2016 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 13–16, 2016, SALERNO, ITALY

SCALABLE TRANSFORMED ADDITIVE SIGNAL DECOMPOSITION BY NON-CONJUGATE GAUSSIAN PROCESS INFERENCE

Vincent Adam[1]*, James Hensman*[2]*, Maneesh Sahani*[1]

[1]University College London
Gatsby Computational Neuroscience Unit
25 Howland Street, London W1T 4JG.

[2] Lancaster University
CHICAS, Faculty of Health and Medicine
Lancaster, LA1 4YB.

ABSTRACT

Many functions and signals of interest are formed by the addition of multiple underlying components, often nonlinearly transformed and modified by noise. Examples may be found in the literature on Generalized Additive Models [1] and Underdetermined Source Separation [2] or other mode decomposition techniques. Recovery of the underlying component processes often depends on finding and exploiting statistical regularities within them. Gaussian Processes (GPs) [3] have become the dominant way to model statistical expectations over functions. Recent advances make inference of the GP posterior efficient for large scale datasets and arbitrary likelihoods [4, 5]. Here we extend these methods to the additive GP case [6, 7], thus achieving scalable marginal posterior inference over each latent function in settings such as those above.

1. INTRODUCTION

We are interested here in settings where a measured signal $y(x)$ defined over a space $\mathcal{X}$ can be modelled as a transformed sum, with arbitrary observation noise, of D independent source functions $f^{(d)}(x^{(d)})$, defined on domains $\mathcal{X}^{(d)}$. In some cases both the domain of each function, and the point of evaluation may be identical. That is $\mathcal{X} = \mathcal{X}^{(1)} = \cdots = \mathcal{X}^{(D)}$ and $x = x^{(1)} = \cdots = x^{(D)}$. (For example, y might be the sum of different functions of time.) In this case, the setting is one of source separation. Alternatively, $\mathcal{X}$ may be the Cartesian product space of possibly unrelated inputs $\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(D)}$, with each $f^{(d)}$ defined on the single input domain $\mathcal{X}^{(d)}$. Then $x = (x^{(1)}, \ldots, x^{(D)})$. This is the setting most often encountered in the literature on Generalized Additive Models (GAMs).

We imagine that the signal has been measured at N points x_i, yielding values y_i. We collect these measurements into a vector $\mathbf{y} = (y_1, \ldots, y_N)$. Similarly the vector $\mathbf{f}^{(d)}$ is formed by the evaluations of $f^{(d)}$ at each $x_i^{(d)}$. We write $f(x) = \sum_d f^{(d)}(x^{(d)})$ and $\mathbf{f} = \sum_d \mathbf{f}^{(d)}$. We focus on conditionally independent observation models where the likelihood factorizes $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^N p_{\text{lik}}(y_i|f_i)$, but each conditional $p_{\text{lik}}(y|f)$ may be arbitrary. Our goal is to recover estimates of the functions $f^{(d)}$.

Bayesian approaches to this problem express beliefs about the function values $p(\mathbf{f}^{(d)})$, which capture assumptions about how the underlying processes might differ, and then compute the posterior $p\left(\mathbf{f}^{(d)}|\mathbf{y}\right)$. Hyperparameters (that is, parameters specifying the form of beliefs) are selected by maximizing the marginal likelihood $p(\mathbf{y})$.

Here, we follow a Bayesian approach and specify a generative model using Gaussian Process (GP) priors for each individual function. We propose an approximate inference algorithm based on sparse variational inducing points to solve the inference problem and hyperparameter optimization.

2. BACKGROUND

2.1. Gaussian Process Regression

Gaussian Processes are infinite collections of random variables, any finite subset of which follows a multivariate Gaussian distribution. They are defined by a mean function m and covariance function k. A draw from a GP defined on a index set $\mathcal{X}$ is a function on the domain $\mathcal{X}$. Given a list of points $X \in \mathcal{X}^N$ and a function draw $f \sim GP(m, k)$, the vector of function evalutions $\mathbf{f}(X)$ is an associated multivariate normal random variable such that $\mathbf{f}(X) \sim \mathcal{N}(\mathbf{m}(X), \mathbf{K}(X, X))$, where $\mathbf{m}$ is a vector of mean function evaluations and $\mathbf{K}$ is a matrix of covariance function values. We will focus here on the case $m = 0$ for simplicity, and without loss of generality.

