Machine Learning (2022) 111:3733–3767
https://doi.org/10.1007/s10994-022-06174-z

Semi-parametric Bayes regression with network-valued covariates

Xin Ma[1] · **Suprateek Kundu**[2] · **Jennifer Stevens**[3]

Received: 1 March 2021 / Revised: 31 January 2022 / Accepted: 3 February 2022 /
Published online: 2 June 2022
© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2022

Abstract
Although there has been an explosive rise in network data in a variety of disciplines, there is very limited development of regression modeling approaches based on high-dimensional networks. The scarce literature in this area typically assume linear relationships between the outcome and the high-dimensional network edges that results in an inflated model plagued by the curse of dimensionality and these models are unable to accommodate non-linear relationships or higher order interactions. In order to overcome these limitations, we develop a novel two-stage Bayesian non-parametric regression modeling framework using high-dimensional networks as covariates, which first finds a lower dimensional node-specific representation for the networks, and then embeds these representations in a flexible Gaussian process regression framework along with supplemental covariates for modeling the continuous outcome variable. Moving from edge-level analysis to node-level model allows us to scale up to high-dimensional networks, and enables node selection via an extension of the Gaussian process framework that involves spike-and-slab priors on the lengthscale parameters. Extensive simulations show a distinct advantage of the proposed approach in terms of prediction, coverage, and node selection. The proposed model achieves considerable gains when predicting posttraumatic stress disorder (PTSD) resilience based on brain networks in our motivating neuroimaging applications, and also identifies important brain regions associated with PTSD. In contrast, existing non-linear approaches that employ the full-edge set or those that use other dimension reduction techniques on the network are not equipped for node selection and results in poor prediction and characterization of predictive uncertainty, while linear approaches using the edge-level features are overly inflated and typically result in poor performance.

Keywords Dimension reduction · Gaussian process regression · Latent scale network models · Manifold · Posttraumatic stress disorder

Editors: João Gama, Alípio Jorge, Salvador García.

✉ Suprateek Kundu
skundu2@mdanderson.org

Extended author information available on the last page of the article

1 Introduction

Recent studies have given rise to a rich variety of network data involving fraud detection (Akoglu et al., 2015), brain networks (Lukemire et al., 2020), social networks (Liben-Nowell and Kleinberg 2007), traffic forecasting (Cui et al., 2019; Li et al., 2017), computer vision (Monti et al., 2017) and so on, that arise in diverse applications. Networks provide inherently richer insights by delving into relationships between nodes that represent different entities in different application domains (nodes can represent brain regions in neuroimaging problems, an individual in social networks, sensors in traffic forecasting, and so on). By using the web of relationships represented by the network, one can obtain more accurate prediction and analysis in neuroimaging applications (Guha and Rodriguez 2020), traffic forecasting (Li et al., 2017), synthetic identity dectection (Zhang et al., 2017), financial fraud detection Zhou et al. (2017), and other scientific applications. Similar to other domains, classification and prediction problems based on networks has gained increasing prominence in our motivating neuroimaging applications involving mental health, where network differences between disease versus control groups resulting from disruptions in the brain functional and structural connectivity are well-established (Higgins et al., 2019; Zhan and Yu 2015). Such brain network disruptions are clearly evident in our motivating Grady Trauma project (GTP) application that shows considerable differences in brain connectivity between individuals with high and low posttraumatic stress disorder (PTSD) resilience (see Fig. 1).

While a limited number of classification methods using high-dimensional network-valued covariates have been proposed (Du et al., 2018; Relión et al., 2019), the literature on prediction models using networks is unfortunately even more sparse and has several crucial pitfalls. For example, most classification and prediction approaches using network valued covariates assume linear relationships, and they often do not account for the complex

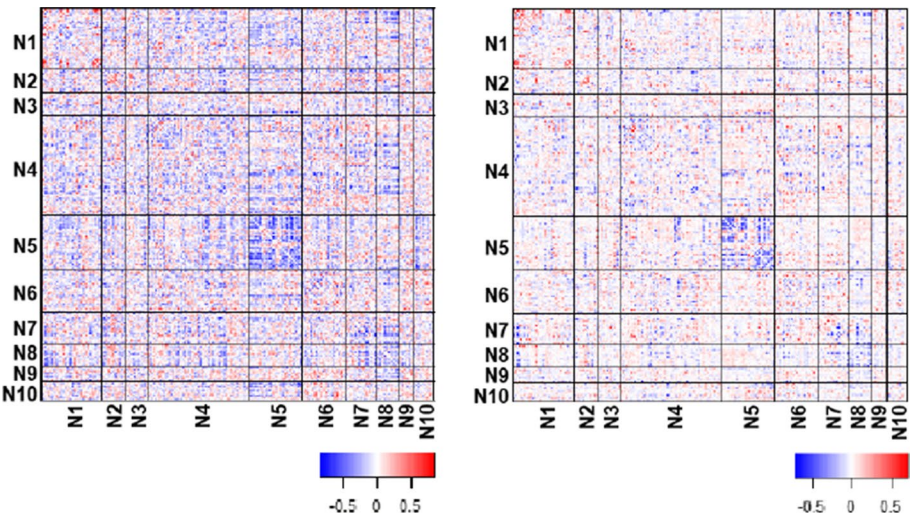

Fig. 1 Differences in absolute correlations (left panel) and fitted edge probabilities (right panel) between participants with highest and lowest resilience score. N1 through N10 respectively denote the following functional networks: motor, cingulo-opercular, auditory, default mode, visual, fronto-parietal, salience, subcortical, ventral attention and dorsal attention

structures inherent in the networks. Linear models are restrictive in terms of not accommodating non-linear relationships between the outcome and the network, and they also do not account for unknown interactions between the network and other supplementary covariates, and even between different network features. The presence of such non-linear associations between the clinical outcome and some edge strengths are evident in our motivating mental health neuroimaging applications (see Fig. 2). One possible avenue for bypassing these constraints involve deep learning approaches such as convolutional neural networks (CNN) as in Meng and Xiang (2018), which extract deep non-linear embeddings from the network that is subsequently used to predict the outcome. However, deep learning methods are data-hungry and require enormous sample sizes that may not always be available especially for neuroimaging problems of interest in this article. Moreover, deep learning techniques typically do not provide the ability for inference and feature selection, which is desirable in our applications of interest. Hence, novel flexible and interpretable non-linear approaches with the added capability to perform inference are required.

The second important limitation with current classification and prediction methods that use network-valued covariates results from the fact the networks are non-Euclidean objects that have an inherently complex structure. This feature results in sub-optimal performance for standard prediction and classification techniques that typically treat network edges/features as interchangeable. In addition, using the entire edge set for the network results in an overly inflated model where the number of candidate parameters increases quadratically with the number of nodes. One potential remedy is to extract lower dimensional

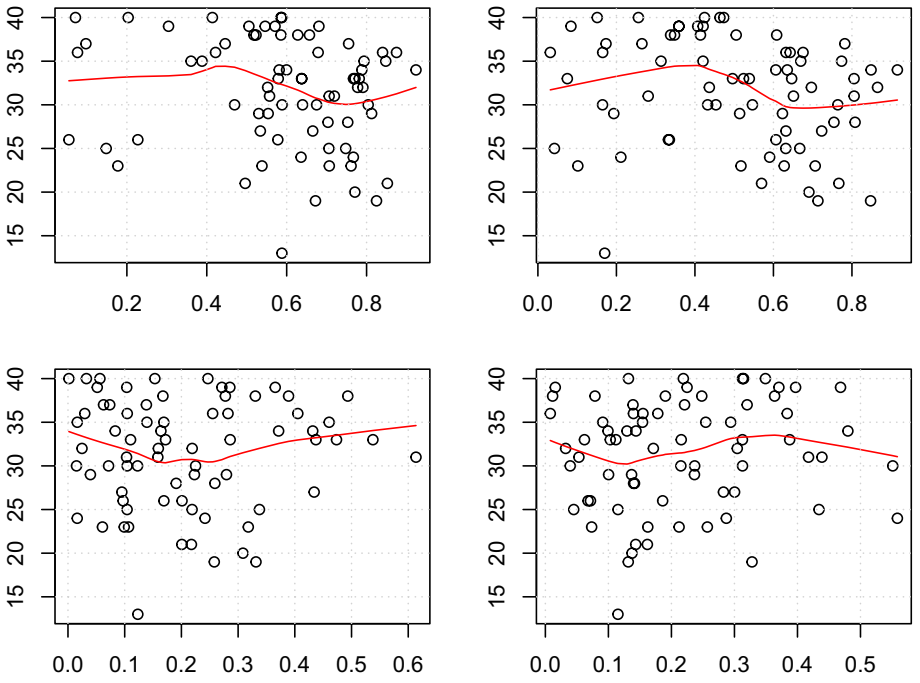

Fig. 2 Scatter plots from the grady trauma project data. The vertical axis represents the resilience score. The horizontal axis represents edge probabilities from four selected edges. The red lines are obtained via Locally Weighted Scatterplot Smoothing (LOWESS) (Cleveland 1979)

embeddings from the high-dimensional networks with minimal loss of information (see Cui et al., 2018 for a survey of graph embedding methods), and subsequently use these embeddings for prediction. However, these graph embedding methods typically result in loss of information and it is not immediately clear how well these embedding approaches are suited for classification and/or prediction purposes and their effect on uncertainty quantification. Moreover, projection of the network onto a lower dimensional manifold leads to loss of interpretability that may result in difficulties in terms of network feature selection. To our knowledge, there is very limited literature on Bayesian classification and prediction methods using lower dimensional projections of high-dimensional networks, which also accommodates network feature selection. Such an approach clearly warrants further consideration and investigation, and is the main focus of this article.

Our goal in this article is to propose a novel two-stage Bayesian methodology to address the aforementioned gaps in network analysis literature. In particular, we develop a flexible Gaussian process modeling framework for scalar continuous outcomes based on lower dimensional manifold projections of high-dimensional networks. Our two stage method uses a manifold learning approach in the first stage to project the high-dimensional network onto lower dimensional latent scale space (Hoff 2005) in order to address the curse of dimensionality. The first stage projection approach results in interpretability at the node level and is remarkable in terms of preserving the original network characteristics given small to moderate number of channels. In the second stage, the projected latent scale features are used to predict continuous scalar outcomes via a flexible Gaussian process regression, which naturally accommodates non-linear relationships and incorporates unknown interactions. In addition to being able to provide a systematic way for extracting the dimension of the underlying manifold space, the proposed approach is able to perform feature selection at the network node level via spike and slab priors on the lengthscale parameters of the Gaussian process, which provides significant advantages over competing network embedding approaches that typically do not preserve interpretability and hence are not equipped to explicitly perform network feature selection.

We denote our approach as latent scale Gaussian process regression (ls-GPR) and develop a fast computational method for implementation. An efficient Expectation-Maximization (EM) algorithm for estimating the latent scales from the given network data is proposed for the first stage model, by leveraging the Polya-Gamma data augmentation scheme in Polson et al. (2013). The Gaussian process regression in the second step is implemented via Markov chain Monte Carlo (MCMC) tools. We note that while it is tempting to propose a joint model that simultaneously updates the latent scales and the regression parameters, such a model runs into computational challenges when updating the latent scales due to the difficulties in deriving closed form posteriors, which is a well-known problem in literature (Yang and Dunson 2016). One can possibly discretize the latent scales to facilitate computational updates, or alternatively use Metropolis-Hastings based strategies or their more efficient variants (Robert 2015) for such a joint analysis. However, both strategies have drawbacks in high dimensions: the former may lead to shrinkage of prior support, while the latter may result in inefficient mixing. Hence we adopt a more practical two-stage implementation that alleviates these computational challenges and is scalable to high dimensional networks. We perform extensive simulations under a variety of network structures and compared the performance to existing linear and non-linear approaches. Our results illustrate the clear advantages of the proposed approach in terms of higher prediction accuracy, superior predictive uncertainty quantification in terms of coverage intervals for test samples that also often have much lower interval widths, and power to detect truly significant predictors while controlling the false positives. Our neuroimaging application

involving PTSD resilience modeling based on brain networks reveals significantly higher prediction accuracy and better predictive uncertainty quantification under the proposed method, and identifies important brain regions that are associated with PTSD.

We would note that this article makes several important contributions. To our knowledge, the proposed approach is one of the first to develop a Bayesian non-parametric regression approach based on network-valued covariates with the added capability of node level feature selection. The superior numerical performance of the approach results from successfully resolving the question of whether accurate lower dimensional manifold representations of high-dimensional networks can be used for improved prediction in conjunction with flexible non-linear regression models (in our case, Gaussian process regression), and whether they can be used to infer important network features. The latent scale representation provides a desirable balance between two extreme scenarios involving an edge-level analysis that is not appealing due to the reasons mentioned previously, and alternate dimension reduction techniques that do not preserve interpretability and hence precludes network feature selection. To our knowledge, given that there is a limited literature on variable selection in Gaussian process latent variable models (GP-LVMs) and limited or no variable selection methods for Bayesian non-parametric regression methods using network-valued covariates, our contribution is of independent interest.

2 Related literature

A common strategy for regression based on network valued covariates is to use summary network measures as explanatory variables (see, for example, Bullmore and Sporns (2009) and references therein). However, the success of such an approach depends heavily on the choice of the network metrics. Moreover, these approaches have reduced exploratory value and potentially sub-optimal performance due to decreased resolution of the summary statistics. Another alternative is to include all the edges in the network as a vectorized predictor, and use these high dimensional features for modeling the clinical phenotype (Craddock et al., 2009). Although penalized regression approaches (Tibshirani 1996) and Bayesian shrinkage (Chang et al., 2018) may be used to model the regression coefficients in these high-dimensional applications, these approaches often treat the edges as interchangeable and fail to respect the inherent network structure that may show properties such as small-worldedness (Bassett and Bullmore 2006) or other patterns of organization. Disregarding these inherent structures and the associated correlations may lead to sub-optimal performance. Recent developments such as Guha and Rodriguez (2020) address this issue by using a tensor-based representation for the regression parameters corresponding to the edges in the network, while Relión et al. (2019) propose a linear classification model that encourages sparsity in the number of nodes and edges of the coefficient matrix. However, these linear approaches still require estimating as many regression coefficients as there are edges, and hence are challenging to implement for high dimensional networks including hundreds of regions containing tens of thousands of edges. Going beyond linear approaches, a contemporary work by Weaver et al. (2021) presented a regularized single index modeling approach that links the outcome to a linear combination of covariates using an unspecified smooth function modeled via splines. Although more flexible than linear regression approaches, single index models do not mitigate the curse of dimensionality presented by a massive number of edges corresponding to our high-dimensional networks of interest (involving 264 nodes and 34716 edges), which may potentially lead to overfitting

and sub-optimal performance. An additional limitation for the approach by Weaver et al. (2021) is that it only reports point estimates without presenting uncertainty quantification and lacks inferential capabilities for feature selection that is often critical in neuroimaging studies.

