ESTIMAÇÃO DE DENSIDADES E APLICAÇÃO EM SEGMENTAÇÃO DE IMAGENS

Giovana Marassi Zambon

Monografia apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Graduado em Física Médica.

BOTUCATU
São Paulo - Brasil
Novembro – 2010

ESTIMAÇÃO DE DENSIDADES E APLICAÇÃO EM SEGMENTAÇÃO DE IMAGENS

Giovana Marassi Zambon

Orientadora: Prof. Dr. **Luzia Aparecida Trinca**

Monografia apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Graduado em Física Médica.

BOTUCATU
São Paulo - Brasil
Novembro – 2010

Ficha Catalográfica

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE

Zambon, Giovana Marassi.
Estimação de densidades e aplicação em segmentação de imagens/ Giovana Marassi Zambon. - Botucatu, 2010

Trabalho de conclusão de curso (bacharelado - Física Médica) - Instituto de Biociências de Botucatu, Universidade Estadual Paulista, 2010
Orientador: Luzia Aparecida Trinca
Capes: 20303025

1. Algoritmos de estimação-maximização. 2. Processos gaussianos. 3. Distribuição (Probabilidades).

Palavras-chave: Algoritmo EM; Histograma; Mistura de gaussianas.

Dedicatória

Aos meus pais e meu irmão,

José Carlos, Lúcia e Franthiesco

Pelo amor, apoio e incentivo em todos os momentos da minha vida.

Agradecimentos

Agradeço primeiramente e acima de tudo a Deus, pois nada seria possível se não fosse pela vontade dele.

Em especial, aos meus pais e meu irmão, exemplo e referência para mim ao longo de toda a minha existência. Agradeço por estarem sempre presentes me apoiando e por toda ajuda, compreensão e estímulos concedidos.

À Profa. Dra. Luzia Aparecida Trinca pela orientação, dedicação e todo ensinamento transmitido ao longo do período de estágio, meus sinceros agradecimentos.

Aos professores, especialmente à Profa. Dra. Diana Rodrigues de Pina, ao Prof. Dr. José Ricardo de Arruda Miranda e ao Prof. Dr. Paulo Fernando de Arruda Mancera, que muito contribuíram para minha formação, através de todos os ensinamentos durante o curso de Física Médica.

À V Turma do curso de Física Médica da UNESP de Botucatu, em especial aos amigos de faculdade Adriana, Anna Luiza, Bruna, Diego, Lucas, Marcela, Mariana, Nayana, Tamara e Victor pela amizade, parcerias nos estudos, apoio e por todos os momentos agradáveis que passamos juntos.

À Micaela Gisoldi Ribeiro por ser a minha família de Botucatu, por todo apoio, companheirismo e por todos os momentos de felicidade.

Aos meus grandes amigos Diego Rodrigues, Gabriela Calgaro, Isabella Birolli e Naiara Pereira por estarem sempre ao meu lado e pelos diversos e imprescindíveis conselhos.

À toda minha família - avós, tios, primos, cunhada- e à todos aqueles que, direta ou indiretamente, colaboraram para minha formação.

Porfim, agradeço o apoio financeiro cedido pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

"A mente que se abre a uma nova idéia jamais voltará ao seu tamanho original"

(Albert Einstein)

Sumário

ESTIMAÇÃO DE DENSIDADES E APLICAÇÃO EM SEGMENTAÇÃO DE IMAGENS

Autora: GIOVANA MARASSI ZAMBON
Orientadora: Prof. Dr. LUZIA APARECIDA TRINCA

RESUMO

Este trabalho tem como objetivo o estudo dos métodos de estimação de densidades e posterior aplicação em segmentação de imagem de tomografia computadorizada. Uma imagem digitalizada pode ser vista como uma matriz de dados cujos valores representam tons de cores em certa escala. O menor elemento no dispositivo de exibição de imagem ao qual é possível atribuir uma cor (um valor) é chamado de pixel. O processo de segmentação consiste na divisão da imagem em regiões através da classificação de cada pixel em categorias. O método mais simples de classificação usa limiares para os valores dos pixels mas exige o conhecimento tanto do número de regiões distintas na imagem como dos próprios limiares. Outra maneira simples é construir um histograma dos valores dos pixels, considerar cada pico como um objeto distinto e definir o limiar como sendo a média entre os valores referentes a dois picos vizinhos. Como o histograma não forma uma curva suave, é difícil decidir

quando um pico realmente representa uma parte distinta ou é apenas uma variação casual. Técnicas mais sofisticadas de estimação de densidades são requeridas. Dados de imagens podem ser considerados como oriundos de mistura de densidades. Neste estudo serão abordadas as técnicas de estimação de densidades paramétrica e não paramétrica. A técnica de estimação por máxima verossimilhança foi aplicada na segmentação de uma imagem de TC.

Palavras-chave: Algoritmo EM; Histograma; Mistura de gaussianas.

DENSITY ESTIMATION AND APPLICATIONS TO IMAGE SEGMENTATION

Author: GIOVANA MARASSI ZAMBON
Adviser: Prof. Dr. LUZIA APARECIDA TRINCA

SUMMARY

The aim of this work is to study some of the density estimation techniques and to apply to the segmentation of medical images. Medical images are used to help the diagnostic of tumor diseases as well as to plan and deliver treatment. A computer image is an array of values representing colors in some scale. The smallest element of the image to which it is possible to assign a value is called pixel. Segmentation is the process of dividing the image in portions through the classification of each pixel. The simplest way of classification is by thresholding, given the number of portions and the threshold values. Another method is constructing a histogram of the pixel values and assign a portion to each pike. The threshold is the mean between two pikes. As the histogram does not form a smooth curve it is difficult to discern between true pikes and random variation. Density estimation methods allow the estimation of a smooth curve. Image data can be considered as mixture of dif-

ferent densities. In this project parametric and nonparametric methods for density estimation will be addressed and some of them are applied to CT image data.

Key word: EM Algorithm; Histogram; Mixture of Gaussians.

1 INTRODUÇÃO

Várias patologias são detectadas e diagnosticadas baseadas na análise de imagens médicas, tais como imagens de tomografia computadorizada (TC). Nestes exames os diferentes tecidos e estruturas que compõem o corpo humano são visualizados nas imagens devido a propriedades relacionadas à interação da radiação com a matéria. Tais propriedades fazem com que ocorram diferentes intensidades de atenuação entre os tipos de tecidos incluindo pequenas diferenciações dentro de um mesmo tecido ou estrutura anatômica. Do ponto de vista de análise, uma imagem é um *array* de valores que representam tons de cores em certa escala. A maioria da imagens médicas usa tons de cinza. Em TC, a imagem é formada de modo que cada tom de cinza corresponde a uma porção específica de atenuação da radiação referente ao tecido irradiado. O menor elemento no dispositivo de exibição de imagem ao qual é possível atribuir uma cor (um valor) é chamado de pixel.

Utilizando imagens, é possível fazer o diagnóstico precoce de tumores em fase inicial de desenvolvimento bem como o acompanhamento do tratamento. Tradicionalmente, esta tarefa é desenvolvida por radiologistas, porém, com o avanço tecnológico, grandes avanços empregados no desenvolvimento de aparelhos para o diagnóstico médico foram obtidos, possibilitando a coleta de informações cada vez mais completas e complexas, visando melhorar a qualidade do diagnóstico por imagem.

Existe uma tendência para a automação ou semi-automação da análise de imagens, tanto como forma de eliminar subjetivismo, como também para agilizar o processo e produzir respostas mais rápidas. Naturalmente, no caso de diagnóstico, o profissional especializado deve continuar visualizando e checando as imagens já que a vista humana talvez seja imbatível na análise qualitativa. A tendência é

automatizar as tarefas mais trabalhosas e de quantificação e talvez produzir sistemas que permitem a intervenção do profissional quando necessário. No caso de doenças tumorais, várias metodologias podem ser empregadas para segmentar e possivelmente quantificar histologicamente a massa tumoral. O objetivo é melhorar a análise radiológica oferecendo medidas mais sofisticadas, precisas e reprodutíveis das informações presentes nas imagens. Neste contexto, o clínico responsável pode ser capaz de promover um melhor planejamento e avaliar melhor o progresso de um dado tratamento através das imagens médicas.