In GP regression, we have a data set $\mathcal{D} = \{x_i, y_i\}_{i=1\ldots N}$, an observation model $y_i|f(x_i)$, and a Gaussian process prior on f. We seek to compute the posterior $p(f|\mathcal{D}) \propto p(f)p(\mathbf{y}|f)$ and the marginal likelihood $p(\mathbf{y}) = \int d\mathbf{f}\, p(\mathbf{y}, \mathbf{f})$.

Two main difficulties arise when trying to compute the posterior. First, even in the simplest case of the conjugate likelihood $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, for which the marginal likelihood can be obtained analytically: $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K_{nn}} + \sigma^2\mathbf{I})$ with $\mathbf{K_{nn}} = \mathbf{K}(X, X)$, computations

Thanks to the Gatsby Charitable Foudation for funding.

require the expensive inversion of an $N \times N$ matrix. Second, when the likelihood is not conjugate estimates are not available in closed form and must be approximated, for example by expectation propagation [8] or variational inference [9]. Such approximations often scale poorly.

Irrespective of the likelihood, sparse approximations to the GP prior have been proposed; see [10, 11]. These approximations extend the GP model with m additional random variables $\mathbf{u}$ drawn from the same GP prior as $\mathbf{f}$ at pseudo-input locations Z, sometimes called *inducing* points. The form of the approximation is to modify the joint prior $p(\mathbf{f}, \mathbf{u})$ by forcing conditional independencies — for example setting $p(\mathbf{f}|\mathbf{u}) = \prod_i p(f_i|\mathbf{u})$. Thus, the $\mathbf{u}$ (and set Z) become parameters, yielding a new parametric prior on $\mathbf{f}$. Writing $\mathbf{K_{mm}} = \mathbf{K}(Z, Z)$ and $\mathbf{K_{nm}} = \mathbf{K}(X, Z)$, the independence assumption induces a low rank form for the covariance matrix $\mathbf{K_{nn}} = \mathbf{K_{nm}}\mathbf{K_{mm}^{-1}}\mathbf{K_{mn}}$, with the rank governed by the size of the inducing set. In this form, the approximation modifies the prior on $\mathbf{f}$, in effect encoding different expectations about the form of the function. An alternative is to treat the inducing points instead as variational parameters in an approximate variational inference framework [4]. In this view it is the joint *posterior* that is approximated rather than prior, assuming the form $p(\mathbf{f}, \mathbf{u}|\mathbf{y}) = p(\mathbf{f}|\mathbf{u})\, q(\mathbf{u})$. This approach proves relatively easy to extend to the non-conjugate case [5]. The approximation does not change the assumed prior model and has a number of appealing theoretical properties [12].

2.2. Additive Gaussian Processes

Nonparametric regression on a high-dimensional domain suffers from the curse of dimensionality: neighbourhoods become more and more local as the dimensionality of the space increases. Thus, estimating high dimensional functions requires either a very large number of data points or prior assumptions of extreme smoothness.

This problem may be alleviated if the target function can be assumed to be formed from the sum of lower-dimensional mappings. [6, 7]. In the GP framework, such additivity can be imposed implicitly by composing an additive kernel. Thus, if $k(x, x') = \sum_d k^{(d)}\left(x^{(d)}, x'^{(d)}\right)$, then the associated GP functions will have additive structure $f(x) = \sum_d f^{(d)}\left(x^{(d)}\right)$, where each $f^{(d)}$ is drawn from a GP with covariance kernel $k^{(d)}$. Thus, one approach to tractable additive GP regression might be to simply combine an additive kernel with a sparse approximation.

