Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Análise de Desempenho em Filas $M/M/1$ Usando uma Abordagem Bayesiana

Márcio A. C. Almeida[1], Frederico R. B. Cruz[2]

Departamento de Estatística, ICEx, UFMG, Belo Horizonte, MG

Resumo. Em teoria de filas, um dos principais interesses dos pesquisadores é estudar o seu comportamento, seu processo de formação e analisar algumas características de desempenho, tais como, por exemplo, a intensidade do tráfego, definida como sendo a razão entre a taxa de chegada e a taxa de atendimento. No entanto, sabe-se que na maioria das vezes os parâmetros envolvidos no processo não são conhecidos e precisam ser estimados através de algum método matemático. Este artigo visa obter estimativas para a intensidade de tráfego em filas markovianas infinitas com um único servidor, denominadas $M/M/1$, na notação de Kendall. Uma análise sobre o enfoque bayesiano é realizada, obtendo-se distribuições a posteriori e preditivas para parâmetros de interesse. Amostras foram obtidas através de simulação e algumas características de desempenho foram analisadas. Verificou-se também o fator de Bayes como critério de seleção de modelo.

Palavras-chave. Filas markovianas, inferência Bayesiana, distribuição a posteriori, características de desempenho.

1 Introdução

A teoria de filas, um ramo da matemática que analisa o comportamento das filas, apesar de ser muito tradicional, ainda atrai interesse dos pesquisadores. Isso porque elas já fazem parte do nosso dia-a-dia. No supermercado, no banco, nos centros lotéricos, nos postos de gasolina, enfim, em qualquer lugar podemos nos deparar com esse sistema. Sabe-se que as filas sempre ocorrem quando a procura por um determinado serviço é maior que a capacidade do sistema de prover este serviço. Um dado sistema de filas pode ser descrito por seis componentes [13], quais sejam: (i) o processo de chegada, (ii) a distribuição do tempo de serviço, (iii) o número de servidores, (iv) a capacidade do sistema, (v) a população de usuários e (vi) a disciplina de atendimento. Em geral os sistemas de filas têm diferentes características, mas suas formas de funcionamento são similares. O grande interesse é otimizar os sistemas de filas, de forma a reduzir seus custos operacionais e melhorar seu desempenho. Sendo assim, o conhecimento de algumas das suas características é importante, tais como a taxa de chegada λ, o tempo que um servidor leva para executar um serviço μ, a intensidade do tráfego ρ, definida como sendo a razão

[1]estatimarcio@gmail.com
[2]fcruz@est.ufmg.br

DOI: 10.5540/03.2015.003.02.0057

entre λ e μ, a esperança do número de clientes no sistema L e o tamanho da fila L_q, entre outras (embora neste artigo seja dada preferência para a denominação *intensidade de tráfego*, note-se que às vezes o ρ é chamado de *fator de utilização da estação de serviço*). Todas esse medidas podem ser quantificadas e avaliadas através de funções paramétricas em modelagem de filas.

A literatura apresenta uma grande quantidades de artigos que estudam diversas características, tanto por métodos inferenciais clássicos, quanto por métodos bayesianos. Entre os métodos clássicos, pode-se citar por exemplo o trabalho de [3], que apresenta estimadores de máxima verossimilhança e aproximações para estimar a taxa de chegada e o tempo de serviço. Em [12] é mostrado que estimar taxas de chegada e taxas de serviços resulta em uma diferença notável entre a distribuição do estado do modelo (parâmetros estimados) e a distribuição de estado para o sistema real (os parâmetros conhecidos). Em [7] é descrita uma aplicação de filas $M/M/1$ e apresentadas estimativas utilizando máxima verossimilhança para filas formadas durante dois períodos, de demanda normal e de férias.

Sob o enfoque bayesiano, um dos primeiros trabalhos foi desenvolvido por [9], que estendeu a metodologia desenvolvida por [3] atribuindo distribuições *a priori* para λ e μ. Em [1] são estudadas distribuições preditivas de medidas de desempenho usuais em filas $M/M/1$. Em [2] foram apresentados momentos da distribuição *a posteriori* e intervalos de credibilidade para ρ.

O objetivo deste trabalho é apresentar estimativas para algumas medidas de desempenho de um sistema de filas $M/M/1$ sob a abordagem bayesiana. São também apresentados resultados computacionais de um critério de seleção de modelos através do fator de Bayes.

