Short course on high-dimensional Bayesian modeling

Joseph (Joey) Antonelli and Antonio (Tony) Linero

June 24th, 2021

Why are we here?

- High-dimensional modeling has vastly grown in popularity over the last couple of decades
- There are a few reasons for this
  - Advancements in computation
    - Especially for Bayesian approaches!
  - Increasing number of large data sets
    - Genomics, imaging, medical records, etc.
  - Advancements in statistical techniques
- Important to have at least a working understanding of these models

Why are we here?

- Bayesian approaches can be particularly useful in this setup
  - Easily account for uncertainty
  - Introduce nonlinearity in a natural way
  - Introduce more complex structures such as hierarchical models
  - Handle missing data
- Many frequentist estimators in high dimensions don't provide inference
  - No confidence intervals for predictions or parameters
- Some work done to alleviate these issues (see Van de Geer et al. (2014); Lee et al. (2016), others)
  - Specific to certain models
  - Rely heavily on strong assumptions and asymptotics

What we hope you take away from this

- At the end of this course, we hope you will be able to
  - Understand prior distributions for high-dimensional models
  - Understand the computational aspects involved with implementing these models
  - Code up your own MCMC using spike-and-slab prior distributions
  - Incorporate nonlinearity into your models
  - Understand the more complex nonlinear models that exist in high dimensions
- We hope that after this course, you will have the tools to try and imbed these ideas into your own research

Disclaimers

- We are not experts in all aspects of high-dimensional Bayesian analysis!
  - We have both used these models in our own research and hope to bestow some of our ideas onto you so you don't run into the same issues that we did
- We mostly focus on spike-and-slab prior distributions and extensions to tree-based models, and do not have time to cover all high-dimensional Bayesian models and therefore certain important concepts will be left out

Roadmap of short course

- We will build from simple models to more complex models

$$\sum_{j=1}^{p} X_j \beta_j \longrightarrow \sum_{j=1}^{p} f_j(X_j) \longrightarrow f(\boldsymbol{X})$$

- We begin with simple linear models to learn foundational concepts
  - Spike-and-slab priors
  - Sensitivity to hyperprior choices
  - How to sample from these models

Roadmap of short course

- We will then alleviate the assumptions of this model
  - Linearity and additivity assumptions
- We will discuss grouped variable selection as a method to introduce nonlinearity as well as fully nonparametric Gaussian process regression
  - How these are used in high-dimensional scenarios
- We finish the course with tree-based models that have been shown to work remarkably well
- All along the way we will be highlighting examples and examining R code to implement these approaches.

Notation and setup

- Y: Outcome of interest
- $\boldsymbol{X}$: P-dimensional covariates
- N: Overall sample size
- Our goal throughout is to use $\boldsymbol{X}$ to predict Y, i.e estimate $E(Y|\boldsymbol{X}) = f(\boldsymbol{X})$
- Two simultaneous goals
  - Good prediction performance
  - Identifying important predictors
- Q: The true number of important predictors in $\boldsymbol{X}$

What do we mean by high dimensions?

- Typically high-dimensional modeling refers to situations where $P > N$
- We won't be discussing asymptotic rates or theoretical results too much, however, it is typically assumed that P grows with N
- Throughout, we will work under the slightly broader definition of any situation where P is large enough to require high-dimensional techniques such as shrinkage or sparsity inducing prior distributions
  - Traditional models either don't apply or perform poorly
  - Interested in learning which predictors affect the outcome

The linear model

- To introduce the foundational concepts, we restrict to the linear model

$$Y = \sum_{j=1}^{p} X_j \beta_j + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- All of the following ideas apply immediately to generalized linear models

$$g^{-1}(E(Y|\boldsymbol{X})) = \sum_{j=1}^{p} X_j \beta_j$$

- If we assume the true parameter is $\boldsymbol{\beta}_0$, then $Q = ||\boldsymbol{\beta}_0||_0 = \sum_{j=1}^{P} 1(\beta_{0j} \neq 0)$

