MODELO DE MISTURA DE PROCESSOS PONTUAIS ESTOCÁSTICOS PARA TEMPOS ENTRE EVENTOS DE SERVIÇOS NA WEB

RODRIGO AUGUSTO DA SILVA ALVES

MODELO DE MISTURA DE PROCESSOS PONTUAIS ESTOCÁSTICOS PARA TEMPOS ENTRE EVENTOS DE SERVIÇOS NA WEB

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Renato Martins Assunção
Coorientador: Pedro Olmo Stancioli Vaz de Melo

Belo Horizonte

Maio de 2015

Alves, Rodrigo Augusto da Silva

A474m Modelo de mistura de processos pontuais estocásticos para tempos entre eventos de serviços na Web / Rodrigo Augusto da Silva Alves. — Belo Horizonte, 2015
xxii, 76 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais. Departamento de Ciência da Computação.

Orientador: Renato Martins Assunção
Coorientador: Pedro Olmo Stancioli Vaz de Melo

1. Computação — Teses. 2. Operadores aleatórios — Teses. 3. Sistemas estocásticos — Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*63(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Modelo de mistura de processos pontuais estocásticos para tempos entre eventos de serviços na Web

RODRIGO AUGUSTO DA SILVA ALVES

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. RENATO MARTINS ASSUNÇÃO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO - Coorientador
Departamento de Ciência da Computação - UFMG

PROFA. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG

PROF. DANIEL RATTON FIGUEIREDO
COPPE - UFRJ

Belo Horizonte, 08 de maio de 2015.

Agradecimentos

Gostaria de primeiramente agradecer à minha mãe e ao meu pai (em memória), pois sem a dedicação que eles sempre tiveram eu não conseguiria chegar onde estou. Aos meus irmãos, Thiago, pelo exemplo, e Diogo, pela dedicação.

Um agradecimento especial para os professores Renato Assunção e Pedro Olmo, meu orientador e co-orientador, respectivamente. Nunca vou esquecer das nossas reuniões e das nossas conversas. Vocês são profissionais e pessoas exemplares.

Ao professor Daniel Figueiredo e à professora Ana Paula pela composição da banca e pelas contribuições propostas.

Aos professores do DCC pela dedicação, com destaque para professora Raquel Prates pelo apoio a mim e a este trabalho.

Aos amigos do CEFET-MG pelo apoio indiscriminado à esta capacitação, em especial, aos professores do DCSA, aos membros da COPEVE e à equipe do *Ciência, Café e Cultura*.

Agradeço à República Federativa do Brasil e a sociedade brasileira por sempre ter fornecido-me um ensino público, gratuito e de excelência.

Por fim, mas com toda a importância, agradeço à Érika pelo carinho e por acreditar em mim até quando eu mesmo não acredito.

"Essentially, all models are wrong, but some are useful."
(George Box)

Resumo

Neste trabalho é proposto um modelo de mistura de processos pontuais, distintos e estocasticamente independentes, para tempos entre eventos de serviços na *Web*. Um deles é o *Self-Feeding Process* (SFP) e o outro é um processo de Poisson homogêneo (PP). O modelo SFP tem se mostrado um excelente descritor para os tempos aleatórios de ocorrências de eventos na *Web*. A motivação para a utilização do PP é a verificação empírica de que os longos períodos de inatividade preditos pelo SFP costumam não ocorrer em alguns exemplos. Para a separação dos processos foi utilizado o Algoritmo EM. No passo E do algoritmo aproximamos o máximo da função de verossimilhança pelo seu valor esperado, uma vez que os rótulos dos processos geradores dos eventos não são conhecidos. Um teste de hipótese foi aplicado a fim de verificar se a variável adicionada no modelo mistura é realmente necessária ou se um dos processos puros, SFP ou PP, é suficiente para descrever o processo estocástico observado. Os resultados foram satisfatórios pois o modelo proposto ajusta-se bem à maioria das bases reais de dados consideradas. Ademais, duas aplicações, baseadas no nosso modelo, foram propostas: detecção de anomalias e detecção de *bursts*.

Palavras-chave: Dinâmica de comunicações, tempos entre eventos, modelo generativo, mistura de processos pontuais.

Abstract

In the present work we propose a mixture of point processes models, distinct and stochastically independent, for Internet services' inter-event times. One is the Self-Feeding Process (SFP) and the other is the homogeneous Poisson process (PP). The SFP model is an excellent descriptor for Web random event times. The motivation for the use of the PP is the empirical verification that the long periods of inactivity predicted by the SFP do not occur in some instances. To disentagle the two processes, we use the EM algorithm. In the E step we approximate the maximum of the likelihood function by its expected value because the events' labels are not known. A hypothesis test was applied to check either the additional free variable in the mixture model is actually needed or a single pure process, SFP or PP, is sufficient to describe the observed stochastic process. The results were satisfactory since the topics are well fitted by the proposed model for nine real data sets. In addition, two applications were proposed: anomaly detection and bursts detection.

Keywords: Communication dynamics, inter-event times, generative model, mixture point processes.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Definição do Problema

Desde a sua popularização, a rede mundial de computadores vem propiciando novas possibilidades para as atividades dos seres humanos. Por exemplo, a *Internet* permite compras em magazines virtuais, agendamentos de consultas médicas e visualizações de vídeos sob demanda. Nos últimos anos, com o advento da *Web* 2.0 e por conseguinte aumento da participação dos usuários da rede, cada vez mais o conteúdo da *Internet* é produzido pelos próprios usuários. Isto pode ser justificado pelo fato dos meios de comunicação estarem mais baratos, mais rápidos e mais disponíveis. Atualmente, milhões de pessoas carregam consigo dispositivos portáteis que as permitem comunicar através da *Internet* quando e onde estiverem.

O presente trabalho busca compreender como se comportam os intervalos de tempo entre eventos de diferentes tecnologias da *Web* modelando-os através de processos pontuais. Foram considerados diversos tipos de serviços: de comunicação, como fóruns de discussão e bate-papo através de hashtags (*AskMe*, *MetaFilter*, *MetaTalk* e *Twitter*); de sistemas de recomendação colaborativos (*Digg*, *Reddit* e *Yelp*) ; de correio eletrônico (*Enron*) e de controle de versão de projetos de software (*Github*). Nossa hipótese é que um modelo geral é capaz de capturar caracterísiscas semelhantes do comportamento temporal de diversas atividades de usuários de diferentes tipos de serviços disponíveis na *Web*.

1.2 Motivação

Com o advento da *Web* 2.0, a participação dos usuários da *Internet* vem ganhando cada vez mais destaque frente a publicação de conteúdos proprietários. Para O'Reilly [2007], na *Web*, páginas pessoais vem se transformando em blogs, conteúdo de manutenção de sistemas em *wikis* e enciclopédias proprietárias em enciclopédias colaborativas. Somando estes episódios à popularização dos meios de acesso à *Internet* a utilização da *Web* tem tornado-se cada vez mais ubíqua.

Além disso, nos últimos anos diversas empresas de tecnologias da informação vem ganhando destaque no mercado financeiro mundial. Segundo o *ranking* da Interbrand 2014 (InterBrand [2014]), organização que mede o valor das marcas de empresas globais, 4 das 10 marcas mais valiosas do mundo são de companhias de alta tecnologia, com destaque para a *Apple* e para a *Google* que aparecem em primeiro em segundo lugar respectivamente. Ainda, segundo este mesmo *ranking*, dentre as empresas que tiveram maiores aumentos nos valores de suas marcas em 2014, encontram-se a *Amazon* e o *Facebook*, este último liderando com 86% de aumento no valor de sua marca. Grande parte das receitas destas empresas estão ligadas aos serviços oferecidos a seus usuários e, por isso, compreender seu comportamento é essencial para a manutenção do negócio.

Neste contexto, compreender padrões do comportamento dinâmico de usuários na *Internet* se mostra promissor em diversas conjunturas pois possui uma gama de aplicações práticas tais como a recomendação de conteúdo, a detecção de eventos em tempo real e métricas de *ranking* em sistemas de buscas. A recomendação de conteúdos tem sido explorada em larga escala em serviços de e-commerce, sítios de notícias e serviços internos de tecnologias da *Web*, com o intuito de promoção dos mesmos (Matos-Júnior et al. [2012] e Schwind & Buder [2012]). Ademais, diversos trabalhos buscam extração de informações em tempo real para usos informativos, comerciais ou científicos. Em Sakaki et al. [2010], por exemplo, foram propostos algoritmos para acompanhamento em tempo real de eventos observando comentários do *Twitter*. Utilizando-se comentários deste mesmo serviço, Ribeiro Jr et al. [2012] propõem um método para georeferência de um evento através de observações de *twetts* relacionados e Liu et al. [2013] avalia a propagação da informação em escala global.

Já o presente trabalho, tem como objetivo compreender o comportamento temporal dos eventos em divervos serviços da *Web* e propor um modelo geral para a captura deste comportamento. A partir deste modelo foi possível sugerir, pelo menos, duas aplicações descritas na Seção 4.6: detecção de anomalias e detecção de *bursts* (rajadas de eventos). Espera-se, futuramente, que este modelo possa ser aplicado em outras situações práticas.

1.3 Trabalhos relacionados

Estudos relacionados a intervalos de tempos entre eventos não são recentes. Uma simples abordagem para este tipo de problema é o clássico processo de Poisson homogêneo (Seção 2.2). No entanto, é comum observar comportamentos que diferem deste tipo de processo pontual. Em diversos casos existem fluxos acelerados de eventos com um intervalo de tempo muito curto entre eles. Estes momentos são chamados de *bursts* ou rajadas. Os momentos de intensidade extrema são seguidos por longos períodos de inatividade, o que contradiz a suposição de uma média constante prevista no processo de Poisson homogêneo (Barabasi [2005]).

Recentes análises de intervalos de tempos entre eventos mostram aparente separação de ideias em dois grupos distintos. O primeiro grupo propõe que as atividades de comunicação humanas podem ser modeladas por variações de processos de Poisson. Um exemplo deste grupo pode ser encontrado em Kleinberg [2003]. O autor propõe que, na realidade, os intervalos entre eventos de envios e recebimentos de emails e publicações de notícias são distribuídos por uma exponencial de parâmetro α_i, quando está no estado i. Uma vez no estado i, o modelo prevê uma probabilidade p_{ij} de transição do estado i para o estado j. Neste estado a distribuição seria uma exponencial com taxa α_j. Basicamente, o autor propõe um processo de Poisson não homogêneo cuja taxa de eventos depende das transições de estados de um autômato. Estas transições seriam a explicação das rajadas de eventos seguidas pelos períodos de baixa atividade. A partir dos dados observados foi possível calcular a sequência de estados que tenha custo mínimo e assim identificar o autômato que representa os dados de forma latente. Os estados encontrados cujas taxas são elevadas tratam-se de *bursts*. Filtrando estes estados pela existência ou não de uma determinada palavra w foi possível construir um arcabouço de detecção de tópicos.

Já Malmgren et al. [2008] utiliza de um processo de Poisson não-homogêneo para modelagem dos tempos entre eventos de um servidor de email de uma universidade. Para eles, os *burts* podem ser explicados como comportamento humano padrão: uma vez que o usuário está utilizando o sistema de email ele está propenso a enviar emails sequencialmente. No entanto, o mesmo realiza esta atividade esporadicamente ao longo do dia, o que explicaria a ausência de eventos por um longo período. Os autores ainda verificam uma peoridicidade deste tipo de comportamento, resultando em processo de Poisson com taxa $\lambda(t) = \rho(t + W)$, onde W é o período do processo. Posteriormente, esta proposta foi atualizada em Malmgren et al. [2009] caracterizando o indivíduo em dois modos distintos: o modo "ativo" no qual os tempos entre envios de email são regidos por um processo de Poisson homogêneo com taxa $\lambda(t) = \rho_a$ e o modo "passivo"

quando estes tempos se comportam como um processo de Poisson não homogêneo com taxa $\lambda(t) = \rho(t + W)$. Em Kim et al. [2012] foi investigado um arcabouço baseado em processos pontuais, para previsão de quais imagens são mais prováveis aparecer no futuro utilizando-se de metadados e imagens do passado. O fluxo de imagens foi modelado como uma mistura de processos pontuais de Poisson não-homogêneos, cada um deles associado com um cluster de imagens de certo tópico. A função intensidade do processo de cada tópico dependia de covariáveis extraídas de metadados e de parâmetros desconhecidos que eram aprendidos de dados observados. A suposição de que cada processo componente é um processo de Poisson não-homogêneo, com intensidade conhecida, torna a tarefa de aprendizagem relativamente simples, embora não trivial. Este modelo é útil para processos com comportamento muito regular.

Em contrapartida, um segundo grupo de pesquisadores acredita que a maior parte das atividades humanas se distanciam sistematicamente do processo de Poisson tradicional. Para Barabasi [2005] tais atividades são sequenciadas por um processo de escolha baseada em uma fila de prioridade sendo muitas vezes melhor aproximada por uma distribuição de cauda pesada (*heavy tailed*) - *power-law* (Faloutsos et al. [1999]), tal que $P(\tau) \propto \tau^{-\alpha}$. A maioria das atividades são executadas muito rapidademente seguidas por tarefas que demandam muito tempo para serem concluídas.

Neste contexto, na literatura podem ser encontrados diversos trabalhos relacionados à dinâmica de interação temporal na *Web*. Em Vázquez et al. [2006] é identificado que atividades como comunicações de email e navegação *Web* seguem padrões diferentes dos previstos em processos de Poisson. Partindo da hipótese de que o escalonamento das atividades dos seres humanos é realizado por uma fila de prioridade, os autores descobriram, para o conjunto de dados estudados, que o tempo entre eventos possui uma distribuição com cauda pesada *power-law* divididas em duas classes universais: $\alpha \approx 1$ e $\alpha \approx 1.5$. A primeira classe estaria associada em geral a situações em que fila tem tamanho fixo. A segunda é associada à fila cujo tamanho é flutuante. Para eles o tamanho da fila é definido por limitações físicas da própria capacidade humana de armazenar as tarefas a serem executadas. Já Dezsö et al. [2006] detectaram que o padrão de visitação de notícias segue esta mesma distribuição com $\alpha \approx 1.2$, em contraste com a previsão exponencial fornecida por modelos simples de visitação a sites.

O desvio de certas atividades humanas em relação ao processo de Poisson, como a comunicação por email, pode impactar o comportamento de outros processos. Em Vazquez et al. [2007] é avaliado o impacto na disseminação de *worms* em duas bases de email. Foi verificado que, na abordagem tradicional, em que os tempos entre eventos são considerados distribuídos por uma exponencial, o número de mensagens infectadas decairia em cerca de 1 dia, o que contradisse as simulações realizadas que constataram

cerca de 21 dias. No entanto, supondo a distribuição dos tempos entre eventos como uma *power-law* este decaimento fica em torno de 25 dias se aproximando dos dados simulados .

O trabalho Vaz de Melo et al. [2013] pertence à perspectiva adotada no segundo grupo. Os autores deste trabalho propõem uma abordagem que utiliza um *Self-Feeding Process* (SFP) (Seção 2.3) como modelagem unificada para tempos entre eventos em serviços de comunicação na *Web*. A ideia básica deste processo é que os eventos sucessivos aumentam a chance de ocorrência de eventos adicionais ocasionando aparição de surtos de eventos seguidos de períodos de quietude. Este processo ajustou-se muito bem a várias das redes estudadas, mas em alguns casos, o ajuste não foi adequado. Neste segundo conjunto de situações, percebe-se a presença de outra fonte geradora de eventos, um processo adicional com características diferentes e misturado ao SFP. Portanto, nesta dissertação, pretende-se expandir o modelo SFP para que ele possa adaptar-se a uma classe maior de fenômenos estocásticos.

Por fim, Vaz de Melo et al. [2014] extende o trabalho anterior propondo aplicações para o SFP em geração de tempos entre eventos para indivíduos sintéticos de diversos serviços da *Web* além da proposição de um modelo de detecção de anomalias.

1.4 Objetivo

O principal objetivo deste trabalho é modelar, por meio de processos pontuais estocásticos, os tempos entre eventos relacionados a atividades de usuários em diversos sistemas disponíveis na *Web*. Ele é uma continuação do trabalho iniciado em Vaz de Melo et al. [2013] onde foi proposto um processo estocástico markoviano para os eventos pontuais de comunicação, o *Self-Feeding Process* (SFP). O SFP ajustou-se muito bem às bases estudadas, mas em alguns casos, o ajuste não foi adequado.

A proposta deste trabalho é usar um modelo de mistura, que visa solucionar o problema encontrado pelo SFP puro, em que os eventos observados sejam uma mistura de dois processos pontuais distintos e estocasticamente independentes. Um deles é o SFP. O outro é um processo de Poisson homogêneo. A motivação para este segundo processo é a verificação empírica de que os longos períodos de inatividade preditos pelo SFP costumam não ocorrer em algumas bases previamente analisadas. Nossa hipótese é que existe uma emissão de eventos aleatórios a uma taxa constante impedindo a ocorrência de períodos muito longos sem eventos. A separação dos processos não é trivial, uma vez que, os eventos emitidos pelos processos são observados sem um rótulo identificador do seu processo-fonte. Esta é a principal dificuldade do ponto de vista da

aprendizagem estatística.

São objetivos específicos deste trabalho:

- Calcular o valor esperado da função de verossimilhança do processo de mistura, considerando os dados observados e os dados incompletos;
- Utilizar a abordagem EM (Expectation Maximization) para separação da mistura;
- Aplicar o modelo em bases reais de serviços da *Web*;
- Verificar e analisar a aderência do modelo às bases reais.

Capítulo 2

Referencial Teórico

O referencial teórico deste trabalho foi baseado nas seguintes referências: Seção 2.1 em Serfozo [1990]; Seção 2.2 em Meyer [1970]; Seção 2.3 em Vaz de Melo et al. [2013] e Vaz de Melo et al. [2014]; Seção 2.4 em Guttorp & Minin [1995]; Seção 2.5 em McLachlan & Krishnan [2007] ; e Seção 2.6 em Casella & Berger [2002].

2.1 Processos pontuais temporais

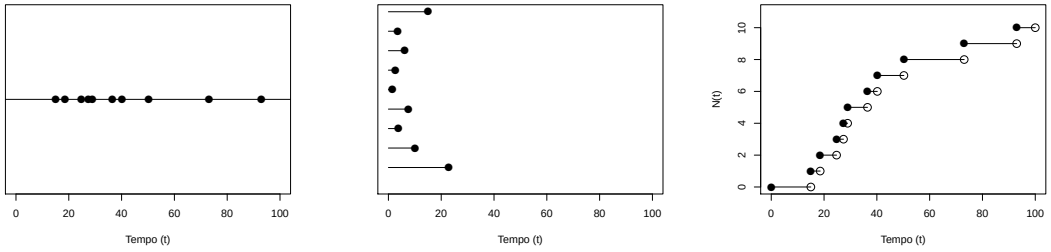
2.1.1 Definição

Um processo pontual pode ser definido como um modelo para descrever um número aleatório de ocorrências de certo evento em intervalos de tempo ou o número de pontos em uma região do espaço. Como neste trabalho analisamos apenas processos pontuais temporais, focaremos a discussão desta seção neste tipo de processo.

Um processo pontual temporal é uma lista dos tempos aleatórios da ocorrência de um determinado evento, portanto, um processo unidimensional. Muitos fenômenos reais podem ser modelados com este tipo de processo como, por exemplo, os intervalos entre terremotos, intervalos entre emergências policiais e intervalos entre envio de mensagens por usuários de um servidor de email.

Neste trabalho, o tempo inicial de um processo pontual temporal será considerado igual a zero. Seja $N^*(a, b)$ o número de eventos no intervalo de tempo $[a, b)$. Uma realização deste tipo de processo pode ser representada graficamente, equivalentemente, através das seguintes maneiras:

- uma sequência de pontos, onde cada ponto representa o tempo exato do evento (Figura 2.1a);



(a) Sequência de Pontos (b) Sequência de Intervalos (c) Função Escada

Figura 2.1: Representações de uma mesma realização de um processo pontual temporal no intervalo $[0, 100)$

- uma sequência ordenada de intervalos, onde cada intervalo representa a distância, em tempo, do evento anterior (Figura 2.1b);
- uma função escada $f(t) = N^*(0, t)$ (Figura 2.1c).

Para fins de simplificação da notação utilizada, adotamos $N^*(0, t) = N(t)$.

2.1.2 Função Intensidade em Processos Pontuais

A densidade média de um intervalo de um processo pontual é definida como a esperança matemática do número de eventos neste intervalo. Portanto, temos que a densidade média $M(a, b)$, do intervalo $[a, b)$ é determinada pela seguinte fórmula:

$$M(x, y) \quad = \quad E[N^*(a, b)] \tag{2.1}$$

Outra medida para a taxa de ocorrência de eventos em processos pontuais é chamada de função intensidade e está relacionada ao número de eventos esperados por uma unidade de tempo. Esta função é definida por:

$$\lambda(t) = \lim_{h\to 0} \frac{Pr\{N^*(t,t+h)>0\}}{h} = \lim_{h\to 0} \frac{M(t,t+h)}{h} \tag{2.2}$$

A função intensidade pode ser não determinística, ou seja, não é formada pela mesma função em todas realizações do processo. Um caso interessante dá-se quando a função intensidade não é constante no tempo, mas muda de forma estocástica como resultado da história anterior do processo. Neste caso, a função intensidade é condicionada ao comportamento do passado do processo sendo então chamada de função intensidade condicionada definida por:

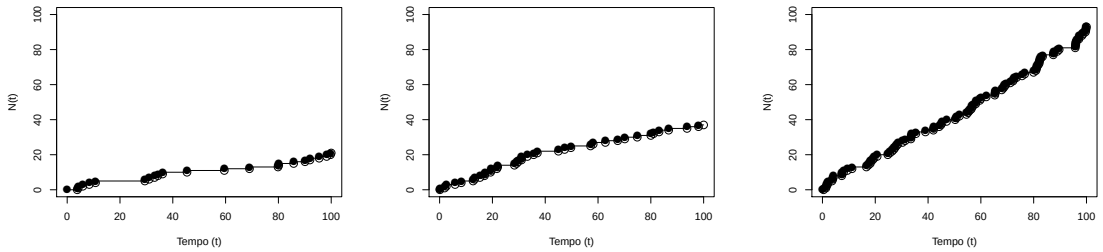
$$\lambda(t|H_t) \quad = \quad \lim_{h\to 0} \frac{M(t,t+h|H_t)}{h} \tag{2.3}$$

A notação $\lambda(t|H_t)$ mostra que a função de intensidade em um ponto t genérico, depende da história H_t do processo até o tempo t, ou seja, dos tempos dos eventos anteriores a t.

2.2 Processo de Poisson Homogêneo

Seja o vetor $T = \{T_1, T_2, T_3, \cdots\}$ composto de variáveis aleatórias independentes e identicamente distribuídas, onde T_i possui distribuição exponencial com taxa λ_{pp} maior que zero. Define-se processo de Poisson homogêneo como um processo pontual onde T_1 é a distância da origem até o primeiro evento e T_i é a distância entre evento i e o evento $(i-1)$.

O processo de Poisson homogêneo é um processo estacionário, isto é, possui função intensidade constante e densidade média, por unidade de tempo, igual a λ_{pp}. Desse modo, tem uma taxa de ocorrência fixa em todo o intervalo e independente da história H_t do processo. Na Figura 2.2, apresenta três exemplos de realizações de processos de Poisson, com diferentes valores de λ_{pp} no intervalo [0,100).



(a) $\lambda_{pp} = 0.20$ (b) $\lambda_{pp} = 0.50$ (c) $\lambda_{pp} = 1.00$

Figura 2.2: Realizações de processos de Poisson homogêneo com diferentes valores λ_{pp} no intervalo [0,100)

Para verificarmos o ajuste de uma realização de um processo pontual ao processo de Poisson homogêneo realizamos uma regressão linear considerando os tempos dos eventos t e $N(t)$, uma vez que a melhor aproximação destas duas variáveis em conjunto é uma reta. O valor do coeficiente de determinação R^2 foi considerado índice de qualidade do ajuste.

2.3 SFP (Self-Feeding Process)

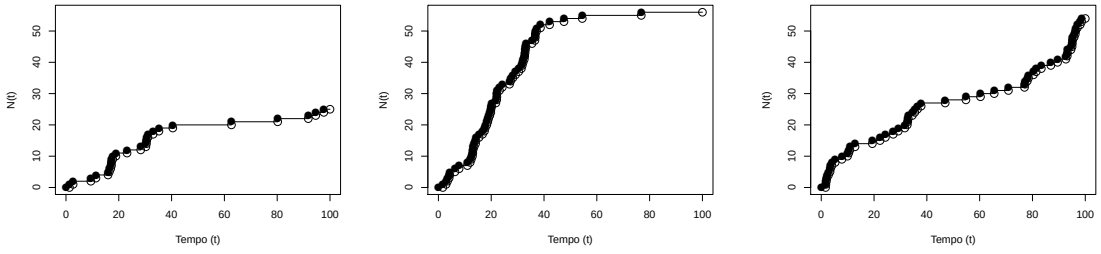
Seja o vetor $T = \{T_1, T_2, T_3, \cdots\}$ de variáveis aleatórias, onde T_i possui distribuição exponencial com taxa $\lambda_{SFP}(t|H_t)$ (2.4), t_i o i-ésimo evento do processo e Δt o último

intervalo entre eventos do processo anteriores a t. Define-se SFP como um processo pontual onde T_1 é a distância da origem até o primeiro evento e T_i é a distância entre evento i e o evento $(i-1)$.

$$\lambda_{SFP}(t|H_t) = \begin{cases} \frac{1}{\mu/e} & \text{se } t \leq t_1 \\ \frac{1}{\mu/e + t_1} & \text{se } t_1 < t \leq t_2 \\ \frac{1}{\mu/e + \Delta t} & \text{se } t > t_2 \end{cases} \tag{2.4}$$

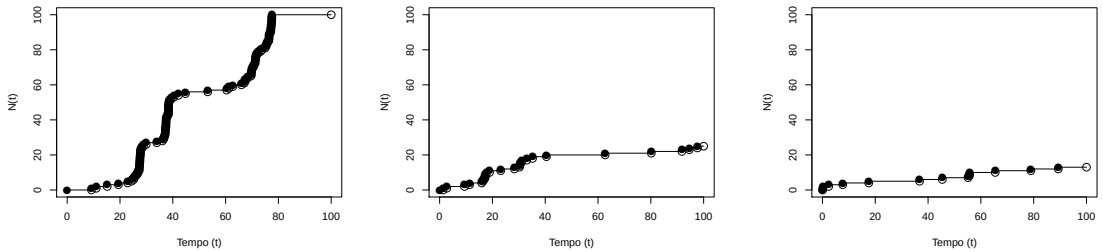
Segundo Vaz de Melo et al. [2013] , o SFP é um processo pontual que gera uma distribuição *power-law* na cauda, para distribuição de probabilidade dos intervalos, e comporta-se como um Processo de Poisson no curto prazo. O processo SFP é estacionário e sua função intensidade $\lambda_{SFP}(t) = \lambda$ é constante em t. O parâmetro μ é aproximadamente igual à mediana dos tempos entre eventos. A função de intensidade condicionada $\lambda_{SFP}(t|H_t)$ muda de forma estocástica como resultado da história anterior do processo, diferindo da função de intensidade do processo $\lambda_{SFP}(t)$ que é constante. Originalmente, o SFP foi definido com o parâmetro adicional ρ que representa o *slope* (inclinação) do processo. Neste trabalho consideraremos $\rho = 1$. Esta decisão foi motivada por duas considerações. Primeiro, pelo fato de que, para várias bases de dados similares às que nós vamos analisar, Vaz de Melo et al. [2013] encontraram $\rho \approx 1$. Segundo, a consideração deste parâmetro tornaria o cálculo da função de verossimilhança da mistura muito complexo, inviabilizando seu uso na prática.

