A Bayesian Reassessment of Nearest-Neighbor Classification (2009)

Cucala, Marin, Robert e Titterington

André Yoshizumi Gomes (IME/USP)

Seminário ministrado no Insper

5 de fevereiro de 2016

O método knn original (determinístico)

k-Nearest Neighbor (Ripley, 1994, 1996): Método de classificação supervisionada, para explorar a conexão funcional entre um grupo de variáveis preditoras (X) e uma variável categórica que representa classes pré-estabelecidas (Y).

Exemplo: classificação de peixes.

- Preditoras (X): peso do peixe, comprimento do peixe.
- Classes (Y): salmão / robalo.

O método knn original (determinístico)

Suponha uma amostra treinamento de n observações, cada uma delas alocada em uma de G classes.

- Dados de treinamento: $(y_i, \mathbf{x}_i)_{i=1}^{n}$.
- $y_i \in \{1, 2, \ldots, G\}$: classe da i-ésima observação.
- $\mathbf{x}_i \in \mathbb{R}^d$: vetor de d preditoras da i-ésima observação.

Uma classe y_{n+1} não observada, associada a um conjunto de preditoras $\mathbf{x}_{n+1}$, é estimada pela classe mais frequente dentre os k vizinhos mais próximos de $\mathbf{x}_{n+1}$ na amostra treinamento.

O método knn original (determinístico)

A vizinhança é definida no espaço das covariáveis $\mathbf{x}_i$:

$$N_k(\mathbf{x}_{n+1}) = \left\{1 \leq i \leq n; d\left(\mathbf{x}_i, \mathbf{x}_{n+1}\right) \leq d\left(., \mathbf{x}_{n+1}\right)_{(k)}\right\}.$$

- $d\left(., \mathbf{x}_{n+1}\right)$: Vetor de distâncias (euclidianas) até $\mathbf{x}_{n+1}$.
- $d\left(., \mathbf{x}_{n+1}\right)_{(k)}$: k-ésima estatística de ordem do vetor.

O valor de k é usualmente obtido ao avaliar o erro de classificação de validação cruzada (*leave one out*).
Toma-se o valor de k que minimiza este erro.

O método knn original (determinístico)

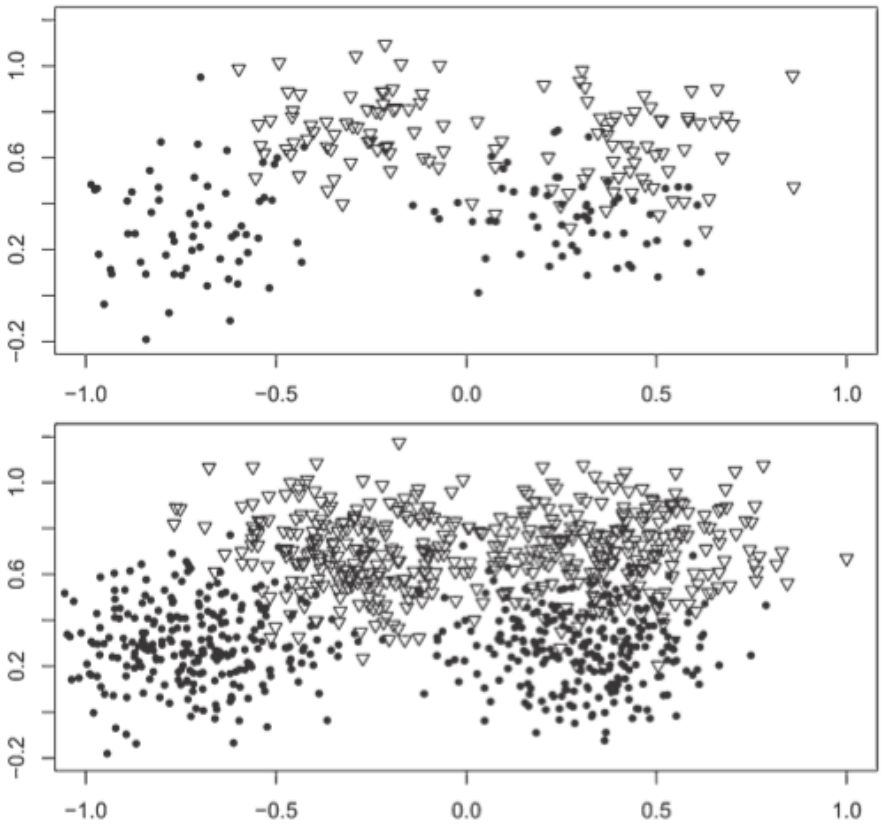
Ilustração: Banco de dados de Ripley (1994). (http://www.stats.ox.ac.uk/pub/PRNN)

- Classificação em duas categorias; populações de mesmo tamanho.
- Amostra treinamento: $n = 250$.
- O modelo é testado em um novo grupo de $m = 1000$ observações.

