Bayesian Aggregation for Extraordinarily Large Data

Guang Cheng[1]

Department of Statistics
Purdue University

Department Seminar
Statistics @ UW – Madison
October, 2015

[1] Acknowledge NSF, ONR and Simons Foundation. A Joint Work with Zuofeng Shang.

The Era of Big Data

At the 2010 Google Atmosphere Convention, Google's CEO Eric Schmidt pointed out that,

"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."

No wonder that the era of Big Data has arrived...

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Bayesian Aggregation

This generic aggregation procedure applies to both finite dimensional parameter and infinite dimensional parameter.

$$
\begin{array}{ccccc}
 & & \text{Subset } 1\ (n_1) & \xrightarrow{\text{Machine } 1} & R_1(\alpha) \\
\text{Big Data } (N) & \xrightarrow{\text{Divide}} & \text{Subset } 2\ (n_2) & \xrightarrow{\text{Machine } 2} & R_2(\alpha) \\
 & & \cdots & \cdots & \cdots \\
 & & \text{Subset } s\ (n_s) & \xrightarrow{\text{Machine } s} & R_s(\alpha) \\
\text{Super}\Big\Downarrow\text{machine} & & & & \text{Aggre}\Big\Downarrow\text{gate} \\
R_{\text{oracle}}(\alpha) & & & & R(\alpha)
\end{array}
$$

$R_{\text{oracle}}(\alpha)$: $(1-\alpha)$ oracle credible region constructed from the entire data (computationally prohibitive in practice, though);
$R_j(\alpha)$: $(1-\alpha)$ credible region constructed from the j-th subset.

A Series of Theoretical Questions...

- How to define an aggregation rule s.t. $R(\alpha)$ covers $(1-\alpha)$ posterior mass, with the same radius as $R_{\text{oracle}}(\alpha)$?
- How to construct a prior s.t. $R(\alpha)$ covers the true parameter (generating the data) with probability $(1-\alpha)$?
- How fast can we allow s to diverge ("splitotics theory")?
- The above tasks are particularly challenging when the parameter in consideration is infinite dimensional, which is the focus of our talk today.

A Series of Theoretical Questions...

- How to define an aggregation rule s.t. $R(\alpha)$ covers $(1-\alpha)$ posterior mass, with the same radius as $R_{\text{oracle}}(\alpha)$?
- How to construct a prior s.t. $R(\alpha)$ covers the true parameter (generating the data) with probability $(1-\alpha)$?
- How fast can we allow s to diverge ("splitotics theory")?
- The above tasks are particularly challenging when the parameter in consideration is infinite dimensional, which is the focus of our talk today.

A Series of Theoretical Questions...

- How to define an aggregation rule s.t. $R(\alpha)$ covers $(1-\alpha)$ posterior mass, with the same radius as $R_{\text{oracle}}(\alpha)$?
- How to construct a prior s.t. $R(\alpha)$ covers the true parameter (generating the data) with probability $(1-\alpha)$?
- How fast can we allow s to diverge ("splitotics theory")?
- The above tasks are particularly challenging when the parameter in consideration is infinite dimensional, which is the focus of our talk today.

A Series of Theoretical Questions...

- How to define an aggregation rule s.t. $R(\alpha)$ covers $(1-\alpha)$ posterior mass, with the same radius as $R_{\text{oracle}}(\alpha)$?
- How to construct a prior s.t. $R(\alpha)$ covers the true parameter (generating the data) with probability $(1-\alpha)$?
- How fast can we allow s to diverge ("splitotics theory")?
- The above tasks are particularly challenging when the parameter in consideration is infinite dimensional, which is the focus of our talk today.

Literature Review

- In the Bayesian community, the existing statistical studies mostly focus on computational or methodological aspects of MCMC-based distributed methods;
- Nonetheless, not much effort has been devoted to *theoretically* understanding scalable Bayesian procedures especially in a general nonparametric context;
- One particular reason is the failure of Bernstein-von Mises theorem in the nonparametric setting found by Cox (1993) and Freedman (1999).

Literature Review

- In the Bayesian community, the existing statistical studies mostly focus on computational or methodological aspects of MCMC-based distributed methods;
- Nonetheless, not much effort has been devoted to *theoretically* understanding scalable Bayesian procedures especially in a general nonparametric context;
- One particular reason is the failure of Bernstein-von Mises theorem in the nonparametric setting found by Cox (1993) and Freedman (1999).

Literature Review

- In the Bayesian community, the existing statistical studies mostly focus on computational or methodological aspects of MCMC-based distributed methods;
- Nonetheless, not much effort has been devoted to *theoretically* understanding scalable Bayesian procedures especially in a general nonparametric context;
- One particular reason is the failure of Bernstein-von Mises theorem in the nonparametric setting found by Cox (1993) and Freedman (1999).

Outline

What is Bernstein-von Mises (BvM) Theorem?

- BvM theorem[2] characterizes *asymptotic shape* of posterior distribution

$$d(\Pi(\cdot|\mathbf{D}_n), P_0(\cdot)) \longrightarrow 0 \text{ as } n \to \infty,$$

where $\Pi(\cdot|\mathbf{D}_n)$ represents a posterior measure based on sample $\mathbf{D}_n$ with size n, $P_0(\cdot)$ is a limiting probability measure, and d denotes a distance measure;

- For example, in parametric models BvM Theorem says

$$\sup_{B\in\mathcal{B}} |\Pi(B|\mathbf{D}_n) - \mathcal{N}(\widehat{\theta}_n, (nI_{\theta_0})^{-1})(B)| = o_{P^n_{\theta_0}}(1),$$

where $\mathcal{B}$ is the Borel algebra on $\mathbb{R}^d$.

[2]Named after two mathematicians: S. Bernstein and R. von Mises.

What is Bernstein-von Mises (BvM) Theorem?