A Figura 2.3 apresenta três realizações deste processo com o parâmetro $\mu = 1$, no intervalo de [0,100). Já na Figura 2.4, é possível observar o processo SFP para diferentes valores de μ.



(a) (b) (c)

Figura 2.3: Realizações de processos SFP com $\mu = 1$ no intervalo [0,100)



(a) $\mu = 0.10$ (b) $\mu = 1.0$ (c) $\mu = 10.0$

Figura 2.4: Realizações de processos SFP com diferentes valores de μ no intervalo [0,100)

A análise do ajuste de uma certa realização de um processo pontual ao SFP foi feita através de uma aproximação da realização à uma distribuição log-logística. A qualidade do ajuste pode ser medida pelo coeficiente de determinação R^2 da regressão linear considerando o logarítmo dos tempos entre eventos versus o logarítmo da *Odds Ratio* (OR) dos mesmos. Da relação entre estas duas variáveis espera-se uma reta. Neste trabalho, para o conjunto de tempos entre eventos, calculamos o OR para os percentis $P_1, P_2, P_3, \cdots P_{100}$. Mais detalhes podem ser encontrados em Vaz de Melo et al. [2014] .

2.4 Estimador de Máxima Verossimilhança

O estimador de máxima verossimilhança (EMV) é usado para inferir o valor do vetor de parâmetros θ de um determinado modelo estatístico paramétrico, a partir dos dados observados. Considerando que sabemos o modelo paramétrico gerador, o EMV é capaz de inferir o vetor $\hat{\theta}$ mais verossímil. Dentre as vantagens deste método encontra-se o fato que, a partir do crescimento da amostra, o EMV tende a ser uma estimativa não viciada e convergente para o valor real. Além disso, o resultado do estimador é ótimo no sentido que sua variância é aproximadamente a menor possível dentre todos os estimadores não viciados. O método consiste em obter a função de verossimilhança $L(\theta)$ e, a partir desta, encontrar o valor de θ que a maximize.

Para processos pontuais a função de máxima verossimilhança $L(\theta)$, calculada no intervalo [a,b), observado os pontos $\{t_1, t_2, t_3, \cdots, t_n\}$ é descrita a seguir:

$$L(\theta) = (\prod_{i=1}^{n} \lambda(t_i|H_{t_i})) \times e^{-\int_a^b \lambda(t|H_t)dt} \tag{2.5}$$

Em termos computacionais, é interessante extrair o logaritmo da função de verossimilhança ($\ell(\theta)$) com o intuito de transformar o máximo de um produto de funções

no máximo de uma soma de funções. Extraindo o logarítmo natural da Equação 2.5 temos:

$$\ell(\theta) = (\sum_{i=1}^{n} \log \lambda(t_i|H_{t_i})) - \int_a^b \lambda(t|H_t)dt \tag{2.6}$$

2.5 Algoritmo EM

O Algoritmo EM é um algoritmo iterativo que tem por objetivo encontrar o vetor de parâmetros $\hat{\theta}$ que atinge o máximo da função de verossimilhança quando existem variáveis não-observadas. Para isto, o algoritmo dispõe de dois passos: o passo E (Expectation) e o passo M (Maximization). No passo E é necessário o cálculo da esperança com respeito às variáveis latentes ou não-observadas da função log-verossimilhança ($Q(\theta|\theta^j, Y)$) condicionada às variáveis observadas Y e às estimativas provisórias de θ (ver (2.7)). No passo M (ver (2.8)), é necessário maximizar a esperança calculada no passo E em relação ao vetor θ.

$$Q(\theta|\theta^j, Y) = E[\ell(\theta|\theta^j, Y)] \tag{2.7}$$

$$\theta^{(j+1)} = \text{argmax}_\theta Q(\theta|\theta^j, Y) \tag{2.8}$$

O algorimo executa alternadamente os passos E e M até que uma condição de convergência seja atendida.

2.6 Teste da razão do máximo da função de verossimilhança

Em vários problemas de inferência, queremos decidir se um conjunto de dados observado segue uma classe de distribuição mais restrita de um modelo mais geral. A classe mais restrita é determinada por uma restrição nos valores permitidos para os parâmetros. Seja Θ_0 este conjunto mais restrito e Θ o conjunto mais geral de parâmetros o teste da razão do máximo da função de verossimilhança é um teste de hipótese que possui uma região de rejeição delimitado por uma constante c. Suponha:

$$Pr\left(\frac{\max L_\Theta(\theta|Y)}{\max L_{\Theta_0}(\theta|Y)} \leq c\right) \leq \alpha \tag{2.9}$$

Para testar a hipótese H_0, se Θ_0 é o modelo mais provável, ou H_1, se Θ_0 não é o modelo mais provável, temos que:

$$2 \times \log \frac{\max L_{\Theta}(\theta|Y)}{\max L_{\Theta_0}(\theta|Y)} \approx {\chi^2}_n, \tag{2.10}$$

onde ${\chi^2}_n$ é uma distribuição qui-quadrado com n graus de liberdade.

O número de graus de liberdade n é a diferença entre o número de parâmetros livres do modelo mais complexo para o modelo mais simples. Neste trabalho, utilizaremos este teste de hipótese para verificarmos se o modelo de mistura, que possui dois parâmetros livres, é realmente necessário ou se um dos processos puros que possuem um único parâmetro livre, SFP ou PP, é suficiente para descrever o processo estocástico observado. Mais detalhes serão discutidos na Seção 3.5.

Capítulo 3

Modelo de Mistura

3.1 Definição

O presente trabalho propõe um modelo baseado em uma mistura de dois processos pontuais estocásticos a fim de inferir sobre o comportamento dos tempos entre eventos em diversos serviços disponíveis na *Web* (Figura 3.1). Um dos processos da mistura é o processo de Poisson homogêneo, com intensidade λ_{pp}, e o outro processo é o SFP, com parâmetro μ. Observamos apenas a mistura, portanto os dados de entrada não possuem rótulo identificador do processo-fonte devendo este ser obtido através de inferência estatística.

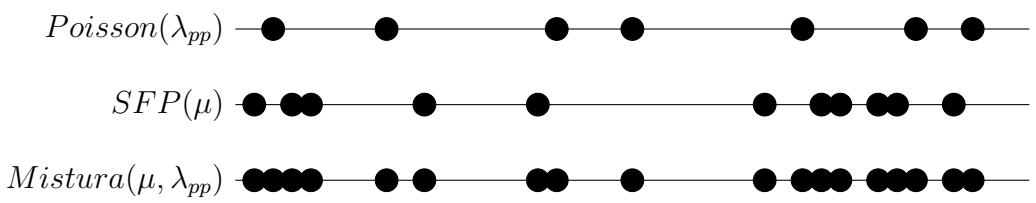
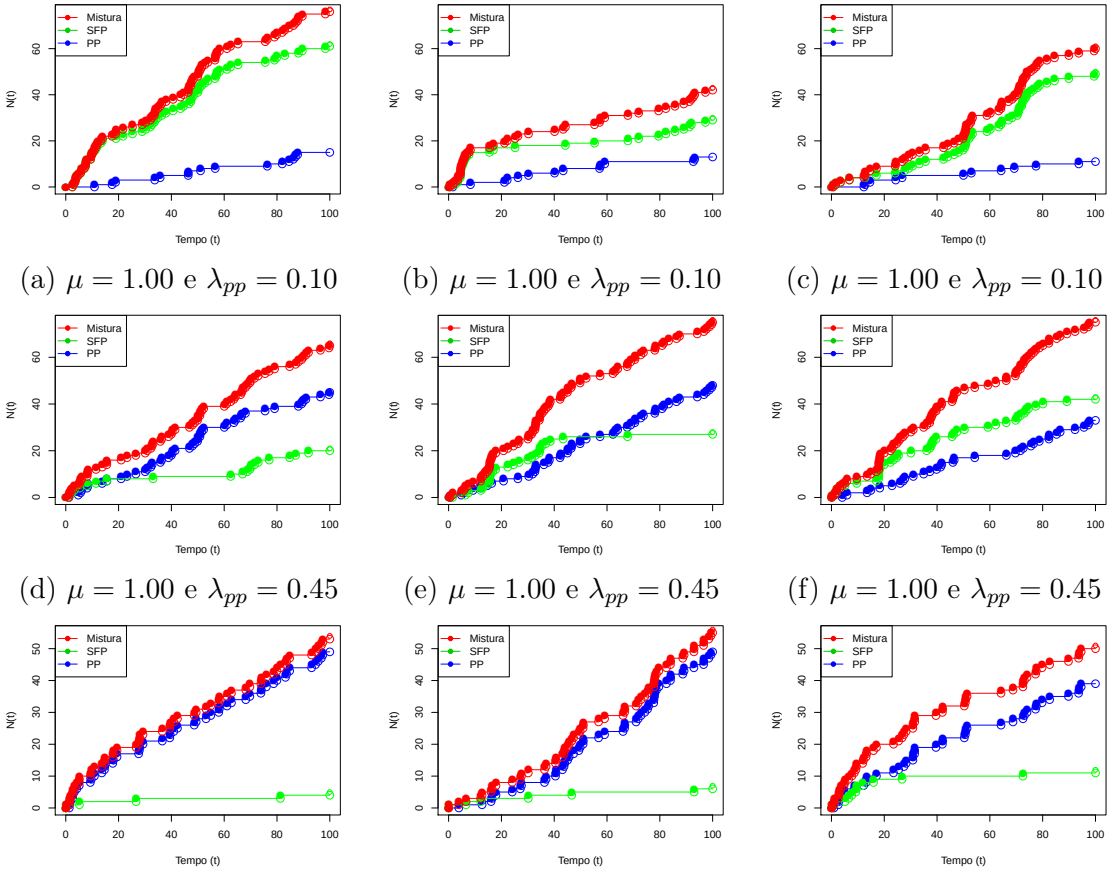


Figura 3.1: Modelo de mistura: $Poisson(\lambda_{pp}) + SFP(\mu)$

O modelo proposto exclui a possibilidade de dois eventos simultâneos na mistura. Além disso, cada evento pode pertencer apenas à um dos processos pontuais geradores, ou seja, ou ele pertence ao processo de Poisson homogêneo ou ele pertence ao SFP. Na Figura 3.2 é possível visualizar diferentes realizações do modelo de mistura no intervalo $[0, 100)$. Em cada gráfico, as curvas mostram o número acumulado de eventos até instante t. A curva em azul representa uma realização de um processo de Poisson homogêneo com parâmetro λ_{PP} e a curva verde representa uma realização de um processo SFP com parâmetro μ. A curva vermelha representa a mistura dos dois processos citados.



(a) $\mu = 1.00$ e $\lambda_{pp} = 0.10$ (b) $\mu = 1.00$ e $\lambda_{pp} = 0.10$ (c) $\mu = 1.00$ e $\lambda_{pp} = 0.10$

(d) $\mu = 1.00$ e $\lambda_{pp} = 0.45$ (e) $\mu = 1.00$ e $\lambda_{pp} = 0.45$ (f) $\mu = 1.00$ e $\lambda_{pp} = 0.45$

(g) $\mu = 8.00$ e $\lambda_{pp} = 0.50$ (h) $\mu = 8.00$ e $\lambda_{pp} = 0.50$ (i) $\mu = 8.00$ e $\lambda_{pp} = 0.50$

Figura 3.2: Realizações da mistura de processos com diferentes valores de μ e λ_{PP} no intervalo [0,100)

3.2 Função de Verossimilhança do Modelo de Mistura

A função de verossimilhança para o modelo de mistura segue o formato especificado em (2.5), proposta para processos pontuais genéricos. A intensidade do processo de mistura pode ser obtida somando as intensidades dos processos misturados, conforme pode ser visto em (3.1).

$$\begin{aligned} \lambda_{Mistura}(t|H_t) &= \lambda_{Poisson}(t|H_t) + \lambda_{SFP}(t|H_t) \\ &= \lambda_{PP} + \lambda_{SFP}(t|H_t) \end{aligned} \tag{3.1}$$

Aplicando (3.1) em (2.6), a função log-Verossimilhança $\ell(\theta)$, calculada no intervalo [a,b), para o processo de mistura é descrita:

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n} \log\left[\lambda_{PP} + \lambda_{SFP}(t_i|H_{t_i})\right] - \int_a^b \left[\lambda_{PP} + \lambda_{SFP}(t|H_t)\right] dt \\ &= \sum_{i=1}^{n} \log\left[\lambda_{PP} + \lambda_{SFP}(t_i|H_{t_i})\right] - \int_a^b \lambda_{SFP}(t|H_t)dt - (b-a)\lambda_{PP}\end{aligned} \tag{3.2}$$

Não é possível calcular (3.2), uma vez que, o cálculo de $\lambda_{SFP}(t|H_t)$ (ver (2.4)) depende do último intervalo proveniente do SFP e não possuímos os rótulos dos processos geradores. Neste caso, usamos o algoritmo EM tomando os rótulos dos eventos (PP ou SFP) como variáveis latentes ou não-observadas. Devemos calcular o valor esperado com respeito a estes rótulos da log-verossmilhança. O valor esperado da função log-verossilhança da mistura dos processos é descrito por:

$$\begin{aligned}E[\ell(\theta)] &= E[\textstyle\sum_{i=1}^{n} \log\left[\lambda_{PP} + \lambda_{SFP}(t_i|H_{t_i})\right]] - E[\int_a^b \lambda_{SFP}(t|H_t)dt] - E[(b-a)\lambda_{PP}] \\ &= \textstyle\sum_{i=1}^{n} E[\log\left[\lambda_{PP} + \lambda_{SFP}(t_i|H_{t_i})\right]] - \int_a^b E[\lambda_{SFP}(t|H_t)]dt - (b-a)\lambda_{PP}\end{aligned} \tag{3.3}$$

Fazendo uma expansão de Taylor de $\log(X)$ e tomando esperança temos que:

$$E[\log(X)] \approx \log E[X] - \frac{V[X]}{2E[X]^2} = \log E[X] - \frac{E[X^2]-E[X]^2}{2E[X]^2} \tag{3.4}$$

Aplicando (3.4) em (3.2):

$$\begin{aligned}E[\ell(\theta)] \approx \textstyle\sum_{i=1}^{n} \left[\log\left(\lambda_{PP} + E[\lambda_{SFP}(t_i|H_{t_i})]\right) - \frac{E[\lambda_{SFP}(t_i|H_{t_i})^2]-E[\lambda_{SFP}(t_i|H_{t_i})]^2}{2(\lambda_{PP}+E[\lambda_{SFP}(t_i|H_{t_i})])^2}\right] - \\ \int_a^b E[\lambda_{SFP}(t_i|H_{t_i})]dt - (b-a)\lambda_{PP}\end{aligned} \tag{3.5}$$

O valor esperado do rótulo de um evento pode ser entendido como a probabilidade do mesmo pertencer a cada um dos processos emissores de eventos. Como $\lambda_{SFP}(t|H_t)$ depende do último intervalo SFP, devemos estimar a probabilidade dos pontos anteriores a t_i pertencerem ao processo SFP e, então, calcular o valor esperado da função no ponto. Por conveniência nos cálculos, e pela necessidade de se iniciar o processo SFP, as seguintes assertivas foram consideradas:

$$Pr\left\{t_1 \in SFP\right\} = 1 \text{ e } Pr\left\{t_2 \in SFP\right\} = 1 \tag{3.6}$$

Para os demais pontos t_i, temos que:

$$Pr\left\{t_i \in SFP\right\} = \frac{\lambda_{SFP}(t_i|H_{t_i})}{\lambda_{MISTURA}(t_i|H_{t_i})} = \frac{\lambda_{SFP}(t_i|H_{t_i})}{\lambda_{SFP}(t_i|H_{t_i})+\lambda_{PP}} \tag{3.7}$$

Por consequência:

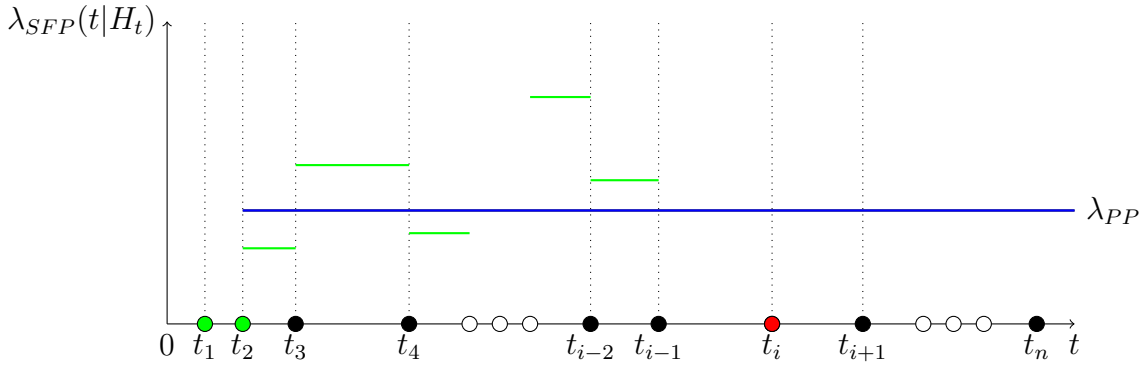
$$Pr\{t_i \in PP\} = \frac{\lambda_{PP}}{\lambda_{SFP}(t_i|H_{t_i})+\lambda_{PP}} = 1 - Pr\{t_i \in SFP\} \tag{3.8}$$

Uma vez que os rótulos de t_1 e t_2 são convencionados em (3.6) podemos calcular $\lambda_{SFP}(t_3|H_{t_3})$, pois o último intervalo SFP é (t_1, t_2). Consequentemente é possível estimar $Pr\{t_3 \in SFP\}$ e $Pr\{t_3 \in PP\}$. No entanto, $\lambda_{SFP}(t_4|H_{t_4})$ não possui cálculo direto já que depende do fato de t_3 pertencer ou não ao processo gerador SFP. Caso t_3 seja SFP, o ultimo intervalo SFP será (t_2, t_3). Entretanto se t_3 for PP, a intensidade $\lambda_{SFP}(t_4|H_{t_4})$ não sofrerá alteração em relação a $\lambda_{SFP}(t_3|H_{t_3})$ pois os dois últimos eventos SFP serão (t_1, t_2). Uma vez que conhecemos as probabilidades do evento t_3 pertencer a cada um dos processos geradores é possivel, então, calcular o valor esperado de $\lambda_{SFP}(t_4|H_{t_4})$. Com o valor esperado de $\lambda_{SFP}(t_4|H_{t_4})$ pode-se extrair $Pr\{t_4 \in SFP\}$ e $Pr\{t_4 \in PP\}$ o que forneceria os subsídios necessários para os cálculos do ponto t_5. Portanto, os valores esperados de $\lambda_{SFP}(t_5|H_{t_5}), \lambda_{SFP}(t_6|H_{t_6}), \lambda_{SFP}(t_7|H_{t_7}), \cdots, \lambda_{SFP}(t_n|H_{t_n})$, podem ser estimados, em sequência, seguindo raciocínio análogo.

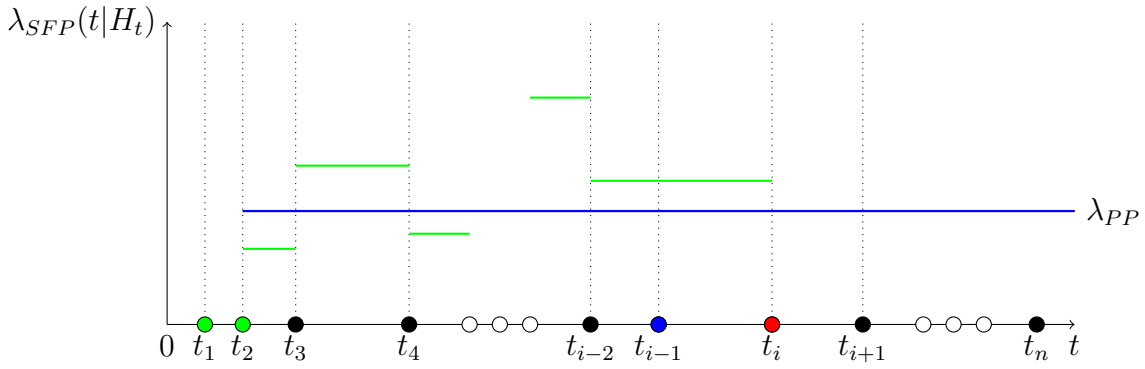
De um modo mais geral, quando desejamos encontrar $E[\lambda_{SFP}(t_i|H_{t_i})]$, conforme descrito na Figura 3.3a, conhecemos todos os valores esperados das intensitades $\lambda_{SFP}(t|H_t)$ até o ponto t_{i-1}. Este valor esperado dependerá de $t_{i-1} \in PP$ ou $t_{i-1} \in SFP$. No caso da primeira opção, o último intervalo SFP não sofrerá alteração, portanto $\lambda_{SFP}(t_i|H_{t_i})$ será igual a $\lambda_{SFP}(t_{i-1}|H_{t_{i-1}})$ (Figura 3.3b). No entanto, caso $t_{i-1} \in SFP$ deve-se levar em consideração o primeiro ponto SFP anterior a t_{i-1} no intervalo $[t_2, t_{i-2}]$. Deste modo, o último intervalo SFP provavelmente será alterado o que ocasionará mudança no valor de $\lambda_{SFP}(t_i|H_{t_i})$ em relação a $\lambda_{SFP}(t_{i-1}|H_{t_{i-1}})$ (Figura 3.3c). Portanto a esperança de $\lambda_{SFP}(t_i|H_{t_i})$ pode ser obtida através da soma de dois fatores, (3.9) e (3.10), de acordo com o rótulo de t_{i-1}. Em (3.10) e (3.11), o parâmetro $H_{(A,B)}$ simboliza que o cálculo da função intensidade do SFP considera que o úlitmo intervalo SFP é $B - A$.

$$Pr\{t_{i-1} \in PP\} \times E[\lambda_{SFP}(t_{i-1}|H_{t_{i-1}})] \tag{3.9}$$

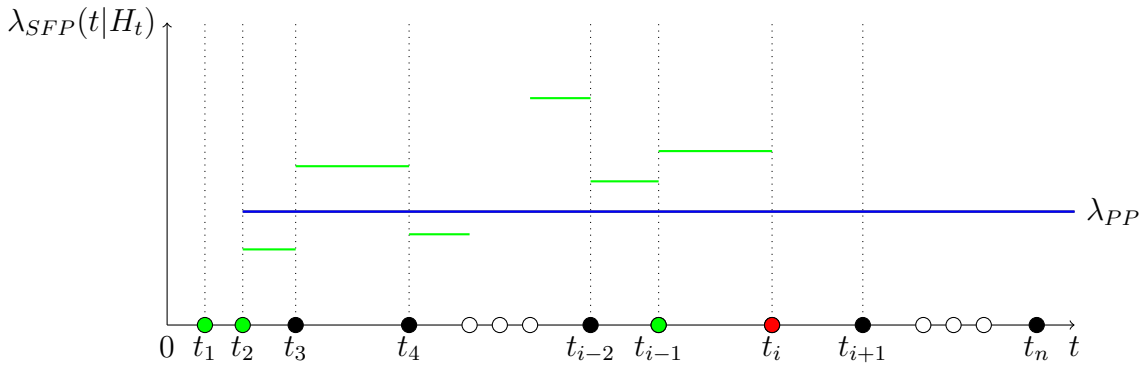
$$\begin{aligned} Pr\{t_{i-1} \in SFP\} \times (&Pr\{t_{i-2} \in SFP\} \times \lambda_{SFP}(t_i|H_{(t_{i-2},t_{i-1})}) + Pr\{t_{i-2} \in PP\} \times \\ &(Pr\{t_{i-3} \in SFP\} \times \lambda_{SFP}(t_i|H_{(t_{i-3},t_{i-1})}) + Pr\{t_{i-3} \in PP\} \times \\ &(\cdots (Pr\{t_3 \in SFP\} \times \lambda_{SFP}(t_i|H_{(t_3,t_{i-1})}) + Pr\{t_3 \in PP\} \times \\ &\lambda_{SFP}(t_i|H_{(t_2,t_{i-1})}))))) \end{aligned} \tag{3.10}$$



(a) Situação Inicial



(b) $t_{i-1} \in Poisson$



(c) $t_{i-1} \in SFP$

Figura 3.3: Influência do rótulo de t_{i-1} no valor esperado de $\lambda(t_i|H_{t_i})$

Simplificando a escrita de (3.10) temos:

$$Pr\left\{t_{i-1} \in SFP\right\} \times \sum_{k=2}^{i-2} \left\{ Pr\left\{t_k \in SFP\right\} \times [\prod_{j=k+1}^{i-2} Pr\left\{t_j \in PP\right\}] \times \lambda_{SFP}(t_i|H_{(t_k,t_{i-1})}) \right\} \tag{3.11}$$

Considerando 3.9 e 3.11 temos que:

$$E[\lambda_{SFP}(t_i|H_{t_i})] = Pr\left\{t_{i-1} \in PP\right\} \times E[\lambda_{SFP}(t_{i-1}|H_{t_{i-1}})] + Pr\left\{t_{i-1} \in SFP\right\} \times \sum_{k=2}^{i-2} \left\{ Pr\left\{t_k \in SFP\right\} \times [\prod_{j=k+1}^{i-2} Pr\left\{t_j \in PP\right\}] \times \lambda_{SFP}(t_i|H_{(t_k,t_{i-1})}) \right\} \tag{3.12}$$

De maneira análoga é possível encontrar $E[\lambda_{SFP}(t_i|H_{t_i})^2]$.

3.3 Separação dos Processos

Para o presente trabalho, a separação dos processos significa encontrar os valores de λ_{PP} e μ, dos processos geradores Poisson homogêneo e SFP, respectivamente. O Algoritmo EM (Seção 2.5) propõe uma abordagem para a solução deste problema. Normalmente, nas iterações deste algoritmo, é necessário saber a proporção dos pontos de cada um dos processos misturados de forma explícita. No entanto, no presente problema esta proporção está diretamente atrelada às intensidades dos processos e por este motivo o cálculo da proporção é realizado de forma implícita. O passo E, descrito em (2.7), foi detalhadamente explicado na Seção 3.2. O valor esperado da função intensidade $\lambda_{SFP}(t_i|H_{t_i})$ é função dos parâmetros λ_{PP}, μ e dos dados observados (Y), ou seja, $E[\lambda_{SFP}(t_i|H_{t_i})] = E[f(\lambda_{PP}, \mu, Y)]$. Seja, $\theta = (\lambda_{PP}, \mu)$, aplicando (3.5) em (2.7) temos:

$$Q(\theta|\theta^j, Y) \approx \sum_{i=1}^{n} \left[\log\left(\lambda_{PP} + E[f(\lambda_{PP}, \mu, Y)]\right) - \frac{E[f(\lambda_{PP},\mu,Y)^2] - E[f(\lambda,\mu,Y)]^2}{2(\lambda_{PP} + E[f(\lambda_{PP},\mu,Y)])^2} \right] - \int_a^b E[f(\lambda_{PP}, \mu, Y)]dt - (b-a)\lambda_{PP} \tag{3.13}$$

No que tange ao passo M (ver 2.8) devemos maximizar (3.13) em relação ao vetor θ. Uma solução seria o cálculo das derivadas parciais para cada um dos parâmetros do vertor θ conforme (3.14).

$$\frac{\partial Q(\theta|\theta^j,Y)}{\partial \lambda_{PP}} = 0 \text{ e } \frac{\partial Q(\theta|\theta^j,Y)}{\partial \mu} = 0 \tag{3.14}$$

O cálculo analítico de (3.14) para o presente problema é de grande complexidade.

