Garrett Frady garrett.frady@uconn.edu

I am a PhD candidate in the Department of Statistics at the University of Connecticut. My advisor is Prof. Dipak K. Dey and my co-advisor is Dr. Shariq Mohammed. The research I conduct with my advisors resides in developing Bayesian methods involving a sparsity inducing prior structure to perform simultaneous estimation and feature extraction in the presence of sparse, high dimensional spatio-temporal data. With the computational ability available to handle massive amounts of data, we are more frequently handling high dimensional data with two-dimensional covariate structures. In these scenarios, correlation exists across multiple dimensions and may not be fully captured in a computationally efficient manner through traditional statistical or machine learning approaches.

Spatio-temporal data analysis has become popular as it can be used to study and model the dynamics of real-world processes that occur in space and time. Measurements from neuroimaging techniques such as electroencephalography (EEG) and magnetic resonance imaging (MRI) may be collected in a spatio-temporal pattern to examine statistical dependencies between the activity of different brain regions in response to a stimulus. As a way of applying our methods to practical situations, we work with multi-subject EEG data to identify active regions of the brain by examining the effects alcoholism has on the functional states of the frontal cortex and parietal lobe. From a statistical standpoint, this may be viewed as a feature extraction process. One of our major goals is to use the extracted active locations of the brain to make accurate subject-level predictions of mental-related illnesses. Understanding how different locations of the brain communicate can play a major role in detecting early onset of mental-related illnesses. Although we apply our methods to neuroimaging data with the goal of classifying subjects according to their EEG measurements, it is worth noting that they can be extended to other application domains with similar structure and characteristics.

Past and Current Work

Often times in practical scenarios, only a fractional subset of the large predictor space may exhibit significant associations with the response; referred to as the *"sparse high dimensional problem"*. Inspired by the previously mentioned concept, numerous studies (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006) have leveraged the penalized loss function (PL) estimator to accomplish simultaneous estimation and feature extraction. The PL estimator in the aforementioned studies was defined as

$$\hat{\boldsymbol{\beta}}_{PL} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}[L(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta})) + Pe(\boldsymbol{\beta}|\lambda)], \tag{1}$$

where $L(\cdot,\cdot)$ is a loss function and $Pe(\cdot|\lambda)$ is a penalty function with regularization parameter $\lambda > 0$. The penalty function and the regularization parameter play a major role in the degree of sparsity introduced into the estimator. The ℓ_0-norm penalty was introduced by Akaike (1974), and further investigated by Schwarz (1978), to induce sparsity into the PL estimator in Eq. (1). The ℓ_0-norm penalty is defined as $\ell_0(\boldsymbol{\beta}|\lambda) = \lambda \sum_{j=1}^{p} \mathbb{1}_{(\beta_j \neq 0)}$, where $\lambda \geq 0$ is a regularization parameter and $\mathbb{1}_{(\cdot)}$ is an indicator function. Several frequentist methods followed with the same goal of reducing the dimension of the problem. In a Bayesian framework, the PL estimator in Eq. (1) can be viewed as the maximum a priori (MAP) estimator of the posterior distribution given by $\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \propto \exp\{-[L(\boldsymbol{y}, h(\boldsymbol{X}\boldsymbol{\beta})) + Pe(\boldsymbol{\beta}|\lambda)]\}$, where $f(\boldsymbol{y}|\boldsymbol{\beta})$ is the likelihood and $\pi(\boldsymbol{\beta})$ is the prior (Tibshirani, 1996). This way, uncertainty in the estimator can be quantified by sampling from the posterior. Moving forward under this paradigm, we construct likelihood, prior, and posterior distributions.