- BvM theorem[2] characterizes *asymptotic shape* of posterior distribution

$$d(\Pi(\cdot|\mathbf{D}_n), P_0(\cdot)) \longrightarrow 0 \ \text{ as } \ n \to \infty,$$

where $\Pi(\cdot|\mathbf{D}_n)$ represents a posterior measure based on sample $\mathbf{D}_n$ with size n, $P_0(\cdot)$ is a limiting probability measure, and d denotes a distance measure;

- For example, in parametric models BvM Theorem says

$$\sup_{B\in\mathcal{B}} |\Pi(B|\mathbf{D}_n) - \mathcal{N}(\widehat{\theta}_n, (nI_{\theta_0})^{-1})(B)| = o_{P^n_{\theta_0}}(1),$$

where $\mathcal{B}$ is the Borel algebra on $\mathbb{R}^d$.

[2] Named after two mathematicians: S. Bernstein and R. von Mises.

A Graphical Illustration

More importantly, BvM theorem implies the frequentist validity of Bayesian credible sets, called as *BvM phenomenon*, as

$$P_{\theta_0}^n(\theta_0 \in (1-\alpha)\text{-th credible set}) \to 1-\alpha.$$

$$\begin{tikzcd}
\Pi(\cdot|\mathbf{D}_n) \arrow[d, "\text{MCMC}"] & P_0(\cdot) \arrow[l, dashed, "\text{BvM Theorem}"'] \arrow[dl, dashed, "\text{BvM Phenomenon}"] \\
(1-\alpha)\text{-th credible set} &
\end{tikzcd}$$

Nonparametric BvM: a negative example

- Consider Gaussian sequence models:

$$Y_i = \theta_{0i} + \frac{1}{\sqrt{n}}\epsilon_i, \quad i = 1, 2, \dots,$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$. The "true" mean sequence $\{\theta_{0i}\}_{i=1}^{\infty}$ is square-summable, i.e., $\sum_{i=1}^{\infty} \theta_{0i}^2 < \infty$;

- Assign a (very innocent) Gaussian Prior:

$$\mathbf{P0}: \quad \theta_i \sim N(0, i^{-2p}) \ \text{ for some } p > 1/2.$$

- Freedman (1999) demonstrated the failure of BvM:

$$P_{\theta_0}^n(\theta_0 \in (1-\alpha) \text{ credible set}) \to 0.$$

The credible set is based on ℓ^2-norm.

Nonparametric BvM: a negative example

- Consider Gaussian sequence models:

$$Y_i = \theta_{0i} + \frac{1}{\sqrt{n}}\epsilon_i, \quad i = 1, 2, \dots,$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$. The "true" mean sequence $\{\theta_{0i}\}_{i=1}^{\infty}$ is square-summable, i.e., $\sum_{i=1}^{\infty} \theta_{0i}^2 < \infty$;

- Assign a (very innocent) Gaussian Prior:

$$\mathbf{P0}: \quad \theta_i \sim N(0, i^{-2p}) \text{ for some } p > 1/2.$$

- Freedman (1999) demonstrated the failure of BvM:

$$P_{\theta_0}^n(\theta_0 \in (1-\alpha) \text{ credible set}) \to 0.$$

The credible set is based on ℓ^2-norm.

Nonparametric BvM: a negative example

- Consider Gaussian sequence models:

$$Y_i = \theta_{0i} + \frac{1}{\sqrt{n}}\epsilon_i, \quad i = 1, 2, \dots,$$

where $\epsilon_i \stackrel{iid}{\sim} N(0,1)$. The "true" mean sequence $\{\theta_{0i}\}_{i=1}^{\infty}$ is square-summable, i.e., $\sum_{i=1}^{\infty} \theta_{0i}^2 < \infty$;

- Assign a (very innocent) Gaussian Prior:

$$\mathbf{P0}: \quad \theta_i \sim N(0, i^{-2p}) \;\text{ for some } p > 1/2.$$

- Freedman (1999) demonstrated the failure of BvM:

$$P_{\theta_0}^n(\theta_0 \in (1-\alpha) \text{ credible set}) \to 0.$$

The credible set is based on ℓ^2-norm.

A Solution: Tuning Prior

- The power of smoothing spline (Wahba, 1990)!
- We will show that nonparametric BvM theorem can be rescured under a new class of Gaussian process (GP) priors motivated by smoothing spline, named as *"tuning prior"*;
- Take Gaussian regression models as an example[3]:

$$Y_i = f_0(X_i) + \epsilon_i,\ i = 1, 2, \dots, n,$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$ and $f \in H^m(0,1)$, a m-th order Sobolev space. Denote its log-likelihood function as

$$\ell_n(f) = -\sum_{i=1}^{n}(Y_i - f(X_i))^2/2.$$

[3] Our nonparametric BvM results hold in a general exponential family that covers both Gaussian and non-Gaussian data, e.g., Poisson data.

A Solution: Tuning Prior

- The power of smoothing spline (Wahba, 1990)!
- We will show that nonparametric BvM theorem can be rescured under a new class of Gaussian process (GP) priors motivated by smoothing spline, named as *"tuning prior"*;
- Take Gaussian regression models as an example[3]:

$$Y_i = f_0(X_i) + \epsilon_i, \ i = 1, 2, \ldots, n,$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$ and $f \in H^m(0,1)$, a m-th order Sobolev space. Denote its log-likelihood function as

$$\ell_n(f) = -\sum_{i=1}^{n} (Y_i - f(X_i))^2/2.$$

[3] Our nonparametric BvM results hold in a general exponential family that covers both Gaussian and non-Gaussian data, e.g., Poisson data.

A Solution: Tuning Prior

- The power of smoothing spline (Wahba, 1990)!
- We will show that nonparametric BvM theorem can be rescured under a new class of Gaussian process (GP) priors motivated by smoothing spline, named as *"tuning prior"*;
- Take Gaussian regression models as an example[3]:

$$Y_i = f_0(X_i) + \epsilon_i, \ i = 1, 2, \ldots, n,$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$ and $f \in H^m(0,1)$, a m-th order Sobolev space. Denote its log-likelihood function as

$$\ell_n(f) = -\sum_{i=1}^{n} (Y_i - f(X_i))^2/2.$$

[3] Our nonparametric BvM results hold in a general exponential family that covers both Gaussian and non-Gaussian data, e.g., Poisson data.