Na literatura são encontradas diferentes aplicações de segmentação de imagens que auxiliam o diagnóstico clínico. A segmentação de imagens de TC pode ser aplicada para detecção e quantificação de pólipos intestinais (Huang & Chau, 2008; Li et al., 2004), sendo este um método não invasivo e, então, mais tolerado pelos pacientes. Em imagens de PET (tomografia computadorizada por emissão de pósitron) a técnica de segmentação é utilizada para a quantificação das regiões tumorais (Amira et al., 2008; Montgomery et al., 2007), o que é de grande utilidade para o estadiamento do tumor e planejamento na radioterapia. Em imagens de TC pulmonar, a segmentação dos pulmões pode ser aplicada para a análise das vias aéreas (Park & nd M Sonka, 1998), detecção de enfisema (Blechschmidt et al., 2001), avaliação da ventilação pulmonar (Brown et al., 1997) e detecção de nódulos pulmonares (Dajnowiec & Alirezaie, 2004; Lee et al., 2001). Tanto imagens de ressonância magnética (RM) quanto de TC fornecem imagens de ótima qualidade que permitem uma análise detalhada do coração, a aplicação da segmentação permite um delineamento do miocárdio em cada imagem no conjunto de dados do paciente, computar o volume, fração de ejeção e analisar a espessura das paredes (Jolly, 2006). O pré-processamento tradicional para detecção de melanomas envolve o delineamento manual das bordas ao redor das áreas lesionadas da imagem, embora seja um método simples, este procedimento exige muito tempo e é muito subjetivo considerando que cada dermatologista possui um nível diferente de experiência, então, é realizada a segmentação para o delineamento automático de melanomas (Tran, 2005). Também,

há aplicações de segmentação intracranial de estruturas de imagens de TC cerebral (Jiang et al., 2007).

A análise de imagens envolve várias etapas, desde a eliminação de possíveis ruídos, classificação de tipos de material, até a estimação de quantidades de interesse. Técnicas estatísticas são aplicadas em, praticamente, todas as fases. Na fase de eliminação de ruídos, alguns tipos de filtragem são usados, para limpar e realçar aspectos de interesse da imagem. O processo de classificação de regiões de interesse na imagem é chamado de segmentação e consiste na divisão da imagem em regiões que correspondem a diferentes objetos ou parte de objetos. Esta divisão envolve a classificação de cada pixel. O método mais simples de classificação usa limiares para os valores dos pixels mas exige o conhecimento tanto do número de regiões distintas na imagem como dos próprios limiares. Outra maneira simples é construir um histograma dos valores dos pixels, considerar cada pico como um objeto distinto e definir o limiar como sendo a média entre os valores referentes a dois picos vizinhos. Como o histograma não forma uma curva suave, é difícil decidir quando um pico realmente representa uma parte distinta ou é apenas uma variação casual. Técnicas mais sofisticadas de estimação de densidades são requeridas e serão abordadas nesta pesquisa.

Este trabalho tem como objetivo principal estudar as técnicas de estimação de densidades. Para a aplicação da técnica foi utilizado uma imagem médica de tomografia computadorizada disponíveis em bancos de dados públicos (por exemplo InVesalius, 2009). Vale ressaltar que as imagens disponíveis em bancos públicos, por questões éticas, são anônimas e não se conhece o estado de saúde dos indivíduos. Vale também ressaltar que o uso destas imagens neste trabalho tem a finalidade apenas de treinamento das técnicas estudadas.

Primeiramente serão apresentadas técnicas de estimação não paramétrica de densidades. Para isto foi feito um levantamento bibliográfico sobre os métodos de estimação de densidades.

Por fim, serão apresentadas técnicas de estimação paramétrica, será

abordado uma técnica de estimação de misturas de gaussianas e o fundamento do algoritmo EM. Para as aplicações foram utilizadas as rotinas *EM image segmentation* de MATLAB (2009), que utiliza o software MATLAB©.

Esta pesquisa está sendo desenvolvida como iniciação científica e conta com o apoio financeiro da FAPESP, processo 2008/10994-5.

2 FUNDAMENTOS TEÓRICOS

2.1 Conceitos Básicos

Para abordar o tema estimação de densidades deve-se primeiramente recapitular o conceito de variável aleatória que por sua vez tem origem num experimento aleatório.

Experimento aleatório é qualquer procedimento, que em teoria pode ser repetido infinitamente, cujo resultado não pode ser previsto com certeza. O conjunto de todos os resultados possíveis de um experimento aleatório é chamado de espaço amostral. Variável aleatória é uma função que possui o espaço amostral como domínio e os números reais como contradomínio. As variáveis aleatórias podem ser classificadas em discretas ou contínuas. As variáveis discretas são aquelas cujos valores assumidos são enumeráveis, já as variáveis contínuas podem assumir infinitos valores, porém não enumeráveis, dentro de um intervalo. O comportamento de uma variável aleatória é caracterizado por sua distribuição de probabilidades.

No caso discreto, uma função (ou distribuição) de probabilidade é uma função $p(x_i)$ que atribui a cada valor possível da variável aleatória sua probabilidade de ocorrência ($0 \leq p(x_i) \leq 1; \sum_{i=1}^{\infty} p(x_i) = 1$). Para variáveis aleatórias contínuas, a distribuição de probabilidades é uma função contínua também chamada de função densidade de probabilidade e é denotada por $f(x)$. $f(x)$ é uma função densidade para uma variável aleatória contínua X, se satisfaz as seguintes condições:

1. $f(x) \geq 0$, para todo $x \in (-\infty, \infty)$

2. A área definida por $f(x)$ é igual a 1.

A probabilidade de uma variável assumir um valor no intervalo $[a; b]$ é dada por

$$P(a \leq X \leq b) = \int_a^b f(x)dx, \tag{1}$$

ou seja, a referida probabilidade corresponde à área sob $f(x)$ no intervalo em questão.

A função de distribuição de probabilidade(fdp) depende de uma ou mais constantes chamadas de parâmetros. Por exemplo, a distribuição exponencial tem apenas um parâmetro, já distribuição Normal ou Gaussiana tem dois parâmetros. Os parâmetros serão, genericamente referidos por θ para escalar e $\boldsymbol{\theta}$ para um vetor de parâmetros.

Esperança de uma variável aleatória X, também denominada valor esperado ou média de uma variável aleatória X, é uma função de X e de sua fdp simbolizada por $E(X)$ ou μ e definida por:

$$E(X) = \mu = \begin{cases} \sum_{i=1}^{\infty} x_i p(x_i) & \text{se } X \text{ é uma variável aleatória discreta;} \\ \int_{-\infty}^{\infty} x f(x)dx & \text{se } X \text{ é uma variável aleatória contínua,} \end{cases} \tag{2}$$

em que $p(x_i)$ é a função de probabilidade da variável discreta ($\Omega_X = (x_1,\ x_2,\ \ldots)$) e $f(x)$ é a fdp da variável contínua ($\Omega_X = R$). Como essa função resulta numa constante dependente dos parâmetros da fdp, ela também é um parâmetro.

A variância de uma variável aleatória X, simbolizada por $Var(X)$ e frequentemente por σ^2, é definida por:

$$Var(X) = E[(X - \mu)^2]. \tag{3}$$

Se X é uma variável aleatória discreta:

$$Var(X) = \sum_{i=1}^{\infty} (x_i - \mu)^2 p(x_i). \tag{4}$$

Se X é uma variável aleatória contínua:

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx. \tag{5}$$

Sejam X e Y duas variáveis aleatórias discretas originárias do mesmo fenômeno aleatório. A distribuição conjunta destas duas variáveis é dada por:

$$p(x, y) = P[(X = x) \cap (Y = y)] = P(X = x, Y = y), \tag{6}$$

em que $p(x, y)$ representa a probabilidade de $(X,\ Y)$ ser igual a $(x,\ y)$. No caso contínuo a densidade conjunta é denotada por $f(x,\ y)$.

