Bayesian inference in high-dimensional models

Subhashis Ghoshal,
North Carolina State University

O' Bayes Meeting, September 6, 2022, Santa Cruz, CA

Acknowledgment

- Based on a review paper with **Sayantan Banerjee** and **Ismaël Castillo**: To appear in *Springer Volume on Data Science* in collaboration with International Indian Statistical Association. Available at arXiv:2101.04491
- Material enriched by collaboration with Eduard Belitser and many former students (Sayantan Banerjee, Weining Shen, Maggie Du, Jami Jackson Mulgrave, Jennifer Wei, Wenli Shi, Rui Zhu)
- **National Science Foundation** and **Army Research Office** for fund

Introduction

High-dimensional statistics

- Dimension of the observed variable X is large.
- The dimension of the parameter θ controlling the distribution of the observed data is large.
- The computational complexity should be sub-linear, linear or slow-polynomial in the dimension, not exponential.
- For theoretical study, like convergence theory, the formulation is asymptotic; $p = \dim(\theta) \to \infty$ as the sample size $n \to \infty$.
- This tutorial will discuss formulation of Bayesian methods, their convergence issues, and computational aspects.

Motivation

- Advances in technology have resulted in massive datasets collected from all aspects of modern life. Very large datasets appear from internet searches, mobile apps, social networking, cloud-computing, wearable devices, as well as from more traditional sources such as bar-code scanning, satellite imaging, air traffic control, banking, finance, and genomics.
- Due to the complexity of such datasets, flexible models are needed involving many parameters, routinely exceeding the sample size.

Key assumption: Hidden low-dimensional structure

In the high-dimensional situation, a meaningful inference is possible only if there is a hidden lower-dimensional structure involving far fewer parameters.

- **Many normal means:** $X_i \sim \mathrm{N}(\theta_i, \sigma^2)$, most θ_i are 0.
- **Linear regression model:** $Y = \langle X, \beta \rangle + \varepsilon$. The vector of regression coefficients β has most entries zero.
- **Generalized linear model:** $Y \sim g(\cdot; \langle X, \beta \rangle)$, most components of β are 0.
- **Change-point model:** Parametric distribution of X_i, such as $\mathrm{N}(\theta_i, \sigma^2)$, changes at a certain points $i_1 < i_2 < \cdots < i_s$ but remains the same in between.

- **Graphical model:** In the interrelationship among a large class of variables, only a few pairs are directly related — typically most pairs of variables are conditionally independent given other variables. The underlying sparsity is very conveniently described by a graph, where the variables are represented by the nodes of a graph, and an edge connecting a pair is present only if they are conditionally dependent given other variables.
- **Gaussian graphical model:** When the variables are jointly Gaussian, the absence of an edge is equivalent to having a zero-entry in the precision (inverse of the covariance) matrix $\Omega = \Sigma^{-1}$.
- **Matrix completion:** Many entries of a matrix are missing and it is assumed that the underlying true matrix has a sparse plus a low-rank structure. Often known as the Netflix Problem.
- **Stochastic block model:** The extent of interaction between two nodes is determined solely by their memberships in certain hidden blocks of nodes: $\pi_{ij} = g(C_i, C_j)$, C_i is the block containing i.

Penalty approach

- Most non-Bayesian approaches use penalty functions in optimizing an objective function that encourages a lower-dimensional structure like sparsity.
- For instance, in a regression problem an ℓ_1-penalty is put in the least square optimization to lead to the LASSO: $\arg\min\{\|Y - X\beta\|^2 + \lambda\|\beta\|_1\}$.
- For a Gaussian graphical model, the graphical LASSO $\arg\min\{n\,\mathrm{tr}(S\Omega) - n\log\det\Omega + \lambda\|\Omega\|_1\}$, where S is the sample dispersion matrix.

Spike-and-slab prior

- Mitchell and Beauchamp (1988, JASA), Ishwaran and Rao (2005, AoS):

$$\pi(\theta) = (1 - w)\phi_0(\theta) + w\phi_1(\theta),$$

where ϕ_0 is a density highly concentrated at 0, ϕ_1 is a density (usually symmetric about 0) allowing intermediate and large values of θ, and w is a small parameter thus inducing sparsity in the mixture.

- For instance, ϕ_0 may be the normal density with mean 0 and a small variance v_0 and ϕ_1 the normal density with mean zero and a relatively large variance v_1.
- Extreme possibility $v_0 = 0$ will be termed as *hard-spike-and-slab prior*; otherwise *soft-spike-and-slab prior*.
- The parameters in ϕ_0 and ϕ_1 as well as w are usually given further priors.

- **Spike-and-slab LASSO**: Ročkovà and George (2018, JASA) — Both ϕ_0 and ϕ_1 are Laplace densities, The posterior mode is a sparse vector as in the LASSO.
- Generally, the spike part of the prior induces a shrinkage towards zero, which can be limited by using a heavier tailed density ϕ_1 for the slab such as a t-density, or at least as heavy-tailed as the Laplace density.
- **Non-local priors**: Johnson and Rossell (2010, JRSS B) — the spike as separated as possible from the slab around zero, by choosing slab distributions that have very little mass close to 0.

Posterior computation with a spike-and-slab prior

- **Stochastic Search Variable Selection (SSVS)**: George and McCulloch (1993, JASA) — The key technique is to introduce latent indicator variables $(\gamma_1, \ldots, \gamma_n)$ for spike $(\gamma = 0)$ or slab $(\gamma = 1)$ and use Gibbs sampling. The posterior automatically visits the models with substantial posterior probabilities out of total 2^n models — visiting all of them is not needed. Non-degenarate spike avoids reversible jump MCMC, but zeros will have to be set using the indicators.
- **EMVS**: Ročkovà and George (2014, JASA) — cheaper alternative to SSVS. Computes the posterior mode model by EM algorithm integrating out the variables, but cannot compute the model posterior probabilities.

