Revisiting the Semi-parametric Latent Factor Model with Regression Trees

Sameer K. Deshpande

March 26, 2019

1 General Setup

As a motivating example, consider modeling q physiological time series, which may be highly interdependent and which may be irregularly sampled (i.e. at any one time we may not observe realizations from each series). The main goal is to impute the value of each time series, to forecast each series several steps into the future, and to provide honest uncertainty quantification about these projections.

Formally, suppose we observe triplets of data $(\mathbf{x}_1, \mathbf{y}_1, \delta_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n, \delta_n)$ where $\mathbf{x}_i \in \mathbb{R}^p$ are covariates (possibly time-dependent), $\mathbf{y}_i \in \mathbb{R}^q$ are the noisy observations from each series, and $\delta_i \in \{0, 1\}^q$ is a vector of indicators with $\delta_{i,k} = 1$ if and only if $y_{i,k}$ is observed. For now, we assume that δ is deterministic. [skd]: in the context of medical time series, this assumption is not particularly realistic. To wit, one series may track the results of a particular diagnostic test that is run only if other vital signs are in some critical region. That is, it's not unrealistic to believe that we observe one series based on the values of previous series. We'll return to this issue later

Independent of δ_i, we model for each $i = 1, \ldots, n$ and $k = 1, \ldots, q$

$$y_{i,k} = f_k(\mathbf{x}_i) + \sigma_k(\mathbf{x}_i)\varepsilon_{i,k}$$

where $\mathbf{f} = (f_1, \ldots, f_q)$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$ are unknown vector valued functions of $\mathbf{x}$ and the $\varepsilon_{i,k}$'s are independent standard normals. For now, we focus on the homoskedastic case where the residual variances are constant and do not depend on inputs $\mathbf{x}_i$; we will return to the heteroskedastic case later.

We take a Bayesian approach, which amounts to specifying a prior over the functions $\mathbf{f}$ and $\boldsymbol{\sigma}$ and updating them with Bayes' theorem to get posterior distributions that reflect our uncertainty about them in light of the data. In principle, we can attempt to learn each f_k independently of one another in an embarrassingly parallel manner. This, of course, precludes "sharing of statistical strength" and we can quite reasonably expect better predictive performance if we take advantage of the potential correlations between the outcomes (see, e.g., Breiman and Friedman, 1997). Improving prediction of one outcome dimension using information from some or all of the other outcome dimensions has been well-studied in machine learning community, under the name of "transfer learning" or "multi-task learning." We borrow from this literature, and refer to each outcome dimensions as a task below.

This is a growing literature that use multi-task GPs regression to model multiple physiological time series (see, e,g., Clifton et al. (2012), Ghassemi et al. (2015), Durichen et al. (2015), Futoma et al. (2017), Colopy et al. (2018), and Cheng et al. (2018)). Essentially, all of these papers represent each underlying f_k as a linear combination of a latent set of independent univariate basis functions, which are assigned GP priors. In this note, we revisit the semi-parametric latent factor model (SLFM) of Teh et al. (2005) but construct our latent basis using additive regression trees in the style of Chipman et al. (2010) instead of realizations of GPs.

2 Proposed Model

2.1 Brief Review of BART

Chipman et al. (2010) consider the single task regression problem, modeling $y_i = f(\mathbf{x}_i) + \sigma\epsilon_i, \epsilon_i \sim N(0, 1)$ and approximate the unknown function f as a sum of regression trees. To set our notation, let T denote a binary decision tree partitioning $\mathbb{R}^p$ that consists of a collection of interior nodes and $L(T)$ terminal or *leaf* nodes. We associate an axis-aligned decision rule of the form $\{x_j < c\}$ or $\{x_j \geq c\}$ to each internal (i.e. non-leaf) node of T. T defines a partition of $\mathbb{R}^p$ into $L(T)$ rectangular cells and we let $\ell(\mathbf{x}, T)$ be the function that returns the index of the cell containing the point $\mathbf{x}$. A *regression* tree is a pair (T, M) consisting of a decision tree T and collection $M = \{\mu_1, \dots, \mu_{L(T)}\}$ of parameters corresponding to each leaf of T. We define the evaluation function $g(\mathbf{x}; T, M) = \mu_{\ell(\mathbf{x},T)}$ which takes as input a point $\mathbf{x} \in \mathbb{R}^d$ and returns the leaf parameters corresponding to the partition cell containing $\mathbf{x}$.

At the heart of BART is the approximation

$$f(\mathbf{x}) \approx \sum_{t=1}^{m} g(\mathbf{x}; T_{(t)}, M_{(t)})$$

and a prior over regression trees $\Pi(T, M)$. By modeling each (T_t, M_t) as *a priori* independent realizations from $\Pi(T, M)$, BART implicitly induces a prior over the space of functions from $\mathbb{R}^p \to \mathbb{R}$. The regression tree prior $\Pi(T, M)$ consists of two parts, a prior $\Pi(T)$ over the space of decisions trees T and a conditional prior $\Pi(M|T)$ of leaf parameters given the decision tree topology. Conditional on the decision tree T, the associated leaf parameters are modeled as i.i.d. $N(\mu_\mu, \sigma^2_\mu)$:

$$\Pi(M|T) = \prod_{\ell=1}^{L(T)} N(\mu_\mu, \sigma^2_\mu).$$

The decision tree prior $\Pi(T)$ corresponds to a branching process and can be described in two parts: the probability that a node at depth d is internal and a distribution over the decision rule at each internal node. Specifically, the probability that node at depth d is internal is $\alpha(1 + d)^{-\beta}$ and conditional on a node being internal, the splitting rule is picked uniformly from the set of all available splitting rules. Together, these parts induce a prior over the space of functions $f : \mathbb{R}^d \to \mathbb{R}$ and we will write $f \sim \text{BART}(m, \alpha, \beta, \mu_\mu, \sigma_\mu)$. Chipman et al. (2010) complete their prior specification by placing a scaled inverse-χ^2 prior over the residual variance $\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}$. We will denote the induced prior on f as $f \sim \text{BART}(m, \alpha, \beta, \mu_\mu, \sigma^2_\mu)$.