A Solution: Tuning Prior

- The power of smoothing spline (Wahba, 1990)!
- We will show that nonparametric BvM theorem can be rescured under a new class of Gaussian process (GP) priors motivated by smoothing spline, named as *"tuning prior"*;
- Take Gaussian regression models as an example[3]:

$$Y_i = f_0(X_i) + \epsilon_i, \; i = 1, 2, \ldots, n,$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$ and $f \in H^m(0,1)$, a m-th order Sobolev space. Denote its log-likelihood function as

$$\ell_n(f) = -\sum_{i=1}^{n} (Y_i - f(X_i))^2/2.$$

[3] Our nonparametric BvM results hold in a general exponential family that covers both Gaussian and non-Gaussian data, e.g., Poisson data.

Tuning Prior: A General Framework

- Assume that f follows a probability measure Π_λ;
- Specify Π_λ through its Radon-Nikodym derivative w.r.t. a base measure Π (also on $H^m(0,1)$) as follows:

$$\frac{d\Pi_\lambda}{d\Pi}(f) \propto \exp\left(-\frac{n\lambda}{2}J(f)\right), \tag{1.1}$$

where $J(f)$ is a type of roughness penalty used in smoothing spline literature.

Tuning Prior: A General Framework

- Assume that f follows a probability measure Π_λ;
- Specify Π_λ through its Radon-Nikodym derivative w.r.t. a base measure Π (also on $H^m(0,1)$) as follows:

$$\frac{d\Pi_\lambda}{d\Pi}(f) \propto \exp\left(-\frac{n\lambda}{2}J(f)\right), \tag{1.1}$$

where $J(f)$ is a type of roughness penalty used in smoothing spline literature.

Tuning Prior: Duality

- Based on (1.1), we have the posterior as

$$\begin{aligned} P(f|\mathbf{D}_n) &:= \frac{\exp(\ell_n(f))d\Pi_{\lambda}(f)}{\int_{H^m(0,1)} \exp(\ell_n(f))d\Pi_{\lambda}(f)} \\ &= \frac{\exp(\ell_{n,\lambda}(f))d\Pi(f)}{\int_{H^m(0,1)} \exp(\ell_{n,\lambda}(f))d\Pi(f)}, \end{aligned}$$

where $\ell_{n,\lambda}(f) = \ell_n(f) - n\lambda J(f)$. Smoothing spline estimate

$$\widehat{f}_{n,\lambda} := \arg \max_{f \in H^m(0,1)} \ell_{n,\lambda}(f);$$

- The name "tuning prior" now makes sense. So, we can employ GCV to select a proper tuning prior (and we did!);
- More importantly, we are able to borrow the recent advances in smoothing spline inference theory (Shang and C., 2013, *AoS*) to build a foundation of nonpara. BvM.

Tuning Prior: Duality

- Based on (1.1), we have the posterior as

$$P(f|\mathbf{D}_n) := \frac{\exp(\ell_n(f))d\Pi_{\lambda}(f)}{\int_{H^m(0,1)} \exp(\ell_n(f))d\Pi_{\lambda}(f)} = \frac{\exp(\ell_{n,\lambda}(f))d\Pi(f)}{\int_{H^m(0,1)} \exp(\ell_{n,\lambda}(f))d\Pi(f)},$$

where $\ell_{n,\lambda}(f) = \ell_n(f) - n\lambda J(f)$. Smoothing spline estimate

$$\widehat{f}_{n,\lambda} := \arg \max_{f \in H^m(0,1)} \ell_{n,\lambda}(f);$$

- The name "tuning prior" now makes sense. So, we can employ GCV to select a proper tuning prior (and we did!);
- More importantly, we are able to borrow the recent advances in smoothing spline inference theory (Shang and C., 2013, *AoS*) to build a foundation of nonpara. BvM.

Tuning Prior: Duality

- Based on (1.1), we have the posterior as

$$P(f|\mathbf{D}_n) := \frac{\exp(\ell_n(f))d\Pi_{\lambda}(f)}{\int_{H^m(0,1)} \exp(\ell_n(f))d\Pi_{\lambda}(f)} = \frac{\exp(\ell_{n,\lambda}(f))d\Pi(f)}{\int_{H^m(0,1)} \exp(\ell_{n,\lambda}(f))d\Pi(f)},$$

where $\ell_{n,\lambda}(f) = \ell_n(f) - n\lambda J(f)$. Smoothing spline estimate

$$\widehat{f}_{n,\lambda} := \arg \max_{f \in H^m(0,1)} \ell_{n,\lambda}(f);$$

- The name "tuning prior" now makes sense. So, we can employ GCV to select a proper tuning prior (and we did!);
- More importantly, we are able to borrow the recent advances in smoothing spline inference theory (Shang and C., 2013, *AoS*) to build a foundation of nonpara. BvM.

Tuning Prior: Gaussian Process Construction

- To satisfy (1.1), we choose Π_λ and Π as two Gaussian measures induced by GP priors as specified below (this can be verified by applying Hájek's Lemma);
- Assign a GP prior on f, i.e., Π_λ, as follows:

$$f \sim G_\lambda(\cdot) = \sum_{\nu=1}^{\infty} w_\nu \varphi_\nu(\cdot),$$

where (recall that m is the smoothness of f_0)

$$w_\nu \sim \begin{cases} N(0,1), & \nu = 1, \ldots, m \\ N\left(0, (\rho_\nu^{1+\beta/2m} + n\lambda\rho_\nu)^{-1}\right), & \nu > m, \end{cases}$$

for a sequence $\rho_\nu \asymp \nu^{2m}$;
- Π is induced by a similar GP (by setting $\lambda = 0$).