Continuous shrinkage priors

- Computation using a spike-and-slab prior involves a latent indicator of the mixture component. Replacing the indicator by a continuous variable leads to the so-called *continuous shrinkage priors*, typically obtained as scale mixtures of normal.
- **Bayesian LASSO**: Park and Casella (2008, JASA), Hans (2009, Biometrika) — Laplace prior which is an exponential scale-mixture of normal. Not sufficient concentration near the value 0 for the entire posterior to concentrate near 0 whenever a coefficient is 0, only the posterior mode may be sparse.
- A more appropriate prior should have higher concentration near 0 while still maintaining a thick tail, by letting the scale parameter to have a more spiked density at 0.

- **Horseshoe prior:** Carvalho, Polson and Scott (2010, Biometrika) — a half-Cauchy scale mixture of normal

$$\theta|\lambda \sim \mathrm{N}(0, \lambda^2), \qquad \lambda \sim \mathrm{Cauchy}^{+}(0, \tau).$$

  The corresponding marginal density of θ has a pole at 0 and Cauchy-like tails.

- **Horseshoe+ prior:** Bhadra, Datta, Polson and Willard (2017, BA) — τ also half-Cauchy.

General features of a continuous shrinkage prior

- $\theta_i|\lambda_i, \tau \sim \mathrm{N}(0, \lambda_i^2\tau^2)$, $i = 1, \ldots, n$.
- λ_i is a local-shrinkage prior corresponding to the specific subject i, whereas τ is a global shrinkage parameter. Hence also called a global-local prior or a one-component prior.
- The local shrinkage parameter should have a heavy tail while the global shrinkage parameter should have a high concentration at 0, respectively controlling the tail and sparsity.
- Different priors on λ_i lead to a variety of interesting priors:
  - Normal-inverse-Gaussian: Caron and Doucet (2008, ICML)
  - Normal-gamma: Griffin and Brown (2010, BA)
  - Generalized double-Pareto: Armagan, Dunson and Lee (2013, BA)
  - Dirichlet-Laplace: Bhattacharya, Pati, Pillai and Dunson (2015, JASA) — $\theta_i|\phi, \tau \sim \mathrm{Lap}(\phi_i\tau)$, $\phi = (\phi_1, \ldots, \phi_p) \sim \mathrm{Dir}(a, \ldots, a)$, τ is given a gamma distribution, $a \in (0, 1)$ leads to a pole at 0.

Slowly growing dimension: Convergence without sparsity

- Consider the following common models with a large number of parameters:
  - **Linear regression:** $Y_i = \theta^{\mathrm{T}} X_i + \varepsilon_i$, independent errors (possibly normal) with variance σ^2, $i = 1, \ldots, n$, $\theta \in \mathbb{R}^p$.
  - **Generalized linear model:** $Y_i \overset{\text{ind}}{\sim} \mathrm{ExpFamily}(g(\theta^{\mathrm{T}} X_i))$, $i = 1, \ldots, n$, g some link function, $\theta \in \mathbb{R}^p$.
  - **Exponential family:** X_i i.i.d. from a p-dimensional standard exponential family $\exp[x^{\mathrm{T}}\theta - \psi(\theta)]$ for a p-dimensional parameter θ.
- Imagine that the dimension p of the parameter space increases with the sample size. At what rate the posterior will concentrate near the true value of the parameter?
- Does a kind of Bernstein-von Mises (BvM) theorem hold, that is, is the posterior of the parameter centered at an efficient estimator and scaled by the concentration rate approximated by a centered normal distribution with dispersion equal to the inverse of the Fisher information?

BvM in increasing dimension

- Intuitively, if p grows sufficiently slowly, classical approximations should be possible.
- Explicit restriction on the growth of p depends on several factors such as data matrix $\sum_{i=1}^{n} X_i X_i^{\mathrm{T}}$ (in regression models) or the Hessian of ψ (in exponential families).
- Commonly the condition $p = o(n^{1/4})$ ensures BvM theorem for regression models [G. (1997, MMS) for GLM, G. (1999, Ber)], and $p = o(n^{1/6})$ for multinomial exponential family [G. (2000, JMVA)].
- The most difficult part is estimating the tail.
- the central part is handled by a Taylor expansion.

Normal sequence model

- $Y_i = \theta_i + \varepsilon_i$, $i = 1, \ldots, n$, $\varepsilon_i \sim \mathrm{N}(0,1)$ i.i.d., $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$.
- True vector belongs to the *nearly-black* class $\ell_0[s] = \{\theta \in \mathbb{R}^n : \#\{i : \theta_i \neq 0\} \leq s\}$, $0 \leq s = s(n) \leq n$.
- Minimax risk for the squared error loss is $2s \log(n/s)$. The logarithmic factor is the penalty for not knowing the locations of the non-zero entries.
- Adaptation: The goal is to achieve the rate without knowing the correct value s_0 of s.
- Bayesian LASSO does not work — the posterior contracts at a suboptimal rate near the true vector.

Recovery using hard-spike-and-slab priors

- A remedy is to assign an additional point-mass at zero using a hard-spike-and-slab prior: for $w \in [0,1]$ and Γ a distribution on $\mathbb{R}$, $\Pi_w = \Pi_{w,\Gamma} = \bigotimes_{i=1}^n \{(1-w)\delta_0 + w\Gamma\}$.
- Choose the weight parameter $w = s/n$ if s is known for the optimal rate $s \log(n/s)$.
- Without knowing s and taking $w = c/n$ leads to a slightly suboptimal rate $s \log n$.
- Empirical Bayes using the marginal maximum likelihood empirical Bayes approach (MMLE) maximizing $\prod_{i=1}^n ((1-w)\phi(Y_i) + wg(Y_i))$, where $g = \gamma * \phi$ gives the optimal rate [Johnstone and Silverman (2004, AoS)],
- Full Bayes for $w \sim \text{Beta}(1, n+1)$, or a *subset-selection prior*

$$s \sim \pi_n, \qquad S|s \sim \text{Unif}(\mathcal{S}_s), \qquad \theta|S \sim \bigotimes_{i \in S} \Gamma \otimes \bigotimes_{i \notin S} \delta_0,$$

with the complexity prior $\pi_n(s) \propto \exp[-as \log(bn/s)]$ [Castillo and van der Vaart (2012, AoS)].