Dado uma distribuição conjunta de $(X,\ Y)$, pode-se obter a função de probabilidade marginal X somando-se (caso discreto) ou integrando-se (caso contínuo) com respeito ao espaço amostral de Y (e vice-versa):

$$P(X = x) = \sum_{y \in \Omega_Y} p(x, y) \tag{7}$$

ou

$$f(x) = \int_{y \in \Omega_Y} f(x, y) dy. \tag{8}$$

A probabilidade de $X = x$, dado que $Y = y$ ocorreu é conhecida como probabilidade condicional e representa a probabilidade de se ter uma saída x dada uma entrada com valor específico y. A probabilidade condicional é dada pela fórmula de Bayes

$$P(X|Y = y) = \frac{P(X \cap Y = y)}{P(Y = y)}$$

no caso discreto e

$$f(x|y) = \frac{f(x,\ y)}{f(y)}$$

no caso contínuo. Se a esperança for tomada tendo como base a fdp condicional tem-se a esperança condicional, ou seja:

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f(x|y) dx,$$

com a integral substituída pela soma no caso discreto. A mesma idéia segue para variância condicional.

Pela fórmula de Bayes tem-se que se as variáveis aleatórias são independentes a distribuição conjunta é dada pelo produto de suas marginais, ou seja

$$f(x,\ y) = f(x) \times f(y).$$

Em geral, para enfatizar a dependência da fdp em seus parâmetros, estas são escritas como $f(x|\theta)$ no caso uniparamétrico ou $f(x|\boldsymbol{\theta})$ no caso multiparamétrico.

No caso de uma amostra aleatória, de tamanho n, $X_1,\ X_2,\ \ldots, X_n$, em que os X_i's são independentes a fdp conjunta de $\mathbf{X} = (X_1,\ X_2,\ \ldots, X_n)$ é dada por

$$f(\mathbf{x}|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta}) \times f(x_2|\boldsymbol{\theta}) \times \ldots \times f(x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

Dentre as possíveis fdp's existe uma classe que pode ser escrita de uma maneira conveniente e cujas propriedades foram exaustivamente estudadas. Esta classe é chamada de família exponencial. Ela inclui muitas das distribuições de probabilidade mais comumente utilizadas em Estatística, tanto contínuas quanto discretas, por exemplo a Gaussiana, Gama, Poisson, Binomial, etc. A função uniparamétrica $f(x|\theta)$ pertence à família exponencial se sua densidade pode ser escrita como

$$f(x|\theta) = \frac{b(x)\exp(C(\theta)t(x))}{a(\theta)}. \tag{9}$$

Se $C(\theta) = \theta$, θ é chamado de parâmetro natural. A extensão para K parâmetros é

$$f(x|\boldsymbol{\theta}) = \frac{b(x)\exp(\mathbf{C}(\boldsymbol{\theta})\mathbf{t}(x)^T)}{a(\boldsymbol{\theta})}, \tag{10}$$

em que $\mathbf{C}(\boldsymbol{\theta})$ e $\mathbf{t}(x)$ são vetores linha de dimensão K e o sobrescrito T representa a operação transposta.

2.2 Estimação não paramétrica de densidades

A estimação de densidades é a construção da estimativa da função de densidade a partir dos dados observados. A estimação de densidades pode ser feita através de métodos paramétricos ou não-paramétricos. O primeiro método assume que os dados são tirados de uma distribuição pertencente a uma família de distribuições paramétricas conhecida e o vetor de parâmetros da função é estimado. Já o outro método estima a função sem que nenhum modelo específico seja adotado. A ideia das técnicas não paramétricas é que, embora seja assumido que uma certa distribuição tenha uma densidade de probabilidade f, ao invés de forçar f a cair em uma dada família paramétrica, os dados poderão falar por si para determinar a estimativa de f. Para isso, vários métodos podem ser utilizados. Neste trabalho

serão abordados os métodos: histograma, estimador *naive*, estimador pelo método do núcleo e estimador de máxima verossimilhança, incluindo um método para estimar densidades de misturas, o algoritmo EM. No desenvolvimento que se segue assumi-se um conjunto de n observações independentes $X_1, ..., X_n$ de uma variável aleatória X e o símbolo $\hat{f}$ é usado para denotar o estimador de densidade em questão.

2.2.1 Histograma

O histograma é o estimador de densidade mais antigo e mais utilizado para apresentação gráfica de resumo de dados. A construção do histograma consiste em dividir a amplitude dos dados em k classes (intervalos) disjuntos e contar o número de observações pertencentes a cada intervalo de classes. Dado uma origem x_0 e um valor para o comprimento (h) do intervalo de classe, define-se os intervalos de classe do histograma pelo intervalo $[x_0 + mh, x_0 + (m + 1)h)$ para valores inteiros de m.

O estimador de $f(x)$, pelo método histograma é definido por

$$\hat{f}(x) = \frac{1}{nh} \times (\text{número de ocorrências de } X_i \text{ no mesmo intervalo de classe que } x). \tag{11}$$

É possível observar na equação (11) que, considerando o conjunto de dados completos, a construção do histograma requer a escolha da origem x_0 e do comprimento h das k classes. O comprimento h do intervalo de classe é escolhida de acordo com os dados e o enfoque que se deseja. Com a finalidade de se obter mais detalhes no histograma, deve-se usar um número maior de classes e consequentemente um menor comprimento h. Um histograma sem muitos detalhes é obtido usando-se um comprimento h maior com um número menor de classes. Contudo, a correta escolha dos parâmetros k e h se dão de forma empírica, não havendo um padrão de escolha pré-estabelecido. Deve-se salientar que um fator que influencia a qualidade da análise do histograma é o próprio tamanho do conjunto de dados e este fato acaba se tornando um fator limitante para o uso dos histogramas. Outra desvantagem que o torna menos útil para análises posteriores é a falta de continuidade.

2.2.2 O estimador *naive*

Por definição tem-se que

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h). \tag{12}$$

Para qualquer h, $P(x - h < X < x + h)$ pode ser estimada pela proporção de observações da amostra pertencentes ao intervalo $(x-h, x+h)$. Então o estimador *naive* de densidade $\hat{f}$ é dado por

$$\hat{f}(x) = \frac{1}{2hn} \times (\text{número de } X_1, ..., X_n \text{ em } (x - h, x + h)), \tag{13}$$

em que h é pequeno.

Para expressar o estimador de uma forma mais clara, considere a função peso w definida por:

$$w(x) = \begin{cases} 0,5 & \text{se } |x| < 1 \\ 0 & \text{caso contrário .} \end{cases}$$

Então, o estimador naive pode ser escrito como

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right). \tag{14}$$

Esta estimativa é construída colocando uma caixa de comprimento $2h$ e altura $(2nh)^{-1}$ em cada observação e então somando-se para obter a estimativa. Para uma interpretação mais construtiva, considere um histograma construído utilizando divisões com um comprimento $2h$. Assuma que nenhuma observação caia exatamente na fronteira de um intervalo. Para x localizado exatamente no centro de uma divisão do histograma, a estimativa *naive* $\hat{f}(x)$ será exatamente a ordenada do histograma no ponto x. Assim, a estimativa *naive* pode ser vista como uma tentativa de se construir um histograma em que todos as observações da amostra estão no centro do intervalo amostral. A escolha da largura de cada compartimento ainda exerce um papel influente na qualidade da estimativa obtida.

Considerando o ponto de vista da estimativa de densidade, o estimador *naive* possui certas características indesejáveis. Segue que, de sua própria definição, o estimador *naive* não é uma função contínua, mas sim, possui saltos em $X_i \pm h$ e derivadas nulas em todos os pontos. Isto resulta em figuras grosseiras de densidades, dificultando a interpretação, principalmente por pessoas não treinadas.

2.2.3 O estimador de densidade pelo método do núcleo (*kernel*)

O estimador pelo método do núcleo pode ser considerado como uma variação do estimador *naive*, contudo, algumas dificuldades encontradas com o estimador *naive* são superadas. Considere a substituição da função peso w pela função núcleo K que satisfaz a condição:

$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{15}$$

Geralmente, K é uma função de densidade de probabilidade simétrica, como exemplo a função de densidade de probabilidade normal. Por analogia com a definição do estimador naive, o estimador núcleo K é definido por:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{16}$$

onde h é o comprimento do espaçamento, também conhecido como largura de banda ou parâmetro de suavização.

