Estatística Espacial (MI418) / Geoestatística (ME907)

Guilherme Ludwig

2019-01-21

Cubic splines

Thin-plate splines

A conexão entre Kriging e thin-plate splines

Splines

Em estatística não paramétrica, um método para resolver problemas de estimação de curvas suaves em $f : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}$, a partir de uma coleção finita de locais $\mathbf{s}_1, \ldots, \mathbf{s}_n$ e valores $y_1, \ldots, y_n$ envolve *splines* (para uma visão geral, veja Wahba, 1990b).

Numa formulação abstrata: *Splines* são funções definidas em $\mathbb{R}^d \to \mathbb{R}$ que resolvem um problema variacional da forma

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{s}_i))^2 + \lambda \|\mathcal{P} f\|_{\mathcal{H}},$$

em que $\lambda > 0$ é uma constante de penalização, $\mathcal{H}$ é uma classe de funções e $\|\mathcal{P} f\|_{\mathcal{H}}$ algum termo de regularização (a teoria por trás deles é bastante complicada!).

Splines

Para dar um exemplo mais concreto, suponha que $d = 1$, e $\mathcal{H} = \mathcal{W}_2^2$ é um espaço de funções $f : \mathbb{R} \to \mathbb{R}$, duas vezes diferenciáveis, com $\|\frac{\partial^2}{\partial t^2} f\|_{\mathcal{H}}$ finita. Então o spline (cúbico) é a solução do problema

$$\min_{f \in \mathcal{W}_2^2} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_0^\infty \left(\frac{\partial^2}{\partial t^2} f(t) \right)^2 \mathrm{d}t,$$

Neste caso, para λ fixo, a solução do problema variacional tem forma fechada, dada por

$$f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \sum_{i=1}^{n} \theta_i (t - t_i)_+^3,$$

em que $(\cdot)_+$ é a função identidade truncada em 0.

Splines

Os parâmetros α_0, α_1, α_2, $\theta_1, \ldots, \theta_n$ podem ser determinados também pelos dados. Basta usar a forma fechada de f no problema variacional para obter

$$\min_{\boldsymbol{\alpha},\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \alpha_0 - \alpha_1 t_i - \alpha_2 t_i^2 - \sum_{j=1}^{n} \theta_i (t_i - t_j)_+^3 \right)^2 + \lambda \boldsymbol{\theta}^t \boldsymbol{\Omega} \boldsymbol{\theta}$$

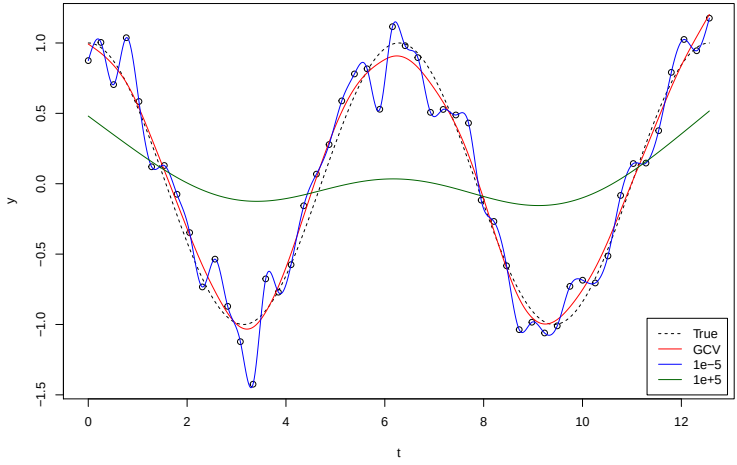
Note que $\boldsymbol{\Omega}$ é uma matriz condicionalmente positiva semi-definida, mas também com forma fechada (integral de polinômios). O parâmetro de regularização λ pode ser obtido através de validação cruzada.

Importante: até agora, não falamos de erros no modelo...

Splines

```
library(mgcv)
t <- seq(0, 4*pi, l = 50)
t0 <- seq(0, 4*pi, l = 1000)
set.seed(1)
y <- cos(t) + rnorm(50, 0, .2)
model <- gam(y ~ s(t, k = 50, bs = "cr"))
f <- predict(model, data.frame(t = t0))
model2 <- gam(y ~ s(t, bs = "cr", k = 50), sp = 1e-5)
f2 <- predict(model2, data.frame(t = t0))
model3 <- gam(y ~ s(t, bs = "cr", k = 50), sp = 1e5)
f3 <- predict(model3, data.frame(t = t0))
```

Splines



Thin-plate splines

Thin-plate splines são uma extensão do spline cúbico no $\mathbb{R}^d$. Suponha que $d = 2$, e $\mathcal{H}$ é um espaço de funções $f : \mathbb{R}^2 \to \mathbb{R}$, duas vezes parcialmente diferenciáveis. Para a escolha de penalização do tipo

$$J[f] = \int_{\mathbb{R}^2} \left(\frac{\partial^2}{\partial s_x^2} f(\mathbf{s}) \right)^2 + 2 \left(\frac{\partial^2}{\partial s_x s_y} f(\mathbf{s}) \right)^2 + \left(\frac{\partial^2}{\partial s_y^2} f(\mathbf{s}) \right)^2 \mathrm{d}\mathbf{s},$$

o thin-plate spline é a solução do problema

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{s}_i))^2 + \lambda J[f],$$

Neste caso, para λ fixo, a solução do problema variacional tem forma fechada, dada por

$$f(\mathbf{s}) = \alpha_0 + \alpha_1 s_x + \alpha_2 s_y + \sum_{i=1}^{n} \theta_i \varphi_i(\mathbf{s} - \mathbf{s}_i),$$

em que $\varphi_i(\mathbf{s}) = \|\mathbf{s} - \mathbf{s}_i\|^2 \log \|\mathbf{s} - \mathbf{s}_i\|$ é uma função base radial.

