BOLETIM *ISBrA*

Volume 1, Número 3 Abril 2007

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

Carta da Presidente

Márcia D'Elia Branco
mbranco@ime.usp.br

Caros Colegas

Após um longo período estamos retomando o boletim da seção brasileira da ISBA. Agradecemos ao Adriano Polpo por ter aceitado o nosso convite para ser o novo editor do boletim. O Adriano também tem trabalhado na remodelagem da nossa página web. Convido a todos a visitá-la no novo endereço *www.ime.usp.br/∼isbra*. Sugestões para o aperfeiçoamento da mesma são bem-vindas.

Neste número são publicados três artigos baseados nos trabalhos de dissertação de mestrado finalistas do último concurso promovido pela ABE. Entre os cinco trabalhos selecionados para apresentação no 17 SINAPE, três foram bayesianos, sendo que destes resultou o primeiro e segundo lugares do concurso. Parabéns ao Erlandson, Marcus e Cristian e seus respectivos orientadores.

Infelizmente em março tivemos uma noticia muito triste para a comunidade bayesiana, o falecimento da professora Pilar Iglesias da PUC da Santiago – Chile. Pilar tinha um vinculo muito estreito com o Brasil. Além de ter estudado e trabalhado aqui, foi a grande incentivadora para a criação deste capítulo. Podemos dizer que ela é a nossa madrinha. Nos últimos anos ela travou uma luta pela vida, onde demonstrou mais uma vez sua garra, dedicação e amor à vida. No dia 15 de março o programa de pós-graduação da estatística da USP realizou uma pequena homenagem a Pilar. Publicamos aqui um resumo desse evento através dos relatos dos docentes que participaram da homenagem. Também publicamos nesta seção algumas das manifestações feitas a lista da ABE sobre o assunto. Essa foi uma maneira que encontramos de registrar o nosso agradecimento a esse ser humano tão especial.

Finalmente algumas novidades sobre o próximo Encontro Brasileiro de Estatística Bayesiana (9EBEB) e informações sobre encontros da área pelo mundo. Também trazemos os relatos do Adriano, sobre o último CLAPEM, realizado em Lima – Peru, e do Josemar sobre a 10^a Escola de Modelos de Regressão, realizada em Salvador, ambos em fevereiro deste ano.

Boa leitura!

Índice

Pilar Loreto Iglesias Zuazola

(04/09/1960 – 03/03/2007)

Por Márcia D'Elia Branco:

Vou apresentar aqui um pequeno resumo da trajetória deste ser humano tão especial. Pilar nasceu em 4 de setembro de 1960 e faleceu em 3 de março de 2007, portanto, com apenas 46 anos. Em Valparaiso, sua cidade natal, além da praia, ela tinha como um dos seus lugares favoritos o Bar Cinzano, local tradicionalmente freqüentado pela boemia da cidade portuária e onde ela fazia questão de levar

SUGESTÕES

Qualquer tipo de sugestão, reclamação, doação, que possa ser utilizada para melhorar a qualidade do boletim é muito bem-vinda.

EXPEDIENTE:
Editor: *Adriano Polpo*
End: Departamento de Estatística – UFSCar / Via Washington Luís, km 235
CEP: 13.565-905 / São Carlos – SP Caixa Postal: 676
e-mail: *polpo@power.ufscar.br*

seus amigos. O amor que ela tinha pela vida se manifestava na paixão pelos bares, pela noite, pela música, dança e longas conversas. Em 1988 Pilar deixa o Chile e inicia seu doutorado no departamento de estatística da USP. Foi nesse período que eu a conheci, na época eu era aluna de mestrado do programa. A amizade foi imediata. A partir daquele ano nosso vinculo se deu em diversas esferas, como colega, como orientadora e posteriormente como colaboradora em trabalhos científicos. Sob a orientação do professor Carlinhos, ela terminou seu doutorado em 1993 com a tese: Formas Finitas do Teorema de De Finetti. Em 1995 retorna a Santiago do Chile, depois de ter feito parte do corpo docente do IME por dois anos. No entanto, o contato com São Paulo é mantido. Ela co-orienta quatro teses de doutorado na USP. Além da minha, a dos colegas Antonio José da Silva, Loretta Gasco (com Heleno Bolfarine), Rosângela Loschi (com Reinaldo Arellano Valle). Ela tem artigos publicados em co-autoria com os seguintes docentes do IME: Carlos Pereira, Heleno Bolfarine, Sérgio Wechsler, Nelson Tanaka, Mônica Sandoval, Luis Gustavo Esteves e Márcia Branco. Em Santiago ela assume diversos cargos importantes, como a presidência da sociedade chilena de estatística e a chefia do departamento. Foi criadora do capítulo chileno da ISBA, em 1997, e madrinha do capítulo brasileiro. Ela estava presente na reunião do SINAPE de 2000, onde se decidiu pela criação do ISBrA. Um dos grandes desafios assumido por ela nos últimos anos, foi a organização do ISBA metting, realizado em 2004 em Viña del Mar. O resultado foi fantástico, sorte daqueles que participaram. Suas diversas visitas ao Brasil, as minhas ao Chile e os encontros em congressos pelo mundo, mantiveram viva a nossa amizade. Eu tenho muito que agradecer a ela, especialmente por ter sempre confiado no meu trabalho, mesmo quando nem eu mesmo acreditava.

Assim era Pilar, meio chilena, meio brasileira, mas sempre latina americana criando um modo alegre de fazer pesquisa, com muito suor e garra, e pouca competição. O seu velório no Chile, foi embalado por canções brasileiras.

Por Carlos Alberto de Bragança Pereira:

No primeiro curso de Inferência Bayesiana da América Latina no Chile, no inicio dos anos 80, encontrei uma menina da graduação que me procurou perguntando se era possível fazer o doutorado naquele assunto. Respondi que sim, era possível, e que nós estávamos, no Brasil, com um programa de bolsas para alunos estrangeiros.

Creio que depois de uns quatro anos, estava na minha frente, aqui na USP, aquela menina alegre pedindo orientação para o doutorado. É claro que aceitei orienta-la, por minha sorte. Ela não deu algum trabalho, pois tudo que foi feito em sua dissertação foi mérito dela. Só tive a incumbência de apresentar os desafios, que foram cumpridos passo a passo, com muita eficiência. Certamente seu sucesso profissional e acadêmico foi independente do orientador. Qualquer que fosse o privilegiado, Pilar iria ter o sucesso e o reconhecimento que teve em sua carreira.

Pilar marcou a nossa comunidade com o espírito latino americano de fazer ciência: mesmo com poucos recursos e muitas barreiras, levou a vida acadêmica com alegria e perseguindo a excelência. Sou grato a ela por permitir que fosse seu orientador.

Poucos dias atrás percebi que eu fui o SEU orientador e não ela a MINHA aluna. Quando fui apresentado a um emérito cientista, disseram "este foi o orientador da Pilar". Senti-me orgulhoso de eu estar ali, naquela condição. Seria dessa forma qualquer que tivesse sido seu supervisor.

Pilar será lembrada por sua competência, por sua alegria, por sua coragem, por sua dedicação e principalmente por sua latinidade; seu sofrimento não evitava a alegria que tinha com a vida e com seu trabalho. Na América Latina, com todo o sofrimento e o subdesenvolvimento, encontramos pessoas do hemisfério norte buscando o prazer e a alegria da dança, dos ritmos e do futebol. Pilar é a representante mais digna da academia latino americana.

Pilar será eterna em nossas lembranças!

Por Heleno Bolfarine:

A presença da Pilar como amiga e colega certamente foi marcante. Qualquer conversa, por mais simples que fosse sempre levava a algo produtivo. Pode-se dizer que ela respirava estatística (bayesiana, claro). Fica em nós um grande vazio. Mas posso me dar por feliz por ter uma vez ministrado em conjunto com ela um curso de teoria das decisões. Foi o melhor curso que já fiz e acredito que para os alunos também.

Por Sérgio Wechsler:

A Pilar amava a vida como poucas pessoas o fizeram. E tornava a vida de quem quer que estivesse por perto muitíssimo mais divertida. Não por coincidência, todas as muitas mensagens na lista da ABE e no livro de condolências no site da PUC-Santiago fazem referência a seu sorriso e sua alegria de viver. E a sua generosidade.

Quanto a seu posicionamento científico, quero lembrar que Pilar defendia ardorosamente o ponto de vista deFinettiano que atualmente é chamado de preditivista: quando regressei do doutoramento dizendo a todos que "Parâmetros tampouco existem", Pilar foi essencialmente a única pessoa na USP a dar ouvidos a essa quase redundância do famoso lema de deFinetti. Ela pediu alguns minutos (para descer e comprar cigarros) e, na volta, já havia decidido alterar radicalmente a índole e a propositura da tese de doutoramento que escrevia na época. Eu acho que nasceu ali a área de pesquisa em teoremas de tipo deFinetti que ela tão brilhantemente liderou.

Nós aqui na ISBrA – que aliás nasceu na cabeça da Pilar – sentiremos falta do seu raciocínio afiado, das suas dicas, de sua paciência para ensinar e discutir, da sua capacidade de colocar todo mundo para trabalhar, mas , acima de tudo, da sua amizade e companhia. Estar perto da Pilar era sempre divertido, a melhor parte do dia ou dos congressos.