2 Metodologia

2.1 Filas $M/M/1$

Um modelo de filas *markovianas*, também conhecidas como filas $M/M/1$, baseia-se nas características dos processos de chegada e de serviço (atendimento) assumidos Poisson e Exponencial, respectivamente. Assim, assume-se que o número de chegadas na unidade de tempo segue uma distribuição de Poisson com taxa λ e o tempo de serviço segue uma distribuição exponencial com taxa μ. Assumindo-se também que o sistema de filas atinge um regime estacionário, isto é, foi observado após um prolongado período de funcionamento (diz-se, também, que o sistema está em equilíbrio), pode-se calcular a intensidade de tráfego ρ. Com relação a ρ, note-se a importância de que se verifique a relação $\rho < 1$ (ou, equivalentemente, $\lambda < \mu$), para garantir-se que o sistema atinja o estado estacionário [13]. Caso contrário a fila só aumentaria. Associada a esta intensidade de tráfego, pode-se encontrar uma distribuição de probabilidade geométrica para o número de clientes no sistema no momento da partida, que, após o equilíbrio, pode ser escrita como [8]:

$$P(M = m) = \begin{cases} \rho^m(1-\rho), & m = 0, 1, 2, \ldots \\ 0, & \text{caso contrário.} \end{cases} \quad (1)$$

A partir da Eq. (1), outras características de desempenho podem também ser encontradas em função de ρ, tais como a probabilidade de servidor ocioso, esperança do número

DOI: 10.5540/03.2015.003.02.0057

de clientes no sistema e tamanho da fila, dados por $P(M = 0) = 1 - \rho$, $L = \rho/(1 - \rho)$, e $L_q = \rho^2/(1 - \rho)$, respectivamente.

2.2 Inferência Bayesiana

Do ponto de vista da inferência bayesiana [4, 10], a informação *a priori* sobre um parâmetro desconhecido pode ser quantificada através de uma distribuição de probabilidade, definida num espaço paramétrico Θ, chamada de distribuição *a priori* $p(\theta)$. Observa-se então uma amostra aleatória da variável de interesse, denominada *função de verossimilhança*, representada por $L(\theta, X)$. Assim, através do Teorema de Bayes, a informação sobre θ é atualizada, obtendo-se então a distribuição *a posteriori* $p(\theta|x)$, dada por:

$$p(\theta|x) = \frac{L(\theta, x)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} \propto L(\theta, x)p(\theta), \tag{2}$$

em que $1/p(x)$ é uma constante normalizadora de $p(\theta|x)$. Assim, toda inferência sobre θ é realizada a partir da distribuição *a posteriori*.

Considerando-se filas $M/M/1$, tem-se que a distribuição do número de clientes no sistema no momento de partida é dada pela Eq. (1). Toma-se então uma amostra aleatória de tamanho n do número de clientes no sistema no momento da partida $X_1, X_2, \ldots, X_n$. Por conseguinte, pode-se escrever a seguinte função de verossimilhança [2]:

$$L(\rho, X) = (1 - \rho)^n \rho^{\Sigma_{i=1}^n x_i}. \tag{3}$$

Este modo de geração de dados garante a independência das observações amostrais, levando em conta a propriedade de ergodicidade da cadeia de Markov. Com isso, se X_m é o número de clientes no sistema no momento da partida do m-ésimo cliente e $p_{ij}^{(k)} = P(X_{m+k} = j|X_m = i)$, a propriedade da ergodicidade garante os limites das probabilidades dadas por $v_j = \lim_{k\to\infty} p_{ij}^{(k)}, j = 0, 1, \ldots$

Para se fazer inferência sobre ρ, partindo da Eq. (3), dois tipos de informações *a priori* podem ser consideradas, em termos de distribuição: uma, denominada *priori conjugada natural*, e outra, *priori não-informativa* [10]. A primeira opção garante que distribuições *a priori* e *a posteriori* pertençam à mesma família paramétrica e a segunda determina que não existe uma informação palpável *a priori* a respeito de ρ. No primeiro caso, a distribuição beta com parâmetros a e b pode ser utilizada, uma vez que ela é flexível, podendo assumir diversas formas, além de ser uma conjugada natural a partir da Eq. (3). Como também temos que ρ pertence ao intervalo $(0, 1)$, uma distribuição uniforme poderia ser utilizada. No entanto, em muitos sistemas, os gerentes operacionais podem oferecer limites mais justos na intensidade do tráfego [2], de tal forma que $0 < c < \rho < d < 1$. Ou seja, apesar de o ρ poder estar entre 0 e 1, sabe-se que são incomuns na prática valores tais que $\rho < c$ ou $\rho > d$. Logo, pode-se assumir também uma distribuição uniforme truncada para ρ. Em outras palavras, tem-se