A possible alternative approach to tackle the curse of dimensionality in regression problems involving high-dimensional networks involves manifold learning, where the network is first projected onto a lower dimensional manifold, which is subsequently embedded within a flexible regression framework. Frequentist examples of reducing the dimension of feature space include principal component analysis and more elaborate methods that accommodate non-linear subspaces, such as isomap (Tenenbaum et al., 2000) and Laplacian eigenmaps (Belkin and Niyogi (2003)). Bayesian manifold approaches characterizing predictive uncertainty have also been developed. Page et al. (2013) proposed a Bayesian nonparametric model for learning of an affine subspace in classification problems, and Kundu and Dunson (2014) proposed a Gaussian process latent variable model to accommodate non-linear subspaces for prediction with scalar covariates. However, there may be a heavy computational price to extricate the number and distribution of the latent variables, and then simultaneously learning the mapping functions while keeping identifiability restrictions. These factors typically restrict such manifold learning based approaches involving Gaussian process models to a small to moderate number of features, although some limited number of scalable approaches are available that avoid having to learn the mapping to the lower-dimensional subspace (Yang and Dunson 2016). The lack of scalabilty under Gaussian process regression to more than a few hundred or thousand predictors is not surprising, given that it is not trivial to sample a large number of lengthscale parameters under an anisotropic Gaussian process model via Metropolis-Hastings updates in such high-dimensions.

Another important limitation of existing manifold learning approaches is that the features in the lower dimensional manifold are typically not interpretable in terms of the original network features. This may be restrictive in neuroimaging applications where it is often important to identify important brain regions that are associated with mental illnesses via localized changes in network configurations. Similar limitations apply for non-linear deep learning approaches for constructing prediction models using network-valued covariates, which have the added complication of requiring large sample sizes to train the deep layers. These efforts are further complicated by the fact that the generalization of deep learning tools such as CNN to irregular or non-Euclidean network data is non-trivial. In fact, there is very limited literature on this topic (Meng and Xiang 2018), and advanced tools such as graph neural networks (Wu et al., 2020) require more testing and validation for brain network applications of interest. Even then, such existing deep learning approaches are not applicable to our problems of interest that focus on interpretable Bayesian methods incorporating network feature selection through inference, and quantification of uncertainty.

3 Proposed methods

Suppose we have data on n participants. For the ith participant ($i = 1, \ldots, n$), the data includes a continuous scalar variable $y_i \in \Re$ representing the clinical outcome of interest, an undirected brain network having p nodes represented as a symmetric binary matrix $G_i(p \times p)$, with $g_i(k, l) = 1/0$ depending on whether the edge $(k, l), k \neq l$, is present or absent in the network, and supplemental covariates $\mathbf{z}_i$ representing environmental

exposures and demographic factors. The binary network G can be based on structural or functional connectivity, while the number of regions depends on the chosen atlas ($p = 264$ under the Power atlas for our applications). We denote the vector of elements in the upper triangle excluding the diagonals for G_i, or edge set, as $\mathbf{e}_i$ of length $p(p-1)/2$. The diagonal elements are excluded since they do not represent connections between distinct nodes and are irrelevant to our problem of interest. The method is described in detail below, and Fig. 3 provides a diagrammatic illustration of our two-stage model.

3.1 Model formulation

3.1.1 First stage: latent scale representation of brain networks

Our goal is to have a parsimonious probability model for the binary networks represented by the edge sets $\mathbf{e}_i, i = 1, \ldots, n$. Clearly there are $2^{p(p-1)/2}$ possible models for the graph space $\mathscr{G}$ that grows exponentially with the number of nodes. In order to tackle this curse of dimensionality, we project the network into a lower dimensional space via a meaningful mapping that avoids restrictive assumptions and fits the data reasonably well. Motivated by the above considerations, we represent the edge probabilities in terms of node level latent scales. In particular, we fit the following latent scale model separately for each sample

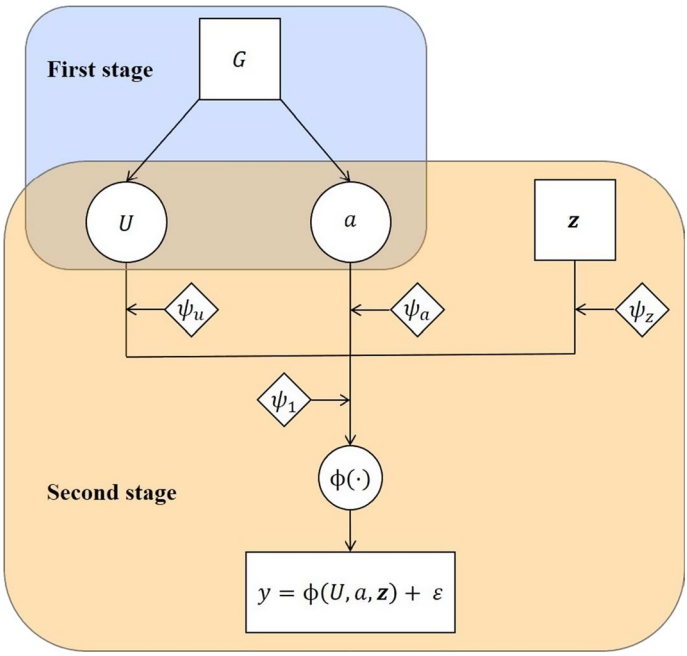

Fig. 3 Schematic diagram of the two-stage model. G represents the given network that is projected onto a lower dimensional manifold involving latent scales U and intercept a in the first stage. These parameters are then combined with observed environmental exposures z to model the outcome via the unknown mean function $\phi(\cdot)$ that is modeled via a Gaussian process regression, in the second stage

$$P(\mathbf{e}_i) = \prod_{k<l,k,l=1}^{p} \pi_{i,kl}^{e_{i,kl}}(1-\pi_{i,kl})^{1-e_{i,kl}}, \log\left(\frac{\pi_{i,kl}}{1-\pi_{i,kl}}\right) = a_i + \mathbf{u}_{ik}^T \Lambda_i \mathbf{u}_{il},$$
$$a_i \sim N(0, \sigma_a^2), u_{ik,r} \sim \mathrm{N}(0, \sigma_u^2), k = 1, \ldots, p, i = 1, \ldots, n, \tag{1}$$

where $\mathbf{u}_{ik} = (u_{ik,1}, \cdots, u_{ik,d})^T$ is the vector of latent scale having d channels for node k, a_i denotes the subject-specific intercept common across edges, and Λ_i represents the $d \times d$ diagonal weight matrix with elements $(\lambda_{i1}, \ldots, \lambda_{id})$ that controls the contribution of the latent scales to the inner product in (1) corresponding to the ith participant. The intercept term controls the overall density of the network and is learnt by pooling data across all edges. The latent scale $\mathbf{u}_l$ captures the importance of node l in the network. If both nodes l and k have activations in the same directions, captured via $u_{k,r}$ and $u_{l,r}$ having the same signs, then they can be construed as functionally connected. In order to preserve identifiability with respect to rotation, we fix the first element of all latent scales (0.5 in our implementations), and the first diagonal element in weight matrix Λ_i is also fixed to be one. The remaining diagonal elements in the weight matrix are either pre-specified, or assigned a Bernoulli prior $\lambda_{ir} \sim Ber(\pi), r = 2, \cdots, d$, that allows one to adaptively select the important channels specific to the network for each individual. The unknown prior inclusion probability π, which controls the number of channels that are expected to be included when fitting the network, is estimated adaptively under a Beta(a_π,b_π) prior.

Model (1) results in a dramatic reduction in the number of parameters from $p(p-1)/2$ to the order of $p \times d + 1$. In particular, the latent scales $U_i = (\mathbf{u}_{i1}, \ldots, \mathbf{u}_{ip})$, for participant i, have dimension $d \times p$ where p is the number of brain regions and d is the intrinsic dimension of the latent scale (or channels) that needs to be determined. In general, finding the intrinsic dimension of the manifold is a difficult problem (Yang and Dunson 2016). However, we are able to systematically choose the manifold dimensions under our approach. The number of channels d can be chosen from a range of possible values, as one that results in the smallest BIC score in the subsequent second stage regression model. Such a choice of d represents a common latent dimension across all samples that results in the best out of sample prediction performance, although the relative importance of individuals channels when fitting the network in the first stage model is expected to vary across samples.

Model (1) is inspired by the latent space modeling of a single network (Hoff 2005) that has been shown to provide a more general characterization of interconnection structures and network properties than stochastic block model (Nowicki and Snijders 2001) and latent distance model (Hoff et al., 2002). A similar model for undirected binary networks appeared in Durante et al. (2017), who used a mixture of latent scale probabilities to model a population of networks. For our neuroimaging applications of interest, it may be too simplistic to assume that groups of participants in the sample share the same latent scales exactly, due to the well-known inherent heterogeneity in PTSD (Lanius et al., 2006). Moreover, our primary objective is to predict the clinical outcome based on the network, which requires a distinct lower dimensional representation for the network corresponding to each participant. Hence for our application, having individual network specific parameters in (1) seems appropriate. In addition, instead of multiplicative gamma priors on the diagonal elements of Λ in Durante et al. (2017) that are specified for adaptive estimation of the weight matrix, our prior construction uses Bernoulli priors on the diagonal elements of Λ that is computationally more straightforward, and allows one to actually select the important channels for fitting the individual level networks.

3.1.2 Second stage: latent scale Gaussian process regression

Once the high-dimensional network has been projected onto the lower dimensional latent scale space in the first stage, we embed these latent scales (along with supplementary covariates $\mathbf{z}$) into a Gaussian process regression framework for predicting the clinical outcome as:

$$y_i = \phi(\Lambda_i^{1/2} U_i, a_i, \mathbf{z}_i) + \epsilon_i, \epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \tau^{-1}), i = 1, \cdots, n. \tag{2}$$

where ϵ_i denotes the residual error normally distributed with precision $\tau \sim \pi(\tau)$, $\phi(\cdot)$ denotes the unknown mean that is a function of the brain network via the estimated weighted latent scales $\Lambda_i^{1/2} U_i$ and intercept a_i derived via model (1), as well as supplementary demographics and environmental exposures $\mathbf{z}_i$. We note that the parameters $(\Lambda_i^{1/2} U_i, a_i)$ capture the entirety of information about the network for the i-th sample, and hence are collectively included in the Gaussian process regression. The function $\phi(\cdot)$ is assumed to have a Gaussian process prior with mean $\mathbf{0}$ and covariance kernel K that has the following squared-exponential structure:

$$K(i, i') = \psi_1 exp\Big\{ -\psi_u \big|\big|\Lambda_i^{1/2} U_i - \Lambda_{i'}^{1/2} U_{i'}\big|\big|_F^2 - \psi_a (a_i - a_{i'})^2 - \psi_z \big|\big|\mathbf{z}_i - \mathbf{z}_{i'}\big|\big|_2^2 \Big\}, i \neq i' \tag{3}$$

where $||\cdot||_F$ and $||\cdot||_2$ denote the Frobenius and L_2 norms respectively, ψ_1 denotes the scale parameter controlling the variance of the mean function, and ψ_u, ψ_a and ψ_z respectively denote the distinct lengthscale parameters corresponding to the latent scales, the intercept term and the supplementary covariates that control the smoothness of the mean function under the Gaussian process. The Gaussian process prior for the mean function allows flexible non-linear relationships between the outcome and the covariates, and can also accommodate unknown interactions between the network and supplemental covariates that is crucial in order to achieve good prediction accuracy. We note that the Frobenius norm in the first term in (3) represents the distance between matrices, whereas the L_2 norm represents the distance between vectors.

3.1.3 Node selection with respect to outcome

Given the node-specific latent scales representation of the network, we are also interested in investigating which nodes in the network contribute to significant differences with respect to the outcome of interest. To achieve this, we extend the proposed model in (2) by using a modified kernel function as

$$K(i, i') = \psi_1 exp\Big\{ -\psi_u \Big[\sum_{j=1}^{p} \beta_j \big|\big|\Lambda_i^{1/2} \mathbf{u}_{ij} - \Lambda_{i'}^{1/2} \mathbf{u}_{i'j}\big|\big|_2^2 \Big] - \psi_a (a_i - a_{i'})^2 - \psi_z \big|\big|\mathbf{z}_i - \mathbf{z}_{i'}\big|\big|_2^2 \Big\}, \tag{4}$$

for $i \neq i'$ where β_j represents the overall contribution of node j in the network towards modeling the outcome. If nodes k and l are both significant related to the outcome (i.e. β_k and β_l are non-zero), then it automatically implies that edge (k, l) is significant. However even then, it is possible for subsets of edges associated with significant nodes $\{k : \beta_k \neq 0\}$ to be unrelated to the outcome, since β_k represents the node level contributions corresponding to the k-th node and is not equipped to directly identify edge-level associations.

