UNIVERSIDADE FEDERAL DE PERNAMBUCO

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

MÉTODO ADAPTATIVO DE *MARKOV CHAIN MONTE CARLO* PARA MANIPULAÇÃO DE MODELOS BAYESIANOS

TESE SUBMETIDA A UFPE
PARA OBTENÇÃO DE GRAU DE DOUTOR
POR

PAULO RENATO ALVES FIRMINO

Orientador: Prof. Enrique López Droguett, Ph. D.

Recife, Agosto de 2009

F525m **Firmino, Paulo Renato Alves.**
Método adaptativo de Markov Chain Monte Carlo para manipulação de modelos Bayesianos / Paulo Renato Alves Firmino. – Recife: O Autor, 2009.
vi, 89 folhas, il : figs., tabs.

Tese (Doutorado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia de Produção, 2009.

Inclui Referências Bibliográficas.

1. Engenharia de Produção. 2. Modelos Bayesianos. 3. Integração de Monte Carlo via cadeias de Markov. 4. Métodos de Quadratura Adaptativos. 5. Métodos de Agrupamento. 6. Rao-Blackwellization. I. Título.

UFPE

658.5 **CDD (22. ed.)** **BCTG/2009-147**

UNIVERSIDADE FEDERAL DE PERNAMBUCO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

PARECER DA COMISSÃO EXAMINADORA
DE DEFESA DE TESE DE
DOUTORADO DE

PAULO RENATO ALVES FIRMINO

"MÉTODO ADAPTATIVO DE *MARKOV CHAIN MONTE CARLO* PARA MANIPULAÇÃO DE MODELOS BAYESIANOS"

ÁREA DE CONCENTRAÇÃO: PESQUISA OPERACIONAL

A comissão examinadora, composta pelos professores abaixo, sob a presidência do(a) primeiro(a), considera **PAULO RENATO ALVES FIRMINO** APROVADO.

Recife, 30 de Julho de 2009.

Prof. ENRIQUE ANDRÉS LÓPEZ DROGUETT, PhD (UFPE)

Prof. LUCIANO NADLER LINS, Doutor (UFPE)

Prof. FERNANDO MENEZES CAMPELLO DE SOUZA, PhD (UFPE)

Prof.ª SILVANA MARIA BASTOS AFONSO DA SILVA, PhD (UFPE)

Prof. MOACYR CUNHA DE ARAUJO FILHO, Docteur (UFPE)

PÁGINA DEDICATÓRIA

Novamente, tenho a oportunidade de fazer nossa essa vitória: aqueles que amo e a mim.

Mestres antigos e recentes: Padre Paulo, Tacio Maciel, Edson Costa de Barros, Maria Cristina Falcão, Sylvio José Pereira, Sidney, Neil, meus examinadores internos para este trabalho, Fernando Campello, Luciano Lins e Ana Paula, e meu orientador desde os tempos do mestrado, Enrique López Droguett.

Amigos antigos e recentes: Roberto, Celício, Roró, Pantuca, Thuca, Marcondes, Dimas, Wanderley, Diogo e Adson.

Companheiros de estrada antigos e recentes: Márcio, Isis, Romero, Paulo Estevão, Ricardo, Rodrigo Bernardo, Ana e Miriam.

Amores antigos e recentes: Aide (minha Mãe), Antônio (meu Pai), David e Renata (meus irmãos), minhas avós e avôs, Ana Isabel e Ana Isadora (minhas sobrinhas), David Iarley e Antony Gabriel (meus sobrinhos), tios e primos e minha família pernambucana: Carlos e Lurdinha (meus compadres), Tavinho (meu afilhado) e sua irmã Laurinha, Elza (minha sogra), Sindelza (minha cunhada) e seu esposo Joca.

As minhas inesgotáveis razões de viver: Sinelza, Sara e Clara, a linhagem dos Vasconcelos Firmino, da qual sou orgulhosamente responsável.

Obrigado, **Meu Pai**, por me permitir viver este momento; até aqui me trouxeste nos braços. O insensato mesmo seria ter consciência sobre incerteza, estudar probabilidade, atentar sobre a ocorrência dos diversos eventos raros que me trouxeram até aqui e, ainda assim, inexplicavelmente creditar tudo isso à minha capacidade tão limitada. A **Ti** a glória, o louvor e o domínio.

RESUMO

Ao longo dos anos, modelos Bayesianos vêm recebendo atenção especial da academia e em aplicações principalmente por possibilitarem uma combinação matemática entre corpos de evidência subjetiva e empírica. A metodologia de integração de Monte Carlo via cadeias de Markov é uma das principais classes de algoritmos para computar estimativas marginais a partir de modelos Bayesianos. Entre os métodos de integração de Monte Carlo via cadeias de Markov, o algoritmo de Metropolis-Hastings merece destaque. Em resumo, para o conjunto de *d* variáveis (ou componentes) do modelo Bayesiano, $\mathbf{X} = (X_1, X_2, ..., X_d)$, tal algoritmo elabora uma cadeia de Markov onde cada estado visitado é uma realização de $\mathbf{X}$, $\mathbf{x} = (x_1, x_2, ..., x_d)$, amostrada das distribuições de probabilidades condicionais das variáveis do modelo, $f(x_i | x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d)$. Quando a simulação é governada por distribuições cuja amostragem direta é viável, o algoritmo de Metropolis-Hastings converge para o método de Gibbs e técnicas de redução de variância tais como *Rao-Blackwellization* podem ser adotadas. Caso contrário, diante de distribuições cuja amostragem direta é inviável, *Rao-Blackwellization* é possível a partir do método de *griddy*-Gibbs, que recorre a funções aproximadas. Esta tese propõe uma variante de *griddy*-Gibbs que pode ser também classificada como uma extensão do algoritmo de Metropolis-Hastings (diferentemente do método de *griddy*-Gibbs tradicional que descarta a possibilidade de se rejeitar os valores amostrados ao longo das simulações). Além disso, algoritmos de integração numérica adaptativos e técnicas de agrupamento, tais como o método adaptativo de Simpson e *centroidal Voronoi tessellations*, são adotados. Casos de estudo apontam o algoritmo proposto como uma boa alternativa a métodos existentes, promovendo estimativas mais precisas sob um menor consumo de recursos computacionais em muitas situações.

Palavras-Chave: Modelos Bayesianos, Integração de Monte Carlo via cadeias de Markov, Métodos de Quadratura Adaptativos, Métodos de Agrupamento, Rao-Blackwellization.

ABSTRACT

Historically, Bayesian models have deserved special attention from academy and applied fields mainly by allowing mathematical combination of human judgments and empirical data. Markov Chain Monte Carlo (MCMC) methodology is one of the main classes of approaches for computing marginal estimates from Bayesian models. Among Markov Chain Monte Carlo methods, Metropolis-Hastings algorithms must be emphasized. In summary, for the set of d variables (or components) of the Bayesian model, $\mathbf{X} = (X_1, X_2, ..., X_d)$, such algorithm elaborate a Markov Chain where each visited state is a random realization of $\mathbf{X}$, $\mathbf{x} = (x_1, x_2, ..., x_d)$, sampled from the full conditional distribution of the variables, $f(x_i | x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d)$. When the sampling process is governed by distributions cheap to be sampled from, Metropolis-Hastings algorithm converge towards the well-known Gibbs sampling and variance reduction techniques such as Rao-Blackwellization can be introduced into the inference. Otherwise, in the face of distributions expensive to be sampled from, Rao-Blackwellization is possible by adopting approximate functions and then a griddy-Gibbs sampling approach, originally a non- Metropolis-Hastings extension since eventual rejections are not taken into account. This thesis is an effort for studying griddy-Gibbs sampling as a Metropolis-Hastings variant. In this way, adaptive rejection Metropolis sampling concepts and simple clustering and numerical integration algorithms (like centroidal Voronoi tessellations and adaptive Simpson's rule, respectively) are introduced into griddy-Gibbs approach. Case studies from literature point out the good performance of the proposed method in comparison with established methods in terms of both accuracy and time consumption.

Keywords: Bayesian Models, Markov Chain Monte Carlo, Adaptive Quadrature Methods, Clustering, Rao-Blackwellization.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

ARMS – *Adaptive rejection* Metropolis *sampling*

ASR – *Adaptive* Simpson*'s rule*

DAG – *Direct Acyclic Graph*

FCD – *Full conditional distribution*

GGS – *Griddy*-Gibbs *sampling*

GS – Gibbs *sampling*

MB – Modelo Bayesiano

MCMC – Markov *Chain* Monte Carlo

MH – Metropolis-Hastings *algorithm*

RB – Rede Bayesiana

RMC – Rede Multiplamente Conectada

RSC – Rede Singularmente Conectada

1. INTRODUÇÃO

Atualmente, a utilização de modelos Bayesianos (MBs) permite o manuseio de grandes conjuntos de variáveis relacionadas e a utilização de qualquer fonte de informação para auxiliar em previsões, diagnósticos e consequentemente tomadas de decisão. Isto é particularmente observado em análises de risco e confiabilidade, onde usualmente os dados empíricos disponíveis são insuficientes para a mensuração de todos os parâmetros requeridos pelos modelos, levando ao uso de opiniões de especialistas, por exemplo.

De fato, com os avanços científicos principalmente na área computacional, a modelagem e inferência estatística sofreram uma grande transformação. Diferentemente de períodos anteriores, nos quais a realidade a ser modelada era simplificada a fim de que desdobramentos analíticos levassem às quantidades de interesse, a elaboração de modelos mais realistas (e conseqüentemente mais complexos) pôde enfim ocorrer. Nos dias atuais, métodos computacionais de otimização, tais como algoritmos genéticos (Michalewicz, 1999) e nuvens de partículas (Kennedy & Eberhart, 2001) têm ocupado papel de destaque na estatística clássica quando da adoção de estimadores tais como de máxima verossimilhança. Firmino *et al.* (2007) e Johansen *et al.* (2008) são um exemplo disso. Por outro lado, a tradicional dificuldade de se extrair estimativas marginais de modelos Bayesianos genéricos devido às complexas operações matemáticas envolvendo inversões e integrais (ou somatórios) tem sido superada, dando espaço à busca por técnicas de manipulação que computem métricas confiáveis e em menor espaço de tempo. Destacam-se neste contexto estudos recentes tais como Langseth *et al.* (2009) e Firmino & Droguett (2009), direcionados à manipulação de redes Bayesianas mistas – estruturas gráficas e de dados resumidos em distribuições de probabilidade envolvendo variáveis discretas e contínuas.

Em linhas gerais, um MB descreve a distribuição de probabilidades conjunta das variáveis envolvidas a partir do produto de suas distribuições condicionais (regra do produto). Pode se tornar um desafio árduo obter distribuições marginais, ou mesmo medidas de posição e dispersão, de MBs de maneira exata. Os modelos ditos não-conjugados são um exemplo disso. Uma das mais flexíveis e consolidadas estratégias para a obtenção de estimativas marginais nesses casos consiste em amostrar uma cadeia de Markov a partir das distribuições condicionais do MB e estudá-la em seu período estacionário. Dado o estado corrente da cadeia, composto pelos valores mais recentemente amostrados para as variáveis do modelo, amostra-se um novo valor para cada variável de acordo com sua distribuição de probabilidades condicionada ao valor corrente das demais variáveis, levando a cadeia a iterar

para um novo estado. Os métodos que seguem tal estratégia são denominados métodos de Monte Carlo via cadeias de Markov (MCMC). Como bem sintetizado por Gilks *et al.* (1996) MCMC é essencialmente integração de Monte Carlo por meio de cadeias de Markov. Os métodos de MCMC têm sido fundamentais para a difusão e desenvolvimento da inferência Bayesiana. Como exemplo, cite-se trabalhos recentes de autores como Kelly & Smith (2009), que elevam o advento de MCMC à condição de estado da arte em áreas tais como avaliação probabilística de riscos.

Muitos são os métodos de MCMC. Dentre os mais flexíveis, encontram-se os algoritmos de Metropolis *et al.* (1953) e Hastings (1970) – MH, *adaptive rejection Metropolis sampling* de Tierney (1991) – ARMS e *slice sampling* de Neal (2003) – SS. Esses métodos permitem que a cadeia seja iterada mesmo diante de uma distribuição condicional não-padrão (cuja amostragem direta é inviável). Nesses casos, tais métodos amostram de uma função alternativa (também chamada de distribuição proposta) à condicional não-padrão. O preço a ser pago por MH e ARMS é a possibilidade de os valores amostrados serem rejeitados, a chamada probabilidade de rejeição de Metropolis-Hastings, presente em ambos os algoritmos. Já SS recorre a uma distribuição uniforme em torno do valor corrente da variável de maneira recursiva até que o valor proposto seja aceito. Adaptativamente, a cada iteração de ARMS os valores amostrados da distribuição proposta rejeitados em uma fase preliminar à avaliação final desta iteração são usados de maneira a aprimorar a distribuição proposta, enquanto que em SS cada ponto rejeitado pode ser usado para delimitar a distribuição uniforme a ser amostrada em seguida. Esta característica adaptativa é uma das principais vantagens de ARMS e SS; a probabilidade de rejeição é inversamente proporcional ao número de valores rejeitados. Por sua vez, trabalhos recentes envolvendo MH adaptativos, tais como Cai *et al.* (2008), Andrieu & Thoms (2008), Gerlach & Chen (2008) e Keith *et al.* (2008), podem também ser encontrados. Em geral MH adaptativos baseiam-se na própria cadeia amostrada até então, não sendo rigorosamente customizados às distribuições condicionais sob estudo. De qualquer maneira, é ainda possível que o valor amostrado da distribuição proposta seja rejeitado em ARMS e MH, assim como pode ocorrer de a distribuição uniforme usada para iterar a cadeia em SS seja muito estreita, requerendo uma quantidade elevada de iterações até que o processo alcance seu período de estacionaridade.

Por outro lado, quando o processo de amostragem envolve apenas distribuições padrões (cuja amostragem direta é possível), uma das variantes de MH com maior prestígio pode ser adotada: *Gibbs sampling* de Geman & Geman (1984) – GS. Além de anular a possibilidade de rejeições, GS permite o uso de técnicas de redução de variância tais como *Rao-*

Blackwellization, levando a inferências mais precisas do que aquelas simplesmente baseadas na cadeia de Markov amostrada para o mesmo número de iterações. Com *Rao-Blackwellization*, GS possibilita estimar distribuições marginais a partir das condicionais que geram os valores amostrados enquanto que em geral outras variantes de MCMC, a exemplo de todas as demais comentadas até aqui, consideram apenas os valores amostrados. Com o objetivo de generalizar o uso de GS aos casos envolvendo condicionais não-padrões, Ritter & Tanner (1992) propuseram o *griddy-Gibbs Sampling* – GGS. A estratégia de GGS é amostrar de uma função aproximada à condicional não-padrão computada a partir de um vetor de pontos nos quais esta foi previamente avaliada. Trata-se de uma idéia simples para promover a aplicação generalizada de GS, porém passível de melhoramentos. Por exemplo, Ritter & Tanner negligenciam o fato de que como a distribuição amostrada geralmente diverge em algum nível da condicional não-padrão original, seus valores amostrados devem ser avaliados segundo a probabilidade de rejeição de Metropolis-Hastings. Embora possua deslizes conceituais desse tipo, poucos são os trabalhos de pesquisa dedicados a um aprimoramento de GGS desde sua publicação, sendo mais comuns aplicações das idéias originais de Ritter & Tanner.

A exemplo de muitos dos métodos propostos ao longo da literatura de MCMC, o presente trabalho faz uma re-leitura de conceitos subjacentes a métodos existentes e introduz novos procedimentos, com o intuito de propor novas variantes. Mais especificamente, apresenta-se uma extensão de GGS que pode ser classificada como um dos poucos métodos de MCMC adaptativos (se não o único) que possibilita rejeições de pontos amostrados e também *Rao-Blackwellization*. Para tanto, o GGS proposto faz uso da probabilidade de rejeição de MH e recorre a algoritmos triviais de integração e agrupamento, tais como o método adaptativo de Simpson apresentado por Mckeeman (1962) e o de MacQueen comentado por Du & Gunzburger (2002), respectivamente. Casos de estudo indicam o potencial do GGS proposto em fornecer estimativas mais acuradas e em menor tempo de simulação em relação ao GGS tradicional.

O principal formalismo adotado no presente trabalho para tratar de MBs é redes Bayesianas (RBs). Inicialmente apresentadas por Pearl (1986) para modelar computacionalmente o raciocínio inferencial humano, RBs são grafos acíclicos e direcionados onde os nós representam variáveis aleatórias e os arcos direcionados implicam em relações de causa e efeito quantificadas a partir de distribuições de probabilidades condicionais. RBs permitem o encapsulamento de vários outros formalismos da estatística clássica e Bayesiana, podendo-se citar cadeias de Markov de tempo discreto, modelos Bayesianos de regressão,

modelos Bayesianos de séries temporais, modelos hierárquicos – Bernardo & Smith (1995) e as metodologias de análise de variabilidade populacional de Droguett *et al.* (2004) e de incerteza de modelos de Droguett & Mosleh (2008). Desde a constatação por Cooper (1990) e Dagum & Luby (1993) de que a manipulação de RBs genéricas é um problema NP-*hard*, este tema tem motivado inúmeras pesquisas; Neil *et al.* (2007), Langseth *et al.* (2009) e Firmino & Droguett (2009) são alguns exemplos recentes direcionados aos casos envolvendo variáveis discretas e contínuas.

1.1. Justificativa

Atualmente, MBs têm ocupado lugar de destaque tanto na área científica quanto em aplicações nos mais diversos segmentos, desde reconhecimento de padrões – Heckerman *et al.* (1995) – até análises probabilísticas de riscos – Firmino *et al.* (2006). É inquestionável para tal tendência a contribuição de formalismos como redes Bayesianas, que vêm facilitando o tratamento e a aplicação de tais modelos. Devido aos métodos de MCMC serem uma das mais atraentes alternativas para lidar com MBs de maneira mais genérica (e RBs mais especificamente), a busca por alternativas que promovam estimativas mais confiáveis e em menor tempo computacional é um dos temas mais pesquisados do segmento. Movimentos recentes apontam a introdução de adaptatividade como uma característica útil neste sentido, um atributo bastante valorizado no presente trabalho.

A realização do trabalho inicia-se com um apanhado geral sobre os principais formalismos subjacentes à inferência Bayesiana, enfatizando-se o poder de generalização de RBs em relação a todos eles e as dificuldades inerentes à sua manipulação. Com o objetivo de lidar com tais dificuldades, introduz-se subsequentemente alguns dos principais algoritmos de MCMC: Metropolis-Hastings (MH), *adaptive rejection Metropolis sampling* (ARMS), *Gibbs sampling* (GS) e *griddy-Gibbs sampling* (GGS). Suas respectivas deficiências são também enumeradas, dando espaço para a introdução de novas variantes.

Em seguida, uma fundamentação teórica sobre o método de integração numérica adaptativo de Simpson e métodos de agrupamento probabilísticos como *centroidal Voronoi tessellations* faz-se necessária, com o intuito de introduzir melhoramentos a GGS, levando a estimativas mais acuradas e em menor tempo de simulação. Ainda neste sentido, métricas de desempenho baseadas no conceito de entropia cruzada são mencionadas, a fim de comparar a acurácia do GGS proposto com a do GGS usualmente adotado pela literatura.

1.1.1. Objetivo Geral

Deseja-se com este trabalho propor uma extensão de MH que promova *Rao-Blackwellization*. A intenção é computar estimativas marginais de MBs mais acuradas com um número reduzido de iterações do método de MCMC, eventualmente levando a um menor consumo de tempo de simulação.

1.1.2. Objetivos Específicos

- Apresentar RBs como principal formalismo para inferência Bayesiana;
- Apresentar alguns dos principais algoritmos de MCMC para manipulação de MBs;
- Enfatizar as limitações e deficiências do método de GGS;
- Introduzir e generalizar o método adaptativo de Simpson (ASR);
- Introduzir e adaptar *Centroidal Voronoi Tessellations* (CVT);
- Sugerir uma variante de GGS baseada em ASR e CVT;
- Comparar o desempenho do algoritmo proposto ao do GGS tradicional e métodos alternativos em termos de acurácia e de consumo de recursos computacionais.

No próximo capítulo, discute-se causalidade como sendo o alvo de modelagem dos principais formalismos subjacentes a MBs, mensurada a partir do cálculo das probabilidades. Apresentar-se-á os conceitos de modelos hierárquicos e algumas metodologias de inferência Bayesiana. Por fim, redes Bayesianas são introduzidas a fim de generalizarem os demais formalismos apresentados, possibilitando a realização mais ampla da inferência Bayesiana. As dificuldades de manipulação de RBs são também explicitadas, dando margem para a introdução de métodos de MCMC.

2. MODELOS BAYESIANOS

2.1. Modelagem da Incerteza Inspirada no Comportamento Humano

Argumentar sobre qualquer domínio normalmente requer simplificações. Quando se deseja enquadrar o conhecimento ou o comportamento de regras como "pássaros voam" ou "onde há fumaça há fogo", sabe-se que há exceções e que não compensa, em muitos casos, enumerá-las por completo. Pearl (1988) buscou estudar e sumarizar uma linguagem de exceções de julgamentos na qual crenças pudessem ser representadas e processadas de acordo com a suposição bastante razoável de que quando não se tem o domínio pleno sobre o problema, o raciocínio sensato deve ser aplicado. Assim, ele resolveu aliar linguagem, objetivos e técnicas do artefato da inteligência artificial e da disciplina de inferência Bayesiana. Segundo Pearl (1988), o principal objetivo da inteligência artificial é prover um modelo matemático computacionalmente viável do comportamento inteligente ou, em outras palavras, do raciocínio sensato. Já a inferência Bayesiana, busca estabelecer uma maneira para que estados de conhecimento sejam alterados à luz de novas evidências, mesmo sendo elas parciais ou incertas.

Além das limitações computacionais superadas apenas nas últimas décadas anteriormente citadas, discutidas em maiores detalhes por Congdon (2006), Martz & Waller (1982) lembram que a inferência Bayesiana se alternou entre períodos de aceitação e rejeição durante muito tempo, principalmente devido à idéia ocasionalmente predominante de que a probabilidade requer uma maciça quantidade de dados, a enumeração de todos os possíveis resultados, e de que é mal estimada por pessoas. Diante disto, muitas das áreas de aplicação da Estatística se direcionaram à inferência empírica, fazendo uso da estatística clássica, baseada em dados propriamente ditos. Pearl absteve-se desta perspectiva empírica e tentou comunicar a idéia de que probabilidade não é de fato sobre números, mas sim sobre a estrutura do raciocínio. Ele tentou exercitar a idéia de que a inferência Bayesiana é única na sua habilidade de processar crenças e o que torna o processamento computacional possível é que a informação necessária para a especificação do contexto de dependências pode ser embutida em grafos e manipulada por propagações locais. Tal linha de pesquisa resultou em um formalismo elegante e abrangente para a modelagem Bayesiana: RBs. A grande maioria dos formalismos para inferência Bayesiana anteriores a RBs pode ser encapsulada em seus moldes. Com isso, não só desenvolvimentos teóricos sobre como manipular MBs mas também uma difusão da inferência Bayesiana puderam ser observados. Embora existam várias denominações para RBs (tais como redes de crenças, redes causais e diagramas de influência) uma busca simples por

"*Bayesian networks*" na página eletrônica www.scopus.com leva à ocorrência de 7918 artigos científicos. Isto é um indício da contribuição de RBs para a difusão de MBs nos contextos científico e aplicado.

2.2. Causalidade

A causalidade é a condição segundo a qual uma causa produz um efeito. Pearl (2000) a traduz como "a consciência do homem acerca do que causa o que a sua volta". Assim, trata-se de causalidades mesmo que involuntariamente, quando se diagnostica o porquê ou prognostica a conseqüência da ocorrência de algum evento. Quando se tenta avaliar o impacto de um novo acessório de determinado produto no mercado, o rendimento de um time de futebol com a entrada de um novo atleta ou no que pode resultar a falta de cuidados com a saúde, o interesse recai sobre os efeitos que tais eventos podem causar. Porém, se o problema é explicar por que as vendas caíram, por que o time não vence ou mesmo diagnosticar uma doença, a perspectiva volta-se às causas das ocorrências destes eventos. Na inferência Bayesiana, a leitura da causalidade pode ainda assumir uma face menos intuitiva em um primeiro momento. Nela, pode-se tentar modelar a influência de parâmetros populacionais intangíveis que regem probabilisticamente a ocorrência de fenômenos observáveis. Desta leitura emerge a possibilidade de atualizar estados de conhecimento sobre algo latente diante da observação dos seus "efeitos". Trata-se de uma tentativa de conectar os parâmetros que regem a natureza (uma entidade superior) com fenômenos físicos observáveis.

Sob esta ótica, o grande desafio foi encontrar uma maneira de traduzir matematicamente a causalidade existente entre os eventos que compõem um dado problema de maneira realista, isto é, sem desprezar as incertezas inerentes. Assim, o cálculo das probabilidades foi adotado.

2.3. Cálculo das Probabilidades

O cálculo das probabilidades permite a representação das dependências e independências entre eventos. Com isto, é possível a abordagem matemática da causalidade, representada através de dependências probabilísticas.

A seguir, faz-se um esboço da aplicação do cálculo das probabilidades na quantificação das incertezas e das relações causais inerentes a um problema.

2.3.1. Axiomas de Kolmogorov

A estrutura matemática da probabilidade é estabelecida por três axiomas apresentados por Kolmogorov, a partir do uso de uma tripla (*U, E, P*), onde *U* se refere ao conjunto cujos elementos são os possíveis resultados do experimento, *E* equivale ao conjunto cujos elementos

envolvem todos os possíveis agrupamentos dos elementos de *U*, os eventos, e *P* é o conjunto de funções, chamadas de probabilidades, que relacionam a cada elemento de *E* um número real.

Axioma 2.1 *A função de P relacionada ao elemento U de E tem o valor 1. Isto pode ser escrito como* $P(U) = 1$.

Axioma 2.2 *As funções de P relacionam apenas valores não-negativos aos eventos de E, isto é,* $P(A) \geq 0. \forall A \in E$.

Axioma 2.3 *Se os elementos de um subconjunto qualquer de E são mutuamente exclusivos, a probabilidade da união de tais elementos equivale à soma das suas probabilidades individuais, ou seja,* $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$.

A realização do cálculo de probabilidades requer que se definam três bases. A primeira, *U*, é o conhecimento sobre todos os possíveis resultados do experimento. Embora pareça algo complexo, *U* é moldável a qualquer nível de detalhamento sobre o problema. Pode-se classificá-lo, por exemplo, como a ocorrência ou não de algum evento; esta categorização abrange todos os possíveis resultados do experimento, mesmo que de forma bastante generalizada. As duas últimas bases para o cálculo de probabilidades, *E* e *P*, permitem a introdução de incerteza quanto aos elementos de *U*, assim como quão verossímil se acredita ser cada combinação entre seus possíveis resultados. No exemplo citado, haveria apenas dois eventos excludentes a terem suas probabilidades quantificadas: a ocorrência ou não do evento em estudo. Note-se que não há sentido em relacionar probabilidades aos elementos de *U*, pois estes não tratam de eventos mas sim de resultados; esta comunicação é realizada através de *E*.

O Axioma 2.1 diz que como *E* contém todas as combinações possíveis dos elementos de *U*, o próprio *U* está contido em *E*, o que significa que existe um evento em *E* que representa o fato de que certamente ocorrerá algo ao se realizar o experimento.

Como as funções de *P* são na verdade probabilidades, não há sentido em considerá-las passíveis da atribuição de valores negativos (Axioma 2.2).

O último axioma apresentado expressa a idéia de que caso não haja qualquer maneira de ocorrer dois ou mais eventos de interesse simultaneamente, a probabilidade de acontecer ao menos um deles equivale à soma de suas probabilidades.

2.3.2. Dependências Probabilísticas

Dois outros conceitos do cálculo das probabilidades fundamentais para MBs, inerentes um ao outro, são o de dependência entre eventos e o de ocorrência condicionada de eventos. Da Figura 2.1, vê-se que a probabilidade de o evento *A* ocorrer dada a ocorrência do evento *B* é a

relação entre a probabilidade conjunta da ocorrência de *A* e *B* (área tracejada) e a probabilidade de ocorrência de *B*. Ou seja, P(*A*|*B*) = P(*A* ∩ *B*)/P(*B*). Desta maneira, diz-se que *A* e *B* são estatisticamente independentes se dada a ocorrência de *B*, a probabilidade de ocorrência de *A* se mantém inalterada e pode ser expresso por P(*A*|*B*) = P(*A*). Esta noção de independência define-se como segue:

Definição 2.1 *Se os elementos de um subconjunto qualquer de E são eventos estatisticamente independentes, a probabilidade da ocorrência conjunta destes iguala-se ao produto das suas probabilidades individuais. Ou, em notação probabilística,* $P\left(\bigcap_i A_i\right) = \prod_i P(A_i)$.

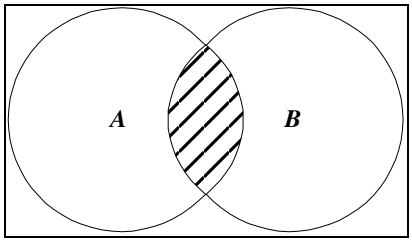


Figura 2.1 Dependência entre eventos.

A notação adotada para expressar a definição de independência entre eventos será *ind*, logo descreve-se a independência entre dois eventos *A* e *B* por *A ind B*. A independência condicional é uma generalização do conceito de independência que pode ser intuitivamente abordada por duas perspectivas: usando o raciocínio probabilístico conceitual e o de análise de dados. Conceitualmente, admite-se que a ocorrência de um subconjunto de eventos qualquer de *E* redistribui as probabilidades de ocorrência de eventos de um outro subconjunto, de forma a torná-los independentes entre si, mesmo que originalmente estes sejam dependentes. Já sob a ótica da análise de dados, considera-se que as informações trazidas pelo conhecimento dos resultados de um subconjunto de eventos podem tornar informações mutuamente irrelevantes, ao se tratar de eventos de um outro subconjunto. Segue a definição de independência estatística condicionada:

Definição 2.2 *Se os elementos de um subconjunto qualquer de E, diga-se C, são eventos estatisticamente independentes quando conhecidos os resultados de um outro subconjunto de E, diga-se O, a probabilidade da ocorrência conjunta dos eventos de C, dado que se sabe sobre a ocorrência dos resultados dos eventos de O, iguala-se ao produto das probabilidades condicionais dos eventos de C, dados os eventos de O. Ou, em notação probabilística,* $P\left(\bigcap_i A_i \mid O\right) = \prod_i P(A_i \mid O)$.

Existe uma grande contribuição dos conceitos de probabilidades condicionais acerca da representação matemática da causalidade. Pode-se começar pela apresentação dos seguintes teoremas:

Teorema 2.1 (Regra do Produto) *Seja C um subconjunto de eventos de E, onde* $C = C_1, C_2, ..., C_n$. $P(C_1 \cap C_2 \cap ... \cap C_n) = P(C_1)P(C_2|C_1)...P(C_n|C_1, C_2, ..., C_{n-1})$.

Teorema 2.2 (Regra de Bayes) *Seja C um subconjunto de eventos de E, onde* $C = C_1, C_2, ..., C_n$. $P(C_j) \neq 0$. *i, j = 1, 2, ..., n:* $P(C_i|C_j) = P(C_i)P(C_j|C_i)/P(C_j)$.

A regra do produto é o principal fundamento de MBs e traduz que a probabilidade da ocorrência conjunta de um subconjunto de eventos pode ser desencadeada em um produto de probabilidades condicionais. Isto tem grande importância, pois permite que um MB com *n* dimensões (a probabilidade conjunta) seja não só modelado mas também manipulado a partir de *n* funções unidimensionais mais acessíveis (as probabilidades condicionais). Desta regra pode-se derivar a de Bayes, que permite que dada a ocorrência de um subconjunto de eventos, se redistribuam as probabilidades dos seus potenciais eventos causadores. As distribuições de probabilidades componentes da regra de Bayes são comumente chamadas de a *priori* [$P(A)$], verossimilhança [$P(B|A)$] e a *posteriori* [$P(A|B)$]. A probabilidade a *priori* é comumente adotada para refletir as incertezas iniciais acerca da ocorrência de *A*, enquanto que a regra de Bayes executa uma tarefa árdua para a mente humana: um raciocínio matematicamente (artificialmente) racional para a atualização desta incerteza inicial, refletido na *posteriori*. Uma vez conhecidos $P(A)$ e $P(B|A)$, tem-se em mãos o primeiro e mais trivial de todos os MBs: a caracterização da regra de Bayes envolvendo dois eventos. MBs mais complexos e até intratáveis matematicamente são facilmente obtidos quando variáveis aleatórias são introduzidas.

2.3.3. Variáveis Aleatórias

Até aqui, as probabilidades têm sido dirigidas a eventos. Porém, para a elaboração de MBs mais sofisticados emerge a necessidade do uso da função variável aleatória. Uma variável aleatória associa a cada possível resultado do experimento um número real e permite, além de uma caracterização mais refinada sobre o fenômeno de interesse, a utilização irrestrita da matemática através do uso das operações aritméticas, de igualdades, desigualdades e do cálculo diferencial e integral, entre outros.

Definição 2.3 *Uma variável aleatória é uma função cujo argumento é um elemento de U e o resultado é um número real.*

De maneira geral, uma variável aleatória pode ser classificada como discreta ou contínua. Uma variável discreta permite que seus possíveis resultados sejam enumerados, assumindo valores na escala dos números inteiros. Complementarmente, variáveis ditas contínuas assumem valores na escala dos números reais. Recomenda-se Ross (2000) para maiores detalhes. De qualquer forma, a função que descreve as incertezas sobre uma dada variável aleatória X discreta ou contínua é denotada no presente trabalho por distribuição de probabilidades de X, $f(x)$, independente de esta ser uma função de massa (direcionada a variáveis discretas) ou de densidade (direcionada a variáveis contínuas) de probabilidade e satisfaz às seguintes condições:

Teorema 2.3 *A função $f(x)$ é a distribuição de probabilidades da variável aleatória X se e só se:*

i. $f(x) \geq 0$, para qualquer x real;
ii. $\sum_x f(x) = 1$, se X for discreta, e $\int_x f(x)dx = 1$, se X for contínua.

