Inference in Gaussian Process Models for Political Science*

JBrandon Duck-Mayr

Abstract

Political scientists often seek to perform inference in settings where knowledge about the functional form mapping predictors to outcomes is imperfect or the traditional assumption of conditional independence of observations does not hold. Recently Gaussian process (GP) models, a family of machine learning techniques, have been used to study politics in such settings; however, many inferential quantities of interest to political science have either not been derived in the statistical and machine learning literature the models hail from or have not been employed in the political science literature yet. I provide practical guidance for applied researchers to implement GP models for more accurate inference in their research, including how to obtain quantities of interest to political scientists, the derivations of which are novel to the GP model literature.

*I would like to thank Jacob Montgomery and participants at Washington University in St. Louis' Political Data Science Lab for their helpful comments.

1 Introduction

Gaussian process (GP) models, a class of machine learning techniques, are increasingly being employed to study politics, from measuring ideology (Gill 2020; Duck-Mayr, Garnett, and Montgomery 2020)[1] to dealing with violations of conditional independence in time-series cross-sectional data (Carlson 2020). GP models are powerful tools to model the relationship between predictors and outcomes when the functional form mapping predictors to response is imperfectly known or observations may not be conditionally independent, common settings in political science. However, inference in the machine learning setting often focuses (sometimes exclusively) on prediction, while other inferential quantities are often of interest to political scientists. I show how these models can be used to obtain quantities of interest to political scientists, including a novel derivation of average marginal effects for GP models.[2] I first provide a primer on GP models, explaining their particular import for political science specifically. I highlight their nascent use in political science, briefly explaining existing approaches to inference with GP models in the political science literature. I then outline how to obtain a number of other quantities of interest and provide practical guidance in the use of these models to enable the discipline to better harness these powerful tools for social scientific inference.

2 A Primer on GP Models

Generally we want to reason about the relationship between predictors X and outcomes y. So, we generally say

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \tag{1}$$

[1] Gill (2020) uses "spatial kriging," which is a type of GP model, to extrapolate ideology measures spatially across the U.S., while Duck-Mayr, Garnett, and Montgomery (2020) develop a novel GP item response theoretic model (GPIRT).

[2] The full derivations can be found in the appendix; in the main body of the paper I present results and practical guidance.

where y_i is our observed outcome for observation i, $\mathbf{x}_i$ is our observed vector of predictors for observation i, and ε_i is the error term—some added random noise. Then, we want to learn about $f(X)$. A stereotypical approach in political science is to assume a functional form for f, and that the noise elements ε_i are independently distributed. In that case our task is to perform inference on the parameters of f. A number of non-parametric approaches are available when the form of f is unknown, and a variety of statistical fixes have been developed for various correlation structures of $\boldsymbol{\varepsilon}$.

A non-parametric approach common in the machine learning literature and now starting to gain traction in political science is to model f as a *Gaussian process* (GP) (Rasmussen and Williams 2006).[3] While there are many methods that have been developed to accomplish non-parameteric inference or handle error correlation, GP models have risen to prominence because in addition to their flexibility, they represent a principled, probabilistic approach that presents a very general framework applicable in a variety of settings (Cheng et al. 2019). Moreover, they can outperform even tailored models for many inferential problems; for example, Carlson (2020) finds GP regression to be more effective at handling error correlation in time-series cross-sectional data than other existing approaches.

A GP is an infinite dimensional generalization of the normal distribution, where any finite subcollection of the process' variables are normally distributed. This is accomplished by specifying the mean and covariance of the distribution as a function of the predictors:

$$f \sim \mathcal{GP}\left(\mu\left(X\right), K\left(X, X\right)\right), \tag{2}$$

where μ is a function (such as, for example, $X\beta$ for a vector of coefficients β) that gives the mean of the distribution of f at X and $K\left(X, X'\right)$ is a matrix-valued function that gives the covariance between values of f at X and X', such that for any finite set of observations $\mathbf{X}$, $\mathbf{f} = f\left(\mathbf{X}\right)$ has a prior distribution

[3]Rasmussen and Williams (2006) is a comprehensive textbook for GP classification and regression. A reader seeking a treatment that is much more in-depth should consult Rasmussen and Williams (2006).

$$\mathbf{f} \sim \mathcal{N}\left(\mu\left(\mathbf{X}\right), K\left(\mathbf{X}, \mathbf{X}\right)\right). \tag{3}$$

In the regression case, where y is continuous and we use a normal likelihood with variance σ_y^2, we can then learn about f after observing $\mathbf{X}$ and $\mathbf{y}$ by appyling Bayes' theorem along with Gaussian identities to derive the exact posterior over $\mathbf{f}$,

$$\mathbf{f} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}\left(\mathbf{m}^*, \mathbf{C}^*\right), \tag{4}$$

$$\mathbf{m}^* = \mu\left(\mathbf{X}\right) + KK_y^{-1}\left(\mathbf{y} - \mu\left(\mathbf{X}\right)\right), \tag{5}$$

$$\mathbf{C}^* = K - KK_y^{-1}K, \tag{6}$$

$$K_y = K + \sigma_y^2 I, \tag{7}$$

where we write $K = K\left(X, X\right)$ for more compact notation.[4] Notice, then, that (for example) Bayesian linear regression is simply a special case of GP regression; GP regression with $\mu\left(X\right) = 0$ and $K\left(X\right) = XX^T$ returns the same solution as linear regression with standard normal priors on the coefficients.

This allows us to learn about *potentially* nonlinear functions of X with very mild assumptions. The assumptions we are making about f are largely through our choice of the mean function μ and the covariance function, or *kernel*, K. The common choice in the machine learning literature is to choose $\mu\left(X\right) = 0$, giving a vague prior over f where all learning about f goes through the kernel. We may also choose a linear mean, $\mu\left(X\right) = X\beta$, encoding an assumption that f should have a linear trend, though perhaps with nonlinear deviations or correlated errors.

An overwhelmingly popular choice for the kernel is the squared exponential covariance function, in which the covariance between $f\left(\mathbf{x}_i\right)$ and $f\left(\mathbf{x}_j\right)$ is given by

[4] Although this derivation is available in multiple sources in varying levels of detail, including Rasmussen and Williams (2006), I provide a derivation in the appendix as well.

$$k\left(\mathbf{x}_i, \mathbf{x}_i\right) = \sigma_f^2 \left(-0.5 \sum_d \frac{\left(x_{id} - x_{jd}\right)^2}{\ell_d^2}\right), \tag{8}$$

where σ_f^2 is a *scale factor* scaling the entire prior covariance matrix and $\boldsymbol{\ell}$ is a vector of *length scales*. This kernel corresponds to assuming that (1) f is smooth,[5] and (2) the correlation between values of f should decrease with distance in the covariate space. Then $\boldsymbol{\ell}$ determines how we define "closeness" along each dimension of X. Often an isotropic version of this kernel is used where ℓ is instead a scalar, treating distance in every dimension the same. This kernel should similarly be most useful in political science; these assumptions match up to problems where we must account for correlated errors across space or time (Carlson 2020; Gill 2020), and notably is also equivalent to a linear model with infinite basis expansion.