Thin-plate splines

Na prática, escreva

$$\mathbf{f} = \mathbf{T}\boldsymbol{\alpha} + \boldsymbol{\Phi}\boldsymbol{\theta},$$

então

$$\hat{\mathbf{f}}_\lambda = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{T}\boldsymbol{\alpha} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 + \lambda\boldsymbol{\theta}^t\boldsymbol{\Omega}_{\mathrm{TPS}}\boldsymbol{\theta},$$

sujeito a $\mathbf{T}^t\boldsymbol{\theta} = \mathbf{0}$. Mas uma propriedade interessante dos *thin-plate splines* é que

$$\boldsymbol{\Omega}_{\mathrm{TPS}} = \begin{pmatrix} \varphi(\|\mathbf{s}_1 - \mathbf{s}_1\|) & \cdots & \varphi(\|\mathbf{s}_1 - \mathbf{s}_n\|) \\ \vdots & \ddots & \vdots \\ \varphi(\|\mathbf{s}_n - \mathbf{s}_1\|) & \cdots & \varphi(\|\mathbf{s}_n - \mathbf{s}_n\|) \end{pmatrix}.$$

Isto é, a solução é, ajeitando alguns termos, da forma

$$\mathbf{h}_T\mathbf{y} + \varphi^t(\boldsymbol{\Phi}^t\boldsymbol{\Phi} + n\lambda\boldsymbol{\Phi})^{-1}(\mathbf{y} - \mathbf{H}_T\mathbf{y})$$

(fica como exercício descrever $\mathbf{H}_T, \mathbf{h}_T$).

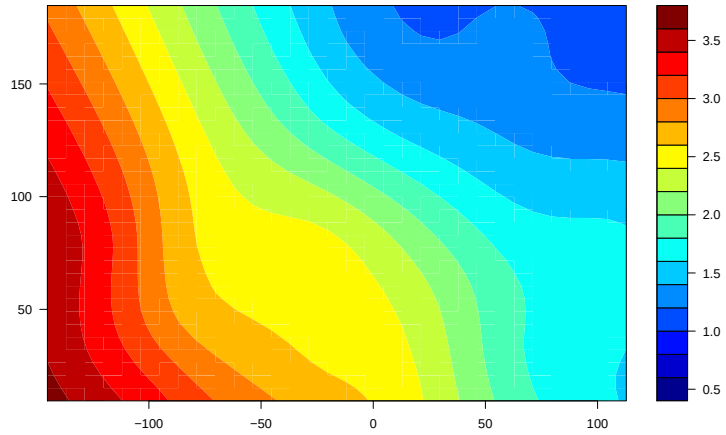
Kriging e thin-plate splines

```
library(mgcv)
library(geoR)
library(fields) # filled.contour
wolfcamp <- read.csv("wolfcamp.csv", skip = 1)
modelTPS <- gam(Data ~ s(X, Y, bs = "tp"),
                data = wolfcamp) # Thin-plate Spline
```

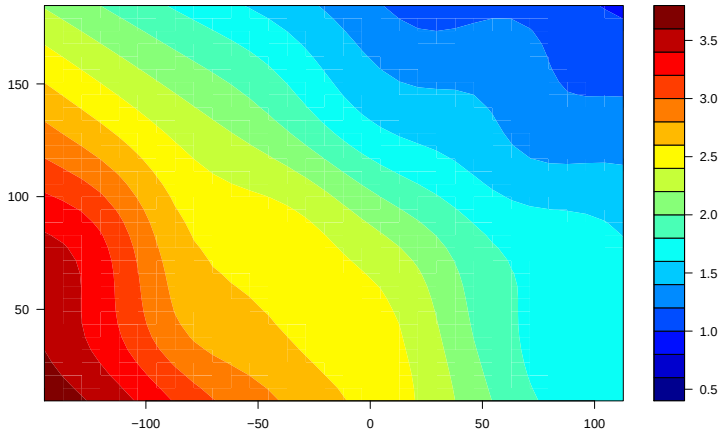
Vou também refazer o ajuste de Universal Kriging...

```
## krige.conv: model with mean given by a 2nd order polynomial on the coor
## krige.conv: Kriging performed using global neighbourhood
```

Kriging e thin-plate splines: TPS



Kriging e thin-plate splines: UK



Conexão entre Kriging e thin-plate spline

Splines

- Modelam efeitos de primeira ordem (funções fixas, desconhecidas).
- Em geral assumem erros independepentes (splines tendem a interpolar erros dependentes).
- Teoria depende de análise funcional
- Em geral bastante versátil (e.g. adequa-se bem a GLM), mas achar λ é difícil

Kriging

- Modelam efeitos de segunda ordem (funções aleatórias, desconhecidas).
- Teoria depende de probabilidade
- Em geral ótimo para processos Gaussianos (dá certo trabalho estender pro caso não-Normal)

Na prática: dão resultados coerentes entre si quase sempre. Veja Cressie (1989), Wahba (1990a), Cressie (1990) e Altman (2000).

Referências I

Altman, N. (2000). Theory & methods: krige, smooth, both or neither? *Australian & New Zealand Journal of Statistics*, 42(4):441–461.

Cressie, N. (1989). Geostatistics. *The American Statistician*, 43(4):197–202.

Cressie, N. (1990). Reply to letter by G. Wahba. *The American Statistician*, 44(3):256–258.

Wahba, G. (1990a). Comment on Cressie. *The American Statistician*, 44:255–256.

Wahba, G. (1990b). *Spline Models for Observational Data*. SIAM, Philadelphia.