Similar to Savitsky et al. (2011), we assume $\beta_j = -\log(\rho_j)$ where $\rho_j \in [0,1]$ for $j = 1, \cdots, p$. Further we assume a spike-and-slab prior on ρ_j's as

$$\pi(\rho_j|\gamma_j) = \gamma_j \mathrm{I}\{0 < \rho_j \leq 1\} + (1-\gamma_j)\delta_1(\rho_j), \gamma_j \sim Ber(\pi^*), \pi^* \sim Beta(a_{\pi^*}, b_{\pi^*}), \quad (5)$$

where $\delta_1(\cdot)$ denotes a point mass at 1, in which case the corresponding β parameter would have a point mass at 0. Equation (5) specifies that with probability $1-\pi^*$ (that is unknown), the j-th network node will have no effect on the regression model (i.e. $\beta_j = 0$), while it will have a non-negligible influence with probability π^*. The unknown prior inclusion probability is estimated under a Beta hyperprior. One can perform feature selection at the node level by including those nodes whose posterior inclusion probabilities lie above a certain threshold, where the threshold can be chosen as 0.5, or can be determined adaptively using a post-hoc approach that controls for false discovery rates as in Kundu et al. (2019). The spike and slab specification is designed to result in dimension reduction via a node selection procedure that completely eliminates the contribution of all network edges related to nodes that are not significantly associated with the outcome. The feature selection at the node level (rather than the edge level), is more suitable in high-dimensional network studies (such as in our neuroimaging applications), and is motivated by the fact that node level summaries are commonly used to study brain network properties (Higgins et al., 2019), which justifies this approach. Hence, the proposed approach works best when the association between the network and the outcome can be primarily characterized via node level network measures.

3.2 Computation framework

The estimation of the first stage model is implemented through an EM algorithm with data augmentation utilizing Theorem 1 in Polson et al. (2013). The EM algorithm is more computationally efficient compared to the previous Markov chain Monte Carlo (MCMC) implementation (Hoff 2005), and leads to significant speed-ups in applications involving high-dimensional networks. The details of the EM algorithm can be found below. For completeness, we have also included details of the MCMC algorithm for the first stage model in the Appendix, but the MCMC based results for the first stage model are not included in the main article. Conditional on the estimated latent scales in the first stage, we use a MCMC sampling scheme to estimate the parameters for the second stage regression model. For the MCMC sampling scheme, the scale parameters $\tau \sim Ga(a_\tau, b_\tau)$ and $\psi_1 \sim Ga(a_{\psi_1}, b_{\psi_1})$ are assigned conjugate priors and updated under closed form posteriors. The lengthscale parameters ψ_u, ψ_a and ψ_z are updated via Metropolis-Hastings steps. For the posterior computation involving node selection in model (4), the MCMC proceeds as in model (2) with additional steps included to update γ's, ρ's, $\beta's$ and π^*, as described in the sequel.

3.2.1 EM algorithm for the first stage model

Denoting $\{\mathbf{e}\} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, $\{a\} = \{a_1, \ldots, a_n\}$, $\{U\} = \{U_1, \ldots, U_n\}$, we can express the likelihood as follows:

$$\pi\big(\{\mathbf{e}\}\big|\{a\},\{U\},\Lambda\big) = \prod_{i=1}^{n}\prod_{k<l,k,l=1}^{p} \frac{\Big[exp\big(a_i + \mathbf{u}_{ik}^T \Lambda_i \mathbf{u}_{il}\big)\Big]^{e_{i,kl}}}{\Big[1 + exp\big(a_i + \mathbf{u}_{ik}^T \Lambda_i \mathbf{u}_{il}\big)\Big]}, k < l, k, l = 1, \ldots, p.$$

Using Theorem 1 in Polson et al. (2013), we can introduce edge-specific latent Pólya-Gamma (PG(0, 1)) variables $\boldsymbol{\omega}_i = \{\omega_{i,kl} : k < l\}, i = 1, \cdots, n$ and denote $\{\boldsymbol{\omega}\}=\{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n\}$, and write the augmented likelihood as:

$$\pi\big(\{\mathbf{e}\}\big|\{a\}, \{U\}, \Lambda_i, \{\boldsymbol{\omega}\}\big) = \prod_{i=1}^{n} \prod_{k<l,k,l=1}^{p} \frac{1}{2} \exp\left\{(e_{i,kl} - 0.5)\big(a_i + \mathbf{u}_{ik}^T \Lambda_i \mathbf{u}_{il}\big) - 0.5\omega_{i,kl}\big(a_i + \mathbf{u}_{ik}^T \Lambda_i \mathbf{u}_{il}\big)^2\right\}.$$

The posterior distribution is proportional to the product of the likelihood and priors $\pi(\{a\})$, $\pi(\{U\})$, $\pi(\Lambda_i)$, $\pi(\{\boldsymbol{\omega}\})$ as specified earlier. The EM algorithm treats the log-posterior as the objective function and maximizes it to obtain the MAP (maximum a posteriori) estimates for a_i, U_i and Λ_i via the M-step, while treating $\{\boldsymbol{\omega}\}$ as missing variables that are imputed via the E-step. The q-th iteration of the EM algorithm is described below.

E step We calculate the conditional expectation of the Pólya–Gamma variables as

$$\omega_{i,kl}^{(q)} = \frac{1}{2\delta_{i,kl}^{(q-1)}}\left[\frac{e^{\delta_{i,kl}^{(q-1)}} - 1}{e^{\delta_{i,kl}^{(q-1)}} + 1}\right], \delta_{i,kl}^{(q-1)} = a_i^{(q-1)} + \mathbf{u}_{ik}^{(q-1)T} \Lambda_i^{(q-1)} \mathbf{u}_{il}^{(q-1)}, k < l, k, l = 1, \ldots, p.$$

and also for the diagonal elements in the weight matrix as

$$\lambda_{ir}^{(q)} = \frac{\pi^{(q-1)} \mathscr{L}_1}{\pi^{(q-1)} \mathscr{L}_1 + (1 - \pi^{(q-1)}) \mathscr{L}_0}$$

where

$$\mathscr{L}_1 = \prod_{k<l,k,l=1}^{p} \frac{1}{2} \exp\left\{(e_{i,kl} - 0.5)\big(a_i^{(q-1)} + \mathbf{u}_{ik}^{(q-1)T} \Lambda_i^{r1} \mathbf{u}_{il}^{(q-1)}\big) - 0.5\omega_{i,kl}^{(q)}\big(a_i^{(q-1)} + \mathbf{u}_{ik}^{(q-1)T} \Lambda_i^{r1} \mathbf{u}_{il}^{(q-1)}\big)^2\right\}$$

and $\Lambda_i^{r1} = diag(1, \lambda_{i2}^{(q)}, \cdots, \lambda_{i(r-1)}^{(q)}, 1, \lambda_{i(r+1)}^{(q-1)}, \cdots, \lambda_{id}^{(q-1)})$; also on the other hand,

$$\mathscr{L}_0 = \prod_{k<l,k,l=1}^{p} \frac{1}{2} \exp\left\{(e_{i,kl} - 0.5)\big(a_i^{(q-1)} + \mathbf{u}_{ik}^{(q-1)T} \Lambda_i^{r0} \mathbf{u}_{il}^{(q-1)}\big) - 0.5\omega_{i,kl}^{(q)}\big(a_i^{(q-1)} + \mathbf{u}_{ik}^{(q-1)T} \Lambda_i^{r0} \mathbf{u}_{il}^{(q-1)}\big)^2\right\}$$

and $\Lambda_i^{r0} = diag(1, \lambda_{i2}^{(q)}, \cdots, \lambda_{i(r-1)}^{(q)}, 0, \lambda_{i(r+1)}^{(q-1)}, \cdots, \lambda_{id}^{(q-1)})$. The diagonal elements are updated in sequence from index 2 to d.

M step We plug in $\{\omega_{i,kl}^{(q)} : k < l\}$ and $\{\lambda_{ir}^{(q)} : r = 2, \cdots, d\}$ to obtain the values for the remaining model parameters as those that maximize the augmented log-posterior. We first find the estimate for a_i $(i = 1, \ldots, n)$ as:

$$a_i^{(q)} = \left(\sum_{k<l,k,l=1}^{p} \left[e_{i,kl} - 0.5 - \omega_{i,kl}^{(q)} \mathbf{u}_{ik}^{(q-1)T} \Lambda_i^{(q)} \mathbf{u}_{il}^{(q-1)}\right]\right)\left(\sigma_a^{-2} + \sum_{k<l,k,l=1}^{p} \omega_{i,kl}^{(q)}\right)^{-1}.$$

As noted earlier, the first element of every latent scale is fixed at a certain pre-specified value b and does not need to be updated. Thus we denote latent scale for the kth node omitting the first element as $\mathbf{u}_{ik(-1)}$. Correspondingly, we fix the first diagonal element of Λ_i at 1 and denote the diagonal weight matrix omitting the first row and column of Λ_i as Λ_{i0}. The latent scales are updated iteratively for $k = 1, \cdots, p$ as $\mathbf{u}_{ik(-1)}^{(q)} = A_{ik}^{-1} B_{ik}$ where

$$A_{ik} = \sum_{1\le j<k} \Big[\omega_{i,jk}^{(q)} \Lambda_{i0}^{(q)} \mathbf{u}_{ij(-1)}^{(q)} \mathbf{u}_{ij(-1)}^{(q)T} \Lambda_{i0}^{(q)} + \sigma_u^{-2} \mathrm{I}_{(d-1)}\Big] + \sum_{k<j\le p} \Big[\omega_{i,jk}^{(q)} \Lambda_{i0}^{(q)} \mathbf{u}_{ij(-1)}^{(q-1)} \mathbf{u}_{ij(-1)}^{(q-1)T} \Lambda_{i0}^{(q)} + \quad \sigma_u^{-2} \mathrm{I}_{(d-1)}\Big]$$

$$B_{ik} = \sum_{1\le j<k} \Big[e_{i,jk} - 0.5 - (a_i^{(q)} + b^2)\omega_{i,jk}^{(q)}\Big] \Lambda_{i0}^{(q)} \mathbf{u}_{ij(-1)}^{(q)}, + \sum_{k<j\le p} \Big[e_{i,jk} - 0.5 - (a_i^{(q)} + b^2)\omega_{i,jk}^{(q)}\Big] \Lambda_{i0}^{(q)} \mathbf{u}_{ij(-1)}^{(q-1)}$$

Finally, the prior inclusion probability π corresponding to the channels is updated as $\pi^{(q)} = (a_\pi - 1 + \sum_{r=2}^{d} \lambda_{ir}^{(q)})/(a_\pi + b_\pi + d - 3)$.

3.2.2 Gibbs sampler for second stage parameters

We denote the Gaussian process atoms as $\boldsymbol{\phi} = \Big(\phi(\hat{a}_1, \hat{\Lambda}_1^{1/2}\hat{U}_1, z_1), \cdots, \phi(\hat{a}_n, \hat{\Lambda}_n^{1/2}\hat{U}_n, z_n)\Big)^T$. The algorithm iterates between the following steps:

1. Update the noise precision τ from Gamma distribution with parameters $(a_\tau + 0.5n)$ and $(b_\tau + 0.5\big|\big|\boldsymbol{y} - \boldsymbol{\phi}\big|\big|_2^2)$.
2. Update the global scale parameter ψ_1 from inverse Gamma distribution with parameters $(a_{\psi_1} + 0.5n)$ and $(b_{\psi_1} + 0.5\boldsymbol{\phi}^T E_0^{-1}\boldsymbol{\phi})$ where $E_0 = \exp\Big(-\psi_u E_u - \psi_a E_a - \psi_z E_z\Big)$. Here we define $E_u(i,i') = \big|\big|\hat{\Lambda}_i^{1/2}\hat{U}_i - \hat{\Lambda}_{i'}^{1/2}\hat{U}_{i'}\big|\big|_F^2, E_a(i,i') = (\hat{a}_i - \hat{a}_{i'})^2, E_z(i,i') = \big|\big|z_i - z_{i'}\big|\big|_2^2, i,i' = 1,\cdots,n$.
3. Draw a candidate ψ_u^* where $log\psi_u^* \sim \mathrm{N}(log\psi_u, 0.01^2)$. Accept the candidate with probability

$$min\left(1, \frac{\big|\psi_1 E_0^* + \tau^{-1}\mathrm{I}_n\big|^{-0.5} \exp\{-0.5\boldsymbol{y}^T(\psi_1 E_0^* + \tau^{-1}\mathrm{I}_n)^{-1}\boldsymbol{y}\}}{\big|\psi 1 E_0 + \tau^{-1}\mathrm{I}_n\big|^{-0.5} \exp\{-0.5\boldsymbol{y}^T(\psi 1 E_0 + \tau^{-1}\mathrm{I}_n)^{-1}\boldsymbol{y}\}}\right)$$

   where $E_0^* = \exp\Big(-\psi_u^* E_u - \psi_a E_a - \psi_z E_z\Big)$. Same procedure for updating ψ_a and ψ_z.
4. Update the Gaussian process atoms $\boldsymbol{\phi}$ from multivariate normal distribution with mean $\Big[\tau^{-1}\psi_1^{-1}E_0^{-1} + \mathrm{I}_n\Big]^{-1}\boldsymbol{y}$ and covariance $\Big[\psi_1^{-1}E_0^{-1} + \tau\mathrm{I}_n\Big]^{-1}$.