To make this more concrete, consider the following example, where

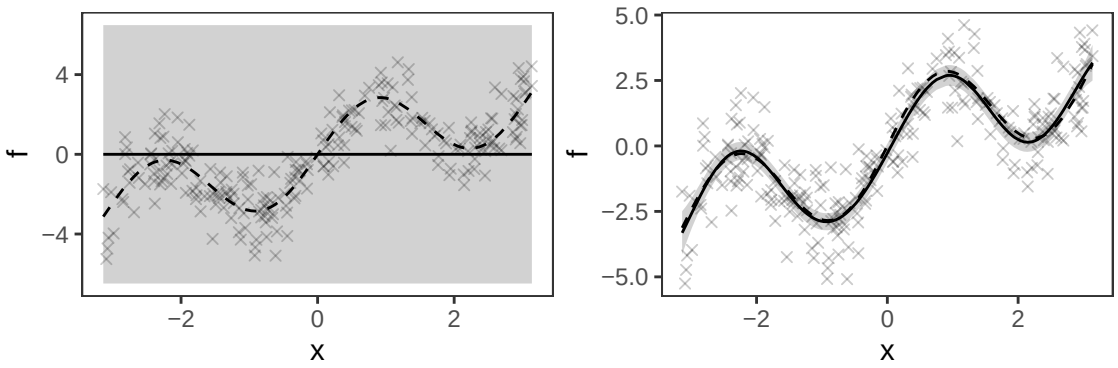
$$f\left(x\right) = 2\sin\left(2x\right) + x, \tag{9}$$

$$y = f\left(x\right) + \varepsilon, \tag{10}$$

$$\varepsilon \sim \mathcal{N}\left(0, 1\right). \tag{11}$$

So, we have a function mapping the single predictor variable x to outcomes y with a linear trend and we have independent noise, but the function also has systematic nonlinear deviations. Figure 1a shows a vague, zero-mean GP prior over f, using the squared exponential covariance function. I simulated 250 x values, drawn from a uniform distribution with bounds $-\pi$ and π (which allows for two full oscillations of f), then simulated corresponding y values with standard normal noise. We can see the posterior from Equation (4) depicted in Figure 1b; essentially we have taken the somewhat mild assumptions encoded by the covariance function that f is smooth and that covariance between function outputs decreases with distance in x to derive a reasonable estimate of f (depicted with the solid line) with a measure of uncertainty

[5]For scholars interested in modeling functions that are not smooth, other kernel options are available with a similar proximity assumption. Please consult Chapter 4 in Rasmussen and Williams (2006) for details.



(a) The GP prior (b) The GP regression posterior

Figure 1: An example GP prior and posterior for the function $f(x) = 2\sin(2x) + x$. Simulated data points are depicted with crosses, the prior (posterior) mean with a solid line and the 95% CI with a shaded region, and the true $f(x)$ with a dashed line.

(depicted with the shaded region).

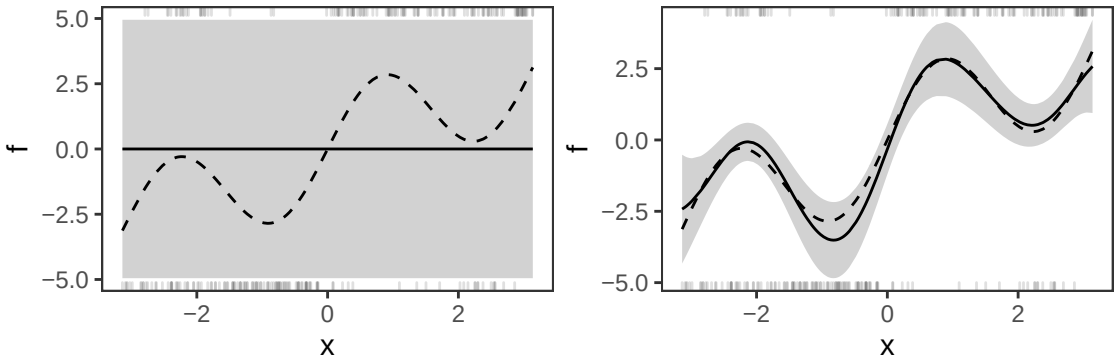
For modeling discrete outcomes, the posterior becomes intractable; however, good approximations of the posterior in important cases have been derived. For example, for dichotomous outcomes, we simply say

$$\Pr(y_i = 1) = \sigma(f(\mathbf{x}_i)), \tag{12}$$

$$f \sim \mathcal{GP}(\mu(X), K(X)), \tag{13}$$

where f is then a latent function fed though σ, which is some sigmoid "squashing" function mapping the reals to $[0, 1]$, such as using a logistic or probit likelihood, to obtain the probability of a postitive response. This gives us a very similar setup to a generalized linear model such as the probit or logit regression, with the difference being that we do not assume as much about the structure of the latent function f. While the posterior does not have a closed form as in the regression case, we can solve for a Laplace approximation to the posterior centered at the posterior mode $\hat{\mathbf{f}}$,

$$\mathbf{f} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}\left(\hat{\mathbf{f}}, \left(K^{-1} + W\right)^{-1}\right), \tag{14}$$



(a) The GP prior (b) Laplace approximation of the posterior

Figure 2: An example GP prior and posterior for the function $f(x) = 2\sin(2x) + x$, where the data were simulated as $x \sim U(-\pi, \pi), \Pr(y = 1) = \sigma(f(x))$, where σ is the logistic function. Observations receiving positive labels are depicted in the rug on the top margin, while observations receiving negative labels are depicted in the rug on the bottom margin; the prior (posterior) mean with a solid line and the 95% CI with a shaded region, and the true $f(x)$ with a dashed line.

$$\hat{\mathbf{f}} = \mu(\mathbf{X}) + K\left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \tag{15}$$

$$W = -\nabla\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right), \tag{16}$$

and other approximations based on minimizing Kullbeck-Liebler divergence are available as well, in addition to being able to simulate the posterior via MCMC sampling, commonly utilizing an elliptical slice sampler (Murray, Adams, and MacKay 2010).

Taking the same example function and simulated x values we used to illustrate GP regression, I draw corresponding dichotomous y values where $\Pr(y = 1) = \sigma(f(x))$, where σ is the logistic function. A zero mean GP prior with squared exponential covariance function is depicted in Figure 2a, and the Laplace approximation of the posterior from Equation (14) is depicted in Figure 2b.