DOI: 10.5540/03.2015.003.02.0057

$$p_1(\rho) \propto \rho^{(a-1)}(1-\rho)^{(b-1)},\ 0<\rho<1,\ a>0,\ b>0, \tag{4}$$

e também

$$p_2(\rho) \propto 1,\ c<\rho<d. \tag{5}$$

Assim, combinando-se as informações da Eq. (3) com as Equações (4) e (5), a partir de Eq. (2), pode ser mostrado que as distribuições a posteriori para ρ são [2]:

$$p_1(\rho|X) = \begin{cases} \frac{1}{\mathcal{B}(a+y,n+b)}\rho^{a+y-1}(1-\rho)^{n+b-1}, & 0<\rho<1, \\ 0, & \text{caso contrário}, \end{cases} \tag{6}$$

$$p_2(\rho|X) = \begin{cases} \frac{1}{\mathcal{B}(c,d,y+1,n+1)}\rho^{y}(1-\rho)^{n}, & c<\rho<d, \\ 0, & \text{caso contrário}, \end{cases} \tag{7}$$

em que $y = \Sigma_{i=1}^{n} x_i$, sendo a primeira a distribuição beta, com parâmetros $a+y$ e $n+b$, e a segunda, a distribuição beta incompleta generalizada, com parâmetros c, d, $y+1$ e $n+1$.

A partir destas distribuições, todas as estimativas podem ser encontradas. De fato, assumindo-se a função de perda quadrática para o erro e tomando-se as médias das distribuições das Equações (6) e (7), é possível deduzir os seguintes estimadores de Bayes [2], para a intensidade de tráfego $\hat{\rho}$, número de clientes no sistema $\hat{L}$ e tamanho da fila $\hat{L}_q$:

$$\hat{\rho}_1 = \tfrac{a+y}{y+a+n+b},\quad \hat{L}_1 = \tfrac{y+a}{n+b},\quad \hat{L}_{q1} = \tfrac{(y+a+1)(y+a)}{(n+b-1)(y+a+n+b)}, \tag{8}$$

$$\hat{\rho}_2 = \tfrac{\mathcal{B}(c,d,y+2,n+1)}{\mathcal{B}(c,d,y+1,n+1)},\quad \hat{L}_2 = \tfrac{\mathcal{B}(c,d,y+2,n)}{\mathcal{B}(c,d,y+1,n+1)},\quad \hat{L}_{q2} = \tfrac{\mathcal{B}(c,d,y+3,n)}{\mathcal{B}(c,d,y+1,n+1)}. \tag{9}$$

Se após observar $X = x$ estivermos interessados na previsão de uma quantidade M, também relacionada com ρ, descrita probabilisticamente por $p(M|\rho)$, então define-se a esperança preditiva a posteriori $E(M|X)$ [4]. Pode-se então analisar as distribuições preditivas para o número de clientes no sistema no momento da partida, em estado estacionário, dadas por:

$$E_1(M|X) = \frac{(r!)\mathcal{B}(y+a+r, n+b-r)}{\mathcal{B}(y+a, n+b)},\ \text{para } r = 1, 2, \ldots \tag{10}$$

$$E_2(M|X) = \frac{(r!)\mathcal{B}(c, d, y+r+1, n-r+1)}{\mathcal{B}(y+1, n+1)},\ \text{para } r = 1, 2, \ldots \tag{11}$$

As densidades preditivas desempenham um papel importante na seleção do modelo bayesiano e na sua comparação [5]. As chances *a posteriori* do modelo M_j, relativo a um modelo M_k, são definidas como $\Pr(M_j = m|X)/\Pr(M_k = m|X)$, o qual é o produto das chances *a priori* $\Pr(M_j)/\Pr(M_k)$ do modelo M_j, relativo ao modelo M_k, chamado de fator de Bayes e dado por:

$$B_{jk} = \frac{f(M|M_j)}{f(M|M_k)} \Rightarrow B_{12} = \frac{p_1(M = m|X)}{p_2(M = m|X)},\ m = 0, 1, 2, \ldots \tag{12}$$

Para decidir entre os modelos j e k, é recomendada a utilização da seguinte regra [6]: considera-se que há provas decisivas, forte ou substancial contra k quando B_{jk} é superior a 100, está entre 10 e 100 ou está entre 3 e 10, respectivamente.