The linear model

- Unknown parameters are $(\boldsymbol{\beta}, \sigma^2)$
- We won't discuss prior distributions for σ^2
  - Assume fixed
  - Standard conjugate inverse-gamma prior for σ^2
- Will discuss prior distributions for $\boldsymbol{\beta}$ that
  - Work when p is large
  - Identify nonzero elements of $\boldsymbol{\beta}_0$
  - Are easy to implement computationally

Prior distributions for β

- We will focus on prior distributions of the following form:

$$P(\beta_j|\gamma_j) \sim (1-\gamma_j)\delta_0 + \gamma_j \mathcal{N}(0, \sigma^2_\beta)$$
$$P(\gamma_j) = \tau^{\gamma_j}(1-\tau)^{1-\gamma_j}$$

- Independent priors for each β_j
- Note that this prior can equivalently be expressed as

$$P(\beta_j) \sim (1-\tau)\delta_0 + \tau \mathcal{N}(0, \sigma^2_\beta)$$

Prior distributions for β

$$P(\beta_j|\gamma_j) \sim (1-\gamma_j)\delta_0 + \gamma_j \mathcal{N}(0, \sigma^2_\beta)$$
$$P(\gamma_j) = \tau^{\gamma_j}(1-\tau)^{1-\gamma_j}$$

- Prior distribution is a two-component mixture distribution (Mitchell and Beauchamp, 1988)
  - Point mass at zero (the spike)
  - Continuous distribution (the slab)
- Intuitively this prior distribution recognizes the fact that some covariates are not important ($\beta_j = 0$), while others are

Prior distributions for β

- We focus on this distribution, but there are many variations of it
- The continuous spike-and-slab prior distribution is commonly used (George and McCulloch, 1993)

$$P(\beta_j|\gamma_j) \sim (1-\gamma_j)\mathcal{N}(0,\sigma_0^2) + \gamma_j\mathcal{N}(0,\sigma_1^2)$$
$$P(\gamma_j) = \tau^{\gamma_j}(1-\tau)^{1-\gamma_j}$$

- Here $\sigma_0^2 < \sigma_1^2$ and is small so that you still have a spike near zero
- Leads to straightforward updates of γ_j, but requires good choices of both (σ_0^2, σ_1^2)

Prior distributions for β

- This model has a number of important features
- Performs variable selection and reduces the number of nonzero parameters
  - Can investigate $P(\gamma_j = 1|\mathcal{D})$ for variable importance
  - Can look at full posterior distribution of $\boldsymbol{\gamma}$ to identify models most supported by the data
- Still performs shrinkage of important coefficients
  - Depends on the magnitude of σ^2_β

Prior distributions for β

- The performance of these prior distributions depends heavily on the choice of hyperpriors
- τ represents the prior probability that a coefficient is nonzero
  - Reflects the underlying sparsity in the model
- σ_β^2 has a large impact on the resulting coefficient estimates
  - Degree of shrinkage
  - Variable selection properties

Updating model parameters

- The most important (and difficult) parameters to update are (β_j, γ_j)
- A traditional Gibbs sampler would update from the following
  - $P(\beta_j|\cdot)$: the full conditional for β_j
  - $P(\gamma_j|\cdot)$: the full conditional for γ_j
  - Repeat for all $j = 1, \ldots, p$
  - $P(\tau|\cdot)$: the full conditional for τ
  - $P(\sigma^2_\beta|\cdot)$: the full conditional for σ^2_β
- This seems easy enough, right?

Updating model parameters

- Unfortunately not
- Let's look at the full conditional distribution for γ_j:

$$P(\gamma_j = 1 | \beta_j, \boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}_{-j}, \tau, \sigma^2_\beta, \mathcal{D}) = 1(\beta_j \neq 0)$$

- The probability is one if $\beta_j \neq 0$ and 0 otherwise
  - Makes sense considering that γ_j is a latent variable indicating whether $\beta_j = 0$ or not
- If we follow this strategy, we will never explore the full model space in our MCMC
  - γ_j can never change from the starting values