Por este motivo, adotamos uma aproximação do passo M, buscando o valor de θ^{j+1} que maximiza a função $Q(\theta|\theta^j, Y)$. O método selecionado para busca do valor de que maximiza é um algoritmo que combina o método da seção áurea com interpolação parabólica sucessivas (Brent [1973]) implementado na função *optimize* do software estatístico R (R Core Team [2014]). O algoritmo foi desenvolvido para funções contínuas, como é o caso da função que desejamos maximizar. Na verdade, o mesmo calcula o mínimo de uma função. Como desejamos obter o máximo foi necessário inverter o valor da função e procurar o seu mínimo.

Para facilitar a convergência e buscar valores factíveis, e não obter resultados puramente matemáticos, foi proposto um intervalo de busca para cada um dos estimadores. O estimador $\hat{\lambda}_{PP} \in [0, n/t_n]$, onde n é o número de pontos da mistura e t_n é o último ponto observado da mistura, assumindo o limite inferior do intervalo quando a mistura se tratar na realidade apenas de um processo SFP e assumindo o limite superior quando a mistura se tratar de um processo de Poisson puro. Já $\hat{\mu} \in [0, t_n]$. O valor de $\mu = 0$ simbolizaria que o SFP teria mediana 0, ou seja, infinitos eventos em um intervalo muito pequeno. Apesar de ser um limite inferior natural, este resultado não é prático. Quando assumimos $\mu = t_n$ o processo trata-se de um Poisson puro. Na realidade, esta última situação acontece quando $\mu = \infty$. No entanto, ao igualarmos $\mu = t_n$, entendemos que metade dos tempos entre eventos possuem valor superior ao tamanho do intervalo, o que aproxima sistematicamente o número de eventos SFP a zero. Portanto o valor de t_n para o parâmetro descrito é suficiente limite superior para a busca de soluções.

3.4 Simulações do EMV da Mistura de Processos

Diversas simulações foram realizadas para verificar se a estimativa dos parâmetros proposta na Seção 3.3 é adequada. Não existe teoria sobre o comportamento do EMV no caso de dados de processos pontuais seguindo um modelo complexo como o nosso modelo de mistura. Para isto, foram gerados dados sintéticos variando o tamanho da amostra e os parâmetros λ_{PP} e μ da mistura. O tamanho da amostra N, considerado para esta análise, varia de 100 a 1000 pontos, em intervalos de 100 pontos, para cada par (λ_{PP},μ). Foram selecionados os parâmetros λ_{PP} e μ afim de variar a porcentagem esperada de pontos provenientes do processo de Poisson ($\%PP$), na proporção de 10% a 90%, em intervalos de 10%. Como os dois processos são estacionários, calculamos, de forma empírica, combinações de λ_{PP} e μ para que se obtenha a porcentagem desejada dos processos.

Para cada dupla (N,$\%PP$) foram realizadas 100 simulações, totalizando 9.000 simulações. O resultado de cada simulação foi medido a partir da distância do estimador para o valor real. Considerando o valor real do parâmetro de cada processo gerador ($\overline{X}$), a função $\Delta(\hat{X})$ é definida como:

$$\Delta(\hat{X}) = \begin{cases} \hat{X}/\overline{X} - 1 & \text{se } \hat{X}/\overline{X} \geq 1 \\ 1 - \overline{X}/\hat{X} & \text{caso contrário} \end{cases} \tag{3.15}$$

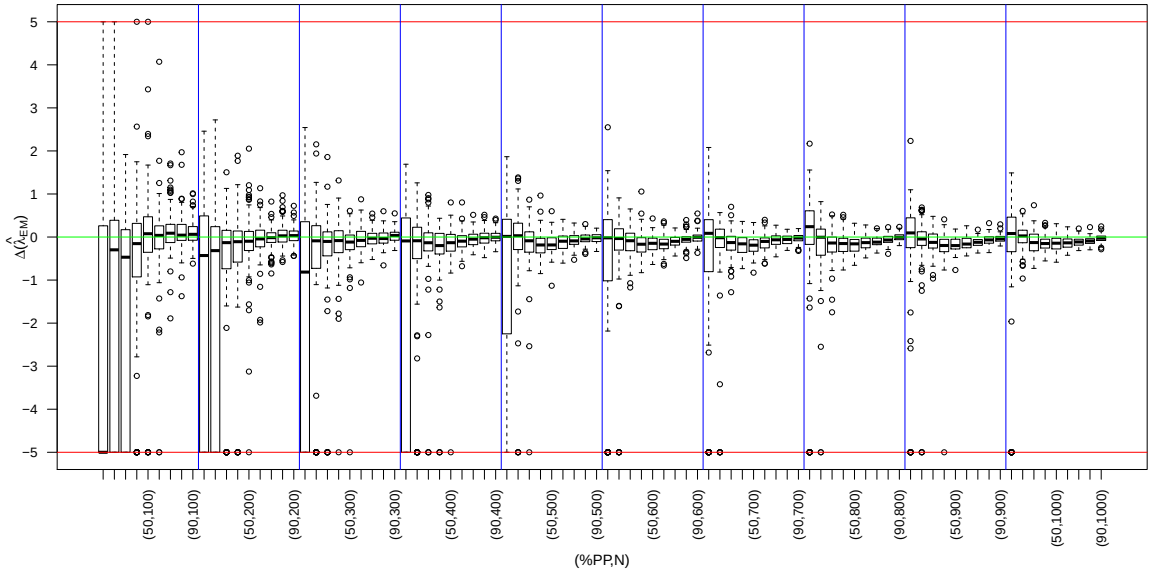
Cada unidade positiva resultante de (3.15) refere-se a um distanciamento relativo positivo de 100% do estimador ao valor real, ou seja $\hat{X} = \overline{X} + \Delta(\hat{X}) \times \overline{X}$. Analogamente, a unidade negativa refere-se a um distanciamento relativo negativo de 100% do estimador ao valor real, ou seja $\hat{X} = \overline{X}/|\Delta(\hat{X}) + 1|$. O valor $\Delta(\hat{X}) = 0$ implica $\hat{X} = \overline{X}$.

3.4.1 Estimador $\hat{\lambda}_{EM}$

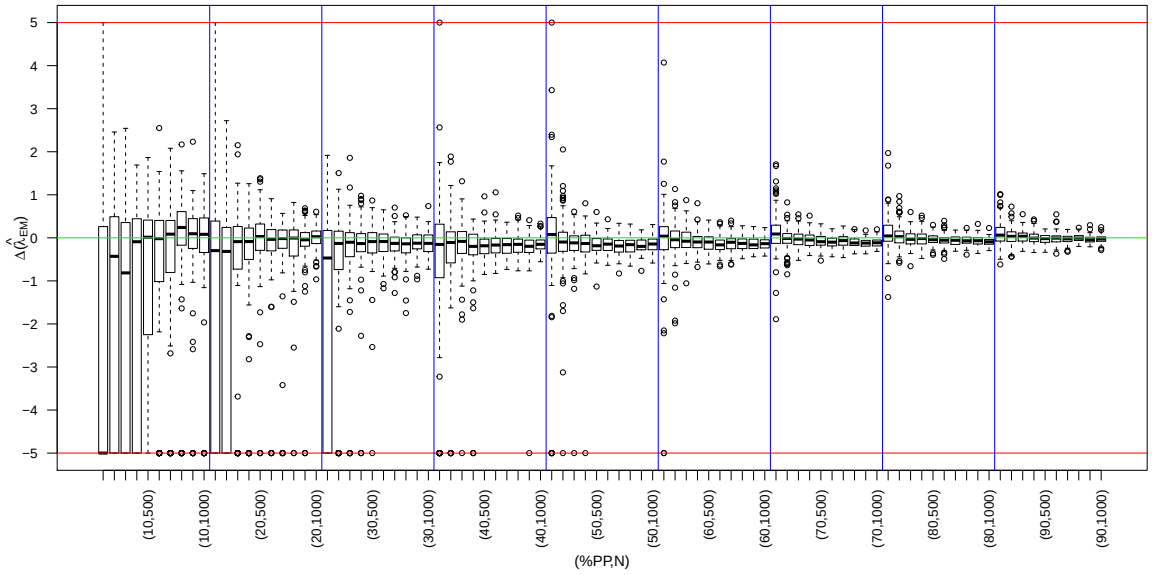
O $\hat{\lambda}_{EM}$ é a estimativa de λ_{PP} através do algoritmo EM conforme proposto na Seção 3.3. Os resultados de $\Delta(\hat{\lambda}_{EM})$ das simulações, agrupados pelo tamanho da amostra e pela porcentagem de pontos do processo de Poisson, podem ser analisados nos boxplots das Figuras 3.4a e 3.4b, respectivamente. Nestas figuras, as linhas vermelhas são chamadas de linhas de censura. Simulações onde $\Delta(\hat{\lambda}_{EM}) \leq -5$ foram censurados em -5 e simulações onde $\Delta(\hat{\lambda}_{EM}) \geq 5$ foram censurados em 5. Já a linha verde significa $\lambda_{EM} = \lambda_{PP}$.

Em algumas situações o EMV detectou ausência do processo de Poisson, mesmo quando a porcentagem esperada para este processo diferia de zero. Nestes casos em que o método considera a mistura como um SFP puro o mesmo aproxima substancialmente o valor de $\hat{\lambda}_{EM}$ de zero o que explicaria os severos distanciamentos do estimador do valor real de λ_{PP} e por conseguinte $|\Delta(\hat{\lambda}_{EM})|$ muito elevado. Este fato pode ser comprovado pela assimetria de alguns boxplots a esquerda da Figura 3.4 quando temos N ou $\%PP$ baixos. Neste cenário, em que o SFP se sobressai perante o processo de Poisson na mistura, percebemos que a influência significativa do PP nos momentos de quietude do SFP auxilia a detecção do mesmo pelo EMV, a exemplo da Figura 3.5a. Nesta realização podemos verificar que o processo de Poisson tem pleno domínio nos momentos de inatividade do SFP enquanto, nos *bursts*, o SFP se destaca. O EMV, neste caso específico, resultou em uma estimativa com erro menor que 1% para λ_{PP}.

Entretanto, nem sempre esta característica é percebida. Duas razões, que esclarecem tal comportamento, foram detectadas. A primeira razão é a baixa quantidade de pontos, tanto do total de pontos da mistura dos processos quanto de pontos pro-

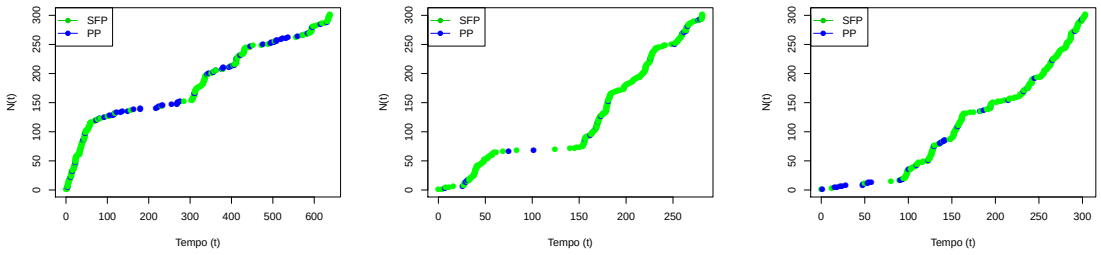


(a) Agrupado por N



(b) Agrupado por $\%PP$

Figura 3.4: Resultado do estimador $\hat{\lambda}_{EM}$ para processos simulados



(a) (b) (c)

Figura 3.5: Exemplos de simulações de processos misturados com $N = 300$ e $\%PP = 20$

venientes do processo de Poisson homogêneo. Por exemplo, na Figura 3.5b podemos vizualizar uma realização de mistura cujo número de pontos provenientes do processo de Poisson homogêneo é tão baixo que aparentemente parece factível ter sido produzido pelo próprio SFP da mistura. Conseguimos identificar os períodos de grande atividade seguidos dos períodos de quietude com quase nenhuma influência do PP.

A segunda razão é o fato do SFP comportar-se como um processo de Poisson em pequenos intervalos de tempo. Por este motivo, em algumas situações nestes pequenos intervalos, a mistura dos processos pode levar a crer que, na realidade, trata-se apenas de um PP formador do SFP. A realização representada na Figura 3.5c é um exemplo desta razão. Apesar do período de quietude no início do processo ser afetado pelo PP, podemos observar entre os tempos 150 e 200, por exemplo, intensidades bem próximas à do processo de Poisson. Nestes dois últimos casos citados a estimativa encontrada para o parâmetro do Processo de Poisson foram bem próximas do valor zero.

É importante destacar que este comportamendo instável do estimador é atenuado com o aumento do número de pontos da realização ou com o aumento da porcentagem de pontos do processo de Poisson.

Em virtude das razões citadas acima, uma análise baseada em médias aritméticas seria inadequada e tendenciosa. O fato de poucos pontos aproximarem de zero faria $|\Delta(\hat{\lambda}_{EM})| \approx \infty$ e, consequentemente, elevaria substancialmente o valor da média de $|\Delta(\hat{\lambda}_{EM})|$. Por este motivo analisar as medianas e os percentis de $\Delta(\hat{\lambda}_{EM})$ é mais adequado. Pela Figura 3.4 podemos verificar que o algoritmo EM se comportou como esperado: quanto maior o tamanho da amosta N melhor o comportamento do estimatidor do parâmetro, ou seja, este está mais próximo do valor real e com menor variância. De forma semelhante quanto maior a porcentagem de pontos provenientes do processo de Poisson mais próximo estava λ_{EM} de λ_{PP}. Os dados tabulados que demonstram este desempenho encontram-se na Tabela 3.1. A despeito dos baixos valores de $\%PP$ ou N, os dados nos mostram que, de um modo geral, λ_{EM} levemente subestima λ_{PP} e varia pouco em torno do valor real.

Tabela 3.1: *Heatmap* de $\Delta(\hat{\lambda}_{EM})$ em função de %PP versus N

N \ %PP	10%	20%	30%	40%	50%	60%	70%	80%	90%
100	-606.1565	-0.2971	-0.4690	-0.1527	0.0768	0.0380	0.0885	0.0434	0.0626
	9041.2117	8787.8752	2819.2986	0.9002	0.4939	0.2700	0.2952	0.3008	0.2474
200	-0.4296	-0.3156	-0.1302	-0.1082	-0.0975	-0.0439	-0.0143	0.0280	0.0335
	6695.2632	1321.8351	0.7333	0.5622	0.3092	0.2106	0.1254	0.1610	0.1432
300	-0.8143	-0.0878	-0.1024	-0.0824	-0.1166	-0.0777	-0.0291	-0.0346	0.0363
	1.27e+4	0.7218	0.4140	0.3539	0.2874	0.2200	0.1564	0.1402	0.1139
400	-0.0903	-0.0864	-0.1325	-0.1984	-0.1315	-0.0961	-0.0494	-0.0165	-0.0057
	1.18e+4	0.4980	0.3221	0.3844	0.3198	0.2283	0.1749	0.1385	0.0877
500	0.0146	0.0301	-0.0873	-0.1846	-0.1840	-0.0973	-0.0895	-0.0431	-0.0175
	2.2209	0.3233	0.3491	0.3659	0.2900	0.2626	0.1700	0.1082	0.1023
600	-0.0190	-0.0350	-0.0861	-0.1673	-0.1457	-0.1633	-0.1002	-0.0541	-0.0070
	0.9751	0.2899	0.3169	0.3395	0.2931	0.2671	0.1866	0.1123	0.0934
700	0.0860	-0.0163	-0.1284	-0.1581	-0.1892	-0.1025	-0.0644	-0.0570	-0.0236
	0.7967	0.2377	0.3038	0.3521	0.3302	0.2517	0.1697	0.1510	0.0870
800	0.2408	-0.0050	-0.1388	-0.1514	-0.1581	-0.1295	-0.1187	-0.0686	-0.0063
	0.6102	0.4221	0.3439	0.3315	0.3255	0.2491	0.1743	0.1222	0.0598
900	0.0954	-0.0454	-0.1244	-0.1965	-0.1980	-0.1631	-0.1220	-0.0668	-0.0439
	0.4483	0.1894	0.2534	0.3389	0.2760	0.2491	0.1917	0.1399	0.0949
1000	0.0816	0.0260	-0.1284	-0.1509	-0.1433	-0.1347	-0.1140	-0.0926	-0.0270
	0.4643	0.1623	0.3028	0.2553	0.2714	0.2347	0.1874	0.1444	0.0798

LEGENDA:

Mediana	-1	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	1
|Percentil|*	1	0.8	0.6	0.4	0.2	0					

*Percentil 25 ou 75, o que tiver maior módulo

3.4.2 Estimador $\hat{\mu}_{EM}$

O $\hat{\mu}_{EM}$ é a estimativa de μ através do algoritmo EM conforme proposto na Seção 3.3. Os resultados de $\Delta(\hat{\mu}_{EM})$ das simulações, agrupados pelo tamanho da amostra e pela porcentagem de pontos do processo de Poisson, podem ser analisados nos boxplots das Figuras 3.6a e 3.6b, respectivamente. Nestas figuras, as linhas vermelhas são chamadas de linhas de censura. Simulações onde $\Delta(\hat{\mu}_{EM}) \leq -5$ foram censurados em -5 e simulações onde $\Delta(\hat{\mu}_{EM}) \geq 5$ foram censurados em 5. Já a linha verde significa $\hat{\mu}_{EM} = \mu$.

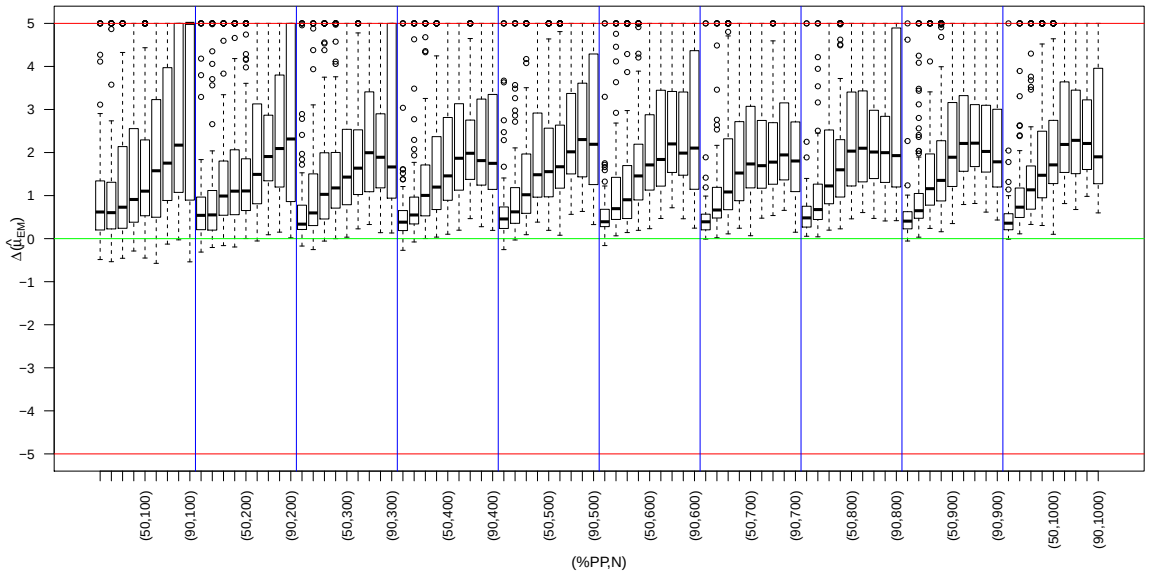
O algoritmo EM superestimou excessivamente o valor de μ na quase totalidade dos casos observados. A proporção de pontos provenientes do processo SFP foi mais significante que o tamanho total da amostra para a qualidade da estimativa (Tabela 3.2). Os melhores resultados foram quando a proporção esperada do SFP foi de 90%.

No entanto, no outro extremo, quando a porcentagem de SFP é baixa, o fato de $\hat{\mu}_{EM}$ superestimar o valor de μ nos faz entender que o método considera que só existe o processo de Poisson na mistura. Esta região é suposta de grande instabilidade. Teoricamente, uma mistura que só tem pontos provenientes do processo de Poisson possui $\mu = \infty$. A presença de poucos pontos pode gerar uma grande variação no valor de $\hat{\mu}_{EM}$ no intervalo que vai da mediana dos tempos entre eventos e infinito.

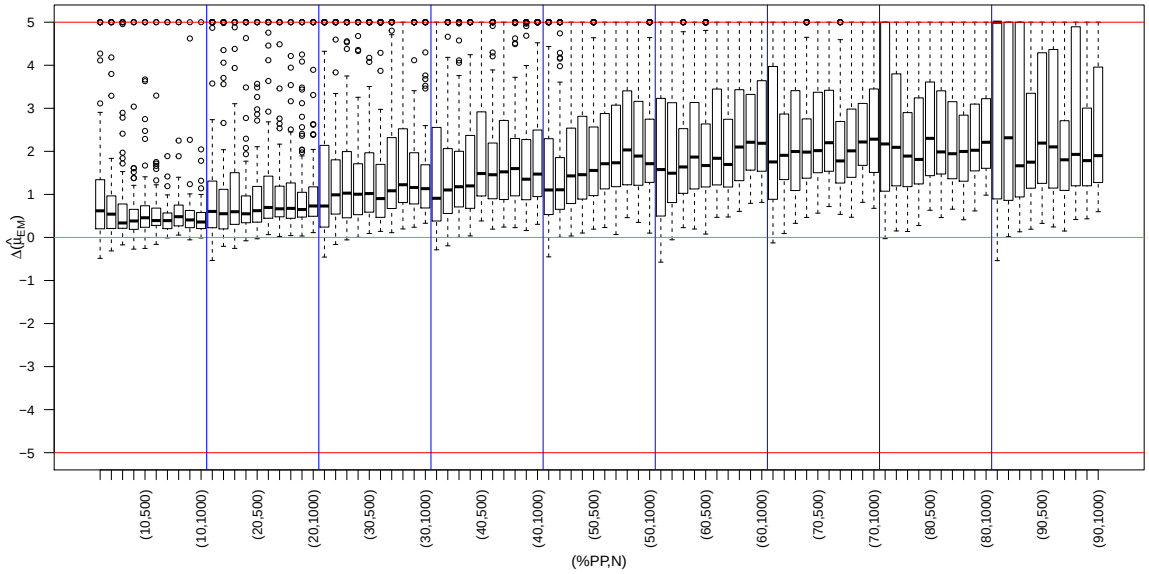
Uma questão a ser levantada é o motivo do algoritmo obter bons resultados para um parâmetro e não para o outro. Acreditamos que a má estimativa de μ pelo algoritmo EM está relacionada com o cálculo do valor esperado da função de verossimilhança. Neste cálculo foram realizadas algumas aproximações como pode ser visto na Seção 3.2. Estas aproximações estão ligadas ao fato de não conhecermos os rótulos dos dados observados, o que influencia diretamente no cálculo estocástico da função intensidade do SFP. No entanto, esta influência é menos significativa no processo de Poisson uma vez que a intensidade deste processo é fixa e determinística durante todo intervalo.

A fim de melhorar a estimativa do parâmetro μ do SFP foi proposto um novo estimador: o $\hat{\mu}_{Median}$. Ele considera o valor de $\hat{\lambda}_{EM}$ suficiente e retira pontos que seriam do processo de Poisson, estimando μ pela mediana através dos pontos que, supostamente, seriam do SFP. Mais detalhes podem ser vistos na Seção 3.4.3.

Na Figura 3.7 podemos analisar gráficos de simulações considerando $\Delta(\hat{\lambda}_{EM})$ versus $\Delta(\hat{\mu}_{EM})$ para diversos valores de $\%PP$ e N. Estes gráficos são úteis para verificar possíveis regiões de concentração dos pontos analisando um impacto no resultado de um parâmetro em outro. O comportamento conjunto corrobora com o comportamento individual descrito na Seção 3.4.1 e nesta seção. Verificamos, em geral, *clusters* acima do eixo das abscissas e variando em volta do eixo das ordenadas, próximo de zero, o que



(a) Agrupado por N



(b) Agrupado por $\%PP$

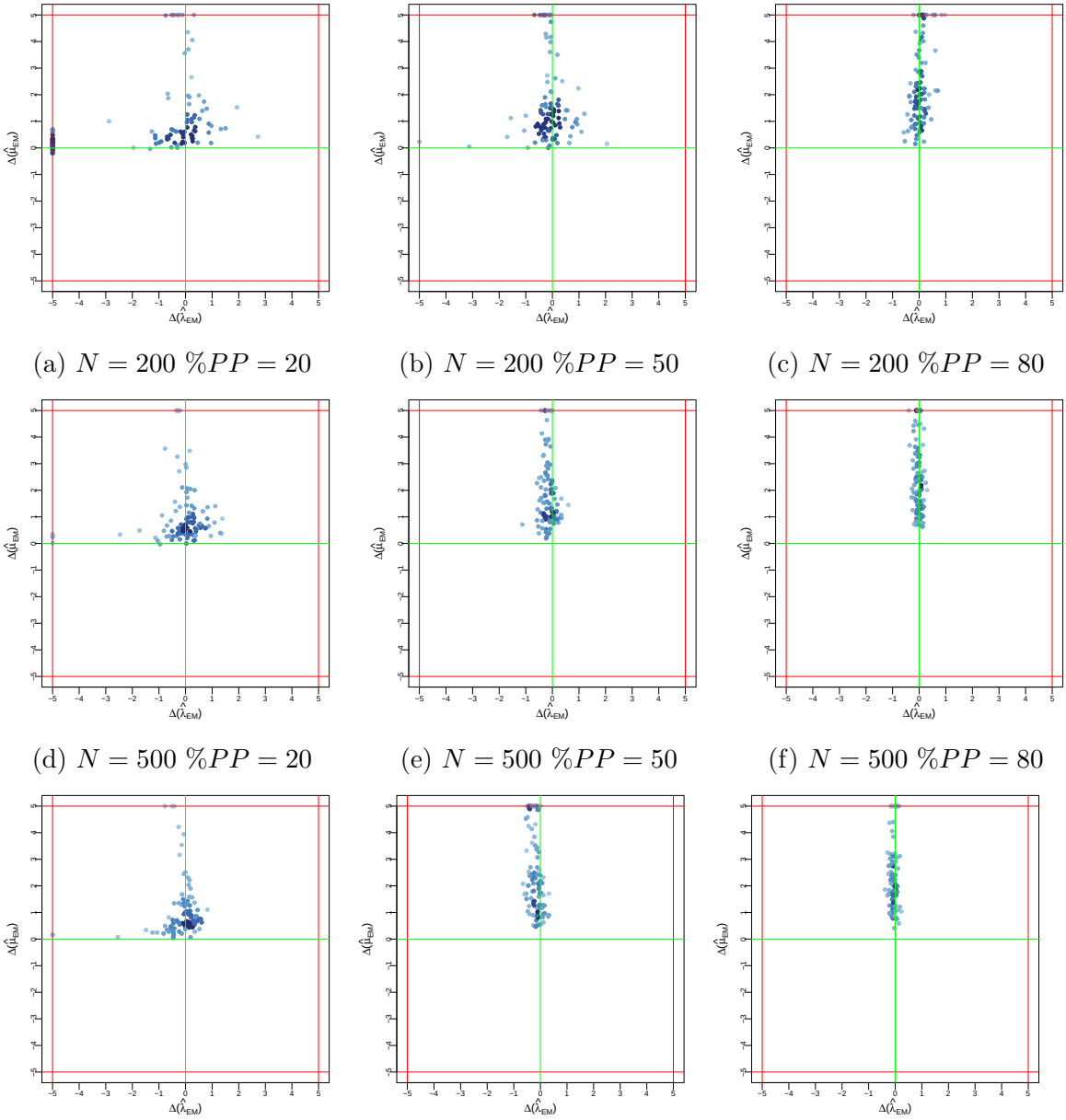
Figura 3.6: Resultado do estimador $\hat{\mu}_{EM}$ para processos simulados

caracteriza a superestimativa de μ e uma boa estimativa para λ_{PP}, respectivamente.