Por Rosângela Loschi:

Não me lembro quando conheci Pilar, mas dois eventos certamente contribuíram para que nos conhecêssemos melhor: o seminário organizado pelo Sérgio Wechsler, nas sextas-feiras de 93, em que discutíamos as idéias de De Finetti sobre probabilidade, e o time de futebol organizado por ela e treinado pelo Luis Renato Fontes. Acho que isto resume um pouco como tem sido o meu caminho ao lado da Pilar: marcado por muito trabalho e, sempre, muita diversão.

Homenagear uma pessoa sempre acaba sendo difícil. Estamos sempre muito acostumados a avaliar a sua produtividade e a falar do seu talento para a ciência, mas, em geral, temos uma dificuldade enorme de olhar dentro de seus olhos e ver o que ela guarda de mais precioso dentro de seu coração e que, para mim, é o que faz dela uma pessoa realmente especial.

Fazer parte de uma homenagem à Pilar me deixa feliz pois me possibilita tornar público o meu agradecimento a ela por tudo o que fez por mim, não só profissionalmente, mas por toda a riqueza que ela trouxe para a minha vida pessoal. Pilar era uma pessoa muito especial com quem aprendi muito não apenas sobre Estatística, como também sobre a vida. Fizemos muitos projetos científicos juntas e nos juntamos a várias pessoas para realizá-los (o Reinaldo Arellano, o Frederico Cruz, o Sergio, alguns de meus alunos, o Fabrizio Ruggeri e tantos outros). Curiosamente, todos os projetos que elaboramos nos bares foram realizados na sua integra.

Mas o que Pilar tinha de tão especial? Varias coisas. Quiséramos nos ter tanto. Tinha um amor incondicional pela vida, o que lhe deu força para lutar por ela até o instante em que seu corpo não resistiu mais. Tinha uma alegria enorme e colocava amor em tudo o que fazia contagiando a todos que estavam ao seu redor. Nunca vi uma pessoa com tanta energia para o trabalho. Seu talento para juntar pessoas completamente diferentes em torno de um bem comum é incontestável. Pude presenciar a mudança nos lugares e nas vidas das pessoas que conviveram com ela. Sempre dizia que a melhor maneira para crescermos enquanto grupo é deixarmos de lado nossas diferenças pessoais e contribuirmos com o que temos de melhor ("Todos temos deficiências, por isto devemos somar as nossas qualidades", dizia ela). Acreditava no potencial das pessoas (muitas vezes mais do que elas próprias) e as apoiava como podia. Fazia sempre com que almejássemos horizontes mais amplos do que aqueles que tínhamos em mente. Curiosamente, quando nos dávamos por nós, já estávamos fazendo o que ela queria e nem nos perguntávamos se queríamos, podíamos ou mesmo se sabíamos como fazer. Acho que Manuel Mendoza foi quem melhor explicou o efeito que Pilar tinha sobre a gente: "Acho que ficamos tão embriagados com o seu entusiasmo e com o amor que põe em tudo que cremos piamente que podemos realizar qualquer coisa". Também valorizava os estudantes e os apoiava em tudo o que podia. Incentivava-os a participar dos congressos sempre. Costumava dizer "Eles são o futuro da Estatística então há que motivá-los."

Uma grande amiga e uma grande companheira. Comemorei varias vitórias ao seu lado e chorei varias vezes no seu ombro. Mas nunca chorava muito, pois ela não deixava. Fazia com que eu cantasse Carinhoso, do Pixinguinha, para que ela aprendesse a letra até que eu me esquecia do porque das lágrimas.

Falei com ela uma semana antes de sua morte e ela já sabia que o fim estava próximo. Também estive com Pilar em outubro de 2006. Falamos sobre tudo: vida, morte, ciência, amores, amigos ... a amizade e sua importância e valor para nossas vidas... "Que Bueno es tener amigos", dizia. Este talvez tenha sido um dos momentos mais marcantes que já vivi. Vi uma amiga que já não mais podia andar e que mesmo assim tentava, que dependia das pessoas para muitas coisas e que ainda assim tinha fé num milagre: "De repente a ciência descobre a cura. Melhor eu estar viva, então". Vi também uma amiga que continuava dando apoio às pessoas como se os problemas delas é que fossem os mais importantes, sempre bem humorada, trabalhando com todos os seus alunos e também comigo e, como sempre, cheia de projetos (acho mesmo que esta, agora, fazendo um projeto para que nós não nos acomodemos aqui na terra.).

Como disse a um amigo muito querido logo após a morte da Pilar, sinto-me triste pois queria ter podido conviver com Pilar mais tempo. Mas me dei conta que, qualquer tempo que se viva ao lado de uma pessoa a quem se ama, mesmo que seja a eternidade, é pouco. Não adianta ficar triste, então. Seria muito egoísmo de minha parte querer Pilar por perto por mais tempo estando ela sofrendo tanto.

Vê-la em outubro mudou a minha vida em vários os aspectos. Sai do Chile com vergonha de ter preguiça e não fazer o que me corresponde fazer, sai de lá com vergonha de reclamar da vida apesar do muito que recebo todos os dias e, acima de tudo, voltei do Chile decidida a não adiar a minha vida mais. Viemos ao mundo para sermos felizes e só temos o agora para isto. O passado, já não podemos mudar; o futuro nem sei se o teremos, agora o presente, este sim, deve ser vivido intensamente.

Espero poder homenagear Pilar colocando em prática, na minha vida, todas as lições que me deu de presente. Termino aqui a minha homenagem com um trecho de uma poesia da Cora Coralina que me

toca profundamente:

"Não sei se a vida é curta ou longa demais para nós. Mas sei que nada faz sentido se não tocarmos o coração das pessoas."

E isto Pilar fez com todas as letras, com toda a intensidade.

Por Josemar Rodrigues:

Eu tive a oportunidade de conviver com a Pilar e o seu entuasiamo e alegria, era contagiante. Uma excelente amiga e uma grande perda para a comunidade bayesiana. A Pilar não está mais com a gente, mas será um referencial que nunca será esquecido.

Por Gauss Cordeiro:

A sua morte prematura é uma grande perda para a Estatística da América Latina e sobretudo para seus inúmeros amigos. Para aqueles que a conheceram de perto, ficará sempre a lembrança do seu sorriso constante e contagiante.

X CLAPEM

Adriano Polpo
(UFSCar)

O X CLAPEM ocorreu de 25 de fevereiro de 2007 a 02 de março de 2007, na cidade de Lima, Perú.

Estive presente no Congresso Latinoamericano de Probabilidade e Estatística onde pude constatar, com muito gosto, a grande presença Bayesiana, mesmo para um congresso que não tem como foco a área Bayesiana. Estiveram presentes no evento alguns nomes como: Alicia Carriquiry, Peter Mueller, Carlos Pereira, Gary Rosner, Teddy Seidenfeld, ...

No total foram 12 Plenary Talks, em que a Estatística Bayesiana foi muito utilizada, com destaque para as seguintes apresentações: Arnoldo Frigessi (Covariate modulated false discovery rate), Montserrat Fuentes (A multivariate nonparametric Bayesian spatialtemporal modeling framework for hurricane surface wind fields) e Alexandra Schmidt (Spatial Stochastic Frontier Models: Accounting for Unobserved Local Determinants of Inefficiency).

Foram também realizadas 13 seções temáticas, sendo uma delas exclusivamente Bayesiana: "Bayesian analysis", organizada por Alicia Carriquiry, com os palestrantes: Tanzi Love, Peter Muller, Gary Rosner e Ernesto San Martín. Também podemos destacar uma seção organizado pelo Carlos Pereira: "Concepts of Independence for Sets of Full Conditional Probability", que contou com a presença de Bayesianos, como: Teddy Seidenfeld (joint work with Fabio G. Cozman) e Ernesto San Martin.

A seção de poster foi dividida em duas partes: Probabilidade e Estatística, contando também com forte presença de trabalhos de Bayesianos brasileiros, como: Heleno Bolfarine, Márcia Branco, Victor Lachos, Hedibert Lopes, Helio Migon, Carlos Pereira, Adriano Polpo, Romy Ravines, Alexandra Schmidt, Ralph Silva, Carlos Valle, ...

O que poderia dizer sobre este evento para os Bayesianos? Mesmo em um evento sem forte apelação Bayesiana, estamos mostrando que a cada dia se torna mais forte e mais comum o uso da estatística Bayesiana. Sendo que os brasileiros não ficam atrás, desenvolvendo grandes trabalhos de pesquisa, engrandecendo o nome dos Estatísticos Bayesianos do Brasil.

E claro não poderia deixar de agradecer a Loretta Gasco, ao Pablo Ferrari e a todos que estiveram envolvidos de alguma forma na organização deste evento. Parabéns pela organização e qualidade dos palestrantes, um grande evento!

10^a EMR

Josemar Rodrigues
(UFSCar)

A 10^a Escola de Modelos de Regressão foi realizada pela primeira vez em numa cidade do Nordeste do País, de 25 a 28 de fevereiro de 2007, em Salvador. A sua programação incluiu dois mini-cursos, nove conferências, quatro sessões temáticas com 14 apresentações orais, oito mini-conferências proferidas por jovens doutores, quatro sessões de comunicações orais totalizando 19 apresentações, um tutorial e uma sessão ampla com apresentação de cerca de 103 posters.

Um aspecto interessante neste evento foi à ênfase dada aos procedimentos bayesianos através do minicurso do professor Dipal K. Dey, seções temáticas, conferências, comunicações orais e posters. O espírito de integração entre bayesianos e não bayesianos durante a realização do evento comprovou que é possível uma discussão científica dos problemas atuais baseados em príncipios totalmente opostos. Neste sentido gostaríamos de dar os parabéns a Comissão Organizadora da 10^a EMR, pela realização deste evento de forma profissional.

