BOLETIM ISBRA

Volume 1 Número 1 Junho 2003

O Boletim oficial do Capítulo Brasileiro da *International Society for Bayesian Analysis*

CARTA DO PRESIDENTE

Josemar Rodrigues
vjosemar@power.ufscar.br

Ao assumirmos a diretoria do ISBRA estávamos conscientes das dificuldades em torná-lo uma realidade brasileira, como um canal de divulgação das aplicações dos métodos Bayesianos. Após alguns meses na diretoria, com a ajuda de vários colegas, conseguimos, dentro de nossas possibilidades, construir a homepage do ISBRA e essa primeira edição do seu boletim. Para que continuemos motivados é fundamental a participação dos associados, enviando sugestões e informações para que melhoremos os trabalhos realizados e estejam a altura de nossa comunidade Bayesiana. Como dizemos na introdução da homepage do ISBRA, a estatística Bayesiana internacional está vivendo uma nova era dentro da Estatística e nossa comunidade não pode ficar indiferente às novas tecnologias. Acreditamos que esta seja a missão do ISBRA, como um capítulo do ISBA. Para comprovar este fato, sugerimos que os associados visitem a homepage www.ams.ucsc.edu/bayes03 do "International Worshop on Bayesian Data Analysis"que será realizado em Santa Cruz, Califórnia, em agosto de 2004. O próximo desafio da diretoria será organizar o VII Encontro Brasileiro de Estatística Bayesiana (EBEB), que será realizado em São Carlos durante o período de 08 a 11 de fevereiro de 2004, com ênfase em aplicações dos métodos Bayesianos semiparamétricos e distribuições assimétricas.

Finalmente, gostaríamos de agradecer aos colegas da diretoria do ISBRA particularmente, ao Professor Hedibert Lopes (UFRJ) na elaboração e edição do primeiro boletim do ISBRA e ao Professor Francisco Louzada-Neto (UFSCar), cuja participação foi fundamental na construção da homepage.

Cordiais saudações bayesianas!

Josemar Rodrigues

CARTA DO EDITOR

Hedibert Lopes
hedibert@im.ufrj.br

É com grande prazer que organizo o primeiro boletim totalmente voltado para a comunidade Bayesiana dentro do Brasil. Como todos vocês sabem, o ISBRA foi criado durante o SINAPE de 2000, em Caxambú, e uma das primeiras idéias que tivemos (leia-se: copiamos) foi a de criar um boletim nos moldes do *ISBA bulletim*. Pode se dizer que esse primeiro boletim foi motivado pelo desejo de vários membros ativos dessa nossa sociedade e se materializou após a última Escola de Séries Temporais, que aconteceu em Conservatória, no interior do Rio de Janeiro. O boletim será publicado semestralmente, nos meses de Junho e Dezembro de cada ano.

Organizei a sessão Bayesiana da Escola e sugeri a criação de um prêmio (ainda sem nome e sem fundos!) para os dois melhores trabalhos orais e para os 4 melhores pôsteres. Além dos resumos desses trabalhos, o boletim também traz uma entrevista muito interessante e historicamente rica com o Professor Hélio Migon, responsável, direta ou indiretamente pela formação de pelo menos 110% dos *Bayesianos do Rio*. Gostaria de convidá-los a participar comigo numa jornada suave pelas próximas páginas.

SUGESTÕES

QUALQUER TIPO DE SUGESTÃO, RECLAMAÇÃO, DOAÇÃO, ETC QUE POSSA SER UTILIZADA PARA MELHORAR A QUALIDADE DO BOLETIM É MUITO BENVINDA.

vjosemar@power.ufscar.br

HÉLIO S. MIGON

por Hedibert Lopes
hedibert@im.ufrj.br

O Professor Hélio S. Migon é titular do Instituto de Matemática, onde vem atuando intensa e significativamente, como essa entrevista mostrará, desde 1978 no Departamento de Métodos Estatísticos e na Pós-Graduação em Estatística. Para quem não sabe, não foi na UFRJ que o Migon, como costumamos chamá-lo, começou sua carreira estatística; carreira essa que não se (de)limitou a vida universitária. O Migon é estatístico de carteirinha, tendo começado calcular suas primeiras médias e variâncias durante seu segundo grau (ensino médio?) da Escola Nacional de Ciências Estatísticas (ENCE). Bom, é melhor eu parar por aqui antes que eu termine toda a estória (e o histórico) antes da entrevista.

1. Migon, conte-nos um pouco dos seus dias de ENCE como aluno secundário e universitário.

Muito cedo, final do ginásio, achei que gostaria de estudar economia. Uma tia muito especial, professora de Matemática, descobriu num anúncio de jornal o curso técnico de estatística. Além disto, me explicou que isto – Estatística – tinha a ver com economia. Lá fui eu! Fiz uma prova de seleção e entrei para o técnico da ENCE. Excelentes professores e um ambiente liberal (fumavam na sala de aula, podiamos entrar e sair a qualquer momento – eramos adultos!!!) que muito me agradava. Vale dizer que eu tinha feito um ginásio num colégio católico (muito severo!) em São Paulo. Após três anos terminamos somente um 12 a 15 alunos. Dentre eles o Basílio, o Zé Ferreira (Carvalho), Carlinhos e espero não ter feito omissões significativas para este papo.

Eles prosseguiram imediatamente para o curso universitário e eu fiz um concurso para o Banco do Brasil no início de 64. Passei e fui trabalhar fora do Rio - em Maringá - PR. Uma grande aventura para um pós-adolescente. Viver numa república, cuidar da propria vida. Foi um período de grandes apredizados. Só retornei a ENCE em 67, quando consegui uma transferência para o recém criado Banco Central. Assim acabei me formando duas turmas depois desse pessoal. O curso de graduação na ENCE exigia uma dedicação intensa. Os fins de semana eram sistematicamente utilizados para estudar.

2. Você se arriscaria em listar alguns fatos marcantes desse período?

As disciplinas eram todas de conteúdo muito pretencioso, mas alguns professores eram jovens e sem muita experiência. Em geral haviam simplesmente sido bons alunos e se propunha a repetir o conteúdo apresentado pelos seus mestres. Desta forma havia um desgaste natural no processo de transmissão do conhecimento, além de uma degenerecência decorrente da alta endogenia do sistema. Por exemplo, em Probabilidade 1 o texto mencionado era o Feller. Num segundo curso as referências que me lembro incluiam o Wilks e o Fiz, ambos muito formais. Outro texto muito mencionado era o Cramer. Finalmente a história: quando me apresentaram aos teste de hipótese diziam que olhando a forma da alternativa eu poderia saber a natureza da região de rejeição. Sacudiam a mão para a direita, quando a alternativa era uma desigualdade do tipo maior do que. Eu jurava que não podia ser assim. Tinha de haver um teorema serio para isto. Descobri por minha conta (”emprestando”um livro do Nelson Chagas -o Freeman) o velho lema do Neyman-Pearson.

Sempre levava para o trabalho (meio dia somente) as notas de aula. Fazia minhas tarefas rapidamente para poder dar uma lida nas aulas. Meu chefe não admitia isto e sistematicamente inventava algo mais para ser feito, mesmo que outros colegas estivessem atentamente examinando o Jornal dos Sports ou o velho Pasquim. Isto foi muito bom pois me ajudou a descobrir que aquela não era minha praia.

O ano de 68, aquele que nunca acabou (risinhos...estou casado faz é tempo), foi muito intenso. Muitas passeatas, algumas com tiroteio na Cinelândia, prisões de colegas próximos e sempre no final um sanduiche no sujinho. Acabei virando o cara do diretório acadêmico. Muito mais pelo fato de que eu satisfazia a um certo decreto lei que exigia no diretórios ”bons”estudantes (com média acima de sete, acho que era isso!) do que por minhas convicçõe políticas. Fiz meu trabalho direitinho: o diretório participava ativamente do chamado ME (movimento estudantil) dos anos 68/69,

tanto que acabou sendo o último daquele período. Mas também mantinha um curso preparatório para o vestibular que se mostrou muito eficiente e fazia propostasatrevidas, como por exemplo a da reforma da grade curricular. Em poucas palavras, dois anos básicos e depois algumas diferentes linhas de aplicação, incluindo economia, computação, etc. Lamento não saber onde anda este documento. A lembrança que tenho é de que pretendíamos, sem ser pretencioso, algo como o MORSE (Mathematics, Operation Research, Statistics and Economics) que vi funcionando em Warwick nos anos 80 ou como o SCORE (Statistics, Computing, Operation Research and Economics) do Impirial College.

Achava que precisava melhorar minha formação. Depois de algumas tentativas mal sucedidas (Matemática moderna! - Jairo Bezerra) descobri, por influência de um grande amigo (Frederico de Carvalho), o caminho do IMPA (Praça Tiradentes). Fiz alguns cursos, incluindo os do Caio da USP de Probabilidade (Gnedenko) e Inferência (um texto do Caio do Colóquio de Poços de Caldas, 1969).

Isto nos anos 69 e 70. Nesta altura o BC era um grande estorvo na minha vida. Queria fazer algo acadêmico.