Key to the success of BART over a wide variety of applied problems has been the existence of useful *default* choices of the associated hyperparameters. Chipman et al. (2010) recommended setting $\alpha = 0.95$ and $\beta = 2$, which essentially regularizes the depth of the decision trees $T_1, \dots, T_m$ in the BART ensemble. Now that the prior marginal distribution of $\mathbf{f}(\mathbf{x})$ is $N(m\mu_\mu, m\sigma^2_\mu)$. Upon centering and scaling the observed outcomes, we would like this prior to assign substantial prior probability to the range of the standardized data. To this end, Chipman et al. (2010) takes $\mu_\mu = 0$ and set

$$\sigma_\mu = \frac{Y_{\max} - Y_{\min}}{2\kappa\sqrt{m}}$$

where $Y_{\max}$ and $Y_{\min}$ are the maximum and minimum values of the standardized responses. With these choices, the extremes of the observed data, $Y_{\max}$ and $Y_{\min}$ are within 2κ marginal prior standard deviations of $f(\mathbf{x})$ Chipman et al. (2010) recommended $\kappa = 2$ as a good default value. All together, the choices $\alpha = 0.95, \beta = 2, \mu_\mu = 0$ and $\sigma_\mu = 0.5 \times \kappa^{-1} m^{-1/2}(Y_{\max} - Y_{\min})$ regularize the regression trees in the BART ensembles so that they are not too deep and so that no individual tree accounts for too large a portion of the variation in the observed data. Put another way, BART seeks to fit the observed data well using an ensemble of "weak learners." With these default choices in mind, abusing our notation slightly, we will take $f \sim \text{BART}(m, \sigma_\mu^2)$ to be a shorthand for $f \sim \text{BART}(m, 0.95, 2, 0, \sigma_\mu^2)$.

2.2 Proposed Model

Throughout, we will assume that the observed data from each task has been centered and scaled to have standard deviation one. In order to perform this scaling, we will require that we observe at least 2 realizations from each task [skd]: This is one disadvantage vis-a-vis using GPs.

In order to induce dependence between the f_k's, we follow the basic idea of Teh et al. (2005) and introduce independent latent basis functions $u_1, \ldots, u_D$ and express each f_k as a linear combination of these basis elements:

$$f_k(\mathbf{x}) = \sum_{d=1}^{D} \phi_{k,d} u_d(\mathbf{x}).$$

More compactly, we will write $\mathbf{f}(\mathbf{x}) = \Phi \mathbf{u}(\mathbf{x})$ where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_q(\mathbf{x}))$ and $\mathbf{u}(\mathbf{x}) = (u_1(\mathbf{x}), \ldots, u_D(\mathbf{x}))$ and $\Phi = (\phi_{k,d}) \in \mathbb{R}^{q \times D}$ records how much each task f_k depends on the basis element u_d. We place independent $\text{BART}(m, \sigma_\mu^2)$ priors on each of the basis elements $u_1, \ldots, u_D$ so that each task f_k is now approximated by a weighted sum of regression trees, which are shared across tasks. This sharing of regression trees induces *a priori* correlation between the tasks. To see this, first note that conditional on Φ,

$$\begin{aligned}
\text{Cov}(f_k(\mathbf{x}), f_{k'}(\mathbf{x}) | \Phi) &= \sum_{d=1}^{D} \phi_{k,d} \phi_{k',d} \text{Var}(u_d(\mathbf{x})) \\
&= m\sigma_\mu^2 \Phi_{k,\cdot} \Phi_{k',\cdot}^\top.
\end{aligned}$$

where $\Phi_{k,\cdot}$ is the $1 \times D$ k$^{\text{th}}$ row vector of Φ.

It is worth noting that the first line above holds regardless of the prior placed on the basis elements $u_1, \ldots, u_D$, while the second line follows directly from the fact that the u_d's are iid BART(m, σ_μ^2). Now so long as the prior on Φ assigns positive probability to the event $\Phi_{k,\cdot}^\top \Phi_{k',\cdot} \neq 0$, $f_k(\mathbf{x})$ and $f_{k'}(\mathbf{x})$ will be marginally correlated as well.

Just like with univariate BART, we aim to set useful default hyperparameter specifications. To this end, we start with $\sigma_\mu = \frac{Y_{\max} - Y_{\min}}{2\kappa\sqrt{mD}}$ where κ is to be specified, where $Y_{\max}$ and $Y_{\min}$ are the maximum and minimum standardized observations across all observations.

We place independent scaled inverse-χ^2 priors on the residual variances $\sigma_1^2, \ldots, \sigma_q^2$, with $\sigma_k^2 \sim \text{Inv. Gamma}\left(\frac{\nu}{2}, \frac{\nu\lambda_k}{2}\right)$. We use Chipman et al. (2010)'s default choice of $\nu = 3$ and pick λ_k so that *a priori* $\mathbb{P}(\sigma_k < \hat{\sigma_k}) = 0.9$ where $\hat{\sigma}_k$ is an initial over-estimate of the residual standard deviation for task k. That is, we take $\lambda_k = \nu^{-1} q_{\nu,0.1} \hat{\sigma}^2$ where $q_{\nu,0.1}$ is the 10% quantile of the χ^2_ν distribution. For $\nu = 3$ and $\hat{\sigma}_k = 1$, which is a reasonable over-estimate of the residual variance once we standardize all responses, we have $\lambda_k \approx 0.195$.

It remains to specify the prior on Φ and to set the number of basis elements D and the number of trees within each basis element m. Recall that one of the key feature of univariate BART is the fact that the marginal prior on $f(\mathbf{x})$ gave substantially probability to the range of the observed responses. While this type of calibration does sacrifice Bayesian coherence, it does ensure that the prior on f is not grossly in conflict with the data. To achieve a similar calibration, notice that conditional on Φ,

$$\mathbf{f}(\mathbf{x}) | \Phi \sim N_q(0_q, m\sigma_\mu^2 \Phi\Phi^\top).$$

One way to ensure that the marginal prior on f_k covers the range of observed realizations of the k$^{\text{th}}$ task is to require

$$\sqrt{m}\sigma_\mu \|\Phi_{k,\cdot}\|_2 \ll Y_{k,\max} - Y_{k,\min}$$

with high prior probability. Recalling our choice of $\sigma_\mu = \frac{Y_{\max} - Y_{\min}}{2\kappa\sqrt{mD}}$, we can achieve this desideratum if

$$\|\Phi_{k,\cdot}\|_2 \leq \sqrt{D} \times \frac{Y_{k,\max} - Y_{k,\min}}{Y_{\max} - Y_{\min}}, \tag{1}$$

with high prior probability.

