Enriched mixtures of generalised Gaussian process experts

Charles W.L. Gadd †
Dept. of Computer Science
Aalto University
Espoo, Finland
cwlgadd@gmail.com

Sara Wade *
School of Mathematics
University of Edinburgh
Edinburgh, United Kingdom
sara.wade@ed.ac.uk

Alexis Boukouvalas
PROWLER.io
Cambridge, United Kingdom
alexis@prowler.io

For the Alzheimer's Disease Neuroimaging Initiative

Abstract

Mixtures of experts probabilistically divide the input space into regions, where the assumptions of each expert, or conditional model, need only hold locally. Combined with Gaussian process (GP) experts, this results in a powerful and highly flexible model. We focus on alternative mixtures of GP experts, which model the joint distribution of the inputs and targets explicitly. We highlight issues of this approach in multi-dimensional input spaces, namely, poor scalability and the need for an unnecessarily large number of experts, degrading the predictive performance and increasing uncertainty. We construct a novel model to address these issues through a nested partitioning scheme that automatically infers the number of components at both levels. Multiple response types are accommodated through a generalised GP framework, while multiple input types are included through a factorised exponential family structure. We show the effectiveness of our approach in estimating a parsimonious probabilistic description of both synthetic data of increasing dimension and an Alzheimer's challenge dataset.

1 INTRODUCTION

The Gaussian process is a powerful and popular prior for nonparametric regression, due to its flexibility, analytic tractability, and interpretable hyperparameters. The GP assumes that the unknown function evaluated at any finite set of inputs has a Gaussian distribution with consistent parameters. It is fully specified by a mean function and symmetric positive definite covariance (or kernel) function, which together encapsulate any prior knowledge and assumptions of the regression function, such as smoothness and periodicity (see Rasmussen and Williams, 2005, for a thorough overview).

In GP regression, the outputs are modelled as noisy observations of an unknown function, which is assigned a GP prior. While GP regression has been successfully applied to various problems, it only allows for flexibility in the regression function, assuming i.i.d. Gaussian errors. Many datasets present departures from this model, such as multi-modality or changing error variance across the input space. Moreover, for computational purposes, a stationary GP is typically employed, which limits the model's ability to recover changing behaviour of the function across the input space, e.g. different smoothness levels.

Density regression refers to the general problem of estimating the conditional density of the targets across the input space, or equivalently, flexible estimation of both the regression function and input-dependent error distribution. Mixtures of experts (Jacobs et al., 1991) address the density regression problem by probabilistically partitioning the input space. Each expert is a conditional model, and a gating network maps experts to local regions of the input space. Scalability is enhanced since each expert considers only its local region, and simplifying assumptions of the regression function need only hold locally in each region.

Experts may range from simple linear models to flexible non-linear approaches. Specifically, GP experts allow the model to infer local non-linearities characterized by different behaviours, such as smoothness and

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. † Corresponding author. ∗ Equal contribution.

variability. In this work, we build upon alternative mixtures of GP experts (Meeds and Osindero, 2006), which employ mixtures to explicitly model the joint distribution of the inputs and targets and GP experts to capture their local relation. We highlight issues of this approach for multi-dimensional inputs, namely, poor scalability and the need for an unnecessary number of experts, and we construct a novel model based on the enriched Dirichlet process (EDP, Wade et al., 2011) to address these issues, that additionally extends the model for multiple response and input types.

The paper is organised as follows. Related work is reviewed in Section 2. In Section 3, we construct a novel mixture of generalised GP experts, that utilises a nested partitioning scheme to improve prediction and uncertainty quantification. Section 4 describes posterior inference. Section 5 illustrates the benefits in a non-linear toy example and a case study to predict cognitive decline in Alzheimer's.

2 RELATED WORK

Mixtures of experts were first combined with GPs in Tresp (2001), resulting in a flexible nonparametric approach for both the experts and gating networks. Infinite mixtures of GP experts were subsequently introduced (Rasmussen and Ghahramani, 2002), allowing the number of experts to be determined by and grow with the data; this is performed by employing the Dirichlet process and kernel classifiers to model the gating network. In the treed-GP (TGP, Gramacy and Lee, 2008), an example of a finite mixture of GP experts, the gating network is defined by partitioning the input space into axis-aligned rectangular regions. More flexible partitioning approaches have also been proposed, such as Voronoi tessellations (Pope et al., 2018).

In contrast to these discriminative approaches, the alternative infinite mixture of GP experts (Meeds and Osindero, 2006) is a generative model for the joint distribution of the inputs and targets. Advantages include the ability to handle missing data and answer inverse problems, as well as more interpretable parameters of the local input model, which implicitly define the gating network, easing prior specification. Moreover, computations are simplified, relying on the different available representations and established algorithms for infinite mixtures of exchangeable data (e.g. Neal, 2000; Kalli et al., 2011; Blei and Jordan, 2006). In Meeds and Osindero (2006), the inputs are modelled with a local multivariate Gaussian distribution, and in multi-dimensions, complexity in the marginal of the inputs may lead to an unnecessarily large number of experts, degrading the predictive performance and increasing uncertainty, due to small sample sizes for each expert. This constraint is removed in Yuan and Neubauer (2009) by using a Gaussian mixture for the local input density; however, a finite approximation to the infinite mixture is used at both levels. Moreover, the local multivariate Gaussian scales poorly with the input dimension D due to the computational cost of dealing with the full D by D matrix.