3. O que você fez logo após se formar?

Quando me formei em 70 o sistema de pós-graduação no Brasil estava principiando. Na área de estatística não se tinha nada firmemente estruturado. Surgiram duas oportunidades: mestrado no CIENES pelo BC (muitas mordomias, bolsa em dólar, futuro garantido, mas talvez na volta a pilotagem de uma máquina de escrever) ou o mestrado em Pesquisa Operacional no Instituto Militar de Engenharaia. A primeira furou por conta do Governo Allende no Chile. O Itamarati não liberou ninguém do Brasil e a segunda não era exatamente o que eu desejava. Acabei sendo aceito, pelo Caio, para o mestrado da USP. Como eu tinha que me manter fui ser auxiliar de ensino na Unicamp (licenciado do BC), onde já trabalhavam o Zé Ferreira, a Gaby e o saudoso Ronaldo Eckstein. Fomos para lá eu, o Wagner Borges e, o também memorável, Maul. Lá conheci entre outros o Paulo Bravo (já aposentado pela UFRJ). Foi um ano de muitas aventuras e aprendizados.

4. Fazer mestrado na USP e lecionarna UNICAMP não gerava muitas horas na estrada?

É verdade! Parte do tempo eu gastava na via Anhanguera. Íamos duas tardes por semana para a USP. Isto foi somente por um ano. Tivemos dificuldades políticas no IMECC e acabamos, praticamente todo o Departamento de Estatística, demitidos pelo reitor Zeferino Vaz. Eu, Ronaldo e Wagner fomos para USP, o Paulo Bravo para o IM e o Maul para a EBAP/FGV.

5. Qual foi seu projeto de tese de mestrado?

Como já mencionei, tudo era experimental. Fiz um apanhado de cursos, incluindo Probabilidade Avançada, um exame de qualificação 'penoso' e uma dissertação em métodos não-paramétricos. A idéia era levantar, em pequenas amostras, a função de poder de alguns testes para análise de modelos de dois fatores de classificação. O popular teste F e algumas alternativas não-paramétricas. Existiam resultados assintóticos e desejavamos estudar a função de poder em amostras pequenas. Assim deveríamos gerar amostras sob a alternativa segundo diversas distribuições e estimar a probabilidade de rejeição. Estas tinham caudas mais pesadas do que a normal e incluiam dentre outras diversas normais contaminadas, a Cauchy a dupla exponencial. Mais detalhes podem ser visto no artigo do *Journal of Computation and Simulation*, pasmem! de 1978. Acho que este foi um dos primeiros artigos publicados no exterior.

Voltando ao assunto: a duplicadora deslocou as colunas do cartão de sorte que a coluna seis (comentários no Fortran) foram deslocadas para a sétima e os *jobs* não rodavam mais! Foram meses ate descobrir isto.

6. E o que você lecionou na UNICAMP?

Como lhe disse fiquei somente um ano na Unicamp. Lembro-me que minha primeira turma foi de Probabilidade e Estatística para Engenharia. Eram cerca de 120 alunos num grande auditório e uma imensa inexperiência do auxiliar de ensino. A começar pela escolha do texto - o livro do P. Meyer. Já na USP (de 1972 a 1975) lecionei diversas disciplinas de serviço para economia, administração e matemática. O melhor todavia foi ministrar a disciplina de Testes Não Paramétricos para a graduação

de Estatística e também para o mestrado. Não vou contar o nome das várias alunas ilustres para não ser deselegante com elas, fornecendo evidências para que vocês infiram com precisão a idade delas.

7. Depois dessas muitas aventuras e aprendizados, onde foi que você pousou?

Acabado o mestrado em outubro de 74 e com a grana muito curta (dois filhos para criar) apareceram duas opções: sair para um doutorado no exterior ou procurar emprego na "indústria". Foi nesse momento que surgiu um convite de um ex-professor da ENCE (Luiz Carlos da Rocha) para uma entrevista na Telebrás em Brasília. Lá fui eu de paleto e gravata. Acabei ficando um ano e poucos meses por lá. Embora o ambiente (o departamente em que trabalhei deu origen ao CPqd da Telebrás) fosse legal, o trabalho não me agradava. Nesta época surgiu a oportunidade de fazer um curso de "economia"no CENDEC/IPEA, onde aprendi, finalmente, um pouco de economia. Numa ida a um congresso de PO no hotel Intercontinental no Rio encontrei o Raul Mourão que me convidou para ir pro SERPRO. Tinhamos vários amigos em comum, incluindo o João Ismael da UFRJ, que conheço desde a época dos cursos do IMPA. Este talvez tenha sido o principal elo de ligação com o SERPRO.

8. É aí que começa sua história na UFRJ?

A vinda para o Rio me reaproximou da UFRJ onde era eu tinha vários amigos: Paulo Bravo, João Ismael, Annibal, e, principalmente, o Basílio, recém chegado do doutorado na Inglaterra. Surgiu um concurso em 78 para o DME. Eram vários candidatos, entre eles o Gauss Cordeiro.

Nessa época as atividades no Serpro eram muito entusiasmantes. Seminário de Análise Multivariada, desenvolvimento de softwares para Amostragem e, sobretudo, um grande sentido prático, pois usavamos tudo isto nos projetos para Receita Federal. Simulador de legislações do Imposto de Renda, Zoneamento agrário do nordeste, Sistemas de detecçãoo de *outliers* para dados da Receita Federal, seleção de empresas para auditagem, etc. De lambuja alguns cursos de verão no IMPA. Me lembro de um curso do Bussab (recem retornado do LES) em Amostragem e outro de Séries Temporais do Vitor Yohai.

O principal projeto, todavia, foi o desenvolviemnto de um modelo econométrico para mercado internacional do café. Este, embora não fosse um projeto típico do nosso Setor de Métodos Quantitativos nos chegou de forma muito interessante e que merece ser contada.

9. Então nos conte, por favor!

Diz a lenda (ou realidade!, já não sei bem!) que numa reunião da OIC (Organização Internacioanl do Café) surgiu a idéia de se estudar a viabilidade de um estoque regulador de preços. Isto é, um organismo neutro que fizesse os preços flutuarem dentro de uma faixa previamente definida evitando grandes perdas para produtores e consumidores. A discussão só prosseguiria se um dos países produtores (subdesenvolvidos) apresentassem um modelo matemático para alimentar as discusssões. Assim, trabalhamos meses (nasceu meu terceiro filho, e na maternidade eu lia um modedo discutita a dinâmica de preços e produção no mercado de cacau) contruíndo um modelo econométrico composto de várias equações estocásticas e algumas equações de balança retratando o mercado internacional do café. Isto e descreviam as relações entre os principais fluxos e preços envolvendo os principais consumidores (países desenvolvidos) e os principais produtores (subdesenvolvidos). Quem quiser saber mais sobre este exemplo deve ler nosso livro (o Hedibert é co-autor) de Análise de Decisões do SINAPE de 2002. Fiz questão de lembrar deste exemplo como sendo um caso típico do famoso erro do tipo III: resolver o problema errado! Mostramos que era viável construir um *buffer stock* no curto prazo e esquecemos que se tudo ficasse cor de rosa (preços variando dentro da tal faixa) todos seriam estimulados a plantar café e o mercado ficaria, no longo prazo, encharcado! E os preços despencariam.

10. Demorou muito para a idéia do doutorado retornar?

As aulas na UFRJ como tempo parcial e o trabalho no Serpro se combinavam harmoniosamente. O tempo parcial me poupava das enfadonhas reunioes de departamento, e os períodos de baixo astral no Serpro eram compensados pelo estimulante contato com os alunos. Nesta época, as conversas com o Basílio acabaram me animando a passar a limpo a questão do doutorado, adiada deste 75.

Estas experiências me levavam a duas paixões na estatística: amostragem e métodos de previsão. Assim apliquei para Southampton e para Warwick. As aceitações chegaram e o desempate se deu por conta de vários fatores. Um deles foi a visita do Adrian Smith ao Departamento de Métodos Estatísticos em 79/80 [leiam sua entrevista ao Boletim ISBA de Marco de 2003!] . Este foi o primeiro contato verdadeiro com o argumento Bayesiano. Na época da USP (72/73) andei pegando o livro do Box e Tiao (recém publicado eu creio) para folear, mas confesso que não conseguia perceber a importância. Desta forma cheguei em Warwick em outubro de 80 com uma mão na frente e outra atraz e um imenso desejo de trabalhar em funções de transferência. Na verdade já vinha tentando ler umas coisas do Zellner e Palm sugeridas pelo Frederico de Carvalho que havia voltado de Louvan (ex-aluno do Drezé). Ele também havia me sugerido ler o livro do Zellner. Falei dessas experiências pro Jeff Harrison e ele, com seu permanente bom humor e vale dizer sua grande auto-confiança, me disse que poderíamos discutir isto mais adiante. Que eu fosse lendo umas notas de aula dele e que atendesse a algumas classes no outono de 80. Fiz um curso do Tony O'Hagan de Inferência Bayesiana e outro do Peter Walley em Análise exploratória de dados. No final deste termo o Jeff me dispensou de fazer cursos e começamos a trabalhar em modelos dinâmicos não lineares. Começamos com um modelo sazonal multiplicativo usando fatores de descontos distintos para as componentes de tendência e de sazonalidade. Não tardou a aparecer um projeto envolvendo dados reais na área de marketing. Desejava-se estabelecer uma relação dinâmica - função de tranferência - entre investimentos em propaganda e memorização da propaganda. A idéia era de que se alguém desejasse mudar de marca de um certo produto de consumo regular, provavelmente optaria por aquela mais lembrada. Isto era útil para se romper a questão de causalidade envolvida na relação entre propaganda e vendas! Novamente não dá pra explicar tudo aqui - quem quiser pode olhar minha tese ou o *paper* no Bayesian Statistics II.