O exemplo a seguir ilustra como a introdução de variáveis aleatórias permite uma modelagem mais rica sobre dado fenômeno e quais desdobramentos matemáticos são adotados de maneira a alcançar distribuições a *posteriori*.

Exemplo 2.1 *Seja P a probabilidade subjacente à ocorrência de dado evento de interesse A e seja X o nº de vezes em que A ocorre em n experimentos aleatórios. Como atualizar o estado de conhecimento sobre P a partir de evidências sobre X?*

O Exemplo 2.1 pode ser visualizado como um MB trivial envolvendo duas quantidades, a probabilidade P contínua compreendida no intervalo [0, 1] e o nº de ocorrências X discreto pertencente ao intervalo [0, n]. Supondo $P\sim$ Beta (a, b) e $[X|P=p]\sim$ Binomial(n, p), para computar a distribuição a *posteriori* $f(p|x)$ tem-se:

$f(p \mid x) = \dfrac{f(p) f(x \mid p)}{f(x)}$ pela regra de Bayes, onde $f(p) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$,

$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$ e $f(x) = \int_p f(p,x)dp = \int_p f(p) f(x \mid p)dp$. Como

$\binom{n}{x} = \dfrac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)}$, tem-se que

$f(x) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \dfrac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \int_p p^{a+x-1}(1-p)^{b+n-x-1} dp$. Note-se que

$\int_p p^{a+x-1}(1-p)^{b+n-x-1} dp =$

$$= \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} \int_p \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)} p^{a+x-1}(1-p)^{b+n-x-1} dp \text{ e que}$$

$\int_p \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)} p^{a+x-1}(1-p)^{b+n-x-1} dp = 1$, já que se trata de uma Beta($a+x$, $b+n-x$) integrada no seu suporte (Teorema 2.3, item ii), levando a

$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)}$. Logo,

$$f(p \mid x) = \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} p^{a+x-1}(1-p)^{b+n-x-1}}{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)}}, \text{ conduzindo em fim a}$$

$$f(p \mid x) = \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)} p^{a+x-1}(1-p)^{b+n-x-1}$$

∎

Assim, conclui-se que diante da evidência $X=x$, a distribuição a *posteriori* de $[P|x]$ é Beta($a+x$, $b+n-x$). Note-se que tal resultado é devido às distribuições adotadas possibilitarem desdobramentos analíticos no cômputo da *posteriori*. De fato, a Beta e a Binomial são distribuições ditas conjugadas. Nesses casos, operações matemáticas tornam possível a obtenção de distribuições condicionais, tais como $f(p|x)$. Porém, a depender da definição das distribuições a *priori* de P e da verossimilhança de $[X|P=p]$, a obtenção analítica de $f(p|x)$ torna-se inviável. Caso, por exemplo, $f(p)$ seja uma distribuição triangular, ou mesmo delineada a partir de um método indireto de edução do conhecimento tal como o proposto por Firmino *et al.* (2008), as integrais envolvidas podem não ter solução fechada. Para maiores detalhes recomenda-se Bolstad (1943). Nestas situações restam duas alternativas: ou molda-se a realidade a um MB analiticamente tratável ou recorre-se a técnicas de inferência aproximadas. Busca-se no presente trabalho estudar alternativas para a segunda opção.

Pode-se notar que no problema descrito acima, o MB delineado representa a distribuição de probabilidades conjunta, $f(P=p, X=x)$, pelo produto das distribuições $f(p)$ e $f(p|x)$. De fato, com o decorrer do trabalho poder-se-á concluir que em geral MBs são descritos por um conjunto de distribuições condicionais cujo produto é proporcional à distribuição de probabilidades conjunta das variáveis envolvidas.

A seguir, alguns dos principais formalismos da inferência Bayesiana são introduzidos. Tais modelos vêm sendo adotados em diversas áreas do conhecimento. Como referências, Hamada *et al.* (2008) – na área de análise de confiabilidade, Clark & Gelfand (2006) – nas ciências

ambientais, Congdon (2003) e Congdon (2006) – em aplicações diversas – e Congdon (2005) – especificamente voltado para a modelagem envolvendo dados categóricos – podem ser citados.

2.4. Modelos Hierárquicos

Modelos hierárquicos são um dos mais importantes formalismos da inferência Bayesiana, ocupando lugar de destaque em áreas tais como nas ciências ambientais, como constatado por Clark & Gelfand (2006). Conceitualmente, Bernardo & Smith (1995) enfatizam que eles tornaram-se também conhecidos como métodos de Bayes empírico por calibrarem as crenças sobre parâmetros diante de dados gerados pelo mecanismo probabilístico neles baseado. Assim, o problema ilustrado no Exemplo 2.1 pode ser entendido como um modelo hierárquico. Em resumo, modelos hierárquicos têm a seguinte forma:

$$f(\underset{\sim}{X} = \underset{\sim}{x} \mid \underset{\sim}{\Lambda} = \underset{\sim}{\lambda}) = f(x_1, x_2, ..., x_k \mid \lambda_1, \lambda_2, ..., \lambda_k) = \prod_{i=1}^{k} f(x_i \mid \lambda_i) \qquad \textit{Equação 2.1}$$

$$f(\Lambda^* = \lambda^*, \underset{\sim}{\Lambda} = \underset{\sim}{\lambda} \mid \underset{\sim}{\Theta} = \underset{\sim}{\theta}) = f(\lambda^*, \lambda_1, \lambda_2, ..., \lambda_k \mid \underset{\sim}{\theta}) = f(\lambda^* \mid \underset{\sim}{\theta}) \prod_{i=1}^{k} f(\lambda_i \mid \underset{\sim}{\theta}) \qquad \textit{Equação 2.2}$$

$$f(\underset{\sim}{\Theta} = \underset{\sim}{\theta}) = \prod_{j=1}^{s_m} f(\theta_{mj}), \qquad \textit{Equação 2.3}$$

onde as variáveis observáveis são $X_1, X_2, ..., X_k$, e os parâmetros Λ_i's podem resumir fenômenos probabilísticos de populações locais geradores das realizações x_i's. A quantidade Λ^* representa a métrica genérica de interesse, enquanto que $\underset{\sim}{\Theta}$ pode expressar o conjunto de parâmetros populacionais que regem probabilisticamente os Λ_i's e também Λ^*. Quando observa-se um conjunto de evidências $x_1, x_2, ..., x_k$, tanto o conjunto de taxas (Λ_i's e Λ^*) quanto $\underset{\sim}{\Theta}$ têm seus graus de incerteza atualizados e representados através das suas distribuições a *posteriori*, calculadas segundo a regra de Bayes (Teorema 2.2). Da Equação 2.1 percebe-se a suposição de independência entre as variáveis observáveis uma vez fixados seus parâmetros subpopulacionais

Logo, modelos hierárquicos são especialmente úteis para problemas de inferência populacional diante de evidências sobre subpopulações não-homogêneas. Um exemplo clássico proveniente da análise de confiabilidade de equipamentos consiste em tentar inferir sobre métricas de confiabilidade de um dado modelo de equipamento que é indiscriminadamente posto em execução em subpopulações cujas condições operacionais e/ou ambientais são diferentes. Tem-se neste contexto informações subpopulacionais em termos de

nº de falhas por ano (ou por quantidade de demandas) ou tempos entre falhas, embora se deseje medir o desempenho do equipamento de maneira genérica; em outras palavras, os dados existem e são obtidos das subpopulações. Pode-se citar como exemplo um problema estudado por muitos autores, entre eles George *et al.* (1993):

Exemplo 2.2 *Trata-se da modelagem estatística de um banco de dados de confiabilidade a partir de um conjunto de 10 bombas postas em operação em uma planta de geração de energia nuclear em condições heterogêneas. Para cada bomba, o nº de falhas em dado tempo de missão,* $N_i(t)$*, é regido por uma taxa de falhas,* L_i ($i = 1, 2, ..., 10$)*, de forma que* $[N_i(t_i)|L_i=l_i] \sim Poisson(l_i t)$*. Por sua vez, as taxas de falhas* L_i *podem ser governadas por um conjunto de parâmetros populacionais* (A, B) *onde* $[L_i|A=a, B=b] \sim Gamma\ (a, b)$*. A priori atribui-se uma distribuição Exponencial*(*1.0*) *para A e uma Gamma* (*0.1, 1.0*) *para B, supostos independentes.*

Note-se que aqui os únicos fenômenos observáveis são os $N_i(t)$ e inferências populacionais tornam-se possíveis neste caso a partir da introdução no MB de uma taxa de falhas genérica L^*, onde $[L^*|a, b] \sim$ Gamma (a, b).

2.5. Metodologia de Análise de Variabilidade Populacional

A análise de variabilidade populacional Bayesiana, também conhecida como o primeiro em dois estágios de atualização Bayesiana ou Bayes hierárquico, é um modelo hierárquico direcionado à inferência sobre métricas de desempenho de equipamentos a partir de várias fontes de informação empíricas e subjetivas direta ou indiretamente relacionadas ao fenômeno de interesse. Como pode ser visto em Droguett *et al.* (2004), esta metodologia é originalmente dedicada à mensuração das incertezas inerentes a parâmetros populacionais de confiabilidade tais como a probabilidade de falha na demanda ou a taxa de falhas de dado componente e podem ser úteis, por exemplo, como distribuições a *priori* genéricas quando do estudo de cenários cuja escassez de dados predomina. Uma das idéias subjacentes à metodologia é a de que itens similares em algum nível de detalhamento trazem consigo informações ao menos parcialmente relevantes acerca do desempenho do equipamento em questão, assim como experiências vividas em situações semelhantes por projetistas, desenvolvedores ou operadores. Logo, como é possível associar uma taxa de falhas ou uma probabilidade de falha na demanda a um dado componente ou sistema de uma população em específico, pode-se também estudar tais parâmetros em relação à superpopulação da qual os componentes fazem parte. Como resultado, tem-se um modelo probabilístico que representa a variabilidade desta população.

Formalmente, as relações probabilísticas do modelo podem ser dadas por

$$f(\underset{\sim}{X}=\underset{\sim}{x},\underset{\sim}{Y}=\underset{\sim}{y}\mid\Lambda=\lambda)=$$

$$f(x_1,x_2,...,x_k,y_1,y_2,...,y_r\mid\lambda)=\prod_{i=1}^{k}f(x_i\mid\lambda)\prod_{j=1}^{r}f(y_j\mid\lambda) \qquad \textit{Equação 2.4}$$

$$f(\Lambda=\lambda\mid\underset{\sim}{\Theta}=\underset{\sim}{\theta})=f(\lambda\mid\underset{\sim}{\theta}) \qquad \textit{Equação 2.5}$$

$$f(\underset{\sim}{\Theta}=\underset{\sim}{\theta})=\prod_{j=1}^{p}f(\theta_j), \qquad \textit{Equação 2.6}$$

onde as variáveis observáveis são (X_1, X_2,..., X_k) e (Y_1, Y_2,..., Y_r). O primeiro conjunto pode se referir a evidências empíricas (nº de falhas por unidade de tempo ou por nº de oportunidades) enquanto que o segundo a evidências subjetivas (estimativas segundo especialistas). A partir da Equação 2.4 percebe-se a propriedade de independência condicional entre as evidências uma vez fixado o parâmetro de confiabilidade populacional Λ, por sua vez governado por hiperparâmetros $\underset{\sim}{\Theta}$. As distribuições a *priori* de $\underset{\sim}{\Theta}$ podem incorporar as crenças iniciais do analista. Assim, tem-se uma alternativa para combinar formalmente (e racionalmente) todas as fontes de informação de forma a inferir sobre a distribuição de probabilidades de Λ. A seguir tem-se um exemplo extraído de Droguett *et al.* (2004).

Exemplo 2.3 *Deseja-se modelar as incertezas envolvendo a taxa de falhas de dado modelo de bomba, Λ. Tem-se em mãos o nº observado de falhas deste modelo em 8 plantas em seus respectivos tempos de missão, $N_i(t_i)=n_i(t_i)$ (i = 1, 2, ...,8), e 4 estimativas provenientes de especialistas, $E_j=e_j$ (j = 1, 2, 3, 4). Supõe-se que $[N_i(t_i)|\Lambda=\lambda]\sim$ Poisson(λt_i) e $[E_j|\Lambda=\lambda]\sim$ Lognormal$(ln\ \lambda, s_j^2)$, onde s_j implica em uma medida de dispersão sobre E_j inversamente proporcional à credibilidade que o analista atribui à opinião emitida pelo j-ésimo especialista. Por fim, o mecanismo probabilístico que rege a taxa de falhas populacional é sintetizado por dois parâmetros também incertos, A e B, tais que $[\Lambda|A=a, B=b]\sim$ Lognormal (a, b^2). O estado de conhecimento inicial do analista sobre o problema indica que $A\sim$ Lognormal(μ_A, σ_A^2) e $B\sim$ Lognormal(μ_B, σ_B^2) a priori, onde μ_A, σ_A, μ_B e σ_B são constantes conhecidas.*

2.6. Metodologia de Incerteza de Modelos

Em geral, modelos matemáticos são uma simplificação da realidade sob estudo, sendo por muitas vezes adequada a introdução de um erro aleatório que reflita a parcela da realidade cujo controle é inviável. Como exemplo, recomenda-se Casella & Berger (2001) para um apanhado geral sobre modelos de regressão. Contudo, uma vez definido o modelo, pode ser de

interesse o seu uso abstraindo-se da sua elaboração. Neste contexto, é possível que as premissas intrínsecas à modelagem em si sejam apenas parcialmente verdadeiras quando da adoção do modelo em um cenário específico, apresentando um comportamento para o erro de predição eventualmente diferente do original. Por outro lado, pode também haver mais do que um modelo total ou parcialmente aplicável a dado fenômeno de interesse, tornando necessária uma metodologia que trate das várias alternativas de maneira acoplada quando da inferência sobre o fenômeno de interesse. Comente-se neste sentido estimativas provindas de opiniões de especialistas, interpretados como modelos de predição.

A metodologia de incerteza de modelos introduz um formalismo para lidar com o problema descrito acima. O desafio passa a ser a mensuração dos erros provenientes dos modelos no contexto do objeto de estudo. Droguett & Mosleh (2008) trabalham com uma variante baseada em dados de desempenho envolvendo as estimativas dos modelos e os respectivos valores observados (verdadeiros), possibilitando o estudo do comportamento do erro no contexto em questão. Formalmente, tem-se

$$f(\underset{\sim}{E_1}=\underset{\sim}{e_1},...,\underset{\sim}{E_r}=\underset{\sim}{e_r} \mid \underset{\sim}{\Lambda_1}=\underset{\sim}{\lambda_1},...,\underset{\sim}{\Lambda_r}=\underset{\sim}{\lambda_r}) =$$

$$f(e_{11},...,e_{1k_1},e_{21},...,e_{rk_r} \mid \underset{\sim}{\lambda_1},...,\underset{\sim}{\lambda_r}) = \prod_{m=1}^{r}\prod_{i=1}^{k_m} f(e_{mi} \mid \underset{\sim}{\lambda_m}) \qquad \textit{Equação 2.7}$$

$$f(\underset{\sim}{U}=\underset{\sim}{u} \mid \underset{\sim}{\Lambda_1}=\underset{\sim}{\lambda_1},...,\underset{\sim}{\Lambda_r}=\underset{\sim}{\lambda_r}, U^*=u^*) =$$

$$= f(u_1,u_2,...,u_r \mid \underset{\sim}{\lambda_1},...,\underset{\sim}{\lambda_r},u^*) = \prod_{m=1}^{r} f(u_m \mid \underset{\sim}{\lambda_m},u^*) \qquad \textit{Equação 2.8}$$

$$f(\underset{\sim}{\Lambda_1}=\underset{\sim}{\lambda_1},...,\underset{\sim}{\Lambda_r}=\underset{\sim}{\lambda_r} \mid \underset{\sim}{\Theta_1}=\underset{\sim}{\theta_1},...,\underset{\sim}{\Theta_r}=\underset{\sim}{\theta_r}) = \prod_{m=1}^{r}\prod_{j=1}^{p_m} f(\lambda_j \mid \underset{\sim}{\theta_m}) \qquad \textit{Equação 2.9}$$

$$f(U^*=u^*,\underset{\sim}{\Theta_m}=\underset{\sim}{\theta_m}) = f(u^*)\prod_{j=1}^{p_m} f(\theta_j), \qquad \textit{Equação 2.10}$$

onde as variáveis observáveis são (E_{11}, E_{12}, ..., E_{1k_1}, ..., E_{r1},..., E_{rk_r}) e (U_1, U_2,..., U_r). O primeiro conjunto refere-se aos erros cometidos por cada um dos r modelos nas k_m oportunidades em que pode-se confrontar suas estimativas com os respectivos valores reais. Já o segundo conjunto associa-se às predições dos modelos envolvidos acerca da quantidade sob estudo, U^*. A partir da Equação 2.7 percebe-se que uma vez conhecidos os parâmetros que sintetizam o mecanismo probabilístico subjacente aos erros dos modelos $\left(\underset{\sim}{\Lambda_m}=\underset{\sim}{\lambda_m}, m=1,...,r\right)$, estes se tornam independentes. O mesmo ocorre com as predições

diante de valores para U^* e $\underset{\sim}{\Lambda}_m$; configura-se assim a suposição de independência entre os modelos. Da Equação 2.9 pode-se modelar problemas envolvendo dados de desempenho não-homogêneos, onde os erros de um dado modelo podem ser supostos como provenientes da mesma família de distribuições, porém com parâmetros diferentes caracterizando a heterogeneidade das subpopulações; um problema de variabilidade populacional tal como descrito na seção anterior. Assim, dados homogêneos podem ser modelados a partir da atribuição de distribuições Delta de Dirac às funções envolvidas na Equação 2.10, enquanto que uma distribuição Uniforme pode refletir dados totalmente heterogêneos.

De maneira a viabilizar um tratamento matemático ao menos parcial, Droguett & Mosleh (2008) consideram tanto os erros quanto as predições como provenientes de uma distribuição Lognormal ou Normal. Para o caso Lognormal, $[E_{mj}| A_m= a_m, B_m= b_m]$~ Lognormal ($log\ a_m$, b_m^2), $[U_m| A_m= a_m, U^*= u^*, B_m= b_m]$~ Lognormal ($log\ u^* + log\ a_m$, b_m^2), $[A_m| T_m=t_m, V_m=v_m]$~ Lognormal (t_m, v_m^2) e $[B_m| W_m=w_m, Z_m=z_m]$~ Lognormal (w_m, z_m^2). Para o caso Normal, $[E_{mj}| A_m= a_m, B_m= b_m]$~ Normal ($a_m$, b_m^2) e $[U_m| A_m= a_m, U^*= u^*, B_m= b_m]$~ Normal ($u^* + a_m$, b_m^2), $[A_m| T_m=t_m, V_m=v_m]$~ Normal (t_m, v_m^2) e $[B_m| W_m=w_m, Z_m=z_m]$~ Lognormal (w_m, z_m^2). Vale ressaltar que as evidências disponíveis influenciam as incertezas sobre a quantidade de interesse U^* a partir do mecanismo probabilístico que rege as predições, $[U_m| A_m= a_m, U^*= u^*, B_m= b_m]$. As distribuições a *priori* de T_m, V_m, W_m e Z_m refletem o estado de conhecimento do analista sobre o quão heterogêneos são os dados de desempenho do modelo; quanto maior a heterogeneidade, mais próxima da distribuição Uniforme deve ser a *priori* dos parâmetros populacionais. Tem-se a seguir um problema inspirado em Droguett & Mosleh (2008).

Exemplo 2.4 *Deseja-se modelar as incertezas envolvendo a temperatura de dada película protetora de um conjunto de cabos exposta ao fogo em ambiente fechado,* U^*. *Para tanto, tem-se em mãos as predições de dois modelos,* u_1 *do modelo* M_1 *e* u_2 *de* M_2, *e 50 dados de desempenho de ambos os modelos. Testes de aderência indicam que os erros provenientes de* M_1 *seguem uma distribuição Normal enquanto que os de* M_2 *seguem uma Lognormal. Devido à escassez de informações e razoável variabilidade dos erros de predição, supõe-se heterogeneidade destes em relação a* M_1 *e* M_2. *Probabilisticamente, tem-se que para* M_1: $[E_{1i}| A_1= a_1, B_1= b_1]$~ *Normal* ($a_1$, b_1^2), $i = 1, 2, ..., 50$. $[U_1| U^*= u^*, A_1= a_1, B_1= b_1]$~ *Normal* ($u^* + a_1$, b_1^2), $[A_1| T_1=t_1, V_1=v_1]$~ *Normal* (t_1, v_1^2) *e* $[B_1| W_1=w_1, Z_1=z_1]$~ *Lognormal* (w_1, z_1^2). *Para* M_2: $[E_{2j}| A_2= a_2, B_2= b_2]$~ *Lognormal* ($log\ a_2$, b_2^2), $j = 1, 2, ..., 50$. $[U_2| U^*= u^*, A_2= a_2, B_2= b_2]$~ *Lognormal* ($log\ u^* + log\ a_2$, b_2^2), $[A_2| T_2=t_2, V_2=v_2]$~ *Lognormal* (t_2, v_2^2) *e* $[B_2| W_2=w_2,$

$Z_2=z_2$]~ Lognormal (w_2, z_2^2). O estado de conhecimento inicial do analista sobre o problema leva a T_k, V_k, W_k, Z_k serem distribuídos uniformemente a priori, onde k=1, 2.

Nestas últimas seções, pôde-se perceber características marcantes de MBs tanto conceitual quanto matematicamente. Conceitualmente, um dos principais interesses de MBs recai sobre o problema de como inferir sobre uma quantidade latente a partir de evidências regidas por fenômenos probabilísticos nela baseados; em outras palavras, enfatiza-se inferências diagnosticas. Matematicamente, MBs podem ser resumidas como um conjunto de distribuições de probabilidades condicionais cujo produto é proporcional à distribuição conjunta das variáveis envolvidas. Assim, torna-se imperativa uma leitura adequada das distribuições condicionais que "codificam" as nuances da realidade modelada, mesmo que isso seja uma tarefa árdua e por muitas vezes inviável para os mais leigos.

Contudo, pode ser também de extremo interesse inferir prognosticamente, mensurando incertezas sobre efeitos. Além disso, a possibilidade de visualizar o MB descrito matematicamente pelas distribuições condicionais pode facilitar não apenas a sua leitura mas também adaptações e mesmo sua elaboração. Tais características naturalmente tornariam a abordagem Bayesiana mais rica e acessível a um maior público e isto de fato ocorreu com a introdução do formalismo de RBs. O passo fundamental de RBs em relação aos demais formalismos Bayesianos apresentados até aqui é o de associar um grafo às distribuições condicionais. Isto possibilita que a elaboração do MB seja iniciada graficamente. Tal visualização das relações causais permite uma modelagem mais intuitiva e sistemática e amplia os horizontes de aplicação de MBs às diversas áreas do conhecimento. Assim, o conceito de causalidade pode ser compreendido de maneira mais clara do que aquele baseado na regência de mecanismos probabilísticos.

2.7. Redes Bayesianas

Percebe-se até aqui que a inferência Bayesiana é relativamente simples quando envolve poucas variáveis. Porém, quando a quantidade de variáveis se eleva, tal inferência torna-se complexa e por muitas vezes desmotivadora. É neste momento em que redes Bayesianas (RBs) se inserem ao problema da inferência Bayesiana, através do relacionamento entre distribuições de probabilidade e grafos, denotado por condição Markoviana.

De maneira a definir a condição Markoviana, faz-se necessária uma introdução à teoria dos grafos. Um grafo direcionado (DG) é um par (V, B), onde V é um conjunto não vazio cujos elementos são chamados de nós (ou vértices) e B é um conjunto de pares ordenados de elementos distintos de V. Os elementos de B são denominados arestas (ou arcos) e se um par

$(Y, X) \in B$, existe um arco de Y para X. Seja um conjunto de nós, $(X_1, X_2, ..., X_k)$, onde $k \geq 2$, tal que $(X_{i-1}, X_i) \in B$, para $2 \leq i \leq k$. O conjunto de arcos conectando os k nós é chamado de caminho de X_1 a X_k e é denotado por $[X_1, X_2, ..., X_k]$. Um grafo acíclico direcionado (DAG) é um grafo direcionado que não possui ciclos, isto é, seus arcos são unidirecionais de forma que partindo-se de qualquer um dos elementos de V é impossível que se retorne ao mesmo. Considere-se aqui que caminho seja lido com o mesmo sentido de caminho unidirecional.

Dados um DAG, $G = (V,B)$, e um par $(Y, X) \in B$, Y é pai de X e X é filho de Y se houver um arco de Y para X. Mais genericamente, Y é ancestral de X e X é descendente de Y se houver um caminho de Y a X. Um nó é chamado de raiz caso não possua qualquer pai. Uma cadeia tem a mesma notação de um caminho, porém representa um fluxo adirecional, ou seja, despreza a direção dos arcos do DAG. Estes conceitos foram extraídos de Neapolitan (2004) e divergem dos conceitos apresentados por outros autores tais como Korb & Nicholson (2003), Pearl (1988) e Edwards (1949), que chamam caminhos e cadeias de caminhos unidirecionais e adirecionais, respectivamente, ou simplesmente caminhos.

A compreensão destes conceitos torna-se simples através de um exemplo. Na Figura 2.2(b), tem-se os seguintes casos: Z é pai e não descendente de X e X é filho e descendente de Z; Z é ancestral e não descendente de W e W é descendente de Z; Z é a única raiz da rede; há apenas um caminho entre Z e W, $[Z, X, W]$; identifica-se apenas uma cadeia entre Z e W, $[Z, X, W] = [W, X, Z]$ e, por fim, o grafo apresentado caracteriza-se como um DAG, pois partindo-se de Z chega-se a X ou a W, saindo-se de X finda-se em W e W não permite qualquer passagem, assim, não há caminhos que permitam o retorno ao nó inicial, seja este Z, X ou W. Os demais grafos da figura serão utilizados ao longo do texto.

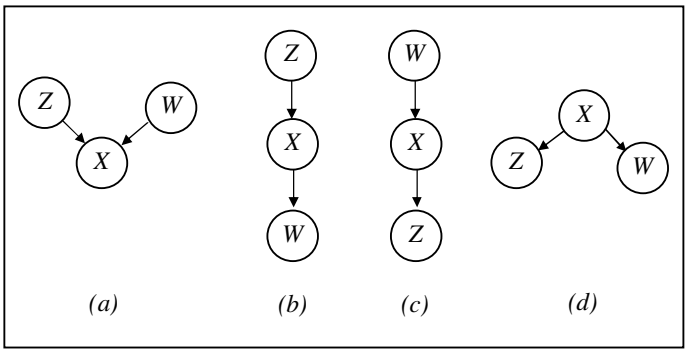


Figura 2.2 DAGs e redes Bayesianas.

2.7.1. Condição Markoviana

A principal característica de processos Markovianos diz respeito à suposição de falta de memória. Quando se sabe sobre o atual estado do processo, informações anteriores são

irrelevantes na quantificação das incertezas envolvendo seus estados futuros (ver capítulo 3). Este é o conceito da condição Markoviana.

Definição 2.4 (Condição Markoviana) *Seja P a distribuição de probabilidades conjunta das variáveis aleatórias em algum conjunto V e seja um DAG G = (V, B). É dito que o par (G, P) satisfaz à condição Markoviana se para cada X ∈ V, X é condicionalmente independente de todas as variáveis que compõem o conjunto dos seus não descendentes, dados os valores das variáveis que compõem seu conjunto de pais. Ou, em notação probabilística, X ind ND(X)|PA(X), onde ND(X) refere-se ao conjunto de não descendentes de X e PA(X) ao seu conjunto de pais.*

Se um nó é raiz, então a condição Markoviana o considera independente dos seus não-descendentes.

Considerando que os nós dos grafos da Figura 2.2 representam variáveis aleatórias, tem-se pela regra do produto que a distribuição de probabilidades conjunta pode ser dada por $P = P(X=x, Z=z, W=w) = f(x)f(z|x)f(w|x, z) = f(x)f(w|x)f(z|x, w) = ... = f(w)f(z|w)f(x|z, w)$. Porém, caso Z *ind* $W|X$ então $f(w|x, z) = f(w|x)$ e $f(z|x, w) = f(z|x)$ e uma associação entre os grafos e a Definição 2.4 leva a concluir que a condição Markoviana é satisfeita em (b), (c) e (d). Em (b) a condição Markoviana é satisfeita, pois dados os pais de W (a variável X) esta se torna independente dos seus não-descendentes (a variável Z), $P = f(z)f(x|z)f(w|x)$. Em (c) e (d) tem-se algo semelhante; uma vez fixados os pais de Z esta se torna independente dos seus não descendentes. Por outro lado, o grafo na Figura 2.2 (a) não satisfaz à condição Markoviana caso Z *ind* $W|X$, pois deste grafo lê-se que Z *ind* W sem a necessidade de que X seja evidenciado.

A condição Markoviana aliada à regra do produto (Teorema 2.1) facilita a obtenção da distribuição de probabilidades conjunta das variáveis do grafo, igualando-a ao produto das distribuições de probabilidades condicionais das mesmas, em relação aos valores dos seus pais. Se na Figura 2.2 (a), deseja-se computar a distribuição de probabilidades conjunta das variáveis do grafo, P, tem-se $P = P(Z=z, W=w, X=x) = f(z)f(w|z)f(x|w,z)$. Como Z e W são nós raiz e conseqüentemente não descendentes um do outro, $f(w|z) = f(w)$ e $P = f(z)f(w)f(x|w, z)$. Este é um teorema que se apresenta da seguinte forma:

Teorema 2.4 *Seja P a distribuição de probabilidades conjunta das variáveis aleatórias em algum conjunto V e seja um DAG G = (V, B). Se o par (G, P) satisfaz à condição Markoviana, então P equivale ao produto das distribuições condicionais de todos os nós, dados os valores dos seus pais, sempre que estas distribuições condicionais existirem.*

O teorema acima diz apenas que uma distribuição conjunta *P* satisfazendo à condição Markoviana com algum DAG *G* pode ser expressa pelo produto das distribuições condicionais das variáveis em relação a seus pais em *G*. Contudo, o interesse maior é iniciar o processo com as distribuições condicionais e então concluir que seu produto é uma distribuição conjunta *P* satisfazendo à condição Markoviana com algum DAG *G*. Neste sentido, seria possível delinear um grafo adequado ao MB e concluir sobre sua distribuição conjunta a partir das condicionais. O teorema a seguir é útil neste sentido.

Teorema 2.5 *Seja um DAG G tal que cada nó é uma variável aleatória discreta para a qual tem-se uma distribuição de probabilidades condicional a seus pais. Então o produto dessas distribuições condicionais equivale à distribuição conjunta dessas variáveis, P, e* (*G, P*) *satisfaz à condição Markoviana.*

Embora o Teorema 2.5 seja relacionado a problemas envolvendo variáveis aleatórias discretas, autores como Neapolitan (2004) enfatizam que ele é notadamente verificado em casos envolvendo variáveis contínuas. No decorrer do presente trabalho, ilustra-se que nos MBs previamente introduzidos isto de fato ocorre.

Neste momento, se está a um passo da definição de RBs a partir da condição Markoviana:

Definição 2.5 *Seja P uma distribuição de probabilidades conjunta das variáveis aleatórias de dado conjunto V e G* = (*B, V*) *um DAG, se* (*G, P*) *satisfaz à condição Markoviana então trata-se de uma RB.*

A partir do Teorema 2.4, *P* é o produto de suas distribuições condicionais em *G* e essa é a maneira pela qual *P* é representada em uma RB. Considerando o Teorema 2.5, uma vez especificado um DAG *G* e um conjunto de distribuições de probabilidades condicionais, tem-se uma RB. Assim, pode-se definir alternativamente uma RB como um DAG onde os nós representam variáveis aleatórias e as conexões direcionadas expressam as relações de causa e efeito entre tais variáveis. Neste contexto, o poder das relações causais inerentes ao problema modelado é descrito por distribuições de probabilidades condicionais sobre cada variável da rede, dados específicos valores do seu conjunto de pais. Ressalte-se, contudo, que caso dada distribuição a *priori* do MB (nó sem pais, ou raiz, no grafo) seja imprópria (não integre 1) uma normalização far-se-á necessária, ou em tal distribuição ou nas distribuições marginais derivadas a fim de que o item ii do Teorema 2.3 seja satisfeito.

A seguir, apresenta-se os formalismos descritos em seções anteriores via RBs.

2.7.2. Modelos Hierárquicos por RBs

O DAG que quando associado ao conjunto de distribuições condicionais da Equação 2.1, Equação 2.2 e Equação 2.3 resulta na RB adequada a modelos hierárquicos, pode ser visualizado na Figura 2.3. Note-se que devido à condição Markoviana tem-se que X_i *ind* $X_j|\Lambda_i$, Λ_i *ind* $\Lambda_j|(\Theta_1, ..., \Theta_p)$ e Θ_i *ind* Θ_j, para $\forall$ i, j no respectivo espaço de possibilidades. Tais propriedades levam aos produtórios encontrados nas três equações.

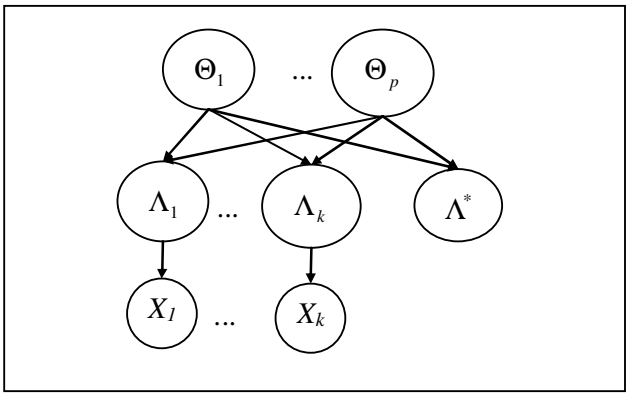


Figura 2.3 Rede Bayesiana que encapsula modelos hierárquicos.