- **Dimension control:** Uniformly in $\ell_0[s]$, for some $M > 0$, $\Pi(|S| > Ms|Y) \to 0$ in probability. Proved using a combination of likelihood and prior tail control. Helps keep the complexity of the problem down.
- **Recovery:** Uniformly in $\ell_0[s]$, for some $M > 0$, $\Pi(\|\theta - \theta_0\| > Ms \log(n/s)|Y) \to 0$.
- **Preventing over-shrinkage:** Need tail at least tail as heavy as Laplace — normal overshrinks towards zero, leading to suboptimal convergence, and lack of uniformity over $\ell_0[s]$.
- **Computation of model posterior probabilities:** $\Pi(S) = Q_n^{-1} \pi_n(s) \binom{n}{s}^{-1} \prod_{i \in S} \psi(X_i) \prod_{i \in S^c} \phi(X_i)$, where $\psi = \phi * g$, g the slab density, and the normalizer Q_n is the **partition function** $Q_n = \sum_{p=0}^{n} \pi_n(p) \binom{n}{p}^{-1} \sum_{S:|S|=p} \prod_{i \in S} \psi(X_i) \prod_{i \in S^c} \phi(X_i)$. The most challenging part is the computation of the partition function, of which, computation of the inner sum. Naive computation has exponential complexity in n. It can be characterized as the coefficient of z^p in the polynomial $\prod_{i=1}^{n} (\phi(X_i) + z\psi(X_i))$, which can be computed in $O(n \log^2 n)$ time.

Oracle approach

- For a given sparsity structure S, the oracle estimator is $\hat{\theta}_S = (Y_S, 0_{S^c})$ with risk $\mathrm{E}\|\hat{\theta}_S - \theta\|^2$ and complexity $|S|$. Define the oracle risk $r(\theta) = \inf_S R(\theta, S)$, where $R(\theta, S) = \mathrm{E}\|\hat{\theta}_S - \theta\|^2 + |S| \log(n/|S|)$, and the intention is to get a procedure (estimator, or posterior distribution) with accuracy matching the oracle risk at all θ without knowing which structure applies — the logarithmic factor in the second term is needed, else the match is not possible.
- Oracle risk at a θ is better than the minimax risk over any class it belongs to, so matching the oracle risk ensures meeting the minimax risk over all possible structure classes.
- The concept is usable not just for sparsity classes, but in any problem with structure classes such as smoothness regimes.
- Developed by Belitser (2017, AoS) and later paper with co-authors.

Normal hard-spike-and-slab prior

- We can use the normal prior. To avoid overshrinkage towards 0, the mean is unspecified, estimated from the data.
- Model posterior probabilities are obtained by applying the Bayes theorem under conjugacy, and plugging-in the estimate.
- Dimension control follows more easily.
- Now the posterior will contract near the correct value at the optimal rate for all possible values of s.
- An advantage of conjugacy is that credible balls are also explicit, and their frequentist coverage can be studied. However, no method — Bayesian or frequentist — can adapt to the optimal size for all parameters maintaining coverage everywhere.
- Only parameter values where the bias arising from ignoring a component is relatively small compared with the estimation error —- called the excessive bias restriction (EBR) condition — can retain coverage with optimal radius credible balls.

Other shrinkage priors

- Sparsity can also be quantified in a weak sense such as using bound on ℓ_p-norm. To address optimal convergence in such classes, one may use a soft-spike-and-slab prior, to get essentially the same posterior contraction rate, provided that the spike distribution is sufficiently concentrated.
  - Spike-and-slab-LASSO: Spike distribution Lap(λ_0) for $\lambda_0 \to \infty$, slab distribution Lap(λ_1) for constant λ_1.
  - Horseshoe: Posterior mean has an explicit expressions in terms of degenerate hypergeometric functions. Posterior contraction at the rate $\sqrt{s \log n}$ [van der Pas et al. (2014 and 2017, EJS), Ghosh and Chakraborti (2017, BA)]; uncertainty quantification by van der Pas et al. (2017, BA).
  - Dirichlet-Laplace: Bhattacharya et al. (2015, JASA) under a growth condition on the norm of the true parameter.
  - Scale-mixture of normals continuous shrinkage: Under known sparsity, unified results are obtained by Van der Pas et al. (2016, EJS).
- Thresholding procedure to be used to identify zero entries.

Multiple testing

- Identifying a sparse structure is a multiple testing problem.
- False discovery rate (FDR) may be used to quantify the accuracy of Bayesian model selection.
- Castillo and Roquain (2020, AoS) derived a uniform FDR control over $\ell_0[s]$ for thresholding procedures when signals are all above the threshold $a\sqrt{2\log(n/s)}$, for some $a > 1$.
- Simultaneous control of FDR and FNR: Averages of ordered ℓ-values $\Pi(\theta_i = 0 | Y_1, \ldots, Y_n)$ to set the rejection threshold.
- Bayes risk for testing [Datta and Ghosh (2013, BA)]: With a hard-spike-and-(normal) slab prior with variance v_1 and known w, the oracle Bayes rule for rejecting $H_{0i} : \theta_i = 0$ is a thresholding procedure
$|Y_i|^2 > (1 + v_1^{-1})[\log(1 + v_1) + 2\log((1 - w)/w)]$.