Modelling the relationship between predictors and outcomes as a GP offers a flexible but principled approach with a number of advantages over other approaches. Unlike parametric approaches, we can be agnostic *a priori* as to the shape of the relationship between predictors and outcomes, accounting for our typical uncertainty over functional form as social scientists.

When compared to many non-parametric approaches that are similarly agnostic, the GP approach provides a more principled probabilistic approach that is more readily extended to varied settings. Finally, as I show in Section 4, the GP approach still allows us to recover and make probabilistic statements about inferential quantities of interest to political scientists.

3 GP Models in Political Science

While GP models have a longer history in statistics and machine learning, they are just beginning to take hold in the study of politics. The GP approach is particularly suited to the study of politics as political scientists often confront situations in which we should acknowledge some uncertainty regarding functional form, or (as may be the modal case in political science) the common assumption of independent errors is violated.

Carlson (2020) considers TSCS data, common in many areas of political science, and recommends GP regression for those settings. Carlson (2020) shows GP regression outperforms a variety of previous approaches such as lagged dependent variable, fixed effects, and random effects specifications, as well as panel-corrected standard errors in the TSCS setting. Gill (2020) similarly utilizes a GP model to handle spatial autocorrelation.

Duck-Mayr, Garnett, and Montgomery (2020) develop an IRT model where, rather than assuming the functional form of the response functions, a GP prior is placed over latent response functions fed into a logistic likelihood. Among their applications demonstrating the method, the authors estimate ideology of members of the House of Representatives in the 116th U.S. Congress. They show this more flexible approach that acknowledges uncertainty over the functional form of the roll call votes' response functions allows for more plausible estimates of extremist members' ideology; while parametric methods that impose monotonicity of responses in ideology force extreme members such as Rep. Alexandria Ocasio-Cortez who often vote against the moderate proposals of her own party to be placed closer to the opposing party, a flexible GP approach can recognize that members such as she should be placed at

the end of the ideological spectrum and instead allow the item response function to bend downwards in such situations.

As political scientists are increasingly taking advantage of the flexible GP approach to handle data that pose difficulties for inference using traditional approaches, I next provide practical guidance for applied researchers and derive distributions of inferential quantities of interest to political scientists that are not covered in the machine learning literature where the lion's share of study of GP models has occurred.

4 Tools for Inference in GP Models

A common goal for those employing GP models is out of sample prediction, so the typical inferential quantity of interest is the distribution of the unknown function at test points. The machine learning literature on GP models has largely focused on this sort of inference in various settings. Political scientists are more often interested in parsing out the effects of the predictors. After explaining the usual process for inference in GP models, I will also show how to derive two types of effects of predictors: the distribution of coefficients if a linear mean function is employed, and the distribution of average marginal effects of predictors whether or not a linear mean function is employed. The former is useful for determining the contribution of a predictor to a linear trend within f, while the latter is necessary for uncovering the full average effect of a predictor on outcomes, since we allow for a potentially nonlinear relationship between predictors and outcomes. A convenient interface to obtain the distribution of all of these quantities is provided in an `R` package.[6]

When the goal is prediction, inference for GP models often follows the following sequence: first, make choices about the GP prior, including the structure of the mean and covariance functions; next, set the parameters of the mean and covariance functions (such as the

[6] A number of packages are available for fitting and predicting out of sample for GP models, both in R and in a number of other programming languages and software environments including python and MATLAB. However, no currently available software implements average marginal effects or provides posterior inference over linear mean function coefficients.

coefficients of a linear mean function, or the scale factor of a squared exponential covariance function) as a model selection step by choosing them to use the GP prior that maximizes the log marginal likelihood of the model given the training data; finally, use the training data and the selected hyperparameters to predict out of sample for test data. This workflow is very effective for generating probabilistic predictions for unknown data generating processes. The posterior predictive distribution for GP regression at test cases $\mathbf{X}^*$ is

$$\mathbf{f}^* \mid \mathbf{y}, \mathbf{X} \sim \mathcal{N}\left(\mu\left(\mathbf{X}^*\right) + K_* K_y^{-1}\left(\mathbf{y} - \mu\left(\mathbf{X}\right)\right), K_{**} - K_* K_y^{-1} K_*^T\right), \tag{17}$$

with a common shorthand of $K_* = K\left(\mathbf{X}^*, \mathbf{X}\right)$ and $K_{**} = K\left(\mathbf{X}^*, \mathbf{X}^*\right)$, and the posterior predictive distribution for classification (using the Laplace approximation) is

$$\mathbf{f}^* \mid \mathbf{y}, \mathbf{X} \sim \mathcal{N}\left(\mu\left(\mathbf{X}^*\right) + K_* \nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right), K_{**} - K_*\left(K + W^{-1}\right)^{-1} K_*^T\right). \tag{18}$$

However, often other inferential quantities are of more interest than out of sample predictions. For example, suppose you believe the important underlying relationship between your predictors and outcomes is in fact linear, but you also want to explicitly model and account for correlated errors, as in Carlson (2020). Then you may use a model of the following form:

$$y \sim \mathcal{N}\left(f\left(X\right), \sigma_y^2 I\right), \tag{19}$$

$$f \sim \mathcal{GP}\left(X\beta, K\left(X, X\right)\right), \tag{20}$$

where your quantity of interest is β. Then it would not be useful to treat the mean function parameters as a model selection problem; rather we want to find the distribution of those parameters themselves instead of the distribution of f. One approach would be to place priors on the mean function *and* covariance function parameters to get the posterior distribution

of all the GP prior's hyperparameters; this approach, taken in Carlson (2020), however, generally requires MCMC sampling, as priors on the covariance function hyperparameters generally result in the posterior being analytically intractible.[7]

If the covariance function parameters are not directly of interest, however, we can set those using Bayesian model selection as in the general prediction workflow; then, with covariance function parameters in hand, the posterior distribution of β is

$$\beta \mid \mathbf{y}, \mathbf{X} \sim \mathcal{N}(\bar{\beta}, \Sigma_\beta), \tag{21}$$

$$\bar{\beta} = (B^{-1} + \mathbf{X}^T K_y^{-1} \mathbf{X})^{-1} (B^{-1}\mathbf{b} + \mathbf{X}^T K_y^{-1} \mathbf{y}), \tag{22}$$

$$\Sigma_\beta = (B^{-1} + \mathbf{X}^T K_y^{-1} \mathbf{X})^{-1}, \tag{23}$$

where $\mathbf{b}$ is the prior mean of β and B is the prior covariance of β.[8]

However, if our motivation for using a GP approach is flexibility in the form of f rather than being interested only in posterior inference over a linear trend contained in f, we should instead perform inference on the *average marginal effect* of our predictors. As this is not a task common in the use of GP models in computer science, the machine learning literature on GP models provides no explicit derivation for this quantity, although for other reasons, building blocks we need for it have previously been derived.

