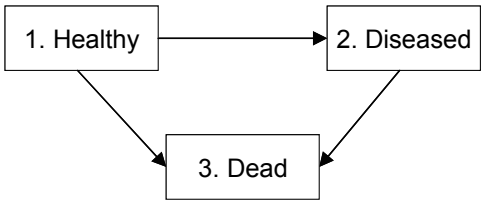

Figure 2: Progressive illness-death model.

The inference in multi-state models is traditionally performed under a Markov assumption for which past and future are independent given its present state (see e.g. [4] and [5]). However, this assumption may fail in some applications, leading to inconsistent estimators. In such cases, alternative (non-Markov) estimators are needed. In this work we review some recent developments in this area, focussing on the estimation of several quantities such as the bivariate distribution function and the transition probabilities. Specifically, we focus on the three-state model of Figure 1 and the illness-death model depicted in Figure 2. In the progressive three-state model, the times between consecutive events (which define states 2 and 3) are often of interest. In Section 2 we present several estimators of the bivariate distribution function of the gap times. Some related problems as estimation of the marginal distribution of the second gap time is discussed. In the framework of the illness-death model, several estimators for the transition probabilities are presented in Section 3. In Section 4, an example of application on bladder tumor recurrence data is re-analyzed to assess the proposed models and methodologies. We also apply our estimation procedures to data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Finally we conclude with a discussion section.

2. ESTIMATION OF THE BIVARIATE DISTRIBUTION

2.1. Notation

Assume the progressive three-state model of Figure 1. Let (T_{12}, T_{23}) be a pair of gap times of successive events, which are observed subjected to random right-censoring. Let C be the right-censoring variable, assumed to be independent of (T_{12}, T_{23}) and let $Y = T_{12} + T_{23}$ be the total time. Because of this, we only

observe $(\widetilde{T}_{12i}, \widetilde{T}_{23i}, \Delta_{1i}, \Delta_{2i})$, $1 \leq i \leq n$, which are n independent replications of $(\widetilde{T}_{12}, \widetilde{T}_{23}, \Delta_1, \Delta_2)$, where $\widetilde{T}_{12} = T_{12} \wedge C$, $\Delta_1 = I(T_{12} \leq C)$, and $\widetilde{T}_{23} = T_{23} \wedge C_2$, $\Delta_2 = I(T_{23} \leq C_2)$ with $C_2 = (C - T_{12})\, I(T_{12} \leq C)$ the censoring variable of the second gap time. Since $\Delta_2 = 1$ implies $\Delta_1 = 1$ then $\Delta_2 = \Delta_2 \Delta_1 = I(Y \leq C)$ is the censoring indicator pertaining to the total time. Define $\widetilde{Y} = Y \wedge C$ and let F_1 and G denote the distribution functions of T_{12} and C, respectively. Since T_{12} and C are independent, the Kaplan–Meier estimator based on the pairs $(\widetilde{T}_{12i}, \Delta_{1i})$'s, consistently estimates the distribution F_1. Similarly, the distribution of the total time may be consistently estimated by the Kaplan–Meier estimator based on $(\widetilde{T}_{12i} + \widetilde{T}_{23i}, \Delta_{2i})$'s. Because T_{23} and C_2 will be in general dependent, the estimation of the marginal distribution for the second gap time is not a simple issue. The same applies to the bivariate distribution function $F_{12}(x, y) = P(T_{12} \leq x, T_{23} \leq y)$. This issue have received much attention recently. Among others it was investigated by Lin *et al.* [6], Van Keilegom [7], de Uña-Álvarez and Meira-Machado [8] or de Uña-Álvarez and Amorim [9].

In this section we present four estimators for the bivariate distribution function of the gap times. All estimator are somehow related since all use (in different ways) the Kaplan–Meier estimator [10].

2.2. Methods

A simple estimator for the bivariate distribution function of the gap times is based on the Kaplan–Meier survival function (Conditional Kaplan–Meier, CKM).

Since $F_{12}(x, y) = P(T_{12} \leq x, T_{23} \leq y) = P(T_{23} \leq y \,|\, T_{12} \leq x)\, P(T_{12} \leq x)$ one simple estimator for the bivariate distribution is given by

$$\widehat{F}_{12}(x, y) \;=\; \widehat{F}_1(x)\, \widehat{F}_{\mathrm{KM}}\big(y \,|\, T_{12} \leq x, \Delta_1 = 1\big) \tag{2.1}$$

where $\widehat{F}_1(x)$ is the Kaplan–Meier product-limit estimator based on the pairs $(\widetilde{T}_{12i}, \Delta_{1i})$'s. The $\widehat{F}_{\mathrm{KM}}(y \,|\, T_{12} \leq x, \Delta_1 = 1)$ is the conditional distribution function for the subset of $T_{12} \leq x$ and $\Delta_1 = 1$ (the Kaplan–Meier estimator based on the subset $(\widetilde{T}_{23i}, \Delta_{2i})$'s such that $\widetilde{T}_{12i} \leq x$ and $\Delta_{1i} = 1$).

Another estimator for the bivariate distribution function was proposed by Lin *et al.* [6]. This estimator is based on inverse probability of censoring weighted (IPCW) and is expressed as

$$\breve{F}_{12}(x, y) \;=\; \breve{H}(x, 0) - \breve{H}(x, y) \tag{2.2}$$

where

$$\breve{H}(x, y) \;=\; \frac{1}{n} \sum_{i=1}^{n} \frac{I\big(\widetilde{T}_{12i} \leq x, \widetilde{T}_{23i} > y\big)}{1 - \widehat{G}\big((\widetilde{T}_{12i} + y)^-\big)}$$

and where $\widehat{G}$ stands for the Kaplan–Meier estimator of the censoring distribution based on the $(\widetilde{Y}_i, 1-\Delta_{2i})$'s.

Recently de Uña-Álvarez and Meira-Machado [8] proposed a simple estimator for the bivariate distribution. The idea behind the estimator is using the Kaplan–Meier estimator pertaining to the distribution of the total time to weight the bivariate data. The proposed estimator (Weighted Kaplan–Meier, WKM) is given by

$$\widetilde{F}_{12}(x,y) \;=\; \sum_{i=1}^{n} W_i \, I\big(\widetilde{T}_{12i} \le x, \widetilde{T}_{23i} \le y\big) \tag{2.3}$$

where

$$W_i \;=\; \frac{\Delta_{2i}}{n-R_i+1} \prod_{R_j=1}^{i-1} \left[1 - \frac{\Delta_{2j}}{n-R_j+1}\right]$$

are the Kaplan–Meier weights attached to $\widetilde{Y}_i$ when estimating the marginal distribution of Y from $(\widetilde{Y}_i, \Delta_{2i})$'s, and for which the ranks of the censored $\widetilde{Y}_i$'s, R_i, are higher than those for uncensored values in the case of ties.

An estimator related to (2.3) was recently proposed by de Uña-Álvarez and Amorim [9]. In this estimator they assume a presmoothed version of the Kaplan–Meier estimator (see [11] and [12] for more details). Presmoothing goes back at least to Dikta (1998) and the idea is to replace the censoring indicators by some smooth fit. This smooth can be based on a certain parametric family (yielding a semiparametric estimator) or using a nonparametric binary regression curve. The term "presmoothing" comes from the fact that smoothing is simply used to get a modified version of the Kaplan–Meier weights, but the final estimator is not smooth itself. Throughout this paper we will assume that the probability of censoring for the second gap time, T_{23}, given the (possibly censored) gap times belongs to a parametric family of binary regression curves. Put $m(x,y) = P(\Delta_2 = 1 \,|\, \widetilde{T}_{12} = x, \widetilde{Y} = y)$, that is, the probability of uncensoring for the total time Y given the observable information on both gap times. Then the new estimator (Smooth Weighted Kaplan–Meier, SWKM) is expressed as

$$\overline{F}_{12}(x,y) \;=\; \sum_{i=1}^{n} W_i^{\star} \, I\big(\widetilde{T}_{12i} \le x, \widetilde{T}_{23i} \le y\big) \tag{2.4}$$

where

$$W_i^{\star} \;=\; \frac{m(\widetilde{T}_{12i}, \widetilde{Y}_i)}{n-R_i+1} \prod_{R_j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{12j}, \widetilde{Y}_j)}{n-R_j+1}\right]$$

are the presmoothed Kaplan–Meier weights where each censoring indicator Δ_{2j} in W_i is replaced by the conditional probability of censoring for the second gap time, given the available information. The m function stands for a (smooth) parametric binary regression model, e.g. logistic. In practice, we assume that

$m(x,y) = m(x,y;\beta)$ where β is a vector of parameters which typically will be computed by maximizing the conditional likelihood of the Δ_2's given $(\widetilde{T}_{12}, \widetilde{T}_{23})$ for those with $\Delta_1 = 1$. Thus, we introduce the parametrically presmoothed Kaplan–Meier weights as

$$W_i^\star(\beta) \;=\; \frac{m(\widetilde{T}_{12i}, \widetilde{Y}_i; \beta)}{n - R_i + 1} \prod_{R_j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{12j}, \widetilde{Y}_j; \beta)}{n - R_j + 1} \right].$$

Note that, unlike (2.3), the SWKM estimator may attach positive mass to pair of gap times with censored second gap time; but only for those with uncensored first gap time. Conditions under which both estimators are consistent is fully discussed in papers by de Uña-Álvarez and Meira-Machado [8] and de Uña-Álvarez and Amorim [9]. Note that without presmoothing, the estimator (2.4) reduces to (2.3). Without censoring both reduce to the empirical estimator.