To provide a unifying framework for multiple output types, alternative infinite mixtures of generalised linear experts are developed in Hannah et al. (2011). In this linear setting, Wade et al. (2014) highlights the issues associated with an overly large number of experts, causing a loss of predictive accuracy and increased uncertainty, particularly as D increases. For mixtures of GP experts, the greater flexibility of GPs over linear experts exacerbates these problems. In the following, we construct a novel mixture of GP experts to overcome these issues, inspired by Wade et al. (2014), that also provides a unifying framework for multiple input and output types.

3 ENRICHED MIXTURES OF GENERALISED GAUSSIAN PROCESS EXPERTS

A mixture model for the joint density of the output $y \in \mathcal{Y}$ and D-dimensional input $x \in \mathcal{X}$ assumes

$$f_Q(y, x) = \int p(y|x, \theta)p(x|\psi)dQ(\theta, \psi). \qquad (1)$$

The three key elements are 1) the **local expert** $p(y|x, \theta)$, a family of densities on $\mathcal{Y}$ for $\theta \in \Theta$; 2) the **local input model** $p(x|\psi)$, a family of densities on $\mathcal{X}$ for $\psi \in \Psi$; and 3) the **mixing measure** Q, a probability measure on $\Theta \times \Psi$. In the following, we define these three key elements for our model.

3.1 Local Experts

We provide a framework for multiple output types by defining the local expert $p(y|x, \theta)$ to be an extension of the generalised linear model (GLM) used in Hannah et al. (2011). Specifically, $p(y|x, \theta)$ belongs to the exponential family, which in canonical form assumes

$$p(y|x, \theta) = \exp\left(\frac{y\eta - b(\eta)}{a(\phi)} + c(y, \phi)\right),$$

where the functions a, b, and c are known and specific to the exponential family with parameters $\theta = (\phi, \nu)$; ϕ is the scale parameter; and η is the canonical parameter with $b'(\eta) = \mu(x) = \mathbb{E}[y|x]$ and $g(\mu(x)) = m(x)$, where g is a chosen link function that maps $\mu(x)$ to the real line. In GLMs (McCullagh and Nelder, 1989),

a linear function of x determines the canonical parameter, i.e. $m(x) = \alpha + x\beta$.

Instead, we consider a general non-linear function and assign a GP prior to the unknown function:

$$m(\cdot)|\beta_0, \lambda \sim \text{GP}(\beta_0, K_\lambda),$$

with constant mean function, $\mathbb{E}[m(x)] = \beta_0$, and kernel function K_λ with hyperparameters λ, defining the covariance of the function at any two inputs, $\text{Cov}[m(x), m(x_*)] = K_\lambda(x, x_*)$. The parameters of this generalized Gaussian process (GGP, Chan and Dong, 2011) are $\theta = (m(\cdot), \beta_0, \lambda, \phi)$. In many examples, it is common to use a zero-centred GP, which is made appropriate by subtracting the overall mean from the response. However, in our case, we must include a constant mean, as the partitioning structure is unknown and the data cannot be centred for each expert. Additionally, by including λ in the set of mixing parameters θ, we can recover non-stationary behaviour, e.g. different length-scales in local regions of the input space. The GGP experts used in Section 5 are 1) **Gaussian** with identity link, $p(y|x, \theta) = \text{N}(y|m(x), \sigma^2)$; and 2) **Ordinal** with probit link for ordered categories $l = 0, \ldots, L$,

$$\mathbb{P}(y \leq l|x, \theta) = \Phi\left[\frac{\varepsilon_l - m(x)}{\sigma}\right],$$

and cutoffs $0 = \varepsilon_0 < \varepsilon_1 < \ldots < \varepsilon_{L-1}$, which may be fixed due to the nonparametric nature of the model (Kottas et al., 2005). The ordinal model can be equivalently formulated through a latent Gaussian response:

$$\tilde{y}|m(x), \sigma^2 \sim \text{N}(m(x), \sigma^2),$$

$$p(y|\tilde{y}) = \begin{cases} \mathbf{1}(\tilde{y} \leq 0) & \text{if } l = 0 \\ \mathbf{1}(\varepsilon_{l-1} < \tilde{y} \leq \varepsilon_l) & \text{if } l = 1, \ldots, L-1 \\ \mathbf{1}(\tilde{y} > \varepsilon_{L-1}) & \text{if } l = L \end{cases},$$

with the ordered probit recovered after marginalisation of the latent $\tilde{y}$. A list of GGP experts is provided in the Supplementary Material (SM), for studies with other output types.

3.2 Local Input Models

We assume a factorised exponential family structure for the local input model. Specifically, it factorises across $d = 1, \ldots, D$, where each $p(x_d|\psi_d)$ belongs to the exponential family, that is,

$$p(x_d|\psi_d) = \exp(\psi_d^T t_d(x_d) - a_d(\psi_d) + b_d(x_d)),$$

and t_d, a_d, and b_d are known functions specified by the choice within the exponential family. The standard conjugate prior for ψ assumes independence of ψ_d across $d = 1, \ldots, D$ with

$$\pi(\psi_d) \propto \exp(\psi_d^T \tau_d - \nu_d a_d(\psi_d)),$$

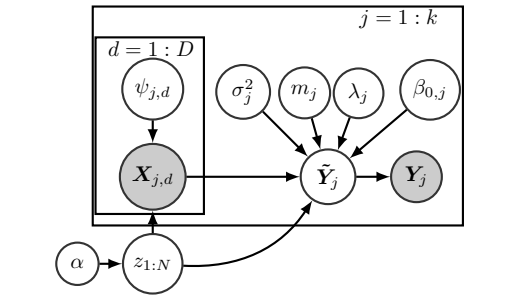
and parameters τ_d and ν_d determining the location and scale of the prior, respectively. In this conjugate setting, ψ can be marginalised, and the local marginal and predictive likelihood of the inputs are available analytically (specific calculations are provided in the SM). Examples used in Section 5 are the 1) **Gaussian** for $x_d \in \mathbb{R}$, with local input model $\text{N}(x_d|u_d, s_d^2)$; 2) **Categorical** for x_d taking *unordered* values $g = 0, 1, \ldots, G_d$, with local input model $\text{Cat}(x_d|\psi_d)$, where $\psi_d = (\psi_{d,0}, \ldots, \psi_{d,G_d})$ is a probability vector; and 3) **Binomial** for x_d taking *ordered* values $g = 0, 1, \ldots, G_d$ with local input model $\text{Bin}(x_g|G_d, \psi_d)$ for $\psi_d \in (0, 1)$.