In the variational inducing point framework, the posterior is a low-dimensional process given by $p(f|\mathbf{y}) \approx q(f) = \int d\mathbf{u}\, p(f|\mathbf{u})\, q(\mathbf{u})$. In the case where $q(\mathbf{u}) = \mathcal{N}(\mu, \Sigma)$, the posterior is a GP with non-additive covariance structure

$$\begin{aligned} q(f) &= \mathcal{N}(\mu_f, \Sigma_f) \\ \mu_f(x) &= K_{xZ}K_{ZZ}^{-1}\mu \\ \Sigma_f(x, x') &= K_{xx'} + K_{xZ}K_{ZZ}^{-1}\left[\Sigma K_{ZZ}^{-1} - I\right]K_{Zx'} \end{aligned}$$

To recover the marginal posterior distributions over the individual functions f^d requires the extra step of computing the joint multivariate posterior GP $q\left(\mathbf{f}^{(1)}, ..., \mathbf{f}^{(d)}\right) = \frac{q(\mathbf{f})}{p(\mathbf{f})}p\left(\mathbf{f}^{(1)}, ..., \mathbf{f}^{(d)}\right)$ and marginalizing. Writing $q\left(\mathbf{f}^{(1)}, ..., \mathbf{f}^{(d)}\right) = \mathcal{N}(\mathbf{v}, \mathbf{V})$, we have

$$\begin{aligned} V^{(d,d')} &= K^{(d)} - K^{(d)}\tilde{K}^{-1}K^{(d')} \\ \nu^{(d)} &= \left(K^{(d)} - K^{(d)}\tilde{K}^{-1}K_{sum}\right)\Sigma_{\mathbf{f}}^{-1}\mu_{\mathbf{f}} \end{aligned}$$

where $K_{sum} = \sum_{d=1}^{D} K^{(d)}$ and $\tilde{K} = \left[\Sigma_{\mathbf{f}}^{-1} - K_{sum}^{-1}\right]^{-1} + K_{sum}$

This marginalisation step is needed because the inducing points do not immediately provide the information to reconstruct the individual components (indeed, they were optimized to reconstruct the summed GP). The posterior cannot be computed separately for individual input points: this would require splitting the mean and variance according to the relative prior variances of each underlying function at that point, and so would be blind to the rest of both the prior and posterior GP structure. Instead the joint must be constructed over a large set of points (ideally the full initial dataset) and those data points should not share any coordinate. Thus, this final step is computationally expensive.

An alternative might be to seek the best additive approximation to the posterior GP according to a chosen metric, but this amounts to solving an additional regression problem.

Here we propose instead to extend the variational inducing point framework for additive GP models by directly constructing a parametric posterior approximation to each component GP. We do so by assuming a factorized approximation to the joint posterior $p\left(\mathbf{f}^{(1)}, ..., \mathbf{f}^{(d)}, \mathbf{u}^{(1)}, ..., \mathbf{u}^{(d)}|Y\right) \approx \prod_d q\left(\mathbf{u}^{(d)}\right)p\left(\mathbf{f}^{(d)}|\mathbf{u}^{(d)}\right)$. A similar approach has been proposed independently by Saul et al [13] to allow the non-linear combination of an arbitrary collection of GPs. However, they did not recognize the computational advantage induced by the additive structure.

Our approach provides the following advantages: Each GP has separate inducing points, allowing the number to be adjusted to the complexity of the specific domain and prior. Furthermore, the inducing variables are readily interpretable as conditional variables for the prediction of each function. This comes however at the cost of losing the covariance structure across functions, but readily applicable methods exist to recover covariance estimates [14].

3. SPARSE ADDITIVE GAUSSIAN PROCESS REGRESSION

To simplify notation, we write $\mathbf{F} = [\mathbf{f}^{(1)}, ..., \mathbf{f}^{(d)}]$ and $\mathbf{U} = [\mathbf{u}^{(1)}, ..., \mathbf{u}^{(d)}]$. Consider an augmented model $p(\mathbf{F}, \mathbf{U}) = \prod_d p(\mathbf{f}^{(d)}, \mathbf{u}^{(d)})$ where the $u^{(d)}$ are associated with pseudo-inputs $Z^{(d)}$, and assume a variational approximation to the posterior of the form $p(\mathbf{F}, \mathbf{U}|\mathbf{y}) \approx \prod_d p(\mathbf{f}^{(d)}|\mathbf{u}^{(d)})q(\mathbf{u}^{(d)})$