It is also important to mention that estimators (2.2), (2.3) and (2.4) are only estimable on $\{(x,y)\colon x + y \leq C_{\max}\}$ where $C_{\max}$ is the maximum follow-up time. This means that consistency of these estimators is only guaranteed on the triangle shown in Figure 3.

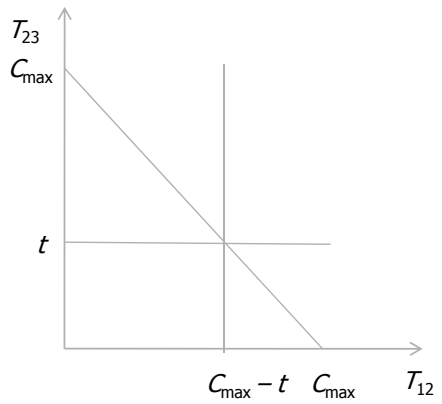

Figure 3: Estimable area of estimators (2.2), (2.3) and (2.4).

We note that the estimates produced via Kaplan–Meier (CKM) may not produce a valid bivariate distribution since it does not guarantee that the bivariate distribution function is monotone. The problem can be explained to the fact that, as the conditioning set $T_{12} \leq x$ changes, the redistribution to the right of the probability mass associated with censored observations also changes. In contrast to the other two methods, the estimators by de Uña-Álvarez and Meira-Machado [8] and de Uña-Álvarez and Amorim [9] are a proper distribution function, in the sense that it attaches positive mass to each observation.

Results of an extensive simulation study comparing the four methods are reported in Meira-Machado and Moreira [13]. The main conclusions are the following:

(**a**) the CKM estimator has larger bias for higher values of the first gap time, but in some cases is one of the estimators with less variance;

(**b**) the WKM estimator has less bias than its smooth version (SWKM); however as expected the later obtained less variance (and less mean square error);

(**c**) the WKM and IPCW estimator are almost unbiased but the last one obtains higher levels of variance for small values of the second gap time.

From the introduced estimators we can obtain an estimator for the marginal distribution of the second gap time, $F_2(y) = P(T_{23} \leq y)$, namely

$$\widehat{F}_2(y) \;=\; \widehat{F}_{12}(+\infty, y) \;=\; \widehat{F}_{\mathrm{KM}}(y \,|\, \Delta_1 = 1) \;, \tag{2.5}$$

$$\widetilde{F}_2(y) \;=\; \widetilde{F}_{12}(+\infty, y) \;=\; \sum_{i=1}^{n} W_i \, I(\widetilde{T}_{23i} \leq y) \;. \tag{2.6}$$

Note that estimator (2.5), obtained from the CKM, is the Kaplan–Meier estimator based on $(\widetilde{T}_{23i}, \Delta_{2i})$'s such that $\Delta_1 = 1$ (i.e., for which the first gap time is uncensored). Estimator (2.6) is different because the Kaplan–Meier weights W_i in this estimator are based on the $\widetilde{Y}_i$-ranks rather than on the $\widetilde{T}_{23i}$-ranks. Indeed, since T_{23} and C_2 are expected to be dependent, the ordinary Kaplan–Meier estimator of F_2 (estimator (2.5)) will be in general inconsistent. The corresponding estimators for (2.2) and (2.4) are obtained using the same ideas.

2.3. Alternative estimators based on the location-scale model

Other estimators were proposed to estimate the bivariate distribution function. A valid estimator of the bivariate distribution function was provided by Van Keilegom [7] which is based on Akritas [14] estimator. However this approach has some limitations since some smoothing is required. Alternative estimators for the above quantities were also given in Van Keilegom *et al.* [15]. This methodology assumes that the vector of gap times (T_{12}, T_{23}) satisfies the nonparametric location-scale regression model $T_{23} = m(T_{12}) + \sigma(T_{12})\varepsilon$, where the functions m and σ are "smooth", and ε is independent of T_{12}. On the basis of the idea of transfer of tail information, the estimator of the error distribution is used to introduce nonparametric estimators for the bivariate distribution function. As shown by the authors, these estimators will be more efficient than the previous, since it allows for the transfer of tail information from lightly censored areas to heavily ones. More details about these methods can be found in the independent paper by Van Keilegom *et al.* [15].

3. ESTIMATION OF THE TRANSITION PROBABILITIES

3.1. Notation

One major goal in longitudinal multi-state studies is the estimation of transition probabilities. Traditionally these quantities are estimated via a nonparametric model (using e.g. the Aalen–Johansen estimator [4]). In a recent paper, Meira-Machado *et al.* [16] introduce a substitute for the Aalen–Johansen estimator in the case of a non-Markov illness-death model. They showed that the new estimator may behave much more efficiently than the Aalen–Johansen when the Markov assumption does not hold. More recently, Amorim *et al.* [17] propose a modification of Meira-Machado *et al.* [16] estimator based on presmoothing ideas which allows for a variance reduction in the presence of censoring. These estimators will be presented in this section, assuming an illness-death model.

In this section we consider the illness-death model depicted in Figure 2 and we assume that all subjects are in state 1 ('healthy') at time $t = 0$. The illness-death model is fully characterized by three transitions: two competing transitions leaving state 1 and one transition to the absorbing 'dead' state for those subjects visiting state 2. Therefore, we have three potential transition times, T_{hj}, from state h to state j. This means that a subject not visiting state 2 will reach the absorbing state at time T_{13}, while this time will be $T_{12} + T_{23}$ if the subject passes through state 2 before. We denote by $\rho = I(T_{12} \leq T_{13})$ the indicator of visiting state 2 at some time. Let $Z = T_{12} \wedge T_{13}$ be the sojourn time in state 1, and let $Y = T_{12} + \rho T_{23}$ be the total survival time of the process. In practice, several issues influence the observation of these variables T_{hj}. Whenever $T_{13} \leq T_{12}$, one gets a right censored value of T_{12} and no information on T_{23} is available. Similarly, the value of T_{13} will be censored for those individuals entering state 2. Further, right censoring may appear due to time limitation in following-up or to other causes. This extra censoring is modeling by considering a censoring variable C which is assumed to be independent of the process; finally, we put $\widetilde{Z} = Z \wedge C$ and $\widetilde{Y} = Y \wedge C$ for the censored versions of Z and Y, and $\Delta_1 = I(Z \leq C)$ and $\Delta_2 = I(Y \leq C)$ for the respective censoring indicators.

3.2. Estimators based on the Kaplan–Meier weights

Meira-Machado *et al.* [16] derived estimators for the transition probabilities $p_{11}(s,t)$, $p_{12}(s,t)$, $p_{22}(s,t)$, for a general non-Markov illness-death process without recovery as follows. Let H denote the survival function for Z then the transition

probabilities are written as

$$p_{11}(s,t) = \frac{P(Z>t)}{P(Z>s)} = \frac{H(t)}{H(s)}\,, \tag{3.1}$$

$$p_{12}(s,t) = \frac{P(s<Z\leq t<Y)}{P(Z>s)} = \frac{E\big[\varphi_{st}(Z,Y)\big]}{H(s)}\,, \tag{3.2}$$

$$p_{22}(s,t) = \frac{P(Z\leq s, t<Y)}{P(Z\leq s<Y)} = \frac{E\big[\widetilde{\varphi}_{st}(Z,Y)\big]}{E\big[\widetilde{\varphi}_{ss}(Z,Y)\big]}\,, \tag{3.3}$$

where $\varphi_{st}(u,v) = I(s<u\leq t,\, v>t)$ and $\widetilde{\varphi}_{st}(u,v) = I(u\leq s,\, v>t)$.