3.4.3 Estimador $\hat{\mu}_{Median}$

Uma proposta alternativa para estimarmos o valor de μ, uma vez que o estimador μ_{EM} não foi satisfatório, é retirar os pontos que hipotéticamente seriam relativos ao processo de Poisson e estimar μ com os pontos que sobrarem, ou seja, os que pertencem ao SFP.

A primeira questão a ser levantada é a maneira de retirar os pontos do processo



(a) $N = 200$ $\%PP = 20$ (b) $N = 200$ $\%PP = 50$ (c) $N = 200$ $\%PP = 80$

(d) $N = 500$ $\%PP = 20$ (e) $N = 500$ $\%PP = 50$ (f) $N = 500$ $\%PP = 80$

(g) $N = 800$ $\%PP = 20$ (h) $N = 800$ $\%PP = 50$ (i) $N = 800$ $\%PP = 80$

Figura 3.7: $\Delta(\hat{\lambda}_{EM})$ versus $\Delta(\hat{\mu}_{EM})$ para diversos valores de $(N, \%PP)$

Tabela 3.2: *Heatmap* de $\Delta(\hat{\mu}_{EM})$ em função de %PP versus N

N \ %PP	10%	20%	30%	40%	50%	60%	70%	80%	90%
100	0.6185	0.6036	0.7302	0.9093	1.1030	1.5767	1.7540	2.1718	9.0642
	1.3720	1.3273	2.1571	2.5875	2.2996	3.3542	4.1349	11.2539	31.0799
200	0.5411	0.5514	0.9898	1.1028	1.1076	1.4922	1.9073	2.0931	2.3148
	0.9747	1.1434	1.8070	2.1541	1.8998	3.2764	3.0628	3.9261	13.1487
300	0.3364	0.5970	1.0284	1.1778	1.4298	1.6374	1.9960	1.8905	1.6651
	0.7829	1.5364	2.0487	2.0344	2.5746	2.5375	3.4948	2.9212	10.5020
400	0.3825	0.5495	1.0040	1.1973	1.4580	1.8669	1.9826	1.8125	1.7488
	0.6548	0.9744	1.7155	2.3809	2.8313	3.1475	2.7784	3.2785	3.6432
500	0.4573	0.6220	1.0196	1.4847	1.5559	1.6701	2.0168	2.3017	2.1915
	0.7519	1.2091	2.0625	2.9246	2.5743	2.6806	3.3949	3.6356	4.3368
600	0.3924	0.6963	0.9031	1.4567	1.7124	1.8388	2.2002	1.9869	2.1034
	0.6861	1.4488	1.7391	2.2068	2.9842	3.4858	3.4485	3.4167	4.9023
700	0.3911	0.6651	1.0841	1.5240	1.7349	1.6949	1.7773	1.9445	1.8036
	0.5692	1.2062	2.3552	2.7439	3.1839	2.7710	2.7125	3.2006	2.7897
800	0.4820	0.6754	1.2247	1.5993	2.0329	2.0999	2.0118	1.9971	1.9287
	0.7653	1.2748	2.5419	2.3041	3.4389	3.6027	2.9968	2.8461	4.9572
900	0.4067	0.6488	1.1606	1.3531	1.8906	2.2112	2.2176	2.0253	1.7845
	0.6373	1.0609	1.9780	2.2784	3.2348	3.3990	3.1640	3.1482	3.0277
1000	0.3589	0.7308	1.1337	1.4720	1.7132	2.1875	2.2829	2.2105	1.9004
	0.5853	1.1763	1.7099	2.5152	2.7542	3.7449	3.5734	3.2417	3.9819

LEGENDA:

Mediana	-1	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	1
|Percentil|*	1	0.8	0.6	0.4	0.2	0					

*Percentil 25 ou 75, o que tiver maior módulo

de Poisson da mistura. O método proposto neste trabalho foi gerar uma sequência aleatória de um processo de Poisson, no intervalo observado, com intensidade $\hat{\lambda}_{EM}$ e retirar os pontos da mistura mais próximos dos pontos do processo de Poisson gerado, respeitando um limite de $2 \times 1/\lambda_{EM}$. Este limite se faz necessário para evitar que retiremos pontos exclusivamente de *burts* na mistura no caso de uma má estimativa de λ_{PP}: apenas cerca de 4% dos intervalos de Poisson são superiores à $2 \times 1/\lambda_{PP}$. Caso nenhum ponto da mistura se aproxime suficientemente de um determinado ponto do processo de Poisson gerado este ponto é desconsiderado. Além disso, vale ressaltar que a retirada é sem reposição, ou seja, uma vez que o ponto é retirado da mistura ele não retorna mais.

De posse do processo SFP puro podemos estimar o parâmetro μ pela mediana dos tempos entre eventos. A este estimador designamos $\hat{\mu}_{Median}$. Esta medida foi selecionada por ser robusta e sofrer menos influência da aleatoriedade do método proposto para retirada dos pontos provenientes do processo de Poisson. Os resultados podem ser visualizados na Figura 3.8 e na Tabela 3.3.

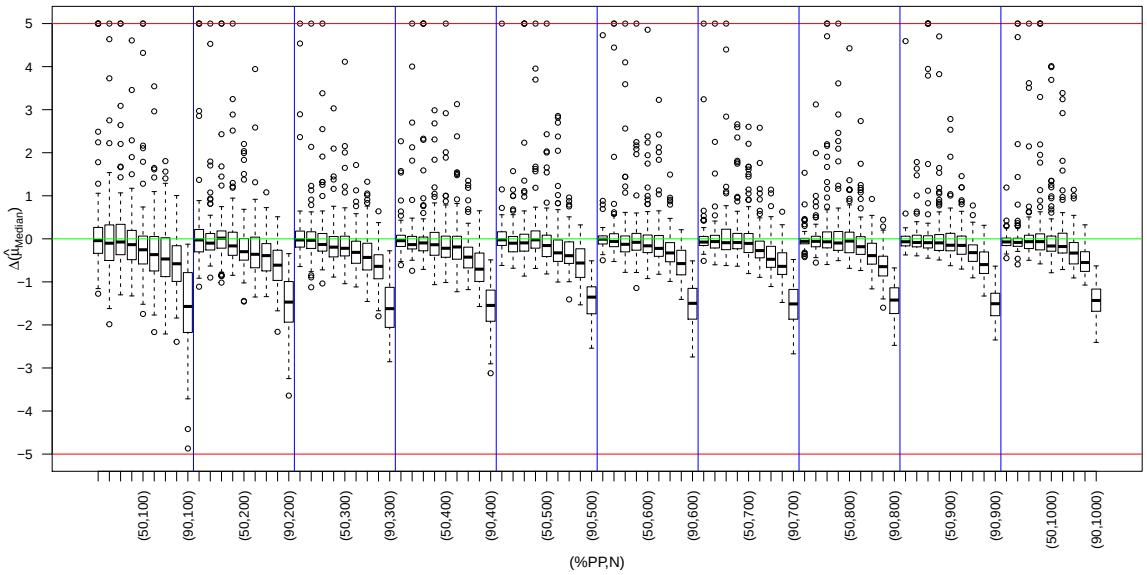
Podemos perceber que o $\hat{\mu}_{Median}$ se comporta melhor que o $\hat{\mu}_{EM}$, subestimando levemente o valor de μ e variando em torno de seu valor real. Como presumível, a estimativa é melhor quando a porcentagem de SFP é alta e também melhora à medida que o número de pontos da amostra aumenta.

Já quando a porcentagem de SFP é baixa, o fato de μ_{Median} subestimar o valor de μ pode ser explicado por uma maior sensibilidade à retirada estocástica dos pontos do PP. Devido a presença de poucos pontos SFP, e consequente diminuição da robustez da mediana, os pontos retirados da mistura se tornam mais relevantes. Por este motivo, podemos notar também maior variabilidade no resultado do estimador.

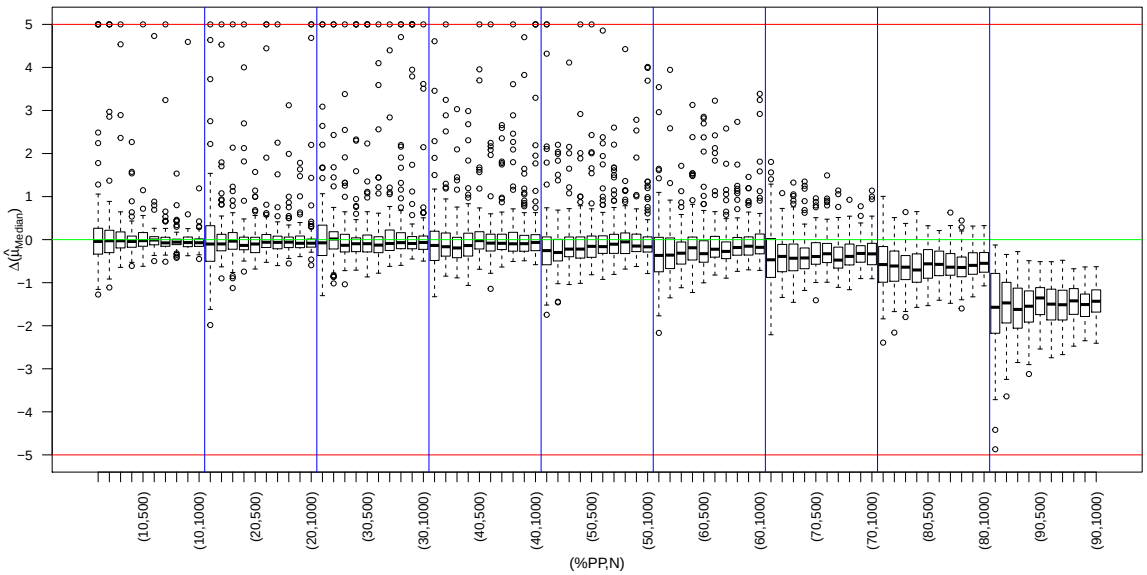
Na Figura 3.9 podemos analisar gráficos de simulações considerando $\Delta(\hat{\lambda}_{EM})$ versus $\Delta(\hat{\mu}_{Median})$ para diversos valores de $\%PP$ e N. O comportamento conjunto corrobora com o comportamento individual descrito nesta seção e na Seção 3.4.1. Verificamos, em geral, *clusters* concentrados próximos da origem. Isto nos diz que os estimadores estão convergindo para valores próximos dos reais de λ_{PP} e μ, conforme desejado.

3.5 Teste de hipótese das simulações

O teste de hipótese utilizado nesta dissertação foi o teste da razão do máximo da função de verossimilhança, descrito na Seção 2.6. Desejamos verificar se a hipótese H_0 de que um modelo mais simples Θ_0 descreve tão bem o processo estocástico quanto o modelo de mistura. Como modelos mais simples vamos considerar alternadamente duas possibilidades: H_0 sendo o modelo PP puro e H_0 sendo o modelo SFP puro. Estas duas situações descritas são casos particulares de uma mistura fixando os parâmetros $\mu = \infty$ e $\lambda_{PP} = 0$, respectivamente. Uma outra ótica para este problema é verificarmos se o parâmetro livre adicional no modelo mais complexo é realmente necessário ou se o modelo mais simples descreve os dados observados igual ou melhor que o modelo de mistura, dispensando o possível parâmetro adicional. Para isto consideramos que se uma das estatísticas descritas em (3.16) e (3.17) for superior a 0.05, o modelo mais complexo, o de mistura, não é necessário e selecionamos o mais simples com maior estatística de aceitação (ϕ). Em ambos os casos o número de parâmetros livres é igual

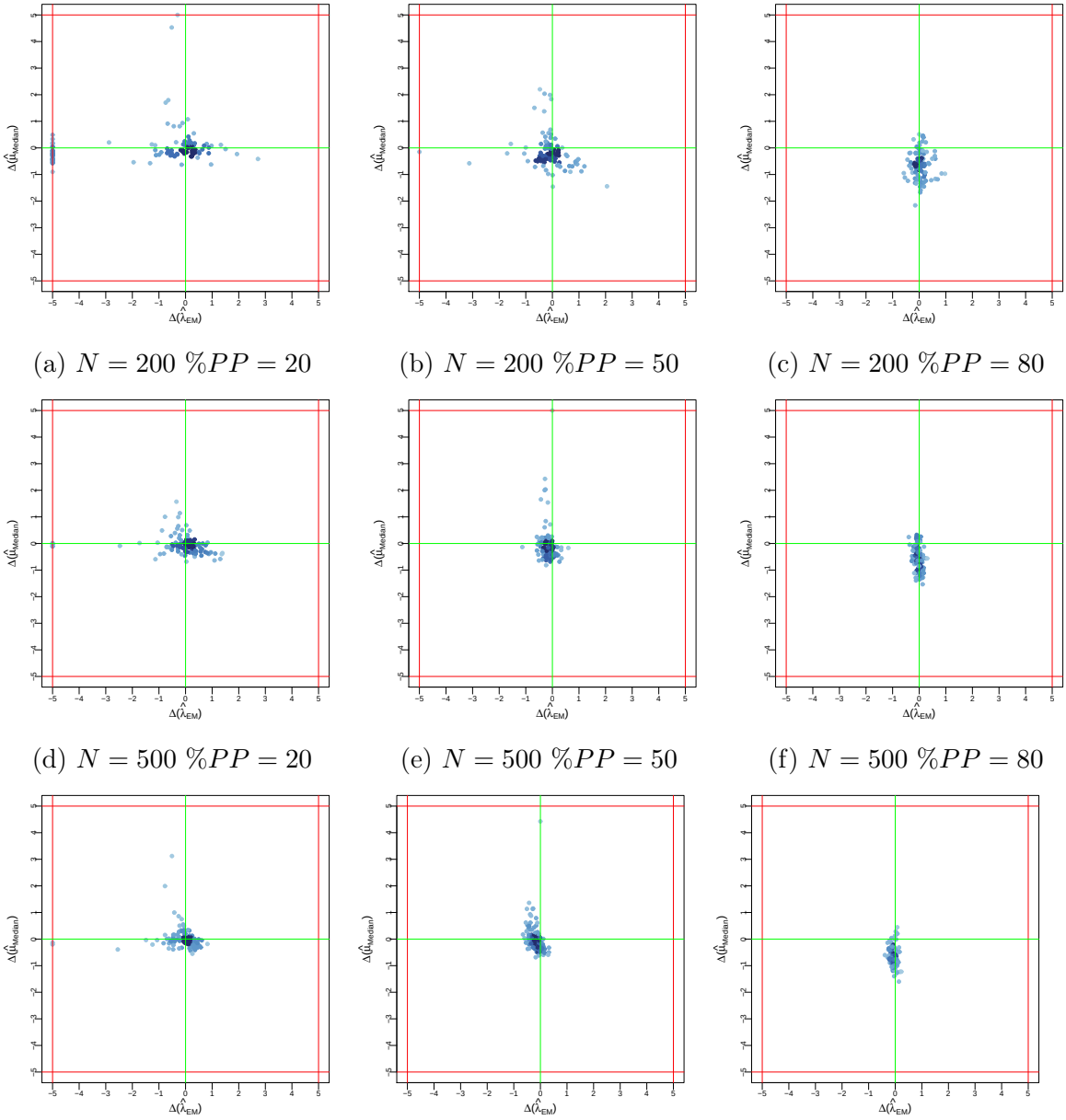


(a) Agrupados por N



(b) Agrupados por $\%PP$

Figura 3.8: Resultado do estimador $\hat{\mu}_{Median}$ para processos simulados



(a) $N = 200$ $\%PP = 20$ (b) $N = 200$ $\%PP = 50$ (c) $N = 200$ $\%PP = 80$

(d) $N = 500$ $\%PP = 20$ (e) $N = 500$ $\%PP = 50$ (f) $N = 500$ $\%PP = 80$

(g) $N = 800$ $\%PP = 20$ (h) $N = 800$ $\%PP = 50$ (i) $N = 800$ $\%PP = 80$

Figura 3.9: $\Delta(\hat{\lambda}_{EM})$ versus $\Delta(\hat{\mu}_{Median})$ para diversos valores de $(N, \%PP)$

Tabela 3.3: *Heatmap* de $\Delta(\hat{\mu}_{Median})$ em função de %PP versus N

N \ %PP	10%	20%	30%	40%	50%	60%	70%	80%	90%
100	-0.0417	-0.1016	-0.0738	-0.1352	-0.2542	-0.3663	-0.4681	-0.5789	-1.5712
	0.3352	0.4946	0.3593	0.4771	0.5768	0.7405	0.8754	0.9883	2.1744
200	-0.0258	-0.0998	0.0150	-0.1644	-0.2987	-0.3594	-0.3897	-0.6124	-1.4700
	0.3020	0.2562	0.2101	0.3803	0.4928	0.6363	0.7417	0.9620	1.9297
300	-0.0266	-0.0347	-0.1317	-0.1933	-0.2198	-0.3148	-0.4322	-0.6398	-1.6215
	0.1917	0.2197	0.2828	0.4109	0.3928	0.5548	0.7217	0.9128	2.0507
400	-0.0429	-0.1343	-0.0950	-0.1371	-0.2192	-0.1888	-0.4254	-0.7046	-1.5468
	0.1713	0.2341	0.2692	0.3719	0.4239	0.4712	0.6747	0.9874	1.8991
500	-0.0233	-0.1015	-0.0943	-0.0236	-0.1561	-0.3257	-0.3918	-0.5621	-1.3548
	0.1684	0.3010	0.2725	0.2096	0.4071	0.5244	0.5619	0.8812	1.7309
600	-0.0089	-0.0605	-0.1274	-0.0810	-0.1621	-0.2224	-0.3278	-0.5749	-1.4985
	0.1184	0.1820	0.3024	0.2617	0.3186	0.4056	0.5314	0.8329	1.8642
700	-0.0754	-0.0634	-0.0895	-0.0852	-0.1058	-0.2736	-0.4755	-0.6398	-1.5120
	0.1547	0.2091	0.2483	0.2286	0.3155	0.4273	0.6561	0.8228	1.8618
800	-0.0578	-0.0579	-0.0693	-0.0966	-0.0523	-0.1826	-0.3897	-0.6463	-1.4224
	0.1101	0.1847	0.2079	0.2675	0.3206	0.3585	0.5826	0.8597	1.7265
900	-0.0678	-0.0844	-0.0894	-0.0940	-0.1489	-0.1529	-0.3208	-0.5984	-1.5073
	0.1630	0.1817	0.2075	0.2689	0.2774	0.3577	0.5206	0.8027	1.7849
1000	-0.0718	-0.0828	-0.0658	-0.0653	-0.1666	-0.1786	-0.3307	-0.5501	-1.4317
	0.1565	0.1719	0.2251	0.2557	0.2836	0.3280	0.5739	0.7538	1.6792

LEGENDA:

Mediana	-1	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	1
|Percentil|*	1	0.8	0.6	0.4	0.2	0					

*Percentil 25 ou 75, o que tiver maior módulo

a um, portanto:

$$\phi(\Theta_{PP}) = 1 - \text{pvalor } \chi^2{}_1 \left(2 \times \log \frac{\max L_{\Theta_{MISTURA}}(\theta|Y)}{\max L_{\Theta}(\theta_{PP}=\{\hat{\lambda}_{PP},\mu=\infty\}|Y)}\right) \tag{3.16}$$

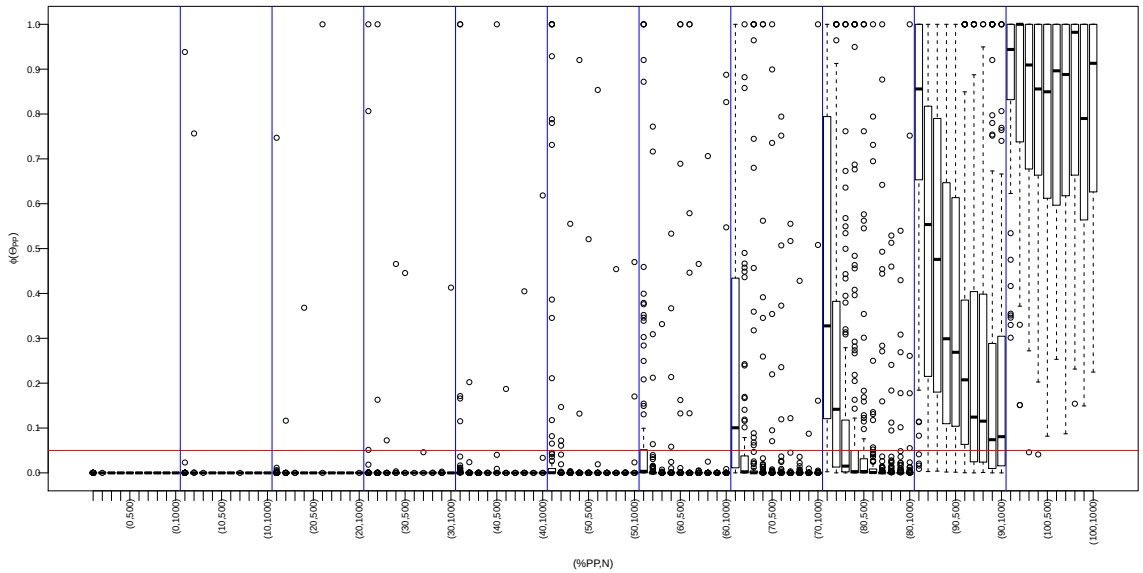
$$\phi(\Theta_{SFP}) = 1 - \text{pvalor } \chi^2{}_1 \left(2 \times \log \frac{\max L_{\Theta_{MISTURA}}(\theta|Y)}{\max L_{\Theta}(\theta_{SFP}=\{\lambda_{PP}=0,\hat{\mu}\}|Y)}\right) \tag{3.17}$$

Como não podemos igualar $\mu = \infty$ no denominador de (3.16), nós consideramos que igualar o valor de μ ao tamanho do intervalo suficientemente grande para o teste de hipótese. Esta consideração corrobora com os limites propostos para busca de valores possíveis para os parâmetros, discutidos na Seção 3.3.

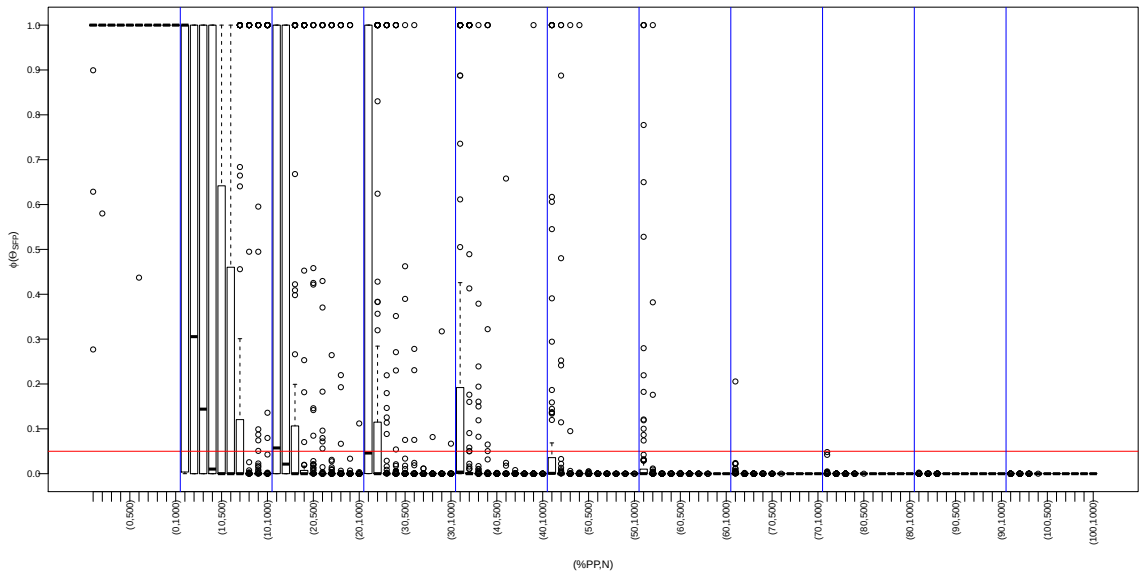
Com o intuito de verificarmos a eficiência deste teste, foram realizadas 2000 simulações adicionais, sendo 1000 provenientes do processos de Poisson puro e as outras 1000 provenientes do SFP puro. Em seguida, calculamos (3.16) e (3.17) para todas as simulações realizadas. Os resultados podem ser analisados nos boxplots da Figura 3.10. As linhas vermelhas representam o valor 0.05.

O comportamento do teste de hipótese vai de acordo com o disticutido nas Seções 3.4.1 e 3.4.2. Em alguns casos, o EMV nos leva a crer que se trata de um processo puro, quando a porcentagem de pontos provenientes de um dos processos é baixa. Estas ocorrências explicam as misturas consideradas processos Poisson: *outliers* e boxplots à medida que aumentam a porcentagem de pontos Poisson, da esquerda para a direita da Figura 3.10a. Situação recíproca é encontrada para o SFP na Figura 3.10b.

Conjuntamente, $\phi(\Theta_{PP})$ e $\phi(\Theta_{SFP})$ estão representadas na Figura 3.11. Nesta figura as linhas vermelhas representam o valor de 0.05. Observamos que não houve casos próximos da linha preta, $f(x) = x$, ou seja, o teste não atribuiu probablidades iguais para SFP e Poisson. O teste da razão de verossimilhança mostrou-se eficaz para verificar as hipóteses pretendidas.