To formalize, we want to reason about the relationship between predictors X and outcomes y, and specifically wish to know the marginal effect of a particular predictor d, i.e., the dth feature of X. In a parametric regression model where we assume $f(X) = X\beta$ and simply estimate β, the marginal effect of X_d is easy to see: $\frac{\partial f(\mathbf{x}_i)}{\partial x_{id}} = \hat{\beta}_d, \forall i$. In GP regression and classification, we gain a much more flexible model that allows for non-independence of observations and non-linear mappings from X to y, but then feature d does not have a constant effect on y. We can summarize the effect of feature d on y with the sample average

[7] A bespoke MCMC sampler for GP regression with priors on all hyperparameters is also offered in the R package, which provides samples *much* faster than the general-purpose Stan implementation used for Carlson (2020).

[8] See Rasmussen and Williams (2006) Section 2.7. A full derivation is also offered in the appendix.

marginal effect,[9] which in the regression context is defined as

$$\gamma_d \triangleq \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\, f\left(\mathbf{x}_i\right)}{\partial\, x_{id}}. \tag{24}$$

For classification, γ_d gives us the sample average partial effect, or the sample average effect on the latent function f, which does not directly translate to the average marginal effect on our dichotomous outcomes y. In this case, often of more interest than γ_d is

$$\pi_d \triangleq \frac{1}{N}\sum_{i}^{N}\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, x_{id}}, \tag{25}$$

which gives the average marginal effect of predictor d on the function $\sigma\left(f\left(X\right)\right)$ that gives the probability of a positive response.

Moreover, when d is discrete, instanteous change in f at our observed points is not meaningful; then we want the average discrete change in f at levels of predictor d. For binary variables, let $\mathbf{X}_1^*$ be a set of test points where all feature observactions are identical to $\mathbf{X}$ except that all observations of feature d have been set to 1, and analogously for $\mathbf{X}_0^*$.[10] Then a more appropriate quantity of interest rather than γ_d is

$$\delta_d \triangleq \frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{x}_{1i}^*\right) - f\left(\mathbf{x}_{0i}^*\right), \tag{26}$$

which gives the average marginal effect on y of taking a 1 vs a 0 value in the regression case, or the average partial effect on f of taking a 1 vs a 0 value in the classification case. For classification, the effect on the probability scale is

$$\psi_d \triangleq \frac{1}{N}\sum_{i=1}^{N} \sigma\left(f\left(\mathbf{x}_{1i}^*\right)\right) - \sigma\left(f\left(\mathbf{x}_{0i}^*\right)\right). \tag{27}$$

For categorical variables, we simply find δ_d or ψ_d for all substantively interesting pairwise comparisons of levels of the categorical variable. (Often this is comparing the various

[9] See Hainmueller and Hazlett (2014) for a similar approach with kernel-regularized least squares.

[10] You can replace 1 and 0 with other binary value labels as needed.

categorical labels to one "baseline" label).

Our starting point for deriving these quantities is noting that "[s]ince differentiation is a linear operator, the derivative of a Gaussian process is another Gaussian process" (Rasmussen and Williams 2006, 191). Let

$$\mathbf{f}_d = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_{1d}} \\ \vdots \\ \dfrac{\partial f_n}{\partial x_{nd}} \end{bmatrix}. \tag{28}$$

Using Equation 9.1 in Rasmussen and Williams (2006),

$$K_d \triangleq \mathbb{C}\left[\mathbf{f}_d, \mathbf{f}\right] = \begin{bmatrix} \dfrac{\partial k\left(\mathbf{x}_1, \mathbf{x}_1\right)}{\partial x_{1d}} & \cdots & \dfrac{\partial k\left(\mathbf{x}_1, \mathbf{x}_n\right)}{\partial x_{1d}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial k\left(\mathbf{x}_n, \mathbf{x}_1\right)}{\partial x_{nd}} & \cdots & \dfrac{\partial k\left(\mathbf{x}_n, \mathbf{x}_n\right)}{\partial x_{nd}} \end{bmatrix}, \tag{29}$$

$$K_{dd} \triangleq \mathbb{C}\left[\mathbf{f}_d, \mathbf{f}_d\right] = \begin{bmatrix} \dfrac{\partial^2 k\left(\mathbf{x}_1, \mathbf{x}_1\right)}{\partial x_{1d} \partial x_{1d}} & \cdots & \dfrac{\partial^2 k\left(\mathbf{x}_1, \mathbf{x}_n\right)}{\partial x_{1d} \partial x_{nd}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 k\left(\mathbf{x}_n, \mathbf{x}_1\right)}{\partial x_{nd} \partial x_{1d}} & \cdots & \dfrac{\partial^2 k\left(\mathbf{x}_n, \mathbf{x}_n\right)}{\partial x_{nd} \partial x_{nd}} \end{bmatrix}, \tag{30}$$

To make the notation more compact, as is usual we set $K = K\left(\mathbf{X}, \mathbf{X}\right)$, and additionally set $\mu = \mu\left(\mathbf{X}\right)$ and

$$\mu_d = \begin{bmatrix} \dfrac{\partial \mu\left(\mathbf{x}_1\right)}{\partial x_{1d}} \\ \vdots \\ \dfrac{\partial \mu\left(\mathbf{x}_n\right)}{\partial x_{nd}} \end{bmatrix}. \tag{31}$$

Then we can describe the joint prior on $\mathbf{f}$ and $\mathbf{f}_d$:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_d \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_d \end{bmatrix}, \begin{bmatrix} K & K_d^T \\ K_d & K_{dd} \end{bmatrix} \right), \tag{32}$$

and for regression and under normal approximations of the posterior for classification,[11] the posterior distribution of $\mathbf{f}_d$ given $\mathbf{X}$ and $\mathbf{y}$ is normal with mean

$$\begin{aligned} \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] &= \int \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{f}, \mathbf{X}\right] p\left(\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right) d\mathbf{f} \\ &= \mu_d + K_d\, K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right), \end{aligned} \tag{33}$$

and variance

$$\begin{aligned} \mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] &= K_{dd} - K_d\, K^{-1}\, K_d^T + \mathbb{E}\left[\left(\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{f}\right] - \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right]\right)^2\right] \\ &= K_{dd} - K_d \left(K^{-1} - K^{-1}\, \mathbb{V}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] K^{-1}\right) K_d^T. \end{aligned} \tag{34}$$

(The full derivations for all results in this section are provided in the appendix for the interested reader). In the regression case,

$$\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = \mu_d + K_d\, K_y^{-1} \left(\mathbf{y} - \mu\right), \tag{35}$$

$$\mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = K_{dd} - K_d\, K_y^{-1}\, K_d^T, \tag{36}$$

For classification, under the Laplace approximation to the posterior,

$$\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = \mu_d + K_d \left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \tag{37}$$

$$\mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = K_{dd} - K_d \left(K + W^{-1}\right)^{-1} K_d^T. \tag{38}$$

[11] When simulating the posterior for classification rather than using an analytical approximation, $\mathbf{f}_d$ is normally distributed given each $\mathbf{f}$ draw (with mean $\mu_d + K_d K^{-1}\left(\mathbf{f} - \mu\right)$ and variance $K_{dd} - K_d K^{-1} K_d^T$), so $\mathbf{f}_d$ samples can simply be taken conditioned on the $\mathbf{f}$ samples.