3.1. Variational lower bound

We apply Jensen's inequality twice. First, we use the standard free-energy lower bound on the evidence under the factorized posterior approximation $q(\mathbf{U}) = \prod_d q\left(\mathbf{u}^{(d)}\right)$:

$$\log p(\mathbf{y}) \geq \mathbb{E}_{\prod_d q\left(\mathbf{u}^{(d)}\right)}\left[\log p(\mathbf{y}|\mathbf{U})\right] - \mathcal{KL}\left(q\right)$$

where $\mathcal{KL}\left(q\right) = \sum_d KL\left[q(\mathbf{u}^{(d)})||p(\mathbf{u}^{(d)})\right]$. Second, we follow [4] to obtain a lower bound on the conditional log likelihood $\log p(\mathbf{y}|\mathbf{U})$

$$\begin{aligned}
\log p(\mathbf{y}|\mathbf{U}) &= \log \mathbb{E}_{\prod_d p\left(\mathbf{f}^{(d)}|\mathbf{u}^{(d)}\right)}\left[\frac{p\left(\mathbf{y},\mathbf{F}|\mathbf{U}\right)}{\prod_d p\left(\mathbf{f}^{(d)}|\mathbf{u}^{(d)}\right)}\right] \\
&\geq \mathbb{E}_{\prod_d p\left(\mathbf{f}^{(d)}|\mathbf{u}^{(d)}\right)}\left[\log \frac{p\left(\mathbf{y},\mathbf{F}|\mathbf{U}\right)}{\prod_d p\left(\mathbf{f}^{(d)}|\mathbf{u}^{(d)}\right)}\right] \\
&= \sum_i \mathbb{E}_{\prod_d p\left(f_i^{(d)}|\mathbf{u}^{(d)}\right)}\left[\log p_{\text{lik}}\Big(y_i\Big|\sum_d f_i^{(d)}\Big)\right] \\
&= \sum_i \mathbb{E}_{p(\rho_i|\mathbf{U})}\left[\log\, p_{\text{lik}}\left(y_i|\rho_i\right)\right]
\end{aligned}$$

with $\rho_i|\mathbf{U} = \sum_d f_i^{(d)}|\mathbf{u}^{(d)}$, the conditional additive predictor. Combining the two inequalities, we obtain:

$$\begin{aligned}
\log p\left(\mathbf{y}\right) &\geq \mathbb{E}_{\prod_d q\left(\mathbf{u}^{(d)}\right)}\left[\log\, p\left(\mathbf{y}|\mathbf{U}\right)\right] - \mathcal{KL}\left(q\right) \\
&\geq \sum_i \mathbb{E}_{\prod_d q\left(\mathbf{u}^{(d)}\right)}\left[\mathbb{E}_{p(\rho_i|\mathbf{U})}\left[\log\, p_{\text{lik}}\left(y_i|\rho_i\right)\right]\right] \\
&\quad - \mathcal{KL}\left(q\right) \\
&\geq \sum_i \mathbb{E}_{q(\rho_i)}\left[\log\, p_{\text{lik}}\left(y_i|\rho_i\right)\right] - \mathcal{KL}\left(q\right)
\end{aligned}$$

where $q\left(\rho_i\right) = \mathbb{E}_{\prod_d q\left(\mathbf{u}^{(d)}\right)}\left[p\left(\rho_i|\mathbf{U}\right)\right]$

Finally, we assume a Gaussian parametric form for $q(\mathbf{u}^{(d)}) = \mathcal{N}(m^{(d)}, S^{(d)})$. This assumption implies that $q(\rho_i)$ will also be Gaussian, and thus expecations can be evaluated rapidly and accurately by Gaussian quadrature methods. Writing $A^{(d)} = K_{nm}^{(d)} K_{mm}^{(d)-1}$, we have

$$\mathbf{f}^{(d)}|\mathbf{u}^{(d)} \sim \mathcal{N}\left(A^{(d)}\mathbf{u}^{(d)}, K_{nn}^{(d)} - A^{(d)}K_{mm}^{(d)}A^{(d)T}\right)$$