Os autores também avaliam o erro de classificação de validação cruzada como uma função de k.

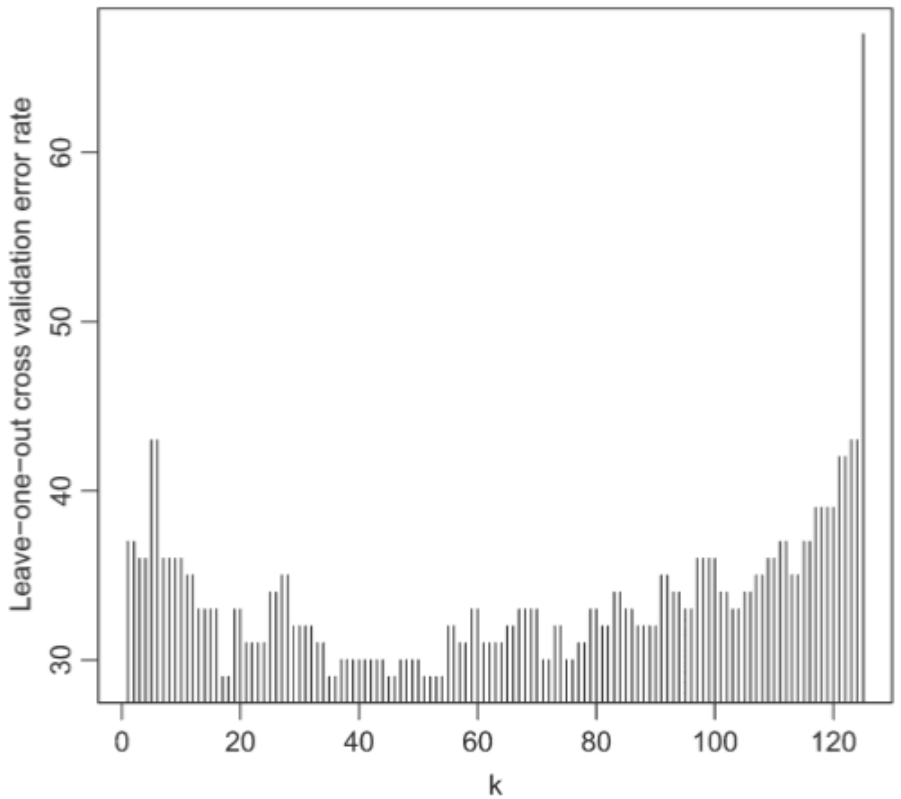
O método knn original (determinístico)

Amostras treinamento e teste (Ripley, 1994).



O método knn original (determinístico)

Taxa de classificação errônea de validação cruzada (leave one out) como função de k.



Objetivo do artigo

Como incorporar a *incerteza* nas previsões das classes de novas observações?

Holmes e Adams (2002, 2003) abordam o problema sob um ponto de vista estatístico.

- Em vez de simplesmente atribuir uma das classes às novas observações, associam uma probabilidade à cada possível classe.
- Utilizam um enfoque Bayesiano: introduzem novos parâmetros, estabelecem distribuições a priori e calculam distribuições preditivas para novas observações.

Objetivo do artigo

Problema: a distribuição condicional conjunta de $\mathbf{y}$ em Holmes e Adams (2002) não está devidamente normalizada.

Os autores corrigem este problema e recalculam as distribuições de interesse.

Ainda comparam diferentes métodos computacionais para fazer inferência sobre β e k.

O modelo knn probabilístico

Para definir a probabilidade conjunta de y_i condicional às preditoras $\mathbf{x}_i$, assumimos que a condicional completa de y_i dados $\mathbf{y}_i = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ e $\mathbf{x}_i$'s depende somente dos k vizinhos mais próximos de $\mathbf{x}_i$.

