Consistency and Adaptation of Gaussian Process Regression, Bayesian Stochastic Block Model and Tail Index

by

Sheng Jiang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____

Surya T. Tokdar, Advisor

_____

David B. Dunson

_____

Alexander Volfovsky

_____

Ismael Castillo

_____

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2021

ABSTRACT

Consistency and Adaptation of Gaussian Process Regression, Bayesian Stochastic Block Model and Tail Index

by

Sheng Jiang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____

Surya T. Tokdar, Advisor

_____

David B. Dunson

_____

Alexander Volfovsky

_____

Ismael Castillo

_____

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2021

Abstract

Bayesian methods offer adaptive inference via hierarchical extensions and uncertainty quantification automatically with corresponding posterior distribution. Frequentist evaluation of Bayesian methods becomes a fundamental and necessary step in Bayesian analysis.

Bayesian nonparametric regression under a rescaled Gaussian process prior offers smoothness-adaptive function estimation with near minimax-optimal error rates. Hierarchical extensions of this approach, equipped with stochastic variable selection, are known to also adapt to the unknown intrinsic dimension of a sparse true regression function. But it remains unclear if such extensions offer variable selection consistency, i.e., if the true subset of important variables could be consistently learned from the data. It is shown here that variable consistency may indeed be achieved with such models at least when the true regression function has finite smoothness to induce a polynomially larger penalty on inclusion of false positive predictors. Our result covers the high dimensional asymptotic setting where the predictor dimension is allowed to grow with the sample size.

Stochastic Block Models (SBMs) are a fundamental tool for community detection in network analysis. But little theoretical work exists on the statistical performance of Bayesian SBMs, especially when the number of communities is unknown. This project studies weakly assortative SBMs whose members of the same community are more likely to connect with one another than with members from other communities. The weak assortativity constraint is embedded within an otherwise weak prior, and, under mild regularity conditions, the resulting posterior distribution is shown to concentrate on the true number of communities and membership allocation as the network size grows to infinity. A reversible-jump Markov Chain Monte Carlo posterior

computation strategy is developed by adapting the allocation sampler of [MMFH13]. Finite sample properties are examined via simulation studies in which the proposed method offers competitive estimation accuracy relative to existing methods under a variety of challenging scenarios.

Tail index estimation has been well studied in the frequentist literature. However, few asymptotic studies on Bayesian tail index estimation are available. This paper works with a transformation based semi-parametric density model by non-parametrically transforming a parametric CDF. The semiparametric density model offers both accurate density estimation and tail index estimation. Compared with frequentist methods, it avoids choosing a high quantile to threshold the data. We provide sufficient conditions on the parametric family and the logistic Gaussian process priors, such that posterior contraction rate of tail index can be established. Limitations of the semiparametric density model are also discussed.

Acknowledgements

First and foremost, I am extremely grateful to my advisor, Dr. Surya T. Tokdar who has been guiding my Ph.D. adventure. In the first few years, I was in the struggle of getting some nice work done. In the difficult times, Surya always offered generous support and insightful suggestions. It is hard to imagine if I could make such significant progress in research and writing without his help.

Secondly, I want to thank my committee members for their valuable intellectual input to my defense and dissertation. I also want to thank Dr. Mike West for encouraging and nudging us to think harder and deeper. Outside Duke, I want to thank Cheng, Sanvesh, and Terrance who have made BNP 12 and ISBA world meeting a great learning experience.

Thirdly, I want to thank my friends who have made my life at Duke colorful. As a big fan of Shengji, I often played the card game. Strategic thinking and educated guessing are fun; group meals are delicious.

Finally, I hope to thank my parents who have been providing unlimited support, both mentally and financially.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Bayesian methods are flexible in modeling and can adapt to different and more complex scenarios via hierarchical extensions. Constraints and prior knowledge can be easily passed to posterior inference by Bayes theorem. The posterior distribution of the model parameters, whether parametric or nonparametric, automatically provides quantitative uncertainty characterization for the unknown quantities of interest.

Suppose data can be repeatedly generated by the statistician who knows exactly the true parameters. A frequentist can compare the sampling distribution of his/her estimators against the truth and see how accurate the estimators are. Obviously, statisticians seldom enjoy the luxury of knowing the truth in real applications. But the thought experiment can be fruitful in understanding how accurate or efficient a statistical procedure is under certain ideal conditions.

How would a Bayesian assess Bayes estimators in the above hypothetical situation? Frequentist evaluation of Bayesian methods typically includes four aspects:

1. Consistency: if the Bayes estimator converges to the true value, as sample size grows to infinity;

2. Posterior contraction rates: how fast the posterior quantity converges to the true value in some metric, as sample size grows to infinity;

3. Bernstein von-Mises (BvM) phenomenon: the posterior distribution is approximately certain Normal distribution, as sample size grows to infinity;

4. Constructing (adaptive) credible sets that have guaranteed asymptotic coverage rates.

Among the aspects of frequentist evaluation, posterior contraction rate calculation is the most fundamental task, as it not only provides a direct comparison against competing frequentist methods, but also serves as a starting point for exploring finer frequentist properties, including BvM phenomena, adaptive credible sets and the ones considered in the dissertation.

Fundamental tools for studying posterior contraction rates, henceforth Schwartz method, have been developed in the past two decades [BSW99, GGvdV00, GvdV07]. A growing number of frequentist evaluations on various Bayesian nonparametric methods have appeared (See [GvdV17] for a textbook treatment).

Nonetheless, many important frequentist properties of many well-known Bayesian methods are left unknown. The dissertation explores some of them via the lens of posterior concentration.

Next, I highlight the specific topics in the follow-up sections. To be specific, the first project studies variable selection consistency of Gaussian process regression which also maintains near minimax-optimal function estimation accuracy; the second

project studies community detection consistency of Bayesian stochastic block model for assortative networks; the third project studies tail index estimation accuracy of a transformation based Bayesian semi-parametric density model. More detailed and broader statistical and scientific background information is introduced later in corresponding chapters.

1.2 Gaussian process regression

In Bayesian nonparametric regression, Gaussian process (GP) priors are a popular choice for the unknown regression function. The estimation error rates for GP regression are shown to be near minimax optimal when the design dimension d is fixed: $n^{-\beta/(2\beta+d)}\log^{\kappa} n$ where n is sample size, β is the *unknown* smoothness parameter of true regression function, and $\kappa > 0$ is some constant [vdVvZ09]. With an hierarchical extension to the GP regression, the error rates can adapt to *unknown* sparsity of the design: $n^{-\beta/(2\beta+d_0)}\log^{\kappa} n$ where $d_0 < d$ is the unknown true sparsity of the design [Tok11, BPD14, YT15]. By a simple comparison, variable selection improves function estimation accuracy by a polynomial factor of n, which is quite significant.

Variable selection can be a by-product of the non-parametric regression problem, as the marginal posterior distribution of the variable inclusion vector can be used for variable selection. Despite near optimal function estimation accuracy, it remains unclear if the variables selected by the posterior mode of the variable inclusion vector are *exactly* truly relevant regressors. If the answer is affirmative, under what condition

variable selection consistency can be guaranteed?

In the Bayesian setting, variable selection consistency refers to the posterior probability on wrong models goes to 0 in probability as sample size grows to infinity. With posterior contraction rate ε_n in L_2, it suffices to study the posterior probability on wrong models *and* the L_2 ball centered at the true regression function f_0 with radius proportional to ε_n.

For regression functions from false negative models that exclude at least one relevant regressor, their L_2 distance to f_0 is greater than $O(\varepsilon_n)$ if some nonparametric beta-min condition holds. The beta-min condition essentially imposes a minimal signal strength constraint to relevant regressors. False negative control is through a reasonable assumption.

For regression functions from false postive models that include *all* relevant regressors *and* at least one irrelevant regressor, the total posterior mass can spread over a much larger L_2 ball centered at f_0 than the L_2 ball also centered at f_0 but with radius proportional to ε_n, such that the posterior mass on the smaller L_2 ball is negligible. Therefore, false positive control is through the sub-optimal posterior contraction rates of false postive models.

1.3 Bayesian stochastic block model

In the context of network analysis, Stochastic block model (SBM), due to its simplicity yet expressiveness, is a popular and powerful modeling tool for community

detection. Under SBM, nodes are connected with probabilities that *only* depend on nodes' community memberships.

Bayesian SBMs with conjugate priors can be inferred with efficient samplers and their empirical performance is successful across different studies [NS01, NF07, vdPvdV18, GBP19]. However, few theoretical guarantees on community detection are available when the number of communities is unknown in advance. (More detailed literature review is deferred to Chapter 3.)

Many networks in social and natural sciences are assortative: nodes in the same community are more likely to be connected with each other than connected with nodes from other communities [See, e.g., New02, New03, ZP20, and the references therein.]. Assortativity provides a channel to identify both the number of communities and the community membership for each node with the nodes' connectivity matrix.

Of course, the nodes' connectivity matrix has to be estimated from data. The hope is the estimation of the connectivity matrix is "accurate" enough so that precise recovery is still possible. As Bayesian SBMs with conjugate priors ignore assortativity, we encode assortativity into the prior as an additional constraint on the connectivity matrix. With Schwartz method, we first establish the posterior contraction rate of connectivity matrix in sup-norm. Then, we use the assortativity assumption to precisely recover the number of communities and the community membership for each node.

A reversible-jump Markov chain Monte Carlo algorithm, by adapting the allo-

cation sampler of [MMFH13], is developed to draw from the posterior distribution. To evaluate finite sample performance, we perform a brief simulation study and the proposed method offers comparable and competitive performance against its competitors.

1.4 Bayesian semi-parametric density model

In the context of heavy-tailed density modeling and extremes forecasting, very limited data on the tail is available and parametric assumptions have to be made. By the Pickands-Balkenma-De Haan theorem, the conditional distribution of excess over certain threshold is asymptotically a generalized Pareto distribution (GPD), as the *threshold* grows to infinity [BDH74, PI75]. The tail index is the inverse of the shape parameter of GPD and can be used to capture the "heaviness" of the tail. Tail index estimation has been extensively studied in the frequentist literature. However, frequentist estimators, parametric or nonparametric, require the practitioners to choose the threshold whose optimal choice is unclear; different choices can lead to different conclusions. (See more detailed literature review in Chapter 4.)

Bayesian approach bypasses the problem of choosing the threshold by modeling the entire density. For instance, [LLD19] consider Pareto mixture models; [TC21] consider a transformation based semi-parametric density model. Both approaches are proved to be consistent. However, posterior contraction rates are unknown. It remains unclear if the Bayesian methods are efficient.

Instead of mixture models, we work with the transformation based density model of [TC21]: the CDF F satisfies $F(y) = \Psi(G_\theta(y))$ where $G_\theta(y)$ is a parametric CDF chosen by the statistician and $\Psi : [0,1] \to [0,1]$ is a non-parametric monotonic transformation. Under the semiparametric density model, the tail index of F is the tail index of G_θ. Regular prior specification of θ can avoid the inconsistency pitfall detailed in [LLD19].

By Schwartz method, we first establish near minimax-optimal density estimation; then, we provide sufficient conditions on the parametric family specification and the prior specification, such that posterior contraction rate of tail index can be established.

Chapter 2

Variable Selection Consistency of Gaussian Process Regression

2.1 Introduction

Sparse estimation and variable selection in high dimensional linear regression has been well studied [Wai09, BvdG11, Ver12, NH14, CSHvdV15, HC15, YWJ16, Wai19]. But an assumption of linearity could be overly restrictive and prone to model misspecification. A natural alternative is to allow the predictor-response relationship to be flexibly determined by an unknown smooth function $f : \mathbb{R}^d \to \mathbb{R}$, leading to the nonparametric regression model

$$Y_i = f(X_i) + \epsilon_i, \qquad \epsilon_i \mid X_i \overset{iid}{\sim} N(0, \sigma^2), \tag{2.1.1}$$

for paired data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$. No theoretical results currently exist on simultaneously estimating f and recovering its sparsity pattern, especially in the high dimensional setting. Earlier results restrict to the low dimensional settings of $d = o(\log n)$ or $d = O(\log n)$ [Zha91, LW08, BL08]. [CD12] present results under more relaxed settings, establishing that variable selection consistency is achievable for

designs of size $\log(d) = o(n)$. However, they focus exclusively on understanding when it is possible to consistently recover the sparsity pattern of f, rather than providing a practicable estimation method of either the function or the sparsity pattern. More recent works have mostly concentrated on sparse additive formulations of f [RWY12, BS17].

In order to estimate f from data, it is attractive to adopt a Bayesian approach where a Gaussian process prior is specified on the unknown regression function. Fairly expressive yet computationally tractable estimation models could be developed by specifying hierarchical extensions of this prior with rescaling and stochastic variable selection in order to infer smoothness and sparsity properties of f, and consequently, improve function estimation and prediction [RW06, GvdV17]. Indeed, under such Bayesian approaches, the posterior distribution of the regression function is known to contract to the true function at a near minimax optimal rate, adapting to both unknown smoothness and sparsity [vdVvZ09, Tok11, BPD14, YT15].

Does this adaptive regression function estimation accuracy translate to correct identification of relevant regressors? We show that the answer is partially *yes*. Specifically, we establish that when the true sparsity pattern has a fixed cardinality d_0 and the true regression function is Sobolev β-smooth, an appropriately specified rescaled Gaussian process model with stochastic variable selection simultaneously estimates the regression function with near minimax optimal estimation accuracy and recovers the correct sparsity pattern with probability converging to one, provided

$\log(d) \leq O(n^{d_0/(2\beta+d_0)})$ and the predictors are independent Gaussian variables with a common variance.

Like [CD12], our result covers the case where the predictor dimension may grow much faster than the sample size, but ties the growth rate with the smoothness level of the true regression function. Larger design dimensions are allowed when the true function is more rough. Both this smoothness dependent bound on the predictor dimension and the independent Gaussian design assumption are necessary for our proof technique which relies on establishing a polynomial slowdown in the posterior contraction rate of a rescaled Gaussian process model when it includes more predictors than necessary.

Establishing this polynomial slowdown in contraction rate requires new and detailed calculations of concentration probabilities of a rescaled Gaussian process around a given function of limited Sobolev smoothness. To prove that a rescaled Gaussian process estimation model achieves near minimax-optimal contraction rate, [vdVvZ09] derived lower bounds on the concentration probabilities for a carefully chosen range of rescaling levels. To establish a polynomial slowdown in rate, we additionally need sharp *upper* bounds on the concentration probabilities at *every* rescaling level. These are derived by an accurate characterization of *small L_2-ball probabilities* of a rescaled Gaussian process at *every* rescaling level. We use the metric entropy method [KL93, LL99, vdVvZ08b] to turn small ball probability calculations into calculating the metric entropy of the unit ball of the Reproducing Kernel Hilbert Space (RKHS)

associated with the Gaussian process. Critically, the Gaussian design assumption allows mapping the RKHS of a squared-exponential Gaussian process to an ℓ_2 ellipsoid whose metric entropy can be bounded accurately.

2.2 GP Regression with Stochastic Variable Selection

Our asymptotic analysis concerns a sequence of experiments, indexed by sample size $n = 1, 2, \ldots$, and with associated sample spaces $\mathcal{D}_n = \{(X_i, Y_i) \in \mathbb{R}^{d_n} \times \mathbb{R}, 1 \leq i \leq n\}$, in which X_i's are taken to be independent realizations from a probability measure Q_n on $\mathbb{R}^{d_n}$ and Y_i's are realized as in the nonparametric regression formulation (2.1.1) for some known $\sigma > 0$ and some unknown $f \in C(\mathbb{R}^{d_n})$, the space of continuous functions on $\mathbb{R}^{d_n}$. We study a Bayesian estimation of f under a rescaled Gaussian process prior with stochastic variable selection. This prior is formalized by hierarchically extending the rescaled Gaussian process prior of [vdVvZ09]; see also [YT15].

2.2.1 Prior specification

We call a stochastic process $W = (W(x) : x \in \mathbb{R}^d)$ a standard, squared exponential Gaussian process on $\mathbb{R}^d$, if for any finite collection $\{x_1, \ldots, x_k\} \subset \mathbb{R}^d$, the random vector $(W(x_1), \ldots, W(x_k))$ has a mean-zero, k-variate Gaussian distribution with covariance matrix $((K(x_i, x_j))_{1 \leq i,j \leq k}$ where $K(x, x') = \exp\{-\|x - x'\|_2^2\}$. It is well

known that such a W could be seen as a random element of $C(\mathbb{R}^d)$ and is infinitely differentiable with probability one.

Now, given a $\gamma \in \{0,1\}^d$ and a probability measure π on $(0,\infty)$, define a γ-sparse, π-rescaled, squared exponential Gaussian process measure, $\mathrm{SEGP}(\mathbb{R}^d;\gamma,\pi)$ for short, as the probability law of the stochastic process $W = (W(x) : x \in \mathbb{R}^d)$ given as $W(x) = W_0(Ax_\gamma)$ where W_0 is a standard, squared exponential Gaussian process on $\mathbb{R}^{|\gamma|}$, $A \sim \pi$ independently of W_0, and $x_\gamma = (x_j : \gamma_j = 1, j = 1,\ldots,d)$ denotes the sub-vector selected according to γ. The rescaling measure $\pi(A)$ plays a key role in smoothness adaptation, facilitating a Bayesian version of fully automated, data-driven bandwidth selection [vdVvZ09].

For our experiment sequence $\mathcal{D}_n$, a rescaled Gaussian process prior on $f \in C(\mathbb{R}^{d_n})$ with stochastic variable selection is defined as the marginal law of f induced by any joint probability measure Π_n on $(\Gamma, f) \in \{0,1\}^{d_n} \times C(\mathbb{R}^{d_n})$ satisfying the following

1. $\Pr(\Gamma = \gamma) = q_n(|\gamma|)/\binom{d_n}{|\gamma|}$, $\gamma \in \{0,1\}^{d_n}$, for some probability vector $(q_n(d) : 0 \le d \le d_n)$ with $q_n(d_n) < 1$.

2. For every $\gamma \in \{0,1\}^{d_n}$, $f \mid (\Gamma = \gamma) \sim \mathrm{SEGP}(\mathbb{R}^{d_n};\gamma,\pi_{n,|\gamma|})$, determined by a collection of probability measures $\pi_{n,d}$ on $(0,\infty)$, $0 \le d \le d_n$.

The sparsity pattern of f is fully encoded by the binary vector Γ. Let $Q_{n,j}$ denote the marginal distribution of the j-th regressor under $X \sim Q_n$. Any $f \in C(\mathbb{R}^{d_n})$ is constant along an axis $j \in \{1,\ldots,d_n\}$ if and only if $\|f - f_j\|_{L_2(Q_n)} = 0$ where $f_j(x) := \int f(x_1,\ldots,x_{j-1},z,x_{j+1},\ldots,x_{d_n})dQ_{n,j}(z)$. Under the prior Π_n, the sparsity

pattern of f given by the subset $\{1 \leq j \leq d_n : f \text{ is not constant along axis } j\}$ is identical to Γ with probability one. Notice that the prior on variable selection is taken to depend only on the cardinality of the included subset.

2.2.2 Connecting selection consistency with estimation accuracy

Given observed data $D_n \in \mathcal{D}_n$, let $\Pi_n(\cdot \mid D_n)$ denote the joint posterior distribution on (Γ, f) under a prior distribution Π_n as above. Assuming D_n was generated from a *true* regression function f_n^* whose sparsity pattern is identified by a $\gamma_n^* \in \{0,1\}^{d_n}$, the issue of variable selection consistency boils down to assessing whether $\Pi_n(\Gamma \neq \gamma_n^* \mid D_n) \to 0$ in some probabilistic manner. As indicated in the Introduction, the main result we prove in this paper is that when the cardinality of γ_n^* remains fixed at a d_0 and f_n^* is of finite Sobolev smoothness of order β, one has $\Pi_n(\Gamma \neq \gamma_n^* \mid D_n) \to 0$ as long as $\log(d_n) \precsim n^{d_0/(2\beta+d_0)}$ and the regressors are independent Gaussian variables with equal variance. Below we give a sketch of the argument of how such a claim can be made based on adaptive function estimation accuracy.

The question of variable selection could be directly related to that of function estimation quality as follows. Consider a small ball around the truth: $E_n = \{f : \rho(f, f_n^*) \leq \varepsilon_n\}$ where ρ is an appropriate metric and $\varepsilon_n > 0$. Notice that

$$\Pi_n(\Gamma \neq \gamma_n^* \mid D_n) \leq \Pi_n(f \in E_n^c \mid D_n) + \Pi_n(\Gamma \neq \gamma_n^*, f \in E_n \mid D_n). \tag{2.2.1}$$

If it were known that the posterior on f contracts to the truth under metric ρ at a rate ε_n or faster, then the first term on the right hand side would eventually vanish in probability. In order for the second term to vanish as well, one needs to establish that the same fast rate of contraction cannot be achieved under a wrong selection of variables.

Partition the space of wrong selections into two parts: $\{0,1\}^{d_n} \setminus \{\gamma_n^*\} = \{\gamma \in \{0,1\}^{d_n} : \gamma_n^* \not\leq \gamma\} \cup \{\gamma \in \{0,1\}^{d_n} : \gamma_n^* < \gamma\} =: \mathrm{FN}(\gamma_n^*) \cup \mathrm{FP}(\gamma_n^*)$, where the defining inequalities are taken to be coordinate-wise. The *false negative* set $\mathrm{FN}(\gamma^*)$ consists of selections that miss at least one true predictor. The *false positive* set $\mathrm{FP}(\gamma^*)$ contains selections that include all important variables and at least one unimportant regressor. Accordingly, the second posterior probability in (2.2.1) splits into two pieces,

$$\begin{aligned}\Pi_n(\Gamma \neq \gamma_n^*, f \in E_n \mid D_n) = \Pi_n(\Gamma \in \mathrm{FN}(\gamma_n^*), f \in E_n \mid D_n) \\ + \Pi_n(\Gamma \in \mathrm{FP}(\gamma_n^*), f \in E_n \mid D_n),\end{aligned} \tag{2.2.2}$$

of which the first term could be expected to be exactly zero for large n as long as one assumes that the signal strength of f_n^* in any of its relevant variables, measured according to ρ, is above a fixed threshold $\delta_n \equiv \delta$ (Assumption 3 in Section 2.3). Such an f_n^* will be at least a δ distance away in metric ρ from any f whose sparsity pattern $\gamma \in \mathrm{FN}(\gamma_n^*)$.

No such separation exists for regression functions associated with selections in the false positive set. For any $\gamma \in \mathrm{FP}(\gamma_n^*)$, one would expect the conditional posterior

$\Pi_n(f \mid \Gamma = \gamma, D_n)$ to place considerable mass around the truth f_n^*. Any hope of the second term on the right hand side of (2.2.2) being small rests on establishing that such a conditional posterior would contract at a slower rate than ε_n. This is a legitimate hope because it is known that the minimax error rate of function estimation usually worsens when additional irrelevant variables are selected in the regression model [Sto82]. Therefore, assuming f_n^* belongs to the class of β-smooth functions for some $\beta > 0$ (Assumption 4 part 1), a reasonable proof strategy would be to consider $\varepsilon_n = M\underline{\varepsilon}_n(\log n)^\kappa$ for some $M, \kappa > 0$ where $\underline{\varepsilon}_n = n^{-\beta/(2\beta+d_0)}$ is the usual minimax rate of estimation of an f^* of β-smooth functions, and establish a polynomial difference in contraction rates between the overall posterior and the conditional posteriors under false positive selection.

In Sections 2.3 and 2.4 we establish the above rate difference result and give a formal proof of variable selection consistency under a rescaled GP prior with stochastic variable selection. Establishing slower posterior contraction rates under false positive selection necessitates working with the $L_2(Q_n)$ metric $\rho(f, f') = \|f - f'\|_{L_2(Q_n)} = \{\int (f - f')^2 dQ_n\}^{1/2}$. This choice of metric is slightly different than those considered in [vdVvZ09, YT15]. Consequently, a new proof is needed to establish overall posterior contraction at a rate of ε_n. To show that posterior contraction rate becomes polynomially slower under false positive selection, we have to make two important assumptions: f_n^* is exactly β-smooth and no smoother (Assumption 4 part 2) and Q_n is the mean-zero Gaussian measure on $\mathbb{R}^{d_n}$ with covariance matrix $\xi^2 \cdot I_{d_n}$ (As-

sumption 1). The finite smoothness of f_n^* is indeed necessary to ensure that the conditional prior $\Pi(f \mid \Gamma = \gamma)$ sits less densely around f_n^* when $\gamma \in \mathrm{FP}(\gamma_n^*)$ than when $\gamma = \gamma_n^*$. But the assumption of an independent Gaussian design with a fixed variance is a technical convenience that enables sharp calculations of the concentration probabilities of rescaled GP laws, which are necessary for establishing the above result. Additionally, the stochastic variable selection prior is assumed to favor small models, to control the total posterior probability of the exponentially growing false positive set (Assumption 7).

2.3 Main Result

Toward a formal and rigorous treatment of the arguments presented above, we first state the necessary assumptions and the main variable consistency result. A lengthy discussion of the assumptions is delayed to Section 4.5. Supporting results on the posterior contraction rates and difference in such rates under correct and false positive selections are presented in Section 2.4, relying upon the sharp small ball probability calculations in Section 2.5.

An important assumption, needed essentially for technical reasons, is that the design distribution is uncorrelated Gaussian.

Assumption 1 (Gaussian random design)**.** *The design measure $Q_n = G_{d_n}$ where G_d denotes the d-variate Gaussian measure $N_d(0, \xi^2 I_d)$. The variance $\xi^2 > 2/e$ may be unknown but does not change with n.*

For each sample size n, the true data generating distribution $\mathbb{P}_n^*$ is taken to be an element of the model space identified by $f = f_n^* \in \mathcal{L}_2(Q_n)$ where $\mathcal{L}_2(Q_n) := C(\mathbb{R}^{d_n}) \cap L_2(Q_n)$.

Additionally, we assume that the sequence $(f_n^* : n \geq 1)$ remains essentially the same across n, formalized as below. For notational convenience, for any $d \in \mathbb{N}$ and $\gamma \in \{0,1\}^d$, let $T_\gamma : C(\mathbb{R}^{|\gamma|}) \to C(\mathbb{R}^d)$ denote the function embedding operator: $(T_\gamma f)(x) = f(x_\gamma)$, $f \in C(\mathbb{R}^{|\gamma|})$, $x \in \mathbb{R}^d$. With T_γ, any high dimensional functions with redundant variables can be decomposed into a low dimensional function without any redundant variables and a variable inclusion vector. Smoothness conditions are directly imposed to the low dimensional functions; sparsity and dimensionality assumptions are made on the variable inclusion vector.

Assumption 2 (Finite sparsity of true regression function)**.** *There exist* $n_0, d_0 \in \mathbb{N}$, $f_0 \in \mathcal{L}_2(Q_0)$ *and a sequence of binary vectors* $\gamma_n^* \in \{0,1\}^{d_n}$, *such that* $d_n \geq d_0$, $|\gamma_n^*| = d_0$ *and* $f_n^* = T_{\gamma_n^*} f_0$ *for all* $n \geq n_0$.

Assumption 2 makes it clear that for all large n, the true function is sparse and the support size d_0 does not grow with sample size. To avoid any ambiguity about the true sparsity level d_0, it is important to identify it as the minimal support size for the sequence $(f_n^* : n \geq 1)$. This is done via the next assumption on signal strength which ensures that each of the d_0 inputs to f_0 results in a variability that is detectable in the L_2 topology. Toward this, for each $j \in \{1, \ldots, d_0\}$, define $f_{0j} \in \mathcal{L}_2(Q_0)$ as the projection of f_0 perpendicular to the j-th axis, given by $f_{0j}(x) :=$

$\int_{\mathbb{R}} f_0(x_1, \ldots, x_{j-1}, z, x_{j+1}, \ldots, x_{d_0}) dG_1(z)$, $x \in \mathbb{R}^{d_0}$.

Assumption 3 (Signal strength is L_2 detectable)**.** *The minimum signal strength in the relevant variables* $\delta := \min_{1 \le j \le d_0} \|f_0 - f_{0j}\|^2_{L_2(G_{d_0})}$ *is strictly positive.*

An immediate consequence of Assumption 3 is that for any $n \ge n_0$, f_n^* is at least a δ distance away from any $f \in \mathcal{L}_2(Q_n)$ that is constant along at least one axis j for which $\gamma^*_{n,j} = 1$. More formally, for any $n \ge n_0$, and any $\gamma \in \{0,1\}^{d_n}$ with $\gamma_n^* \not\le \gamma$, $\inf\{\|f_n^* - f\|^2_{L_2(Q_n)} : f \in T_\gamma C(\mathbb{R}^{|\gamma|}) \cap \mathcal{L}_2(Q_n)\} \ge \delta$. To see why, without loss of generality, consider the toy example of $f_0(x_1, x_2)$ and $f(x_1, x_3)$ where x_1 and x_2 are correct variables and x_3 is the redundant variable. Then we can compute $\|f_0 - f\|^2_{L_2(Q)} = \mathbb{E}[(f_0(X_1, X_2) - f(X_1, X_3))^2] = \mathbb{E}[(f_0(X_1, X_2) - f_{0,2}(X_1) + f_{0,2}(X_1) - f(X_1, X_3))^2]$ whose cross-term $\mathbb{E}[(f_0(X_1, X_2) - f_{0,2}(X_1))(f_{0,2}(X_1) - f(X_1, X_3))] = 0$ holds because X_i's are independent.

Next, we formalize the notion that the true regression function is β-smooth but no smoother. For any $d \in \mathbb{N}$, $\beta > 0$, let $H^\beta(\mathbb{R}^d)$ denote the Sobolev space of functions $h : \mathbb{R}^d \to \mathbb{R}$ with norm $\|h\|_{H^\beta(\mathbb{R}^d)}$ given by

$$\|h\|^2_{H^\beta(\mathbb{R}^d)} := \int_{\mathbb{R}^d} |\hat{h}(\lambda)|^2 (1 + \|\lambda\|_2^2)^\beta d\lambda < \infty \tag{2.3.1}$$

where $\hat{h}$ is the Fourier transform[1] of h. Recall that functions $h \in H^\beta(\mathbb{R}^d)$ have square-integrable, (weak) derivatives $D^{(k)} f$, $k = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$, of order $|k| \le \beta$.

[1] $\hat{h}(\lambda) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-i\langle \lambda, t \rangle} h(t) dt$, $\lambda \in \mathbb{R}^d$.

Assumption 4 (Smoothness of f_0)**.** *The true function f_0 satisfies*

1. *There exists a $\beta > d_0/2$ such that $f_0 \in H^\beta(\mathbb{R}^{d_0}) \cap L_2(Q_0)$.*

2. *There also exists an $\alpha \in (\beta, \beta(1 + 1/d_0))$ such that $|\widehat{f_0\sqrt{g_{d_0}}}(\lambda)| \gtrsim \|\lambda\|_2^{-(\alpha+d_0/2)}$ for every $\lambda \in \mathbb{R}^{d_0}$ with $\|\lambda\|_2 \geq 1$, where g_{d_0} is the probability density function of G_{d_0}.*

Part 2 of the assumption ensures that $f_0 \notin H^b(\mathbb{R}^{d_0})$ for any $b > \alpha$ and hence has limited regularity which is important in establishing that the posterior contraction rate at f_n^* is polynomially slower under false positive inclusion.

The posterior contraction rates also depend on the rescaling measures $\pi_{n,d}$, $0 \leq d \leq d_n$. The following assumption is mildly adapted from [vdVvZ09]. The modification is needed in part because in determining sharp upper bounds on the concentration probabilities of rescaled GP priors, one needs to integrate over the entire range of the rescaling parameter. Below, with a slight abuse of notation, we let $\pi_{n,d}$ also denote the probability density function underlying the eponymous rescaling measure.

Assumption 5 (Rescaling measures)**.** *For each $d \in \mathbb{N}$, there exist constants C_1, C_2, C_3, D_1 and D_2, all independent of d, such that*

1. *for all sufficiently large a, $\pi_{n,d}(a) \geq D_1 e^{-C_1 a^d \log^{d+1}(a)}$;*

2. *for every $a > 1/\xi$, $\pi_{n,d}(a) \leq D_2 a^{d-1} e^{-C_2 a^d (\log^{d+1}(a) \vee 1) + C_3 \log(d)}$;*

3. *$\pi_{n,d}(0, \xi^{-1}) = 0$.*

This assumption is satisfied, for example, when $\pi_{n,d}$ is the truncation to (ξ^{-1}, ∞) of the probability law of a random variable A for which $A^d \log^{d+1}(A)$ has an exponential distribution with a rate parameter that is constant in n and d.

Our next assumption regulates how fast the design dimension d_n can grow in n. If the support γ_n^* were known, the $L_2(Q_n)$ minimax estimation error of a β-smooth function would be of the order of $\underline{\varepsilon}_n = n^{-1/(2+d_0/\beta)}$ [Sto82]. Not knowing the support means that one incurs an additional error of the order of $d_0 \log(d_n/d_0)/n$ for having to carry out variable selection. We require this additional error to not overwhelm the original estimation error.

Assumption 6 (Growth of d_n)**.** *The design dimension d_n satisfies* $\log(d_n) \precsim n\underline{\varepsilon}_n^2 \asymp n^{d_0/(2\beta+d_0)}$.

A final assumption is needed on the sparsity induced by the prior distribution. In particular it is needed that the prior on Γ favors small selection sizes and heavily penalizes extremely large selections.

Assumption 7 (Prior sparsity)**.** *For all sufficiently large n,*

1. $q_n(d_0) \geq \exp\{-n\underline{\varepsilon}_n^2\} \asymp \exp\{-n^{d_0/(2\beta+d_0)}\}$,
2. $q_n(d) \leq \exp\{-Cd^\rho\}$ *for every* $d \succsim n^{2\beta/\{\alpha(2\beta+d_0)\}}$, *for some constants* $C > 0$ *and* $\rho \geq (d_0+1)/2$.

Assumption 7 seemingly requires the knowledge of the true support size d_0, but one can relax this by letting ρ grow slowly as sample size increases. The assumption

would hold, for instance if one chose a prior that caps the selection size $|\Gamma|$ at an $m_n \leq d_n$ and let m_n grow slowly with n, e.g., $m_n \asymp n^{1/\log\log n}$. Formally, $q_n(d) \propto I(d < n^{1/\log\log n})$, $d = 0, \ldots, d_n$. An alternative is to not use a cap, but employ aggressive penalization of larger selections: $q_n(d) \propto d^{k\log\log n - 1} \exp\{-d^{k\log\log n}\}$, $0 \leq d \leq d_n$, for some constant k. The latter choice is equivalent to an appropriately tuned Beta-Binomial prior on individual regressor inclusion.

Building upon these formal assumptions, we are able to offer the following rigorous statement and proof of variable selection consistency.

Theorem 2.3.1. *Under Assumptions 1-7,* $\mathbb{P}_n^* \left[\Pi_n(\Gamma \neq \gamma_n^* \mid D_n)\right] \to 0$, *as* $n \to \infty$.