Eventos

- 9^o **Encontro Brasileiro de Estatística Bayesiana (9 EBEB)**, Maresias Beach Hotel/São Sebastião – SP, Brasil, 24 a 27 de fevereiro de 2008. (*http://www.ime.usp.br/∼isbra/ebeb/*)

  – *Conferencistas*:

  Alan Gelfand (Duke University, USA)

  Carlos Alberto de Bragança Pereira (USP, Brasil)

Dani Gamerman (UFRJ, Brasil)

Marilena Barbieri (Università di Roma3, Itália)

Marina Vannucci (Texas A&M University, USA)

Peter Muller (University of Texas/M. D. Anderson Cancer Center, USA)

Renato Martins Assunção (UFMG)

Sonia Petrone (Università Bocconi di Milano, Itália)

– *Mini-conferências*:

Patrícia Klarmann Ziegelmann (UFRGS)

Victor Hugo Lachos (UNICAMP)

Vera L. Damasceno Tomazella (UFSCAR)

– *Minicurso*:

Hedibert Freitas Lopes (University of Chicago)

– *Sessão Especial Pilar Iglesias*:

Fernando Andrés Quintana (PUC, Chile)

Ignácio Vidal Garcia (Universidad de Talca, Chile)

Sérgio Wechsler (USP, Brasil)

- Spatial and Spatio-Temporal Statistics, Fayetteville, Arkansas, USA, April 12th-14th, 2007. (*http://comp.uark.edu/∼jjsong/SLS2007/*)
- Sixth International Workshop on Objective Bayesian Analysis, Universit'a "La Sapienza", Piazzale Aldo Moro, 5 Roma – ITALY, June 9-12th, 2007. (*http://3w.eco.uniroma1.it/OB07*)
- Bayesian Inference in Stochastic Processes (BISP5), Valencia, Spain, June 14th-16th, 2007. (*http://www.uv.es/bisp5/*)
- International Workshop on New Direction in Monte Carlo Methods, Fleurance, France, June 25th – 29th, 2007. (*http://www.adapmc07.enst.fr/*)
- 5th International Symposium on Imprecise Probability: Theories and Applications, Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic, July, 16th-19th, 2007. (*http://www.sipta.org/isipta07/*)
- Tenth IMS Meeting of New Researchers in Statistics and Probability University of Utah, Salt Lake City, UT, USA, July 24 – 28, 2007. (*http://www.bios.unc.edu/∼gupta/NRC*)
- Ninth Workshop on Case Studies of Bayesian Statistics, Carnegie Mellon University, Pittsburgh, PA, USA, October 19th and 20th, 2007. (*http://workshop.stat.cmu.edu/bayes9*)

Artigos

Métodos Estatísticos Aplicados à Análise da Expressão Gênica

Saraiva, E.F.[1], Milan, L.A., e Dias, T.C.M.[2]

Resumo

Os arranjos de DNA são ferramentas utilizadas para medir os níveis de expressão de uma grande quantidade de genes ou fragmentos de genes simultaneamente, em situações diferentes. Comparando estas medidas é possível identificar genes envolvidos em doenças de origem genética. Neste texto, apresentamos quatro mé todos estatísticos que podem ser aplicados à análise da expressão gênica. O primeiro, é o teste t proposto por Baldi e Long (2001). A partir do trabalho de Baldi e Long (2001) desenvolvemos três abordagens. A primeira, considera o ajuste de modelos com uma e duas médias, seguido da seleção de modelos via fator de Bayes e DC. Na segunda utilizamos o modelo de mistura de processo Dirichlet e na terceira abordagem utilizamos o modelo com mistura infinita de distribuições.

1 Introdução

Com o desenvolvimento da genética foram criados os termos transcriptoma, que representa o conjunto completo dos transcritos (RNA's) e proteoma, que representa o conjunto completo das proteínas. À medida que mais genes vão sendo conhecidos, tem-se a possibilidade de passar para fases de análises seguintes: saber quando e onde estes genes são expressos, ou seja, o funcionamento do genoma, genoma funcional.

Segundo Felix *et al.*, (2002), "O fluxo de informação gênica do DNA nos cromossomos (genoma) até o proteoma, é intermediado pelo

[1]Primeiro lugar no concurso de dissertação do 17º SINAPE.
[2]DEs/UFSCar

conjunto das moléculas de RNA (transcriptoma). Assim, a concentração relativa de transcritos de um determinado gene em uma célula é um indicativo do quanto esse gene está sendo expresso, isto é, do quanto a célula está investindo do seu maquinário bioquímico para produzir a proteína codificada pelo gene".

Com isso, pesquisadores voltaram suas atenções ao desenvolvimento de tecnologias, visando medir a concentração relativa dos transcritos (RNA's) dos genes em células. Uma das principais ferramentas para este tipo de estudo são os arranjos de DNA.

Os arranjos de DNA são lâminas, comumente de vidro ou náilon, utilizadas para medir os níveis de expressão de uma grande quantidade de genes ou fragmentos de genes simultaneamente, em situações diferentes. Comparando estas medidas é possível identificar os genes que apresentam evidências para níveis de expressão diferentes entre uma situação de interesse e uma situação de controle. Como os dados numéricos, relacionados às medidas dos níveis de expressão dos genes são obtidos com variabilidade, métodos estatísticos são importantes para a análise dos dados com o objetivo de identificar os genes diferencialmente expressos entre as situações em estudo. E o interesse em identificar estes genes é que eles podem estar envolvidos na origem e/ou evolução de alguma doença de origem genética.

Neste texto apresentamos os resultados obtidos com a pesquisa de mestrado, desenvolvida com o objetivo de comparar o desempenho de métodos estatísticos, capazes de identificar genes diferencialmente expressos entre uma situação de interesse e uma situação de controle. O texto está organizado da seguinte forma: Na seção 2, apresentamos os métodos estatísticos utilizados na pesquisa. Na Seção 2.1 descrevemos o teste t, proposto por Baldi e Long (2001). Na Seção 2.2 utilizamos o fator de Bayes e o DIC para selecionar entre modelos com médias iguais ou diferentes. Na Seção 2.3 propomos a utilização do modelo de mistura de processo Dirichlet. Na Seção 2.4, utilizamos um modelo bayesiano com mistura infinita de distribuições. Para cada método desenvolvemos um estudo de simulação, para verificar seu comportamento na identificação dos genes diferencialmente expressos, e aplicamos a dados reais, obtidos do experimento realizado com a bactéria *Escherichia Coli* (ver Arfin *et al.*, 2000), e fazemos uma comparação de seus desempenhos. Na Seção 3 fazemos algumas considerações finais sobre os métodos propostos.

2 Análise da Expressão Gênica

Para as análises, consideramos as situações controle e tratamento e determinamos os logaritmos das medidas dos níveis de expressão observadas para cada gene em cada situação e supomos que estas medidas transformadas foram geradas segundo uma distribuição normal. Esta suposição de normalidade é muito utilizada na literatura, ver por exemplo, Baldi e Long (2001), Efron *et al.* (2001), Do *et al.* (2002).

Assim, para cada gene g, temos um conjunto de variáveis observáveis $x^c_{g1},...,x^c_{gn_c}$ e $x^t_{g1},...,x^t_{gn_t}$, independentes, representando o logaritmo dos níveis de expressão do gene g na situação controle (c) e tratamento (t), para $g = 1,2,...,G$, onde G é quantidade de genes em estudo.

2.1 Teste t

Uma abordagem utilizada para determinar se o gene g apresenta evidências para níveis de expressão diferentes, é a utilização de um teste de hipóteses sob a forma $H_0 : \mu_{gc} = \mu_{gt}$ *versus* $H_1 : \mu_{gc} \neq \mu_{gt}$, para $g = 1,...,G$.

Baldi e Long (2001) utilizam o teste t para determinar se o gene g apresenta ou não evidências para diferença, em que fixado um nível de significância α, se $|t_g|$ é maior que o valor de referência, $t_{1-\frac{\alpha}{2},p}$, então há evidências de que o gene g apresenta níveis de expressão significativamente diferentes, quando comparamos tratamento com controle, para $g = 1,2,...,G$.

Um problema que surge para a aplicação do teste t aos dados de expressão gênica, é o tamanho das amostras n_c e n_t que geralmente são pequenas.

Para verificar o comportamento do teste t na identificação dos genes diferencialmente expressos, desenvolvemos um estudo de simulação considerando diferentes afastamentos na média e na variância da distribuição de tratamento com relação a distribuição de controle.

Com este estudo observamos que o teste t identifica eficientemente os genes com evidências para diferença de médias quando as variâncias de tratamento e controle são razoavelmente estáveis. Se temos diferenças de médias acompanhada de aumento na variância de tratamento com relação a variância de controle, o teste t não identifica adequadamente os genes diferencialmente expressos.

Aplicamos o teste t a dados reais, obtidos do experimento realizado com a bactéria *Escherichia Coli* e este apresentou o mesmo comportamento observado na simulação. Ou seja, o teste t se mostra como uma ferramenta estatística inadequada para identificar alterações na média quando temos um aumento na variância de tratamento com relação a variância de controle.