Updating model parameters

- The main way to avoid this issue is to integrate out β_j when updating γ_j
- A common strategy used in the model averaging literature is to integrate out all unknown parameters other than $\boldsymbol{\gamma}$ and update from

$$P(\boldsymbol{\gamma}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\gamma})P(\boldsymbol{\gamma})}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\boldsymbol{\gamma})P(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}} P(\mathcal{D}|\boldsymbol{\gamma})P(\boldsymbol{\gamma})}$$

- Can then follow an MCMC strategy of successfully sampling from
  - $P(\boldsymbol{\gamma}|\mathcal{D})$
  - $P(\boldsymbol{\beta}, \tau, \sigma^2_\beta|\boldsymbol{\gamma}, \mathcal{D})$

Updating model parameters

- This requires knowledge of the marginal likelihood of the data
- In certain settings this is analytically tractable
  - Linear regression
- In other settings, this does not have a closed form expression
  - Approximations are available in some cases
- Even when the marginal likelihood has a closed form solution, it can be computationally prohibitive

Updating model parameters

- We will instead follow a Gibbs sampling style strategy by iterating through
  - $P(\beta_j, \gamma_j|\cdot)$: the full conditional for (β_j, γ_j)
  - Repeat for all $j = 1, \ldots, p$
  - $P(\tau|\cdot)$: the full conditional for τ
  - $P(\sigma^2_\beta|\cdot)$: the full conditional for σ^2_β
- The key is how we sample (β_j, γ_j)
  - Easy computationally
  - Doesn't get stuck at a particular γ_j value

Updating model parameters

- We will sample (β_j, γ_j) successively from
  - $P(\gamma_j | \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \tau, \sigma^2_\beta, \mathcal{D})$
  - $P(\beta_j | \gamma_j, \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \tau, \sigma^2_\beta, \mathcal{D})$
- Note that we don't condition on β_j in the update for γ_j
  - Integrates over possible values of β_j and avoids earlier problem
- The update for β_j is straightforward and is simply the full conditional we would use to update this parameter

Updating model parameters

- Now how do we update from the conditional of γ_j that doesn't condition on β_j?

$$P(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \tau, \sigma^2_\beta, \mathcal{D})$$

- This isn't straightforward, but a simple probability trick will facilitate computation

- For simplicity, let's denote all parameters with the exception of β_j and γ_j as $\boldsymbol{\theta}$

Updating model parameters

- We can re-write the quantity of interest as

$$P(\gamma_j = 1|\boldsymbol{\theta}, \mathcal{D}) = \frac{P(\beta_j = 0, \gamma_j = 1|\boldsymbol{\theta}, \mathcal{D})}{P(\beta_j = 0|\gamma_j = 1, \boldsymbol{\theta}, \mathcal{D})}$$

- The quantity on the right is simpler to work with
  - No longer averaging over β_j
- We'll see that we can re-write this in terms of quantities that are straightforward to calculate

Updating model parameters

- We can re-write the quantity of interest as

$$\begin{aligned}\frac{P(\beta_j = 0, \gamma_j = 1|\boldsymbol{\theta}, \mathcal{D})}{P(\beta_j = 0|\gamma_j = 1, \boldsymbol{\theta}, \mathcal{D})} &= \frac{P(\boldsymbol{\theta}, \mathcal{D}|\beta_j = 0, \gamma_j = 1)P(\beta_j = 0, \gamma_j = 1)}{P(\boldsymbol{\theta}, \mathcal{D})P(\beta_j = 0|\gamma_j = 1, \boldsymbol{\theta}, \mathcal{D})} \\ &= \frac{P(\boldsymbol{\theta}, \mathcal{D}|\beta_j = 0)P(\beta_j = 0, \gamma_j = 1)}{P(\boldsymbol{\theta}, \mathcal{D})P(\beta_j = 0|\gamma_j = 1, \boldsymbol{\theta}, \mathcal{D})} \\ &\propto \frac{P(\beta_j = 0, \gamma_j = 1)}{P(\beta_j = 0|\gamma_j = 1, \boldsymbol{\theta}, \mathcal{D})}\end{aligned}$$