   For our additional node selection analysis using modified kernel function, we need two more steps in the Gibbs sampler to update γ_j, ρ_j and β_j for $j = 1,\cdots,p$ in sequence, as well as update the hyperparameter π^* as illustrated below:
5. Update γ_j, ρ_j and β_j for $j = 1,\cdots,p$ in sequence. While keeping parameters for $k \neq j$ fixed, we have two moves to make.
   - Between-model move: if currently $\gamma_j = 1$, we propose to have $\gamma_j^{'} = 0$, $\rho_j^{'} = 1$ and $\beta_j^{'} = 0$; otherwise if currently $\gamma_j = 0$, we propose to have $\gamma_j^{'} = 1$, draw $\rho_j^{'}$ from $Unif(0,1)$ and let $\beta_j^{'} = -\log(\rho_j^{'})$. We accept the proposal with probability

$$\min\left\{1, \frac{\pi(\gamma_j^{'})\big|\psi_1 E_j^{'} + \tau^{-1}\mathrm{I}_n\big|^{-0.5} \exp\{-0.5\boldsymbol{y}^T(\psi_1 E_j^{'} + \tau^{-1}\mathrm{I}_n)^{-1}\boldsymbol{y}\}}{\pi(\gamma_j)\big|\psi_1 E_j + \tau^{-1}\mathrm{I}_n\big|^{-0.5} \exp\{-0.5\boldsymbol{y}^T(\psi_1 E_j + \tau^{-1}\mathrm{I}_n)^{-1}\boldsymbol{y}\}}\right\}$$

where $E_j = \exp\Big(- \psi_u E_u^* - \psi_a E_a - \psi_z E_z \Big)$, $E_u^*(i,i') = \big|\big|\hat{\Lambda}_i^{1/2}\hat{U}_i B - \hat{\Lambda}_{i'}^{1/2}\hat{U}_{i'} B\big|\big|_F^2$ and $B = diag(\sqrt{\beta_1}, \cdots, \sqrt{\beta_p})$ is a diagonal matrix with the j-th diagonal element as $\sqrt{\beta_j}$. On the other hand, we have $E_j^{'} = \exp\Big(- \psi_u E_u^{*'} - \psi_a E_a - \psi_z E_z \Big)$, $E_u^{*'}(i,i') = \big|\big|\hat{\Lambda}_i^{1/2}\hat{U}_i B' - \hat{\Lambda}_{i'}^{1/2}\hat{U}_{i'} B'\big|\big|_F^2$ and $B' = diag(\sqrt{\beta_1}, \cdots, \sqrt{\beta_p})$ where the j-th diagonal element is $\sqrt{\beta_j^{'}}$.

– Within-model move: this move is only triggered when we sample $\gamma_j^{'} = 1$. Then we further draw another $\rho_j^{''}$ from $Unif(0, 1)$ and $\beta_j^{''} = -\log(\rho_j^{''})$. And we accept this proposal with probability

$$\min\left\{1, \frac{\left|\psi_1 E_j^{''} + \tau^{-1}\mathrm{I}_n\right|^{-0.5}\exp\{-0.5\boldsymbol{y}^T(\psi_1 E_j^{''} + \tau^{-1}\mathrm{I}_n)^{-1}\boldsymbol{y}\}}{\left|\psi_1 E_j^{'} + \tau^{-1}\mathrm{I}_n\right|^{-0.5}\exp\{-0.5\boldsymbol{y}^T(\psi_1 E_j^{'} + \tau^{-1}\mathrm{I}_n)^{-1}\boldsymbol{y}\}}\right\}$$

and here

$$E_j^{''} = \exp\Big(- \psi_u E_u^{*''} - \psi_a E_a - \psi_z E_z \Big), E_u^{*''}(i,i') = \big|\big|\hat{\Lambda}_i^{1/2}\hat{U}_i B'' - \hat{\Lambda}_{i'}^{1/2}\hat{U}_{i'} B''\big|\big|_F^2$$

and $B'' = diag(\sqrt{\beta_1}, \cdots, \sqrt{\beta_p})$ where the j-th diagonal element is $\sqrt{\beta_j^{''}}$.

6. Update the prior inclusion probability π^* from Beta distribution with parameters $(a_{\pi^*} + \sum_{j=1}^p \gamma_j)$ and $(b_{\pi^*} + p - \sum_{j=1}^p \gamma_j)$.

3.3 Prediction for testing samples

The prediction for additional subjects involves two steps. First step takes the upper triangular vector of binary network matrices $\boldsymbol{e}_1^*, \cdots, \boldsymbol{e}_m^*$ as input of first stage model and obtain the estimates of intercept and latent scales. These estimates are then used in the second step, together with other supplementary covariates, as input of second stage model and eventually obtain the predicted values of the response variable. From the property of Gaussian process, the response of the additional subjects $\boldsymbol{y}^* = (y_1^*, \cdots, y_m^*)^T$ and the response of the original n subjects $\boldsymbol{y}$ jointly follow a multivariate normal distribution as

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y}^* \end{pmatrix} \sim \mathrm{N}\Bigg(\mathbf{0}_{(n+m)}, \begin{pmatrix} K + \tau^{-1}\mathrm{I}_n & K_*^T \\ K_* & K_{**} + \tau^{-1}\mathrm{I}_n \end{pmatrix}\Bigg),$$

$$K_*(q,i) = \psi_1 exp\Big\{ - \psi_u\big|\big|\Lambda_q^{1/2}U_q^* - \Lambda_i^{1/2}U_i\big|\big|_F^2 - \psi_a(a_q^* - a_i)^2 - \psi_z\big|\big|\boldsymbol{z}_q^* - \boldsymbol{z}_i\big|\big|_2^2\Big\},$$

$$K_{**}(q,q') = \psi_1 exp\Big\{ - \psi_u\big|\big|\Lambda_q^{1/2}U_q^* - \Lambda_{q'}^{1/2}U_{q'}^*\big|\big|_F^2 - \psi_a(a_q^* - a_{q'}^*)^2 - \psi_z\big|\big|\boldsymbol{z}_q^* - \boldsymbol{z}_{q'}^*\big|\big|_2^2\Big\},$$

$$i = 1, \cdots, n, q, q' = 1, \cdots, m$$

When working with the modified kernel (4), we need to replace the $\Lambda^{1/2}U$ in the covariance kernel with $\Lambda^{1/2}UB$ where B is a diagonal matrix of $\sqrt{\beta_1}, \cdots, \sqrt{\beta_p}$. Then the conditional mean can be expressed as:

$$\mathbf{E}(\boldsymbol{y}^*|\boldsymbol{y}) = K_*(K + \tau^{-1}\mathrm{I}_n)^{-1}\boldsymbol{y}$$

The expression above is based on the marginal distribution of the outcome. We can also base the prediction expression on the Gaussian process atoms $\boldsymbol{\phi}$. Then the computation would be

$$\begin{pmatrix} \boldsymbol{\phi} \\ y^* \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}_{(n+m)}, \begin{pmatrix} K & K_*^T \\ K_* & K_{**} + \tau^{-1}\mathrm{I}_n \end{pmatrix}\right), \mathbf{E}(y^*|\boldsymbol{\phi}) = K_* K^{-1}\boldsymbol{\phi}$$

This expression can be used for in-sample prediction with the Gibbs sampler algorithm.

3.4 Hyper-parameter selection

An optimal implementation of our method requires a careful choice of several hyperparameters in order to optimize performance. In the first stage EM algorithm, we set $\sigma_a^2 = \sigma_u^2 = 2$ for the priors on a and elements of U, which worked well across a range of hyper-parameter combinations and several network structures. We also present (in Appendix) the sensitivity analysis for different combinations of (σ_a^2, σ_u^2) in the first stage model, which illustrates the robustness in performance under reasonable choices of these hyper-parameters as long as these values are not extremely large or small. For the weight matrix Λ, we can (i) fix it at identity matrix or (ii) update it under Bernoulli priors and examine the performance under two different hyperparameter choices on the prior inclusion probability π that is chosen to be *Beta*(1, 10) or *Beta*(1, 25) in order to facilitate a sensitivity analysis for channel selection. The prior on the spike and slab probability π^* set at *Beta*(1, 1) that translated to a prior mean inclusion probability of 0.5 and hence it does not favor either inclusion or exclusion of the nodes.

For simulation studies, we performed sensitivity analysis for different choices of d in the working model, whereas for the GTP data analysis, we choose the number of channels using a data adaptive approach via cross-validation, in a manner that minimizes the prediction error under the second stage regression model in a validation sample. For the second stage Gibbs sampler, the prior on ψ_1 is set to inverse gamma (0.1, 80) that results in overall good prediction performance and coverage at the risk of wider predictive intervals in some settings. However in the analysis of GTP data, we also implemented the method under $\psi_1 \sim Inv - Ga(0.1, 10)$ for sensitivity analysis. The prior on τ is pre-specified as Gamma (0.1, 80) that mimics a non-informative variance, which enables good prediction and coverage.

4 Empirical experiments

4.1 Simulation studies

We considered different scenarios corresponding to varying true network structures. Scenarios 1 and 2 correspond to scale-free and small-world networks respectively that are frequently encountered in real-world neuroimaging applications, whereas Scenario 3 uses the networks obtained from resting state fMRI data from our GTP application. When generating the data, a latent scale model (1) was first fit to the true network in order to obtain estimates of latent scales U, the intercept term a and the weight matrix Λ. Subsequently, the response was generated as $y = 0.5a + \exp(\mathbf{b}_1 \Lambda \mathbf{u}_{s_1}) + \sum_{j=2}^{n_{act}} \mathbf{b}_j \Lambda \mathbf{u}_{s_j} + \epsilon_i^*, \epsilon_i^* \sim \mathrm{N}(0, 0.5^2)$, where each vector in the set $\{\mathbf{b}_1, \ldots, \mathbf{b}_{n_{act}}\}$ had length d and was sparse with non-zero

values as 1 or −1. The set $\{s_1, \cdots, s_{n_{act}}\}$ contained the indices of the active nodes, where the number of active nodes were chosen at three levels 10, 30, 50 for the simulations.

The total number of channels d for the true model was chosen as 3, while different values of d were used in the working model for sensitivity analysis. Moreover, we only present the simulation results for Λ fixed to identity matrix in the working model, since the performance with Bernoulli priors on the weight matrix for channel selection resulted in inferior prediction, although we did present a sensitivity analysis with channel selection in the analysis of GTP data reported in the next section. Supplementary covariates were not included for our simulation examples in Scenario 1–3, although we did include additional covariates (such as trauma exposure) in our analysis of GTP data as detailed in the sequel. For Scenarios 1, 2, the training and test sample sizes were 50 each, whereas for Scenario 3, the training and test sample sizes correspond to 41 and 40 respectively corresponding to the GTP dataset.

4.1.1 True network generation

We used 100 nodes for the simulated networks in Scenarios 1 and 2 that is common in brain connectome literature (Lukemire et al., 2020), while we used 264 nodes from the Power atlas for Scenario 3. Hence our studies showcase the generalizability of the proposed method across with varying network sizes. The network generation is described in detail below.

Scenario 1 Here, scale-free networks were used, which were generated by function `sample_pa` in `R` package `igraph` (Csardi and Nepusz 2006). The number of nodes was set to be 100 (p) with size of initial graph varying between 2 and 10, and number of edges to add in each time step varying between 4 and 10.

Scenario 2 Here, the network was generated with small-world structure using function `watts.strogatz.game` in `R` package `igraph`. The size of starting lattice was kept at 1 and the total size of lattice was set to 100 (p). The rewiring probability was kept at 0.5 while the neighborhood size varied between 4 and 12.

Scenario 3 Here, we generated the binary network based on resting state fMRI connectivity using GTP data. A description of the study can be found in the PTSD Data Application section. We calculated the resting state network for each participant corresponding to the Power atlas with 264 nodes using the graphical lasso algorithm (Friedman et al., 2008) under varying sparsity levels corresponding to regularization parameter values $\lambda = 0.05, 0.10$, and 0.15, with a larger λ value corresponding to a sparser network.

We note that the current data generation settings in Scenario 1 yields simulated networks with density lying between 0.07 and 0.19; while for Scenario 2, the network density varies between 0.08 and 0.24. For Scenario 3 involving the GTP data, the network density varies between [0.18, 0.25] corresponding to $\lambda = 0.05$ and it decreases for higher values of λ. These simulated network densities represent acceptable levels of sparsity in brain networks that are encountered in literature (see Hallquist and Hillary (2018)), which typically vary between 5% to 25%.

4.1.2 Competing methods and performance metrics

We denote the proposed latent scale Gaussian process regression approach without node selection as lsGPR, and denote the corresponding version with node selection via spike and slab priors as sparse lsGPR or sp-lsGPR. Our proposed method was compared to the

linear shrinkage methods including lasso (Tibshirani 1996), ridge (Hoerl and Kennard 1970), elastic net (Zou and Hastie 2005), and Bayesian horseshoe prior (Carvalho et al., 2010), which all used the full edge set as the predictors. The models were implemented through `R` packages `glmnet` (Friedman et al., 2010) and `monomvn` (Gramacy 2018). We also compared with non-linear approaches that used GPR on the full edge set (edge-GPR), on the reduced representation from principal component analysis (pca-GPR), and on the reduced representation from Laplacian Eigenmap (Belkin and Niyogi 2003) (mf-GPR). The dimension reduction of the pca-GPR and mf-GPR methods were implemented using `R` function `prcomp` and `R` package `dimRed` (Kraemer et al., 2018) respectively. We used the squared exponential kernel (Rasmussen and Williams 2006) for all GPR approaches, with similar priors on the scale parameter of the kernel for fair comparisons. We evaluated prediction performance in terms of out of sample predictive MSE, as well as coverage and 95% predictive interval widths for the test samples under all Bayesian approaches. Here the coverage was defined as the ratio of test samples where the predictive intervals covered the observed value of these test samples to the total number of test samples. We also evaluated the node selection performance for our method and compared to Bayesian horseshoe prior method in terms of the area under the curve (AUC). For computing the AUC, we varied the threshold for posterior inclusion probabilities in order to detect significant nodes under the proposed sparse lsGPR method. Similarly for the Bayesian horseshoe that is a continuous shrinkage approach, we denoted a node as significant if one or more edges associated with that particular node was significant, where an edge is defined as significant if the estimated regression coefficient had absolute value greater than some threshold. In addition to AUC, the ROC curves themselves are also presented for the simulation examples in the Appendix. The total number of MCMC iterations for the second stage method was 10,000 with a burn in of 5000 under all Bayesian methods.

We note that the above competing methods are designed to cover a gamut of potential competitors that are state-of-the art in existing literature. The linear regression approaches are perhaps most widely used in literature. Moreover, the comparisons with alternate Gaussian process regression approaches with and without dimension reduction are natural in the context of the proposed method that involves dimension reduction via latent scales coupled with a Gaussian process regression framework. Such comparisons are particularly pertinent in the context of previous literature that suggests that Gaussian process regression methods which involve overly large number of features without dimension reduction perform sub-optimally for high-dimensional regression problems (Jiang 2007).