2.2 Fator de Bayes e DIC

Se o gene g não apresenta evidências para diferença entre tratamento e controle, então consideramos que as medidas dos níveis de expressão observadas foram geradas de uma mesma distribuição normal, $x^c_{g1},...,x^c_{gn_c},x^t_{g1},...,x^t_{gn_t} \sim N\left(\mu_g,\sigma^2_g\right)$, e definimos esta situação como modelo M_0, para $g = 1,2,...,G$.

Caso o gene g apresente evidências para diferença, então consideramos que as medidas dos níveis de expressão observadas foram geradas de distribuições normais diferentes, $x^c_{g1}, ..., x^c_{gn_c} \sim N\left(\mu_{gc}, \sigma^2_{gc}\right)$ e $x^t_{g1}, ..., x^t_{gn_t} \sim N\left(\mu_{gt}, \sigma^2_{gt}\right)$, e definimos esta situação como modelo M_1, para $g = 1, 2, ..., G$.

Considerar se há ou não evidências para níveis de expressão diferentes para um determinado gene g equivale a considerar o modelo M_0 ou M_1 mais adequado às medidas de níveis de expressão observadas. Dessa forma consideramos a abordagem bayesiana para a modelagem e os métodos de seleção de modelos fator de Bayes e DIC[3] para selecionar entre os modelos M_0 e M_1 o que melhor representa os dados observados.

Para o cálculo do fator de Bayes utilizamos a aproximação via método MCMC (ver Kass e Raftery, 1995). Como o fator de Bayes é influenciado pelas distribuições *a priori*, utilizamos uma equalização dos hiperparâmetros de forma a reduzir esta influência.

Para verificar o comportamento do fator de Bayes e do DIC na identificação dos genes diferencialmente expressos, desenvolvemos um estudo de simulação similar ao realizado para o teste t. Aplicamos o fator de Bayes e o DIC aos dados da bactéria *Escherichia Coli*.

Na simulação e na aplicação o fator de Bayes e o DIC identificam evidências para diferença tanto com relação à variação na média quanto com relação à variação na variância, ou ambos, das medidas de tratamento com relação as medidas de controle.

Assim, acreditamos que a utilização do fator de Bayes ou do DIC tenha uma melhor performance na identificação dos genes diferencialmente expressos, que o teste t.

Para detalhes sobre a aplicação do fator de Bayes e do DIC para à análise da expressão gênica ver Saraiva et al., (2007).

2.3 Modelo MPD

Buscando diminuir as restrições para as análises propomos a utilização da abordagem bayesiana semi-paramétrica, conhecida como modelo de mistura de processo Dirichlet (modelo MPD).

De modo a facilitar o desenvolvimento das análises, introduzimos a seguinte notação.

Considere que para cada gene g em estudo, temos um conjunto de variáveis observáveis $x^{t_0}_{g1}, ..., x^{t_0}_{gn_0}$, $x^{t_1}_{g1}, ..., x^{t_1}_{gn_1}$, ..., $x^{t_K}_{g1}, ..., x^{t_K}_{gn_K}$, independentes, representando o logaritmo dos níveis de expressão do gene na situação de controle (t_0), tratamento 1 (t_1) até a situação de tratamento K (t_K) e que $\bar{x}_{gt_0}, \bar{x}_{gt_1}, ..., \bar{x}_{gt_K}$ são as médias observadas em cada situação.

Considerar se há ou não evidências para diferença equivale a considerar se a média observada, $\bar{x}_{gt_k}$ na condição de tratamento t_k, para $k = 1, 2, ..., K$, é gerada ou não da mesma distribuição da média $\bar{x}_{gt_0}$ das medidas de controle t_0, $g = 1, 2, ..., G$.

Para isso, consideramos um modelo de mistura de processo Dirichlet (ver Antoniak, 1974).

Desenvolvemos um estudo de simulação para o modelo MPD, para verificar seu comportamento na identificação dos genes diferencialmente expressos e aplicamos aos dados da bactéria *Escherichia Coli*.

Comparado aos resultados obtidos com o teste t, fator de Bayes e DIC, o modelo MPD, com relação ao teste t, possui um melhor desempenho, pois identifica diferenças de médias independentemente se temos diferença de variâncias, o mesmo não acontecendo com a aplicação do teste t. Com relação ao fator de Bayes e o DIC somente os genes com apenas diferenças de variâncias, identificados pelo fator de Bayes e pelo DIC, não são identificados pelo modelo MPD.

Para detalhes sobre a utilização do modelo MPD para a análise da expressão gênica ver Saraiva *et al.*, (2007).

2.4 Modelo com Mistura Infinita

Ao invés da análise de um gene por vez podemos estar interessados em identificar e analisar grupos de genes. Para esta finalidade, propomos um modelo com mistura infinita.

Aqui consideramos apenas a situação de tratamento e controle. Logo, vamos utilizar a notação descrita inicialmente, isto é, $x^c_{g1}, ..., x^c_{gn_c} \sim N\left(\mu_{gc}, \sigma^2_{gc}\right)$ e $x^t_{g1}, ..., x^t_{gn_c} \sim N\left(\mu_{gt}, \sigma^2_{gt}\right)$, para $g = 1, 2, ..., G$.

Seja τ_g o efeito de tratamento para o gene g, dado por $\tau_g = \mu_{gt} - \mu_{gc}$, e $d_g = \bar{x}_{gt} - \bar{x}_{gc}$ a estatística observada, que consideramos como sendo gerada de uma distribuição normal com média τ_g e variância σ^2_g, onde $\sigma^2_g = \frac{\sigma^2_{gt}}{n_t} + \frac{\sigma^2_{gc}}{n_c}$, para $g = 1, 2, ..., G$.

Assumimos um modelo bayesiano com mistura infinita de distribuições normais para a identificação dos grupos de genes.

Para identificar quais grupos são compostos por genes com medidas de níveis de expressão com evidências para diferença, consideramos os modelos: M_0, se as estatísticas observadas d_g dos genes pertencentes ao grupo não apresentam evidências para diferença; e M_1 se as estatísticas observadas d_g dos genes pertencentes ao grupo apresentam evidências para diferença.

Considerar se um grupo é composto ou não por genes com medidas de níveis de expressão com evidências para diferença equivale a considerar o modelo M_0 ou M_1 como sendo o modelo que melhor explica as medidas observadas dos genes pertencentes ao grupo. Para selecionar M_0 ou M_1, utilizamos o DIC.

Comparado aos métodos anteriores, todos os genes identificados pelo teste t e pelo modelo

[3] ver Kass e Raftery (1995) e Spiegelhalter *et al.*, (2002).

MPD também foram identificados pelo modelo com mistura infinita e DIC. Dos identificados pelo fator de Bayes e DIC, somente os que apresentam diferença de variâncias não foram identificados pelo modelo com mistura infinita e DIC.

Para detalhes sobre a utilização do modelo com mistura infinita de distribuições para a análise da expressão gênica ver Saraiva *et al.*, (2007).

3 Considerações Finais

Neste texto descrevemos os métodos estatísticos aplicados à análise da expressão gênica com objetivo de identificar os genes que apresentam evidências para níveis de expressão diferentes entre tratamento e controle.

Com os resultados obtidos temos que o teste t não se mostra eficiente para se obter resultados satisfatórios. O fator de Bayes e o DIC identificam evidências para diferença quando temos diferença de médias e/ou varâncias. O modelo MPD identifica evidências para diferença de médias, independentemente se temos diferença de variâncias entre tratamento e controle. O mesmo acontece com o modelo com mistura infinita de distribuições e DIC.

A utilização do Fator de Bayes e DIC e do Modelo MPD é interessante para uso prático, pois apresentam resultados semelhantes quando o interesse é apenas na diferença de médias, e o tempo de simulação para se obter os resultados é menor do que quando utilizamos o modelo com mistura infinita e DIC.

4 Agradecimento

Aos meus orientadores que conduziram a pesquisa de forma coerente para que fosse realizada com sucesso. A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro.

Referências Bibliográficas

[1] Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S. and Hatfield, G. W. (2000) Global Gene Expression Profiling in *Escherichia Coli* K12. *J. Biol. Chem.*, **275**, 29672-29684.

[2] Baldi, P. and Long, D. A. (2001) A Bayesian Framework for the Analysis of Microarray Expression Data: regularized t-test and statistical inferences of gene changes. *Bioinformatcs*, **17**, 509-519.

[3] Dahl, D. B. (2002) Modeling Diferential Gene Expression Using a Dirichlet Process Mixture Model. http://www.stat.tamu.edu/~dahl/em4ged/paper.pdf.

[4] Do, K.A; Müller, P. Tang, F.(2002) A Bayesian Mixture for Differential Gene Expression. http://odin.mdacc.tmc.edu/~pm/pap/DMT02.pdf.

[5] Efron, B., Tibishirani, R., Storey, J. D., and Tusher V. (2001) Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

[6] Felix, J. M., Drummond, R. D.; Nogueira, F. T. S.; Junior, V. E. R.; Jorge, R. A.; Arruda, P.; Menossi, M. (2002) Genoma Funcional.*Biotecnólogia Ciência e desenvolvimento*, **24**, 60-67.

[7] Kass, R., and Raftery, A. (1995) Bayes Factor. *Journal of the American Statistical Association*, **90**, 773-795.

[8] Saraiva, E. F., Milan, L. A. e Dias, T. C. M. (2007) Applying the Bayes Factor in the Analysis of Gene Expression (Submetido).