Proof. As before, let $E_n = \{f \in \mathcal{L}_2(Q_n) : \|f - f_n^*\|_{L_2(Q_n)} \leq \varepsilon_n\}$ where $\varepsilon_n = \underline{\varepsilon}_n(\log n)^\kappa$ with $\kappa = (d_0 + 1)/(2 + d_0/\beta)$. Consider the upper bound on $\Pi_n(\Gamma \neq \gamma_n^* \mid D_n)$ jointly given in (2.2.1) and (2.2.2). By Theorem 2.4.1 in the next section, $\Pi_n(f \in E_n^c \mid D_n) \to 0$ in probability as $n \to \infty$. By Assumption 3, the first term of the bound given by (2.2.2) is exactly zero for all large n because the prior probability $\Pi_n(\Gamma \in \mathrm{FN}(\gamma_n^*), f \in E_n) = 0$ whenever $\varepsilon_n < \delta$. The second piece of this bound vanishes in probability by Proposition 1, which leverages on detailed calculations of concentration properties of Gaussian process laws presented in Section 2.5. □

2.4 Posterior Concentration via Schwartz Theory

This Section presents supporting results for the proof of Theorem 3.3.2. As the proof technique is standard, details of the proofs are in Appendix A.

Theorem 2.4.1. *Under Assumptions 1, 2, 4-1, 6 and 5, let* $\varepsilon_n = \underline{\varepsilon}_n(\log n)^{\kappa}$ *with* $\kappa = (d_0+1)/(2+d_0/\beta)$ *for* $n \geq 1$*, then for any sufficiently large constant* M*,*

$$\mathbb{P}_n^*[\Pi_n(f \in \mathcal{L}_2(Q_n) : \|f - f_n^*\|_{L_2(Q_n)} > M\varepsilon_n\} \mid D_n)] \to 0, \text{ as } n \to \infty.$$

A proof of this result is presented in Appendix A by verifying Theorem 2.1 of [GGvdV00]. In the proof, we first verify the Kullback-Leibler prior mass condition that for all sufficiently large n,

$$\Pi_n(f \in B_n(f_n^*, \varepsilon_n)) \geq e^{-n\varepsilon_n^2} \tag{2.4.1}$$

where for any $g \in \mathcal{L}_2(Q_n)$ and $\epsilon > 0$, one defines $B_n(g,\epsilon) = \{f : K(\mathbb{P}_g^1, \mathbb{P}_f^1) \leq \epsilon^2, V(\mathbb{P}_g^1, \mathbb{P}_f^1) \leq \epsilon^2\}$, with $\mathbb{P}_f^1$ denoting the probability distribution of a single observation pair (X_1, Y_1) under the model element with regression function f, and, $K(\mathbb{P}_g^1, \mathbb{P}_f^1) = \mathbb{P}_g^1 \log(d\mathbb{P}_g^1/d\mathbb{P}_f^1)$, $V(\mathbb{P}_g^1, \mathbb{P}_f^1) = \mathbb{P}_g^1[(\log(d\mathbb{P}_g^1/d\mathbb{P}_f^1))^2] - K(\mathbb{P}_g^1, \mathbb{P}_f^1)^2$. Next, a sieve $\mathbb{B}_n \subset L_2(Q_n)$, $n \in \mathbb{N}$ is produced such that $\Pi_n(f \notin \mathbb{B}_n) \leq \exp(-4n\varepsilon_n^2)$ and $\log N(\varepsilon_n, \mathbb{B}_n, \|\cdot\|_{L_2(Q_n)}) \leq n\varepsilon_n^2$ for all sufficiently large n. Here, $N(\epsilon, S, \rho)$ is used to denote the ϵ-covering number of a subset S in a metric space with metric ρ.

Our statement and proof technique for Theorem 2.4.1 mirror the near minimax-

optimal posterior contraction results and proofs of [vdVvZ09, YT15], but are slightly novel in its use of the $L_2(Q_n)$ topology on the space of regression functions. In particular, under the stochastic design assumption, one has the simplification that $B_n(g,\epsilon) = \{f \in \mathcal{L}_2(Q_n) : \|f-g\|_{L_2(Q_n)} \leq \epsilon\})$. Thus the Kullback-Leibler prior mass condition translates to a simpler $L_2(Q_n)$ prior mass condition:

$$\Pi_n(\{f \in \mathcal{L}_2(Q_n) : \|f - f_n^*\|_{L_2(Q_n)} \leq \varepsilon_n\}) \geq e^{-n\epsilon_n^2}. \tag{2.4.2}$$

Proposition 1. *Under Assumptions 1, 2, 4-2, 5 and 7, let* $E_n = \{f \in \mathcal{L}_2(Q_n) : \|f - f_n^*\|_{L_2(Q_n)} \leq M\varepsilon_n\}$, $\varepsilon_n = \underline{\varepsilon}_n(\log n)^\kappa$ *with* $\kappa = (d_0+1)/(2+d_0/\beta)$ *for* $n \geq 1$, *then for all sufficiently large constant* M,

$$\mathbb{P}_n^*[\Pi_n(\Gamma \in \mathrm{FP}(\gamma_n^*), f \in E_n \mid D_n)] \to 0, \ \text{as } n \to \infty.$$

A proof is given in Appendix A. It follows Lemma 1 of [Cas08] and establishes that

$$\frac{\Pi_n(\Gamma \in \mathrm{FP}(\gamma_n^*), f \in E_n)}{\Pi_n(f \in B_n(f_n^*, \varepsilon_n))} \leq e^{-2n\varepsilon_n^2}, \tag{2.4.3}$$

which, according to Lemma 1 in [GvdV07], along with (2.4.1), guarantees that the posterior probability of the set in the numerator vanishes in probability.

The message of Proposition 1 is that the posterior contraction rates of false positive models are significantly slower than minimax rates. With Theorem 2.4.1, it further implies false positive models eventually receive negligible posterior mass. The

use of posterior contraction rates adaptation echoes [GLvdV08] who establishes model selection consistency among density models with different regularity levels.

The key inequalities in (2.4.2) and (2.4.3) require establishing lower and upper bounds on the prior concentration in $L_2(Q_n)$ balls around f_n^*. Such concentrations are completely governed by the *f_n^*-shifted small ball probabilities* of the underlying sparse Gaussian processes at appropriate rescaling levels. The lower bound calculations are similar to those in [vdVvZ09] and require working with rescaling levels that appropriately grow to infinity as determined by the true smoothness level β. But, as mentioned earlier, the upper bound calculations are much more technically involved and require dealing with all rescaling levels. Furthermore, unlike [vdVvZ09] we calculate small ball probabilities by viewing the Gaussian process as a random element in $L_2(Q_n)$, necessitating new characterization of the associated reproducing kernel Hilbert space. These details are presented in the next section.

2.5 Small Ball Probability of Rescaled GP

2.5.1 Series representation of $W^{a,\gamma}$

Let $W^{a,\gamma} \sim \text{SEGP}(\mathbb{R}^{d_n}; \gamma, \delta_a)$ where $\gamma \in \{0,1\}^{d_n}$, $|\gamma| > 0$, and $a > 0$ is a fixed rescaling level, i.e., the rescaling measure is the Dirac delta measure at level a. In this section, we obtain a series representation of $W^{a,\gamma}$ via the Karhunen-Loève expansion of the covariance kernel of $W^{a,\gamma}$.

The covariance kernel of $W^{a,\gamma}$ is

$$K_{a,\gamma}(s,t) = e^{-a^2\|s_{[\gamma]}-t_{[\gamma]}\|_2^2} = \prod_{\{i:\gamma_i=1\}} e^{-a^2(s_i-t_i)^2}. \tag{2.5.1}$$

To obtain an eigen-expansion of $K_{a,\gamma}(\cdot,\cdot)$, first recall the eigen-expansion of univariate Squared Exponential kernel function under Gaussian design. For $s,t \in \mathbb{R}$, $K_{a,1}(s,t) = e^{-a^2(s-t)^2} = \sum_{j=0}^{\infty} \lambda_j \varphi_j(s)\overline{\varphi_j(t)}$, where eigenvalues $\lambda_j = \sqrt{2v_1/V}B^j$, eigenfunctions $\varphi_j(x) = e^{-(v_3-v_1)x^2} H_j\left(\sqrt{2v_3}x\right)$ with physicists' Hermite polynomial $H_j(x) = (-1)^j e^{x^2}\frac{d^j}{dx^j}e^{-x^2}$, and the constants in the eigen-expansion are defined as follows,

$$v_1^{-1} = 4\xi^2, v_2 = a^2, v_3 = \sqrt{v_1^2 + 2v_1v_2}, V = v_1 + v_2 + v_3, B = v_2/V. \tag{2.5.2}$$

(See Chapter 4.3 of [RW06] for more details). The eigenfunctions $\{\varphi_j\}$ form an orthonormal basis under the inner product

$$\langle \varphi_j, \varphi_k \rangle_G = \int_{\mathbb{R}} \varphi_j(s)\,\varphi_k(s)\,g(s)\,ds = \delta_0(j-k),$$

where G denotes the univariate Normal distribution $N(0,\xi^2)$ and g is its density function relative to Lebesgue measure. The Gaussian measure G corresponds to Assumption 1.

With the univariate expansion, $K_{a,\gamma}$ is a tensor product of univariate SE kernels

and admits the following expansion: for $s, t \in \mathbb{R}^{d_n}$,

$$\begin{aligned} K_{a,\gamma}(s,t) &= \prod_{\{i:\gamma_i=1\}} \left(\sum_{j=0}^{\infty} \lambda_j^{(i)} \varphi_j^{(i)}(s_i) \overline{\varphi_j^{(i)}(t_i)} \right) \\ &\equiv \sum_{k=0}^{\infty} \mu_k^{(\gamma)} \psi_k^{(\gamma)}(s) \overline{\psi_k^{(\gamma)}(t)}, \end{aligned} \tag{2.5.3}$$

where $\lambda_j^{(i)}$ is the j^{th} eigenvalue of i^{th} univariate SE kernel, $\varphi_j^{(i)}$ is the j^{th} eigenfunction of i^{th} univariate SE kernel, eigenfunctions $\left\{\psi_k^{(\gamma)}\right\}$ and eigenvalues $\left\{\mu_k^{(\gamma)}\right\}$ are ordered by collecting lower order terms first.

The eigenvalue $\mu_k^{(\gamma)} = (2v_1/V)^{|\gamma|/2} B^m$ for some $m \in \mathbb{N}$ and m is weakly increasing in k. Note the number of $k-$tuples of positive integers whose sum is m is $\binom{m-1}{k-1}$, the number of terms involving B^m for $m \geq 1$ is $\sum_{k=1}^{m} \binom{m-1}{k-1}\binom{|\gamma|}{k} = \binom{|\gamma|+m-1}{m}$ by Vandermonde's identity. By ratio test, the eigenvalues are summable: $\sum_{m=1}^{\infty} \binom{|\gamma|+m-1}{m} B^m < \infty$. As we collect low order terms first, first few terms are $\mu_0^{(\gamma)} = (2v_1/V)^{|\gamma|/2}$, $\mu_k^{(\gamma)} = (2v_1/V)^{|\gamma|/2} B$ for $k = 1 : |\gamma|$, $\mu_k^{(\gamma)} = (2v_1/V)^{|\gamma|/2} B^2$ for $k = |\gamma|+1 : |\gamma| + \binom{|\gamma|+1}{2}$. (Recall the constants in (2.5.2).)

Eigenfunctions are defined accordingly depending on eigenvalues. The ordering of the eigenfunctions with the same eigenvalues does not matter. The eigenfunctions are orthogonal if the base measure is $Q_n \equiv N(0, \xi^2 I_{d_n})$ where isotropy is assumed without loss of generality.

With the eigen-expansion of $K_{a,\gamma}(\cdot,\cdot)$ and the summable eigenvalues, $W^{a,\gamma}$ has

the series representation: for $Z_j \overset{iid}{\sim} N(0,1)$, $j = 0, 1, ...,$

$$W_t^{a,\gamma} = \sum\nolimits_{j=0}^{\infty} Z_j \sqrt{\mu_j^{(\gamma)}} \psi_j^{(\gamma)}(t). \tag{2.5.4}$$

By orthonormality of $\{\psi_j^{(\gamma)}\}$, $||W^{a,\gamma}||^2_{L_2(Q_n)} = \sum_{j=0}^{\infty} Z_j^2 \mu_j^{(\gamma)}$. Then, $W^{a,\gamma} \in L_2(Q_n)$, $a.s.$.

2.5.2 Concentration Function

With the series representation (2.5.4), we can treat the Gaussian process $W^{a,\gamma}$ as a Borel measurable map into the Banach space $(L_2(Q_n), ||\cdot||_{L_2(Q_n)})$ such that the random variable $b^*(W^{a,\gamma})$ is Gaussian for every continuous, linear map $b^* : L_2(Q_n) \to \mathbb{R}$. Since $W^{a,\gamma}$ is continuous with probability 1, the sample paths under consideration are continuous versions. For any set $U \subset L_2(Q_n)$, the probability $\Pr(W^{a,\gamma} \in U)$ equals the probability $\Pr\left(W^{a,\gamma} \in U \cap C(\mathbb{R}^{d_n})\right)$. In the end, the Banach space where $W^{a,\gamma}$ lives is essentially $\mathcal{L}_2(Q_n) \equiv L_2(Q_n) \cap C(\mathbb{R}^{d_n})$.

The RKHS $\mathcal{H}^{a,\gamma}$ associated with $W^{a,\gamma}$ is the completion of the range $S[\mathcal{L}_2(Q_n)^*]$ where $S[b^*] = \mathbb{E}[W^{a,\gamma} b^*(W^{a,\gamma})]$, $b^* \in \mathcal{L}_2(Q_n)^*$ in the sense of Pettis integral; see [vdVvZ08b] for more details. The corresponding RKHS norm is determined by the inner product $\langle Sb_1^*, Sb_2^* \rangle_{\mathcal{H}^{a,\gamma}} = \mathbb{E}\left(b_1^*\left(W^{a,\gamma}\right) b_2^*\left(W^{a,\gamma}\right)\right)$. Using the same argument as (2.4) of [vdVvZ08b], the RKHS norm is stronger than the $L_2(Q_n)$ norm. Hence, $\mathcal{H}^{a,\gamma}$ is seen as a dense subset of $\mathcal{L}_2(Q_n)$ and can be isometrically identified with an ℓ_2 sequence space. In our case, with the eigen-expansion (2.5.3), the RKHS unit ball

$\mathcal{H}_1^{a,\gamma}$ of $W^{a,\gamma}$ can be isometrically identified with the following ellipsoid:

$$\left\{\{\theta_j\}_{j=1}^{\infty} : \sum\nolimits_{j=1}^{\infty} \theta_j^2/\mu_j^{(\gamma)} \leq 1\right\} \subseteq \ell^2(\mathbb{N}). \tag{2.5.5}$$

This isometry provides a different route to compute the metric entropy of the unit ball of $\mathcal{H}^{a,\gamma}$.

Prior mass calculations in Theorem 2.4.1 and Proposition 1 require both upper and lower bounds for $\Pr(\|W^{a,\gamma} - f_n^*\|_{L_2(Q_n)} < \varepsilon_n)$, the f_n^*-shifted ε_n-ball probability of $W^{a,\gamma}$.

With the above abstract formulation, the small ball probability, in log scale, can be bounded as

$$\phi_{f_n^*}^{a,\gamma}(\varepsilon_n) \leq -\log \Pr(\|W^{a,\gamma} - f_n^*\|_{L_2(Q_n)} < \varepsilon_n) \leq \phi_{f_n^*}^{a,\gamma}(\varepsilon_n/2), \tag{2.5.6}$$

with $\phi_{f_n^*}^{a,\gamma}$ denoting the concentration function

$$\phi_{f_n^*}^{a,\gamma}(\varepsilon_n) = \inf_{h \in \mathcal{H}^{a,\gamma}: \|h - f_n^*\|_{L_2(Q_n)} < \varepsilon_n} \|h\|_{\mathcal{H}^{a,\gamma}}^2 - \log \Pr(\|W^{a,\gamma}\|_{L_2(Q_n)} < \varepsilon_n), \tag{2.5.7}$$

where $\|\cdot\|_{\mathcal{H}^{a,\gamma}}$ is the canonical norm of the Hilbert space $\mathcal{H}^{a,\gamma}$.

With inequality (2.5.6), bounding shifted small ball probability is essentially bounding the concentration function. The concentration function has two parts: the decentering part and the centered small ball probability exponent. The decentering

part measures the position of the centering f_n^* relative to the RKHS.

2.5.3 Centered Small Ball Probability Bounds via Metric Entropy

The centered small ball probability is bounded using the metric entropy method [KL93, LL99, vdVvZ08b]. The metric entropy method links the metric entropy of the RKHS unit ball with the centered small ball probability. Section 6 of the review paper [vdVvZ08b] summarizes the quantitative relationship between bounds of metric entropy and bounds of small ball probability. In our analysis, we first calculate the metric entropy of the RKHS unit ball and then use the relationship to derive bounds of centered small ball probabilities as a corollary.

With isometry, the metric entropy of the RKHS unit ball is the metric entropy of the ℓ_2 ellipsoid (2.5.5). It is well known that the metric entropy of ℓ_2 ellipsoid in the fashion of (2.5.5) depends on the decay rate of $\{\mu_j^{(\gamma)}\}$. Lemma 13 gives bounds for the metric entropy of the RKHS unit ball.

Lemma 1. *Suppose $\mathcal{H}_1^{a,\gamma}$ is the RKHS unit ball associated to the GP SEGP($\mathbb{R}^{d_n}; \gamma, \delta_a$) with the design measure $Q_n \equiv N(0, \xi^2 I_{d_n})$, constants v_1 and V are defined in (2.5.2), constants a, ε and $|\gamma|$ satisfy $\varepsilon^{-2} \geq C_H(a\xi)^{|\gamma|}$ for some constant C_H such that $\log\left(\frac{1}{\varepsilon}\right) - \frac{|\gamma|}{4}\log\left(\frac{V}{2v_1}\right) \asymp \log\left(\frac{1}{\varepsilon}\right)$, and $a\xi\log(1/\varepsilon) > |\gamma|$, then the metric entropy of*

$\mathcal{H}_1^{a,\gamma}$ *satisfies*

$$\log N\left(\mathcal{H}_1^{a,\gamma}, \varepsilon, \|\cdot\|_{L_2(Q_n)}\right) \precsim a^{|\gamma|}\log(1/\varepsilon)^{|\gamma|+1}/|\gamma|!,$$
$$\log N\left(\mathcal{H}_1^{a,\gamma}, \varepsilon, \|\cdot\|_{L_2(Q_n)}\right) \succsim a^{|\gamma|}\log(1/\varepsilon)^{|\gamma|}/|\gamma|!.$$

The proof of Lemma 13 follows standard technique of metric entropy of ℓ_2 sequence spaces. The lower bound is smaller than the upper bound by a logarithmic factor. This gap affects the bounds for small ball probabilities, but it does not jeopardize Theorem 2.4.1 and Proposition 1.

Then the centered small ball probability bounds are obtained in Lemma 2 as a corollary.

Lemma 2. *Suppose $\phi_0^{a,\gamma}$ is the concentration function associated to the Gaussian process SEGP($\mathbb{R}^{d_n}; \gamma, \delta_a$) with the design measure $Q_n \equiv N(0, \xi^2 I_{d_n})$, constants a and ε satisfy $\varepsilon^{-2} \geq C_H(a\xi)^{|\gamma|}$ and $a\xi\log(1/\varepsilon) > |\gamma|$, then there exists a constant C independent of a, ξ and $|\gamma|$ such that*

$$\phi_0^{a,\gamma}(\varepsilon) \leq Ca^{|\gamma|}\log(a/\varepsilon)^{|\gamma|+1}/|\gamma|!,$$

and there exists a constant C' such that

$$\phi_0^{a,\gamma}(\varepsilon) \geq C'a^{|\gamma|}\log(1/\varepsilon)^{|\gamma|}/|\gamma|!.$$

Proof. With the assumptions, the metric entropy calculation in Lemma 13 holds.

Following the same idea as in Lemma 4.6 of [vdVvZ09], the first assertion holds.

Using the first inequality of Lemma 2.1 in [AILvZ09], let $\lambda = 2$ in the Lemma, $\phi_0^{a,\gamma}(\varepsilon_n) \geq \log N\left(\mathcal{H}^{a,\gamma}, \varepsilon_n, \|\cdot\|_{L_2(Q_n)}\right) - 2 \geq C' a^{|\gamma|} \log(1/\varepsilon)^{|\gamma|}/|\gamma|!$ holds for some constant C'.

□

2.5.4 Shifted Small Ball Probability Estimates

With centered small ball probability calculation, bounds for shifted small ball probability are readily available if the bounds of the decentering part are obtained. Lemma 3 and Lemma 4 give upper and lower bounds of the decentering part as a function of rescaling level and model index parameter. Lemma 3 is used to *lower* bound the prior mass on $\{\|W^{a,\gamma_n^*} - f_n^*\|_{L_2(Q_n)} < \varepsilon_n\}$; Lemma 4 is used to *upper* bound the prior mass on $\{\Gamma \in \mathrm{FP}(\gamma_n^*)\} \cap \{\|W^{a,\Gamma} - f_n^*\|_{L_2(Q_n)} < \varepsilon_n\}$.

Lemma 3. *Suppose $\mathcal{H}^{a,\gamma_0}$ is the RKHS of the GP SEGP($\mathbb{R}^{d_n}; \gamma_n^*, \delta_a$) with the design measure $Q_n \equiv N(0, \xi^2 I_{d_n})$, $\gamma_n^* \in \{0,1\}^{d_n}$ encodes the indices of true variables, $d_0 \equiv |\gamma_n^*|$, $f_0 \in H^\beta(\mathbb{R}^{d_0}) \cap L_2(Q_0)$, $f_n^* = T_{\gamma_n^*} f_0$, then for every $a > 0$, there exists a constant C, ε_0 such that for all $\varepsilon < \varepsilon_0$*

$$\inf_{h \in \mathcal{H}^{a,\gamma_0} : \|h - f_n^*\|_{L_2(Q_n)} < \varepsilon} \|h\|^2_{\mathcal{H}^{a,\gamma_0}} \leq C\left(2\sqrt{\pi}\right)^{d_0} a^{d_0} e^{C\varepsilon^{-2/\beta}/a^2}.$$

The proof of Lemma 3 leverages the representation theorem of the RKHS and

uses the squared RKHS norm of a special element in the true model neighborhood as an upper bound. In light of Lemma 4.1 of [vdVvZ09] and Lemma 7.1 of [vdVvZ08b], after embedded into $L_2(Q_n)$, elements in $\mathcal{H}^{a,\gamma}$ admit the representation: for $t \in \mathbb{R}^{d_n}$, $h_\psi(t) = \int_{\mathbb{R}^{|\gamma|}} e^{i(\lambda,t_0)}\psi(\lambda)\, m_{a,\gamma}(\lambda)\, d\lambda$, where $t \equiv (t_0, t_1)$ with $t_0 \in \mathbb{R}^{|\gamma|}$ and $t_1 \in \mathbb{R}^{d_n}\backslash\mathbb{R}^{|\gamma|}$, $m_{a,\gamma}(\cdot)$ is the spectral density of the $\gamma-$dimensional Gaussian process with rescaling level a, and ψ is in the complex Hilbert space $L_2(m_{a,\gamma})$; its RKHS norm is defined as $\|h_\psi\|^2_{\mathcal{H}^{a,\gamma}} = \int_{\mathbb{R}^{|\gamma|}} |\psi(\lambda)|^2 m_{a,\gamma}(\lambda)\, d\lambda$. The spectral density $m_{a,\gamma}(\lambda) = a^{-|\gamma|} m_\gamma(\lambda/a)$, where $m_\gamma(\lambda) = e^{-\|\lambda\|_2^2/4} / \left(2^{|\gamma|}\pi^{|\gamma|/2}\right)$ and $\lambda \in \mathbb{R}^{|\gamma|}$.

Lemma 4. *Suppose $\mathcal{H}^{a,\gamma}$ is the RKHS of the GP SEGP$(\mathbb{R}^{d_n}; \gamma, \delta_a)$ with the design measure $Q_n \equiv N(0, \xi^2 I_{d_n})$, $\gamma_n^* \in \{0,1\}^{d_n}$ encodes the indices of true variables, $d_0 \equiv |\gamma_n^*|$, $f_0 \in H^\beta(\mathbb{R}^{d_0}) \cap L_2(Q_0)$ with Fourier transform satisfying $|\widehat{f_0\sqrt{g_{d_0}}}(\lambda)| \gtrsim \|\lambda\|_2^{-(\alpha+d_0/2)}$ for all $\|\lambda\|_2 \geq 1$ and some constant $\alpha > 0$, $f_n^* = T_{\gamma_n^*} f_0$, and $c_\xi \equiv \xi/\sqrt{2}$. Pick a $\gamma \in \mathrm{FP}(\gamma_n^*)$, then for all $a > 0$, there exist constants C, C' and ε_0 only dependent of f_0, such that for all $\varepsilon < \varepsilon_0$,*

$$\inf_{h \in \mathcal{H}^{a,\gamma}: \|h - f_n^*\|_{L_2(Q_n)} < \varepsilon} \|h\|^2_{\mathcal{H}^{a,\gamma}} \geq C\varepsilon^2 (c_\xi a)^{|\gamma|} e^{C'\varepsilon^{-2/\alpha}\left(\xi^2 \wedge a^{-2}\right)}.$$

The strategy of the proof follows the proof of Theorem 8 in [vdVvZ11]. In Lemma 4, the RKHS norm is expected to explode as the neighborhood in the false positive model space shrinks. Also note the effect of rescaling level a: larger a means better approximation and hence larger RKHS norm; small a means flatness in the prior sample path and hence smaller RKHS norm.

2.6 Discussion

We have shown here that a GP regression model equipped with stochastic variable selection can simultaneously offer adaptive regression function estimation and consistent recovery of its sparsity pattern. This result is derived under several assumptions, some of which are mathematical formalizations of reasonable statistical considerations while others are needed more for technical reasons than statistical ones. Below we offer a detailed discussion of the reasonability and limitations of the formal assumptions and explore possible relaxations.

2.6.1 Gaussian design assumption

Assumption 1 requires a Gaussian random design and is quite restrictive, but is needed for a specific technical reason: it leads to a Karhunen-Loève eigen expansion of the squared exponential kernel in closed form. With the Karhunen-Loève expansion, the RKHS unit ball is isometric to an ℓ_2 sequence space whose metric entropy can be accurately bounded. Further, small ball probability estimates are available via the metric entropy method [KL93, LL99, vdVvZ08b]. In particular, it permits an upper bound of the prior mass condition that is crucial for Proposition 1. Also, the assumption $\xi^2 > 2/e$ is made to simplify the calculations for Proposition 1. It holds in most applications as the design matrix can be standardized.

Without the Gaussian assumption, it is not tractable to work out sharp upper bounds of the prior mass condition needed for Proposition 1. An alternative approach

to metric entropy calculation of the RKHS unit ball without assuming Gaussian random design is to extend [AILvZ09] where the RKHS unit ball is identified as a set of "well-behaved" entire functions whose metric entropy can be accurately bounded. However, direct extension carrying over the rescaling parameter gives sub-optimal lower bounds of the metric entropy, and the resulting upper bounds of the prior mass condition become meaningless.

A natural relaxation is to assume the Radon-Nikodym derivative of Q_n with respect to G_{d_n} is bounded away from 0 and ∞, uniformly across all $n \geq 1$. This uniform absolute continuity implies that convergence in $\|\cdot\|_{L_2(Q_n)}$ is equivalent to convergence in $\|\cdot\|_{L_2(G_{d_n})}$. Clearly, the uniform boundedness assumption renders the relaxation quite limited.

2.6.2 Fixed sparsity assumption

Assumption 2 makes it clear that for all large n, the true function is sparse and the support size d_0 does not grow with sample size. While the latter condition may appear too restrictive, it is shown in [YT15] that in the case d_n grows nearly exponentially in n, one cannot hope to consistently estimate a sparse, smooth regression function nonparametrically unless the true support size remains essentially constant.

Assumption 3 identifies the true sparsity level d_0 as the minimal support size for the sequence $(f_n^* : n \geq 1)$. Under this assumption, each of the d_0 inputs to f_0 results in a variability that is detectable in the L_2 topology. This is essentially a

nonparametric version of the β-min assumption in sparse linear regression literature.

2.6.3 Nonparametric beta-min condition

Assumption 3 requires the minimal signal strength of relevant regressors. In particular, (1) the signal strength is measured in L_2; (2) the minimal strength is lower bounded by a strictly positive constant δ. By inspecting the proof of our main result, the constant δ can be relaxed to δ_n such that $\varepsilon_n \prec \delta_n \precsim 1$. In other words, the GP regression is able to discover the regressors whose signal strengths are stronger than $O(\varepsilon_n)$. In principle, it is debatable that regressors with signal strengths weaker than $O(\varepsilon_n)$ are considered as "relevant".

2.6.4 Limited smoothness assumption

Assumption 4 imparts only a limited amount of smoothness on f_0. The first part of the assumption requires true f_0 to be in the Sobolev space $H^\beta(\mathbb{R}^{d_0})$ in which functions satisfy (2.3.1). Analogous to Hölder smoothness, functions in $H^\beta(\mathbb{R}^{d_0})$ have square-integrable (weak) partial derivatives up to order β. The second part of Assumption 4 is adapted from [vdVvZ11] and combines the probability density of the random design. It encodes a lower bound for the regularity of f_0 in the spirit of the self-similarity assumption in [GN10, B$^+$12, SvdVvZ15, Ray17]. A direct consequence of the assumption is lower bounds for the RKHS norm of the functions in a L_2 neighborhood of the true function (Lemma 4). Further, the resulting lower

bounds are necessary for upper bounding the prior mass condition in Proposition 1. The self-similarity assumption in the Gaussian sequence model can be written as the convolution of $\hat{f}_0$ and a sum of delta functions evaluated at some frequencies. This convolution reflects the fixed design nature of Gaussian sequence model. In contrast, with random design, it is reasonable to weigh "signals" at different frequencies differently.

Note that the decay rate of a convolution is bounded above by the sum of the decay rates of the functions convolved when they are in both L_1 and L_∞. As $\widehat{\sqrt{g_{d_0}}}$ is a Gaussian function decaying exponentially, $\widehat{\sqrt{g_{d_0}}} \in L_1 \cap L_\infty$. Part 1 implies $\hat{f}_0 \in L_1 \cap L_\infty$. Therefore, part 2, coupled with part 1, requires $|\widehat{\hat{f}_0\sqrt{g_{d_0}}}(\lambda)|$ to decay approximately at the same rate as $|\hat{f}_0(\lambda)|$ as $\|\lambda\| \to \infty$, that is, $|\hat{f}_0(\lambda)| \gtrsim \|\lambda\|_2^{-(\alpha+d_0/2)}$ as $\|\lambda\| \to \infty$.

One concern is whether there exists an f_0 satisfying Assumption 4. The answer is yes. As an example, let $\hat{f}_0(\lambda) = (1 + \|\lambda\|)^{-r}$ with $r = \alpha + d_0/2 > \beta + d_0/2$. Then $f_0 \in H^\beta(\mathbb{R}^{d_0})$ and $\widehat{f_0\sqrt{g_{d_0}}}(\lambda) = \hat{f}_0 * \widehat{\sqrt{g_{d_0}}}(\lambda) \geq \int_{\|t\|\leq 1} \hat{f}_0(\lambda - t)\widehat{\sqrt{g_{d_0}}}(t)dt \gtrsim (2 + \|\lambda\|)^{-r} \asymp \|\lambda\|_2^{-(\alpha+d_0/2)}$. More such functions could be constructed as long as the tail decay rate of $\hat{f}_0$ is maintained.

The limited smoothness assumption rules out true functions that are infinitely differentiable. This assumption is key to our proof strategy which exploits a polynomial slowdown in posterior contraction rate when spurious predictors are included in the model. Similar rate slowdown also manifests for infinitely smooth true functions,

but the depreciation is only logarithmic [vdVvZ09]. While our proof strategy fails to exploit such subtler rate drops, variable selection consistency may still hold in these cases. Indeed, for Bayesian linear regression with variable selection, posterior contraction rates slow down only by a multiplicative constant when spurious variables are included, and yet variable selection consistency holds in such cases.

Curiously, the results presented in this paper could be applied to construct a modified nonparametric regression model with guaranteed variable selection consistency, no matter the level of smoothness of the true function. One can augment the design matrix with an additional, independent Gaussian "covariate" Z, and create a new response $Y_i' = Y_i + g_0(Z_i)$ where g_0 is a known function of limited smoothness. Now, for the new design (X, Z) and the modified response data Y', the underlying true function has limited smoothness, and hence our results guarantee asymptotically accurate recovery of the important coordinate variables of X in addition to the the synthetic variable Z. However, the resulting posterior convergence rate may be slower than optimal if f_0 were smoother than g_0.

2.6.5 Restricted design dimension growth assumption

Assumption 6 restricts the applicability of our result only up to a design size $\log(d_n) \precsim n^{d_0/(2\beta+d_0)}$. Consequently, we are restricted to $\log(d_n) \ll n$ when β is large. This is a serious limitation because as shown in [CD12], with true sparsity fixed at d_0, variable selection consistency should hold for larger design sizes, up to the limit

$\log(d_n) = O(n)$. Notice however that we prove results for a Bayesian estimation method that simultaneously infers the sparsity patter Γ and the regression function f, and offers a near minimax optimal estimation of the latter by adapting to the true smoothness level. It is this adaptation that imposes the stricter bound on d_n, beyond which the estimation error rate is dominated by the variable selection penalty which does not worsen polynomially between correct selection and false positive selections.

It is unclear at the moment whether the Bayesian estimation model studied here, or any other model which offers smoothness adaptive function estimation, could actually achieve variable selection consistency with d_n growing faster than our bound. However, if one were to sacrifice on adaptive function estimation, variable selection consistency may be achieved even when $n^{d_0/(2\beta+d_0)} \ll \log(d_n) \precsim n$ under a variation of our GP regression model where the random rescaling component (Assumption 5) is replaced with a dimension-specific deterministic rescaling

$$\pi_{n,d}(a) = \mathbb{1}(a = n^{1/(2\underline{\beta}+d)}),$$

for every $n \in \mathbb{N}$, $d \in \{1, \ldots, d_n\}$, where $\underline{\beta}$ is a fixed, small positive scalar. This modification grants variable selection consistency up to design dimensions d_n with $\log(d_n) \precsim n^{d_0/(2\underline{\beta}+d_0)}$ as long as the true smoothness $\beta > \underline{\beta}$. However, the posterior contraction rate under the deterministically rescaled GP prior is $O(n^{-\underline{\beta}/(2\underline{\beta}+d_0)})$ (up to a logarithmic factor of n), which could be much slower than the optimal rate $n^{-\beta/(2\beta+d_0)}$. That is, in the extreme case $\log(d_n) \precsim n$, we can pick a small positive $\underline{\beta}$

for the deterministic rescaling to guarantee variable selection consistency at the cost of estimation accuracy.