Estes resultados foram apresentados num seminário conjunto dos departamentos de estatística de Warwick e Birmingham e serviu como exame de candidatura ao doutorado. Nesta época está chegando ao Departamento o Mike West, recém doutorado em Nottingham sob a orientação do Adrian. Após minha apresentação discutimos várias vezes e tive a oportunidade de mostrar a ele os primeiros resultados que vinha obtendo com modelos dinamigos generalizados. Especialmente o beta-binomial e o Poisson-Gama. Daí pintou uma grande parceria que nos enriqueceu mutuamente. O velho Jeff nos oferecendo genereosamente suas brilhantes idéias e um monte de experiência, o Mike com sua urgência em construir uma carreira, que acabou se mostrando meteórica, além de brilhante, e eu correndo atrás de resultados para finalizar a tese de doutorado no prazo de três anos. Isto tudo junto acabou gerando, em 1985, um paper no JASA. Minha correria, todavia, acabou caracterizando um pequeno engano. Deixei de preparar alguns outros artigos da tese para finalizar tudo em três anos e alguns meses e voltar correndo pro Brasil.

11. Acredito que os alunos de mestrado, agora espalhados por vários institutos pelo Brasil, gostariam muito de saber como se desenvolvia a 20 anos o processo de doutoramento no exterior, desde os primeiros contatos até o dia da volta passando pelas experiencias profissionais e pessoais (como por exemplo, contactar o pessoal ou ter notícias do Brasil)

Não era muito diferente de hoje em dia. As exigências costumavam ser as mesmas: exame de proficiência em inglês, teste de conhecimentos (tipo GRE), etc. As exigências variam entre universidades e países. As principais diferenças ficavam por conta de um sistema de bolsas muito incerto e contatos internacionais difíceis, em razão de que tinhamos poucos doutores formados.

12. Meu primeiro contato com você foi durante um curso de inferência estatística na graduação da UFRJ em 1988. Se bem me lembro você ainda mantinha uma posição paralela no BC (ou seria BB?) e já estava no Brasil a 4 anos. Como foi a vida pós-doutorado? Fale-nos sobre a estatística Bayesiana naquele momento no Brasil.

Quando voltei do doutorado fique algum tempo no Serpro. Embora eu tenha ficado três anos

somente no exterior, o retorno foi sofrido. Tinham acontecido mudanças significativas na sociedade brasileira. Foi nesta época que inauguramos um contato estreito com o IPEA. Eu e o Gutemberg (doutorando da PUC) fizemos em 86 algumas previsões de indicadores econômicos usando modelos dinâmicos e também Box e Jenkins. Tudo direitinho com intervalo de confiança e tudo, o que não era muito comum na época. Saí do Serpro antes que o setor de métodos quantitativos se esfacelasse. Andei pela PUC, pela ENCE e acabei ficando em dedicação exclusiva no IM/UFRJ. Foi por aí que nos conhecemos. Vou falar mais de como as coisas evoluíram na UFRJ. Numa visita a Inglaterra em 86/87 propuz ao Dani solicitar uma bolsa de recém doutor para a UFRJ. Ele topou e veio para cá em 87. A partir daí passamos a trabalhar juntos e com um sucesso relativo. Embora sem muito planejamento incluímos o Ajax (IPEA) no processo de formação de nossos alunos de mestrado. A UFRJ já tinha uma certa inclinação Bayesiana pelas influências do Basílio e do Marlos Viana e não foi difícil consolidá-la naturalmente. Passamos a ministrar as disciplinas de Inferência e Modelos dinâmicos, além de modelos lineares generalizados.

13. Um desses dias você me disse ter contado uns 30 alunos de mestrado nos últimos 20 anos (ou 30 anos?). Como essa vasta experiência contribuiu a implementação da ousada (e bem sucedida) idéia de um doutorado em estatística na UFRJ?

Orientamos diversos doutorandos na Coppe - eu, Dani, Basílio e Anibal. Acho que chega a uns 25. Isto nos deu um grande respaldo para propor nosso programa de doutorado em estatística. O processo de reconhecimento do curso de doutorado foi longo e bastante exigente. Conversamos muito com o Coordenador da Matemaática na CAPES (na época o Mário Jorge da UFMG) e fomos incorporando todas as sugestões. Nosso regulamento contempla várias regras interessantes. Por exemplo temos uma comissão de avaliação externa. Vale destacar todavia que o grande fator para o sucesso deste projeto foi o empenho do Dani.

Agradeço ao Migon, tenho certeza em nome de todos os membros e muitos outros, por gentilmente nos conceder essa cativante entrevista.

BAYESIAN COVARIATE SELECTION IN MULTIVARIATE HIERARCHICAL MODELS

Patrícia Ziegelmann
patyz@mat.ufrgs.br

1. Introduction

Pharmacokinetics (PK) are studies designed to characterise the time course of a drug after its introduction into the human body. Typical data of PK studies involve measurements of drug concentration over time for a number of individuals. Individual and treatment characteristics are also available. Population PK studies play an important role in drug development since its correct understanding provide guidance to specify effective dosage regimen for the target population. A suitable statistical model for population PK studies should accommodate the two sources of variability intrinsically observed in the data: within and between individuals. This structure is naturally adjusted by using hierarchical two stage models. In the first stage, the relationship between concentration and time is modelled for each individual, addressing the within individual variation. In the second stage, the among individuals variation is accommodated through the specification of a model for the individual parameters. In PK population studies, the functional form used in stage 1 usually comes from compartmental model theory and it is nonlinear in its parameters. Because of this nonlinearity, the entire model is referred to as a hierarchical nonlinear model. To model the variability between individuals in the second stage one can consider the naive approach where the variability is considered due to random variation only. In this work we also consider a sistematic component through the use of covariates (individual and treatment characteristics). The decision of how many and which covariates should be included in the model is made through a covariate selection procedure.

The hierarchical nonlinear model we use involves complex covariance structures to accommodate the variability of observations within and between individuals and a nonlinear function in stage 1. Also population PK studies using clinical PK data usually involve measurements of drug concentration at sparse time points. These characteristics make the process of inference a very difficult challenge. The recent advances in using MCMC procedures have allowed practical implementation of Bayesian methods to deal with complex models. Also, the great flexibility of Bayesian methods to incorporate complex covariance structures reveals Bayesian procedures as a powerful methodology for PK population models.

In this work we concentrate in the multivariate case where different PK responses are considered in the same model. To accommodate the multivariate response we use a combined model where possible correlation among responses can be taken into account without require the same number of observations across responses. The inference process is carried out using a fully parametric Bayesian approach. Our main concern is to suggest a Bayesian procedure to the covariate selection step in the second stage. In this regarding we introduce an extension of the "Gibbs variable selection" procedure proposed by Dellaportas, Forster and Ntzoufras (2002). This extension is appropriated to the multivariate case.

2. Growth Hormone Data

A set of data from the "Growth Hormone - Olympic Games 2000" project has been used to illustrate the methodology described in this work. The data result from a double-blind, randomized and placebo controlled study. Around one hundred healthy volunteers of both genders with ages between 18 and 35 years old and physically active participated in the study. Each one of 29 individuals was randomized to receive either single or double dose of recombinant human Growth Hormone for 28 days, followed by a 56-day wash-out period. The data we use are the measurements of the concentration of two different markers (IGF-I and PIIIP) on the days 0 (before starting the treatment), 21 and 28 (during the treatment) and 30,33,42 and 84 (during the wash-out period). Information on characteristics like age, height, weight and gender are also available.

To model the relationship between concentration and time at stage 1 of the hierarchical nonlinear model we use a one-compartment model with linear-first-order elimination (Gibaldi, 1982). We assume the principle of superposition. Hence, considering that the concentrations are observed immediately after each dose is taken and each

individual receives the same fixed dose between dose intervals of one day, the predicted concentration for the j^{th} observation of response k of individual i is given by

$$Bas_{ik} + \frac{Dose_i}{Vol_{ik}} \times \frac{1-\exp\left(-d_{ikj}\times\frac{Cl_{ik}}{Vol_{ik}}\right)}{1-\exp\left(-\frac{Cl_{ik}}{Vol_{ik}}\right)} \times \exp\left(-t_{ikj}\times\frac{Cl_{ik}}{Vol_{ik}}\right)$$

where $Dose_i$, d_{ikj} and t_{ikj} are observed dose history covariates and Bas_{ik}, Cl_{ik} and Vol_{ik} are unknown PK parameters to be estimated. These parameters have clinical interpretation.

3. Statistical Model

The statistical methodology behind PK studies involves mixed effects models for longitudinal data. The natural framework is hierarchical models with two stages. To follow a Bayesian framework, a third stage, with the prior for the population parameters from the first and the second stages, is incorporated. To accommodate the multivariate response we construct a combined model (Ziegelmann and Brown, 2001). The great advantage of this approach is that the possible correlation between markers can be taken into account without requiring the same number of observations for each marker. The idea is to build a vector $\boldsymbol{y}_i$ for each individual i consisting of the observations of IGF-I and the observations of PIIIP stacked. In this section we present the three stage hierarchical model used in this work.

