O uso de ondaletas para ANOVA funcional

Airton Kist, Aluisio de Souza Pinheiro
Departamento de Estatística - Universidade Estadual de Campinas
13083-859 Campinas - SP, Brasil,

1 Introdução

Sofisticados equipamentos de detecção e monitorização são rotineiramente utilizados em pesquisas na medicina, sismologia, meteorologia, fisiologia, mercado financeiro e muitos outros campos. Isso permite aos pesquisadores coletar um grande número de observações em particular ao longo do tempo, que pode ser modelado, essencialmente, como curvas contínuas (funções).

A análise de variância (ANOVA) é uma ferramenta amplamente utilizada nas aplicações estatísticas. Apesar de muito útil para a manipulação de dados de baixa dimensão, tem suas limitações em análise de respostas funcionais. A análise de variância funcional (FANOVA) fornece alternativas aos métodos clássicos de ANOVA e ainda permite uma interpretação simples. Devido ao grande conjunto de aplicações que envolvem dados funcionais, nos últimos anos os modelos FANOVA tornaram-se populares o que tem aumentado a literatura a respeito do assunto. Ramsay and Silverman (2002) e Ramsay and Silverman (2005) apresentam técnicas estatísticas para lidar com dados funcionais. Um dos novos desafios decorrentes da análise estatística de dados funcionais é a comparação entre curvas ou conjuntos de curvas. Fan and Lin (1998) propôs testes para (testar a hipótese funcional) comparação de dois grupos de curvas baseado em testes Neyman adaptativos e procedimentos de limiarização ondaleta de Fan (1996) aplicados a coeficientes de Fourier e ondaletas empíricos dos dados, respectivamente. Abramovich et al. (2004) aplica a estrutura do teste de hipótese funcional minimax assintoticamente originado por Ingster (1982) para testes FANOVA de efeitos fixos. Em particular, adapta os procedimentos de teste baseados em ondaletas de Spokoiny (1996) para testar um sinal zero em um modelo "sinal + ruído branco"e mostra sua otimalidade assintótica (no sentido minimax) para testes em modelos FANOVA de efeitos fixos para uma ampla classe de alternativas. Abramovich and Angelini (2006) estende os testes ótimos de Abramovich et al. (2004) para modelos FANOVA de efeitos mistos e derivara testes de taxa ótima corrrespondentes.

Estamos interessados em estudar modelos em que as observações são caminhos amostrais de um processo estocástico conduzido por

$$dY_i(t) = f_i(t)dt + \epsilon dX_i(t), \quad t \in [0, T],$$

ou da sua versão discreta

$$dY_i(t) = f_i(t)dt + \epsilon dX_i(t), \quad t = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, r. \tag{1.1}$$

em que ϵ é um parâmetro de difusão, r é um inteiro positivo finito, $f_i(\cdot)$ são funções desconhecidas de $\mathbb{R} \to \mathbb{R}$ e $\{X_i(t) : t \in \mathbb{R}\}$ representam processos estocásticos independentes, para $i = 1, \ldots, r$.

2 Teste de média normal em alta dimensão

Fan (1996) estudou o modelo em que $\mathbf{X} \sim \mathrm{N}(\theta, \mathrm{I}_n)$ é um vetor aleatório normal n-dimensional. O objetivo é testar

$$\mathrm{H}_0 : \theta = 0 \quad \text{versus} \quad \mathrm{H}_1 : \theta \neq 0. \tag{2.1}$$

A estatística do teste da razão de máxima verossimilhança do problema (2.1) é $\|\mathbf{X}\|^2$, que testa todas as componentes de $\mathbf{X}$. Este teste tem poder aproximadamente de

$$1 - \Phi\left(\frac{z_{1-\alpha} - \|\theta_1\|^2/\sqrt{2n}}{\sqrt{1 + 2\|\theta_1\|^2/n}}\right) \approx 1 - \Phi(z_{1-\alpha} - \|\theta_1\|^2/\sqrt{2n}) \tag{2.2}$$

na alternativa $\theta = \theta_1$, desde que $\|\theta_1\|^2 = o(n)$, em que α é o nível de significância e $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. O poder tende α mesmo quando $\|\theta_1\| \to \infty$ com $\|\theta_1\|^2 = o(\sqrt{n})$. Mas como a dimensionalidade é alta, testando todas as dimensões, acumula-se erro estocástico o que deteriora o desempenho (2.2) do procedimento de teste. O preço está refletido no fator $1/\sqrt{n}$ do lado direito de (2.2).