2.6.6 Separating variable selection from estimation

The discussion in the above two subsections indicates that one could gain on variable selection consistency by sacrificing on optimal function estimation. This leads to a much broader question. Should the two inference goals, optimal function estimation and consistent variable selection, be separated, and perhaps approached in a two-stage manner? We believe the answer is *yes*. A two-stage approach may sound contradictory to our adopted position of Bayesian modeling and inference. But in reality, the Bayesian paradigm, when augmented with decision theoretic considerations, may offer an incredibly fertile ground for exploring a rigorous two-stage approach. For example, one may adapt [HC15] and estimate γ by the formal Bayes estimate $\hat{\gamma}_B$ under the loss function $L((f,\gamma),(\hat{f},\hat{\gamma})) = \|f - \hat{f}\|^2_{L_2(Q_n)} + \lambda J(\hat{\gamma})$, for some model complexity penalty function $J(\gamma)$ and penalty tuning parameter $\lambda > 0$. The Bayes estimate may be computed as $\hat{\gamma}_B = \arg\min_\gamma \{\|\bar{f} - \bar{f}_\gamma\|^2_{L_2(Q_n)} + \lambda J(|\gamma|)\}$ where $\bar{f}$ is the posterior mean of f and $\bar{f}_\gamma = \int \bar{f}(x) \prod_{j:\gamma_j=0} G_1(dx_j)$ is the projection of $\bar{f}$ along γ under $L_2(Q_n)$. Whether such computations are feasible and result in consistent estimation of γ remain to be seen. It seems likely that the additional penalty component may help resolve the true model from false positive models, even if the posterior weights assigned to the latter were not polynomially smaller.

Besides potential theoretical gains, such a Bayesian decision theoretic approach is appealing on philosophical grounds alone. Any statistical analysis may have any number of inference goals and no single estimation method may be universally optimal. For example, cross-validation based LASSO is known to offer great prediction accuracy while suffering from enhanced false detection [BVDBS$^+$15]. A Bayesian decision theoretic approach may help unify such disparate goals where the analyst fits a single model to the data, but produces different estimates for different quantities by utilizing appropriately chosen loss function for each task. For example, in the adaptation described above, one may still report the posterior mean $\bar{f}$ as the Bayes estimate of f (under the integrated square loss) while producing a sparse estimate $\hat{\gamma}_B$ for variable selection according the loss function discussed above. It will be exciting to carry out rigorous posterior asymptotic behavior of such formal Bayes estimates.

Chapter 3

Consistent Bayesian Community Detection for Assortative Networks

3.1 Introduction

Community detection is a fundamental goal of many statistical analyses of network data [GN02, YAT16, Abb17]. To determine the number of communities, various tests have been constructed based on modularity [ZLZ11], random matrix theory [BS16, Lei16], and likelihood ratio [WB17]. Methods based on information criteria [SYF17] and network cross-validation [CL18, LLZ20] have also been designed. In the Bayesian realm, a stochastic block model (SBM) is often employed to jointly infer the number of communities, the connectivity probability matrix, and the membership assignment [NS01, MMFH13, GBP19].

Despite clear empirical evidence of good statistical performance [MMFH13, GBP19, ZP20], few theoretical guarantees are available regarding community detection of Bayesian SBMs when the number of communities is *unknown*. As a few exceptions, [GBP19] and [vWK20] show that the community count may be consistently estimated under the restrictive assumptions of a homogeneous SBM. It is unclear if

their calculations generalize to more realistic scenarios, for which the best theoretical result known currently states that the posterior probability of community count concentrates at or below the truth [GvdVZ20].

We examine and establish exact asymptotic recovery in community detection for the special sub-class of *weakly assortative* SBMs. Weak assortativity offers a stronger encoding of the notion of *communities* in networks than general SBMs in the sense that nodes within the same community are *strictly* more likely to connect with each other than with nodes from other communities. Despite its limitations, assortativity is taken to be a salient feature for many real world social and biological networks; see, for example, [New02, New03, ZP20] and the references therein. Efficient algorithms have been developed in the context of assortative mixed-membership SBM [GGF$^+$12, LAW16]. [AL$^+$18] formally introduce weak assortativity under which they show exact recovery for semi-definite programming relaxations for SBMs with equal sized communities.

Crucially, the assortativity assumption enables an exact mathematical recovery of the community structure from the node-wise connectivity probabilities, as long as each community contains at least two nodes. Of course, the node-wise connectivity probability matrix is estimated from data with statistical error. But as long as it is sufficiently "close" to the truth, it is still possible to precisely recover the membership allocation and the community count. We investigate a Bayesian estimation of a weakly assortative SBM under a modified Nowicki-Snijders prior [NS01], and

establish that the posterior on the node-wise connectivity matrix contracts to the truth in the sup-norm topology. Posterior contraction under sup-norm is necessary to the identification strategy detailed above. [GPB19, GvdVZ20] establish (near) minimax optimal posterior contraction rates in the L_2 norm. However, posterior contraction in L_2 or other norms that are weaker than the sup-norm do not grant the identification of the number of communities or the membership assignment from node-wise connectivity probabilities. Our sup-norm posterior contraction calculation applies the Schwartz method [GGvdV00, GvdV07, GvdV17]. The key observation is that the sup-norm is dominated by the Hellinger distance in the special context of SBMs, so the tests required by the Schwartz method exist.

The theoretical gains of the weakly assortative SBMs come at the price of losing conjugacy with respect to the original Nowicki-Snijders prior. But posterior computation may be carried out with a reasonably efficient reversible-jump Markov chain Monte Carlo (MCMC) algorithm based of the allocation sampler in [MMFH13]. Results from extensive numerical studies show that our Bayesian weakly assortative SBM offers comparable and competitive statistical performance against various alternatives in estimating the community count and membership assignment.

3.2 The weakly assortative stochastic block model

Suppose an $n \times n$ binary adjacency matrix A is observed, with entry $A_{ij} = 1$ if node i and node j are connected and $A_{ij} = 0$ otherwise. The stochastic block

model (SBM) assumes there are $K \in \mathbb{Z}_+$ communities among the n nodes and the connection between nodes *exclusively* depends on their community membership. The community assignment Z partitions nodes $\{1, ..., n\}$ into K non-empty groups and assigns each node with a community label. Let the community-wise connectivity probability matrix be $P \in [0, 1]^{K \times K}$. Then,

$$A_{ij}|Z \overset{ind}{\sim} Ber(P_{Z(i)Z(j)}) \text{ for } 1 \leq i < j \leq n, \tag{3.2.1}$$

and $P(A_{ii} = 0|Z) = 1$ for $i \in \{1, ..., n\}$, assuming no self-loops. We denote the above SBM model as $SBM(Z, P, n, k)$. Due to its simplicity and expressiveness, SBM and its variants are fundamental tools for community detection [e.g., KN11, ABFX08, Pei14].

To measure the precision of community detection, this paper focuses on "exact recovery" which is defined as

$$\mathbb{P}(\mathcal{A}(\hat{Z}, Z) = 1) = 1 - o(1) \tag{3.2.2}$$

for some estimator $\hat{Z}$, where $\mathcal{A}(x, y) = \max_\pi \frac{1}{n} \sum_{i=1}^n 1(x_i = \pi(y_i))$ and the maximum is taken over all relabelings of y [Abb17].

3.2.1 Bayesian SBM with conjugate priors

For Bayesian estimation of the SBM, [NS01] propose the following conjugate prior: given K,

$$\begin{aligned} P_{ab} &\overset{iid}{\sim} U(0,1), a, b = 1, ..., K \\ Z_i &\overset{iid}{\sim} MN(\pi), i = 1, ..., n \\ \pi &\sim Dir(\alpha). \end{aligned} \tag{3.2.3}$$

This prior is widely used and adapted to more complicated cases in the Bayesian SBM literature [GPB19, vdPvdV18, GBP19, MMFH13].

For the unknown K case, to maintain conjugacy, it is natural to place a Poisson prior on K [MMFH13, GBP19]. With conjugacy, [MMFH13] marginalize out P from the posterior $\Pi_n(Z, K, P|A)$ and develop an efficient "allocation sampler" to directly sample from $\Pi_n(Z, K|A)$; [GBP19] adapt the idea of mixture of finite mixture (MFM) sampler of [MH18] to the SBM case: marginalize out K from the posterior $\Pi_n(Z, K, P|A)$, and develop a Gibbs sampler sampling from $\Pi_n(Z, P|A)$.

3.2.2 Bayesian weakly assortative SBM

In this paper, we propose to modify the conjugate specification of Nowicki and Snijders' prior on the connectivity matrix P by imposing a weak assortativity constraint. The constraint is imposed in two steps: first specify a prior distribution for the diagonal entries of P, then conditional on the diagonal entries, specify a prior distribution

on the off-diagonal entries such that the off-diagonal entries are strictly less than their corresponding diagonal entries.

For instance, we specify the following prior:

$$\begin{aligned}
&P_{aa}|K,\delta \overset{iid}{\sim} U(\delta,1], a \in \{1,...,K\},\\
&P_{ab}|K,\delta,\{P_{aa}\}_{a\in\{1,...,K\}} \overset{ind}{\sim} U(0, P_{aa}\wedge P_{bb}-\delta), a<b \in \{1,...,K\},\\
&\delta \propto \sqrt{\log(n)/n},\\
&K \propto Pois(1),
\end{aligned} \tag{3.2.4}$$

where the hyperparameter δ is chosen to be a deterministic sequence that goes to 0 as the network size grows to infinity. Uniform distributions and Poisson distribution in (3.2.4) are used for simplicity and can be replaced with other distributions.

In contrast to the Nowicki and Snijders' priors, our prior specification directly imposes conditional dependence between diagonal entries and off-diagonal entries. The dependence requires weak assortativity in the posterior distribution of connectivity matrix and matches the idea of "community" at the price of losing conjugacy.

The modification is mainly for two reasons. First, the prior constraint of weak assortativity offers a neat identification of the number of communities, and allows us to consistently estimate the number of communities and membership. (See more details in section 3.3.1.) Furthermore, the resulting posterior under the modified prior is more interpretable for assortative networks. Though the prior specification following [NS01] is conjugate, off-diagonal entries can be greater than diagonal entries

under the prior, that is, nodes can be more likely to be connected to nodes from other communities than nodes from their own community. Such configurations violate the idea of "community". Consequently, posterior samples of connectivity matrices can violate weak assortativity and are hard to interprete within the framework of SBM.

3.2.3 L_2 minimax rate

As it is a special sub-class of SBM, one may wonder if the weakly assortative SBM (a-SBM) actually solves a simpler community detection problem. To answer this question, we calculate the L_2 minimax rate of estimation for a-SBM and compare it with the minimax rates derived in [GLZ15]. With Definition 4.1 of [AL$^+$18], the a-SBM has the following space of connectivity matrix

$$S_{k,\delta} = \left\{ P \in [0,1]^{k\times k} : P^T = P, P_{ii} > \delta + \max_{j\neq i} (P_{ij}), i \in \{1,...,k\} \right\}, \tag{3.2.5}$$

where $\delta \in [0,1)$ is a separation parameter that captures the signal strength of the a-SBM. The key departure from the general SBM is the weak assortativity constraint: $P_{ii} > \delta + \max_{j\neq i}(P_{ij})$, for all $i \in \{1,...,k\}$. Under this constraint, between-community connection probabilities are smaller than within-community connection probabilities by at least a δ amount. The gap is inherited by the node-wise connectivity probability matrix.

With Z denoting the membership assignment, we can define the space for node-

wise connectivity probability matrix:

$$\Theta_{k,\delta} = \left\{ T(ZPZ^T) \in [0,1]^{n\times n} : P \in S_{k,\delta}, Z \in \mathcal{Z}_{n,k} \right\}, \tag{3.2.6}$$

where $\mathcal{Z}_{n,k}$ is the collection of all possible assignment of n nodes into k communities which have *at least two* elements, and $T(M) := M - \text{diag}(M)$ for any square matrix M. The node-wise connectivity probability matrix inherits the structural assumption of weak assortativity. The minimum community size assumption allows recovering community membership from node-wise connectivity probability matrix. It is worthwhile to emphasize that singleton communities are ruled out.

The following L_2 minimax result implies that a-SBM estimation is as difficult as the original SBM estimation problem, as long as the separation parameter δ is shrinking at certain rate. In our calculation, the separation squared (δ^2) is dominated by the "clustering rate" $\log(k)/n$ [GLZ15, GM18, KTV17].

Proposition 2. *For any* $k \in \{1, .., n\}$ *and* $\delta \precsim \sqrt{\log(k)/n}$,

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_{k,\delta}} \mathbb{E}\left[||\hat{\theta} - \theta||_2^2 \right] \asymp \frac{k^2}{n^2} + \frac{\log(k)}{n}. \tag{3.2.7}$$

Proof. The upper bound follows theorem 2.1 of [GLZ15] as the weakly assortative connectivity matrix space is a subset of the unconstrained connectivity matrix space. The lower bound follows the proof of theorem 2.2 of [GLZ15] but their construction violates the weakly assortative constraint. However, a weakly assortative version of

their construction is available. For brevity, we only highlight the differences from the proof in [GLZ15].

For the nonparametric rate, we construct the Q^{ω} matrix of [GLZ15] by $Q^{\omega}_{ab} = Q^{\omega}_{ba} = \frac{1}{2} - \delta - \frac{c_1 k}{n}\omega_{ab}$, for $a > b \in \{1, ..., k\}$ and $Q^{\omega}_{aa} = \frac{1}{2}$, for $a \in \{1, ..., k\}$. The rest of the proof for the nonparametric rate remains the same. Next, for the clustering rate, we construct the Q matrix of [GLZ15] with the following form

$$Q = \begin{bmatrix} D_1 & B \\ B^T & D_2 \end{bmatrix},$$

where $D_1 = \frac{1}{2} I_{k/2}$, B follows the same construction of [GLZ15] except that $B_a = 1/2 - \delta - \sqrt{c_2 \log k/n} \cdot \omega_a$ for $a \in \{1, ..., k/2\}$, $D_2 = (1/2 - \delta - \sqrt{\log k/n}) \cdot 1_{k/2} 1_{k/2}^T + (\delta + \sqrt{\log k/n}) I_{k/2}$. As $\delta \precsim \sqrt{\log(k)/n}$, the KL divergence upper bound remains the same. The rest of the proof for the clustering rate remains the same as the entropy calculation and the volume argument are unaffected.

□

3.3 Consistent Bayesian community detection

3.3.1 Identification Strategy

The first consequence of weak assortativity is that the node-wise connectivity probability matrix spaces of different ranks are non-overlapping. This observation offers

a neat partition of the parameter space by the number of communities.

Lemma 5. *Suppose $k \neq k' \in \mathbb{N}$, then $\Theta_{k,\delta} \cap \Theta_{k',\delta'} = \emptyset$ for any $\delta, \delta' \geq 0$.*

Secondly, with weak assortativity, it is possible to exactly identify the number of communities, the membership of every node and the community-wise connectivity probability matrix from node-wise connectivity probability matrix under mild conditions. A more rigorous statement is presented in Lemma 6. The recovery is based on checking each node's connectivity probabilities with other nodes, as each node is connected with nodes from its own community with the highest probability.

Lemma 6. *Suppose $P \in S_{k,\delta}$ for some constant $\delta > 0$, $\theta = T(ZPZ^T)$ for some $Z \in \mathcal{Z}_{n,k}$, T^{-1} recovers both community assignment Z and connectivity matrix P from θ.*

Proof. Without loss of generality, assume the nodes are ordered by community and we can write $Z = [\mathbf{1}_{n_1}, ..., \mathbf{1}_{n_k}]$ where n_j denotes the number of nodes in community j and $\mathbf{1}_{n_j}$ is a $n \times 1$ vector with entries in the j^{th} block being 1. Therefore, the off-diagonal terms of θ are the off-diagonal terms of ZPZ^T.

Suppose we hope to pin down i^{th} node's community membership. We take i^{th} row of θ and it contains the connectivity probabilities of node i and all other nodes. As $Z \in \mathcal{Z}_{n,k}$ whose minimum community size is two, $\mathcal{C}_i \equiv \{j : \theta_{ij} = \max_{\ell} \theta_{i\ell}\}$ is *exactly* the set of node(s) from the community of node i. If $\mathcal{C}_i$ contains node(s) from other communities, then the connectivity probabilities of node i with those node(s)

are cross-community which are strictly less than the within-community connectity probabiilty of node i, contradicting the construction of $\mathcal{C}_i$. If $\mathcal{C}_i$ misses node(s) from the community of node i, then the connectivity probabilities of node i with those node(s) are within-community which have to match the connectivity probabilities of nodes in $\mathcal{C}_i$. Therefore, by enumerating the above procedure for all rows of θ, Z is identified up to a permutation of columns.

To recover P from θ, it suffices to use Z and plug in corresponding values from θ.

□

In practice, the *exact* knowledge of node-wise connectivity probability matrix is not available. However, the precise recovery in Lemma 6 is possible with the estimated node-wise connectivity probability matrix. This is formalized in Lemma 7. We use sup-norm to characterize the accuracy of the knowledge of node-wise connectivity probability matrix. For any node-wise connectivity matrix θ^0, there exists Z_0 and P^0 such that $\theta^0 = T(Z_0 P^0 Z_0^T)$. Without loss of generality, we can fix the column ordering of Z_0 so that P^0 is consequently defined.

Lemma 7. *Suppose* $\theta^0 = T(Z_0 P^0 Z_0^T)$ *for some* $Z_0 \in \mathcal{Z}_{n,k_0}$, $P^0 \in S_{k_0,\delta}$ *and* $\delta > 0$. *Then,* $\{\theta = T(ZPZ^T) : ||\theta - \theta^0||_\infty \leq r, Z \in \mathcal{Z}_{n,k}, P \in S_{k,\delta}\} = \{T(Z_0 P Z_0^T) : ||P - P^0||_\infty \leq r, P \in S_{k_0,\delta}\}$ *holds for all* $r < \delta/2$.

Proof. Pick any $\theta \in \{\theta = T(ZPZ^T) : ||\theta - \theta^0||_\infty \leq r, Z \in \mathcal{Z}_{n,k}, P \in S_{k,\delta}\}$, define $\mathcal{C}_i = \{j : \theta_{ij} = \max_\ell \theta_{i\ell}\}$; similarly, for θ^0, define $\mathcal{C}_i^0 = \{j : \theta_{ij}^0 = \max_\ell \theta_{i\ell}^0\}$. The statement is equivalent to $\mathcal{C}_i = \mathcal{C}_i^0$ for all $i \in \{1, ..., n\}$ and all θ.

First, note for any $j \in \mathcal{C}_i^0$ and $\ell \in \{1, ..., n\} \backslash \mathcal{C}_i^0$, $\theta_{ij} - \theta_{i\ell} = \theta_{ij} - \theta_{ij}^0 + \theta_{ij}^0 - \theta_{i\ell}^0 + \theta_{i\ell}^0 - \theta_{i\ell} > \delta - 2r > 0$. That is, $\mathcal{C}_i^0$ identifies a set of nodes with higher connectivity probabilities with node i relative to nodes from $\{1, ..., n\} \backslash \mathcal{C}_i^0$. Recall $\mathcal{C}_i$ is the collection of nodes with the highest connectivity probability. Then, $\mathcal{C}_i \subseteq \mathcal{C}_i^0$ for all $i \in \{1, ..., n\}$.

If $\mathcal{C}_i^0$ contains nodes from at least two communities of θ, then there exist $j_1, j_2 \in \mathcal{C}_i^0$, such that $|\theta_{ij_1} - \theta_{ij_2}| > \delta$ as $P \in S_{k,\delta}$. Note for all $j_1, j_2 \in \mathcal{C}_i^0$, $\theta_{ij_1}^0 = \theta_{ij_2}^0$, then it follows $|\theta_{ij_1} - \theta_{ij_2}| = |\theta_{ij_1} - \theta_{ij_1}^0 + \theta_{ij_1}^0 - \theta_{ij_2}^0 + \theta_{ij_2}^0 - \theta_{ij_2}| \leq |\theta_{ij_1} - \theta_{ij_1}^0| + |\theta_{ij_2}^0 - \theta_{ij_2}| \leq 2r < \delta$. Then, the contradiction implies $\mathcal{C}_i = \mathcal{C}_i^0$ for all i. As θ is arbitrary, $\mathcal{C}_i = \mathcal{C}_i^0$ for all $i \in \{1, ..., n\}$ and for all θ. □

3.3.2 Posterior Concentration

To study the asymptotic behavior of the weakly assortative SBM, we make the following assumptions on the prior specification. The prior specification in Assumption 8 and 14 is indexed by n, the number of nodes in the network , and can be interpreted as a sequence of prior distributions.

Assumption 8. *(Prior mass on the parameter space) There exists $\bar{\delta} \in (0, 1)$ such that for all $0 < \delta < \bar{\delta}$ and $k > 1$, $\Pi_n(S_{k,\delta}|K = k) \geq 1 - e^{-n^2\delta^2}$.*

Assumption 8 requires that the prior specification is essentially weakly assortative. Under Nowicki and Snijders' prior, conditional on k communities, the prior probability of weak assortativity is $1/k^k$. Therefore, Nowicki and Snijders' prior does

not satisfy Assumption 8.

Assumption 9. *(Prior decay rates)*

1. *(Prior on P conditional on K and δ)*

   For $a \in \{1, ..., k\}$, diagonal entries $\{P_{aa}\}$ are independent with prior density $\pi_n(P_{aa}|K, \delta) \geq e^{-C\log(n)P_{aa}} 1_{\{P_{aa} \in (\delta, 1)\}}$ for some positive constant C independent of $a \in \{1, ..., k\}$.

   For $a < b \in \{1, ..., k\}$, off-diagonal entries $\{P_{ab}\}_{a \in \{1,...,k\}}$ are conditionally independent on diagonal entries with conditional prior density

$$\pi_n\left(P_{ab}|\{P_{aa}\}_{a\in\{1,...,k\}}, \delta, K\right) \geq e^{-C\log(n)(P_{aa}\wedge P_{bb})} 1_{\{P_{ab}\in[0, P_{aa}\wedge P_{bb}-\delta]\}} \tag{3.3.1}$$

   for some positive constant C independent of $a, b \in \{1, ..., k\}$.

2. *(Prior on Z conditional on K)*

   The prior on the membership assignment Z satisfies $\Pi_n(Z = z|K = k) \geq e^{-Cn\log(k)}$ for all $z \in \mathcal{Z}_{n,k}$ and for some universal positive constant C.

3. *(Prior on K)*

   The support of K is $[K_n]$ with $K_n \precsim \sqrt{n}$. For $k \in [K_n]$, the prior on K satisfies $\Pi_n(K = k) \geq e^{-Ck\log(k)}$ for some universal positive constant C.

Assumption 14 makes more specific decay rate assumptions on the prior mass of connectivity matrix P, the assignment Z, and the number of communities K. The

rate assumption of the prior on P given K and δ essentially requires the prior density on P is lower bounded away from 0. For instance, the uniform prior on P and the Poisson prior on K in (3.2.4) satisfy Assumption 14.

Theorem 3.3.1. *Suppose adjacency matrix $A \sim SBM(Z_0, P^0, n, k_0)$, let $\theta^0 = T(Z_0 P^0 Z_0^T)$, $P^0 \in \Theta_{k_0,\delta_0}$ for some $k_0 \precsim \sqrt{n}$ and $\sqrt{\log(k_0)/n} \prec \delta_0 \prec 1$, and the number of zero and one entries of θ^0 is at most $O(n^2\varepsilon_n)$ where $\varepsilon_n^2 \asymp \frac{\log(k_0)}{n}$. The prior Π_n satisfies Assumption 8 and 14. Then, for all sufficiently large M,*

$$\mathbb{P}_{0,n}\Pi_n\left(\theta : ||\theta - \theta^0||_\infty \geq M\varepsilon_n | A\right) \to 0.$$

The proof of Theorem 3.3.1 follows Schwartz method [Sch65, BSW99, GGR99, GvdV07]. Details of the proof are deferred to Section 3.3.3.

As L_∞ minimax rates of SBM or a-SBM are unknown, it is unclear if the posterior contraction rate in Theorem 3.3.1 is minimax-optimal or not. The L_2 minimax rate calculation in Proposition 2 can be suggestive on the sharpness of the posterior contraction rate in L_∞. However, the conditions on the separation parameter differ: Proposition 2 requires $\delta_0 \precsim \sqrt{\log(k_0)/n}$, while Theorem 3.3.1 requires $\delta_0 \succ \sqrt{\log(k_0)/n}$. As suggested by Lemma 7, our identification strategy relies on sufficient separation measured by posterior contraction rates.

Despite the potentially sub-optimal posterior contraction rates, we are able to establish exact recovery for community detection. The main result is summarized as follows.

Theorem 3.3.2. *Under the same assumptions of Theorem 3.3.1,*

$$\mathbb{P}_{0,n}\left[\Pi_n\left(\{K = k_0\} \cap \{Z = Z_0\}|A\right)\right] \to 1.$$

Proof. In light of Theorem 3.3.1, the posterior mass is essentially on $\{\theta : ||\theta - \theta_n^0||_\infty \leq \varepsilon_n\}$. Therefore, we leverage Lemma 7 to identify k_0 and Z_0 on the set.

Define $E_0 = \{K = k_0\} \cap \{Z = Z_0\}$. Note the decomposition

$$E_0^c = \left(E_0^c \cap \{||\theta - \theta^0||_\infty \leq \varepsilon_n\}\right) \cup \left(E_0^c \cap \{||\theta - \theta^0||_\infty > \varepsilon_n\}\right)$$

for some ε_n, then

$$\Pi_n\left(E_0^c|A\right) \leq \Pi_n\left(E_0^c, ||\theta - \theta^0||_\infty \leq \varepsilon_n|A\right) + \Pi_n\left(||\theta - \theta^0||_\infty > \varepsilon_n|A\right) \tag{3.3.2}$$

where $\varepsilon_n \to 0$ is chosen to match the posterior contraction rate in sup-norm.

Then, the posterior probability of choosing wrong number of communities or wrong membership assignment can be upper bounded via the identification assumption and convergence of the posterior distribution of θ. For the first part of Equation (3.3.2), the δ separation assumption of θ^0 satisfies $\delta_0 \succ \varepsilon_n$. Then, by Lemma 7, for all sufficiently small ε_n, $\{||\theta - \theta^0||_\infty \leq \varepsilon_n\}$ is the same as its Z_0 slice where the implied number of communities is k_0.

For the second part, Theorem 3.3.1 implies $\mathbb{P}_0[\Pi_n\left(||\theta - \theta^0||_\infty > \varepsilon_n|A\right)] \to 0$.

□

Theorem 3.3.2 states that the posterior mass on Z_0 converges to 1 in probability as sample size grows to infinity. Recall the definition of exact recovery requires $\mathbb{P}(\mathcal{A}(\hat{Z}, Z_0) = 1) = 1 - o(1)$. Posterior mode of Z (up to some column permutations) is the Bayesian estimator and achieves exact recovery asymptotically.

A weaker recovery concept "almost exact recovery" requires $\mathbb{P}(\mathcal{A}(\hat{Z}, Z_0) = 1 - o(1)) = 1 - o(1)$. Estimators slightly different from the posterior mode can achieve almost exact recovery as long as the misclassification rate $1 - \mathcal{A}(\hat{Z}, Z_0) = o(1)$.

3.3.3 Proof of Theorem 3.3.1

Pioneered by [Sch65] and further developed by [BSW99, GGR99, GvdV07], Schwartz method is the major tool to study posterior concentration properties of Bayesian procedures as sample size grows to infinity [GvdV17]. Schwartz method seeks for two sufficient conditions to guarantee posterior concentration: the existence of certain tests and prior mass condition. The existence of certain tests often reduces to the construction of certain sieves and an entropy condition associated with the sieve, if the metric under which we wish to obtain posterior contraction is dominated by Hellinger distance. The prior mass condition requires sufficient amount of prior mass on some KL neighborhood near the truth.

Establishing convergence in $||\cdot||_\infty$ via the general framework of Schwartz method requires $||\cdot||_\infty$ to be dominated by Hellinger distance. In general, $||\cdot||_\infty$ is (weakly)

stronger than Hellinger distance and not dominated by Hellinger distance. However, in the special case of SBM, the parameter space is constrained and the desired dominance holds. This observation is shown in Lemma 8.

Lemma 8. *Suppose* $A_{ij}|\theta \overset{IND}{\sim} Ber(\theta_{ij})$ *for* $i < j$ *and* $i, j \in \{1, ..., n\}$*, then* $||\cdot||_\infty$ *is dominated by Hellinger distance:* $||\theta^0 - \theta^1||_\infty \leq 2H\left(\mathbb{P}_{\theta^0}, \mathbb{P}_{\theta^1}\right)$.

With the norm dominance, the existence of certain tests reduces to construct a suitable sieve which charges sufficient prior mass and whose metric entropy is under control. In our proof, the sieve is constructed as the set of all well separated node-wise connectivity probability matrices: $\bigcup_{k=1}^{K_n} \Theta_{k,\delta_n}$ for some carefully chosen δ_n and K_n.

In light of Lemma 5, the metric entropy of the sieve can be neatly bounded. The entropy calculation is summarized in Lemma 9.

Lemma 9. *Suppose* $\varepsilon_n \to 0$ *as* $n \to \infty$*, and* $\varepsilon_n \precsim \delta_n$*, then metric entropy satisfies*

$$\log N\left(\varepsilon_n, \bigcup_{k=1}^{K_n} \Theta_{k,\delta_n}, ||\cdot||_\infty\right) \precsim (n+1)\log K_n + \frac{1}{2}K_n\left(K_n+1\right)\log\left(1/\varepsilon_n\right). \quad (3.3.3)$$

The prior mass condition in terms of KL divergence can be reduced to a prior mass condition in terms of $||\cdot||_\infty$ norm. This observation is summarized in Lemma 10.

Lemma 10. *The observation model is* $A_{ij}|\theta^0 \overset{IND}{\sim} Ber(\theta^0_{ij})$ *for* $i < j$ *and* $i, j \in \{1, ..., n\}$*. Suppose* $C_0 = \min_{i<j:0<\theta^0_{ij}<1} \theta^0_{ij}\left(1-\theta^0_{ij}\right) > 0$*, and the number of zero and*

one entries of θ^0 *is less than* $O(n^2\varepsilon_n)$ *for some* $\varepsilon_n \to 0$ *such that* $n^2\varepsilon_n \to \infty$. *If* $||\theta - \theta^0||_\infty \leq \varepsilon_n$, *then* $KL\left(\mathbb{P}_{\theta^0}, \mathbb{P}_\theta\right) \precsim C_0^{-1} n^2 \varepsilon_n^2$, *and* $V_{2,0}\left(\mathbb{P}_{\theta^0}, \mathbb{P}_\theta\right) \precsim C_0^{-1} n^2 \varepsilon_n^2$.

Lemma 10 simplifies the prior mass condition to element-wise probability calculation. Immediately with Assumption 14, we obtain the following prior mass calculation.

Lemma 11. *Suppose* $P^0 \in S_{k_0,\delta_0}$ *for some* $k_0 \precsim \sqrt{n}$ *and* $\tau_n \prec \delta_0 \prec 1$, *where* $\varepsilon_n \prec \tau_n \prec 1$ *and* $\varepsilon_n^2 \asymp \log(k_0)/n$, *then under Assumption 14, there exists a constant* C *only dependent on* P^0 *and* C_0 *such that*

$$\Pi_n\left(P \in S_{k_0,\tau_n} : ||P - P^0||_\infty < C_0\varepsilon_n; Z = Z_0; K = k_0\right) \geq e^{-Cn^2\varepsilon_n^2} \tag{3.3.4}$$

holds for all sufficiently large n.

With the above preparation, the proof of Theorem 3.3.1 is as follows. The structure of the proof follows [GvdV07].

Proof. We first verify prior mass condition. By Lemma 10, the set

$$\left\{\theta \in \bigcup_{k=1}^{K_n} \Theta_{k,0} : KL\left(\mathbb{P}_{\theta_n^0}, \mathbb{P}_\theta\right) < n^2\varepsilon_n^2, V_{2,0}\left(\mathbb{P}_{\theta_n^0}, \mathbb{P}_\theta\right) < n^2\varepsilon_n^2\right\}$$

contains a sup-norm ball $\left\{\theta \in \bigcup_{k=1}^{K_n} \Theta_{k,0} : ||\theta - \theta_n^0||_\infty < C_0\varepsilon_n\right\}$ for some constant C_0 only dependent on θ^0. Choose $\varepsilon_n \prec \tau_n \prec \delta_0$, the sup-norm ball further contains the following sup-norm ball $\{\theta \in \Theta_{k_0,\tau_n} : ||\theta - \theta_n^0||_\infty < C_0\varepsilon_n\}$. By Lemma 7, the sup-

norm ball is essentially its Z_0 slice which reduces to

$$\Pi_n \left(P \in S_{k_0,\tau_n} : ||P - P^0||_\infty < C\varepsilon_n; Z = Z_0; K = k_0 \right).$$

By Lemma 12, the prior mass is further lower bounded by $e^{-Cn^2\varepsilon_n^2}$ for some constant C only dependent on P^0 and C_0.

Next, we check the existence of tests. The existence of tests boils down to metric entropy condition and prior mass condition of the sieve. The sieve is constructed as $\bigcup_{k=1}^{K_n} \Theta_{k,\delta_n}$ with $1 \succ \delta_n \gtrsim \varepsilon_n$.

Metric entropy condition of the sieve requires the metric entropy is upper bounded by $Cn^2\varepsilon_n^2$. Clearly, this is satisfied by Lemma 9.

It is left to show the prior mass on the sieve. Note $\Pi_n \left(\left(\bigcup_{k=1}^{K_n} \Theta_{k,\delta_n} \right)^c \right) \leq \Pi_n \left(\Theta^c_{K_n,\delta_n} \right) = \Pi_n \left(\Theta^c_{K_n,\delta_n} | K = K_n \right) \Pi_n(K = K_n)$, then the prior mass on the sieve is also satisfied by a union bound:

$$\begin{aligned}
\Pi_n \left(\Theta^c_{k,\delta_n} | K = k \right) &\leq \textstyle\sum_{z \in \mathcal{Z}_{n,k}} \Pi_n \left(\Theta^c_{k,\delta_n} | Z = z, K = k \right) \Pi_n \left(Z = z | K = k \right) \\
&\leq \max_{z \in \mathcal{Z}_{n,k}} \Pi_n \left(\Theta^c_{k,\delta_n} | Z = z, K = k \right) \\
&= \max_{z \in \mathcal{Z}_{n,k}} \Pi_n \left(T(zPz^T) : P \in S^c_{k,\delta_n} | Z = z, K = k \right) \\
&\leq \Pi_n \left(S^c_{k,\delta_n} | K = k \right) \\
&\leq e^{-n^2\delta_n^2} \\
&\precsim e^{-Cn^2\varepsilon_n^2}
\end{aligned}$$

for some constant C, where the last inequality holds by Assumption 1.