- The second quality held because γ_j is irrelevant once we condition on β_j
- The third step held because neither $P(\boldsymbol{\theta}, \mathcal{D}|\beta_j = 0)$ or $P(\boldsymbol{\theta}, \mathcal{D})$ are functions of γ_j

Updating model parameters

- We can further decompose this quantity as

$$\frac{P(\beta_j = 0, \gamma_j = 1)}{P(\beta_j = 0 | \gamma_j = 1, \boldsymbol{\theta}, \mathcal{D})} = \frac{P(\beta_j = 0 | \gamma_j = 1) P(\gamma_j = 1)}{P(\beta_j = 0 | \gamma_j = 1, \boldsymbol{\theta}, \mathcal{D})}$$
$$= \frac{\tau \Phi(0; 0, \sigma^2_\beta)}{\Phi(0; m, v)}$$

- Where m and v are the mean and variance of the full conditional posterior distribution for β_j
- $\Phi(0; a, b)$ is the density at zero for a normal distribution with mean a and variance b

Updating model parameters

- We can do the same decomposition for $\gamma_j = 0$:

$$P(\gamma_j = 0|\boldsymbol{\theta}, \mathcal{D}) \propto \frac{P(\beta_j = 0|\gamma_j = 0)P(\gamma_j = 0)}{P(\beta_j = 0|\gamma_j = 0, \boldsymbol{\theta}, \mathcal{D})} = 1 - \tau$$

- So we can sample γ_j from a bernoulli distribution with probability given by

$$\frac{\frac{\tau\Phi(0;0,\sigma_\beta^2)}{\Phi(0;m,v)}}{\frac{\tau\Phi(0;0,\sigma_\beta^2)}{\Phi(0;m,v)} + (1-\tau)}$$

Updating model parameters

- This is extremely easy to calculate
- The only computation is in the calculation of m and v
  - Already need to calculate these anyways when updating β_j!
- The only thing this relied on was having a closed-form update for the conditional distribution of β_j given all other parameters, $P(\beta_j = 0 | \gamma_j = 1, \boldsymbol{\theta}, \mathcal{D})$
  - True for linear regression and generalized linear models
  - True in many other settings as well

Updating model parameters

- Now we've done the hard part!
- We now need to update the remaining parameters (τ, σ^2_β)
- One approach is to assign hyperprior distributions to each of these parameters
  - Relatively straightforward
  - Conjugate priors

Updating model parameters

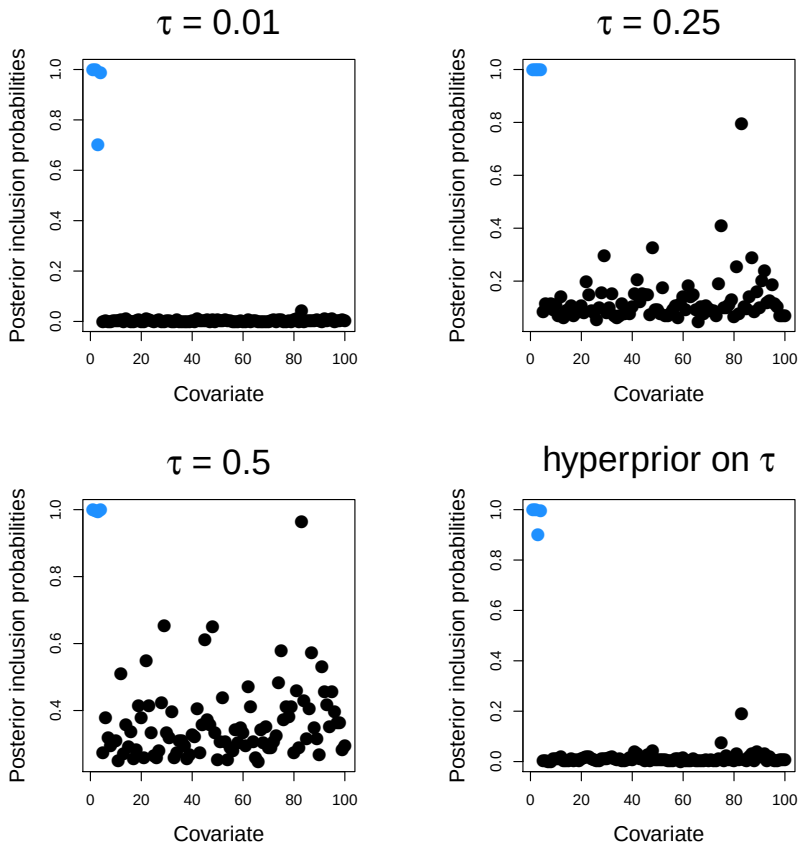
- Typically a beta prior distribution is assigned to τ
  - Conjugate
- A common choice in high-dimensional settings is to let the prior depend on the number of covariates