Portanto o teste de Neyman

$$\mathrm{H}_0 : \theta = 0 \quad \text{versus} \quad \mathrm{H}_1 : \theta = \theta_1.$$

apresenta problemas de dimensionalidade.

Então Fan (1996) propôs um procedimento de teste Neyman adaptativo truncado para os primeiros m dos n testes com a seguinte estatística do teste

$$T^*_{AN} = \max_{1 \leq m \leq n} \left\{ \frac{1}{\sqrt{2m}} \sum_{j=1}^{m} (X_j^2 - 1) \right\}.$$

Este teste é equivalente a rejeitar H_0 quando

$$T_{AN} = \sqrt{2 \log\log n} T^*_{AN} - (2 \log\log n + 0,5 \log\log\log n - 0,5 \log(4\pi))$$

é muito grande.

A distribuição assintótica de T_{AN} sob H_0 é dada por

$$P(T_{AN} \leq x) \to \exp(-\exp(-x)),$$

mas salienta também que a convergência é muito lenta.

Fan and Lin (1998) propôs um teste de comparação de múltiplos grupos de curvas. Considerou o modelo

$$Y_{ij}(t) = f_i(t) + \epsilon_{i,j}(t), \quad t \in [0, 1],$$

em que $j = 1, \ldots, n_i,\ i = 1, \ldots, I$ e t é discretamente observado em T pontos igualmente espaçados. Além disso, os $\epsilon' s$ são estacionários com média zero. O interesse é testar

$$\mathrm{H}_0 : f_i(t) = f(t), \quad i = 1, \ldots, I,\ t \in [0, 1].$$

Este problema pode ser tratado de uma forma semelhante ao problema de análise de variância de alta dimensinalidade (HANOVA).

A solução assintótica para o problema HANOVA é o mesmo se os $\epsilon' s$ são independentemente distribuídos ou simplesmente estacionários. A aplicação de testes clássicos irá resultar em procedimento de baixo poder, dando erros acumulados devido à configuração de alta dimensionalidade.

De modo análogo como em Fan (1996), Fan and Lin (1998) propõe um procedimento de teste Neyman adaptativo truncado com a seguinte estatística de teste

$$F_{\hat{m}} = \max_{1\leq m\leq T^*} \frac{1}{\sqrt{2(I-1)m}} \times \left\{ \sum_{k=1}^{m} \sum_{i=1}^{I} n_i \sigma_i^*(k)^{-2} (\overline{Y}_i^*(k) - \overline{Y}^*(k))^2 - (I-1)m \right\}$$

em que $\overline{Y}_i^*(k)$ é a curva média do i-ésimo grupo no domínio frequência, i.e.,

$$\overline{Y}_i^*(k) = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}^*(k),$$

sendo $Y_{ij}^*(\cdot)$ a transformada de Fourier discreta das observações Y_{ij}, e

$$\sigma_i^{*2}(k) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij}^*(k) - \overline{Y}_i^*(k))^2.$$

A versão normalizada é dada por [1]

$$T_{HANOVA} = \sqrt{2 \log\log T^*} F_{\hat{m}}^* - (2 \log\log T^* + 0,5 \log\log\log T^* - 0,5 \log(4\pi))$$

Uma tabela de p valores simulada (baseada em 1000000 de simulações) é fornecida por Fan and Lin (1998). Questões de poder são também abordadas. Também mostrou a convergência em distribuição da estatística do teste para a distribuição extrema III, mas salienta também que a convergência é lenta.

3 Modelo de análise de variância em espaços funcionais

O problema de estimação funcional vem sendo estudado de formas variadas na literatura. Uma possibilidade bastante promissora se dá pela utilização de bases ortonormais de wavelets (ondaletas). Essa solução é interessante por sua: frugalidade; otimalidade assintótica; e velocidade computacional.

Um modelo muito estudado é aquele em que as observações (dados) são caminhos amostrais de um processo estocástico conduzido por

$$dY_i(t) = f_i(t)dt + \epsilon dX_i(t), \quad t \in [0, T], \tag{3.1}$$

ou da sua versão discreta

$$dY_i(t) = f_i(t)dt + \epsilon dX_i(t), \quad t = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, r. \tag{3.2}$$

em que ϵ é um parâmetro de difusão, r é um inteiro positivo finito, $f_i(\cdot)$ são funções desconhecidas de $\mathbb{R} \to \mathbb{R}$ e $\{X_i(t) : t \in \mathbb{R}\}$ representam processos estocásticos independentes, para $i = 1, \ldots, r$.