A estrutura desta probabilidade condicional tem a forma de uma distribuição de Boltzmann (Moller e Waagepetersen, 2003).

Supondo apenas duas classes (0 e 1), a expressão da condicional completa para $i = 1, 2, \ldots, n$ é dada por:

O modelo knn probabilístico

$$f_{Y_i|Y_{-i},X,B,K}\left(y_i|\mathbf{y}_{-i},\mathbf{x},\beta,k\right) = \frac{\exp\left(\left(\frac{\beta}{k}\right)\sum_{j\in N_k(\mathbf{x}_i)} I_{\{y_i\}}(y_j)\right)}{\sum_{t=0,1}\exp\left(\left(\frac{\beta}{k}\right)\sum_{j\in N_k(\mathbf{x}_i)} I_{\{t\}}(y_j)\right)}, \quad (1)$$

onde $I_{\{y_i\}}(y_j) = 1$ se $y_i = y_j$ e 0 caso contrário, e $\beta > 0$ é um parâmetro que regula a "influência" que os vizinhos mais próximos exercem sobre a probabilidade de $y_i = t$, $t \in \{0, 1\}$.

O modelo knn probabilístico

- $\beta = 0$ corresponde a uma distribuição uniforme sobre todas as classes. As probabilidades independem dos vizinhos.
- $\beta \to \infty$, por outro lado, corresponde a atribuir probabilidade 1 à classe de maior frequência dos vizinhos mais próximos, e 0 às demais. Indica dependência extrema dos vizinhos.

O modelo knn probabilístico

Problema: dificilmente existirá uma distribuição conjunta cujas condicionais possuam a forma apresentada em (1).

O sistema knn geralmente é assimétrico. Enquanto $\mathbf{x}_i$ é um dos k vizinhos mais próximos de $\mathbf{x}_j$, $\mathbf{x}_j$ não necessariamente será um dos k vizinhos mais próximos de $\mathbf{x}_i$.

Dessa forma, a distribuição condicional de $\mathbf{x}_i$ não dependeria de $\mathbf{x}_j$, mas a de $\mathbf{x}_j$ dependeria de $\mathbf{x}_i$, o que é impossível de um ponto de vista probabilístico (Besag, 1974; Cressie, 1993).

O modelo knn probabilístico

Holmes e Adams (2002) procuraram definir uma probabilidade condicional conjunta com esta questão em mente. A constante de normalização, porém, não foi devidamente calculada.

A condicional conjunta proposta pelos autores tem a seguinte forma:

$$f_{Y|X,B,K}\left(\mathbf{y}|\mathbf{x},\beta,k\right) = \exp\left(\left(\frac{\beta}{k}\right)\sum_{i=1}^{n}\sum_{j\in N_k(\mathbf{x}_i)} I_{\{y_i\}}(y_j)\right)/Z(\beta,k). \tag{2}$$

O modelo knn probabilístico

Dessa forma, as condicionais completas de cada Y_i relativas a (2) têm a seguinte forma simetrizada:

$$f_{Y_i|Y_{-i},X,B,K}\left(y_i|\mathbf{y}_{-i},\mathbf{x},\beta,k\right) \propto$$
$$\exp\left\{\frac{\beta}{k}\left(\sum_{j\in N_k(\mathbf{x}_i)} I_{\{y_i\}}(y_j) + \sum_{i\in N_k(\mathbf{x}_j)} I_{\{y_j\}}(y_i)\right)\right\} \quad (3)$$

$(i,j = 1,2,\ldots,n; i \neq j)$

Distribuição preditiva de Y_{n+1}

Baseado em (3), a distribuição preditiva de uma nova observação y_{n+1} dado suas preditoras $\mathbf{x_i}$ e a amostra treinamento $(\mathbf{y}, \mathbf{x})$ é, para t = 0,1:

$$P_{Y_{n+1}|Y,X_{n+1},X,B,K}\left(y_{n+1} = t|\mathbf{x}_{n+1}, \mathbf{y}, X, \beta, k\right) \propto$$

$$\exp\left\{\frac{\beta}{k}\left(\sum_{j \in N_k(\mathbf{x}_{n+1})} I_{\{y_{n+1}\}}(y_j) + \sum_{(n+1) \in N_k(\mathbf{x}_j)} I_{\{y_j\}}(y_{n+1})\right)\right\} \quad (4)$$

$(j = 1, 2, \ldots, n)$.