Since then γ_d is a constant ($1/N$) times the sum of correlated normal random variables,

$$\gamma_d \sim \mathcal{N}\left(\frac{1}{N}\sum_{i=1}^{N} m_{\gamma_d i}\,,\ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} c_{\gamma_d ij}\right), \tag{39}$$

$$\mathbf{m}_{\gamma_d} = \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] \tag{40}$$

$$\mathbf{C}_{\gamma_d} = \mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] \tag{41}$$

Importantly, we may also get the average marginal effect of feature d within subgroups of $\mathbf{X}$ rather than the full sample average by simply altering the indices of summation in Equation (39). The distribution of δ_d is analogous:

$$\delta_d \sim \mathcal{N}\left(\frac{1}{N}\sum_{i=1}^{N} m_{\delta_d i}\,,\ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} c_{\delta_d ij}\right), \tag{42}$$

$$\mathbf{m}_{\delta_d} = \mathbb{E}\left[\mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right] - \mathbb{E}\left[\mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right], \tag{43}$$

$$\mathbf{C}_{\delta_d} = \mathbb{V}\left[\mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{V}\left[\mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{C}\left[\mathbf{f}_1^*, \mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{C}\left[\mathbf{f}_0^*, \mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right]. \tag{44}$$

Unfortunately, the distribution π_d cannot be analytically expressed, though we can readily simulate from it. First note that

$$\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, x_{id}} = \frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, f}\frac{\partial\, f\left(\mathbf{x}_i\right)}{\partial\, x_{id}}. \tag{45}$$

Generally the sigmoid function σ has a known derivative; for example, in the logistic case,

$$\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, f} = \sigma\left(f\left(\mathbf{x}_i\right)\right)\left(1 - \sigma\left(f\left(\mathbf{x}_i\right)\right)\right). \tag{46}$$

Since we have an approximation to the posterior on f, and given f, the posterior over $\mathbf{f}_d$ is

$$\mathbf{f}_d \mid \mathbf{f} \sim \mathcal{N}\left(\mu_d + K_d\, K^{-1}\left(\mathbf{f} - \mu\right), K_{dd} - K_d\, K^{-1}\, K_d^T\right), \tag{47}$$

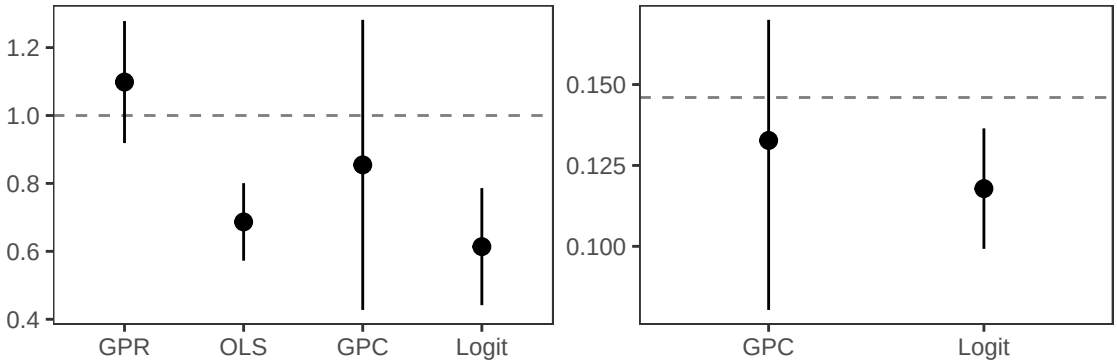
we can obtain M samples of π_d by

- drawing $\mathbf{f}^t$ from the chosen posterior approximation, such as the Laplace approximation $\mathcal{N}\left(\mu + K\left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), (K^{-1} + W)^{-1}\right)$,[12]
- drawing $\mathbf{f}_d^t$ from $\mathcal{N}\left(\mu_d + K_d K^{-1}\left(\mathbf{f}^t - \mu\right), K_{dd} - K_d K^{-1} K_d^T\right)$,
- and calculating $\pi_d^t = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \frac{\partial \sigma\left(f_i^t\right)}{\partial f} f_{id}^t$,

so that we can summarize the distribution of π_d using the M draws, similar to the CLARIFY procedure (King, Tomz, and Wittenberg 2000). We can also similarly simulate the distribution of ψ_d by simply generating values of $f\left(\mathbf{X}_1^*\right)$ and $f\left(\mathbf{X}_0^*\right)$ from the posterior approximation, pushing those samples through the chosen sigmoid σ, and calculate the average of the differences to get the average marginal effect for each sample so that we can summarize the distribution of average marginal effects.

We can illustrate the use of average marginal effects in GP models by returning to our previous example. For the function $f(x) = 2\sin(x) + x$ where $x \sim U(-\pi, \pi)$, the true average marginal effect of x on $f(x)$ is 1—the oscillations cancel out and we are left with the effect of the linear term. In other words, on average, $f(x)$ increases with x at a rate of 1, which is often the type of information we want to have about functions of interest as political scientists. For the dichotomous simulated data we may also be interested in the average slope of $\sigma(f(x))$; here, the true average marginal effect of x on $\sigma(f(x))$ is 0.146. Figure 3 shows the posterior means and 95% CIs for the average marginal effects in the regression and classification cases, with linear model baselines depicted for comparison; we can see the average marginal effect on both $f(x)$ and $\sigma(f(x))$ is captured well by the GP models. Even though there is a true linear trend in f and the nonlinear deviations are designed to cancel out over the range of the simulated x values, the linear models by contrast poorly estimate the average marginal effect of x on both the link and probability scales.

[12] At this point, since we have resorted to simulation, it may be tempting to use ESS to draw from the posterior. This can be done, but then we lose the ability to perform simple model selection for GP prior hyperparameters, resulting in markedly increased computation time.



(a) AME of x on $f(x)$ (b) AME of x on $\sigma(f(x))$

Figure 3: Average marginal effect (AME) of x. The left panel shows the AME of x on $f(x)$ for both regression and classification; the right panel shows the AME of x on the probability of a positive outcome in classification. The true theoretical AME is given by the dashed line in both panels.