Assim como o estimador *naive* pode ser considerado como uma soma de "caixas"centradas nas observações, o estimador pelo método do núcleo é a soma dos "picos"localizados nas observações, gerados pela função núcleo. A função núcleo K determina a forma dos picos enquanto que h determina sua largura. Novamente, a escolha do valor h implica na qualidade da estimação. A escolha de um valor h baixo resulta em uma estimativa da densidade de probabilidade mais detalhada enquanto que um valor h alto fornece uma estimativa mais suavizada.

O problema na escolha da intensidade de suavização, ou seja, escolha do valor de h é muito importante na estimação de densidade. A escolha apropriada do valor de h dependerá sempre da finalidade que se deseja na estimação. Pode-se escolher este valor por simples inspeção visual, contudo muitas aplicações requerem uma escolha automática deste parâmetro. Esta escolha automática, também pode ser utilizada como ponto de partida para outras tentativas posteriores afim de encontrar o melhor valor. Se a estimação de densidade é utilizada habitualmente para um grande conjunto de dados ou como parte de um procedimento, então o método automático é essencial.

Uma abordagem fácil e natural é usar uma família de distribuição padrão, deste modo, se é usado o estimador pelo método do núcleo Gaussiano, o valor de h ótimo é dado por:

$$h_o = 1.06\sigma n^{-1/5}. \tag{17}$$

Silverman (1986) mostra que tal escolha otimiza algumas propriedades do estimador. Para alguns casos, um melhor resultado pode ser obtido reescrevendo a Fórmula 17 em termos da amplitude interquartística R,

$$h_o = 0.79Rn^{-1/5}. \tag{18}$$

Existe uma vasta literatura sobre outras propostas de otimização da escolha de h que não será apresentada aqui.

2.3 Estimação paramétrica de densidades

2.3.1 Estimador de Máxima Verossimilhança

Um dos métodos paramétricos de estimação de densidades mais usados é o de Máxima Verossimilhança (MV). A teoria mostra que os estimadores de MV apresentam muito boas propriedades como, por exemplo, não tendenciosidade e consistência se tamanhos de amostras razoáveis são utilizados.

Primeiramente, será introduzido brevemente o conceito de função de verossimilhança.

Sejam $X_1, ..., X_n$ uma amostra aleatória, e X_i's independentes, de tamanho n da variável aleatória X com função de densidade de probabilidade $f(x|\theta)$, com $\theta \in \Theta$, em que Θ é o espaço paramétrico. Após observar a amotra, ou seja, obter $x_1, ..., x_n$ a função de verossimilhança de θ é a densidade conjunta dos X_i's, dada por

$$L(\theta; \mathbf{x}) = f(x_1; \theta) f(x_2; \theta) ... f(x_n; \theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{19}$$

O estimador de máxima verossimilhança de θ é o valor $\hat{\theta} \in \Theta$ que maximiza a função de verossimilhança $L(\theta; \mathbf{x})$ (Bolfarine, 2001). Note que a função de verossimilhança é considerada uma função do parâmetro θ fixado o vetor de dados observados $\mathbf{x} = (x_1, x_2, ..., x_n)$ já que as observações já foram feitas. Em muitos casos é mais fácil encontrar o máximo de $\ln L(\theta; \mathbf{x})$ do que de L e como a função log é monótona, de tal modo que estas duas funções alcançarão seu valor máximo para o mesmo valor de θ, costuma-se usar o logaritmo natural da função de verossimilhaça denotado por

$$l(\theta; \mathbf{x}) = \ln L(\theta; \mathbf{x}). \tag{20}$$

Sob condições gerais, admitindo-se que θ seja um número real e que $L(\theta; \mathbf{x})$ seja uma função derivável em θ, o estimador de máxima verossimilhança pode ser encontrado através da resolução de:

$$l'(\theta; \mathbf{x}) = \frac{dl(\theta; \mathbf{x})}{d\theta} = 0 \tag{21}$$

que é conhecida como equação de verossimilhança. Para se concluir que a solução da equação (21) é um ponto de máximo, é necessário verificar se

$$l''(\hat{\theta}; \mathbf{x}) = \frac{d^2 l(\theta; \mathbf{x})}{d\theta^2}\Big|_{\theta=\hat{\theta}} < 0. \tag{22}$$

No decorrer desta seção serão apresentados alguns exemplos para melhor ilustrar a teoria abordada.

Exemplo 1: Sejam $X_1, ..., X_n$ uma amostra aleatória da distribuição da variável aleatória $X \sim N(\mu, 1)$. Nesse caso, a função de verossimilhança é dada por

$$L(\mu; \mathbf{x}) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big\}, \tag{23}$$

com $\Theta = \{\mu; -\infty < \mu < \infty\}$. Como

$$l(\mu; \mathbf{x}) = \ln L(\mu; \mathbf{x}) = -n \ln \sqrt{2\pi} - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2, \tag{24}$$

utilizando (21) a equação de verossimilhança para este caso é dada por

$$\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0, \tag{25}$$

logo o estimador de máxima verossimilhança de μ é dado por

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}, \tag{26}$$

verifica-se que (22) é satisfeita pois

$$\frac{d^2 l(\theta; \mathbf{x})}{d\theta^2}\Big|_{\theta=\hat{\theta}} = -n < 0. \tag{27}$$

Até então foi considerado o caso em que a função de verossimilhança depende apenas de um parâmetro. Agora será considerado situações em que a verossimilhança depende de dois ou mais parâmetros: $\boldsymbol{\theta} = (\theta_1, ..., \theta_r)$. Portanto, caso a estimação seja feita para múltiplos parâmetros, a equação (21) deve ser utilizada considerando as derivadas parciais sob cada parâmetro

$$\frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_i} = 0 \tag{28}$$

e as soluções das equações serão obtidas individualmente.

Exemplo 2: Sejam $X_1, ..., X_n$ uma amostra aleatória de $X \sim N(\mu_x, \sigma^2)$ e $Y_1, ..., Y_m$ uma amostra aleatória de $Y \sim N(\mu_y, \sigma^2)$. Nesse caso, $\boldsymbol{\theta} = (\mu_x; \mu_y; \sigma^2)$. Portanto a verossimilhança correspondente à amostra observada é dada por

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}) = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^m \exp\Big\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_x)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mu_y)^2\Big\}, \tag{29}$$

logo

$$l(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}) = -\frac{(n+m)}{2} \ln 2\pi - \frac{(m+n)}{2} \ln \sigma^2 - \sum_{i=1}^{n} \frac{(x_i - \mu_x)^2}{2\sigma^2} - \sum_{i=1}^{m} \frac{(y_i - \mu_y)^2}{2\sigma^2}. \tag{30}$$

Derivando (30) com relação a $\mu_x; \mu_y$ e σ^2, encontramos as equações

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y})}{\partial \mu_x} = \sum_{i=1}^{n} (x_i - \hat{\mu_x}) = 0; \tag{31}$$

$$\frac{\partial l(\boldsymbol{\theta};\mathbf{x},\mathbf{y})}{\partial \mu_y} = \sum_{j=1}^{m}(y_j - \hat{\mu_y}) = 0 \tag{32}$$

e

$$\frac{\partial l(\boldsymbol{\theta};\mathbf{x},\mathbf{y})}{\partial \sigma^2} = -\frac{(m+n)}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\left\{\sum_{i=1}^{n}(x_i - \hat{\mu_x})^2 + \sum_{j=1}^{m}(y_j - \hat{\mu_y})^2\right\} = 0, \tag{33}$$

cuja solução apresenta os estimadores

$$\hat{\mu_x} = \bar{X}, \quad \hat{\mu_y} = \bar{Y} \tag{34}$$

e

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2}{m+n}. \tag{35}$$

Estes foram exemplos simples com soluções analíticas mas na prática surgem problemas cujos estimadores só podem ser obtidos iterativamente.