Inferência Bayesiana para o modelo knn

Como obter a distribuição preditiva de y_{n+1} incondicional a β e k?

$$P_{Y_{n+1}|Y,X_{n+1},X}\left(y_{n+1} = t|\mathbf{x}_{n+1}, \mathbf{y}, X\right) = \sum_k \int P\left(y_{n+1} = t|\mathbf{x}_{n+1}, \mathbf{y}, X, \beta, k\right) \pi(\beta, k|\mathbf{y}, X), \quad (5)$$

onde $\pi(\beta, k|\mathbf{y}, \mathbf{x}) \propto f(\mathbf{y}|\mathbf{x}, \beta, k)\pi(\beta, k)$ é a distribuição a posteriori de (β, k) dada a amostra de treinamento $(\mathbf{y}, \mathbf{x})$.

Inferência Bayesiana para o modelo knn

Priori para (β, k): distribuição Uniforme no suporte compacto $[0, \beta_{max}] \times \{1, \ldots, K\}$.

- A limitação em β, $\beta \leq \beta_{max}$ é costume em modelos de Boltzmann, devido ao fenômeno de "transição de fase" (Moller, 2003).
- Além disso, K é, no máximo, igual ao tamanho amostral da classe menos prevalente: $K \leq \min(n_0, n_1)$.

A forma de (3) torna a posteriori $\pi(\beta, k|\mathbf{y}, \mathbf{x})$ difícil de ser obtida. Mas ainda é possível gerar observações (β, k) desta distribuição via Metropolis-Hastings.

Metropolis-Hastings para (β, k)

Considere a seguinte reparametrização para β:

$$\beta = \beta_{max} \frac{\exp(\theta)}{1 + \exp(\theta)}.$$

A distribuição proposta para θ foi um passeio aleatório Normal: $\theta' \sim N\left(\theta^{(t)}, \tau^2\right)$.

Para k, foi proposta uma Uniforme nos $2r$ vizinhos de $k^{(t)}$: $\left\{k^{(t)} - r, \ldots, k^{(t)} - 1, k^{(t)} + 1, \ldots, k^{(t)} + r\right\} \bigcap \{1, 2, \ldots, K\}$, com densidade $Q_r(k, .)$ tal que $k' \sim Q_r(k^{(t-1)}, .)$.

Metropolis-Hastings para (β, k)

A taxa de aceitação do algoritmo MH é

$$\rho = \frac{f\left(\mathbf{y}|\mathbf{x}, \beta', k'\right)\pi(\beta', k')/Q_r(k^{(t-1)}, k')}{f\left(\mathbf{y}|\mathbf{x}, \beta^{(t-1)}, k^{(t-1)}\right)\pi(\beta^{(t-1)}, k^{(t-1)})/Q_r(k', k^{(t-1)})} \times \frac{\exp(\theta')/(1+\exp(\theta'))^2}{\exp(\theta^{(t-1)})/(1+\exp(\theta^{(t-1)}))^2}. \quad (6)$$

Com isso, podemos gerar uma cadeia $\left\{(\beta, k)^{(1)}, \ldots, (\beta, k)^{(M)}\right\}$ da distribuição a posteriori de (β, k).

O problema da constante de normalização

Como o cálculo explícito de (5) é impossível, poderíamos utilizar a seguinte aproximação para a distribuição, a partir da cadeia gerada pelo MH:

$$\frac{1}{M}\sum_{i=1}^{M} P\left(y_{n+1} = t|\mathbf{x}_{n+1}, \mathbf{y}, X, (\beta, k)^{(i)}\right). \qquad (7)$$

Porém, todas as funções de probabilidade obtidas até aqui envolvem, ainda, a intratável constante de normalização em (2), $Z(\beta, k)$. *Como lidar?*

O problema da constante de normalização

Foram consideradas três abordagens:

- *Pseudo-likelihood*: substitui a verdadeira distribuição conjunta pela pseudo-verossimilhança, definida como o produto das condicionais associadas a (2), eliminando assim a intratável constante de normalização.
- *Path sampling*: estima a constante normalizadora através de técnicas de Monte Carlo.
- *Perfect Sampling*: calcula uma distribuição conjunta artificial, obtida por meio de uma variável auxiliar z, que desconsidera as constantes normalizadoras (a distribuição de z é arbitrária).