Conforme Pinheiro and Vidakovic (2009), a utilidade potencial dos modelos mais simples, como o (3.2), com $X_i(t) = W_i(t)$, pode ser justificada pela teoria da equivalência assintótica (no sentido de Le Cam (1986)), desenvolvida por, entre outros Brown and Low (1996) e Nussbaum (1996). Nesta configuração Klemelä (2006) mostra que métodos ondaletas adaptativos produzirão resultados

[1] T^* é o número máximo de dimensões a serem testadas

minimax ótimos para funcionais quadráticos no sentido de Donoho and Johnstone (1998) e Donoho and Johnstone (1999).

Deseja-se testar a hipótese nula

$$\mathrm{H}_0 : f \equiv 0$$

contra a alternativa não paramétrica composta que a função f está longe de zero em norma L^2, $\|f\| \geq \rho(\epsilon)$.

O problema está em descrever a taxa mínima (ótima) para a distância $\rho(\epsilon)$ para que o teste com probabilidades de erro prevista seja ainda possível. O resultado depende muito do tipo de suposições de suavidade que são impostas para a função f. Na realidade quer-se testar a hipótese nula $\mathrm{H}_0 : f \equiv 0$ contra uma classe de alternativas tão grande quanto possível.

Quando em (3.1)

$$X_i(t) = W_i(t),$$

em que $\{W(t) : t \in \mathbb{R}\}$ representa um movimento Browniano, existem muitos trabalhos nesse sentido. Por exemplo, Ingster (1982), Ingster (1984), Ingster (1993), Lepski and Spokoiny (1999),Spokoiny (1996), Abramovich et al. (2004), Abramovich and Angelini (2006), entre outros, assumiram somente que f obedece algumas condições de suavidade. Mais precisamente, a função f é assumida pertencer a alguma bola de Besov $B^s_{p,q}(C)$.

4 Modelo FANOVA de efeitos fixos

Abramovich et al. (2004) estudou o modelo em que os caminhos amostrais do processo estocástico são impulsionados por

$$dY_i(t) = f_i(t)dt + \epsilon dW_i(t), \quad i = 1, 2, \ldots, r; \quad t \in [0, 1], \tag{4.1}$$

em que $\epsilon > 0$ é o parâmetro de difusão, r é um inteiro positivo finito, $f_i(\cdot)$ são funções desconhecidas de $\mathbb{R} \to \mathbb{R}$ e $\{W_i(t) : t \in \mathbb{R}\}$ representam movimentos Brownianos padrão independentes, para $i = 1, 2, \ldots, r$.

Abramovich et al. (2004) propõe que as observações dos caminhos amostrais de (4.1) sejam estudadas pela decomposição a seguir (única com probabilidade um)

$$f_i(t) = m_0 + \mu(t) + a_i + \gamma_i(t), \quad i = 1, \ldots, r; \quad t \in [0, 1] \tag{4.2}$$

em que: m_0 é a média geral (constante); $\mu(t)$ representa o efeito principal em t; a_i representa o efeito principal em i; $\gamma_i(t)$ representa o efeito da interação entre i e t.

As componentes da decomposição (4.2) satisfazem as seguintes condições de identificabilidade:

$$\int_0^1 \mu(t)dt = 0; \quad \sum_{i=1}^r a_i = 0;$$
$$\sum_{i=1}^r \gamma_i(t) = 0; \quad \int_0^1 \gamma_i(t)dt = 0, \quad \forall i = 1. \ldots, r; \quad t \in [0, 1].$$

Algumas hipóteses de interesse são

$$\mathrm{H}_0 : \mu(t) \equiv 0, \quad t \in [0, 1]; \tag{4.3}$$
$$\mathrm{H}_0 : a_i \equiv 0, \quad i = 1, \ldots, r; \text{ e} \tag{4.4}$$
$$\mathrm{H}_0 : \gamma_i(t) \equiv 0, \quad i = 1, \ldots, r \quad t \in [0, 1]. \tag{4.5}$$

A hipótese dada em (4.4) equivale às clássicas hipóteses de análise de variância e pode ser tratada por técnicas usuais. Por outro lado as hipóteses (4.3) e (4.5) são funcionais em essência, e suas alternativas devem ser formuladas cuidadosamente para que testes consistentes possam ser realizados.

