Revisiting the Semi-parametric Latent Factor Model with Regression Trees

Sameer K. Deshpande

March 31, 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.

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 involves specifying priors over the functions $\mathbf{f}$ and $\boldsymbol{\sigma}$ and updating them with Bayes' theorem to obtain a posterior distribution that reflect our

uncertainty about $\mathbf{f}$ and $\boldsymbol{\sigma}$ 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.

Gaussian processes (GPs) are a major workhorse for Bayesian non-parametric regression in the machine learning literature. A particularly popular approach to multi-output regression in this community is to represent each task f_k as a linear combination of a latent set of independent realizations from GP priors and much of the recent literature on modeling multiple physiological time series uses such models (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)). In this note, we consider an alternative, in which we represent the basis functions are regression trees in the style of Chipman et al. (2010).

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 $M = \{\mu_1, \ldots, \mu_{L(T)}\}$ 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. $N(\mu_\mu, \sigma_\mu^2)$:

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

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 $f \sim \text{BART}(m, \alpha, \beta, \mu_\mu, \sigma_\mu^2)$.

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 $T_1, \ldots, T_m$ in the BART ensemble. Now that the prior marginal distribution of $\mathbf{f}(\mathbf{x})$ is $N(m\mu_\mu, m\sigma_\mu^2)$. 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 for each $i, \sum_k \delta_{i,k} \geq 1$ (i.e. at least one observed outcome per observation) and for each $k, \sum_i \delta_{i,k} \geq 2$ (i.e. at least two observations of each task).

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

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

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. We complete our prior construction by taking $\phi_{k,d} \sim N(0, \sigma_{\phi,k}^2)$ for each k and $\sigma_k^2 \sim \text{Inv. Gamma}\left(\frac{\nu}{2}, \frac{\nu\lambda_k}{2}\right)$, where the $\sigma_{\phi,k}$'s and λ_k's are to be specified.

Observe 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^{th} row vector of Φ. Further, since $\Phi_{k,\cdot}^\top \Phi_{k',\cdot} \neq 0$ almost surely, $f_k(\mathbf{x})$ and $f_{k'}(\mathbf{x})$ will be marginally correlated *a priori*. It is worth noting that the first line above holds regardless of the prior placed on the basis elements u_d, while the second following

directly from the fact that the u_d are iid BART(m, σ_μ^2).

It remains to specify the hyper-parameters $\sigma_\mu, \sigma_{\phi,k}$ and λ_k. Recall that our aim is to provide useful *default* values so that our proposed method may be applied to a wide variety of problems. 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 tasks. We use Chipman et al. (2010)'s default choice of $\nu = 3$ and select λ_k so that *a prior* $\mathbb{P}(\sigma_k < \hat{\sigma}_k) = 0.9$, where $\hat{\sigma}_k$ is an initial over-estimate of the residual standard deviation for task k. It is straightforward to compute $\lambda_k = \nu^{-1} q_\nu(0.1)\hat{\sigma}_k^2$, where $q_\nu(0.1)$ is the 10% quantile of the χ_ν^2 distribution. As we standardize our data to have variance 1, we will always take $\hat{\sigma}_k = 1$ so that with the choice $\nu = 3$, we have $\lambda_k \approx 0.195$.

Now, conditional on Φ, we have

$$f_k(\mathbf{x}) \sim N_q(0, m\sigma_\mu^2 \|\Phi_{k,\cdot}\|_2^2).$$

Further, this conditional prior covers the range $[Y_{k,\min}, Y_{k,\max}]$ with at least 2κ standard deviations whenever

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

Following Chipman et al. (2010) we will take $\kappa = 2$. One way to ensure that the *marginal* prior of $f_k(\mathbf{x})$ assigns substantial prior probability to the range of the observed data is to ensure that the above inequality holds with high prior probability. Since $\phi_{k,d} \sim N(0, \sigma_{\phi,k}^2)$, we know $\|\Phi_{k,\cdot}\|_2^2 \sim \sigma_{\phi,k}^2 \chi_D^2$, and we set

$$\sigma_{\phi,k} = \sqrt{\frac{D}{q_D(0.9)}} \times \frac{Y_{k,\max} - Y_{k,\min}}{Y_{\max} - Y_{\min}}.$$

Note that *a priori*, the marginals of $\mathbf{f}(\mathbf{x})$ are independent. To ensure that all marginals cover the corresponding range of the observed data simultaneously with high prior probability, we can choose

$$\sigma_{\phi,k} = \sqrt{\frac{D}{q_D(0.9^{1/q})}} \times \frac{Y_{k,\max} - Y_{k,\min}}{Y_{\max} - Y_{\min}}$$

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 very close 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.