4.1.3 Results

The prediction and coverage results for Scenarios 1, 2 are presented in Fig. 4. The horizontal axis represents the different levels for number of active nodes involved in generating the response. From the boxplots, it is immediately clear that both the lsGPR and the sparse lsGPR have superior predictive performance that is significantly improved compared to existing methods under Scenarios 1–2. Although the lsGPR including all nodes has greater predictive accuracy compared to sparse lsGPR for some cases, it is important to note that the sparse lsGPR method consistently has greater coverage accuracy compared to lsGPR, that points to a better characterization of predictive uncertainty under node selection via the spike and slab prior compared to when all nodes are included. The sparse lsGPR consistently has greater coverage compared to different approaches with varying number of active nodes in almost all settings under Scenarios 1, 2. Moreover, the credible interval

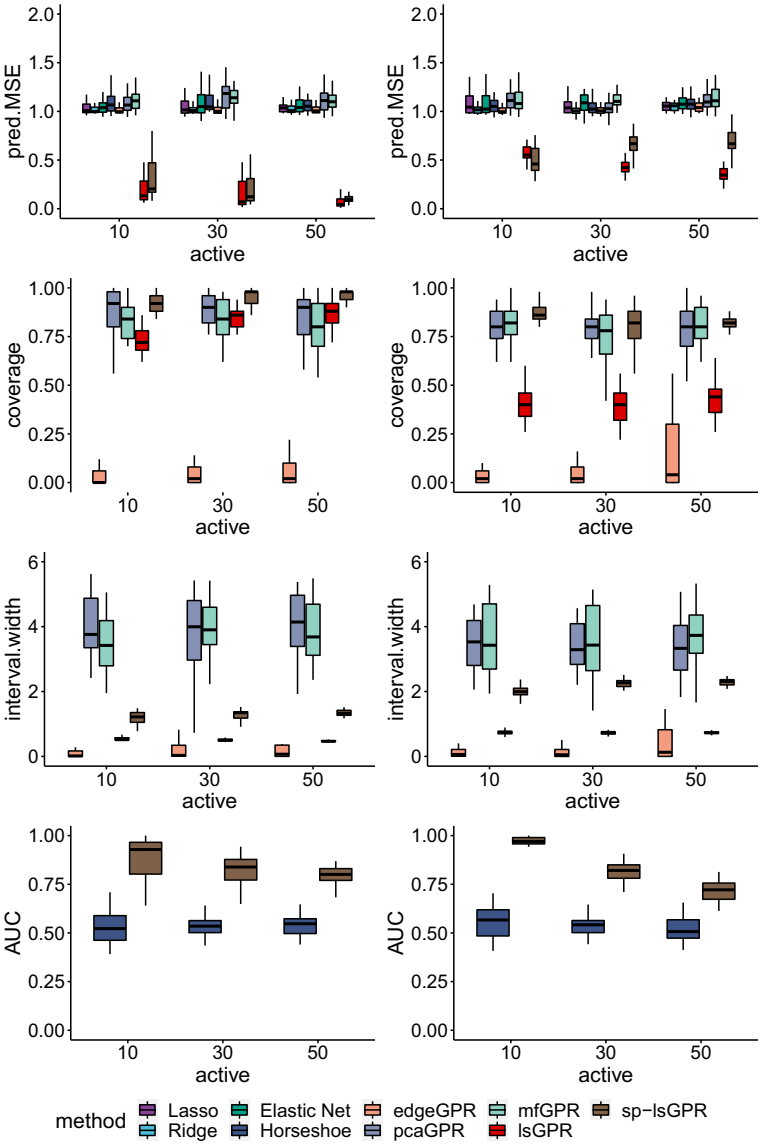

Fig. 4 Boxplots for MSE, coverage, credible interval width and node selection AUC with varying number of activated nodes, under simulation Scenario 1 (left column) and Scenario 2 (right column)

width is reasonable and often lower than competing GP regression approaches that use alternate dimension reduction methods. In contrast, the edge level GPR has extremely narrow predictive intervals that is impractical and results in very poor coverage. Finally, the sparse lsGPR method has a significantly higher AUC compared to the Bayesian horseshoe consistently across different number of active nodes, although the node selection performance becomes less accurate as the number of active nodes increases, which is expected.

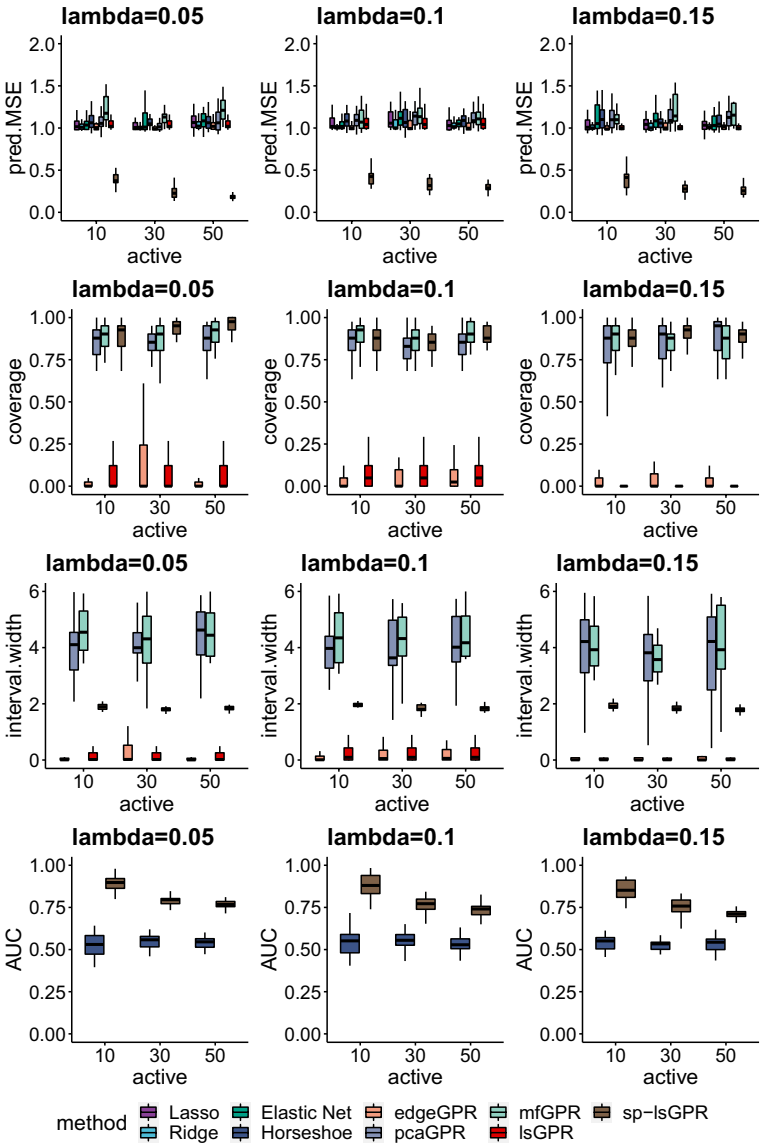

Fig. 5 Boxplots for MSE, coverage, credible interval width , and AUC under Scenario 3 involving GTP networks with varying sparsity levels and different number of activated nodes

Figure 5 illustrates the results of Scenario 3 where the horizontal axis represents the number of active nodes as in Fig. 4, and the three columns correspond to the three levels of the regularization parameter λ of the graphical lasso in obtaining the binary networks, with higher values implying a sparser network. The sparse lsGPR method results in significantly improved prediction accuracy compared to all approaches, including the lsGPR method involving all nodes. Further, the coverage under the sparse lsGPR method is significantly improved compared to all other methods for the majority of settings, although the

pcaGPR and mfGPR methods have slightly improved coverage in a few cases that results from extremely wide predictive intervals under these methods, which may not be desirable. Moreover, both the edgeGPR and the lsGPR involving all nodes have extremely poor predictive coverage that results from narrow predictive intervals, which is not desirable. In addition to superior performance in prediction and coverage, our proposed method also shows a consistent advantage in terms of node selection based on the AUC metric over the Bayesian horseshoe prior method, as in Scenarios 1, 2. Furthermore, the ROC curves presented in Fig. 12 in the Appendix also clearly show the superior sensitivity and specificity for variable selection under the proposed approach across all network sparsity levels in Scenario 3.

In summary, the lsGPR and the sparse lsGPR with node selection has significantly improved prediction performance compared to all approaches for Scenarios 1, 2, but the performance of lsGPR involving all nodes declines under Scenario 3 involving high-dimensional networks derived from real fMRI data. Moreover, the sparse lsGPR has a decided advantage over lsGPR involving all nodes in terms of characterizing predictive uncertainty as reflected by higher coverage under predictive intervals for test samples that have reasonable width. The edgeGPR method suffers from an overwhelmingly large number of edges, and the performance of the lsGPR method involving all nodes deteriorates when the number of nodes is increased in Scenario 3 (compared to Scenarios 1, 2). Competing dimension reduction techniques relying on PCA or other manifold projections have inferior prediction performance, as well as inferior coverage for the majority of settings, in spite of having sufficiently wide predictive intervals, which suggests a poor ability to characterize predictive uncertainty. All linear methods have inferior predictive performance as expected, and the variable selection under Bayesian horseshoe results in inaccurate node selection that is potentially due to an overwhelmingly large number of edges in the high-dimensional networks considered.

We conducted a sensitivity test for the hyper-parameters σ_a^2 and σ_u^2 needed in the first stage EM algorithm. Table 5 in the Appendix reports the performance metrics for different hyper-parameter combinations under simulation scenario 1. From the table we can see that the suggested hyper-parameter values ($\sigma_a^2 = 2, \sigma_u^2 = 2$) enjoy a generally good performance in terms of prediction and variable selection. Also in general, the overall prediction and variable selection are not sensitive to the choice of hyper-parameters in the first stage EM algorithm, as long as some extreme choices are not used, for example, $\sigma_u^2 = 0.2$ or $\sigma_u^2 = 20$.

We also compared the performance of the proposed approach using EM algorithm in the first stage versus using MCMC for fitting the first stage model, under simulation scenario 1. The scatterplots of Fig. 10 in the Appendix show high concordance between prediction and uncertainty quantification results under the latent scales derived from the EM algorithm and the MCMC approach for fitting the first stage model. Since both methods for estimating the first stage model leads to a comparable performance under the second stage regression model (which is our main focus of interest), we recommend using the EM algorithm in the first stage since it is computationally scalable for high-dimensional networks.

Finally, a sensitivity analysis with varying number of channels in Fig. 6 suggests that the performance of the sparse lsGPR method remains consistently better than competing approaches even when the value of d used in the working model is different than the number of true channels used when generating the data. However, the performance of the proposed approach in terms of node selection (AUC) seems to deteriorate significantly when the number of channels used in the working model is less than the true number of channels used for generating the data. Additional numerical examples are included in the Appendix section in Fig. 11, which illustrate the effect of mis-specification of the number of

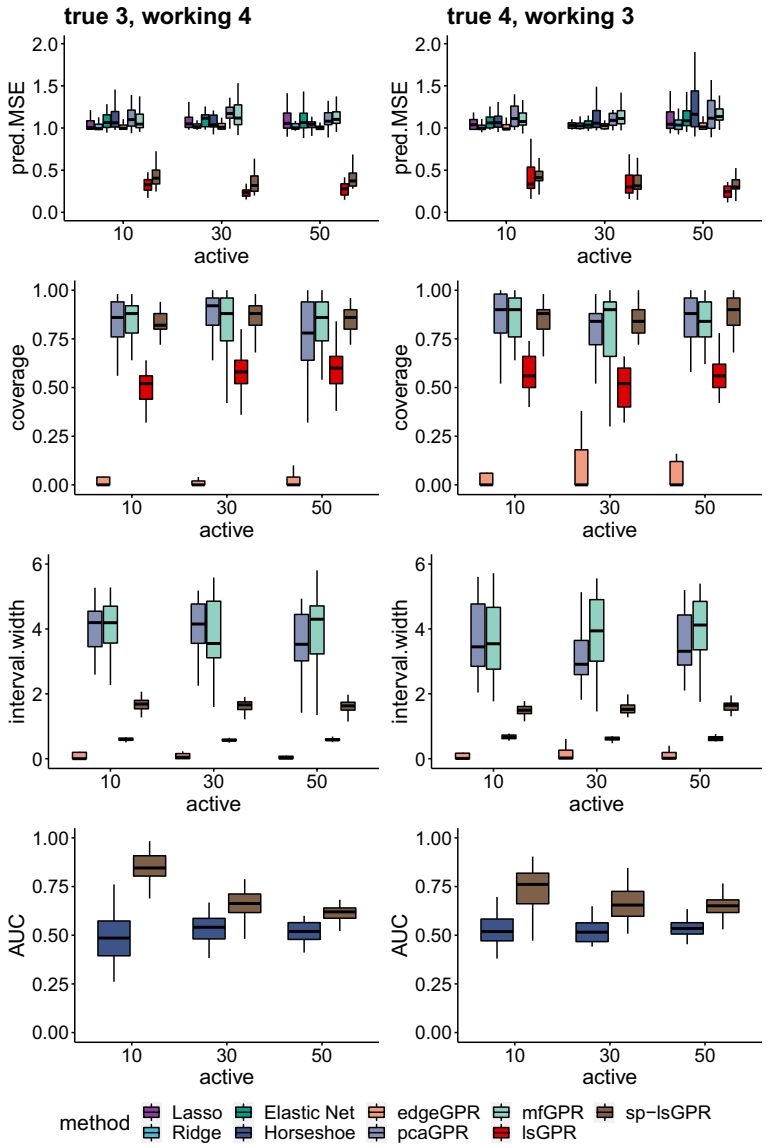

Fig. 6 Sensitivity analysis on the number of channels under Scenario 1. The left column is under situation where the data generation uses 3 channels while estimation uses 4 channels. The right column is under situation where data generation uses 4 channels and estimation with 3 channels

channels. Based on these results, it is clear that the proposed approach performs best when the value of d used in the working model is at least as large as the true number of channels.