Advantages of this factorised exponential form include improved scalability, inclusion of multiple input types, and richer parametrisation. Indeed, the mixtures of GP experts in Meeds and Osindero (2006); Yuan and Neubauer (2009); Nguyen and Bonilla (2014) consider only continuous inputs with a local multivariate Gaussian density and conjugate inverse Wishart prior on the covariance matrix. However, even for moderately large D, this approach becomes unfeasible. Specifically, the computational cost of dealing with the full covariance matrix is $O(D^3)$, which is reduced to $O(D)$ in our factorised form. Furthermore, Consonni and Veronese (2001) highlight the poor parametrisation of the Wishart prior; in particular, there is a single parameter to control variability. In our model, flexibility of the conjugate prior is enhanced, as it includes a scale parameter ν_d for each of the D variances. We emphasize that although the inputs are locally independent, globally, they may be dependent. For example, a highly-correlated, elliptically-shaped Gaussian can be accurately approximated with a mixture of several smaller spherical Gaussians.

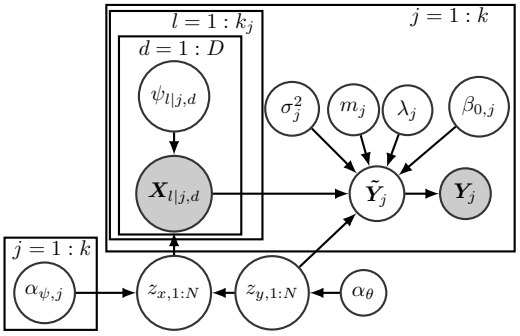
3.3 Mixing Measure

The Bayesian model is completed with a prior on the mixing measure Q, and the Dirichlet process (DP, Ferguson, 1973) is a popular nonparametric choice. Indeed, it is utilised in Rasmussen and Ghahramani (2002); Meeds and Osindero (2006); Hannah et al. (2011), among many others. Instead, we propose to use the enriched Dirichlet process (EDP, Wade et al., 2011) and highlight its advantages for improved prediction, better uncertainty quantification, and more interpretable clustering.

Dirichlet process. The parameters of the DP consist of the concentration parameter $\alpha > 0$ and the base measure Q_0, a diffuse probability measure on $\Theta \times \Psi$. The DP is discrete with probability one, and realisations place positive mass on a countably infinite number of atoms. When utilised as a prior for the mixing measure $Q \sim \text{DP}(\alpha, Q_0)$, this implies a countably in-



(a) DP MoE



(b) EDP MoE

Figure 1: Mixture of experts (MoE) with 1(a) DP prior and 1(b) EDP prior on the mixing measure Q. Here, $\boldsymbol{Y}_j$ and $\tilde{\boldsymbol{Y}}_j$ denote the observed and latent outputs in cluster j, with $\boldsymbol{X}_j$ denoting the inputs in cluster j for the DP and $\boldsymbol{X}_{l|j}$ denoting the inputs in x-cluster l nested in y-cluster j for the EDP.

finite mixture for the joint density in (1). For N data points (y_n, x_n), $n = 1, \ldots, N$, this induces a random partition of the data points into clusters, where data points belong to same cluster if they are generated from the same mixture component. Introducing the latent variable z_n denoting the cluster allocation of data point n, in order of appearance, and the parameters (θ_j, ψ_j) denoting the parameters of the j^{th} observed cluster, the mixing measure Q can be marginalised. In this case, the model can be expressed as

$$(y_n, x_n)|z_n = j, \theta_j, \psi_j \overset{ind}{\sim} p(y_n|x_n, \theta_j)p(x_n|\psi_j),$$

where $(\theta_j, \psi_j) \overset{iid}{\sim} Q_0$. The law of allocation variables is defined by the predictive distributions (Blackwell and MacQueen, 1973):

$$z_{N+1}|z_{1:N} \sim \frac{\alpha}{\alpha + N}\delta_{k+1} + \sum_{j=1}^{k} \frac{N_j}{\alpha + N}\delta_j,$$

where k is the number of clusters and N_j is the number of data points allocated to cluster j. In this setting, the number of clusters is determined by and can grow with the data.