Path Sampling

A vantagem do *path sampling* é nos permitir trabalhar com a "verdadeira" distribuição de probabilidade.

Ao estimar a constante normalizadora, podemos usar a aproximação MCMC dada por (7) sem maiores problemas.

A partir da expressão em (2), se temos que

$$S(\mathbf{y}) = \sum_i \sum_{j \in N_k(\mathbf{x}_i)} I_{\{y_i\}}(y_j)/k,$$

então

$$Z(\beta, k) = \sum_{\mathbf{y}} \exp[\beta S(\mathbf{y})]. \tag{8}$$

Path Sampling

Além disso,

$$\begin{aligned} \partial Z(\beta, k)/\partial\beta &= \sum_{\mathbf{y}} S(\mathbf{y}) \exp[\beta S(\mathbf{y})] \\ &= Z(\beta, k) \sum_{\mathbf{y}} S(\mathbf{y}) \exp[\beta S(\mathbf{y})]/Z(\beta, k) \\ &= Z(\beta, k)\mathbb{E}_{\beta}[S(\mathbf{y})]. \end{aligned} \tag{9}$$

Assim, a razão $Z(\beta, k)/Z(\beta', k)$ pode ser obtida pela seguinte integral:

$$\log\left\{Z(\beta, k)/Z(\beta', k)\right\} = \int_{\beta'}^{\beta} \mathbb{E}_{u,k}[S(\mathbf{y})]du. \tag{10}$$

Path Sampling

Como obter a aproximação para $Z(\beta, k)$?

Para o caso de amostras balanceadas (ambas as classes têm o mesmo número de observações), sabemos que para $\beta = 0$, $\mathbb{E}_{0,k}[S(\mathbf{y})] = n/2$. Assim,

$$\log Z(\beta, k) = n \log 2 + \int_0^\beta \mathbb{E}_{u,k}[S(\mathbf{y})] du, \qquad (11)$$

que pode ser calculado via integração numérica.

Estudo: Dados Simulados (Ripley, 1994)

Vamos utilizar o mesmo conjunto de dados já apresentado para comparar os métodos.

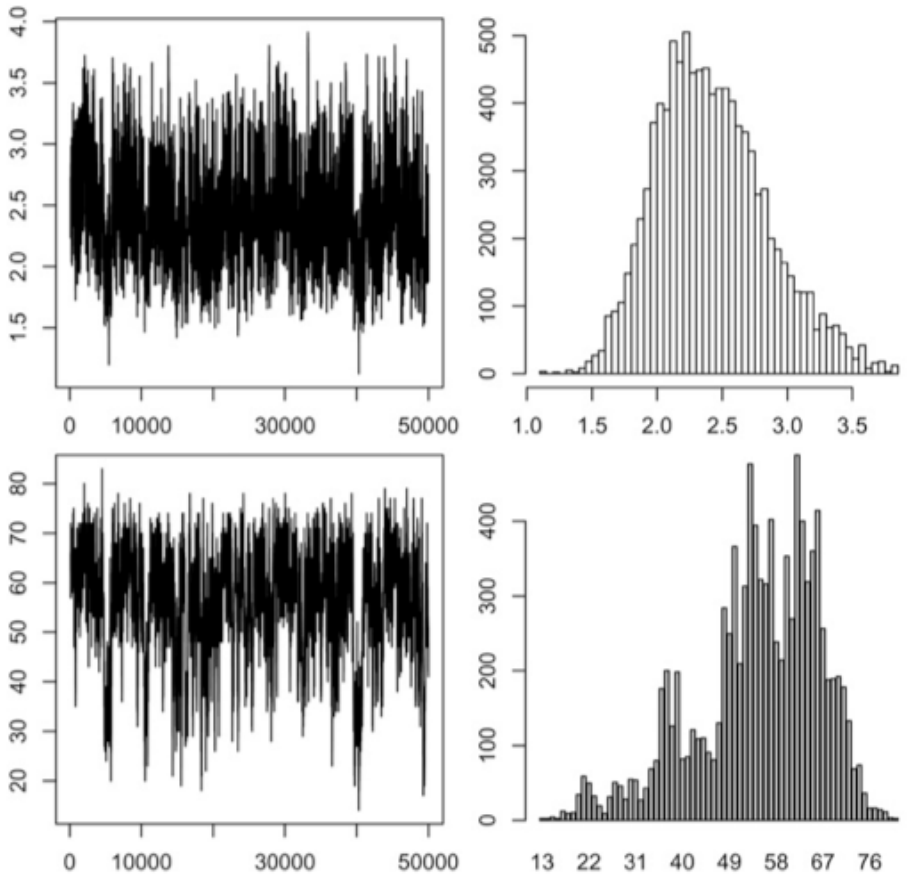
Objetivo: inferência sobre β e k.