□

3.4 Posterior sampler and inference

3.4.1 Reversible-jump MCMC algorithm

Under the truncation based weakly assortative Nowicki-Snijders prior (3.2.4), the posterior distribution factorizes as

$$\Pi_n\left(Z,K,P|A\right) \propto \Pi\left(A|Z,P\right)\Pi_n\left(P|Z\right)\Pi_n\left(Z|K\right)\Pi_n\left(K\right) \tag{3.4.1}$$

with

$$\begin{aligned}
\Pi\left(A|Z,P\right) &= \prod_{1\le a\le b\le K} P_{ab}^{O_{ab}(Z)}(1-P_{ab})^{n_{ab}(Z)-O_{ab}(Z)} \\
\Pi_n\left(P|Z,K,\delta_n\right) &= \prod_{1\le a<b\le K} \frac{1_{(0\le P_{ab}\le (P_{aa}\wedge P_{bb})-\delta_n)}}{(P_{aa}\wedge P_{bb})-\delta_n} \\
\Pi_n\left(Z|K\right) &= \frac{\Gamma(K)}{\Gamma(n+K)}\prod_{1\le c\le K}\Gamma\left(n_c\left(Z\right)+1\right) \\
\Pi_n\left(K\right) &\propto \frac{1}{K!}1_{1\le K\le K_n}.
\end{aligned}$$

For comparison, the original Nowicki-Snijders prior is conjugate and the community-wise connectivity probability matrix P can be marginalized out in the posterior distribution. Therefore, posterior inference on K is directly based on posterior draws from $\Pi_n(Z,K|A)$. However, the truncated Nowicki and Snijders' prior loses conjugacy. Our posterior inference needs to sample from $\Pi_n(P,Z,K|A)$.

We devise a Metropolis-Hastings algorithm to simulate a Markov chain of draws from (3.4.1), where in each iteration of the chain a proposed state (Z^*,K^*,P^*) is

accepted with probability

$$\min\left(1, \frac{\Pi_n\left(Z^*, K^*, P^*|A\right)}{\Pi_n\left(Z, K, P|A\right)} \frac{\Pi_{prop}(Z, K, P|Z^*, K^*, P^*)}{\Pi_{prop}(Z^*, K^*, P^*|Z, K, P)}\right) \tag{3.4.2}$$

where Π_{prop} denotes the density function of the proposal distribution and (Z, K, P) denotes the current state. The proposal distribution is adapted from the allocation sampler developed in [MMFH13]. For each iteration of the sampler, we first sample (Z, K) in the spirit of the allocation sampler, and then sample P given (Z, K). That is, the proposal distribution is decomposed into two parts: conditional on the previous draw (P, Z, K) and data matrix A,

$$\Pi_{prop}\left(Z^*, K^*, P^*|Z, K, P, A\right) \propto \Pi_{prop}\left(P^*|Z^*, A\right) \Pi_{prop}\left(Z^*, K^*|Z, K, P, A\right)$$

where $P^*_{ab}|Z^*, A \overset{ind}{\sim} Beta\left(O^*_{ab} + 1, n^*_{ab} - O^*_{ab} + 1\right)$ with $O^*_{ab} \equiv O_{ab}(Z^*)$ and $n^*_{ab} \equiv n_{ab}(Z^*)$, and $(Z^*, K^*)|(Z, K, P, A)$ are simulated in the spirit of the allocation sampler developed in [MMFH13, NF07], but with key differences due to the way the connectivity probability matrix P is involved and used in likelihood evaluation under the current model. In contrast, the allocation sampler of [MMFH13] explores the (Z, K) space with P marginalized out. Details of the posterior sampler are in Appendix B.

The expectation of the proposal distribution $\Pi_{prop}(P^*|(Z^*, A))$ is the ordinary block constant least squares estimator which is widely used to estimate the con-

nectivity probability matrix in the literature [see GLZ15, KTV17, vdPvdV18, for instance]. As the proposal density matches the likelihood component $\Pi(A|P^*, Z^*)$, the acceptance rate is a product of prior density ratios and proposal density ratios.

3.4.2 Posterior Inference

Under the 0-1 loss function $\ell(k, k_0) = 1_{k=k_0}$, the Bayes estimate of K is simply the posterior mode. As in the Metropolis-Hastings sampler, K communities may contain empty communities, we compute the effective number of communities based on samples of Z. The community assignment is identified up to a label switching. In our matrix formulation, the assignment Z is identified up to a column permutation. That is, ZZ^T is invariant to column permutations. If the $(i, j)^{th}$ entry of ZZ^T is 1, node i and node j are classified into the same community by Z. In addition, the node-wise connectivity θ is also identified without relabelling concerns. With the 0-1 loss function $\ell(Z, Z_0) = 1_{(ZZ^T=Z_0Z_0^T)}$, Bayes estimate of Z is its posterior mode. To pin down the posterior mode of Z, we can find the posterior mode of ZZ^T and the corresponding Z is the posterior mode of Z.

3.5 Numerical experiments

Section 3.3 presents asymptotic properties of Bayesian SBM with weakly assortative priors which is henceforth abbreviated as "a-SBM". This section assesses finite sample properties of a-SBM under a variety of challenging settings.

3.5.1 Simulation design

We perform simulation studies for different configurations of the number of communities, network size, and overall sparsity of connectivity. In particular, we choose $(k_0, n, \rho) \in \{3, 5, 7\} \times \{50, 75\} \times \{\frac{1}{2}, 1\}$, and for each (k_0, n, ρ) configuration, 100 networks are generated from $SBM(Z_0, \rho P^0, n, k_0)$. To control the source of variation in the synthetic networks, the 100 networks share the same community structure Z_0 where nodes are deterministically and uniformly assigned to k_0 communities; the 100 networks also share the same connectivity matrix ρP^0. The randomness in the 100 synthetic networks is only from the stochastic generation of Bernoulli trials of $SBM(Z_0, \rho P^0, n, k_0)$.

We choose the following cases for P^0.

- Case 1: $P^0 = 0.6 \times I_{k_0} + 0.2 \times 1_{k_0} 1_{k_0}^T$,
- Case 2: $P^0 = 0.2 \times I_{k_0} + 0.6 \times 1_{k_0} 1_{k_0}^T$,
- Case 3: $P^0 = 0.4 \times I_{k_0} + 0.4 \times 1_{k_0} 1_{k_0}^T$,
- Case 4: $P^0 = 0.2 \times I_{k_0} + 0.2 \times 1_{k_0} 1_{k_0}^T + 0.4 \times 1_{k_0, \lceil k_0/2 \rceil} 1_{k_0, \lceil k_0/2 \rceil}^T$,

where I_k denotes identity matrix of rank k, 1_k denotes the $k-$dimensional vector of ones, and $1_{n,k}$ denotes the $n-$dimensional vector with the first k elements being 1 and the rest $(n-k)$ elements being 0.

In the four cases, within-community connectivity probabilities are all 0.8. For simplicity, the between-community connectivity probabilities are the same for Cases

1-3; in Case 1, cross community connectivity is weak; in Case 2, cross community connectivity is strong; and in Case 3, cross community connectivity is medium. Case 4 combines the structure of Case 1 and Case 3 and half of the cross community connectivity is strong.

The reasons for choosing $n \in \{50, 75\}$ are as follows. Firstly, many networks in natural and social sciences are often of moderate size. Secondly, asymptotically consistent estimators can perform poorly when sample size is moderate. It is more informative to compare methods for networks of moderate size than that for networks with thousands of nodes. Thirdly, MCMC algorithms are computationally expensive, and the computation bottleneck prevents us from networks with more than thousands of nodes.

As the number of parameters in the SBM grows in the order of $O(k_0^2)$, the difficulty of community detection increases as k_0 grows. The case of $k_0 = 7$ imitates the situation of many communities, while the cases of $k_0 \in \{3, 5\}$ imitate networks with moderately many communties.

3.5.2 Simulation results

For comparison, we also implement Bayesian SBM with the Nowicki and Snijders' prior [NF07, GBP19], composite likelihood BIC method [SYF17], and network cross-validation [CL18]. Two posterior samplers for the Nowicki and Snijders' prior are available in the literature: the allocation sampler of [MMFH13], and the MFM

adapted MCMC algorithm of [GBP19]. We use the code provided in the supplementary materials of [GBP19] and choose default values for the hyperparameters in their algorithm. The Bayesian SBM of [GBP19, MMFH13] is henceforth denoted as "c-SBM" (Bayesian SBM with conjugate priors). [SYF17] propose composite likelihood BIC to choose the number of communities, and this method is henceforth denoted as "CLBIC". [CL18] design a cross-validation strategy to choose the number of communities for SBM, and it is henceforth denoted as "NCV".

Table 3.1: RMSE of $\hat{K}$.

			Case 1		Case 2		Case 3		Case 4	
k_0	n	Method	$\rho = \frac{1}{2}$	$\rho = 1$	$\rho = \frac{1}{2}$	$\rho = 1$	$\rho = \frac{1}{2}$	$\rho = 1$	$\rho = \frac{1}{2}$	$\rho = 1$
3	50	a-SBM	1.8	1.9	1.8	1.3	0.3	2.0	0.3	1.0
		c-SBM	0.8	1.9	1.9	1.0	0.2	1.9	0.6	0.9
		CLBIC	0.5	1.3	1.3	1.3	0.0	1.4	0.6	1.0
		NCV	0.9	2.0	2.0	2.0	0.0	2.0	0.9	0.9
	75	a-SBM	1.0	2.0	1.6	1.1	0.1	1.9	0.2	0.9
		c-SBM	0.5	2.0	1.6	1.0	0.3	1.8	0.4	0.9
		CLBIC	0.0	1.0	0.9	1.0	0.0	1.0	0.0	1.0
		NCV	0.1	2.0	1.9	2.0	0.0	2.0	0.0	1.0
5	50	a-SBM	3.0	3.9	3.9	2.3	1.2	4.0	3.6	2.8
		c-SBM	3.7	3.9	4.0	3.0	1.4	3.9	3.8	2.9
		CLBIC	3.1	3.4	3.3	3.5	1.9	3.4	3.2	2.9
		NCV	4.0	4.0	4.0	4.0	2.0	4.0	4.0	3.2
	75	a-SBM	2.0	3.9	3.9	2.6	0.5	4.0	2.3	2.9
		c-SBM	2.7	4.0	4.0	3.0	0.8	4.0	2.3	2.9
		CLBIC	2.6	3.0	3.0	2.9	0.0	3.0	2.8	2.7
		NCV	3.9	4.0	4.0	3.9	0.0	4.0	3.9	2.7
7	50	a-SBM	5.6	5.9	5.9	4.1	3.5	6.0	6.0	4.6
		c-SBM	5.9	5.9	6.0	5.1	4.8	6.0	5.9	5.0
		CLBIC	5.2	5.3	5.3	5.5	4.9	5.3	5.3	4.9
		NCV	6.0	6.0	6.0	6.0	6.0	6.0	6.0	5.6
	75	a-SBM	4.7	6.0	6.0	4.4	2.0	6.0	5.5	4.8
		c-SBM	5.4	5.9	5.9	5.0	2.6	5.9	5.4	5.0
		CLBIC	4.9	5.0	5.0	4.9	3.5	5.0	5.0	4.7
		NCV	6.0	6.0	6.0	6.0	3.5	6.0	6.0	4.8

Table 3.2: Bias of $\hat{K}$.

			Case 1		Case 2		Case 3		Case 4	
k_0	n	Method	$\rho = \frac{1}{2}$	$\rho = 1$	$\rho = \frac{1}{2}$	$\rho = 1$	$\rho = \frac{1}{2}$	$\rho = 1$	$\rho = \frac{1}{2}$	$\rho = 1$
3	50	a-SBM	1.3	1.9	−1.6	0.0	0.1	−1.9	0.1	−0.6
		c-SBM	−0.5	−1.9	−1.8	−1.0	−0.0	−1.9	−0.1	−0.8
		CLBIC	−0.2	−1.2	−1.2	−1.1	0.0	−1.3	−0.3	−0.9
		NCV	−0.6	−2.0	−2.0	−2.0	0.0	−2.0	−0.3	−0.8
	75	a-SBM	0.5	−1.9	−1.1	−0.6	0.0	−1.9	0.0	−0.7
		c-SBM	−0.1	−1.9	−1.3	−1.0	0.0	−1.6	0.0	−0.8
		CLBIC	0.0	−1.0	−0.8	−0.9	0.0	−1.0	0.0	−0.9
		NCV	0.0	−2.0	−1.9	−1.9	0.0	−2.0	0.0	−0.9
5	50	a-SBM	−2.5	−3.9	−3.8	−2.0	0.7	−4.0	−3.6	−2.7
		c-SBM	−3.7	−3.9	−4.0	−3.0	−1.0	−3.9	−3.7	−2.9
		CLBIC	−3.1	−3.4	−3.3	−3.4	−1.6	−3.3	−3.2	−2.8
		NCV	−4.0	−4.0	−4.0	−4.0	−1.5	−4.0	−4.0	−3.0
	75	a-SBM	−1.1	−3.9	−3.9	−2.4	0.0	−4.0	−2.0	−2.8
		c-SBM	−2.5	−4.0	−4.0	−3.0	−0.3	−3.9	−2.0	−2.9
		CLBIC	−2.5	−3.0	−3.0	−2.9	0.0	−3.0	−2.8	−2.7
		NCV	−3.8	−4.0	−4.0	−3.9	0.0	−4.0	−3.8	−2.6
7	50	a-SBM	−5.5	−5.9	−5.9	−3.9	−3.1	−6.0	−6.0	−4.5
		c-SBM	−5.9	−5.9	−6.0	−5.0	−4.7	−5.9	−5.9	−4.9
		CLBIC	−5.2	−5.3	−5.3	−5.4	−4.8	−5.3	−5.3	−4.8
		NCV	−6.0	−6.0	−6.0	−6.0	−6.0	−6.0	−6.0	−5.5
	75	a-SBM	−4.6	−6.0	−5.9	−4.3	−1.4	−5.9	−5.5	−4.8
		c-SBM	−5.4	−5.9	−5.9	−5.0	−2.3	−5.9	−5.3	−5.0
		CLBIC	−4.8	−5.0	−5.0	−4.8	−3.4	−5.0	−5.0	−4.7
		NCV	−6.0	−6.0	−6.0	−6.0	−3.2	−6.0	−6.0	−4.8

Compared with c-SBM, a-SBM achieves similar accuracy across different configurations. To be specific, when $k_0 = 3$, a-SBM tends to over-estimate the number of communities; when $\rho = \frac{1}{2}$ and $k_0 \in \{5, 7\}$, a-SBM is slightly more accurate than c-SBM in Case 1 and 3 and similarly accurate to c-SBM in Case 2 and 4. When the posterior samples of connectivity matrix of c-SBM are also weakly assortative,

c-SBM is essentially a-SBM. Therefore, it is reasonable to expect a-SBM and c-SBM have similar accuracy in networks generated from weakly assortative SBM.

Compared with CLBIC, a-SBM is less accurate in most cases. This is due to the design of P^0 in Case 1 - 3, such that the working likelihood of CLBIC is close to the true likelihood. In Case 4, the true likelihood is more complicated than the working likelihood of CLBIC, and the advantage of CLBIC over a-SBM is less obvious.

Compared with NCV, a-SBM is more accurate in most cases. To be specific, when $k_0 = 3$ and $\rho = \frac{1}{2}$, a-SBM tends to over-estimate the number of communities; in other configurations, a-SBM is more accurate than NCV.

Case 2 is the most difficult as the between community connectivity probability is very close to within community connectivity probability. Indeed, the methods nearly uniformly choose one big community, except that CLBIC sometimes chooses two communities.

To assess the membership assignment accuracy, we use the Hubert-Arabie adjusted Rand index [HA85, Ran71] to measure the agreement between two clustering assignments. The index is expected to be 0 if two independent assignments are compared, and is 1 if two equivalent assignments are compared. Though the adjusted Rand index tends to capture the disagreement among large clusters, community sizes in our simulation study are about the same and the adjusted Rand index is still a meaningful metric.

Given a synthetic network A and draws from the posterior distribution $\Pi(\cdot|A)$,

we can compute the adjusted Rand index of posterior draws of Z against Z_0 and use their mean as the accuracy metric for $\Pi(\cdot|A)$. Like the adjusted Rand index for two clustering assignments, the averaged index assesses the agreement of the posterior distribution of Z against the truth Z_0.

Table 3.3 presents the average of adjusted Rand indices of the 100 synthetic networks under different (k_0, ρ, n) configurations in the four cases. Overall, the average adjusted Rand index of a-SBM is similar to that of c-SBM. This echoes the similar estimation accuracy of k of a-SBM and c-SBM, as community detection is highly sensitive to the number of communities. When $\rho = 1/2$ and $k_0 \in \{5, 7\}$, a-SBM is slightly better than c-SBM in Case 1 and 3. When data is less informative, the regularity in the prior of a-SBM improves estimation accuracy over c-SBM. The advantage disappears in Case 2 and 4 where cross community connectivity is close to within community connectivity.

Table 3.3: Adjusted Rand index

			Case 1		Case 2		Case 3		Case 4	
k_0	ρ	n	a-SBM	c-SBM	a-SBM	c-SBM	a-SBM	c-SBM	a-SBM	c-SBM
3	$\frac{1}{2}$	50	0.62	0.63	0.00	0.00	0.05	0.01	0.37	0.42
		75	0.87	0.91	0.00	0.00	0.12	0.05	0.47	0.51
	1	50	0.97	0.98	0.01	0.01	0.86	0.86	0.56	0.58
		75	0.99	0.99	0.03	0.03	0.96	0.97	0.58	0.57
5	$\frac{1}{2}$	50	0.10	0.03	0.00	0.00	0.01	0.00	0.20	0.22
		75	0.25	0.11	0.00	0.00	0.01	0.00	0.28	0.30
	1	50	0.83	0.86	0.00	0.00	0.06	0.03	0.33	0.34
		75	0.94	0.99	0.00	0.00	0.32	0.17	0.35	0.36
7	$\frac{1}{2}$	50	0.03	0.00	0.00	0.00	0.01	0.00	0.12	0.12
		75	0.07	0.01	0.00	0.00	0.00	0.00	0.19	0.20
	1	50	0.24	0.12	0.00	0.00	0.01	0.00	0.24	0.24
		75	0.57	0.43	0.00	0.00	0.04	0.02	0.27	0.27

3.6 Sparse networks

The framework in Section 3.3 can be extended to sparse networks whose overall connectivity probability shrinks to 0 as network size increases [e.g. KTV17, GM18]. We state the posterior contraction rates and the posterior consistency results for those sparse networks as follows. Their proofs follow exactly the same argument except that the derivations involve the sparse factor ρ_n.

Theorem 3.6.1. *Suppose adjacency matrix $A \in \{0,1\}^{n\times n}$ is generated from the SBM with $\theta_n^0 = \rho_n T(Z_0 P^0 Z_0^T)$, $\sqrt{\log(k_0)/n} \prec \rho_n \precsim 1$, $P^0 \in \Theta_{k_0,\delta_0}$ for some $k_0 \precsim \sqrt{n}$ and $\delta_0 > 0$ such that $\rho_n\delta_0 \succ \sqrt{\log(k_0)/n}$, and the number of zero and one entries of $T(Z_0 P^0 Z_0^T)$ is at most $O(n^2\varepsilon_n)$ where $\varepsilon_n^2 \asymp \frac{\log(k_0)}{n}$. The prior Π_n satisfies Assumption 14. Then, for all sufficiently large M,*

$$\mathbb{P}_{0,n}\Pi_n\left(\theta : ||\theta - \theta_n^0||_\infty \geq M\varepsilon_n|A\right) \to 0.$$

The posterior contraction rate in Theorem 3.6.1 is independent of the sparsity level. In contrast, L_2 minimax rates of error derived in [KTV17, GM18] are proportional to the sparsity level. We conjecture that L_∞ minimax rates of error are also proportional to the sparsity level. It is likely that the posterior contraction rate in Theorem 3.6.1 is sub-optimal.

Theorem 3.6.2. *Under the same assumptions of Theorem 3.6.1, then*

$$\mathbb{P}_{0,n}\left[\Pi_n\left(\{K = k_0\} \cap \{Z = Z_0\}|A\right)\right] \to 1.$$

In the sparse network setting, the separation parameter δ_0 also vanishes at the rate of ρ_n. Our identification strategy for the number of communities requires $\rho_n\delta_0 \succ \varepsilon_n \asymp \sqrt{\log(k_0)/n}$ to guarantee exact recovery. In contrast, some work in the sparse network literature works for networks with sparser sparsity levels [e.g. Abb17, for a recent survey]. The Bayesian model outlined in (3.2.4) may need additional modifications to adapt to networks at various sparse levels.

The number of true communities k_0 is allowed to be sub-linear in n. For the sparse SBMs $SBM(Z_0, \log(n)/nP^0, n, k_0)$, it is unclear when exact recovery is possible for k_0 sub-linear in n [Abb17]. It is pitiful that our sparse case is denser than the sparse SBM.

3.7 Discussion

In this paper, we have shown Bayesian SBM can consistently estimate the number of communities and achieve exact recovery. Towards this end, we propose the weakly assortative Nowicki-Snijders' prior and trade conjugacy of Nowicki-Snijders' prior for simpler and clearer asymptotic analysis. A brief simulation study is performed to illustrate finite sample performance.

3.7.1 Bayesian SBM with conjugate priors

In the simulation studies, c-SBM has similar finite sample estimation accuracy to a-SBM. We conjecture that c-SBM can also consistently estimate the number of communities and the membership assignment for networks generated from weakly assortative SBM. However, the proof technique adopted in this paper cannot be applied to c-SBM.

3.7.2 Efficient sampler

The price of losing conjugacy is on the computation side. The posterior sampler in [GBP19] is much faster than our allocation sampler as they successfully adapt the idea of MFM sampler of [MH18] to the SBM case. It remains unclear if the MFM idea can be applied to the non-conjugate case.

Chapter 4

Bayesian Tail Index Estimation

4.1 Introduction

In scientific applications, many random quantities of interest tend to exhibit more "extreme" values than Gaussian random variables, with examples including stock returns, file sizes, city sizes, transmission durations and traffic of data networks [Res07, Gab09]. More empirical examples can be found in, just to name a few, [KPN02, Cas12, Mar07, EKM13].

In the context of heavy-tailed density modeling and extremes forecasting, a mathematically succinct representation of the departure from normality is through regularly varying functions. A random variable X has a heavy (right) tail if its CDF $F(x)$ satisfies, as $x \to \infty$,

$$1 - F(x) = x^{-\alpha} L(x)$$

for some constant $\alpha > 0$ and some slowly varying function $L(x)$. The parameter α is the tail index that characterizes tail heaviness.

Tail index estimation has been extensively studied in the frequentist literature. By the Pickands-Balkenma-De Haan theorem, the conditional distribution of excess over certain threshold is asymptotically a generalized Pareto distribution (GPD), as the

threshold grows to infinity [BDH74, PI75]. This justifies the peaks-over-threshold estimation methods that fit a GPD[dZBK10]. Nonparametric estimators are also developed: Hill's estimator [Hil75] and Pickand's estimator [PI75].

The peaks-over-threshold estimation only uses high quantiles and ignores substantial amount of data. To use the whole dataset, several methods estimate the entire density by adding a GPD tail to the bulk [TAO06, MSL$^+$11, dNGL12]; transformation based density model via two-stage estimation is also proposed [Mar07].

However, the aforementioned frequentist estimators all require the practitioners to choose the threshold or some order statistics whose optimal choice is unclear and challenging [SM12, Res07]. Also, analysis needs to account for the threshold choice uncertainty.

Bayesian approach bypasses the problem of choosing the threshold or order statistics by modeling the entire density. For instance, [LLD19] consider Pareto mixture models; [TC21] consider a transformation based semi-parametric density model. Both approaches are proved to be consistent. However, posterior contraction rates are unknown. It remains unclear if the Bayesian methods are efficient. It is surprising that asymptotic analysis on Bayesian tail index estimation is rare.

Instead of mixture models, we work with the transformation based density model of [TC21]: the CDF F satisfies $F(y) = \Psi(G_\theta(y))$ where $G_\theta(y)$ is a parametric CDF chosen by the statistician and $\Psi : [0,1] \to [0,1]$ is a non-parametric monotonic transformation. Under the semiparametric density model, the tail index of F is the

tail index of G_θ. Regular prior specification of θ can avoid the inconsistency pitfall detailed in [LLD19].

As a working example, we choose generalized Pareto distribution (GPD) as the parametric family and put a uniform prior on the bounded parameter space. Further, we put a logistic conditional Gaussian process prior on the nonparametric transformation. Under the prior, the exponentially powerful test required by Schwartz method [GvdV17] exists. Therefore, we establish posterior contraction rate of tail index for the transformation based semiparametric density model.

As a side result, the transformation based semi-parametric density model achieves minimax optimal error rate (up to a log factor) for density estimation. The proof also follows Schwartz method where a novel sieve is constructed by viewing sample paths of the Gaussian process prior as Borel measurable maps in $C^2[0,1]$.

4.2 Bayesian semi-parametric density model

4.2.1 The model

The semi-parametric density model takes the following form

$$p_{\theta,\psi}(y) = g_\theta(y)\psi(G_\theta(y)), \tag{4.2.1}$$

where $G_\theta(\dot{)}$ is a parametric CDF indexed by θ, transformation $\psi(\cdot) \in C^2[0,1]$ [Mar07]. Under the model, the tail index of $p_{\theta,\psi}$ is the tail index of g_θ.

The class of CDF $\{G_\theta(\cdot) : \theta \in \Theta\}$ with Θ being compact is chosen by the statistician. For instance, one can choose student$-t$ distributions with θ being the degree of freedom; alternatively, one can choose generalized Pareto distribution with θ being a scale parameter and a shape parameter.

The Pickands-Balkenma-De Haan theorem justifies the use of a generalized Pareto Distribution (GPD) to model excesses conditional on some high quantile. So it is natural to non-parametrically transform the GPD to fit the data. In the followup analysis, the parametric family is set to be a generalized Pareto Distribution (GPD). Specifically, we set $\theta = (\alpha, \sigma)$ and work with

$$g_\theta(x) = \frac{1}{\sigma}(1 + \frac{x}{\alpha\sigma})^{-(\alpha+1)} 1_{x\geq 0} \tag{4.2.2}$$

where $\sigma \in \mathbb{R}_{++}$ is the scale parameter, and $\alpha \in \mathbb{R}_{++}$ is the inverse of the shape parameter. The CDF of (4.2.2) is

$$G_\theta(x) = (1 - (1 + \frac{x}{\alpha\sigma})^{-\alpha}) 1_{x\geq 0}. \tag{4.2.3}$$

The Bayesian approach is to specify some prior on θ and place a logistic Gaussian process (LGP) prior on ψ. More detailed prior specification is presented in Section 4.3 and Section 4.4.

4.2.2 Tail expansion

The exact Hall condition for CDF F is, for $\alpha, \beta, C, C' > 0$, as $x \to \infty$,

$$|1 - F(x) - Cx^{-\alpha}| = C'x^{-\alpha(1+\beta)} + o(x^{-\alpha(1+\beta)}). \tag{4.2.4}$$

Denote the class of distribution satisfying the exact Hall condition as $\mathbb{H}_{\alpha,\beta,C,C'}$, tail index of distribution F as $\alpha_+(F)$, and the second order stochastic parameter of F as $\beta(F)$.

In estimating the tail index, the minimax error rate of tail index estimation is $n^{-\beta/(2\beta+1)}$ if β is known [Dre01, HW84]. When β is unknown, the minimax error rate may have additional log factor depending on the specific tail decay assumption [HW85, GS08, CK15].

As the optimality of tail index estimation depends on the second order stochastic parameter, it is of crucial importance to examine it under model (4.2.1). Write $F(y) = \Psi(G_\theta(y))$, then by Taylor expansion of $\bar{\Psi}(1-u)$ around $u = 0$,

$$\bar{F}(y) = \bar{\Psi}(1 - \bar{G}_\theta(y)) = \psi(1)\bar{G}_\theta(y) - \frac{1}{2}\dot{\psi}(1)\bar{G}_\theta(y)^2 + o(\bar{G}_\theta(y)^2).$$

The second order stochastic parameter of model (4.2.1) depends on the first and the second order stochastic parameter of $G_\theta(\cdot)$. Suppose $G_\theta \in \mathbb{H}_{\alpha,\beta,C,C'}$, then as $y \to \infty$,

- if $\beta > 1$, $F \in \mathbb{H}_{\alpha,1,\psi(1)C,\frac{1}{2}\dot{\psi}(1)C^2}$;

- if $\beta = 1$, $F \in \mathbb{H}_{\alpha,1,\psi(1)C,|\psi(1)C'-\frac{1}{2}\dot{\psi}(1)C^2|}$;

- if $\beta \in (0,1)$, $F \in \mathbb{H}_{\alpha,\beta,\psi(1)C,\psi(1)C'}$.

The second order stochastic parameter of model (4.2.1) is in $(0,1]$, which might artificially impose an adverse effect on estimation efficiency.

Now, suppose G_θ is the CDF of generalized Pareto distribution, as $y \to \infty$,

$$\bar{G}_\theta(y) = (1 + \frac{y}{\alpha\sigma})^{-\alpha} = (\alpha\sigma)^\alpha y^{-\alpha}[1 - \alpha^2\sigma y^{-1} + o(y^{-1})].$$

Further with $C_{\theta,\psi} = (\alpha\sigma)^\alpha\psi(1)$, as $y \to \infty$,

- $\alpha \in (0,1)$, $|\bar{F}(y) - C_{\theta,\psi}y^{-\alpha}| = \frac{1}{2}C_{\theta,\psi}y^{-2\alpha}\dot{\phi}(1)(\alpha\sigma)^\alpha(1+o(1))$, $\beta(\mathbb{P}_{\theta,\psi}) = 1$;

- $\alpha = 1$, $|\bar{F}(y) - C_{\theta,\psi}y^{-\alpha}| = \frac{1}{2}C_{\theta,\psi}y^{-2\alpha}(2\alpha^2\sigma + \dot{\phi}(1)(\alpha\sigma)^\alpha)(1+o(1))$, $\beta(\mathbb{P}_{\theta,\psi}) = 1$;

- $\alpha \in (1,\infty)$, $|\bar{F}(y) - C_{\theta,\psi}y^{-\alpha}| = C_{\theta,\psi}y^{-\alpha-1}\alpha^2\sigma(1+o(1))$, $\beta(\mathbb{P}_{\theta,\psi}) = \alpha^{-1} \in (0,1)$.

With G_θ being generalized Pareto distribution, $\beta(\mathbb{P}_{\theta,\psi})$ deterministically depends on $\alpha_+(\mathbb{P}_{\theta,\psi})$ if $\alpha_+(\mathbb{P}_{\theta,\psi}) > 1$.

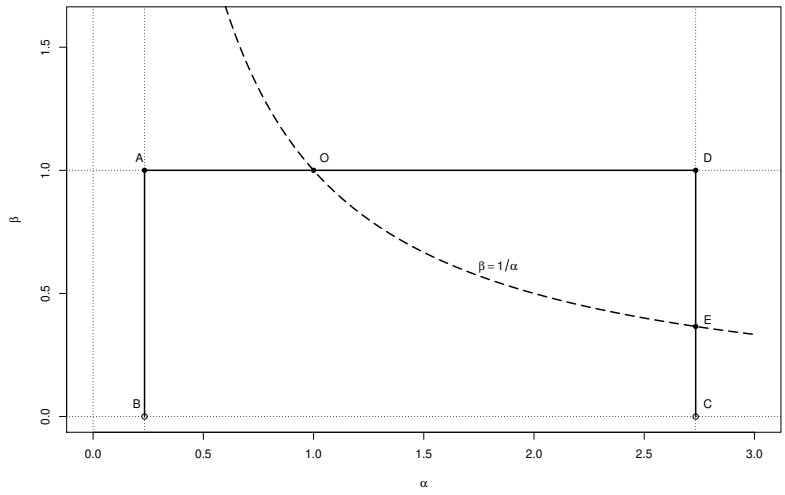

Figure 4.1: Illustration of (α, β) space of $\mathbb{P}_{\theta,\psi}$

For a better illustration, the (α, β) space of $\mathbb{P}_{\theta,\psi}$ is presented in Figure 4.1. Suppose $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, the (α, β) space of $\mathbb{P}_{\theta,\psi}$ with $\beta(G_\theta)$ being unrestricted is the rectangle $ABCD$; the (α, β) space of $\mathbb{P}_{\theta,\psi}$ with G_θ being generalized Pareto distribution is the curve AOE.

4.3 Density estimation

To be a reasonable density model, the Bayesian semi-parametric model (4.2.1) should provide precise density estimation. This section provides minimal assumptions such that density estimation under our model is near minimax optimal.

Assumption 10. *For all multi-index α with $|\alpha| = 2$, there exists an envelope function $M(X)$ such that*

- $\sup_{\theta\in\Theta} |\partial^\alpha \log g_\theta(x)| \leq M(x)$ *holds for all x, and*

- $\sup_{\theta\in\Theta,\psi\in C^2[0,1]} \mathbb{P}_{\theta,\psi}[M(X)] < \infty$.

Assumption 11. *For all multi-index α with $1 \leq |\alpha| \leq 2$, there exists an envelope function $M_G(X)$ such that*

- $\sup_{\theta\in\Theta} |\partial^\alpha G_\theta(x)| \leq M_G(x)^{|\alpha|}$ *holds for all x, and*

- $\sup_{\theta\in\Theta,\psi\in C^2[0,1]} \mathbb{P}_{\theta,\psi}[|M_G(X)|^2] < \infty$.

Assumption 12. *For a multi-index α whose $\dim(\alpha) = \dim(\theta)$, there exist a function $M_{e,1}(X)$ and a function $M_{e,2}(X)$ such that*

- $\sup_{\theta\in\Theta} \sum_{|\alpha|=1} |\partial^\alpha \log g_\theta(x)| \leq M_{e,1}(x)$, *and* $\sup_{\theta\in\Theta} \mathbb{P}_{\theta,\psi^*}[e^{|M_{e,1}(X)|}] < \infty$;

- $\sup_{\theta\in\Theta} \sum_{|\alpha|=1} |\partial^\alpha G_\theta(x)| \leq M_{e,2}(x)$, *and* $\sup_{\theta\in\Theta} \mathbb{P}_{\theta,\psi^*}[e^{C|M_{e,2}(X)|}] < \infty$ *for all $C > 0$.*

Assumption 10-12 impose regularity conditions for g_θ in its differentiability and integrability. In the case of Generalized Pareto distribution with Θ being compact, these Assumptions are satisfied: as $\theta \in \Theta$ and Θ is compact, Assumption 10 is satisfied with $M(x)$ being a constant; Assumption 11 is satisfied with $M_G(x)$ being some constant; Assumption 12 is satisfied with $M_{e,1}(x) = \log(1 + cx)$ for some universal constant c and $M_{e,2}(x)$ being some constant.

Assumption 13. *Let $\phi^*(x) = \log(\psi^*(x))$, $||\dot{\phi}^*||_\infty < \infty$, and $||\ddot{\phi}^*||_\infty < \infty$.*

Assumption 13 imposes regularity condition for the transformation ψ and requires its logarithm to be at least twice differentiable.

Assumption 14 (Prior specification).