[9] Saraiva, E. F., Milan, L. A. e Dias, T. C. M. (2007) Analysis of the Expression Gene Data Using Dirichlet Process Mixture Model (Submetido).

[10] Saraiva, E. F., Milan, L. A. e Dias, T. C. M. (2007) Bayesian Model with Infinite Mixture applied to Analysis of Gene Expression (Submetido).

[11] Spiegelhalter, D. J., Best, N. G., Carlim, B. P., van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, B, **64** (3) 583-639.

Modelos para Processos Espaço-Temporais Inflacionados de Zeros

Fernandes, M.V.M.[1], Schmidt, A.M. e Migon, H.S.

1 Introdução

Neste trabalho consideramos modelos para processos que tipicamente assumem valores não-negativos e, frequentemente, são observados como zero; ou que assumem valores discretos, mas são inflacionados de zeros. Além disso, assume-se que esses processos são observados ao longo do tempo, em diferentes localizações. Portanto, é preciso considerar modelos que suportem a presença do excesso de zeros e, também, descreva a correlação espaço-temporal inerente as observações. Os modelos propostos são baseados na idéia originalmente proposta por Velarde *et al* (2004). Todo o procedimento de inferência é feito seguindo o paradigma bayesiano.

Generalizamos o modelo proposto em Velarde *et al* (2004) trabalhando com misturas de distribuições mas, utilizando, de acordo com o contexto, tanto distribuições contínuas (criando uma variável mista) quanto distribuições discretas. No âmbito da modelagem espacial, consideramos tanto dados de área quanto observações distribuídas continuamente no espaço, permitindo, no caso de variáveis contínuas, diferentemente de Velarde *et al* (2004), previsões para quaisquer localizações no espaço (medidas ou não).

2 Modelos para Processos Espaço-Temporais Inflacionados de Zeros

Seja $\{Y_t(s) : s \in D \subset R^2; t = 1, 2, ...\}$ um campo aleatório espacial no tempo t e localização s. Considerando o excesso de zeros nos dados e variáveis não negativas, modelamos a distribuição de probabilidades para $Y_t(s)$, da seguinte forma:

$$p(Y_t(s) \mid \theta_t(s), \lambda_t(s)) = \begin{cases} (1-\theta_t(s)) + \theta_t(s)p(Y_t(s) \mid \lambda_t(s)) & \text{se } Y_t(s) = 0, \\ \theta_t(s)p(Y_t(s) \mid \lambda_t(s)) & \text{se } Y_t(s) > 0. \end{cases} \tag{2.1}$$

Trata-se da mistura de uma distribuição de Bernoulli, com uma função densidade de probabilidade (fdp) ou função de probabilidade (fp) $p(Y_t(s) \mid \lambda_t(s))$. Note que $1 - \theta_t(s)$ representa a probabilidade de se obter um valor 0 e $\theta_t(s)$ de se obter um valor proveniente de $p(Y_t(s) \mid \lambda_t(s))$. Neste contexto, a probabilidade total de um valor nulo, é dado por $1 - \theta_t(s) + p(Y_t(s) = 0 \mid \lambda_t(s))$.

No segundo nível de hierarquia do modelo, podemos introduzir covariáveis que acreditamos influenciam θ_t ou a média de $Y_t(s) \mid \lambda_t(s)$ e aplicar estruturas dinâmicas aos parâmetros $\theta_t(s)$ e $\lambda_t(s)$. Desta forma, propomos

$$logit(\theta_t(s)) = \log \frac{\theta_t(s)}{1-\theta_t(s)} = F'_{1t}\boldsymbol{\gamma}_t + S_{1t}(s) \tag{2.2}$$

$$\boldsymbol{\gamma}_t = G\boldsymbol{\gamma}_{t-1} + w_{\boldsymbol{\gamma}t}, \quad w_{\gamma t} \sim N(0, W_{\boldsymbol{\gamma}}) \tag{2.3}$$

$$g(E(Y_t(s) \mid \lambda_t(s))) = g(\lambda_t(s)) = F'_{2t}\boldsymbol{\alpha}_t + S_{2t}(s) \tag{2.4}$$

$$\boldsymbol{\alpha}_t = H\boldsymbol{\alpha}_{t-1} + w_{\boldsymbol{\alpha}t}, \quad w_{\boldsymbol{\alpha}t} \sim N(0, W_{\boldsymbol{\alpha}}), \tag{2.5}$$

onde $g(\lambda_t(s))$ é a função de ligação da distribuição $p(Y_t(s) \mid \cdot)$.

As componentes $S_{1t}(s)$ e $S_{2t}(s)$ representam os efeitos espaço-temporais que estão presentes para capturar estruturas que as componentes em F_{it} não captam . Essa estrutura vai estar diretamente relacionada com o tipo de referência espacial dos dados em estudo. Se forem dados de área, utilizamos uma priori baseada num campo aleatório markoviano (distribuição auto-regressiva condicional (CAR)), caso estejamos trabalhando com observações feitas em pontos fixos em uma região, usaremos processos gaussianos. As estruturas F_{1t} e F_{2t} representam as covariáveis existentes e seus efeitos podem seguir uma evolução dinâmica. Após atribuir distribuições a priori para todos os parâmetros do modelo e, seguindo o teorema de Bayes, a distribuição a posteriori não possui forma analítica fechada. Para obtenção de amostras da distribuição a posteriori, utilizaremos métodos de Monte Carlo via Cadeias de Markov (MCMC). Na próxima seção discutimos esses pontos brevemente e descrevemos dois exemplos em que utilizamos o modelo proposto.

3 Aplicações

Nesta seção o modelo proposto é ajustado tanto para observações contínuas e não-negativas (nível de chuva no Rio de Janeiro), como para observações discretas que apresentam excesso de zeros (casos de dengue no Rio de Janeiro).

3.1 Modelando a chuva na cidade do Rio de Janeiro

A chuva é uma variável que assume valores positivos, mas que freqüentemente observamos período de seca (zero). Os dados utilizados são índices pluviométricos da cidade do Rio de Janeiro, que compreendem 75 semanas entre os anos de 2001 e

[1]Segundo lugar no concurso de dissertação do 17^o SINAPE.

2002. As observações foram coletadas em 32 estações monitoradoras da Geo-Rio.

Aqui $p(Y_t(s) \mid \lambda_t(s))$ em (2.2) é uma densidade que modelará os níveis de chuva maiores que zero. Distribuições como a Gama, Exponencial e a Lognormal, são prováveis candidatas neste caso. O parâmetro $\theta_t(s)$ representa a chance de obtermos um valor não-nulo (em modelos para variáveis contínuas) e neste caso $\theta_t(s)$ indica a probabilidade de ocorrência de chuva na semana t e localização s. Para a modelagem de $\theta_t(s)$ consideramos

$$\begin{aligned} logit\ (\theta_t(s)) &= \log \frac{\theta_t(s)}{1-\theta_t(s)} \\ &= \gamma_{0t} + \sum_{j=1}^{p} \gamma_j I(Y_{t-j}(s) > 0) + S_{1t}(s), \end{aligned} \tag{3.1}$$

onde γ_{0t} representa um nível que varia suavemente no tempo, e γ_j $j = 1, ..., p$ são fatores que visam incorporar a influência da existência de chuva, em p semanas anteriores, sobre a probabilidade de chuva na semana corrente. E, finalmente, $S_{1t}(s)$ representa o efeito espaço-temporal, da semana t, na localização s. Analogamente, consideramos que

$$g(\lambda_t(s)) = \alpha_{0t} + S_{2t}(s). \tag{3.2}$$

Aqui, $g(\lambda_t(s))$ é a função de ligação que dependerá da distribuição, $p(Y_t(s) \mid \lambda_t(s))$, adotada. Novamente α_{0t} representa um nível variando suavemente no tempo e $S_{2t}(s)$ um efeito espaço-temporal que vai capturar qualquer estrutura não explicada por α_{0t}. Em particular, assumimos, *a priori*, que

$$S_{jt}(\mathbf{s}) \sim GP(0, \sigma_j^2 \exp(-\phi_j \mid\mid s - s' \mid\mid)), \ j = 1, 2$$

segue um processo Gaussiano com média 0, variância σ_j^2 e função de correlação exponencial, onde ϕ_j controla quão rápido a correlação desses efeitos no espaço, decai para zero. Aqui $\mid\mid s - s' \mid\mid$ denota distância euclideana entre os pontos s e s'.

Amostras da posteriori são obtidas através de MCMC. Em particular, utilizamos o amostrador de Gibbs com alguns passos do Metropolis-Hastings e o método da rejeição adaptativo. O maior desafio é sortear os parâmetros α_{0t} e γ_{0t}, devido a correlação temporal imposta pelo modelo. Para obtenção de um algoritmo eficiente, propomos uma extensão do algoritmo proposto por Gamerman (1998) para o contexto espaço-temporal. Para maiores detalhes veja Fernandes(2006).

Para exemplificar um possível resultado desta modelagem, a Figura 1 apresenta as previsões ao longo do tempo para uma estação monitoradora que não foi considerada no procedimento de inferência, quando $p(Y_t(s) \mid \lambda_t(s))$ segue uma distribuição gama com média $\lambda_t(s)$.