Outro conceito importante para o próximo tópico a ser abordado é de *estatística suficiente*. Chama-se de *estatística* qualquer função da amostra, por exemplo $X_1 + X_2 + \ldots + X_n$ é uma estatística. Quando uma estatística contém a mesma informação sobre um determinado parâmetro que a amostra completa contém, ela é dita ser suficiente para estimar o parâmetro. Por exemplo, no Exemplo 1 tem-se que para estimar μ não é necessário ter cada observação individualmente, basta saber $\sum_{i=1}^{n} X_i$. Então $\sum_{i=1}^{n} X_i$ é uma estatística suficiente para μ. Uma estatística $t(X_1,\ X_2,\ \ldots,\ X_n)$ é suficiente para θ se e somente se a distribuição condicional de $\mathbf{X}$ dado $t(X_1,\ X_2,\ \ldots,\ X_n)$ não depende de θ. O conceito se extende para fdp multiparamétrica. Um teorema útil sobre estatística suficiente é o da fatoração: Seja $X_1,\ X_2,\ \ldots,\ X_n$ uma amostra aleatória de X. O conjunto de estatísticas $t_1(X_1,\ X_2,\ \ldots,\ X_n)$, $t_2(X_1,\ X_2,\ \ldots,\ X_n)$, $t_r(X_1,\ X_2,\ \ldots,\ X_n)$ é conjuntamente suficiente se e somente se a

$$f(\mathbf{x}|\boldsymbol{\theta}) = g(t_1(\mathbf{x}),\ t_2(\mathbf{x}),\ \ldots,\ t_r(\mathbf{x})) \times h(\mathbf{x}) = g(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}) \times h(\mathbf{x}),$$

ou seja, a conjunta fatora em duas funções, uma dependente apenas da amostra e a outra dependente apenas das estatísticas suficientes e dos parâmetros. Assim, se

$f(x|\boldsymbol{\theta})$ pertence à família exponencial tem-se que

$$\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = b(\mathbf{x}) \exp(C(\boldsymbol{\theta})\mathbf{t}(\mathbf{x})^T) a(\boldsymbol{\theta}).$$

2.4 Algoritmo EM

O algoritmo EM é visto como uma abordagem geral para a computação iterativa dos estimadores de máxima-verossimilhança quando as observações podem ser vistas como dados incompletos. Dados incompletos podem ser gerados devido a observações ou parte de observações perdidas (*missing*), devido a presença de características ou variáveis não observáveis ou puramente como um artefato para tornar o processo de estimação por máxima verossimilhança mais tratável. O método foi utilizado independentemente por vários autores desde a década de 1950 para resolver problemas específicos, mas colocado como uma alternativa de estimação geral por Dempster et al. (1977). O termo "dados incompletos"geralmente implica na existência de dois espaços amostrais Ω_y e Ω_x. Os dados observados $\mathbf{y}$ são realizações de Y. E a correspondência $\mathbf{x}$ em Ω_x não é observada diretamente, mas apenas indiretamente através de $\mathbf{y}$, através da equação $\mathbf{y} = \mathbf{y}(\mathbf{x})$. Referimos a $\mathbf{x}$ como dados completos.

Seja $f(\mathbf{x}|\boldsymbol{\theta})$ a família de distribuição de X com vetor de parâmetros $\boldsymbol{\theta}$, através da qual pode-se obter sua família de densidades correspondente, considerando os dados observados $g(\mathbf{y}|\boldsymbol{\theta})$. A especificação dos dados completos $f(\mathbf{x}|\boldsymbol{\theta})$ é relacionada com a especificação dos dados incompletos $g(\mathbf{y}|\boldsymbol{\theta})$ por:

$$g(\mathbf{y}|\boldsymbol{\theta}) = \int_{\Omega_{x(y)}} f(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}. \tag{36}$$

em que $\Omega_{x(y)}$ é um subespaço de Ω_x. O algoritmo EM tem como objetivo encontrar $\boldsymbol{\theta}$ que maximixa $g(\mathbf{y}|\boldsymbol{\theta})$ fixando $\mathbf{y}$ nos valores observados. A busca se torna mais fácil através da relação entre $f(\mathbf{x}|\boldsymbol{\theta})$ e $g(\mathbf{y}|\boldsymbol{\theta})$. Em muitas situações o procedimento é iterativo pois não há solução analítica. Cada iteração do algoritmo EM envolve duas etapas: a etapa da obtenção da esperança (E-*step*) e a etapa de maximização (M-*step*).

A etapa E-*step* obtém a estatística suficiente para os dados completos $\mathbf{x}$, dado os dados observados $\mathbf{y}$ e a etapa M-*step* utiliza a estimativa dos dados completos e estima $\boldsymbol{\theta}$ por máxima verossimilhança como se a estimativa dos dados completos fossem os dados observados.

Para definir o algoritmo EM, primeiramente suponha que $f(\mathbf{x}|\boldsymbol{\theta})$ pertence a uma família exponencial regular:

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{b(\mathbf{x})\exp(\boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T)}{a(\boldsymbol{\theta})}, \tag{37}$$

onde $\boldsymbol{\theta}$ é o vetor de parâmetros naturais, $\mathbf{t}(\mathbf{x})$ é o vetor da estatística suficiente dos dados completos e o sobrescrito T representa a matriz transposta. Como a integral de $f(\mathbf{x}|\boldsymbol{\theta})$ em Ω_x é 1 tem-se que

$$a(\boldsymbol{\theta}) = \int b(\mathbf{x})\exp(\boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T)dx, \tag{38}$$

Suponha que $\boldsymbol{\theta}^{(k)}$, onde k é o número de iterações realizadas, denota o valor atual de $\boldsymbol{\theta}$ depois de k ciclos do algoritmo. O próximo ciclo compreende duas etapas:

1. E-*step*: Calcula a esperança do log da verossimilhança dos dados completos, dado os dados observados e uma estimativa de $\boldsymbol{\theta}$. Para a família exponencial isto se reduz a esperança da estatística suficiente, $\mathbf{t}(\mathbf{x})$, dos dados completos, ou seja,
$$\mathbf{t}^{(k)}(\mathbf{x}) = E(\mathbf{t}(\mathbf{x})|\mathbf{y}, \boldsymbol{\theta}^{(k)}). \tag{39}$$

2. M-*step*: Determina $\boldsymbol{\theta}^{(k+1)}$ como uma solução das equações
$$E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}) = \mathbf{t}^{(k)}(\mathbf{x}). \tag{40}$$

As equações (40) são uma forma familiar de equações para estimação de máxima-verossimilhança dado os dados de uma família exponencial regular. Essas equações usualmente definem um estimador de máxima-verossimilhança de $\boldsymbol{\theta}$ (Dempster et al., 1977). Pode-se mostrar que para a família exponencial,

$$\ln f(\mathbf{x}|\boldsymbol{\theta}) = -\ln a(\boldsymbol{\theta}) + \boldsymbol{\theta}\mathbf{t}(\mathbf{x}) + \ln b(\mathbf{x})$$

e a condição que a maximiza é $\mathbf{t}(\mathbf{x}) = E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta})$, ou seja, a condição em (40).

Agora, segue uma breve explicação de como a repetição das etapas E-*step* e M-*step* induzem ao valor $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ que maximiza

$$l(\boldsymbol{\theta}) = \ln g(\mathbf{y}|\boldsymbol{\theta}). \tag{41}$$

Pela definição de densidade condicional temos a seguinte relação entre a densidade de $\mathbf{y}$, a densidade dos dados completos e a densidade condicional para $\mathbf{x}$ dado $\mathbf{y}$ e $\boldsymbol{\theta}$:

$$g(\mathbf{y}|\boldsymbol{\theta}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})}{k(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})}. \tag{42}$$

Então a equação (41) pode ser escrita como

$$l(\boldsymbol{\theta}) = \ln f(\mathbf{x}|\boldsymbol{\theta}) - \ln k(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}). \tag{43}$$

Para a família exponencial, nota-se que:

$$k(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) = b(\mathbf{x}) \exp(\boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T)/a(\boldsymbol{\theta}|\mathbf{y}) \tag{44}$$

em que

$$a(\boldsymbol{\theta}|\mathbf{y}) = \int_{\Omega_{x(y)}} b(\mathbf{x}) \exp(\boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T) d\mathbf{x}, \tag{45}$$

ou seja, a distribuição de $\mathbf{x}$ condicional em $\mathbf{y}$ também é da família exponencial com os mesmos parâmetros naturais, mas com espaço amostral em $\Omega_{x(y)}$.