As a starting point, consider the relatively simply prior $\phi_{k,d} \sim N(0, \sigma_{\phi,k}^2)$. With this choice, we note that $\|\Phi_{k,\cdot}\|_2^2 \sim \sigma_{\phi,k}^2 \chi_D^2$, so one way to achieve (1) *marginally* for each $k = 1, \ldots, q$ is to set

$$\sigma_{\phi,k} = \left(\frac{D}{q_{D,0.9}} \right)^{1/2} \times \frac{Y_{k,\max} - Y_{k,\,min}}{Y_{\max} - Y_{\min}}$$

where $q_{D,0.9}$ is the 90% quantile of a χ_D^2 distribution. Note that we can easily use a different quantile.

A slightly stronger requirement may be to satisfy (1) simultaneously for each task, which motivates taking

$$\sigma_{\phi,k} = \left(\frac{D}{q_{D,0.9^{1/q}}} \right)^{1/2} \times \frac{Y_{k,\max} - Y_{k,\min}}{Y_{\max} - Y_{\min}}.$$

It is worth pointing out that as $D \to \infty$, the ratio $\frac{D}{q_{D,\tilde{\alpha}}} \leftrightarrow 1$ for any $\tilde{\alpha} > 0.5$ We note that larger choices of $\sigma_{\phi,k}^2$ induce *less* regularization on the fit of f_k since it disperses more marginal probability mass of f_k outside the interval $[Y_{k,\min}, Y_{k,\max}]$.

In a certain sense, our goal is to find a D-dimensional representation of the q-dimensional tasks. Up to this point, we have described only fixing D to some, possibly larger value. Another goal may be to learn a lower dimensional representation of the q tasks. To achieve this, consider the following spike-and-slab prior:

$$\begin{aligned} \phi_{k,d} | \gamma_{k,d} &\sim \gamma_{k,d} N(0, \sigma_{\phi,k}^2) + (1 - \gamma_{k,d}) \delta_0 \\ \gamma_{1,d}, \ldots, \gamma_{q,d} | \theta_d &\overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\theta_d) \;\text{ for } d = 1, \ldots, D \\ \theta_1, \ldots, \theta_D &\sim \text{Beta}(a, b) \end{aligned}$$

Conditional on $\Phi_{k,\cdot}$, we still have that $f_k(\mathbf{x}) | \Phi_{k,\cdot} \sim N(0, \sigma_\mu^2 \|\Phi_{k,\cdot}\|_2^2)$. However, the random variable $\|\Phi_{k,\cdot}\|_2^2$ is now a scaled χ^2-random variable with a *random* number of degrees of freedom. In light of this, if we took σ_ϕ as before, the value of $\sigma_\mu^2 \|\Phi_{k,\cdot}\|_2^2$ will tend to be much smaller, resulting in much more aggressive shrinkage of f_k to $\overline{Y}_k$. Instead, we can set

$$\sigma_{\phi,k} = \left(\frac{D}{q_{aD/(a+b),0.9}} \right)^{1/2} \times \frac{Y_{k,\max} - Y_{k,\,min}}{Y_{\max} - Y_{\min}}.$$

2.3 Connection to Existing Work

Our model is closely related to (and in fact, inspired by) the semi-parametric latent factor (SFLM) of Teh et al. (2005), who introduce independent basis functions $u_1, \ldots, u_D$ and loading matrix $\Phi \in \mathbb{R}^{q \times D}$ and model $\mathbf{f} = \Phi \mathbf{u}$. In that work, they place Gaussian process priors on the basis functions, $u_d \sim \text{GP}(k_d)$ while have placed BART priors on them. In fact, the SLFM is a special case of the liner model of coregionalization (LMC) that is commonly used for fitting multi-output GPs (see, e.g. Álvarez et al., 2012). At a high-level, the LMC works by introducing several independent collections of basis functions which are independent draws from a common GP prior and then expressing the functions $f_1, \ldots, f_q$ Typically, the

In the latent factor model $\mathbf{f} = \Phi \mathbf{u}$, we immediately compute the cross-covariance

$$\text{Cov}(f_k(\mathbf{x}), f_{k'}(\mathbf{x}') | \Phi) = \sum_{d=1}^{D} \phi_{k,d} \phi_{k'd} \text{Cov}(u_d(\mathbf{x}), u_d(\mathbf{x}')).$$

When the basis functions are drawn from Gaussian process priors, $\text{Cov}(u_d(\mathbf{x}), u_d(\mathbf{x}'))$ is just an evaluation of the relevant kernel function. In fact, even in our setting, this covariance also has a convenient kernel representation. Specifically, if $u \sim \text{BART}(m, \sigma_\mu^2)$, then

$$\text{Cov}(u(\mathbf{x}), u(\mathbf{x}')) = m\sigma_\mu^2 \mathbb{P}(\mathbf{x} \sim \mathbf{x}') := m\sigma_\mu^2 k_{\text{BART}}(\mathbf{x}, \mathbf{x}')$$

where we write $\mathbf{x} \sim \mathbf{x}'$ iff $\mathbf{x}$ and $\mathbf{x}'$ are assigned to the same partition cell in a randomly drawn decision tree. This connection between BART and kernel methods was first made in Linero (2017), who further stated a heuristic theorem that so long as $m\sigma_\mu^2 \to \tilde{\sigma}_\mu^2$ when $m \to \infty$, a realization from a $\text{BART}(m, \sigma_\mu^2)$ prior converge weakly to a realization from a $\text{GP}(\tilde{\sigma}_\mu^2 k_{\text{BART}})$ prior. So in a certain sense, we may view posterior inference with BART as approximating posterior inference with a $\text{GP}(\tilde{\sigma}_\mu^2 k_{\text{BART}})$ prior.