$$\boldsymbol{\rho}|\mathbf{U} \sim \mathcal{N}\left(\sum_d A^{(d)}\mathbf{u}^{(d)}, \sum_d K_{nn}^{(d)} - A^{(d)}K_{mm}^{(d)}A^{(d)T}\right)$$

and so

$$q(\boldsymbol{\rho}) = \mathbb{E}_{\prod_d q\left(\mathbf{u}^{(d)}\right)}\left[p\left(\boldsymbol{\rho}|\mathbf{U}\right)\right] = \mathcal{N}\left(\sum \mu_{add}^{(d)}, \sum_d \Sigma_{add}^{(d)}\right),$$

with,

$$\begin{aligned}
\mu_{add}^{(d)} &= A^{(d)}m^{(d)} \\
\Sigma_{add}^{(d)} &= K_{nn}^{(d)} + A^{(d)}\left(S^{(d)} - K_{mm}\right)A^{(d)T}
\end{aligned}$$

3.2. Optimization

The lower bound may be optimized numerically with respect to $m^{(d)}$, $S^{(d)}$, the inducing point locations and any kernel hyperparameters. The structure of the bound lends itself to stochastic gradient descent using minibatches, which allows massively scaled inference.

The KL divergences in the bound take $O\left(DM^3\right)$ operations to compute. Computing predictions $q(\mathbf{f})$ to evaluate the likelihood term of the bound takes $O\left(DM^2N\right)$.

The expectation computation $\mathbb{E}_{q(\rho_i)}\left[\log\, p_{\text{lik}}\left(y_i|\rho_i\right)\right]$ can be performed either by Monte-Carlo sampling or by quadrature methods. One key feature here is that the expectation is under a one-dimensional random variable (the additive predictor) irrespective of the number of latent functions, thus avoiding the poor scaling of quadrature methods and Monte Carlo sampling for Gaussian expectations.

We implemented and optimized the objective function using the GPflow[1] framework based on Tensorflow [15].

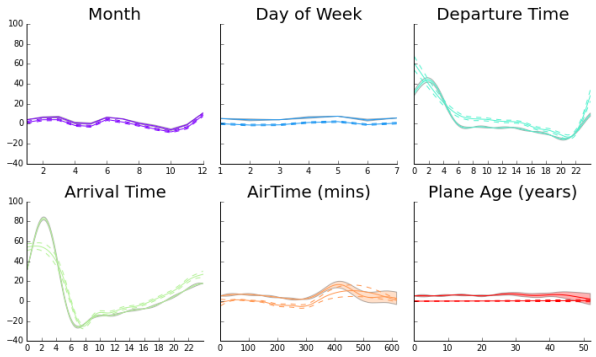
3.3. A GAM perspective

In the GAM literature, parametric assumptions on the form of the individual components are typically based on regression smoothing splines, leading to functions of the form $f(x) = \sum_k w_k \phi_k(x)$. Parameters are estimated by optimizing a penalized likelihood of the form $l(\mathbf{f}) + \sum_d \lambda^{(d)}\|\mathbf{f}^{(d)}\|^2$. The regularization parameters $\lambda^{(d)}$ provide a trade-off between goodness of fit and smoothness. This approach yields a point estimate for the best fitting functions. Asymptotic arguments are then used to derive approximate 'Bayesian' confidence intervals or to generate objectives to select the regularization parameters [16]. The form of the inferred functions is thus set in advance. Recent work extended this approach by predetermining the form of the posterior on the functions [17], adding priors on the spline parameters and computing the marginal posterior on those parameters. However, the priors in this approach are difficult to interpret, and the approximate inference schemes required do not scale well. For examples, see [18, 19].

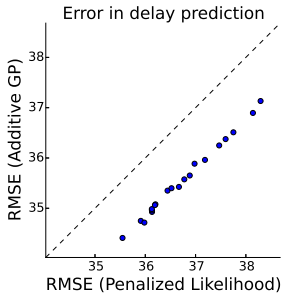
Our approach is similar in spirit, in that we assume independent GP priors for each function $f^{(d)} \sim \mathcal{GP}$ and choose a parametric form for the marginal posterior on each GP of the form $q(\mathbf{f}^{(d)}) = \int d\mathbf{u}^{(d)}\, p(\mathbf{f}^{(d)}|\mathbf{u}^{(d)})q(\mathbf{u}^{(d)})$. However, by contrast to the spline-based methods of [17], our method exploits the rich expressive power of GPs, and scales very well.