2.7.3. Metodologia de Variabilidade Populacional por RBs

Considerando agora a RB que molda a metodologia de variabilidade populacional, chega-se ao DAG da Figura 2.4. Da condição Markoviana, lê-se que X_i *ind* $X_j|\Lambda$, Y_i *ind* $Y_j|\Lambda$, X_i *ind* $Y_j|\Lambda$, e que Θ_i *ind* Θ_j, para $\forall$ i, j no respectivo espaço de possibilidades. Logo, obtem-se os mesmos desdobramentos exibidos da Equação 2.4 até a Equação 2.6.

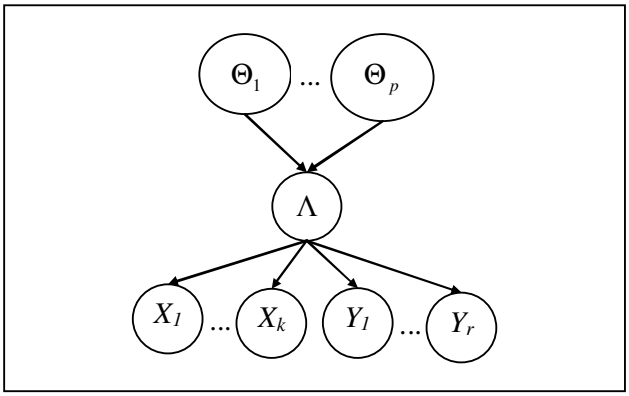


Figura 2.4 Rede Bayesiana que molda a metodologia de variabilidade populacional.

2.7.4. Metodologia de Incerteza de Modelos por RBs

O DAG associado à metodologia de incerteza de modelos é apresentado na Figura 2.5. Da condição Markoviana, lê-se por exemplo que E_{mj} *ind* $E_{mk}|(\Lambda_{m1}, ..., \Lambda_{mp_m})$, Λ_{mi} *ind* $\Lambda_{mj}|(\Theta_{m1}, ..., \Theta_{ms_m})$ e que Θ_{mi} *ind* Θ_{mj}, para $\forall$ *m*, *i*, *j* no respectivo espaço de possibilidades. Logo, obtem-se os mesmos desdobramentos exibidos da Equação 2.7 até a Equação 2.9.

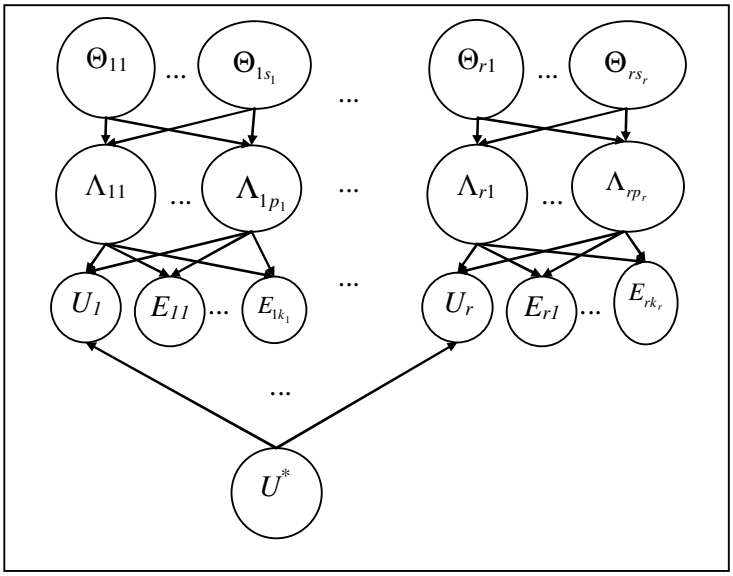


Figura 2.5 Rede Bayesiana que se estende à metodologia de incerteza de modelos.

Como dito anteriormente, além da adequação a outros formalismos da inferência Bayesiana, RBs permitem a modelagem da causalidade de maneira mais abrangente, possibilitando análises prognósticas além das diagnósticas. Para tanto vale revisitar o desenvolvimento analítico seguido para resolução do Exemplo 2.1, quando enfatizou-se a regra de Bayes. De fato, a regra de Bayes é útil em extrair distribuições marginais de variáveis causais quando as variáveis evidenciadas são seus efeitos. Quando o interesse recai sobre os efeitos opta-se por regras de marginalização. De qualquer forma, como se tem a distribuição de probabilidades conjunta, busca-se sempre marginalizar as demais variáveis resultando na distribuição da variável de interesse. Tal distribuição marginal pode ser condicionada às variáveis eventualmente evidenciadas. Assim, a partir deste momento distribuições a *posteriori* podem ser também denotadas por distribuições marginais. O exemplo a seguir serve para ilustrar os argumentos aqui apresentados:

Exemplo 2.5 *A qualificação dos profissionais responsáveis pela transformação de dada matéria-prima em produto* (*Z*) *influencia a qualidade do produto* (*X*) *que, conseqüentemente, afeta o índice de sua aprovação no mercado* (*W*). *A RB correspondente a este problema é tal qual o DAG da Figura 2.2*(*b*) *aliado às distribuições de probabilidades condicionais f*(*z*), *f*(*x*|*z*) *e f*(*w*|*x*).

De maneira a simplificar os cálculos, considera-se apenas duas categorias para as variáveis, favorável (0) ou desfavorável (1) para a empresa. Caso a diretoria da empresa deseje calcular as incertezas acerca da qualidade dos seus produtos (*X*):

$$f(x) = \sum_z \sum_w f(z,x,w) = \sum_z \sum_w f(z) f(x \mid z) f(w \mid x) = \sum_z f(z) f(x \mid z) \sum_w f(w \mid x).$$

Como $\sum_w f(w \mid x) = 1$, $f(x) = \sum_z f(z) f(x \mid z)$.

Agora, se for de interesse obter a distribuição marginal de *W*:

$f(w) = \sum_z \sum_x f(z,x,w) = \sum_z \sum_x f(z) f(x \mid z) f(w \mid x)$. A última igualdade decorre da condição Markoviana (*W ind Z*|*X*).

Seguindo o Exemplo 2.5, se a empresa desejar atualizar o seu estado de conhecimento em relação às variáveis que compõem o problema a partir da evidência de que seus empregados são bem qualificados (*Z*=0), as inferências prognósticas sobre [*X*|*Z*=*0*] são diretas a partir do grafo, *f*(*x*|*Z*=*0*). Já a predição sobre a variável *W* requer mais sofisticação.

$f(w \mid z) = \sum_x f(x,w \mid z) = \sum_x f(x \mid z) f(w \mid x)$. Note-se que esta última igualdade é novamente devido à condição Markoviana, de onde *W ind Z*|*X*.

A atualização a respeito das probabilidades das causas de um determinado efeito também pode ser realizada diante do conhecimento sobre tal efeito, a partir do teorema de Bayes e de regras de marginalização. A diretoria da empresa está diante de uma nova evidência, obteve-se um bom/mau índice de aceitação do produto no mercado ($W=w$). Assim:

$f(x \mid w) = f(x) f(w \mid x) / f(w)$, onde $f(x)$ e $f(w)$ são tais quais as funções obtidas anteriormente.

2.7.5. Algoritmos para a Obtenção de Distribuições Marginais

Observando a maneira pela qual as distribuições marginais das variáveis do exemplo utilizado no tópico anterior são obtidas, pode-se perceber a complexidade dos cálculos subjacentes a depender da topologia da rede. Contudo, devido às independências condicionais, permitindo que propagações de evidências globais possam ser feitas a partir de operações locais, Pearl (1988) descreve alguns dos primeiros métodos de manipulação de RBs. O primeiro deles permite a atualização em RBs singularmente conectadas (RSCs). Uma RSC possui a característica de existir no máximo uma cadeia entre dois nós quaisquer da rede. Quando uma RB não possui esta característica ela é chamada rede multiplamente conectada (RMC). Um exemplo de RSC é a rede associada à metodologia de variabilidade populacional (Figura 2.4), enquanto que aquela que engloba a de incerteza de modelos (Figura 2.5) ilustra RMCs.

A principal dificuldade de se inferir sobre RMCs é devido à presença de laços na atualização das crenças. Pearl (1988) comenta que se a presença dos laços é ignorada, as atualizações podem ser realizadas indefinidamente nestes e o processo não converge para um equilíbrio estável. Ele acrescenta que como o equilíbrio assintótico não é coerente, as probabilidades marginais de todas as variáveis da rede não são adequadamente representadas.

Sob esta ótica, Pearl (1988) sugere três métodos para a manipulação de laços: o agrupamento de variáveis, o condicionamento de variáveis e a simulação estocástica. Os dois primeiros foram originalmente elaborados para lidar com problemas envolvendo variáveis discretas com espaço de possibilidades reduzido e permitem uma inferência exata sobre as distribuições marginais, enquanto que o último realiza uma inferência aproximada via MCMC podendo ser utilizado para manipular RBs envolvendo variáveis de qualquer natureza.

A Figura 2.6 (a) mostra uma RMC, já que existem no mínimo duas cadeias entre os nós A e D, tais como [A, B, D] e [A, C, D]. O método dos agrupamentos cria uma variável Z referente à conjunção das variáveis B e C, tal qual a Figura 2.6 (b). Desta forma, quando se deseja avaliar $f(z|a)$ está-se na verdade trabalhando com $f(b, c|a)$ que, devido à condição

Markoviana, iguala-se a $f(b|a)\ f(c|a)$. Já quando o interesse recai sobre $f(e|z)$ tem-se $f(e|b, c)$ que resulta em $f(e|c)$.

O método dos condicionamentos elimina os laços através da fixação de variáveis localizadas estrategicamente na rede. Para cada valor da variável *A*, a rede da Figura 2.6 (a) torna-se singularmente conectada. Supondo que a variável *A* possui duas categorias, diga-se 0 e 1, as Figura 2.6 (c) e Figura 2.6 (d) exibem as suas respectivas RSCs. Ao final, as marginais se dão pela média ponderada dos cenários construídos. Se, por exemplo, deseja-se calcular $f(e)$, tem-se que

$$f(e) = \sum_{a=0}^{1} f(a,e) = \sum_{a=0}^{1} f(a) f(e \mid a), \text{ onde } f(e \mid a) = \sum_{c} f(c,e \mid a) = \sum_{c} f(c \mid a) f(e \mid c).$$

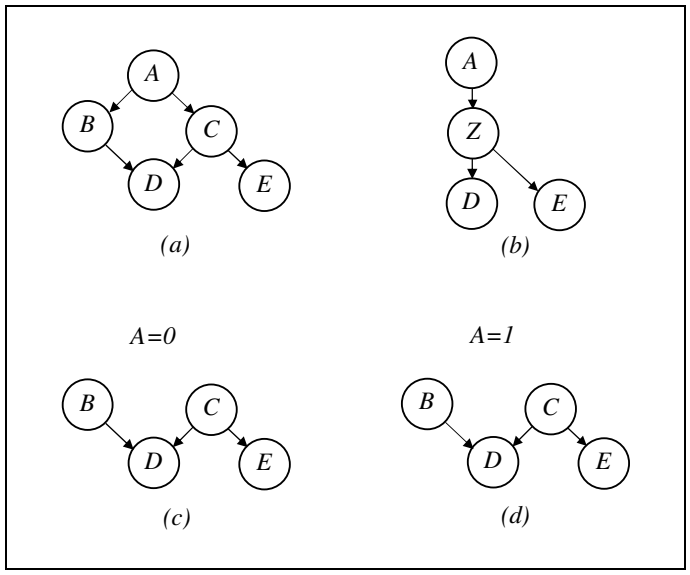


Figura 2.6 Métodos exatos para a marginalização em RMCs.

Pode-se citar alguns autores que se dedicam aos algoritmos de inferência exata. Suermondt & Cooper (1991) estudam o método dos condicionamentos. Diéz (1996) propõe uma otimização de tal método, denotando-a de condicionamento local. Já o método dos agrupamentos é estendido por Huang & Darwiche (1994) e Ladeira *et al.* (1999), por exemplo. Em trabalhos paralelos da área, pesquisadores vêm se debruçando sobre o problema de manipular RBs ditas mistas ou híbridas (que envolvem variáveis contínuas e discretas) através de métodos de inferência exata. A principal estratégia tem sido discretizar as variáveis cujo espaço de possibilidades é intratável pelos métodos de inferência exata antes de adotá-los. Kozlov & Koller (1997) apresentam as vantagens de se adotar uma discretização

dinâmica e não-uniforme que se adapte às características locais do MB bem como às evidencias atribuídas. Koller *et al.* (1999) propõem um hibridismo entre métodos de agrupamento e de simulação que generaliza algoritmos tais como o de discretização dinâmica de Kozlov & Koller (1997). Neil *et al.* (2007) estudam e implementam o método de discretização dinâmica de Kozlov & Koller (1997) através do pacote ©AgenaRisk e Neil *et al.* (2008) avaliam seu desempenho no contexto de análise de confiabilidade. Há, ainda, os métodos baseados em misturas de Exponenciais truncadas, que permitem uma generalização dos métodos de discretização no sentido de que ao invés de fazerem uso de funções lineares por partes em cada um dos intervalos definidos na discretização (aproximação usualmente adotada), recorrem a misturas de Exponenciais. Como trabalhos nesta vertente, Rumí *et al.* (2006) e Rumí & Salmerón (2007) podem ser citados.

Na prática, a velocidade das inferências exatas depende de fatores como a estrutura da rede, em termos de quão conectados estão os nós, quantos são os laços e onde se localizam os nós evidenciados (Korb & Nicholson, 2003). Em outras palavras, os métodos de inferência exata, além de serem naturalmente dedicados a RBs compostas por variáveis discretas com espaço de possibilidades reduzido, podem requerer simplificações no MB de maneira a possibilitarem sua manipulação.

Por outro lado, casos específicos geralmente envolvendo distribuições conjugadas, têm também sido abordados por alguns autores. Pearl (1988) apresenta RBs cujas variáveis possuem distribuições condicionais normais, Heckerman *et al.* (1995) tratam do caso Multinomial, Gilks *et al.* (1996) estudam diversas distribuições conjugadas e Castillo *et al.* (1997) e Castillo & Kjærulff (2003) analisam a sensibilidade e a propagação simbólica em RBs Gaussianas.

Como um dos principais objetivos do presente trabalho consiste em permitir uma modelagem mais realista, opta-se por estudar alternativas que lidem com o modelo sem alterar nem a topologia da rede correspondente nem a natureza de suas variáveis. Assim, os métodos de inferência aproximada, principalmente os baseados em MCMC, são enfatizados nos próximos capítulos.

3. MÉTODOS DE MCMC

Do capítulo anterior pôde-se perceber que as dificuldades dos métodos de inferência exata são proporcionais à estrutura da rede e à natureza de suas variáveis. Por outro lado, notar-se-á no presente capítulo que as limitações dos métodos de inferência aproximada são ligadas às distribuições de probabilidades condicionais. De fato, Cooper (1990) prova que a inferência probabilística por métodos de inferência exata em RMCs é NP–*hard* e Dagum & Luby (1993) chega à mesma conclusão ao estudar os métodos de simulação. Isto significa que um algoritmo genérico de inferência probabilística eficiente, precisa e acurada mostra-se improvável. De qualquer forma, além de manterem a integridade da natureza das variáveis, os métodos de simulação podem se tornar os únicos capazes de manipular dada rede a depender de sua topologia; eles são mais amplamente aplicáveis do que os métodos de inferência exata.

Neste capítulo, alguns dos principais métodos de MCMC são apresentados, assim como suas vantagens e desvantagens. Trata-se de uma oportunidade para constatar como as grandes contribuições para o desenvolvimento da área de amostragem de distribuições de probabilidade se deram a partir de pequenos incrementos sobre algoritmos existentes. Para tanto, faz-se necessária uma introdução aos conceitos da integração de Monte Carlo e de cadeias de Markov.

3.1. Integração de Monte Carlo

O principal objetivo da integração de Monte Carlo é inferir sobre um valor esperado a partir de uma amostra da distribuição subjacente. Aqui o conceito de valor esperado deve ser visto em um sentido amplo. Se, por exemplo, deseja-se inferir sobre a probabilidade de dado sistema operar sem falhas durante t unidades de tempo (trata-se da definição clássica de confiabilidade de equipamentos), pode-se definir uma variável de Bernoulli X onde o valor 1 implica no evento de interesse e 0 no seu complementar. Neste caso, a partir da amostra infere-se sobre $E(X)$. Note-se que com esse raciocínio é possível inferir sobre toda a distribuição de probabilidades de uma variável ao invés de um evento em particular. O grande desafio passa a ser então amostrar das distribuições envolvidas no processo de simulação. A seguir, algumas das principais alternativas de amostragem para a integração de Monte Carlo são apresentadas. A título de facilitar a compreensão dos algoritmos, apenas o caso univariado será abordado, no qual se deseja amostrar de uma dada função de interesse $f(x)$ que modela as incertezas sobre uma variável aleatória X. Mais adiante poder-se-á perceber que a partir das

funções univariadas subjacentes ao MB sua amostragem conjunta torna-se possível. A função univariada *f* será também denotada por distribuição-alvo, seu suporte será dado pelo intervalo [*a*, *b*] e seu valor máximo será representado por *c*.

3.1.1. Método da Transformação Inversa

O método da transformação inversa é o mais fundamental dos métodos de integração de Monte Carlo. Ele se baseia no seguinte teorema:

Teorema 3.1 *Seja U uma variável aleatória distribuída uniformemente no intervalo [0, 1]. Seja F uma função não-decrescente e X uma variável aleatória tal que* $X = F^{-1}(U)$. *Então F(x) é a função de distribuição de probabilidades acumulada da variável aleatória X.*

Do Teorema 3.1, vê-se que o método da transformação inversa permite que a amostragem seja realizada a partir da distribuição de probabilidades acumulada da variável, uma vez que seja possível isolar seu argumento diante de um valor aleatório entre 0 e 1. O exemplo a seguir ilustra a adequação de tal método para lidar com a distribuição Exponencial.

Exemplo 3.1 *Seja U uma variável aleatória distribuída uniformemente no intervalo [0, 1]. Seja F a função de distribuição de probabilidades acumulada de uma variável aleatória X~ Exponencial (λ). Assim,* $F(x) = 1 - e^{-\lambda x}$. *Para uma dada realização de U, u, tem-se* $1-e^{-\lambda x} = u \therefore \lambda x = \log(1-u) \therefore x = \log(1-u)/\lambda$.

Assim, uma vez amostrado um valor da distribuição Uniforme(0, 1), tem-se uma realização de uma Exponencial(λ). Logo, caso os valores gerados a partir da distribuição Uniforme possam ser supostos independentes (o que é geralmente o caso), tem-se em mãos uma amostra de valores independentes da Exponencial. Ressalte-se que a relação direta entre a distribuição Uniforme e a Exponencial foi possível devido à distribuição acumulada da Exponencial possibilitar a inversão. Como exposto anteriormente, um MB descreve uma distribuição de probabilidades conjunta a partir da distribuição de probabilidades condicionais de cada variável. Assim, tem-se cronologicamente dois desafios para a aplicação do método da transformação inversa: obter a distribuição acumulada associada a *f* e subsequentemente sua inversa. Infelizmente, muitas são as distribuições das quais a obtenção da acumulada ou a inversão é impossível e, como constatado adiante neste capitulo, elas se tornam cada vez mais freqüentes à medida que a complexidade do modelo associado aumenta. Uma das alternativas mais flexíveis são os métodos de aceitação-rejeição.

3.1.2. Métodos de Aceitação-rejeição

Os métodos de aceitação-rejeição são uma poderosa alternativa para lidar com distribuições cuja amostragem direta é inviável. Tais algoritmos amostram de uma distribuição alternativa a f, também chamada de distribuição proposta, $g(x)$, definida de tal forma que $\beta g(x) \geq f(x)$, $\beta \geq 1$. Eles partem do teorema fundamental de simulação:

Teorema 3.2 *Seja X uma variável aleatória cuja distribuição de probabilidades é dada por f. Amostrar de f equivale a amostrar de (X, U)~ Uniforme [(x, u): $0 \leq u \leq f(x)$].*

Do Teorema 3.2, a variável U é chamada de auxiliar por não ser diretamente associada a f. Como bem justificado por Robert & Casella (2004), a introdução de uma variável auxiliar permite que a amostragem de f seja dada em uma perspectiva diferente. O par (X, U) pode ser amostrado a partir da geração de coordenadas aleatórias no conjunto $D' = ((x, u): 0 \leq u \leq f(x))$. Note-se que D' representa a área sob f descrita em termos de segmentos de retas associadas ao seu suporte. Além disso, como a distribuição marginal de X é a distribuição-alvo, f, cada realização de X, x, que compõe D' é de fato proveniente de f, mesmo que tal realização tenha sido produzida com apenas a avaliação de f no ponto x. Neste momento, o grande desafio a ser superado é amostrar do conjunto D'. Uma alternativa seria amostrar X de f e em seguida $[U|X=x]$~ Uniforme$(0, f(x))$. Porém, amostrar de f é justamente o problema sob estudo. De fato, o curso natural seria amostrar U de sua distribuição marginal e em seguida $[X|U=u]$, levando a um novo problema de amostragem. A solução mais viável torna-se, portanto, simular o par (X, U) de um conjunto mais acessível, D, tal que $D \supseteq D'$.

Uma vez conhecido um valor que delimite f e seu suporte, $\beta \geq c$ e $[a, b]$, tem-se em mãos a configuração mais trivial para o conjunto D. $D = ((w, u): a \leq w \leq b, 0 \leq u \leq \beta)$. Esta abordagem é a mais rudimentar dos métodos de aceitação-rejeição. Nela, os pontos são amostrados do retângulo D que envolve f em todo o seu suporte. A idéia é gerar um valor de W~ Uniforme(a, b) e outro de U~ Uniforme$(0, \beta)$. Se $u \leq f(w)$, então w é considerado como proveniente de f; caso contrario w é rejeitado. Ao final de um número de repetições deste algoritmo, uma amostra composta apenas pelos pontos aceitos permite que inferências sobre o valor esperado da quantidade de interesse sejam realizadas. Note-se que os valores amostrados podem ser considerados independentes, supondo-se que os valores gerados a partir da Uniforme o são. Uma iteração do algoritmo é exibida abaixo.

Algoritmo 3.1: Iteração do Algoritmo de Aceitação-Rejeição Rudimentar

1. Gere uma realização de W~ Uniforme(a, b), w;
2. Gere uma realização de U~ Uniforme$(0, \beta)$, u;

3. Se $u \leq f(w)$ então aceite z como proveniente de f;
4. Caso contrário rejeite z.

Vale comentar que o bom desempenho do algoritmo está intrinsecamente associado à probabilidade de aceitação dos pontos propostos, P_a. Aqui, $P_a = \frac{\int_x f(x)dx}{\beta(b-a)} = \frac{1}{\beta(b-a)}$. Note-se que P_a assume seu maior valor caso $\beta = c$, onde c é por muitas vezes dispendioso de se obter. Por outro lado, caso $\beta < c$ configura-se uma situação na qual a amostra gerada já não pode ser considerada como proveniente de f, mas sim da função *Mínimo* [β, $f(x)$ $\forall$ *x no suporte de f*]. De qualquer forma, nesse processo de amostragem a distribuição proposta é Uniforme (*a, b*) ponderada por uma constante $(b\text{-}a)\beta$. A Figura 3.1 esboça os resultados da aplicação do algoritmo para uma função qualquer, *f*. Apenas os pontos sob *f* são aceitos como provenientes desta.

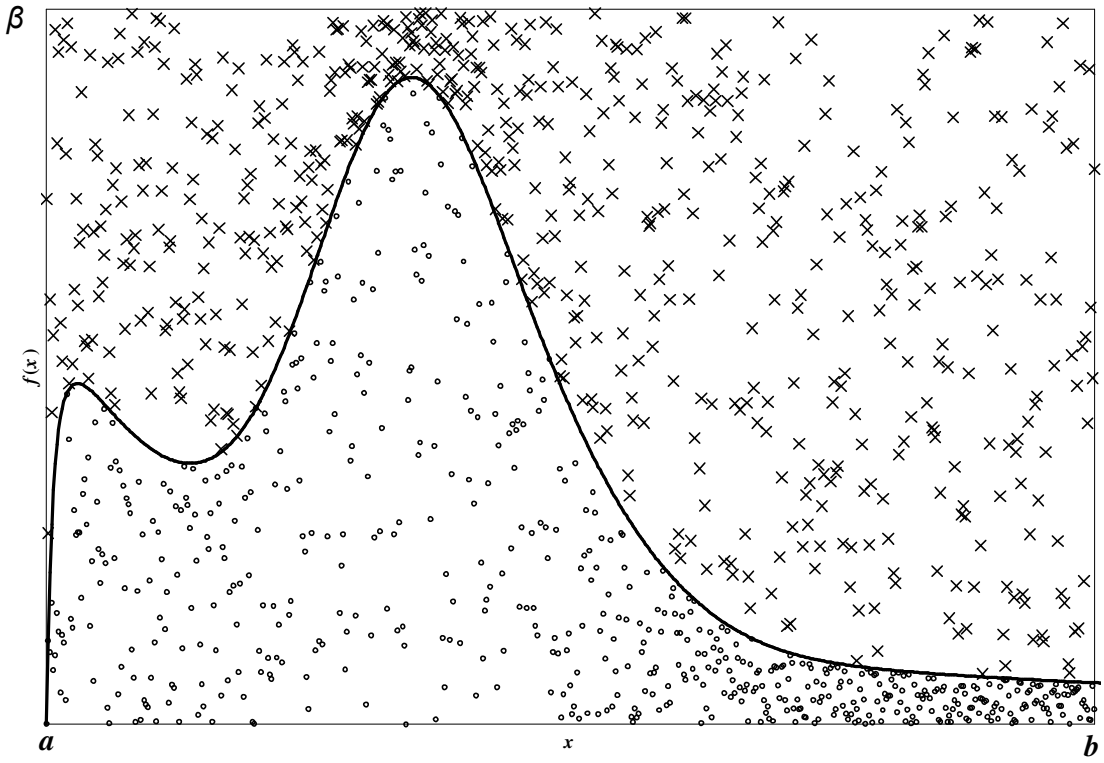


Figura 3.1 Ilustração do algoritmo rudimentar de aceitação-rejeição baseado em uma distribuição candidata proporcional a uma Uniforme.

De maneira a melhorar o desempenho das simulações, uma função mais semelhante à distribuição-alvo pode ser considerada, reduzindo assim a área do conjunto D. Para uma distribuição proposta qualquer, g, ponderada por uma constante β, onde $\beta \geq 1$, $P_a = \frac{\int_x f(x)dx}{\beta \int_x g(x)dx} = \frac{1}{\beta}$. O algoritmo é tal como segue

Algoritmo 3.2: Iteração do Método de Aceitação-Rejeição

1. A partir do método da transformação inversa, gere w da acumulada associada a g, G(x);
2. Gere uma realização de U~ Uniforme[0, $\beta g(w)$], u;
3. Se $u \leq f(w)$ então aceite w como proveniente de f;
4. Caso contrário rejeite w e retorne ao passo 1.

A Figura 3.2 ilustra uma distribuição proposta cuja integral é menor que aquela associada ao algoritmo rudimentar (Algoritmo 3.1), reduzindo assim a probabilidade de rejeição dos pontos gerados desta como provenientes da função-alvo.

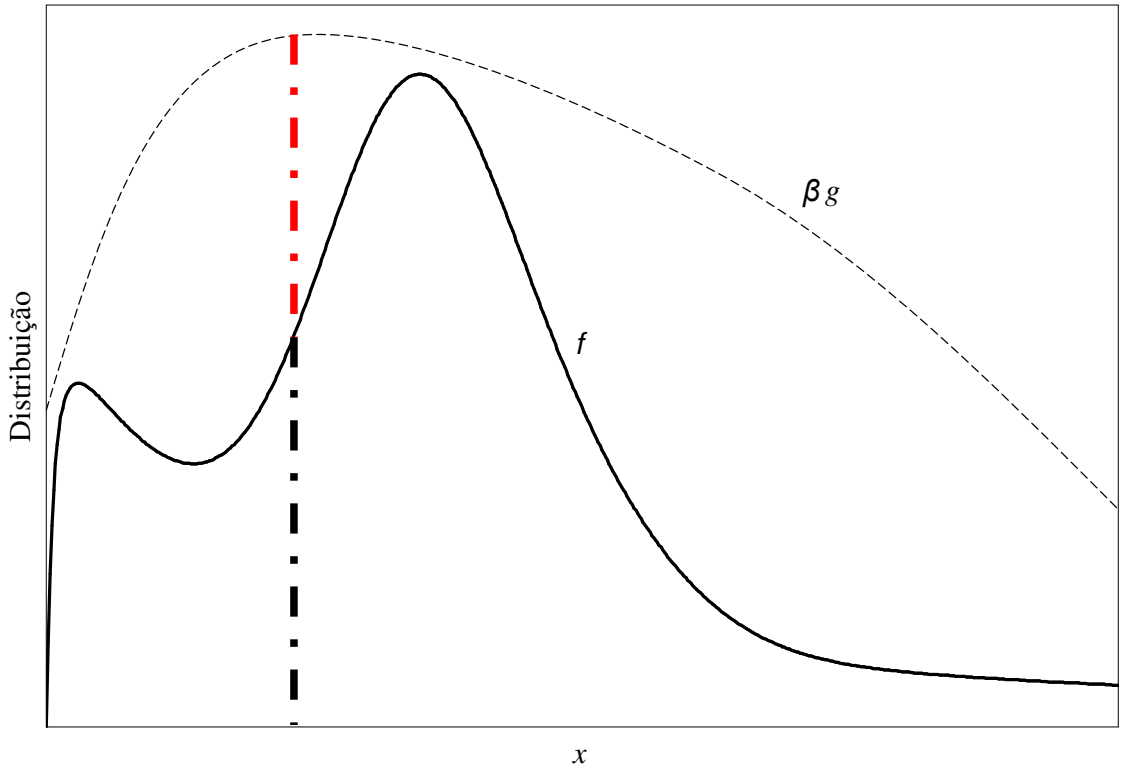


Figura 3.2 Ilustração do algoritmo geral de aceitação-rejeição baseado em uma distribuição candidata que envolve a alvo.

Além de promover a amostragem de valores independentes, uma outra característica marcante deste método de aceitação-rejeição é que ele não faz qualquer uso dos valores rejeitados. Tais pontos poderiam ser utilizados no sentido de aprimorar a aderência da distribuição proposta à alvo, reduzindo assim a área sob D. Isto naturalmente aumentaria a probabilidade de aceitação à medida que pontos amostrados fossem rejeitados. Tal característica adaptativa pode ser encontrada no método adaptativo de aceitação-rejeição (ARS do inglês *Adaptive Rejection Sampling*).

3.1.3. Método Adaptativo de Aceitação-Rejeição (ARS)

Este algoritmo foi desenvolvido para lidar com os casos nos quais f é log-côncava; isto é, $\frac{d^2}{dxdx} log\, f(x) \leq 0$. Nestas condições, pode-se elaborar uma função proposta $g(x)$ onde $log\, g(x) \geq log\, f(x)$, através de retas secantes a pontos de um conjunto **y** nos quais f tenha sido previamente avaliada. Note-se que aqui $g(x)$ é uma função exponencial por partes no suporte de f. Pode-se então iniciar $log\, g(x)$ com um vetor envolvendo alguns poucos pontos nos quais f é avaliada, gerar pontos a partir de g e usá-los para aprimorar adaptativamente a aderência de g a f a cada rejeição. A Figura 3.3 esboça o ajuste linear por partes em função das retas secantes no vetor de pontos $\mathbf{y} = (y_1, y_2, y_3, y_4)$.

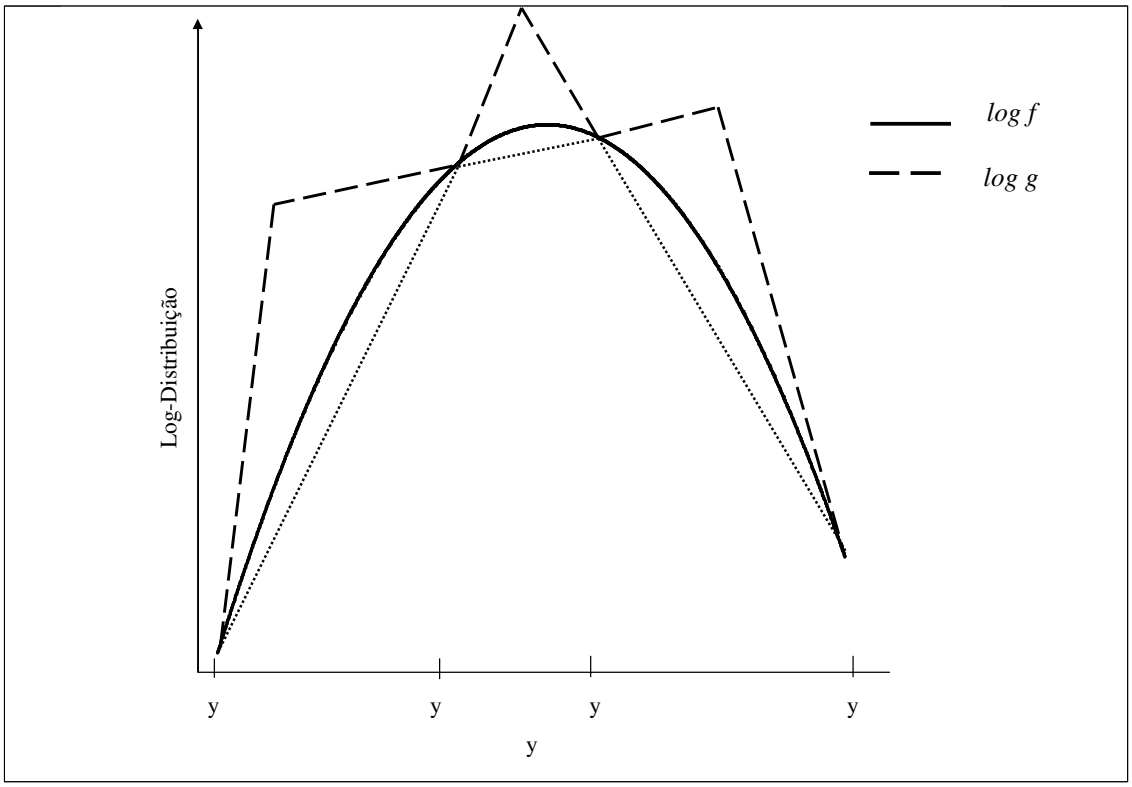


Figura 3.3 Ilustração do algoritmo ARS, originalmente direcionado para funções-alvo log-côncavas.