Multiple change-point model

- $X_i \sim \mathrm{N}(\theta_i, \sigma^2)$, change-points $1 < i_1 < \cdots < i_{s-1} \leq n$, θ same value between change points.
- Alternative lower-dimensional structure.
- Belitser and G. (2022) pursue the oracle approach with independent normal prior for each common mean, given the configuration of the change-points, selecting prior means by block-average.
- Oracle rate for the Euclidean norm with the s change-points is $s \log(n/s)$.
- Posterior contraction achieves the oracle rate.
- Modal model's complexity bounded by a multiple of s in high probability.
- Inflated confidence ball centered at modal estimate and size of estimated oracle gives coverage at all parameters satisfying the EBR condition.
- Correct change-points are identified with high true probability if the gap between means in successive blocks are sufficiently

Linear regression

- $Y_i = \beta^{\mathrm{T}} X_i + \varepsilon_i$, $i = 1, \ldots, p$, $X_i \in \mathbb{R}^p$
- Let $\|X\| = \max \|X_{\cdot,j}\|$.
- Prior $\pi_p(s) \propto c^{-s} p^{-as}$, $S||S|$ is uniform, given S, g_S the product of $|S|$ Laplace densities $\beta \mapsto (\lambda/2)\exp(-\lambda|\beta|)$, $\lambda = \mu\|X\|$, for $p^{-1} \leq \mu \leq 2\sqrt{\log p}$ and the remaining entries of β are set to 0.

Compatibility condition

- For $p > n$, β cannot be uniquely recovered from $X\beta$.
- However, if β is assumed to be sparse (with only $s \ll n$ components non-zero) and the submatrix of X corresponding to the active predictors is full rank s, β can be recovered.
- Define the *compatibility number of model* $S \subset \{1, \ldots, p\}$ by $\overline{\phi}(S) := \inf \left\{ \|X\beta\| |S|^{1/2} / \|X\| \, \|\beta_S\|_1 : \|\beta_{S^c}\|_1 \leq 7\|\beta_S\|_1, \, \beta_S \neq 0 \right\}$, where $\beta_S = (\beta_i : i \in S)$, and the *$\ell_r$-compatibility number* in vectors of dimension s by $\phi_r(s) := \inf \left\{ \|X\beta\| |S_\beta|^{1-r/2} / \|X\| \, \|\beta\|_r : 0 \neq |S_\beta| \leq s \right\}$, $r = 1, 2$, where $S_\beta = \{j : \beta_j \neq 0\}$, the support of a sparse vector β.
- For β_0 the true vector of regression coefficients and $S_0 = S_{\beta_0}$ and $s_0 = |S_0|$, we assume that, with $r = 1$ or 2 depending on the context, $\min\left(\phi_r(S_0), \overline{\phi}(Cs_0)\right) \geq d > 0$, where C is a suitably large constant depending on S_0, a, μ.

Recovery

Findings from Castillo, Schmidt-Hieber and van der Vaart (2015, AoS)

- Dimension s at most a constant multiple of the true dimension s_0, with high posterior probability in true probability.
- When the ℓ_1-compatibility numbers are bounded away from 0, $\mathrm{E}_{\beta_0}\Pi(\|\beta - \beta_0\|_1 > Ms_0\sqrt{\log p}/\|X\| | Y_1, \ldots, Y_n) \to 0$.
- The corresponding rate in terms of the Euclidean distance is $\sqrt{s_0 \log p}/\|X\|$, assuming that the ℓ_2-compatibility numbers are bounded away from 0.
- The rate for prediction, i.e., bound for $\|X\beta - X\beta_0\|$, is of the order $\sqrt{s_0 \log p}$ under a slightly adapted compatibility condition.
- The convergence results are uniform over the parameter space under the boundedness conditions on the compatibility numbers, and match with those of celebrated estimators in the frequentist literature.

Variable selection

- To address variable selection, they developed a technique based on a distributional approximation under relatively low choices of the parameter λ, known as the small-λ regime.
- This is similar to the normal approximation to the posterior distribution in the Bernstein-von Mises theorem, but the difference is that in this context, the approximating distribution is a mixture of sparse normal distributions over different dimensions.
- Then the problem of variable selection can be transferred to that for the approximate posterior distribution, and it can be shown that no proper superset of the correct model can be selected with appreciable posterior probability.
- It is, however, possible to miss signals that are too small in magnitude. If all non-zero signals are assumed to be at least of the order $\sqrt{s_0 \log p}/\|X\|$, then none of these signals can be missed, because missing any of them introduces an error of the magnitude of the contraction rate. Thus, under this situation, distributional approximation reduces to a single normal component with sparsity exactly as in the true coefficient.

Group sparsity and correlated response

- Ning, Jeong and G. (2020, Bernoulli) considered response variable Y to have dimensional $d \to \infty$ slowly, and with completely unknown $d \times d$-covariance matrix Σ.
- The regression has group-sparsity — certain non-overlapping groups of variables are simultaneousy active or not active.
- They used a prior on Σ using a Cholesky decomposition and used a hard-spike-and-slab prior with multivariate Laplace slab on the group of regression coefficients selected together.
- General theory of posterior contraction is applied using exponentially consistent tests for separation based on the Réyni divergence.
- Squared posterior contraction rate $\epsilon_n^2 = \max\{(s_0 \log G)/n, (s_0 p_{\max} \log n)/n, (d^2 \log n)/n\}$, where G stands for the total number of groups of predictors, s_0 the number of active groups and $p_{\max}$ the maximum number of predictors in a group, provided that the regression coefficients and the covariance matrix are appropriately norm-bounded.
- Variable selection consistency using the BvM technique.