References

Carlson, David. 2020. "Modeling without conditional independence: Gaussian process regression for time-series cross-sectional analyses." Working paper, available at https://mysite.ku.edu.tr/dcarlson/research/.

Cheng, Lu, Siddharth Ramchandran, Tommi Vatanen, Niina Lietzén, Riitta Lahesmaa, Aki Vehtari, and Harri Lähdesmäki. 2019. "An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data." *Nature Communications* 10 (1): 1–11.

Duck-Mayr, JBrandon, Roman Garnett, and Jacob Montgomery. 2020. "GPIRT: A Gaussian Process Model for Item Response Theory." In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI),* edited by Jonas Peters and David Sontag, 124:520–529. Proceedings of Machine Learning Research. PMLR.

Gill, Jeff. 2020. "Measuring constituency ideology using Bayesian universal kriging." *State Politics & Policy Quarterly,* https://doi.org/10.1177/1532440020930197. eprint: https://doi.org/10.1177/1532440020930197. https://doi.org/10.1177/1532440020930197.

Hainmueller, Jens, and Chad Hazlett. 2014. "Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach." *Political Analysis* 22 (2): 143–168.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the most of statistical analyses: Improving interpretation and presentation." *American Journal of Political Science* 44 (2): 341–355.

Murray, Iain, Ryan Prescott Adams, and David J. C. MacKay. 2010. "Elliptical slice sampling." *The Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* 9:541–548.

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning.* MIT Press.

Appendix

A Derivation of posterior over β in GP regression

In Gaussian process regression, we have the following model specification:

$$y \sim \mathcal{N}(f(X) + X\beta, \sigma_y^2 I), \tag{48}$$

$$\beta \sim \mathcal{N}(b, B), \tag{49}$$

$$f \sim \mathcal{GP}(0, K(X)). \tag{50}$$

Suppose we want the posterior distribution of β. We find

$$p(\beta \mid y, X) \propto p(y \mid \beta)p(\beta) \tag{51}$$

$$= N(y; X\beta, K_y) \times N(\beta; b, B), \tag{52}$$

where

$$K_y = K(X) + \sigma_y^2 I. \tag{53}$$

We can show

$$\beta \mid y \sim \mathcal{N}(\bar{\beta}, \Sigma_\beta), \tag{54}$$

$$\bar{\beta} = (B^{-1} + X^T K_y^{-1} X)^{-1}(B^{-1} b + X^T K_y^{-1} y), \tag{55}$$

$$\Sigma_\beta = (B^{-1} + X^T K_y^{-1} X)^{-1} \tag{56}$$

with a bit of tedious algebra:

$$p(\beta \mid y, X) \propto \exp\left(-\frac{1}{2}\left[(y - X\beta)^T K_y^{-1}(y - X\beta) + (\beta - b)^t B^{-1}(\beta - b)\right]\right) \tag{57}$$

$$= \exp\Bigg(-\frac{1}{2}\Bigg[y^T K_y^{-1} y - \beta^T X^T K_y^{-1} y - y^T K_y^{-1} X\beta + \beta^T X^T K_y^{-1} X\beta \tag{58}$$

$$+ \beta^T B^{-1}\beta - b^T B^{-1}\beta - \beta^T B^{-1} b + b^T B^{-1} b\Bigg]\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\Bigg[\beta^T B^{-1}\beta + \beta^T X^T K_y^{-1} X\beta - \beta^T B^{-1} b - \beta^T X^T K_y^{-1} y \tag{59}$$

$$- b^T B^{-1}\beta - y^T K_y^{-1} X\beta\Bigg]$$

$$- \frac{1}{2}\Bigg[y^T K_y^{-1} y + b^T B^{-1} b\Bigg]\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\Bigg[\beta^T (B^{-1} + X^T K_y^{-1} X)\beta - \beta^T (B^{-1} b + X^T K_y^{-1} y) \tag{60}$$

$$- (b^T B^{-1} + y^T K_y^{-1} X)\beta\Bigg]$$

$$- \frac{1}{2}\Bigg[y^T K_y^{-1} y + b^T B^{-1} b\Bigg]\Bigg)$$

$$= \exp\Bigg(-\frac{1}{2}\Bigg[\Big(\beta - (B^{-1} + X^T K_y^{-1} X)^{-1}(B^{-1} b + X^T K_y^{-1} y)\Big)^T \tag{61}$$

$$\times (B^{-1} + X^T K_y^{-1} X)$$

$$\times \Big(\beta - (B^{-1} + X^T K_y^{-1} X)^{-1}(B^{-1} b + X^T K_y^{-1} y)\Big)\Bigg]$$

$$- \frac{1}{2}\Bigg[y^T K_y^{-1} y + b^T B^{-1} b\Bigg]\Bigg)$$

$$\propto \exp\Bigg(-\frac{1}{2}\Bigg[\Big(\beta - \bar{\beta}\Big)^T \Sigma_\beta^{-1}(\beta - \bar{\beta})\Bigg]\Bigg), \tag{62}$$

which is clearly the core of a normal distribution.

B Distribution of derivatives of Gaussian processes

Luckily, "[s]ince differentiation is a linear operator, the derivative of a Gaussian process is another Gaussian process" (Rasmussen and Williams 2006, 191).

Let

$$\mathbf{f}_d = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_{1d}} \\ \vdots \\ \dfrac{\partial f_n}{\partial x_{nd}} \end{bmatrix}. \tag{63}$$

Using Equation 9.1 in Rasmussen and Williams (2006),

$$K_d \triangleq \mathbb{C}\left[\mathbf{f}_d, \mathbf{f}\right] = \begin{bmatrix} \dfrac{\partial k\left(\mathbf{x}_1, \mathbf{x}_1\right)}{\partial x_{1d}} & \cdots & \dfrac{\partial k\left(\mathbf{x}_1, \mathbf{x}_n\right)}{\partial x_{1d}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial k\left(\mathbf{x}_n, \mathbf{x}_1\right)}{\partial x_{nd}} & \cdots & \dfrac{\partial k\left(\mathbf{x}_n, \mathbf{x}_n\right)}{\partial x_{nd}} \end{bmatrix}, \tag{64}$$

$$K_{dd} \triangleq \mathbb{C}\left[\mathbf{f}_d, \mathbf{f}_d\right] = \begin{bmatrix} \dfrac{\partial^2 k\left(\mathbf{x}_1, \mathbf{x}_1\right)}{\partial x_{1d} \partial x_{1d}} & \cdots & \dfrac{\partial^2 k\left(\mathbf{x}_1, \mathbf{x}_n\right)}{\partial x_{1d} \partial x_{nd}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 k\left(\mathbf{x}_n, \mathbf{x}_1\right)}{\partial x_{nd} \partial x_{1d}} & \cdots & \dfrac{\partial^2 k\left(\mathbf{x}_n, \mathbf{x}_n\right)}{\partial x_{nd} \partial x_{nd}} \end{bmatrix}, \tag{65}$$