Enriched Dirichlet process. The EDP defines a prior for the joint measure Q on $\Theta \times \Psi$ by decomposing it in terms of the marginal Q_θ and conditional $Q_{\psi|\theta}(\cdot|\theta)$. The parameters consist of the base measure Q_0 on $\Theta \times \Psi$, with marginal $Q_{0\,\theta}$ and conditional $Q_{0\,\psi|\theta}$, and concentration parameters α_θ and $\alpha_\psi(\theta)$ for $\theta \in \Theta$. The EDP assumes 1) $Q_\theta \sim \text{DP}(\alpha_\theta Q_{0\,\theta})$; 2) $Q_{\psi|\theta}(\cdot|\theta) \sim \text{DP}(\alpha_\psi(\theta)Q_{0\,\psi|\theta}(\cdot|\theta))$ for all $\theta \in \Theta$; and 3) independence of $Q_{\psi|\theta}(\cdot|\theta)$ across $\theta \in \Theta$ and from Q_θ. When utilised as a prior for the mixing measure $Q \sim \text{EDP}(\alpha_\theta, \alpha_\psi(\theta), Q_0)$, this induces a random nested partition of data points in y-clusters and x-subclusters within each y-cluster. The latent cluster allocation of each data point consists of two terms $z_n = (z_{y,n}, z_{x,n})$, where $z_{y,n} = j$ if the nth data point belongs to jth y-cluster with parameter θ_j and $z_{x,n} = l$ if the nth data point belongs to lth x-cluster with parameter $\psi_{l|j}$ within the jth y-cluster. After marginalising Q, the model can be expressed as

$$(y_n, x_n)|z_n = (j, l), \theta_j, \psi_{l|j} \overset{ind}{\sim} p(y_n|x_n, \theta_j)p(x_n|\psi_{l|j}),$$

where $\theta_j \overset{iid}{\sim} Q_{0\,\theta}$ and $\psi_{l|j}|\theta_j \overset{iid}{\sim} Q_{0\,\psi|\theta}(\cdot|\theta_j)$. The law of allocation variables is defined by:

$$z_{N+1}|z_{1:N} \sim \frac{\alpha_\theta}{\alpha_\theta + N}\delta_{(k+1,1)} + \sum_{j=1}^{k} \frac{N_j}{\alpha_\theta + N}\left(\frac{\alpha_{\psi,j}}{\alpha_{\psi,j} + N_j}\delta_{(j,k_j+1)} + \sum_{l=1}^{k_j} \frac{N_{l|j}}{\alpha_{\psi,j} + N_j}\delta_{(j,l)}\right),$$

where k denotes the number of y-clusters of sizes N_j and k_j denotes the number x-clusters within the j^{th} y-cluster of sizes $N_{l|j}$. Further, hyperpriors on the concentration parameters assume $\alpha_\theta \sim \text{Gam}(u_\theta, v_\theta)$ and $\alpha_{\psi,j} = \alpha_\psi(\theta_j)$ are independent with $\alpha_{\psi,j} \sim \text{Gam}(u_\psi, v_\psi)$.

A graphical comparison of the MoE with the DP and EDP priors is provided in Figure 1. The DP mixture of GGP experts allocates data points to similar groups to obtain a good approximation to the joint density, with similarity measured by the local expert and input model. The local factorised exponential family for the inputs is crucial for scaling to multi-dimensions and inclusion of multiple input types. However, this results in a rigid similarity measure between inputs, and as D increases x tends to dominate the partitioning structure, typically requiring many small clusters to capture increasing departures from the local input model. This occurs despite the flexible nature of GPs, often requiring only a few GP experts to approximate the conditional of y given x, and results in degradation of regression and conditional density estimates, wide credible intervals, and uninterpretable clustering due to the small sample sizes for each expert. By replacing the DP with the EDP, the nested partition-

Algorithm 1 Non-conjugate collapsed Gibbs sampler

Input: data $(y_n, x_n)_{n=1}^N$
Initialize: $(z_{1:N}^{(0)}, \sigma_{1:k}^{2\,(0)}, \beta_{0,1:k}^{(0)}, \lambda_{1:k}^{(0)}, \alpha_\theta^{(0)}, \alpha_{\psi,1:k}^{(0)}, \tilde{y}_{1:N}^{(0)})$
▷ *by sampling from the prior.*
for $m = 1$ **to** M **do**
  for $n = 1$ **to** N **do**
    Local updates to:

$$z_n^{(m)} | z_1^{(m)}, \ldots, z_{n-1}^{(m)}, z_{n+1}^{(m-1)}, \ldots, z_N^{(m-1)}$$

    ▷ *extending and combining Algorithm 3 and Algorithm 8 of Neal (2000) for the nested partition.*
  end for
  Global split-merge y-cluster updates to: $z_{1:N}^{(m)}$
  ▷ *Metropolis-Hastings step to move an x-cluster to be nested within a new or different y-cluster, with 'smart' proposals.*
  Global split-merge x-cluster updates to: $z_{x,1:N}^{(m)}$
  ▷ *extending Jain and Neal (2004); Wang and Russell (2015) to split or merge x-clusters nested within a common y-cluster, with 'smart-dumb' proposals.*
  Sample y-cluster parameters $(\sigma_{1:k}^{2\,(m)}, \beta_{0,1:k}^{(m)}, \lambda_{1:k}^{(m)})$
  ▷ *using Hamiltonian Monte Carlo (Duane et al., 1987).*
  Sample concentration parameters $(\alpha_\theta^{(m)}, \alpha_{\psi,1:k}^{(m)})$
  ▷ *with auxiliary variable techniques (Escobar and West, 1995).*
  Sample latent outputs $\tilde{y}_{1:N}^{(m)}$
  ▷ *if present, through Gibbs sampling and CDF inversion (Kotecha and Djuric, 1999).*
end for

ing scheme allows the data to determine if the conditional of y given x can be recovered with fewer experts. The y-clustering is determined by similarity measured through the local expert and a more flexible local input model, which can itself be a mixture. Moreover, a simple analytically computable allocation rule is maintained, allowing the construction of efficient inference algorithms.

4 POSTERIOR INFERENCE

For inference, we resort to Markov chain Monte Carlo (MCMC) and derive a collapsed Gibbs algorithm to sample the latent allocation variables $z_{1:N}$ and unique y-cluster parameters (θ_j), with the x-cluster parameters $(\psi_{l|j})$ marginalised. Additionally, we focus on the case when the functions $m_j(\cdot)$ can be marginalised; this includes the Gaussian likelihood, but also the ordered probit, among others, through data augmentation. In the latter, the data is augmented with latent Gaussian outputs $\tilde{y}_{1:N}$, which have a deterministic relationship with the observed outputs. Algorithm 1 gives an overview of the MCMC scheme.