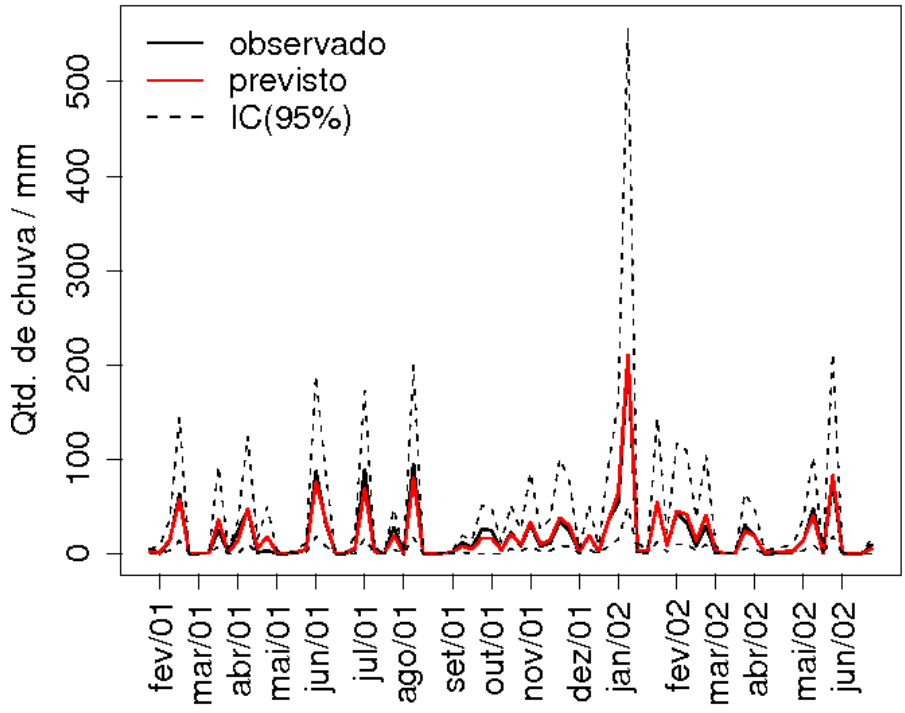

Figure 1: Previsões para a estação monitoradora de Laranjeiras - Média e Intervalo de Credibilidade

3.2 Modelando a dengue na cidade do Rio de Janeiro

O Rio de Janeiro sofreu nos anos de 2001 e 2002 uma grave epidemia de dengue. Ferreira (2003), faz uma análise espaço-temporal dos casos de dengue no Rio de Janeiro, focando esse período. Neste estudo é citado a existência de semanas onde há pouca ou nenhuma notificação de casos de dengue (períodos pré e pós-epidêmicos) tornando impossível o ajuste de um modelo Poisson. Analisamos dados correspondentes a 77 semanas obtidos em 156 bairros do Rio de Janeiro, utilizando o modelo aqui proposto.

Seja Y_{it} a notificação de casos de dengue na semana t e bairro i. Seguindo o modelo em (2.2), assumimos que $p(Y_{it} \mid \theta_{it}, \lambda_{it})$ segue uma distribuição de Poisson com média λ_{it}. Além disso, consideramos que

$$\log \frac{\theta_{it}}{1-\theta_{it}} = \gamma_0 + \gamma_1 I(Y_{it-1} > 0). \tag{3.3}$$

Vale lembrar que nesse caso $1-\theta_{it}$, representa a probabilidade de obtermos valores iguais a zero, não provenientes da distribuição $p(Y_{it} \mid \lambda_{it})$. E, também,

$$Y_{it} \mid \lambda_{it}, e_{it} \sim Po(\lambda_{it} e_{it}), \quad i = 1, \cdots, 156 \quad t = 1, \cdots, 77, \tag{3.4}$$

onde λ_{it} representa o risco relativo de dengue e e_{it}, o número esperado de casos de dengue, no bairro i, e semana t. Considerando pop_i o contingente populacional do bairro i, o número de casos esperados de dengue pode ser obtido da seguinte maneira:

$$e_{it} = \frac{\sum_{i=1}^{n} Y_{it}}{\sum_{i=1}^{n} pop_i} \times pop_i. \tag{3.5}$$

Para a modelagem do risco, propomos a seguinte estrutura:

$$\log(\lambda_{it}) = \alpha_{0t} + S_{it}, \tag{3.6}$$

Para os efeitos espaço-temporais, como estamos trabalhando com dados de área, atribuímos uma priori CAR com estrutura de vizinhança do tipo 0-1 e variância σ_S^2. Para obtenção de amostras da posteriori, propomos a utilização do mesmo algoritmo descrito na subseção anterior.

A Figura 2 apresenta os riscos de dengue para uma determinada semana do estudo nos bairros do Rio de Janeiro.

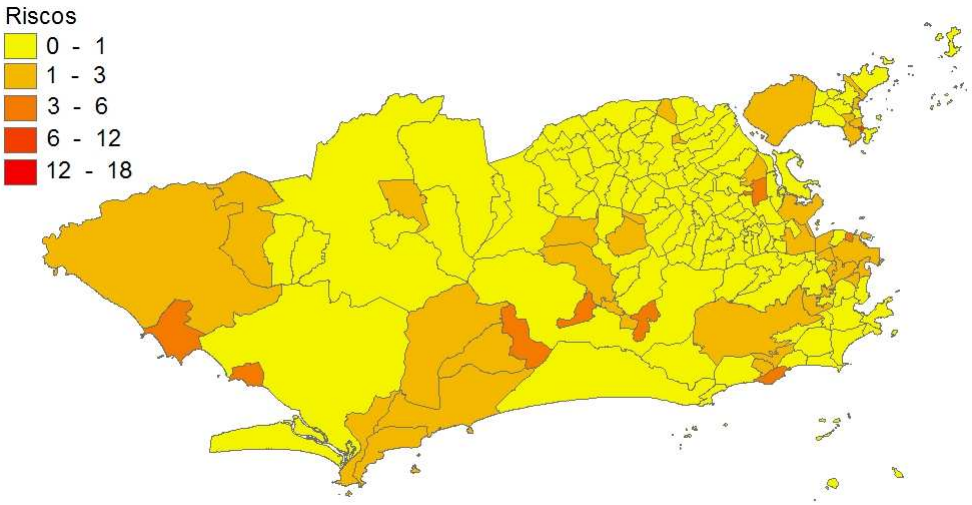

Figure 2: Riscos relativos de dengue para a 62^a semana do estudo

4 Conclusões e Projeto Futuro

O presente trabalho teve como objetivo explorar uma nova classe de modelos para observações inflacionadas de zeros e com dependência espaço-temporal. Basicamente, trabalhou-se com mistura de distribuições, sempre uma Bernoulli, modelando a existência ou excesso do valor zero, com uma outra distribuição de interesse, que foi considerada discreta ou contínua. Acreditamos que esta é uma classe flexível de modelos. O procedimento de inferência foi feito através do Paradigma de Bayes. Amostras da posteriori foram obtidas através de MCMC e um algoritmo eficiente, que leva em conta a correlação temporal dos parâmetros, foi proposto.

Uma alternativa para obtenção de amostras da posteriori dos parâmetros que evoluem no tempo é o uso do método baseado no *Linear Bayes*, proposto por Ravines *et al.* (2007). Pretendemos investigar esse ponto.

Como projeto futuro pretendemos também modelar os casos de dengue como função do nível de chuva, já que o ciclo de vida do mosquito transmissor é afetado pela quantidade de água disponível. Vários desafios surgem neste contexto, já que a resolução espacial das observações de chuva e dengue são diferentes e, além disso, o efeito da chuva sobre o risco relativo de dengue não deve ser instantâneo.

Referências Bibliográficas

Fernandes, M. V. M. (2006). Modelos para processos espaço-temporais inflacionados de zeros. Dissertação de Mestrado, Instituto de Matemática, UFRJ, Brasil.

Ferreira, G. (2003). Análise Espaço-Temporal da Distribuição dos Casos de Dengue na Cidade do Rio de Janeiro no Período de 1986 a 2002. Dissertação de Mestrado. Instituto de Matemática, UFRJ, Brasil.

Gamerman, D. (1998). Markov chain Monte Carlo for Dynamic Generalized Linear Models. *Biometrika*, 85, 215-227.

Ravines, R., Migon, H. S. e Schmidt, A. M. (2007). An Efficient Sampling Scheme for Dynamic Generalized Linear Models. Relatório Técnico, Depto de Métodos Estatísticos, IM-UFRJ.

Velarde, L. G. C., Migon, H. S. and Pereira, B. B. (2004). Space-time modeling of rainfall data. *Environmetrics*, 15, 561-576.

Inferência Bayesiana no Modelo Normal Assimétrico

Cristian L. Bayes [1] e Márcia D. Branco

1 Introdução

A distribuição normal-assimétrica introduzida por Azzalini (1985) é uma classe ùtil de distribuições que preserva algumas propriedades da distribuição normal e inclui distribuições assimétricas unimodais. Uma variável aleatória Z tem distribuição normal-assimétrica padrão se sua função de densidade de probabilidade é dada por

$$f_Z(z) = 2\phi(z)\Phi(\lambda z) \quad (-\infty < z < \infty), \tag{1.1}$$

onde $\phi(.)$ e $\Phi(.)$ são as funções de densidade de probabilidade e de distribuição de uma normal padrão, respectivamente. O parâmetro λ caracteriza a forma da distribuição e também é denominado parâmetro de assimetria, pois valores negativos de λ indicam assimetria negativa e valores positivos de λ assimetria positiva. Se $\lambda = 0$ a densidade acima coincide com a densidade da distribuição normal padrão e portanto é simétrica. Utilizaremos a seguinte notação $Z \sim SN(\lambda)$.