- *There exists a universal constant C such that for every measurable $E \subseteq \Theta$, $\Pi_n(\theta \in E) \geq C|E|^{-1}$.*
- *For the nonparametric part, W^A follows a rescaled SE GP and further induces a prior on ψ: $\psi(t) = \frac{e^{W^A(t)}}{\int_0^1 e^{W^A(s)}ds}$.*
- *The rescaling parameter A of W^A follows a Gamma-like distribution such that there exist constants $C, D > 0$, for all sufficiently large a, its prior density $\pi(a) \geq De^{-Ca\log^2(a)}$.*

Theorem 4.3.1 (Density estimation). *Suppose $Y_i \sim p_0 \equiv p_{\theta^*,\psi^*}$ for $i = 1, ..., n$, the parametric family $\{g_\theta : \theta \in \Theta\}$ satisfies Assumption 10-12. the nonparametric part $\psi^* \in C^\gamma[0,1]$ and satisfies Assumption 13, the prior specification satisfies Assumption 14. Let $\varepsilon_n \propto n^{-\gamma/(1+2\gamma)}\log^\kappa(n)$ with $\kappa = 2\gamma/(2\gamma+1)$. Then for every sufficiently large M,*

$$\mathbb{P}_0[\Pi_n(h(p_{\theta,\psi}, p_0) > M\varepsilon_n|Y^n)] \to 0$$

The proof follows the Schwartz method [GvdV17, vdVvZ09]. Specifically, we first verify the prior mass condition and then construct a sieve that receives sufficient prior mass and whose metric entropy in Hellinger distance is under control. The complication of the sieve construction is due to the fact that $KL(p_{\theta_1,\psi}, p_{\theta_2,\psi}) \precsim (||\dot{\phi}||_\infty + ||\ddot{\phi}||_\infty)||\theta_1 - \theta_2||_\infty^2$ which requires additional regularity on the derivatives. As the rescaled SEGP W^A has sample paths that are infinitely smooth, we can view W^A as a map in $C^2[0,1]$ with norm $||w||_{C^2} = ||w||_\infty + ||\dot{w}||_\infty + ||\ddot{w}||_\infty$ and specify the

sieve in terms of $||\cdot||_{C^2}$.

The prior mass condition in terms of KL is simplified to distances in (θ, ψ) due to the following Lemma.

Lemma 12 (Prior mass). *Suppose $X_i \sim p_0 \equiv p_{\theta^*,\psi^*}$ for $i = 1, ..., n$. Let $\psi(t) = \frac{e^{w(t)}}{\int_0^1 e^{w(s)}ds}$ and $\psi^* \in C^\gamma[0,1]$ for some $\gamma \geq 2$. If $||\theta - \theta^*||_\infty \leq \varepsilon_n$ and $||w - w^*||_\infty \leq \varepsilon_n$ for some $\varepsilon_n < 1/2$, then there exists a constant C_0 only dependent on $\theta^*, \psi^*, \{G_\theta\}$, such that $KL(p_{\theta^*,\psi^*}, p_{\theta,\psi}) \leq C_0\varepsilon_n^2$, and $V_{2,0}(p_{\theta^*,\psi^*}, p_{\theta,\psi}) \leq C_0\varepsilon_n^2$.*

As the sieve is in $||\cdot||_{C^2}$, small ball probability in $||\cdot||_{C^2}$ becomes necessary. To this end, we follow the metric entropy method [KL93, LL99, vdVvZ08b] to derive lower bounds for the centered small ball probability.

Lemma 13 (Metric entropy). *Suppose $W^a \sim$ SEGP at rescaling level $a > a_0$ for some positive constant a_0, $\mathcal{H}^a$ is the RKHS associated with W^a, and $\mathcal{H}_1^a$ is the RKHS unit ball. There exists a universal constant C such that for $\varepsilon < 1/2$,*

$$\log N(\varepsilon, \mathcal{H}_1^a, ||\cdot||_{C^2}) \leq Ca\log(a/\varepsilon)^2.$$

The proof of Lemma 13 generally follows Lemma 4.5 of [vdVvZ09] and constructs piecewise polynomials to approximate the functions in $\mathcal{H}_1^a$. But our polynomials are different because (1) the spectral density of the rescaled SEGP has thinner tail than the one assumed in Lemma 4.5 of [vdVvZ09]; (2) the Banach space norm $||\cdot||_{C^2}$ requires a finer approximation than the $||\cdot||_\infty$ norm. The complete proof of Lemma

13 is presented in section C.1.3

Lemma 13 studies rescaled SEGP on $[0,1]$, which suffices for the current paper. For rescaled SEGP $\{W_t^a : t \in [0,1]^d\}$, the metric entropy of its RKHS unit ball $\mathcal{H}_1^a$ satisfies $\log N(\varepsilon, \mathcal{H}_1^a, ||\cdot||_{C^2}) \leq Ca^d \log(a/\varepsilon)^{d+1}$ for some universal constant C.

Lemma 14 (Centered small ball probability)**.** *Suppose $W^a \sim$ SEGP at rescaling level $a > a_0$ for some positive constant a_0, then there exists a universal constant C such that*

$$-\log \mathbb{P}(||W^a||_{C^2} \leq \varepsilon) \leq Ca \log(a/\varepsilon)^2$$

Proof. The proof follows the metric entropy argument outlined in Lemma 4.6 of [vdVvZ09].

□

4.4 Tail index estimation

4.4.1 Prior specification

The prior on (θ, ψ) induces a prior on the tail behavior of $p_{\theta,\psi}$. The induced prior distribution of $C_{\theta,\psi}$ should be essentially supported on a bounded set. Because otherwise, the tail probability at certain locations are mainly driven by $C_{\theta,\psi}$ instead of α, and then Type II error of certain tests dramatically increases. For the same reason, the induced prior distribution on the coefficient of $y^{-\alpha-1\wedge\alpha}$ should also be essentially supported on a bounded set.

Assumption 15 (Prior specification). • *There exist constants* $\bar{\alpha}, \underline{\alpha}, \bar{\sigma}, \underline{\sigma} > 0$ *such that* $\alpha \in [\underline{\alpha}, \bar{\alpha}]$ *and* $\sigma \in [\underline{\sigma}, \bar{\sigma}]$ *with prior probability 1. There exists a universal constant* C *such that for every measurable* $E \subseteq [\underline{\alpha}, \bar{\alpha}] \times [\underline{\sigma}, \bar{\sigma}]$, $\Pi_n((\alpha, \sigma) \in E) \geq C|E|^{-1}$.

- *For the nonparametric part, there exists a constant* M, $\psi(t) = \frac{e^{W^A(t)}}{\int_0^1 e^{W^A(s)} ds}$ *where* $W^A = \tilde{W}^A|_{||\tilde{W}^A||_{C^1} < M}$ *follows a conditional GP prior and* $\tilde{W}^A$ *follows a rescaled SE GP on* $[0, 1]$.

- *The rescaling parameter* A *of* $\tilde{W}^A$ *follows a Gamma-like distribution such that there exist constants* $C, D > 0$, *for all sufficiently large* a, *its prior density* $\pi(a) \geq De^{-Ca\log^2(a)}$.

The conditional GP W^A is well-defined as $\Pi(||\tilde{W}^A||_{C^1} < M) > 0$ for any $M > 0$. The lower bound of the prior mass of W^A is lower bounded by the prior mass of $\tilde{W}^A$ if M is large enough. Too see it,

$$\begin{aligned}
\Pi(||W^A - w_0||_{C^1} \leq \varepsilon) &= \Pi(||\tilde{W}^A - w_0||_{C^1} \leq \varepsilon | ||\tilde{W}^A||_{C^1} \leq M) \\
&= \frac{\Pi(||\tilde{W}^A - w_0||_{C^1} \leq \varepsilon, ||\tilde{W}^A||_{C^1} \leq M)}{\Pi(||\tilde{W}^A||_{C^1} \leq M)} \\
&\geq \Pi(||\tilde{W}^A - w_0||_{C^1} \leq \varepsilon, ||\tilde{W}^A||_{C^1} \leq M)
\end{aligned}$$

where the last display equals $\Pi(||\tilde{W}^A - w_0||_{C^1} \leq \varepsilon)$ if $||w_0||_{C^1} + \varepsilon \leq M$. Note the norm dominance, $\Pi(||W^A - w_0||_\infty \leq \varepsilon) \geq \Pi(||W^A - w_0||_{C^1} \leq \varepsilon)$ which is further lower bounded by $\Pi(||\tilde{W}^A - w_0||_{C^1} \leq \varepsilon)$ if $||w_0||_{C^1} \leq M + 1$. By choosing a reasonably large M, the above argument holds for a reasonably large set of $w_0 \in C^2[0, 1]$.

4.4.2 Existence of tests

Under Assumption 15, Hall condition (4.2.4) is satisfied with C and C' both bounded with prior probability 1. In the follow-up analysis, we focus on the class of distributions $\mathbb{H}_{\alpha,C,x_0}$ such that for all $x > x_0$,

$$|1 - F(x) - Cx^{-\alpha}| \leq \frac{1}{2}Cx^{-\alpha}. \tag{4.4.1}$$

The factor $\frac{1}{2}$ is chosen for convenience and can be changed to any number in $(0, 1)$.

Under Assumption 15, there exists a universal constant x_0 dependent on $\underline{\alpha}, \bar{\alpha}$, $\underline{C}, \bar{C}, \underline{C}'$ and $\bar{C}'$, such that $\mathbb{P}_{\theta,\psi} \in \mathbb{H}_{\alpha,C,x_0}$ for all $\alpha \in [\underline{\alpha}, \bar{\alpha}]$, $C \in [\underline{C}, \bar{C}]$, and $C' \in [\underline{C}', \bar{C}']$. That is, the prior support of $\mathbb{P}_{\theta,\psi}$ is a subset of $\cup_{\alpha,C}\mathbb{H}_{\alpha,C,x_0}$ for some constant x_0.

For $k \in \mathbb{N}$, define $S_{\theta,\psi}(k) \equiv \mathbb{P}_{\theta,\psi}(Y > e^k) = 1 - \Psi(G_\theta(e^k))$, where $Y \sim \mathbb{P}_{\theta,\psi}$. Suppose $\mathbb{P}_{\theta,\psi} \in \mathbb{H}_{\alpha,C,x_0}$, then as $k \to \infty$,

$$|S_{\theta,\psi}(k) - Ce^{-k\alpha}| \leq \frac{1}{2}Ce^{-k\alpha}, \tag{4.4.2}$$

which further implies $\frac{1}{2}Ce^{-k\alpha} \leq S_{\theta,\psi}(k) \leq \frac{3}{2}Ce^{-k\alpha}$.

Lemma 15 (existence of tests)**.** *Suppose the null hypothesis H_0 is $\mathbb{P}_{\theta_0,\psi_0}$ and $\mathbb{P}_{\theta_0,\psi_0} \in \mathbb{H}_{\alpha_0,\beta_0,C_0,C_0'}$ for some $\alpha_0, \beta_0, C_0, C_0' > 0$; the alternative hypothesis H_1 is $\{\mathbb{P}_{\theta,\psi} : |\alpha_+(p_{\theta,\psi}) - \alpha_0| > M\varepsilon_n\}$. Let $\mathbb{H}_n = \cup_{\underline{\alpha} \leq \alpha \leq \bar{\alpha}, \underline{C} \leq C \leq \bar{C}, x_0}\mathbb{H}_{\alpha,C,x_0}$ for constants $\underline{\alpha}, \bar{\alpha}, \underline{C}, \bar{C}, x_0 >$*

0. *Consider the test*

$$T_n = 1(|\hat{p}_k - S_{0,n}| > \varepsilon_n S_{0,n})$$

where $\hat{p}_k \equiv \frac{1}{n}\sum_{i=1}^n 1(Y_i > e^k)$, $k = \text{loglog}(n)$, $S_{0,n} \equiv S_{\theta_0,\psi_0}(k)$, *and* $\varepsilon_n \asymp n^{-\kappa}$ *for some constant* $\kappa \in (0,1)$. *Then,*

$$\mathbb{P}_{\theta_0,\psi_0}[T_n] \leq 2e^{-2n\varepsilon_n^2 S_{0,n}^2}, \text{ and } \sup_{\mathbb{P}_{\theta,\psi} \in H_1 \cap \mathbb{H}_n} \mathbb{P}_{\theta,\psi}[1 - T_n] \leq 2e^{-n\varepsilon_n^2 S_{0,n}^2}$$

hold for all sufficiently large n.

The test statistic is the ratio $\frac{\hat{p}_k}{S_{0,n}}$. The key step in bounding Type II error is to bound the ratio $\frac{S_{\theta,\psi}}{S_{0,n}}$. The sieve $\mathbb{H}_n$ allows us to focus on the first order stochastic of the tail probability expansion and then, the ratio becomes tractable. Without $\mathbb{H}_n$, the ratio $\frac{S_{\theta,\psi}}{S_{0,n}}$ can be 1 by choosing a suitable C'.

4.4.3 Posterior conatrction rate of tail index

With the existence of the test, we can establish posterior contraction rate for tail index estimation by Schwartz theory. The proof of Theorem 4.4.1 is standard.

Theorem 4.4.1. *Suppose* $Y_1, ..., Y_n \overset{iid}{\sim} \mathbb{P}_{\theta_0,\psi_0} \in \mathbb{H}_{\alpha_0,\beta_0,C_0,C_0'}$ *with* $\log(\psi_0) \in C^\gamma[0,1]$ *and* $\gamma_0 > 2$, *prior specification satisfies Assumption 15, and* $\varepsilon_n \asymp n^{-\gamma_0/(1+2\gamma_0)}\log^{\alpha_0+\kappa}(n)$ *with* $\kappa = 2\gamma_0/(2\gamma_0+1)$, *and* $\gamma_0 = \gamma - 1$. *Then, for every sufficiently large* M,

$$\mathbb{P}_{\theta_0,\psi_0}[\Pi_n(|\alpha - \alpha_0| > M\varepsilon_n | Y^n)] \to 0.$$

Proof. The proof follows the proof of Theorem 8.9 of [GvdV17]. Let $E_n = \{\alpha : |\alpha - \alpha_0| > M\varepsilon_n\}$ and $J_n = \{Y^n : \int \frac{p_{\theta,\psi}}{p_{\theta_0,\psi_0}}(Y^n)d\Pi(\theta,\psi) \geq e^{-n\bar{\varepsilon}_n^2}\}$ where $\bar{\varepsilon}_n \prec \varepsilon_n(\log n)^{-\alpha_0}$.

Next we work with the following inequality

$$\Pi(E_n|Y^n) \leq T_n + 1(J_n^C) + (1-T_n)e^{n\bar{\varepsilon}_n^2}\int_{E_n}\frac{p_{\theta,\psi}}{p_{\theta_0,\psi_0}}(Y^n)d\Pi(\theta,\psi),$$

where $\mathbb{P}_{\theta_0,\psi_0}[T_n] \leq 2e^{-2n\varepsilon_n^2 S_{0,n}^2}$ by Lemma 15, $\mathbb{P}_{\theta_0,\psi_0}[1(J_n^C)] \precsim 1/(n\bar{\varepsilon}_n^2)$ by Lemma 12, Lemma 8.10 of [GvdV17], and the argument below Assumption 15, and

$$\mathbb{P}_{\theta_0,\psi_0}\left[(1-T_n)\int_{E_n}\frac{p_{\theta,\psi}}{p_{\theta_0,\psi_0}}(Y^n)d\Pi(\theta,\psi)\right] \leq \sup_{\mathbb{P}_{\theta,\psi}\in E_n\cap\mathbb{H}_n}\mathbb{P}_{\theta,\psi}[1-T_n]+\Pi(\mathbb{H}_n^C) \leq 2e^{-n\bar{\varepsilon}_n^2 S_{0,n}^2}$$

by Lemma 15 and Assumption 15.

□

The posterior contraction rate of tail index roughly matches the posterior contraction rate for density estimation–up to a log factor and the smoothness level. As the error rates of test T_n require $||\tilde{W}||_{C^1}$ to be bounded, we work with a conditional Logistic GP prior such that sample paths' C^1 norm is bounded almost surely. The prior mass condition induced by the conditional Logistic GP prior becomes a small ball probability in C^1 norm for the unconditional GP prior, which incurs a slow-down in posterior contraction rate.

Though the minimax rate of tail index estimation depends on the second order stochastic parameter β, the posterior contraction rate can be much faster than the

minimax rate. This is due to the assumption that $Y_i \overset{iid}{\sim} \mathbb{P}_{\theta_0,\psi_0}$ which constitutes a much smaller parameter space than the space used in minimax rate calculation. (See Figure 4.1 for an illustration.)

4.5 Discussion

We have established posterior contraction rate for density estimation and tail index estimation for the transformation based Bayesian semiparametric density model. The proofs follow Schwartz method by verifying a prior mass condition and existence of certain tests. Below, we offer some discussion on the reasonability and limitations of the asymptotic analysis.

4.5.1 GPD as the parametric family

Since the second order stochastic parameter of GPD depends on its tail index, the model space generated by the class of GPD is quite limited. As shown in Figure 4.1, the model space in (α, β) is a curve. In contrast, the whole possible space is a rectangle containing the curve.

In principle, reasonable choice of G_θ should have more flexible second order stochastic parameter, so that the resulting model space covers the whole possible space. But common parametric distribution families generally do not have flexible second order stochastic parameters. Among them, GPD is computationally feasible and empirically successful [TC21].

4.5.2 Optimality of the posterior contraction rate

As the model space is different from the one used in minimax rate calculation, minimax rate derived in the literature is not a meaningful benchmark. In fact, the posterior contraction rate can be much faster than the minimax rate when the true density is smooth. It remains unclear if our rate calculation obtains the best possible rate.

4.5.3 Model misspecification

As illustrated in Figure 4.1, it might be more reasonable to assume the true model is not exactly on the curve AOE. One possibility is the truth is still in the rectangle $ABCD$. In this case, correct specification is possible for suitable choice of the parametric family. The other possibility is the truth is above the rectangle $ABCD$. In this case, the transformation based semiparametric density model is misspecified in terms of (α, β). The current work needs to be extended along this line.

Chapter 5

Conclusion

5.1 Concluding remarks

In this dissertation, I have studied variable selection consistency of Gaussian process regression, community detection consistency of Bayesian stochastic block model, and posterior contraction rate of tail index of a transformation based Bayesian semiparametric density model. These projects contribute to the growing literature of asymptotic analysis of Bayesian methods.

Schwartz method is a neat tool in establishing posterior contraction rates, despite that a log factor often appears in the posterior contraction rates. Pursuing exact minimax optimal posterior contraction rates is challenging, especially for Bayesian nonparametric methods. In our applications, near minimax optimal posterior contraction rates suffice.

With posterior mass concentrated around the truth in some topology, we further explore if finer properties can be supported on the posterior distribution. The trick is in choosing a suitable topology, such that posterior concentration can distinguish different latent "structures", e.g., relevant variables selected and community assignment for each node.

The frequentist adventure of Bayesian methods has just begun. In the next section, I highlight some relevant extensions and directions to explore in the future.

5.2 Future Work

Variable selection for Bayesian nonparametric regression In the context of nonparametric regression, an alternative to Gaussian process priors is tree priors, e.g., BART and Bayesian CART [CGM98, CGM10, HLM20]. Recent work has shown optimal estimation accuracy of Bayesian CART [see, e.g., CR19, and the references therein.] Variable selection in regression tree has not been explored yet.

The hierarchical GP prior is essentially a spike-and-slab prior, so that irrelevant variables can be exactly zero'ed out. In principle, continuous shrinkage type of prior is also possible. One can specify anisotropic rescaling to the GP prior and determines the relevance of each variable by its corresponding rescaling parameter. Though it is called "automatic relevance determination" [RW06], optimal "automatic" thresholding rule in variable selection remains unclear.

Though the hierarchical GP prior is shown to achieve variable selection consistency, efficient algorithms for posterior sampling need further development.

Bayesian stochastic block model The actual performance of a-SBM under model misspecification is unclear. Indeed, real world networks could generally follow the pattern of assortativity but do not exactly obey assortativity. To remedy this, the

weakly assortative Bayesian SBM can be extended to degree corrected SBM to allow for influential communities or nodes.

Bayesian semiparametric density model The asymptotic analysis requires the truth to be generated from our model which constitutes a limited class of distributions. One direction is to study consistency and posterior contraction rate of our model, assuming the truth is outside our model space. Another direction is to extend the model to allow for more distributions by working with a more flexible parametric family.

Appendix A

Appendix for Chapter 2

The Appendix contains complete proofs for Theorem 2.4.1, Proposition 1, Lemma 13, Lemma 3, Lemma 4 in Chapter 2.

A.1 Proof of Theorem 2.4.1

Proof. It suffices to show there exist sets (sieve) $\mathbb{B}_n$, such that the following three conditions hold for all sufficiently large n:

$$\begin{aligned}
&\Pi_n\left(\|W^{A,\Gamma} - f_n^*\|_{L_2(Q_n)} \leq \varepsilon_n\right) \geq e^{-n\varepsilon_n^2} \\
&\Pi_n\left(W^{A,\Gamma} \notin \mathbb{B}_n\right) \leq e^{-4n\varepsilon_n^2} \\
&\log N\left(\varepsilon_n, \mathbb{B}_n, \|\cdot\|_{L_2(Q_n)}\right) \leq n\varepsilon_n^2.
\end{aligned}$$

Prior mass condition

Assumption $\beta > d_0/2$ implies $1/\varepsilon_n \geq C_H(a\xi)^{d_0/2}$ and $a\xi\log(1/\varepsilon_n) > d_0$ hold for every $a \in [C(1/\varepsilon_n)^{1/\beta}, 2C(1/\varepsilon_n)^{1/\beta}]$ with some constant C. We can apply Lemma 2 and Lemma 3 to obtain the following lower bound

$$\begin{aligned}
& \Pi_n\left(\|W^{A,\gamma_n^*} - f_n^*\|_{L_2(Q_n)} \leq 2\varepsilon_n\right) \\
\geq\ & \int_{K_n}^{2K_n} \Pi_n\left(\|W^{a,\gamma_n^*} - f_n^*\|_{L_2(Q_n)} \leq 2\varepsilon_n | a\right) \Pi_n\left(da\right) \\
\geq\ & \int_{K_n}^{2K_n} e^{-\phi_{f_n^*}^{a,\gamma_n^*}(\varepsilon_n)} \pi_n\left(a\right) da \\
\geq\ & \int_{K_n}^{2K_n} e^{-\left(Ca^{d_0} + C' a^{d_0}\log(a/\varepsilon_n)^{d_0+1}\right)} \pi_n\left(a\right) da \\
\geq\ & e^{-C\varepsilon^{-d_0/\beta}\log(1/\varepsilon_n)^{d_0+1}}
\end{aligned}$$

where $K_n = C(1/\varepsilon_n)^{1/\beta}$, $\varepsilon_n \asymp n^{-1/(2+d_0/\beta)}(\log n)^{\kappa_1}$ and $\kappa_1 = \frac{d_0+1}{2+d_0/\beta}$ such that quantity in the exponent satisfies $\varepsilon_n^{-d_0/\beta}\log(1/\varepsilon_n)^{d_0+1} \precsim n\varepsilon_n^2$. We can achieve

$$\Pi_n\left(\|W^{A,\gamma_n^*} - f_n^*\|_{L_2(Q_n)} \leq \varepsilon_n\right) \geq e^{-\frac{1}{2}n\varepsilon_n^2}$$

by choosing ε_n to be a large multiple of $n^{-1/(2+d_0/\beta)}(\log n)^{\kappa_1}$.

Therefore, with prior mass of model γ_n^* satisfying $\Pi_n\left(\Gamma = \gamma_n^*\right) = \Pi_n(|\Gamma| = d_0)\binom{d_n}{d_0}^{-1} \geq e^{-\frac{1}{2}n\varepsilon_n^2}$,

$$\begin{aligned}
& \Pi_n\left(\|W^{A,\Gamma} - f_n^*\|_{L_2(Q_n)} \leq \varepsilon_n\right) \\
\geq\ & \Pi_n\left(\Gamma = \gamma_n^*\right) \Pi_n\left(\|W^{A,\gamma_n^*} - f_n^*\|_{L_2(Q_n)} \leq \varepsilon_n\right) \\
\geq\ & e^{-n\varepsilon_n^2}.
\end{aligned}$$

Sieve construction

The sieve $\{\mathbb{B}_n\}$ is constructed as

$$\mathbb{B}_n = \bigcup\nolimits_{\gamma\in\{0,1\}^{d_n}:|\gamma|\leq \underline{d}_n} \mathbb{B}_{n,\gamma}$$

where $\underline{d}_n = C(n\varepsilon_n^2)^{1/\rho}$ for some constant C and $\mathbb{B}_{n,0^{d_n}} = [-M_n, M_n]$; for $\gamma \neq 0^{d_n}$,

$$\mathbb{B}_{n,\gamma} = M_n\sqrt{r_n}\mathcal{H}_1^{r_n,\gamma} + \varepsilon_n\mathbb{B}_{1,\gamma},$$

where $\mathbb{B}_{1,\gamma}$ is the unit ball in the Banach space $T_\gamma^{d_n}L_2(\mathbb{R}^{|\gamma|})$ indexed by γ. M_n, and r_n are specified such that $M_n^2 \asymp n\varepsilon_n^2$, and $r_n^{|\gamma|}\log^{|\gamma|+1}(n) \asymp n\varepsilon_n^2$. The choice of r_n depends on γ but for ease of notation γ is dropped. To apply Lemma 13 and Lemma 2, it requires $r_n^{|\gamma|} \precsim \varepsilon_n^{-2}$. Clearly, the choice of r_n satisfies this requirement.

Verifying condition $\Pi_n\left(W^{A,\Gamma} \notin \mathbb{B}_n\right) \leq e^{-4n\varepsilon_n^2}$

For $\gamma = 0^{d_n}$, the prior on the regression function is $N(0,1)$,

$$\begin{aligned}\Pi_n\left(W^{0^{d_n}} \notin \mathbb{B}_n\right) &\leq \Pi_n\left(W^{0^{d_n}} \notin \mathbb{B}_{n,0^d}\right)\\ &= 2\left(1-\Phi\left(M_n\right)\right)\\ &\leq \tfrac{2}{\sqrt{2\pi}M_n}e^{-M_n^2/2}\end{aligned}$$

where Φ denotes standard Normal cdf. By choosing M_n to be a large multiple of $n\varepsilon_n^2$, $\Pi_n\left(W^{0^{d_n}} \notin \mathbb{B}_n\right) \leq e^{-4n\varepsilon_n^2}$ holds for all sufficiently large n.

In light of Lemma 4.7 of [vdVvZ09], the nesting property

$$M_n\mathcal{H}_1^{a,\gamma} + \varepsilon_n\mathbb{B}_{1,\gamma} \subseteq B_{n,\gamma}$$

holds for every $a \in [1/\xi, r_n]$. By Borell's inequality, for every $a \in [1/\xi, r_n]$ and $\gamma \neq 0^{d_n}$,

$$\begin{aligned}\Pi_n\left(W^{a,\gamma} \notin \mathbb{B}_n\right) &\leq \Pi_n\left(W^{a,\gamma} \notin \mathbb{B}_{n,\gamma}\right)\\ &\leq \Pi_n\left(W^{a,\gamma} \notin M_n\mathcal{H}_1^{a,\gamma} + \varepsilon_n\mathbb{B}_{1,\gamma}\right)\\ &\leq 1-\Phi\left(\Phi^{-1}\left(e^{-\phi_0^{a,\gamma}(\varepsilon_n)}\right)+M_n\right)\\ &\leq 1-\Phi\left(\Phi^{-1}\left(e^{-\phi_0^{r_n,\gamma}(\varepsilon_n)}\right)+M_n\right)\end{aligned}$$

where last inequality is because $e^{-\phi_0^{a,\gamma}(\varepsilon_n)} = \Pi_n\left(\|W^{a,\gamma}\|_{L_2(Q_n)} \leq \varepsilon_n\right)$ is decreasing in a.

In light of Lemma 2, $M_n \geq 4\sqrt{\phi_0^{r_n,\gamma}(\varepsilon_n)}$ holds for all sufficiently large n. Since $e^{-\phi_0^{r_n,\gamma}(\varepsilon_n)} < 1/4$ holds for all small enough ε_n, then it follows $M_n \geq -2\Phi^{-1}\left(e^{-\phi_0^{r_n,\gamma}(\varepsilon_n)}\right)$ and the above inequality is further upper bounded by

$$1-\Phi\left(M_n/2\right) \leq e^{-M_n^2/8}.$$

So for every $\gamma \in \{0,1\}^{d_n} \backslash 0^{d_n}$, the following

$$\begin{array}{rl} & \Pi_n\left(W^{A,\gamma} \notin \mathbb{B}_n\right) \\ \leq & \Pi_n\left(W^{A,\gamma} \notin \mathbb{B}_{n,\gamma}\right) \\ \leq & \int_{1/\xi}^{r_n} \Pi_n\left(W^{a,\gamma} \notin M_n \mathcal{H}_1^{a,\gamma} + \varepsilon_n \mathbb{B}_{1,\gamma}\right) \pi_n\left(a|\gamma\right) da + \Pi_n\left(A > r_n|\gamma\right) \\ \leq & e^{-M_n^2/8} + e^{-C_2 r_n^{|\gamma|} \log^{|\gamma|+1}(r_n) + C_3 \log(|\gamma|)} \\ \leq & \frac{1}{2} e^{-4n\varepsilon_n^2} \end{array}$$

holds for all sufficiently large n. Hence,

$$\begin{array}{rcl} \Pi_n\left(W^{A,\Gamma} \notin \mathbb{B}_n\right) & = & \sum_{\gamma \in \{0,1\}^d} \Pi_n\left(W^{A,\gamma} \notin \mathbb{B}_n\right) \Pi_n\left(\Gamma = \gamma\right) \\ & \leq & \frac{1}{2} e^{-4n\varepsilon_n^2} + \sum_{\gamma \in \{0,1\}^{d_n} : |\gamma| > \underline{d}_n} \Pi_n\left(\Gamma = \gamma\right) \\ & \leq & e^{-4n\varepsilon_n^2}. \end{array}$$

Verifying condition $\log N\left(\varepsilon_n, \mathbb{B}_n, \|\cdot\|_{L_2(Q_n)}\right) \leq n\varepsilon_n^2$

Clearly, $N\left(\mathbb{B}_{n,0^{d_n}}, \varepsilon_n, \|\cdot\|_{L_2(Q_n)}\right) = 2M_n/\varepsilon_n$.

In light of Lemma 13, for $M_n\sqrt{r_n} > 2\varepsilon_n$, the metric entropy of $M_n\sqrt{r_n}\mathcal{H}_1^{r_n,\gamma} + \varepsilon_n \mathbb{B}_{1,\gamma}$ is bounded above:

$$\begin{array}{rl} & \log N\left(2\varepsilon_n, M_n\sqrt{r_n}\mathcal{H}_1^{r_n,\gamma} + \varepsilon_n \mathbb{B}_{1,\gamma}, \|\cdot\|_{L_2(Q_n)}\right) \\ \leq & \log N\left(\varepsilon_n, M_n\sqrt{r_n}\mathcal{H}_1^{r_n,\gamma}, \|\cdot\|_{L_2(Q_n)}\right) \\ \precsim & r_n^{|\gamma|} \log\left(M_n\sqrt{r_n}\varepsilon_n^{-1}\right)^{|\gamma|+1} / |\gamma|!. \end{array}$$

Since $\log\left(M_n\sqrt{r_n}\varepsilon_n^{-1}\right) \asymp \log\left(1/\varepsilon_n\right) \asymp \log(n)$, the above metric entropy is further bounded above by

$$\log N\left(2\varepsilon_n, M_n\sqrt{r_n}\mathcal{H}_1^{r_n,\gamma} + \varepsilon_n \mathbb{B}_{1,\gamma}, \|\cdot\|_{L_2(Q_n)}\right) \precsim r_n^{|\gamma|} \log(1/\varepsilon_n)^{|\gamma|+1} \asymp n\varepsilon_n^2.$$

The bound holds for all γ due to the choice of r_n.

Then note the bound $\log\left(\sum_{i=1}^n x_i\right) \leq \log n + \log\left(\max_i x_i\right)$, and the bound $\log(x \vee y) \leq \log(x) + \log(y)$ for $x, y > 1$, it follows,

$$\log N\left(\varepsilon_n, \mathbb{B}_n, \|\cdot\|_2\right) \precsim \log(2^{\underline{d}_n} - 1) + \frac{1}{2} n\varepsilon_n^2 + \log(\tfrac{2M_n}{\varepsilon_n})$$

where $\underline{d}_n \prec n\varepsilon_n^2$ by construction and ε_n is some multiple of $n^{-1/(2+d_0/\beta)} \log^{\kappa}(n)$.

□

A.2 Proof of Proposition 1

Before the proof of Proposition 1, first review some basic properties of Lambert W function. Lambert W function defines the product log function:

$$xe^x = y \Leftrightarrow x = W(y).$$

In the proof of Proposition 1, we need to solve an equation like $(x - c_1)e^x = c_2$ for x. With Lambert W function, the solution is in closed form: $x^* = W(c_2 e^{-c_1}) + c_1$. The following lemma is useful for our calculations; a proof can be found in the end of Appendix A.

Lemma 16. *The Lambert $W(\cdot)$ function is strictly increasing. If $x = W(y) > 1$, then $y > e^{W(y)} > y/\log(y)$. Furthermore, if $x = W(y) \in (0, 1]$, then $y \geq W(y) \geq y/e$.*

With the above preparation, here comes the proof of Proposition 1.