Foram realizadas 50 mil iterações dos algoritmos, após 40 mil iterações de *burn-in*.

Os gráficos de cima dizem respeito a β, e os de baixo a k.

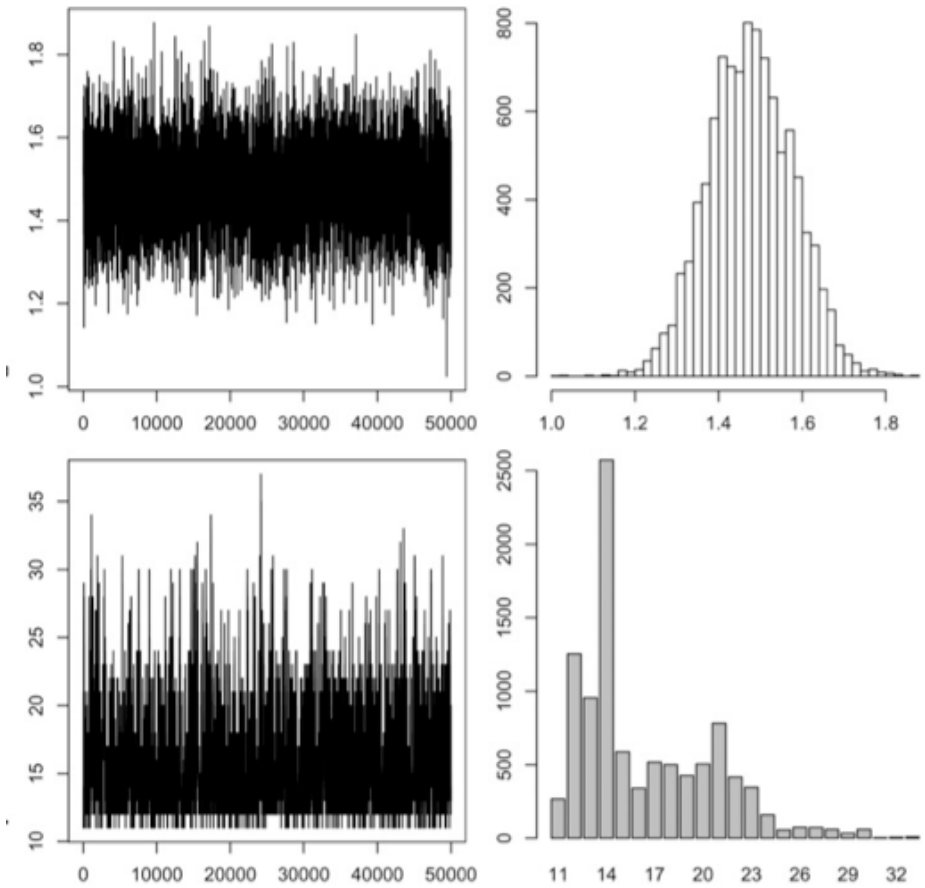
Estudo: Dados Simulados (Ripley, 1994)