In the first two chapters of my dissertation, we conquer the task of feature extraction and estimation in high dimensional problems through a local Bayesian modeling approach utilizing the Gaussian and Diffused-gamma (GD) prior (Goh and Dey, 2018). To bypass the computational complexity, we build local binary classification models of subject level responses at each time point. Following the formulation of $\hat{\boldsymbol{\beta}}_{PL}$ in Eq. (1), we choose the wide-ranging class of Bregman divergence (BD) measures as our loss function. To develop a general likelihood function, we highlight the *duality property* between the loss function and the likelihood function, which affirms that the negative log-likelihood function can be thought of as a loss function. Extending the aforementioned dual relationship, we assume the negative log-likelihood function is a BD measure. In particular, we define our likelihood function as $f_\psi(\boldsymbol{y}|\boldsymbol{\beta}) \propto exp\{-BD_\psi(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta}))\}$, which corresponds to the loss function in Eq. (1). The GD prior was developed to closely emulate the ℓ_0-norm penalty defined above, which corresponds to the

penalty function in Eq. (1). To make this connection, Goh and Dey (2018) defined a precise approximation of the ℓ_0-norm penalty, $\ell_0(\boldsymbol{\beta}|\lambda)$, given by $\tilde{\ell}_0(\boldsymbol{\beta}|\lambda,\tau_0) = \lambda\sum_{j=1}^{L}\frac{\{(\boldsymbol{X}_{[,j]})^T\boldsymbol{X}_{[,j]}\}\beta_j^2}{\tau_0^2+\{(\boldsymbol{X}_{[,j]})^T\boldsymbol{X}_{[,j]}\}\beta_j^2}$, where $\tau_0 > 0$ is a constant chosen to be sufficiently small and $\boldsymbol{X}_{[,j]}$ is the j^{th} column of design matrix $\boldsymbol{X}$.

The GD prior is defined as follows: $\pi(\boldsymbol{\beta},\boldsymbol{d}) \propto \pi_G(\boldsymbol{\beta}|\boldsymbol{d})\pi_D(\boldsymbol{d})$, where $\pi_G(\boldsymbol{\beta}|\boldsymbol{d}) \propto \prod_{j=1}^{L}[d_j^{1/2}exp\{-\frac{d_j}{2}\beta_j^2\}]$ and $\pi_D(\boldsymbol{d}) \propto \prod_{j=1}^{L}[d_j^{\lambda-1/2}exp\{-\frac{\tau_0^2}{2\{(\boldsymbol{X}_{[,j]})^T\boldsymbol{X}_{[,j]}\}}d_j\}]$, where $\lambda \geq 0$ is a tuning parameter and $\tau_0 > 0$ is a constant chosen to be sufficiently small. In the first paper, we assigned λ a non-informative gamma prior, i.e., $\lambda \sim Gam(\alpha_1,\alpha_2)$, where α_1 and α_2 are shape and rate parameters, respectively. Since $\pi_D(\boldsymbol{d})$ is independent of $\boldsymbol{\beta}$, $\pi_{GD}(\boldsymbol{\beta},\boldsymbol{d}) \propto \pi_G(\boldsymbol{\beta}|\boldsymbol{d})$ with respect to $\boldsymbol{\beta}$, and thus, for any $\lambda \geq 0$ and $\tau_0 > 0$, we can show $\arg\max_{\boldsymbol{\beta}}\{f_\psi(\boldsymbol{y}|\boldsymbol{\beta})\pi_{GD}(\boldsymbol{\beta},\hat{\boldsymbol{d}})\} = \arg\min_{\boldsymbol{\beta}}\{BD_\psi\{\boldsymbol{y},\boldsymbol{h}(\boldsymbol{X\beta})\} + \tilde{\ell}_0(\boldsymbol{\beta}|\lambda,\tau_0)\}$, where $\hat{\boldsymbol{d}}$ is the MAP estimator of $\boldsymbol{d}$. Hence, the MAP estimator for $\boldsymbol{\beta}$ is a nice approximation of the PL estimator with a BD measure as the loss function and $\tilde{\ell}_0(\cdot|\lambda,\tau)$ defined above as the penalty function.