Linero et al. (2018) is closest in spirit to our model. In that work, they model multiple tasks using a common set of regression trees but introduce a different set of leaf parameters for each task. For a given leaf, the q associated parameters are drawn from a multivariate prior, thereby inducing dependence between the modeled tasks. [skd]: I haven't thought about this too much yet but this model might be a really special case of the one we've described above. In any case, I think the correlation between tasks is fixed throughout and is not learned as it is in our proposal.

3 A Gibbs sampler

Given the data $\mathbf{y}$ and our prior specification, we have a posterior

$$\Pi((T_1^{(1)}, M_1^{(1)}), \ldots, (T_m^{(d)}, M_m^{(d)}), \Phi, \boldsymbol{\sigma} | \mathbf{y})$$

on all of the unknown parameters in our semi-parametric latent factor model. We sample from this posterior using a slight elaboration on the backfitting procedure of Chipman et al. (2010). Like their procedure, our's is, at a high level, a Gibbs sampler, that iterates between sequentially updating the basis functions $u_1, \ldots, u_D$, the loading matrix Φ, and the residual variances $\boldsymbol{\sigma}^2$, while keeping the other parameters fixed. We describe each of these conditional updates in the next three subsections.

3.1 Updating the basis functions

Keeping Φ and $\boldsymbol{\sigma}$ fixed, we update each regression tree sequentially holding the remaining $mD - 1$ trees fixed. Each update of a regression tree proceeds in two steps: first, we update the decision tree with a Metropolis-Hastings steps and then draw new leaf parameters conditional on this new decision tree. As we detail below, the MH step is facilitated by an easy-to-compute acceptance probability and the leaf parameters are updated in conjugate fashion.

Recall for each $i = 1, \ldots, n$, we have the triplet $(\mathbf{x}_i, \mathbf{y}_i, \delta_i)$ with likelihood model

$$p(\mathbf{y}_i | \Phi, \mathbf{x}_i, \delta_i, \mathbf{u}, \boldsymbol{\sigma}^2) = \prod_{k=1}^{q} \left[\left(2\pi\sigma_k^2\right)^{-\frac{\delta_{i,k}}{2}} \exp\left\{ -\frac{\delta_{i,k} \left(y_{i,k} - \sum_{d=1}^{D} \phi_{k,d} u_d(\mathbf{x}_i)\right)^2}{2\sigma_k^2} \right\} \right],$$

where $\delta_{i,k} = 1$ if we observe $y_{i,k}$ and 0 otherwise.

Suppose we are updating $(T_t^{(d)}, M_t^{(d)})$, the t^{th} tree of the d^{th} basis function. Define

$$r_{i,k} = y_{i,k} - \sum_{d' \neq d} \phi_{k,d'} u_{d'}(\mathbf{x}_i) - \phi_{k,d} \sum_{t' \neq t} g(\mathbf{x}_i; T_t^{(d)}, M_t^{(d)})$$

when $\delta_{i,k} = 1$ and $r_{i,k} = 0$ otherwise. When $\delta_{i,k} = 1$, $r_{i,k}$ is the partial residual based on the fit of the remaining $mD - 1$ trees. Note that when we do not observe $y_{i,}$ (i.e. $\delta_{i,k} = 0$), the specific value of $r_{i,k}$ is immaterial and we have, for convenience, set it to zero. Given the remaining $mD-1$ trees, $\Phi, \boldsymbol{\sigma}$, and $\mathbf{Y}$, the collection of partial residuals $\{r_{i,k}\}$ are sufficient for $(T_t^{(d)}, M_t^{(d)})$:

$$\begin{aligned} p(T, M|\mathbf{y}, \Phi, \boldsymbol{\sigma}, ...) = & \prod_{\ell=1}^{L(T)} \prod_{i \in I(\ell,T)} \prod_{k=1}^{q} \left(2\pi\sigma_k^2\right)^{-\frac{\delta_{i,k}}{2}} \exp\left\{-\frac{(r_{i,k}-\phi_{k,d}\mu_\ell)^2}{2\sigma_k^2}\right\} \\ & \times \prod_{\ell=1}^{L} \left(2\pi\sigma_\mu^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{\mu_\ell^2}{2\sigma_\mu^2}\right\} \\ & \times p(T) \end{aligned}$$

We may write this somewhat more concisely as

$$\begin{aligned} p(T, M|\mathbf{y}, \Phi, \boldsymbol{\sigma}, ...) = p(T) \times & \prod_{\ell=1}^{L(T)} \left[\left(2\pi\sigma_\mu^2\right)^{-\frac{1}{2}} \times \prod_{k=1}^{q} \left(2\pi\sigma_k^2\right)^{-\frac{n_{\ell,k}(T)}{2}}\right] \\ & \times \prod_{\ell=1}^{L(T)} \exp\left\{-\frac{1}{2}\left[\sigma_\mu^{-2}\mu_\ell^2 + \sum_{k=1}^{q}\sum_{i\in I(\ell,T)} \delta_{i,k}\sigma_k^{-2}(r_{i,k}-\phi_{k,d}\mu_\ell)^2\right]\right\} \end{aligned}$$

where $n_{\ell,k}$ counts the number of observations of task k that are associated to leaf ℓ (i.e. $n_{\ell,k} = \sum_{i\in I(\ell,T)} \delta_{i,k}$).

Completing the square, we compute

$$\sigma_\mu^{-2}\mu_\ell^2 + \sum_{k=1}^{q}\sum_{i\in I(\ell,T)} \delta_{i,k}\sigma_k^{-2}(r_{i,k}-\phi_{k,d}\mu_\ell)^2 = V_\ell^{-1}(\mu - M_\ell)^2 + \sum_{i\in I(\ell,T)}\sum_{k=1}^{q} \delta_{i,k}\sigma_k^{-2} r_{i,k}^2 - V_\ell^{-1}M_\ell^2,$$

where

$$\begin{aligned} V_\ell &= \left(\sigma_\mu^{-2} + \sum_{k=1}^{q} n_{\ell,k}\phi_{k,d}^2\sigma_k^{-2}\right)^{-1} \\ M_\ell &= V_\ell \times \sum_{i\in I(\ell,T)}\sum_{k=1}^{q} \delta_{i,k}\phi_{k,d} r_{i,k}\sigma_k^{-2}. \end{aligned}$$