2.4 Posterior Inference

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 the Gibbs sampler in detail in Appendix 5.

3 Illustration

We now illustrate the proposed procedure, which we for now call SLFM-BART, using two benchmark datasets from the multi-output GP literature. These datasets were considered in Nguyen and Bonilla (2014) and Requeima et al. (2019). We compare the performance of SLFM-BART to fitting independent BART models.

3.1 Foreign Exchange Rate

In our first experiment, model the exchange rates between USD and three precious metals (gold, silver, and platinum) and the top 10 international currencies (CAD, EUR, JPY, GBP, CHF, AUD, HKD, NZD, KRW, and MXN). We follow the experimental setup of Álvarez et al. (2010) and use data from the 251 working days of the year 2007[1]. We remove data from the exchange rates of CAD for days 50-100, JPY for days 100-150, and AUD for days 150-200. Our goal is to recover these long contiguous realizations from the three tasks and we measure performance using standardized MSE. We fit SLFM-BART using a range of values of m and D and the suggested hyper-parameter settings above. For each combination of m and D we generated 1500 MCMC samples, discarding the first 500 as burn-in. Figure 1 shows the SMSE on each task and the overall average SMSE for each combination considered.

[1]The data is available from http://fx.sauder.ubc.ca

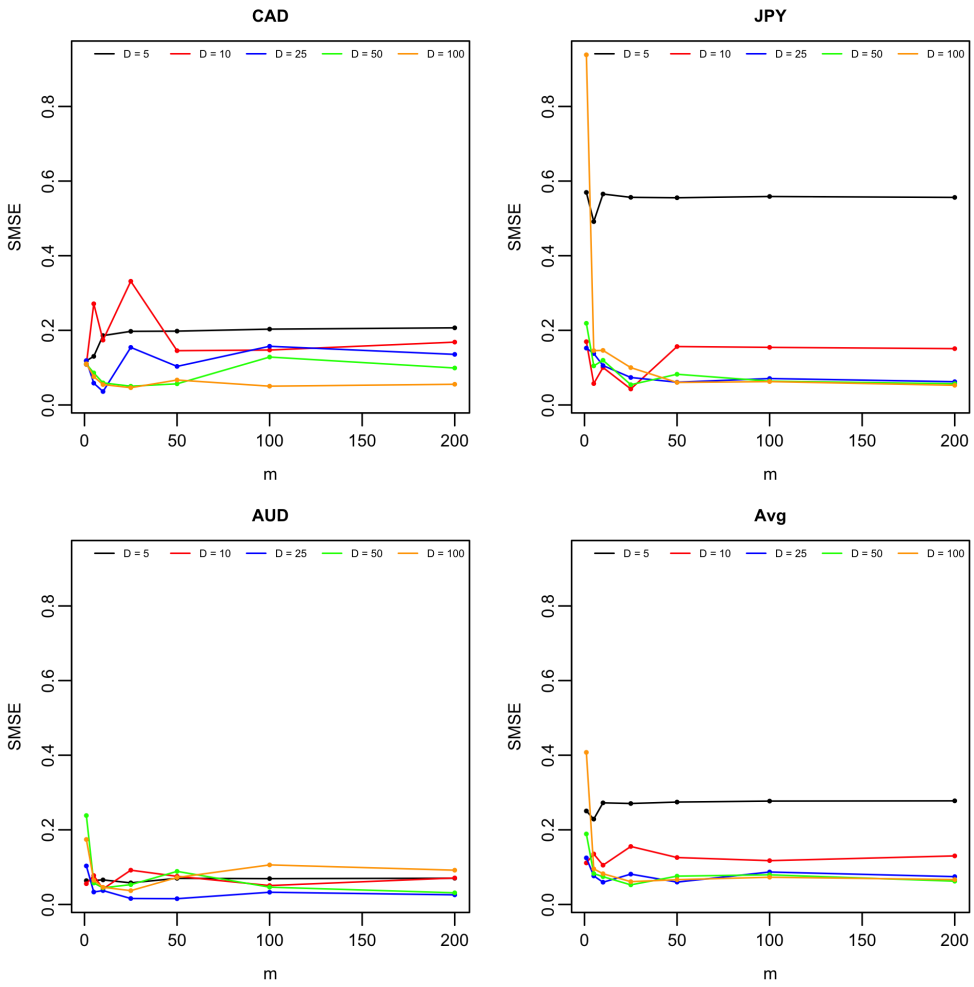

Figure 1: Standardized mean square error for SLFM-BART for several combinations of m and D.