General class of linear regression models

- A general setup of sparse linear regression [Jeong and G., (2021, EJS)]: $Y_i = X_i\beta + \xi_{\eta,i} + \varepsilon_i$,
  - $\varepsilon_i \sim \mathrm{N}_{m_i}(0, \Delta_{\eta,i})$ independently with possibly varying dimension m_i and covariance matrices $\Delta_{\eta,i}$, depending on a nuisance parameter η.
  - The additional term $\xi_{\eta,i}$ incorporates various departure from a simple linear model
- Examples
  - Multiple response models with missing components
  - Multivariate measurement error models
  - Parametric correlation structure
  - Mixed effects models
  - Graphical structure with sparse precision matrices
  - Nonparametric heteroskedastic regression models
  - Partial linear models

- Euclidean distance on β and $n^{-1}\sum_{i=1}^{n}\{\|\xi_{\eta,i}-\xi_{\eta',i}\|^2+\|\Delta_{\eta,i}-\Delta_{\eta',i}\|_{\mathrm{F}}^2\}$ on η are used for contraction rates.
- Compatibility conditions are needed to recover β from $X\beta$ as in a simple linear regression model.
- Contraction rate derived by the general theory of posterior contraction in terms of the Réyni divergence with the required test obtained from local likelihood ratio tests.
- Extending the BvM technique, variable selection consistency can be established.

Oracle approach for linear regression

- Belitser and G. (2020, AoS) extend the oracle approach from normal mean to linear regression.
- Oracle risk at β is
  $R(\beta, I) = \mathrm{E}_\beta \|X\hat{\beta}_I - X\beta_I\| + \sigma^2 |I| \log(ep/|I|)$, where
  $I = \{i : \beta_i \neq 0\}$ is the model used.
- Hard-spike-and-slab prior with normal slab:
  $\beta|I \sim \mathrm{N}_{|I|}(\mu_I, \kappa\sigma^2(X_I^{\mathrm{T}} X_I)^{-1})$, $\beta_{I^c} = 0_{|I^c|}$,
  $\mu_I = (X_I^{\mathrm{T}} X_I)^{-1} X_I^{\mathrm{T}} Y$ (empirical Bayes); $I||I|$ is uniform, $|I|$ is given the complexity prior.
- Conjugacy given I allows explicit expressions.

- **Fundamental oracle inequality:** $\|X\beta - X\beta_0\| \leq Mr(\beta_0)$ with high posterior probability in P_{β_0}-probability.
- **Dimension control:** within a multiple of the oracle dimension (bounded by the true dimension) with high posterior probability in true probability.
- **Prediction ($X\beta$'s) accuracy**: At the oracle rate $\leq \sqrt{s \log(ep/s)}$ — no compatibility condition is needed.
- **Estimation rate**: $r(\beta_0)\sqrt{s}/\|X\|$ if the concerned compatibility numbers are bounded away from 0 — at least as good as using Laplace slab.
- **Uncertainty quantification:** Let $\hat{I}$ be the posterior modal model. Consider the posterior $(1-\alpha)$-credible ball centered at $\hat{\beta}_{\hat{I}}$. The uniformly for all parameter value satisfying an EBR condition, the credible ball radius inflated by some M has high frequentist coverage.
- **Model selection**: Modal model can be identified by a simulated annealing algorithm with moves adding a predictor or removing, and computing $X_I^{\mathrm{T}} X_I$ recursively from the previous one.

Other Bayesian methods for linear regression

- Narisetty and He (2014, AoS) modified the posterior measure by sparsifying the precision matrix appearing in the spike-and-slab posterior representation to accelerate SSVS. Called the *skinny Gibbs posterior*, the optimal rate of contraction is shown.
- Song and Liang (2017, Preprint) derived posterior contraction and variable selection properties for continuous shrinkage priors illustrating for the horseshoe, Dirichlet-Laplace, normal-gamma, and t-mixtures.
- Gao, Zhou and van der Vaart (2020, AoS) considered a structured linear model $Y = L_X\beta + \varepsilon$, where L_X is a linear operator depending on X and obtained minimax contraction rates using a block-Laplace prior. Result applies to linear regression, stochastic block models, biclustering, group sparsity, multi-task learning, dictionary learning.

Generalized linear model (GLM)

- Conjugate prior [Chen and Ibrahim (2003, Sinica)] used by Chen et al. (2008, BA) for variable selection.
- Posterior contraction for a hard-spike-and-slab prior in terms of Hellinger distance on the density is studied by Jiang (2007, AoS), under $\log p = o(n^{1-\xi})$ for some $\xi > 0$, using the general theory of posterior contraction rate.
- Atchadé (2017, AoS) studied contraction rates of pseudo-posterior distributions in a general setting assuming only a certain local expansion of the pseudo-likelihood ratio and derived posterior contraction rates for hard-spike-and-slab priors by constructing certain test functions.
- Jeong and G. (2021, Biometrika) obtained results on recovery rates for the regression coefficients for the usual posterior distribution using hard-spike-and-slab priors.

Logistic regression

- For logistic regression, Wei and G. (2017, JSPI) obtained posterior contraction rate and selection consistency were established for continuous shrinkage priors.
- For multidimensional logistic regression model, Jeong (2021, JSPI) obtained posterior contraction rate and selection consistency using hard spike-and-slab priors.

Additive nonparametric regression

Additive structure $f(x_1, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j)$.

- The minimax rate is $\max\{\sqrt{(s \log p)/n}, \sqrt{s} n^{-\alpha/(2\alpha+1)}\}$, where s is the number of active predictors and α is the smoothness of individual functions. Yang and Tokdar (2015, AoS) put Gaussian process priors for each selected component function aided by a hard spike-and-slab variable selection prior on the set of active predictors. Attained the minimax rate up to a logarithmic factor adaptively.
- Belitser and G. (2020, AoS) derived the same rate adaptively without the logarithmic factor by converting the model to a linear regression model using limited term orthogonal basis expansion and applying the oracle inequalities. Frequentist coverage for the posterior credible ball was obtained under an EBR condition for random covariates.
- Wei et al. (2018, Sinica) used a B-spline basis expansion prior with a multivariate version of the Dirichlet-Laplace continuous shrinkage prior on the coefficients, and obtained non-adaptive posterior contraction and variable selection consistency.

Single index regression

Roy et al. (2020, BA): Application in brain atrophy $f(t, X, Z) = a_0(X^{\mathrm{T}}\beta, Z^{\mathrm{T}}\eta) - F(t)a_0(X^{\mathrm{T}}\beta, Z^{\mathrm{T}}\eta)$, with X a high-dimensional predictor, Z a low-dimensional predictor and a_0, a_1, F smooth functions.