3 Resultados

Para avaliar a metodologia descrita, foram simulados no *software* estatístico R [11] dados do número de clientes no sistema no momento da partida, a partir da Eq. (3), utilizando amostras de tamanho 10, 100 e 200, para $\rho = 0,2$, 0,5 e 0,9. O procedimento foi realizado 5.000 vezes. Dois tipos de distribuições *a priori* foram utilizadas, uma distribuição *a priori* beta, com parâmetros 1,5 e 2,5, e uma distribuição uniforme truncada, com parâmetros 0,05 e 0,95, conforme mostrado na Figura 1.

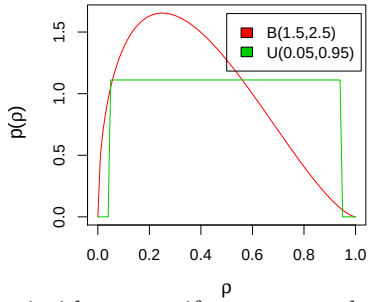

Figura 1: Distribuição a priori beta e uniforme truncada para a intensidade de tráfego.

A Tabela 1 mostra as estimativas para cada tamanho de amostra e valor de ρ. Observa-se que com o aumento do tamanho amostral, as estimativas se aproximam do verdadeiro valor de ρ utilizado, onde as estimativas utilizando a distribuição uniforme truncada parecem estar mais próximas.

Tabela 1: Estimativas da intensidade de tráfego com respectivo erro padrão entre parênteses.

		n		
ρ	Distribuição	10	100	200
0,2	Beta	0,2318 (0,0011)	0,2051 (0,0005)	0,2021 (0,0004)
	Uniforme	0,2380 (0,0011)	0,2043 (0,0005)	0,2016 (0,0004)
0,5	Beta	0,4623 (0,0014)	0,4944 (0,0005)	0,4973 (0,0004)
	Uniforme	0,4806 (0,0015)	0,4968 (0,0005)	0,4986 (0,0004)
0,9	Beta	0,8690 (0,0006)	0,8969 (0,0001)	0,8985 (0,0001)
	Uniforme	0,8806 (0,0005)	0,8982 (0,0001)	0,8991 (0,0001)

A Tabela 2 mostra algumas características de operação da fila. Observa-se que com o aumento do tamanho amostral as estimativas se aproximam dos valores exatos para o número de clientes no sistema e o tamanho da fila. As estimativas baseadas na distribuição *a priori* uniforme truncada parece ter um desempenho um pouco melhor. Para determinar qual modelo melhor se ajustou aos dados (simulados), fixou-se um tamanho amostral de

DOI: 10.5540/03.2015.003.02.0057

200 e obteve-se a distribuição preditiva do número de clientes no sistema no momento de partida, para o número de clientes de 0 a 5, e calculou-se o fator de Bayes para cada um deles. A partir da Tabela 3 observa-se que o fator de Bayes ficou próximo de 1 para todos os valores de ρ, não evidenciando para este parâmetro que algum dos dois modelos seja superior ao outro.

Tabela 2: Valores exatos e estimativas para o número de clientes no sistema e tamanho da fila.

				n					
				10		100		200	
ρ	L	L_q	Distribuição	$\hat{L}$	$\hat{L_q}$	$\hat{L}$	$\hat{L_q}$	$\hat{L}$	$\hat{L_q}$
0,2	0,25	0,05	Beta	0,3161	0,1117	0,2602	0,0578	0,2544	0,0537
			Uniforme	0,3584	0,1204	0,2618	0,0574	0,2552	0,0535
0,5	1,00	0,50	Beta	0,9263	0,5445	0,9874	0,5027	0,9944	0,5020
			Uniforme	1,1081	0,6275	1,0071	0,5103	1,0043	0,5057
0,9	9,00	8,10	Beta	7,3334	7,1021	8,7931	7,9828	8,8948	8,0404
			Uniforme	8,7268	7,8462	9,0079	8,1097	9,0034	8,1043

Tabela 3: Probabilidade preditiva a posteriori do número de clientes no sistema e o fator de Bayes.

