Convergence of latent mixing measures in finite and infinite mixture models

Long Nguyen

Department of Statistics
University of Michigan

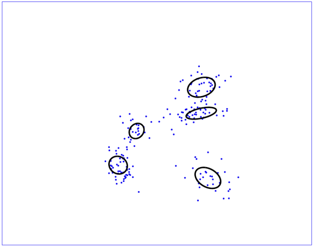
BNP Workshop, ICERM 2012

Outline

1. Identifiability and consistency in mixture model-based clustering
   - convergence of mixing measures
2. Wasserstein metric
3. Posterior concentration rates of mixing measures
   - finite mixture models
   - Dirichlet process mixture models
4. Implications and proof ideas

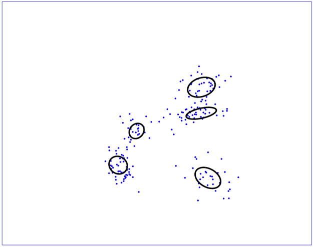
Clustering problem

How do we subdivide $D = \{X_1, \ldots, X_n\}$ in $\mathbb{R}^d$ into clusters?



Clustering problem

How do we subdivide $D = \{X_1, \ldots, X_n\}$ in $\mathbb{R}^d$ into clusters?



Assume that data $X_1, \ldots, X_n$ are iid sample from a mixture model

$$p_G(x) = \sum_{i=1}^{k} p_i f(x|\theta_i)$$

How do we guarantee consistent estimates for mixture components $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ and $\mathbf{p} = (p_1, \ldots, p_k)$?

Bayesian nonparametric approach

Define mixing distribution:

$$G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$$

Endow G with a prior distribution on the space of probability measures $\bar{\mathcal{G}}(\Theta)$

- for finite mixtures, k is given
  use parametric priors on mixing probabilities $\mathbf{p}$ and $\boldsymbol{\theta}$

- for infinite mixtures, k is unknown
  use a nonparametric prior such as the Dirichlet process:

$$G \sim \mathrm{DP}(\gamma, H)$$

Bayesian nonparametric approach

Define mixing distribution:

$$G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$$

Endow G with a prior distribution on the space of probability measures $\bar{\mathcal{G}}(\Theta)$

- for finite mixtures, k is given
  use parametric priors on mixing probabilities $\mathbf{p}$ and $\boldsymbol{\theta}$

- for infinite mixtures, k is unknown
  use a nonparametric prior such as the Dirichlet process:

$$G \sim \mathrm{DP}(\gamma, H)$$

Compute posterior distribution of G given data, $\Pi(G|X_1, \ldots, X_n)$

We are interested in concentration behavior of the posterior of G

Posterior concentration of mixing measure G

Let $X_1, \ldots, X_n$ be an iid sample from the mixture density

$$p_G(x) = \int f(x|\theta)G(d\theta)$$

f is known, while $G = G_0$ unknown discrete mixing measure

Questions

- when does the posterior distribution $\Pi(G|X_1, \ldots, X_n)$ concentrate most of its mass around the "truth" G_0?
- what is the rate of concentration (convergence)?

Related Work

Significant advances in posterior asymptotics (i.e., posterior consistency and convergence rates)

- general theory: Barron, Shervish & Wasserman (1999), Ghosal, Ghosh & van der Vaart (2000), Shen & Wasserman (2000), Walker (2004), Ghosal & van der Vaart (2007), Walker, Lijoi & Prunster (2007), ... going back to work of Schwarz (1965) and Le Cam (1973)

- mixture models: Ghosal, Ghosh & Ramamoorthi (1999), Genovese & Wasserman (2000), Ishwaran & Zarepour (2002), Ghosal & van der Vaart (2007), ...

These work focus mostly on the posterior concentration behavior of the data density p_G, not mixing measure G per se

Related Work on mixture models

Convergence of parameters $\mathbf{p}$ and θ in certain finite mixture settings:

- polynomial-time learnable settings: Kalai, Moitra, and Valiant (2010), Belkin & Sinha (2010); going back to Dasgupta (2000)
- overfitted setting: Rousseau & Mengersen (JRSSB, 2011)

Convergence of mixing measure G in a univariate finite mixture:

- settled by Jiahua Chen (AOS, 1995), who established optimal rate $n^{-1/4}$
- Bayesian asymptotics by Ishwaran, James and Sun (JASA, 2001)

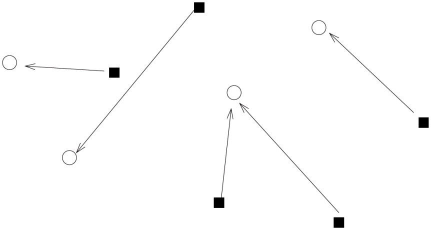
Literature on deconvolution in kernel density estimation, in '80 and early '90 (Hall, Carroll, Fan, Zhang, ...)