$$\tau \sim \mathcal{B}(C, p)$$

  where C is a pre-specified constant
- Mean of this distribution is $\frac{C}{C+p}$
  - More sparsity as p grows
- See Scott and Berger (2010) for a great paper on the impact of these different choices on variable selection

Importance of τ

- Results can be fairly sensitive to the choice of τ
  - Fully Bayes approach does well at finding a good solution
- True nonzero coefficients in blue, others in black



Importance of σ^2_β

- Fairly intuitive that τ impacts performance of variable selection
  - Prior probability of inclusion for each covariate
- Less clear is what impact σ^2_β has on the resulting model
- The most obvious utility of σ^2_β is for shrinkage of the resulting coefficients
  - Reduce variability of resulting estimates
- Does it impact variable selection?

Importance of σ^2_β

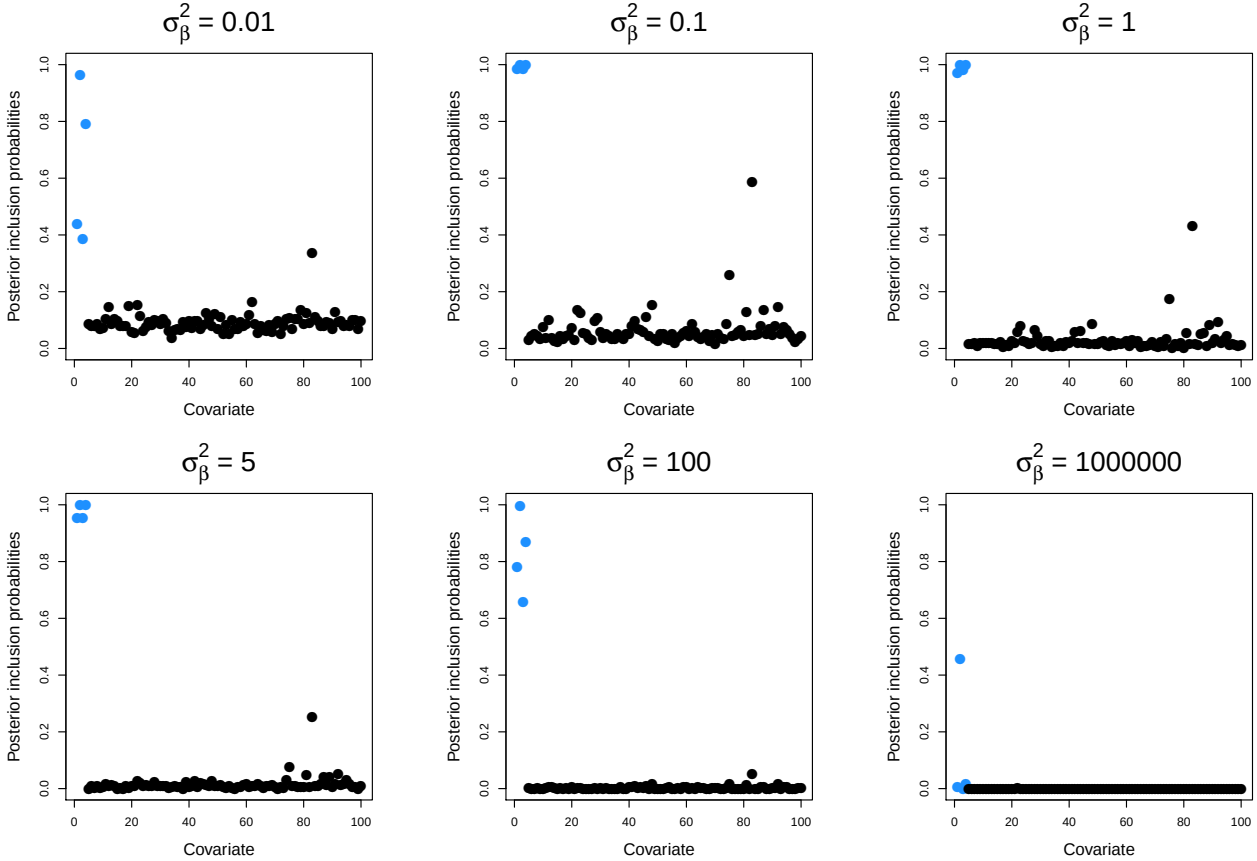
- To gain intuition, let's look at the term we used to update γ_j