A key component of our GD approach is to replace the latent vector $\boldsymbol{d}$ by its MAP estimator, rather than integrating it out. Since $\pi_{GD}(\boldsymbol{\beta},\boldsymbol{d}) \propto \pi_G(\boldsymbol{\beta}|\boldsymbol{d})$ with respect to $\boldsymbol{\beta}$, the posterior distribution can be viewed as $\pi(\boldsymbol{\beta}|\boldsymbol{y},\hat{\boldsymbol{d}}_{MAP}) \propto f_\psi(\boldsymbol{y}|\boldsymbol{\beta})\pi_{GD}(\boldsymbol{\beta},\hat{\boldsymbol{d}}_{MAP}) \propto f_\psi(\boldsymbol{y}|\boldsymbol{\beta})\pi_G(\boldsymbol{\beta}|\hat{\boldsymbol{d}}_{MAP})$. Using the fact that the GD prior is proper and the likelihood function is non-degenerate and bounded everywhere, we can show that the GD posterior distribution is proper. The full conditional posterior distributions are given by $\pi(\boldsymbol{d}|\boldsymbol{y},\boldsymbol{\beta}) \propto \pi_G(\boldsymbol{\beta}|\boldsymbol{d})\pi_D(\boldsymbol{d})$ and $\pi(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{d}) \propto f_\psi(\boldsymbol{y}|\boldsymbol{\beta})\pi_G(\boldsymbol{\beta}|\boldsymbol{d})$. We represent our prior knowledge about β_j by using the fact that $\beta_j|\hat{d}_j \overset{ind.}{\sim} N(0,1/\hat{d}_j)$. To bypass the computational difficulty in updating all of the coefficients simultaneously, we make use of the univariate full conditional distributions given by $\pi(d_j|\text{others}) \propto d_j^{\lambda}\ exp\{-\frac{\tau_0^2/\{(\boldsymbol{X}_{[,j]})^T\boldsymbol{X}_{[,j]}\}+\beta_j^2}{2}d_j\}$ and $\pi(\beta_j|\text{others}) \propto \exp\{-BD_\psi(\boldsymbol{y},\boldsymbol{h}(\boldsymbol{X\beta})) - \frac{d_j}{2}\beta_j^2\}$, $j = 1,\ldots,L$. The full conditional posterior distribution of λ is $\pi(\lambda|others) \propto \lambda^{\alpha_1-1}\exp\{-\sum_{j=1}^{L}(\frac{\hat{d}_j}{2}) - \alpha_2\lambda\}$.

The estimation procedure is conducted using the Markov chain Monte Carlo (MCMC) sampling algorithm at each local model. We additionally implement a component-wise Gibbs sampler by iteratively generating β_j's from their component-wise full conditionals. Since a closed form of the full conditional distribution of β_j may not exist, we use the Metropolis-Hastings algorithm within the Gibbs sampler. Once we obtain the estimates from the MCMC sampling procedure, we implement a two-stage feature extraction process at each local model to induce sparsity, and determine which locations are active. In the numerical studies, we made use of the **R** package `rstan` (Stan Development Team, 2023), which implements a Hamiltonian Monte Carlo (HMC) sampling procedure. This approach is more computationally efficient, especially for the real data and larger simulation settings. HMC uses the derivatives of the density function being sampled to generate efficient transitions spanning the posterior (Betancourt and Girolami, 2015; Neal et al., 2011). In the first stage of the two-stage feature extraction process, we use a Bayesian averaging FDR approach as a way of forcing coefficients of smaller magnitude to 0, while maintaining the estimates of coefficients of higher magnitude. In the second stage, we compute the area under the curve formed by the beta estimates obtained from stage 1 and perform k-means clustering with $k = 2$ on the resulting areas under the curve. We expect the model to select locations of the brain which are significantly correlated to predicting a genetic disposition caused by alcohol. Once the feature extraction process is complete, at each time point, the ith subject is assigned a prediction probability, $\hat{p}_{it} = h(\boldsymbol{x}_{it}^T\hat{\boldsymbol{\beta}}_t)$, where $\hat{\boldsymbol{\beta}}_t$ is the local estimated coefficient vector; $\hat{p}_{it}$ represents the likeliness of the subject being classified as a 1 over a 0. In the first paper, $h(\cdot)$ is the inverse logit link function. We construct local weights, w_{it}, based on the local prediction probabilities with the goal of allocating higher priority to local models with more certainty of classifying the subject as a 0 or 1. The weights for subject i are defined as $w_{it} = \frac{(\hat{p}_{it}-0.5)^2}{\sum_{t=1}^{\tau}(\hat{p}_{it}-0.5)^2}$, $t = 1,\ldots,\tau$. The final predicted response for subject i is given as $\hat{y}_i = 1$ if $\sum_{t=1}^{\tau} w_{it}\hat{p}_{it} > 0.5$; and $\hat{y}_i = 0$ otherwise.