Integrating out the μ_ℓ,'s, we have $p(T|\mathbf{y}, \Phi, \boldsymbol{\sigma}, \ldots) \propto$

$$p(T) \times \prod_{\ell=1} \left[V_\ell^{\frac{1}{2}} \sigma_\mu^{-1} \times \left(\prod_{k=1}^{q} (2\pi\sigma_k^2)^{-\frac{n_{\ell,k}}{2}} \right) \exp\left\{ \frac{1}{2} \left[V_\ell^{-1} M_\ell^2 - \sum_{i \in I(\ell,T)} \sum_{k=1}^{q} \delta_{i,k} \phi_{k,d} r_{i,k}^2 \right] \right\} \right]$$

Many of the terms above can actually be absorbed into a constant that does not depend on T, leaving us with the less cluttered expression

$$p(T|\mathbf{y}, \Phi, \boldsymbol{\sigma}, \ldots) \propto p(T) \times \prod_{\ell=1}^{L(T)} V_\ell^{1/2} \sigma_\mu^{-1} \exp\left\{ \frac{V_\ell^{-1} M_\ell^2}{2} \right\}$$

We also note that

$$p(M|T, \mathbf{y}, \Phi, \boldsymbol{\sigma}, \ldots) \propto \prod_{\ell=1}^{L(T)} \exp\left\{ -\frac{(\mu_\ell - M_\ell)^2}{2V_\ell} \right\},$$

enabling us to draw the new leaf parameters $\mu_\ell \sim N(M_\ell, V_\ell)$.

Given a decision tree T, we update it with a simplified version of the transition kernel introduced in Chipman et al. (1998). Specifically, we propose a new tree T^* by either growing T at one terminal node or by pruning T by collapsing two adjacent leafs to their common parent. For a grow proposal T^*, let M_0 and V_0 be the posterior mean and variance of the leaf parameter corresponding to the leaf in T that is split to produce two children in T^*. Similarly, let M_l, M_r, V_l and V_r be the posterior means and variances of the leaf parameter corresponding to these two new leaves in T^*. Then the MH ratio is given by

$$\alpha(T^*, T) = \min\left\{ 1, \frac{q(T^*|T)p(T^*)}{q(T|T^*)p(T)} \times \frac{\left(V_l^{\frac{1}{2}} \sigma_\mu^{-1} \exp\left\{ \frac{M_l^2}{2V_l} \right\} \right) \left(V_r^{\frac{1}{2}} \sigma_\mu^{-1} \exp\left\{ \frac{M_r^2}{2V_r} \right\} \right)}{V_0^{\frac{1}{2}} \sigma_\mu^{-1} \exp\left\{ \frac{M_0^2}{2V_0} \right\}} \right\}$$

We can write down a similar expression for the acceptance probability for prune proposals.

3.2 Updating Φ

We now describe the updates of Φ when we have a Gaussian or spike-and-slab prior on $\phi_{k,d}$.

3.2.1 Gaussian Prior

Observe that

$$\begin{aligned}p(\Phi_{k,\cdot}|\mathbf{u},\mathbf{y},\boldsymbol{\sigma}^2) &\propto \exp\left\{-\frac{1}{2\sigma_{\phi,k}^2}\sum_{d=1}^{D}\phi_{k,d}^2 - \frac{1}{2\sigma_k^2}\sum_{i=1}^{n}\delta_{i,k}\left(y_{i,k}-\sum_{d=1}^{D}\phi_{k,d}u_d(\mathbf{x}_i)\right)^2\right\}\\ &\propto \exp\left\{-\frac{1}{2}\left[\Phi_{k,\cdot}^\top V_\Phi^{-1}\Phi_{k,\cdot} - 2\sigma_k^{-2}\sum_{i=1}^{n}\delta_{i,k}y_{i,k}\Phi_{k,\cdot}^\top\mathbf{u}(\mathbf{x}_i)\right]\right\}\end{aligned}$$

where $V_\Phi = \left[\sigma_{\phi,k}^{-2}I_d + \sigma_k^{-2}\sum_{i=1}^n \delta_{i,k}\mathbf{u}(\mathbf{x}_i)\mathbf{u}^\top(\mathbf{x}_i)\right]^{-1}$. From here, we immediately conclude that $\Phi_{k,\cdot} \mid \mathbf{y},\mathbf{u},\boldsymbol{\sigma} \sim N(M_\Phi, V_\Phi)$ where

$$M_\Phi = \sigma_k^{-2}V_\Phi\sum_{i=1}^{n}\delta_{i,k}y_{i,k}\mathbf{u}(\mathbf{x}_i)$$

While this expression is tidy, compute V_Φ and M_Φ at each iteration of our Gibbs sampler may be slightly tedious.

Alternatively (and more similar to what is developed in the next subsection), we may update each $\phi_{k,d}$ sequentially holding the remaining elements constant. To this end, define $b_i = y_{i,k} - \sum_{d'\neq d}\phi_{k,d'}u_{d'}(\mathbf{x}_i)$. We have

$$\begin{aligned}p(\phi_{k,d}|\mathbf{u},\mathbf{y},\boldsymbol{\sigma}^2,\Phi_{-k,-d}) &\propto \exp\left\{-\frac{1}{2}\left[\sigma_{\phi,k}^{-2}\phi_{k,d}^2 - \sigma_k^{-2}\sum_{i=1}^{n}\delta_{i,k}(b_i-\phi_{k,d}u_d(\mathbf{x}_i))^2\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}\left[\phi_{k,d}^2 v_\phi^{-1} - 2\phi_{k,d}\sigma_k^{-2}\sum_{i=1}^{n}\delta_{i,k}b_iu_d(\mathbf{x}_i)\right]\right\}\end{aligned}$$

where $v_\phi^{-1} = \sigma_{\phi,k}^{-2} + \sigma_k^{-2}\sum_{i=1}^n \delta_{i,k}u_d(\mathbf{x}_i)^2$. Hence the conditional posterior distribution of $\phi_{k,d}$ given all other parameters is $N(m_\phi, v_\phi)$ where

$$m_\phi = v_\phi \times \sigma_k^{-2}\sum_{i=1}^{n}\delta_{i,k}b_iu_d(\mathbf{x}_i)$$