Proof of Proposition 1

Proof. As in Lemma 1 of [Cas08], Lemma 1 in [GvdV07], or Lemma 5 in [BSW99], it suffices to show for some ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, the following holds

$$\frac{\Pi_n(\Gamma \in \mathrm{FP}(\gamma_n^*), E_n)}{\Pi_n(B_n(f_n^*, \varepsilon_n))} = o\left(e^{-4n\varepsilon_n^2}\right) \tag{A.2.1}$$

where $E_n = \left\{f : \|f - f_n^*\|_{L_2(Q_n)} \leq M\varepsilon_n\right\}$ as defined in the main text, and $B_n(f_n^*, \varepsilon_n) = \left\{f : KL\left(P_{f_n^*}, P_f\right) \leq \varepsilon_n^2, V_{2,0}\left(P_{f_n^*}, P_f\right) \leq \varepsilon_n^2\right\}$ is the KL neighborhood of the truth. To prove (A.2.1), we show the following:

$$\begin{aligned}
&\textstyle\sum_{\gamma \in \mathrm{FP}(\gamma_n^*), |\gamma| \precsim \varepsilon_n^{-2/\alpha}} \Pi_n\left(\|W^{A,\gamma} - f_n^*\|_{L_2(Q_n)} \leq M\varepsilon_n\right) \leq e^{-5n\varepsilon_n^2} \\
&\textstyle\sum_{\gamma \in \mathrm{FP}(\gamma_n^*), |\gamma| \succsim \varepsilon_n^{-2/\alpha}} \Pi_n(|\Gamma| = |\gamma|) \leq e^{-5n\varepsilon_n^2} \\
&\Pi_n(B_n(f_n^*, \varepsilon_n)) \geq \Pi_n(B_n(f_n^*, \varepsilon_n) \cap \{\Gamma = \gamma_n^*\}) \geq e^{-n\varepsilon_n^2}
\end{aligned}$$

Denominator

In the nonparametric regression with Normal error, KL divergence is equivalent to $L_2(Q_n)$ norm: $K\left(P_{f_n^*}, P_f\right) = \frac{1}{2\sigma^2}\|f_n^* - f\|_{L_2(Q_n)}^2$ and the KL variation $V_{2,0}\left(P_{f_n^*}, P_f\right) = \frac{1}{\sigma^2}\|f_n^* - f\|_{L_2(Q_n)}^2$. Therefore, the denominator has the lower bound in $L_2(Q_n)$ norm:

$$\Pi_n(B_n(f_n^*, \varepsilon_n)) \geq \Pi_n\left(\|W^{A,\gamma_n^*} - f_n^*\|_{L_2(Q_n)} \leq \sigma\varepsilon_n\right)\Pi_n(\Gamma = \gamma_n^*)$$

where $\Pi_n(\Gamma = \gamma_n^*)$ is the prior probability of the true model. In light of the concen-

tration inequality, Lemma 2 and Lemma 3,

$$\begin{aligned}
& \Pi_n\left(\|W^{A,\gamma_n^*} - f_n^*\|_{L_2(Q_n)} \leq \sigma\varepsilon_n\right) \\
\geq & \int_{K_n}^{2K_n} \Pi_n\left(\|W^{a,\gamma_n^*} - f_n^*\|_{L_2(Q_n)} \leq \sigma\varepsilon_n | a\right) \Pi_n\left(da\right) \\
\geq & \int_{K_n}^{2K_n} e^{-\phi_{f_n^*}^{a,\gamma_n^*}(\sigma\varepsilon_n/2)} \Pi_n\left(da\right) \\
\geq & \int_{K_n}^{2K_n} e^{-\left(Ca^{d_0} + C'a^{d_0}\log(a/\varepsilon_n)^{d_0+1}\right)} \pi_n\left(a\right) da \\
\geq & e^{-C\varepsilon_n^{-d_0/\beta}\log(1/\varepsilon_n)^{d_0+1}}
\end{aligned}$$

where $K_n = C(1/\varepsilon_n)^{1/\beta}$, $\varepsilon_n \asymp n^{-1/(2+d_0/\beta)}(\log n)^{\kappa_1}$ and $\kappa_1 = \frac{|\gamma_n^*|+1}{2+d_0/\beta}$. Note for $\gamma_n^* = 0^{d_n}$, the prior on the mean is $W \sim N(0,1)$. Denote standard Normal density function as $\phi(\cdot)$, then

$$\begin{aligned}
\Pi_n\left(\|W - f_n^*\|_{L_2(Q_n)} \leq \sigma\varepsilon_n\right) &= \Pi_n\left(f_n^* - \sigma\varepsilon_n \leq f \leq f_n^* + \sigma\varepsilon_n\right) \\
&\geq 2\sigma\varepsilon_n \phi\left(\left(f_n^* + \sigma\varepsilon_n\right) \vee \left(f_n^* - \sigma\varepsilon_n\right)\right) \\
&\geq e^{-n\varepsilon_n^2}
\end{aligned}$$

holds for every sufficiently large n.

Numerator

In light of the concentration inequality (2.5.6),

$$\begin{aligned}
& \Pi_n\left(\|W^{A,\gamma} - f_n^*\|_{L_2(Q_n)} \leq M\varepsilon_n\right) \\
= & \int \Pi_n\left(\|W^{A,\gamma} - f_n^*\|_{L_2(Q_n)} \leq M\varepsilon_n | a\right) \Pi_n\left(da|\gamma\right) \\
\leq & \int_0^\infty e^{-\phi_{f_n^*}^{a,\gamma}(M\varepsilon_n)} \Pi_n\left(da|\gamma\right), \qquad \text{(A.2.2)}
\end{aligned}$$

where $\Pi_n(\cdot|\gamma)$ denotes the prior on A conditional on model index parameter γ. The sufficiently large constant M can be absorbed into ε_n by choosing it to be a large multiple of $n^{-1/(2+d_0/\beta)}(\log n)^{\kappa_1}$. For the ease of notation, henceforth M is absorbed into ε_n.

The integral (A.2.2) is further bounded by three parts:

$$\int_0^{1/\xi} e^{-\phi_{f_n^*}^{a,\gamma}(\varepsilon_n)} \Pi_n\left(da|\gamma\right) + \int_{1/\xi}^{\tau_n} e^{-\phi_{f_n^*}^{a,\gamma}(\varepsilon_n)} \Pi_n\left(da|\gamma\right) + \Pi_n\left(A > \tau_n|\gamma\right) \qquad \text{(A.2.3)}$$

where $\tau_n = C{\varepsilon_n}^{-1/\beta}$ for some constant C, ξ is a parameter of the dominating measure Q_n. The three quantities are bounded respectively as follows.

The first quantity of (A.2.3) is 0 due to the prior on A.

The third quantity of (A.2.3) has the desired upper bound $e^{-4n\varepsilon_n^2}$, because $\tau_n^{|\gamma|} \succ n\varepsilon_n^2$ for every $\gamma \in FP(\gamma_n^*)$.

For the second quantity of (A.2.3), by Lemma 4 and Lemma 2, the integral is

bounded above by

$$D_2\tau_n^{|\gamma|}\sup_{a:1/\xi\le a\le\tau_n}\exp\left\{-C(c_\xi a)^{|\gamma|}\varepsilon_n^2 e^{c\varepsilon_n^{-2/\alpha}/a^2}-C_2a^{|\gamma|}+C_3\log(|\gamma|)\right\} \tag{A.2.4}$$

where the prior density of $A|\Gamma=\gamma$ by assumption is upper bounded by the term $D_2\tau_n^{|\gamma|-1}e^{-C_2a^{|\gamma|}+C_3\log(|\gamma|)}$.

The solution to the supremum problem is $\tilde{a}=\sqrt{\frac{c_n}{W_n}}$ where $c_n=c\varepsilon_n^{-2/\alpha}$, $C_n=C_2C^{-1}c_\xi^{-|\gamma|}\varepsilon_n^{-2}$, $W_n=W\left(C_ne^{-|\gamma|/2}|\gamma|/2\right)+|\gamma|/2$, and W is Lambert W function. This is solved by solving first order condition and verifying second order condition. Then the quantity (A.2.4) is proportional to

$$\tau_n^{|\gamma|}\exp\left\{-C\varepsilon_n^2(c_\xi\tilde{a})^{|\gamma|}\left(e^{W_n}+C_n\right)+C_3\log(|\gamma|)\right\}. \tag{A.2.5}$$

In light of Lemma 16, there are two cases for $W(\cdot)$ and W_n. To apply Lemma 16, we need to solve $C_ne^{-|\gamma|/2}|\gamma|/2=e$ for $|\gamma|$. When $c_\xi^2e\le1$, $C_ne^{-|\gamma|/2}|\gamma|/2\ge e$ holds for all $|\gamma|>d_0$. When $c_\xi^2e>1$, there exists only one solution larger than d_0 for all sufficiently large $n>N(d_0,\xi)$. The solution is upper bounded by $\bar{\zeta}_n\equiv\frac{4}{1+2\log c_\xi}(1+\Delta)\log\left(C_2C^{-1}\varepsilon_n^{-1}\right)$ for any constant $\Delta>0$, and is lower bounded by $\underline{\zeta}_n\equiv\frac{4}{1+2\log c_\xi}\log\left(C_2C^{-1}\varepsilon_n^{-1}\right)$ for all sufficiently large $n>N(\Delta)\vee N(d_0,\xi)$.

The followup analysis is divided into three regimes. Regime I deals with the fixed dimension and slowly growing dimension case. Regime II and regime III take care of growing dimension cases with different growth rates.

Regime I: $W\left(C_ne^{-|\gamma|/2}|\gamma|/2\right)\ge1$**.**

In this regime, $C_ne^{-|\gamma|/2}|\gamma|/2\ge e$ and $d_0<|\gamma|<\bar{\zeta}_n\asymp\log(\varepsilon_n^{-1})$. By Lemma 16,

$$\frac{1}{2}|\gamma|C_n>e^{W_n}>\frac{1}{2}|\gamma|C_n/\left(\log\left(|\gamma|C_n/2\right)\right),$$

and $\log(C_n)+\log(|\gamma|)\gtrsim W_n\gtrsim\log(C_n)+\log(|\gamma|)-\log(\log\left(C_n\right)+\log(|\gamma|/2))$. Then the supremum (A.2.5) is upper bounded by

$$\exp\left\{-C_2\left(\frac{\varepsilon_n^{-2/\alpha}}{\log C_n+\log\left(|\gamma|\right)}\right)^{|\gamma|/2}+|\gamma|\log(\tau_n)+C_3\log(|\gamma|)\right\}.$$

Note $\bar{\zeta}_n>|\gamma|\ge d_0+1$, the integral (A.2.4) is further bounded above by

$$D_2\exp\left\{-C_2'\big(\varepsilon_n^{-2/\alpha}\log^{-1}(\varepsilon_n^{-1})\big)^{(d_0+1)/2}+\bar{\zeta}_n\log(\tau_n)+C_3\log(\bar{\zeta}_n)\right\}$$

for some constant C_2' and all sufficiently large n.

Since $\alpha<\beta\left(1+1/d_0\right)$ and $\bar{\zeta}_n\asymp\log\tau_n\asymp\log\left(\varepsilon_n^{-1}\right)$, the following holds for all

sufficiently large n

$$\begin{aligned}
&\textstyle\sum_{\gamma\in \mathrm{FP}(\gamma_n^*):W\left(C_n e^{-|\gamma|/2}|\gamma|/2\right)\geq 1} \Pi_n\left(\|W^{A,\gamma}-f_n^*\|_{L_2(Q_n)}\leq M\varepsilon_n\right)\\
\leq\ & 2^{\bar{\zeta}_n-d_0} D_2 e^{-C_2''\left(\varepsilon_n^{-2/\alpha}\log^{-1}\left(\varepsilon_n^{-1}\right)\right)^{(d_0+1)/2}}\\
\leq\ & \tfrac{1}{2}e^{-5n\varepsilon_n^2}.
\end{aligned}$$

Regime II: $W\left(C_n e^{-|\gamma|/2}|\gamma|/2\right)<1$ **and** $|\gamma|\precsim \varepsilon_n^{-2/\alpha}$

In this regime, $C_n < e^{|\gamma|/2+1}|\gamma|/2$, $|\gamma|>\underline{\zeta}_n$, and $|\gamma|/2\leq W_n\leq |\gamma|/2+1$. Then plugging in bounds on W_n, the supremum in (A.2.5) is upper bounded by $\exp\{-C_2\left(\frac{\varepsilon_n^{-2/\alpha}}{|\gamma|/2+1}\right)^{|\gamma|/2}+|\gamma|\log(\tau_n)+C_3\log(|\gamma|)\}$. Note $\varepsilon_n^{-2/\alpha}\succsim |\gamma|>\underline{\zeta}_n>d_0$, the integral (A.2.4) is further bounded above by

$$\exp\left\{-C_2'\left(\varepsilon_n^{-2/\alpha}/\left(\underline{\zeta}_n/2\right)\right)^{\underline{\zeta}_n/2}+\varepsilon_n^{-2/\alpha}\log(\tau_n)+C_3'\log(\varepsilon_n^{-1})\right\}$$

for some constant C_2' and all sufficiently large n.

Since $\alpha<\beta\left(1+1/d_0\right)$, $|\gamma|\precsim \varepsilon_n^{-2/\alpha}\prec n\varepsilon_n^2$, and $\underline{\zeta}_n\asymp \log\tau_n\asymp \log\left(\varepsilon_n^{-1}\right)$, the following holds for all sufficiently large n

$$\begin{aligned}
&\textstyle\sum_{\gamma\in \mathrm{FP}(\gamma_n^*):W\left(C_n e^{-|\gamma|/2}|\gamma|/2\right)<1} \Pi_n\left(\|W^{A,\gamma}-f_n^*\|_{L_2(Q_n)}\leq M\varepsilon_n\right)\\
\leq\ & 2^{C\varepsilon_n^{-2/\alpha}-\underline{\zeta}_n} D_2 e^{-C_2'\left(\varepsilon_n^{-2/\alpha}/(\underline{\zeta}_n/2)\right)^{\underline{\zeta}_n/2}}\\
\leq\ & \tfrac{1}{2}e^{-5n\varepsilon_n^2}.
\end{aligned}$$

Regime III: $W\left(C_n e^{-|\gamma|/2}|\gamma|/2\right)<1$ **and** $|\gamma|\succsim \varepsilon_n^{-2/\alpha}$

In this regime, $\Pi_n\left(|\Gamma|=|\gamma|\right)\leq e^{-C|\gamma|^\rho}$ for some $\rho\geq (d_0+1)/2$, the following holds for some constant C' and all sufficiently large n

$$\Pi_n\left(\Gamma\in \mathrm{FP}(\gamma_n^*):|\Gamma|\geq \varepsilon_n^{-2/\alpha}\right)\leq \Pi_n\left(|\Gamma|\geq \varepsilon_n^{-2/\alpha}\right)\leq e^{-C'\varepsilon_n^{-2\rho/\alpha}}\leq e^{-5n\varepsilon_n^2}.$$

Therefore, combining the three regimes and the rest ingredients,

$$\Pi_n\left(\Gamma\in \mathrm{FP}(\gamma_n^*),\{\|W^{A,\Gamma}-f_n^*\|_{L_2(Q_n)}\leq M\varepsilon_n\}\right)\leq 2e^{-5n\varepsilon_n^2}$$

and the ratio (A.2.1) holds.

□

A.3 Proof of Lemma 13

Proof. With some abuse of notation, denote (2.5.5) by $\mathcal{H}^{a,\gamma}$. By isometry, it suffices to compute the metric entropy of (2.5.5) with respect to the ℓ_2 metric $||\cdot||_2$.

The strategy of the proof is to construct a $\tilde{\mathcal{H}} \subset \mathcal{H}^{a,\gamma}$ and use the metric entropy of $\tilde{\mathcal{H}}$ as bounds for the metric entropy of $\mathcal{H}^{a,\gamma}$. For the ease of notation, superscript " (γ)" of the eigenvector and eigenvalue notation is dropped as the dependence on γ is clear.

The eigenvalue μ_j takes the form $(2v_1/V)^{|\gamma|/2}B^m$ for some $m \in \mathbb{N}$. Solve

$$(2v_1/V)^{|\gamma|/2}B^m = \varepsilon^2$$

for $m \in \mathbb{R}$ and the solution denoted by m^* is

$$m^* = (\log(1/B))^{-1}\left(2\log(1/\varepsilon) - \frac{|\gamma|}{2}\log\left(\frac{V}{2v_1}\right)\right).$$

The assumption $1/\varepsilon \geq C_H(\xi a)^{|\gamma|/2}$ guarantees $m^* \asymp (\log(1/B))^{-1}\log(1/\varepsilon)$.

Define $\tau = \sum_{j=0}^{\lfloor m^* \rfloor}\binom{|\gamma|+j-1}{|\gamma|-1} = \binom{\lfloor m^* \rfloor + |\gamma|}{|\gamma|}$ where $\lfloor m^* \rfloor$ is the greatest integer less than m^*. By construction, $\sqrt{\mu_j} \geq \varepsilon$ for all $j \leq \tau$ and $\sqrt{\mu_j} \leq \varepsilon$ for all $j > \tau$.

Define $\tilde{\mathcal{H}}_\varepsilon = \{\theta \in \mathcal{H}^{a,\gamma} : \theta_j = 0, \forall j > \tau\}$. $\tilde{\mathcal{H}}_\varepsilon$ contains the elements in $\mathcal{H}^{a,\gamma}$ whose components after τ vanish to 0. Any $\varepsilon-$cover of $\tilde{\mathcal{H}}_\varepsilon$ forms a $\sqrt{2}\varepsilon-$cover of $\mathcal{H}^{a,\gamma}$. To see this, suppose $\{\theta^k\}_1^N$ forms a $\varepsilon-$cover of $\tilde{\mathcal{H}}$, then for any $\theta \in \mathcal{H}^{a,\gamma}$,

$$\begin{aligned}\min_k ||\theta - \theta^k||_2^2 &= \min_k \sum_{j=1}^{\tau}\left(\theta_j - \theta_j^k\right)^2 + \sum_{j=\tau+1}^{\infty}\theta_j^2 \\ &\leq \varepsilon^2 + \mu_\tau \sum_{j=\tau+1}^{\infty}\theta_j^2/\mu_j \\ &\leq 2\varepsilon^2.\end{aligned}$$

As the $\sqrt{2}$ scaling does not matter in metric entropy calculation, it suffices to work with $\tilde{\mathcal{H}}$.

Upper bound

By construction, $\tilde{\mathcal{H}}_\varepsilon \supseteq B_2^{\tau+1}(\varepsilon)$ where $B_q^k(\varepsilon)$ denotes $k-$dimensional ℓ_q-ball of radius ε: $B_q^k(\varepsilon) = \left\{x : \sum_{i=1}^k x_i^q \leq \varepsilon^q; x_j = 0, \forall j > k\right\}$. $\tau + 1$ is due to eigenvalues start from μ_0,

Note $\tilde{\mathcal{H}}_\varepsilon + B_2^{\tau+1}(\varepsilon) \subseteq 2\tilde{\mathcal{H}}_\varepsilon$, standard volume argument yields

$$N_{[]}\left(\varepsilon, \tilde{\mathcal{H}}_\varepsilon, ||\cdot||_2\right) vol\left(B_2^{\tau+1}(\varepsilon/2)\right) \leq vol\left(2\tilde{\mathcal{H}}_\varepsilon\right)$$

where the volume $vol\left(2\tilde{\mathcal{H}}_\varepsilon\right) = 2^{\tau+1}vol\left(\tilde{\mathcal{H}}_\varepsilon\right) \leq 2^{\tau+1}\prod_{i=0}^{\tau}\sqrt{\mu_i}$ and the volume $vol\left(B_2^{\tau+1}(\varepsilon/2)\right) = (\varepsilon/2)^{\tau+1}vol\left(B_2^{\tau+1}(1)\right)$. Then, it follows

$$\log N_{[]}\left(\varepsilon, \tilde{\mathcal{H}}_\varepsilon, ||\cdot||_2\right) \leq 2(\tau+1)\log 2 + (\tau+1)\log(1/\varepsilon) + \frac{1}{2}\sum_{i=0}^{\tau}\log\mu_i$$

where $\log(\mu_0) = \frac{|\gamma|}{2}\log(2v_1/V)$, and $\log\mu_i = \log(\mu_0) + h\log B$ for $h = 1, ..., \lfloor m^* \rfloor$ and

$i \in [\binom{|\gamma|+h-1}{|\gamma|}, \binom{|\gamma|+h}{|\gamma|})$. The exponent h is the multiplicity of eigenvalues.

Note $B < 1$ and total multiplicity is $\sum_{j=0}^{\lfloor m^* \rfloor} j\binom{|\gamma|-1+j}{|\gamma|-1} \geq \sum_{j=1}^{\lfloor m^* \rfloor} \binom{|\gamma|-1+j}{|\gamma|-1} = \tau - 1$, it follows

$$\begin{aligned} \sum_{i=0}^{\tau} \log \mu_i &= \tfrac{|\gamma|}{2} \log\left(2v_1/V\right)(\tau+1) - \sum_{j=0}^{\lfloor m^* \rfloor} j\binom{|\gamma|-1+j}{|\gamma|-1} \log\left(1/B\right) \\ &\leq \tfrac{|\gamma|}{2} \log\left(2v_1/V\right)(\tau+1) - \log\left(1/B\right)(\tau-1). \end{aligned}$$

Note $a\xi \log(1/\varepsilon) > |\gamma|$ implies $m^* \asymp m^* + |\gamma|$, the with the definition of τ and $1/\varepsilon \geq C_H(\xi a)^{|\gamma|/2}$,

$$\begin{aligned} & \log N_{[]}\left(\varepsilon, \tilde{\mathcal{H}}_\varepsilon, ||\cdot||_2\right) \\ \precsim\ & (\tau+1)\left[\log(1/\varepsilon) - \tfrac{|\gamma|}{4}\log\left(\tfrac{V}{2v_1}\right)\right] - \tfrac{1}{2}(\tau-1)\log(1/B) \\ \precsim\ & (m^* + |\gamma|)^{|\gamma|} \log(1/\varepsilon)/|\gamma|! \\ \precsim\ & (\log(1/B))^{-|\gamma|} \log(1/\varepsilon)^{|\gamma|+1}/|\gamma|!. \end{aligned}$$

Lower bound

Since $\tilde{\mathcal{H}}_\varepsilon \subset \mathcal{H}^{a,\gamma}$, lower bound of $\tilde{\mathcal{H}}_\varepsilon$'s metric entropy also lower bounds $\mathcal{H}^{a,\gamma}$'s metric entropy. The fact $\tilde{\mathcal{H}}_\varepsilon \supseteq B_2^{\tau+1}(\varepsilon)$ implies

$$\begin{aligned} \log N\left(\tilde{\mathcal{H}}_\varepsilon, \varepsilon/2, ||\cdot||_2\right) &\geq \log N\left(B_2^{\tau+1}(\varepsilon), \varepsilon/2, ||\cdot||_2\right) \\ &\geq (\tau+1)\log 2 \\ &\asymp (\log(1/B))^{-|\gamma|} \log(1/\varepsilon)^{|\gamma|}/|\gamma|, \end{aligned}$$

and $\log N\left(\mathcal{H}^{a,\gamma}, \varepsilon, ||\cdot||_2\right) \gtrsim (\log(1/B))^{-|\gamma|} \log(1/\varepsilon)^{|\gamma|}/|\gamma|!$.

Bounding $\log(1/B)$

Note

$$\begin{aligned} \log\left(1/B\right) &= \log\left(1 + v_1/v_2 + v_3/v_2\right) \\ &= \log\left(1 + v_1/v_2 + \sqrt{\left(v_1/v_2\right)^2 + 2v_1/v_2}\right) \end{aligned}$$

where $v_1/v_2 = 1/\left(4a^2\xi^2\right)$.

When a is close to 0, $\log(1/B) \asymp \log(v_1/v_2) \asymp \log(1/a) \precsim 1/a$.

When a is sufficiently large, v_1/v_2 is close to 0 and $\sqrt{\left(v_1/v_2\right)^2 + 2v_1/v_2} \succ v_1/v_2$. With the relation $\log(1+x) \approx x$ for $x \approx 0$, it follows $\log(1/B)^{-1} \asymp a\xi$.

For more precise characterization, there exists constants C_1, C_2 and C_3 such that

$$C_1 a \leq \log(1/B)^{-1} \leq C_2 + C_3 a$$

holds for all $a > 0$. The constants only depend on ξ.

□

A.4 Proof of Lemma 3

Proof. The proof extends Lemma 7 of [vdVvZ11] and allows for rescaling and different dimensions. Following the representation theorem, for $\lambda_0 \in \mathbb{R}^{d_0}$, let

$$\psi(\lambda_0) = \frac{\hat{f}_0(\lambda_0)}{m_{a,\gamma_n^*}(\lambda_0)} 1_{(\lambda_0 \in \mathbb{R}^{d_0} : ||\lambda_0||_2 < K)},$$

then for every $t \in \mathbb{R}^{d_n}$, $f_n^*(t) - h_\psi(t) = \int_{\lambda_0 \in \mathbb{R}^{d_0} : ||\lambda_0||_2 \geq K} e^{i(\lambda_0, t_0)} \hat{f}_0(\lambda_0) \, d\lambda_0$. By Hölder 's inequality,

$$\begin{aligned} |h_\psi(t) - f_n^*(t)|^2 &\leq \int_{\mathbb{R}^{d_0}} |\hat{f}_0(\lambda_0)|^2 1(||\lambda_0||_2 \geq K) \, d\lambda_0 \\ &\leq ||f_0||^2_{H^\beta(\mathbb{R}^{d_0})} \sup_{\lambda_0 \in \mathbb{R}^{d_0}} (1 + ||\lambda_0||_2^2)^{-\beta} 1(||\lambda_0||_2 \geq K) \\ &\precsim ||f_0||^2_{H^\beta(\mathbb{R}^{d_0})} K^{-2\beta}. \end{aligned}$$

Therefore,

$$\begin{aligned} ||h_\psi - f_n^*||^2_{L_2(Q_n)} &= \int_{\mathbb{R}^{d_n}} |h_\psi(t) - f_n^*(t)|^2 g_{d_n}(t) \, dt \\ &\precsim ||f_0||^2_{H^\beta(\mathbb{R}^{d_0})} K^{-2\beta}. \end{aligned}$$

Choose $K^{-\beta} \propto \varepsilon$ such that $||h_\psi - f_n^*||_{L_2(Q_n)} \leq \varepsilon$.

$$\begin{aligned} ||h_\psi||^2_{\mathcal{H}^{a,\gamma_n^*}} &= \int_{\lambda_0 \in \mathbb{R}^{d_0} : ||\lambda_0||_2 < K} |\hat{f}_0(\lambda_0)|^2 m_{a,\gamma_n^*}(\lambda_0)^{-1} d\lambda_0 \\ &= \int_{\lambda_0 \in \mathbb{R}^{d_0} : ||\lambda_0||_2 < K} |\hat{f}_0(\lambda_0)|^2 (2\sqrt{\pi})^{d_0} a^{d_0} e^{\frac{1}{4a^2}||\lambda_0||_2^2} d\lambda_0 \\ &\leq ||f_0||^2_{H^\beta(\mathbb{R}^{d_0})} (2\sqrt{\pi})^{d_0} a^{d_0} e^{\frac{1}{4a^2}K^2}. \end{aligned}$$

Then it holds that

$$||h_\psi||^2_{\mathcal{H}^{a,\gamma_n^*}} \precsim \left(2\sqrt{\pi}\right)^{d_0} a^{d_0} e^{C\varepsilon^{-2/\beta}/a^2}$$

□

A.5 Proof of Lemma 4

Proof. By Parseval's identity, for $h_\psi \in \mathcal{H}^{a,\gamma}$,

$$||h_\psi - f_n^*||^2_{L_2(Q_n)} = ||(h_\psi - f_n^*)\sqrt{g_{d_n}}||_2^2 = ||\widehat{h_\psi \sqrt{g_{d_n}}} - \widehat{f_n^* \sqrt{g_{d_n}}}||_2^2.$$

Therefore, $||h_\psi - f_n^*||_{L_2(Q_n)} < \varepsilon$ implies $||\widehat{h_\psi \sqrt{g_{d_n}}} \chi_K - \widehat{f_n^* \sqrt{g_{d_n}}} \chi_K||_2 < \varepsilon$, where $\chi_K = \left\{\lambda \in \mathbb{R}^{|\gamma|} : ||\lambda||_2 > K\right\}$. Triangle inequality implies

$$||\widehat{h_\psi \sqrt{g_{d_n}}} \chi_K||_2 > ||\widehat{f_n^* \sqrt{g_{d_n}}} \chi_K||_2 - \varepsilon. \tag{A.5.1}$$

For the right hand side of inequality (A.5.1), denote $\mathbb{R}^{|\gamma|}\backslash\mathbb{R}^{d_0}$ as $\mathbb{R}^{d_1}$, with the assumption on f_0,

$$\begin{aligned}
& ||\widehat{f_n^*\sqrt{g_{d_n}}}\chi_K||_2^2 \\
\propto & \int\int_{(\lambda_0,\lambda_1)\in\mathbb{R}^{|\gamma|}:||\lambda_0||_2^2+||\lambda_1||_2^2>K^2} |\widehat{f_0\sqrt{g_{d_0}}}(\lambda_0)|^2|\widehat{\sqrt{g_{d_1}}}(\lambda_1)|^2 d\lambda_0 d\lambda_1 \\
\geq & \int\int_{(\lambda_0,\lambda_1)\in\mathbb{R}^{|\gamma|}:||\lambda_0||_2>K} |\widehat{f_0\sqrt{g_{d_0}}}(\lambda_0)|^2|\widehat{\sqrt{g_{d_1}}}(\lambda_1)|^2 d\lambda_0 d\lambda_1 \\
\gtrsim & \int_{\lambda\in\mathbb{R}^{d_0}:||\lambda_0||_2>K} ||\lambda_0||^{-(2\alpha+d_0)} d\lambda_0 \\
\gtrsim & (1/K)^{2\alpha}
\end{aligned}$$

where $K \propto \varepsilon^{-1/\alpha}$ is chosen such that $||\widehat{f_n^*\sqrt{g_{d_n}}}\chi_K||_2$ is lower bounded by 2ε.

By Lemma 16 of [vdVvZ11] and the representation of h_ψ,

$$\begin{aligned}
& ||\widehat{h_\psi\sqrt{g_{d_n}}}\chi_{2K}||_2 \\
= & ||((\psi m_{a,\gamma}) * \widehat{\sqrt{g_{d_n}}})\chi_{2K}||_2 \\
\leq & ||\psi m_{a,\gamma}\chi_K||_2||\widehat{\sqrt{g_{d_n}}}(1-\chi_K)||_1 + ||\psi m_{a,\gamma}||_2||\widehat{\sqrt{g_{d_n}}}\chi_K||_1.
\end{aligned}$$

The terms on the above right hand side are bounded in the following way,

$$\begin{aligned}
||\psi m_{a,\gamma}\chi_K||_2^2 &= \int_{\lambda\in\mathbb{R}^{|\gamma|}:||\lambda||_2>K} |\psi(\lambda)|^2 m_{a,\gamma}(\lambda)^2 d\lambda \\
&\leq m_{a,\gamma}(K)\int_{\lambda\in\mathbb{R}^{|\gamma|}:||\lambda||_2>K} |\psi(\lambda)|^2 m_{a,\gamma}(\lambda)\, d\lambda \\
&\leq (2\sqrt{\pi})^{-|\gamma|}a^{-|\gamma|}e^{-\frac{1}{4}K^2/a^2}||h_\psi||_{\mathcal{H}^{a,\gamma}}^2
\end{aligned}$$

$$\begin{aligned}
||\widehat{\sqrt{g_{d_n}}}(1-\chi_K)||_1 &= \int_{\lambda\in\mathbb{R}^{|\gamma|}:||\lambda||_2\leq K} \left(2\sqrt{2\pi}\xi\right)^{|\gamma|/2} e^{-\xi^2||\lambda||_2^2} d\lambda \\
&< \int_{\mathbb{R}^{|\gamma|}} \left(2\sqrt{2\pi}\xi\right)^{|\gamma|/2} e^{-\xi^2||\lambda||_2^2} d\lambda \\
&\precsim (2\sqrt{2\pi})^{|\gamma|/2}\xi^{-|\gamma|/2} < \infty
\end{aligned}$$

$$\begin{aligned}
||\psi m_{a,\gamma}||_2^2 &= \int_{\mathbb{R}^{|\gamma|}} |\psi(\lambda)|^2 m_{a,\gamma}(\lambda)^2 d\lambda \\
&\leq (2\sqrt{\pi})^{-|\gamma|}a^{-|\gamma|}\int_{\mathbb{R}^{|\gamma|}} |\psi(\lambda)|^2 m_{a,\gamma}(\lambda)\, d\lambda \\
&= (2\sqrt{\pi})^{-|\gamma|}a^{-|\gamma|}||h_\psi||_{\mathcal{H}^{a,\gamma}}^2
\end{aligned}$$

$$\begin{aligned}
||\widehat{\sqrt{g_{d_n}}}\chi_K||_1 &= \int_{\lambda\in\mathbb{R}^{|\gamma|}:||\lambda||_2>K} \left(2\sqrt{2\pi}\xi\right)^{|\gamma|/2} e^{-\xi^2||\lambda||_2^2} d\lambda \\
&\leq e^{-\xi^2K^2/8}\int_{\lambda\in\mathbb{R}^{|\gamma|}:||\lambda||_2>K} \left(2\sqrt{2\pi}\xi\right)^{|\gamma|/2} e^{-\frac{7}{8}\xi^2||\lambda||_2^2} d\lambda \\
&\leq e^{-\xi^2K^2/8}\int \left(2\sqrt{2\pi}\xi\right)^{|\gamma|/2} e^{-\frac{7}{8}\xi^2||\lambda||_2^2} d\lambda \\
&\precsim (2\sqrt{2\pi})^{|\gamma|/2}\xi^{-|\gamma|/2}e^{-\xi^2K^2/8}
\end{aligned}$$

Since a factor of 2 does not matter, combining bounds of RHS and LHS of inequality (A.5.1) yields the following inequality

$$a^{-|\gamma|/2}(\sqrt{2}/\xi)^{|\gamma|/2}\left(e^{-\frac{1}{8}K^2/a^2} + e^{-\frac{1}{8}\xi^2K^2}\right)||h_\psi||_{\mathcal{H}^{a,\gamma}} \precsim ||\widehat{h_\psi\sqrt{g_{d_n}}}\chi_{2K}||_2 \geq \varepsilon$$

which is

$$||h_\psi||^2_{\mathcal{H}^{a,\gamma}} \gtrsim \varepsilon^2 a^{|\gamma|} c_\xi^{|\gamma|} e^{\frac{1}{4}K^2\left(a^{-2}\wedge\xi^2\right)}$$

where $c_\xi = \xi/\sqrt{2}$.

□

A.6 Proof of Lemma 16

Proof. Let $f(x) = xe^x$ and f is a strictly increasing, smooth map: $[0,\infty) \to [0,\infty)$. So its inverse is also strictly increasing.

For the second claim, the first inequality follow because by the definition of $W(\cdot)$ and the assumption $x > 1$,

$$y = W(y)e^{W(y)} > e^{W(y)}.$$

That is, $\log(y) > W(y)$. Then, $\log(y)e^{W(y)} > W(y)e^{W(y)} = y$, which is the second inequality.

Finally, as $W(y) \in (0,1]$, by monotonicity, $W(y) \leq W(y)\, e^{W(y)} \equiv y \leq W(y)\, e$, that is,

$$y \geq W(y) \geq y/e.$$

□

Appendix B

Appendix for Chapter 3

The Appendix contains complete proofs for Lemma 5, 8, 9, 10 and 12, details of the sampler, and complete simulation results for all configurations.

B.1 Proofs

B.1.1 Proof of Lemma 5

Proof. Suppose $\theta \in \Theta_{k,\delta}$, it suffices to show $\theta \notin \Theta_{k',\delta'}$ for all $k' < k$ and $\delta' \geq 0$. Now prove the statement by contradiction.

If $\theta \in \Theta_{k',\delta'}$ for some $k' < k$ and $\delta' \geq 0$, then some nodes from some communities implied by θ are merged. But by construction of $\Theta_{k,\delta}$, between-community connectivity probabilities of θ are strictly less than corresponding within community connectivity probabilities. Therefore, once merged, the connectivity probabilities of the merged block are not identical. This is a contradiction.

□

B.1.2 Proof of Lemma 8

Proof. The Hellinger distance between two Bernoulli random variables satisfies

$$\begin{aligned} H^2\left(\mathbb{P}_{\theta_{ij}^0}, \mathbb{P}_{\theta_{ij}^1}\right) &= \tfrac{1}{2}\left[\left(\sqrt{\theta_{ij}^0} - \sqrt{\theta_{ij}^1}\right)^2 + \left(\sqrt{1-\theta_{ij}^0} - \sqrt{1-\theta_{ij}^1}\right)^2\right] \\ &= \tfrac{1}{2}\left[\left(\tfrac{1}{2}2|\sqrt{\theta_{ij}^0} - \sqrt{\theta_{ij}^1}|\right)^2 + \left(\tfrac{1}{2}2|\sqrt{1-\theta_{ij}^0} - \sqrt{1-\theta_{ij}^1}|\right)^2\right] \\ &\geq \tfrac{1}{4}\left(\theta_{ij}^0 - \theta_{ij}^1\right)^2 \end{aligned}$$

as θ_{ij}^0 and θ_{ij}^1 are in $[0,1]$, $|\sqrt{\theta_{ij}^0} + \sqrt{\theta_{ij}^1}| \leq 2$ and $|\sqrt{1-\theta_{ij}^0} + \sqrt{1-\theta_{ij}^1}| \leq 2$.