Pseudo-likelihood



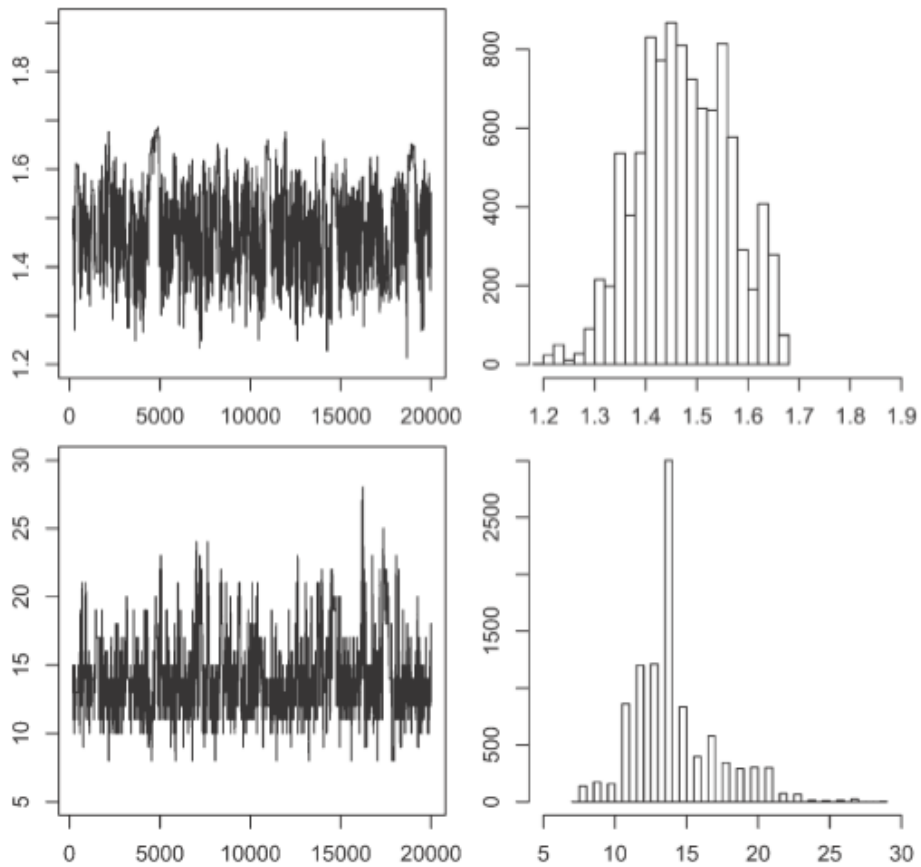
Estudo: Dados Simulados (Ripley, 1994)

Path sampling



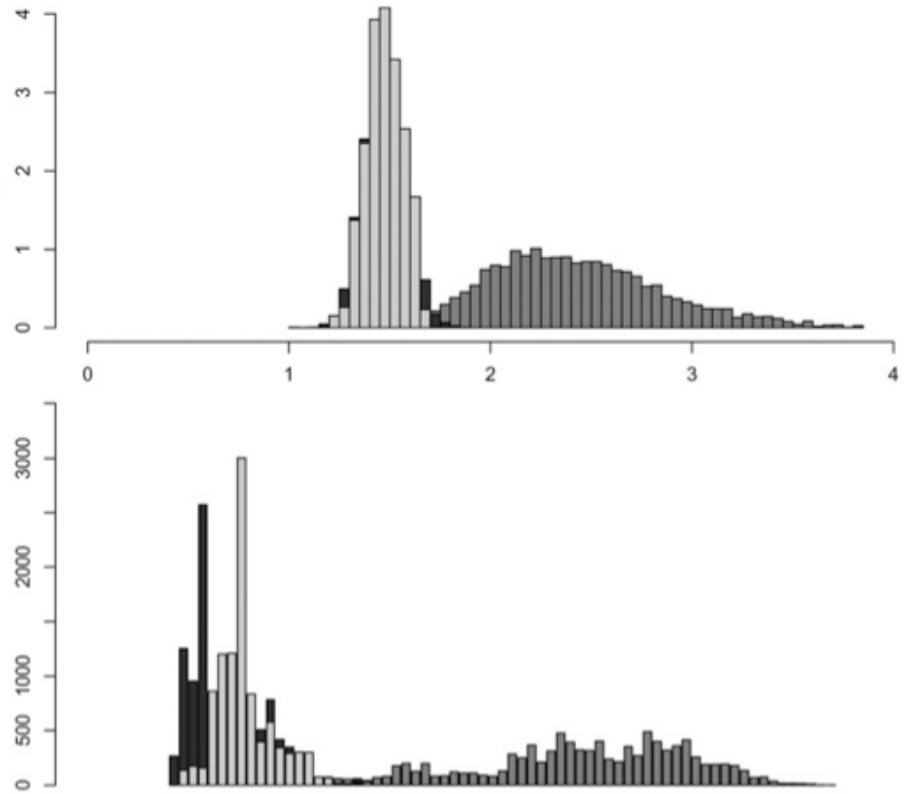
Estudo: Dados Simulados (Ripley, 1994)