Tuning Prior: Gaussian Process Construction

- To satisfy (1.1), we choose Π_λ and Π as two Gaussian measures induced by GP priors as specified below (this can be verified by applying Hájek's Lemma);
- Assign a GP prior on f, i.e., Π_λ, as follows:

$$f \sim G_\lambda(\cdot) = \sum_{\nu=1}^{\infty} w_\nu \varphi_\nu(\cdot),$$

where (recall that m is the smoothness of f_0)

$$w_\nu \sim \begin{cases} N(0,1), & \nu = 1, \ldots, m \\ N\left(0, (\rho_\nu^{1+\beta/2m} + n\lambda\rho_\nu)^{-1}\right), & \nu > m, \end{cases}$$

for a sequence $\rho_\nu \asymp \nu^{2m}$;
- Π is induced by a similar GP (by setting $\lambda = 0$).

Tuning Prior: Gaussian Process Construction

- To satisfy (1.1), we choose Π_λ and Π as two Gaussian measures induced by GP priors as specified below (this can be verified by applying Hájek's Lemma);
- Assign a GP prior on f, i.e., Π_λ, as follows:

$$f \sim G_\lambda(\cdot) = \sum_{\nu=1}^{\infty} w_\nu \varphi_\nu(\cdot),$$

where (recall that m is the smoothness of f_0)

$$w_\nu \sim \begin{cases} N(0,1), & \nu = 1, \ldots, m \\ N\left(0, (\rho_\nu^{1+\beta/2m} + n\lambda\rho_\nu)^{-1}\right), & \nu > m, \end{cases}$$

for a sequence $\rho_\nu \asymp \nu^{2m}$;
- Π is induced by a similar GP (by setting $\lambda = 0$).

- Our construction of GP prior is motivated from Wahba' Bayesian view on smoothing spline (Wahba, 1990);
- The RKHS induced by G_λ is essentially $H^{m+\beta/2}(0,1)$, where β adjusts the prior support;
- In addition, we need to assume $\beta \in (1, 2m+1)$ to guarantee $E\{J(G_\lambda, G_\lambda)\} < \infty$ such that the sample path of G_λ belongs to $H^m(0,1)$ a.s..

- Our construction of GP prior is motivated from Wahba' Bayesian view on smoothing spline (Wahba, 1990);
- The RKHS induced by G_λ is essentially $H^{m+\beta/2}(0,1)$, where β adjusts the prior support;
- In addition, we need to assume $\beta \in (1, 2m+1)$ to guarantee $E\{J(G_\lambda, G_\lambda)\} < \infty$ such that the sample path of G_λ belongs to $H^m(0,1)$ a.s..

- Our construction of GP prior is motivated from Wahba' Bayesian view on smoothing spline (Wahba, 1990);
- The RKHS induced by G_λ is essentially $H^{m+\beta/2}(0,1)$, where β adjusts the prior support;
- In addition, we need to assume $\beta \in (1, 2m+1)$ to guarantee $E\{J(G_\lambda, G_\lambda)\} < \infty$ such that the sample path of G_λ belongs to $H^m(0,1)$ a.s..

Underlying Eigensystem $(\varphi_\nu(\cdot), \rho_\nu)$

- Under mild conditions, f admits a Fourier expansion:

$$f(\cdot) = \sum_{\nu=1}^{\infty} f_\nu \varphi_\nu(\cdot),$$

  where $\varphi_\nu(\cdot)$'s are basis functions in $H^m(0,1)$.

- An example for (φ_ν, ρ_ν) is the following ODE solution:

$$\varphi_\nu^{(2m)}(\cdot) = \rho_\nu \varphi_\nu(\cdot), \ \varphi_\nu^{(j)}(0) = \varphi_\nu^{(j)}(1) = 0, \ j = 2, \ldots, 2m-1,$$

  where φ_ν's have closed forms. This is also called as "uniform free beam problem" in physics.

Underlying Eigensystem $(\varphi_\nu(\cdot), \rho_\nu)$

- Under mild conditions, f admits a Fourier expansion:

$$f(\cdot) = \sum_{\nu=1}^{\infty} f_\nu \varphi_\nu(\cdot),$$

  where $\varphi_\nu(\cdot)$'s are basis functions in $H^m(0,1)$.
- An example for (φ_ν, ρ_ν) is the following ODE solution:

$$\varphi_\nu^{(2m)}(\cdot) = \rho_\nu \varphi_\nu(\cdot), \;\; \varphi_\nu^{(j)}(0) = \varphi_\nu^{(j)}(1) = 0, \;\; j = 2, \ldots, 2m-1,$$

  where φ_ν's have closed forms. This is also called as "uniform free beam problem" in physics.

Nonparametric BvM theorem

Theorem 1

Given that $\lambda \asymp n^{-2m/(2m+\beta)}$, *we have*

$$\sup_{S \subset H^m(0,1)} |P(S|D_n) - \Pi_W(S)| = o_{P_{f_0}^n}(1),$$

where $\Pi_W(\cdot)$ *is the probability measure induced by a GP* W.

Specifications of the Limiting GP W

- Suppose that $\widehat{f}_{n,\lambda}(\cdot) = \sum_{\nu=0}^{\infty} \widehat{f}_{n,\nu}\varphi_\nu(\cdot)$;
- The mean function of W (also the approximate posterior mode of $P(\cdot|D_n)$) is

$$\widetilde{f}_{n,\lambda} := \sum_{\nu=0}^{\infty} a_{n,\nu}\widehat{f}_{n,\nu}\varphi_\nu(\cdot).$$

  Hence, $\widetilde{f}_{n,\lambda} \neq \widehat{f}_{n,\lambda}$;
- The mean-zero GP $W_n := W - \widetilde{f}_{n,\lambda}$ is expressed as

$$W_n(\cdot) = \sum_{\nu=0}^{\infty} b_{n,\nu} z_\nu \varphi_\nu(\cdot) \ \text{ and } \ z_\nu \overset{iid}{\sim} N(0,1);$$

- Here, $a_{n,\nu}$ and $b_{n,\nu}$ are both non-random sequences.