Then, (3.1) and the denominator of (3.2) only involve the Z variable, and they can be estimated by the ordinary Kaplan–Meier estimator, $\widehat{H}$, based on the pairs $(\widetilde{Z}_i, \Delta_{1i})$'s. The transition probability (3.3) and the numerator of the (3.2) involve expectations of particular transformations of the pair (Z,Y) that can be estimated in different ways. In this section we present two methods to empirically approximate these expectations from the data $\big\{\big(\widetilde{Z}_i, \widetilde{Y}_i, \Delta_{1i}, \Delta_{2i}, \Delta_{1i}\rho_i\big), 1\leq i\leq n\big\}$, which are assumed to form a random sample of the vector $(\widetilde{Z}, \widetilde{Y}, \Delta_1, \Delta_2, \Delta_1\rho)$.

In Meira-Machado *et al.* [16], the expectations $E\big(\varphi_{st}(Z,Y)\big)$ and $E\big(\widetilde{\varphi}_{st}(Z,Y)\big)$ were estimated by Kaplan–Meier integrals of the form

$$\sum_{i=1}^{n} W_i\, \varphi_{st}(\widetilde{Z}_i, \widetilde{Y}_i)$$

where W_i are the Kaplan–Meier weight attached to $\widetilde{Y}_i$ when estimating the marginal distribution of Y from the $(\widetilde{Y}_i, \Delta_{2i})$'s.

Note that, without right-censoring, the estimator of the transition probabilities reduces to the relative frequency of processes in state j at time t among those in state h at time $s<t$. Meira-Machado *et al.* [16] derived large sample properties of these estimators which may be generalized to more complicated non-Markov processes.

The main weakness of this method [16] is that it provides large standard errors in estimation, specially when there is a large proportion of censored data. In order to overcome this issue Amorim *et al.* [17] propose a modification of Meira-Machado *et al.* (2006)'s estimator based on presmoothing ideas, in the presence of censoring. The implementation of these ideas is straightforward in the case of the progressive three-state model (see Section 2) but not so simple for the illness-death model (as explained below).

In the presmoothed version [17], the expectations in (3.2) and (3.3) are estimated by

$$\sum_{i=1}^{n} W_i^{\star}\, \varphi_{st}(\widetilde{Z}_i, \widetilde{Y}_i)$$

where

$$W_i^\star \;=\; \frac{m(\widetilde{Z}_i,\widetilde{Y}_i)}{n-R_i+1}\;\prod_{R_j=1}^{i-1}\left[1-\frac{m(\widetilde{Z}_j,\widetilde{Y}_j)}{n-R_j+1}\right]$$

and where $m(z,t)$ stands for an estimator of the binary regression function

$$m(z,t) \;=\; P\big(\Delta_2=1\,|\,\widetilde{Z}=z,\widetilde{Y}=t\big)\,.$$

The problem in the illness-death model is that the function $m(z,t)$ will typically be discontinuous along the line $t=z$, that is, for those values $(\widetilde{Z},\widetilde{Y})$ corresponding to subjects who are censored while being in state 1 or who suffer a direct transition to the absorbing state. To construct $m(z,t)$ the authors propose to estimate independently two functions: $m_1(z,t)$ such that $m_1(\widetilde{Z},\widetilde{Y})$ is the conditional probability of censoring on Y given $(\widetilde{Z},\widetilde{Y})$ and given that a transition to state 2 is observed; and $m_2(t)$ which is the conditional probability of observing a direct transition from state 1 to state 3 given $\widetilde{Z}=t$ (or $\widetilde{Y}=t$) and given that a transition to state 2 is never observed. These functions can be fitted by some smooth models, so we finally have

$$m(z,t) \;=\; m_1(z,t)\,I(z<t)+m_2(t)\,I(z=t)\;.$$

The estimator $m_1(z,t)$ is based on the subsample $\{i:\Delta_{1i}\rho_i=1\}$, while $m_2(t)$ is computed from $\{i:\Delta_{1i}\rho_i=0\}$. The only condition which is assumed on these two functions is that they should approximate well their targets in a uniform sense (see [17] for more details).

Results from a simulation study comparing the two methods is reported in Amorim *et al.* [17], revealing that the semiparametric estimator is more efficient.

4. EXAMPLES OF APPLICATION

The methods described in Section 2 and Section 3 are illustrated through two real data sets. First, we use data from a bladder cancer study (Byar (1980)) conducted by the Veterans Administration Cooperative Urological Research Group. In addition to this data set we also use the well-known and widely studied colon cancer database. In both data sets, a nonfatal event (recurrence) is observed during the disease course. Also, in both data sets, recurrence is a time-dependent covariate that can be re-expressed as a multi-state model, with states based on the values of the covariate. In the first database all deceased patients died after having a recurrence making it possible for the progressive three-state model to be used (Figure 1). In the second database some subjects died without having a recurrence, making feasible for the illness-death model, depicted in Figure 2, to be used.

4.1. Bladder cancer data

In this study, patients had superficial bladder tumors that were removed by transurethral resection. Many patients had multiple recurrences (up to a maximum of 9) of tumors during the study, and new tumors were removed at each visit. For illustration purposes we re-analyze data from 85 individuals in the placebo and thiotepa treatment groups; these data are available as part of the R survival package. Here, only the first two recurrence times and the corresponding gap times T_{12} and T_{23} are considered. From the total of 85 patients, 47 relapsed at least once and, among these, 29 experienced a new recurrence. We have a total amount of censoring of 66% from which 44.7% is obtained from censored observations on the first gap time. We have about 38% of censored Y's among the uncensored first gap time.

We computed the estimated values for all the estimators of the bivariate distribution function, $F_{12}(x, y)$, introduced in Section 2, for x equal to 3, 13, 29 and 49 and y values 3, 10, 17.75 and 36.75, corresponding to marginal survival probabilities of 0.25, 0.5, 0.75 and 0.95. The estimated values of $F_{12}(x, y)$ are reported in Table 1. In this case it is clearly seen that the four methods can provide quite different results, specially at the right tail of the bivariate distribution, where the censoring effects are stronger.

Table 1: Estimated values of the bivariate distribution function $F_{12}(x, y)$ for different pairs of values. Bladder cancer data.

x	Estimator	y			
		3	10	17.75	36.75
3	CKM	0.0364	0.0607	0.1261	0.1746
	IPCW	0.0320	0.0432	0.1240	0.1726
	WKM	0.0128	0.0427	0.1045	0.1167
	SWKM	0.0328	0.0556	0.1089	0.1203
13	CKM	0.0763	0.1684	0.2533	0.3284
	IPCW	0.0668	0.1510	0.2540	0.3154
	WKM	0.1036	0.1742	0.2511	0.2633
	SWKM	0.1193	0.1814	0.2565	0.2679
29	CKM	0.1513	0.2703	0.3680	0.4499
	IPCW	0.1677	0.2902	0.3830	0.4932
	WKM	0.1729	0.2436	0.3205	0.3482
	SWKM	0.2331	0.2952	0.3704	0.3913
49	CKM	0.1571	0.2801	0.3803	0.4764
	IPCW	0.1556	0.2336	0.4355	0.5457
	WKM	0.2294	0.3001	0.3932	0.4209
	SWKM	0.2652	0.3273	0.4109	0.4318

4.2. Colon cancer data

For illustration, we apply the proposed methods of Section 3 to data from a large clinical trial on patients affected by colon cancer. All subjects underwent a curative surgery for colo-rectal cancer. Unfortunately, some of these patients have residual cancer, which lead to disease recurrence and death (in some cases). From the total of 929 patients, 468 (about 50%) developed recurrence and among these 414 (88%) died. Only 38 patients died without recurrence. The rest of the patients (423) remained alive and disease-free up to the end of the follow-up. The presence of patients that experienced a direct transition from the initial state to the absorbing state leads to the need of using the illness-death model with states "Alive and disease-free" (State 1), "Alive with recurrence" (State 2) and "dead" (State 3). Using Cox proportional hazards models, we verified that the transition rate from state 2 to state 3 is affected by the time spent in the previous state [19]. This allowed us to conclude that the Markov assumption may be unsatisfactory for the colon cancer data set. Note that both methods presented in Section 3 do not make use of the Markov information. We will present estimated transition probabilities calculated using these two approaches.