The Gaussian Process approach may be seen as a generalization of a fixed basis expansion. For example, assuming $f(x) = \sum_k^K w_k \phi_k\left(x\right)$ amounts to setting a prior GP of the form $k(x, x') \propto \Phi(x)\Phi(x')^T$ where $\Phi = [\phi_1, ..., \phi_K]^T$. Such kernels define finite dimensional function spaces, the dimensionality of which determines an upper bound to the number of inducing points required for the associated approximate

[1] https://github.com/GPflow/GPflow



(a) Inferred functions



(b) Prediction error

Fig. 1. (a) Posterior over individual latent function in the additive regression analysis of the airline data (shaded). Maximum Penalized likelihood estimate with 95% confidence interval (dashed). (b) RMSE of predicted delays for both methods.

posterior. As a concrete example, if a component with a one-dimensional domain is linear, one can use a linear kernel and only two inducing points.

4. EXPERIMENTS

We investigated the validity of our approach in two experiments, corresponding to two different application domains. In the first experiment, a GAM model was fit to a large dataset, to yield interpretable posteriors over the individual functions composing the additive predictor. In a second experiment, we used the approach to efficiently separate sources mixed into a single time-series.

4.1. Generalized additive model

We performed an additive regression analysis on a dataset derived from every commercial flight in the USA for the year 2008. The dataset comprised more than 2 million entries, from which we selected a training set and a test set both containing $N = 50000$ data points. Following [20], we attempted to predict the delay of flights using the following 6 covariates: age of the aircraft, planned airtime, planned departure time, planned arrival time, day of the week, month.

We used an additive Gaussian Process prior (one function per covariate) with RBF kernel, assuming a conjugate likelihood with independent Gaussian noise in each measured delay. The inducing-point variational approximation was based on $M = 10$ inducing points per covariate, and we learned both the variational parameters (the inducing values) and the noise parameter of the likelihood. The pseudo-input locations were fixed to a homogenous grid for each dimension, with kernel hyperparameters controlling the prior variance and the smoothness of each GP.

The posterior marginals, shown in Figure 1(a), provide a visualisation of the impact of the individual covariates on delay. Note that additivity makes the offset of each function arbitrary: the critical features are the pattern of variation and the relative magnitude of variation across the different functions. We found the dominant predictors to be, in decreasing order of importance: arrival time, departure time, and airtime. The marginal posteriors suggest, for instance, that greater delays can be expected for night flights.

We compared our method with *bam*, the scalable version of the gam toolbox contained in the R package mgvc [21], using the same model structure and setting all parameters to their default values. Inferred functions shown in Figure 1(a) are broadly similar, as is the reconstruction error. However, root mean squared error in predicted delays for 20 repetitions of the experiment are shown in Figure 1(b), and demonstrate a distinct advantage for the proposed GP approach.

4.2. Additive source separation

The second experiment investigated the capacity of the additive GP approach to: handle non-homogenous sampling grids, incorporate rich priors over sources, allocate a variable number of inducing points to different sources, and handle observation non-linearities.

We created an artificial mixture of partially overlapping locally periodic sources with a slowly varying background. This mixture was mapped through a monotone non-linearity and corrupted by Gaussian white noise. The two sources had the Gabor form $s_{x_c,\sigma,f}(x) = e^{-\frac{|x-x_c|}{\sigma}} \cos(2\pi f x)$, while the slowly varying background source was taken to be the function $\cos(5x)$. The combination yielded the mixture signal $m(x)$, which was passed through a compressing non-linearity $g(x) = sign(x)|x|^{1/2}$ and combined with white noise. The resulting signal can be seen in Figure 2(a).

For inference, the locally periodic sources were assumed to be sampled from stationary GPs with kernel $k_{\tau,\omega}(x, x') = e^{-\frac{|x-x'|}{\tau}} \cos(2\pi\omega|x - x'|)$. We sampled N=5000 input points from a uniform distribution between 0 and 1. For each we computed a corresponding output $y_i = g(m(x_i)) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}\left(0, \sigma_n^2\right)$.