Specifications of the Limiting GP W

- Suppose that $\widehat{f}_{n,\lambda}(\cdot) = \sum_{\nu=0}^{\infty} \widehat{f}_{n,\nu}\varphi_\nu(\cdot)$;
- The mean function of W (also the approximate posterior mode of $P(\cdot|D_n)$) is

$$\widetilde{f}_{n,\lambda} := \sum_{\nu=0}^{\infty} a_{n,\nu}\widehat{f}_{n,\nu}\varphi_\nu(\cdot).$$

Hence, $\widetilde{f}_{n,\lambda} \neq \widehat{f}_{n,\lambda}$;

- The mean-zero GP $W_n := W - \widetilde{f}_{n,\lambda}$ is expressed as

$$W_n(\cdot) = \sum_{\nu=0}^{\infty} b_{n,\nu} z_\nu \varphi_\nu(\cdot) \text{ and } z_\nu \overset{iid}{\sim} N(0,1);$$

- Here, $a_{n,\nu}$ and $b_{n,\nu}$ are both non-random sequences.

Specifications of the Limiting GP W

- Suppose that $\widehat{f}_{n,\lambda}(\cdot) = \sum_{\nu=0}^{\infty} \widehat{f}_{n,\nu}\varphi_\nu(\cdot)$;
- The mean function of W (also the approximate posterior mode of $P(\cdot|D_n)$) is

$$\widetilde{f}_{n,\lambda} := \sum_{\nu=0}^{\infty} a_{n,\nu}\widehat{f}_{n,\nu}\varphi_\nu(\cdot).$$

  Hence, $\widetilde{f}_{n,\lambda} \neq \widehat{f}_{n,\lambda}$;
- The mean-zero GP $W_n := W - \widetilde{f}_{n,\lambda}$ is expressed as

$$W_n(\cdot) = \sum_{\nu=0}^{\infty} b_{n,\nu} z_\nu \varphi_\nu(\cdot) \quad \text{and} \quad z_\nu \overset{iid}{\sim} N(0,1);$$

- Here, $a_{n,\nu}$ and $b_{n,\nu}$ are both non-random sequences.

Specifications of the Limiting GP W

- Suppose that $\widehat{f}_{n,\lambda}(\cdot) = \sum_{\nu=0}^{\infty} \widehat{f}_{n,\nu}\varphi_\nu(\cdot)$;
- The mean function of W (also the approximate posterior mode of $P(\cdot|D_n)$) is

$$\widetilde{f}_{n,\lambda} := \sum_{\nu=0}^{\infty} a_{n,\nu}\widehat{f}_{n,\nu}\varphi_\nu(\cdot).$$

  Hence, $\widetilde{f}_{n,\lambda} \neq \widehat{f}_{n,\lambda}$;
- The mean-zero GP $W_n := W - \widetilde{f}_{n,\lambda}$ is expressed as

$$W_n(\cdot) = \sum_{\nu=0}^{\infty} b_{n,\nu} z_\nu \varphi_\nu(\cdot) \ \text{ and } \ z_\nu \overset{iid}{\sim} N(0,1);$$

- Here, $a_{n,\nu}$ and $b_{n,\nu}$ are both non-random sequences.

Recall "Bayesian Aggregation"

$$
\begin{array}{ccccc}
 & & \text{Subset } 1\ (n) & \xrightarrow{\text{Machine } 1} & R_{1,n}(\alpha) \\
\text{Big Data } (N) & \xrightarrow{\text{Divide}} & \text{Subset } 2\ (n) & \xrightarrow{\text{Machine } 2} & R_{2,n}(\alpha) \\
 & & \cdots & \cdots & \cdots \\
 & & \text{Subset } s\ (n) & \xrightarrow{\text{Machine } s} & R_{s,n}(\alpha) \\
\text{Super}\Big\Downarrow\text{machine} & & & & \text{Aggre}\Big\Downarrow\text{gate} \\
R_{\text{oracle}}(\alpha) & & & & R_N(\alpha)
\end{array}
$$

Note that $N = s \times n$.
Both n and s are allowed to diverge.

Uniform Nonparametric BvM Theorem

Uniform BvM theorem characterizes limit shapes of a sequence of s nonparametric posterior distributions (under proper tuning priors) as long as s does not grow too fast.

Theorem 2

Given that $\lambda \asymp N^{-2m/(2m+\beta)}$ (used in each subset with size n), we have

$$\sup_{S \subset H^m(0,1)} \max_{1 \le j \le s} |P(S|D_{j,n}) - \Pi_{W_j}(S)| = o_{P^n_{f_0}}(1)$$

as long as s does not grow faster than $N^{(\beta-1)/(2m+\beta)}$.

Aggregated Credible Interval

- The j-th credible ball is defined as

$$R_{j,n}(\alpha) = \{f \in H^m(0,1) : \|f - \widetilde{f}_{j,n}\|_2 \leq r_{j,n}(\alpha)\},$$

  where the radius $r_{j,n}(\alpha)$ is directly obtained via MCMC;

- The aggregated credible ball is constructed as

$$R_N(\alpha) = \{f \in H^m(0,1) : \|f - \bar{f}_{N,\lambda}\|_2 \leq \bar{r}_N(\alpha)\};$$

- As will be seen, the aggregation step is through weighted averaging Fourier frequencies and weighted averaging individual radii. No additional computation is needed;
- In theory, uniform BvM shows that $R_N(\alpha)$ (asymptotically) covers $(1-\alpha)$ posterior mass and also possesses frequentist validity when $\lambda \asymp N^{-2m/(2m+\beta)}$ and $s = o(N^{(\beta-1/(2m+\beta))})$.

Aggregated Credible Interval

- The j-th credible ball is defined as

$$R_{j,n}(\alpha) = \{f \in H^m(0,1) : \|f - \widetilde{f}_{j,n}\|_2 \le r_{j,n}(\alpha)\},$$

  where the radius $r_{j,n}(\alpha)$ is directly obtained via MCMC;
- The aggregated credible ball is constructed as

$$R_N(\alpha) = \{f \in H^m(0,1) : \|f - \bar{f}_{N,\lambda}\|_2 \le \bar{r}_N(\alpha)\};$$

- As will be seen, the aggregation step is through weighted averaging Fourier frequencies and weighted averaging individual radii. No additional computation is needed;
- In theory, uniform BvM shows that $R_N(\alpha)$ (asymptotically) covers $(1-\alpha)$ posterior mass and also possesses frequentist validity when $\lambda \asymp N^{-2m/(2m+\beta)}$ and $s = o(N^{(\beta-1/(2m+\beta))})$.