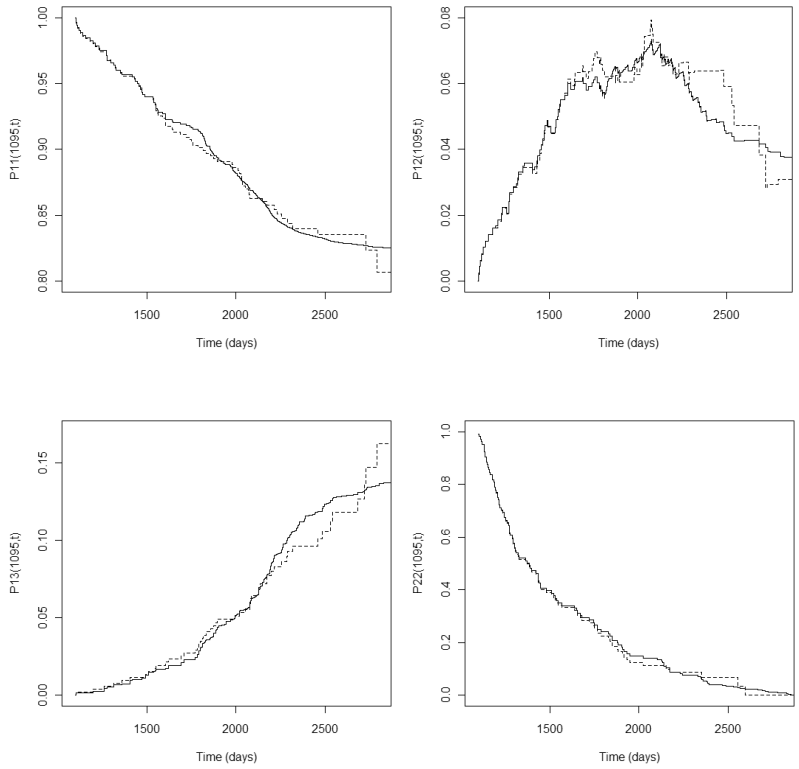

Figure 4: Estimated transition probabilities for $p_{hj}(s,t)$ with $s = 1095$ based on the Kaplan–Meier weights (dashed line) and based on presmoothed Kaplan–Meier weights (solid line).

In Figure 4 we illustrate differences between the estimated transition probabilities, $p_{hj}(s,t)$, $1 \leq h \leq j \leq 3$, based on presmoothing the Kaplan–Meier weights (semiparametric) and the estimator corresponding to no presmoothing [16]. The semiparametric estimator was obtained using a standard logistic model for the parametric estimation of m. The value s was chosen to be as 3 years (1095 days). From this figure we see that the semiparametric estimator have more jump points but with smaller steps. The additional jump points correspond to patients with censored values of the total time that underwent a transition from state 1 to state 2 before time s (uncensored sojourn time in state 1). The number of jump points and the size of the steps are strictly related to the amount of censoring and to the sample size. As expected, both methods provide similar point estimates at small time values while some departures are appreciated for higher time values. In sum, the semiparametric approach provides more reliable curves with less variability, specially in the right tail.

5. DISCUSSION

In this paper we present nonparametric and semiparametric estimators for quantities of interest in multi-state survival modeling. The interest is focused on the estimation of the bivariate distribution function for censored gap times and the estimation of transition probabilities. For both quantities we present two methods based on the Kaplan–Meier estimator pertaining to the distribution of the total time to weight the data. One of these methods is based on presmoothing the Kaplan–Meier estimator. For this, we assume that the probability of censoring for total time given the (possibly censored) gap times belongs to a parametric family of binary regression curves. Some of these estimators may behave much more efficiently than the competing ones. These methods are illustrated using data from two cancer studies.

ACKNOWLEDGMENTS

Thanks to a referee for careful reading of the paper. Luís F. Meira-Machado acknowledges financial support by Grant PTDC/MAT/104879/2008 (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education. The research was also partially funded by CMAT and FCT under the POCI 2010 program.

REFERENCES

[1] Andersen, P.K.; Borgan, O.; Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

[2] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.

[3] Meira-Machado, L.; de Uña-Álvarez, J.; Cadarso-Suárez, C. and Andersen, P.K. (2009). Multi-state models for the analysis of time to event data, *Statistical Methods in Medical Research*, **18**, 195–222.

[4] Aalen, O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations, *Scandinavian Journal of Statistics*, **5**, 141–150.

[5] Andersen, P.K. and Keiding, N. (2002). Multi-state models for event history analysis, *Statistical Methods in Medical Research*, **11**, 91–115.

[6] Lin, D.Y.; Sun, W. and Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data, *Biometrika*, **86**, 59–70.

[7] Van Keilegom, I. (2004). A note on the nonparametric estimation of the bivariate distribution under dependent censoring, *Journal of Nonparametric Statistics*, **16**, 659–670.

[8] de Uña-Álvarez, J. and Meira-Machado, L. (2008). A simple estimator of the bivariate distribution function for censored gap times, *Statistics & Probability Letters*, **78**, 2440–2445.

[9] de Uña-Álvarez, J. and Amorim, A.P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap times, *Biometrical Journal*, **53**(1), 113–127.

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

[11] de Uña-Álvarez, J. and Rodríguez-Campos, C. (2004). Strong consistency of presmoothed Kaplan–Meier integrals when covariables are present, *Statistics*, **38**, 483–496.

[12] Dikta, G. (1998). On semiparametric random censorship models, *Journal of Statistical Planning and Inference*, **66**, 253–279.

[13] Meira-Machado, L. and Moreira, A. (2010). Estimation of the bivariate distribution function for censored gap times, *Proceedings of the 19th International Conference on Computational Statistics*, 1367–1374.

[14] Akritas, M.G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring, *The Annals of Statistics*, **22**, 1299–1327.

[15] Van Keilegom, I.; de Uña-Álvarez, J. and Meira-Machado, L. (2011). Nonparametric location-scale models for censored successive survival times, *Journal of Statistical Planning and Inference*, **141**(3), 1118–1131.

[16] Meira-Machado, L.; de Uña-Álvarez, J. and Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model, *Lifetime Data Analysis*, **12**, 325–344.

[17] Amorim, A.P.; de Uña-Álvarez, J. and Meira-Machado, L. (2011). Presmoothing the transition probabilities in the illness-death model, *Statistics & Probability Letters*, *in Press*.

[18] Byar, D.P. (1980). Veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine and topical thiotepa, *Bladder Tumors and Other Topics in Urological Oncology*, **18**(36), 363–370.

[19] Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies, *Biometrics*, **42**, 855–865.

REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 99–114

MIXTURES OF FACTOR MODELS FOR MULTIVARIATE DISEASE RATES

Authors: T.C. Bailey
– College of Engineering, Mathematics and Physical Sciences, University of Exeter, U.K.
t.c.bailey@exeter.ac.uk

P.J. Hewson
– School of Mathematics and Statistics, University of Plymouth, U.K.
paul.hewson@plymouth.ac.uk

Abstract:

- A range of different approaches have been suggested for the multivariate modelling of the geographical distribution of different but potentially related diseases. We suggest an addition to these methods which incorporates a discrete mixture of latent factors, as opposed to using CAR or MCAR random effect formulations. Our proposal provides for a potentially richer range of dependency structures than those encompassed in previously used models in that it is capable of representing an enhanced range of correlation structures between diseases at the same time as implicitly allowing for less restrictive spatial correlation structures between geographical units. We illustrate results of using the model on data taken from cancer registries on four carcinomas in some 300 UK geographical areas.

Key-Words:

- *mixture models; factor analysis; multivariate disease rates.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

The literature in spatial epidemiology contains a growing number of references to multivariate modelling of the geographical distribution of morbidity or mortality rates for potentially related diseases. Dabney and Wakefield [9] suggest that the two main motivations for this interest are firstly, to explore similarities or dissimilarities in the geographical risk distribution for the different diseases, and, secondly, to 'borrow strength' across disease rates to shrink the uncertainty in geographical risk assessment for any one of the individual diseases. Regardless of the relative balance of interest between these two motivations, it is clear that achievement of either objective will be limited if the structure of the multivariate model used is inadequate and/or its related assumptions are unrealistic. In particular, the possible dependency structure (either between diseases, or across geographical units) should not be overly constrained by the model structure. For example, it may not be the case that relationships between diseases are the same in, say, rural versus urban environments, nor that dependency between disease rates in neighbouring small areas will be homogeneous at a larger geographical scale where spatial discontinuities may well be present. The model structure should provide a sufficiently rich range of dependency structures to encompass such possibilities.

Various approaches to spatial modelling of multivariate disease rates have been proposed. Many of these may be characterised as generalised linear mixed models (GLMMs) of varying descriptions in which the dependence structure not explained by covariates is represented in terms of random effects which are correlated between diseases and across geographical units. The Multivariate Conditional Autoregressive (MCAR) model is one popular approach for dealing with multivariate disease rates in small areas (e.g. [13, 26, 10]) but some have commented that MCAR formulations remain difficult to fine tune because the correlations in random effects between diseases and/or across spatial units are not easy to disentangle. Models which incorporate more explicit latent structure also feature in multivariate modelling of disease rates. Held *et al.* [22] review a range of approaches to joint disease modelling including shared latent processes. Early examples include that used by Knorr-Held and Best [24] to identify a shared spatial component in the geographical distribution of bivariate disease rates and the simple latent variable formulation employed by Wang and Wall [36] to model Minnesota cancer rates. More recently, Liu *et al.* [27] have proposed a Structural Equation Model (SEM) for cancer rates where the three cancers have a single shared spatially structured latent variable. Latent structure models are not restricted to area applications. Christensen and Amemiya [6] have suggested an approach applicable to point data which has been illustrated by Minozzo and Fruttini [30] who examined bivariate point measures of types of diabetes morbidity.