The results from the numerical studies in the first chapter of my dissertation proved to be comparable to preexisting methods in literature with respect to subject-level prediction accuracy and the ability to detect significant (l,t) pairs. Over and above that, our GD approach determines which locations of the brain are active. Therefore, we can additionally identify patterns of functional connectivity associated with cognitive or behavioral states. It is worth noting that the electrodes which were selected most often are located in the temporal lobe and parietal lobe of the brain, which both play a crucial role in a wide range of sensory and cognitive processes. The parietal lobe is responsible for processes such as sensory integration and spatial awareness, and has been referred to as the association region of the brain Ackerman et al. (1992). The temporal lobe, which is part of the frontal cortex, is responsible for processes such as memory formation and object recognition, and is often referred to as the neocortex Ackerman et al. (1992).

In the second chapter of my dissertation, we are incorporating structure into the GD prior formulation and exploring the impact of using different link functions. The GD prior in the first paper lacked structure in that the spatial and temporal dimensions of the data were not directly exploited. The covariates in each local model were electrical signals observed at locations which were assumed to be independently distributed. Since local models were constructed at each time point, we ignored the temporal structure and did not reintroduce it back into the model through feature extraction or prediction. We are aware that correlation exists across multiple dimensions, so we want to introduce correlation into GD prior to account for such behavior in the EEG data. Intuitively, locations in the brain which are closer in distance will be more highly correlated than those which are spaced further apart. As a way of introducing spatial correlation into the model, we define the covariance term of the multivariate normal distribution assigned to $log(\boldsymbol{\lambda}_t) = (log(\lambda_{1t}), \ldots, log(\lambda_{Lt}))^T \in \mathbb{R}^L$ as a function of a distance-based correlation matrix. That is, the regularization parameters $\boldsymbol{\lambda}_t = (\lambda_{1t}, \ldots, \lambda_{Lt})^T \in \mathbb{R}^L$ are modeled as multivariate log normal. With this setup, if location i and location j are close together, λ_i and λ_j will be highly correlated, which will result in d_i and d_j having a similar mean and variance. Thus, the estimates of d_i and d_j will coincide. Consequently, the estimates of model coefficients β_i and β_j will be associated. Similar to the spatial structure, the previous time point will be highly correlated with the current time point. To account for temporal correlation, we borrow information, through the mean of the multivariate log normal distribution on $\boldsymbol{\lambda_t}$, from the previous time point to construct the model at the current time point. Using this formulation, both the spatial and the temporal structure of the data are directly accounted for in the model.

With the choice of BD measures as a loss function, our model is able to account for a variety of likelihood functions. The response variable in the EEG data is binary, but we can handle responses from any member of the natural exponential family. In non-linear regression, deciding on a link function is an important step. Of course the distribution of the response variable plays a major role in selecting a link function, but there are multiple other characteristics of a data set to be considered in this decision process. In the first paper, we used the logit link function, however, due to the lopsided frequency of 0's and 1's in the data, it makes sense to consider asymmetrical link functions. In fact, we consider two competing families of skewed link functions, generalized extreme value (GEV) family and power family. Specifically, we are investigating the complementary log-log (c-log-log) link function from the GEV family (Wang and Dey, 2010), which is commonly used in binary response models, and the generalized logit link function from the power family (Ordoñez et al., 2023). The choice of link function will influence which assumptions from the first paper can be relaxed.

Future Work

Looking ahead, we will investigate a more advanced and refined feature extraction approach. In the pursuit of an enhanced and more effective statistical approach to solve multidimensional optimization problems, it is imperative to continue improving the methods of feature extraction and prediction. To better capture the complicated characteristics within the data, we will introduce a non-parametric component into the model. Thus, we will no longer fit conventional generalized linear models (GLMs). Instead, the data will be modeled as a Gaussian process. However, the parameters of interest, $\boldsymbol{\beta}$, will still be modeled with the structured GD prior. That is, the new model will be a sum of the Bayesian GLM previously described and a Gaussian process. The flexibility of the model to handle different response variables will be preserved. Building on the approach seen in Hu and Dey (2023), we will propose a feature extraction algorithm for a Gaussian process model using artificial nuisance columns as a baseline to extract active features and determine structural relationships between features. The feature extraction algorithms for Gaussian process models will work for both regression and classification problems Hu and Dey (2023). Additional information obtained from the Gaussian process feature extraction approach will explain more of the variability in the response and assist in making more accurate prediction. Keeping in mind that a major objective of our research is to predict the presence of neurological illnesses, accuracy and interpretability are highly sought after.