A equação (43) pode ser desenvolvida, resultando em:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \ln b(\mathbf{x}) + \boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T - \ln(a|\boldsymbol{\theta}) - (\ln(b(\mathbf{x})) + \boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T - \ln(a(\boldsymbol{\theta}|\mathbf{y}))) && (46) \\ &= -\ln a(\boldsymbol{\theta}) + \ln a(\boldsymbol{\theta}|\mathbf{y}). && (47) \end{aligned}$$

Por diferenciação paralela de (38) e (45) obtem-se:

$$D \ln a(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta}) \ln a(\boldsymbol{\theta}) = E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}), \tag{48}$$

desde que

$$\frac{\partial \ln a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{a(\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\Omega_x} b(\mathbf{x}) \exp(\boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T) d\mathbf{x} \tag{49}$$

$$= \frac{1}{a(\boldsymbol{\theta})} \int_{\Omega_x} \frac{\partial}{\partial \boldsymbol{\theta}} b(\mathbf{x}) \exp(\boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T) d\mathbf{x} \quad (50)$$

$$= \frac{1}{a(\boldsymbol{\theta})} \int_{\Omega_x} b(\mathbf{x}) \exp(\boldsymbol{\theta}\mathbf{t}(\mathbf{x})^T) \mathbf{t}(\mathbf{x}) d\mathbf{x} \quad (51)$$

$$= \int_{\Omega_x} \mathbf{t}(\mathbf{x}) f(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \quad (52)$$

$$= E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}). \quad (53)$$

Similarmente,

$$D \ln a(\boldsymbol{\theta}|\mathbf{y}) = E(\mathbf{t}(\mathbf{x})|\mathbf{y}, \boldsymbol{\theta}), \quad (54)$$

consequentemente

$$Dl(\boldsymbol{\theta}) = -E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}) + E(\mathbf{t}(\mathbf{x})|\mathbf{y}, \boldsymbol{\theta}), \quad (55)$$

Então a derivada do log da verossimilhança tem uma representação atrativa como a diferença das esperanças incondicional e condicional da estatística suficiente. A equação (55) é a chave para a compreensão dos passos E-*step* e M-*step* do algoritmo EM e de que a repetição destes passos converge para um $\hat{\boldsymbol{\theta}}$, que maximiza a verossimilhança, pois fazendo $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, $Dl(\boldsymbol{\theta}) = 0$.

A metodologia acima foi descrita especificamente para famílias exponenciais. Dempster et al. (1977) generalizaram-na para qualquer tipo de distribuição através da introdução da função

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = E(\ln f(\mathbf{x}|\boldsymbol{\theta}')|\mathbf{y}, \boldsymbol{\theta}). \quad (56)$$

Cada ciclo do algoritmo envolve:

E-*step*: Cálculo de $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$

M-*step*: Busca de $\boldsymbol{\theta}^{k+1}$ que maximiza $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$.

O objetivo é maximizar $\ln f(\mathbf{x}|\boldsymbol{\theta})$ mas como $f(\mathbf{x}|\boldsymbol{\theta})$ é desconhecida, maximiza-se sua esperança condicionada aos dados observados $\mathbf{y}$ e a estimativa $\boldsymbol{\theta}^k$.

Pode-se mostrar que para a família exponencial

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = -\ln a(\boldsymbol{\theta}) + \boldsymbol{\theta}\mathbf{t}^{(k)}(\mathbf{x})^T + E(b(\mathbf{x})\mathbf{y}, \boldsymbol{\theta}^k) \tag{57}$$

tal que maximizar $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ é equivalente a maximizar $-\ln(a(\boldsymbol{\theta})) + \boldsymbol{\theta}\mathbf{t}^{(k)}(\mathbf{x})^T$ cuja condição que o faz esta colocada em (40).

2.4.1 Estimação de Misturas de Gaussianas

Nesta seção será descrito o algoritmo EM para estimação de misturas de gaussianas.

Seja $\mathbf{y} = (y_1,\ y_2,\ ...,\ y_n)$ uma amostra de observações independentes de uma mistura de J distribuições normais univariadas, cada uma com média μ_j e variância σ_j^2, e $\mathbf{z} = (z_1,\ z_2,\ ...,\ z_n)$ as variáveis classificatórias que determinam qual componente da mistura gerou cada observação. Assim, $\mathbf{x} = (y,\ z)$ forma o vetor de dados completos, com fdp dada por $p(x) = p(z)f(y|z)$. Desta forma, tem-se que a densidade de probabilidade marginal de Y é expressa como:

$$p(y|\boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j f(y|\mu_j, \sigma_j^2) \tag{58}$$

em que $\boldsymbol{\pi} = (\pi_1,\ \pi_2,\ ...,\ \pi_j)$ corresponde ao vetor de proporções desconhecidas das distribuições normais na mistura, de forma que $\sum_{j=1}^{J} \pi_j = 1$, $f(y|\boldsymbol{\theta})$ é função de densidade de probabilidade da distribuição normal e $\boldsymbol{\theta}$ é o vetor dos parâmetros desconhecidos da mistura $\boldsymbol{\theta}_j = (\pi_j,\ \mu_j,\ \sigma_j^2)$.

O ln da verossimilhança de $\mathbf{y}$ é dado por

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} \ln \sum_{j=1}^{J} \pi_j f(y_i|\mu_j, \sigma_j^2), \tag{59}$$

que é difícil de ser maximizado devido ao ln de uma soma envolvido.

Porém a distribuição conjunta de $x_i = (y_i,\ z_i),\ i = 1,\ 2,\ \ldots,\ n$, em que z_i é a categoria (componente) a qual y_i pertence, pode ser escrita como

$$f(x_i|\boldsymbol{\theta}) = I(z_i = j)\pi_j f(y_i|z_i, \mu_j, \sigma_j^2) \tag{60}$$

em que $I(z_i = j) = 1$ se $z_i = j$ e zero caso contrário e $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_J)$. A verossimilhança para os dados completos é dada por:

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} \pi_{z_i} f(y_i|z_i; \mu_{z_i}, \sigma^2_{z_i}). \tag{61}$$

O ln de $L(\boldsymbol{\theta}; \mathbf{x})$ é

$$l(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \Bigg(\ln \pi_{z_i} + \ln f(y_i|z_i; \mu_{z_i}, \sigma^2_{z_i}) \Bigg). \tag{62}$$

Substituindo a $f(y_i|z_i, \mu_{z_i}, \sigma_{z_i})$ em (62) tem-se:

$$l(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \Big[\ln(\pi_{z_i}) - \frac{1}{2} \Big(\ln \sigma^2_{z_i} - \frac{(y_i - \mu_{z_i})^2}{\sigma^2_{z_i}} - \ln 2\pi \Big) \Big]. \tag{63}$$

Então, para misturas de gaussianas a etapa da obtenção da esperança (E-*step*) e a etapa de maximização (M-*step*) do algoritmo EM são:

E-*step*:

Dada a estimativa atual dos parâmetros em $\boldsymbol{\theta}^k$, em que k corresponde ao número de iterações do algoritmo, a distribuição condicional de Z_i, pelo Teorema de Bayes, é dada por:

$$T^k_{j,i} = P(Z_i = j|Y_i = y_i; \boldsymbol{\theta}^k) = \frac{\pi^k_j f_j(y_i|\boldsymbol{\theta}^k)}{\sum_{j=1}^{J} \pi^k_j f_j(y_i|\boldsymbol{\theta}^k)}, \tag{64}$$

que é conhecida como a proporção penalizada da densidade normal.