Perfect sampling



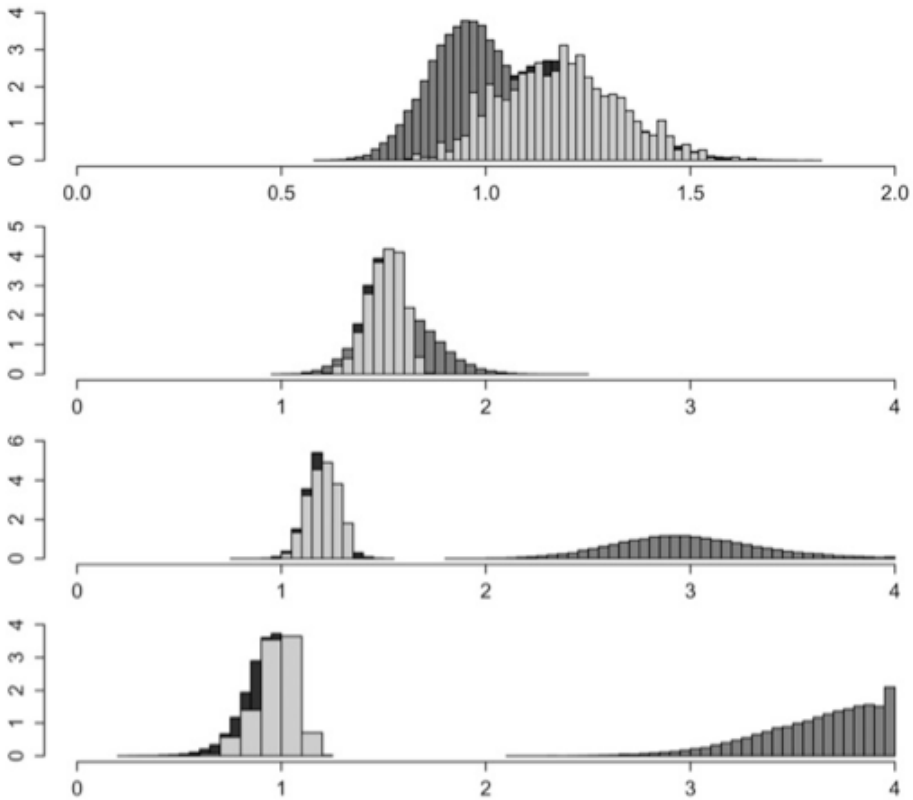
Estudo: Dados Simulados (Ripley, 1994)

Sobreposição dos três métodos (claro: perfect sampling; médio: pseudo-likelihood; escuro: path sampling).



Estudo: Dados Simulados (Ripley, 1994)

Aproximação da posteriori de β para $k = 1, 10, 70$ e 125 (claro: perfect sampling; médio: pseudo-likelihood; escuro: path sampling).



Considerações Finais

O artigo tratou de corrigir um problema observado no paper de Holmes e Adams (2002, 2003), em que a constante normalizadora da distribuição não estava propriamente calculada.

Os autores comparam diferentes metodologias que lidam com a constante e viabilizam a inferência sobre os parâmetros do modelo.

A metodologia de *pseudo-likelihood* forneceu inferências bastante suspeitas para os dados simulados, enquanto *path sampling* e *perfect sampling* pareceram ser mais robustos. Aqui, *path sampling* leva vantagem por ser computacionalmente muito mais rápido.

Referências

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), Journal of the Royal Statistical Society, Ser. B, 36, 192–236.

Cressie, N. A. C. (1993), Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: John Wiley & Sons.

Holmes, C. C., and Adams, N. M. (2002), "A Probabilistic Nearest Neighbour Method for Statistical Pattern Recognition," Journal of the Royal Statistical Society, Ser. B, 64, 295–306.

Holmes, C. C., and Adams, N. M. (2003), "Likelihood Inference in Nearest-Neighbour Classification Models," Biometrika, 90, 99–112.

Referências

Moller, J., and Waagepetersen, R. (2003), Statistical Inference and Simulation for Spatial Point Processes, Boca Raton, FL: Chapman and Hall/CRC.

Ripley, B. D. (1994), "Neural Networks and Related Methods for Classification" (with discussion)," Journal of the Royal Statistical Society, Ser. B, 56, 409–456.

Ripley, B. D. (1996), Pattern Recognition and Neural Networks, Cambridge: Cambridge University Press.