- For identifiability of the model, the coefficients β and η need to be unit vectors in respective dimensions, and β should be sparse as well in an appropriate sense.
- Sparse prior in the polar co-ordinate system using a soft spike-and-slab prior with a uniform slab and spike distributions with spikes at appropriate multiples of $\pi/2$.
- Posterior contraction rate in terms of the average squared distance on the functions using the general theory of posterior contraction as $\max\{n^{-\iota/(2\iota+2)}, n^{-\iota'/(2\iota'+1)}, \sqrt{(s \log p)/n}\}$, up to a logarithim factor, where ι is the smoothness of the functions (a_0, a_1), ι' is the smoothness of the function F, and s is the sparsity of the true β.

Density regression

- Response variable $Y \in [0,1]$, predictor $X \in [0,1]^p$, conditional density $f(y|x)$. Only s variables are active, so constant in the remaining arguments. Location of the these variables, s and the smoothness indexes of the function in the variables are unknown.
- A prior is put on conditional density using a basis expansion in tensor products of B-splines, and inserting a spike-and-slab mechanism while putting prior on the coefficients.
- Shen and G. (2016, Bernoulli) obtained the posterior contraction rate $(n/\log n)^{-\beta^*/(2\beta^*+s)}$, β^* (harmonic) average smoothness index, under $\log p = O(n^\alpha)$ for some $\alpha < 1$.

Oracle approach for general structure

Belitser and Nurushev (2019, arXiv)

- Any structured problem: For a structure I, $X|(\theta, I) \sim P_{\theta,I}$, satisfying some general conditions.
- Oracle rate: $r^2(\theta) = \inf_I \{\mathrm{E}\|\hat{\theta}_I - \theta\|^2 + \rho(I)\}$, where ρ the model complexity (at least as big as dimension of the model $d(I)$).
- Normal working model, conjugate normal prior on θ given I with unspecified mean estimated, $\pi(I) \propto e^{-\kappa\rho(I)}$.
- Posterior contracts at the rate $r(I)$ uniformly, giving minimax rate for all possible classes.
- Frequentist coverage of inflated $(1-\alpha)$ credible ball centered at the posterior mean of the posterior modal model $\hat{I}$, under the EBR condition.

Learning structural relationship

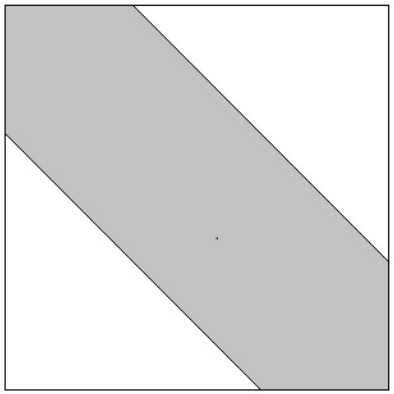
- A lower-dimensional structure is that most variables are pairwise independent, so most off-diagonal entries of the covariance matrix are (approximately) zero.
  - **Banding:** When the variables are arranged in some natural order, the pairwise covariances are zero (or decay quickly) if the lag between the corresponding two indexes is larger than some value. Holds exactly in a moving average (MA) process, while in an autoregressive (AR) or autoregressive moving average (ARMA) process, the covariances exponentially decay with the lag.
  - **No pattern:** In a $p \times p$ covariance matrix, only s of the $\binom{p}{2}$ off-diagonal entries are non-zero.
- Another important low-dimensional structure is sparse plus low-rank: $\Sigma = D + \Lambda\Lambda^{\mathrm{T}}$, where D is diagonal matrix and Λ is $p \times r$, $r \ll p$, with possibly sparse columns. Arises in structural linear models $X_i = \Lambda\eta_i + \varepsilon_i$, where Λ is a $p \times r$ sparse factor loading matrix and $\eta_i \sim \mathrm{N}_p(0, I)$ are independent latent factors.

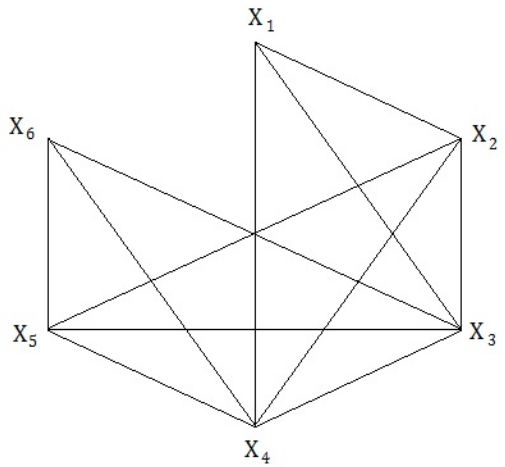
Graphical model

- An intrinsic relation among a set of variables is described by conditional dependence of a pair when the effects of the remaining variables are eliminated by conditioning on them.
- It is convenient to describe this structure using a graph, where each variable stands for a node and an edge connects a pair of nodes if and only if the two are conditionally dependent. Therefore such models are popularly called graphical models.
- If X_i and X_j are conditionally independent given $X_{-i,-j} := (X_k : k \neq i, j)$, then it follows that $\omega_{ij} = 0$, where $((\omega_{ij})) = \Sigma^{-1}$ is the precision matrix of X, to be denoted by Ω.
- In a Gaussian graphical model (GGM), i.e., when X is distributed as jointly normal, $X \sim \mathrm{N}_p(0, \Omega^{-1})$, then the converse also holds, namely $\omega_{ij} = 0$ implies that X_i and X_j are conditionally independent given $X_{-i,-j}$.

Sparse precision matrix

- In a Gaussian graphical model, learning the graph of dependence is learning the structure of Ω under sparsity, i.e., most off-diagonal entries of Ω are 0.
- Banding structure holds for autoregressive processes.