Aggregated Credible Interval

- The j-th credible ball is defined as

$$R_{j,n}(\alpha) = \{f \in H^m(0,1) : \|f - \widetilde{f}_{j,n}\|_2 \leq r_{j,n}(\alpha)\},$$

  where the radius $r_{j,n}(\alpha)$ is directly obtained via MCMC;
- The aggregated credible ball is constructed as

$$R_N(\alpha) = \{f \in H^m(0,1) : \|f - \bar{f}_{N,\lambda}\|_2 \leq \bar{r}_N(\alpha)\};$$

- As will be seen, the aggregation step is through weighted averaging Fourier frequencies and weighted averaging individual radii. No additional computation is needed;
- In theory, uniform BvM shows that $R_N(\alpha)$ (asymptotically) covers $(1-\alpha)$ posterior mass and also possesses frequentist validity when $\lambda \asymp N^{-2m/(2m+\beta)}$ and $s = o(N^{(\beta-1/(2m+\beta))})$.

Aggregated Credible Interval

- The j-th credible ball is defined as

$$R_{j,n}(\alpha) = \{f \in H^m(0,1) : \|f - \widetilde{f}_{j,n}\|_2 \leq r_{j,n}(\alpha)\},$$

  where the radius $r_{j,n}(\alpha)$ is directly obtained via MCMC;
- The aggregated credible ball is constructed as

$$R_N(\alpha) = \{f \in H^m(0,1) : \|f - \bar{f}_{N,\lambda}\|_2 \leq \bar{r}_N(\alpha)\};$$

- As will be seen, the aggregation step is through weighted averaging Fourier frequencies and weighted averaging individual radii. No additional computation is needed;
- In theory, uniform BvM shows that $R_N(\alpha)$ (asymptotically) covers $(1-\alpha)$ posterior mass and also possesses frequentist validity when $\lambda \asymp N^{-2m/(2m+\beta)}$ and $s = o(N^{(\beta-1/(2m+\beta))})$.

Aggregation Details

- Aggregted center:

$$\bar{f}_{N,\lambda}(\cdot) = \sum_{\nu=1}^{\infty} a_{N,\nu} \bar{f}_\nu \varphi_\nu(\cdot) \text{ and } \bar{f}_\nu = (1/s) \sum_{j=1}^{s} \widehat{f}_{n,\nu}^{(j)};$$

- Aggregted radius:

$$\bar{r}_N(\alpha) = \sqrt{\frac{1}{N}\left[\zeta_{1,N} + \sqrt{\frac{\zeta_{2,N}}{\zeta_{2,n}}}\left(\frac{n}{s}\sum_{j=1}^{s} r_{j,n}^2(\alpha) - \zeta_{1,n}\right)\right]},$$

where

$$\zeta_{k,n} = \sum_{\nu=1}^{\infty}\left(\frac{n}{\tau_\nu^2 + n(1+\lambda\rho_\nu)}\right)^k.$$

- In fact, the aggregated radius $\bar{r}_N$ is (asymptotically) the same as that of oracle credible ball; see simulations.

Aggregation Details

- Aggregted center:

$$\bar{f}_{N,\lambda}(\cdot) = \sum_{\nu=1}^{\infty} a_{N,\nu} \bar{f}_\nu \varphi_\nu(\cdot) \quad \text{and} \quad \bar{f}_\nu = (1/s) \sum_{j=1}^{s} \widehat{f}_{n,\nu}^{(j)};$$

- Aggregted radius:

$$\bar{r}_N(\alpha) = \sqrt{\frac{1}{N}\left[\zeta_{1,N} + \sqrt{\frac{\zeta_{2,N}}{\zeta_{2,n}}}\left(\frac{n}{s}\sum_{j=1}^{s} r_{j,n}^2(\alpha) - \zeta_{1,n}\right)\right]},$$

where

$$\zeta_{k,n} = \sum_{\nu=1}^{\infty}\left(\frac{n}{\tau_\nu^2 + n(1+\lambda\rho_\nu)}\right)^k.$$

- In fact, the aggregated radius $\bar{r}_N$ is (asymptotically) the same as that of oracle credible ball; see simulations.

Aggregation Details

- Aggregted center:

$$\bar{f}_{N,\lambda}(\cdot) = \sum_{\nu=1}^{\infty} a_{N,\nu} \bar{f}_{\nu} \varphi_{\nu}(\cdot) \ \text{ and } \ \bar{f}_{\nu} = (1/s) \sum_{j=1}^{s} \widehat{f}_{n,\nu}^{(j)};$$

- Aggregted radius:

$$\bar{r}_N(\alpha) = \sqrt{\frac{1}{N}\left[\zeta_{1,N} + \sqrt{\frac{\zeta_{2,N}}{\zeta_{2,n}}}\left(\frac{n}{s}\sum_{j=1}^{s} r_{j,n}^2(\alpha) - \zeta_{1,n}\right)\right]},$$

where

$$\zeta_{k,n} = \sum_{\nu=1}^{\infty}\left(\frac{n}{\tau_\nu^2 + n(1+\lambda\rho_\nu)}\right)^k.$$

- In fact, the aggregated radius $\bar{r}_N$ is (asymptotically) the same as that of oracle credible ball; see simulations.

Aggregated Credible Interval

We can also aggregate individual credible intervals for linear functionals of f, denoted as $F(f)$.