We compute

$$\sum_{i=1}^{n} \delta_{i,k}(b_i - \phi_{k,d}u_d(\mathbf{x}_i))^2 = \sum_{i=1}^{n} \delta_{i,k}(b_i^2 - 2b_i u_d(\mathbf{x}_i)\phi_{k,d} - \phi_{k,d}^2 u_d(\mathbf{x}_i))$$

3.2.2 Spike-and-slab prior

We re-parametrize like in Titsias and Lázaro-Gredilla (2011) so that $\phi_{k,d} = \gamma_{k,d}\tilde{\phi}_{k,d}$ where *a priori* $\tilde{\phi}_{k,d} \sim N(0, \sigma^2_{\phi,k})$. For each k, let $\boldsymbol{\gamma}_{k,-d}$ be the set of all indicators associated with task k except for $\gamma_{k,d}$ and similarly define $\tilde{\Phi}_{k,-d}$. We update the pairs $(\tilde{\phi}_{k,d}, \gamma_{k,d})$ holding $\mathbf{u}, \boldsymbol{\sigma}$ and $\boldsymbol{\gamma}_{k,-d}$ and $\tilde{\Phi}_{k,-d}$ fixed. To this end, let $b_i = y_{i,k} - \sum_{d' \neq d} \phi_{k,d} u_d(\mathbf{x}_i)$ and note

$$\begin{aligned} p(\tilde{\phi}_{k,d}, \gamma_{k,d}, | \cdots) \propto \exp\left\{ -\frac{1}{2\sigma_k^2} \sum_{i=1}^{n} \delta_{i,k}(b_i - \gamma_{k,d}\tilde{\phi}_{k,d}u_d(\mathbf{x}_i))^2 \right\} \\ \times \exp\left\{ -\frac{\tilde{\phi}^2_{k,d}}{2\sigma^2_{\phi,k}} \right\} \times \theta_k^{\gamma_{k,d}}(1-\theta_k)^{1-\gamma_{k,d}} \end{aligned}$$

We can define

$$\begin{aligned} v_\phi(\gamma) &= \left(\sigma_\phi^{-2} + \gamma\sigma_k^{-2} \sum_{i=1}^{n} \delta_{i,k} u_d(\mathbf{x}_i)^2 \right)^{-1} \\ m_\phi(\gamma) &= \gamma v_\phi(\gamma)\sigma_k^{-2} \sum_{i=1}^{n} b_i u_d(\mathbf{x}_i)\delta_{i,k} \end{aligned}$$

and marginalize out $\phi_{k,d}$ to obtain the conditional mass function of $\gamma_{k,d}$ given all of the other parameters

$$p(\gamma_{k,d}| \dots) \propto \theta_d^{\gamma_{k,d}}(1-\theta_d)^{1-\gamma_{k,d}} v_\phi(\gamma_{k,d})^{\frac{1}{2}} \exp\left\{ \frac{m_\phi(\gamma_{k,d})^2}{2v_\phi(\gamma_{k,d})} \right\}$$

So we may draw a new $\gamma_{k,d}$ according to this distribution. Once we do that, we can draw a new $\tilde{\phi}_{k,d} \sim N(m_\phi(\gamma_{k,d}), v_\phi(\gamma_{k,d}))$

Having updated each $\gamma_{k,d}$, we have a conjugate update for $\theta_1, \dots, \theta_d$:

$$p(\theta_1, \dots, \theta_d| \dots) \propto \prod_{d=1}^{D} \theta_d^{a_\theta - 1 + \sum_k \gamma_{k,d}}(1-\theta_d)^{b_\theta - 1 + \sum_k (1-\gamma_{k,d})}$$

[skd]: the expression of $p(\gamma | \ldots)$ above is somewhat different than what Titsias and Lázaro-Gredilla (2011) derived, which is a bit concerning... it's possible I don't follow their notation

3.3 Updating $\boldsymbol{\sigma}$

Once we have updated $\mathbf{u}$ and Φ, we can update the residual variances for each task σ_k^2. We follow Chipman et al. (2010) and model *a priori* $\sigma_k^2 \sim \lambda_k \nu / \chi_\nu^2$. In other words, we have $\sigma_k^2 \sim$ Inv. Gamma $\left(\frac{\nu}{2}, \frac{\lambda_k \nu}{2}\right)$. Just as with the rows of Φ, because our likelihood model factorizes over the tasks and because we have independent priors on the task-specific residual variances, we can update the σ_k^2's independently. Observe that

$$p(\sigma_k^2 | \mathbf{y}, \mathbf{u}, \Phi) \propto \left(\sigma_k^2\right)^{-\frac{\nu}{2}-1} \exp\left\{-\frac{\nu\lambda_k}{2\sigma_k^2}\right\} \times \prod_{i=1}^{n} \left(\sigma_k^2\right)^{-\frac{\delta_{i,k}}{2}} \exp\left\{-\frac{\delta_{i,k}(y_{i,k} - \Phi\mathbf{u}(\mathbf{x}_i))^2}{2\sigma_k^2}\right\}$$

and we immediately have

$$\sigma_k^2 | \mathbf{y}, \mathbf{u}, \Phi \sim \text{Inv. Gamma}\left(\frac{\nu + \sum \delta_{i,k}}{2}, \frac{\nu\lambda_k + \sum \delta_{i,k}(y_{i,k} - \Phi\mathbf{u}(\mathbf{x}_i))^2}{2}\right)$$

4 Illustrations

To illustrate the proposed procedure, consider a toy dataset with $n = 1000, p = 2$ and $q = 10$.

4.1 Cars Dataset

Here we predict city miles per gallon, highway mpg, and price using the remaining features. Note that most of these features are categorical and have been re-coded to 0/1 dummy variables.

Again, we see the advantage of a multivariate approach.

4.2 Foreign Exchange Data

4.3 EEG Data

4.4 ...