$$\frac{\tau \Phi(0; 0, \sigma^2_\beta)}{\Phi(0; m, v)}$$

- Clearly this is an increasing function of τ
- σ^2_β shows up in both the numerator and denominator
  - m and v are both functions of σ^2_β

Importance of σ^2_β

- When σ^2_β is too small, we overly shrink coefficients and can't distinguish between the spike and slab leading to bad posterior inclusion probabilities
- When σ^2_β is too big, posterior inclusion probabilities go down



Updating σ^2_β

- Can place a conjugate prior on σ^2_β

$$\sigma^2_\beta \sim \text{InverseGamma}(a, b)$$

- Can also allow for a separate slab variance for each covariate (Mitra and Dunson, 2010)
  - Reduces shrinkage of larger coefficients

$$\sigma^2_{\beta_j} \sim \text{InverseGamma}(a, b)$$

- τ and σ^2_β can both be estimated with empirical Bayes as well

Updating model parameters

- Now let's take a look at some R code to see one way in which MCMC with these models is performed

Extending to nonlinear models

- Let's see how we can extend these ideas to the nonlinear model
- Now our goal will be to estimate the following

$$E(Y|X) = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$$

- Intuitively we want to place a spike and slab prior on this function somehow
  - Either the function is a flat function at zero, or something else
  - We will see two ways to do this

Extending to nonlinear models

- Easiest way is to make parametric assumption about $f_j(\cdot)$

$$\begin{aligned} f_j(X_j) &= \sum_{k=1}^{K} b_k(X_j)\beta_{jk} \\ &= \widetilde{X}_j \boldsymbol{\beta}_j \end{aligned}$$

- Here, $b_k(\cdot)$ are basis functions such as polynomials, natural cubic splines, wavelets, etc.
- $\boldsymbol{\beta}_j$ is a $K-$dimensional vector of parameters for covariate j

Extending to nonlinear models

- If $\boldsymbol{\beta}_j = \mathbf{0}$, then $f(X_j) = 0$ and the covariate is dropped from the model
- Therefore we can use a multivariate version of the spike-and-slab prior

$$P(\boldsymbol{\beta}_j|\gamma_j) \sim (1-\gamma_j)\delta_{\mathbf{0}} + \gamma_j \mathcal{N}_K(\mathbf{0}, \Sigma_\beta)$$

- The prior distribution is now a mixture between a point mass at the vector $\mathbf{0}$ and a multivariate normal distribution
- Similar to other grouped variable selection approaches
  - Either all in or all out

Extending to nonlinear models

- There are effectively no differences between this and the univariate approach seen earlier
- The one difference is the choice of slab variance
  - Now a covariance matrix
- There are a couple of natural choices
  - $\Sigma_\beta = \sigma^2_\beta (\boldsymbol{X_j}^T \boldsymbol{X_j})^{-1}$
  - $\Sigma_\beta = \sigma^2_\beta I_K$
- Can simply choose the scaled identity matrix for simplicity, and should work reasonably well