Stage 1: Suppose $i = 1, \ldots, I$ individuals, $k = 1, \ldots, K$ and $j = 1, \ldots, n_{ik}$ observations for each individual i and response k. A nonlinear regression model to address the within individual variability is specified as $y_{ikj} = f_k(\boldsymbol{x}_{ikj}, \boldsymbol{\theta}_{ik}) + \varepsilon_{ikj}$, where $f_k(\boldsymbol{x}_{ikj}, \boldsymbol{\theta}_{ik})$ takes the form of the one-compartment model showed in Section 2 for all k, $\boldsymbol{x}_{ikj}$ are the known dose history vectors, $\boldsymbol{\theta}_i$ are vectors of individual PK parameters and ε_{ikj} are within individual random errors.

Stage 2: Consider the (KL) dimensional vector $\boldsymbol{\theta}_i$ which is the combined vector of individual PK parameters. Suppose that all individual covariates are constant over time within an individual. A multivariate linear regression model to address the variability among individuals is specified as $\boldsymbol{\theta}_i = \boldsymbol{A}_i\boldsymbol{\mu} + \boldsymbol{\delta}_i$, where $\boldsymbol{A}_i$ is a fixed known design matrix, $\boldsymbol{\mu}$ is a vector of fixed parameters and $\boldsymbol{\delta}_i$ are vectors of between individual random errors.

Stage 3: A hyperprior distribution for all the parameters in the model is specified as $f(\boldsymbol{\tau}, \boldsymbol{\mu}, \Sigma_M^{-1}, \Sigma_\theta^{-1}) = \prod_{k=1}^{K} f(\tau_k)\ f(\boldsymbol{\mu})\ f(\Sigma_M^{-1})\ f(\Sigma_\theta^{-1})$, using conjugated distributions.

4. Gibbs Variable Selection for Multiple Linear Regression

In this section we introduce an extension of the Gibbs variable selection procedure proposed by Dellaportas, Forster and Ntzoufras (2002). The procedure involves including a latent variable to identify the possible models. Our extension is appropriate to select a subset of covariates from a set of potential covariates in the case of multivariate linear regression. The idea is to have a subset of covariates on which to base the multivariate linear regression in stage 2 of the multivariate hierarchical nonlinear model presented in Section 3.

Consider the combined vector of individual PK parameters $\boldsymbol{\theta}_i^T = (\boldsymbol{\theta}_{i1}^T, \ldots, \boldsymbol{\theta}_{iK}^T)$ as the dependent variable. Recall that $k = 1, \ldots, K$ represents the response variables of the hierarchical model. Each vector $\boldsymbol{\theta}_{ik}$ has dimension L, which is the number of individual PK parameters associated with response k. That is, the dependent variable $\boldsymbol{\theta}_i$ is a vector with dimension KL and elements θ_{kl}. Consider a set of individual covariates $A_1^*, \ldots, A_P^*$. Also, consider that each parameter θ_{kl} can be modelled as $\theta_{kl} = \mu_{kl0} + \mu_{kl1}\ A_1^* + \cdots + \mu_{klP}\ A_P^*$. Further, let $\boldsymbol{\mu}$ be a vector of regression coefficients, $\boldsymbol{\mu}^T = (\ \mu_{110}, \mu_{111}, \ldots,\ \mu_{11P},\ \ldots, \mu_{1L0}, \mu_{1L1}, \ldots, \mu_{1LP},\ \ldots, \mu_{K10}, \mu_{K11}, \ldots, \mu_{K1P},\ \ldots,\ \mu_{KL0}, \mu_{KL1}, \ldots, \mu_{KLP}\ )$ and $\boldsymbol{\gamma}$ a vector of indicator variables, $\boldsymbol{\gamma}^T = (1, \gamma_{111}, \ldots,\ \gamma_{11P},\ \ldots,\ 1, \gamma_{1L1}, \ldots, \gamma_{1LP},\ \ldots, 1, \gamma_{K11}, \ldots, \gamma_{K1P},\ \ldots,\ 1, \gamma_{KL1}, \ldots, \gamma_{KLP}\ )$. Here $\gamma_{klp} = 1$ indicates that the covariate p is included in the model for the parameter θ_{kl}. Notice that this specification for $\boldsymbol{\gamma}$ assumes that the constant μ_{kl0} is a fixed parameter for each θ_{kl}. Then, the regression equation at stage 2 of the hierarchical model takes the form $\boldsymbol{\theta}_i = \boldsymbol{A}_i\ \boldsymbol{\vartheta} + \boldsymbol{\delta}_i$, where $\boldsymbol{\vartheta}$ is a vector of dimension $KL(P+1)$ and elements $\vartheta_{klp} = \gamma_{klp}\ \mu_{klp}$, for $k = 1, \ldots, K$, $l = 1, \ldots, L$ and $p = 0, \ldots, P$. The matrix $\boldsymbol{A}_i$ is a fixed known design matrix associated with individual i.

It is defined by $\boldsymbol{A}_i = \text{diag}\,[\underbrace{\boldsymbol{A}_i^*\ ,\ \ldots\ ,\ \boldsymbol{A}_i^*}_{KL\ \text{elements}}]$, where $(\boldsymbol{A}_i^*)^T$ is a vector of covariates associated with individual i. That is, $\boldsymbol{A}_i^* = (1, A_{i1}^*, \ldots, , A_{iP}^*)$, where A_{ip}^* is the value of the covariate p for individual i for $p = 1, \ldots, P$.

The hyperprior distributions defined at stage 3 of the hierarchical model have now to incorporate $\boldsymbol{\gamma}$. Here, the parameter $\boldsymbol{\mu}$ is assumed to depend on $\boldsymbol{\gamma}$ as $f(\boldsymbol{\gamma}, \boldsymbol{\mu}) = f(\boldsymbol{\mu}|\boldsymbol{\gamma})\ f(\boldsymbol{\gamma})$. Thus, considering the partition of $\boldsymbol{\mu}$ into $(\boldsymbol{\mu}_\gamma, \boldsymbol{\mu}_{\setminus\gamma})$, the hyperprior $f(\boldsymbol{\gamma}, \boldsymbol{\mu})$ is also partitioned into $f(\boldsymbol{\gamma}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\mu}_{\setminus\boldsymbol{\gamma}}) = f(\boldsymbol{\mu}_{\setminus\boldsymbol{\gamma}}|\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})\ f(\boldsymbol{\mu}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\ f(\boldsymbol{\gamma})$. The elements of this equation is called model prior and pseudoprior respectively. The pseudoprior will be used to update the $\boldsymbol{\mu}$´s which are not included in the model. Therefore the performance of the Gibbs variable selection is closely related to the choice of the pseudopriors. Therefore, the multivariate nonlinear hierarchical model modified to incorporate a covariate selection procedure has the following three stages:

Stage 1: The same nonlinear regression model as described in Section 3.

Stage 2: The multivariate linear regression model described in this section.

Stage 3: The same hyperpriors defined in Section 3 where $f(\boldsymbol{\mu})$ is replaced by $f(\boldsymbol{\gamma}, \boldsymbol{\mu})$ as above.

Therefore, in order to complete the model, the hyperpriors $f(\boldsymbol{\gamma})$, $f(\boldsymbol{\mu}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})$ and $f(\boldsymbol{\mu}_{\setminus\boldsymbol{\gamma}}|\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ are specified in stage 3 using conjugated distributions. Assuming conditional independence of μ_p given $\boldsymbol{\gamma}$ we can see that the prior $f(\boldsymbol{\mu}|\boldsymbol{\gamma})$ is given by $f(\boldsymbol{\mu}|\boldsymbol{\gamma}) = \prod f(\mu_{klp}|\gamma_{klp})$, where each $f(\mu_{klp}|\gamma_{klp})$ is given by $f(\mu_{klp}|\gamma_{klp}) = \gamma_{klp}\ f(\mu_{klp}|\gamma_{klp} = 1)\ +\ (1 - \gamma_{klp})\ f(\mu_{klp}|\gamma_{klp} = 0)$.

5. Model Implementation

Bayesian inference procedures are based on the joint posterior distribution of all unknown quantities in the model. Analytical procedures are not possible for our model due to the nonlinearity on the parameters $\boldsymbol{\theta}$. Also, numerical integration is affected by the large number of parameters (which increases linearly with the number of individuals. In this work we use a Metropolis-within-Gibbs procedure with a covariate selection step appropriated to the multivariate case. As a practical problem of implementation notice that the number of possible covariate models can be very large (262,144 in our example). As a consequence, the percentage of visits to each model tend to be very small even for the most probable ones. Also, the computation time to run one iteration of the algorithm is not very low. Based on this, we propose a model building strategy with three steps. In the preliminary step a saturated model (all potential covariates are included for all the different PK parameters) is fitted as a pilot run. Then, ergodic summaries for the mean population parameters are calculated based on the final iterations. After, these summaries are used to specify the parameters of the pseudopriors. The second is the covariate selection step. Here, the strategy described in Section 4, aiming to identify promising subsets of covariates, is implemented. So, in the final step, a covariate model as described in Section 3 is fitted using the promising models.

6. Example

The example involves the data from the Growth Hormone study. The analysis was carried out using C codes written by the author. We consider treatment (single or double dose), age and body surface (function of height and weight) as potential covariates. Also, we assume that all individual covariates are constant across time within each individual (using mean values when necessary).

Preliminary step: A pilot run with 500,000 iterations of the saturated model was carried out. The posterior mean and standard deviation for all μ_{klp} were calculated over 25,000 values taken from the last 250,000 iterations (taking one value every ten iterations). These summaries were considered as the respective mean and standard deviation of the normal pseudopriors.