5 Next Directions

There are several potential directions we may pursue, summarized below

- Heteroskedasticity: in our model, we assumed the residual error in each task did not depend on $\mathbf{X}$. We can relax this and model the residual errors using the product-of-trees formulation in Pratola et al. (2017)
- Mixed type data: Frequently we observe both continuous and discrete data an wish to model both jointly. We can use ideas from Pourmohamad and Lee (2016) or Murray (2017)
- Sparsity in Φ and discovering the latent dimension D: throughout, we have focused on a fixed D, which must be specified by the user. Implicit in our development thus far is the assumption that variation in the q-dimensional response is adequately captured using D latent basis functions. In some cases, we have taken $D > q$, in the hopes that the u_d's form an over-complete basis. But in many applications, when q is larger, there may be substantial interest in learning a *lower*-dimensional representation where the latent dimension $D \ll q$ must be learned. We can elaborate the model above to accommodate this setting by placing a sparsity-inducing prior on the elements of Φ. One example is a spike-and-slab prior in the style of Titsias and Lázaro-Gredilla (2011). Rather than using a fixed Bernoulli prior for the indicators, however, we may resort to an Indian Buffet Process style prior. This will in fact place a mild identifiability constraint on the u_d's and is quite similar to ideas from sparse factor analysis.

6 Illustration

To illustrate the proposed procedure, consider the following toy dataset with $n = 1000, p = 2$ and $q = 2$. We drew the two predictors X_1 and X_2 uniformly from the unit interval and considered four true basis functions, depicted in Figure 1. All but u_2 is a function of X_1 alone and u_2 depends on X_2 only through the indicator $\mathbf{1}(X_2 > 0.5)$:

$$u_2 = 3X_1 + (2 - 5 \times \mathbf{1}(X_2 > 0.5)) \sin(\pi X_1) - 2 \times \mathbf{1}(X_2 > 0.5).$$

The functions u_3 and u_4 were drawn from Gaussian processes. u_3 was drawn from a GP with squared exponential kernel with length scale 0.25, while u_4 came from a GP whose kernel was a product of a squared exponential and periodic kernel, both with length scales 0.1.

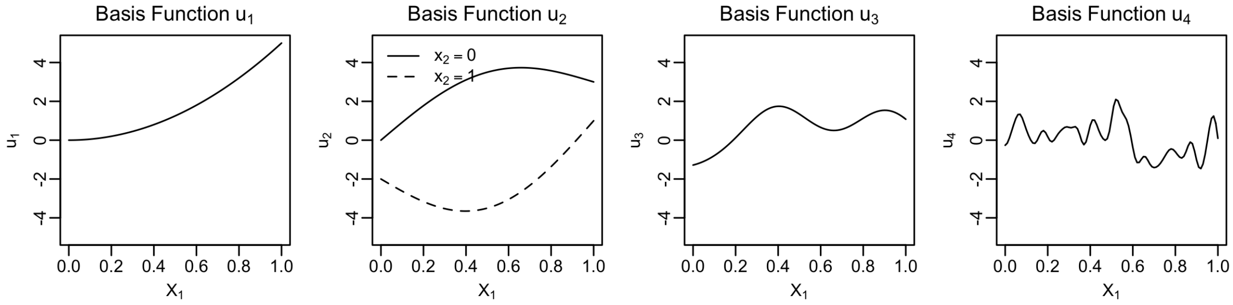

Figure 1: Basis functions used to generate toy dataset

To construct f_1 and f_2, we drew each $\phi_{k,d} \sim N(0,1)$ and then set $y_{i,1} = f_1(\mathbf{x}_i) + 0.75\epsilon_{i,1}$ and $y_{i,2} = f_2(\mathbf{x}_i) + 0.5\epsilon_{i,2}$ where $\epsilon_{i,1}, \epsilon_{i,2} \sim N(0,1)$. Figure 2 plots f_1 and f_2 along with the observed data from both tasks.

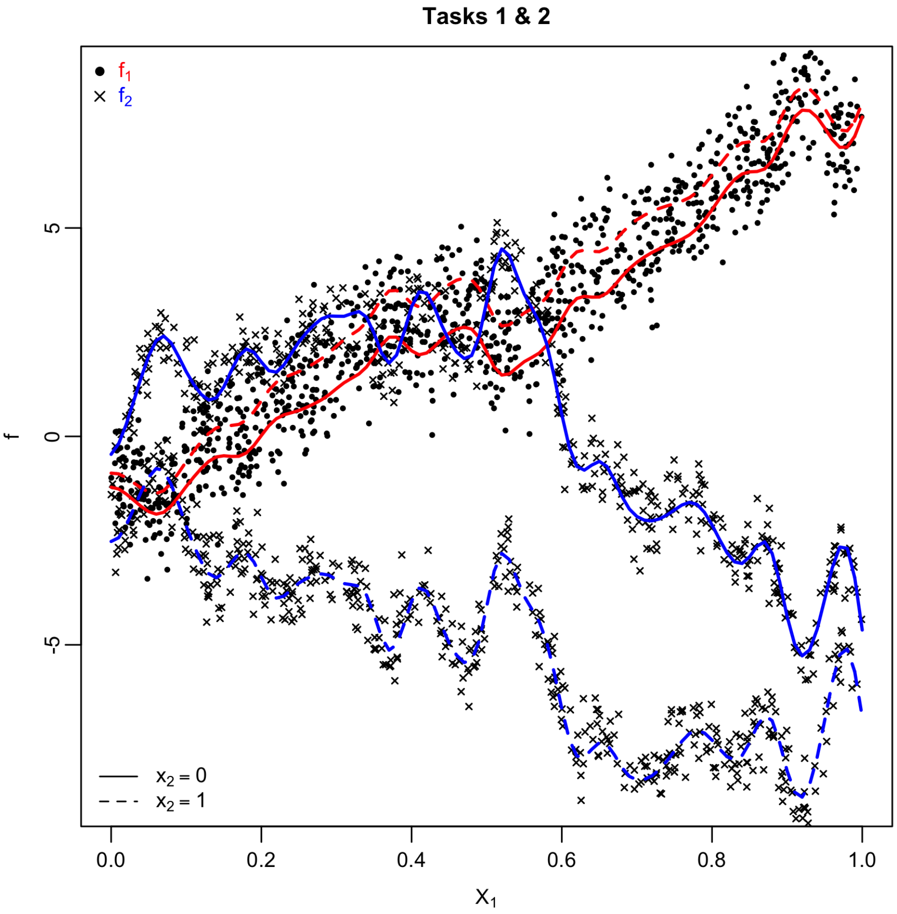

Figure 2: Two tasks, f_1 (red) and f_2 (blue). Realizations from f_1 are plotted with $\bullet$'s while realizations from f_2 are plotted with $\times$'s.