N° de	$\rho = 0,2$			$\rho = 0,5$			$\rho = 0,9$		
Clientes	Beta	Uniforme	B_{12}	Beta	Uniforme	B_{12}	Beta	Uniforme	B_{12}
0	0,7979	0,7983	0,9995	0,5027	0,5014	1,0024	0,1015	0,1009	1,0064
1	0,1600	0,1597	1,0015	0,2487	0,2487	1,0000	0,0911	0,0906	1,0057
2	0,0331	0,0330	1,0031	0,1237	0,1240	0,9976	0,0818	0,0840	1,0050
3	0,0070	0,0070	1,0042	0,0618	0,0621	0,9951	0,0734	0,0731	1,0043
4	0,0015	0,0015	1,0050	0,0311	0,0313	0,9927	0,0659	0,0657	1,0035
5	0,0003	0,0003	1,0055	0,0157	0,0158	0,9903	0,0592	0,0590	1,0028

4 Conclusões e Recomendações

Métodos inferenciais sob uma abordagem bayesiana foram utilizados para estimar a intensidade de tráfego ρ em modelos de filas $M/M/1$. O modelo apresentado mostrou-se robusto para se fazer predição, uma vez que o pesquisador tem a vantagem de poder atribuir seu conhecimento prévio a respeito da operação de um sistema simples em teoria de filas. Mesmo sem o conhecimento da taxa de chegada e do tempo de serviço da fila, é possível fazer inferências para a intensidade do tráfego, o número de clientes no sistema no momento da partida e o tamanho da fila, bem como calcular probabilidades no estado estacionário. Utilizando-se valores simulados e duas formas de informação *a priori*, foram obtidas estimativas muitos próximas dos valores verdadeiros, principalmente com o aumento do tamanho amostral. O fator de Bayes não evidenciou o modelo mais adequado para os dados, uma vez que os resultados foram bastante próximos. Como recomendação para futuros trabalhos podem-se utilizar outras distribuição de probabilidade *a priori* além da beta, outros tipos de dados, bem como sistemas de filas mais gerais, tais como $M/G/1$, $M/M/c$, $G/M/1$, entre outros.

DOI: 10.5540/03.2015.003.02.0057

Agradecimentos

Este trabalho foi parcialmente financiado pelo CNPq (processo 304671/2014-2) e FAPEMIG (CEX-PPM-00013-14), aos quais os autores expressam seus agradecimentos.

Referências

[1] C. Armero and M. J. Bayarri, Bayesian prediction in $M/M/1$ queues, Queueing Systems, vol. 15, 401–417, (1994).

[2] A. Choudhury and A. C. Borthakur, Bayesian inference and prediction in the single server Markovian queue, Metrika, vol. 67, 371–383, (2008).

[3] A. B. Clarke, Maximum likelihood estimates in a simple queue, The Annals of Mathematical Statistics, vol. 28, 1036–1040, (1957).

[4] R. S. Ehlers, Introdução à Inferência Bayesiana, (2003), URL `http://www.leg.ufpr.br/%7Epaulojus/CE227/ce227.pdf`.

[5] D. Gamerman and H. F. Lopes, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman and Hall/CRC, London, UK, 2 ed., (2006).

[6] H. Jeffreys, The Theory of Probability, Oxford University Press, Oxford, (1998).

[7] K. S. Kannan and A. Jabarali, Parameter estimation of single server queue with working vacations, Research & Reviews: Journal of Statistics (Special Issue on Recent Statistical Methodologies and Applications), vol. 2, 94–98, (2014).

[8] J. Medhi, Stochastic Models in Queueing Theory, Academic Press, 2 ed., (2003).

[9] M. V. Muddapur, Bayesian estimates of parameters in some queueing models, Annals of the Institute of Statistical Mathematics, vol. 24, 327–331, (1972).

[10] C. D. Paulino, M. A. A. Turkman e B. Murteira, Estatística Bayesiana, Fundação Calouste Gulbenkian, Lisboa, (2003).

[11] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, (2013). URL `http://www.R-project.org/`.

[12] L. Schruben and R. Kulkarni, Some consequences of estimating parameters for the $M/M/1$ queue, Operations Research Letters, vol. 1, 75–78, (1982).

[13] H. M. Wagner, Pesquisa Operacional, Prentice-Hall do Brasil Ltda., Rio de Janeiro, 2 ed., (1986).

DOI: 10.5540/03.2015.003.02.0057