MCAR versus explicit latent structure aside, there are two assumptions which dominate most of these previously described multivariate models. First, that spatial dependency across small areas is essentially 'smooth' and not subject to global spatial discontinuities. Second, that it is reasonable to assume that a single relationship between diseases counts applies to all types of areas. In some (perhaps many) applications the dependence structure may well be more complex than that implied by these rather broad assumptions. Normand *et al.* [31] highlight that in the absence of adequate covariate information, simple exchangeability assumptions across areas may not be valid in many of the GLMMs used to analyse healthcare provision. In a multivariate setting these exchangeability concerns across areas remain, but are compounded by additional concerns over whether the dependence structure between diseases varies geographically.

We therefore propose a model for use in such contexts which potentially provides a richer range of dependency structures than those encompassed by previous approaches. Rather than representing the dependence structure not explained by covariates in terms of correlated random effects, we suggest that it is preferable to formulate correlations in terms of an explicit latent structure similar to that arising in factor analysis. Our model is based on latent structure mixtures and we argue that incorporating a discrete mixture into the latent structure loadings in the model simultaneously provides potential to represent an enhanced range of correlation structures between diseases, at the same time as allowing for less restrictive spatial correlation structures between geographical units.

The structure we propose could be considered similar to a 'mixture of factor analysers'. Such models (mostly for Gaussian responses) have been reported in other contexts in the statistical literature and elsewhere. For example, Mclachlan and Peel [29] discuss mixtures of factor analysers, Lee and Song [25] report on mixtures in relation to Structural Equation Modelling, and Viroli [35] describes 'independent factor analysis' based upon approaches developed in the signal processing literature. Many insights into the properties of GLMMs for multivariate disease rates can be gained from studying recent developments in factor analysis which has been enjoying somewhat of a methodological renaissance in a Bayesian setting [1] with a number of useful results emerging. The development of our latent structure mixture model for joint disease modelling in this paper is encouraged by these results and draws upon our belief in the value of viewing correlated random effects in a factor analysis framework.

In Section 2 we develop our model and describe fitting strategy. In Section 3 we introduce an illustrative data set on which to demonstrate results which concerns four cancers in some 300 geographical units in England, Scotland and Wales. We present model results for these data in Section 4 and then go on to discuss conclusions in Section 5.

2. MODEL FORMULATION

The basic structure of the problem we consider is that we have data, y_{ij}, representing the number of cases in area i for disease j $(i = 1, ..., n;\ j = 1, ..., p)$. The corresponding expected number of cases e_{ij} is also known, this being based on age/sex standardised rates for the whole of the study region, or for some appropriate alternative reference population (equivalently, we may know y_{ij} along with the standard morbidity ratio (SMR) y_{ij}/e_{ij}, for disease j in area i). Where appropriate we will refer to the vector of disease counts in each area as $\boldsymbol{y}_i = (y_{i1}, ..., y_{ip})$ and the corresponding vector of expected counts as $\boldsymbol{e}_i = (e_{i1}, ..., e_{ip})$. In a practical setting we may well also have a vector of covariates $\boldsymbol{x}_i$ $(i = 1, ..., n)$ measured in each area, but for simplicity of exposition we will assume throughout this paper that such covariates are not available. If required these can be included into the models we develop in an obvious and straightforward fashion.

It is usual to assume disease counts are Poisson distributed viz: $y_{ij}|\lambda_{ij} \sim \text{Pois}(e_{ij}\lambda_{ij})$, with the mean vector, $\boldsymbol{\lambda}_i = (\lambda_{i1}, ..., \lambda_{ip})$, in each area then being modelled through an appropriate link function by a suitable linear predictor. In developing our modelling framework we build upon proposals made by Wang and Wall [36] mentioned in Section 1 which used a log link and a simple linear predictor involving a single area specific latent variable with a disease specific loading, so that: $\log(\lambda_{ij}) = \phi_i \delta_j$, where δ_j is the disease specific loading and ϕ_i is the area specific latent (unmeasured) variable which was in turn assumed to follow a Conditional Auto-Regressive Gaussian (CAR) distribution over the areas. In this model correlation between diseases within an area is reflected through the shared latent variable and spatial correlation across areas is achieved via the CAR). However, the simple structure only provides for a limited range of correlation structures between diseases (same for all areas) and makes possibly unrealistic assumptions about the spatial dependence (it is 'smooth' — there is no possibility of global spatial discontinuity). We therefore consider ways to provide more complex possibilities for dependencies between diseases and across areas.

First, to allow potential for a more complex dependence structure between diseases, we include q latent variables. So that the model becomes

$$\log(\lambda_{ij}) \ = \ \sum_{h=1}^{q} \phi_{ih}\, \delta_{jh} \ , \tag{2.1}$$

where δ_{jh} is a disease specific loading for area specific latent variable ϕ_{ih} $(h = 1, ..., q)$. We can express this more succinctly as

$$\log(\boldsymbol{\lambda}_i) \ = \ \boldsymbol{\phi}_i \boldsymbol{\Delta} \ , \tag{2.2}$$

where it is understood that $\log(\boldsymbol{\lambda}_i) = (\log \lambda_{i1}, ..., \log \lambda_{ip})$ and where $\boldsymbol{\phi}_i = (\phi_{i1}, ..., \phi_{iq})$

is the vector of latent variables for area i and $\boldsymbol{\Delta}$ is the $q \times p$ matrix of loadings:

$$\boldsymbol{\Delta} = \begin{pmatrix} \delta_{11} & \cdots & \delta_{p1} \\ \vdots & \ddots & \vdots \\ \delta_{1q} & \cdots & \delta_{pq} \end{pmatrix}.$$

This formulation raises identifiability problems (e.g. rotational indeterminacy) so we follow Lopes and West [28] and constrain the loading matrix $\boldsymbol{\Delta}$ so that it is upper triangular with the diagonal strictly positive, i.e.:

$$\boldsymbol{\Delta} = \begin{pmatrix} \delta_{11} & \delta_{21} & \cdots & \delta_{p1} \\ 0 & \delta_{22} & \ddots & \delta_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_{pq} \end{pmatrix}.$$

To incorporate a richer range of dependency across areas we then further extend model (2.2) to a mixture model across s sets of q latent variables in each area, with $\boldsymbol{\phi}_i^{(k)} = \left(\phi_{i1}^{(k)}, ..., \phi_{iq}^{(k)}\right)$ denoting the k^{th} set of latent variables. So that the model becomes

$$y_{ij} \mid \lambda_{ij}^{(1)}, ..., \lambda_{ij}^{(s)} \sim \sum_{k=1}^{s} \pi_{ik} \operatorname{Pois}\left(e_{ij}\lambda_{ij}^{(k)}\right),$$

with π_{ik} denoting mixing probabilities ($\sum_k \pi_{ik} = 1$) and with

$$\log\left(\boldsymbol{\lambda}_i^{(k)}\right) = \boldsymbol{\phi}_i^{(k)} \boldsymbol{\Delta}^{(k)},$$

where $\boldsymbol{\Delta}^{(k)}$ is a $q \times p$ matrix of loadings for the the k^{th} set of latent variables ($k = 1, ..., s$) and with each such matrix is subject to the constraints described earlier. The latent variables $\phi_{ih}^{(k)}$ are assumed to follow independent Gaussian distributions with means $\mu_h^{(k)}$ for $k = 1, ..., s,\ h = 1, ..., q$ and $i = 1, ..., n$.

So each area is now a mixture of s types of areas, with each type of area being associated with a different set of q latent variables and corresponding loadings. Note there is no explicit spatial dependence (e.g. CAR or MCAR) in the above formulation. However, implicit spatial dependence arises through groups of areas being free to share a similar pattern of mixing probabilities over the sets of latent variables and loadings. This type of spatial dependence is potentially very flexible since it does not necessarily impose undue global spatial smoothness.

Finally, we incorporate additional unstructured area and disease specific random effects into the linear predictor of our formulation above in order to deal with possible overdispersion. These additional random effects are effectively equivalent to 'uniqueness' in the traditional factor analysis literature.