Desta forma, a expressão resultante no E-*step* é:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = E[\ln(L(\boldsymbol{\theta}; \mathbf{x}))] = \sum_{i=1}^{n} \sum_{j=1}^{J} T^k_{j,i} \left[\ln \pi_j - \frac{1}{2} \Big(\ln \sigma^2_j - \frac{(y_i - \mu_j)^2}{\sigma^2_j} - \ln 2\pi \Big) \right]. \tag{65}$$

M-*step*:

Considerando k iterações, na etapa de maximização é necessário a obtenção do argumento que maximiza $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$, ou seja:

$$\left(\pi_j^{(k+1)}, \sigma_j^{2(k+1)}, \mu_j^{(k+1)}\right) = \text{argmax}_{\boldsymbol{\theta}}\left\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^k)\right\}. \tag{66}$$

Sendo assim, tem-se que:

$$\pi_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} T_{j,i}^{(k)} \tag{67}$$

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} y_i T_{j,i}^{(k)}}{\sum_{i=1}^{n} T_{j,i}^{(k)}} \tag{68}$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{i=1}^{n} T_{j,i}^{(k)}\left(y_i - \mu_j^{(k+1)}\right)^2}{\sum_{i=1}^{n} T_{j,i}^{(k)}} \tag{69}$$

As expressões (67), (69) e (68) fornecem as estimativas dos parâmetros desconhecidos de uma mistura de gaussianas baseada na maximização da função de verossimilhança (Pawitan, 2001).

3 MATERIAL E MÉTODOS

Os métodos de estimação de densidades foram aplicados a diversos conjuntos de dados. Para os métodos mais simples como estimador *naive* e do núcleo foi escrita funções no R para os cálculos e gráficos. Para a estimação de misturas de gaussianas pelo método EM foi utilizado o pacote do MATLAB (2009). Por questões de limitação de espaço serão apresentados apenas os resultados da estimação de densidade para uma imagem de tomografia computadorizada da região pulmonar disponível em InVesalius (2009), usando o algoritmo EM. Após a estimação da densidade, a imagem foi segmentada. Como a imagem é formada por pixels cujos tons variam de acordo com as estruturas, assume-se que cada estrutura gera tonalidades de acordo com uma gaussiana. A segmentação de imagens consiste de duas etapas. Na primeira etapa o algoritmo EM é usado para estimar os parâmetros da mistura. Na segunda etapa os diferentes tons de cinza da imagem são indexados por distintas gaussianas.

O número de regiões de interesse a ser segmentada na imagem é considerado ser igual à quantidade de Gaussianas da mistura que é pré-especificado a priori. Desta forma, a imagem segmentada é composta de tantos tons de cinza quanto o número de Gaussianas usadas na mistura. Isto ocorre porque é possível fazer uma transição imediata entre a representação de imagem utilizando um modelo de mistura gaussiana e a segmentação de imagem probabilística.

A correspondência direta pode ser feita entre a representação de mistura e a imagem plana. Cada pixel da imagem original é associado com a gaussiana mais provável. Uma vez estimado o vetor $\boldsymbol{\theta}$, o processo de segmentação implementado foi baseado na máxima densidade de probabilidade de cada tom de cinza pertencer a

uma dada gaussiana, ou seja, a indexação de cada pixel foi feito da seguinte maneira.

$$Ind(y) = \text{argmax}_j f(y|\boldsymbol{\theta}_j). \tag{70}$$

4 RESULTADOS E DISCUSSÃO

A imagem de tomografia computadorizada referente à região pulmonar utilizada para ilustrar a técnica de estimação de misturas gaussianas e segmentação está representada na Figura 1.

Figura 1: Imagem de TC referente à região pulmonar.

Utilizando a imagem original, Figura 1, foi aplicado o algoritmo EM e realizada a segmentação considerando números distintos de gaussianas para compor a mistura. Os dados constituem uma matriz de 512×512 de valores entre 0 e 255. Inicialmente, o primeiro teste considerou a imagem como sendo composta por 4 gaussianas. As Figuras 2 e 3 representam, respectivamente, a imagem segmentada e a densidade estimada para mistura de 4 gaussianas. Como resultado, temos que a região dos pulmões foi bem destacada.

Figura 2: Imagem da região pulmonar segmentada utilizando 4 gaussianas.

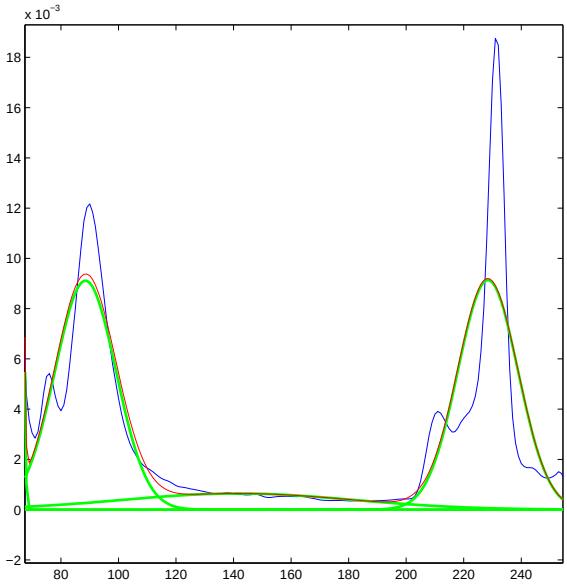

Figura 3: Densidade estimada da região pulmonar segmentada utilizando 4 gaussianas. Curvas em verde representam as densidades das componentes da mistura; a curva em vermelho se refere à soma das densidades e a curva em azul representa a estimativa suavizada pelo estimador de núcleo.

O segundo teste apresentado aqui considerou 7 gaussianas, cujos resultados são apresentados nas figuras 4 e 5. A imagem mostra um pouco mais de detalhes das estruturas, porém como a matriz de dados inclui a área em preto ao redor e externa do corpo, decidimos recortar apenas os pulmões com o objetivo de tentar ressaltar suas estruturas. Os resultados considerando apenas os pulmões e 10 gaussianas estão apresentados nas figuras 6 e 7, que fornecem mais detalhes das estruturas dos pulmões.

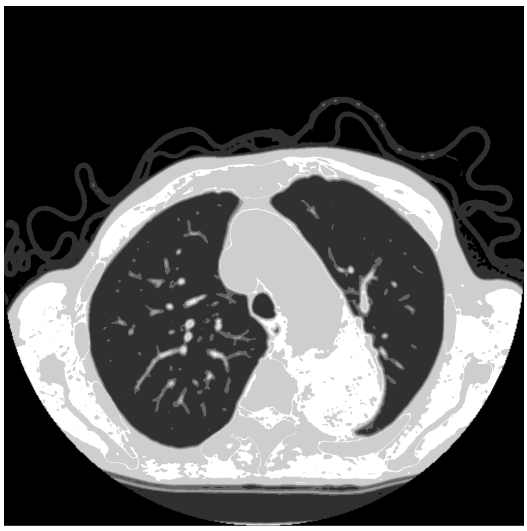

Figura 4: Imagem da região pulmonar segmentada utilizando 7 gaussianas.

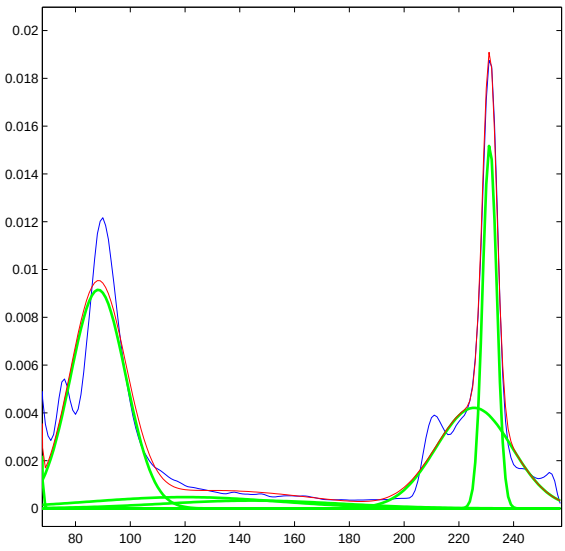

Figura 5: Densidade estimada da região pulmonar segmentada utilizando 7 gaussianas. Curvas em verde representam as densidades das componentes da mistura; a curva em vermelho se refere à soma das densidades e a curva em azul representa a estimativa suavizada pelo estimador do núcleo.