To make the notation more compact, as is usual we set $K = K\left(\mathbf{X}, \mathbf{X}\right)$, and additionally set $\mu = \mu$ and

$$\mu_d = \begin{bmatrix} \dfrac{\partial \mu\left(\mathbf{x}_1\right)}{\partial x_{1d}} \\ \vdots \\ \dfrac{\partial \mu\left(\mathbf{x}_n\right)}{\partial x_{nd}} \end{bmatrix}. \tag{66}$$

Then we can describe the joint prior on $\mathbf{f}$ and $\mathbf{f}_d$:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_d \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_d \end{bmatrix}, \begin{bmatrix} K & K_d^T \\ K_d & K_{dd} \end{bmatrix} \right). \tag{67}$$

Then the posterior distribution of $\mathbf{f}_d$ given $\mathbf{X}$ and $\mathbf{y}$ is normal with mean

$$\begin{aligned} \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] &= \int \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{f}, \mathbf{X}\right] p\left(\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right) d\mathbf{f} && (68) \\ &= \int \mu_d + K_d K^{-1} \left(\mathbf{f} - \mu\right) p\left(\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right) d\mathbf{f} && (69) \\ &= \mu_d + K_d K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right), && (70) \end{aligned}$$

and variance (using the law of total variance)

$$\begin{aligned} \mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] &= K_{dd} - K_d K^{-1} K_d^T + \mathbb{E}\left[\left(\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{f}\right] - \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right]\right)^2\right] && (71) \\ &= K_{dd} - K_d K^{-1} K_d^T \\ &\quad + \mathbb{E}\left[\left(\mu_d + K_d K^{-1} \left(\mathbf{f} - \mu\right)\right.\right. \\ &\qquad \left.\left. - \left(\mu_d + K_d K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right)\right)\right)^2\right] && (72) \\ &= K_{dd} - K_d K^{-1} K_d^T \\ &\quad + \mathbb{E}\left[\left(K_d K^{-1} \left(\mathbf{f} - \mu\right) - K_d K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right)\right)^2\right] && (73) \\ &= K_{dd} - K_d K^{-1} K_d^T \\ &\quad + \mathbb{E}\left[\left(K_d K^{-1}\mathbf{f} - K_d K^{-1}\mu\right.\right. \\ &\qquad \left.\left. - K_d K^{-1} \mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] + K_d K^{-1}\mu\right)^2\right] && (74) \\ &= K_{dd} - K_d K^{-1} K_d^T + \mathbb{E}\left[\left(K_d K^{-1}\mathbf{f} - K_d K^{-1} \mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right]\right)^2\right] && (75) \end{aligned}$$

$$= K_{dd} - K_d\, K^{-1}\, K_d^T + \mathbb{E}\left[K_d\, K^{-1}\Big(\mathbf{f} - \mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right]\Big)^2 K^{-1} K_d^T\right] \tag{76}$$

$$= K_{dd} - K_d\, K^{-1}\, K_d^T + K_d\, K^{-1}\, \mathbb{E}\left[\Big(\mathbf{f} - \mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right]\Big)^2\right] K^{-1} K_d^T \tag{77}$$

$$= K_{dd} - K_d\, K^{-1}\, K_d^T + K_d\, K^{-1}\, \mathbb{V}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] K^{-1} K_d^T \tag{78}$$

$$= K_{dd} - K_d \left(K^{-1} - K^{-1}\, \mathbb{V}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] K^{-1}\right) K_d^T. \tag{79}$$

B.1 The regression case

In the regression case,

$$\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = \mu_d + K_d\, K^{-1}\left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right) \tag{80}$$

$$= \mu_d + K_d\, K^{-1}\left(\mu + K\, K_y^{-1}\left(\mathbf{y} - \mu\right) - \mu\right) \tag{81}$$

$$= \mu_d + K_d K^{-1}\left(K\, K_y^{-1}\left(\mathbf{y} - \mu\right)\right) \tag{82}$$

$$= \mu_d + K_d\, K_y^{-1}\left(\mathbf{y} - \mu\right), \tag{83}$$

where, as usual, $K_y = K + \sigma_y^2 I$. Then

$$\mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = K_{dd} - K_d \left(K^{-1} - K^{-1}\, \mathbb{V}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] K^{-1}\right) K_d^T \tag{84}$$

$$= K_{dd} - K_d \left(K^{-1} - K^{-1}\left(K - K K_y^{-1} K\right) K^{-1}\right) K_d^T \tag{85}$$

$$= K_{dd} - K_d \left(K^{-1} - \left(I - K_y^{-1} K\right) K^{-1}\right) K_d^T \tag{86}$$

$$= K_{dd} - K_d \left(K^{-1} - \left(K^{-1} - K_y^{-1}\right)\right) K_d^T \tag{87}$$

$$= K_{dd} - K_d\, K_y^{-1}\, K_d^T. \tag{88}$$

Note the similarity to the predictive distribution of $\mathbf{f}_*$ for new cases $\mathbf{X}_*$.

B.2 The classification case

In the classification case, under the Laplace approximation to the posterior,

$$\mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = \mu_d + K_d\, K^{-1} \left(\mathbb{E}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] - \mu\right) \tag{89}$$

$$= \mu_d + K_d\, K^{-1} \left(\mu + K\left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right) - \mu\right) \tag{90}$$

$$= \mu_d + K_d K^{-1} \left(K\, \left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right)\right) \tag{91}$$

$$= \mu_d + K_d\, \left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \tag{92}$$

$$\mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] = K_{dd} - K_d \left(K^{-1} - K^{-1}\, \mathbb{V}\left[\mathbf{f} \mid \mathbf{X}, \mathbf{y}\right] K^{-1}\right) K_d^T \tag{93}$$

$$= K_{dd} - K_d \left(K^{-1} - K^{-1} \left(K^{-1} + W\right)^{-1} K^{-1}\right) K_d^T \tag{94}$$

$$= K_{dd} - K_d \left(K + W^{-1}\right)^{-1} K_d^T, \quad \text{(by the matrix inversion lemma)} \tag{95}$$

$$W = -\nabla\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right). \tag{96}$$

C Distribution of average marginal effects

Since then γ_d is a constant ($1/N$) times the sum of correlated normal random variables,

$$\gamma_d \sim \mathcal{N}\left(\frac{1}{N}\sum_{i=1}^{N} m_{\gamma_d i}\,,\ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} c_{\gamma_d ij}\right), \tag{97}$$

$$\mathbf{m}_{\gamma_d} = \mathbb{E}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] \tag{98}$$

$$\mathbf{C}_{\gamma_d} = \mathbb{V}\left[\mathbf{f}_d \mid \mathbf{X}, \mathbf{y}\right] \tag{99}$$

Importantly, we may also get the average marginal effect of feature d within subgroups of $\mathbf{X}$ rather than the full sample average by simply altering the indices of summation in Equation 97.