O algoritmo ARS é dado por:

Algoritmo 3.3: Iteração do Método Adaptativo de Aceitação-Rejeição

1. A partir do método da transformação inversa, gere y da acumulada, G(x), associada à função exponencial por partes, g;
2. Gere uma realização de U~ Uniforme[0, $g(y)$], u;
3. Se $u \leq f(y)$ então aceite y como proveniente de f;
4. Caso contrário rejeite y, introduza y e $f(y)$ em **y** e g, respectivamente, de maneira a aumentar a aderência de g a f e retorne ao passo 1.

Apesar da característica adaptativa deste algoritmo, sua aplicabilidade limita-se apenas aos casos nos quais $log\ f(x)$ é uma função côncava. Embora isto permita a amostragem de muitos dos modelos paramétricos, dentre eles a distribuição Normal, Lognormal, t – Student e muitos outros da família Exponencial, funções com mais do que um máximo, por exemplo, não podem ser abordadas. Com o intuito de propor uma variante mais amplamente aplicável, o

Método Adaptativo de Aceitação-Rejeição por Metropolis (ARMS do inglês *Adaptive Rejection Metropolis Sampling*) foi proposto por Tierney (1991).

3.1.4. Método Adaptativo de Aceitação-Rejeição por Metropolis (ARMS)

De maneira a tratar das funções que não satisfazem à log-concavidade, uma condição adicional é incluída no ARS. Trata-se da probabilidade de aceitação de Metropolis-Hastings $P_{z,w} = Mínimo\left[1, \frac{f(w)g(z)}{f(z)g(w)}\right]$, onde z é o último valor aceito para X na simulação e w o valor mais recentemente proposto pelo algoritmo. Tal probabilidade permite um relaxamento da condição de que $g(x)\beta \geq f(x)$, imposta por todos os algoritmos introduzidos até aqui. Agora, é necessário apenas que o suporte de g seja igual ou contenha o de f. Denotando por z' o próximo valor aceito como proveniente de f , o algoritmo resulta em:

Algoritmo 3.4: Iteração do Método Adaptativo de Aceitação-Rejeição por Metropolis

1. A partir do algoritmo ARS, gere w da acumulada associada a g, G(x);
2. Gere uma realização de U~ Uniforme[*0*, *1*], u;
3. Se $u \leq P_{z,w}$ então aceite w como proveniente de f e faça $z' = w$;
4. Caso contrário rejeite w e faça $z' = z$.

Vale ressaltar que este é o primeiro dos algoritmos apresentados que promove uma relação explícita de dependência entre os valores amostrados. A probabilidade de o próximo valor amostrado, w, ser aceito depende exclusivamente do balanço de probabilidades envolvendo uma transição deste para o último valor aceito, z, e vice-versa. Contudo, pode-se delinear as distribuições propostas de maneira mais genérica do que a sugerida por ARS e ARMS. Tal generalidade é uma característica do algoritmo de Metropolis-Hastings (MH).

3.1.5. Método de Metropolis - Hastings (MH)

O algoritmo de MH é a principal generalização de ARMS. Ele permite que a distribuição proposta assuma qualquer forma, sendo necessário apenas que $g(x)$ possua ao menos o mesmo suporte que $f(x)$. Assim, g pode ser desenvolvido a partir de ARMS, pode ser dado por uma Uniforme, uma Normal e assim por diante. O algoritmo é tal como segue:

Algoritmo 3.5: Iteração do Algoritmo de Metropolis-Hastings

1. Gere w da acumulada associada a uma distribuição proposta g, G(x);
2. Gere uma realização de U~ Uniforme[*0*, *1*], u;

3. Se $u \leq P_{z,w}$ então aceite w como proveniente de f e faça $z' = w$;
4. Caso contrário rejeite w e faça $z' = z$.

Recordando que $P_{z,w} = Mínimo\left[1, \frac{f(w)g(z)}{f(z)g(w)}\right]$, vê-se que $\frac{f(w)g(z)}{f(z)g(w)} = \frac{f(w)/g(w)}{f(z)/g(z)}$.

Assim, o algoritmo sempre aceitará valores w tais que $f(w)/g(w) > f(z)/g(z)$ e nem sempre rejeitará w quando $f(w)/g(w) < f(z)/g(z)$. Quando $g = f$, $P_{z,w} = 1$ e o algoritmo converge para o método da transformação inversa. Por outro lado, caso a discrepância entre g e f se acentue, $P_{z,w} \rightarrow 0$. Note-se que aqui g pode não ser necessariamente uma distribuição de probabilidades, pois não é preciso conhecer qualquer fator de normalização que equivalha à integral de g. Além disso, o infortúnio de se requerer a constante de ponderação considerada nos métodos anteriores pode ser negligenciada no algoritmo de MH.

Para maiores detalhes sobre métodos de amostragem, recomenda-se Ross (2002), Robert & Casella (2004), Devroye (1986) e Gilks *et al.* (1996).

Neste momento é útil ressaltar que diferentemente dos métodos anteriores, MH (e naturalmente ARMS) não promovem uma amostra de valores independentes. A aceitação do próximo ponto proposto como proveniente de f sofre uma influência direta do último ponto aceito. Em uma análise mais detalhada, pode-se concluir que diante de todo o histórico de pontos aceitos como provenientes de f, a probabilidade de o próximo ponto proposto ser aceito é função apenas do ponto aceito corrente. Esta é uma característica de processos estocásticos Markovianos (vista em maiores detalhes a seguir), que quando presente em dado método de integração de Monte Carlo o promove à condição de método de MCMC.

3.2. Cadeias de Markov

Mais genericamente, cadeias de Markov são classificadas como processos estocásticos.

Definição 3.1 *Um processo estocástico X_t, $t \in T$, é uma coleção de variáveis aleatórias, onde T é o conjunto de índices, ou parâmetros, e X_t é o estado do processo no instante t.*

Freqüentemente, T refere-se ao tempo e X_t, o estado do processo no instante t, à sua condição observada em relação a algum parâmetro. Assim, X_t pode representar o valor do dólar em relação ao real no instante t do dia ou a quantidade de pacientes atendidos em uma unidade hospitalar em um dado horário da tarde. Mais especificamente no contexto de simulação de distribuições, X_t pode se referir ao valor amostrado de dada distribuição f na t^a iteração do algoritmo de MH. Cryer (1986) e Souza (2002) apresentam maiores detalhes sobre este tema.

Uma cadeia de Markov é um processo estocástico que não possui memória, isto é, a probabilidade de o processo se encontrar em determinado estado futuramente depende unicamente de onde ele se encontra atualmente, desprezando o seu histórico. Uma cadeia de Markov é graficamente representada por um DG, onde os arcos representam a possibilidade de transição entre os estados, ou nós, quantificada através de taxas ou probabilidades. Conceitualmente tem-se que:

Definição 3.2 *Uma cadeia de Markov é um processo estocástico onde*

$$P(X_{t+\Delta t}=j| X_0= i_0, X_1= i_1, ..., X_t= i) = P(X_{t+\Delta t} =j| X_t= i) = P_{ij}(\Delta t).$$

Uma cadeia de Markov é dita *irredutível* se existe ao menos um caminho provável entre todos os pares de nós e chamada de *aperiódica* se a probabilidade de se manter em qualquer dos seus estados em uma transição imediatamente seguinte é não nula. Estas duas características estão seguramente presentes em cadeias de Markov referentes a redes Bayesianas envolvendo variáveis discretas nas quais não há probabilidades condicionais nulas. Quando uma cadeia de Markov é irredutível e aperiódica pode-se aproximar P_j, a probabilidade de o sistema se encontrar no estado j (P_j é também conhecida como probabilidade estacionária do estado j, relativa à proporção relativa de tempo em que o sistema permanece no estado j para um tempo de missão indeterminado), pela probabilidade de transição $P_{ij}(t)$, quando t cresce. Recomenda-se Ross (2000) e Ross (2002) para um apanhado geral sobre o assunto.

Retornando à seção anterior, vê-se que o algoritmo de MH promove a elaboração de uma cadeia de Markov, pois verifica-se a igualdade da Definição 3.2. Nestes casos, tem-se um método de integração de Monte Carlo via cadeias de Markov, como já mencionado.

3.3. Amostrando de MBs Multidimensionais

Vale ressaltar que os estados da cadeia de Markov quando da aplicação de um método de MCMC podem ser descritos no espaço multidimensional. Assim, a simulação de um MB pode ser orientada de acordo com a distribuição de probabilidades condicional de cada variável dados os valores das demais variáveis do modelo. Para cada variável tem-se a oportunidade de amostrar da distribuição de probabilidades condicional subjacente, de acordo com os algoritmos apresentados na seção anterior, promovendo uma mudança de estados da cadeia.

Considerando um MB que descreve a distribuição de probabilidades conjunta $P(X_1, X_2,..., X_d)$ e um estado inicial para a cadeia $\boldsymbol{x_0} = (x_1^{(0)}, x_2^{(0)},..., x_d^{(0)})$, o método de MCMC opera a t^a iteração como a seguir:

Algoritmo 3.6: t^a Iteração de um Método de MCMC

1. Gere $x_1^{(t)}$ de acordo com $f(x_1|x_2^{(t-1)}, x_3^{(t-1)},..., x_d^{(t-1)})$;
2. Gere $x_2^{(t)}$ de acordo com $f(x_2|x_1^{(t)}, x_3^{(t-1)},..., x_d^{(t-1)})$;
3. Gere $x_3^{(t)}$ de acordo com $f(x_3|x_1^{(t)}, x_2^{(t)}, x_4^{(t-1)},..., x_d^{(t-1)})$;

...

d. Gere $x_d^{(t)}$ de acordo com $f(x_d|x_1^{(t)}, x_2^{(t)},..., x_{d-1}^{(t)})$;

A cada uma das sub-iterações a cadeia de Markov subjacente transita de um estado para outro. A questão fundamental é como obter e subsequentemente amostrar das distribuições de probabilidades condicionais envolvidas $f(x_i|x_1^{(t)}, x_2^{(t)},..., x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, ..., x_d^{(t-1)})$, denotadas daqui em diante por $f(x_i)$. De maneira a tratar do primeiro desafio, os conceitos da teoria dos grafos são também úteis. Do Algoritmo 3.6, pode-se perceber a eventual dificuldade em computar $f(x_i)$. Porém, devido à introdução da teoria dos grafos promovida por RBs, tal tarefa torna-se menos árdua a partir do conceito de cobertura de Markov. A cobertura de Markov de dada variável é um conjunto de variáveis que quando instanciadas tornam esta independente das demais, promovendo uma simplificação na elaboração das distribuições de probabilidades condicionais.

Uma rede Bayesiana pode ter um grande número de variáveis e a probabilidade de uma variável pode ser afetada pelo conhecimento (ou instanciação) de uma variável que esteja distante dela (distância aqui significa que as cadeias entre as duas variáveis possuem um grande número de nós). Porém, pode ser demonstrado (Neapolitan, 2004) que a instanciação de um conjunto de nós próximos interrompe a comunicação entre um nó e os demais. Tal conjunto é chamado de cobertura de Markov.

Definição 3.3 *Seja V um conjunto de variáveis aleatórias, P a sua distribuição de probabilidades conjunta e $X \in V$. Então a cobertura de Markov $M(X)$ é qualquer conjunto de variáveis tal que X é condicionalmente independente de todas as demais variáveis dados os valores das variáveis de $M(X)$. Isto é denotado por X ind $V \cap [M(X) \cup X]^c | M(X)$.*

A cobertura de Markov permite que inferências sobre as variáveis que compõem a RB sejam realizadas localmente; trata-se de uma generalização da condição Markoviana apresentada anteriormente, pois além dos pais da variável, seus filhos e respectivos pais são também considerados. Assim, independente do tamanho da rede, qualquer variável é passível de atualização através de propagações locais. Matematicamente, tem-se então que $f(x_i \mid x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d) = f(x_i \mid M(x_i))$, facilitando assim as sub-iterações do método de MCMC. O teorema a seguir apresenta os componentes de uma cobertura de Markov.

Teorema 3.3 *Se o par* (G, P) *satisfaz à condição de Markov, então para cada X,* $M(X)$ *é composto por* $PA(X)$, $C(X)$ *e* $PA[C(X)]$, *onde* $PA(X)$ *é o conjunto de pais de X,* $C(X)$ *é o seu conjunto de filhos e* $PA[C(X)]$ *é composto pelos conjuntos dos pais de cada filho de X.*

Quando uma cobertura de Markov não possui subconjuntos que são, também, coberturas de Markov, esta é definida como uma fronteira de Markov. Pode-se provar (Neapolitan, 2004) que se P, a distribuição de probabilidades conjunta das variáveis de V, é estritamente positiva, então para cada $X \in V$ existe uma única cobertura de Markov para X.

Uma vez computada a distribuição de probabilidades condicional da variável X_i, $f(x_i)$, dá-se início ao problema de amostrar desta. Além dos algoritmos introduzidos anteriormente (seção 3.1) outros métodos igualmente importantes podem ser adotados. De fato, um dos mais importantes métodos de amostragem multidimensional é verificado quando todas as distribuições-alvo, $f(x_i)$ para $i = 1, 2, ..., d$, são passíveis de amostragem direta: o método de Gibbs (GS do inglês *Gibbs sampling*). Nele, $P_{z,w} = 1$, isto é, a probabilidade de aceitação de cada valor proposto é 1 e MH converge para o método da transformação inversa. A próxima seção introduz GS com maiores detalhes.

3.3.1. Método de Gibbs (GS)

O método de Gibbs inicialmente proposto por Geman & Geman (1984) é um dos mais importantes algoritmos de MCMC. Supondo que é sempre possível amostrar da distribuição de probabilidades condicional de cada variável dados os valores das variáveis componentes da sua cobertura de Markov no MB, GS amostra diretamente destas a partir do método da transformação inversa e promove uma cadeia de Markov que não envolve rejeições de pontos propostos.

GS pode ser compreendido como uma aplicação multidimensional do método da transformação inversa. Quando aplicado a casos envolvendo apenas variáveis discretas, Pearl (1988) denotou GS como método de simulação estocástica. Pelas palavras do próprio Pearl, trata-se de um método de cálculo de probabilidades por contagem da freqüência em que os eventos ocorrem em uma série de iterações executadas. A RB que envolve o MB é usada para gerar amostras aleatórias de suas configurações hipotéticas prováveis a partir das distribuições de probabilidades condicionais subjacentes. Ao final das simulações, pode-se inferir sobre a distribuição de probabilidades marginal de cada variável de interesse a partir de duas fontes. A primeira consiste na freqüência relativa de cada valor ou intervalo amostrado. Esta alternativa é justamente aquela adotada pelos algoritmos de aceitação-rejeição apresentados nas seções anteriores. A segunda opção envolve a ponderação das distribuições de probabilidades

condicionais das quais as amostras foram obtidas, conhecida como *Rao-Blackwellization* por se basear no teorema de Rao-Blackwell:

Teorema 3.4 *Sejam W e Y duas variáveis aleatórias quaisquer. Pode-se provar que V(W) = E(V(W|Y)) + V(E(W|Y)), onde E(Z) e V(Z) implicam respectivamente na esperança e variância da variável Z.*

Do Teorema 3.4, pode-se ver que $V(E(W|Y)) \leq V(W)$, já que $E(V(W|Y)) \geq 0$. Assim, o estimador da esperança de $E(W|Y)$ – denotado por T_1 – é estatisticamente mais eficiente do que o de $E(X)$ – denotado por T_2 – ou seja, para o mesmo tamanho amostral $V(T_1) \leq V(T_2)$. Para uma dada variável, X_i, cuja amostra em S repetições do Algoritmo 3.6 (após o período estacionário da cadeia) é dada por $\boldsymbol{x}_i = (x_{i1}, x_{i2},... x_{iS})$, onde $\boldsymbol{x}_i$ foi definida de acordo com S distribuições condicionais $\boldsymbol{f}_i = (f_{i1}, f_{i2},... f_{iS})$, *Rao-Blackwellization* infere sobre dada distribuição marginal de interesse $\pi(x_i)$ a partir de

$$\hat{\pi}_i(x) = g_i^*(x) = \frac{1}{S}\sum_{j=1}^{S} f_{ij}(x)\,. \qquad \textit{Equação 3.1}$$

A Figura 3.4 ilustra o processo de estimação via *Rao-Blackwellization*. A cada ponto x do suporte da distribuição marginal, π, tem-se uma estimativa para $\pi(x)$ a partir da média das distribuições condicionais envolvidas nas simulações durante o período de estacionaridade da série. Recomenda-se Gelfand & Smith (1990) para maiores detalhes sobre este tema.

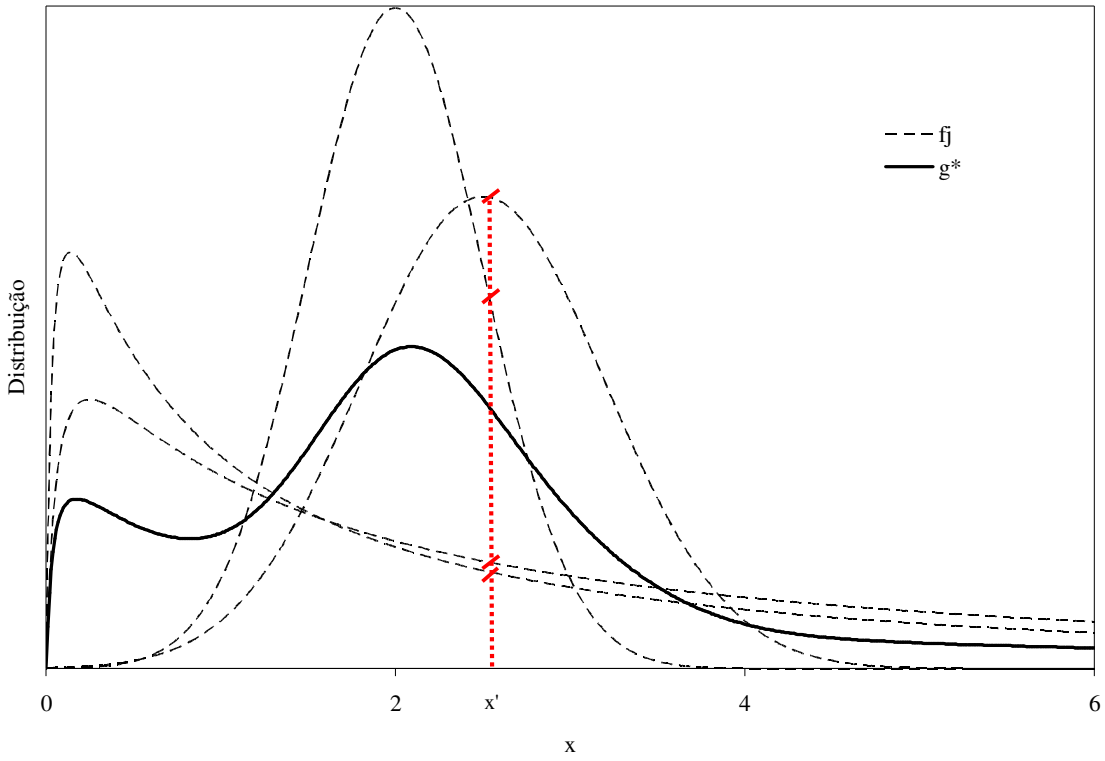


Figura 3.4 Ilustração do processo de estimação via Rao-Blackwellization. As linhas tracejadas são as distribuições condicionais e a sólida a distribuição marginal estimada.

Segundo York (1992), quando GS é aplicado a MBs envolvendo variáveis discretas com espaço de possibilidades reduzido não é preciso que a distribuição do processo seja estritamente positiva para que se alcance a convergência às distribuições de interesse, como entende-se de Hrycej (1990); o que é necessário é que a cadeia de Markov a ser percorrida durante as simulações seja irredutível.

Retornando ao Exemplo 2.5, que trata da qualidade do produto de determinada empresa, onde $Z \rightarrow X \rightarrow W$, tem-se 8 possíveis estados para o sistema a ser simulado (Tabela 3.1). O estado 0, por exemplo, condiz a uma situação onde a qualificação dos profissionais (Z), a qualidade do produto (X) e a aceitação do produto pelo mercado (W) mostram-se favoráveis a uma boa imagem externa da organização.

Tabela 3.1 Estados da cadeia de Markov referente à rede Bayesiana que trata da qualidade dos produtos de determinada organização.

Z	*X*	*W*	*Estado da cadeia*
0	*0*	*0*	*0*
0	*0*	*1*	*1*
0	*1*	*0*	*2*
0	*1*	*1*	*3*
1	*0*	*0*	*4*
1	*0*	*1*	*5*
1	*1*	*0*	*6*
1	*1*	*1*	*7*

A simulação estocástica via MCMC faz uso das coberturas de Markov (Definição 3.3) e do método da transformação inversa apresentado anteriormente para realizar suas inferências. Supondo que se deseja conhecer as incertezas acerca das variáveis *Z*, *X* e *W*, começa-se com a atribuição inicial das variáveis da rede e segue-se com uma atualização de cada variável em uma ordem predeterminada. Aqui, adota-se a ordem *Z*, *X, W*, com todas as variáveis igualadas a zero inicialmente. Assim, a cadeia de Markov referente às simulações parte do estado *0*. Se $f(Z=0)=0.7$, $f(X=0|Z=0)= 0.8$, $f(X=0|Z=1)= 0.3$, $f(W=0|X=0)= 0.7$ e $f(W=0|X=1)= 0.2$:

Passo 1, atualiza-se *Z*: $f(z|X=0, W=0) = \alpha\, P(z)P(X=0|z) = \alpha\, [(0.7)(0.8); (0.3)(0.3)]=$
$= (0.86; 0.14)$.

Gera-se um número Uniforme (0, 1), $u = 0.4$. Como $u \leq f(Z=0|X=0, W=0)$, adota-se $Z=0$ (método da transformação inversa para funções discretas). A cadeia de Markov se mantem no estado *0*.

Passo 2, atualiza-se *X*: $f(x|Z=0, W=0) = \alpha\, f(x|\, Z=0)f(W=0|x) = \alpha[(0.8)(0.7); (0.2)(0.2)] =$
$= (0.93; 0.07)$.

Note-se que a 1ª igualdade se dá pela condição Markoviana (Z *ind* $W|X$). Gera-se um número uniforme entre 0 e 1, $u = 0.95$. Como $u \leq f(X=0|Z=0, W=0) + f(X=1|Z=0, W=0)$, adota-se $X=1$. A cadeia de Markov transita para o estado *2*.

Passo 3, atualiza-se *W*: $f(w|Z=0, X=1) = f(w|\, X=1) = (0.2; 0.8)$.

Gera-se um número uniforme entre 0 e 1, $u = 0.3$. Como $u \leq f(W=0|Z=0, X=1) + f(W=1|Z=0, X=1)$, adota-se $W=1$. A cadeia de Markov transita para o estado *3*.

Ao final do passo 3 conclui-se uma iteração do método de MCMC e um indivíduo hipotético é gerado. Quanto maior o número de iterações, mais próximas dos verdadeiros parâmetros tornam-se as freqüências relativas de cada categoria. Alternativamente, ao final

das simulações é possível computar a média das distribuições condicionais de cada variável utilizadas para amostrar a cadeia, o procedimento de inferência por *Rao-Blackwellization* que é estatisticamente mais eficiente do que aquele simplesmente baseado nos valores amostrados, como comentado anteriormente (Teorema 3.4).

Ressalte-se que a cadeia de Markov subjacente a este problema é irredutível e aperiódica, o que garante que estimativas baseadas no período estacionário do processo amostrado tenderão às distribuições verdadeiras à medida que tal período cresça.

Infelizmente, a possibilidade de se adotar GS torna-se cada vez menor à medida que o MB sob estudo cresce em complexidade. Naturalmente existirão distribuições de probabilidades condicionais para as quais a amostragem é inviável. Retornando ao Exemplo 2.1, no qual o MB envolve duas variáveis, P e X, onde $f(p)$ e $f(x|p)$ desencadeiam a distribuição de probabilidades conjunta de P e X. Embora que a amostragem direta de $f(x|p)$ seja possível a partir da distribuição Binominal, amostrar de $f(p|x)$ pode não ser possível caso $f(p)$ seja delineada por uma função não-paramétrica que reflita as incertezas iniciais acerca de P, ou, mais genericamente, caso $f(p)$ seja não-conjugada a $f(x|p)$. Com o objetivo de permitir o uso mais geral de GS, Ritter & Tanner (1992) propuseram o método de *Griddy-Gibbs* (GGS do inglês *Griddy-Gibbs sampling*).

3.3.2. Método de *Griddy*-Gibbs (GGS)

Diante de uma distribuição cuja amostragem direta é inviável, f, GGS basicamente promove uma iteração da cadeia de Markov por meio dos seguintes passos:

Algoritmo 3.7: Iteração do Método de GGS

1. Avalie a distribuição-alvo, f, em um conjunto de pontos do seu suporte $\mathbf{y} = (y_1, y_2, \ldots, y_n)$;
2. Compute $\mathbf{g} = (g_x(y_1, y_2), g_x(y_2, y_3), \ldots, g_x(y_{n-1}, y_n))$, um conjunto de funções que aproximam f em cada intervalo $[y_k, y_{k+1}]$, $k = 1, 2, \ldots, n-1$;
3. Gere x de acordo com $\mathbf{G}(x)$, a função proporcional à acumulada associada a $\mathbf{g}$, e considere-o como proveniente de f.

A principal vantagem de adotar GGS na impossibilidade de GS é que ao final das iterações da cadeia de Markov, *Rao-Blackwellization* é ainda possível. Contudo, um cuidado adicional é necessário: uma normalização através de $\mathbf{G}(y_n)$, resultando na seguinte estimativa para a distribuição marginal de interesse, $\pi(x)$:

$$\hat{\pi}(x) = g^{*}(x) = \frac{1}{S}\sum_{j=1}^{S}\frac{g_j(x)}{G(y_{n_j})}. \qquad \textit{Equação 3.2}$$

O uso de toda a distribuição condicional ao invés de apenas um ponto amostrado desta promove uma redução do nº de iterações da simulação e distribuições marginais mais suaves, geralmente melhor ajustadas às distribuições de interesse. Os demais métodos apresentados anteriormente envolvendo aceitação-rejeição (adaptativos ou não) permitem inferências apenas a partir da cadeia de Markov amostrada. Toda informação adicional acerca de *f* é descartada uma vez que sua amostragem é realizada. Note-se que a depender da quantidade de rejeições realizadas por dado método adaptativo tal como ARMS, pode se alcançar um nível de aderência da função proposta a *f* cujo descarte é lamentável.

Por outro lado, pode-se perceber o quão sensível é GGS a ambos: (i) o método pelo qual o vetor de pontos **y** é definido e (ii) as funções adotadas para aproximar *f* entre os pontos de **y**. Tem sido um procedimento-padrão em exercícios envolvendo GGS a adoção de funções lineares para compor o conjunto **g**. Além de Ritter & Tanner (1992), Bauwens & Lubrano (1998) e Ausín & Galeano (2007) podem ser citados. Contudo, funções pontuais por partes resultando em uma distribuição discreta no suporte de *f* podem também ser encontradas em trabalhos como o de Ardia *et al.* (2009). De fato, é raro ou não existe qualquer aplicação de GGS considerando outro tipo de aproximação. As funções lineares por partes são adotadas de duas maneiras. A primeira e mais adotada na prática baseia-se na regra de integração trapezoidal, isto é, $g_x(y_k, y_{k+1})$ é a reta que passa pelos pontos $(y_k, f(y_k))$ and $(y_{k+1}, f(y_{k+1}))$. A segunda envolve uma função uniforme por partes no intervalo $[a_k, a_{k+1}]$ do qual x_k é o ponto médio, ou seja, $g_x(a_k, a_{k+1}) = f(y_k)$.

Sobre o vetor **y**, dois critérios são apresentados pela literatura de GGS. O primeiro é simplesmente computar *n* pontos igualmente espaçados no intervalo [*a*, *b*]. Baseando-se neste algoritmo, autores como Ardia *et al.* (2009) manipulam modelos heterocedásticos autoregressivos (ARCH do inglês *autoregressive conditional heteroskedastic*) enquanto Bauwens & Lubrano (1998) e Ausín & Galeano (2007) trabalham com modelos GARCH generalizados (GARCH do inglês *generalized autoregressive conditional heteroskedastic*) e Wei (2002) estuda modelos GARCH com dados censurados. Um dos principais desafios a superar quando da adoção deste método é determinar *n*. Se *n* é subestimado, as inferências podem ser viciadas; se *n* é superestimado *f* é desnecessariamente avaliado conduzindo a ineficiência em termos de consumo de memória e tempo de simulação. A principal regra que

tem sido adotada consiste em determinar n a partir de um julgamento subjetivo; usualmente em um procedimento de setar-e-testar até que julgamentos favoráveis sejam alcançados.

Uma alternativa adaptativa encontrada na literatura de GGS para determinar **y** à medida que a simulação itera é considerar um vetor do qual a massa probabilística sob a aproximação de f entre pontos sucessivos é aproximadamente constante. Tal procedimento gera um vetor que concentra mais pontos em regiões com maior massa e menos pontos em regiões com menor massa. Este raciocínio é mencionado por Ritter & Tanner (1992) como a principal alternativa para gerar um bom vetor de pontos, embora raramente usado na prática, talvez devido à difícil tarefa de identificar eficientemente os pontos que particionem equiprovavelmente **g**, um ponto deixado em aberto desde então. De qualquer forma, defende-se aqui que estudar o comportamento da função derivada de f conduz a uma melhor aproximação adaptativa. O seguinte exemplo torna-se útil na argumentação a favor de tal idéia.

Exemplo 3.2 *Deseja-se aproximar a seguinte distribuição:* $f(x) \propto \begin{cases} e^{-10}, x \in [0,1] \\ e^{-10x}, x > 1 \\ 0, x < 0. \end{cases}$

Pode-se perceber que embora o intervalo [0, 1] contenha a maior massa de probabilidade de f, avaliar $f(0)$ e $f(1)$ é suficiente para precisamente aproximar f em tal intervalo através de uma função linear. Por outro lado, o intervalo com menor massa de probabilidade ($x > 1$) vai requerer um maior número de avaliações de f de maneira a ajustar adequadamente uma função linear por partes a f neste intervalo. O raciocínio proposto por Ritter & Tanner negligenciaria partes com menor massa de probabilidade de f, regiões que ocupam papel de suma importância em áreas tais como análise probabilística de riscos, que usualmente envolvem eventos raros, sendo fontes de estudo de pesquisadores tais como Zio & Pedroni (2009). De fato, levar em consideração o comportamento da função derivada de f, f', seria um critério naturalmente melhor. Intervalos com maior variação de f' teriam a maior quantidade de pontos. Contudo, estudar o comportamento de f' pode ser uma tarefa também difícil. No presente trabalho, propõe-se superar este desafio a partir da introdução de métodos adaptativos de quadratura.

3.4. Período de *Burn-in*

Quando um método de MCMC gera uma cadeia de Markov durante as simulações, a escolha de valores iniciais não afeta a distribuição estacionária de probabilidades dos estados (Gilks *et al.* 1996) caso a irredutibilidade seja verificada. Mesmo que o cenário inicial do

processo seja pouco provável, quando o número de iterações cresce suficientemente a convergência é garantida. Porém, como as inferências baseiam-se em médias aritméticas, a escolha inapropriada do cenário inicial pode levar a maiores imprecisões. O que se sugere é um período de *burn-in* que descarte os cenários inicialmente gerados. Gilks *et al.* (1996) acrescentam que mesmo sendo o estado inicial bem definido, um período de *burn-in* contribui para melhores estimativas.

Formalmente, a identificação do melhor período de *burn-in* é realizada por ferramentas de diagnóstico de convergência, porém, Geyer (1992) considera tais métodos desnecessários. Segundo o autor, caso valores extremos sejam evitados, o período de *burn-in* pode ser estimado como estando entre *1%* e *2%* da quantidade de iterações executadas. O problema a ser resolvido seria evitar tais valores extremos. Neste sentido, Gilks *et al.* (1996) recomendam a utilização de valores iniciais esparsos, ou seja, a realização de várias simulações a partir de cenários iniciais diferentes; a intenção é verificar a sensibilidade do processo estocástico a distintos cenários iniciais. Uma outra alternativa é definir como estado inicial do processo aquele mais verossímil; para isto, seleciona-se o estado mais provável segundo a distribuição de probabilidades conjunta das variáveis da rede condicionadas a eventuais evidências.