So the final model then becomes

$$\log\left(\boldsymbol{\lambda}_i^{(k)}\right) = \boldsymbol{\phi}_i^{(k)} \boldsymbol{\Delta}^{(k)} + \boldsymbol{\zeta}_i, \tag{2.3}$$

where the random effects, $\boldsymbol{\zeta}_i = (\zeta_{i1}, ..., \zeta_{ip})$, are independent zero mean Gaussian with variances drawn from inverse Gamma hyperpriors.

Using mixture models as in the above formulation, raises a number of fitting and identifiability issues Mclachlan and Peel [29]. We use an MCMC fitting approach with a flat hyperprior for the group means $\mu_h^{(k)}$ of the latent variables; and with loadings δ_{jh} (subject to the identifiability constraints described earlier) given zero mean Gaussian priors with inverse Gamma hyperpriors for their distinct associated variances. The model involves both unknown numbers of latent variables and mixture components, but there is considerable complexity in using dimension changing methods (e.g. RJMCMC) even with just unknown numbers of latent variables (e.g. [28]) let alone when this is compounded with an unknown number of mixture components. We therefore follow suggestions made by Green [19] and use a strategy whereby distinct models are fitted to distinct dimensionalities. In selecting the number of mixture components we draw on similarities between our latent structure mixture and the 'mixture of factor analysers' model [29, Chap 8]. We note that in the machine learning literature, 'Variational Bayes' approaches are used to fit 'mixtures of factor analysers' [17, 3] which are essentially equivalent to minimising the Kullback–Leibler distance between the factorised approximation and the joint posterior. We have therefore selected a strategy based on the Kullback–Leibler distance in order to select the number of mixture components. We use this measure to assess the distance between our fitted model and a model which assumes the two closest mixture components have been merged. An approximation to the Kullback–Liebler distance can be generated as a byproduct of Gibbs sampling [32] and we use this to guide model selection with respect to the number of substantive mixture components which may be supported by the data. We follow Celeux *et al.* [5] by not placing any constraints on the ordering of the mixture group means and deal with label-switching by post-processing the output of our MCMC sampler. Mixture group memberships for loadings and latent variables (where applicable) were assumed to be categorical variables with a Dirichlet prior for the probabilities, $\boldsymbol{\pi}_i = (\pi_{i1}, ..., \pi_{is})$, of belonging to the different groups, i.e.

$$p(\pi_{i1}, ..., \pi_{is}) \;=\; \frac{\Gamma(\alpha_1 \cdots \alpha_s)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_s)} \; \pi_{i1}^{\alpha_1 - 1} \cdots \pi_{is}^{\alpha_s - 1} \,,$$

where α_k represents the prior group weights for each of the $k = 1, ..., s$ mixing components.

Routine checks for MCMC convergence were used involving Gelman and Rubin's R [15], Geweke's statistic [16] and Heidelberger and Welch's statistic [21]. These are slow models to fit, running multiple chains is therefore somewhat time consuming but it is essential given the level of cross-correlation possible due the model formulation. Whilst somewhat slower than using customised code, we used the widely accessible WinBUGs software package [34] to fit models.

Overall model fit was assessed by a number of measures. In addition to assessing mixture group membership in terms of Kullback–Leibler distance, we also considered Posterior Predictive Loss as proposed by Gelfand and Ghosh [14] as well as a variant on this proposed by Gneiting and Raftery [18] which enjoys the advantage of being a proper scoring rule and has been examined specifically in respect of count data by Czado *et al.* [8]. We refer to this proper score as the 'Dawid and Sebastiani score' following earlier work reported in Dawid and Sebastiani [11].

We also validated the performance of our model by means of out of sample predictions using the posterior predictive distribution of the relative risks for each disease for each area studied. We removed one tenth of the observations at random, and fitted the model to these data. The posterior predictive density for these deleted observations was collected during the model fitting process, and compared with the observed data.

3. DATA

As an illustrative application to demonstrate model performance, we consider data on reported numbers of cases of four types of cancer in some 300 geographical units covering England, Wales and Scotland. These data were obtained from the 9 cancer registries in England as well as the cancer registries in Scotland and Wales and comprise the number of cases reported between 1999 and 2001 of, 'Lung cancer' (ICD-10 classified sites C33-C34, i.e. Trachea, Bronchus and Lung cancer), 'Oral cancer' (C00-C14, i.e. Lip, Oral Cavity and Pharynx), Breast cancer (C50) and Cancer of the Cervix (C53). Data were collected on prevalence for males and females for the first two cancers, but only for females in respect of breast and cervical cancer. Direct standardisation [7] was used to estimate associated expected morbidity based on quinary age bands for the whole of the study region.

These data refer to the smallest administrative geographical unit available, i.e. the 303 Primary Care Trusts (PCT) in England, the 22 Local Health Boards in Wales and the 14 National Health Service Boards in Scotland. For convenience we will subsequently refer to all such units by the name given to the majority, namely 'PCTs'. The English and Welsh entities are comparable in size, for example the mean population within an English PCT is 163,000 with a minimum 63,700 and a maximum 372,600 whereas the Welsh Local Health Board mean population was 131,900 with a minimum of 56,500 and a maximum of 310,300. Scotland is dominated by a couple of very large NHS Boards, the mean population was 720,320 with a minimum of 38,400 and a maximum of 1,736,300. Some caution may therefore be needed when comparing results from England and Wales with

those of Scotland due to aggregation effects alone. We concentrate here on those 335 'PCTs' which are entirely based on the mainland, i.e we exclude islands. It should be noted for later reference that one of the mainland PCTs in Cornwall (the far South West of our maps) contains an aggregate of data from the Isles of Scilly.

Basic information about the number of cancers registered under each diagnosis in each PCT are contained in Table 1. As is usual with administratively collected data, there are some provisos over accuracy. Particular problems with UK cancer registry data are documented in Best and Wakefield [4], and we note that it may not be entirely reasonable to assume that each cancer registry collects the data in exactly the same way.

Table 1: Summary information on the mean, standard deviation, minimum and maximum number of cancer cases registered for each of the six diagnosis groups in each of the 335 non-island PCTs in England, Scotland and Wales.

	Oral (F)	Lung (F)	Oral (M)	Lung (M)	Breast	Cervix
Mean	16.25	132.81	28.65	205.06	359.99	26.68
Std. Dev.	11.62	107.06	23.04	136.67	184.47	17.62
Minimum	1	29	3	41	107	4
Maximum	133	1333	276	1653	1832	175

Table 2 gives the observed correlation coefficients between the various cancer rates. It can be seen for example that Lung cancer rates are highly correlated between males and females (0.88), the same is not so true of oral cancer rates (0.31). Figure 1 provides the same information in graphical form.

Table 2: Observed correlation between cancer rates for the four cancers, male and female data shown separately.

	Oral (F)	Lung (F)	Oral (M)	Lung (M)	Breast	Cervix
Oral (F)	1.00	0.23	0.31	0.24	0.04	0.22
Lung (F)	0.23	1.00	0.52	0.88	−0.32	0.47
Oral (M)	0.31	0.52	1.00	0.52	−0.18	0.35
Lung (M)	0.24	0.88	0.52	1.00	−0.39	0.49
Breast	0.04	−0.32	−0.18	−0.39	1.00	−0.18
Cervix	0.22	0.47	0.35	0.49	−0.18	1.00

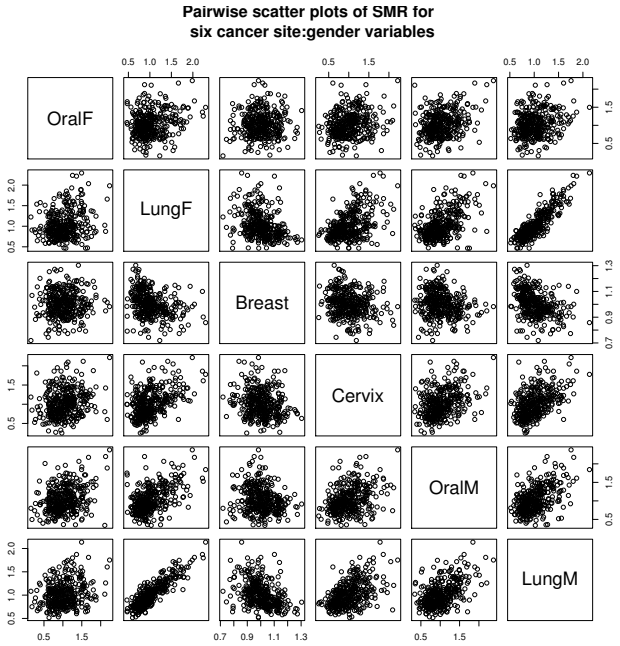

Figure 1: Pairwise scatter plots for the six age standardised cancer rates.