Moreover, Table 1 shows the computation time under stage 1 for different methods, corresponding to an implementation on a high performance computing (HPC) environment utilizing Intel Xeon CPU at 2.80GHz. The results in Table 1 are averaged over several simulations and only serve as approximation. The computation time for the Bayesian

Table 1 Computation time summary for different number of channels under Scenario 1 with Stage1 and Stage2 referring to the two stages in our lsGPR method

Method	Lasso	Ridge	Elastic Net	Horseshoe
CompTime	<1s	4s	10s	4 min
Method	*edgeGPR*	*pcaGPR*	*mfGPR*	
CompTime	0.5 min	0.5 min	0.5 min	
Method	*Stage1 (d=3)*	*Stage1 (d=5)*	*Stage1 (d=7)*	*Stage1 (d=10)*
CompTime	7.5 min	9 min	10.5 min	13 min
Method	*Stage1 (d=25)*	*Stage1 (d=50)*	*Stage2*	
CompTime	26 min	58 min	0.5 min	

approaches are reported for 10,000 iterations. We observe that although the computation time increases almost linearly with the number of channels, the overall approach is computationally feasible for up to $d = 25$ or even $d = 50$ channels. However we don't anticipate requiring so many channels in order to obtain a good prediction performance under our approach. In our experience, a small to moderate number of channels in the first stage model is often adequate to deal with prediction involving high dimensional networks. We note that although the additional computational burden for fitting the first stage latent scale model results in an increase in the overall computation time that is higher than competing methods, the proposed approach is still scalable to high-dimensional networks of interest.

4.2 PTSD data application

4.2.1 Grady trauma project and resilience scores

The Grady Trauma Project (GTP) recruited African American females to study the risk factors for PTSD in a low-socioeconomic status (Stevens et al., 2013). The resting state functional magnetic resonance imaging (rs-fMRI) brain scans for each individual were obtained on a 3.0T Siemens Trio with echo-planar imaging (Siemens, Malvern, PA). T1-weighted anatomical scans were gathered with 176 contiguous 1 mm sagittal slices using 3D MP-RAGE sequence (TR/TE/TI=2000/3.02/900 MS, $1mm^3$ voxel size). The functional images were collected in an ascending interleaved sequence with 37 3 mm axial slices and no gap between slices (TR/TE=2000/30 ms, FA=90°, $3mm^3$ voxel size).

We used the preprocessing script released from the 1000 Functional Connectomes Project to preprocess the brain images using standard steps. We first performed skull stripping on the T1-weighted images. Then we removed the first four volumes of the functional scans for signal stabilization, with 146 volumes remaining for the downstream preprocessing. We registered the T1-weighted image and functional images to the MNI standard space with a 6 mm FWHM Gaussian kernel for registration and smoothing. Motion corrections and removing of nuisance signals were also performed on the images. Finally, we put the functional images through band-pass filter to retain frequencies between 0.01 and 0.1 Hz.

After removing data with movement or drowsiness issues, we have 81 participants with available rs-fMRI data. For our analysis, we use the whole brain parcellation presented in Power et al. (2011) for the brain images, involving 264 region of interest (ROIs). These regions are further organized into ten functional modules including motor, cingulo-opercular (CON), auditory, default mode (DMN), visual, fronto-parietal (FPN), salience (SAN),

sub-cortical, ventral attention (VAN) and dorsal attention (DAN) (Cole et al., 2013). These functional modules have been assigned based on resting state fMRI studies (Power et al., 2011), which is well-suited for our data. Using these 264 regions (nodes), a network was computed separately for each individual using the graphical lasso algorithm (Friedman et al., 2008) with regularization parameter λ (higher λ represents networks with greater sparsity) and subsequently these networks were used for analysis. The GTP study has also acquired data on the Connor-Davidson Resilience Scale (Connor and Davidson 2003) for measuring resilience as individual's ability to thrive in the face of adversity. Our goal is to model resilience as a continuous clinical measure of well-being in PTSD using resting state functional connectivity as well as demographic factors such as the participants' age, and environmental exposure including traumatic events inventory (TEI) score (Sprang 1997) and the childhood trauma questionnaire (CTQ) total score (Scher et al., 2001). The resilience score of interest is only available for 73 participants, and hence we focus our analysis on this subset.

For our analysis, we used 10 channels for the latent scales in the first stage model that was chosen via cross-validation and yielded desirable results for prediction and uncertainty quantification under the second stage model. We note that the ability of the latent scales to reconstruct the network can be quantified via the area under the ROC curve (AUC) as illustrated in Fig. 7. In particular, one can first fit the latent scale model to the given network and subsequently reconstruct the edge probabilities under the fitted latent scale model. These edge probabilities can be thresholded under varying cut-offs to inform whether an edge was present or absent under the reconstructed edge set as per the latent scales model. Then, this reconstructed edge set can be compared to the observed network edges and the sensitivity and specificity are computed for a series of thresholds that can then be used for computing the AUC. We note that while an increasing number of channels is expected to lead to greater AUC, it may not necessarily translate to gains in predictive accuracy in the second stage model, and instead may result in an inflated number of parameters without providing any tangible benefits in terms of prediction and feature selection. This is evident from our GTP analysis, where the predictive accuracy and uncertainty quantification with 10 channels was often superior compared to an alternate analysis with 25 channels in the first stage model (results not presented due to space constraints).

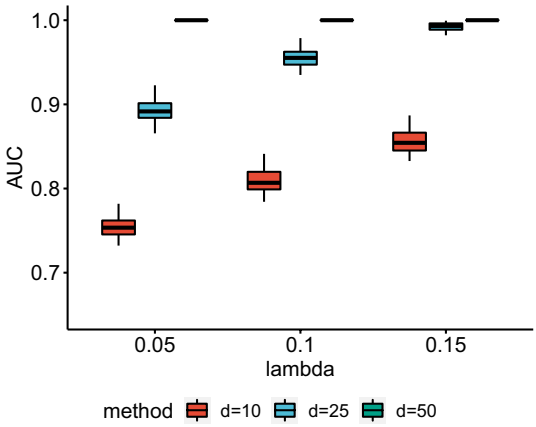

Fig. 7 Network recovery comparisons for the GTP Data under varying number of channels and network densities

Table 2 GTP study analysis results for predictive mean squared error (MSE), coverage and (interval) width over 50 random splits. The sp-lsGPR and lsGPR methods here have fixed the weight matrix to identity matrix in first stage and set prior on ψ_1 at inverse Gamma (0.1, 80)

		sp-lsGPR	lsGPR	mfGPR	pcaGPR	edgeGPR	Horseshoe	Elastic net	Ridge	Lasso
$\lambda = 0.05$	MSE	0.885	0.926	0.973	1.090	1.013	1.092	1.052	1.005	1.046
	Coverage	0.986	0.859	0.766	0.774	0.110				
	Width	5.230	4.001	3.872	4.014	0.323				
$\lambda = 0.1$	MSE	0.864	0.890	0.912	1.067	1.013	1.099	1.107	1.010	1.081
	Coverage	0.965	0.882	0.822	0.691	0.099				
	Width	5.086	3.964	4.273	3.508	0.299				
$\lambda = 0.15$	MSE	0.875	0.873	0.968	1.042	1.014	1.105	1.086	1.024	1.057
	Coverage	0.963	0.859	0.752	0.651	0.098				
	Width	5.053	3.561	3.840	3.284	0.291				

Table 3 Sensitivity analysis for GTP study analysis results with alternative hyperparameter settings

		sp-lsGPR1	*sp-lsGPR2*	*sp-lsGPR3*	*sp-lsGPR4*	*sp-lsGPR5*	*lsGPR1*	*lsGPR2*
$\lambda = 0.05$	MSE	0.927	0.901	0.933	0.925	0.911	0.963	0.943
	Coverage	0.977	0.975	0.723	0.794	0.799	0.691	0.844
	Width	5.170	5.188	2.267	2.689	2.645	2.916	3.733
$\lambda = 0.1$	MSE	0.890	0.924	0.916	0.931	0.937	0.890	0.979
	Coverage	0.990	0.978	0.756	0.811	0.758	0.871	0.745
	Width	5.295	5.175	2.382	2.749	2.546	3.939	3.744
$\lambda = 0.15$	MSE	0.911	0.905	0.926	0.943	0.933	1.001	1.008
	Coverage	0.991	0.990	0.739	0.790	0.804	0.775	0.552
	Width	5.228	5.360	2.363	2.686	2.690	3.340	2.264

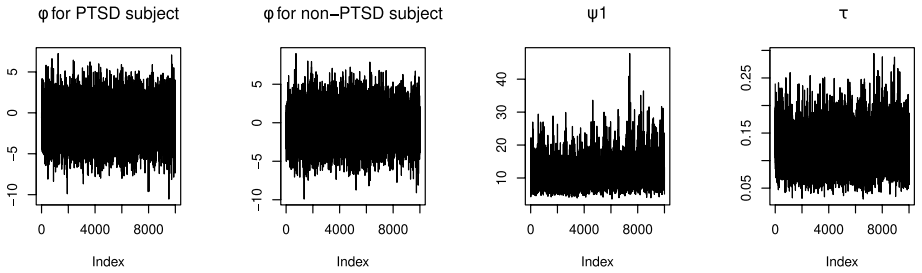

Fig. 8 Trace plots for gaussian process atom ϕ of one PTSD and one non-PTSD subjects, and hyperparameters ψ_1 and τ

4.2.2 Results

The MCMC in stage 2 of the proposed model converged rapidly as evident from Geweke's convergence diagnostic test (Geweke 1992). Fig. 8 shows the trace plots for different model parameters. Moreover, Table 2 shows the prediction performance of different methods when modeling the resilience score. The results are obtained from 50 random splits into training and testing samples. Both the lsGPR and the sparse lsGPR yield considerable lower predictive MSE compared to all competing methods across varying network densities represented via different λ settings ($\lambda = 0.05, 0.1, 0.15$). Permutation tests show that both the lsGPR and the sparse lsGPR method have significant reduction in MSE at 5% level of significance compared to all competing methods, and across all levels of network sparsity. Moreover, the prediction under the sparse lsGPR is significantly lower compared to the lsGPR involving all nodes for network densities represented by $\lambda = 0.05, 0.1$, while the predictive MSE is similar for both methods when $\lambda = 0.15$. However, the coverage of the sparse lsGPR is significantly higher compared to the lsGPR method involving all nodes as well as all other competing methods, across all settings. This suggests the superior ability for characterization of predictive uncertainty using node selection under the proposed sparse lsGPR method, which results from superior predictive ability combined with sufficiently wide predictive intervals.

In contrast, approaches that employ the full edge set, including the linear models as well as the edge-GPR method have extremely inferior prediction performance,

illustrating the perils of regression using high dimensional edge space that does not respect the inherent dependency structure of the network. Moreover, the edge-GPR method often has poor prediction performance that is sometime subdued even compared to linear models, which highlights the drawbacks of using the full edge set in linear or non-linear models. Moreover, the edge-GPR approach has considerably poor coverage due to tight intervals compared to other non-linear regression approaches. These results suggest a strong justification for non-linear regression modeling with dimension reduction using networks that is able to accommodate unknown interactions with additional exposure variables, and has orders of magnitudes improvements over linear models. Importantly, among all the dimension reduction approaches considered, the advantage of dimension reduction using a latent scale manifold approach coupled with sparse node selection (i.e. the sparse lsGPR method) is most prominent that highlights its' considerable benefits over standard dimension reduction techniques.

The performance of the lsGPR and the sparse lsGPR methods were evaluated under different hyperparameter settings as reported in Table 3, which serves as sensitivity analysis of our proposed methods. We consider the following settings: sp-lsGPR1 refers to updating the channels with prior on π at $Beta(1, 10)$ and setting the prior on ψ_1 at inverse Gamma (0.1, 80); sp-lsGPR2 refers to updating the channels with prior on π at $Beta(1, 25)$ and setting the prior on ψ_1 at inverse Gamma (0.1, 80); sp-lsGPR3 refers to fixing the values of the weight matrix to identity while using the prior on ψ_1 as inverse Gamma (0.1, 10); sp-lsGPR4 refers to updating the channels with prior on π at $Beta(1, 10)$ and specifying the prior on ψ_1 at inverse Gamma (0.1, 10); sp-lsGPR5 refers to updating the channels with prior on π at $Beta(1, 25)$ and specifying the prior on ψ_1 at inverse Gamma (0.1, 10); lsGPR1 refers to updating the channels with prior on π at $Beta(1, 10)$ and setting the prior on ψ_1 at inverse Gamma (0.1, 80) but without node selection; lsGPR2 refers to updating the channels with prior on π at $Beta(1, 25)$ and setting the prior on ψ_1 at inverse Gamma (0.1, 80) and without node selection. This sensitivity analysis reveals that although the out of sample predictive accuracy is less under these hyperparameter settings compared to the performance reported in Table 2, the characterization of predictive uncertainty as reflected by the coverage of test samples is significantly higher when using channel selection under different priors on the channel selection probability (see first two columns in Table 3). We also discovered that changing the prior on the scale parameter of the Gaussian process from $\psi_1 \sim InverseGamma(0.1, 80)$ as in Table 2 to $\psi_1 \sim InverseGamma(0.1, 10)$ in Table 3 results in a decrease in predictive accuracy as well as coverage of test samples, due to the tapering of predictive intervals for test samples. These results suggest that overall, the hyperparameter choices made in Table 2 work well for the proposed approaches.