Uma variável mais flexível pode ser construída introduzindo-se parâmetros de posição e escala. Assim, $Y = \xi + \tau Z$. Então, Y tem função de distribuição de probabilidade dada por

$$f_Y(y) = 2\frac{1}{\tau}\phi\left(\frac{y-\xi}{\tau}\right)\Phi\left(\lambda\frac{y-\xi}{\tau}\right), \xi \in \mathbb{R}, \tau > 0. \tag{1.2}$$

Utilizaremos a notação $Y \sim SN(\mu, \sigma^2, \lambda)$, chamado modelo de três parâmetros. Neste caso a média e a variância de Y estão dadas por$E[Y] = \xi + \tau\delta\sqrt{\frac{2}{\pi}}$ e $\text{var}[Y] = \tau^2[1 - \frac{2}{\pi}\delta^2]$, sendo $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$. Notemos que, δ pode ser utilizado como uma parametrização alternativa com interpretação similar a λ, com a caraterística de ser limitado $[|\ \delta\ | < 1]$.

O coeficiente de assimetria da normal-assimétrica é dado por

$$\gamma = \sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi} - 1\right)\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^3\left(1 - \frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}\right)^{-\frac{3}{2}}. \tag{1.3}$$

Esta medida caracteriza como e quanto a distribuição se afasta da condição de simetria. γ é uma função crescente em $|\lambda|$, e se $\lambda = 0$ então $\gamma = 0$. Da expressão em (1.3) obtemos que $\gamma \in [-0.99527, 0.99527]$. O fato do coeficiente de assimetria ser limitado indica que a normal-assimétrica não consegue modelar dados com grandes assimetrias.

Inferência sob os parâmetros desta distribuição baseada em máxima verossimilhança apresenta alguns problemas, tais como:

(a) o e.m.v. para λ pode ser infinito;

(b) a informação de Fisher é singular quando $\lambda = 0$;

(c) existência de um ponto de sela.

O problema de singularidade da matriz de informação de Fisher acontece em geral quando a média do modelo está superparametrizada, como é o caso da normal-assimétrica onde $E[Y] = \xi +$

[1] Menção honrosa no prêmio de dissertação do 17^o SINAPE.

$\tau\delta\sqrt{\frac{2}{\pi}}$ depende dos três parâmetros. Para resolver este problema, Azzalini (1985) sugere a seguinte parametrização do modelo, $\varsigma = \mu + \sigma\frac{2}{\pi}\frac{\lambda}{\sqrt{1+\lambda^2}}$, $\omega = \sigma\frac{2}{\pi}\frac{\lambda^2}{1+\lambda^2}$ e γ o coeficiente de assimetria do modelo dado em (1.3), para uma maior discussão ver Pewsey (2000).

Liseo & Loperfido (2006) e Sartori (2006) propuseram diferentes métodos para resolver o problema de viés do estimador de máxima verossimilhança para o parâmetro de assimetria λ, utilizando os enfoques bayesiano e clássico, respectivamente.

Seja $y_1, \ldots, y_n$ uma amostra aleatória de $SN(\xi, \tau^2, \lambda)$. Neste caso a função de verossimilhança é dada por

$$L(\xi, \tau^2, \lambda) = \prod_{i=1}^{n} \phi\left(\frac{y_i - \xi}{\tau}\right) \Phi\left(\lambda\frac{y_i - \xi}{\tau}\right). \quad (1.4)$$

Observe que, considerando $\xi = 0, \tau = 1$ e $y_i > 0$, $\forall i$ a função de verossimilhança dada em (1.4) é uma função monótona crescente e portanto o estimador de máxima verossimilhança (e.m.v.) é infinito. Similarmente, se $y_i < 0$, $\forall i$ o e.m.v. será menos infinito. Liseo & Loperfido (2006) mostraram que os casos apontados acima caracterizam totalmente as amostras cujo e.m.v. não é finito, e a probabilidade de obter-se uma amostra com e.m.v. infinito é dada por $\left(\frac{1}{2} - \frac{1}{\pi}arctan\lambda\right)^n + \left(\frac{1}{2} + \frac{1}{\pi}arctan\lambda\right)^n$.

Sartori (2006) propos um estimador alternativo ao e.m.v. de λ baseado numa correção de viés dada por Firth (1993). Este estimador sempre é finito e será denotado por $\widetilde{\lambda}$. Para obtermos o estimador $\widetilde{\lambda}$ devemos solucionar a equação

$$l'(\lambda) + M(\lambda) = 0$$

sendo U a função escore e $M(\lambda) = -\frac{\lambda}{2}\frac{a_4(\lambda)}{a_2(\lambda)}$, com $a_k = E_Z\left[z^k\left(\frac{\phi(\lambda z)}{\Phi(\lambda z)}\right)^2\right]$, $k = 0, 1, 2, ...,$ onde os valores esperados são avaliados na distribuição normal-assimétrica padrão e é necessário utilizar métodos numéricos.

Liseo & Loperfido (2006) consideram a inferência bayesiana sob uma priori de referência, baseados no método de Berger & Bernardo (1992). No caso uniparamétrico a priori de referência coincide com a priori de Jeffreys, e é dada pela raiz quadrada da informação de Fisher. Então,

$$f^J(\lambda) \propto \sqrt{\int 2x^2\phi(x)\frac{\phi^2(\lambda x)}{\Phi(\lambda x)}dx}. \quad (1.5)$$

Liseo & Loperfido (2006) obtiveram as seguintes propriedades:

(a) $f^J(\lambda)$ é simétrica em torno de $\lambda = 0$ e decrescente em $|\lambda|$;

(b) a cauda de $f^J(\lambda)$ é da ordem $O\left(\lambda^{-\frac{3}{2}}\right)$.

Neste trabalho propomos uma boa aproximação para o fator de correção de viés M e para a priori de Jeffreys. Estas aproximações facilitaram o trabalho computacional para ambos métodos. Também, propomos uma analise bayesiana não informativa alternativa utilizando uma priori uniforme para uma reparametrização do parâmetro de assimetria.

2 Especificação a priori e aproximações

A representação estocástica para uma variável aleatória normal-assimétrica $Z \sim SN(\lambda)$ (Henze, 1986) é dada por

$$Z = \sqrt{(1-\delta^2)}V + \delta U, \quad (2.1)$$

sendo V tem distribuição normal padrão e U *half*-normal. Uma possibilidade de estabelecermos uma priori objetiva e própria é considerarmos a parametrização $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$, o qual, é limitado no intervalo $[-1, 1]$. Neste caso, a escolha natural seria $\delta \sim U(-1, 1)$, uma uniforme no intervalo $[-1, 1]$. Esta induz no espaço paramétrico de λ uma distribuição t-Student com os parâmetros especificados a seguir

$$\lambda \sim t\left(0, \frac{1}{2}, 2\right). \quad (2.2)$$

onde $t(\mu, \sigma^2, v)$ denota a função de densidade de probailidade t-Student com parâmetro de posição μ, escala σ^2 e graus de liberdade v.

A segunda priori considerada, é a priori de Jeffreys dada por Liseo & Loperfido (2006). Podemos observar que é difícil trabalhar com a expressão (1.5), razão pela qual, estamos propondo neste trabalho a seguinte aproximação para a priori de Jeffreys

$$f^J(\lambda) \approx t\left(0, \frac{\pi^2}{4}, \frac{1}{2}\right). \quad (2.3)$$

É importante notar que a cauda desta aproximação tem a mesma ordem da priori de Jeffreys, isto é, $O\left(\lambda^{-\frac{3}{2}}\right)$.
Uma boa aproximação para o fator M de Sartori é dada por $M(\lambda) = -\frac{\lambda}{2}\frac{a_4(\lambda)}{a_2(\lambda)} \approx -\frac{3\lambda}{2}\left[1 + \frac{2\lambda^2}{(\pi^2/4)}\right]^{-1}$.
Para maiores detalhes sob a obtenção destas aproximações ver Bayes & Branco (2007) e Bayes (2005).

3 Inferência à Posteriori

Para a obtenção da distribuição à posteriori consideraremos a representação estocástica dada por (2.1), assim temos a seguinte representação hierárquica

$$\begin{array}{l} Y_i|U_i, \lambda, \xi, \tau \sim N(\xi + \tau\delta U_i, \tau^2(1-\delta^2)) \\ U_i \sim HN(0,1)\ , \quad i = 1, \ldots, n \end{array} \quad (3.1)$$

sendo $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

Considerando a priori de Jeffreys para o parâmetro de locação (ξ) e escala (τ) e assumindo independência a priori dos parâmetros, obtemos a seguinte especificação a priori

$$p(\lambda, \xi, \tau) \propto \frac{1}{\tau} p(\lambda), \tag{3.2}$$

sendo $p(\lambda)$ a densidade de uma $t(0, b; d)$. Quando $d = \frac{1}{2}$ e $b = \frac{\pi^2}{4}$ obtemos a aproximação da priori de Jeffreys e para $d = 2$ e $b = \frac{1}{2}$ obtemos a priori induzida pela uniforme, $\delta \sim U(-1, 1)$.