Single-site Gibbs updates of the allocation variables can result in sticky chains, especially in high-dimensions. To improve mixing, two novel split-merge updates are developed to allow for global changes to the allocation variables. The first set of moves proposes to move an x-cluster to be nested within a new or different y-cluster, with 'smart' proposals to increase the acceptance probability. The second set proposes to split or merge x-clusters nested within a common y-cluster; 'smart' proposals are again employed but, in this case, need to be paired with corresponding 'dumb' proposals (i.e. random allocations), in order to increase the acceptance probability due to reversibility constraints of the Metropolis-Hastings algorithm (Wang and Russell, 2015). The SM contains a full description of the algorithm for the interested reader.

Predictions. From the M MCMC samples, we compute predictions for the new output y_* given x_*. In the Gaussian case, the posterior expectation of y_* is approximated by

$$\mathbb{E}[y_* | x_*, y_{1:N}, x_{1:N}] = C^{-1} \left(\sum_{m=1}^{M} p_{k^{(m)}+1}^{(m)}(x_*) \mu_\beta + \sum_{j=1}^{k^{(m)}} p_j^{(m)}(x_*) \widehat{m}_j^{(m)}(x_*) \right),$$

where C is the normalising constant; and for a new cluster, the predictive mean is simply the prior expectation of $\beta_{0,k^{(m)}+1}$, denoted by μ_β, while for an existing cluster, the GP predictive mean in cluster j is denoted by $\widehat{m}_j^{(m)}(x_*)$. Thus, for each sample, the expectation is a weighted average of the GP predictions from each cluster and a new cluster, with input-dependent weights that flexibly measure the similarity between the new input and the inputs of each cluster through a mixture. Specifically, the weights of a new cluster and existing cluster j, for $j = 1, \ldots, k^{(m)}$, are, respectively,

$$\begin{aligned} p_{k^{(m)}+1}^{(m)}(x_*) &= \frac{\alpha_\theta^{(m)}}{\alpha_\theta^{(m)} + N} h(x_*), \\ p_j^{(m)}(x_*) &= \frac{N_j^{(m)}}{\alpha_\theta^{(m)} + N} \left[\frac{\alpha_{\psi,j}^{(m)}}{\alpha_{\psi,j}^{(m)} + N_j^{(m)}} h(x_*) + \sum_{l=1}^{k_j^{(m)}} \frac{N_{l|j}^{(m)}}{\alpha_{\psi,j}^{(m)} + N_j^{(m)}} h(x_* | \boldsymbol{X}_{l|j}^{(m)}) \right]. \end{aligned} \quad (2)$$

Here, $h(x_*)$ is the marginal density of x^* and $h(x_* | \mathbf{X}_{l|j})$ is the predictive marginal density of x_* given $\mathbf{X}_{l|j}$, which contains the x_n such that $z_n = (j, l)$.

In contrast, for the DP, the weight of an existing cluster in (2) is more rigidly defined based on a single predictive marginal density, arising from the factorised exponential family. The predictive density or appropriate quantities for other output types are similarly computed. The joint approach also allows calculation of predictions based only on a subset of inputs, which are useful for visualisation and for test cases with missing inputs. Full derivations are provided in the SM.

Clustering. The enriched MoE induces a nested clustering of the data points into y-clusters and nested x-clusters. This latent clustering may be of interest to identify similar groups of data points and to improve understanding of the model. The MCMC samples from the posterior over this nested clustering; to summarise the samples, we obtain the point estimate from minimising the posterior expected variation of information (VI, Wade and Ghahramani, 2018), first estimating the y-level partition and then the nested x-level partition. To visualise uncertainty in the clustering structure, we also compute the posterior similarity matrix with elements $p(z_{y,n} = z_{y,n'}|y_{1:N}, x_{1:N})$ representing the posterior probability that two points are clustered together and approximated by the fraction of times this occurred in the chain.

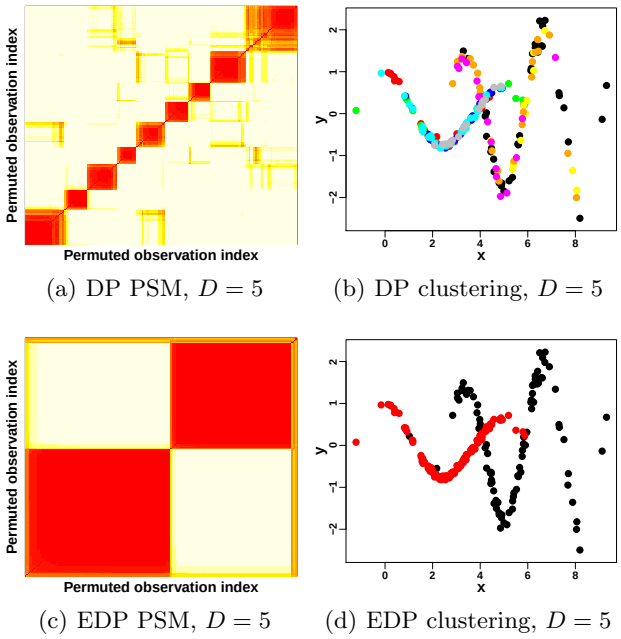
5 EXAMPLES

We demonstrate the advantages of the enriched MoE in two examples. Namely, improved predictive accuracy, smaller credible intervals, and more interpretable clustering. The first demonstrates increasing improvement over the DP as D increases, whilst the second shows the range of applicability of our model for ordinal outputs with multiple input types. Code to implement the model and reproduce the results is available at `github.com/cwlgadd/MixtureOfExperts`. Prior parameter specification, algorithm details and epoch times are detailed in the SM.