We also examined the node selection results from our sparse lsGPR method based on the posterior inclusion probabilities. Table 4 includes the nodes appear in the top ten percent of the ranked posterior inclusion probabilities from analysis of the binary networks with shrinkage parameters λ at 0.05, 0.1 and 0.15. A total of 12 nodes were selected including one from auditory module, three from default mode, one from visual module, one from fronto-parietal module, one from salience, one from subcortical, and four from unknown module. These regions are visually illustrated in the top panel of Fig. 9, whereas the bottom panel highlights the largest edge differences among these regions between most and least resilient participants. We note that the behavioral patterns listed above have all been shown to be compromised in trauma exposed individuals (Sripada et al., 2013; Falconer et al., 2008 and hence our analysis has a direct relevance in linking the brain network with behavior in trauma exposed individuals. These

Table 4 Information on selected nodes in predicting resilience for the GTP study using sp-lsGPR method. (L) and (R) represent left and right cerebrum respectively

Index Power atlas	Functional module	ROI location	Average posterior inclusion probability
64	Auditory	(L) Sub-lobar, insula	0.1139
91	Default mode	(L) Limbic lobe, posterior cingulate	0.1138
102	Default mode	(R) Frontal lobe, superior frontal gyrus	0.1139
137	Default mode	(L) Frontal lobe, inferior frontal gyrus	0.1139
140	Unknown	(R) Occipital lobe, lingual gyrus	0.1136
172	Visual	(L) Occipital lobe, middle occipital gyrus	0.1137
177	Fronto-parietal	(L) Parietal lobe, inferior parietal lobule	0.1137
185	Unknown	(R) Cerebellum posterior lobe, tuber	0.1137
210	Salience	(R) Frontal lobe, inferior frontal gyrus	0.1143
233	Subcortical	(R) Sub-lobar, extra-nuclear	0.1139
243	Unknown	(L) Cerebellum posterior lobe, declive	0.1139
248	Unknown	(L) Limbic lobe, uncus	0.1142

network differences support our node selection results by highlighting the connectivity differences between individuals with high and low resilience levels. In contrast, the Bayesian horseshoe method has edge-level posterior inclusion probabilities no greater than 0.04 thus we choose not to report its selection results.

5 Conclusion and future direction

In this article, we make major contributions in the network analysis literature by proposing one of the first flexible non-linear Bayesian non-parametric methods for regression with network-valued covariates under a Gaussian process framework, which also has the added advantage of network node selection via spike and slab priors. The proposed latent scale Gaussian process regression approach has the advantage of preserving interpretability at the node level that also facilitates node level feature selection, while allowing flexible non-linear relationships between the network and the outcome while accommodating unknown interactions between the network and supplementary covariates that is simply not feasible under existing linear approaches. Hence, the proposed method provides a desirable middle ground between interpretable linear models equipped with feature selection, and highly flexible non-linear models lacking interpretability and feature selection capabilities, in the context of regression with high-dimensional networks. Our extensive numerical studies revealed the significant advantages under the proposed approach over competing linear and non-linear methods. Interestingly, the version of the proposed method incorporating node selection had better predictive performance as well as higher accuracy in characterising predictive uncertainty compared to the version of the proposed method without node selection, which reinforces existing evidence in literature regarding the merits of feature selection and dimension reduction under a Gaussian process framework (Jiang 2007).

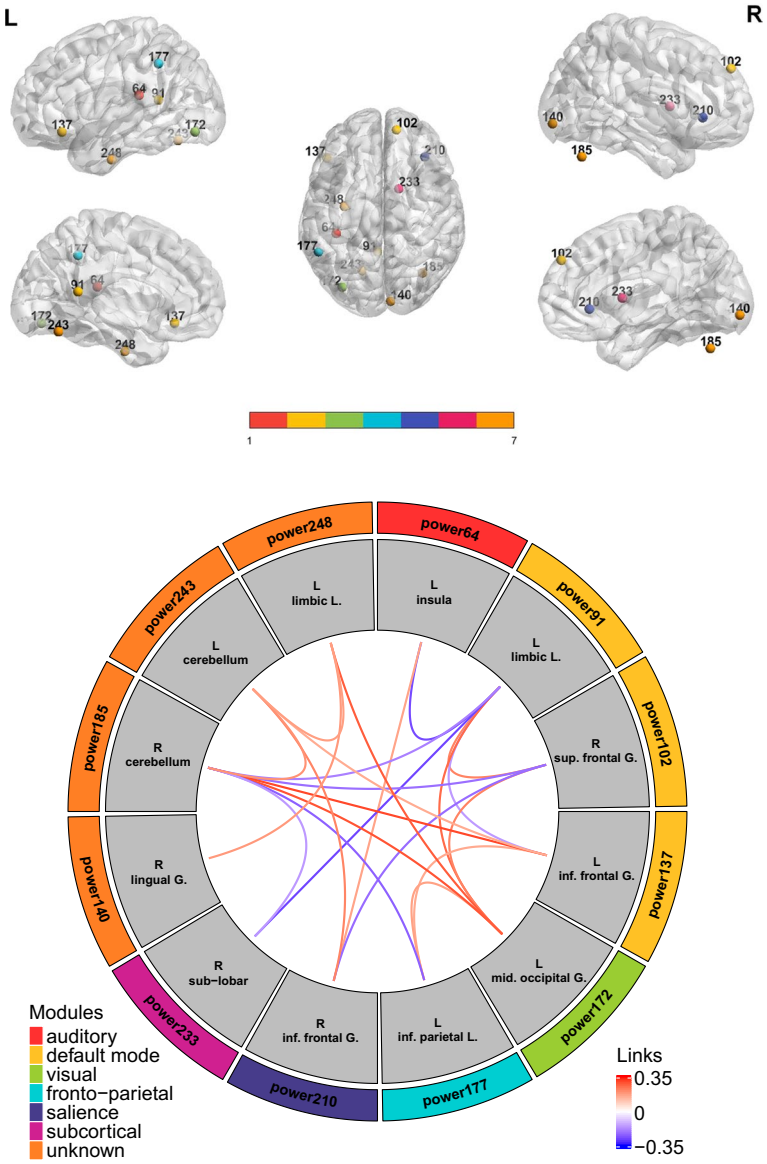

Fig. 9 Brain maps of selected nodes with respect to resilience using our proposed method (upper) and circular plot for differences in functional connectivity between most and least resilient participants among selected nodes (lower). The connections in the lower panel are red or blue depending on whether the difference in edge strengths between the most and least resilient participants is positive or negative

Although we adopt a two stage approach for computational ease and scalability to high-dimensional networks, we note that it is possible that the two-step approach could lead to error propagation. However, our primary goal is to regress the clinical outcome on lower dimensional representations of the brain network, and some inadequacies in

the estimation of the latent scales are tolerated as long as it does not lead to significant decrease in prediction performance in the second stage. Future research may involve more flexible and scalable approaches that jointly update the mapping to the manifold as well as the regression parameters. Given the scarcity of flexible regression approaches involving high-dimensional covariates in literature, the proposed method is expected to have a significant impact in network analysis literature.

Appendix

MCMC algorithm for first stage model

The Gibbs Sampler for the first stage model iterates between the following steps:

1. Update the Pólya–Gamma variables $\omega_{i,kl} \sim PG(1, \delta_{i,kl})$ where $PG(\cdot)$ denotes the Pólya–Gamma distribution and $\delta_{i,kl} = a_i + \mathbf{u}_{ik}^T \Lambda_i \mathbf{u}_{il}$ for $1 \le k < l \le p$.
2. Update the diagonal elements in the weight matrix

$$\lambda_{ir} \sim Bernoulli\left(\frac{\pi \mathscr{L}_1}{\pi \mathscr{L}_1 + (1-\pi)\mathscr{L}_0}\right)$$

where for $r = 2, \cdots, d$,

$$\mathscr{L}_1 = \prod_{k<l,k,l=1}^{p} \frac{1}{2} \exp\left\{(e_{i,kl} - 0.5)\left(a_i + \mathbf{u}_{ik}^T \Lambda_i^{r1} \mathbf{u}_{il}\right) - 0.5\omega_{i,kl}\left(a_i + \mathbf{u}_{ik} T \Lambda_i^{r1} \mathbf{u}_{il}\right)^2\right\}$$

and $\Lambda_i^{r1} = diag(1, \lambda_{i2}, \cdots, \lambda_{i(r-1)}, 1, \lambda_{i(r+1)}, \cdots, \lambda_{id})$, also

$$\mathscr{L}_0 = \prod_{k<l,k,l=1}^{p} \frac{1}{2} \exp\left\{(e_{i,kl} - 0.5)\left(a_i + \mathbf{u}_{ik}^T \Lambda_i^{r0} \mathbf{u}_{il}\right) - 0.5\omega_{i,kl}\left(a_i + \mathbf{u}_{ik} T \Lambda_i^{r0} \mathbf{u}_{il}\right)^2\right\}$$

and $\Lambda_i^{r0} = diag(1, \lambda_{i2}, \cdots, \lambda_{i(r-1)}, 0, \lambda_{i(r+1)}, \cdots, \lambda_{id})$.
3. Update the intercept term $a_i \sim N(\mu_i, \sigma_i^2)$ where $\sigma_i^2 = (\sigma_a^{-2} + \sum_{k<l} \omega_{i,kl})^{-1}$ and $\mu_i = \sigma_i^2 \sum_{k<l} (e_{i,kl} - 0.5 - \omega_{i,kl} \mathbf{u}_{ik}^T \Lambda_i \mathbf{u}_{il})$.
4. Update the latent scales $\mathbf{u}_{ik(-1)} \sim N(A_{ik}^{-1} B_{ik}, A_{ik}^{-1})$ where $A_{ik} = \sum_{j \ne k} \left[\omega_{i,jk} \Lambda_{i0} \mathbf{u}_{ij(-1)} \mathbf{u}_{ij(-1)}^T \Lambda_{i0} + \sigma_u^{-2} \mathrm{I}_{(d-1)}\right]$ and $B_{ik} = \sum_{j \ne k} \left[e_{i,jk} - 0.5 - (a_i + b^2)\omega_{i,jk}\right] \Lambda_{i0} \mathbf{u}_{ij(-1)}$.
5. Update $\pi \sim Beta(a_\pi + \sum_{r=2}^{d} \lambda_{ir}, b_\pi + d - 1 - \sum_{r=2}^{d} \lambda_{ir})$.

Figure 10 includes the scatterplots between the prediction and uncertainty quantification results corresponding to the latent scales derived from fitting the EM algorithm and the MCMC to the first stage model fitting.

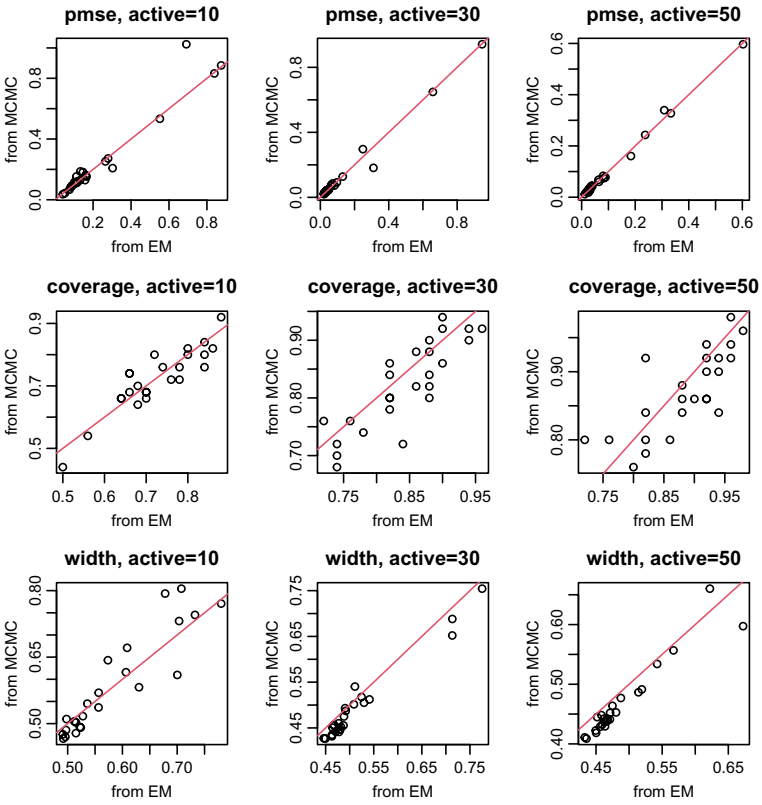

Fig. 10 Scatterplots comparing the prediction and uncertainty quantification results corresponding to the latent scales derived from fitting the EM algorithm and the MCMC to the first stage model fitting

Additional results for sensitivity analysis of channel number selection

The following Fig. 11 includes additional results when the difference between the channel number in the true model and the working model is larger than 1.

Sensitivity analysis for hyper-parameters

The following Table 5 illustrates the performance of the proposed approach under varying choices of hyperparameters (σ_a^2, σ_u^2) in the EM algorithm for fitting the first stage model.

