GeoRDD manuscript

Maxime Rischard

December 5, 2016

Contents

1 Introduction

1.1 Motivation

1.2 Prior attempts

2 Model Specification

2.1 Notation

- 2-dimensional coordinate space $\mathscr{S}$
- treatment units are in region $\mathscr{S}_T \subset \mathscr{S}$ and control units are in non-overlapping $\mathscr{S}_C$ outside of the treatment region, so that $\mathscr{S}_C = \mathscr{S}_T^c$ and $\mathscr{S}_T \cup \mathscr{S}_C = \mathscr{S}$

- Observed outcomes for units in treatment region $s \in \mathscr{S}_T$ are labeled $Y_T(\mathbf{s})$, and units in control region $Y_C(\mathbf{s})$.
- Potential outcomes framework: Each unit has a potential outcome under treatment $Y_T(\mathbf{s})$ and a potential outcome under control $Y_C(\mathbf{s})$. If $s \in \mathscr{S}_T$, then $Y_T(\mathbf{s})$ is observed, otherwise $Y_C(\mathbf{s})$ is observed.

2.2 1GP solution

Most straightforwardly, we model the observed outcomes Y at locations S as the sum of an intercept μ, linear trend $S\beta$, a spatial Gaussian process $f(S)$, a constant treatment effect τ in the treatment region, and iid normal noise ϵ.

$$Y_i(\mathbf{s}) = \mu + \mathbf{s}^\intercal \beta + f(\mathbf{s}) + \tau \mathbb{I}\left\{\mathbf{s} \in \mathscr{S}_T\right\} + \epsilon_i \tag{1}$$

$$f(S) \sim \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \tag{2}$$

$$k(\mathbf{s}, \mathbf{s}') = \sigma_{\mathrm{GP}}^2 \exp\left(-\frac{(\mathbf{s}-\mathbf{s}')^\intercal (\mathbf{s}-\mathbf{s}')}{2\ell^2}\right) \tag{3}$$

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right) \tag{4}$$

$f(S)$ is a smooth surface covering all of $\mathscr{S}$, specificed as a Gaussian Process with squared exponential covariance kernel k with lengthscale ℓ and variance σ_{GP}^2. The squared exponential kernel is frequently used in spatial settings. The constant treatment effect implies the assumption that $Y_T(\mathbf{s}) = \tau + Y_C(\mathbf{s})$ for all units at all locations.

2.3 2GP solution

The constant treatment effect is a strong assumption that will be hard to justify in many applications. To allow the treatment effect to vary spatially, an alternative is to specify two independent Gaussian processes for the treatment response and the control response.

$$Y_{T,i}(\mathbf{s}) = \underbrace{\mu_T + \mathbf{s}^\intercal \beta_T + f_T(\mathbf{s})}_{g_T(\mathbf{s})} + \epsilon_i \tag{5}$$

$$Y_{C,i}(\mathbf{s}) = \underbrace{\mu_C + \mathbf{s}^\intercal \beta_C + f_C(\mathbf{s})}_{g_C(\mathbf{s})} + \epsilon_i \tag{6}$$

$$f_T(S), f_C(S) \overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \tag{7}$$

$$k(\mathbf{s}, \mathbf{s}') = \sigma_{\mathrm{GP}}^2 \exp\left(-\frac{(\mathbf{s}-\mathbf{s}')^\intercal (\mathbf{s}-\mathbf{s}')}{2\ell^2}\right) \tag{8}$$

$$\tag{9}$$

Here, the treatment effect τ is no longer included explicitly in the model. Instead, the treatment effect at a location s is derived as the difference between the two (noise-free) surfaces g_T and g_C.

$$\tau(\mathbf{s}) = \left[\mu_T + \mathbf{s}^\intercal \beta_T + f_T(\mathbf{s})\right] - \left[\mu_C + \mathbf{s}^\intercal \beta_C + f_C(\mathbf{s})\right]$$

In this specification, the kernel parameters ℓ and σ_{GP} are the same in the treatment and control regions, so we assume that the spatial smoothness of the responses isn't affected by the treatment. This assumption will be reasonable in most applications, but can be easily relaxed. Inference on the hyperparameters proceeds as in the 1GP case, using the sum of the likelihood in the treatment and control regions.

2.4 Discussion

- different assumptions
- will stick to 2GP from now on

3 Inference

By specifying the spatial variation as Gaussian processes, we can leverage the properties of multivariate normals to obtain analytical forms for the estimate of the treatment effect.