(a) H_0 sendo o modelo Processo de Poisson Homogêneo puro



(b) H_0 sendo o modelo SFP puro

Figura 3.10: Testes de hipóteses aplicados em processos simulados

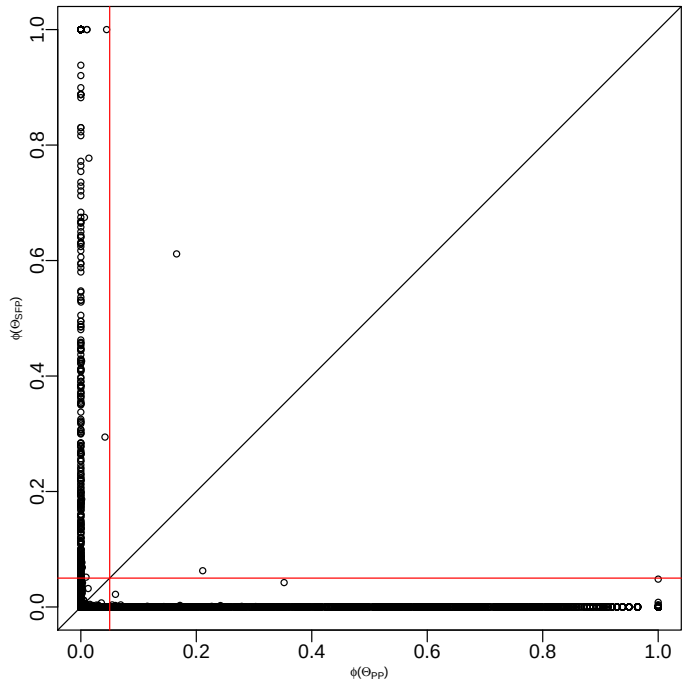


Figura 3.11: $\phi(\Theta_{PP})$ versus $\phi(\Theta_{SFP})$

Capítulo 4

Resultados

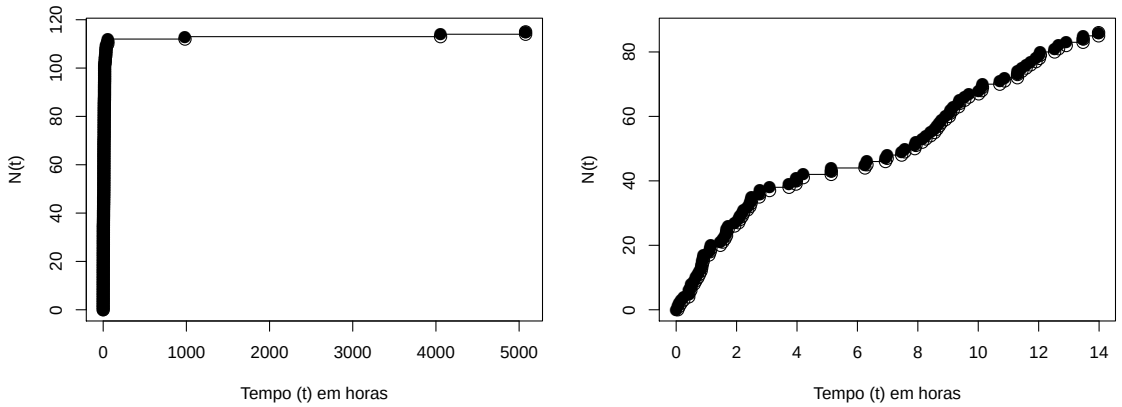
4.1 Descrição dos dados

Neste trabalho foram utilizados dados coletados de nove conjuntos, que foram divididos em quatro grupos distintos. O primeiro grupo contém eventos que são comentários em tópicos de diversos serviços da *Web*, tais como fóruns de discussão(*AskMe*, *MetaFilter*, *MetaTalk*) e sistemas de recomendação colaborativos (*Digg*, *Reddit*). O segundo grupo, por sua vez, contém os eventos de comunicações entre usuários: envios e recebimentos de emails (*Enron*) e bate-papos através de *twetts* com utilização de *hashtags* (*Twitter*). O terceiro grupo possui eventos de um sistema de controle de versão de projetos de software (*Github*). No quarto grupo os eventos são críticas a um restaurante cadastrado em um serviço de recomendação deste tipo de estabelecimento (*Yelp*).

Algumas destas bases são públicas e foram disponibilizadas na *Internet* para *download*. Os dados das bases *AskMe*, *MetaFilter* e *MetaTalk* foram publicados pelo *Metafilter Infodump Project2*[1]. De modo semelhante a base *Digg*[2] foi baixada, mas não se encontra mais disponível para *download*. Os eventos de *Enron* foram disponibilizados pelo *CALO Project (A Cognitive Assistant that Learns and Organizes)*[3] da Universidade Carnegie Mellon. Os dados do *Yelp*[4] foram retirados de um desafio proposto pela empresa proprietária do serviço denominado *Yelp Dataset Challenge*.

Os dados provenientes do *Reddit* e do *Twitter* foram coletados utilizando as respectivas APIs construídas pelas empresas proprietárias dos serviços. Finalmente, os eventos da base *Github* foram adquiridos do arquivo público do serviço e represen-

[1] `http://stuff.metafilter.com/infodump/` - Acessado em Setembro de 2013
[2] `http://www.infochimps.com/datasets/diggcom-data-set` - Acessado em Setembro de 2013
[3] `https://www.cs.cmu.edu/~./enron/` - Acessado em Setembro de 2013
[4] `http://www.yelp.com/dataset_challenge` - Acessado em Agosto de 2014



(a) 100% dos dados observados (b) 75% dos dados observados

Figura 4.1: Representação do indivíduo 29665 da base AskMe

tam linhas de tempo públicas de projetos e de pessoas. Estes dados foram cedidos ao trabalho por pesquisadores parceiros da UFMG.

Em uma primeira análise das bases de dados percebemos um comportamento comum e bem particular em indivíduos de algumas delas: uma alta concentração de eventos em uma pequena faixa de tempo quando considerado todo o intervalo observado. Este cenário apresenta dois fatores desinteressantes à abordagem deste trabalho. Primeiro, tais indivíduos praticamente não possuem eventos do processo de Poisson na mistura, o que o torna um SFP puro. Este cenário já foi amplamente estudado em Vaz de Melo et al. [2013] e Vaz de Melo et al. [2014]. Segundo, indivíduos assim são provavelmente comunicações sobre um tópico que já foi encerrado. Isso é muito comum em fóruns de discussão em que o tópico é criado, discutido por um curto período e encerrado. Caso o assunto volte a ser tratado, um novo tópico é aberto e, na maioria das vezes, sem ligação com o anterior. Portanto, observamos que, quando ocorre tal comportamento, eles não descrevem como o tópico se comportou na maior parte dos eventos. Um exemplo deste comportamento está representado na Figura 4.1. Aparentemente, com 100% dos dados observados, apenas duas situações distintas podem ser consideradas: uma grande rajada de eventos durante as primeiras 20 horas seguidas por um enorme período de baixa atividade nas mais de 4000 horas seguintes(Figura 4.1a). Ao censurarmos os 25% últimos eventos, percebemos um comportamento bem distinto. Detectamos que 75% dos pontos encontram-se nas primeiras 14 horas (Figura 4.1b), menos de 1% de todo o intervalo observado. Mais importante, nota-se a presença de um processo de Poisson homogêneo com uma taxa mais significativa para a mistura e também com maior representatividade no momento ativo do tópico.

Indivíduos com características similares as estas são presentes nas seguintes bases:

Tabela 4.1: Estatísticas descritivas das bases de dados consideradas neste trabalho

Base	Número de Indivíduos	Número de Eventos		
		Mínimo	Médio	Máximo
AskMe	490	74	99.30	699
Digg	974	39	90.41	296
Enron	145	55	1,541.35	14258
Github	66159	52	598.84	8775
MetaFilter	8243	72	131.10	4148
MetaTalk	2460	73	151.92	2714
Reddit	102	37	535.43	4706
Twitter	17088	50	969.68	8564
Yelp	1929	50	127.84	1646

AskMe, *Digg*, *MetaFilter*, *MetaTalk* e *Reddit*. Assim, nestas bases consideramos para análise os primeiros 75% dos dados observados. Vale ressaltar que grande parte dos indivíduos destes conjuntos de dados não apresentaram tal comportamento citado. No entanto, para que não haja perda de generalidade, aplicamos o filtro a todos os indivíduos das bases citadas.

Na Tabela 4.1, apresentamos algumas estatísticas descritivas das bases de dados consideradas neste trabalho.

4.2 Resumo do modelo

Nesta seção será proposto um fluxo para os cálculos e testes. Uma representação gráfica simplificada pode ser analisada no fluxograma da Figura 4.2. Apresentaremos um exemplo detalhando as etapas descritas neste fluxograma para um indivíduo específico. O indivíduo selecionado foi o 10019 da base de dados do *Twitter*. A representação deste indivíduo pode ser vista na Figura 4.3a.

Primeiramente, executamos o Algorimo EM e obtemos os estimadores $\hat{\lambda}_{EM}$ e $\hat{\mu}_{EM}$. Isto significa que deixamos os parâmetros λ_{PP} e μ livres. Como resultados obtivemos que $\hat{\lambda}_{EM} = 1.09 \times 10^{-4}$ e $\hat{\mu}_{EM} = 1591.69$ maximizam a função de verossimilhança (Seção 3.2) considerando os tempos entre eventos observados. Para chegar a estes

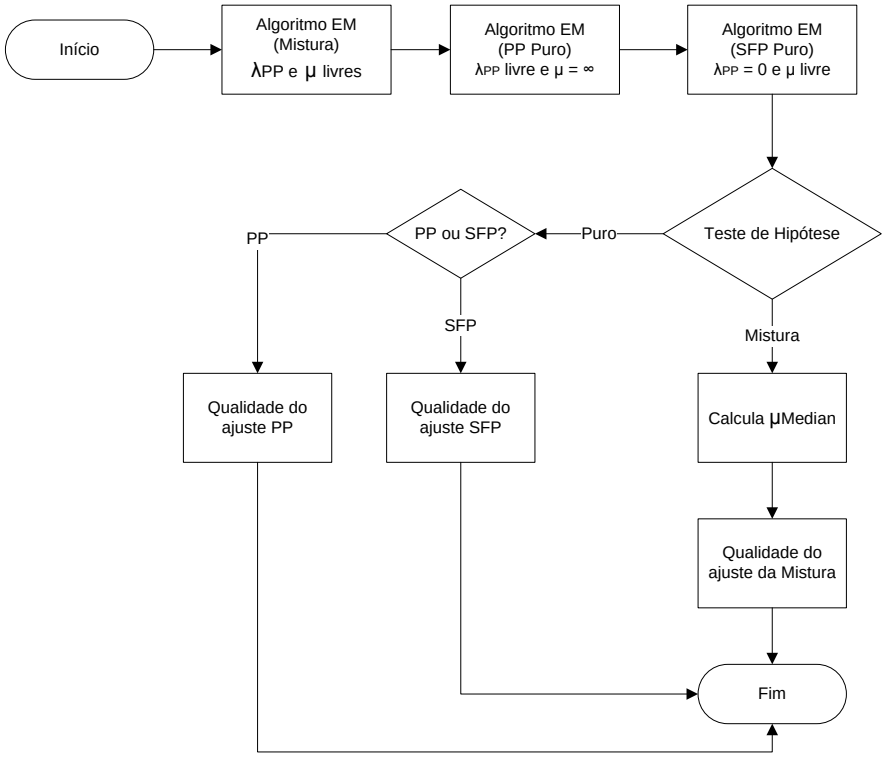


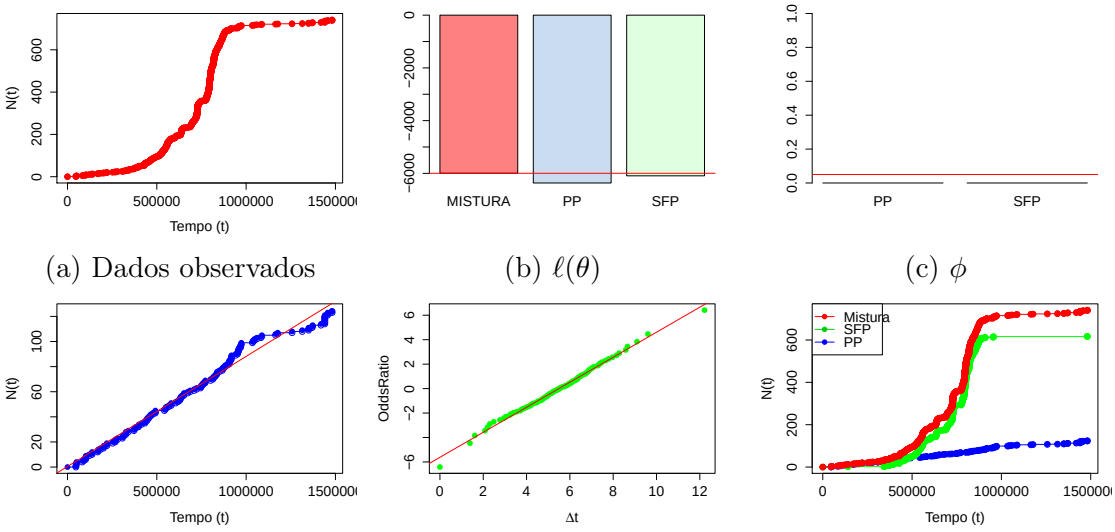
Figura 4.2: Fluxograma resumo do modelo

valores três iterações dos passos E e M foram suficientes.

Igualando o valor de μ a 1481027, tamanho do intervalo observado, encontramos que o valor mais verossímel de λ_{PP} é 5.03×10^{-4}. Como se trata de uma variável livre, um passo do algoritmo é suficiente para encontrar este a valor. Este cenário supõe que os dados são provenientes de um processo de Poisson homogêneo. Uma terceira situação é considerar os dados como um SFP puro. Neste caso igualamos λ_{PP} a zero e obtemos o valor de $\hat{\mu}_{SFP}$ igual a 1978.26. Em todos os casos são utilizadas abordagens do tipo Expectation Maximization para extração dos valores dos estimadores.

Através destas estimativas dos parâmetros, calculamos o valor correspondente da função de verossimilhança (3.3) para cada um dos três cenários considerados (Figura 4.3b). Para saber se o indivíduo analisado é, mais provavelmente, um processo puro ou uma mistura, realizamos o teste de hipótese (Seção 3.5). Observe na Figura 4.3c os valores de $\phi(\Theta_{PP})$ e $\phi(\Theta_{SFP})$ que, para este caso, são inferiores a 0.05 (linha em vermelho) e, portanto, o indivíduo é melhor representado por um processo de mistura.

Caso a hipótese de processo de Poisson fosse a mais provável, o valor de $\phi(\Theta_{PP})$ seria maior igual a 0.05 e como resultado teríamos o valor do parâmetro $\hat{\lambda}_{PP} = 5.03 \times 10^{-4}$. Analogamente, para a hipótese SFP teríamos $\hat{\mu} = 1978.26$



(a) Dados observados (b) $\ell(\theta)$ (c) ϕ

(d) Ajuste PP (e) Ajuste SFP (f) Processos Separados

Figura 4.3: Resumo do modelo para o indivíduo 10019 da base Twitter

e o valor de $\phi(\Theta_{SFP})$ seria maior ou igual 0.05. A qualidade dos ajustes nestes dois casos seriam realizados conforme descrito na Seção 2.2 e na Seção 2.3, respectivamente, considerando todos os pontos do indivíduo para o cálculo. Caso estas duas estatísticas fossem superiores a 0.05 selecionariamos a de maior valor. Persistindo a igualdade, o indivíduo é considerado processo de Poisson, por ser a hipótese mais simples dentre as três.

Como o caso em questão é uma mistura, o próximo passo é calcular o valor de $\hat{\mu}_{Median}$ (Seção 3.4.3). O valor encontrado para este estimador foi 253.55. Portanto o resultado final da estimativa de λ_{PP} e μ da mistura foram 1.09×10^{-4} e 253.55, respectivamente.

Para verificarmos a qualidade do ajuste da mistura necessitaríamos conhecer o rótulo do processo gerador de cada evento. Como isto não é possível, aproximamos as medidas de qualidade retirando pontos provenientes do processo Poisson de modo idêntico ao proposto na Seção 3.4.3: simulamos uma sequência aleatória de um processo de Poisson com intensidade $\hat{\lambda}_{EM}$ e retiramos os pontos da mistura mais próximos dos pontos do processo de Poisson gerado, respeitando um limite de $2 \times 1/\lambda_{EM}$. Caso nenhum ponto da mistura se aproxime suficientemente de um determinado ponto do processo de Poisson gerado este ponto é desconsiderado. Repetimos este processo de separação 100 vezes, retornando os processos cuja soma dos R^2, do PP e do SFP separados, foi a maior encontrada. A qualidade dos ajustes foram 0.98 e 0.99, respectivamente. Os gráficos dos ajustes encontram-se nas Figuras 4.3d e 4.3e. O resultado da separação pode ser visto na Figura 4.3f.

Tabela 4.2: Resultado dos testes de hipóteses por base

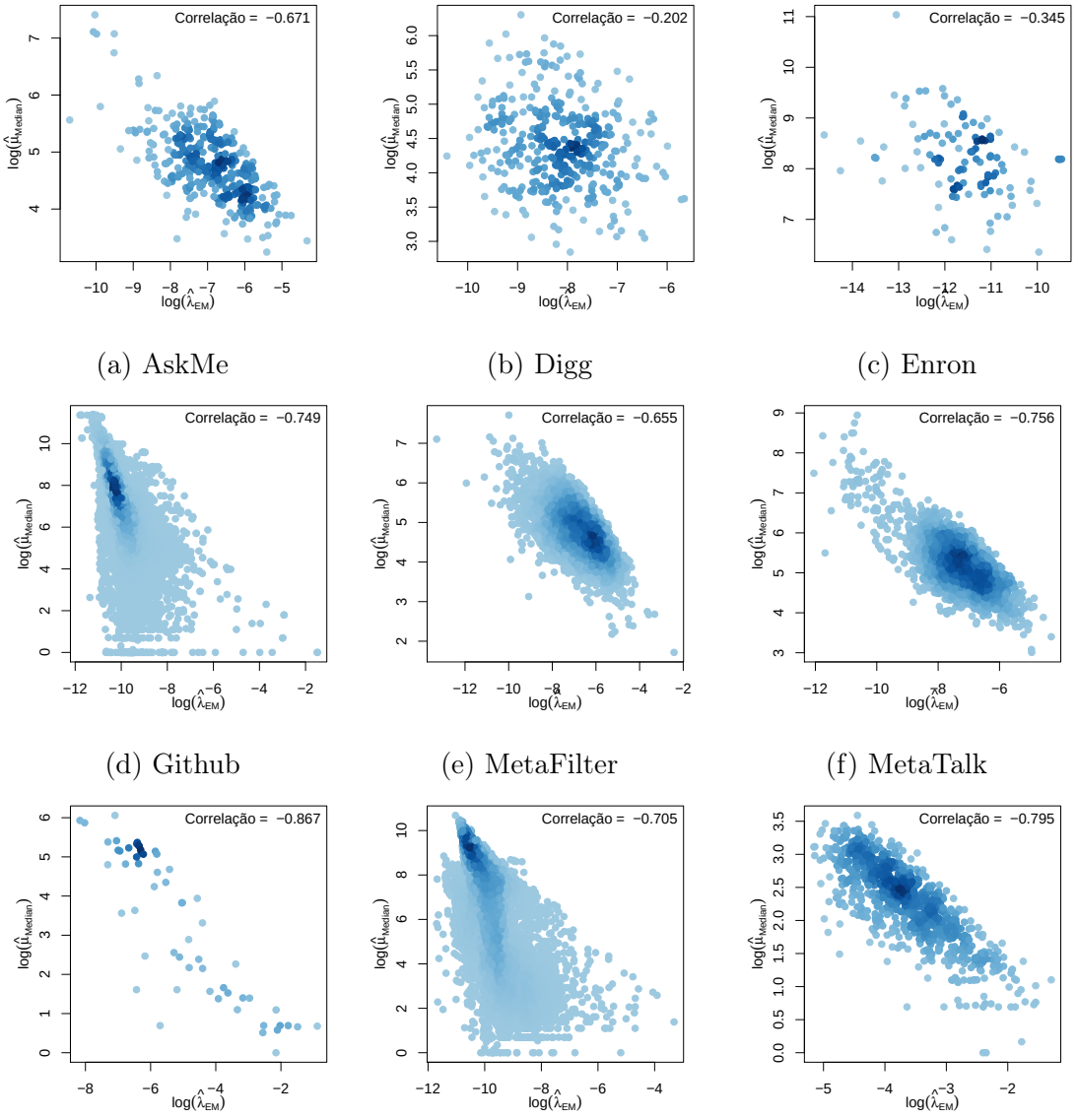
Base	Número de Indivíduos	Teste de Hipótese		
		Mistura	PP	SFP
AskMe	490	333 (67.96%)	43 (8.78%)	114 (23.26%)
Digg	974	353 (36.24%)	2 (0.21%)	619 (63.55%)
Enron	145	106 (73.1%)	0 (0%)	39 (26.9%)
Github	66159	61495 (92.95%)	3650 (5.52%)	1014 (1.53%)
MetaFilter	8243	5625 (68.24%)	1279 (15.52%)	1339 (16.24%)
MetaTalk	2460	1691 (68.74%)	271 (11.02%)	498 (20.24%)
Reddit	102	58 (56.86%)	21 (20.59%)	23 (22.55%)
Twitter	17088	15913 (93.12%)	72 (0.42%)	1103 (6.46%)
Yelp	1929	774 (40.12%)	927 (48.06%)	228 (11.82%)

4.3 Comportamento conjunto dos parâmetros dos processos

Na Figura 4.4 foram plotados os gráficos de $\log \hat{\lambda}_{EM}$ versus $\log \hat{\mu}_{Median}$. Regiões mais escuras significam maior concentração de pontos. Como podemos observar, excetuando *Reddit* e *Twitter*, o aspecto do comportamento conjunto das variáveis dos processos se assemelha a uma normal bivaridada. Por possuir poucos indivíduos, a visualização da base *Reddit* não apresenta padrão claro. Os indivíduos do *Twitter*, por sua vez, parecem formar uma distribuição desconhecida e complexa.

4.4 Testes de hipóteses

Na Tabela 4.2 podemos ver os resultados dos testes de hipóteses para todas as bases consideradas. Na Figura 4.5 é apresentado um gráfico para cada base das estatísticas $\phi(\Theta_{PP})$ versus $\phi(\Theta_{SFP})$. Podemos observar que, para as bases reais, o comportamento foi similar ao demonstrado na Seção 3.5, o que nos leva a crer que os testes de hipóteses propostos foram também suficientes para separação dos modelos para as bases reais.

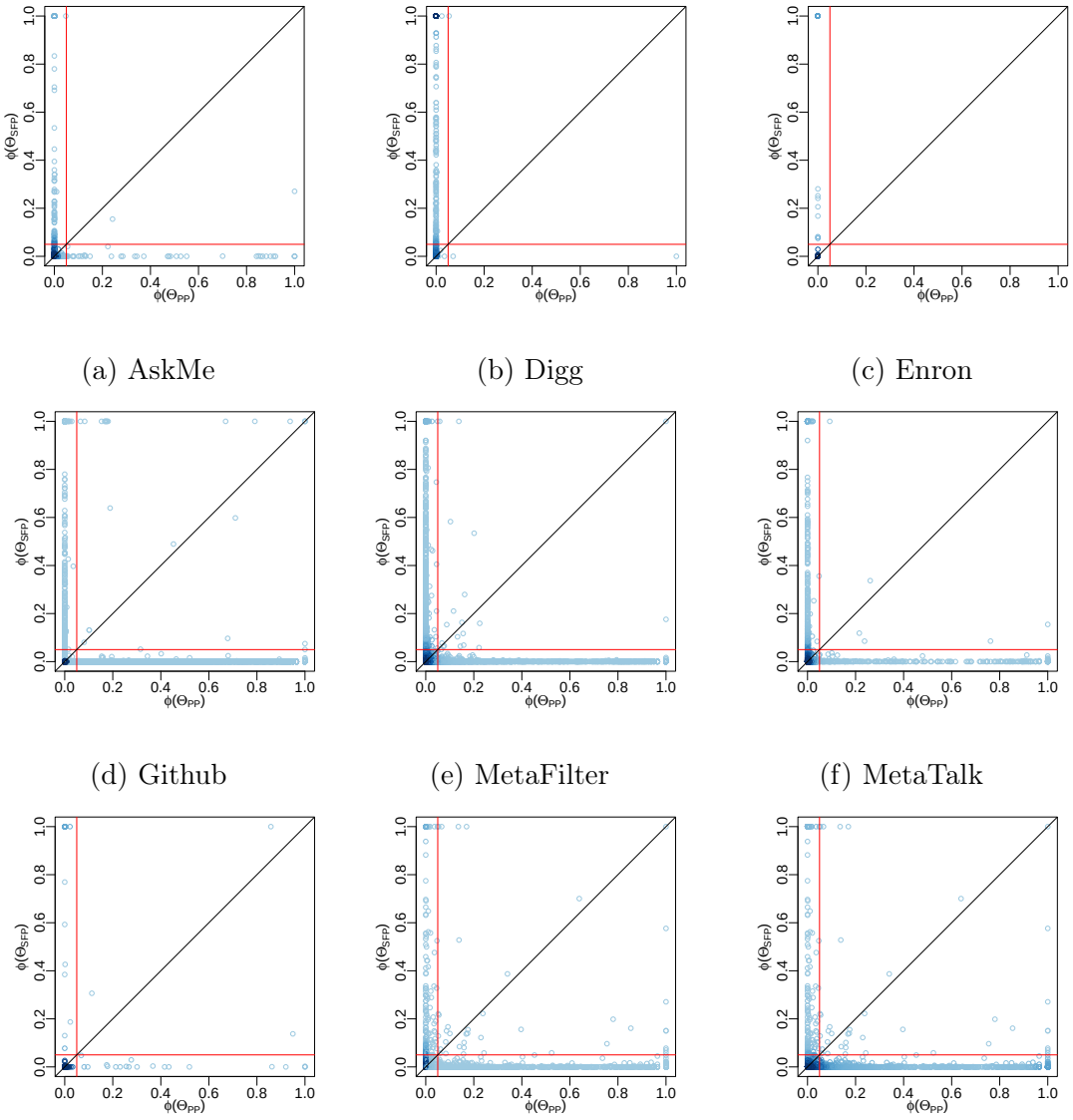


(a) AskMe (b) Digg (c) Enron

(d) Github (e) MetaFilter (f) MetaTalk

(g) Reddit (h) Twitter (i) Yelp

Figura 4.4: Comportamento conjunto dos parâmetros dos processos: $\log(\hat{\lambda}_{EM})$ versus $\log(\hat{\mu}_{Median})$



(a) AskMe (b) Digg (c) Enron

(d) Github (e) MetaFilter (f) MetaTalk

(g) Reddit (h) Twitter (i) Yelp

Figura 4.5: Testes de hipóteses por base: $\phi(\Theta_{PP})$ versus $\phi(\Theta_{SFP})$

4.5 Ajuste dos dados ao Modelo

Nesta seção exibiremos a qualidade dos ajustes aos dados. Os resultados dos ajustes foram obtidos conforme fluxograma da Figura 4.2. Os valores possíveis de R^2 estão contidos no intervalo $[0, 1]$. Quanto mais próximo do valor 1, melhor o ajuste dos dados ao modelo.