Graphical Wishart prior

- A conjugate prior for a precision matrix with a given graphical structure is given by a graphical Wishart distribution — certain block marginals are Wishart.
- Banerjee and G. (2014, EJS) obtained contraction rate under operator norm for approximately banded matrices: $\epsilon_n = k^{3/2} \max\{k\sqrt{(\log p)/n}, \gamma(k)\}$, where $\gamma(k)$ is the approximation rate using k-banded matrices.
- Explicit expression for Bayes estimates and model probabilities are available.

General sparse precision matrix

Only information: s off-diagonal entries are non-zero, s unknown, $s \ll \binom{p}{2}$, $p \ll n$.

- Graphical LASSO: Negative log-likelihood minimized with ℓ_1-penalty on Ω.
- Bayesian analogue is Bayesian graphical LASSO by Wang (2012, BA), converting the ℓ_1-penalty to Laplace prior on off-diagonals, and exponential prior on diagonals.
- Restriction to positive definiteness is non-trivial, but Wang developed a trick for fast posterior update called 'scaling-it-up'.
- Posterior concentration not possible without an edge selection mechanism in the prior.
- Thresholding posterior samples lets edge selection.

Recovery rate using edge selection

- Put an independent edge selection indicator in the prior through a hard spike.
- Under some assumptions on prior parameters and the true precision matrix, the posterior contracts at the rate $n^{-1/2}\sqrt{(p+s)\log n}$ with respect to the Frobenius norm [Banerjee and G. (2015, JMVA)], which is optimal.
- Computation is much harder, but a Laplace approximation method can compute the posterior modal model pretty fast.
- Continuous shrinkage prior like horseshoe instead of Laplace has been used [Li et al. (2019, JCGS)] with a data-augmentation Gibbs sampling for the half-Cauchy.
- This should have the same contraction rate under appropriate conditions, but a formal result is not available yet.

High dimensional discriminant analysis

- Classification based on a high-dimensional predictor $X = (X_1, \ldots, X_p) \sim \mathrm{N}_p(\mu, \Omega^{-1})$, where $(\mu, \Omega) = (\mu_1, \Omega_1)$ for the first group and $(\mu, \Omega) = (\mu_2, \Omega_2)$ for the second.
- A Bayesian procedure with priors on μ_1, μ_2 and Ω_1, Ω_2, the performance can nearly match the oracle if the posterior distributions of μ_1, μ_2 and Ω_1, Ω_2 contract near the true values sufficiently fast.
- Du and G. (2018, Sankhya) put a prior on Ω based on a sparse modified Cholesky decomposition $\Omega = LDL^{\mathrm{T}}$.
- Positive definiteness is automatic, but prior is dependent on ordering.
- If the probability of a non-zero entry is decreased as $i^{-1/2}$, then the probability of a non-zero at the ith and jth rows of Ω are roughly equal for large $i \asymp j$.
- Misclassification rate converges to that of the oracle Bayes classifier for a general class of shrinkage priors when $p^2(\log p)/n \to 0$, if the number of off-diagonal entries in the true Ω is $O(p)$.

Ising model

- Bernoulli random variables defined over the nodes of the graph $G = (V, E)$
  $p(X; \theta) = \exp\{\sum_{r \in V} \theta_r X_r + \sum_{(r,t) \in E} \theta_{rt} X_r X_t - A(\theta)\}$.
- Evaluation of the log-partition function $A(\theta)$ is challenging — variational methods have been developed.
- The conditional distribution of the nodes gives a logistic regression model.
- This gives a pseudo-likelihood function, by multiplying all one-dimensional conditionals.
- Atchade's rate result applies in this case.

Nonparanormal model

- The distribution of $X = (X_1, \ldots, X_p) \in [0,1]^p$ reduces to a multivariate normal vector through p monotone increasing transformations: for some monotone functions $f_1, \ldots, f_p$, $f(X) := (f_1(X_1), \ldots, f_p(X_p)) \sim \mathrm{N}_p(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^p$ and positive definite matrix Σ.
- The model is not identifiable and needs to fix the location and scale of the functions or the distribution.
- Mulgrave and G. (2020, BA): A finite random series based on a B-spline basis expansion is used to construct a prior on the transformation.
- The B-spline basis maintains monotonicity if the coefficients only needs to be made increasing.
- Usual priors on μ and Σ.

- A multivariate normal prior truncated to the cone of ordered values can be conveniently used.
- Constraints $f(0) = 1/2$ and $f(3/4) - f(1/4) = 1$ are imposed, which translate to linear constraints, and hence the prior remained multivariate normal before imposing the order restriction.
- Samples from the posterior distribution of the ordered multivariate normal coefficients can be efficiently obtained using the exact Hamiltonian MCMC (Packman and Paninski, 2014, JCGS).
- Learning of f and (μ, Σ) are (separately) consistent.
- Other approaches are regression-based (Mulgrave and G., 2022 SADM) and rank-likelihood based (Mulgrave and G., 2022 JSPI).

Graphical model with measurement error

- $(X_1, \ldots, X_p)$ follows a Gaussian graphical model with sparse precision matrix Ω, but X_i are observed only with a measurement error, that is, we observe $Y_i = X_i + Z_i$, $i = 1, \ldots, p$, where each Z_{ij} is independent $\mathrm{N}(0, \nu)$, ν known.
- The precision matrix for an observation is $(\Omega^{-1} + \nu I)^{-1}$, which need not be sparse, but has a lower-dimensional structure.
- Shi, G. and Martin (2021, EJS) showed that the contraction rate for the no-measurement error case under the Frobenius norm is preserved, by extracting Ω from $(\Omega^{-1} + \nu I)^{-1}$.
- Interestingly, ν need not be small.
- If ν is unknown, a replication $Y_{ij} = X_i + Z_{ij}$ does the job.
- Non-Gaussian measurement errors can also be handled, but the results are not so neat.