Nenhuma forma paramétrica é especificada para $\mu(t)$ e $\gamma_i(t)$ sob a hipótese alternativa e deseja-se testar a hipótese nula contra uma classe de alternativas tão grande quanto possível. Em particular assume-se que $f_i(t)$ (e portanto, $\mu(t)$ e $\gamma_i(t)$), $i = 1, \ldots, r$ pertencem a uma mesma bola de Besov $B^s_{p,q}(C)$ (com raio $C > 0$ em $[0, 1]$, $s > 0$ e $1 \leq p, q \leq \infty$). Então considera-se a seguinte sequência de alternativas, respectivamente:

$$\mathrm{H}_1 : \mu \in \mathcal{F}(\rho); \tag{4.6}$$

$$\mathrm{H}_1 : \gamma_i \in \mathcal{F}(\rho), \quad \text{para pelo menos um } i = 1, \ldots, r, \tag{4.7}$$

em que $\mathcal{F}(\rho) = \{f \in B^s_{p,q}(C) : \|f\|_2 \geq \rho\}$, sendo $\|\cdot\|_2$ a norma $L^2([0, 1])$.

Em Abramovich et al. (2004) são apresentados procedimentos de testes funcionais não adaptativos e adaptativos baseados em ondaletas.

Nos testes não adaptativos supõe-se que todos os parâmetros da bola de Besov, s, p, q e o raio C são conhecidos, enquanto que nos procedimentos adaptativos os testes independem do conhecimento prévio dos parâmetros s, p, q e C, o que acontece em geral nas aplicações práticas.

Ambos os testes derivados são ótimos assintoticamente do ponto de vista minimax, eles diferem a menos de um fator $\log \log$ para testar as hipóteses nulas (4.3) e (4.5) contra as sequências de alternativas (4.6) e (4.7).

5 Resultados

Neste trabalho estendemos os resultados de Abramovich et al. (2004) para modelos FANOVA de efeitos fixos para o caso em que o ruído segue um processo de Ornstein-Uhlenbeck. Ou seja, para o modelo de efeitos fixos com erro autoregressivo a tempo contínuo de primeira ordem (CAR(1)).

Abramovich et al. (2004) estima as funções $f_i(\cdot)$ de equações do tipo

$$dY_i(t) = f_i(t)dt + \epsilon dX_i(t), \quad t \in [0, T],$$

quando temos r caminhos amostrais observados, em que: ϵ é um parâmetro de difusão, r é um inteiro positivo finito, $f_i(\cdot)$ são funções desconhecidas de $\mathbb{R} \to \mathbb{R}$ e $\{X_i(t) : t \in \mathbb{R}\}$ representam processos estocásticos independentes, para $i = 1, \ldots, r$, com $X_i = W_i$. Adaptamos os estimadores baseados em ondaletas propostos em Abramovich et al. (2004) para estudar o comportamento para amostras finitas dos testes ótimos minimax para amostras não iid. Esse estudo foi feito para o caso em que o ruido $X_i(t)$ é autocorrelacionado, definido como um processo autoregressivo a tempo contínuo de primeira ordem.

Referências

Abramovich, F. and Angelini, C. (2006). Testing in mixed-effects FANOVA models. *J. Statist. Plann. Inference*, 136(12):4326–4348.

Abramovich, F., Antoniadis, A., Sapatinas, T., and Vidakovic, B. (2004). Optimal testing in a fixed-effects functional analysis of variance model. *Int. J. Wavelets Multiresolut. Inf. Process.*, 2(4):323–349.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398.

Donoho, D. and Johnstone, I. (1999). Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica*, 9:1–32.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921.

Fan, D. (1996). The uniformly most powerful unbiased test of the normal mean. *Gongcheng Shuxue Xuebao*, 13(1):58–66.

Fan, J. and Lin, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.*, 93(443):1007–1021.

Ingster, Y. I. (1982). Minimax nonparametric detection of signals in white gaussian noise. *Problems Inform. Transmission*, 18(2):130–140.

Ingster, Y. I. (1984). Asymptotically minimax testing of nonparametric hypotheses on the density of the distribution of an independent sample. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 136:74–96. Studies in mathematical statistics, VI.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Math. Methods Statist.*, 2(4):249–268.

Klemelä, J. (2006). Sharp adaptive estimation of quadratic functionals. *Probab. Theory Related Fields*, 134(4):539–564.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

Lepski, O. V. and Spokoiny, V. G. (1999). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Annals of Statistics*, 25(4):2399–2430.

Pinheiro, A. and Vidakovic, B. (2009). *Wavelets in Functional Data Analysis*. 13a Escola de Séries Temporais e Econometria, São Carlos.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York. Methods and case studies.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6):2477–2498.