4. RESULTS

Models were fitted as described in Section 2, using the priors and convergence criteria indicated there. A standard burn in period of 50,000 iterations was used, a further 100,000 samples thinned by a factor of 20 were used for posterior inference. As mentioned previously, all results reported here were obtained using the WinBUGs software package [34].

We fitted a range of models with differing numbers of latent variables and differing numbers of mixture components. Given $p = 6$ disease counts we considered all the classically identifiable possibilities, i.e. $k = 1, ..., 3$ latent variables. It was found feasible to fit a three latent variable model and we prefer that both because it has the greatest potential to model a complex dependence structure and because it has the lowest posterior predictive score. In general, the posterior predictive Dawid and Sebastiani score tends to favour models with a larger number of mixture components. However, Kullback–Leibler tends to favour a two component solution. Figure 2 contains a density plot of the sampled values for the approximate Kullback–Leibler distance between a two component mixture and a one component model and indicates a considerable distance between the two and one component means of the second latent variable. Given this support from the Kullback–Leibler distance we accept a two component solution despite

the fact that this fits slightly less well than higher numbers of components on the Dawid and Sebastiani score.

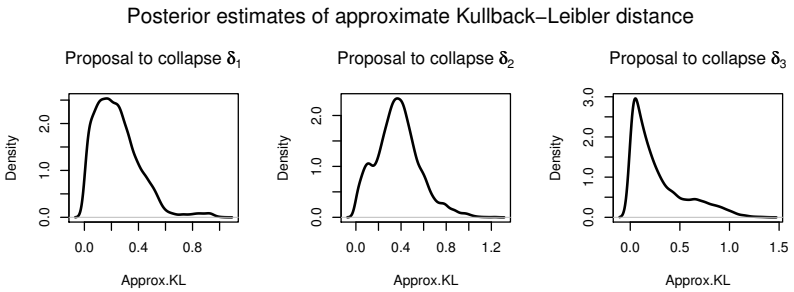

Figure 2: Approximate Kullback–Leibler distance between two and one component latent structure mixture model with three latent variables when considering each of the latent variables.

Figure 3 presents maps of the geographical distribution of raw and model estimated posterior mean relative risk for Breast cancer. The model achieves a degree of shrinkage in terms of the posterior mean of the relative risks when compared with the raw data. Maps for the other cancer counts reveal a similar story.

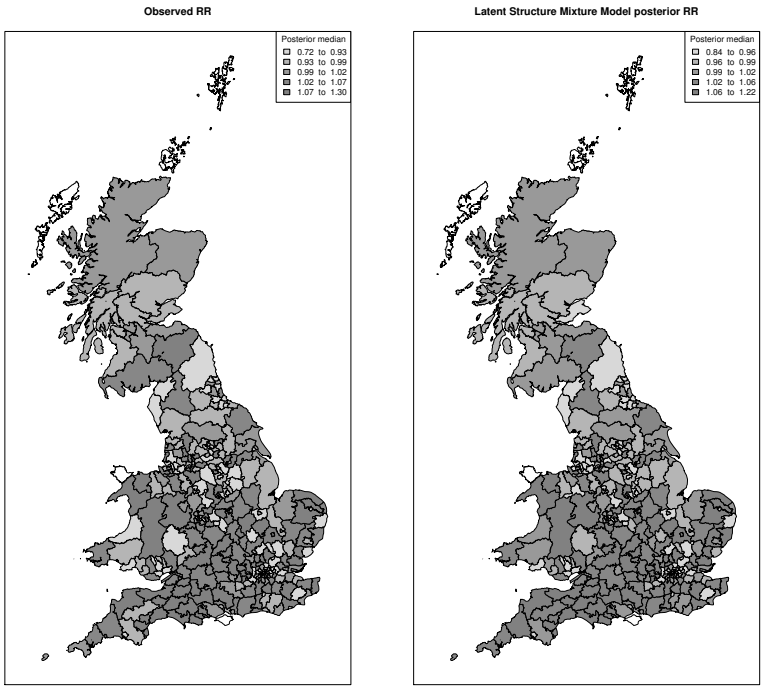

Figure 3: Breast cancer raw rates and posterior mean relative risk from latent structure mixture model.

Perhaps more interesting, is that with this model it is possible to examine posterior mean mixing probabilities for each PCT. Figure 4 gives the posterior groupings of PCTs associated with this measure. It is clear that although no explicit spatial structure is imposed in this model, the mixture groups appear to be highlighting a spatial pattern that has a substantive interpretation (Scotland and industrial areas in England and Wales). There does therefore appear appear to be some interesting possibilities in using this type of model formulation.

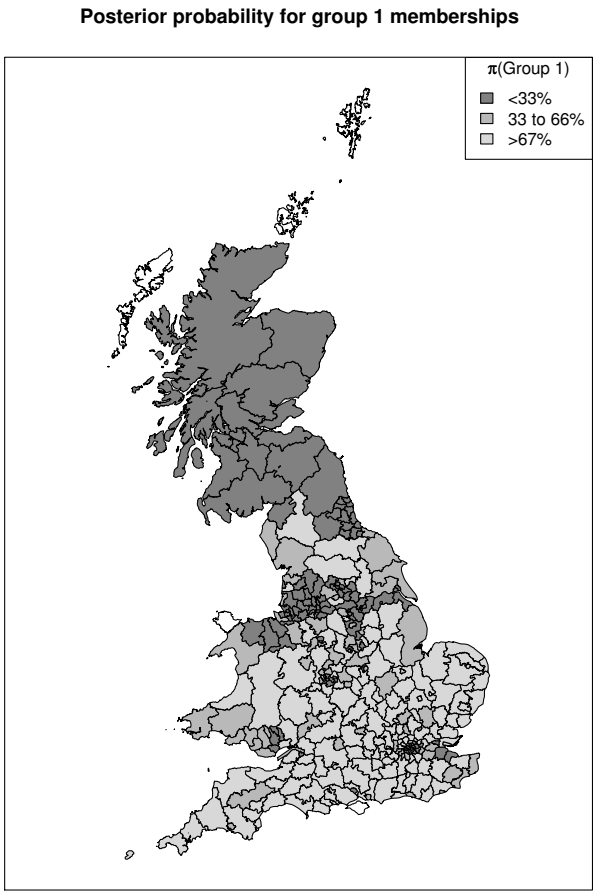

Figure 4: Posterior probability of PCT group membership for latent structure mixture model with two components and three latent variables.

Finally, we present illustrative results demonstrating the out of sample performance of the model. A random 35 PCTs had data removed for a a randomly selected cancer site (Female Lung Cancer). Results are depicted in Figure 5 which contrasts the posterior predictive density for the omitted data with the actual data that had been excluded from the model.

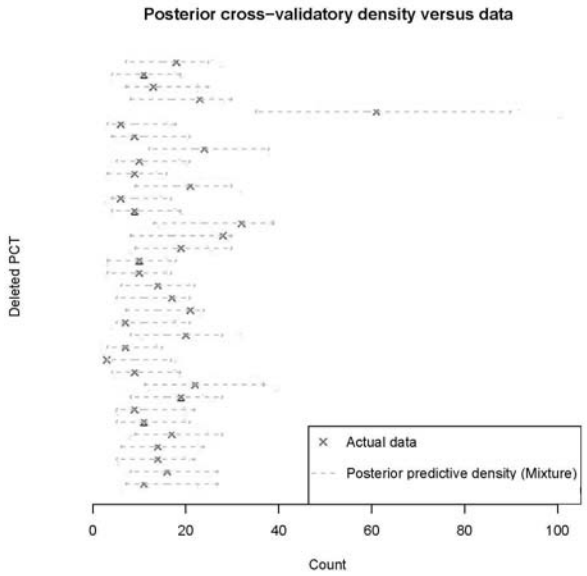

Figure 5: Out of sample posterior predictive density for 35 'PCTs' randomly removed from the Female Lung Cancer site. The actual removed data points have been superimposed.

5. DISCUSSION

We believe that the latent structure mixture model developed in this paper provide a tractable approach to handling situations in joint disease modelling where it may be anticipated that a single dependence structure, either between diseases or across geographical units, is overly restrictive. We also believe that such situations are not uncommon and that aspects of the illustrative cancer morbidity data we have examined substantiate the argument for employing mixture models as a way of avoiding unreasonable exchangeability assumptions.