Extending to nonlinear models

- What if we don't want to make parametric assumptions about $f_j(\cdot)$?
- What if we don't want to specify basis functions, $b_k(\cdot)$?
- The nonparametric Bayesian approach would be to place a prior on the function $f_j()$
- The most natural choice is to place a Gaussian process prior on this function
  - Very flexible
  - Been shown to work well empirically

Brief intro to Gaussian processes

- To place a Gaussian process prior we can write

$$f_j \sim GP(\mu_j(X_j), K_j(X_j, X_j'))$$

- Here $\mu_j(\cdot)$ is the mean function
  - Could be a linear function
  - Could be the zero function
- $K_j(X_j, X_j')$ is a kernel function that reflects the similarity/distance between X_j and X_j'
  - Smaller distances mean larger values

Brief intro to Gaussian processes

- This formulation implies that for any finite collection of points, we have

$$(f_j(X_{j1}), \ldots, f_j(X_{jn}))' \sim \mathcal{N}((\mu_j(X_{j1}), \ldots, \mu_j(X_{jn}))', \Sigma_j)$$

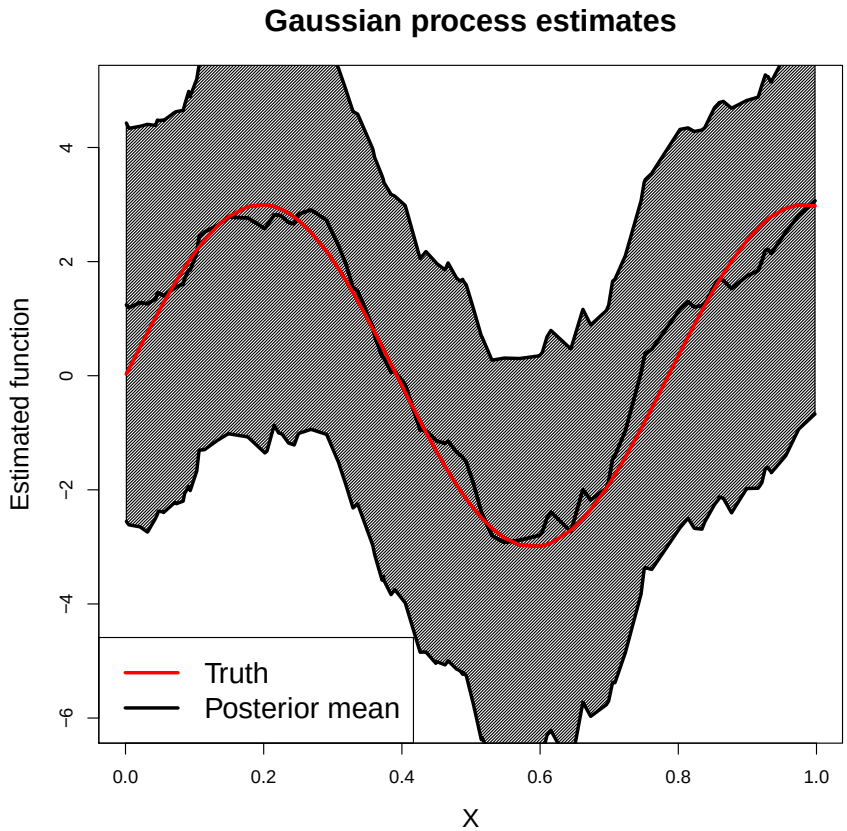
  where the (a, b) element of Σ_j is $K(X_{ja}, X_{jb})$

- This allows the function to deviate from the pre-specified mean function $\mu_j(\cdot)$

- The main assumption is smoothness
  - Nearby points have similar values of $f_j(\cdot)$
  - Degree of smoothness controlled by kernel function
  - Similar in spirit to local or kernel regression

Brief intro to Gaussian processes

- Gaussian processes can approximate nonlinear functions well without having to specify any functional form of the true function



Gaussian processes and model averaging

- Letting $f_j(\boldsymbol{X}_j)$ be the vector of n observed locations, we can specify the following prior distribution (Reich et al., 2009)

$$f_j(\boldsymbol{X}_j) \sim \mathcal{N}(\mathbf{0}, \sigma_j \Sigma_j)$$
$$\sigma_j \sim (1 - \gamma_j)\delta_0 + \gamma_j G$$