Com o objetivo de possibilitar uma abordagem condicionada ao número de simulações do método de amostragem multidimensional e às características de cada variável do MB, Bauwens & Lubrano (1998) se desvinculam dessas linhas discussões e interpretam estacionaridade como o período no qual a média adotada para resumir a serie, seja esta associada à amostra ou às respectivas distribuições condicionais geradoras, encontra-se em dado intervalo de confiança; um procedimento essencialmente idêntico àquele subjacente a cartas de controle. Para uma dada variável, X_i, e um número de simulações T, se CS_t permanece dentro de um intervalo $[-\gamma, \gamma]$ para todo t maior que $K(\gamma)$, então interpreta-se que a série convergiu após $K(\gamma)$ realizações, possibilitando a estimação de $E(X_i)$ sob um erro relativo de $100\gamma\%$. CS_t é dado por :

$$CS_t = \frac{\frac{1}{t}\sum_{i=1}^{t}(x_i - \mu_X)}{\sigma_X}, \; t = 50, 100, 150, \ldots, T \qquad \textit{Equação 3.3}$$

Onde μ_X e σ_X representam respectivamente a média e desvio-padrão empíricos das T realizações de X_i.

4. MÉTODO DE MCMC PROPOSTO

No capítulo anterior buscou-se enfatizar as vantagens de se adotar GGS em complemento a GS. A principal crítica acerca dos métodos de aceitação-rejeição (adaptativos ou não) repousa sobre o fato de estes não fazerem total uso da informação gerada à medida que novas avaliações sobre as distribuições-alvo são efetuadas. Tais métodos possuem como característica marcante inferir sobre distribuições marginais a partir dos valores amostrados apenas, enquanto que GGS (a exemplo de GS) possibilita *Rao-Blackwellization*. Contudo, pontos a melhorar no algoritmo de GGS foram também levantados, restando ainda questões a serem discutidas no presente capítulo.

O principal objetivo deste capítulo é propor uma variante de GGS que promova inferências mais precisas e em menor tempo de simulação quando comparada ao algoritmo de GGS tradicionalmente aplicado, isto é, baseado em vetores de pontos igualmente espaçados e funções lineares por partes. A principal motivação para aprimorar GGS ao invés de outras vertentes dos métodos de MCMC reside no fato de GGS possibilitar *Rao-Blackwellization*, fazendo uso exaustivo das distribuições de probabilidades condicionais utilizadas durante as simulações. O número de iterações do método MCMC pode então ser reduzido, sem perda de acurácia das estimativas.

As contribuições sobre o GGS proposto pela literatura se darão em duas perspectivas, a teórica e a numérica. Teoricamente, conceitos de ARMS são introduzidos em GGS, levando à superação de uma deficiência fundamental sua: a negligência da probabilidade aceitação de MH, uma característica marcante dos métodos de MCMC. Como visto nas primeiras seções do capítulo anterior, quando amostra-se de uma distribuição proposta ao invés da distribuição-alvo propriamente dita, os pontos amostrados podem ser rejeitados. Nos métodos de MCMC tal possibilidade é descrita através da probabilidade de aceitação de MH. Isto pode ser imperativo para a obtenção de estimativas acuradas através de GGS, já que as distribuições consideradas para a fase de *Rao-Blackwellization* são delineadas de acordo com os estados visitados pela cadeia de Markov amostrada. Numericamente, algumas alternativas simples e computacionalmente baratas são introduzidas em GGS, possibilitando melhores ajustes para as funções propostas e também um melhor ajuste final por *Rao-Blackwellization*. Para o primeiro caso, métodos adaptativos de quadratura triviais são introduzidos, tais como a regra adaptativa de Simpson (ASR do inglês *adaptive Simpsons' rule*). Isto possibilita uma definição adaptativa do vetor de pontos nos quais a função-alvo, f, é avaliada. Além disso, similarmente a ARMS propõe-se ajustar uma função linear por partes a $\log f(x)$ ao invés de

diretamente a $f(x)$ (tal como o GGS tradicional). Tal estratégia reduz a probabilidade de se deparar com problemas envolvendo aritmética de ponto flutuante, tais como as dificuldades em operar valores muito próximos de zero ou muito elevados. Para determinar de maneira inteligente o vetor de pontos nos quais *Rao-Blackwellization* se dará, propõe-se a introdução de conceitos de métodos de agrupamento probabilístico, como *centroidal Voronoi tessellations* (CVT). Como mencionado por Du & Gunzburger (2002), CVT têm sido adotados em diversas áreas de aplicação, incluindo compressão de dados, análises de agrupamento, biologia celular, comportamento territorial de animais e alocação ótima de recursos. Dada uma região Ω, uma função de medida de distância e uma partição de Ω composta por r subconjuntos de interesse, CVT busca identificar o centro de massa dos r subconjuntos que mais eficientemente representam Ω de acordo com a função de distância. Propõe-se então utilizar os conceitos inerentes a CVT para elaborar o melhor vetor de pontos a ser usado para *Rao-Blackwellization*.

Baseando-se nos argumentos apresentados acima, o resto deste capítulo é dado como segue: na próxima seção um algoritmo adaptativo probabilístico que generaliza ASR é proposto; em seguida conceitos elementares de CVT são introduzidos; por fim, propõe-se a variante de GGS baseada em ARMS, ASR e CVT.

4.1. Regra Probabilística Adaptativa de Simpson

Um algoritmo adaptativo é aquele que se adapta à forma da função sob estudo. A regra de Simpson é um dos mais simples algoritmos para a integração numérica adaptativa. Para dados pontos l e u ($l < u$), o algoritmo aproxima a integral de $h(x)$ em $[l, u]$, $F_h(l, u) = \int_l^u h(x)dx$, por

$$A_h(l, u) = [h(l) + 4h(m) + h(u)](u - l) / 6, \qquad \textit{Equação 4.1}$$

onde $m = (l + u) / 2$ e $A_h(l, u)$ é a área sob a parábola, $g_x(l, u)$, que passa por $(l, h(l))$, $(m, h(m))$ e $(u, h(u))$. Press *et al.* (1992) apresentam maiores detalhes. É fácil ver que se $h(m) = (h(l) + h(u))/2$ então a parábola é convertida ao polinômio interpolador de Lagrange de grau 1 enquanto que se $h(l) = h(m)$ e $h(u) = h(m)$ então a parábola transforma-se em um polinômio interpolador de Lagrange de grau 0.

Baseando-se em Mckeeman (1962), uma regra adaptativa para aproximar a área $F_h(l, u)$ condicionando-se a um erro tolerável atribuído pelo usuário ε (> 0.0) é considerada aqui. Contudo, diferentemente de autores como Malcolm & Simpson (1975), sugere-se utilizar um erro relativo ao invés de um absoluto. Este último pode levar a um proibitivo número de

avaliações de funções-alvo não normalizadas que envolvem uma massa total razoavelmente maior que 1. Assim, considera-se aqui um erro relativo local no intervalo $[l, u]$ ε_{lu}:

$$\varepsilon_{lu} = \frac{u-l}{b-a}\left(1-\frac{v_{min}}{v_{max}}\right), \text{ onde } v_{max}=\text{Máximo}[A_h(l, m) + A_h(m, u), A_h(l, u)] \text{ e} \qquad \textit{Equação 4.2}$$

v_{min} = Mínimo$[A_h(l, m) + A_h(m, u), A_h(l, u)]$.

Emerge desta forma o critério recursivo para continuar a análise nos intervalos $[l, m]$ e $[m, u]$ caso $\varepsilon_{lu} > \varepsilon$. Caso contrário, $F_h(l, m)$ e $F_h(m, u)$ são aproximadas por $A_h(l, m)$ e $A_h(m, u)$, respectivamente, e os valores l, m e u são incluídos (nesta ordem) no vetor de pontos **y**. Sob esta regra adaptativa simples, intervalos nos quais h apresenta um comportamento irregular (com expressivas variações da sua função derivada h') são mais provavelmente detalhados do que os intervalos nos quais h é mais suave (com pequenas variações de h'). Trata-se de uma alternativa simples e computacionalmente barata para suprir a idéia de adaptatividade sugerida por Ritter & Tanner (1992) comentada anteriormente (seção 3.3.2), porém enfatizando variações da função derivada ao invés da eqüiprobabilidade entre intervalos.

Ao final deste processo, a função-alvo é conhecida nos pontos do vetor **y** e baseando-se na Equação 4.1, a função acumulada correspondente pode ser estimada por

$$G(y_k) = \sum_{j=2}^{k} A_f(x_{k-1}, x_k) = G(y_{k-1}) + A_f(y_{k-1}, y_k). \qquad \textit{Equação 4.3}$$

Considerando $h(x) = log\ f(x)$, funções lineares por partes podem aproximar f e h (este último resultando em funções exponenciais por partes para estimar f) e algoritmos computacionais elementares possibilitam a obtenção tanto da função acumulada $G(y)$ em qualquer ponto não pertencente a **y**, quanto $G^{-1}(u)$, a função inversa para qualquer quantidade u no intervalo $[0, G(b)]$. Assim, caso dada função f não possibilite uma amostragem direta, realiza-se uma aproximação a esta e sua subseqüente amostragem.

Além da aproximação linear sobre h, autores como Meyer *et al.* (2008) têm estudado o impacto de se adotar aproximações parabólicas. De maneira a promover melhoramentos ao ARMS, os autores aplicam tal aproximação nas regiões onde h é côncava, resultando em funções Gaussianas por partes alternativamente às exponenciais. Naturalmente, métodos mais elaborados para computar tanto $G(y)$ quanto $G^{-1}(u)$ são demandados nestes casos. No presente trabalho, apenas aproximações lineares para f e h são levados em consideração. Além da maior simplicidade de implementação e eficiência computacional, isto também facilita a introdução de aleatoriedade ausente no ASR original quando da composição do vetor **y**. ASR deterministicamente particiona um dado intervalo $[l, u]$ de acordo com o seu ponto médio, m. Aqui, por sua vez, sugere-se selecionar um ponto aleatoriamente em torno de m, $m' = l + \xi(u-$

l), onde ξ é uniformemente distribuído no intervalo $[0.5-\tau, 0.5+\tau]$, $0 \leq \tau \leq 0.5$. O ASR proposto será denotado por ASR2(τ). Pode-se perceber que ASR2(0) coincide com o ASR tradicional enquanto que ASR2(0.5) leva a um procedimento de particionamento totalmente aleatório do intervalo $[l, u]$. Em resumo, uma vez conhecidos ε, $h(l)$ e $h(u)$, o algoritmo proposto prossegue como dado abaixo:

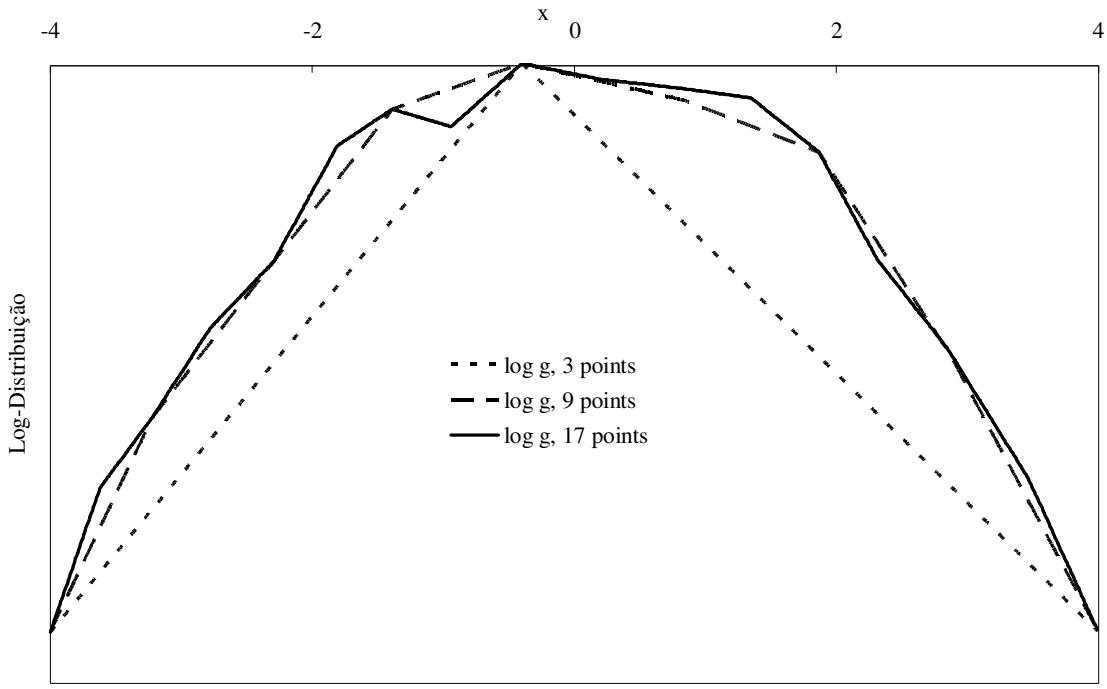
Algoritmo 4.1: Iteração da Regra Probabilística Adaptativa de Simpson

1. Gere um valor ξ de uma distribuição Uniforme($0.5-\tau$, $0.5+\tau$) e faça $m' = l + \xi(u-l)$;
2. Avalie o logaritmo da distribuição-alvo no ponto m', $h(m') = log\ f(m')$;
3. Ajuste duas funções lineares entre l e u, $g_x(l, m')$ e $g_x(m', u)$ e suas respectivas integrais nos intervalos de interesse, $A_h(l, m')$ e $A_h(m', u)$;
4. Compute ε_{lu} de acordo com a Equação 4.3;
5. Se $\varepsilon_{lu} > \varepsilon$ retorne à etapa 1, considerando primeiramente o intervalo $[l, m']$ e em seguida o intervalo $[m', u]$.

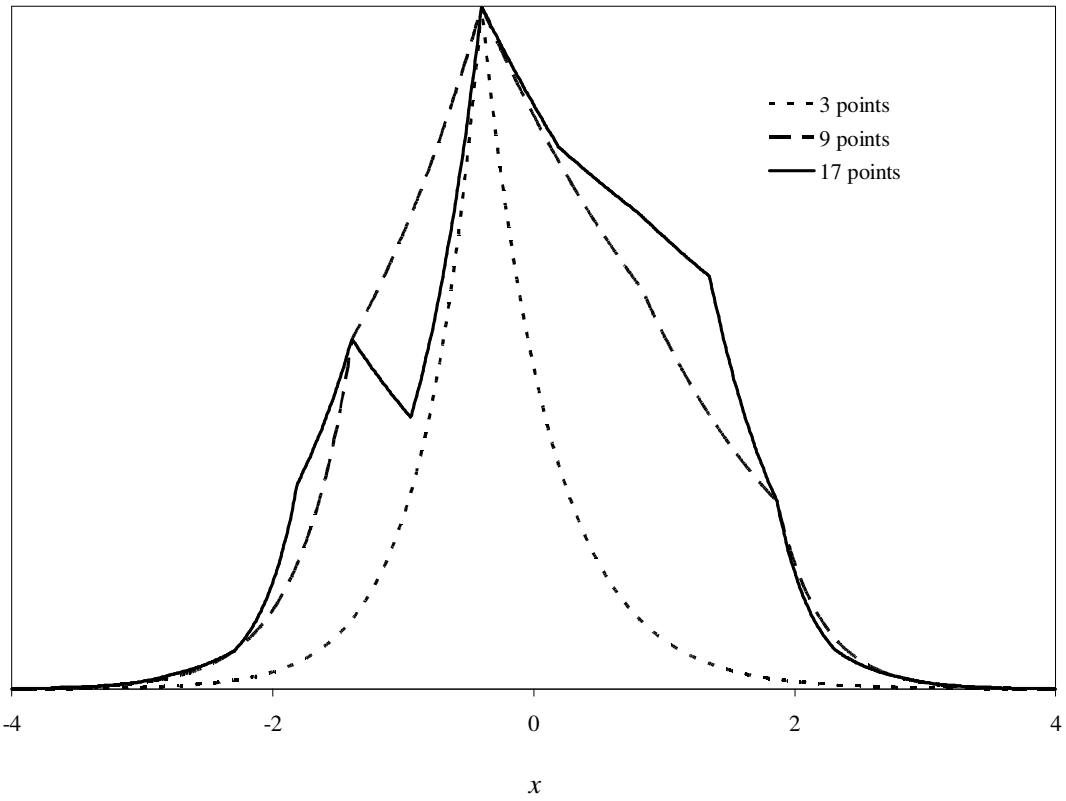
O desempenho do algoritmo é ilustrado através do seguinte exemplo:

Exemplo 4.1 *Deseja-se amostrar da seguinte distribuição condicional:* $f(x) \propto exp(-x^2/2)\ [sin(3x)^2 + 1]\ [cos(5x)^4 + 1]$, *uma tarefa árdua caso opte-se por adotar o método da transformação inversa.*

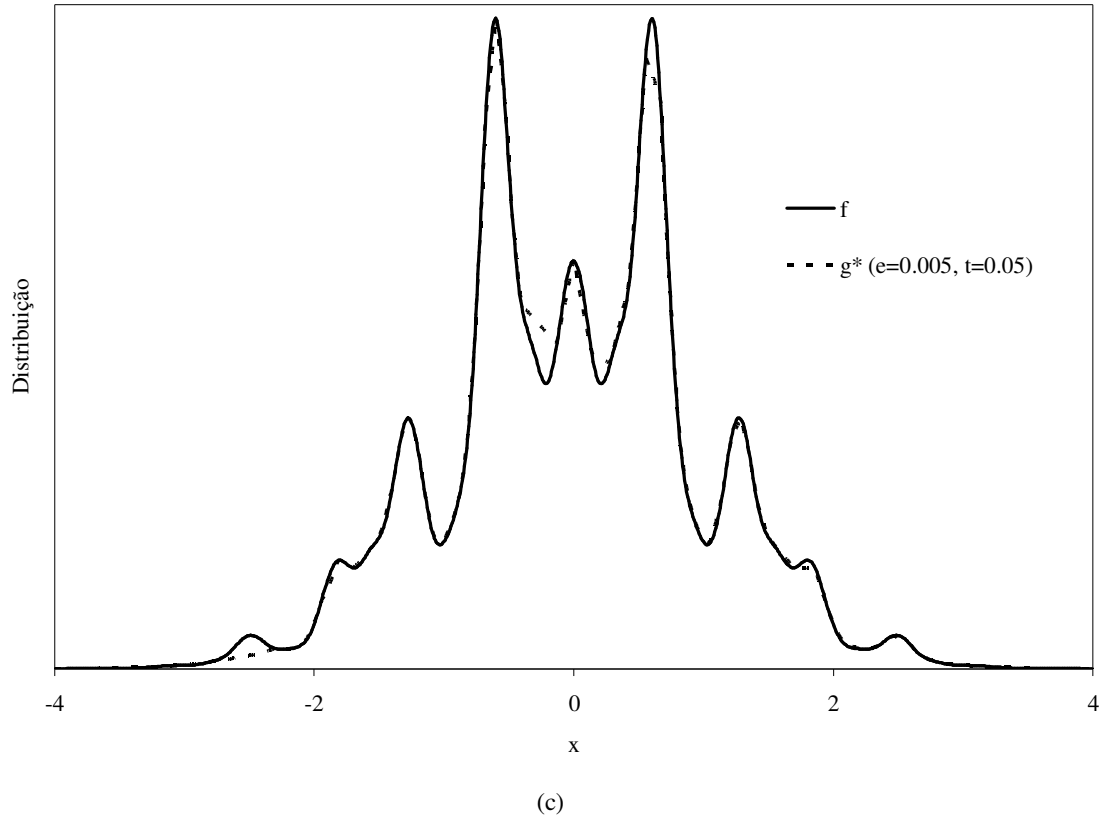
A Figura 4.1 esboça uma execução do algoritmo ASR2(τ) considerando $\tau = 0.05$ e $\varepsilon = 0.005$ na tentativa de inferir sobre $f(x)$, cuja amostragem direta mostra-se uma tarefa ao menos árdua. Devido à característica adaptativa de ASR2 de acordo com a comparação de áreas, este possibilita uma avaliação grosseira sobre o comportamento da função derivada de *log f*. Como no presente trabalho considera-se aproximações *log*-lineares, tal derivada coincide com o coeficiente angular das retas. A variação das derivadas do logaritmo da função em estudo rudimentarmente capturada por ASR2 pode ser visualizada ao comparar as curvas delineadas envolvendo 3, 9 e 17 pontos em (a). A Figura 4.1 (b) exibe as respectivas funções na escala original de *f*. Ao final de 61 avaliações de *f*, o erro relativo tolerável atribuído ($\varepsilon = 0.005$) é satisfeito em todos os intervalos envolvidos, resultando no ajuste final apresentado na Figura 4.1(c).


x
-4
-2
0
2
4
Log-Distribuição
log g, 3 points
log g, 9 points
log g, 17 points

(a)


3 points
9 points
17 points
-4
-2
0
2
4
x

(b)



(c)

Figura 4.1 Desempenho do algoritmo ASR2(0.05), sob ε = 0.005, para ajustar a curva f(x) ∝ exp(-x²/2) [sin(3x)² + 1] [cos(5x)⁴ + 1]: (a) as primeiras iterações na escala logarítmica, (b) as primeiras iterações na escala original e (c) a curva estimada resultante (com 61 pontos).

Vale comentar que o algoritmo recursivamente particiona o intervalo de interesse [*l, u*] de tal forma que o vetor de pontos **y** é definido de maneira ordenada, sem qualquer esforço adicional. Além disso, como trabalha-se no eixo logaritmo, as chances de haver problemas de aritmética de pontos flutuantes é sensivelmente reduzida, requerendo apenas a computação das áreas sob a função exponencial por partes resultante a fim de computar uma estimativa inicial para a função acumulada de interesse. Comente-se, também, a adequação do algoritmo para modelar funções associadas a quantidades discretas, requerendo apenas um ajuste na definição de *m*', que pode assumir o valor inteiro mais próximo daquele calculado no primeiro passo do Algoritmo 4.1, por exemplo. Notar-se-á no próximo capítulo que o esforço computacional adicional de ASR2(τ) é de fato desprezível quando comparado à redução do número de avaliações das funções-alvo que este promove em relação aos métodos tradicionais adotados para GGS, a fim de se alcançar a mesma medida de acurácia.

4.2. Centroidal Voronoi Tessellations

Em aplicações de GGS, uma fase importante consiste em determinar o vetor de pontos $\mathbf{z} = (z_1, z_2, \ldots, z_r)$ nos quais *Rao-Blackwellization* se baseia. Tal etapa é suprimida pelos algoritmos tradicionais de GGS uma vez que estes consideram o mesmo vetor de pontos para cada variável durante toda a simulação. Contudo, quando métodos adaptativos tais como ASR2(τ) são adotados, os vetores $\mathbf{y}_j$ ($j = 1, 2, \ldots, S$) usualmente diferem entre si e a definição apropriada do vetor $\mathbf{z}$ passa a ser um desafio. Um vetor inadequado pode levar a uma estimativa final grosseira ou desnecessariamente complexa (e custosa) da distribuição marginal de interesse, $\pi(x)$. De maneira a lidar com este problema, sugere-se aqui adotar técnicas de agrupamento tais como *centroidal Voronoi tessellations* (CVT). Dado um conjunto de pontos de entrada $\mathbf{w}=(w_1, w_2, \ldots, w_r)$ pertencentes a um domínio Ω, a região de Voronoi, V_i, correspondente ao ponto w_i consiste de todos os pontos em Ω que são mais próximos de w_i do que de qualquer outro ponto do conjunto $\mathbf{w}$, de acordo com uma medida de distância. O conjunto $\mathbf{V}=(V_1, V_2, \ldots, V_r)$ compõe uma partição de Ω e é conhecido como rede de Voronoi onde w_j é o centro de massa de V_j. Grosseiramente falando, CVT possui a característica de mais eficientemente, em termos estatísticos, distribuir r pontos de maneira a inferir sobre π. Eles promovem o estimador de menor variância para π.

A Figura 4.2 ilustra a proposta de CVT para um dado modelo bidimensional envolvendo as variáveis, X_1 e X_2. A figura exibe a distribuição dos pontos amostrados do modelo nas coordenadas (x_1, x_2). Considerando uma partição do espaço de possibilidades Ω em cinco conjuntos, $(V_1, V_2, \ldots, V_5)$, obtém-se as regiões envoltas pelas linhas tracejadas e as estimativas dos seus respectivos centros de gravidade (as bolas fechadas), $(w_1, w_2, \ldots, w_5)$.

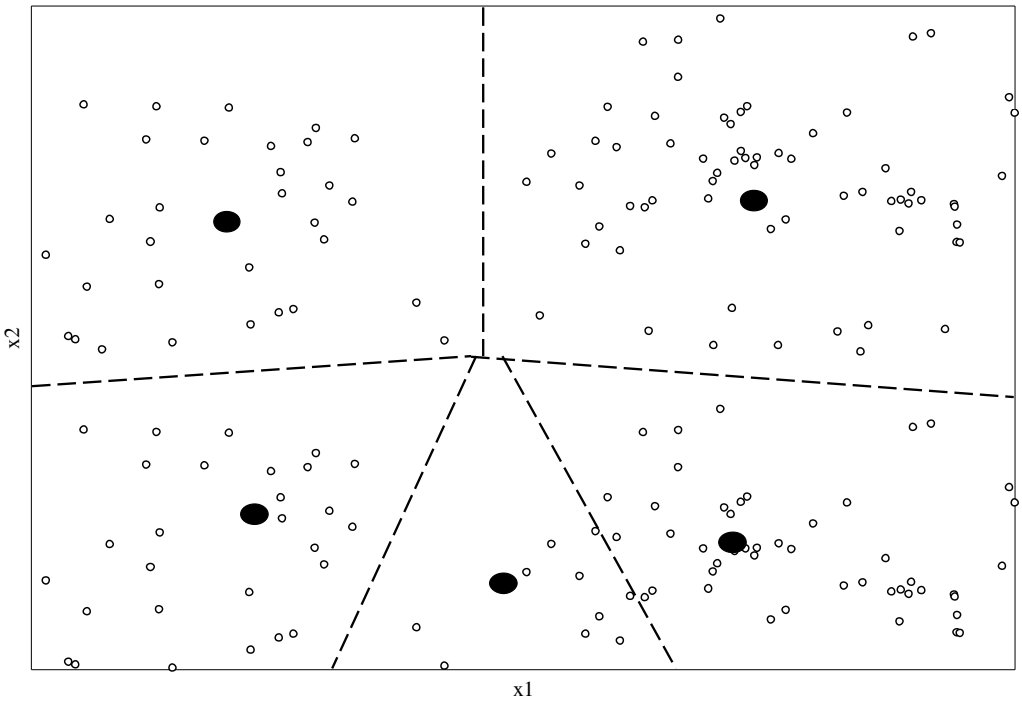


Figura 4.2 Esboço da proposta de CVT envolvendo 5 sub-regiões (r=5) do espaço de possibilidades de dado modelo bidimensional.

Recomenda-se Du & Gunzburger (2002) para maiores detalhes. Um dos principais algoritmos para computar CVT é o método de MacQueen. Dados Ω, r e π, o algoritmo se desenvolve como a seguir:

Algoritmo 4.2: Método de MacQueen

1. Selecione um conjunto inicial de r pontos $\mathbf{w}=(w_1, w_2, \ldots, w_r)$ de Ω de acordo com π; inicie os contadores de cada região $d_l = 1$, $l = 1, 2, \ldots, r$;
2. Amostre v de π;
3. Encontre o ponto w_l de $\mathbf{w}$ que mais se aproxima de v; denote o índice de w_l por l^*;
4. Atribua $w_{l^*} = \dfrac{d_{l^*} w_{l^*} + v}{d_{l^*} + 1}$ e incremente $d_{l^*} = d_{l^*} + 1$;
5. Se este novo conjunto de pontos não satisfizer ao critério de convergência adotado, retorne ao passo 2 do algoritmo.

Pode-se notar que o algoritmo de MacQueen é naturalmente aplicável aos pontos amostrados que compõem a cadeia de Markov subjacente ao método de MCMC. Após S iterações do período de estacionaridade da cadeia poder-se-ia adotar tal método a fim de

elaborar um modelo não-paramétrico customizado à amostra de cada variável. Contudo, o objetivo aqui é adotar o método de agrupamento de forma a identificar o melhor vetor de pontos para inferir sobre a distribuição marginal de interesse, π, a partir das distribuições condicionais ajustadas no decorrer das S iterações do método de MCMC. Neste contexto, adotar a amostra elaborada levaria a um desperdício de informações acerca do comportamento da função derivada de cada aproximação, grosseiramente obtidas durante as aplicações de ASR2(τ).

Diante de tais argumentos, propõe-se uma adaptação ao método de MacQueen. A idéia é elaborar um vetor do qual uma função exponencial por partes $\mathbf{g}^*$ possa estimar acuradamente a distribuição marginal de interesse π, a um baixo custo computacional. Neste sentido, Ω equivale ao suporte de π e $\mathbf{w}$ ao vetor de interesse z, enquanto que r é função do tamanho dos vetores delineados durante as simulações para a variável em estudo, isto é $r = Máximo\ [n_j\ (j = 1, 2, \ldots, T)]$ onde T é o número total de iterações executadas pelo método de MCMC. Com o objetivo de reduzir o consumo de recursos computacionais e fazer uso das informações acerca do comportamento das funções derivadas das aproximações adotadas, os passos 1 e 2 do algoritmo são adaptados. Ao invés de amostrar de π, os vetores computados $\mathbf{y}_j$ ($j = 1, 2, \ldots, S$) são considerados. Assim, no passo 1 são atribuídos a $\mathbf{w}$ os primeiros r elementos distintos de ($\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_S$) e o passo 2 baseia-se nos valores restantes deste conjunto. Desta forma, além de aumentar a eficiência computacional, promove-se uma elaboração inteligente do vetor de pontos para o qual *Rao-Blackwellization* ocorrerá, a partir de uma inferência sobre os pontos que levam a uma melhor representação das variações da função derivada de π.

A Figura 4.3 esboça um passo-a-passo do algoritmo a partir das distribuições condicionais ajustadas por ASR2(τ) no decorrer das simulações, representadas aqui por $\mathbf{g}_k$ e $\mathbf{g}_l$, e seus respectivos vetores de pontos de suporte. Ressalte-se que devido ao caráter adaptativo de ASR2(τ) os vetores de suporte de $\mathbf{g}_k$ e $\mathbf{g}_l$ são oportunamente diferentes. A estimativa final faz uso destes vetores quando da aplicação de *Rao-Blackwellization*, representada pelos retângulos pontilhados na última linha da figura, denotada por vetor ($\mathbf{g}^*$).

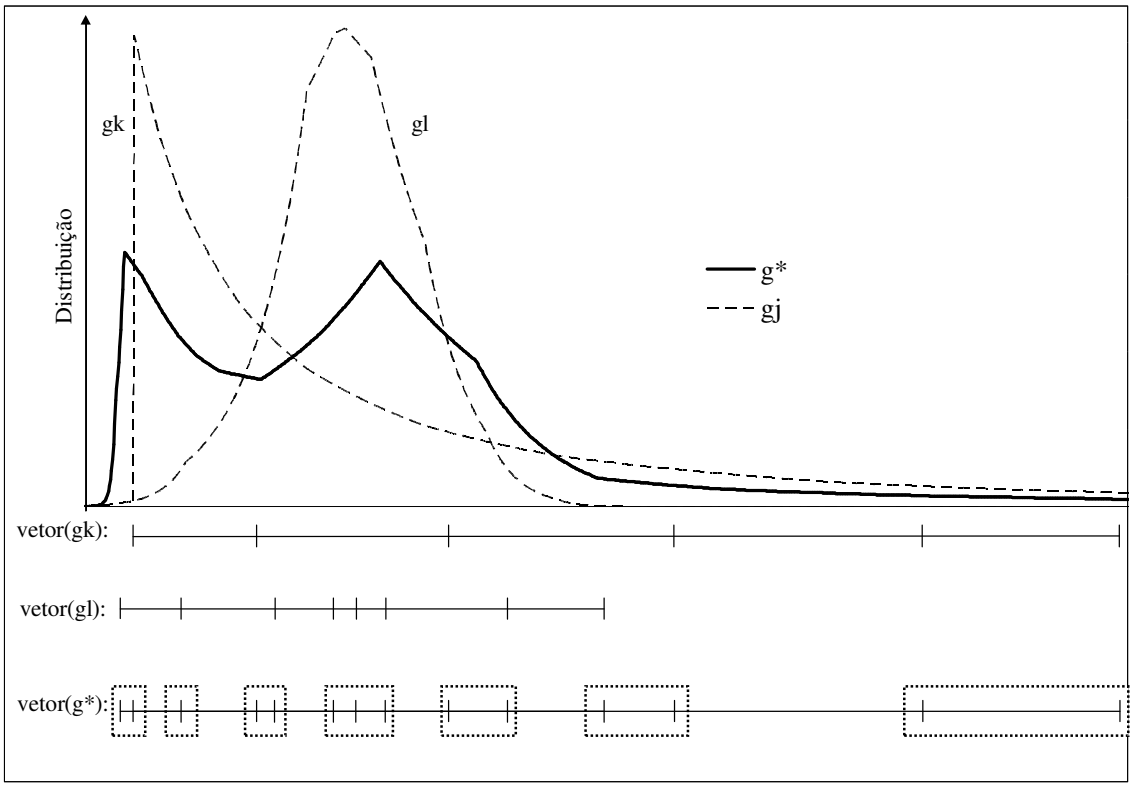


Figura 4.3 Ilustração do algoritmo Rao-Blackwellization adaptado para o presente trabalho. Os retângulos pontilhados indicam os grupos elaborados a partir dos vetores delineados pelo método ASR2(τ) durante as simulações.