4.5.1 Processo de Poisson homogêneo puro

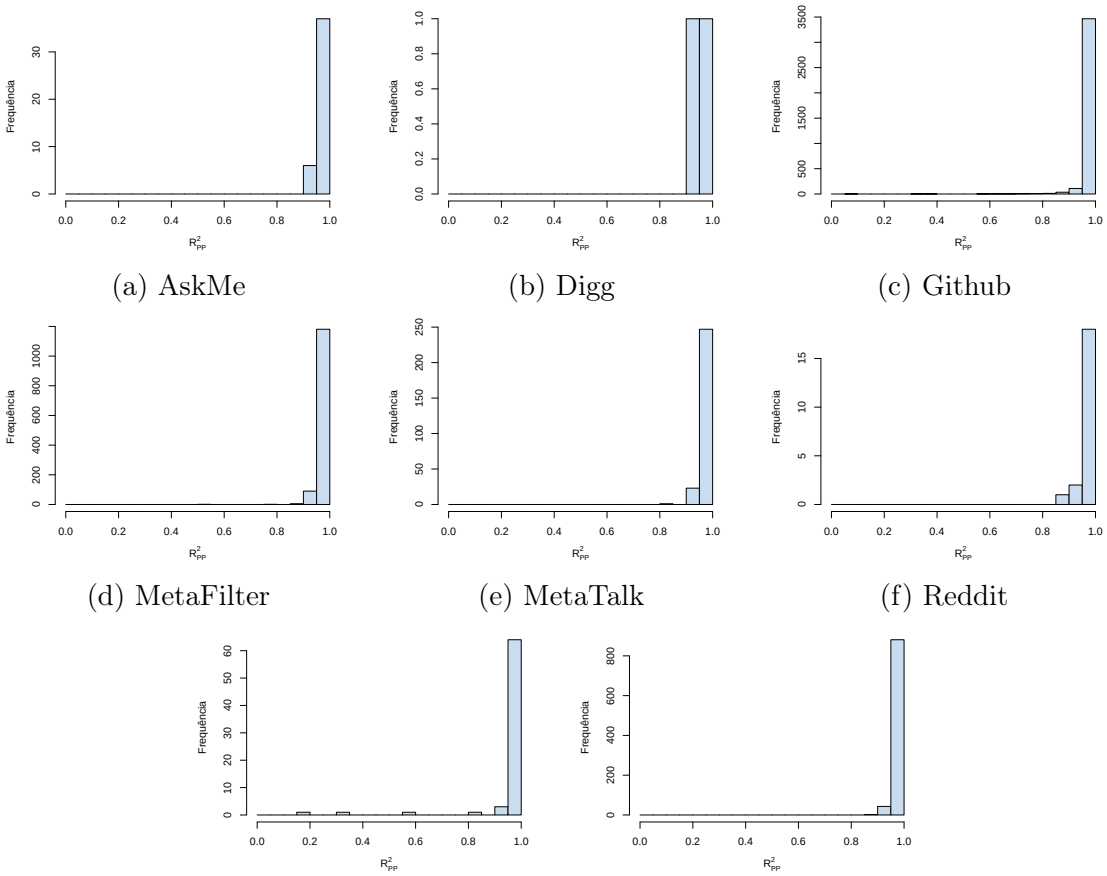
Nesta seção apresentaremos os resultados dos ajustes dos dados naqueles casos em que o teste da razão de máxima verossimilhança levou à aceitação do modelo de Poisson puro. A Figura 4.6 contém histogramas da qualidade do ajuste medido pelo R^2_{PP}. O valor de R^2_{PP} para ajuste teórico perfeito possui valor igual a 1. Nota-se uma grande concentração de valores entre 0.95 e 1.00, com esporádicas variações. Apesar da base *Digg* ser uma exceção, como ela é composta por apenas dois indivíduos PP, não é significativa a sua análise como um processo de Poisson puro, ou seja, seu comportamento é predominantemente de um SFP ou de uma mistura dos dois processos.

No Apêndice A podem ser visualizados um exemplo para cada base de indivíduos considerados processo de Poisson puro.

4.5.2 SFP puro

Nesta seção apresentaremos os resultados dos ajustes dos dados naqueles casos em que o teste da razão de máxima verossimilhança levou à aceitação do modelo de SFP puro. Mais uma vez, o valor da qualidade do ajuste teórico perfeito possui valor igual a 1. A Figura 4.7 contém histogramas da qualidade do ajuste medido pelo R^2_{SFP}. Observe que, para a maioria das bases, há uma grande concentração de valores entre 0.95 e 1.00, com esporádicas variações. Destoam deste comportamento *Github*, *Reddit* e *Twitter*, com considerável número de indivíduos com ajustes inferiores a 0.95. Acreditamos que, na realidade, estes indivíduos não são de nenhuma das três hipóteses consideradas. Vale ressaltar que, para a primeira e a última base, a porcentagem de indivíduos considerados SFP puro é baixa, ou seja, o comportamento padrão de ambas é de mistura de processos.

No Apêndice B podem ser visualizados um exemplo para cada base de indivíduos considerados SFP puro.



(a) AskMe (b) Digg (c) Github

(d) MetaFilter (e) MetaTalk (f) Reddit

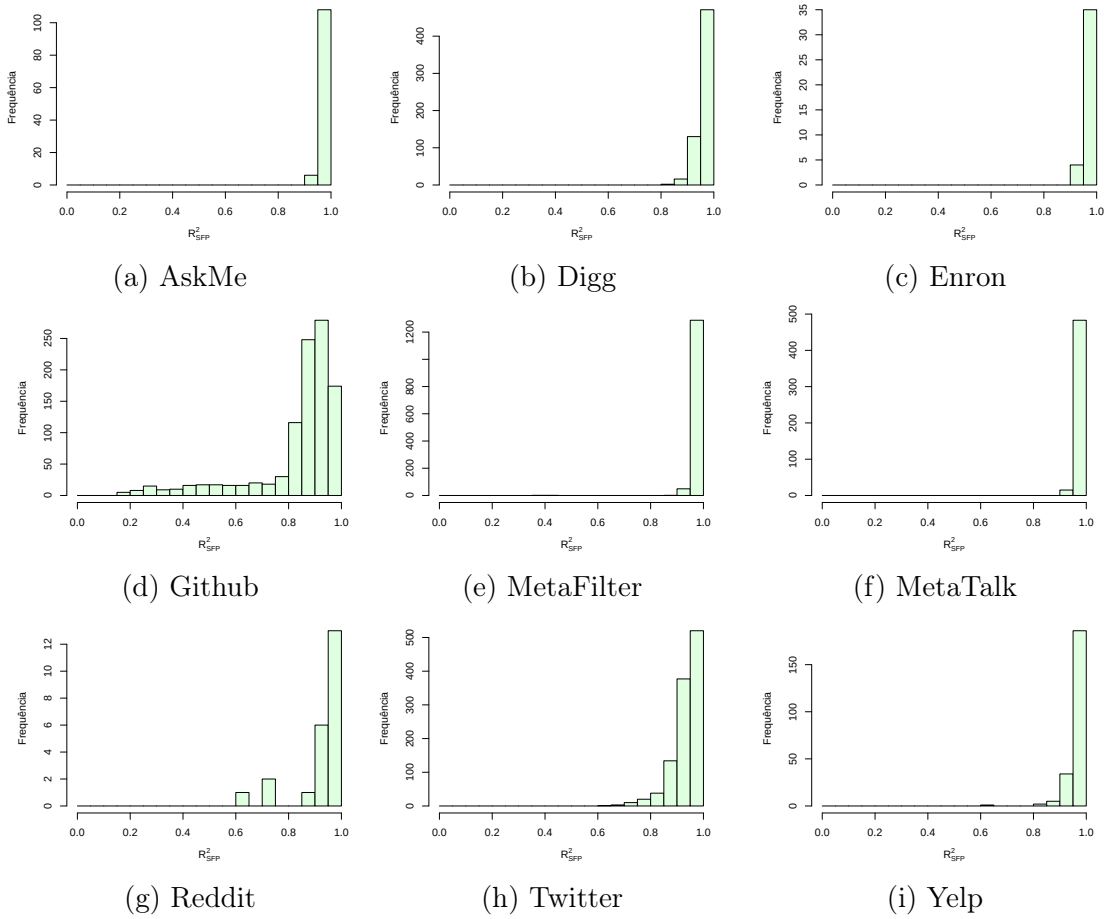
(g) Twitter (h) Yelp

Figura 4.6: Histogramas por base da qualidade dos ajustes dos indivíduos aceitos como modelo PP puro

4.5.3 Modelo de Mistura

Nesta seção apresentaremos os resultados dos ajustes dos dados que são considerados, pelos testes de hipótese, misturas dos processos Poisson homogêneo e SFP. Na Figura 4.8, o resultado apresentado para a base *AskMe* é subdivido em três subfiguras: Figura 4.8a, que é composta por um gráfico de R^2_{PP} versus R^2_{SFP} (as linhas em vermelho significam o valor 0.95) em que é possível observar o comportamento conjunto da qualidade dos ajustes dos processos separados; Figura 4.8b, que é um histograma da qualidade do ajuste do processo Poisson homogêneo separado da mistura; e Figura 4.8c, que é um histograma da qualidade ajuste do SFP separado da mistura. Da Figura 4.9 à Figura 4.16 apresentamos os resultados análogos para as demais bases.

Os valores de R^2_{PP} das misturas estão concentrados entre os valores de 0.95 e 1.00, havendo poucos indivíduos com valores inferiores a este intervalo para todas as bases. A base *Digg* é a que possui maior variabilidade dentre todas as consideradas, porém



(a) AskMe (b) Digg (c) Enron

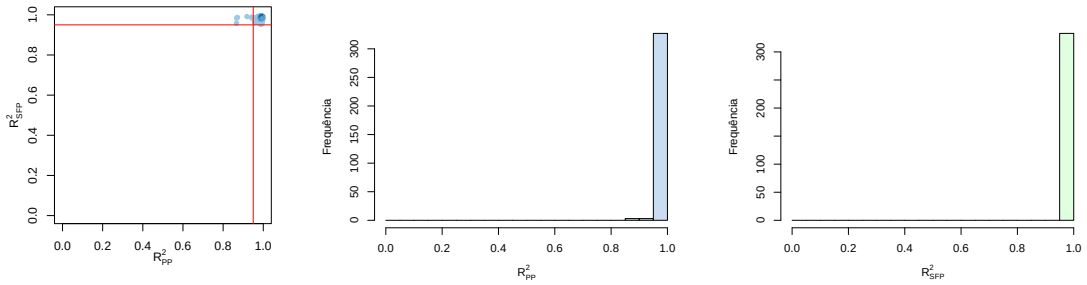
(d) Github (e) MetaFilter (f) MetaTalk

(g) Reddit (h) Twitter (i) Yelp

Figura 4.7: Histogramas por base da qualidade dos ajustes dos indivíduos aceitos como modelo SFP puro

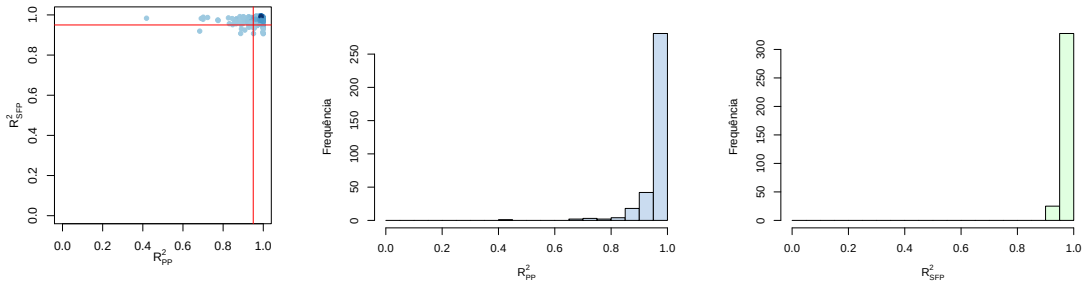
esta variabilidade é pouco significativa. Do mesmo modo, os valores da qualidade dos ajustes dos SFPs pertencentes à mistura possuem maioria dos valores próximos a 1.00, no entanto com maior variação. Esta variação pode ser explicada pelo método estocástico utilizado para separação dos processos. Por se tratar de um método não determinístico, não pode-se garantir que não haja um ajuste melhor que o apresentado. Entretanto, não se descarta que, em alguns casos, o processo gerador de *bursts* que se junta ao processo de Poisson homogêneo seja uma terceira alternativa não tratada neste trabalho.

Na Figura 4.17 pode ser visto, para cada base estudada, um histograma da porcentagem de eventos esperados provenientes do processo de Poisson. Para o cálculo desta porcentagem, consideramos o valor esperado de pontos de um PP com taxa $\hat{\lambda}_{EM}$ no intervalo observado de cada indivíduo e dividimos pelo número de pontos observados. Alguns comportamentos referentes à porcentagem dos eventos provenientes do processo de Poisson merecem destaque: as bases *AskMe*, *Diff*, *MetaFilter*, *MetaTalk*,



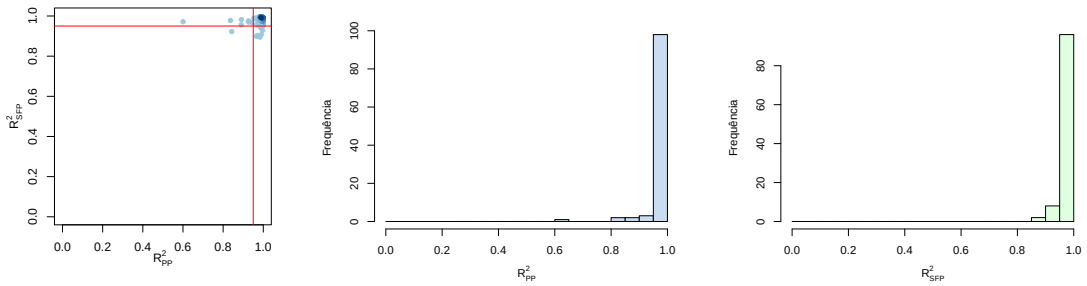
(a) (b) (c)

Figura 4.8: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base AskMe



(a) (b) (c)

Figura 4.9: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base Digg

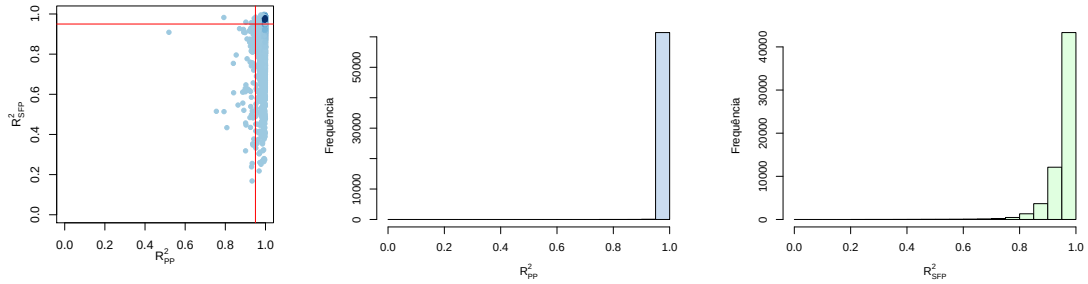


(a) (b) (c)

Figura 4.10: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base Enron

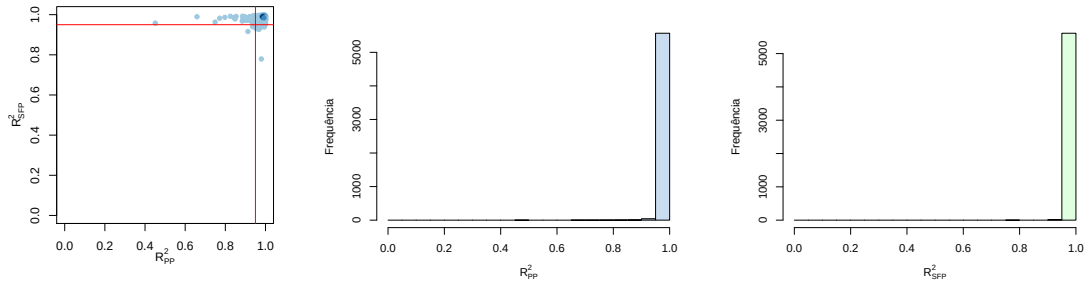
Reddit e *Yelp* possuem uma distribuição semelhante a uma distribuição normal; já as bases *Github* e *Twitter* possuem uma moda bem definida; a base *Enron* possui duas modas bem definidas.

No Apêndice C podem ser visualizados um exemplo para cada base de indivíduos considerados mistura dos processos de Poisson e SFP.



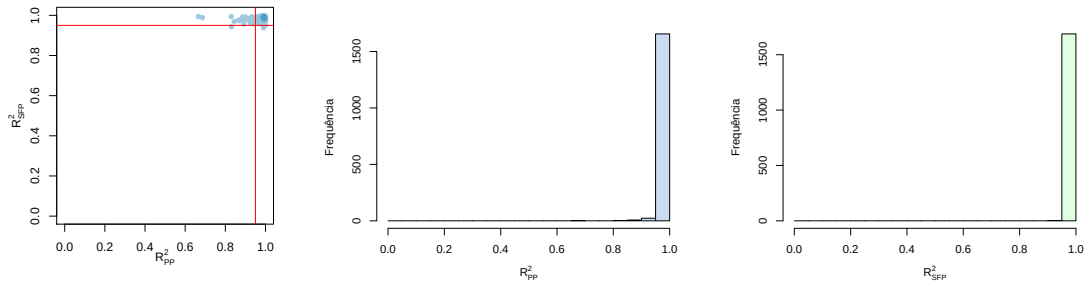
(a) (b) (c)

Figura 4.11: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base Github



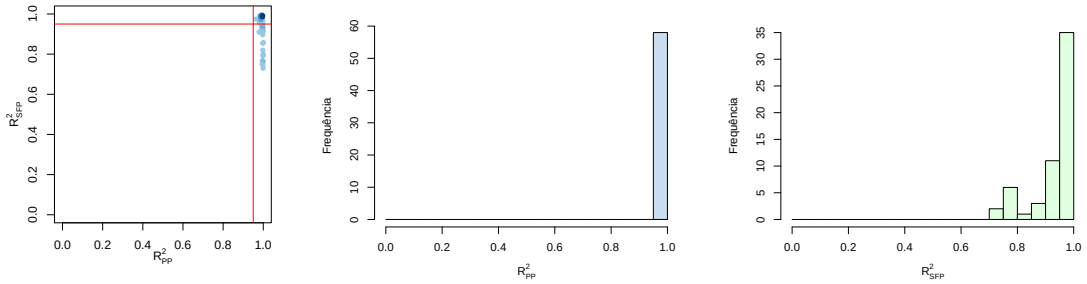
(a) (b) (c)

Figura 4.12: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base MetaFilter



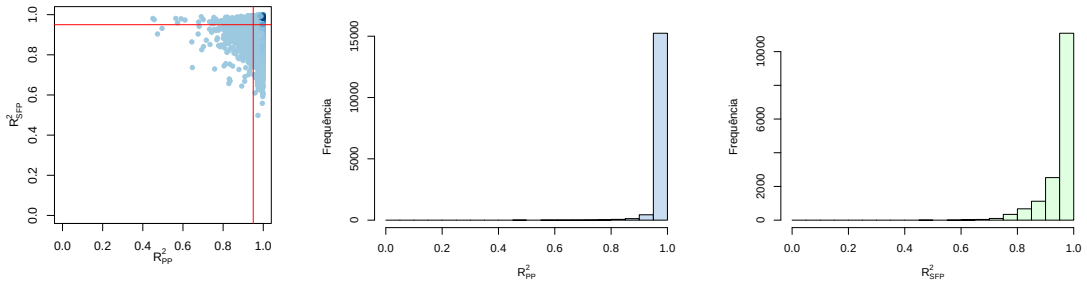
(a) (b) (c)

Figura 4.13: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base MetaTalk



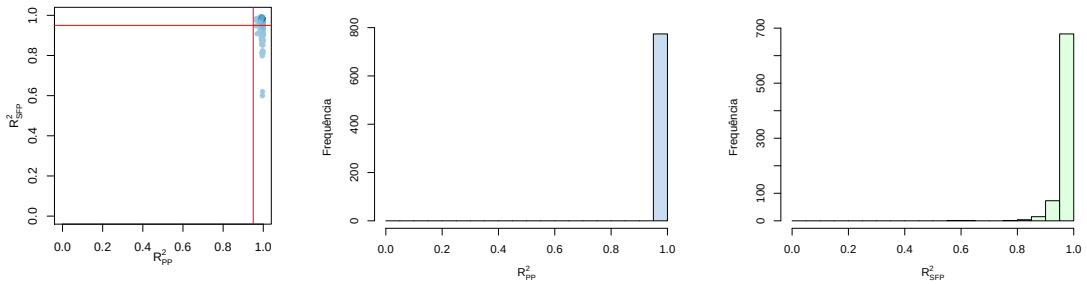
(a) (b) (c)

Figura 4.14: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base Reddit



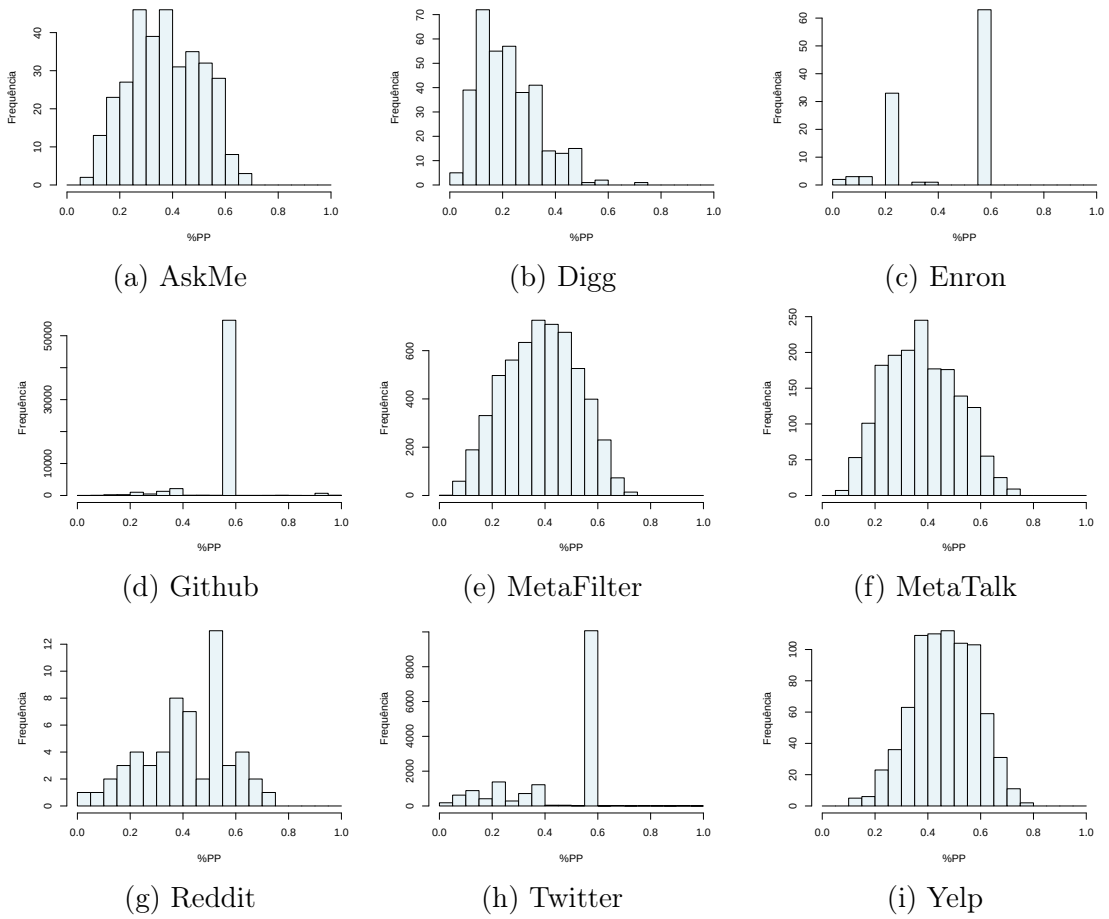
(a) (b) (c)

Figura 4.15: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base Twitter



(a) (b) (c)

Figura 4.16: Qualidade dos ajustes dos indivíduos aceitos como modelo de mistura para a base Yelp



(a) AskMe (b) Digg (c) Enron

(d) Github (e) MetaFilter (f) MetaTalk

(g) Reddit (h) Twitter (i) Yelp

Figura 4.17: Histograma de %PP por base

4.6 Aplicações

Antes de apresentar esta seção, é importante ressaltar que aqui consideramos apenas os indivíduos cujos valores da qualidade dos ajustes foram superiores à 0.95.

4.6.1 Detecção de Anomalias

Com o intuito de detectar possíveis comportamentos anômalos, nesta seção propomos um modelo para ajustar a distribuição do estimador bivariado x=($\log \hat{\lambda}_{EM}$, $\log \hat{\mu}_{Median}$). A partir de observações iniciais, selecionamos a distribuição normal bivariada. Para ajustar os dados à normal bivariada consideramos, para cada conjunto de dados separadamente, $\overline{x}$, que é o valor médio do parâmetro x dos indivíduos, e S, que é a matriz de covariância dos dois parâmetros de x. Posteriormente, calculamos a distância de Mahalanobis entre cada indivíduo e o centro da normal bivariada que foi melhor ajustada. O método para o cálculo desta distância D_M é descrito na Equação 4.1.

$$D_M = \sqrt{(x - \overline{x})^T S^{-1} (x - \overline{x}))} \tag{4.1}$$

Para o caso de dois parâmetros livres, $\log \hat{\lambda}_{EM}$ e $\log \hat{\mu}_{Median}$, D_M é distribuída por uma qui-quadrado com dois graus de liberdade. Assim, consideramos como comportamentos destoante do padrão aqueles indivíduos que possuem baixa probabilidade de pertencer à normal bivariada ajustada para o conjunto de parâmetros do seu grupo. A conjectura é que se a distância D_M do individuo é um valor cuja probabilidade do mesmo pertencer a uma normal com parâmetros $\overline{x}$ e S é inferior à 1%, ou seja, que atenda a inequação descrita na Equação 4.2, este possuirá um comportamento anômalo com relação aos demais indivíduos. As anomalias detectadas foram destacadas como pontos vermelhos na Figura 4.19.

$$D_M > \text{qvalor}{\chi^2}_2(0.99) \tag{4.2}$$

A fim de investigar que tipo de indivíduos foram identificados como anômalos, analisamos os tópicos que os mesmos tratavam. O indivíduo 219940 da base *AskMe*, por exemplo, possui o menor valor de $\hat{\lambda}_{EM}$ dentre todos os indivíduos de seu grupo. Quando analisamos o assunto do mesmo, constatamos que é relacionado à perda de um animal de estimação da esposa do criador do tópico. Ele pede ajuda para encontrá-lo aos demais membros do fórum. Considere que o comportamento padrão de um indivíduo desta base de dados é de uma grande rajada de eventos no início seguida por um período de taxa constante similar a um processo de Poisson, com possíveis *burts* de intensidades menores ocasionados por algum determinado comentário. Posteriormente a taxa de eventos do tópico decresce até que o mesmo é fechado e, caso o assunto volte a ser discutido, é criado um novo tópico. No entanto, não foi o que aconteceu com o indivíduo em questão. Como pode ser visto na Figura 4.18, todo o comportamento padrão explicado acontece do surgimento do *post* até a linha vermelha em destaque. Após a sua criação, diversos usuários ofereceram sugestões de como encontrar o animal desaparecido e enviaram mensagens de apoio ao proprietário com grande frequência, o que desencadeou a rajada de eventos no início do tópico. Posteriormente, alguns comentários similares com menor taxa aconteceram até a aparente morte do tópico, que é o evento imediatamente anterior ao marcado pela linha em vermelho. Este longo período de inatividade foi quebrado por um *post* do proprietário que informa que havia recebido um telefonema com informações a respeito do animal desaparecido. Este grande intervalo sem eventos também foi responsável pela estimativa baixa de $\hat{\lambda}_{EM}$. Novamente o proprietário receberia sugestões e mensagens de apoio, o que elevou a taxa de eventos até que o animal foi encontrado (evento postado pelo proprietário destacado

em azul). A partir daí é desencadeada uma nova rajada de eventos com conteúdo diferente dos anteriores. Desta vez os usuários postaram mensagens felicitando os donos pelo ocorrido. Percebemos, portanto, um tópico dentro de outro, que apesar de serem complementares, possuem dinâmicas claramente diferentes.