The periodicity parameter of the kernel was fixed to the

true generating value. Inducing point pseudo-inputs were initialized by randomly selecting corresponding coordinates from the actual data points: 30 pseudo-inputs were used for each of the two local sources and 10 for the smooth one. We sought to learn the kernel hyperparameters, the pseudo-input locations, the likelihood noise parameter and the variational parameters.

Figure 2(b) shows how the inferred marginals match the sources well, with inducing points spread across the signal support. Inducing points and variational parameters shape the variational posterior both on the 'active' portion of the sources and where it is close to zero. On this latter part, they ensure the posterior variance does not fall back to the prior variance, which must be set to a high value for all sources to capture their range.

The identified mixture contains local tonal components. A common method suited to the extraction of similar components is called Empirical Mode Decomposition (EMD). Although it is not designed to deal with observation non linearity, it is regarded as being robust to noise. We applied the ensemble version of EMD (EEMD) as implemented in the package libeemd [22] to decompose the signal. For this example, using 6 Intrinsic Mode Functions (IMF) led to the best results, but these were still inferior to the GP approach.

5. CONCLUSION

We have demonstrated a scalable algorithm to infer an underlying set of signals or functions, given noisy, transformed observations of their sum. The approach depends on capturing the statistical structure of the underlying signals using Gaussian processes. Scalability is addressed by using an inducing-point based variational inference approach, tailored specifically to provide robust reconstructions of the individual processes. The same approach has applications to a range of problems, from signal decomposition to generalised additive regression modelling.

Acknowledgments

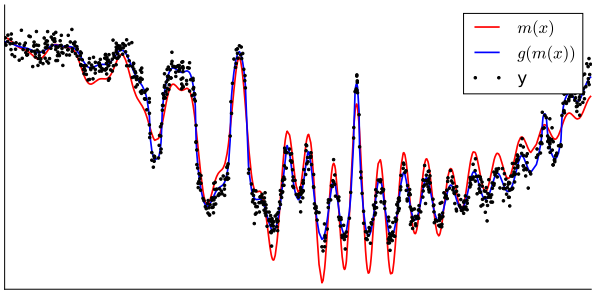
We would like to thank Wittawat Jitkrittum and Heiko Strathmann for helpful early contributions to this project.

6. REFERENCES

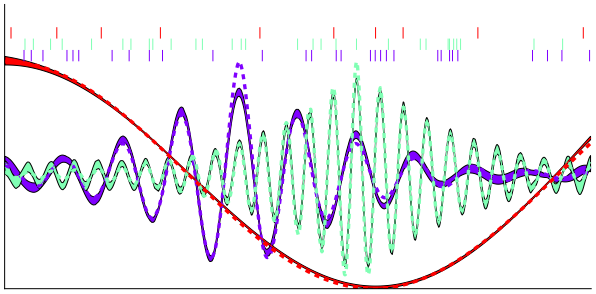
[1] Trevor J Hastie and Robert J Tibshirani, *Generalized additive models*, vol. 43, CRC Press, 1990.

[2] Antoine Liutkus, Roland Badeau, and Gäel Richard, "Gaussian processes for underdetermined source separation," *Signal Processing, IEEE Transactions on*, vol. 59, no. 7, pp. 3155–3167, 2011.

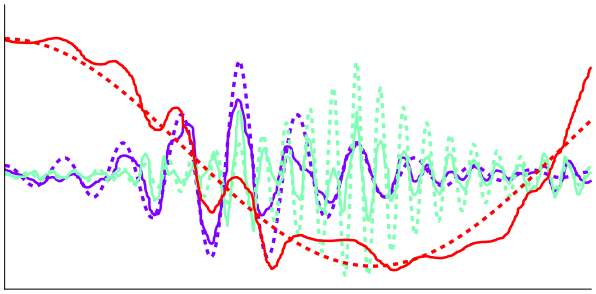
[3] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive*