Posterior concentration behavior of mixing measures in multivariate finite mixtures, and infinite mixtures remains unresolved

Outline

Optimal transportation problem (Monge/Kantorovich, cf. Villani, '03)

How to optimally transport to goods from a collection of producers to a collection of consumers, all of which are located in some space?



squares: locations of producers; circles: locations of consumers

The optimal cost of transportation defines a (Wasserstein) distance between "production density" and "consumption density".

Wasserstein metric (cont)

Let $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$, $G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$. A coupling between $\mathbf{p}$ and $\mathbf{p'}$ is a joint distribution $\mathbf{q}$ on $[1, \ldots, k] \times [1, \ldots, k']$ whose marginals are $\mathbf{p}$ and $\mathbf{p'}$.
That is, for any $(i, j) \in [1, \ldots, k] \times [1, \ldots, k']$,

$$\sum_{i=1}^{k} q_{ij} = p_j; \quad \sum_{j=1}^{k'} q_{ij} = p_i.$$

Definition

Let ρ be a distance function on Θ, the Wasserstein distance is defined by:

$$d_\rho(G, G') = \inf_{\mathbf{q}} \sum_{i,j} q_{ij} \rho(\theta_i, \theta'_j).$$

When $\Theta \subset \mathbb{R}^d$, ρ is Euclidean metric on $\mathbb{R}^d$, for $r \geq 1$, use ρ^r as a distance function on $\mathbb{R}^d$ to obtain L_r Wasserstein metric:

$$W_r(G, G') := \left[\inf_{\mathbf{q}} \sum_{i,j} q_{ij} \|\theta_i - \theta'_j\|^r \right]^{1/r}.$$

Examples and Facts

Wasserstein distance W_r metrizes weak convergence in the space of probability measures on Θ.

Examples and Facts

Wasserstein distance W_r metrizes weak convergence in the space of probability measures on Θ.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$.

Examples and Facts

Wasserstein distance W_r metrizes weak convergence in the space of probability measures on Θ.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$.

If $G_0 = \delta_{\theta_0}$ and $G = \sum_{i=1}^k p_i \delta_{\theta_i}$, then

$$W_1(G_0, G) = \sum_{i=1}^{k} p_i \|\theta_0 - \theta_i\|.$$

Examples and Facts

Wasserstein distance W_r metrizes weak convergence in the space of probability measures on Θ.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$.

If $G_0 = \delta_{\theta_0}$ and $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$, then

$$W_1(G_0, G) = \sum_{i=1}^{k} p_i \|\theta_0 - \theta_i\|.$$

If $G = \sum_{i=1}^{k} \frac{1}{k} \delta_{\theta_i}$, $G' = \sum_{j=1}^{k} \frac{1}{k} \delta_{\theta'_j}$, then

$$W_1(G, G') = \inf_{\pi} \sum_{i=1}^{k} \frac{1}{k} \|\theta_i - \theta'_{\pi(i)}\|,$$

where π ranges over the set of permutations on $(1, \ldots, k)$.

Relations between Wasserstein distances and divergences

If $W_2(G, G') = 0$, then clearly $G = G'$ so that

$$V(p_G, p_{G'}) = h(p_G, p_{G'}) = K(p_G, p_{G'}) = 0.$$

It can be shown that an f-divergence (e.g., variational distance V, Hellinger h, Kullback-Leibler distance K) between $p_G, p_{G'}$ is always bounded from above by a Wasserstein distance

- if $f(x|\theta)$ is Gaussian with mean parameter θ, then

$$h(p_G, p_{G'}) \leq W_2(G, G')/2\sqrt{2}.$$

- if $f(x|\theta)$ is Gamma with location parameter θ, then

$$K(p_G||p_{G'}) = O(W_1(G, G')).$$

Relations between Wasserstein distances and divergences

If $W_2(G, G') = 0$, then clearly $G = G'$ so that

$$V(p_G, p_{G'}) = h(p_G, p_{G'}) = K(p_G, p_{G'}) = 0.$$

It can be shown that an f-divergence (e.g., variational distance V, Hellinger h, Kullback-Leibler distance K) between $p_G, p_{G'}$ is always bounded from above by a Wasserstein distance

- if $f(x|\theta)$ is Gaussian with mean parameter θ, then

$$h(p_G, p_{G'}) \leq W_2(G, G')/2\sqrt{2}.$$

- if $f(x|\theta)$ is Gamma with location parameter θ, then

$$K(p_G || p_{G'}) = O(W_1(G, G')).$$

Conversely: if the distance between p_G, p'_G is small, can we ensure that $W_2(G, G')$ (or $W_1(G, G')$, etc) be small?