Covariate selection step: A Metropolis-within-Gibbs strategy was applied for 50,000 iterations. During these iterations a total of 3,341 out of 262,144 different models were visited. Figure 1 shows the evolution of the posterior model probability for the models whose posterior probabilities were greater than 0.01. As an example, the most frequently selected models have age for the parameter θ_{11}, age and BodyS for θ_{21} and treatment for θ_{23}. Altogether the six most frequently models were visited around 29% of the total of iterations. It is quite a high percentage considering the huge number (262,144) of possible models.

Final model step: In this step a covariate model is fitted assuming the "best" model found in the earlier step. A single long run of a Metropolis-within-Gibbs strategy was performed with 1,000,000 iterations. The first 600,000 iterations were discarded as the burn-in period. The sample was taken from the remaining iterations (keeping one from every ten iterations). As an example of analysis, Figure 2 shows estimated predictive means with ± 2 predictive standard deviation curves for two individuals on the single dose arm.

7. Discussion

A formal approach to covariate selection is particularly important when the covariate model requires parsimony. In the multivariate hierarchical model that we have described, the correlations of the dependent variables (the individual PK parameters) between themselves and with the potential covariates are likely to be very complex. These aspects increase the care needed to decide about which covariates should be included in the model. In this work we have described a Bayesian variable selection procedure appropriated to multivariate linear regression. The implementation is rather straightforward and seems to work quite well as it is shown in our example. The procedure is specially appealing when the number of possible models is very high such as the multivariate model described.

Referências

[1] Carlin, B.P. and Chib, S. (1995) Bayesian Model Choice via Markov Chain Monte Carlo. *Journal of the Royal Statistical Society* B, 57(3), pp. 473-484.

[2] Dellaportas, P., Forster, F.F and Ntzoufras, I. (2002) On Bayesian Model and Variable Selection using Gibbs Sample. *Statistics and Computing* 12, pp. 27-36.

[3] Gibaldi M. and Perrier D. (1982) Pharmacokinetics. New York:Marcel Dekker, second edition.

[4] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J.(1996) Markov Chain Monte Carlo in Practice. London: Chapman and Hall.

[5] Wakefield J. (1996) The Bayesian Analysis of Population Pharmacokinetic Models. *Journal of the American Statistical Association* 91(433), pp. 62-75.

[6] Brown, P. and Ziegelmann, P.K. (2000) Bayesian Approach in Pharmacokinetic Models. *Bayesian Methods with Applications to Science, Policy and Oficial Statistics - Selected Papers from ISBA 2000*, pp 583-592.

MALÁRIA × CHUVA NO PARÁ: UMA ANÁLISE ESPAÇO-TEMPORAL

Aline Nobre

anobre@dme.ufrj.br

1. Introdução

O estudo de padrões geográficos em razões de incidência ou mortalidade de doença é bastante utilizado em epidemiologia e é útil na formulação de hipóteses etiológicas. Nos últimos anos, o aumento no interesse público e científico nesse tipo de estudo ocorreu junto com um crescimento da disponibilidade de dados de saúde espacialmente referenciados. Quando a doença de interesse é rara e/ou as áreas geográficas são pequenas, a contagem tende a ser baixa e a variação amostral associada alta. O mapeamento das razões de incidência ou mortalidade destas doenças não é confiável, devido à heterogeneidade do tamanho da população. Os métodos clássicos utilizam taxas que são dadas pela razão entre o número de casos observados e o número de casos esperados. Portanto, o desenvolvimento de métodos estatísticos mais sensíveis para a análise e interpretação de dados de saúde espacialmente referenciados tem sido requerido. Sob o enfoque Bayesiano, os riscos relativos são estimados através de uma média, representada por meio de covariáveis e de efeitos aleatórios que podem ou não acomodar uma estrutura espacial. Na modelagem Bayesiana, essa informação sobre a dependência geográfica local é especificada através da priori.

Este trabalho tem como objetivos observar as similaridades entre as ocorrências de malária em alguns municípios do Pará, estudar a relação entre a incidência de malária e a quantidade de chuva, assim como verificar a existência de estrutura espacial e temporal. Os modelos de mapeamento de doenças utilizados na literatura incluem efeitos aleatórios não estruturados, bem como efeitos espacialmente estruturados. Além disso, incorporamos a presença de uma componente temporal. Este artigo está organizado da seguinte forma: na seção 2 descrevemos os dados. A seção seguinte mostra os modelos de mapeamento de doença comumente utilizados na literatura. Os resultados e conclusões são mostrados na seção 4.

2. Descrição dos dados

As informações deste trabalho referem-se ao número de casos de malária em alguns municípios do estado do Pará através de dados mensais observados durante os anos de 1996 à 1998. O estado do Pará contém aproximadamente 140 municípios, alguns de grande extensão territorial. Devido a falta de coleta dos dados para alguns municípios, a região de estudo foi definida de acordo com a informação disponível, de forma que alguns municípios próximos, mesmo sem informação, também foram incluídos na análise. Por essa razão, a região de estudo foi reduzida para 69 municípios, dentre os quais 34 não apresentam informação sobre casos de malária. Um dos objetivos do trabalho é fazer inferência sobre esses municípios que não possuem informação sobre o número de casos de malária. Sob o enfoque Bayesiano, esse objetivo é atingido naturalmente, com os modelos propostos na seção 3.

Existem tanto fatores naturais quanto sociais que afetam a dinâmica da transmissão da infecção. Dentre os fatores naturais, a única informação disponível é a quantidade de chuva medida em *mm*. Os dados para chuva foram coletados em 78 estações pluviométricas monitoradoras espalhadas pelos diversos municípios no período de interesse, 1996 à 1998.

3. Modelos para mapeamento de doenças

Suponha que a região de interesse seja dividida em 69 municípios contíguos. Além disso, sejam e_i a contagem esperada da doença na área i baseada nos fatores de risco conhecidos (sexo, idade, etc) e, r_i, o risco relativo no município i, $i = 1, \ldots, n$. Na literatura de mapeamento de doenças é comum assumir que, condicional a e_i e r_i, o número de casos de malária no município i, y_i, é modelado como variáveis aleatórias independentes com distribuição de Poisson, isto é,

$$y_i | r_i, e_i \sim Poisson(e_i r_i), \quad i = 1, \ldots, 69. \qquad (1)$$

A contagem esperada é uma quantidade conhecida baseada nos fatores de riscos e é utilizada como padronização dos dados. Devido a ausência de variáveis de confundimento, em nossas análises o valor esperado, e_i foi calculado com base no tamanho da população, ou seja, $e_i = P_i P^*$, onde $P^* = \sum_i y_i / \sum_i P_i$ e P_i representa a população da área i num período determinado.

Nesses modelos, a inferência sobre os riscos relativos é desenvolvida sob o enfoque Bayesiano. Devido a natureza intratável das distribuições a posteriori marginais, a inferência é possibilitada através do uso de algoritmos Monte Carlo via Cadeia de Markov (MCMC). O amostrador de Gibbs Gilks *et al.* (1996) é utilizado para obtenção de amostra das distribuições a posteriori dos parâmetros.

O modelo mais simples que considera a variabilidade dos riscos relativos assume a presença de um efeito aleatório não estruturado, ou seja, a localização geográfica não é considerada. Neste caso, r_i é modelado por

$$\log(r_i) = \alpha_0 + \alpha_1 chuva_i + u_i, \qquad (2)$$

onde α_0 é o intercepto representando uma média geral de $\log(r_i)$ comum a todas os municípios, α_1 mede o efeito da chuva no município i no $\log(r_i)$ representando o incremento na média, e por fim $u_i \sim Normal(0, \tau_u^2)$. Os u_i's são efeitos aleatórios que representam a heterogeneidade não estruturada presente nos dados e τ_u^2 é o hiperparâmetro utilizado para controlar a variabilidade desses efeitos.

Outra forma de considerar a variabilidade dos riscos relativos é através dos modelos de campos aleatórios de Markov (MRF) Gaussianos. Neste caso, a distribuição condicional do risco relativo da área i dado todos os outros riscos relativos, depende apenas das suas áreas vizinhas (ver Besag, York e Mollié (1991), Cressie (1993) e Besag e Kooperberg (1995)). Neste caso, o logaritmo dos riscos relativos é modelado da seguinte forma,

$$\log(r_i) = \alpha_0 + \alpha_1 chuva_i + b_i, \qquad (3)$$

onde os b_i's são os efeitos aleatórios que representam a componente espacialmente estruturada, diferentemente dos u_i's da equação (**??**), que representam efeitos aleatórios não-estruturados.

A distribuição condicional de b_i é dada por

$$b_i|b_j, j \neq i \sim N\left(\frac{\sum_{j\in\delta_i} w_{ij}b_j}{\sum_{j\in\delta_i} w_{ij}}, \frac{\tau_b^2}{\sum_{j\in\delta_i} w_{ij}}\right),$$

onde δ_i representa o conjunto de áreas vizinhas da i-ésima área, w_{ij} representa os pesos dos vizinhos com $w_{ii} = 0$ e $w_{ji} = w_{ij}$ tal que a matriz de pesos é simétrica.