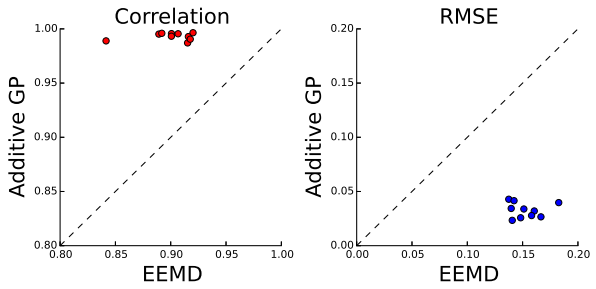
(a) Observed signal



(b) Inferred sources (Additive GP)



(c) Inferred sources (EEMD)



(d) Reconstruction error

Fig. 2. (a) The mixture signal. (b) Output of Source separation with our method. Dashed line are the true sources, tubes represent the posterior GP for each individual source. Vertical bars show the positions of the pseudo inputs associated to the inducing points (c) IMF recovered using EEMD. (d) Averaged RMSE and Correlation coefficient between true sources and recovered sources for multiple resampling of the data.

Computation and Machine Learning), The MIT Press, 2005.

[4] Michalis K Titsias, "Variational learning of inducing variables in sparse gaussian processes," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 567–574.

[5] James Hensman, Alexander Matthews, and Zoubin Ghahramani, "{Scalable Variational Gaussian Process Classification}," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015, pp. 351–360.

[6] David K Duvenaud, Hannes Nickisch, and Carl E Rasmussen, "Additive gaussian processes," in *Advances in neural information processing systems*, 2011, pp. 226–234.

[7] Nicolas Durrande, David Ginsbourger, Olivier Roustant, and Laurent Carraro, "Additive covariance kernels for high-dimensional gaussian process modeling," *arXiv preprint arXiv:1111.6233*, 2011.

[8] Matthias Seeger, "Expectation propagation for exponential families," Tech. Rep., 2005.

[9] Manfred Opper and Cédric Archambeau, "The variational gaussian approximation revisited," *Neural Comput.*, pp. 786–792, 2009.

[10] Hannes Nickisch and Carl Edward Rasmussen, "Approximations for binary gaussian process classification," *Journal of Machine Learning Research*, vol. 9, no. Oct, pp. 2035–2078, 2008.

[11] Joaquin Quiñonero-Candela and Carl Edward Rasmussen, "A unifying view of sparse approximate gaussian process regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[12] Alexander G de G Matthews, James Hensman, Richard E Turner, and Zoubin Ghahramani, "On sparse variational methods and the Kullback-Leibler divergence between stochastic processes," *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*, 2016.

[13] Alan Saul, James Hensman, Aki Vehtai, and Neil D Lawrence, "Chained Gaussian processes," *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*, 2016.

[14] Ryan J Giordano, Tamara Broderick, and Michael I Jordan, "Linear response methods for accurate covariance estimates from mean field variational bayes," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 1441–1449. Curran Associates, Inc., 2015.

[15] Martín Abadi and al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[16] Simon Wood, *Generalized additive models: an introduction with R*, CRC press, 2006.

[17] Stefan Lang and Andreas Brezger, "Bayesian p-splines," *Journal of Computational & Graphical Statistics*, vol. 13, no. 1, pp. 183–213, 2004.

[18] Jan Luts, Shen Wang, John Ormerod, and Matt Wand, "Semiparametric regression analysis via infer. net," 2014.

[19] Tung H Pham and Matt P Wand, "Generalized additive mixed model analysis via gammslice," 2014.

[20] James Hensman, Nicolo Fusi, and Neil D Lawrence, "Gaussian processes for big data," in *Conference on Uncertainty in Artificial Intellegence*. auai.org, 2013, pp. 282–290.

[21] Simon N Wood, Yannig Goude, and Simon Shaw, "Generalized additive models for large data sets," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 64, no. 1, pp. 139–155, 2015.

[22] PJJ Luukko, Jouni Helske, and Esa Räsänen, "Introducing libeemd: A program package for performing the ensemble empirical mode decomposition," *Computational Statistics*, vol. 31, no. 2, pp. 545–557, 2016.