A Gaussian process is a collection of random variables, any finite number of which have joint Gaussian distributions (Rasmussen et al., 2006). Gaussian processes are entirely specified by a mean function $m(\boldsymbol{x})$ and a covariance function $C(\boldsymbol{x}, \boldsymbol{x}')$: $f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), C(\boldsymbol{x}, \boldsymbol{x}'))$. For convenience, the mean function is chosen to be 0. Then for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^K$, $(f(\boldsymbol{x}_1, \ldots, f(\boldsymbol{x}_n))^T \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma_{i,j} = C(\boldsymbol{x}_i, \boldsymbol{x}_j)$. A common choice of $C(\cdot, \cdot)$ is the radial basis function (RBF) kernel. Hu and Dey (2023) adopted a new covariance function referred to as the reparameterized inverse-RBF kernel. We will explore a more general kernel known as the Matern kernel; the RBF kernel is a special case. We will further investigate how the

kernel changes from a Gaussian process to a log Gaussian process. The choice of link function will depend on the distribution of the response variable.

A traditional Gaussian process model has the form

$$E[y_i|\boldsymbol{x}_i] = \mu_i = h^{-1}(f(\boldsymbol{x}_i)), \qquad i = 1, \ldots, n,$$
$$f(\boldsymbol{x}) \sim \mathcal{GP}(\boldsymbol{0}, C(\cdot, \cdot|\boldsymbol{\theta})),$$

where $\boldsymbol{x}_i$ is the feature vector of the i^{th} observation and $h(\cdot)$ is the link function. The new model we will be proposing has the form

$$E[y_i|\boldsymbol{x}_i] = \mu_i = g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}) + h^{-1}(f(\boldsymbol{x}_i)), \qquad i = 1, \ldots, n,$$
$$f(\boldsymbol{x}) \sim \mathcal{GP}(\boldsymbol{0}, C(\cdot, \cdot|\boldsymbol{\theta}),$$

where $g(\cdot)$ is the link function connecting the response variable to the linear predictor. This is different from $h(\cdot)$, which connects the response directly to the features. In the EEG case study, a feature vector corresponds to a location of the brain and entries of the vector correspond to the electrical signal in the brain at that location over all time points.

References

Ackerman, S. et al. (1992), *Discovering the brain*, National Academies Press.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723.

Betancourt, M. and Girolami, M. (2015), 'Hamiltonian monte carlo for hierarchical models', *Current trends in Bayesian methodology with applications* **79**(30), 2–4.

Fan, J. and Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American statistical Association* **96**(456), 1348–1360.

Goh, G. and Dey, D. K. (2018), 'Bayesian map estimation using gaussian and diffused-gamma prior', *Canadian Journal of Statistics* **46**(3), 399–415.

Hu, Z. and Dey, D. K. (2023), 'Generalized variable selection algorithms for gaussian process models by lasso-like penalty', *Journal of Computational and Graphical Statistics* **0**(0), 1–10.
URL: *https://doi.org/10.1080/10618600.2023.2256802*

Neal, R. M. et al. (2011), 'Mcmc using hamiltonian dynamics', *Handbook of markov chain monte carlo* **2**(11), 2.

Ordoñez, J. A., Prates, M. O., Bazán, J. L. and Lachos, V. H. (2023), 'Penalized complexity priors for the skewness parameter of power links', *Canadian Journal of Statistics* .

Rasmussen, C. E., Williams, C. K. et al. (2006), *Gaussian processes for machine learning*, Vol. 1, Springer.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The annals of statistics* pp. 461–464.

Stan Development Team (2023), 'RStan: the R interface to Stan'. R package version 2.26.23.
URL: *https://mc-stan.org/*

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Wang, X. and Dey, D. K. (2010), 'Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption'.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American statistical association* **101**(476), 1418–1429.