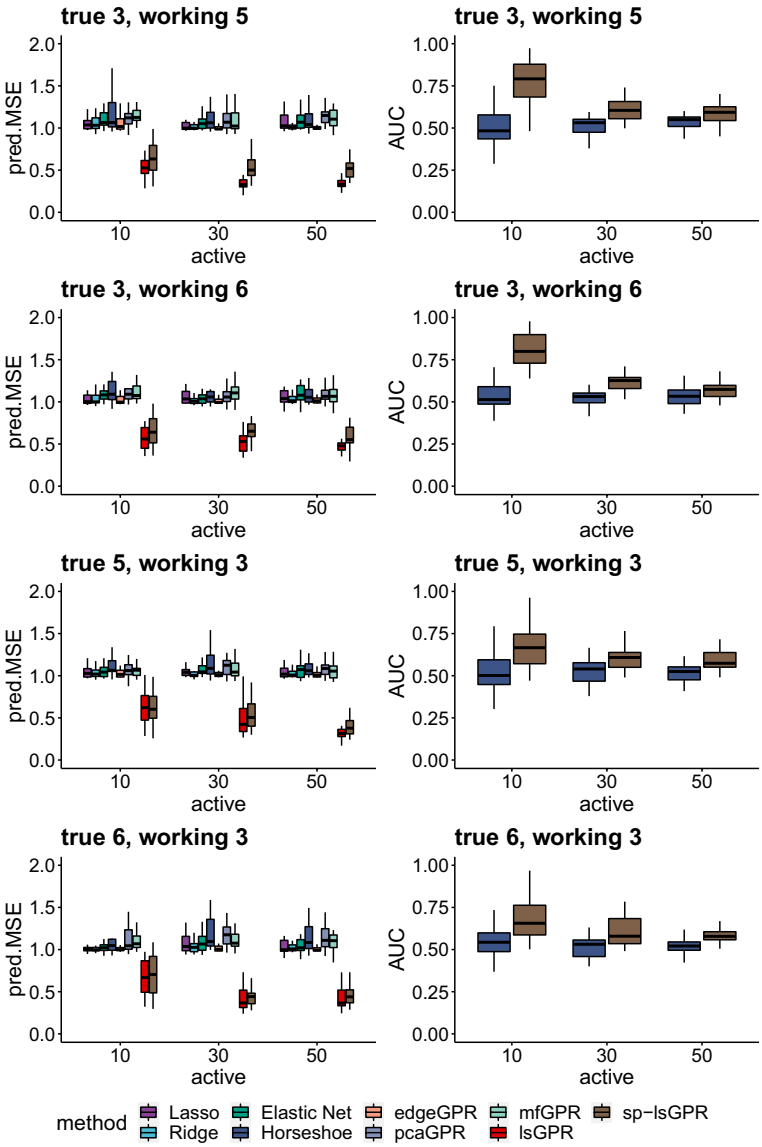

Fig. 11 Additional results for channel number sensitivity tests

Table 5 Sensitivity analysis corresponding to hyperparameters (σ_a^2, σ_u^2) for the first stage EM algorithm

$[\sigma_a^2,\sigma_u^2]$	Active=10			Active=30			Active=50		
	lsGPR PMSE	sp-lsGPR PMSE	sp-lsGPR AUC	lsGPR PMSE	sp-lsGPR PMSE	sp-lsGPR AUC	lsGPR PMSE	sp-lsGPR PMSE	sp-lsGPR AUC
[2, 2]	0.35	0.32	0.92	0.11	0.16	0.79	0.07	0.11	0.76
[0.2, 2]	0.25	0.25	0.90	0.10	0.18	0.80	0.12	0.13	0.74
[0.5, 2]	0.74	0.27	0.91	0.11	0.14	0.82	0.07	0.13	0.77
[1, 2]	0.28	0.25	0.85	0.19	0.22	0.80	0.08	0.13	0.76
[4, 2]	0.42	0.34	0.85	0.12	0.17	0.78	0.09	0.13	0.78
[8, 2]	0.26	0.21	0.87	0.11	0.16	0.81	0.07	0.13	0.79
[20, 2]	0.29	0.22	0.87	0.09	0.13	0.83	0.09	0.12	0.77
[2, 0.2]	0.61	0.66	0.72	0.29	0.37	0.64	0.26	0.29	0.58
[2, 0.5]	0.41	0.39	0.79	0.12	0.16	0.75	0.11	0.17	0.68
[2, 1]	0.26	0.26	0.91	0.10	0.13	0.83	0.15	0.18	0.76
[2, 4]	0.39	0.25	0.86	0.17	0.18	0.79	0.08	0.13	0.76
[2, 8]	0.30	0.32	0.83	0.07	0.16	0.73	0.07	0.14	0.68
[2, 20]	0.35	0.46	0.77	0.14	0.20	0.69	0.13	0.19	0.64

ROC curves in supporting the node selection results

We present here in Fig. 12 ROC curves from simulation scenario 3 under different network sparsity levels and different number of active nodes in generating the response variable.

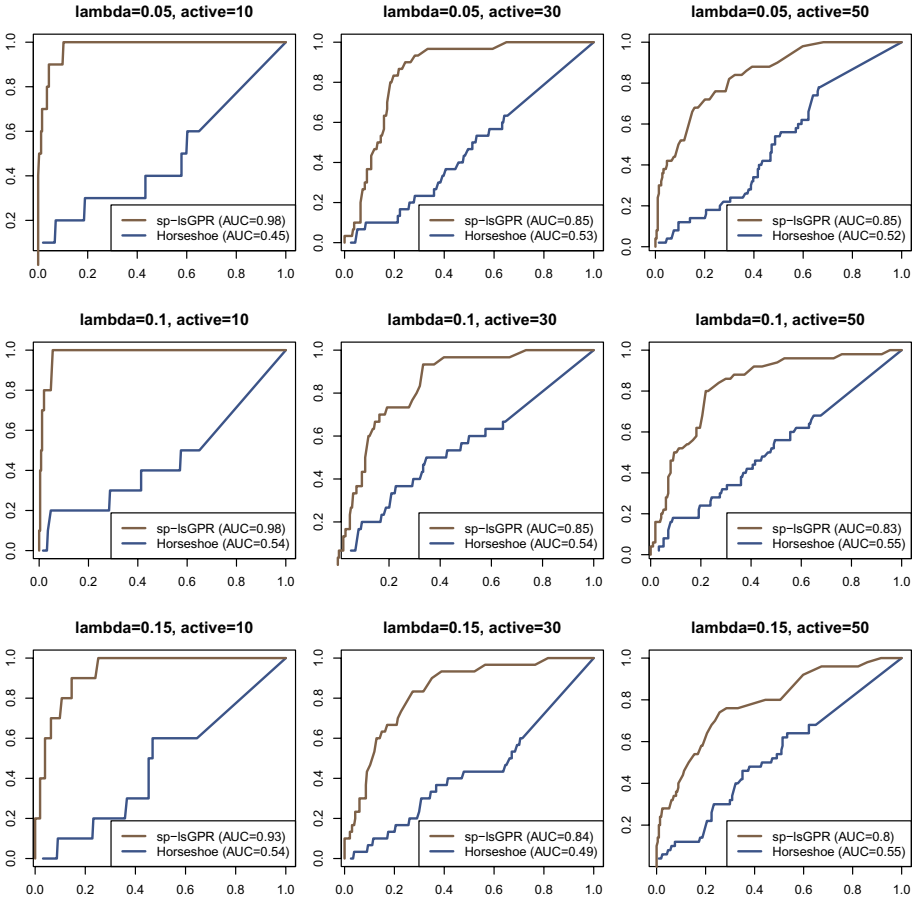

Fig. 12 Selected ROC curves from simulation scenario 3. The Y-axis represents sensitivity and the X-axis represents 1-specificity

Funding This work was partially supported by NIMH award 1R01MH120299. The Grady Trauma Project was supported by Grant/Award Nos.: MH098212, MH071537, MH101380, MH101976, MH096764, M01RR00039, UL1TR000454, HD085850, T32HL130025.

Code Availability Statement The R scripts used for analysis is provided here: https://github.com/Xin-Ma-1/lsGPR

Materials Availability Statement The network data used for GTP analysis is provided here: https://github.com/Xin-Ma-1/lsGPR

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Study procedures were approved by the institutional review board of Emory University, and procedures were consistent with the Declaration of Helsinki. The study protocol involved automatic consent from participants regarding the use of the study data in a deidentified form for secondary analysis. Use of

experimental animals, and human participants - The article does not report experiments on live vertebrates and/or higher invertebrates, and only involves secondary analysis of deidentified data from ADNI.

References

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery, 29*(3), 626–688.

Bassett, D. S., & Bullmore, E. (2006). Small-world brain networks. *The Neuroscientist, 12*(6), 512–523.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation, 15*(6), 1373–1396.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience, 10*(3), 186.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika, 97*(2), 465–480.

Chang, C., Kundu, S., & Long, Q. (2018). Scalable bayesian variable selection for structured high-dimensional data. *Biometrics, 74*(4), 1372–1382.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*(368), 829–836.

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience, 16*(9), 1348.

Connor, K. M., & Davidson, J. R. (2003). Development of a new resilience scale: The connor-davidson resilience scale (cd-risc). *Depression and Anxiety, 18*(2), 76–82.

Craddock, R. C., Holtzheimer, P. E., III., Hu, X. P., & Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine: An Official Journal of the Int Soc for Magnetic Resonance in Medicine, 62*(6), 1619–1628.

Csardi, G., Nepusz, T., et al. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems, 1695*(5), 1–9.

Cui, P., Wang, X., Pei, J., & Zhu, W. (2018). A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering, 31*(5), 833–852.

Cui, Z., Henrickson, K., Ke, R., & Wang, Y. (2019). Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems, 21*(11), 4883–4894.

Du, Y., Fu, Z., & Calhoun, V. D. (2018). Classification and prediction of brain disorders using functional connectivity: Promising but challenging. *Frontiers in Neuroscience, 12,* 525.

Durante, D., Dunson, D. B., & Vogelstein, J. T. (2017). Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association, 112*(520), 1516–1530.

Falconer, E., Bryant, R., Felmingham, K. L., Kemp, A. H., Gordon, E., Peduto, A., et al. (2008). The neural networks of inhibitory control in posttraumatic stress disorder. *Journal of Psychiatry & Neuroscience: JPN, 33*(5), 413.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432–441.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: IN BAYESIAN STATISTICS, Citeseer

Gramacy, R.B. (2018). monomvn: Estimation for Multivariate Normal and Student-t Data with Monotone Missingness. https://CRAN.R-project.org/package=monomvn, r package version 1.9-8

Guha, S., & Rodriguez, A. (2020). Bayesian regression with undirected network predictors with an application to brain connectome data. *Journal of the American Statistical Association, 16*(534), 581–593.

Hallquist, M. N., & Hillary, F. G. (2018). Graph theory approaches to functional network organization in brain disorders: A critique for a brave new small-world. *Network Neuroscience, 3*(1), 1–26.

Higgins, I. A., Kundu, S., Choi, K. S., Mayberg, H. S., & Guo, Y. (2019). A difference degree test for comparing brain networks. *Human Brain Mapping, 40*(15), 4518–4536.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association, 100*(469), 286–295.
Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association, 97*(460), 1090–1098.
Jiang, W., et al. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *The Annals of Statistics, 35*(4), 1487–1511.
Kraemer, G., Reichstein, M., & D, M.M. (2018). dimRed and coRanking—unifying dimensionality reduction in r. *The R Journal 10*(1), 342–358, https://journal.r-project.org/archive/2018/RJ-2018-039/index.html, coRanking version 0.2.2
Kundu, S., & Dunson, D. B. (2014). Latent factor models for density estimation. *Biometrika, 101*(3), 641–654.
Kundu, S., Mallick, B. K., Baladandayuthapani, V., et al. (2019). Efficient bayesian regularization for graphical model selection. *Bayesian Analysis, 14*(2), 449–476.
Lanius, R., Bluhm, R., Lanius, U., & Pain, C. (2006). A review of neuroimaging studies in ptsd: Heterogeneity of response to symptom provocation. *Journal of Psychiatric Research, 40*(8), 709–729.
Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926
Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for Information Science and Technology, 58*(7), 1019–1031.
Lukemire, J., Kundu, S., Pagnoni, G., & Guo, Y. (2020). Bayesian joint modeling of multiple brain functional networks. *Journal of the American Statistical Association, 116*(534), 518–530.
Meng, L., & Xiang, J. (2018). Brain network analysis and classification based on convolutional neural network. *Frontiers in Computational Neuroscience, 12,* 95.
Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., & Bronstein, M.M. (2017). Geometric deep learning on graphs and manifolds using mixture model cnns. *In: Proceedings of the IEEE conference on computer vision and pattern recognition,* pp 5115–5124
Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association, 96*(455), 1077–1087.
Page, G., Bhattacharya, A., & Dunson, D. (2013). Classification via bayesian nonparametric learning of affine subspaces. *Journal of the American Statistical Association, 108*(501), 187–201.
Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya-gamma latent variables. *Journal of the American Statistical Association, 108*(504), 1339–1349.
Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., et al. (2011). Functional network organization of the human brain. *Neuron, 72*(4), 665–678.
Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MA: MIT Press Cambridge.
Relión, J. D. A., Kessler, D., Levina, E., & Taylor, S. F. (2019). Network classification with applications to brain connectomics. *The Annals of Applied Statistics, 13*(3), 1648.
Robert, C.P. (2015). The metropolis-hastings algorithm. arXiv preprint arXiv:1504.01896
Savitsky, T., Vannucci, M., & Sha, N. (2011). Variable selection for nonparametric gaussian process priors: Models and computational strategies. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics, 26*(1), 130.
Scher, C. D., Stein, M. B., Asmundson, G. J., McCreary, D. R., & Forde, D. R. (2001). The childhood trauma questionnaire in a community sample: Psychometric properties and normative data. *Journal of Traumatic Stress, 14*(4), 843–857.
Sprang, G. (1997). The traumatic experiences inventory (tei): A test of psychometric properties. *Journal of Psychopathology and Behavioral Assessment, 19*(3), 257–271.
Sripada, R. K., Garfinkel, S. N., & Liberzon, I. (2013). Avoidant symptoms in ptsd predict fear circuit activation during multimodal fear extinction. *Frontiers in Human Neuroscience, 7,* 672.
Stevens, J. S., Jovanovic, T., Fani, N., Ely, T. D., Glover, E. M., Bradley, B., & Ressler, K. J. (2013). Disrupted amygdala-prefrontal functional connectivity in civilian women with posttraumatic stress disorder. *Journal of Psychiatric Research, 47*(10), 1469–1478.
Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319–2323.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.
Weaver, C., Xiao, L., & Lindquist, M.A. (2021). Single-index models with functional connectivity network predictors. Biostatistics https://doi.org/10.1093/biostatistics/kxab015

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24.

Yang, Y., Dunson, D. B., et al. (2016). Bayesian manifold regression. *The Annals of Statistics, 44*(2), 876–905.

Zhan, X., & Yu, R. (2015). A window into the brain: advances in psychiatric fMRI. *BioMed Research International*. https://doi.org/10.1155/2015/542467.

Zhang, S., Zhou, D., Yildirim, M.Y., Alcorn, S., He, J., Davulcu, H., & Tong, H. (2017). Hidden: hierarchical dense subgraph detection with application to financial fraud detection. *In: Proceedings of the 2017 SIAM International Conference on Data Mining,* SIAM, pp 570–578

Zhou, D., Zhang, S., Yildirim, M.Y., Alcorn, S., Tong, H., Davulcu, H., & He, J. (2017). A local algorithm for structure-preserving graph cut. *In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp 655–664

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Xin Ma[1] · Suprateek Kundu[2] · Jennifer Stevens[3]

[1] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, USA

[2] Department of Biostatistics, The University of Texas at MD Anderson Cancer Center, Houston, USA

[3] Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, USA