A maioria dos estudos envolvendo modelos para mapeamentos de doenças, não tratam da componente espacial e temporal conjuntamente. Em algumas situações é importante modelar essa interação espaço-tempo conjuntamente. O modelo (**??**) é reescrito como

$$y_{it}|r_{it}, e_{it} \sim Poisson(e_{it}r_{it}) \quad i = 1, \ldots, 69 \quad t = 1, \ldots, 36. \tag{4}$$

ou seja, agora y_{it} representa o número de casos de malária no município i no mês t, e_{it} e r_{it} representam, respectivamente, a contagem esperada e o risco relativo do município i no mês t.

Assunção, Reis e Oliveria (2001) adotam um modelo de tendência polinomial de segunda ordem no tempo e neste trabalho investigaremos essa modelagem. Neste caso, o logaritmo do risco relativo é modelado através de uma tendência polinomial de segunda ordem, ou seja,

$$\log(r_{it}) = \alpha_i + \beta_i(t-1) + \gamma_i(t-1)^2$$

onde α_i representa o logaritmo do risco no município i no primeiro mês e $\beta_i + \gamma_i(2t-1)$ o incremento sobre o logaritmo do município i quando mudamos do mês t para $t+1$. Nesta aproximação os parâmteros α_i's, β_i's e os γ_i's têm estrutura espacial CAR com hiperparâmetros τ_α^2, τ_β^2 e τ_γ^2. Logo os parâmetros do polinômio são diferentes para cada município. Isto significa que o modelo considera a interação entre as componentes espaciais e temporais, já que os efeitos temporais é diferente para cada área. Vale destacar também que covariáveis não estão presentes.

Dois modelos adicionais são verificados para descrever a evolução no tempo, o modelo linear descrito por um polinômio de primeira ordem, $\log(r_{it}) = \alpha_i^* + \beta_i^*(t-1)$, e um modelo constante no tempo, $\log(r_{it}) = \alpha_i^{**}$.

4. Resultados e Conclusões

Neste trabalho utilizamos como variável resposta o número de casos de malária em 69 municípios do Pará. Portanto os dados são agregados por área. A quantidade de chuva, utilizada como covariável, foi medida em 78 estações monitoradoras espalhadas pelos municípios, e consequentemente temos dados pontuais. Esse problema é conhecido na literatura como troca de suporte, onde temos variáveis medidas em unidades espaciais diferentes. Com o objetivo de agregar essas informações, um modelo de interpolação espacial Bayesiana foi utilizado para prever a chuva na sede dos municípios. Em seguida, essa informação foi utilizada como covariável para modelar o número de casos de malária no município.

Inicialmente, foi ajustado para cada um dos 36 meses um modelo não estruturado e um CAR. Os resultados obtidos através dos modelos de mapeamento de doenças descritos na seção 3 revelam que existe uma estrutura espacial da doença, ou seja, municípios próximos possuem riscos de malária similares. Os resultados da análise visual feita através dos mapas, quando comparamos os modelos com e sem estrutura, não são confirmados pelo método de comparação de modelos utilizado, o Critério de Informação da Deviance (DIC). Um dos motivos para esse resultado, é o fato de que a região de estudo é composta por municípios de grande extensão territorial, e talvez a agregação da malária a nível municipal não seja capaz de detectar possíveis estruturas espaciais.

As figuras 3 e 4 apresentam o boxplot para o coeficiente da chuva ao longo dos meses no período de 1996 à 1998. Os resultados indicam uma mudança no comportamento destes coeficientes ao longo dos meses. Entretanto, a média a posteriori da maioria deles está próxima de zero. O sinal negativo dessas estimativas indica que quanto maior a quantidade de chuva, menor o risco da doença. Esse resultado provavelmente deve-se ao fato de que a chuva que cai em um determinado mês só afeta o desenvolvimento do mosquito nos meses seguintes. A proliferação do mosquito se dá no início do período chuvoso e alguns meses após seu final. No início, dada a abundância de água limpa, as condições de procriação tornam-se mais adequadas.

Em seguida, a chuva continuada ajuda a limpar os focos dos mosquitos. Entretanto, quando vem a estiagem, a situação volta a ser apropriada para o desenvolvimento da larva e, portanto, a proliferação se acentua juntamente com o número de casos infectados.

A figura 5 apresenta a mediana a posteriori dos riscos relativos estimados pelo modelo de tendência polinomial de segunda ordem para cada município analisado para o ano de 1996. Nota-se que esse modelo identifica a região sudoeste do Pará, assim como o modelo CAR, juntamente com o centro do estado como sendo uma região de ocorrência média-alta de casos de malária.

A chuva foi utilizada no modelo com o objetivo de explicar a incidência da malária. Como essas variáveis encontram-se em unidades espaciais diferentes, essas informações foram agregadas de forma que a informação de chuva utilizada, na modelagem da malária, foi a informação prevista na sede dos municípios. Portanto estamos utilizando essa informação para todo o município, já que o número de casos de malária é proveniente do município como um todo. Além disso, não estamos levando em conta a variabilidade dessa informação, pois utilizamos a média prevista pelo modelo de krigagem Bayesiana. Portanto uma possível extensão, é a modelagem conjunta da malária e da chuva, evitando assim a subestimação das incertezas associadas a ambas medições.

Referências

[1] Assunção, R. M., Reis, I. A. e Oliveira, C. D. L. (2001) Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model, *Statistics in Medicine*, 20, 2319–2335.

[2] Besag, J. e Kooperberg, C. (1995) On Conditional and Intrinsic Autoregression, *Biometrika*, 82, 733–746.

[3] Besag, J., York, J. e Mollié, A.(1991) Bayesian image restoration, with two applications spatial statistics, *Annals of the Institute of Statistical Mathematics*, 43,1–59.

[4] Cressie, N. A. C. (1993) *Statistical for Spatial Data*, Wiley, New York.

[5] Gilks, W.R., Richardson, S. e Spiegelhalter, D. J. (1996) - *Markov Chain Monte Carlo in Practice*, Chapman & Hall.

MODELAGEM DINÂMICA BAYESIANA DE FUNÇÕES DE TRANSFERÊNCIA PARA DADOS EPIDEMIOLÓGICOS

Mariane Alves, Dani Gamerman e Marco Ferreira

mariane@dme.ufrj.br, dani@im.ufrj.br e marco@im.ufrj.br

1. Introdução

Diversos estudos, em todo o mundo, têm apontado elevações nos níveis de poluição atmosférica como fatores potencializadores de doenças que podem levar a internações hospitalares e, até mesmo, a óbitos. Em particular, nas grandes metrópoles brasileiras, este efeito já se faz sentir. Uma das questões de relevância, nesse contexto, é a inferência sobre a forma de propagação, ao longo do tempo, dos efeitos danosos de elevações nos níveis de poluentes.

Na aplicação desenvolvida neste trabalho, objetiva-se, primordialmente, estudar o impacto (instantâneo e defasado) de poluentes sobre as contagens de óbitos infantis devido a doenças respiratórias em São Paulo. O banco de dados utilizado foi disponibilizado pelo projeto PASSO-USP e contém registros diários das variáveis de interesse, compreendendo o período de janeiro/1994 a dezembro/1997.

2. Modelo Proposto

O modelo proposto para análise das séries temporais em questão pertence à classe dos Modelos Lineares Generalizados Dinâmicos (West, Harrison e Migon, 1985). A contagem de óbitos é tratada como um processo de Poisson, cuja intensidade é modelada a partir de dois blocos estruturais: um de tendência e outro de efeito da covariável poluente. A persistência do efeito defasado dos poluentes é inserido nos modelos por meio da adoção de funções de tranferência (Box, 1994) com formas

paramétricas específicas. Uma vantagem dessa abordagem é a possibilidade de inferir sobre a duração do efeito dos poluentes, sem que seja necessário especificar um número máximo de defasagens para estes. A escolha de parametrizações adequadas para as funções de transferência torna a modelagem bastante parcimoniosa.

Inicialmente, adotou-se uma função de transferência que preconiza o decaimento exponencial do efeito dos poluentes sobre as contagens de óbitos. A abordagem inferencial é Bayesiana, sendo utilizados métodos de Monte Carlo via Cadeias de Markov para obtenção de informações a posteriori. Denote-se por Y_t a contagem de óbitos e por X_t o nível do poluente no dia t. O modelo inicialmente proposto tem equação observacional dada por $Y_t \sim Poisson(\lambda_t)$, com a intensidade do processo modelada por $\log(\lambda_t) = \mu_t + \xi_t$. O bloco de tendência, μ_t, segue a equação de evolução $\mu_t = \mu_{t-1} + \omega_t$, enquanto o bloco de efeito do poluente é modelado segundo $\xi_t = \rho\xi_{t-1} + \beta X_t$. As densidaes a priori para os parâmetros envolvidos no modelo são: $\omega_t \sim N(0, W)$, $W \sim IG(\frac{n}{2}, \frac{ns}{2})$, $\rho \sim U[0,1]$, $\beta \sim N(m_\beta, C_\beta)$. Observe-se que o parâmetro β mede o impacto instantâneo do poluente sobre os óbitos, enquanto o parâmetro ρ mede a memória do processo quanto ao efeito passado desta covariável.