- Two examples:
  - Evaluation functional: $F_z(f) = f(z)$;
  - Integral functional: $F_\omega(f) = \int_0^1 f(z)\omega(z)dz$ for a known function $\omega(\cdot)$ such as an indicator function;
- Individual credible interval for $F(f)$:

$$CI_{j,n}^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\widetilde{f}_{j,n})| \leq r_{F,j,n}(\alpha)\};$$

- The aggregated version is constructed as

$$CI_N^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\bar{f}_{N,\lambda})| \leq \bar{r}_{F,N}(\alpha)\},$$

  where $\bar{r}_{F,N}(\alpha)$ is a weighted ℓ_2 average of $r_{F,j,n}(\alpha)$'s.

Aggregated Credible Interval

We can also aggregate individual credible intervals for linear functionals of f, denoted as $F(f)$.

- Two examples:
  - Evaluation functional: $F_z(f) = f(z)$;
  - Integral functional: $F_\omega(f) = \int_0^1 f(z)\omega(z)dz$ for a known function $\omega(\cdot)$ such as an indicator function;
- Individual credible interval for $F(f)$:

$$CI_{j,n}^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\widetilde{f}_{j,n})| \le r_{F,j,n}(\alpha)\};$$

- The aggregated version is constructed as

$$CI_N^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\bar{f}_{N,\lambda})| \le \bar{r}_{F,N}(\alpha)\},$$

where $\bar{r}_{F,N}(\alpha)$ is a weighted ℓ_2 average of $r_{F,j,n}(\alpha)$'s.

Aggregated Credible Interval

We can also aggregate individual credible intervals for linear functionals of f, denoted as $F(f)$.

- Two examples:
  - Evaluation functional: $F_z(f) = f(z)$;
  - Integral functional: $F_\omega(f) = \int_0^1 f(z)\omega(z)dz$ for a known function $\omega(\cdot)$ such as an indicator function;
- Individual credible interval for $F(f)$:

$$CI_{j,n}^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\widetilde{f}_{j,n})| \le r_{F,j,n}(\alpha)\};$$

- The aggregated version is constructed as

$$CI_N^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\bar{f}_{N,\lambda})| \le \bar{r}_{F,N}(\alpha)\},$$

where $\bar{r}_{F,N}(\alpha)$ is a weighted ℓ_2 average of $r_{F,j,n}(\alpha)$'s.

Aggregated Credible Interval

We can also aggregate individual credible intervals for linear functionals of f, denoted as $F(f)$.

- Two examples:
  - Evaluation functional: $F_z(f) = f(z)$;
  - Integral functional: $F_\omega(f) = \int_0^1 f(z)\omega(z)dz$ for a known function $\omega(\cdot)$ such as an indicator function;
- Individual credible interval for $F(f)$:

$$CI_{j,n}^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\widetilde{f}_{j,n})| \leq r_{F,j,n}(\alpha)\};$$

- The aggregated version is constructed as

$$CI_N^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\bar{f}_{N,\lambda})| \leq \bar{r}_{F,N}(\alpha)\},$$

where $\bar{r}_{F,N}(\alpha)$ is a weighted ℓ_2 average of $r_{F,j,n}(\alpha)$'s.

Aggregated Credible Interval

We can also aggregate individual credible intervals for linear functionals of f, denoted as $F(f)$.

- Two examples:
  - Evaluation functional: $F_z(f) = f(z)$;
  - Integral functional: $F_\omega(f) = \int_0^1 f(z)\omega(z)dz$ for a known function $\omega(\cdot)$ such as an indicator function;
- Individual credible interval for $F(f)$:

$$CI_{j,n}^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\widetilde{f}_{j,n})| \le r_{F,j,n}(\alpha)\};$$

- The aggregated version is constructed as

$$CI_N^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\bar{f}_{N,\lambda})| \le \bar{r}_{F,N}(\alpha)\},$$

where $\bar{r}_{F,N}(\alpha)$ is a weighted ℓ_2 average of $r_{F,j,n}(\alpha)$'s.

Aggregated Credible Interval

We can also aggregate individual credible intervals for linear functionals of f, denoted as $F(f)$.

- Two examples:
  - Evaluation functional: $F_z(f) = f(z)$;
  - Integral functional: $F_\omega(f) = \int_0^1 f(z)\omega(z)dz$ for a known function $\omega(\cdot)$ such as an indicator function;
- Individual credible interval for $F(f)$:

$$CI_{j,n}^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\widetilde{f}_{j,n})| \le r_{F,j,n}(\alpha)\};$$

- The aggregated version is constructed as

$$CI_N^F(\alpha) := \{f \in S^m(\mathbb{I}) : |F(f) - F(\bar{f}_{N,\lambda})| \le \bar{r}_{F,N}(\alpha)\},$$

where $\bar{r}_{F,N}(\alpha)$ is a weighted ℓ_2 average of $r_{F,j,n}(\alpha)$'s.

A Series of Theoretical Questions...

- How to define an aggregation rule s.t. $R(\alpha)$ covers $(1-\alpha)$ posterior mass, with the same radius as $R_{\text{oracle}}(\alpha)$?

  Weighted averaging individual centers (in terms of their Fourier coefficients) and radii by *analytical formula.*

- How to construct a prior s.t. $R(\alpha)$ covers the true parameter (generating the data) with probability $(1-\alpha)$?

  Pick a proper tuning prior by GCV.

- How fast can we allow s to diverge?

  s cannot grow faster than a rate jointly determined by the smoothness of f_0 and the smoothness of GP prior.

A Series of Theoretical Questions...

- How to define an aggregation rule s.t. $R(\alpha)$ covers $(1-\alpha)$ posterior mass, with the same radius as $R_{\text{oracle}}(\alpha)$?

  Weighted averaging individual centers (in terms of their Fourier coefficients) and radii by *analytical formula.*

- How to construct a prior s.t. $R(\alpha)$ covers the true parameter (generating the data) with probability $(1-\alpha)$?

  Pick a proper tuning prior by GCV.

- How fast can we allow s to diverge?

  s cannot grow faster than a rate jointly determined by the smoothness of f_0 and the smoothness of GP prior.

A Series of Theoretical Questions...

- How to define an aggregation rule s.t. $R(\alpha)$ covers $(1-\alpha)$ posterior mass, with the same radius as $R_{\text{oracle}}(\alpha)$?