Identifiability in mixture models

The family $\{f(\cdot|\theta), \theta \in \Theta\}$ is identifiable if for any $G, G' \in \mathcal{G}(\Theta)$, $|p_G(x) - p_{G'}(x)| = 0$ for almost all x implies that $G = G'$.

- $\mathcal{G}(\Theta)$ is space of discrete measures with finite support points on Θ, $\bar{\mathcal{G}}(\Theta)$ is space of all discrete measures on Θ

Identifiability in mixture models

The family $\{f(\cdot|\theta), \theta \in \Theta\}$ is identifiable if for any $G, G' \in \mathcal{G}(\Theta)$, $|p_G(x) - p_{G'}(x)| = 0$ for almost all x implies that $G = G'$.

- $\mathcal{G}(\Theta)$ is space of discrete measures with finite support points on Θ, $\bar{\mathcal{G}}(\Theta)$ is space of all discrete measures on Θ

Stronger notion of identifiability (due to Chen (1995) for univariate case)

Strong identifiability

Let $\Theta \subseteq \mathbb{R}^d$. The family $\{f(\cdot|\theta), \theta \in \mathbb{R}^d\}$ is strongly identifiable if $f(x|\theta)$ is twice differentiable in θ, and for any finite k and distinct $\theta_1, \ldots, \theta_k$, the equality

$$\sup_{x \in \mathcal{X}} \left| \sum_{i=1}^{k} \alpha_i f(x|\theta_i) + \beta_i^T Df(x|\theta_i) + \gamma_i^T D^2 f(x|\theta_i) \gamma_i \right| = 0 \quad (1)$$

implies that $\alpha_i = 0$, $\beta_i = \gamma_i = \mathbf{0} \in \mathbb{R}^d$ for $i = 1, \ldots, k$. Here, $Df(x|\theta_i)$ and $D^2 f(x|\theta_i)$ denote the gradient and the Hessian at θ_i of $f(x|\cdot)$, resp.

Wasserstein identifiability: finite mixtures

Suppose that

- Θ is a compact subset of $\mathbb{R}^d$
- the family $\{f(\cdot|\theta)\}$ is strongly identifiable
- the Hessian matrix $D^2 f(x|\theta)$ satisfies a uniform Lipschitz condition
- $\mathcal{G}_k(\Theta)$ denotes the space of discrete measures with at most $k < \infty$ support points in Θ.

Theorem 1 (Nguyen, 2012)

For any $G_0 \in \mathcal{G}_k(\Theta)$, there is a constant $C_0 = C_0(k, G_0) > 0$ such that

$$W_2^2(G_0, G) \leq C_0 \times V(p_{G_0}, p_G) \;\; \forall G \in \mathcal{G}_k(\Theta)$$

$V(\cdot, \cdot)$ denotes the variational distance between two densities.

Wasserstein identifiability: infinite mixtures

Let $G \in \bar{\mathcal{G}}(\Theta)$ (i.e., G has potentially unbounded number of support points)

We are restricted to convolution mixture models, i.e., $f(x|\theta)$ takes the form $f(x-\theta)$ for some multivariate density function f on $\mathbb{R}^d$, so that

$$p_G(x) = G * f(x) = \sum_i p_i f(x - \theta_i).$$

Suppose that

- Θ is a bounded subset of $\mathbb{R}^d$
- f is a density function on $\mathbb{R}^d$ that is symmetric around 0.
- Fourier transform $\tilde{f}(\omega) \neq 0$ for all $\omega \in \mathbb{R}^d$.

Theorem 2 (Nguyen, 2012)

Given assumptions on Θ and f in the previous page.

(1) **Ordinary smooth likelihood.** If $|\tilde{f}(\omega)\prod_{j=1}^{d}|\omega_j|^{\beta}| \geq d_0$ as $\omega_j \to \infty$, $(j = 1, \ldots, d)$ for some positive constants d_0 and β.

Then for any $m < 4/(4 + (2\beta + 1)d)$, there is some constant $C_1 = C_1(d, \beta, m) > 0$ such that for any $G, G' \in \bar{\mathcal{G}}(\Theta)$,

$$W_2^2(G, G') \leq C_1 \times V(p_G, p_{G'})^m.$$

(2) **Supersmooth likelihood.** If $|\tilde{f}(\omega)\prod_{j=1}^{d}\exp(|\omega_j|^{\beta}/\gamma)| \geq d_0$ as $\omega_j \to \infty$, $(j = 1, \ldots, d)$ for some positive constants β, γ, d_0.