On this dataset, we achieved the best overall average SMSE of 0.0526 using $D = 50$ basis functions consisting of $m = 25$ trees each. There does not seem to be a discernible pattern in the relationship between SMSE and D or m. Figure 2 shows the posterior predictive means and 95% prediction intervals for SLFM-BART with $(m, D) = (25, 50)$. We see immediately that both independent BART and SLFM-BART essentially interpolate the training data (gray points). Further, in general, the prediction intervals for independent BART are wider than for SLFM-BART, especially on the green testing points. Interestingly, we see that SLFM-BART and independent BART overestimate the held-out CAD data and that in fact, independent BART is less biased.

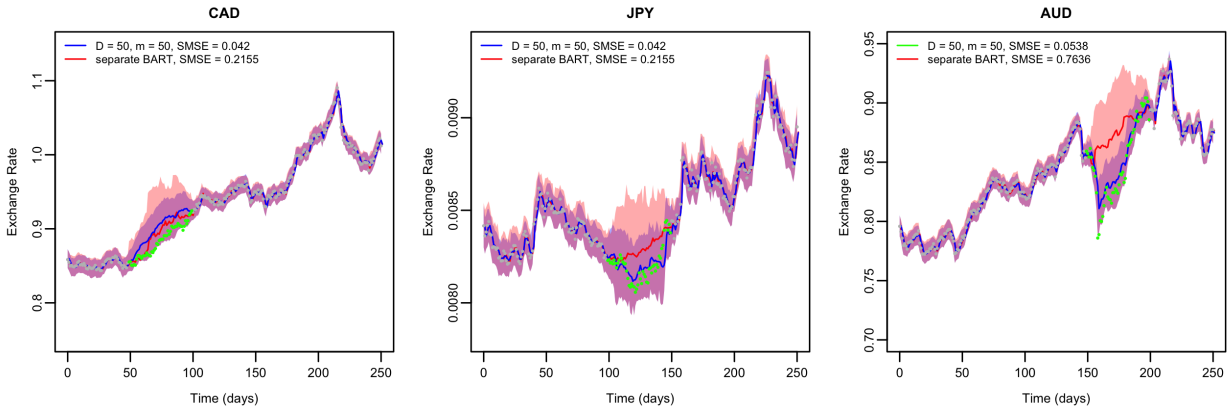

Figure 2: Posterior predictive mean and 95% intervals for CAD, JPY, and AUD for SLFM-BART with $(m, D) = (25, 50)$ (blue) and independent BART (red). Green points were held out for testing purposes.

3.2 Air Temperature

Our next experiment uses air temperature data recorded between 10 and 15 July 2013 from 4 different networks of weather sensors (named Bramblemet, Sotonment, Cambermet, and Chimet) on the south coast of England[2]. Measurements were taken every 5 minutes. Following Nguyen and Bonilla (2014), we held out contiguous observations from the Cambermet and Chimet data to serve as training. Figure 3 is the analog of Figure 1 and shows the SMSE for SLFM-BART with various combinations of m and D.

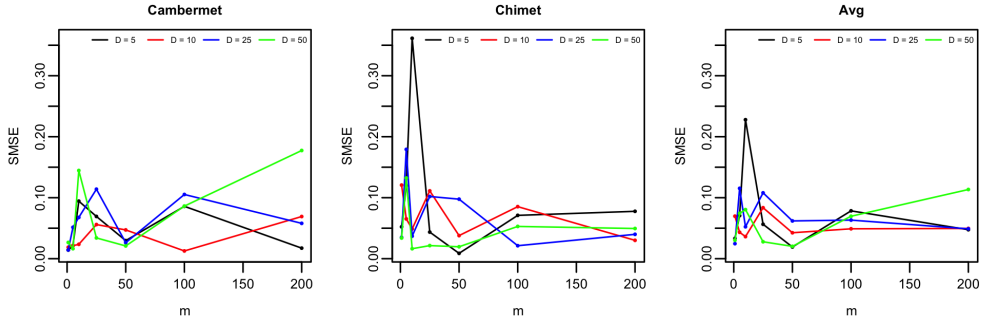

Figure 3: Standardized mean square error for SLFM-BART for several combinations of m and D.