3.1 1GP

We proceed by placing normal priors on μ, β and τ. The model specification can then be used to obtain covariances between the observations and these parameters. In fact, $(Y, f(S), \tau, \mu, \beta) \mid \ell, \sigma_{\mathrm{GP}}$ is multivariate normal with variance-covariance given by

$$
\begin{aligned}
\tau &\sim \mathcal{N}\left(0, \sigma_\tau^2\right) && (10)\\
\mu &\sim \mathcal{N}\left(0, \sigma_\mu^2\right) && (11)\\
\beta &\sim \mathcal{N}\left(0, \sigma_\beta^2\right) && (12)\\
\operatorname{cov}(Y_i(\mathbf{s}), \tau) &= \sigma_\tau^2 \, \mathbb{I}\left\{\mathbf{s} \in \mathscr{S}_T\right\} && (13)\\
\operatorname{cov}(Y_i(\mathbf{s}), \mu) &= \sigma_\mu^2 && (14)\\
\operatorname{cov}(Y_i(\mathbf{s}), \beta) &= \sigma_\beta^2 \mathbf{s}^\intercal \mathbf{s} && (15)\\
\operatorname{cov}(Y_i(\mathbf{s}), Y_i(\mathbf{s}')) &= \sigma_\mu^2 + \sigma_\tau^2 \, \mathbb{I}\left\{\mathbf{s} \in \mathscr{S}_T\right\} \mathbb{I}\left\{\mathbf{s}' \in \mathscr{S}_T\right\} + \sigma_\beta^2 \mathbf{s}^\intercal \mathbf{s}' + k(\mathbf{s}, \mathbf{s}') + \delta_{ij}\sigma_\epsilon^2 && (16)\\
\operatorname{cov}(Y(\mathbf{s}), f(\mathbf{s}')) = \operatorname{cov}(f(\mathbf{s}), f(\mathbf{s}')) &= k(\mathbf{s}, \mathbf{s}') && (17)
\end{aligned}
$$

Multi-variate theory then allows us to condition any of these objects on the others. We are particularly interested in the posterior distribution $\tau \mid Y, \ell, \sigma_{\mathrm{GP}}$ which is given by

$$
\tau \mid Y, \ell, \sigma_{\mathrm{GP}} \sim \mathcal{N}\left(\operatorname{cov}\left(Y, \tau\right)^\intercal \operatorname{cov}\left(Y\right)^{-1} Y, \sigma_\tau^2 - \operatorname{cov}\left(Y, \tau\right)^\intercal \operatorname{cov}\left(Y\right)^{-1} \operatorname{cov}\left(Y, \tau\right)\right) \tag{18}
$$

To proceed computationally, we define the treatment indicator vector $\mathbb{I}_T$ with ith entry equal to 0 when $\mathbf{s}_i$ is in the control region, and 1 in the treatment region, and the $n \times n$ kernel covariance matrix $\mathbf{K}$ having entries $\mathbf{K}_{ij} = k(\mathbf{s}_i, \mathbf{s}_j)$. The posterior mean and variance are then easily computed.

$$
\begin{aligned}
\mathbb{E}\left(\tau \mid Y, \ell, \sigma_{\mathrm{GP}}, \sigma_\epsilon\right) &= \sigma_\tau^2 \, \mathbb{I}_T^\intercal \left\{\sigma_\mu^2 + \sigma_\tau^2 \, \mathbb{I}_T \, \mathbb{I}_T^\intercal + \sigma_\beta^2 S S^\intercal + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right\}^{-1} Y && (19)\\
\operatorname{var}\left(\tau \mid Y, \ell, \sigma_{\mathrm{GP}}, \sigma_\epsilon\right) &= \sigma_\tau^2 - \sigma_\tau^2 \, \mathbb{I}_T^\intercal \left\{\sigma_\mu^2 + \sigma_\tau^2 \, \mathbb{I}_T \, \mathbb{I}_T^\intercal + \sigma_\beta^2 S S^\intercal + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right\}^{-1} \mathbb{I}_T && (20)
\end{aligned}
$$

What remains is the inference on the hyperparameters $\sigma_\epsilon, \sigma_{\mathrm{GP}}$ and ℓ. The two approaches typically taken in modern spatial statistics are either to maximize the marginal likelihood of Y as a function of those three parameters, or to assign them a prior and take a Bayesian approach, requiring that the posterior of τ be integrated over those parameters. The compromise is clear: the Bayesian approach incorporates the uncertainty in the hyperparameters, thus giving more reliable inference on τ, but maximizing the marginal likelihood has a much lower computation cost. Therefore, we recommend taking the Bayesian approach whenever computationally possible, and maximizing the marginal likelihood when the data is larger.