4.3. GGS Proposto

Este capítulo se conclui com a introdução de uma variante de GGS que envolve conceitos de ASR, ARMS e CVT. ASR é adotado para determinar o vetor de pontos nos quais uma função exponencial por partes, **g**, é utilizada como proposta a uma função-alvo, *f*, cuja amostragem direta é inviável. Em seguida, amostra-se de **g** até que a fase de rejeição de ARMS acabe. A cada rejeição, **g** é aprimorado tal como feito em ARMS. Finalmente, amostra-se de **g** e, diferentemente do GGS tradicional, avalia-se se o valor obtido deve ser ou não aceito como proveniente de *f*. Todas as avaliações de *f* realizadas neste processo são incorporadas a **g**, de maneira a aumentar sua aderência a *f*. Ao final das simulações, identifica-se o período de estacionaridade do processo associado a cada variável de acordo com o método de Bauwens & Lubrano (1998), apresentado no capítulo anterior (seção 3.4). Para as *S* iterações do período de estacionaridade da variável de interesse *X*, aplica-se *Rao-Blackwellization* tal como descrito na seção anterior. Para dado erro relativo atribuído pelo

usuário (ε), nível de aleatorização de ASR2, τ, limites para o suporte de *f*, [*a, b*], valor corrente de *X* na cadeia de Markov subjacente (x_t) e denotando por x_{t+1} o próximo valor assumido por *X* após a iteração de GGS, o algoritmo é dado por:

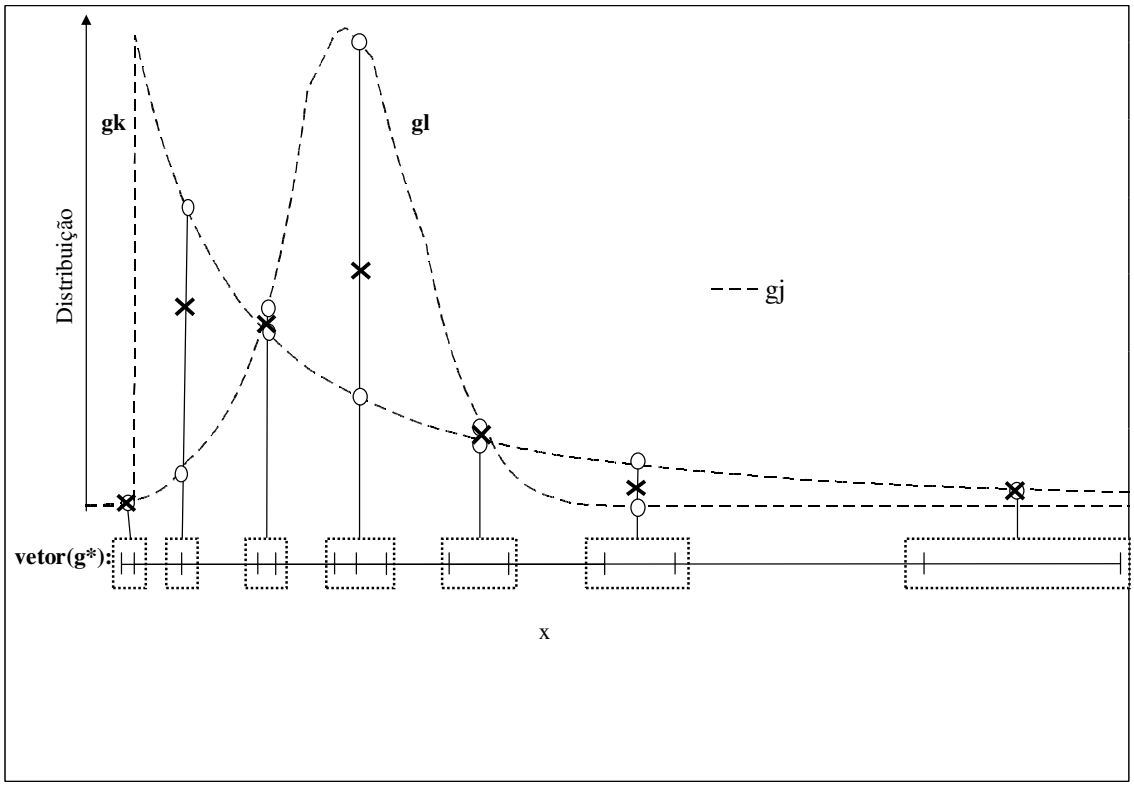
Algoritmo 4.3: Iteração do GGS Proposto

1. A partir do método ASR2(τ) determine o vetor de pontos, $\mathbf{y}_{t+1}$, no qual baseia-se a função proposta exponencial por partes, **g**;
2. A partir do método da transformação inversa, gere *w'* da acumulada associada a **g**, **G**;
3. Gere uma realização de *U*~ Uniforme[0, $\boldsymbol{g}(w')$], *u'*;
4. Se $u' > f(w')$ então introduza o par $(w', f(w'))$ no vetor $\mathbf{y}_{t+1}$ e aprimore **g** e retorne ao passo 2.
5. A partir do método da transformação inversa, gere *w* da acumulada associada a **g**, **G**;
6. Gere uma nova realização de *U*~ Uniforme[*0, 1*], *u*;
7. Se $u \leq P_{x_t,w} = Mínimo\left[1, \frac{f(w)g(x_t)}{f(x_t)g(w)}\right]$,então aceite *w* como proveniente de *f* e faça $x_{t+1} = w$;
8. Caso contrário rejeite *w* e faça $x_{t+1} = x_t$.

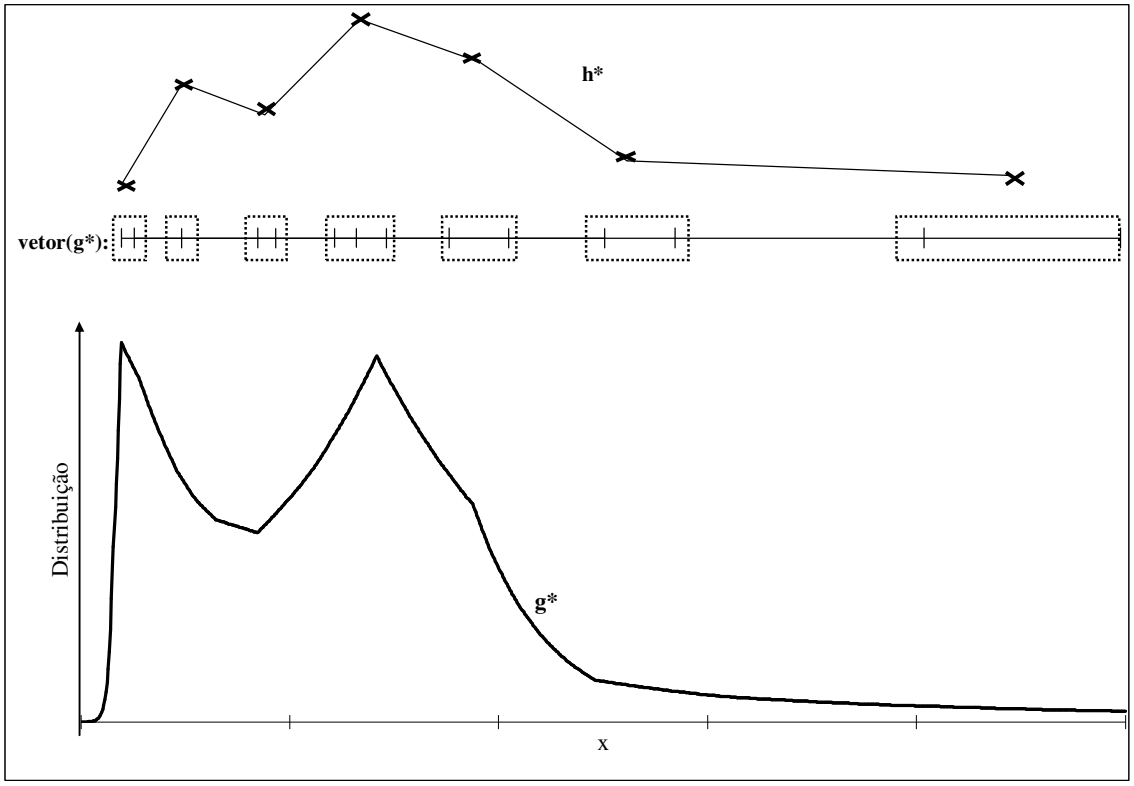
Como mencionado anteriormente, quanto maior a aderência de **g** à função *f*, maior a probabilidade de aceitação de MH, fazendo MH convergir para GS. Assim, é intuitivo perceber o potencial de ASR2(τ) na promoção de tal convergência, sob um natural aumento de esforço computacional. Em outros termos, além de dar suporte a GGS, ASR2(τ) pode ser embutido em MH, possibilitando uma ponderação entre a taxa de rejeição de MH e o tempo despendido para elaborar as distribuições propostas, através do erro relativo tolerável atribuído pelo usuário. Devido à sua simplicidade de implementação e liberdade contra suposições restritivas, ASR2(τ) pode emergir como uma alternativa robusta para o uso genérico de MH.

Por outro lado, embora não seja mencionado na literatura de GGS, considerar a possibilidade de rejeição dos pontos propostos pode ser crucial para o sucesso de GGS, uma vez que as distribuições propostas posteriormente usadas para *Rao-Blackwellization* são função dos valores assumidos pelas variáveis no decorrer do passeio aleatório realizado pela cadeia de Markov.

Ao final das simulações, um procedimento adicional mostra-se necessário a fim de se inferir sobre as distribuições marginais de interesse. A distribuição marginal, π, é estimada por $\mathbf{g}^*$, a distribuição média do conjunto das S aproximações computadas no período de estacionaridade da série associado à variável de interesse, X. Como o GGS proposto trabalha na escala logarítmica, tem-se um conjunto de funções $\mathbf{h}_j(z) = log(\mathbf{g}_j(z)) = \alpha_j + \beta_j x$, $j = 1, 2, \ldots, S$. Uma possível alternativa é elaborar $\mathbf{g}^*(z)$ a partir da função média, $\mathbf{g}'(z) = \frac{1}{S}\sum_{j=1}^{S}\frac{exp(\mathbf{h}_j(z))}{G_j(y_{n_j})}$, avaliada no vetor de pontos definido pelo CVT adaptado. Esta etapa é esboçada na Figura 4.4 (a), onde as bolas abertas indicam os valores de duas funções envolvidas de acordo com o vetor de pontos do CVT adaptado. Para cada um desses pontos, assume-se a função logarítmica $\mathbf{h}'(z) = log(\mathbf{g}'(z))$ e, em seguida, ajusta-se uma função linear por partes, $\mathbf{h}^*(z)$, sobre os pares $((z_i, \mathbf{h}'(z_i)), (z_{i+1}, \mathbf{h}'(z_{i+1})))$, como ilustrado na parte superior da Figura 4.4 (b). Finalmente, pode-se atribuir $\mathbf{g}^*(x) = exp(\mathbf{h}^*(x))$, para qualquer x no suporte de π (ver parte inferior da Figura 4.4 (b)).



(a) *Identificando o valor médio,* $\mathbf{g}'(z)$, *para cada z definido pelo CVT adaptado (as marcações cruzadas).*



(b) *Ajuste final (parte inferior da figura) a partir da função linear por partes ajustada (parte superior da figura) sobre o logaritmo dos pontos médios definidos pelo CVT adaptado exibido em (a).*

Figura 4.4 Cômputo das distribuições marginais do GGS proposto a partir da conversão logarítmica adotada.

O próximo capítulo dedica-se a avaliar o desempenho do algoritmo proposto em relação a métodos alternativos da literatura, enfatizando-se neste contexto o GGS tradicional.

5. CASOS DE ESTUDO

Neste capítulo, o desempenho do método de MCMC proposto é comparado a métodos alternativos tais como GGS tradicional (baseado em pontos igualmente espaçados e funções lineares por partes) em duas perspectivas. Primeiramente, a acurácia das distribuições marginais segundo GGS proposto e tradicional são comparadas. Para este propósito, MH e métodos de MCMC alternativos componentes do pacote © WinBUGS (versão 1.4) são adotados como referência. O *software* WinBUGS vem sendo desenvolvido pelo Colégio Imperial do Reino Unido desde 1996 a partir da versão (denominada BUGS) elaborada pelo Conselho de Pesquisa Médica, também do Reino Unido. O WinBUGS encapsula uma série de métodos para MCMC oportunamente escolhidos para manipular o modelo Bayesiano sob estudo, a depender de suas propriedades subjacentes. Dentre as propriedades consideradas pelo programa encontram-se a conjugação das distribuições condicionais que parametrizam o modelo, a log-concavidade das distribuições condicionais envolvidas nas iterações do método de MCMC e a natureza do espaço de possibilidades das variáveis do modelo. A medida de acurácia adotada deriva da distância por entropia cruzada entre a distribuição marginal estimada por GGS, $\mathbf{g}^*$, e a distribuição marginal esperada, π. Por outro lado, a eficiência computacional de GGS é estudada em termos de consumo de tempo e de número de avaliações das funções-alvo no processo de simulação. A principal motivação para este último reside no fato de que o primeiro é fortemente dependente dos recursos computacionais exigidos para a simulação, que em se tratando de GGS é por sua vez uma função direta da complexidade de avaliação das funções-alvo e da quantidade de aproximações componentes de cada função linear ou exponencial por partes. De qualquer forma, de maneira a promover uma plataforma para comparações, GGS foi implementado em linguagem de programação C++ e executado em um computador portátil com sistema operacional ©Windows com 1.83 GHz de capacidade.

Adicionalmente a uma comparação entre o GGS proposto e o tradicional, estuda-se a qualidade de inferência do primeiro em relação a métodos alternativos encontrados na literatura de RBs. O critério de convergência de séries adotado para GGS é aquele proposto por Bauwens & Lubrano (1998) sob $\gamma = 0.03$, comentado anteriormente (final do capítulo 3).

5.1. Medida de Distância de Kullback-Leibler

A distância por entropia cruzada possibilita a mensuração das divergências pontuais entre duas funções. Assim, seja π a distribuição marginal esperada e $\mathbf{g}^*$ sua estimativa, a distância

por entropia cruzada, também conhecida como informação de Kullback-Leibler (KL) ou distância de KL de $\mathbf{g}^*$ com respeito a π, tem sido adotada para avaliar a acurácia de estimativas às suas respectivas funções esperadas. Autores como Neil *et al.* (2007) e Keith *et al.* (2008) são alguns exemplos recentes. A distância de KL é definida como o valor esperado

$$\mathrm{KL}_\pi(\mathbf{g}^*) = E_\pi\left\{log\left[\frac{\pi(X)}{\mathbf{g}^*(X)}\right]\right\} = \int log\left[\frac{\pi(x)}{\mathbf{g}^*(x)}\right]\pi(x)dx. \qquad \textit{Equação 5.1}$$

É fácil ver que a distância de KL se anula caso $\mathbf{g}^* = \pi$, assim quanto maior a aderência de $\mathbf{g}^*$ a π menor será a distância de KL associada. Contudo, pode-se também perceber que porções negativas de $log(\pi/\mathbf{g}^*)$ decrescem as positivas quando do cálculo de KL_π $(\mathbf{g}^*)$, possibilitando a anulação da distância de KL mesmo para estimativas sensivelmente discrepantes da função esperada. De maneira a contornar este problema, considera-se aqui uma medida de variabilidade ao invés da esperança de $log[\pi(X)/\mathbf{g}^*(X)]$:

$$\mathrm{VKL}_\pi(\mathbf{g}^*) = E_\pi\left\{\left\{log\left[\frac{\pi(X)}{\mathbf{g}^*(X)}\right]\right\}^2\right\} - [\mathrm{KL}_\pi(\mathbf{g}^*)]^2. \qquad \textit{Equação 5.2}$$

Um bom estimador para $\mathrm{VKL}_\pi(\mathbf{g}^*)$ pode ser baseado em um vetor de K+1 pontos igualmente espaçados no suporte de π (limitado por $[a, b]$), resultando na estimativa

$$\hat{vkl}_\pi(\mathbf{g}^*) = \frac{b-a}{K(N+1)}\left\{\sum_{\substack{j=0,\\ f(v_j)>0}}^{K}\left\{log\left[\frac{\pi(v_j)}{\mathbf{g}^*(v_j)}\right]\right\}^2\pi(v_j) - \hat{KL}_\pi(\mathbf{g}^*)^2\right\} \qquad \textit{Equação 5.3}$$

onde $\hat{KL}_\pi(\mathbf{g}) = \sum_{\substack{j=0,\\ f(v_j)>0}}^{K} log\left[\frac{\pi(v_j)}{\mathbf{g}^*(v_j)}\right]\pi(v_j)$, $v_j = a + j(b\text{-}a)/K$, e N é o número de vezes em que $\pi(v_j) > 0$, $j = 0, 1, \ldots K$. Para os casos de estudo considerados, adota-se $K = 1000$.

Por fim, de maneira a avaliar o nível de variabilidade das estimativas provenientes do GGS proposto em relação ao tradicional, trinta triagens do método de MCMC baseado no GGS proposto e daquele baseado no GGS tradicional são replicadas e a relação entre as medidas de acurácia dos métodos são computadas em cada triagem:

$$vkl_{ratio}(i) = \frac{vkl_{trad}(i)}{vkl_{prop}(i)} \ (i = 1, 2, ..., 30), \qquad \textit{Equação 5.4}$$

possibilitando o estudo de medidas estatísticas de dispersão e posição associadas.

5.2. MBs Unidimensionais

Antes de dar início às comparações, um estudo inicial sobre o parâmetro τ do ASR2(τ) mostra-se pertinente. Para tanto, apresenta-se o caso a seguir.

Caso 5.1 *Deseja-se estimar a distribuição-alvo f introduzida no Exemplo 3.2, página 45.*

A Figura 5.1 exibe a medida de acurácia associada ao GGS tradicional e ao proposto sob $\tau = 0$ e $\tau = 0.1$ como uma função do tamanho do vetor de pontos (n) e do erro relativo atribuído pelo usuário, ε. Pode-se ver que o GGS tradicional apresenta os melhores resultados apenas para os menores valores de n, quando o nível de imprecisão das estimativas é razoavelmente elevado. Por outro lado, ASR2(0.1) promove índices ao menos tão bons quanto aqueles baseados em ASR, principalmente para pequenos valores de n. Tal característica é especialmente útil quando a distribuição-alvo é custosa de se avaliar, tal como aquelas subjacentes a MBs com um grande número de evidências empíricas.

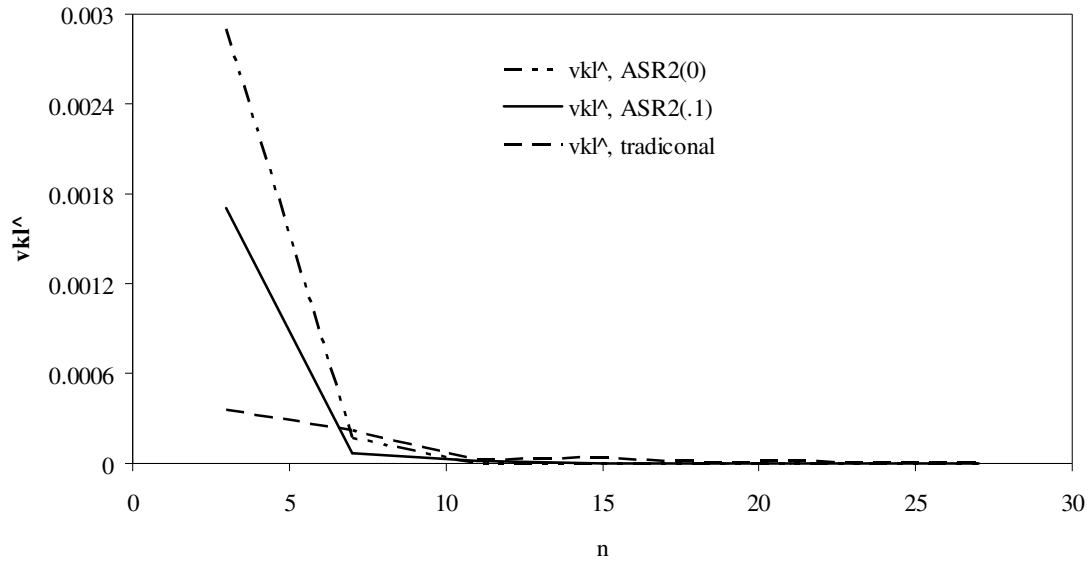


Figura 5.1 Desempenho do GGS tradicional (baseado em funções lineares por partes e pontos igualmente espaçados) e proposto (com τ=0 e τ=0.1) para estimar a distribuição relacionada ao Caso 5.1 (K = 1000).

Os próximos casos permitem uma comparação entre o GGS tradicional e o proposto em termos de acurácia.

Caso 5.2 *O modelo univariado considerado por Keith et al. (2008) é estudado entre os limites (-15, 20):*

$f(x) = 0.25 \cdot f_1(x) + 0.7 \cdot f_2(x) + 0.05 \cdot f_3(x)$, onde
$f_1(x) \equiv$ distribuição Normal, (média, variância) = (-6.0, 2.0);
$f_2(x) \equiv$ distribuição Normal, (média, variância) = (0.0, 1.0);
$f_3(x) \equiv$ distribuição Normal, (média, variância) = (15.0, 0.1).

Considerando ε = 5E-03 e τ = 1E-04, f é avaliado 43 vezes de acordo com dada aplicação de ASR2(1E-04), resultando em uma medida de discrepância igual a 8.27E-04. Tal discrepância aumenta para 7.54E-03 quando GGS tradicional é considerado para o mesmo número de avaliações de f , um valor mais que nove vezes maior que o primeiro. Para MBs univariados, as distribuições propostas pelo GGS tradicional não variam, sendo também idênticas à estimativa final do método. Por outro lado, devido ao fator aleatório de ASR2(τ) e ARMS, aprimoramentos sobre cada distribuição condicional estimada pelo GGS proposto podem ocorrer. Assim, no presente exemplo a discrepância associada ao GGS tradicional é de 7.54E-03. Por sua vez, uma simulação baseada em 100 iterações do GGS proposto (resultando em 4308 avaliações de f) reduz a discrepância para 8.12E-04 – uma discrepância 2% menor que a de uma única aplicação de ASR2(1E-04). A Figura 5.2 exibe os ajustes para f baseados nas 100 iterações de GGS.

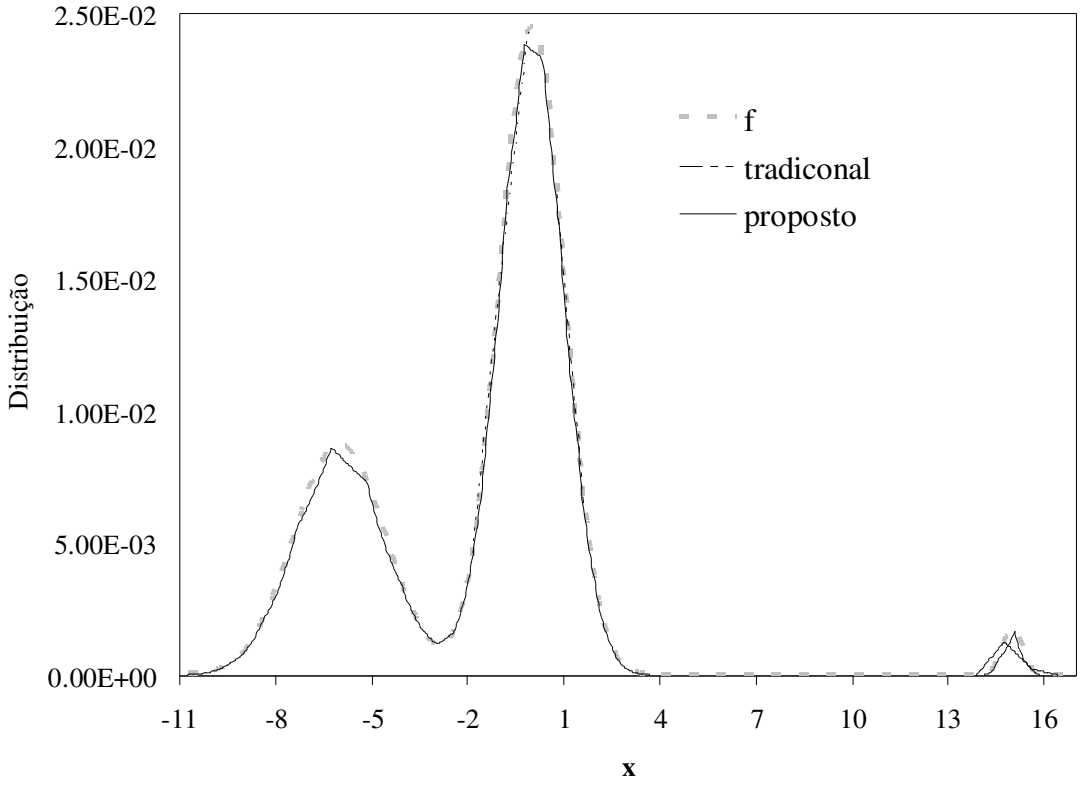


Figura 5.2 GGS tradicional (43 pontos) e proposto (sob ε = 5E-3 e τ =1E-4) para inferior sobre o Caso 5.2 após 100 iterações.

De maneira a avaliar a variabilidade dos resultados provenientes do GGS proposto em relação ao tradicional (que para casos univariados não apresenta variabilidade nos resultados), trinta triagens do método de MCMC baseado no GGS proposto são replicadas e a relação exposta na Equação 5.4 é aplicada a cada uma. Aqui, a média amostral de vkl_{ratio} foi de 9.39, indicando que espera-se que a discrepância do GGS tradicional equivalha a 9.39 vezes a do proposto. O desvio-padrão associado foi de 0.19, com razões variando de 8.995 a 9.821, levando ao intervalo de confiança para a razão média entre [9.32, 9.46] sob um nível de significância de 5%. Isto indica uma precisão do GGS proposto sempre expressivamente superior à do tradicional.

Caso 5.3 *O modelo univariado extraído de Robert & Casella (2004) é estudado entre os limites (-4, 4):*

$$f(x) \propto exp(-x^2/2)[sin(6x)^2 + 3cos(x)^2 sin(4x)^2 + 1]$$

A partir de ε = 7E-3 e τ = 0.01, uma dada iteração do GGS proposto avaliou f 42 vezes, resultando em uma medida de discrepância de 4.79E-04. Por sua vez, a medida de discrepância do GGS tradicional para o mesmo número de avaliações de f foi de 5.28E-04, quase 10% maior que o primeiro. Após 100 iterações (resultando em 4301 avaliações de f), a discrepância da distribuição marginal estimada pelo GGS proposto decresceu para 4.66E-04. A Figura 5.3 mostra os ajustes para f de acordo com o GGS tradicional e proposto.

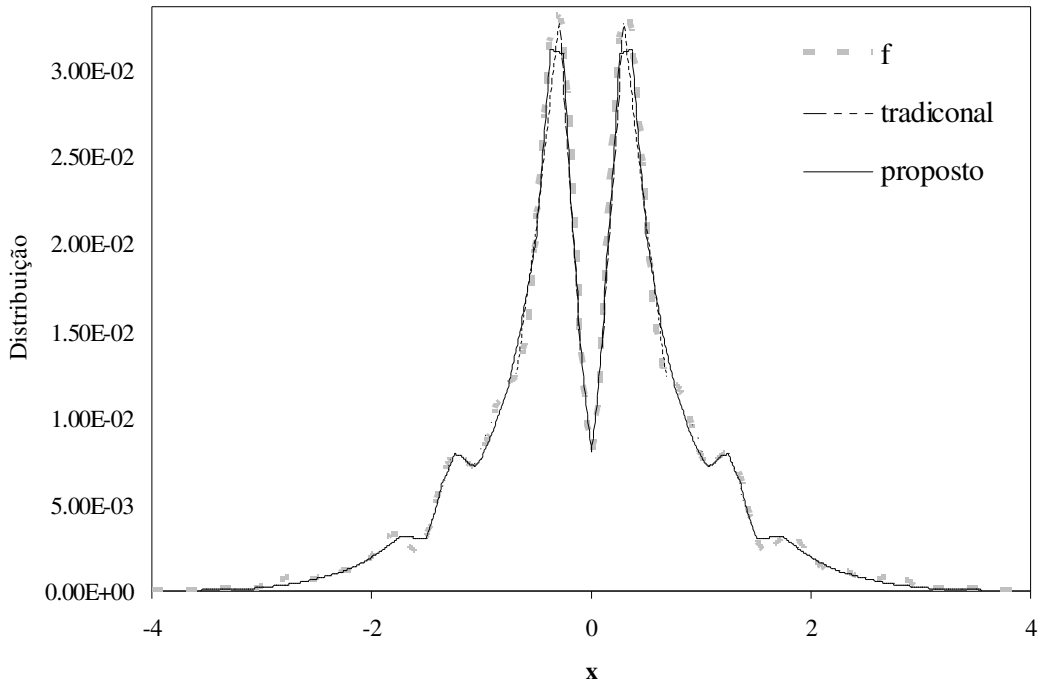


Figura 5.3 Estimativas do GGS tradicional (42 pontos) e proposto (sob ε = 7E-3 e τ = 0.01) para o Caso 5.3 após 100 iterações.

Em relação à variabilidade nas estimativas do GGS proposto de acordo com a Equação 5.4, a média de vkl_{ratio} foi de 1.31, sugerindo um erro 31% maior do GGS tradicional em média quando comparado ao GGS proposto. O desvio-padrão amostral da medida foi de 0.11, com valores variando entre 1.13 e 1.46, refletindo que em todas as trinta triagens o GGS proposto se mostrou melhor que o tradicional. O intervalo de confiança com 5% de significância para a razão média é [1.27, 1.35].

Nos próximos casos, um estudo sobre MBs multidimensionais é realizado.

5.3. MBs Multidimensionais

Aborda-se aqui MBs multidimensionais extraídos da literatura.

Caso 5.4 *A RB mista envolvendo o problema de previsão do consumo de energia, estudada por Brewer et al. (1996) e exibida na Figura 5.4, é considerada. Aqui, uma amostragem direta das distribuições condicionais das variáveis contínuas P (preço da energia) e G (consumo doméstico) não é possível, pois estas envolvem o produto entre as distribuições Normal e Gamma. Por outro lado, amostrar da distribuição condicional de D (demanda de energia) é possível, assim como ocorre com as variáveis categóricas C (introdução de taxas) e T (melhorias na eficiência técnica).*

Com o propósito de avaliar o desempenho de GGS, todas as distribuições condicionais são aproximadas. Considerando os limites [0, 10] para todas as variáveis contínuas e sob ε = 8E-3 e τ = 0.05, uma execução do GGS proposto baseado em 1000 iterações, após um período de *burn-in* de 500 iterações é observada. As distribuições-alvo foram avaliadas em um total de 125824 vezes (não mais que 31 avaliações de cada distribuição-alvo) durante o processo de simulação, 88 pontos foram rejeitados devido à probabilidade de MH (6% do total), e registrou-se uma probabilidade mínima de aceitação de 0.59.

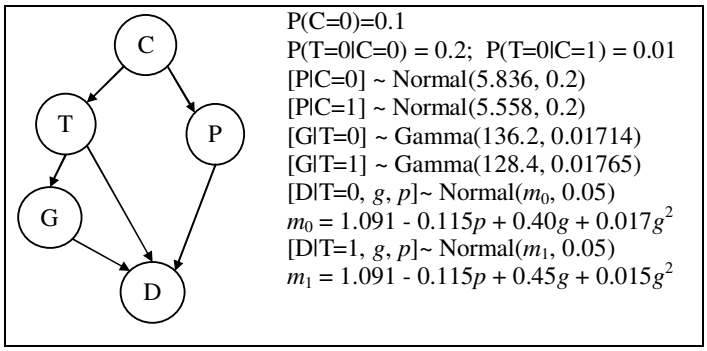


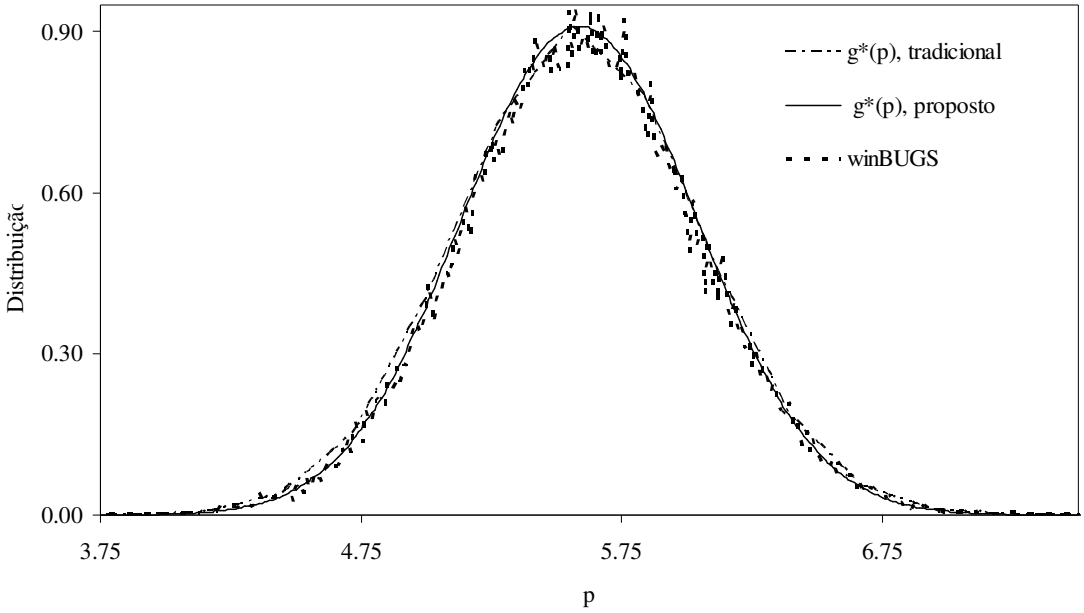
Figura 5.4 RB extraída de Brewer et al. (1996).

Assumindo o maior número de avaliações do GGS proposto (31 pontos) para a aplicação do GGS tradicional, observou-se 145500 avaliações das distribuições-alvo (cerca de 16% maior que o do GGS proposto). A Tabela 5.1 apresenta algumas medidas estatísticas das distribuições marginais estimadas pelo GGS proposto e tradicional. Os valores esperados são provenientes de simulações realizadas no pacote ©WinBUGS, sob 65536 iterações após um período de *burn-in* de 10000. Pode-se perceber a maior acurácia do GGS proposto mesmo sob um menor número de avaliações das distribuições-alvo envolvidas.