5.1 Simulated Mixture of Damped Cosine Functions

In the first example, $N = 200$ points are generated with only the first input as a predictor. The true output model is a highly non-linear regression obtained as a mixture of two damped cosines (Santner et al., 2003). The inputs are independently sampled from a multivariate normal. The additional inputs, which reduce the data's signal, are positively correlated among each other but independent of the first input.

We consider our EDP model with automatic relevance determination (ARD) squared exponential kernels for the GP experts. Benchmarks include the DP model; EDP model with linear experts; Lasso; GP; and TGP. This example demonstrates the robustness and ability of the EDP model to recover an underlying sparse data-generating structure, while the other methods fail to do so. Additional experiments with isotropic kernels are provided in the SM, showcasing the scalability of the GP experts with respect to D.



(a) DP PSM, $D = 5$ (b) DP clustering, $D = 5$

(c) EDP PSM, $D = 5$ (d) EDP clustering, $D = 5$

Figure 2: Simulated example for $D = 5$. *Left*: Heat map of the posterior similarity matrix (PSM), with observations indices permuted based on hierarchical clustering to improve visualisation. *Right:* VI clustering estimate with data points $(y_n, x_{n,1})$ coloured by cluster membership. Rows correspond to DP and EDP MoE, respectively. For the EDP, plots correspond to the y-level clustering.

The heat map of the posterior similarity matrix from the DP MoE in Figure 2(a) highlights data points with a high probability of clustering together in red; even for $D = 5$, the need for a large number of clusters is clearly evident. Indeed, the VI clustering estimate in Figure 2(b) contains nine clusters for $D = 5$. Conversely, the corresponding plots for the y-level clustering of the EDP in Figures 2(c)-2(d) highlight two y-clusters. Figure 3(a) emphasizes this improvement of the EDP in recovering the true number of y-clusters for increasing D (red line), while also employing a large number of x-clusters (red dashed line) to recover the marginal of x, in line with the DP (black line). Instead, the EDP with linear experts requires several y-clusters (pink line) for a local linear approximation to the highly non-linear function; however, for higher D it simply models the data as noise due to the inability to recover the sparse structure with small sample

sizes for each high-dimensional linear expert.

The improvement in the clustering leads to more accurate predictions and tighter credible intervals. We quantify the predictive accuracy in density regression with the L_1 error, that is the approximate L_1 distance between the estimated predictive response density and true data generating density, averaged across test samples. These errors are depicted in Figure 3(b), alongside the average length of the 95% credible intervals in Figure 3(c). As expected, the Lasso, an effective tool for sparse linear regression, performs quite poor in this highly non-linear example. However, the GP and TGP perform just as bad due to the inability to cope with bimodality in the non-axis aligned clustering. While the L_1 errors of the DP and EDP generally increase with D (as expected due to the increased input noise), the EDP is the most robust. Moreover, the EDP produces tighter credible intervals across D, compared to the other methods, while maintaining similar coverage (in the SM).

5.2 Alzheimer's Challenge

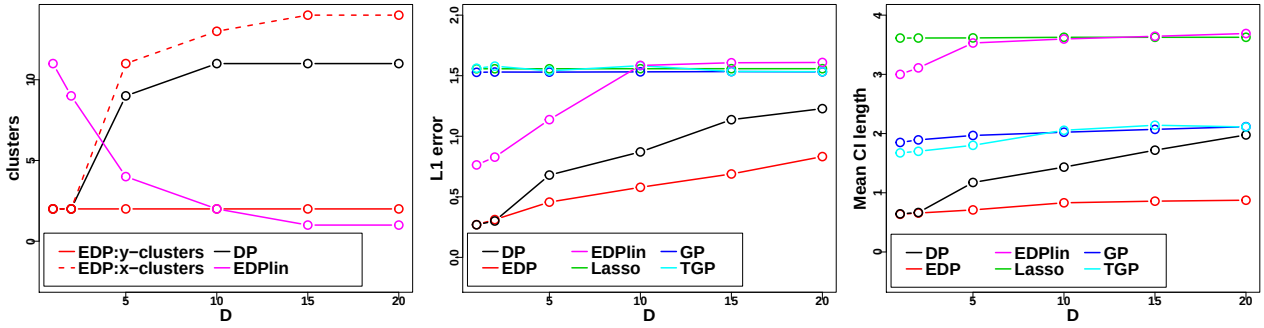
Motivated by the Alzheimer's Disease Big Data DREAM Challenge (`https://www.synapse.org/#!Synapse:syn2290704/wiki/60828`), this study aims to predict cognitive scores 24 months after initial assessment. This can potentially assist in early diagnosis of Alzheimer's disease (AD) and provide personalised predictions with uncertainty for patients and their families. Training data is extracted from the Alzheimer's Disease Neuro-Initiative (ADNI) database (`www.adni-info.org`). We emphasise that the competition test data can no longer be accessed, and the *test* results presented here are based on a random split of the data into training and test sets of sizes $N = 384$ and $N^* = 383$. The ordinal response y_n is the mini-mental state exam (MMSE) score at a 24 month follow-up visit; MMSE is an extensively used clinical measure of cognitive decline, defined on a 30 point scale with lower scores reflecting increased impairment. The $D = 6$ inputs include baseline age (in fraction of years); gender; baseline MMSE; education; APOE genotype, with values 0, 1, or 2, reflecting the number of copies of the type 4 allele; and diagnosis at baseline of cognitively normal (CN), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and AD, respectively. The winners of this subchallenge were GuanLab (GL) and ADDT. GL (Zhu and Guan, 2014) trained separate support vector machines (SVM) within each group of CN, MCI or AD; SVMs provide non-probabilistic predictions, and for comparison, we also train linear regression models within group to obtain prediction intervals in GL2. ADDT (Hwang et al., 2014) used robust regression based on M-estimation, optimally combined diagnosis and APOE4, and included interactions.