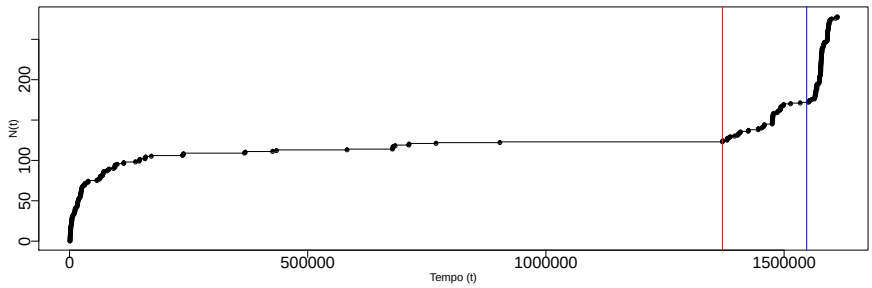
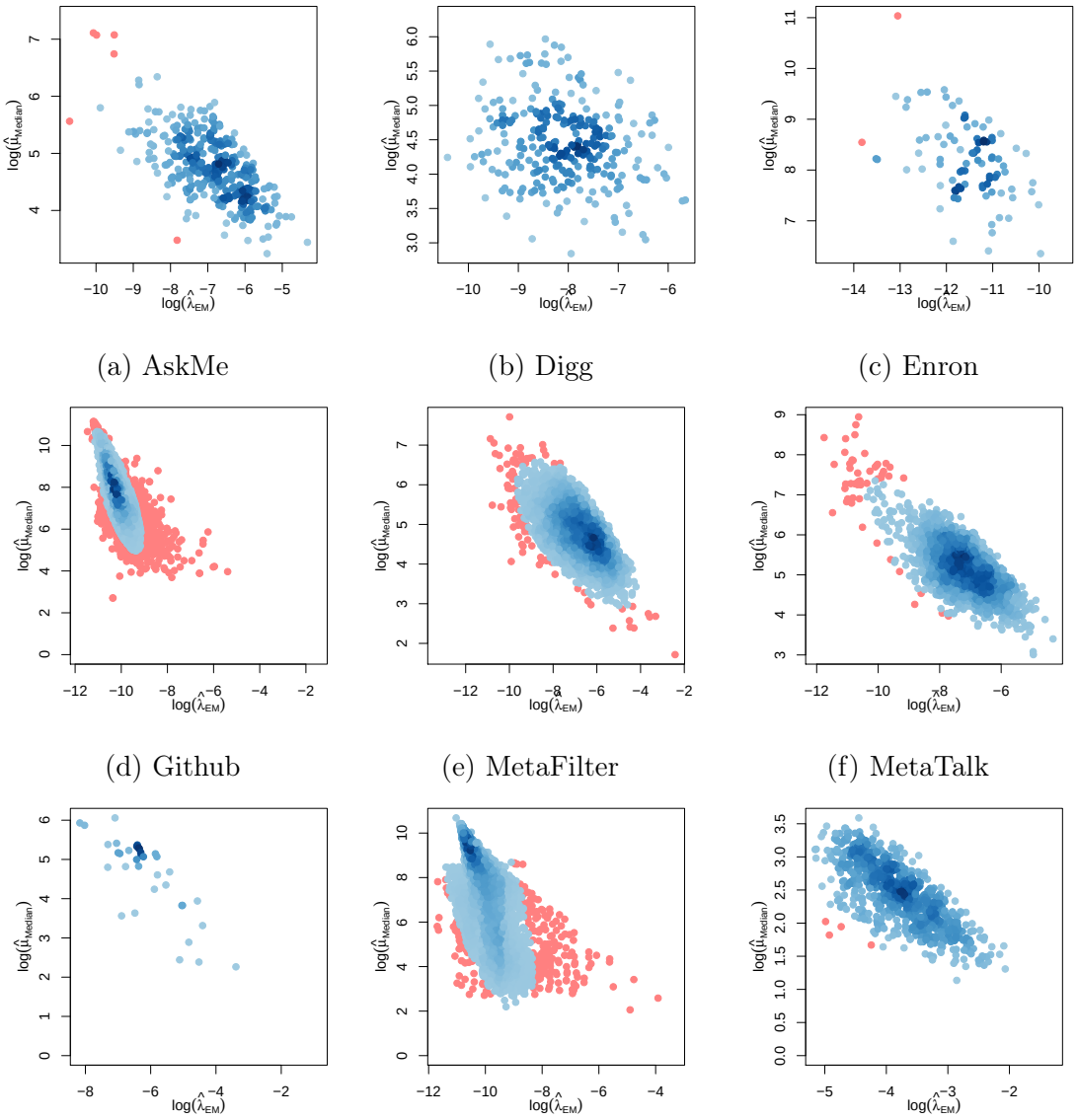


Figura 4.18: Representação do indivíduo 219940, considerado anomalia na base AskMe

Considere também, por exemplo, os indivíduos 18067 e 21900 da base *MetaTalk*, que tratam de um assunto pouco comum no grupo em questão. Ambos são lembretes da data de término para postagem em um evento realizado semestralmente entre os usuários do serviço intitulado *MeFiSwap*. Isso faz com que vários usuários justifiquem o atraso de participação ou comentem algo em relação ao evento. O *MeFiSwap* é uma maneira que os usuários encontraram para compartilhar suas listas de músicas preferidas. O primeiro é em relação ao evento que aconteceu no verão de 2009 e o segundo no inverno de 2012. Por serem lembretes, eles não adicionam conteúdos, mas referem-se a outros *posts* que aconteceram no fórum, o que configura uma promoção.

No *Twitter*, o indivíduo 1088 também apresentou comportamento fora do padrão. Oposto ao primeiro caso, seu destaque foi pela alta taxa do processo de Poisson ajustada pelo modelo proposto. Ao analisarmos os seus *tweets*, concluímos que se tratava de uma promoção que distribuía ingressos para um determinado evento cultural. Para concorrer aos ingressos, os interessados deveriam colocar a *hashtag* *#iwantisatickets* em algum *post* de seu perfil. Isto desencadeou um grande número de *posts* relacionados, elevando o valor de $\hat{\lambda}_{EM}$ no período observado. Depois do período de promoção, os eventos relativos a essa *hashtag* cessaram.

Finalmente, outro caso interessante é evento 65232 do serviço *MetaFilter*, que foi considerado inapropriado e deletado da página inicial por um moderador. O autor do tópico propõe que as compras de supermercado devem ser feita por mulheres pois sua esposa havia encontrado cupons de descontos dos quais ele seria incapaz de pensar. O assunto foi considerado pouco relevante e rapidamente gerou críticas entre os próprios usuários da ferramenta, que consideraram o tópico inútil.



(a) AskMe (b) Digg (c) Enron

(d) Github (e) MetaFilter (f) MetaTalk

(g) Reddit (h) Twitter (i) Yelp

Figura 4.19: Aplicação de detecção de anomalias por base

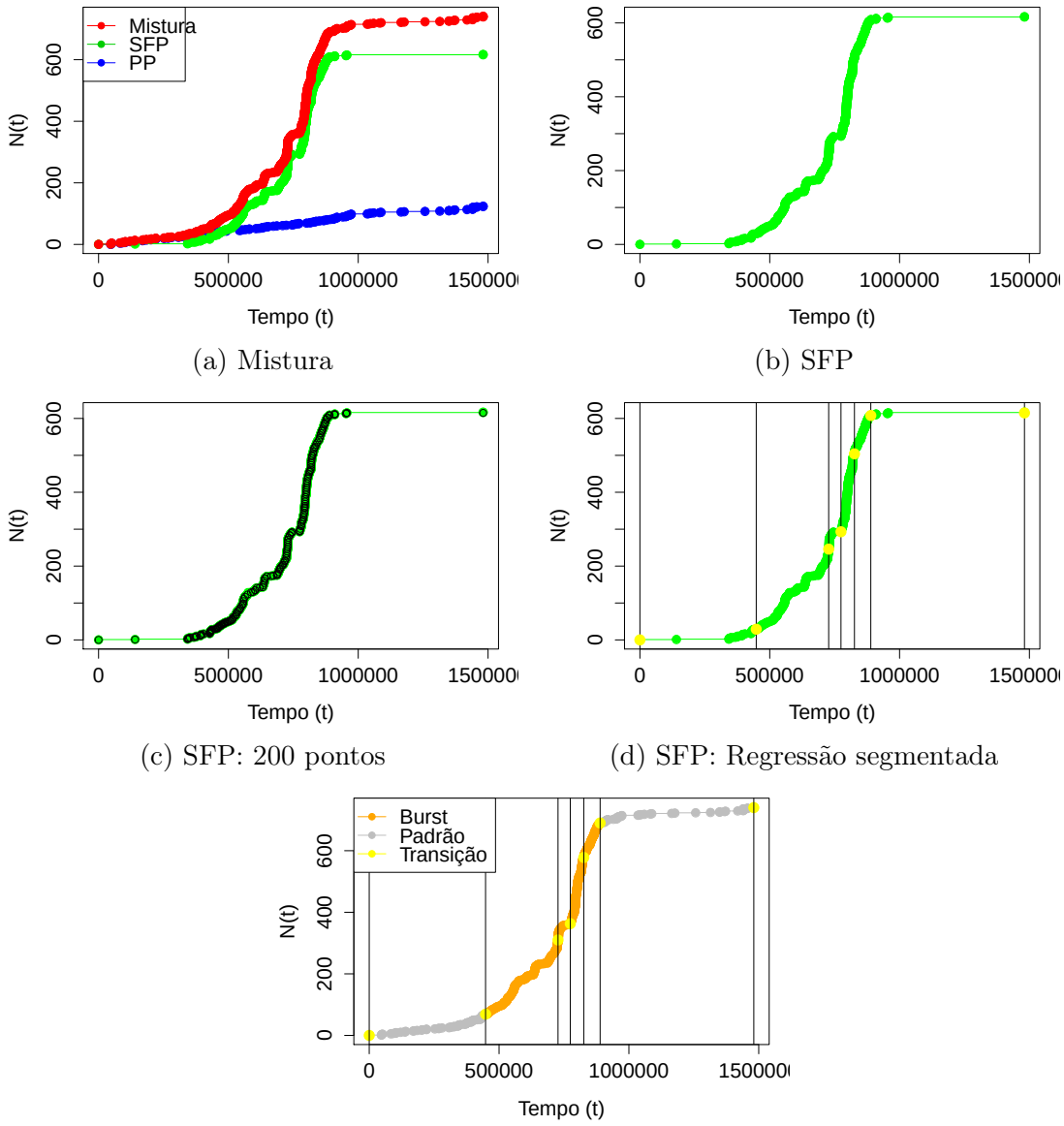
4.6.2 Detecção de Bursts

Outra aplicação prática desenvolvida neste trabalho é a de detecção de períodos de *bursts* nos indivíduos. Esta aplicação se torna interessante por diversos motivos, dentre eles destacamos: (i) possibilitar novos estudos que visam a compreensão das razões do acontecimento dos períodos de alta taxa de eventos; (ii) verificar padrões que nos permitam prever se o período atual de um tópico é de quietude ou de alta atividade; (iii) identificar possíveis subtópicos dentro de *burts* separados.

Vamos exemplificar o método para o individuo 10019 (o mesmo da Figura 4.20a) do conjunto de dados do *Twitter*. Como o processo de Poisson homogêneo possui taxa constante de eventos em qualquer intervalo, os eventos provenientes do mesmo podem ser considerados como de comportamento esperado quando nenhum fator externo atua e influencia o tópico em questão. Por outro lado, o SFP possui características distintas: períodos de longa inatividade seguidos por períodos de atividade intensa. Este último processo pontual, portanto, é responsável pelos *burts* da mistura.

Considerando o SFP puro (Figura 4.20b) extraído da mistura, aplicamos o algoritmo *Segmented Least Squares: Multi-way Choices* descrito em Kleinberg & Tardos [2006], que busca reduzir o erro quadrático mínimo de uma regressão por segmentos. Este algoritmo tem ordem de complexidade cúbica e, portanto, é altamente ineficiente para grandes quantidades de dados. Por isso, reduzimos os SFPs a no máximo 200 pontos, mantendo as características iniciais que permitam a regressão por segmentos. Para isto, consideramos os percentis de 1 a 100, que sozinhos podem não ser suficientes, pois tais pontos podem estar altamente concentrados nos *bursts*. Assim, selecionamos 100 novos pontos, que são uniformemente esparsados pelo intervalo de tempo de toda a série temporal. Além disso, o primeiro e o último ponto da mistura evento são obrigatoriamente adicionados. O resultado desta redução de pontos pode ser visto na Figura 4.20c e o resultado da segmentação na Figura 4.20d. Na Figura 4.20c, os pontos pretos foram os considerados para a execução do algoritmo. Já na Figura 4.20d, os pontos amarelos foram denominados pontos de transição, que indicam o começo ou fim de um novo segmento.

A determinação do que pode ser considerado um *burst* depende de cada aplicação. Consideramos que cada segmento encontrado possui uma potência τ, que é a razão entre o número de pontos SFP observados em um seguimento e a média esperada de eventos Poisson para o mesmo. Portanto, um segmento possui o valor esperado de $\tau+1$ vezes a média de eventos PP. Neste trabalho consideramos que $\tau = 1$ é suficiente para se determinar uma rajada pois, nestes casos, o segmento considerado terá o dobro do número de eventos esperados pelo comportamento do processo Poisson. Na Figura 4.20e



(a) Mistura

(b) SFP

(c) SFP: 200 pontos

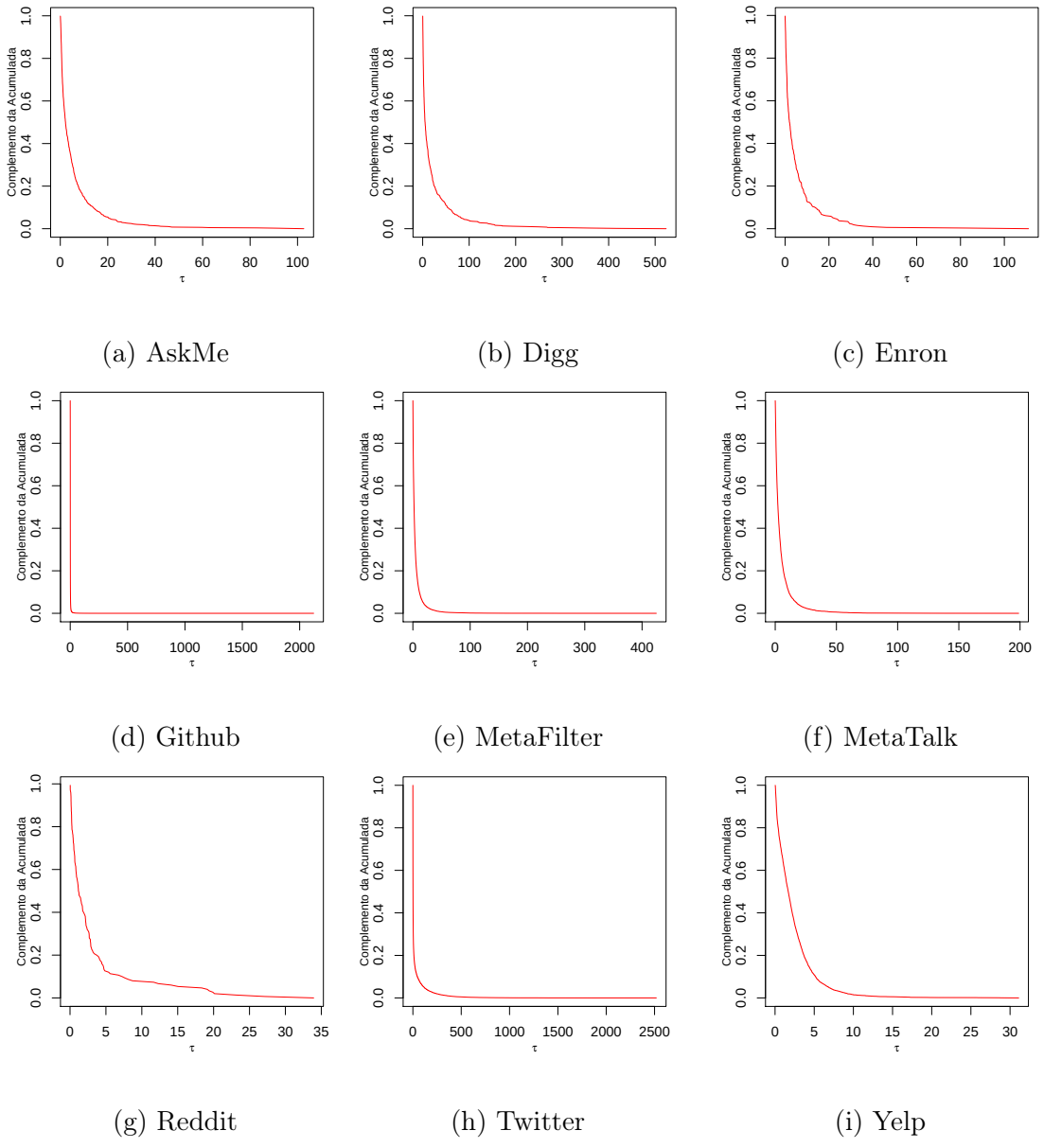
(d) SFP: Regressão segmentada

(e) Mistura: Detecção de burts

Figura 4.20: Passos da aplicação de detecção de burts para o indivíduo 10019 da base Twitter

podemos verificar os bursts detectados. Os pontos em laranja simbolizam os *bursts* enquanto os pontos em cinza a atividade padrão.

Da detecção dos seguimentos de todos os indivíduos considerados nas aplicações verificamos que a distribuição da potência de *bursts* τ dos seguimentos é de cauda pesada, o que vai de acordo com alguns trabalhos relacionados citados (Barabasi [2005], Vázquez et al. [2006], Vaz de Melo et al. [2013]). No entanto, concluímos também que há uma alta concentração de segmentos com potências baixas, sendo que os valores extraordinariamente elevados de evento por segmentos possuem baixas probabilidades.



(a) AskMe (b) Digg (c) Enron

(d) Github (e) MetaFilter (f) MetaTalk

(g) Reddit (h) Twitter (i) Yelp

Figura 4.21: Complemento da distribuição acumulada de τ para cada base

O complemento da distribuição acumulada de τ pode ser analisado na Figura 4.21.

4.7 Discussão dos Resultados

O principal objetivo deste trabalho é modelar, por meio de processos estocásticos pontuais, o comportamento temporal de eventos de usuários quando estes fazem uso de diversos sistemas na *Web*. O modelo de mistura proposto na Seção 3 foi implementado com o intuito de satisfazer tal objetivo.

Dentre as bases de dados consideradas encontram-se diferentes tipos de serviços: de comunicação, como fóruns de discussão e bate-papo através de hashtags (*AskMe*, *MetaFilter*, *MetaTalk* e *Twitter*); de recomendação colaborativos (*Digg*, *Reddit* e *Yelp*); de correio eletrônico (*Enron*) e de controle de versão de projetos de software (*Github*). Foram analisados 97590 indivíduos divididos nas nove bases citadas. O conjunto de dados com o maior número de indivíduos é o *Github* com 66159 e o menor é o Reddit com 102. A média geral, por indivíduo, foi de aproximadamente 600 eventos.

Na Seção 4.3, observamos que o comportamento conjunto dos parâmetros extraídos do algoritmo EM e ajustado pelo método proposto na Seção 3.4.3, $\hat{\lambda}_{EM}$ e $\hat{\mu}_{Median}$, possuem correlação negativa. Isto significa que, para valores elevados de μ, temos baixos valores de λ_{PP}. Este fato nos mostra que processos com grandes rajadas de eventos tendem a ter um comportamento constante maior que os processos onde os *bursts* são menos significativos, ou seja, processos que tenham alto valor esperado de eventos SFP são acompanhados de processos de Poisson de alta taxa de eventos. Situação análoga pode ser constatada em processos de baixa atividade. Como podemos observar na Figura 4.4 , excetuando *Reddit* e *Twitter*, o aspecto do comportamento conjunto das variáveis dos processos se assemelha a uma normal bivaridada. Por possuir poucos indivíduos, a base *Reddit* não possui um comportamento claro. Já os indivíduos do *Twitter* parecem formar uma distribuição desconhecida e complexa.

A despeito dos testes de hipóteses expostos na Seção 4.4, podemos observar que, para as bases reais, o comportamento foi similar ao demonstrado na Seção 3.5. Na grande maioria dos casos, quando um teste de hipótese atribuiu um alto valor para $\phi(\Theta_{PP})$, o outro atribui um baixo valor para $\phi(\Theta_{SFP})$, e vice-versa. O número de casos em que este comportamento não ocorreu é baixo frente ao total de indivíduos analisados. Estes casos podem estar atrelados à própria variação estocástica do método proposto para avaliar as hipóteses. Os fatos aqui citados nos levam a crer que os testes de hipóteses propostos foram também suficientes para separação dos modelos para as bases reais.

As hipóteses consideradas foram: processo de Poisson homogêneo puro, SFP puro ou mistura dos dois anteriores. Na Tabela 4.2 pode-se analisar a distribuição destas hipóteses por bases. Notamos que a mistura de processos se mostrou mais provável que os processos puros para cerca de 70% dos indivíduos e para a maioria das bases. Destas, a maior porcentagem de misturas encontra-se na base *Github* com mais de 90% dos indivíduos sendo provenientes de mistura dos processos. Apenas no caso do *Digg* e do *Yelp* que os tópicos concentraram em processos puros: SFP e PP respectivamente. O *Enron* não apresentou indivíduos considerados processos de Poisson puros. A principal justificativa para o desenvolvimento do modelo de mistura foi que, no caso do SFP

puro, apesar de ter ajustado muito bem a diversas bases, existiram casos que o ajuste não foi adequado. Observamos que os dados nos mostram que o modelo proposto estende o SFP e ainda sugere uma terceira alternativa, o processo de Poisson puro.

Quanto à qualidade dos ajustes, podemos perceber que foram satisfatórios (Seções 4.5.1, 4.5.2 e 4.5.3). Os histogramas dos valores dos ajustes, apresentados nestas seções, em geral estão concentrados próximo ao valor de 1, que é o alvo desejado. Merecem destaques os resultados das bases *Github*, *Twitter* e *Reddit*, que obtiveram considerável número de indivíduos com ajustes inferiores a 0.95 quando considerados SFP puro. Acreditamos que tais indivíduos provavelmente não se adéquam a nenhuma das três hipóteses propostas neste trabalho. Os ajustes dos processos de Poisson das misturas estão altamente concentrados entre os valores de 0.95 e 1.00, excetuando a base *Digg*. No que tange aos SFPs pertencentes à mistura, pode ser constatada a concentração próximo do valor 1, no entanto com maior variação. Esta variação pode ser explicada pelo método estocástico utilizado para separação dos processos. Por se tratar de um método não determinístico, não se pode garantir que não exista alguma separação que se ajuste melhor aos dados. Não se descarta, no entanto, a possibilidade de uma terceira alternativa, não tratada pelo modelo, responsável pelo processo gerador de *burts*, diferente do SFP, que se junta ao processo de Poisson.

Uma outra análise que merece destaque é a porcentagem de pontos provenientes de cada um dos processos quando misturados (Figura 4.17). Este cálculo foi aproximado considerando o valor esperado de pontos de um PP com taxa $\hat{\lambda}_{EM}$ no intervalo observado de cada indivíduo dividido pelo número de pontos observados. Alguns comportamentos merecem destaques em relação à porcentagem de pontos PP: as bases *AskMe*, *Diff*, *MetaFilter*, *MetaTalk*, *Reddit* e *Yelp* possuem uma distribuição semelhante à uma distribuição normal; já as bases *Github* e *Twitter* possuem uma moda e pequena variação em torno da mesma; a base *Enron* possui o comportamento mais atípico de todas com duas modas bem definidas.

Quanto às aplicações, duas foram implementadas: detecção de anomalias e detecção de *bursts*. A primeira trata da detecção de comportamentos anômalos quando considerados os processos em conjunto (Seção 4.6.1). Aproximando o $\log \hat{\lambda}_{EM}$ versus $\log \hat{\mu}_{Median}$ de uma normal bivariada foi possível calcular a distância de Mahalanobis dos indivíduos da média dos valores do parâmetros. Indivíduos cuja probabilidade de pertencer ao grupo foi inferior a 1% foram considerados anômalos. Apenas as bases *Digg* e *Reddit* não apresentaram anomalias. Na seção relacionada foram apresentados cinco exemplos detectados como anômalos de diferentes bases: dois são de *metaposts*; um foi deletado pelo moderador da ferramenta por comportamento inadequado; um fugiu do padrão comum do grupo, reativando o tópico quando se esperava o fecha-

mento; e o último tratou de uma promoção de ingressos para um evento cultural (para participação era necessário um *post* o que elevou consideravelmente a atividade).

Já a segunda aplicação, a de detecção de *bursts*, é descrita na Seção 4.6.2. Consideramos o SFP como responsável pelo processo gerador de burts. Realizando uma regressão linear por segmentos extraímos a razão entre o número de eventos SFP e o número eventos esperados para o processo de Poisson homogêneo dos mesmos. A esta razão denominamos potência do segmento τ. Assim temos um fator indicativo relativo entre período de alta atividade e o comportamento padrão do indivíduo. Esta situação se mostra interessante pois permite diferentes definições do que se trata ou não um *burst* em função da aplicação. Para o trabalho atual consideramos o valor de $\tau = 1$, que significa que o seguimento tem o dobro do número de eventos padrão esperado, suficiente para a definição de *bursts*.

Em virtude do que foi mencionado neste capítulo, acreditamos que o trabalho cumpriu seus objetivos propostos. Os resultados são considerados satisfatórios ao que foi proposto. O modelo de mistura desenvolvido nesta dissertação aderiu bem às bases reais.

Capítulo 5

Conclusões e trabalhos futuros

Nesta dissertação, foi proposto um modelo de mistura que incorpora duas fontes de geração de eventos: o SFP e o processo de Poisson homogêneo. Dentre as contribuições deste trabalho encontra-se a metodologia desenvolvida para a separação dos processos uma vez que não existe teoria sobre o comportamento do EMV no caso de dados de processos pontuais seguindo um modelo complexo como o nosso. Foi utilizado o algoritmo EM para encontrar os estimadores e propostos testes de hipóteses com o intuito de verificar se o SFP puro ou PP puro descrevem os dados observados igual ou melhor que o modelo de mistura.

O modelo proposto possui apenas dois parâmetros mas é capaz de descrever diversos comportamentos. Primeiramente, a taxa do processo de Poisson nos indica o comportamento esperado quando nenhum fator externo atua e influencia o tópico em questão. Já o parâmetro do SFP nos indica a mediana dos tempos entre eventos do processo gerador de *bursts*. Foram utilizadas nove bases reais para validação, de diversos tipos de serviços disponíves na *Web*, nas quais consideramos que o modelo foi bem ajustado. Além disso, quando considerado o comportamento coletivo de cada base verificamos que a porcentagem de eventos provenientes do processo de Poisson segue aparente distribuição normal ou com modas bem definidas. No que tange ao comportamento conjunto das variáveis do processos observamos que possuem correlação negativa.

Ademais, implementamos duas aplicações. A primeira trata da detecção de anomalias. Nesta aplicação, descatamos que a investigação de indivíduos identificados como anômalos resultou em diferentes exemplos de comportamentos destoantes do comportamento padrão das base de dados. Já a segunda aplicação foi a de detecção de *bursts*. Esta nos permite particionar os indivíduos e propor uma relação entre os períodos de alta atividade e os períodos de atividade padrão.

Uma primeira possibilidade de trabalho futuro é a extensão do modelo proposto aos indivíduos que não se ajustaram bem. Para isto é necessário um estudo dos indivíduos em questão com o intuito de elaborar uma hipótese que explique o motivo destes ajustes insuficientes e, a partir de então, propor uma solução para este problema.

Um outro caso é possibilidade de usar outros processos pontuais estocásticos utilizando a metodologia proposta neste trabalho. O processo de Poisson não-homogêneo, por exemplo, poderia explicar comportamentos sazonais proporcionando condutas distintas dos indivíduos em função do tempo. Neste caso, seria necessário adaptar a função de verossimilhança da mistura a esta nova realidade. Testes dos estimadores e testes de hipóteses sugeridos nesta dissertação podem ser reaproveitados.