3. Aplicação

A princípio, foram feitos alguns exercícios com dados simulados segundo o modelo acima. Ainda não obtivemos sucesso na estimação do modelo com nível variando no tempo. Atualmente, esforços estão sendo feitos no sentido de solucionar tais questões. Para a aplicação com dados reais, então, o modelo foi simplificado, tornando-se o nível μ constante ao longo do tempo. Os resultados obtidos encontram-se na tabela abaixo. Como se pode perceber, o poluente que apresenta maior impacto sobre os óbitos infantis em São Paulo, no período da análise, é CO, seguido por PM_{10} e, finalmente, SO_2. Ao analisar estes mesmos dados,por meio de modelos aditivos generalizados, Conceição *et al.* (2001) constataram a existência de associação significante entre mortalidade infantil por doenças respiratórias e níveis diários de CO, SO_2 e PM_{10}. Ao incluir os poluentes simultaneamente no modelo, observaram relação significante apenas para o poluente CO e significância marginal do poluente SO_2. Ferreira e Gamerman (2000) avaliaram a associação entre poluentes e mortalidade infantil durante o ano de 1991, utilizando, como variáveis explicativas, os níveis de NO_2 e CO, relacionados às contagens de óbitos por meio de Modelos Dinâmicos Lineares Generalizados. O modelo ali adotado não considera os efeitos de médio e longo prazo das covariáveis, medindo, assim, o impacto instantâneo dos poluentes sobre o número de óbitos (tabela 1).

	Estatísticas Sumárias								
	CO			PM_{10}			$S0_2$		
	IC(5%)	média	IC(95%)	IC(5%)	média	IC(95%)	IC(5%)	média	IC(95%)
β	0.026	0.036	0.047	0.019	0.030	0.040	0.010	0.023	0.038
ρ	0.892	0.920	0.943	0.851	0.896	0.931	0.398	0.693	0.859
μ	0.797	0.827	0.856	0.818	0.846	0.875	0.835	0.862	0.889

Tabela 1: Estatísticas sumárias da amostra da densidade a posteriori.

A figura 6 ilustra a função de transferência entre CO e as contagens de óbitos. O nível médio de CO observado no período de análise foi 4,43 ppm. Na figura, ilustramos o efeito (em termos de aumento percentual na contagem de óbitos) de elevações dos níveis de CO em relação a este nível médio.

4. Conclusões e Extensões

As aplicações feitas revelaram que o efeito de poluentes sobre os óbitos infantis em São Paulo é significativo e persistente ao longo do tempo. Atualmente, estamos aprimorando estas aplicações, inserindo na modelagem variáveis metereológicas tais como temperatura e umidade. Temos trabalhado, ainda, na inserção simultânea dos diversos poluentes como variáveis regressoras.

Referências

[1] Box, G. E. P., Jenkins, G. M. e Reinsel, G. C. (1994) - *Time Series Analysis - Forecasting and Control*, Prentice-Hall, Inc., 3a. ed.

[2] Conceição, G. M. S., Miraglia, S. G. E. K., Kishi, H. S., Saldiva, P. H. N. e Singer, J. M. (2001) - Air pollution and children mortality - A time series study in São Paulo, Brazil, *Environmental Health Perspectives*, **109**, suplemento 3.

[3] Ferreira, M. A. e Gamerman, D. (2000) - Dynamic Generalized Linear Models, em DEY, D. K., GHOSH, S. K. e MALLICK, B. K., *Generalized Linear Models*, Marcel Dekker Inc., pp 57-72.

[4] West, M., Harrison, J. e Migon, H. S. (1985) - Dynamic generalized linear models and Bayesian forecasting, *J. Am. Statist. Assoc.*, **80**, pp. 73-96.

PROCESSOS ARFIMA(0,D,0) COM ERROS T-STUDENT E HIPERBÓLICO

Ralph Silva e Helio Migon

rphss@dme.ufrj.br e migon@im.ufrj.br

1. Introdução

Os modelos de longa depedência descritos pelos processos auto-regressivos fracionalmente integrados médias móveis - ARFIMA(p,d,q) têm mostrado sua relevância em diversos campos de aplicação devido sua capacidade de captar a forte estrutura de dependência existente entre os dados, mesmo para as obsevações distantes entre si. Em particular, tem-se os processos ARFIMA(0,d,0), também chamado de ruído fracionário. A hipótese de erros t-Student e hiperbólicos se justificam pela presença de mais massa de probabilidade nas caudas das duas distrinuições se comparada ao modelo com erros normais. A função de verossimilhança é obtida por uma aproximação (Li e McLeod, 1986) e mistura de escala normal-gama inversa para os erros t-Student e média-escala norma-gaussiana inversa generalizada (GIG) para os hiperbólicos.

2. Processo ARFIMA(0,d,0)

2.1 Definição

Uma série temporal $\{y_t\}$ gerada por um ARFIMA(0,d,0) é descrita por

$$(1-B)^d y_t = \epsilon_t$$

onde B é o operador diferença e ϵ_t são variáveis aleatórias independentes e identicamente distribuídas com média 0 e variância finita σ^2 e

$$(1-B)^d = \sum_{j=1}^{\infty} \binom{d}{j} (-1)^j (-B)^j$$

2.2 Processo ARFIMA(0,d,0) com erros t-Student

A função de verossimilhança é dada por

$$p(\mathbf{Y}_T \mid \Psi) = (2\pi\delta^2)^{-\frac{T}{2}} \exp\left[-\frac{1}{2\delta^2}\sum_{t=1}^{T}(y_t-\mu_t)^2\right]$$

onde

$$\mu_t = \sum_{j=1}^{m} \varphi_j y_{t-j} \;\text{ e }\; \Psi = (d, \sigma^2, \nu, \delta^2, Y_0)$$

com

$$(\delta^2 \mid \nu, \sigma^2) \sim GI(\nu/2, \sigma^2\nu/2)$$

A distribuição a priori é própria e vaga para os parâmetros do modelo.

2.3 Processo ARFIMA(0,d,0) com erros hiperbólicos

A função de verossimilhança é dada por (Barndorff-Nielsen, 1997),

$$p(\mathbf{Y}_T \mid \Psi) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t-\mu_t)^2\right]$$

onde

$$\mu_t = \sum_{j=1}^{m} \varphi_j y_{t-j} \;\text{ e }\; \Psi = (d, \sigma^2, \gamma, \delta, Y_0)$$

com

$$(\sigma^2 \mid \gamma, \delta) \sim GIG(0,5; \gamma; \delta)$$

A distribuição a priori é própria e vaga para os parâmetros do modelo.

2.4 Inferência

A inferência é feita através de uma amostra da distribuição conjunta a posteriori. O amostrador de Gibbs (Gelfand and Smith, 1990) é utilizado, e para a condicional completa do parâmetro d utiliza-se o amostrador da fatia, ou *slice sampler* (Neal, 1997).

Referências

[1] Barndorff-Nielsen, O. (1977) Exponetially decreasing distrinutions for the logarithm of particle size. *Proceedings of the Royal Society London, Series A*, 353, 401–419.

[2] Gelfand, A. and Smith, A. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

[3] Li, W.K. and McLeod, A.I (1986), Fractional time series modelling, *Biometrika*, 73, 1,217–221

[4] Neal, R (1997) Markov chain Monte methods based slicing the density function (Tech. Rep.) Toronto, Canada. *http://www.cs.utoronto.ca/radford*.

UMA ABORDADEM BAYESIANA PARA MODELOS DE FRAGILIDADE ESPACIAL

Leonardo Bastos e Dani Gamerman

bastos@dme.ufrj.br e dani@im.ufrj.br

1. Introdução

Os modelos de fragilidade são caracterizados pela introdução de um efeito eleatório (fragilidade) na função de risco do modelo de taxas de falha proporcionais (Cox, 1972), com o objetivo de controlar a heterogeneidade não observada das unidades em estudo. Quando incorporamos à fragilidade uma estrutura de correlação espacial temos os Modelos de Fragilidade Espacial.

Carlin *et al.* (2003) tratam a dependência espacial usando modelos CAR, Henderson *et al.* (2003) utilizam um modelo de fragilidade Gama Multivariado. Neste trabalho propomos a utilização de processos Gaussianos usados em Geoestatística para modelar a dependência espacial das fragilidades.

Devido à complexidade das distribuições a posteriori obtidas, foram utilizados métodos de Monte Carlo via Cadeia de Markov (MCMC).

2. Modelos de Fragilidade Espacial

A função de risco para os modelos de fragilidade espacial é dada por:

$$h(t_i; \mathbf{X}_i, \mathbf{s}_i) = h_0(t_i) e^{\mathbf{X}_i \beta + W(\mathbf{s}_i)} \quad (5)$$

onde para o indivíduo i, t_i é o tempo de falha, $\mathbf{X}_i$ é um vetor de variáveis explicativas, $\mathbf{s}_i$ é a posição espacial e $h_0(t_i)$ é a função de risco de base.

A fragilidade espacial $\mathbf{W}(s) = (W(\mathbf{s}_1), \ldots W(\mathbf{s}_n))^T$ é modelada segundo o seguinte processo gaussiano:

$$\mathbf{W}(s) \sim N(\mathbf{0}, \Sigma) \quad (6)$$

onde $\Sigma = \sigma^2 \rho(D, \Theta)$, D é a matriz de distâncias dos indivíduos e Θ é o conjunto de parâmetros da função de correlação utilizada. Neste trabalho utilizamos a função de correlação exponencial: $\rho(d_{ij}, \phi) = \exp\left[-\frac{d_{ij}}{\phi}\right]$, $d_{ij} = ||s_i, s_j||$.