For this dataset, we achieves the best overall SMSE of 0.0192 using $m = 50$ and $D = 5$.

[2]Data for Bramblemet and links for the other networks available www.bramblemet.co.uk and links therein

Figure 4 shows the predictive means and 95% intervals for the Chimet and Cambermet data. We saw that SLFM-BART overestimates the held-out data while independent BART underestimates the held-out data. It is quite concerning to see that the 95% intervals from SLFM-BART completely miss the heldout data.

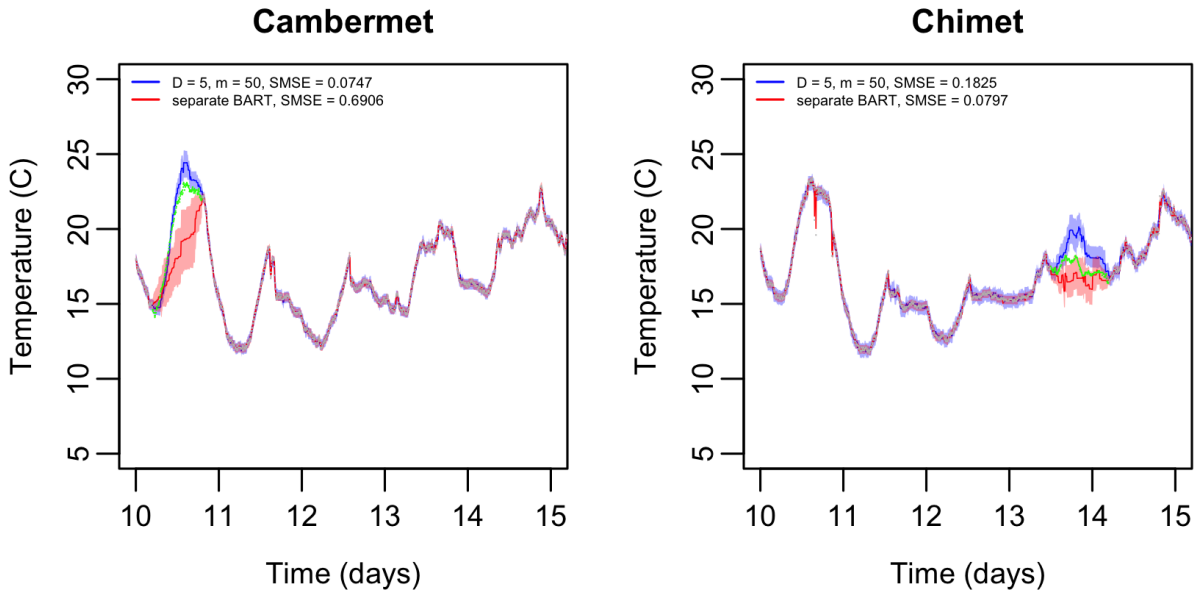

Figure 4: Posterior predictive mean and 95% intervals for Cambermet and Chimet for SLFM-BART with $(m, D) = (50, 5)$ (blue) and independent BART (red). Green points were held out for testing purposes.

Table **??** compares the overall SMSE of SLFM-BART for several combinations of m and D along with the following competitors based on GPs.

- IGP: Fit independent GP with squared exponential kernels, as reported by Nguyen and Bonilla (2014)
- CGP: convolved GP model, as reported by Nguyen and Bonilla (2014)
- COGP: collaborative multi-output GP, as reported by Nguyen and Bonilla (2014). This model is closest in spirit to ours as it is also expresses each task as a linear combination of basis functions
- GPAR: Autoregressive Gaussian processes (Requeima et al., 2019). This is a compositional model as it expresses each $y_k = f_k(\mathbf{x}, y_1, \ldots, y_{k-1})$ with GP prior placed on each f_k

It is interesting to note that for most combinations of D and m, SLFM-BART outperforms COGP and CGP.