However, for binary classification, we may be more interested in the distribution of

$$\pi_d \triangleq \frac{1}{N}\sum_{i}^{N} \frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, x_{id}} \tag{100}$$

than γ_d. Unfortunately, that distribution cannot be analytically expressed, though we can readily simulate from it. First note that

$$\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, x_{id}} = \frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, f}\frac{\partial\, f\left(\mathbf{x}_i\right)}{\partial\, x_{id}}. \tag{101}$$

Generally the sigmoid function σ has a known derivative; for example, in the logistic case,

$$\frac{\partial\, \sigma\left(f\left(\mathbf{x}_i\right)\right)}{\partial\, f} = \sigma\left(f\left(\mathbf{x}_i\right)\right)\left(1 - \sigma\left(f\left(\mathbf{x}_i\right)\right)\right). \tag{102}$$

Since we have an approximation to the posterior on f, and given f, the posterior over $\mathbf{f}_d$ is

$$\mathbf{f}_d \mid \mathbf{f} \sim \mathcal{N}\left(\mu_d + K_d\, K^{-1}\left(\mathbf{f} - \mu\right), K_{dd} - K_d\, K^{-1}\, K_d^T\right), \tag{103}$$

we can obtain M samples of π_d by

- drawing $\mathbf{f}^t$ from $\mathcal{N}\left(\mu + K\left(\nabla \log p\left(\mathbf{y} \mid \hat{\mathbf{f}}\right)\right), \left(K^{-1} + W\right)^{-1}\right)$,
- drawing $\mathbf{f}_d^t$ from $\mathcal{N}\left(\mu_d + K_d K^{-1}\left(\mathbf{f}^t - \mu\right), K_{dd} - K_d K^{-1} K_d^T\right)$,
- and calculating $\pi_d^t = \dfrac{1}{N}\sum_{i=1}^{N} \dfrac{\partial \sigma\left(f_i^t\right)}{\partial f} f_{id}^t$,

so that we can summarize the distribution of π_d using the M draws, similar to the CLARIFY procedure (King, Tomz, and Wittenberg 2000).

Moreover, another issue to consider is discrete variables in $\mathbf{X}$. For binary variables, let $\mathbf{X}_1^*$ be a set of test points where all feature observactions are identical to $\mathbf{X}$ except that all observations of feature d have been set to 1, and analogously for $\mathbf{X}_0^*$.[13] Then a more appropriate quantity of interest rather than γ_d is

$$\delta_d = \frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{x}_{1i}^*\right) - f\left(\mathbf{x}_{0i}^*\right), \tag{104}$$

which gives the average marginal effect on y of taking a 1 vs a 0 value in the regression case, or the average partial effect on f of taking a 1 vs a 0 value in the classification case. Letting $\mathbf{f}_1 = f\left(\mathbf{X}_1^*\right)$ and analogously for $\mathbf{f}_0$, δ_d is distributed

$$\delta_d \sim \mathcal{N}\left(\frac{1}{N}\sum_{i=1}^{N} m_{\delta_d i}\,,\ \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} c_{\delta_d ij}\right), \tag{105}$$

$$\mathbf{m}_{\delta_d} = \mathbb{E}\left[\mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right] - \mathbb{E}\left[\mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right], \tag{106}$$

$$\mathbf{C}_{\delta_d} = \mathbb{V}\left[\mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{V}\left[\mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{C}\left[\mathbf{f}_1^*, \mathbf{f}_0^* \mid \mathbf{X}, \mathbf{y}\right] + \mathbb{C}\left[\mathbf{f}_0^*, \mathbf{f}_1^* \mid \mathbf{X}, \mathbf{y}\right]. \tag{107}$$

For classification we can also simulate the distribution of

$$\psi_d = \frac{1}{N}\sum_{i=1}^{N} \sigma\left(f\left(\mathbf{x}_{1i}^*\right)\right) - \sigma\left(f\left(\mathbf{x}_{0i}^*\right)\right) \tag{108}$$

by simply generating values of $f\left(\mathbf{X}_1^*\right)$ and $f\left(\mathbf{X}_0^*\right)$ from the posterior approximation, pushing those samples through the chosen sigmoid σ, and calculate the average of the differences to

[13] You can replace 1 and 0 with other binary value labels as needed.

get the average marginal effect for each sample so that we can summarize the distribution of average marginal effects, similar to the continuous case. In the case of a categorical variable that has been one-hot encoded into $\mathbf{X}$, we can simply follow the above procedures for all the substantively interesting pairwise comparisons between category labels. Moreover, in some cases using this procedure to find the distribution of difference in MAP predictions at two discrete values of a continuous variable may be more readily interpretable than π_d.

D Derivatives of mean and covariance functions

Regarding the derivatives of the mean function,

$$\mu\left(X\right)=0\Rightarrow\frac{\partial\,\mu\left(\mathbf{x}_i\right)}{\partial\,\mathbf{x}_{id}}=0, \tag{109}$$

$$\mu\left(X\right)=X\beta\Rightarrow\frac{\partial\,\mu\left(\mathbf{x}_i\right)}{\partial\,\mathbf{x}_{id}}=\beta_d, \tag{110}$$

to cover a couple of popular choices.

Note that calculating the mean and variance of the distribution of γ_d requires a twice-differentiable covariance function. For the squared exponential covariance function with automatic relevance determination,

$$k\left(\mathbf{x}_i,\mathbf{x}_j\right)=\sigma_f^2\exp\left(-\frac{1}{2}\sum_d\left[\frac{\left(x_{id}-x_{jd}\right)^2}{\ell_d^2}\right]\right). \tag{111}$$

Then,

$$\frac{\partial\,k\left(\mathbf{x}_i,\mathbf{x}_j\right)}{\partial\,\mathbf{x}_{id}}=k\left(\mathbf{x}_i,\mathbf{x}_j\right)\frac{x_{jd}-x_{id}}{\ell_d^2} \tag{112}$$

$$\frac{\partial^{\,2}k\left(\mathbf{x}_i,\mathbf{x}_j\right)}{\partial\,\mathbf{x}_{id}\partial\,\mathbf{x}_{jd}}=k\left(\mathbf{x}_i,\mathbf{x}_j\right)\left(\frac{1}{\ell_d^2}+\frac{\left(x_{jd}-x_{id}\right)\left(x_{id}-x_{jd}\right)}{\ell_d^4}\right), \tag{113}$$

Note that when $i=j$, the cross partial simplifies to σ_f^2/ℓ_d^2.