  Weighted averaging individual centers (in terms of their Fourier coefficients) and radii by *analytical formula.*

- How to construct a prior s.t. $R(\alpha)$ covers the true parameter (generating the data) with probability $(1-\alpha)$?

  Pick a proper tuning prior by GCV.

- How fast can we allow s to diverge?

  s cannot grow faster than a rate jointly determined by the smoothness of f_0 and the smoothness of GP prior.

Simulations

- Gaussian regression models:

$$Y = f_0(X) + \epsilon,$$

where $\epsilon \sim N(0,1)$ and

$$f_0(x) = 3\beta_{30,17}(x) + 2\beta_{3,11}(x),$$

where $\beta_{a,b}$ is the pdf of Beta distribution. Set $m = 2$;

- Assign a tuning prior with $\beta = 2$ and λ being selected by GCV as follows;
- Let λ_{GCV} be the GCV-selected tuning parameter with the order $N^{-2m/(2m+1)}$ by applying to the entire data (A practical formula needs to be developed here). Set λ as $\lambda_{GCV}^{(2m+1)/(2m+\beta)}$ to match with the order $\asymp N^{-2m/(2m+\beta)}$.

Simulations

- Gaussian regression models:

$$Y = f_0(X) + \epsilon,$$

where $\epsilon \sim N(0,1)$ and

$$f_0(x) = 3\beta_{30,17}(x) + 2\beta_{3,11}(x),$$

where $\beta_{a,b}$ is the pdf of Beta distribution. Set $m = 2$;

- Assign a tuning prior with $\beta = 2$ and λ being selected by GCV as follows;
- Let λ_{GCV} be the GCV-selected tuning parameter with the order $N^{-2m/(2m+1)}$ by applying to the entire data (A practical formula needs to be developed here). Set λ as $\lambda_{GCV}^{(2m+1)/(2m+\beta)}$ to match with the order $\asymp N^{-2m/(2m+\beta)}$.

Simulations

- Gaussian regression models:

$$Y = f_0(X) + \epsilon,$$

where $\epsilon \sim N(0,1)$ and

$$f_0(x) = 3\beta_{30,17}(x) + 2\beta_{3,11}(x),$$

where $\beta_{a,b}$ is the pdf of Beta distribution. Set $m = 2$;

- Assign a tuning prior with $\beta = 2$ and λ being selected by GCV as follows;
- Let λ_{GCV} be the GCV-selected tuning parameter with the order $N^{-2m/(2m+1)}$ by applying to the entire data (A practical formula needs to be developed here). Set λ as $\lambda_{GCV}^{(2m+1)/(2m+\beta)}$ to match with the order $\asymp N^{-2m/(2m+\beta)}$.

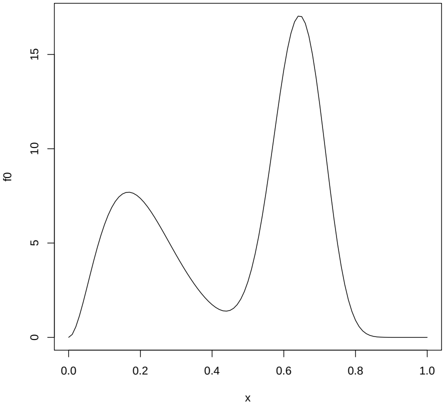

Figure 1: *Plot of the true function f_0.*

Computing Time

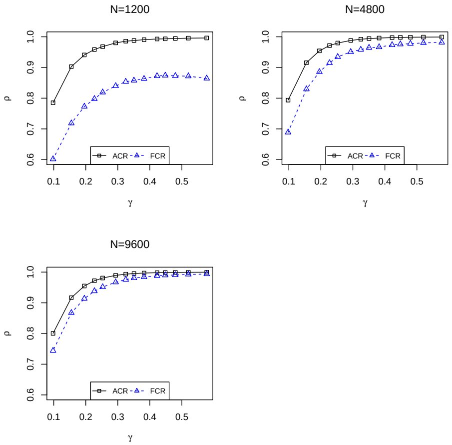

Figure 2: *ρ versus γ based on FCR and ACR, where $\rho = (T_0 - T)/T_0$, T_0 is computing time based on big data and T is the D&C time. And, $\gamma = \log s / \log N$ describes the growth of s.*

Phase Transition: Coverage Probability

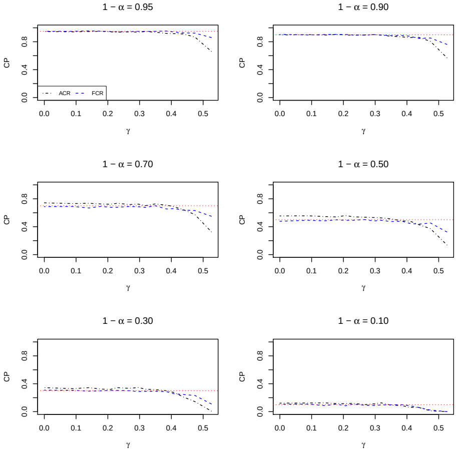

Figure 3: *Frequentist coverage probability (CP) of* $R_N(\alpha)$ *against* γ *for* $N = 2400$. *Red-dotted line indicates the position of* $1 - \alpha$.

Phase Transition: Radius

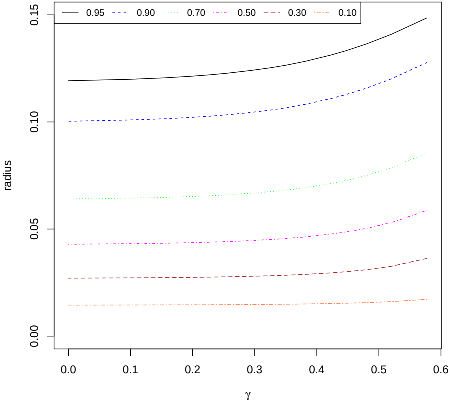

Figure 4: *Radius of $R_N(\alpha)$ against γ for various α.*