3. Simulação e Principais Resultados

Foi gerado uma amostra de tamanho 100 de uma distribuição Weibull com função de risco dada por (**??**), com duas covariáveis e função de risco de base $h_0(t) = \alpha\lambda t^{\alpha-1}$. As distribuições a priori utilizadas são próprias e relativamente vagas. Foram geradas 15000 valores, utilizando um *Burn-in* de 7500 e *lag* igual a 40. Os resultados obtidos foram os seguintes:

Pode-se perceber que o método utilizado gerou bons resultados para os dados simulados, gerando estimativas bem próximas dos valores verdadeiros. O próximo passo é aplicar a dados reais: dados de câncer.

Em trabalhos futuros vamos modelar a função de risco usando processos gama independentes, usar outras famílias de funções de correlação espacial, incluir uma estrutura espacial nos coeficientes da regressão e introduzir no modelo de regressão uma estrutura de dependência temporal para testar não-proporcionalidade das taxas de falhas.

Referências

[1] Carlin, B.P. e Banerjee S. (2003). Hierarchical Multivariate CAR Models for Spatio-Temporally Correlated Survival Data, *Bayesian Statistics 7*, Oxford, a ser publicado.

[2] Cox, D.R. (1972). Regression models and life tables. *Journal of Royal Statistical Society, Serie B*, 34, p. 187-220.

[3] Henderson, R., Shimakura S. e Grost, D. (2003). Modelling Spatial Variation in Leukaemia Survival Data, *JASA*, a ser publicado.

SAMPLE SIZE DETERMINATION FOR DICOTOMIC FINITE POPULATIONS: BAYESIAN PERSPECTIVE

Cléber Figueiredo, Daniela Ramires, Marcos Oliveira e Carlos Pereira

figuecl@usp.br

1. Introduction

The objective of this work is the determination of the sample size, n, for the estimation of the number, θ, of items with a specified characteristic C in a finite population with known size N. The perspective is Bayesian and the model is presented in Basu & Pereira (1982).

Notation: $Be(a,b)$, $Bebi(n;a,b)$, represent, respectively, the Beta and Beta-binomial distributions, with parameter (a,b) and sample size n. Moreover, we consider $n_0 = a+b$, $A = a+x$, $B = b+(n-x)$, $\tilde{n} = n+n_0$, $\tilde{N} = N+n_0$, $m = \frac{A}{\tilde{n}}$, $E(x)$ = mean of x and $V(x)$ = variance of x.

2. Model

Items are produced according to a Bernoulli process, where π is the failure rate. The items are stored in lots of size N. For a specific lot, we want to estimate the number of defective items, θ. A sample of size n ($< N$) is selected from this lot. Consider x and $\theta - x$ the number of defective items in the sample, with size n, and in the unsample, with size $N-n$, respectively. According to Basu & Pereira (1982), we have:

If $\pi \sim Be(a,b)$, then: (i) $\pi|x \sim Be(A,B)$; (ii) $\theta \sim Bebi(N;a,b)$; (iii) $x \sim Bebi(n;a,b)$; (iv) $(\theta - x) \sim Bebi(N-n;a,b)$; (v) x and θ are condicionally independent given π and (vi) $(\theta - x)|x \sim Bebi(N-n;A,B)$.

Also from the Beta-binomial properties it follows that $E(\theta - x|x) = (N-n)m$ and

$$\begin{aligned} V(\theta - x|x) &= \frac{(N-n)\tilde{N}m(1-m)}{(\tilde{n}+1)} \\ &\leq \frac{(N-n)\tilde{N}}{4(\tilde{n}+1)}. \end{aligned}$$

Note that the superior variance boundary is reached when $A = B$, or equivalently, when $m = 1-m = \frac{1}{2}$. In this case the distribution of $(\theta - x)|x$ is symmetric around m.

3. Sample size determination

In the inferential process, described above, the values a, b e N are specified and x shall be observed after the determination of n. The parameter of interest is θ, or equivalently $\rho = \frac{\theta}{N}$. Note also that if $r_0 = \frac{n_0}{N}$ e $r = \frac{n}{N}$ are sample ratios, a priori and a posteriori, then $E = E(\rho|x) = (1-r)m + \frac{x}{N}$ and

$$\begin{aligned} V = V(\rho|x) &= \frac{(1-r)(1+r_0)m(1-m)}{(\tilde{n}+1)} \\ &\leq \frac{(1-r)(1+r_0)}{4(\tilde{n}+1)}. \end{aligned}$$

In order to determine the sample size, let us concentrate in this proportion ρ, of defective items. For the final inference, the objective is to obtain the smallest interval $[I_1, I_2]$ in such a way that ρ belongs to this interval with probability at least $1-\alpha$, i.e., $Pr\{I_1 < \rho < I_2|x\} \geq 1-\alpha$.

For simplicity, let us fix $\alpha = 0.0455$. The least favorable case will occur when the largest posterior variance is obtained. This is the case where $m = 1-m = \frac{1}{2}$. In this situation, we have a symmetric posterior distribution around of $(1-r)m + \frac{x}{N}$. Using the standard approach, based on the approximation to the standard normal distribution, the interval length will be near $4D = 0.1$, where $D = V^{\frac{1}{2}}$. Considering the case where the prior distribution is uniform, i.e., $a = b = 1$, and for a finite population of $N = 5,000$ units, we would have $n = 365$.

Table 1 presents the values of n for several situations where we considered various values of D (or $I_2 - I_1$) and of N, keeping $\alpha = 0.0455$. The S-Plus software (see Krause & Olson, 1997) was used for these calculations.

4. Final Comments

A Bayesian alternative for the sample size determination was presented. Note that the 95.45% credibility was used only for illustrative purposes. However, in addition to simplicity, we have the fact that, if we keep the fixed credibility, for non-symmetric cases ($A < B$ or $A > B$) the precision increases ($I_2 - I_1$ decreases) whenever m approximates zero or one. On the other hand, if we keep the interval length fixed, the credibility will increase according to the skewness, and this does not hold under the standard sampling theory. For instance, let $N = 1,000$, $n = 284$ and $x = 142$, a symmetric case. The interval for ρ with 95.55% of credibility is $[0.4997; 0.5022]$. On the other hand, if $x = 80$, a non-symmetric case, the interval would be $[0.2816; 0.2836]$, a shorter one.

Referências

[1] Basu, D & Pereira, CAB (1982) *On the Bayesian analysis of categorical data: The problem of non-response*. Journal of Planning and Inference 6: 345 – 362.

[2] Krause, A & Olson, M (1997) *The Basics of S and S-Plus*. Springer.

Population	$D = 0.01$	$D = 0.025$	$D = 0.05$	$D = 0.07$	$D = 0.1$
5,000	1,655	**365**	93	48	22
2,000	1,108	329	90	47	22
1,000	712	**284**	86	46	22
500	416	221	80	44	21
300	268	171	73	41	20
200	185	133	65	39	20
100	96	83	49	33	18
50	49	44	33	24	15

Table 1: Sampling sizes obtained for fixing the standard deviation and population sizes based on a not skew posterior distribution *Beta – Binomial*.

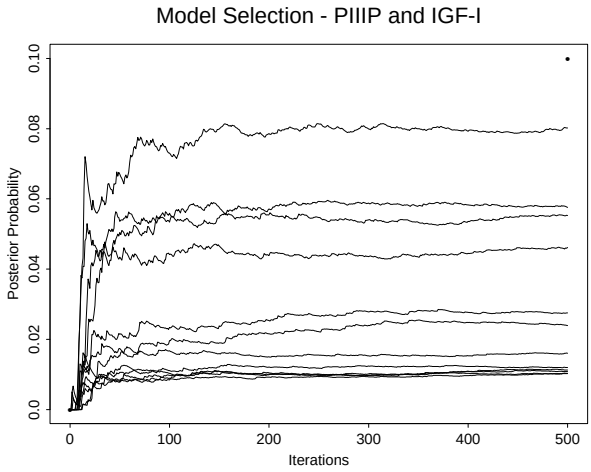

Figura 1: Posterior Probability for the most frequently selected models.

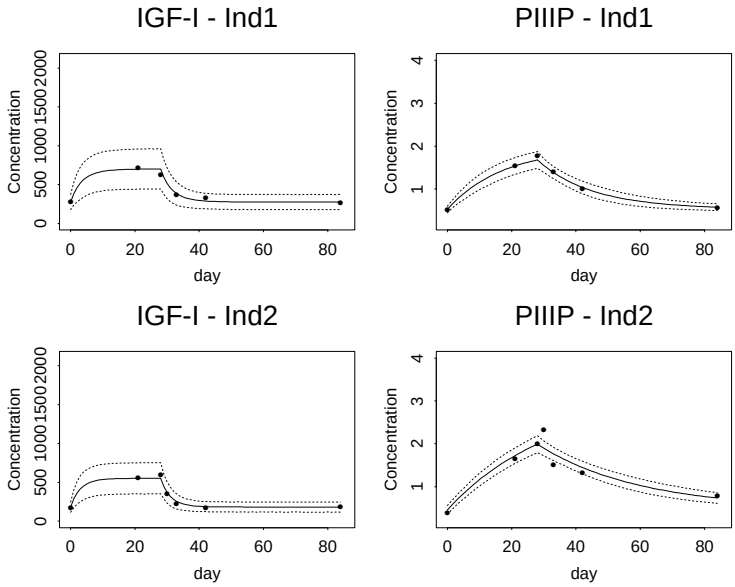

Figura 2: IGF-I and PIIIP estimated profiles. Observed concentrations are represented by the dots.