- G is any continuous distribution that lives on the positive real line
- We use the zero mean function here and the covariance matrix Σ_j is the kernel matrix from before
- The variance is either zero and the covariate is not in the model, or it is positive and the covariate is included using a GP

Gaussian processes and model averaging

- Alternatively, we could specify the spike-and-slab prior directly on the observed functions

$$f_j(\boldsymbol{X}_j) \sim (1-\gamma_j)\delta_{\mathbf{0}} + \gamma_j \mathcal{N}_n(\mathbf{0}, \sigma_j \Sigma_j)$$

- The n values are either all zero together, or all nonzero
- γ_j has the same interpretation as in the simpler models
  - Importance of covariate j
  - $P(\gamma_j = 1|\mathcal{D})$ shows the strength of this importance

Gaussian processes and model averaging

- This looks substantially more complicated, but updating γ_j is equally straightforward!
- We can use the same trick to see that

$$P(\gamma_j = 1|\mathcal{D}, \boldsymbol{\theta}) = \frac{P(f_j(\boldsymbol{X}_j) = \mathbf{0}, \gamma_j = 1|\mathcal{D}, \boldsymbol{\theta})}{P(f_j(\boldsymbol{X}_j) = \mathbf{0}|\gamma_j = 1, \mathcal{D}, \boldsymbol{\theta})}$$
$$\propto \frac{\tau\ \Phi(\mathbf{0}; \mathbf{0}, \sigma_j \boldsymbol{\Sigma_j})}{\Phi(\mathbf{0}; \boldsymbol{M}, \boldsymbol{V})}$$

  where $\Phi(\cdot)$ now corresponds to a multivariate normal density of dimension n

- $\boldsymbol{M}$ and $\boldsymbol{V}$ are now the full conditional mean and variance for $f_j(\boldsymbol{X}_j)$
  - This full conditional distribution is a multivariate normal distribution

Pros and cons of Gaussian processes

- As discussed earlier, GPs are very flexible
  - Can capture basically any true function
- The main drawback is the heavy computational burden
  - Calculation of $\boldsymbol{M}$ and $\boldsymbol{V}$ requires inversion of an $n \times n$ matrix
  - Extremely slow for even moderate sample sizes
  - Have to do this for each covariate!
- A number of computational speedups and approximations have been proposed to alleviate this issue
  - See Gramacy and Lee (2008); Banerjee et al. (2008, 2013)

Overview of nonlinear spike-and-slab models

- If computation time is a big concern, use the basis function approach

$$f_j(X_j) = \widetilde{X}_j \boldsymbol{\beta}_j$$

  - Essentially equal in computation speed as the linear model
  - Not quite as flexible as a GP, but still very flexible

- We have still made the assumption of additivity in all of these models

- Some work done to alleviate this assumption in GPs (Qamar and Tokdar, 2014)

$$E(Y|X) = f_1(\boldsymbol{X}) + \ldots, + f_k(\boldsymbol{X})$$

  where each f_j is made up of a subset (though not necessarily just one) covariate

References

Banerjee, A., Dunson, D. B. and Tokdar, S. T. (2013). Efficient gaussian process regression for large datasets. *Biometrika* **100** 75–89.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 825–848.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association* **88** 881–889.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* **103** 1119–1130.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E. et al. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics* **44** 907–927.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association* **83** 1023–1032.

Mitra, R. and Dunson, D. (2010). Two-level stochastic search variable selection in glms with missing predictors. *The international journal of biostatistics* **6**.

Qamar, S. and Tokdar, S. T. (2014). Additive gaussian process regression. *arXiv preprint arXiv:1411.7009* .

Reich, B. J., Storlie, C. B. and Bondell, H. D. (2009). Variable selection in bayesian smoothing spline anova models: Application to deterministic computer codes. *Technometrics* **51** 110–120.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* 2587–2619.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R. et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics* **42** 1166–1202.