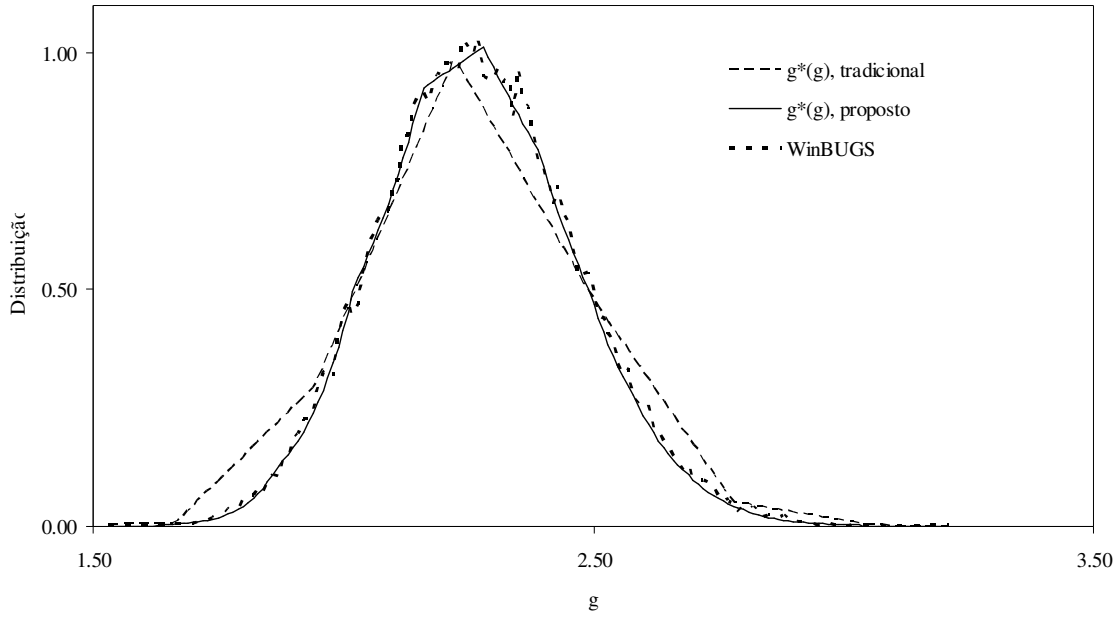
Tabela 5.1 Medidas estatísticas relacionadas às estimativas provenientes do GGS proposto e tradicional, baseando-se em uma amostra de 1000 pontos, após um período de burn-in de 500 iterações, para a RB extraída de Brewer et al. (1996).

			GGS	
Variável	Parâmetro	Esperado	Tradicional	Proposto
C	Pr(C=0)	0.10	0.09	0.10
T	Pr(T=0)	0.03	0.01	0.01
	Média	5.59	5.59	5.59
P	Variância	0.21	0.23	0.21
	Média	2.27	2.26	2.27
G	Variância	0.04	0.06	0.04
	Média	1.54	1.53	1.55
D	Variância	0.06	0.09	0.06

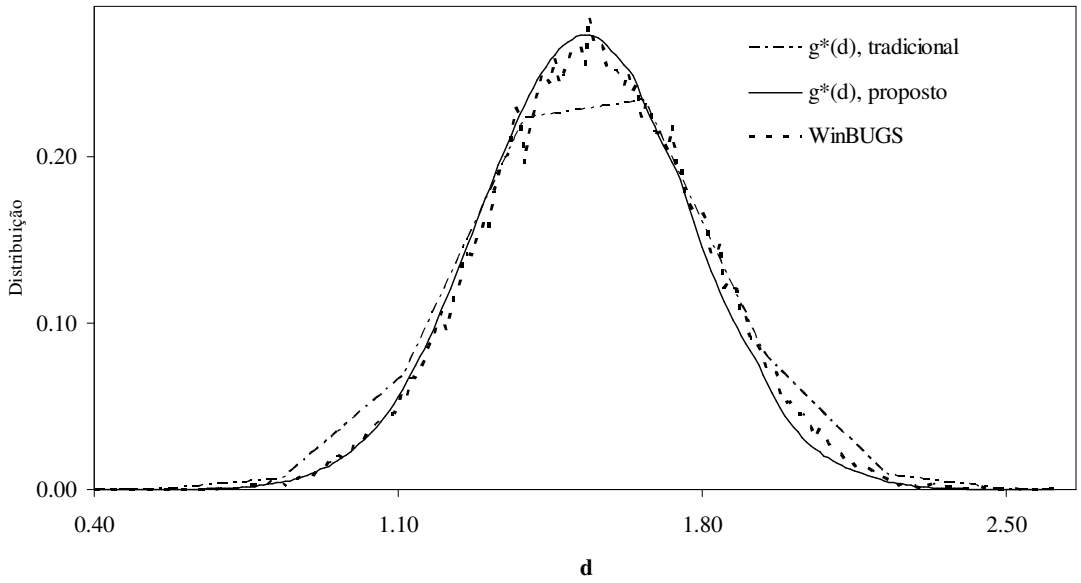
A Figura 5.5 exibe as distribuições marginais das variáveis contínuas estimadas pelos GGS proposto e tradicional e também a distribuição esperada de acordo com o pacote © WinBUGS. Em relação à acurácia, constata-se a maior precisão do GGS proposto sobre o tradicional nesta triagem dos métodos. Comente-se a maior suavidade obtida pelo GGS proposto em relação aos resultados do próprio WinBUGS. Mesmo envolvendo um menor número de iterações do método de MCMC, aplicar *Rao-Blackwellization* possibilita um ajuste mais suave (e provavelmente mais plausível) que aquele do próprio WinBUGS, cuja estimativa fundamenta-se nos valores amostrados durante as simulações. Para promover tal nível de suavidade, o GGS proposto requereu 3.781 segundos e o WinBUGS 4.021 segundos.



(a) GGS tradicional: *burn-in* de 500 pontos, *vkl*^ = 0.01; GGS proposto: *burn-in* de 500 pontos, *vkl*^ = 0.006.



(b) tradicional: *burn-in* de 500 pontos, *vkl*^ = 0.079; GGS proposto: *burn-in* de 500 pontos, *vkl*^ = 0.005.



(c) tradicional: *burn-in* de 500 pontos, *vkl*^ = 0.046; GGS proposto: *burn-in* de 500 pontos, *vkl*^ = 0.006.

Figura 5.5 Distribuições marginais das variáveis contínuas envolvidas na RB extraída de Brewer et al. (1996) de acordo com GGS tradicional e proposto e WinBUGS.

A Tabela 5.2 exibe o comportamento da medida de variabilidade dos resultados provenientes dos GGS proposto e tradicional, a partir da Equação 5.4, para as variáveis contínuas. Note-se que em média, o GGS proposto apresentou sempre um melhor desempenho. Embora que para a variável P tenham sido observadas simulações nas quais o GGS tradicional apresentou resultados melhores (mais precisamente 14% melhores de acordo

com o mínimo registrado), para as demais variáveis o desempenho do GGS proposto mostrou-se expressivamente melhor, onde o GGS tradicional alcança uma discrepância quase que doze vezes a do proposto. Os intervalos de confiança a 5% de significância para a razão média indicam que mesmo para P tem-se um melhor desempenho do método proposto.

Tabela 5.2 Estudo de variabilidade de resultados dos GGS tradicional e proposto (a partir da Equação 5.4) para as variáveis contínuas da RB extraída de Brewer et al. (1996).

Variável	Métrica	vkl_{ratio}
P	Média	1.12
	Desvio-padrão	0.13
	Mínimo	0.86
	Máximo	1.47
	IC(5%, média)	1.08 1.17
G	Média	8.52
	Desvio-padrão	2.46
	Mínimo	3.04
	Máximo	11.95
	IC(5%, média)	7.64 9.39
D	Média	5.27
	Desvio-padrão	1.21
	Mínimo	1.42
	Máximo	6.68
	IC(5%, média)	4.84 5.70

Caso 5.5 *O MB de confiabilidade humana encapsulado em uma RB mista por Langseth et al. (2009) é estudado. Aqui, a relação causal entre as variáveis contínuas Z_1 e Z_2 (usadas para modelar características ambientais tais como nível de iluminação e ruído) e as variáveis categóricas T_1, T_2, T_3 e T_4 (adotadas para representar a habilidade de o indivíduo realizar (in)adequadamente a i^a tarefa) é descrita por meio de funções logísticas, onde p denota a probabilidade de T_i assumir o valor 0 (o indivíduo realizar adequadamente a i^a tarefa). Este problema é especialmente importante devido a Langseth et al. (2009) terem o adotado para comparar métodos alternativos de manipulação de RBs mistas. A RB correspondente é exibida na Figura 5.6.*

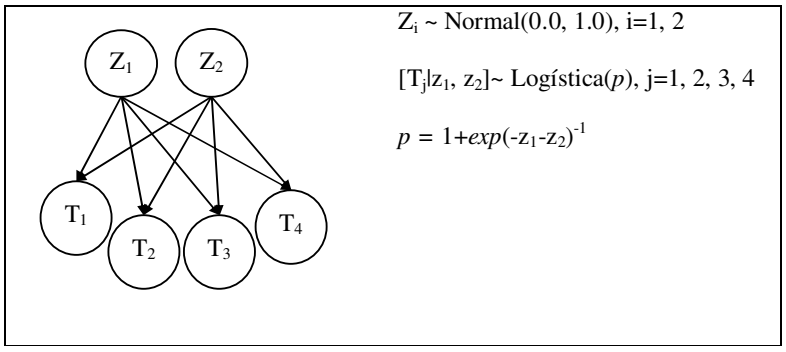


Figura 5.6 RB mista extraída de Langseth et al. (2009).

Quatro métodos foram considerados por Langseth *et al.* (2009): discretização, misturas de exponenciais truncadas (MTEs do inglês *mixtures of truncated exponentials*), métodos variacionais e o algoritmo de MH assumindo distribuições propostas Normais. Duas medidas de confiabilidade foram consideradas, a probabilidade conjunta $Pr(T_1=0, T_2=0, T_3=0, T_4=0)$ e a distribuição conjunta a posteriori $f(z_1, z_2| T_1=0, T_2=0, T_3=0, T_4=0)$. De maneira a comparar o GGS proposto às alternativas consideradas por Langseth *et al.* (2009), a regra do produto é aplicada à primeira métrica: $Pr(T_1=0, T_2=0, T_3=0, T_4=0)= Pr(T_1=0)\ Pr(T_2=0| T_1=0)\ Pr(T_3=0| T_1=0, T_2=0)\ Pr(T_4=0| T_1=0, T_2=0, T_3=0)$. Cada probabilidade é estimada baseando-se em ε = 2E-3 e τ = 0.01 e 1500 iterações do GGS proposto (as primeiras 500 iterações foram descartadas). Z_1 e Z_2 são limitados por [-5, 5]. Quatro execuções do método de MCMC baseado no GGS proposto (uma execução por probabilidade) envolveram funções aproximadas com um vetor de pontos com tamanho menor que 24, resultando em 143 rejeições segundo a probabilidade de MH (cerca de 0.53% dos valores propostos).

A Tabela 5.3 mostra os resultados a partir do GGS proposto e dos métodos alternativos. Pode-se perceber que neste caso, a acurácia do GGS proposto é superada apenas pelo algoritmo de MH. Vale ressaltar que o bom desempenho do método de MH adotado por Langseth *et al.* (2009) é principalmente devido ao fato de eles amostrarem realizações de Z_1 e Z_2 de maneira independente a partir de uma Normal-padrão, anulando a probabilidade de rejeição, e em seguida aplicarem *Rao-Blackwellization* sobre $Pr(T_1=0, T_2=0, T_3=0, T_4=0)$; tal procedimento pode ser entendido como uma extensão do *logic sampling*, um método cujo uso é geralmente adequado a RBs sem variáveis instanciadas, apenas. Para maiores detalhes sobre este algoritmo, recomenda-se Henrion (1988).

Tabela 5.3 Desempenho dos métodos alternativos para manipulação de RBs mistas considerados por Langseth et al. (2009) e do GGS proposto na estimação de $\theta=Pr(T_1=0, T_2=0, T_3=0, T_4=0)$.

Método	θ	Erro relativo
Discretização	0.176999	0.01803
MTE	0.176819	0.01699
MH	0.174660	0.00457
GGS proposto	0.172480	0.00797
Esperado	0.173865	0.00000

Em relação ao segundo parâmetro de confiabilidade, uma alternativa a $f(z_1, z_2| T_1=0, T_2=0, T_3=0, T_4=0)$ é considerada, uma vez que o GGS proposto foi desenvolvido para estimar

distribuições marginais univariadas. Assim, retornando à RB esboçada na Figura 5.6, o desempenho do GGS proposto é avaliado ao inferir sobre $f(z_i|\ T_1=0, T_2=0, T_3=0, T_4=0)$, i=1, 2, sob $\varepsilon = \tau =$ 1E-02 e 1500 iterações (as primeiras 800 e 850 iterações para Z_1 e Z_2, respectivamente, foram descartadas). Não mais que 21 avaliações da mesma distribuição-alvo ocorreram (46618 avaliações ao longo das simulações) e 154 pontos foram rejeitados de acordo com a probabilidade de MH. De maneira a favorecer o GGS tradicional, este foi baseado em vetores de 21 pontos (ao todo, 26% a mais que o GGS proposto). Como referência, o pacote ©WinBUGS é adotado a fim de computar as distribuições a *posteriori* esperadas. Uma execução do WinBUGS baseada em 75536 iterações (as primeiras 10000 foram descartadas) é considerada. A Figura 5.7 apresenta a acurácia dos GGS proposto e tradicional ao estimar a distribuição a *posteriori* de Z_1. Aqui, a acurácia de ambos os métodos se equiparou, embora o GGS proposto tenha recorrido às distribuições-alvo um número sensivelmente menor de vezes. Vale ressaltar, mais uma vez, o nível de suavidade de GGS em comparação com as inferências a partir do ©WinBUGS. Nas execuções observadas, o tempo de simulação do GGS proposto foi de 3.213 segundos, o do GGS tradicional de 3.906 segundos e do WinBUGS de 3.974 segundos.

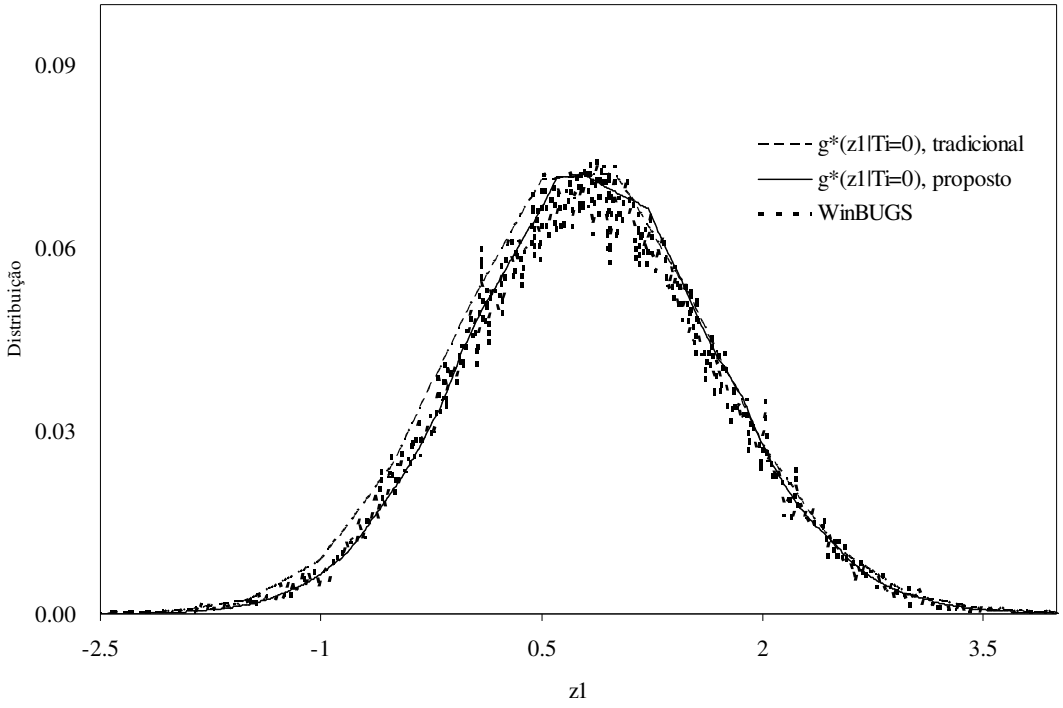


Figura 5.7 Distribuição a posteriori, $f(z_1|\ T_1=0, T_2=0, T_3=0, T_4=0)$, para a RB mista extraída de Langseth et al. (2009). GGS tradicional: burn-in de 350 pontos, vkl^ = 0.0097; GGS proposto: burn-in de 800 pontos, vkl^ = 0.0097

Em termos de variabilidade dos resultados, considerando Z_1, obteve-se vkl_{ratio} = 1.10 em média, com desvio-padrão de 0.11, tendo valores variando entre 0.91 e 1.45, indicando que embora o GGS tradicional (envolvendo um maior consumo computacional) tenha apresentado melhores resultados em alguns casos, espera-se que um melhor desempenho seja alcançado caso opte-se pelo GGS proposto. Isto é também indicado pelo intervalo de confiança (sob um nível de significância de 5%) calculado para a razão média das trinta replicações dos métodos de GGS: [1.06, 1.14].

Caso 5.6 *O modelo de regressão não-linear Bayesiano associado à demanda de oxigênio bioquímico apresentado por Ritter & Tanner (1992) é estudado. Neste caso, amostra-se de duas distribuições de probabilidades condicionais:*

$f(\theta_i|\theta_j) \propto (\sum(y_k - \mu_k)^2)^{-2}$, onde $i \neq j = 1, 2$,

$\mu_k=\theta_1(1-exp(-\theta_2 x_k))$, x = (1, 2, 3, 4, 5, 7), y = (8.3, 10.3, 19.0, 16.0, 15.6, 19.8).

Considerando os mesmos limites propostos por Ritter & Tanner (1992) ([-20, 50] para θ_1 e [-2, 6] para θ_2), a Figura 5.8 ilustra as estimativas dos GGS proposto e tradicional para a distribuição marginal de θ_2. Como referência, 70000 iterações do algoritmo de MH sob um período de *burn-in* de 4750 pontos foram também computadas. Vale comentar que as distribuições propostas consideradas no algoritmo de MH foram delineadas pelo método ASR2(0.05) sob ε = 3E-4, resultando em uma taxa de rejeição de 1.9%. Considerando τ = 0.05, ε = 5E-4 e 2150 amostras, o GGS proposto descartou os primeiros 450 pontos de θ_2, de acordo com o critério de convergência de Bauwens & Lubrano (1998), e avaliou as distribuições-alvo envolvidas 117421 vezes (com vetores variando de 7 a 35 pontos), resultando em 125 rejeições segundo a probabilidade de MH (2.9% dos valores propostos). A respectiva medida de acurácia da distribuição marginal de θ_2 assumiu o valor $vkl\hat{}$ = 0.017. Com o propósito de alcançar a mesma precisão em função da mesma quantidade de iterações do método de MCMC (2150 ao todo), aplicou-se o GGS tradicional baseado em 45 pontos igualmente espaçados, resultando em 193500 avaliações das distribuições-alvo (65% mais que o GGS proposto). Foram descartadas as primeiras 1150 amostras de θ_2. O tempo de simulação do GGS tradicional foi de 12.375 segundos enquanto que o do GGS proposto foi de 8.156 segundos.

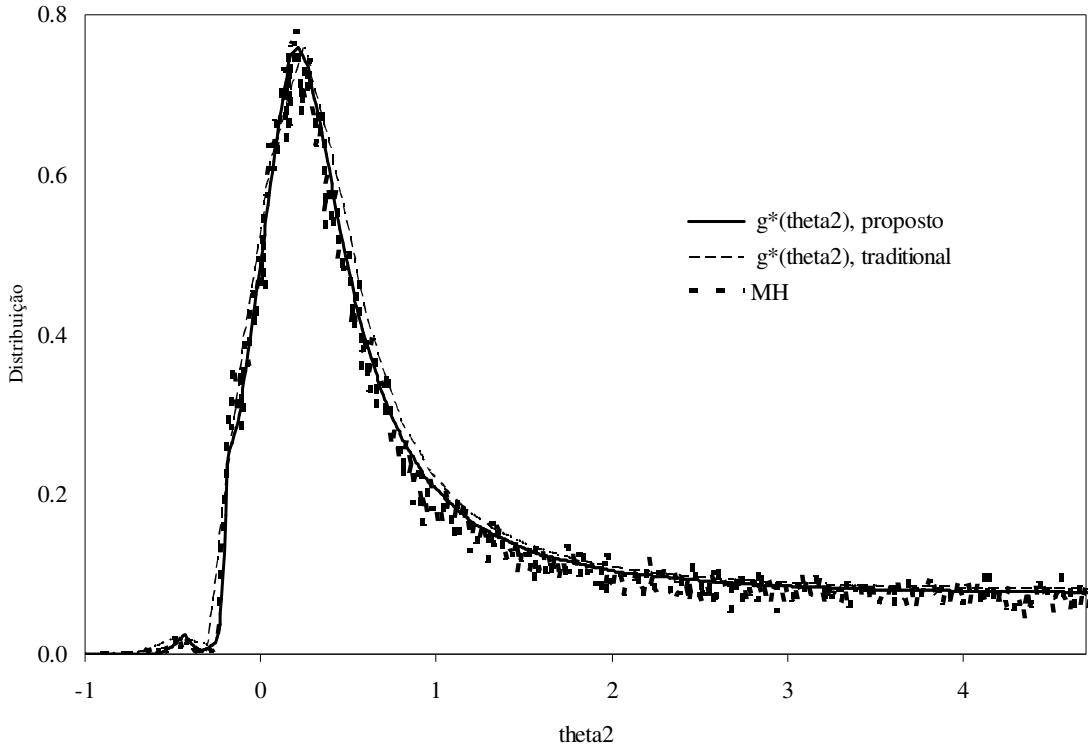


Figura 5.8 Distribuição marginal estimada de θ_2, Caso 5.6, de acordo com 70000 iterações de MH (sob um período de burn-in de 4750 pontos), 2150 iterações do GGS tradicional (com um vetor de 45 pontos e período de burn-in de 1150 pontos), e 2150 execuções do GGS proposto (sob $\tau = 0.05$, $\varepsilon = 5E\text{-}4$, e um período de burn-in de 450 pontos).

Por fim, a análise de variabilidade dos resultados de GGS para θ_2 em trinta triagens, considerando os mesmos parâmetros apresentados anteriormente para ambos os algoritmos, levou a um valor médio amostral para vkl_{ratio} de 1.08 e desvio-padrão igual a 1.16 (valores entre 0.14 e 6.05), o que indica o relativo equilíbrio na acurácia de ambos os métodos, embora que o GGS proposto seja sensivelmente mais eficiente computacionalmente. O intervalo de confiança associado para a razão média (sob um nível de significância de 5%) é [0.67, 1.50].

Caso 5.7 *O MB descrito no Exemplo 2.2 (página 14), estudado por George et al. (1993), é considerado. A Figura 2.3 pode representar a RB associada aos dados descritos na Tabela 5.4. Aqui, L_0 representa a taxa de falhas genérica.*

Neste caso, a comparação de desempenho entre o GGS tradicional e o proposto se deu por meio de um paralelo com os resultados do ©WinBUGS baseado em 70000 iterações e sob um período de *burn-in* de 5000 valores. Quando da aplicação de GGS, todas as variáveis foram limitadas ao intervalo (0, 5]. O GGS tradicional fez uso de 95 pontos igualmente espaçados e 2100 iterações foram executadas (levando a 27300000 avaliações das distribuições-alvo

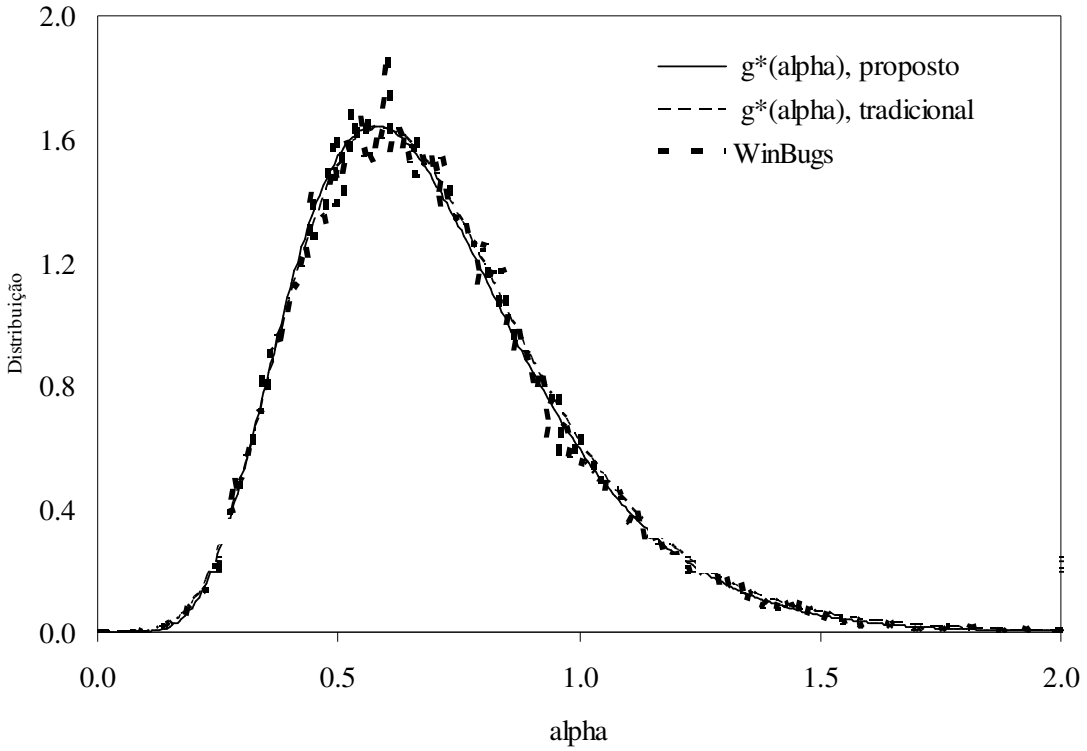
envolvidas) para se alcançar *vkl^* = 0.007 para a distribuição marginal estimada de A. Os períodos de *burn-in* para (A, B, L_0, L_1) foram respectivamente (1100, 1100, 700, 800). O GGS proposto, por outro lado, alcançou tal nível de acurácia com τ = 0.05, ε = 3E-4 e 2100 iterações, conduzindo a uma taxa de rejeição de 2% segundo a probabilidade de MH e 11739000 avaliações das distribuições condicionais envolvidas (com vetores envolvendo uma quantidade de pontos variando de 17 a 91) – apenas 43% do número de avaliações do GGS tradicional. Os períodos de *burn-in* para (A, B, L_0, L_1) foram respectivamente (1100, 1350, 800, 700). De fato, embora o GGS proposto tenha avaliado as distribuições-alvo um número de vezes sensivelmente menor que o GGS tradicional, sua acurácia foi ao menos tão boa quanto a do último para todas as variáveis envolvidas. Este caso de estudo é útil para expor as dificuldades do GGS tradicional em estimar distribuições marginais com curtose elevada, isto é, mais compactas (Figura 5.9-d). Nestes casos, tal método requer um elevado número de avaliações das distribuições-alvo para adequadamente aproximá-las. O tempo demandado pelo GGS tradicional foi de 23.843 segundos enquanto que o GGS proposto consumiu 17.671 segundos e o WinBUGS 5.783 segundos. Vale comentar que o WinBUGS faz uso de propriedades particulares do problema, tais como a conjugação de algumas distribuições subjacentes ao MB, aumentando assim sua eficiência computacional neste caso. A Figura 5.8 exibe os ajustes para as distribuições marginais de A, B, L_0 e L_1. Destaque-se mais uma vez o nível de suavidade derivado do método proposto em relação às distribuições marginais oriundas do WinBUGS, principalmente para as quantidades A e B.

Tabela 5.4 Descrição das distribuições condicionais e dados empíricos relacionados ao Exemplo 2.2 de acordo com George et al. (1993).

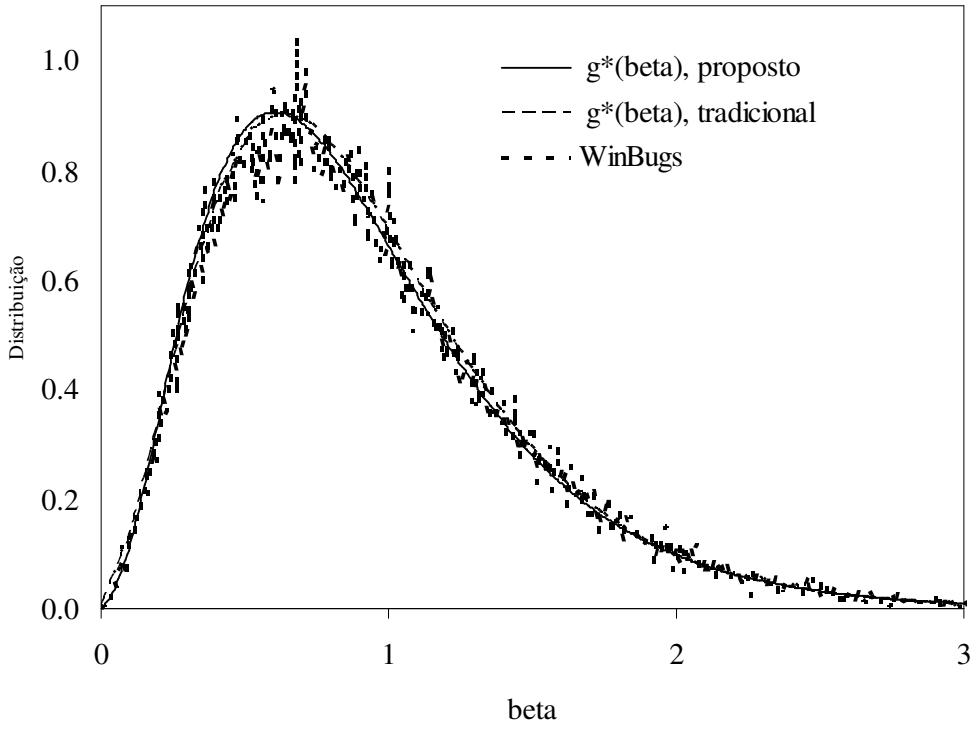
A ~ Exponencial (1.0);
B ~ Gamma(0.1, 1.0)
$[L_0|A=a, B=b] \sim$ Gamma(a, b)
$[L_i|A=a, B=b] \sim$ Gamma(a, b)
$[N_i|L_i=\lambda_i] \sim$ Poisson($t_i \cdot \lambda_i$)
i = 1, 2, ..., 10.

Evidencias:

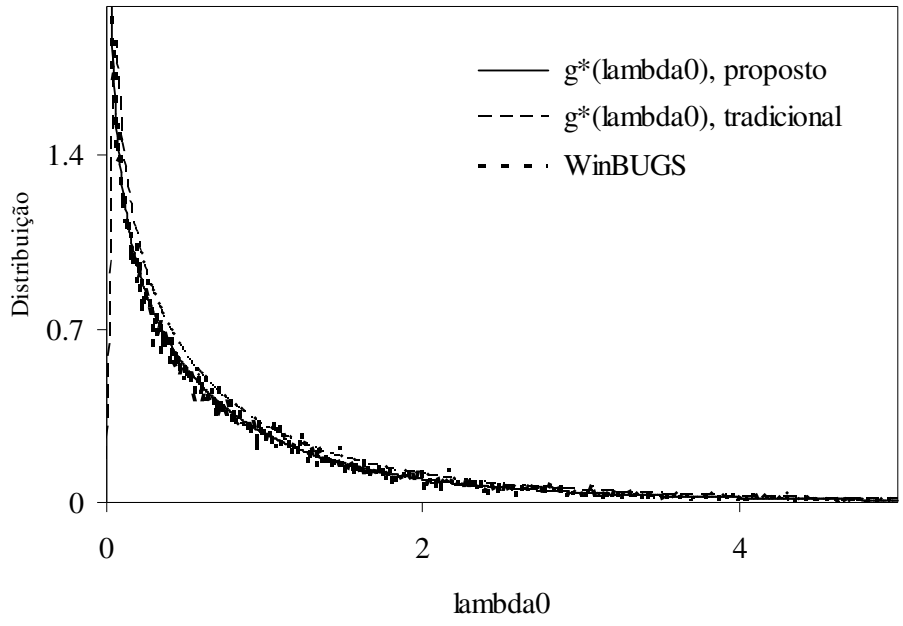
i	1	2	3	4	5	6	7	8	9	10
t_i	94.5	15.7	62.9	126.0	5.24	31.4	1.05	1.05	2.1	10.5
N_i	5	1	5	14	3	19	1	1	4	22



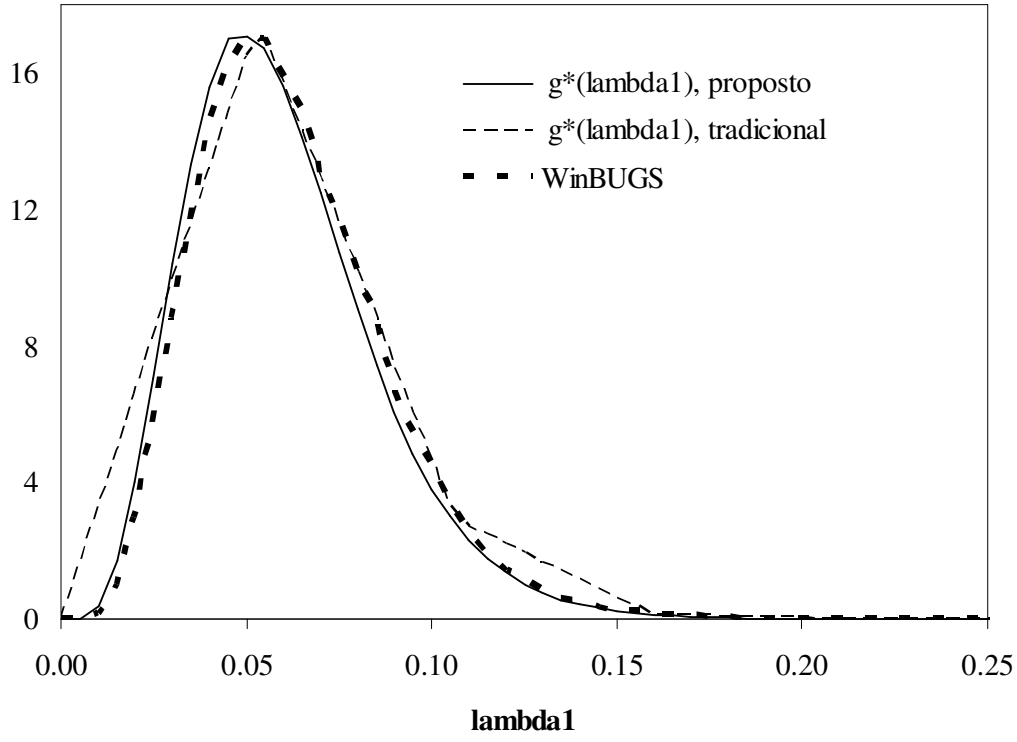
(a) GGS tradicional: *burn-in* de 100 pontos, *vkl*^ = 0.007; GGS proposto: *burn-in* de 1100 pontos, *vkl*^ = 0.007.