Our primary focus has been on statistical methodology, rather than identifying any substantive epidemiological issues arising from the particular cancer morbidity data we have examined. That said, there could well be interesting epidemiological distinctions between the areas discriminated using our approach as reported in Section 4. It is quite striking that the areas with lowest probability of group 1 membership tending to correspond to former industrial areas of Scotland, North England and Wales, those PCTs with the highest probability of group 1 membership tending to correspond to more affluent and rural areas in Southern England. This fits well with the epidemiology of these diseases, lifestyle factors such as alcohol and tobacco consumption being more dominant in the

non-group 1 areas (hence greater lung and oral cancer) and other factors being responsible for greater breast cancer risk in the group 1 areas. However, we are cautious of over-interpretation in this regard. Further work is needed to deal with these models in such a way that the mixtures on the loadings can be disentangled from the mixtures on the latent variables, but it does appear from our results that the two structures do act differently.

We have concentrated on modelling dependency structure and not explicitly addressed use of additional covariate information on geographical units other than routine standardisation for age/sex population structure. We appreciate that in practice it is very likely that relevant additional covariate information will be available on the geographical units concerned. If so, then this can easily be handled by simply including relevant fixed effects into the linear predictor of the model we have proposed and does not present any additional methodological challenges.

In summary, we believe that incorporating a mixture distribution into a latent structure model has considerable potential in modelling multivariate disease rates. The advantages of using a latent structure model relate to the transparent way in which correlation structure is represented in the model allowing the modeller to tune this accordingly. It is less obvious how to do this within, say, the MCAR formulation where the latent structure is not explicit. We appreciate that in this paper we have not carried out any formal comparison of the fit of our proposed model to other formulations such as the MCAR. This topic is taken up and reported elsewhere in an expanded version of this paper (see [23]).

REFERENCES

[1] Aitkin, M. and Aitkin, I. (2005). *Bayesian inference for factor scores.* In "Contemporary Advances in Psychometrics" (A. Maydeu-Olivares and J.J. McArdle, Eds.), New York, Erlbaum.

[2] Banerjee, S.; Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, Chapman and Hall / CRC.

[3] Beal, M.J. and Ghahramani, Z. (2003). *The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures.* In "Bayesian Statistics 7" (J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, Eds.), Oxford University Press.

[4] Best, N. and Wakefield, J. (1999). Accounting for inaccuracies in population counts and case registration in cancer mapping studies, *J.R. Statist. Soc. A*, **162**, 363–382.

[5] Celeux, G.; Hurn, M. and Robert, C.P. (2000). Computational and inferential difficulties with mixture posterior distributions, *J. Am. Statist. Ass.*, **95**, 957–970.

[6] Christensen, W.F. and Amemiya, Y. (2002). Latent variable analysis of multivariate spatial data, *J. Am. Statist. Ass.*, **97**, 302–317.

[7] Clayton, D.G. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardised relative risks for use in disease mapping, *Biometrics*, **43**, 671–681.

[8] Czado, C.; Gneiting, T. and Held, L. (2007). Predictive model assessment for count data, *University of Washington Department of Statistics Technical Report*, Number 518.

[9] Dabney, A. and Wakefield, J. (2005). Issues in the mapping of two diseases, *Statistical Methods in Medical Research*, **14**, 83–112.

[10] Daniels, M.J. and Kass, R.E. (2001). Shrinkage estimators for covariance matrices, *Biometrics*, **57**, 1173–1184.

[11] Dawid, A.P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design, *Annals of Statistics*, **27**, 66–81

[12] Dreassi, E. (2007). Polytomous disease mapping to detect uncommon risk factors for related diseases, *Biometrical Jnl.*, **49**, 520–529.

[13] Gelfand, A. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis, *Biostatistics*, **4**, 11–15.

[14] Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach, *Biometrika*, **85**(1), 1–11.

[15] Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457–511.

[16] Geweke, J. (1992). *Evaluating the accuracy of sampling-based approaches to calculating posterior moments.* In "Bayesian Statistics 4" (J. Bernado, J. Berger, A. Dawid and A. Smith, Eds.), Oxford, Clarendon Press.

[17] Ghahramani, Z. and Beal, M. (2000). *Variational Inference for Bayesian Mixtures of Factor Analysers*, MIT Press.

[18] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction and estimation, *J. Am. Statist. Ass.*, **102**, 359–378.

[19] Green, P.J. (2003). *Trans-dimensional Markov chain Monte Carlo.* In "Highly Structured Stochastic Systems" (P. Green, N.L. Hjort and S. Richardson, Eds.), Number 27 in Oxford Statistical Science Series, pp. 179–198, Oxford, Oxford University Press.

[20] Hardy, G.; Littlewood, J. and Pólya, G. (1988). *Some Theorems Concerning Monotonic Functions* (2nd ed.), pp. 83–84, Cambridge, UK, Cambridge University Press.

[21] Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient, *Opns. Res.*, **31**, 1109–44.

[22] Held, L.; Natario, I.; Fenton, S.; Rue, H. and Becker, N. (2005). Towards joint disease mapping, *Statistical Methods in Medical Research*, **14**, 61–82.

[23] Hewson, P.J. and Bailey, T.C. (2010). Modelling multivariate disease rates with a latent structure mixture model, *Statistical Modelling*, **10**, 241–264.

[24] Knorr-Held, L. and Best, N. (2001). A shared component model for joint and selective clustering of two diseases, *J. R. Statist. Soc. A*, **164**, 73–86.

[25] Lee, S.-Y. and Song, X.-Y. (2003). Bayesian model selection for mixtures of structural equation models with an unknown number of components, *British Journal of Mathematical and Statistical Psychology*, **56**, 145–165.

[26] Leonard, T. and Hsu, J.S. (1992). Bayesian inference for a covariance matrix, *The Annals of Statistics*, **20**(4), 1669–1696.

[27] Liu, X.; Wall, M.M. and Hodges, J.S. (2005). Generalized spatial structural equation models, *Biostatistics*, **6**(4), 539–557.

[28] Lopes, H. and West, M. (2004). Bayesian model assessment in factor analysis, *Statistica Sinica*, **4**, 41–67.

[29] McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*, New York, J. Wiley and Sons.

[30] Minozzo, M. and Fruttini, D. (2004). Loglinear spatial factor analysis: an application to diabetes mellitus complications, *Environmetrics*, **15**, 423–434.

[31] Normand, S.-L.T.; Glickman, M.E. and Gatsonis, C.A. (1997). Statistical methods for profiling providers of medical care: issues and applications, *J. Am. Statist. Ass.*, **92**, 803–814.

[32] Sahu, S. and Cheng, R. (2003). A fast distance based approach for determining the number of components in mixtures, *The Canadian Journal of Statistics*, **31**, 3–22.

[33] Sargent, D.J.; Hodges, J.S. and Carlin, B.P. (2006). Structured Markov chain Monte Carlo, *Journal of Computational and Graphical Statistics*, **9**, 217–234.

[34] Spiegelhalter, D.; Thomas, A. and Best, N. (1998). WinBUGS version 1.1.1 User Manual, *MRC Biostatistics Unit Technical Report*, Cambridge.

[35] Viroli, C. (2004). Choosing the number of factors in independence factor analysis model, *Metodoloski zvezki*, **1**, 407–418.

[36] Wang, F. and Wall, M.M. (2003). Generalized common spatial factor model, *Biostatistics*, **4**, 569–582.

REVSTAT – Statistical Journal

Background

Statistical Institute of Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT – STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

— The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.

— All research articles will be refereed by at least two persons, one from the Editorial Board and another, external.

— The only working language allowed will be English.

— Three volumes are scheduled for publication, one in April, one in June and the other in November.

— On average, four articles will be published per issue.

Aims and Scope

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

Abstract / indexed in

REVSTAT is expected to be abstracted / indexed at least in *Current Index to Statistics*, *Statistical Theory and Method Abstracts* and *Zentralblatt für Mathematik*.

Instructions to Authors, special-issue editors and publishers

Papers may be submitted in two different ways:

— By sending a paper copy to the Executive Editor and one copy to one of the two Editors or Associate Editors whose opinion the author(s) would like to be taken into account, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

— By sending a paper copy to the Executive Editor, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Submitted manuscripts (text, tables and figures) should be typed only in black, on one side, in double spacing, with a left margin of at least 3 cm and not have more than 30 pages.

The first page should include the name, affiliation and address of the author(s) and a short abstract with the maximum of 100 words, followed by the key words up to the limit of 6, and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style.

This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to *PC Windows System* (Zip format), *Mackintosh*, *Linux* and *Solaris Systems* (StuffIt format), and *Mackintosh System* (BinHex Format), are available in the REVSTAT link of the National Statistical Institute's Website: http://www.ine.pt/revstat/inicio.html

Additional information for the authors may be obtained in the above link.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: liliana.martins@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

Also send always the final paper version to:

Maria José Carrilho
Executive Editor, REVSTAT – STATISTICAL JOURNAL
Instituto Nacional de Estatística, I.P.
Av. António José de Almeida
1000-043 LISBOA
PORTUGAL

Copyright