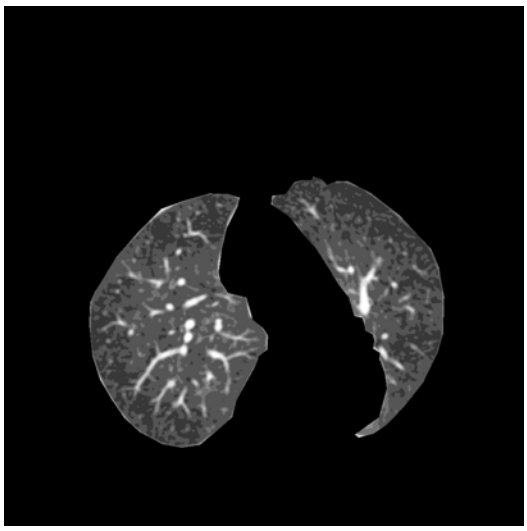

Figura 6: Imagem dos pulmões segmentados utilizando 10 gaussianas.

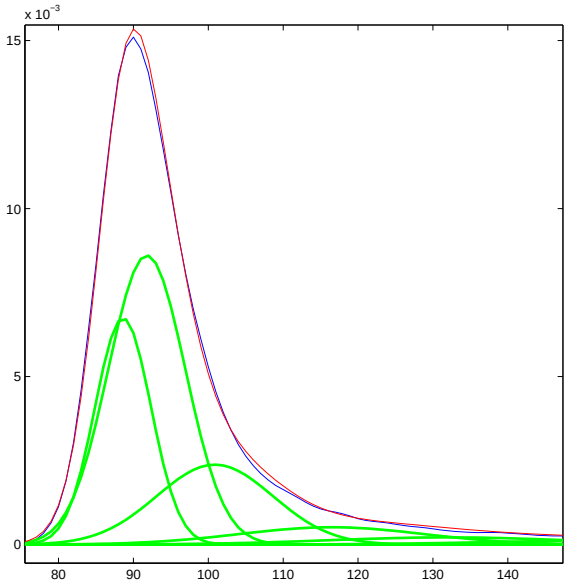

Figura 7: Densidade estimada dos pulmões segmentados utilizando 10 gaussianas. Curvas em verde representam as densidades das componentes da mistura; a curva em vermelho se refere à soma das densidades e a curva em azul representa a estimativa suavizada pelo estimador do núcleo.

Analisando a Figura 2 juntamente com seu histograma pode ser observado que utilizando um número baixo de gaussianas foi possível destacar os pulmões, já na Figura 4 com um número um pouco maior de gaussianas maiores detalhes da imagem foram ressaltados. Para a Figura 6, que apenas os pulmões foram segmentados, a estimação utilizou 10 gaussianas, fornecendo alto detalhe das estruturas dos pulmões, melhorando assim a visualização de detalhes na imagem em questão.

A escolha do número de gaussianas a ser utilizado irá depender totalmente dos objetivos a serem alcançados com a segmentação da imagem, assim como da supervisão de um especialista. Se desejar obter maiores detalhes da imagem ou distinguir pequenas estruturas será necessário um número razoavelmente alto de gaussianas para compor a mistura. Já, se o objetivo for a localização e identificação de certas estruturas extensas (como órgãos, tecidos) possivelmente poderá ser utilizado um número baixo de gaussianas.

Portanto, dependendo do interesse da aplicação da segmentação utiliza-se um número específico de gaussianas e a imagem segmentada poderá auxiliar no diagnóstico médico, pois resulta em imagem mais clara para possíveis análises, detecções e avaliações.

5 CONCLUSÕES

A metodologia aplicada na imagem de TC resultou em imagens devidamente segmentadas. Foi mostrado que com diferentes números de gaussianas pode-se obter resultados diferentes e o que irá determinar a escolha do número de gaussianas é o interesse que se tem na imagem a ser segmentada.

Visando detectar diferentes tipos de estruturas de interesse em imagens, a aplicação da técnica apresentada é de grande utilidade. A imagem de TC é formada a partir da diferença significativa de contraste entre as estruturas presentes na corpo humano. Assim, a distribuição dos tons de cinza da imagem pode ser vista como uma mistura de gaussianas, sendo que cada gaussiana (ou um conjunto de gaussianas) representam uma estrutura ou tecido biológico. Dessa forma, a aplicação da técnica de segmentação baseada na estimação de misturas de gaussianas resulta em uma imagem mais nítida, com características mais evidentes.

Vale ressaltar que para uso prático na área médica é indispensável a análise de especialista na área para que o processo de segmentação seja direcionado de acordo com a finalidade do uso da imagem.

REFERÊNCIAS BIBLIOGRÁFICAS

AMIRA, A.; CHANDRASEKARAN, S.; MONTGOMERY, D. W. G.; SERVAN UZUN, I. A segmentation concept for positron emission tomography imaging using multiresolution analysis. **Neurocomput.**, v.71, n.10-12, p.1954–1965, 2008.

BLECHSCHMIDT, R.; WERTHSCHÜTZKY, R.; LÖRCHER, U. Automated CT image evaluation of the lung: a morphology-based concept. In: , 2001. **IEEE Trans Med Imaging**; resumos. , 2001. 434–442.

BOLFARINE, H. **Introdução a inferência estatística**. Rio de Janeiro: Sociedade Brasileira de Matemática, 2001.

BROWN, M.; MCNITT-GRAY, M.; MANKOVICH, N.; GOLDIN, J.; ABERLE, J. H. L. W. D. Method for segmenting chest CT image data using an anatomical model: preliminary results. In: , 1997. **IEEE Trans Med Imaging**; resumos. , 1997. 828–839.

DAJNOWIEC, M.; ALIREZAIE, J. Computer simulation for segmentation of lung nodules in CT images. In: , 2004. **International Conference on Systems, Man and Cybernetics**; resumos. , 2004.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **J Roy Statist Soc Ser B**, v.39, n.1, p.1–38, 1977.

HUANG, Z.; CHAU, K. A new image thresholding method based on Gaussian mixture model. **Applied Mathematics and Computation**, v.205, p.899–907, 2008.

INVESALIUS. Portal do Software Público Brasileiro, 2009. URL: http://www.softwarepublico.gov.br; [Online, acessado em 18 de janeiro de 2009].

JIANG, S.; CHEN, W.; FENG, Q.; CHEN, Z. Automatic Segmentation of Cerebral Computerized Tomography Based on Parameter-Limited Gaussian Mixture Model. In: , 2007. **Bioinformatics and Biomedical Engineering, ICBBE**; resumos. , 2007. 644-647.

JOLLY, M. Automatic Segmentation of the Left Ventricle in Cardiac MR and CT Images. **International Journal of Computer Vision**, v.70, p.151–163, 2006.

LEE, Y.; HARA, T.; FUJITA, H.; ITOH, S.; ISHIGAKI, T. Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique.. In: , 2001. **IEEE Trans Med Imaging**; resumos. , 2001. 595–604.

LI, X.; LIANG, Z.; ZHANG, P.; KUTCHER, G. J. Accurate colon residue detection algorithm with partial volume segmentation. In: , 2004. **Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series**; resumos. , 2004. 1419-1426.

MATLAB. EMseg, 2009. URL: http://www.mathworks.com/matlabcentral/fileexchange; [Online, acessado em 13 de outubro de 2009].

MONTGOMERY, D. W. G.; AMIRA, A.; ZAIDI, H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. **Medical Physics**, v.34, p.722–736, 2007.

PARK, W.; ND M SONKA, E. H. Segmentation of intrathoracic airway trees: a fuzzy logic approach. In: , 1998. **IEEE Trans Med Imaging**; resumos. , 1998. 489–497.

PAWITAN, Y. **In All Likelihood**. USA: Oxford University Press, 2001.

SILVERMAN, B. W. **Density Estimation for Statistics and Data Analysis**. Chapman & Hall/CRC, 1986.

TRAN, J. Bayesian Segmentation of Dermatological Images using Mixture Models and Markov Random Fields. online, Bachelor of Software Engineering Honours, Clayton Campus, 2005. "URL: http://www.csse.monash.edu.au/hons/seprojects/2005/Ji.Yong.Tran/litreview.pdf; [Online, acessado em 21 de agosto de 2009]".