3.2 2GP

In the 2GP setting, we begin by modeling the treatment and control units with two independent Gaussian processes with shared hyperparameters. Because the treatment and control regions do not overlap, inference on the treatment effect is only measurable near the boundary. In the classical one-dimensional regression discontinuity design, the estimand is therefore defined at the boundary $x = b$:

$$
\tau = \lim_{x \downarrow b} \mathbb{E}\left[y \mid X = s\right] - \lim_{x \uparrow b} \mathbb{E}\left[y \mid X = x\right] = \mathbb{E}\left[Y_T \mid X = b\right] - \mathbb{E}\left[Y_C \mid X = b\right] \tag{21}
$$

Analogously, we focus on the treatment effect at the boundary ∂ between the treatment and control regions. ∂ is therefore a one-dimensional subset of $\mathscr{S}$. We will proceed by extrapolating both Gaussian processes to the boundary, and then subtracting the predictions to obtain the estimated treatment effect. Computationally, we need to represent this boundary as a set of k "sentinel" units distributed along the boundary $\boldsymbol{\partial} = \left\{\partial_1, \ldots, \partial_k\right\}, \ \partial_i \in \partial$. The extrapolation step then proceeds mechanically through multivariate-normal theory.

$$
\begin{aligned}
g_T(\boldsymbol{\partial} \mid Y_T, S_T, \ell, \sigma_{\mathrm{GP}}, \sigma_\epsilon) &\sim \mathcal{N}\left(\mu_{\boldsymbol{\partial}|T}, \Sigma_{\boldsymbol{\partial}|T}\right) && (22)\\
\mu_{\boldsymbol{\partial}|T} &\equiv \operatorname{cov}\left(g_T(\boldsymbol{\partial}), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} Y_T && (23)\\
\Sigma_{\boldsymbol{\partial}|T} &\equiv \operatorname{cov}\left(g_T(\boldsymbol{\partial})\right) - \operatorname{cov}\left(f_T(\boldsymbol{\partial}), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} \operatorname{cov}\left(Y_T, g_T(\boldsymbol{\partial})\right) && (24)\\
& && (25)
\end{aligned}
$$

All the covariance terms can be derived from the model similarly to what we saw in the 1GP procedure. Analogously, we also generate predictions for $g_C(\boldsymbol{\partial})$ using the data in the control region, and denote their posterior mean and covariance as $\mu_{\boldsymbol{\partial}|C}$ and $\Sigma_{\boldsymbol{\partial}|C}$. Since the two surfaces are modeled as independent, the treatment effect $\tau(\boldsymbol{\partial}) = g_T(\boldsymbol{\partial}) - g_C(\boldsymbol{\partial})$ along the boundary is also multivariate normal with posterior mean $\mathbb{E}\left(\tau(\boldsymbol{\partial}) \mid Y\right) = \mu_{\boldsymbol{\partial}|T} - \mu_{\boldsymbol{\partial}|C}$ and covariance $\operatorname{cov}\left(\tau(\boldsymbol{\partial}) \mid Y\right) = \Sigma_{\boldsymbol{\partial}|T} + \Sigma_{\boldsymbol{\partial}|C}$.

4 Handling covariates

The Gaussian Process specification makes it easy to incorporate a linear model on non-spatial covariates, both mathematically and computationally. The model is modified by the addition of the linear regression term $D\gamma$ on the $n \times p$ matrix of covariates D. In the spirit of ridge regression, we recommend placing a normal prior $\mathcal{N}(0, \sigma_\gamma^2)$ on the regression coefficients. This preserves the multivariate normality of the problem, with the simple addition of a term $\sigma_\gamma^2 D^\intercal D$ to the covariance of Y.

With the 1GP model, covariates can therefore be handled at very little additional cost, except that the additional hyperparameter σ_γ^2 needs to be fitted.

5 2GP: Testing for non-zero effect

6 Average treatment effect

6.1 Linear average

6.2 Inverse variance

7 Spatial advantage

- spreading units along a boundary doesn't necessarily reduce power
- multiple experiments interpretation

8 Example: NYC school districts

9 Conclusion