Recall that the main idea is to leverage the fact that f_1 and f_2 are dependent while learning them. As a baseline, we see how well we can learn these two functions separately, using individual BART fits. Figure 3 shows the posterior mean f_1 and f_2 using the default hyperparameter specifications.

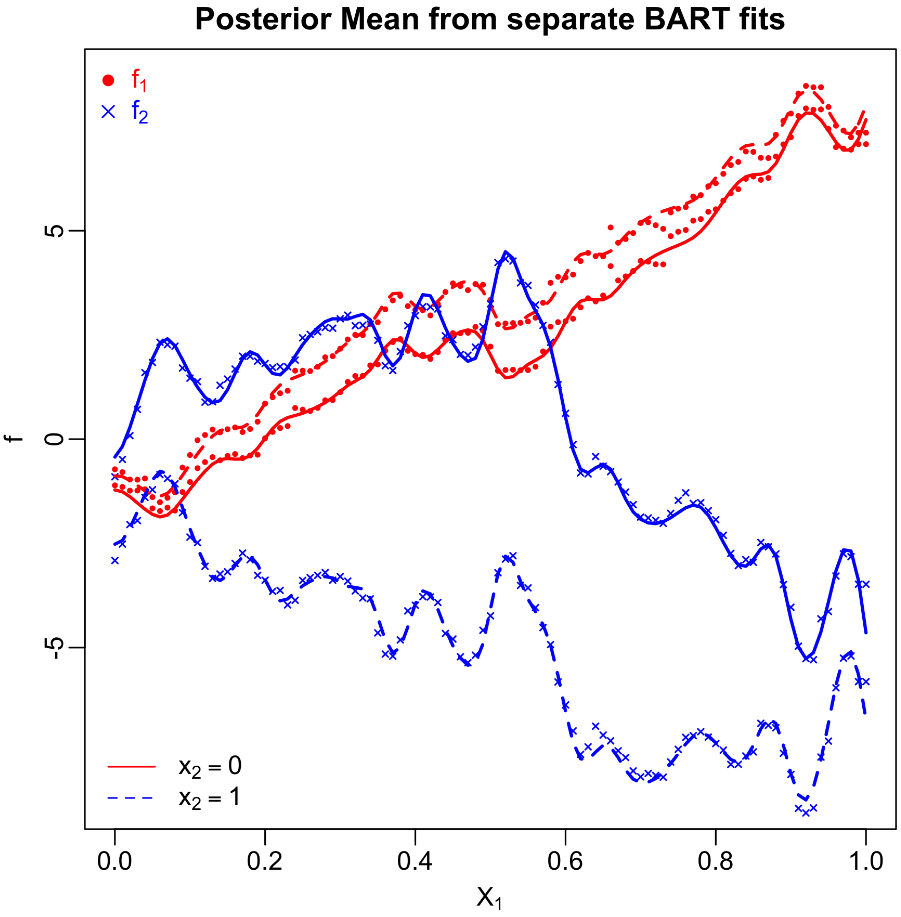

Figure 3: Points are the posterior means of f_1 and f_2 resulting from separate BART fits with default hyper-parameters. The true functions are plotted as lines.

The in-sample root mean-squared errors $\left(n^{-1}\sum\left(\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)\right)^2\right)^{1/2}$ are 0.169 for task 1 and 0.168 for task 2. The out-of-sample RMSEs are 0.181 and 0.202, respectively. The hope is that our proposed models can achieve better performance than these baselines.

References

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society (Series B)*, 59(1):3 – 54.

Cheng, L.-F., Darnell, G., Dumitrascu, B., Chivers, C., Draugelis, M. E., Li, K., and Engelhardt, B. E. (2018). Sparse multi-output gaussian processes for medical time series prediction. arXiv:1703.09112v2.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935 – 948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266 – 298.

Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J., and Tarassenko, L. (2012). Gaussian process regression in vital-sign early warning systems. In *Proceedings of the 34th Annual International Conference of the IEEE EMBS*, pages 6161 – 6164.

Colopy, G. W., Roberts, S. J., and Clifton, D. A. (2018). Bayesian optimizations of personalized models for patient vital-sign monitoring. *IEEE Journal of Biomedical and Health Informatics*, 22(2):301 – 310.

Durichen, R., Pimentel, M. A., Clifton, L., Schweikard, A., and Clifton, D. A. (2015). Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering*, 62(1):314 – 322.

Futoma, J., Hariharan, S., Heller, K., Sendak, M., Brajer, N., Clement, M., Bedoya, A., and O'Brien, C. (2017). An improved multi-output gaussian process rnn with real-tile validation for early sepsis detection. In *Proceedings of Matchine Learning for Healthcare*.

Ghassemi, M., Pimentel, M. A., Naumann, T., Brennan, T., Clifton, D. A., Szolovits, P., and Feng, M. (2015). A multivariate timeseries approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proceedings of the 29th AAAI Conference on Articial Intelligence (AAAI 2015)*.

Linero, A. R. (2017). A review of tree-based Bayesian methods. *Communicatinos for Statistical Applications and Methods*, 24(6):543 – 559.

Linero, A. R., Sinha, D., and Lipsitz, S. R. (2018). Semiparametric mixed-scale models using shared bayesian forests. arXiv:1809.08521.

Murray, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count response. arXiv:1701.01503.

Pourmohamad, T. and Lee, H. K. H. (2016). Multivariate stochastic process models for correlated responses of mixed type. *Bayesian Analysis*, 11(3):797 – 820.

Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2017). Heteroscedastic bart using multiplicative regression trees. arXiv:1709.07542.

Teh, Y. W., Seeger, M., and Jordan, M. I. (2005). Semiparametric latent factor models. In *AISTATS 2005 – Proceedings of the 10th International Workshop on Artifical Intelligence and Statistics*, pages 333 – 340.

Titsias, M. K. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2339–2347. Curran Associates, Inc.