By independence, $P_\theta = \otimes_{i<j} P_{\theta_{ij}}$. Then, the Hellinger distance between $\mathbb{P}_{\theta^0}$ and

$\mathbb{P}_{\theta^1}$ satisfies

$$\begin{aligned} H^2\left(P_{\theta^0}, P_{\theta^1}\right) &= 2 - 2\prod_{i<j}\left(1 - \tfrac{1}{2}H^2\left(P_{\theta^0_{ij}}, P_{\theta^1_{ij}}\right)\right) \\ &\geq 2 - 2\prod_{i<j}\left(1 - \tfrac{1}{8}\left(\theta^0_{ij} - \theta^1_{ij}\right)^2\right) \\ &\geq 2 - 2\min_{i<j}\left(1 - \tfrac{1}{8}\left(\theta^0_{ij} - \theta^1_{ij}\right)^2\right) \\ &= \tfrac{1}{4}\max_{i<j}\left(\theta^0_{ij} - \theta^1_{ij}\right)^2 \\ &= \tfrac{1}{4}||\theta^0 - \theta^1||^2_\infty. \end{aligned}$$

□

B.1.3 Proof of Lemma 9

Proof. Note $\Theta_{k,\delta_n} = \cup_{Z\in\mathcal{Z}_{n,k}}\Theta^Z_{k,\delta_n}$, where $\Theta^Z_{k,\delta_n} = \left\{T(ZPZ^T) : P \in S_{k,\delta_n}\right\}$ denotes the Z slice of the parameter space.

By Lemma 7 and the assumption on δ_n and ε_n, node-wise connectivity probability matrix space can be simplified via

$$\left\{\theta : ||\theta - \theta^0||_\infty < \varepsilon_n\right\} = \left\{T(Z_0PZ_0^T) : ||P - P^0||_\infty < \varepsilon_n\right\}.$$

This relation implies the covering number $N\left(\varepsilon_n, \Theta^Z_{k,\delta_n}, ||\cdot||_\infty\right) \leq (1/\varepsilon_n)^{k(k+1)/2}$, and then union bound implies the covering number $N\left(\varepsilon_n, \Theta_{k,\delta_n}, ||\cdot||_\infty\right) \leq k^n(1/\varepsilon_n)^{k(k+1)/2}$.

By Lemma 5, $\Theta_{k,\delta}$ are non-overlapping for different k, then another union bound implies the statement (3.3.3).

□

B.1.4 Proof of Lemma 10

Proof. First recall some basic expansions from calculus. For $x_0 \in (0,1)$, define $f(x) = -x_0\log\frac{x}{x_0} - (1-x_0)\log\frac{1-x}{1-x_0}$ for $x \in [0,1]$. Taylor expand $f(x)$ around x_0:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) + \tfrac{1}{2}f''(x_0)(x - x_0)^2 + O\left(|x - x_0|^3\right) \\ &= \tfrac{1}{2x_0(1-x_0)}(x - x_0)^2 + O\left(|x - x_0|^3\right). \end{aligned}$$

For $x_0 = 0$, the above $f(x) = -\log(1-x)$ with the convention $0\log 0 = 0$. Its Taylor expansion around 0 is $f(x) = -\log(1-x) = x + O(x^2)$. For $x_0 = 1$, the above $f(x) = -\log(x)$ also with the convention $0\log 0 = 0$. Its Taylor expansion around 1 is $f(x) = -\log(x) = 1 - x + O\left((1-x)^2\right)$.

With $||\theta - \theta^0||_\infty \leq \varepsilon_n$ and the assumption on θ^0, expand KL divergence at θ^0,

$$\begin{aligned} KL\left(\mathbb{P}_{\theta^0}, \mathbb{P}_\theta\right) &= -\sum_{i<j:\theta^0_{ij}>0} \theta^0_{ij} \log \frac{\theta_{ij}}{\theta^0_{ij}} - \sum_{i<j:\theta^0_{ij}<1} \left(1-\theta^0_{ij}\right) \log \frac{1-\theta_{ij}}{1-\theta^0_{ij}} \\ &\leq (N_0 + N_1)\left(\varepsilon_n + O\left(\varepsilon_n^2\right)\right) + \frac{n(n-1)}{2} C_0^{-1}\left(\varepsilon_n^2 + O\left(|\varepsilon_n|^3\right)\right) \\ &\precsim n^2 \varepsilon_n^2 / C_0 \end{aligned}$$

where $N_0 = \#\left\{(i,j) : \theta^0_{ij} = 0, i<j\right\}$ denotes the number of zero entries in θ^0, and $N_1 = \#\left\{(i,j) : \theta^0_{ij} = 1, i<j\right\}$ denotes the number of one entries in θ^0.

To bound $V_{2,0}$, note the Taylor expansion of $f(x) = \log \frac{x}{1-x}$ around $x_0 \in (0,1)$ satisfies $f(x) = \log \frac{x}{1-x} = \log \frac{x_0}{1-x_0} + \frac{1}{x_0(1-x_0)}(x - x_0) + O\left((x-x_0)^2\right)$.

By independence of different entries and with $||\theta - \theta^0||_\infty \leq \varepsilon_n$, KL variation can be bounded similarly by an expansion of $f(x) = \log(x/(1-x))$:

$$\begin{aligned} V_{2,0}\left(\mathbb{P}_{\theta^0}, \mathbb{P}_\theta\right) &= \mathbb{P}_0 \left\{ \left[\sum_{i<j} \left(A_{ij} \log \frac{\theta_{ij}}{\theta^0_{ij}} + (1 - A_{ij}) \log \frac{1-\theta_{ij}}{1-\theta^0_{ij}} \right) + KL\left(\mathbb{P}_{\theta^0}, \mathbb{P}_\theta\right) \right]^2 \right\} \\ &= \sum_{i<j} \mathbb{P}_0 \left\{ \left[\left(A_{ij} \log \frac{\theta_{ij}}{\theta^0_{ij}} + (1 - A_{ij}) \log \frac{1-\theta_{ij}}{1-\theta^0_{ij}} \right) + KL\left(\mathbb{P}_{\theta^0_{ij}}, \mathbb{P}_{\theta_{ij}}\right) \right]^2 \right\} \\ &= \sum_{i<j} \theta^0_{ij}\left(1-\theta^0_{ij}\right) \left(\log \frac{\theta_{ij}}{1-\theta_{ij}} - \log \frac{\theta^0_{ij}}{1-\theta^0_{ij}} \right)^2 \\ &\precsim \sum_{i<j} \frac{1}{\theta^0_{ij}\left(1-\theta^0_{ij}\right)} \varepsilon_n^2 \\ &\precsim n^2 \varepsilon_n^2 / C_0 \end{aligned}$$

□

B.1.5 Proof of Lemma 12

Proof. By the dependence assumption made in Assumption 14, the prior mass has the factorization

$$\Pi_n\left(P \in S_{k_0,\tau_n} : ||P - P^0||_\infty < C_0 \varepsilon_n | K = k_0\right) \Pi_n\left(Z = Z_0 | K = k_0\right) \Pi_n\left(K = k_0\right). \tag{B.1.1}$$

Next, we bound individual components of (B.1.1) respectively.

To bound the first component of (B.1.1), the conditional indepence of the off-diagonal entries of P on the diagonal entries of P suggests the following factorization,

$$\begin{aligned} & \Pi_n\left(P \in S_{k_0,\tau_n} : ||P - P^0||_\infty < C_0 \varepsilon_n | K = k_0\right) \\ = \; & \Pi_n\left(\bigcap_{1 \leq a \leq b \leq k_0} E_{n,ab} | K = k_0\right) \\ = \; & \prod_{1 \leq a \leq k_0} \left\{ \int_{E_{n,aa}} \left[\prod_{1 \leq a < b \leq k_0} \Pi_n\left(E_{n,ab} | \{P_{aa}\}, K = k_0\right) \right] d\Pi_n(P_{aa} | K = k_0) \right\} \end{aligned}$$

where $E_{n,ab} = \{P_{ab} : |P_{ab} - P^0_{ab}| < C_0 \varepsilon_n\}$. As $P^0 \in S_{k_0,\delta_0}$, $P^0_{ab} < P^0_{aa} \wedge P^0_{bb} - \delta_0$. On the other hand, if $0 < P_{aa} \wedge P_{bb} - P^0_{aa} \wedge P^0_{bb} < C_0 \varepsilon_n$, $P^0_{ab} < P^0_{aa} \wedge P^0_{bb} - \delta_0 < P_{aa} \wedge P_{bb} - \tau_n$,

and (conditional) prior density of P_{ab} given $P_{aa} \wedge P_{bb}$ is positive on $(P^0_{ab} - C_0\varepsilon_n, P^0_{ab})$ for all $a, b \in [k_0]$.

By Assumption 14–(2), for $a < b \in [k_0]$, the prior probability $\Pi_n(E_{n,ab}|\{P_{aa}\}, K = k_0) \geq \frac{1}{2}|E_{n,ab}| \min_{P_{ab} \in E_{n,ab}} \pi_n(P_{ab}|\{P_{aa}\}, K = k_0, \delta) \gtrsim \varepsilon_n e^{-C\log(n)(P_{aa} \wedge P_{bb})}$ for some universal constant C. As $P_{aa} \in E_{n,aa}$ for $a \in [k_0]$, $P_{aa} \wedge P_{bb} \leq (P^0_{aa} \wedge P^0_{bb}) + C_0\varepsilon_n \leq ||P^0||_\infty + C_0\varepsilon_n$, which gives a bound independent of $\{P_{aa}\}$.

Similarly, Assumption 14 (2) implies $\Pi_n(E_{n,aa}|K = k_0) \gtrsim \varepsilon_n e^{-C\log(n)(P^0_{aa} + C_0\varepsilon_n)}$.

Therefore, combining the bounds for P_{ab}'s gives

$$\Pi_n\left(P \in S_{k_0,\tau_n} : ||P - P^0||_\infty < C_0\varepsilon_n | K = k_0\right) \gtrsim e^{Ck_0^2\log(\varepsilon_n) - Ck_0^2\log(n)(||P^0||_\infty + C_0\varepsilon_n)}$$

where k_0^2 has the same order as $\frac{1}{2}k_0(k_0+1)$ and is used for simpler notation, and the constant C is universal.

As $\varepsilon_n^2 \asymp \log(k_0)/n$ and $1 \gtrsim \log(k_0)/n$, $\log(n) \gtrsim -\log(\varepsilon_n)$. As $k_0 \precsim \sqrt{n}$, $k_0^2\log(n) \precsim n\log(k_0)$. Then, $\Pi_n\left(P \in S_{k_0,\tau_n} : ||P - P^0||_\infty < C_0\varepsilon_n | K = k_0\right) \gtrsim e^{-Cn\log(k_0)}$ for some constant C dependent on P^0.

To bound the second and the third component of (B.1.1), by Assumption 14 (3) and (4), there exists a universal constant C such that $\Pi_n\left(Z = Z_0 | K = k_0\right) \geq e^{-Cn\log(k_0)}$ and $\Pi_n\left(K = k_0\right) \geq e^{-Cn\log(k_0)}$.

Note $n^2\varepsilon_n^2 \asymp n\log(k_0)$, the right hand side of the inequality (3.3.4) can be replaced with $e^{-Cn\log(k_0)}$ and (3.3.4) holds for some constant C dependent on P^0.

□

B.2 Posterior Sampler

This section presents details of the Metropolis-Hastings algorithm used to draw posterior samples from (3.4.1). The proposal has two stages: in the first stage, sample (Z, K); in the second stage, sample P given (Z, K). The first stage is adapted from the allocation sampler [MMFH13].

At t^{th} iteration, the proposal $\Pi_{prop}\left(Z^*, K^* | A, Z^{(t)}, K^{(t)}, P^{(t)}\right)$ consists of the four steps MK, GS, M3 and AE with equal probability $\frac{1}{4}$. With proposal (Z^*, K^*), sample $P^*|(Z^*, A)$ by independently sampling each entry of P^* from the Beta distribution $Beta\left(O^*_{ab} + 1, n^*_{ab} - O^*_{ab} + 1\right)$. With proposal (P^*, Z^*, K^*), the acceptance rates in the allocation sampler regimes are computed.

B.2.1 MK

MK: choose "add" or "delete" one empty cluster with probability 1/2. If "add" move is chosen, randomly pick one community identifier from $[K + 1]$ for the new empty community and rename the others as necessary; if "delete" move is chosen, randomly

pick one community from $[K]$, delete the community if it is empty and abandon the MK move if it is not empty.

In the step MK, if "add" one empty community is chosen, accept the proposal with probability $\min\left(1, \frac{\Pi_n(P^*|Z^*)}{\Pi_n\left(P^{(t)}|Z^{(t)}\right)} \frac{K}{K^*} \frac{1}{n+K}\right)$; if "delete" one empty community is chosen, accept the proposal with probability

$$\min\left(1, \frac{\Pi_n\left(P^*|Z^*\right)}{\Pi_n\left(P^{(t)}|Z^{(t)}\right)} \frac{K}{K^*}(n+K-1)\right).$$

B.2.2 GS

GS: relabel a random node. First randomly pick i then generate $Z^*(i)$ according to $\Pi_{prop}(Z^*(i) = k) \propto \beta(Z^*, A)^{-1}\Pi(Z^*|K^*)$ where $K^* = K^{(t)}$, the prior probability $\Pi(Z^*|K^*) = \int \Pi(Z^*|\alpha, K^*)\Pi(\alpha|K^*)d\alpha = \frac{\Gamma(K^*)}{\Gamma(n+K^*)} \prod_{1\le c\le K} \Gamma(n_c^* + 1)$ due to multinomial-Dirichlet conjugacy, and $\beta(Z^*, A) = \prod_{1\le a\le b\le K} \frac{\Gamma(n_{ab}^*+2)}{\Gamma(O_{ab}^*+1)\Gamma(n_{ab}^*-O_{ab}^*+1)}$ is the coefficient corresponding to the proposal distribution of P. Clearly, $Z^*(j) = Z(j)$ for all $j \neq i \in [n]$.

In the step GS, suppose node i is chosen and its original label c_1 is relabeled with c_2, then accept the proposal with probability $\min\left(1, \frac{\Pi_n(P^*|Z^*)}{\Pi_n\left(P^{(t)}|Z^{(t)}\right)}\right)$.

B.2.3 M3

M3: randomly pick two communities $c_1, c_2 \in [K]$, reassign nodes $\{i : Z(i) \in \{c_1, c_2\}\}$ to $\{c_1, c_2\}$ sequentially according to the following scheme. Start with $B_0 = B_1 = \emptyset$ and A_0 being the sub-network without nodes from community c_1 and c_2, define the assignment $B_h = \{Z^*(x_i)\}_{i=1}^{h-1}$ with x_i being the node index of the i^{th} element in $\{i : Z(i) \in \{c_1, c_2\}\}$, define the sub-network $A_h = A_{h-1} \cup \{x_h\}$ by appending one more node, define the assignment $Z_{B_h}^{c_j}$ for the sub-network A_h as the assignment with the node x_h assigned to c_j, and define the size of communities in the sub-network A_{h-1} as $\{n_{h,c}\}_{c\in[K]}$. For $i \in [n_c]$, assign the i^{th} node of $\{i : Z(i) \in \{c_1, c_2\}\}$ to c_1 with probability $p_{B_i}^{c_1}$ and to c_2 with probability $p_{B_i}^{c_2} \equiv 1 - p_{B_i}^{c_1}$, where $\frac{p_{B_i}^{c_1}}{p_{B_i}^{c_2}} = \frac{\Pi(A_i, Z_{B_i}^{c_1}, K, P)}{\Pi(A_i, Z_{B_i}^{c_2}, K, P)} = \frac{\Pi(A_i|P, Z_{B_i}^{c_1})\Pi(P|Z_{B_i}^{c_1}, K)\Pi(Z_{B_i}^{c_1}|K)\Pi(K)}{\Pi(A_i|P, Z_{B_i}^{c_2})\Pi(P|Z_{B_i}^{c_2}, K)\Pi(Z_{B_i}^{c_2}|K)\Pi(K)} = \frac{\Pi(A_i|P, Z_{B_i}^{c_1})(n_{i,c_1}+1)}{\Pi(A_i|P, Z_{B_i}^{c_2})(n_{i,c_2}+1)}$.

To improve mixing, once c_1 and c_2 are drawn, shuffle $\{i : Z(i) \in \{c_1, c_2\}\}$ before the sequential reassignment. Therefore, the ordering of node indices in the sequential reassignment is random.

In the step M3, suppose community c_1 and c_2 are chosen, then accept the proposal with probability $\min\left(1, \frac{\Pi_n(P^*|Z^*)}{\Pi_n\left(P^{(t)}|Z^{(t)}\right)} \frac{\prod_{i=1}^{n_c} p_{B_i}^{Z(i)}}{\prod_{i=1}^{n_c} p_{B_i}^{Z^*(i)}} \frac{\Gamma\left(n_{c_1}^*+1\right)\Gamma\left(n_{c_2}^*+1\right)}{\Gamma\left(n_{c_1}^{(t)}+1\right)\Gamma\left(n_{c_2}^{(t)}+1\right)} \frac{\beta\left(Z^{(t)}, A\right)}{\beta(Z^*, A)}\right)$, where $n_c = n_{c_1} + n_{c_2}$.

B.2.4 AE

AE: merge two random clusters or split one cluster into two clusters with probability 1/2. If "merge" is chosen, randomly merge two clusters c_1 and c_2 with $Z^*(i) = c_1$ for all $i \in \{j : Z(j) \in \{c_1, c_2\}\}$ and $Z^*(i) = Z(i)$ for all $i \notin \{j : Z(j) \in \{c_1, c_2\}\}$. The proposal probability is $\binom{K}{2}^{-1}$. If "split" is chosen, randomly pick two cluster identifiers $\{c_1, c_2\}$ from $[K+1]$, renaming others' identifiers as necessary, and assign the nodes in cluster c_1 to the cluster c_2 with the random probability $p_c \sim U(0,1)$. By integrating out p_c, the proposal probability is $\frac{\Gamma\left(n_{c_1}+1\right)\Gamma\left(n_{c_2}+1\right)}{K(K+1)\Gamma(n_c+2)}$.

In the step AE, if "merge" two communities is chosen, accept the proposal with probability $\min\left(1, \frac{\Pi_n(P^*|Z^*)}{\Pi_n\left(P^{(t)}|Z^{(t)}\right)} \frac{K^{(t)}}{K^*} \frac{\beta\left(Z^{(t)},A\right)}{\beta(Z^*,A)} \frac{K^*+n}{n_{c_1}^{(t)}+1}\right)$; if "split" is chosen, accept the proposal with probability $\min\left(1, \frac{\Pi_n(P^*|Z^*)}{\Pi_n\left(P^{(t)}|Z^{(t)}\right)} \frac{K^{(t)}}{K^*} \frac{\beta\left(Z^{(t)},A\right)}{\beta(Z^*,A)} \frac{n_{c_1}^{(t)}+1}{K+n}\right)$.

B.3 Complete simulation results

This section provides complete simulation results. We choose $(k_0, n, \rho) \in \{3,5,7\} \times \{50,75\} \times \{\frac{1}{2}, 1\}$, and for each (k_0, n, ρ) configuration, 100 networks are generated from $SBM(Z_0, \rho P^0, n, k_0)$.

To reduce Monte Carlo error and reach reasonable mixing, the Metropolis-Hastings algorithm and the allocation sampler collect 2×10^4 posterior draws for each synthetic dataset after discarding first 10^4 draws as burn-in. Both algorithms are initialized at $K = 2$ and random membership assignment.

Table B.1: Bias and RMSE comparison, Case 1

			$\rho = 1/2$		$\rho = 1$	
k_0	n	Method	Bias	RMSE	Bias	RMSE
3	50	a-SBM	1.34	1.76	0.07	0.26
		c-SBM	-0.45	0.83	-0.03	0.22
		CLBIC	**-0.23**	**0.48**	**0.00**	**0.00**
		NCV	-0.55	0.95	**0.00**	**0.00**
	75	a-SBM	0.54	0.96	0.01	0.10
		c-SBM	-0.12	0.45	0.02	0.28
		CLBIC	**0.00**	**0.00**	**0.00**	**0.00**
		NCV	0.01	0.10	**0.00**	**0.00**
5	50	a-SBM	**-2.50**	**2.98**	0.71	1.20
		c-SBM	-3.68	3.72	-1.04	1.39
		CLBIC	-3.07	3.10	-1.56	1.81
		NCV	-3.96	3.96	-1.46	2.04
	75	a-SBM	**-1.08**	**1.99**	0.04	0.51
		c-SBM	-2.45	2.65	-0.31	0.79
		CLBIC	-2.51	2.58	**0.00**	**0.00**
		NCV	-3.78	3.85	**0.00**	**0.00**
7	50	a-SBM	-5.51	5.58	**-3.07**	**3.43**
		c-SBM	-5.88	5.89	-4.69	4.81
		CLBIC	**-5.20**	**5.23**	-4.82	4.85
		NCV	-6.00	6.00	-5.95	5.96
	75	a-SBM	**-4.57**	**4.72**	**-1.40**	**1.99**
		c-SBM	-5.36	5.40	-2.30	2.58
		CLBIC	-4.84	4.86	-3.41	3.50
		NCV	-5.98	5.98	-3.16	3.49

Table B.2: Bias and RMSE comparison, Case 2

			$\rho = 1/2$		$\rho = 1$	
k_0	n	Method	Bias	RMSE	Bias	RMSE
3	50	a-SBM	-1.90	1.93	-1.94	1.96
		c-SBM	-1.90	1.93	-1.90	1.93
		CLBIC	**-1.24**	**1.33**	**-1.32**	**1.41**
		NCV	-2.00	2.00	-2.00	2.00
	75	a-SBM	-1.94	1.95	-1.90	1.92
		c-SBM	-1.93	1.95	-1.62	1.84
		CLBIC	**-0.95**	**1.00**	**-0.96**	**1.00**
		NCV	-2.00	2.00	-2.00	2.00
5	50	a-SBM	-3.89	3.90	-3.98	3.98
		c-SBM	-3.91	3.93	-3.92	3.93
		CLBIC	**-3.35**	**3.39**	**-3.33**	**3.37**
		NCV	-4.00	4.00	-4.00	4.00
	75	a-SBM	-3.93	3.94	-3.97	3.97
		c-SBM	-3.96	3.97	-3.94	3.95
		CLBIC	**-2.97**	**2.97**	**-3.03**	**3.05**
		NCV	-4.00	4.00	-4.00	4.00
7	50	a-SBM	-5.91	5.92	-5.99	5.99
		c-SBM	-5.92	5.93	-5.94	5.95
		CLBIC	**-5.25**	**5.27**	**-5.31**	**5.33**
		NCV	-6.00	6.00	-6.00	6.00
	75	a-SBM	-5.95	5.95	-5.94	5.95
		c-SBM	-5.89	5.90	-5.87	5.89
		CLBIC	**-5.04**	**5.05**	**-4.99**	**5.00**
		NCV	-6.00	6.00	-6.00	6.00

Table B.3: Bias and RMSE comparison, Case 3

			$\rho = 1/2$		$\rho = 1$	
k_0	n	Method	Bias	RMSE	Bias	RMSE
3	50	a-SBM	-1.59	1.84	**0.08**	**0.35**
		c-SBM	-1.80	1.85	-0.09	0.57
		CLBIC	**-1.19**	**1.28**	-0.29	0.56
		NCV	-1.98	1.98	-0.33	0.88
	75	a-SBM	-1.11	1.55	0.04	0.20
		c-SBM	-1.32	1.55	0.03	0.39
		CLBIC	**-0.82**	**0.92**	**0.00**	**0.00**
		NCV	-1.91	1.94	**0.00**	**0.00**
5	50	a-SBM	-3.84	3.87	-3.57	3.63
		c-SBM	-3.96	3.96	-3.72	3.76
		CLBIC	**-3.30**	**3.33**	**-3.18**	**3.22**
		NCV	-4.00	4.00	-3.99	3.99
	75	a-SBM	-3.92	3.93	-1.98	**2.27**
		c-SBM	-3.95	3.96	**-1.97**	2.32
		CLBIC	**-2.96**	**2.98**	-2.81	2.84
		NCV	-4.00	4.00	-3.83	3.86
7	50	a-SBM	-5.92	5.93	-5.96	5.96
		c-SBM	-5.95	5.95	-5.94	5.94
		CLBIC	**-5.29**	**5.31**	**-5.30**	**5.32**
		NCV	-6.00	6.00	-6.00	6.00
	75	a-SBM	-5.94	5.95	-5.48	5.52
		c-SBM	-5.89	5.90	-5.34	5.38
		CLBIC	**-4.95**	**4.96**	**-4.96**	**4.97**
		NCV	-6.00	6.00	-6.00	6.00

Table B.4: Bias and RMSE comparison, Case 4

			$\rho = 1/2$		$\rho = 1$	
k_0	n	Method	Bias	RMSE	Bias	RMSE
3	50	a-SBM	**0.03**	1.33	**-0.62**	1.04
		c-SBM	-0.99	1.01	-0.85	0.94
		CLBIC	-1.11	1.28	-0.93	0.97
		NCV	-1.98	1.99	-0.80	0.93
	75	a-SBM	**-0.56**	1.08	-0.68	0.91
		c-SBM	-0.96	0.98	**-0.77**	**0.88**
		CLBIC	-0.89	**0.95**	-0.89	0.95
		NCV	-1.92	1.95	-0.92	0.97
5	50	a-SBM	**-1.99**	**2.29**	**-2.67**	**2.77**
		c-SBM	-2.99	3.00	-2.93	2.94
		CLBIC	-3.41	3.47	-2.80	2.85
		NCV	-3.99	3.99	-3.01	3.17
	75	a-SBM	**-2.39**	**2.62**	-2.76	2.85
		c-SBM	-2.95	2.96	-2.88	2.91
		CLBIC	-2.86	2.92	-2.69	2.74
		NCV	-3.91	3.93	**-2.63**	**2.73**
7	50	a-SBM	**-3.89**	**4.07**	**-4.51**	**4.64**
		c-SBM	-5.04	5.05	-4.94	4.95
		CLBIC	-5.44	5.48	-4.81	4.87
		NCV	-5.99	5.99	-5.53	5.59
	75	a-SBM	**-4.28**	**4.41**	-4.79	4.83
		c-SBM	-4.95	4.96	-4.98	4.98
		CLBIC	-4.83	4.87	**-4.66**	**4.69**
		NCV	-5.95	5.96	-4.80	4.83

Appendix C

Appendix for Chapter 4

C.1 Proofs

C.1.1 Proof of Theorem 4.3.1

Proof. The proof needs to establish the prior mass condition

$$\Pi_n\left((\theta, w) : KL(p_{\theta^*,\psi^*}, p_{\theta,\psi}) < \varepsilon_n^2, V_{2,0}(p_{\theta^*,\psi^*}, p_{\theta,\psi}) < \varepsilon_n^2\right) \geq e^{-Cn\varepsilon_n^2}$$

for some constant C, and construct a sieve $\mathcal{F}_n$ such that

$$\log N(\varepsilon_n, \mathcal{F}_n, h(\cdot,\cdot)) \leq n\varepsilon_n^2, \text{ and } \Pi_n(\mathcal{F}_n^C) \leq e^{-(C+4)n\varepsilon_n^2}.$$

By Lemma 12, verifying the prior mass condition is reduced to verify $\Pi_n((\theta, W^A) : ||\theta-\theta^*||_\infty \leq C\varepsilon_n, ||W^A-w^*||_\infty \leq C\varepsilon_n) \geq e^{-Cn\varepsilon_n^2}$ for some constant C only dependent on (θ^*, w^*). By Assumption 14, $\Pi_n(\theta : ||\theta-\theta^*||_\infty \leq C\varepsilon_n) \gtrsim \varepsilon_n^{-p} \geq e^{-Cn\varepsilon_n^2}$ holds for every sufficiently large n. As W^A follows a rescaled SE GP with rescaling parameter A, in a similar argument of the proof of Theorem 3.1 of [vdVvZ09], $\Pi_n(W^A : ||W^A - w^*||_\infty < C\varepsilon_n) \geq e^{-Cn\varepsilon_n^2}$ holds for every sufficiently large n.

The sieve is constructed as $\mathcal{F}_n = B_n^W \otimes B_n^\Theta$ where $B_n^W = M_n\sqrt{r_n/\delta}\mathcal{H}_1^{r_n} + \varepsilon_n\mathbb{B}_1 \cap \mathcal{B}_1$, $\mathbb{B}_1$ is the unit ball of $C[0,1]$ in $||\cdot||_\infty$, $\mathcal{B}_1$ is the unit ball of $C^2[0,1]$ in $||\cdot||_{C^2}$, and B_n^Θ is an ε_n/M_n-net of Θ in $||\cdot||_\infty$. For any $w \in B_n^W$, $||w||_{C^2} \leq \varepsilon_n + M_n\sqrt{r_n/\delta}$.

For (θ_i, w_i) and (θ_j, w_j) in $\mathcal{F}_n$ such that $||\theta_i - \theta_j||_\infty \leq M_n/\varepsilon_n$ and $||w_i - w_j||_\infty \leq \varepsilon_n$, $h(p_{\theta_i,w_i}, p_{\theta_j,w_j}) \leq 2\varepsilon_n$. Then, $\log N(\mathcal{F}_n, 3\varepsilon_n, h(\cdot,\cdot)) \leq \log N(B_n^\Theta, \varepsilon_n, ||\cdot||_\infty) + \log N(B_n^W, 2\varepsilon_n, ||\cdot||_\infty) \leq p\log(M_\Theta/\varepsilon_n) + \log N(M_n\sqrt{r_n/\delta}\mathcal{H}_1^{r_n}, \varepsilon_n, ||\cdot||_\infty) \leq 4n\varepsilon_n^2$.

By viewing W^a as a map in $C^2[0,1]$, $\Pi_n(W^a \notin B_n^W) = \Pi_n(W^a \notin M_n\sqrt{r_n/\delta}\mathcal{H}_1^{r_n} + \varepsilon_n\mathcal{B}_1)$. By Borell inequality, it follows $\Pi_n(W^a \notin B_n^W) \leq 1 - \Phi(\Phi^{-1}(e^{-\phi_0^r(\varepsilon_n)}) + M_n) \leq e^{-(C+4)n\varepsilon_n^2}$ for every sufficiently large n.

□

C.1.2 Proof of Lemma 12

Proof. Note $KL(p_{\theta^*,\psi^*}, p_{\theta,\psi}) = \mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi}}] = \mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}}]$ becomes the sum of the two terms: $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}]$ and $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}}]$. Next we bound them respectively.

In the first term $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}]$, ψ^* is fixed, and $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}] = \mathbb{P}_{\theta^*,\psi^*}[\log\frac{g_{\theta^*}(X)}{g_\theta(X)} +$

$\log\frac{\psi^*(G_{\theta^*}(X))}{\psi^*(G_\theta(X))}]$. For the parametric part, with Taylor expansion of $\log g_\theta(x)$ at θ^*,

$$\log\frac{g_{\theta^*}(x)}{g_\theta(x)} = -\sum\nolimits_{|\alpha|=1} \partial^\alpha \log g_{\theta^*}(x)(\theta-\theta^*)^\alpha - \mathcal{R}^g_{\theta^*,1}(\theta-\theta^*;x)$$

where α is a multi-index and $\mathcal{R}$ denotes the remainder term in Lagrange's form: $\mathcal{R}^g_{a,k}(h;x) \equiv \sum_{|\alpha|=k+1} \partial^\alpha \log g_{a+ch}(x) h^\alpha/\alpha!$ for some $c \in (0,1)$. As $X \sim \mathbb{P}_{\theta^*,\psi^*}$ and by Assumption 10, $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{g_{\theta*}(X)}{g_\theta(X)}] = -\mathbb{P}_{\theta^*,\psi^*}[\mathcal{R}^g_{\theta^*,1}(\theta-\theta^*;X)] \le \frac{1}{2}\mathbb{P}_{\theta^*,\psi^*}[M(X)]||\theta-\theta^*||_1^2$.

For the non-parametric part, with Taylor expansion of $\log\psi^*(G_\theta(x))$ at θ^*,

$$\log\frac{\psi^*(G_{\theta^*}(x))}{\psi^*(G_\theta(x))} = -\sum\nolimits_{|\alpha|=1} \partial^\alpha \log\psi^*(G_{\theta^*}(x))(\theta-\theta^*)^\alpha - \mathcal{R}^{\psi^*}_{\theta^*,1}(\theta-\theta^*;x)$$

where the remainder $\mathcal{R}^{\psi^*}_{a,k}(h;x) \equiv \sum_{|\alpha|=k+1} \partial^\alpha \log\psi^*(G_{a+ch}(x)) h^\alpha/\alpha!$ for some $c \in (0,1)$. As $X \sim \mathbb{P}_{\theta^*,\psi^*}$, by chain rule and Assumption 11, $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{\psi^*(G_{\theta^*}(x))}{\psi^*(G_\theta(x))}] = -\mathbb{P}_{\theta^*,\psi^*}\mathcal{R}^{\psi^*}_{\theta^*,1}(\theta-\theta^*;X) \le \frac{1}{2}(||\dot\phi^*||_\infty^2 + ||\ddot\phi^*||_\infty)\mathbb{P}_{\theta^*,\psi^*}[M_G(X)^2]||\theta-\theta^*||_1^2$ where $\phi^* \equiv \log\psi^*$.

Combining the above two parts, the first item $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}] \le C_0\varepsilon_n^2$, where the constant C_0 depends on $\theta^*, \psi^*, \{G_\theta\}$ and $\dim(\theta)$.

In the second term $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}}]$, θ is fixed and $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}}] = \mathbb{P}_{\theta,\psi^*}[\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}}] = \mathbb{P}_{\theta,\psi^*}[(\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}-1)\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}}] + KL(p_{\theta,\psi^*}, p_{\theta,\psi})$. By Cauchy-Schwartz inequality,

$$\mathbb{P}_{\theta,\psi^*}[(\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}-1)\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}}] \le \sqrt{\mathbb{P}_{\theta,\psi^*}[(\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}-1)^2]}\sqrt{\mathbb{P}_{\theta,\psi^*}[(\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}})^2]}.$$

Next, we bound the RHS respectively.