Para facilitar o trabalho computacional consideraremos a seguinte reparametrização $\beta = \tau\delta$ and $\eta = \tau\sqrt{1-\delta^2}$. Assim, temos o seguinte modelo hierárquico

$$\begin{array}{l} Y_i|U_i, \xi, \eta, \beta \sim N(\xi + \beta U_i, \eta^2) \\ U_i \sim HN(0,1), \quad i = 1, \ldots, n. \end{array} \tag{3.3}$$

Considerando a especificação a priori dada em (3.2) e utilizando métodos usuais de transformação de variáveis, induzimos a seguinte distribuição a priori na nova parametrização,

$$f(\xi, \beta, \eta, w) \propto \frac{1}{\eta^2} exp\left(-\frac{1}{2}\frac{\beta^2 w}{\eta^2 b}\right) w^{\frac{d+1}{2}-1} exp\left(-\frac{d}{2}w\right). \tag{3.4}$$

De (3.3) e (3.4), obtemos as distribuições condicionais à posteriori, dadas por

$$w|\xi, \beta, \eta, U, Y \sim Gamma\left(\frac{d+1}{2}, \frac{1}{2}\left(\frac{\beta^2}{b\eta^2} + d\right)\right) \tag{3.5}$$

$$u_i|\xi, \beta, \eta, w, Y \sim N\left(\frac{(y_i - \mu)\beta}{\eta^2 + \beta^2}, \frac{\eta^2}{\eta^2 + \beta^2}\right) \forall i, i = 1, ..., n$$

$$\xi|\beta, \eta, w, U, Y \sim N\left(\frac{\sum_{i=1}^{n}(y_i - \beta u_i)}{n}, \frac{\eta^2}{n}\right)$$

$$\beta|\xi, \eta, w, U, Y \sim N\left(\frac{\sum_{i=1}^{n}(y_i - \xi)u_i}{\frac{w}{b} + \sum_{i=1}^{n} u_i^2}, \frac{\eta^2}{\frac{w}{b} + \sum_{i=1}^{n} u_i^2}\right)$$

$$\frac{1}{\eta^2}|\mu, \beta, w, U, Y \sim Gamma\left(\frac{n+1}{2}, \frac{1}{2}\left(\frac{w\beta^2}{b} + \sum_{i=1}^{n}(y_i - \mu - \beta u_i)^2\right)\right)$$

As distribuições condicionais à posteriori são utilizadas para implementar o algoritmo de Gibbs e obter amostras das distribuições marginais à posteriori. Este algoritmo é utilizado nas próximas seções.

4 Aplicação aos dados de fronteira

Nesta seção utilizaremos os dados de fronteira apresentados por Azzalini em sua página web *http* : *//azzalini.stat.unipd.it/SN/*, que consistem de uma amostra de 50 observações de uma $SN(0, 1, 5)$. Estes dados são interessantes pois o e.m.v. para o parâmetro de assimetria λ é infinito, embora pelo histograma (ver figura 3) dos dados não pareça que a distribuição *half*-normal ($\lambda = \infty$) seja a mais adequada para ajustá-los. Sartori (2006) obteve um valor de estimativa mais adequado, $\widetilde{\lambda} = 9.14$. Utilizando a nossa aproximação para o fator de correção de viés M obtemos $\widetilde{\lambda} = 8.67$. Estes valores são muito próximos, especialmente se consideramos que Sartori utilizou métodos numéricos para avaliar seu estimador .

Na Tabela 1 apresentamos as estimativas de máxima verossimilhança (e.m.v.), a média, a mediana e o máximo à posteriori, considerando as duas prioris para λ especificadas na seção 2, para os parâmetros do modelo.

Verifica-se que as estimativas de ξ e τ não apresentam muita diferença entre si e em relação aos verdadeiros valores dos parâmetros. No entanto, para λ essas estimativas diferem muito. Para este parâmetro, a mediana a posteriori sob a priori uniforme apresentou o melhor resultado. Sob a priori de Jeffreys o melhor resultado foi obtido pelo máximo à posteriori. Também podemos observar a assimetria positiva da distribuição à posteriori, e no caso da priori de Jeffreys uma cauda pesada a direita.

Para o cálculo do máximo a posteriori foram utilizados algoritmos de maximização numérica fornecidos pela rotina *optim* do programa R, os valores da média e da mediana à posteriori foram obtidos através do amostrador de Gibbs.

		Priori I: Jeffreys			Priori II: Uniforme		
Param	e.m.v.	Máximo	Média	Mediana	Máximo	Média	Mediana
λ	∞	6.85	∞	31.27	3.97	7.61	5.26
ξ	-0.11	-0.04	-0.10	-0.11	0.02	0.06	-0.02
τ	1.51	1.35	1.61	1.58	1.47	1.30	1.27

Table 1: Estimativas pontuais para os dados de fronteira, sob duas diferentes especificações à priori

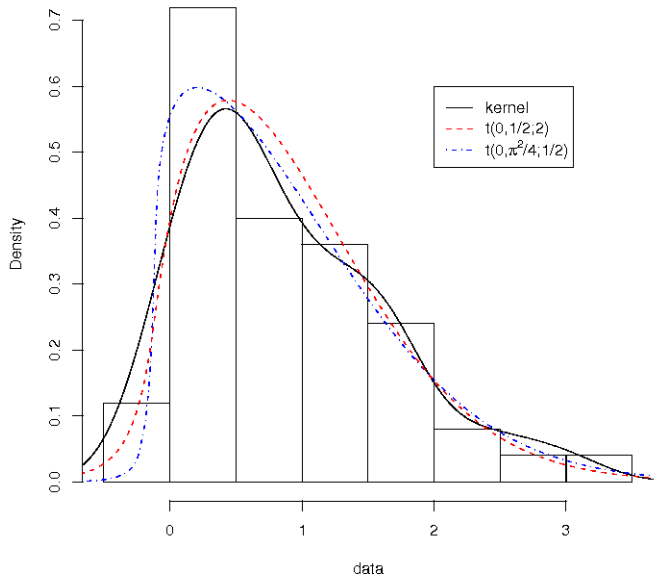

Figure 3: Histograma dos dados, distribuição estimada via kernel e preditivas

A Figura 3 confirma os resultados apresentados na tabela 1, pode-se observar que a distribuição preditiva sob distribuição à priori uniforme ajusta melhor aos dados.

5 Estudo de Simulação

Para o nosso estudo de simulação consideramos o caso uniparamétrico, sendo que para cada valor de λ foram geradas 10000 amostras de tamanho n de uma $SN(\lambda)$, com $n = 30, 50, 100, 200$. Para cada conjunto de dados calculamos o e.m.v., a média, a mediana e o máximo à posteriori sob as duas distribuições à priori apresentadas anteriormente, assim como intervalos de credibilidade com caudas iguais e HPD de probabilidade 95%.

Os resultados do estudo de simulação mostraram que na maioria dos casos a média à posteriori sob a priori uniforme e o máximo à posteriori sob a priori de Jeffreys tem um melhor desempenho que os outros estimadores, em relação ao viés e o erro quadrático médio. Em relação a estimação intervalar, os intervalos de credibilidades sob a priori uniforme mostraram-se superiores aos outros intervalos. No contexto de teste de hipóteses, o fator de Bayes sob priori uniforme conseguiu detectar melhor a assimetria dos dados do que o teste de razão de verossimilhanças e o fator de Bayes sob priori de Jeffreys. Assim, concluímos que o uso da priori uniforme é uma boa alternativa para se fazer inferência sob o parâmetro de assimetria do modelo. Para uma maior discussão ver Bayes & Branco (2007) e Bayes (2005).

6 Comentários Finais

Neste trabalho apresentamos boas aproximações para a distribuição à priori de Jeffreys, a matriz de informação de Fisher e para os termos $a_k = E[Z^k(\frac{\phi(\lambda Z)}{\Phi(\lambda Z)})^2]$, as quais acreditamos serão úteis para novas pesquisas nesta área. Também propomos uma nova reparametrização, que na abordagem clássica permite obter formas fechadas na construção do algoritmo EM; e na abordagem bayesiana, obter formas conhecidas para as distribuições condicionais à posteriori, o que facilita a implementação do algoritmo de Gibbs. O enfoque bayesiano considerando priori uniforme na parametrização δ é uma solução para o desafio proposto por Azzalini para os dados de fronteira. Além disso, o estudo de simulação confirma a vantagem do uso desta distribuição a priori.

References

Azzalini, A. (1985). A class of distributions wich includes the normal ones, *Scandinavian Journal of Statistics* **12**: 171–178.

Bayes, C. L. (2005). *Inferência bayesiana no modelo normal assimétrico*, Master's thesis, IME-USP.

Bayes, C. L. & Branco, M. D. (2007). Bayesian inference for the skewness parameter, *Brazilian Journal of Probability and Statistics* . *To appear*.

Berger, J. & Bernardo, J. (1992). On the development of reference priors, *Bayesian Statistics* **4**: 35–60.

Firth, D. (1993). Bias reduction of maximum likelihood estimates, *Biometrika* **82**: 27–38.

Henze, N. (1986). A probabilistic representation of the skew-normal distribution, *Scandinavian Journal of Statistics* **13**: 271–275.

Liseo, B. & Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution, *Journal of Statistical Planning and Inference* **136(2)**: 373–389.

Pewsey, A. (2000). Problems of inference for Azzalini's skew-normal distribution, *Journal of Applied Statistics* **27**: 859–870.

Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions, *Journal of Statistical Planning and Inference* **136(12)**: 4259–4275.

Convidamos a todos vocês a tornarem-se membros do ISBrA. O procedimento é simples, basta fazer o pagamento da anuidade do ISBA (*http://www.bayesian.org*) e depois enviar o comprovante de pagamento para *isbra@ime.usp.br*.