Table 1: Alzheimer's challenge. Comparison of EDP with DP and competition winners by 1) number of clusters; 2) mean absolute *test* error; 3) empirical coverage and 4) average length of 95% credible intervals.

	$\widehat{k}$	MAE_{test}	EC_{95}	$\bar{\text{CI}}_{95}$
EDP	3	**2.112**	0.948	8.96
DP	7	2.149	0.950	9.10
GL	3	2.153	-	-
GL2	3	2.208	0.945	11.06
ADDT	-	2.158	0.867	8.29

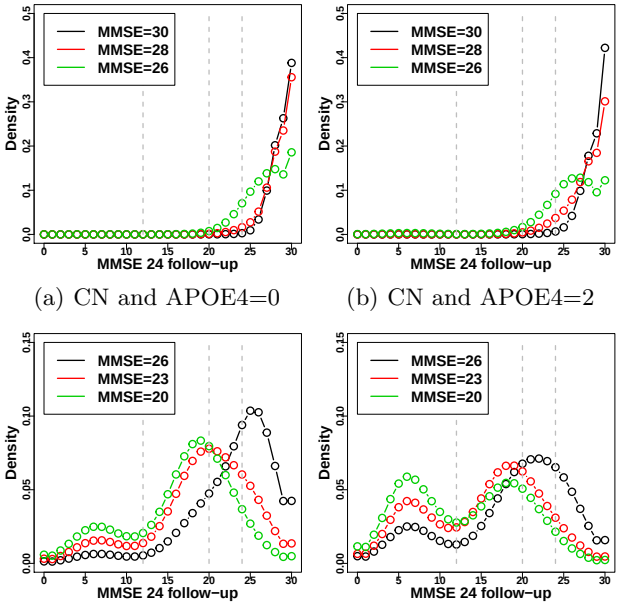
The EDP MoE can flexibly recover non-linear trajectories of the cognitive decline, while also clustering patients into input-dependent groups of similar trajectories. We employ ARD kernels to identify the relevant inputs within each cluster and consider the ordered probit GGP with fixed cutoffs $0 = \varepsilon_0 < \varepsilon_1 = 1 < \varepsilon_2 = 2 \ldots < \varepsilon_{29} = 29$. Table 1 summarises the *test* performance of the methods via the mean absolute error and empirical coverage and average length of the 95% credible intervals. Compared to the DP, the EDP performs slightly better in mean absolute *test* error and has smaller uncertainty, reflected in a reduced average credible interval length, while maintaining good coverage (≈ 0.95). This improvement is due to its ability to capture the relationship between y and x with fewer clusters, also leading to more interpretable clustering. Indeed, the VI estimate of the y-clustering of the EDP has only three clusters, while the DP has seven. Similar to GL, the EDP identifies three clusters of mostly CN, MCI, and AD individuals, with some adjustments for other variables, particularly, MMSE scores.

The EDP MoE produces flexible nonparametric density estimates of MMSE follow-up scores that change smoothly with the inputs. Specifically, Figure 4 shows how the densities become less peaked with larger variability for decreased baseline MMSE, increased APOE4, and increased severity in diagnosis. Instead, GL and ADDT are not able to capture this behaviour, e.g. with a minimum prediction interval length of 8 for ADDT, despite the high probability of follow-up MMSE close to 30 for CN individuals with a baseline MMSE of 30 in Figure 4. Also, note the apparent difference across APOE genotype for AD patients, with an increased probability of progressing to severe dementia for carriers in Figure 4. Thus, the EDP provides much improved uncertainty in predictions, which is particularly important in clinical settings and in relation to established cutoffs for MMSE. The SM con-



(a) Number of clusters (b) L_1 error (c) Average CI length

Figure 3: Simulated example. Comparison of the EDP MoE with the DP MoE, EDP model with linear experts, Lasso, GP, and TGP in terms of number of clusters in the VI estimate, approximate L_1 distance between the estimated and true conditional densities, and average length of the 95% credible intervals (CI).



(a) CN and APOE4=0 (b) CN and APOE4=2

(c) AD and APOE4=0 (d) AD and APOE4=2

Figure 4: Alzheimer's challenge. Marginal predictive density of MMSE 24-month follow-up scores for different combinations of MMSE baseline, APOE4, and baseline diagnosis from the EDP mixture of experts. Dashed lines indicate established cutoffs for MMSE: ≥ 25 suggests no dementia; $20-24$ suggests mild dementia; $13-19$ suggests moderate dementia; ≤ 12 suggests severe dementia.

tains a deeper discussion on the clustering and clinical relevance of the findings.

6 DISCUSSION

Infinite mixtures of GP experts are flexible models, that can capture non-stationary functions and departures from the typical homoscedastic normality assumptions on the errors. In this work, we proposed a novel enriched mixture of GGP experts, with local independence of the inputs, to increase scalability and allow inclusion of multiple input types, and a nested partitioning scheme, to improve predictive accuracy, uncertainty, and interpretability of the clustering. Moreover, through the generalised GP framework, we can account for different output types.