To bound the first component of the RHS, note $\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}} - 1 = e^{\log(p_{\theta^*,\psi^*})-\log(p_{\theta,\psi^*})} - 1$. Zeroth order Taylor expansion of $\log(p_{\theta,\psi^*})$ at θ^* is $\log(p_{\theta,\psi^*}(x)) = \log(p_{\theta^*,\psi^*}(x)) + \mathcal{R}^p_{\theta^*,0}(\theta-\theta^*;x)$ where the remainder term

$$\begin{aligned}\mathcal{R}^p_{a,k}(h;x) &\equiv \textstyle\sum_{|\alpha|=k+1} \frac{1}{\alpha!}\partial^\alpha \log(p_{a+ch,\psi^*}(x)) h^\alpha \\ &= \textstyle\sum_{|\alpha|=k+1} \frac{1}{\alpha!}\left(\partial^\alpha\log(g_{a+ch}(x)) + \partial^\alpha\log(\psi^*(G_{a+ch}(x)))\right) h^\alpha\end{aligned}$$

for some $c \in (0,1)$. By Assumption 12, $|\mathcal{R}^p_{\theta^*,0}(\theta-\theta^*;x)| \le (M_{e,1}(x) + ||\dot\phi^*||_\infty M_{e,2}(x))||\theta-\theta^*||_1 \le (M_{e,1}(x) + ||\dot\phi^*||_\infty M_{e,2}(x))\varepsilon_n$. By the inequality $e^x - 1 \le xe^x$ for all $x \ge 0$, $|\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}} - 1| \le |\mathcal{R}^p_{\theta^*,0}(\theta-\theta^*;x)| e^{|\mathcal{R}^p_{\theta^*,0}(\theta-\theta^*;x)|}$. Then by the integrability part of Assumption 12, $\mathbb{P}_{\theta,\psi^*}[(\frac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}-1)^2] \le C_0\varepsilon_n^2$ where C_0 depends on ψ^*, $\dot\phi^*$ and $\dim(\theta)$.

To bound the second component of the RHS, by Lemma 3.1 of [vdVvZ08a], $\mathbb{P}_{\theta,\psi^*}[(\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}})^2] = V_2(p_{\theta,\psi^*}, p_{\theta,\psi}) \precsim ||\log(p_{\theta,\psi^*}) - \log(p_{\theta,\psi})||_\infty^2 = ||\phi^* - \phi||_\infty^2 \precsim \varepsilon_n^2$.

For the KL divergence $KL(p_{\theta,\psi^*}, p_{\theta,\psi})$, by Lemma 3.1 of [vdVvZ08a], $KL(p_{\theta,\psi^*}, p_{\theta,\psi}) \precsim ||\log(p_{\theta,\psi^*}) - \log(p_{\theta,\psi})||_\infty^2 = ||\phi^* - \phi||_\infty^2 \precsim \varepsilon_n^2$.

Combining the Cauchy-Schwartz inequality and the KL divergence $KL(p_{\theta,\psi^*}, p_{\theta,\psi})$, the second item $\mathbb{P}_{\theta^*,\psi^*}[\log\frac{p_{\theta,\psi^*}}{p_{\theta,\psi}}] \leq C_0\varepsilon_n^2$, where the constant C_0 depends on ψ^*, ϕ^*, and $\dim(\theta)$.

Finally, with Lemma 3.1 of [vdVvZ08a],

$$\begin{aligned} V_{2,0}(p_{\theta^*,\psi^*}, p_{\theta,\psi}) &\leq V_2(p_{\theta^*,\psi^*}, p_{\theta,\psi}) \\ &\leq 2\left(\mathbb{P}_{\theta^*,\psi^*}[\log\tfrac{p_{\theta^*,\psi^*}}{p_{\theta,\psi^*}}]^2 + \mathbb{P}_{\theta^*,\psi^*}[\log\tfrac{p_{\theta,\psi^*}}{p_{\theta,\psi}}]^2\right) \\ &\leq 2\left(V_2(p_{\theta^*,\psi^*}, p_{\theta,\psi^*}) + ||\log\tfrac{\psi^*}{\psi}||_\infty^2\right) \\ &\precsim \varepsilon_n^2. \end{aligned}$$

□

C.1.3 Proof of Lemma 13

Proof. The spectral density of $1-d$ SEGP is $\mu(\lambda) = e^{-\lambda^2/4}/(2\sqrt{\pi})$ and the spectral density of W^a is $\mu_a(\lambda) = \mu(\lambda/a)/a$. By Lemma 4.1 of [vdVvZ09], elements in $\mathcal{H}^a$ are the real part of the following representation:

$$h(t) = \int_{-\infty}^{\infty} \eta(\lambda) e^{-it\lambda} \mu_a(\lambda) d\lambda \tag{C.1.1}$$

for $t \in [0,1]$ and some $\eta \in L_2(\mathbb{R}, \mu_a)$, and $||h||_{\mathcal{H}^a} = \inf ||\eta||_{2,\mu_a}$ where the infimum is over all η such that the representation (C.1.1) holds. The proof is to construct an ε-net over $\mathcal{H}_1^a$ with piecewise polynomials whose metric entropy is used as the upper bound.

Let $R = a_0/a$, $\{t_1, ..., t_m\}$ be an $R/2$-net of $[0,1]$ in $||\cdot||_\infty$, and $\cup_i B_i$ be a partition of $[0,1]$ such that $t_i \in B_i$ for $i = 1, ..., m$. Define the piecewise polynomials as $P = \sum_{i=1}^m P_{i,a_i} 1_{B_i}$ where each component $P_{i,a_i}(t) = \sum_{0\leq n\leq k} a_{i,n}(t - t_i)^n$ and the coefficients $a_{i,n}$ form an ε/R^{n-2}-net of $[-C_{a_0}/R^n, C_{a_0}/R^n]$ where $C_{a_0} = \sqrt{2}e^{2a_0^2}$. The order of $P_{i,a_i}(t)$ is $k \asymp \log(a/\varepsilon)$ such that $a^2k^2/2^k < \varepsilon$.

Clearly, the log cardinality of the piecewise polynomials is bounded above by

$$\log(\prod_i \prod_{n:n\leq k} \#\{a_{i,n}\}) \leq m\log(\prod_{n:n\leq k} \frac{2C_{a_0}/R^n}{\varepsilon/R^{n-2}}) \leq mk(\log(2C_{a_0}/\varepsilon) + 2\log(1/R)).$$

As $m \asymp R^{-1}$ and $R = a_0/a$, the log cardinality is further bounded above by $Ca\log(1/\varepsilon)\log(a/\varepsilon)$ for some universal constant C.

Next, we show for any $h \in \mathcal{H}_1^a$, there exists some P such that $||h - P||_{C^2} \leq 3K\varepsilon$ for some universal constant K, which is equivalent to show $||h - P||_\infty \leq K\varepsilon$, $||\dot{h} - \dot{P}||_\infty \leq K\varepsilon$, and $||\ddot{h} - \ddot{P}||_\infty \leq K\varepsilon$ hold simultaneously.

Pick an element $h \in \mathcal{H}_1^a$, apply Hölder inequality to the representation (C.1.1), it follows

$$|h(z)| \leq \int_{-\infty}^{\infty} |\eta(\lambda)| e^{|\operatorname{Im}(z)||\lambda|} \mu_a(\lambda) d\lambda \leq ||\eta||_{2,\mu_a} \sqrt{\int_{-\infty}^{\infty} e^{2|\operatorname{Im}(z)||\lambda|} \mu_a(\lambda) d\lambda} \leq \sqrt{2} e^{2a^2 |\operatorname{Im}(z)|^2}.$$

Therefore, $\{h : h \in \mathcal{H}_1^a\}$ can be extended to entire analytic functions on $\mathbb{C}$. For our purpose, we consider the extension to $\Omega = \{z \in \mathbb{C} : |\operatorname{Re}(z)| \leq 1, |\operatorname{Im}(z)| \leq R\}$. Let C_R be a circle of radius R in the complex plane around t_i, then by Cauchy formula,

$$|\frac{D^n h(t_i)}{n!}| = |\frac{1}{2\pi i} \oint_{C_R} \frac{h(z)}{(z-t_i)^{n+1}} dz| \leq \frac{\sqrt{2} e^{2a^2 R^2}}{R^n} \equiv \frac{C_{a_0}}{R^n}.$$

As an entire analytic function on Ω, h admits a Taylor expansion at t_i: $h(z) = \sum_{n \geq 0} \frac{D^n h(t_i)}{n!} (z - t_i)^n$. Its first derivative $\dot{h}(z) = \sum_{n \geq 1} \frac{D^n h(t_i)}{(n-1)!} (z - t_i)^{n-1} = \sum_{n \geq 0} \frac{D^n \dot{h}(t_i)}{n!} (z - t_i)^n$ and second derivative $\ddot{h}(z) = \sum_{n \geq 2} \frac{D^n h(t_i)}{(n-2)!} (z - t_i)^{n-2} = \sum_{n \geq 0} \frac{D^n \ddot{h}(t_i)}{n!} (z - t_i)^n$.

Therefore, for suitably chosen $\{a_{i,n}\}$ and t_i, for all $z \in B_i$,

$$|\sum_{n>k} \frac{D^n h(t_i)}{n!} (z - t_i)^n| \leq \sum_{n>k} \frac{C_{a_0}}{R^n} (R/2)^n = C_{a_0}/2^k;$$

$$|\sum_{n \leq k} \frac{D^n h(t_i)}{n!} (z - t_i)^n - P_{i,a_i}(z)| \leq \sum_{n \leq k} \frac{\varepsilon}{R^{n-2}} (R/2)^n \leq R^2 \varepsilon;$$

$$|\sum_{n \geq k} \frac{D^{n+1} h(t_i)}{n!} (z - t_i)^n| \leq \sum_{n \geq k} (n+1) \frac{C_{a_0}}{R^{n+1}} (R/2)^n = C_{a_0} R^{-1} (k+2)/2^{k-1};$$

$$|\sum_{1 \leq n \leq k} \frac{D^n h(t_i)}{n!} n (z - t_i)^{n-1} - \dot{P}_{i,a_i}(z)| \leq \sum_{1 \leq n \leq k} \frac{n\varepsilon}{R^{n-2}} (R/2)^{n-1} \leq KR\varepsilon;$$

$$|\sum_{n \geq k-1} \frac{D^{n+2} h(t_i)}{n!} (z - t_i)^n| \leq \sum_{n \geq k-1} (n+2)^2 \frac{C_{a_0}}{R^{n+2}} (R/2)^n \leq KR^{-2} k^2 / 2^k;$$

$$|\sum_{2 \leq n \leq k} \frac{D^n h(t_i)}{n!} n(n-1)(z - t_i)^{n-2} - \ddot{P}_{i,a_i}(z)| \leq \sum_{2 \leq n \leq k} \frac{n^2 \varepsilon}{R^{n-2}} (R/2)^{n-2} \leq K\varepsilon.$$

As k is chosen such that $a^2 k^2 / 2^k < \varepsilon$ and $R < 1$, there exists a universal constant K such that $||h - P||_{C^2} \leq 3K\varepsilon$. □

C.1.4 Proof of Lemma 15

Proof. By Hoeffding's inequality, for all n and $\varepsilon > 0$,

$$\mathbb{P}_{\theta,\psi}\left(|\hat{p}_k - S_{\theta,\psi}(k)| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}. \tag{C.1.2}$$

Therefore, $\mathbb{P}_{\theta_0,\psi_0}[T_n] \leq 2e^{-2n\varepsilon_n^2 S_{0,n}^2}$ holds for every n.

For Type II error, define $H_1^+ \equiv \{\mathbb{P}_{\theta,\psi} : \alpha_+(p_{\theta,\psi}) - \alpha_0 > M\varepsilon_n\}$ and $H_1^- \equiv \{\mathbb{P}_{\theta,\psi} : \alpha_+(p_{\theta,\psi}) - \alpha_0 < -M\varepsilon_n\}$. Hence, $\sup_{\mathbb{P}_{\theta,\psi}\in H_1\cap\mathbb{H}_n} \mathbb{P}_{\theta,\psi}[1-T_n] = \sup_{\mathbb{P}_{\theta,\psi}\in H_1^+\cap\mathbb{H}_n} \mathbb{P}_{\theta,\psi}[1-T_n] \vee \sup_{\mathbb{P}_{\theta,\psi}\in H_1^-\cap\mathbb{H}_n} \mathbb{P}_{\theta,\psi}[1-T_n]$.

For $\mathbb{P}_{\theta,\psi} \in H_1 \cap \mathbb{H}_n$, denote $\alpha = \alpha_+(p_{\theta,\psi})$,

$$\begin{aligned}
\mathbb{P}_{\theta,\psi}[1-T_n] &= \mathbb{P}_{\theta,\psi}(|\hat{p}_k - S_{0,n}| \leq \varepsilon_n S_{0,n}) \\
&= \mathbb{P}_{\theta,\psi}(|\hat{p}_k - S_{0,n}| \leq \varepsilon_n S_{0,n}, |\hat{p}_k - S_{\theta,\psi}(k)| \leq \varepsilon_n S_{0,n}) \\
&\quad + \mathbb{P}_{\theta,\psi}(|\hat{p}_k - S_{0,n}| \leq \varepsilon_n S_{0,n}, |\hat{p}_k - S_{\theta,\psi}(k)| > \varepsilon_n S_{0,n}) \\
&\leq 1\left(|S_{0,n} - S_{\theta,\psi}(k)| \leq 2\varepsilon_n S_{0,n}\right) + \mathbb{P}_{\theta,\psi}(|\hat{p}_k - S_{\theta,\psi}(k)| > \varepsilon_n S_{0,n}) \\
&\leq 2e^{-2n\varepsilon_n^2 S_{0,n}^2}
\end{aligned}$$

where the supremum of the first term over $H_1 \cap \mathbb{H}_n$ is 0 for all sufficiently large n, and the second item, by Hoeffding's inequality, is bounded above by $2e^{-n\varepsilon_n^2 S_{0,n}^2}$ for all $\mathbb{P}_{\theta,\psi} \in H_1 \cap \mathbb{H}_n$ and n.

Now we bound the first term. As $S_{0,n} > 0$, the first term becomes the indicator function $1\left(|1 - \frac{S_{\theta,\psi}(k)}{S_{0,n}}| \leq 2\varepsilon_n\right)$. By (4.4.2), $\frac{1}{3}\frac{C}{C_0}(\log(n))^{\alpha_0-\alpha} \leq \frac{S_{\theta,\psi}(k)}{S_{0,n}} \leq 3\frac{C}{C_0}(\log(n))^{\alpha_0-\alpha}$. Note C/C_0 is bounded for all $\mathbb{P}_{\theta,\psi} \in \mathbb{H}_n$, the ratio $\frac{S_{\theta,\psi}(k)}{S_{0,n}}$ is in the same order as $(\log(n))^{\alpha_0-\alpha}$ with proportionality constants involved universal for all $\mathbb{P}_{\theta,\psi} \in \mathbb{H}_n$.

If $\underline{\alpha} \leq \alpha \leq \alpha_0 - M\varepsilon_n$, $(\log(n))^{M\varepsilon_n} \precsim \frac{S_{\theta,\psi}(k)}{S_{0,n}} \precsim (\log(n))^{\alpha_0-\underline{\alpha}}$. Since $(\log(n))^{M\varepsilon_n} = e^{M\varepsilon_n \text{loglog}(n)} = 1 + M\varepsilon_n \text{loglog}(n) + o(\varepsilon_n)$, and $(\log(n))^{\alpha_0-\underline{\alpha}} \to \infty$ as $n \to \infty$,

$$\inf_{\mathbb{P}_{\theta,\psi}\in H_1^-\cap\mathbb{H}_n} |1 - \frac{S_{\theta,\psi}(k)}{S_{0,n}}| \succsim \text{loglog}(n)\varepsilon_n.$$

If $\alpha_0 + M\varepsilon_n \leq \alpha \leq \bar{\alpha}$, $(\log(n))^{-M\varepsilon_n} \succsim \frac{S_{\theta,\psi}(k)}{S_{0,n}} \succsim (\log(n))^{\alpha_0-\bar{\alpha}}$. Since $(\log(n))^{-M\varepsilon_n} = e^{-M\varepsilon_n \text{loglog}(n)} = 1 - M\varepsilon_n \text{loglog}(n) + o(\varepsilon_n)$, and $(\log(n))^{\alpha_0-\bar{\alpha}} \to 0$ as $n \to \infty$,

$$\inf_{\mathbb{P}_{\theta,\psi}\in H_1^+\cap\mathbb{H}_n} |1 - \frac{S_{\theta,\psi}(k)}{S_{0,n}}| \succsim \text{loglog}(n)\varepsilon_n.$$

Combining the above two cases, there exists a constant N_0 such that for all $n > N_0$, $\inf_{\mathbb{P}_{\theta,\psi}\in H_1\cap\mathbb{H}_n} |1 - \frac{S_{\theta,\psi}(k)}{S_{0,n}}| > 2\varepsilon_n$ and $\sup_{\mathbb{P}_{\theta,\psi}\in H_1\cap\mathbb{H}_n} 1\left(|1 - \frac{S_{\theta,\psi}(k)}{S_{0,n}}| \leq 2\varepsilon_n\right) = 0$.

□

C.1.5 Derivatives of GPD

First order derivatives of $\log g_\theta(x)$ w.r.t. θ:

- $\partial_\theta^{(1,0)} \log g_\theta(x) = -\log(1+\frac{x}{\alpha\sigma}) + (1+\alpha)(1+\frac{x}{\alpha\sigma})^{-1}\frac{x}{\sigma\alpha^2} = -\log(1+\frac{x}{\alpha\sigma}) + (1+\frac{1}{\alpha})(1-\frac{1}{1+\frac{x}{\alpha\sigma}})$.
- $\partial_\theta^{(0,1)} \log g_\theta(x) = -\sigma^{-1} + (1+\alpha)(1+\frac{x}{\alpha\sigma})^{-1}\frac{x}{\sigma^2\alpha} = -\sigma^{-1} + \sigma^{-1}(1+\alpha)(1-(1+\frac{x}{\alpha\sigma})^{-1}) = \sigma^{-1}(\alpha - (1+\alpha)(1+\frac{x}{\alpha\sigma})^{-1})$.

Second order derivatives of $\log g_\theta(x)$ w.r.t. θ:

- $\partial_\theta^{(2,0)} \log g_\theta(x) = -\frac{1}{\alpha^2} + \frac{1}{1+\frac{x}{\alpha\sigma}}\frac{x}{\sigma\alpha^2}(1+\frac{\sigma}{x} + (1+\frac{1}{\alpha})\frac{-1}{1+\frac{x}{\alpha\sigma}}) = -\frac{1}{\alpha^2} + \frac{1}{1+\frac{x}{\alpha\sigma}}\frac{x}{\sigma\alpha^2}(1+\frac{\sigma}{x} + (1+\frac{1}{\alpha})\frac{-1}{1+\frac{x}{\alpha\sigma}}) = -\frac{1}{\alpha^2} + \frac{1}{\alpha}\frac{(x/\sigma)^2+\alpha}{(x/\sigma+1)(x/\sigma+\alpha)}$.
- $\partial_\theta^{(1,1)} \log g_\theta(x) = \frac{1}{\sigma}\frac{(x/\sigma-1)x/\sigma}{(x/\sigma+\alpha)^2}$.
- $\partial_\theta^{(0,2)} \log g_\theta(x) = -\sigma^{-2}(\alpha - (1+\alpha)(1+\frac{x}{\alpha\sigma})^{-1}) + \sigma^{-1}(1+\alpha)(1+\frac{x}{\alpha\sigma})^{-2}\frac{-x}{\alpha\sigma^2} = \frac{1}{\sigma^2}(\frac{\alpha+1}{(\frac{x}{\sigma\alpha}+1)^2} - \alpha)$.

First order derivatives of $G_\theta(x)$ are

- $\partial_\theta^{(1,0)} G_\theta(x) = (1+\frac{x}{\alpha\sigma})^{-\alpha}\left(\log(1+\frac{x}{\alpha\sigma}) - \frac{x}{\sigma\alpha+x}\right)$
- $\partial_\theta^{(0,1)} G_\theta(x) = -(1+\frac{x}{\sigma\alpha})^{-(\alpha+1)}\frac{x}{\sigma^2}$.

Second order derivatives of $G_\theta(x)$ are

- $\partial_\theta^{(2,0)} G_\theta(x) = -(1+\frac{x}{\alpha\sigma})^{-\alpha}\left(\left(\log(1+\frac{x}{\alpha\sigma}) - \frac{x}{\sigma\alpha+x}\right)^2 + \frac{x^2}{\alpha(x+\sigma\alpha)^2}\right)$.
- $\partial_\theta^{(1,1)} G_\theta(x) = \frac{x}{\sigma^2}(1+\frac{x}{\sigma\alpha})^{-(\alpha+1)}\left(\log(1+\frac{x}{\alpha\sigma}) - \frac{\alpha+1}{\alpha}\frac{x}{\sigma\alpha+x}\right)$.
- $\partial_\theta^{(0,2)} G_\theta(x) = \frac{x}{\sigma^4}(1+\frac{x}{\sigma\alpha})^{-(\alpha+2)}\left(x(\alpha^{-1}-1)+2\sigma\right)$.

Bibliography

[Abb17] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[ABFX08] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

[ACBL13] Arash A Amini, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

[ADL13] Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013.

[AILvZ09] Frank Aurzada, Ildar A Ibragimov, MA Lifshits, and JH van Zanten. Small deviations of smooth stationary Gaussian processes. *Theory of Probability & Its Applications*, 53(4):697–707, 2009.

[AL+18] Arash A Amini, Elizaveta Levina, et al. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

[Alv01] MI Fraga Alves. A location invariant hill-type estimator. *Extremes*, 4(3):199–217, 2001.

[B+12] Adam D Bull et al. Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics*, 6:1490–1516, 2012.

[Bar89] Andrew R Barron. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, pages 107–124, 1989.

[BB04] Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *The annals of statistics*, 32(3):870–897, 2004.

[BC09] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[BDH74] August A Balkema and Laurens De Haan. Residual life time at great age. *The Annals of probability*, pages 792–804, 1974.

[Ber92] Simeon M Berman. The tail of the convolution of densities and its application to a model of HIV-latency time. *The Annals of Applied Probability*, 2(2):481–502, 1992.

[BHP98] Adrian Bowman, Peter Hall, and Tania Prvan. Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85(4):799–808, 1998.

[Bir84] L Birgé. Sur un théorème de minimax pour des variables indépendantes équidistribuées. *Probab. Math. Statist*, 3:259–282, 1984.

[BL08] Karine Bertin and Guillaume Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241, 2008.

[BPD14] Anirban Bhattacharya, Debdeep Pati, and David Dunson. Anisotropic function estimation using multi-bandwidth Gaussian processes. *Annals of statistics*, 42(1):352, 2014.

[BPPD15] Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

[BS16] Peter J Bickel and Purnamrita Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):253–273, 2016.

[BS17] Cristina Butucea and Natalia Stepanova. Adaptive variable selection in nonparametric sparse additive models. *Electronic Journal of Statistics*, 11(1):2321–2357, 2017.

[BSW99] Andrew Barron, Mark J Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

[BVDBS+15] Małgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103, 2015.

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[Cas08] Ismaël Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

[Cas12] Enrique Castillo. *Extreme value theory in engineering*. Elsevier, 2012.

[CD12] Laëtitia Comminges and Arnak S Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696, 2012.

[CGM98] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

[CGM10] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[CK15] Alexandra Carpentier and Arlene KH Kim. Adaptive and minimax optimal estimation of the tail coefficient. *Statistica Sinica*, pages 1133–1144, 2015.

[CKP14] Ismaël Castillo, Gérard Kerkyacharian, and Dominique Picard. Thomas Bayes' walk on manifolds. *Probability Theory and Related Fields*, 158(3-4):665–710, 2014.

[CL18] Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

[CPS10] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[CR19] Ismael Castillo and Veronika Rockova. Uncertainty quantification for Bayesian CART. *arXiv preprint arXiv:1910.07635*, 2019.

[CSHvdV15] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[DEDH89] Arnold LM Dekkers, John HJ Einmahl, and Laurens De Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, pages 1833–1855, 1989.

[DJVZ10] R De Jonge and JH Van Zanten. Adaptive nonparametric Bayesian inference using location-scale mixture priors. *The Annals of Statistics*, 38(6):3300–3320, 2010.

[dNGL12] Fernando Ferraz do Nascimento, Dani Gamerman, and Hedibert Freitas Lopes. A semiparametric bayesian approach to extreme value estimation. *Statistics and Computing*, 22(2):661–675, 2012.

[DPR08] J-J Daudin, Franck Picard, and Stéphane Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008.

[Dre01] Holger Drees. Minimax risk bounds in extreme value theory. *The Annals of Statistics*, 29(1):266–294, 2001.

[dZBK10] P de Zea Bermudez and Samuel Kotz. Parameter estimation of the generalized pareto distribution—part ii. *Journal of Statistical Planning and Inference*, 140(6):1374–1388, 2010.

[EKM13] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

[ET96] David Eric Edmunds and Hans Triebel. *Function spaces, entropy numbers, differential operators*, volume 120. Cambridge University Press, 1996.

[FSR19] Jairo Fúquene, Mark Steel, and David Rossell. On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(5):809–837, 2019.

[Gab09] Xavier Gabaix. Power laws in economics and finance. *Annu. Rev. Econ.*, 1(1):255–294, 2009.

[GBP19] Junxian Geng, Anirban Bhattacharya, and Debdeep Pati. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526):893–905, 2019.

[GGF+12] Prem K Gopalan, Sean Gerrish, Michael Freedman, David Blei, and David Mimno. Scalable inference of overlapping communities. *Advances in Neural Information Processing Systems*, 25:2249–2257, 2012.

[GGR99] Subhashis Ghosal, Jayanta K Ghosh, and RV Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist*, 27(1):143–158, 1999.

[GGvdV00] Subhashis Ghosal, Jayanta K Ghosh, and Aad W van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

[GLMZ16] Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. Optimal estimation and completion of matrices with biclustering structures. *The Journal of Machine Learning Research*, 17(1):5602–5630, 2016.

[GLvdV08] Subhashis Ghosal, Jüri Lember, and Aad van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.

[GLZ15] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

[GM93] Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[GM97] Edward I George and Robert E McCulloch. Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

[GM18] Chao Gao and Zongming Ma. Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *arXiv preprint arXiv:1811.06055*, 2018.

[GN02] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[GN10] Evarist Giné and Richard Nickl. Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170, 2010.

[GPB19] Prasenjit Ghosh, Debdeep Pati, and Anirban Bhattacharya. Posterior contraction rates for stochastic block models. *Sankhya A*, pages 1–29, 2019.

[GS08] Ion Grama and Vladimir Spokoiny. Statistics of extremes by oracle estimation. *The Annals of Statistics*, 36(4):1619–1648, 2008.

[GV21] Andrew Gelman and Aki Vehtari. What are the most important statistical ideas of the past 50 years? *Journal of the American Statistical Association*, (just-accepted):1–29, 2021.

[GvdV07] Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

[GvdV17] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[GvdVZ20] Chao Gao, Aad W van der Vaart, and Harrison H Zhou. A general framework for Bayes structured linear models. *The Annals of Statistics*, 48(5):2848–2878, 2020.

[GZ16] Chao Gao and Harrison H Zhou. Rate exact Bayesian adaptation with modified block priors. *The Annals of Statistics*, 44(1):318–345, 2016.

[GZFA10] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[HC15] P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.

[Hen85] Jõgi Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Annals of the Institute of Statistical Mathematics*, 37(2):235–240, 1985.

[HHW10] Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282, 2010.

[Hil75] Bruce M Hill. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174, 1975.

[HKK16] Kohei Hayashi, Takuya Konishi, and Tatsuro Kawamoto. A tractable fully Bayesian method for the stochastic block model. *arXiv preprint arXiv:1602.02256*, 2016.

[HLL83] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[HLM20] Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian additive regression trees: a review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278, 2020.

[HW84] Peter Hall and Alan H Welsh. Best attainable rates of convergence for estimates of parameters of regular variation. *The Annals of Statistics*, pages 1079–1084, 1984.

[HW85] Peter Hall and Alan H Welsh. Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, pages 331–341, 1985.

[IJS01] Hemant Ishwaran, Lancelot F James, and Jiayang Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456):1316–1332, 2001.

[JMS96] M Chris Jones, James S Marron, and Simon J Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.

[JPM01] Lancelot F James, Carey E Priebe, and David J Marchette. Consistent estimation of mixture complexity. *Annals of Statistics*, pages 1281–1296, 2001.

[Ker00] Christine Keribin. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66, 2000.

[KL93] James Kuelbs and Wenbo V Li. Metric entropy and the small ball problem for Gaussian measures. *Journal of Functional Analysis*, 116(1):133–157, 1993.

[KN11] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[KPN02] Richard W Katz, Marc B Parlange, and Philippe Naveau. Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304, 2002.

[KRvdV10] Willem Kruijer, Judith Rousseau, and Aad van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

[KT61] Andrei N Kolmogorov and Vladimir M Tihomirov. ε-entropy and ε-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.(Ser. 2)*, 17:277–364, 1961.

[KTV17] Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

[LAW16] Wenzhe Li, Sungjin Ahn, and Max Welling. Scalable MCMC for mixed membership stochastic blockmodels. In *Artificial Intelligence and Statistics*, pages 723–731. PMLR, 2016.

[LBA12] Pierre Latouche, Etienne Birmele, and Christophe Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115, 2012.

[LC86] Lucien M Le Cam. *Asymptotic methods in statistical theory*. Springer-Verlag New York, Inc., 1986.

[LC12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[LeC73] Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[Lei16] Jing Lei. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1):401–424, 2016.

[Ler92] Brian G Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, pages 1350–1360, 1992.

[LL99] Wenbo V Li and Werner Linde. Approximation, metric entropy and small ball estimates for Gaussian measures. *The Annals of Probability*, 27(3):1556–1578, 1999.

[LLD19] Cheng Li, Lizhen Lin, and David B Dunson. On posterior consistency of tail index for bayesian kernel mixture models. *Bernoulli*, 25(3):1999–2028, 2019.

[LLW06] Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284, 2006.

[LLZ20] Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020.

[LP09] Hannes Leeb and Benedikt M Pötscher. Model selection. In *Handbook of Financial Time Series*, pages 889–925. Springer, 2009.

[LPM+08] Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

[LW08] John Lafferty and Larry Wasserman. Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63, 2008.

[Mar07] Natalia Markovich. *Nonparametric analysis of univariate heavy-tailed data: research and practice*, volume 753. John Wiley & Sons, 2007.

[MH14] Jeffrey W Miller and Matthew T Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370, 2014.

[MH18] Jeffrey W Miller and Matthew T Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

[MMFH13] Aaron F McDaid, Thomas Brendan Murphy, Nial Friel, and Neil J Hurley. Improved Bayesian inference for the stochastic block model with application to large networks. *Computational Statistics & Data Analysis*, 60:12–31, 2013.

[MSL+11] A MacDonald, Carl John Scarrott, Dominic Lee, Brian Darlow, Marco Reale, and Glynn Russell. A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157, 2011.

[Nea96] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1996.

[New02] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[New03] Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003.

[New06] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[NF07] Agostino Nobile and Alastair T Fearnside. Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162, 2007.

[NG04] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[NH14] Naveen Naidu Narisetty and Xuming He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

[NS01] Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[PB15] Debdeep Pati and Anirban Bhattacharya. Optimal Bayesian estimation in stochastic block models. *arXiv preprint arXiv:1505.06794*, 2015.

[Pei14] Tiago P Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, 2014.

[Pei17] Tiago P Peixoto. Nonparametric Bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317, 2017.

[PI75] James Pickands III. Statistical inference using extreme order statistics. *Annals of statistics*, 3(1):119–131, 1975.

[PS10] Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian statistics*, 9:501–538, 2010.

[Ran71] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[Ray17] Kolyan Ray. Adaptive Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 45(6):2511–2536, 2017.

[RCY11] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[Res07] Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

[Ric84] John Rice. Bandwidth choice for nonparametric regression. *The Annals of Statistics*, pages 1215–1230, 1984.

[Rou10] Judith Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics*, 38(1):146–180, 2010.

[Rud87] Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 1987.

[RW06] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[RWY12] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427, 2012.

[SB10] James G Scott and James O Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619, 2010.

[Sch65] Lorraine Schwartz. On Bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26, 1965.

[SM12] Carl Scarrott and Anna MacDonald. A review of extreme value threshold es-timation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60, 2012.

[Ste04] Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3):386, 2004.

[STFP12] Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

[STG13] Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

[Sto82] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[SvdVvZ15] Botond Szabó, Aad W van der Vaart, and JH van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015.

[SYF17] D Franco Saldana, Yi Yu, and Yang Feng. How many communities are there? *Journal of Computational and Graphical Statistics*, 26(1):171–181, 2017.

[TAO06] Andrea Tancredi, Clive Anderson, and Anthony O'Hagan. Accounting for threshold uncertainty in extreme value estimation. *Extremes*, 9(2):87, 2006.

[TC21] Surya T Tokdar and Erika L Cunningham. Heavy-tailed density estimation. *XXXX*, 2021.

[TK93] VM Tikhomirov and AN Kolmogorov. ε-entropy and ε-capacity of sets in functional spaces. In *Selected works of AN Kolmogorov*, pages 86–170. Springer, 1993.

[Tok11] Surya T Tokdar. Dimension adaptability of Gaussian process models with variable selection and projection. *arXiv preprint arXiv:1112.0716*, 2011.

[vdPvdV18] SL van der Pas and AW van der Vaart. Bayesian community detection. *Bayesian Analysis*, 13(3):767–796, 2018.

[vdVvZ07] Aad van der Vaart and Harry van Zanten. Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics*, 1:433–448, 2007.

[vdVvZ08a] Aad W van der Vaart and J Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, pages 1435–1463, 2008.

[vdVvZ08b] Aad W van der Vaart and J Harry van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics, 2008.

[vdVvZ09] Aad W van der Vaart and J Harry van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, pages 2655–2675, 2009.

[vdVvZ11] Aad van der Vaart and Harry van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119, 2011.

[Ver12] Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[VW98] Isabella Verdinelli and Larry Wasserman. Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *The Annals of Statistics*, 26(4):1215–1241, 1998.

[vWK20] J van Waaij and BJK Kleijn. Uncertainty quantification in the stochastic block model with an unknown number of classes. *arXiv preprint arXiv:2005.01362*, 2020.

[Wai09] Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[WB17] YX Rachel Wang and Peter J Bickel. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2):500–528, 2017.

[Weg03] Marten Wegkamp. Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273, 2003.

[Yan99] Yuhong Yang. Model selection for nonparametric regression. *Statistica Sinica*, pages 475–499, 1999.

[Yan05] Yuhong Yang. Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[YAT16] Zhao Yang, René Algesheimer, and Claudio J Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, 6(1):1–18, 2016.

[YT15] Yun Yang and Surya T Tokdar. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674, 2015.

[YWJ16] Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

[Zha91] Ping Zhang. Variable selection in nonparametric regression with continuous covariates. *The Annals of Statistics*, pages 1869–1882, 1991.

[ZLZ11] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.

[ZLZ12] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

[ZP20] Lizhi Zhang and Tiago P Peixoto. Statistical inference of assortative community structures. *Physical Review Research*, 2(4):043271, 2020.

[ZZ+16] Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

Biography

Sheng Jiang received a Bachelor's degree in economics and a Bachelor's degree in English from Tsinghua University, Beijing, China, in July 2012. In May 2015, he received a Master's degree in Statistical and Economic Modeling from Duke University, Durham, NC, USA. Since August 2015, he started doctoral study in statistics at Duke University.

Sheng's undergraduate project "Estimating Returns to Education in Urban China: Evidence from a Natural Experiment in Schooling Reform" is published in *Journal of Comparative Economics*, 2020. Sheng provides statistical analysis to the paper "Occurrence and distribution of hexavalent chromium in groundwater from North Carolina, USA" which is published in *Science of The Total Environment*, 2020. The first project of the dissertation is accepted at *The Annals of Statistics*.

In 2019, he received Travel Award for Bayesian Nonparametrics Conference (BNP12). In the academic year of 2017-2018, he received graduate fellowship from The Statistical and Applied Mathematical Sciences Institute (SAMSI).