Sugerimos, ainda como um possível trabalho futuro, a associação dos eventos provenientes do SFP e do PP, de forma separada, à eventuais tópicos relacionados nas bases. Uma opção seria a agregação dos parâmetos λ_{PP} e μ à um modelo gráfico probabilístico, por exemplo, a Alocação Latente de Dirichlet (*LDA, Latent Dirichlet Allocation*, em inglês), o que poderia ser uma solução bayesiana para alocação latentes dos tópicos.

Finalmente, a potência do segmento τ, relacionada à aplicação de detecção de *bursts*, nos parece promissora à aplicação de previsão de comportamento futuro. O número mínimo de eventos esperados estaria ligado a uma possível associação de baixas potências e a taxa de eventos do processo de Poisson homogêneo. Como a potência nos oferece uma taxa relativa entre o número de eventos do segmento e o número de eventos esperados pelo padrão do indivíduo considerado, ela pode auxiliar na sugestão do número médio de eventos futuros. Para isto, em um estudo mais detalhado, a distribuição das potências em cada uma das bases bem como a verificação de um possível padrão de aparecimento surtos de eventos podem ser considerados parâmetros de um possível modelo preditivo.

Referências Bibliográficas

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207--211.

Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Prentice-Hall, New Jersey. ISBN 0-13-022335-2.

Casella, G. & Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Dezsö, Z.; Almaas, E.; Lukács, A.; Rácz, B.; Szakadát, I. & Barabási, A.-L. (2006). Dynamics of information access on the web. *Physical Review E*, 73(6):066132.

Faloutsos, M.; Faloutsos, P. & Faloutsos, C. (1999). On power-law relationships of the internet topology. Em *ACM SIGCOMM Computer Communication Review*, volume 29, pp. 251--262. ACM.

Guttorp, P. & Minin, V. N. (1995). *Stochastic modeling of scientific data*. CRC Press.

InterBrand (2014). The interbrand ranking of the best global brands 2014. `http://www.bestglobalbrands.com/2014/ranking/`, Acessado em: 09/02/2015.

Kim, G.; Fei-Fei, L. & Xing, E. P. (2012). Web image prediction using multivariate point processes. Em *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1068--1076. ACM.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373--397.

Kleinberg, J. & Tardos, E. (2006). *Algorithm design*. Pearson Education.

Liu, P.; Tang, J. & Wang, T. (2013). Information current in twitter: which brings hot events to the world. Em *Proceedings of the 22nd international conference on World*

Wide Web companion, pp. 111--112. International World Wide Web Conferences Steering Committee.

Malmgren, R. D.; Hofman, J. M.; Amaral, L. A. & Watts, D. J. (2009). Characterizing individual communication patterns. Em *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 607--616. ACM.

Malmgren, R. D.; Stouffer, D. B.; Motter, A. E. & Amaral, L. A. (2008). A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153--18158.

Matos-Júnior, O.; Ziviani, N.; Botelho, F.; Cristo, M.; Lacerda, A. & da Silva, A. S. (2012). Using taxonomies for product recommendation. *Journal of Information and Data Management*, 3(2):85.

McLachlan, G. & Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Meyer, P. L. (1970). Probabilidade: aplicações à estatística. Em *Probabilidade: aplicações à estatística*. Livro Técnico.

O'Reilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. *Communications and Strategies*, 65(1):17--37.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr, S. S.; Rennó, D.; Gonçalves, T. S.; Davis Jr, C. A.; Meira Jr, W. & Pappa, G. L. (2012). Observatório do trânsito: sistema para detecção e localização de eventos de trânsito no twitter. *Simpósio Brasileiro de Bancos de Dados*.

Sakaki, T.; Okazaki, M. & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. Em *Proceedings of the 19th international conference on World wide web*, pp. 851--860. ACM.

Schwind, C. & Buder, J. (2012). Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not? *Computers in Human Behavior*, 28(6):2280--2290.

Serfozo, R. F. (1990). *Point Processes em: Heyman, D. P. e Sobel, MJ(eds)*, volume 2. Elsevier, North-Holland, Amsterdam.

Vaz de Melo, P. O.; Faloutsos, C.; Assunção, R.; Alves, R. & Loureiro, A. A. (2014). Universal and distinct properties of communication dynamics: How to generate realistic inter-event times. *arXiv preprint arXiv:1403.4997*.

Vaz de Melo, P. O. S.; Faloutsos, C.; Assunção, R. & Loureiro, A. (2013). The self-feeding process: a unifying model for communication dynamics in the web. Em *Proceedings of the 22nd international conference on World Wide Web*, pp. 1319--1330. International World Wide Web Conferences Steering Committee.

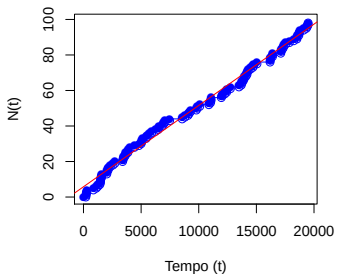
Vázquez, A.; Oliveira, J. G.; Dezsö, Z.; Goh, K.-I.; Kondor, I. & Barabási, A.-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127.

Vazquez, A.; Racz, B.; Lukacs, A. & Barabasi, A.-L. (2007). Impact of non-poissonian activity patterns on spreading processes. *Physical review letters*, 98(15):158702.

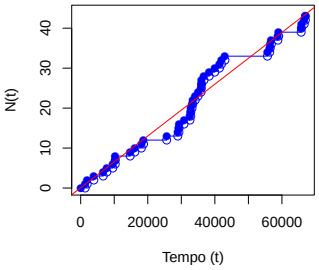
Apêndice A

Exemplos indivíduos aceitos como modelo PP puro

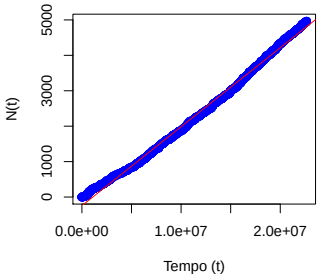
Na Figura A.1, a linha vermelha é considerada linha de ajuste. Como o processo de Poisson homogêneo é estacionário é esperado que a função do tempo versus o número de eventos seja uma reta. Todos os exemplos apresentados possuem $R^2{}_{PP} = 0.99$, excetuando a Figura A.1b que possui $R^2{}_{PP} = 0.98$



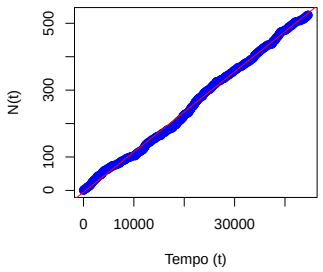
(a) AskMe



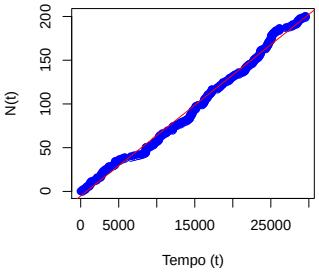
(b) Digg



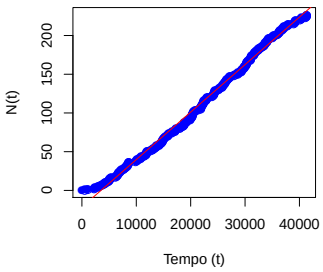
(c) Github



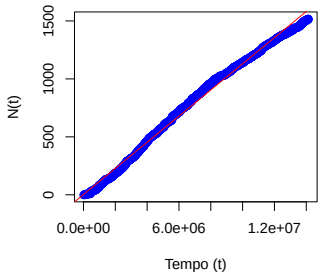
(d) MetaFilter



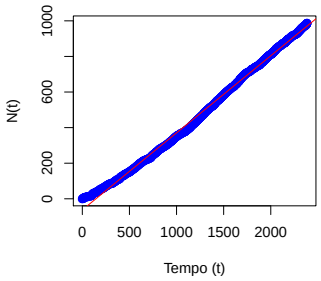
(e) MetaTalk



(f) Reddit



(g) Twitter



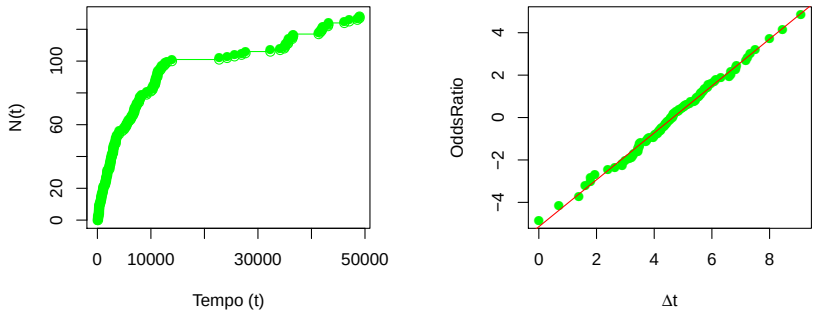
(h) Yelp

Figura A.1: Exemplos indivíduos aceitos como modelo PP puro por base

Apêndice B

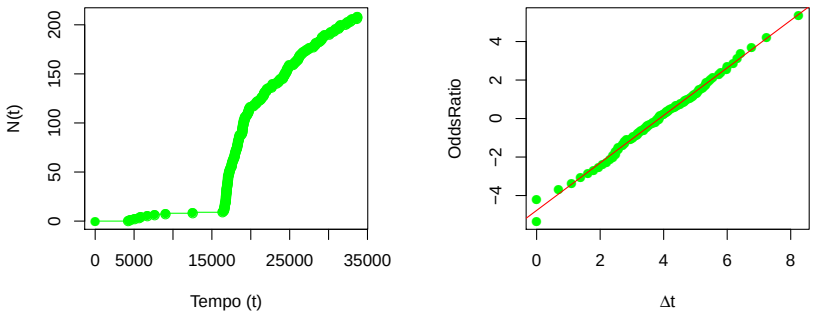
Exemplos indivíduos aceitos como modelo SFP puro

A Figura B.1 representa um indivíduo da base AskMe considerado um SFP puro. Na Figura B.1a pode ser analisado o comportamento temporal onde consegue-se observar os períodos de baixa atividade seguidos por rajadas de eventos, e vice versa. Já na Figura B.1b, a linha vermelha é considerada linha de ajuste. Bons ajustes possuem Δt versus $OddsRatio$ como uma reta. As demais bases estão representadas da Figura B.2 à Figura B.9. Todos os exemplos apresentados possuem $R^2{}_{SFP} = 0.99$, excetuando a Figura B.4 que possui $R^2{}_{SFP} = 0.98$.



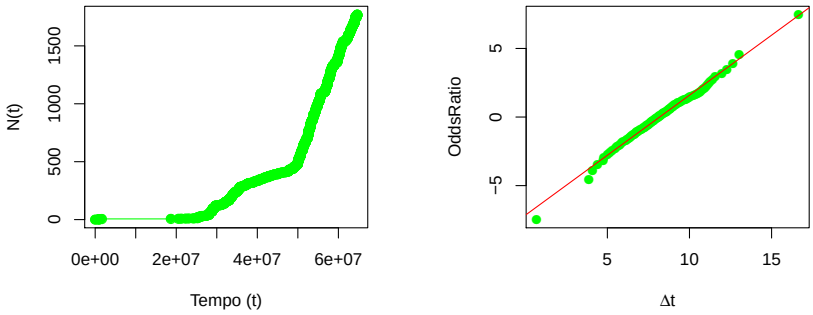
(a) (b)

Figura B.1: Exemplo de indivíduo aceito como modelo SFP puro na base AskMe



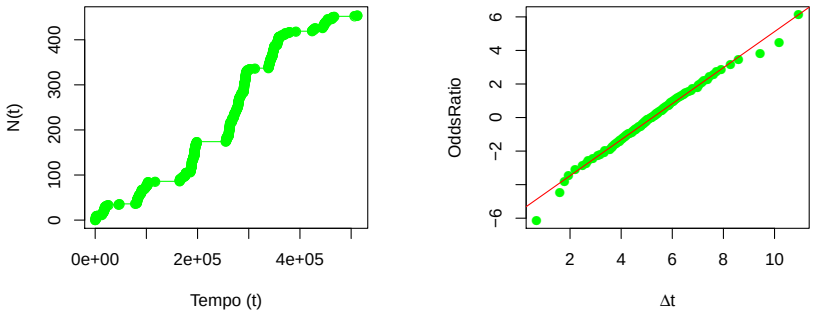
(a) (b)

Figura B.2: Exemplo de indivíduo aceito como modelo SFP puro na base Digg



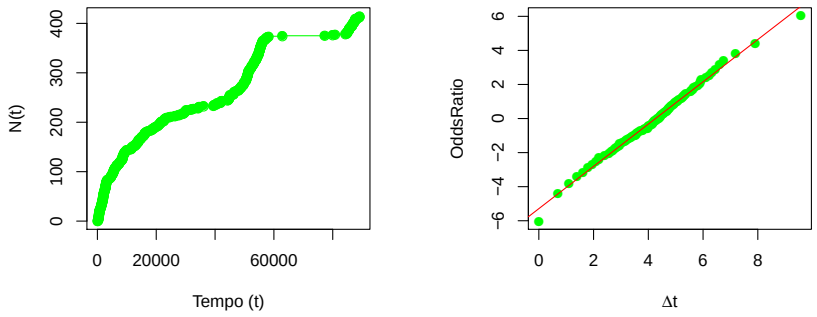
(a) (b)

Figura B.3: Exemplo de indivíduo aceito como modelo SFP puro na base Enron



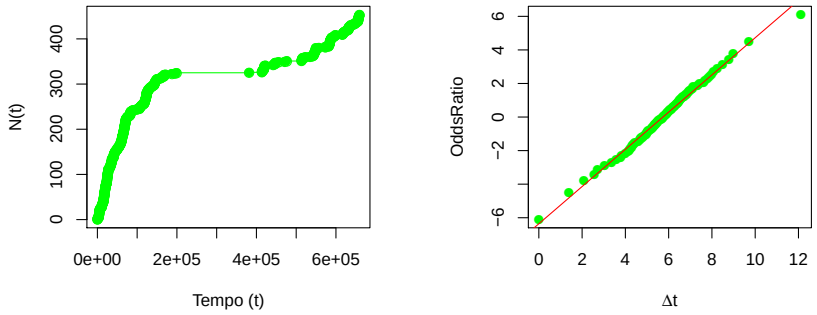
(a) (b)

Figura B.4: Exemplo de indivíduo aceito como modelo SFP puro na base Github



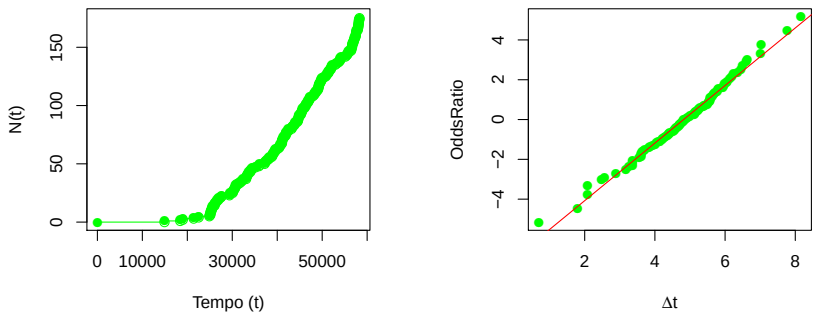
(a) (b)

Figura B.5: Exemplo de indivíduo aceito como modelo SFP puro na base MetaFilter



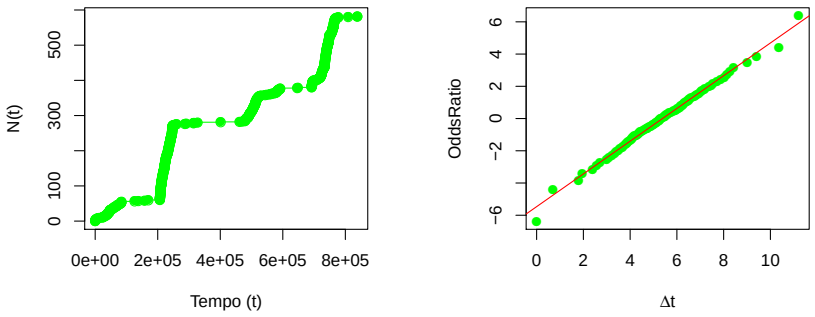
(a) (b)

Figura B.6: Exemplo de indivíduo aceito como modelo SFP puro na base MetaTalk



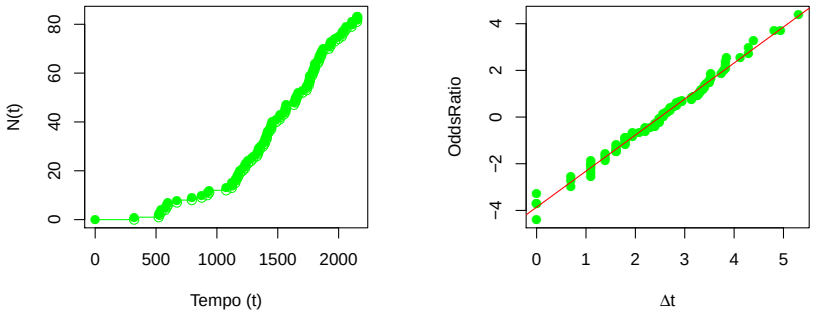
(a) (b)

Figura B.7: Exemplo de indivíduo como modelo SFP puro na base Reddit



(a) (b)

Figura B.8: Exemplo de indivíduo como modelo SFP puro na base Twitter



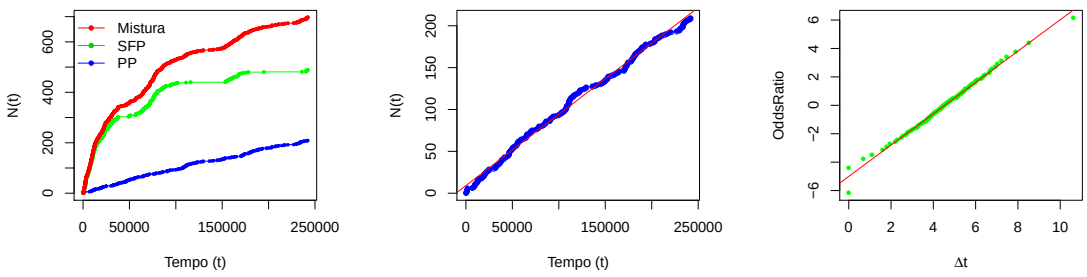
(a) (b)

Figura B.9: Exemplo de indivíduo aceito como modelo SFP puro na base Yelp

Apêndice C

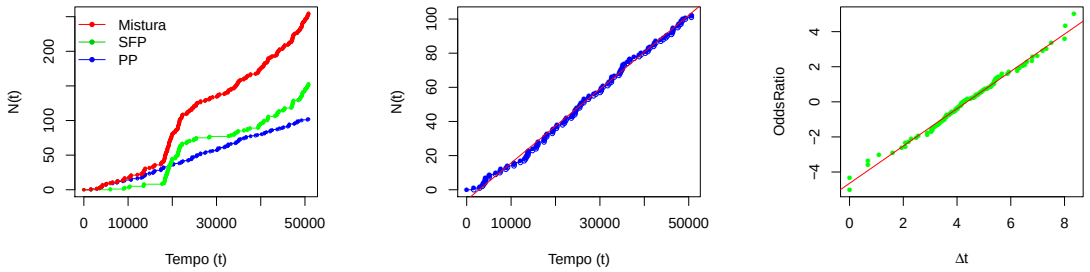
Exemplos indivíduos aceitos como modelo de mistura

A Figura C.1 representa um indivíduo da base AskMe considerado mistura dos processos SFP e PP. Na Figura C.1a pode ser analisado o comportamento temporal onde consegue-se observar a mistura dos processos. Já nas Figuras C.1b e C.1c, as linhas vermelhas são consideradas linhas de ajuste dos processos PP e SFP, respectivamente. As demais bases estão representadas da Figura C.2 à Figura C.9. Todos os exemplos apresentados possuem $R^2{}_{PP} = 0.99$ e $R^2{}_{SFP} = 0.99$.



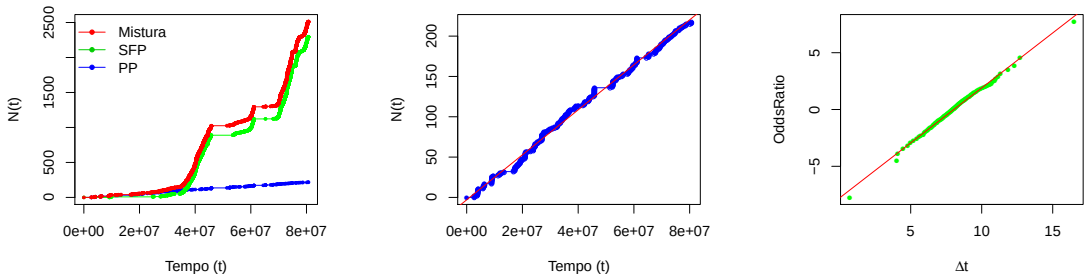
(a) (b) (c)

Figura C.1: Exemplo de indivíduo aceito como modelo de mistura na base AskMe



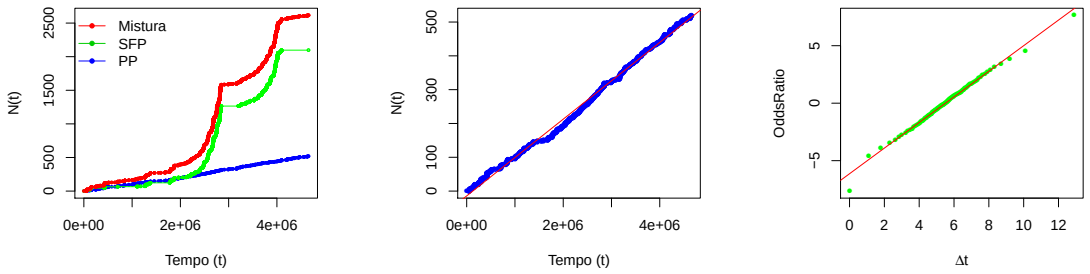
(a) (b) (c)

Figura C.2: Exemplo de indivíduo aceito como modelo de mistura na base Digg



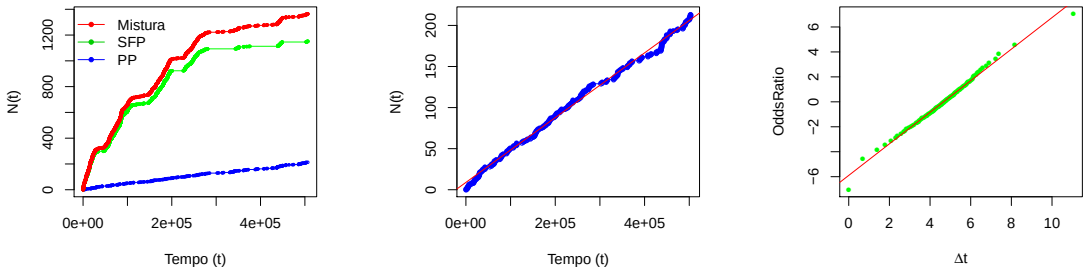
(a) (b) (c)

Figura C.3: Exemplo de indivíduo aceito como modelo de mistura na base Enron



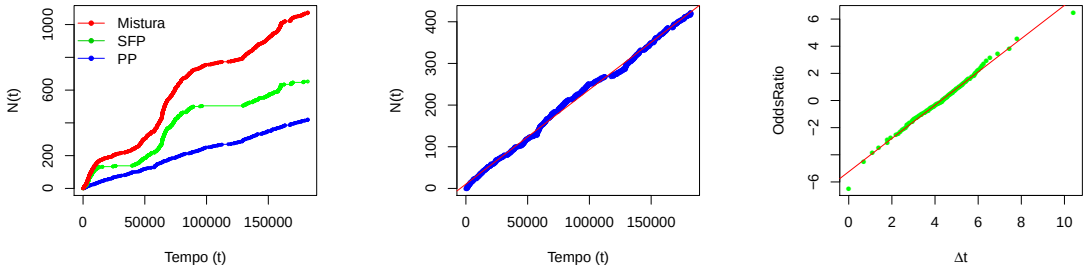
(a) (b) (c)

Figura C.4: Exemplo de indivíduo aceito como modelo de mistura na base Github



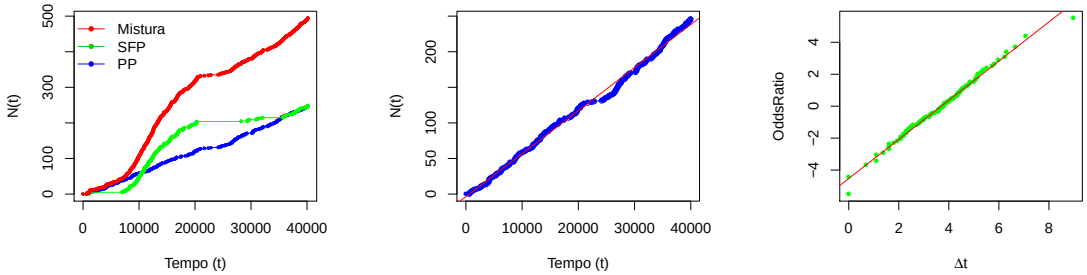
(a) (b) (c)

Figura C.5: Exemplo de indivíduo aceito como modelo de mistura na base MetaFilter



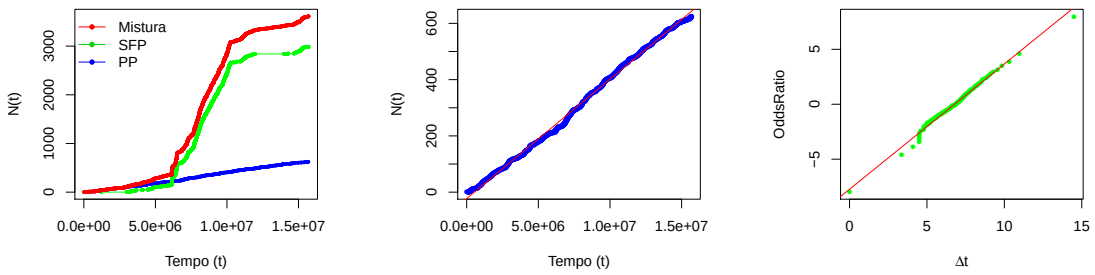
(a) (b) (c)

Figura C.6: Exemplo de indivíduo aceito como modelo de mistura na base MetaTalk



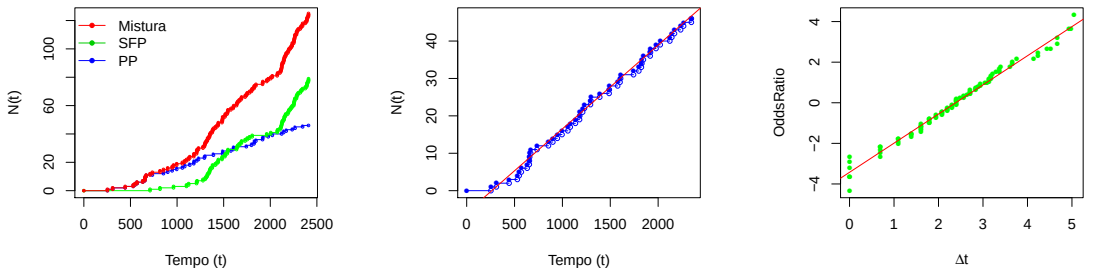
(a) (b) (c)

Figura C.7: Exemplo de indivíduo aceito como modelo de mistura na base Reddit



(a) (b) (c)

Figura C.8: Exemplo de indivíduo aceito como modelo de mistura na base Twitter



(a) (b) (c)

Figura C.9: Exemplo de indivíduo aceito como modelo de mistura na base Yelp