Estimating a long vector smoothly varying over a graph

- $X_i = \theta_i + \varepsilon_i$, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ independently, $i \in V = \{1, \ldots, n\}$.
- There is a natural notion of neighbors for elements of V, giving an undirected graph.
- $f = (\theta_1, \ldots, \theta_n)$ is assumed to "smoothly vary" over V.
- Mathematically quantified through the graph Laplacian $L = D - A$, where D is the diagonal matrix of node degrees, and A stands for the adjacency matrix.
- f is said to belong to a Hölder ball of β-smoothness of radius Q if $\langle f, (I + (n^{2/r} L)^{\beta}) f \rangle \leq Q^2$, r graph's dimension.
- The minimax rate of recovery is $n^{-\beta/(2\beta + r)}$ [Kirichenko and van Zanten (2018, EJS)].
- For prior $f \sim \mathrm{N}_n(0, (n/c)^{(2\alpha + r)/r} (L + n^{-2} I)^{(2\alpha + r)/r})$, c exponential, the posterior contracts nearly at the optimal rate if $\beta \leq \alpha + r/2$ [Kirichenko and van Zanten (2017, EJS)].

Functional observations over a graph

- The last result can be extended to observations taking values in a Hilbert space as well, so that functional observations can be treated [Roy and G., (2021, JMVA)].
- They introduced a notion of Sobolev smoothness classes indexed by two smoothness indexes, β for the graphical smoothness and γ for the functional smoothness.
- They designed Bayes procedures which achieve the minimax rate of recovery is $n^{-\beta\gamma/(2\beta\gamma+\beta+r\gamma)}$, where r is the dimension of the graph.
- If the smoothness index γ is unknown, the rate can be achieved by a Bayes procedure adaptively within a logarithmic factor.
- In a non-adaptive setting, they also showed that a slightly inflated Bayesian credible ball of optimal size has adequate frequentist coverage.

Matrix models

A number of high-dimensional problems involve unknown matrices instead of vectors. Some examples include multiple linear regression with group sparsity, multi-task learning, matrix completion, stochastic block model and biclustering.

Generic results in structured linear models

- Gao et al. (2015, AoS) obtained posterior contraction rates for many models simultaneously in a general 'signal plus noise' model when the signal is a vector or a matrix having some specific structure.
- One example is multiple linear regression with group sparsity, where one observes $Y = XB + W$, with X being an $n \times p$ matrix, B a $p \times m$ matrix, and the columns of B are assumed to share common support of size s.
- They obtained the minimax rate $s(m + \log(ep/s))$ for the prediction problem of estimating XB in Frobenius norm.

- This model is a special case of so-called multi-task learning problems, where the columns of B share some specific structure.
- For instance, instead of assuming a joint sparsity pattern among columns, one may assume that columns of B can only be chosen from a given list of $k \ll m$ possible columns. The corresponding posterior contraction rate is $pk + m \log k$.
- Dictionary learning assumes that the signal matrix of size $n \times d$ is $\theta = QZ$, for Q an $n \times p$ dictionary matrix and Z a discrete matrix of size $p \times d$ with sparse columns. The corresponding posterior contraction rate is $np + ds \log(ep/s)$.

Stochastic block model

- k groups, $Y = \theta + W$ with $\theta_{ij} = Q_{z(i)z(j)} \in [0,1]^{n \times n}$ for some matrix Q of size $k \times k$ of edge probabilities, a labeling map $z \in \{1, \ldots, k\}^n$ and W a centered Bernoulli noise.
- A Bayes procedure contracts adaptively at the minimax rate $k^2 + n \log k$ with unknown k [Belitser and Nurushev (2019, arXiv), Pati and Bhattacharya (2015, arXiv)]
- **Biclustering model:** An asymmetric extension of the SBM, where θ is an $n \times m$ rectangular matrix and rows and columns of Q have their own labeling, with k and l groups respectively. Adaptive posterior contraction rate is $kl + n \log k + m \log l$ [Gao et al. (2020, AoS), Belitser and Nurushev (2019, arXiv)].

Community detection

- Recovery of labels $z_.$.
- Van der Pas and van der Vaart (2018, BA) showed that the posterior mode corresponding to a beta prior on edge probabilities and Dirichlet prior probabilities for the label proportions asymptotically recovers the labels when the number of groups k is known, provided that the mean degree of the graph is at least of order $\log^2 n$.
- Their approach relies on studying the *Bayesian modularity*, that is, the marginal likelihood of the class labels given the data, when the edge probabilities are integrated out.
- Kleijn and van Waaij (2018, Preprint) and van Waaij and Kleijn (2020, Preprint) derived rates for Bayesian estimation and uncertainty quantification in the special case of the planted multi-section model, where the matrix Q has only two different values, one for diagonal elements and another for off-diagonal ones.

Matrix completion

- We observed n randomly selected noisy entries of an unknown $m \times p$ matrix M typically assumed to be of low-rank r (or well-approximated by a low-rank matrix).
- The nearly optimal recovery rate $(m+p)r \log \max(m,p)$was obtained by Mai and Alquier (2015, EJS) using a PAC-Bayesian procedure with a prior sitting close to small-rank matrices.
- Suzuki (2015, Proc of MLR) derived similar results for posterior distributions, and generalized to tensors.

High-dimensional Variational Bayes

- Variational Bayes is an optimization-based fast posterior computing tool.
- However, it does not actually compute the posterior distribution, not even approximately — just obtains a random measure that tracks the actual posterior within a predetermined family in the best possible way.
- From a frequentist-Bayes point of view, we should not mind if the resulting random measure has (nearly) the same contraction rate under essentially the same conditions.
- For high-dimensional regression, spike-and-slab prior based mean-field variational posterior computation techniques have been developed [Carbonetto and Stephens (2012, BA), Ormerod et al. (2017, EJS)].
- Ray and Szabó (2020, JASA) obtained contraction rate using the class $\otimes_{i=1}^{p}\{\gamma_i \mathrm{N}(\mu_i, \sigma_i^2) + (1-\gamma_i)\delta_0\}$.

Thank you