Table 1: Standardized mean square error for predicting the held out continuous segments for CAD, JPY, and AUD. The notation SLFM(m, D) indicates that the D basis functions were comprised of m trees each.

Method	FOREX	AIRTEMP
(5, 1)	0.2505	0.0332
(5, 25)	0.2705	0.0564
(5, 50)	0.2744	**0.0192**
(25, 1)	0.1248	0.0247
(25, 25)	0.0813	0.1080
(25, 50)	0.0599	0.0616
(50, 1)	0.1890	0.0303
(50, 25)	**0.0526**	0.0278
(50, 50)	0.0759	0.0202
Ind. BARTs	0.4379	0.3055
Ind. GPs	0.5996	0.8944
COGP	0.2125	0.1077
CGP	0.2427	0.1125
GPAR	0.0302	–

4 Next Steps

There are several potential directions, which we briefly outline now.

We have so far only considered the homoskedastic setting where the residual variance for each task was constant. It is straightforward to consider the heteroskedastic case using Pratola (2016)'s "product-of-trees" formulation. More substantively, recall that our choice of $\phi_{k,d} \sim N(0, \sigma^2_{\phi,k})$ ensures that Φ is dense almost surely. This in turn implies that every task depends on every basis function, which may be an overly restrictive imposition. Indeed, in the FOREX example, while our SLFM-BART fit the JPY and AUD data better than individual BART regressions, it did substantially worse on the CAD data. This would suggest that the CAD exchange rate is somewhat less dependent on the other series.

A more flexible model would introduce sparsity to Φ, enabling us to adapt to situations where some tasks are independent of certain basis functions and of other tasks. One approach would

be to use a spike-and-slab prior:

$$\phi_{k,d}|\gamma_{k,d} \sim \gamma_{k,d} N(0, \sigma^2_{\phi,k}) + (1 - \gamma_{k,d})\delta_0$$
$$\gamma_{1,d}, \ldots, \gamma_{q,d}|\theta_d \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\theta_d) \text{ for } d = 1, \ldots, D$$

We may place independent Beta(a, b) priors on the θ_d's or we may enforce an ordering $\theta_1 > \cdot\theta_D$ using the stick-breaking construction of the Indian Buffet Process.

In the AIRTEMP example, we saw that both SLFM-BART and independent BART were able to detect the daily rise and fall of temperature without pre-specifying any periodicity. In the context of time series data it may be worthwhile to consider adding shape constraints to the general BART framework. Currently, to split at a given node we draw a covariate index v and cut-point c and split observations according to $x_v \leq c$ or $x_v > c$. To allow for periodicity, we may similarly select v but before choosing a cut-point, we can draw a period P_v and split observations according to $x_v - P_v \times \lfloor x_v/P_v \rfloor \leq c_P$ or not, where c_P is a cutpoint drawn from the interval $[0, P_v]$.

Finally, we have assumed throughout that each task was continuous. A fruitful direction to pursue would be the joint modeling of mixed-type data, in which some of the tasks were categorical or count data. The approach of Pourmohamad and Lee (2016) may be useful.

References

Álvarez, M. A., Luengo, D., Titsias, M. K., and Lawrence, N. D. (2010). Efficient multioutput Gaussian processes through variational inducing kernels. In *Proceedings of the 13th International Conference of Artifical Intelligence and Statistics (AISTATS)*.

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.

Nguyen, T. V. and Bonilla, E. V. (2014). Collaborative multi-output Gaussian processes. In *Proceedings of the Thirtieth Conference of Uncertainty in Artifical Intelligence*, UAI '14, pages 643 – 652, Arlington, Virginia, United States of America. AUAI Press.

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. (2016). Efficient metropolis-hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Analysis*, 11(3):885 – 911.

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

Requeima, J., Tebbutt, W., Bruinsma, W., and Turner, R. E. (2019). The Gaussian process autoregressive model (gpar). In *AISTATS 2019 – Proceedings of the 22nd International Conference on Artifical Intelligence and Statistics*, volume 89.

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.

5 A Gibbs sampler

We describe each of these conditional updates in the next three subsections.

5.1 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.

5.2 Updating Φ

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

5.2.1 Gaussian Prior

Observe that

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

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

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

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

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

5.2.2 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

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

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

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

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 $\theta_1, \ldots, \theta_d$:

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

[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

5.3 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

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