(b) GGS tradicional: *burn-in* de 1100 pontos, *vkl*^ = 0.016; GGS proposto: *burn-in* de 1350 pontos, *vkl*^ = 0.013.



(c) GGS tradicional: *burn-in* de 700 pontos, *vkl^* = 0.052; GGS prpposto: *burn-in* de 800 pontos, *vkl^* = 0.015.



(d) GGS tradicional: *burn-in* de 800 pontos, *vkl^* = 0.049; GGS prpposto: *burn-in* de 700 pontos, *vkl^* = 0.012.

Figura 5.9. distribuição estimada de A, B, L_0 e L_1 de acordo com ©WinBUGS baseado em 70000 iterações (sob um período de burn-in de 5000 pontos), 2100 iterações do GGS tradicional (com vetores envolvendo 95 pontos igualmente espaçados) e 2100 execuções do GGS proposto (sob $\tau = 0.05$ e $\varepsilon = 5E\text{-}4$).

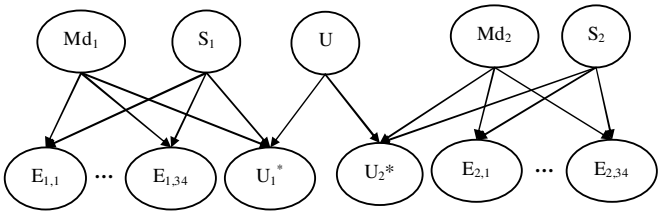
Em relação à análise de variabilidade dos resultados (Equação 5.4), a Tabela 5.5 exibe algumas métricas estatísticas associadas às quantidades A, B e L_0. Vê-se que para as trinta triagens, o desempenho do GGS proposto sempre superou o tradicional em média, embora que para os hiper-parâmetros (A e B) tenha ocorrido de o GGS tradicional se mostrar mais preciso em alguns casos. Por outro lado, a distribuição marginal da taxa de falhas genérica, L_0, foi sempre expressivamente melhor estimada pelo GGS proposto, onde o GGS tradicional apresentou um erro ao menos 7.79 vezes o do proposto. De qualquer maneira, os intervalos de confiança (sob 5% de significância) indicam um desempenho expressivamente melhor do GGS proposto para L_0 e uma acurácia levemente maior para A e B.

Tabela 5.5 Estudo de variabilidade de resultados dos GGS tradicional e proposto (a partir da Equação 5.4) para a taxa de falhas genérica (L_0) e seus hiper-parâmetros, A e B, em relação ao problema estudado por George et al. (1993).

Variável	Parâmetro	*Vkl*
	Média	1.58
	Desvio-padrão	1.36
	Mínimo	0.03
	Máximo	6.48
		1.09
A	IC(5%, média)	2.06
	Média	1.37
	Desvio-padrão	1.17
	Mínimo	0.08
	Máximo	5.68
		0.95
B	IC(5%, média)	1.79
	Média	8.87
	Desvio-padrão	0.79
	Mínimo	7.79
	Máximo	10.54
		8.59
L0	IC(5%, média)	9.16

Caso 5.8 *Um MB semelhante ao descrito no Exemplo 2.4 (página 17), estudado por Droguett & Mosleh (2008), é considerado. A RB relacionada pode ser tal como a da Figura 2.5, mais especificamente esboçada na Figura 5.10. Têm-se em mãos 34 evidências do erros cometidos por dois modelos utilizados para inferir sobre dado parâmetro de interesse. Supõe-se homogeneidade entre as subpopulações geradoras dos erros.*

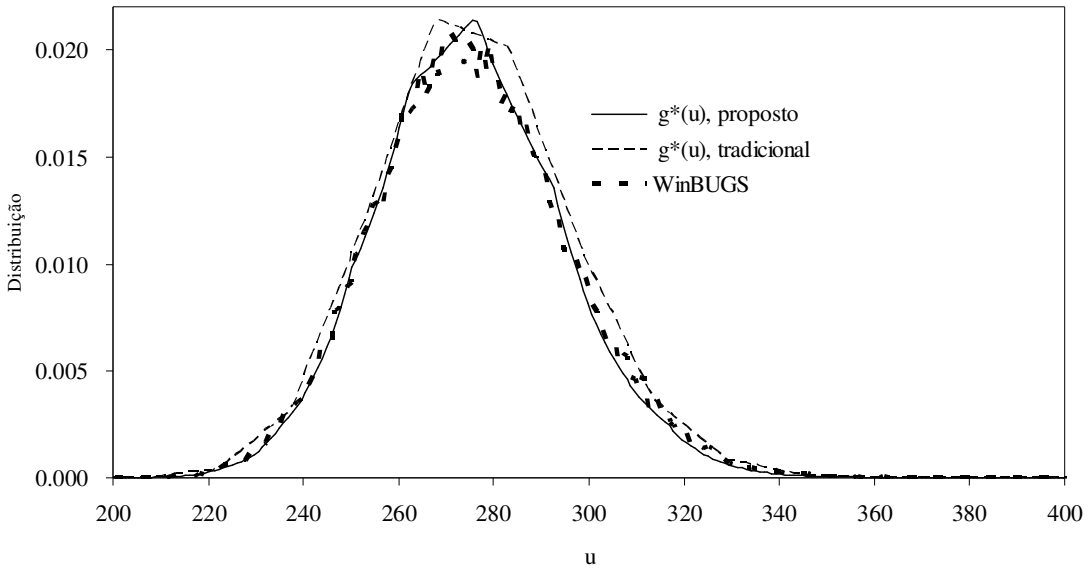
Para o processo de simulação, foi considerado $\varepsilon = 6E\text{-}3$, $\tau = 0.05$ e 2100 iterações do GGS proposto, resultando em 776 rejeições devido à probabilidade de MH (7% dos valores propostos) e 300278 avaliações das distribuições-alvo envolvidas (variando de 15 a 41 pontos para cada uma). Os períodos de *burn-in* para (U, Md_1, S_2) foram respectivamente (400, 700, 1050). De maneira a alcançar uma acurácia similar à do GGS proposto ao menos em relação à variável U, foram executadas 2100 iterações do GGS tradicional com vetores envolvendo 60 pontos, o que resultou em 630000 avaliações das distribuições-alvo (mais que duas vezes a quantidade de avaliações do GGS proposto). Os períodos de *burn-in* para (U, Md_1, S_2) foram respectivamente (750, 1200, 450). Com o intuito de comparar o GGS proposto com o tradicional, o pacote ©WinBUGS foi mais uma vez aplicado sob 100000 iterações e um período de *burn-in* de 30000 iterações. A Figura 5.11 apresenta as estimativas para as distribuições a posteriori de U, Md_1 e S_2. Mais uma vez, o GGS proposto apresentou melhor desempenho, principalmente para as distribuições de alta curtose. O tempo de simulação requerido pelo GGS proposto foi de 14.063 segundos enquanto o GGS tradicional consumiu 31.36 segundos e o WinBUGS 27.354 segundos. Comente-se que neste caso, os resultados promovidos pelo GGS proposto sugerem resultados melhores do que o próprio WinBUGS para Md_1, devido ao comportamento sensivelmente instável deste para tal variável.



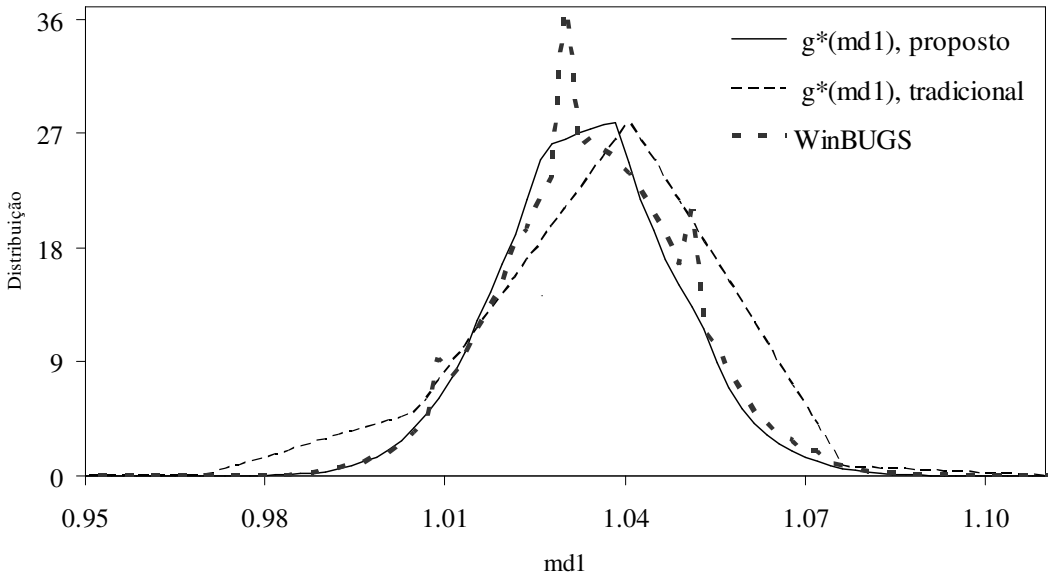
U ~ Uniforme (100, 1000); Md_i ~ Uniforme (0.4, 2.5); S_i~Uniforme (1E-5, 0.5); $[E_{i,j}|Md_i=a, S_i=b]$ ~ Lognormal (log(a), b^2); $[U^*_i|U=u, Md_i=a, S_i=b]$ ~ Lognormal (log(u) + log(a), b^2); i = 1, 2; j=1, …, 34

Evidências:
$U_1^* = 291$; $U_2^* = 269$;
$E_{1j}=e_{1j}=$
(1.20, 1.19, 1.22, 1.02, 1.04, 0.96, 1.01, 1.05, 1.07, 0.99, 1.03, 1.06, 1.05, 1.06, 1.08, 1.05, 1.06, 1.10, 0.99, 1.10, 0.98, 1.02, 0.80, 0.95, 1.01, 1.00, 1.05, 0.92, 0.96, 0.99, 0.90, 1.15, 1.01, 1.14);
$E_{2j}=e_{2j}=$
(1.13, 1.15, 1.10, 1.10, 1.14, 1.03, 1.07, 1.09, 1.12, 1.07, 1.08, 1.21, 1.17, 1.18, 1.13, 1.16, 1.15, 1.26, 0.86, 1.14, 1.17, 0.85, 1.33, 0.89, 1.09, 1.17, 1.21, 1.07, 0.77, 0.81, 0.92, 0.98, 0.96, 0.94).

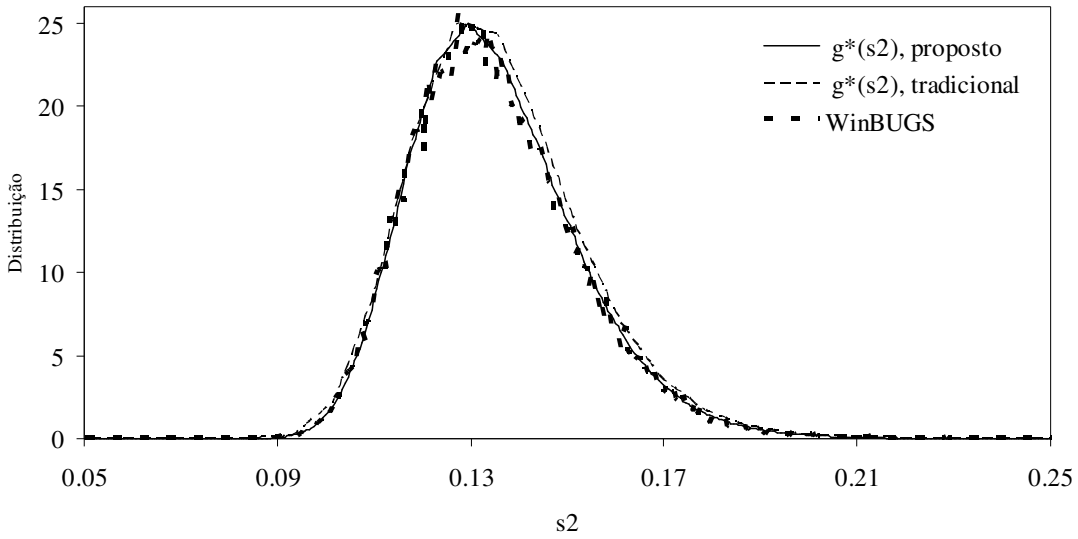
Figura 5.10 RB relacionada a uma instância da metodologia da incerteza de modelos proposta por Droguett & Mosleh (2008) (ver seção 2.6, capítulo 2) envolvendo dois modelos independentes e 34 evidências do erro cometido por cada modelo.



(a) GGS tradicional: *vkl^* = 0.006; GGS proposto: *vkl^* = 0.006.



(b) GGS tradicional: *vkl^* = 0.140; GGS proposto: *vkl^* = 0.033.



(c) GGS tradicional: *vkl^* = 0.007; GGS proposto: *vkl^* = 0.006.

Figura 5.11. Distribuições a posteriori estimadas de U, Md_1, and S_2 de acordo com ©WinBUGS baseado em 100000 iterações (sob um período de burn-in de 30000 pontos), 2100 iterações do GGS tradicional (com vetores de 60 pontos) e 2100 iterações do GGS proposto (sob τ = 0.05, ε = 6E-3).

Em termos de variabilidade dos resultados, considerando a quantidade U, obteve-se vkl_{ratio} = 1.15 em média, com desvio-padrão de 0.15, tendo valores variando entre 0.84 e 1.44, conduzindo a um intervalo de confiança (com nível de significância de 5%) para a razão média [1.09, 1.20]. Isto indica que, além de promover uma maior eficiência computacional, o GGS proposto é sensivelmente melhor que o tradicional em termos de acurácia para este caso.

5.4. Síntese

Nesta subseção, busca-se resumir algumas características envolvendo as simulações realizadas no presente capítulo.

Quanto aos parâmetros de entrada do GGS proposto, um erro relativo em torno de ε = 5E-3 e um fator de aleatoriedade τ = 5E-2, envolvendo 1500 iterações do método de MCMC subjacente, mostrou-se em geral adequado para realização de inferências.

Quanto aos resultados, de maneira geral, pôde-se perceber que a capacidade adaptativa do GGS proposto, aliada a seus fatores de aleatoriedade devido a ASR2(τ) e à probabilidade de rejeição de Metropolis-Hastings, possibilitaram resultados encorajadores a seu favor quando comparados aos do GGS tradicional, baseado em pontos igualmente espaçados e funções lineares por partes. Isto se verificou principalmente para os casos onde as distribuições

marginais de interesse possuem elevada curtose, requerendo do GGS tradicional relativo esforço para realizar ajustes razoáveis. Os casos 5.7 e 5.8 são um exemplo disso.

Em relação a alguns dos demais métodos encontrados na literatura, o GGS proposto apresentou bons resultados. Considerando sua proposta inicial de ser aplicável a MBs genéricos, o caso 5.5 pôde indicar seu melhor desempenho em relação a todos os métodos avaliados, com exceção da extensão *logic sampling*, cuja aplicação é usualmente limitada a MBs que não possuem variáveis evidenciadas. Ressalte-se ainda o bom desempenho do método proposto quando comparado ao próprio WinBUGS, adotado como referência na maioria dos casos. Este último se mostrou por muitas vezes mais dispendioso em termos de tempo computacional sem, no entanto, superar o primeiro em termos de suavidade das estimativas marginais.

Por outro lado, verificaram-se as dificuldades de GGS em lidar com problemas envolvendo funções cuja variação da função derivada se mostra expressiva em pequenas regiões do seu suporte apenas, tais como no caso 5.6. Nestes casos, o GGS tradicional pode requerer um número de pontos demasiadamente grande para adequadamente representar tal variação e o GGS proposto pode simplesmente não captá-la.

6. CONCLUSÕES

Este trabalho foi, de acordo com o conhecimento do autor, uma primeira tentativa de aprimorar o método de MCMC de *Griddy-Gibbs* (GGS). A principal motivação para tanto residiu na possibilidade de se adotar regras de redução de variância, tais como *Rao-Blackwellization*, que reduzem o número de iterações do MCMC necessárias para obtenção de estimativas marginais confiáveis a partir de modelos Bayesianos (MBs). O método resultante mostra-se aplicável para a realização da inferência Bayesiana independente do contexto em questão.

Os primeiros capítulos foram úteis para ilustrar a abrangência de MBs nas diversas áreas do conhecimento, com ênfase especial aos problemas de análise de confiabilidade e riscos. Argumentou-se sobre a importância de formalismos tais como redes Bayesianas (RBs) neste contexto, tanto para a aplicação de MBs existentes quanto para o desenvolvimento de novos modelos. A possibilidade de visualização do modelo introduzida formalmente por RBs, por exemplo, impulsionou não apenas a aplicação da inferência Bayesiana mas também a compreensão de modelos matematicamente complexos por uma maior malha de profissionais. Mencionou-se, também, as dificuldades de obtenção de estimativas marginais que acompanham a liberdade de se modelar mais realisticamente a partir de RBs. Devido ao problema de manipular RBs ser *NP-hard*, a pesquisa sobre métodos direcionados para tal propósito está em constante desenvolvimento.

Neste sentido, os métodos de MCMC foram mais formalmente abordados pelo trabalho. Enfatizou-se que tais métodos podem ser direcionados à manipulação de MBs sem que a natureza do problema ou das variáveis envolvidas seja alterada; prática usualmente adotada por métodos alternativos aos de MCMC. Neste percurso, pôde-se destacar como as re-interpretações ou mesmo leves alterações em algoritmos existentes promoveram os avanços mais significativos dos métodos de MCMC. Tal etapa conclui-se com a apresentação do método de Gibbs (GS) que possibilita o uso de *Rao-Blackwellization* e de GGS, que busca aproximar GS quando este não pode ser diretamente aplicado ao MB.

Nos últimos capítulos, estudos acerca de novas variantes de GGS foram elaborados. Deficiências conceituais e de aproximação propagadas desde sua introdução em 1992 foram discutidas, dando suporte para a introdução de algoritmos alternativos. Neste sentido, métodos numéricos adaptativos e técnicas de agrupamento probabilísticas foram estudadas e adaptadas a fim de melhorar o desempenho de GGS tanto em termos de acurácia quanto de eficiência computacional. Além disso, defendeu-se aqui a tese de que GGS deve ser interpretado como

uma extensão de MH (cuja probabilidade de rejeição pode ser não-nula) ao invés de uma derivação de GS. Caso contrário, argumentou-se que os resultados de GGS podem ser questionáveis, já que a presença de uma probabilidade de rejeição é uma característica fundamental dos métodos de MCMC que recorrem a funções alternativas àquelas cuja amostragem direta é inviável. De maneira a suprir tal deficiência, conceitos de métodos de MCMC consolidados foram também incorporados a GGS. A soma de todas estas características incluídas em GGS o promoveu à condição de único método de aceitação-rejeição adaptativo que permite *Rao-Blackwellization*. Resultados baseados em alguns casos de estudos extraídos da literatura, que datam desde a década passada até o presente ano, indicam o seu grande potencial para a promoção de inferências Bayesianas mais acuradas e com menor consumo computacional, principalmente em se tratando de distribuições com elevada curtose (mais compactas). O GGS tradicional requer um número elevado de avaliações da função de interesse nestes casos. Ressaltou-se, ainda, o desempenho do GGS proposto quando comparado aos métodos embutidos no pacote WinBUGS ©. Em algumas situações, o primeiro alcançou resultados mais confiáveis, mesmo recorrendo a menos recursos computacionais. Destaque-se neste contexto os casos nos quais a conjugação de distribuições é menos freqüente.

6.1. Limitações do Trabalho

A princípio inevitavelmente, o infortúnio de requerer uma cadeia de Markov subjacente ao menos irredutível é uma das limitações do algoritmo proposto (característica dos métodos de MCMC). Um outro ponto passível de críticas consiste da necessidade de se ter intervalos que delimitem o suporte das funções estudadas. Embora que por muitas vezes seja possível atribuir restrições neste sentido, tal tarefa pode se tornar bastante árdua ou mesmo inviável a depender do MB. Por fim, as dificuldades para se determinar parâmetros cruciais tanto do GGS proposto (tais como o nível de tolerância atribuído pelo usuário e o nível de aleatoriedade do método adaptativo) quanto da própria metodologia de MCMC (número de iterações e período de *burn-in*) podem também ser destacadas como fontes de questionamentos.

6.2. Trabalhos Futuros

Trabalhos futuros devem ser direcionados ao estudo de métodos numéricos adaptativos alternativos ao adotado no presente trabalho e à adoção de aproximações polinomiais por partes, ao invés das lineares atualmente consideradas. Em complemento, algoritmos para a

definição de valores que delimitem o suporte das funções estudadas devem também ser desenvolvidos. Ainda neste sentido, uma comparação mais exaustiva em termos de acurácia e consumo computacional, envolvendo demais alternativas de manipulação de modelos Bayesianos, mostra-se também pertinente. Tal tarefa torna-se custosa devido à necessidade de implementação dos métodos estudados para a realização de comparações imparciais.

Por outro lado, em uma perspectiva mais genérica, o aprimoramento de MBs existentes será trabalhado. Cite-se como exemplo, a modelagem de parâmetros de confiabilidade em função do tempo subjacentes a processos não-homogêneos de Poisson ou Gamma em detrimento do processo homogêneo de Poisson considerado no formalismo de análise de variabilidade populacional (seção 2.5) ou, em outra vertente, a modelagem de dependências entre modelos na metodologia de incerteza de modelos (seção 2.6).

REFERÊNCIAS BIBLIOGRÁFICAS

Andrieu, C. e Thoms, J. A tutorial on adaptive MCMC. *Stat Comput*, v.18, p.343–373. 2008.

Ardia, D., Hoogerheide, L. F. e Dijk, H. K. V. Adaptive Mixture of Student-t Distributions as a Flexible Candidate Distribution for Efficient Simulation: The R Package AdMit. *Journal of Statistical Software*, v.29, n.3, p.1-32. 2009.

Ausín, M. C. e Galeano, P. Bayesian estimation of the Gaussian mixture GARCH model. *Computational Statistics & Data Analysis*, v.51, n.5, p.2636 – 2652. 2007.

Bauwens, L. e Lubrano, M. Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics Journal*, v.1, n.1, p.C23–C46. 1998.

Bernardo, J. M. e Smith, A. F. *Bayesian theory*. Chichester: John Wiley & Sons Ltd. 1995.

Bolstad, W. M. *Introduction to Bayesian statistics*. New Jersey: John Wiley & Sons. 1943.

Brewer, M. J., Aitken, C. G. G. e Talbot, M. A comparison of hybrid strategies for Gibbs sampling in mixed graphical models. *Computational Statistics & Data Analysis*, v.21, n.3, p.343-365. 1996.

Cai, B., Meyer, R. e Perron, F. Metropolis Hastings algorithms with adaptive proposals. *Stat Comput*, v.18, n.421–433. 2008.

Casella, G. e Berger, R. L. *Statistical inference*. California: Duxbury. 2001.

Castillo, E., Gutiérrez, J. M., Hadi, A. S. e Solares, C. Symbolic propagation and sensitivity analysis in Gaussian Bayesian networks with application to damage assessment. *Artificial Intelligence in Engineering*, v.11, p.173-181. 1997.

Castillo, E. e Kjærulff, U. Sensitivity analysis in Gaussian Bayesian networks using a symbolic-numerical technique. *Reliability Engineering and System Safety*, v.79, p.139–148. 2003.

Clark, J. S. e Gelfand, A. E. *Hierarchical modelling for the environmental sciences: statistical methods and applications*. Oxford: Oxford University Press. 2006.

Congdon, P. *Applied Bayesian Modelling*. Chichester: John Wiley & Sons, Ltd. 2003.

Congdon, P. *Bayesian Models for Categorical Data*. Chichester: John Wiley & Sons, Ltd. 2005.

Congdon, P. *Bayesian Statistical Modelling*. Chichester: John Wiley & Sons. 2006.

Cooper, G. F. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, v.42, n.2-3, p.393-405. 1990.

Cryer, J. D. *Time series analysis*. Boston: PWS Publishers. 1986.

Dagum, P. e Luby, M. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, v.60, n.1, p.141-153. 1993.

Devroye, L. *Non-uniform random variate generation*. New York: Springer-Verlag. 1986.

Diéz, F. J. Local conditioning in Bayesian networks. *Artificial Intelligence*, v.87, p.1-20. 1996.

Droguett, E. L., Groen, F. e Moslehb, A. The combined use of data and expert estimates in population variability analysis. *Reliability Engineering and System Safety*, v.83, p.311–321. 2004.

Droguett, E. L. e Mosleh, A. Bayesian Methodology for Model Uncertainty Using Model Performance Data. *Risk Analysis*, v.28, n.5, p.1457-1476. 2008.

Du, Q. e Gunzburger, M. Grid generation and optimization based on centroidal Voronoi tessellations. *Applied Mathematics and Computation*, v.133, n.2-3, p.591–607. 2002.

Edwards, D. *Introduction to Graphical Modelling*. New York: Springer. 1949.

Firmino, P. R. A. e Droguett, E. L. *A numerical procedure for handling mixed Bayesian networks*. *In*: European Safety and Reliability Annual Conference (ESREL). Prague, 2009.

Firmino, P. R. A., Filho, R. L. M. S. e Droguett, E. L. *An Expert Opinion Elicitation Method Based on Bayesian Intervals Estimation and Computational Searching Algorithms: an Application to Oil Refinery Risk Analysis. In*: International Probabilistic Safety Assessment and Management Conference, PSAM 9. Hong Kong, 2008.

Firmino, P. R. A., Menêzes, R. D. C. S., Droguett, E. L. e Duarte, D. C. D. L. *Eliciting Engineering Judgments in Human Reliability Assessment*. *In*: 52nd Annual Reliability & Maintainability Symposium. Newport Beach: IEEE, 2006.

Firmino, P. R. A., Moura, M. D. C., Pontual, A. D. O., Lins, I. D. e Droguett, E. L. Política ótima de manutenção preventiva de sistemas reparáveis baseada em confiabilidade. *XX Congresso Panamericano de Engenharia Naval Transporte Marítimo e Engenharia Portuária (COPINAVAL)*. São Pau 2007.

Gelfand, A. E. e Smith, A. F. M. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, v.85, n.410, p.398-409. 1990.

Geman, S. e Geman, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.* , v.6, p.721-741. 1984.

George, E., Makov, U. e Smith, A. Conjugate likelihood distributions. *Scandinavian J. Statist.*, v.20, n.2, p.147–156. 1993.

Gerlach, R. e Chen, C. S. Bayesian inference and model comparison for asymmetric smooth transition heteroskedastic models. *Stat Comput*, v.18, p.391–408. 2008.

Geyer, C. J. Practical Markov chain Monte Carlo. *Statictic Science*, v.7, p.473-511. 1992.

Gilks, W. R., Richardson, S. e Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall/CRC. 1996.

Hamada, M. S., Wilson, A. G., Reese, C. S. e Martz, H. F. *Bayesian Reliability*. New York: Springer. 2008.

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, v.57, n.97-109. 1970.

Heckerman, D., Mamdani, A. e Wellman, M. P. Realworld applications of Bayesian networks. *Communications of the ACM* v.38, n.3, p.24-68. 1995.

Henrion, M. *Propagating uncertainty in Bayesian networks by probabilistic logic sampling*. *In*: Uncertainty in Artificial Intelligence 2. North-Holland, 1988.

Hrycej, T. Gibbs Sampling in Bayesian Networks. *Artificial Intelligence*, v.46, p.351-363. 1990.

Huang, C. e Darwiche, A. Inference in belief network: A procedural guide. *International Journal of Approximate Reasoning*, v.11, n.1, p.1-158. 1994.

Johansen, A. M., Doucet, A. e Davy, M. Particle methods for maximum likelihood estimation in latent variable models. *Stat Comput*, v.18, p.47–57. 2008.

Keith, J. M., Kroese, D. P. e Sofronov, G. Y. Adaptive independence samplers. *Statistics & Computation*, v.18, n.4, p.409–420. 2008.

Kelly, D. L. e Smith, C. L. Bayesian inference in probabilistic risk assessment - The current state of the art. *Reliability Engineering and System Safety*, v.94, p.628-643. 2009.

Kennedy, J. e Eberhart, R. C. *Swarm intelligence*. San Diego: Morgan Kaufmann. 2001.

Koller, D., Lerner, U. e Angelov, D. *A General Algorithm for Approximate Inference and Its Application to Hybrid Bayes Nets*. *In*: 15th Annual Conference on Uncertainty in Artificial Intelligence. Stockholm, 1999.

Korb, K. B. e Nicholson, A. E. *Bayesian artificial intelligence*. Florida: Chapman & Hall/CRC. 2003.

Kozlov, A. V. e Koller, D. *Nonuniform dynamic discretization in hybrid networks*. *In*: 13th Conference on Uncertainty in Artificial Intelligence. Rhode Island, 1997.

Ladeira, M., Vicari, R. M. e Coelho, H. *Redes Bayesianas Multiagentes*. *In*: Encontro Nacional de Inteligência Artificial. Rio de Janeiro, 1999.

Langseth, H., Nielsen, T. D., Rumí, R. e Salmerón, A. Inference in hybrid Bayesian networks. *Reliability Engineering and System Safety*, v. , p.doi:10.1016/j.ress.2009.02.027. 2009.

Malcolm, M. A. e Simpson, R. B. Local Versus Global Strategies for Adaptive Quadrature. *ACM Transactions on Mathematical Software*, v.1, n.2, p.129-146. 1975.

Martz, H. F. e Waller, R. A. *Bayesian reliability analysis*. Florida: Krieger Publishing Company. 1982.

Mckeeman, W. M. Adaptive numerical integration by Simpson's rule, Algorithm 145. *Communications of the ACM*, v.5, n.12, p.604. 1962.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. e Teller, E. Equations of state calculations by fast computing machine. *Chem. Phys.*, v.21, n.1087-1091. 1953.

Meyer, R., Cai, B. e Perron, F. Adaptive rejection Metropolis sampling using Lagrange interpolation polynomials of degree 2. *Computational Statistics and Data Analysis*, v.52, n.7, p.3408–3423. 2008.

Michalewicz, Z. *Genetic algorithms + data structures = evolution programs*. Berlin: Springer. 1999.

Neal, R. M. Slice sampling. *The Annals of Statistics*, v.31, n.3, p.705–767. 2003.

Neapolitan, R. E. *Learning Bayesian Networks*. New Jersey: Pearson Prentice Hall. 2004.

Neil, M., Tailor, M. e Marquez, D. Inference in hybrid Bayesian networks using dynamic discretization. *Statistics & Computation*, v.17, n.3, p.219–233. 2007.

Neil, M., Tailor, M., Marquez, D., Fenton, N. e Hearty, P. Modelling dependable systems using hybrid Bayesian networks. *Reliability Engineering and System Safety*, v.93, p.933–939. 2008.

Pearl, J. Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence*, v.29, p.241-288. 1986.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. California: Morgan Kaufmann. 1988.

Pearl, J. *Causality, Reasoning, and Inference*. New York: Cambridge University Press. 2000.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. e Flannery, B. P. *Numerical recipes in C: the art of scientific computing*. Cambridge: Cambridge University Press. 1992.

Ritter, C. e Tanner, M. A. Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, v.87, n.419, p.861-868. 1992.

Robert, C. P. e Casella, G. *Monte Carlo statistical methods*. New York: Springer. 2004.

Ross, S. M. *Introduction to Probability Models*. Florida: Harcourt Academic Press. 2000.

Ross, S. M. *Simulation*. San Diego: Academic Press. 2002.

Rumí, R. e Salmerón, A. Approximate probability propagation with mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, v.45, p.191–210. 2007.

Rumí, R., Salmerón, A. e Moral, S. Estimating Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. *Test*, v.15, n.2, p.397–421. 2006.

Souza, F. M. C. D. *Sistemas Probabilísticos*. Recife: Talus. 2002.

Suermondt, H. J. e Cooper, G. F. Initialization for the method of conditioning in Bayesian belief networks. *Artificial Intellingence*, v.50, p.83-94. 1991.

Tierney, L. *Exploring posterior distributions using Markov chains. In*: Computer Science and Statistics: Proc. 23rd Symp. Interface Fairfax Station, 1991.

Wei, S. X. A censored–GARCH model of asset returns with price limits. *Journal of Empirical Finance*, v.9, n.2, p.197–223. 2002.

York, J. Use of the Gibbs Sampler in expert systems. *Artificial Intelligence*, v.56, p.115-130. 1992.

Zio, E. e Pedroni, N. Estimation of the functional failure probability of a thermal–hydraulic passive system by Subset Simulation. *Nuclear Engineering and Design*, v.239, p.580-599. 2009.