A number of proposals extend mixtures of linear experts for high-dimensional inputs using regularisation or variable selection, e.g. Peralta and Soto (2014); Barcella et al. (2017). Here, we consider GP experts with ARD kernels, that allow determination of the local relevance of each input, as well as with isotropic kernels, that result in improved scalability in high-dimensions. An important future research direction will incorporate methods to scale the GP experts to higher-dimensions, while also allowing local relevance determination, through dimension reduction techniques (Snelson and Ghahramani, 2006). We also note that high-dimensional examples with ARD kernels will likely require careful hyperprior specification on the length-scale parameters (van der Vaart and van Zanten, 2009) to recover the sparse structure.

To scale to larger datasets, future research will also focus on fast approximate inference such as MAP techniques (Raykov et al., 2016) that maintain a non-degenerate likelihood, enabling out-of-sample predictions and the use of standard tools such as cross-validation. For large sample sizes, further computational gains can be made through sparse or low-rank assumptions on the GP experts, (see e.g. Rasmussen and Williams, 2005, Chapter 8). Parallelisation (Chang and Fisher, 2013; Zhang et al., 2019) is also relevant to scale with large N, and other ideas in Zhang et al. (2019) may additionally be incorporated, to increase the acceptance of new clusters in single-site Gibbs, for high D and vague priors.

Acknowledgements

Sara Wade acknowledges support through the Warwick Academic Returners Fellowship. Charles Gadd acknowledges the Warwick Centre for Predictive Modelling where part of this work was completed.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.ucla.edu`). As such, the investigators within the ADNI provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD).

Data collection and sharing for the Alzheimer's disease study was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). We acknowledge the funding contributions of ADNI supporters (`http://adni.loni.usc.edu/about/#fund-container`).

References

W. Barcella, M. De Iorio, and G. Baio. A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models. *Canadian Journal of Statistics*, 45:254–273, 2017.

D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1(2):353–355, 1973.

D.M Blei and M.I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

A.B. Chan and D. Dong. Generalized Gaussian process models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2681–2688, 2011.

J. Chang and J.W. Fisher. Parallel sampling of DP mixture models using sub-cluster splits. In *Advances in Neural Information Processing Systems*, pages 620–628, 2013.

G. Consonni and P. Veronese. Conditionally reducible natural exponential families and enriched conjugate priors. *Scandinavian Journal of Statistics*, 28:377–406, 2001.

S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

R.B. Gramacy and H.K. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

L.A. Hannah, D.M. Blei, and W.B. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12:1923–1953, 2011.

J. Hwang, X. Shen, and Y. Pawitan. Mini-mental state examination change score prediction for early diagnosis of Alzheimer's disease. 2014. URL `https://www.synapse.org/#!Synapse:syn2759392/wiki/69612`.

R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2004.

M. Kalli, J.E. Griffin, and S.G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

J.H. Kotecha and P.M. Djuric. Gibbs sampling approach for generation of truncated multivariate Gaussian random variables. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1757–1760, 1999.

A. Kottas, P. Müller, and F. Quintana. Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14:610–625, 2005.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. London, UK, 1989.

E. Meeds and S. Osindero. An alternative infinite mixture of Gaussian process experts. In *Advances in Neural Information Processing Systems*, pages 883–890, 2006.

R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

T. Nguyen and E. Bonilla. Fast allocation of Gaussian process experts. In *Proceedings of the 31st International Conference on Machine Learning*, pages 145–153, 2014.

B. Peralta and A. Soto. Embedded local feature selection within mixture of experts. *Information Sciences*, 269:176–187, 2014.

C.A. Pope, J.P. Gosling, S. Barber, J. Johnson, T. Yamaguchi, G. Feingold, and P. Blackwell. Modelling spatial heterogeneity and discontinuities using Voronoi tessellations. *arXiv preprint arXiv:1802.05530*, 2018.

C.E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems*, pages 881–888, 2002.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Y.P. Raykov, A. Boukouvalas, and M.A. Little. Simple approximate MAP inference for Dirichlet processes mixtures. *Electronic Journal of Statistics*, 10(2):3548–3578, 2016.

T.J. Santner, B.J. Williams, and W.I. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York, 2003.

E. Snelson and Z. Ghahramani. Variable noise and dimensionality reduction for sparse Gaussian processes. In *Uncertainty in Artificial Intelligence*, pages 461–468, 2006.

Volker Tresp. Mixtures of Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 654–660, 2001.

A. W. van der Vaart and J. H. van Zanten. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.

S. Wade and Z. Ghahramani. Bayesian cluster analysis: point estimation and credible balls. *Bayesian Analysis*, 13(2):559–626, 2018.

S. Wade, S. Mongelluzzo, and S. Petrone. An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, 6:359–386, 2011.

S. Wade, D.B. Dunson, S. Petrone, and L. Trippa. Improving prediction from Dirichlet process mixtures via enrichment. *Journal of Machine Learning Research*, 15 (1):1041–1071, 2014.

W. Wang and S.J. Russell. A smart-dumb/dumb-smart algorithm for efficient split-merge MCMC. In *Uncertainty in Artificial Intelligence*, pages 902–911, 2015.

C. Yuan and C. Neubauer. Variational mixture of Gaussian process experts. In *Advances in Neural Information Processing Systems*, pages 1897–1904, 2009.

M.M. Zhang, S.A. Williamson, and F. Perez-Cruz. Accelerated inference for latent variable models. *arXiv preprint arXiv:1705.07178*, 2019.

F. Zhu and Y. Guan. Guanlab's solution to the 2014 DREAM Alzheimer's disease big data challenge (1st place). 2014. URL `https://www.synapse.org/#!Synapse:syn2527678/wiki/69937`.