Then there is some constant $C_1 = C_1(d, \beta) > 0$ such that for any $G, G' \in \bar{\mathcal{G}}(\Theta)$,

$$W_2^2(G, G') \leq C_1 \times (-\log V(p_G, p_{G'}))^{-2/\beta}.$$

Examples.

If f is the standard normal density on $\mathbb{R}^d$, $\tilde{f}(\omega) = \prod_{j=1}^{d} e^{-\omega_i^2/2}$, we obtain that

$$W_2^2(G, G') \lesssim \frac{1}{\log(1/V(p_G, p_{G'}))}.$$

If f is a Laplace density on $\mathbb{R}$, e.g., $\tilde{f}(\omega) = \frac{1}{1+\omega^2}$, then

$$W_2^2(G, G') \lesssim V(p_G, p_{G'})^m$$

for any $m < 4/9$.

Outline

Main result: Finite mixtures

$k < \infty$ is known, Π is a prior distribution of mixing measures in $\mathcal{G}_k(\Theta)$.

Suppose that the "truth" $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$. Moreover,

(A1) Θ is compact subset of $\mathbb{R}^d$, and the family of likelihood functions $f(\cdot|\theta)$ is strongly identifiable.

(A2) under prior Π, all p_i are bounded away from 0, and all pairwise distances $\|\theta_i - \theta_j\|$ are bounded away from 0.

(A3) some additional mild conditions on Π

Theorem 3

Let $X_1, \ldots, X_n$ be an iid sample from P_{G_0}, where $G_0 \in \mathcal{G}_k(\Theta)$. Under Assumptions (A1–A3), there is a constant $M > 0$ such that

$$\Pi(W_2(G_0, G) \geq Mn^{-1/4} | X_1, \ldots, X_n) \to 0$$

in P_{G_0}-probability, as $n \to \infty$.

Main result: Dirichlet process mixtures

Given the "true" discrete measure $G_0 = \sum_{i=1}^{k} p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$, but k is unknown (potentially infinite)

Endow $G \in \bar{\mathcal{G}}(\Theta)$ with Dirichlet process prior $G \sim \mathrm{DP}(\nu, P_0)$ for some $\nu > 0$ and non-atomic $P_0 \in \mathcal{P}(\Theta)$.

Furthermore,

(B1) $\Theta \subset \mathbb{R}^d$ is compact, and P_0 has a Lebesgue density that is bounded away from zero.

(B2) For some constants $C_1, m_1 > 0$, $K(f_i, f_j') \leq C_1 \rho^{m_1}(\theta_i, \theta_j')$ for any $\theta_i, \theta_j' \in \Theta$. For any $G \in \mathrm{spt}(\Pi)$, $\int p_{G_0}(\log(p_{G_0}/p_G))^2 \leq C_2 K(p_{G_0}, p_G)^{m_2}$ for some constants $C_2, m_2 > 0$.

Theorem 4

Let $X_1, \ldots, X_n$ be an iid sample from P_{G_0}, where $G_0 \in \bar{\mathcal{G}}(\Theta)$. Given Assumptions (B1) and (B2) and the smoothness conditions for the likelihood family, there is a sequence $\beta_n \searrow 0$ such that

$$\Pi(W_2(G_0, G) \geq \beta_n | X_1, \ldots, X_n) \to 0$$

in P_{G_0} probability. Specifically,

(1) for ordinary smooth likelihood functions, take $\beta_n \asymp (\log n/n)^{\frac{2}{(d+2)(4+(2\beta+1)d)+\delta}}$, for any small $\delta > 0$.

(2) for supersmooth likelihood functions, take $\beta_n \asymp (\log n)^{-1/\beta}$.

Outline

Key elements of proof

We follow standard method of proof (cf., Ghosal, Ghosh & van der Vaart (2000), Ghosh & Ramamoorthi (2002))

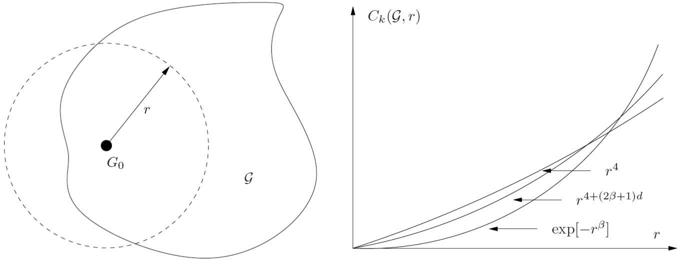
- existence of tests that discriminate a mixing measure G from the complement of a ball
- the (induced) prior distribution on p_G is sufficiently dense in Kullback-Leibler distance

Technical challenges: the analysis has to be done in Wasserstein metric W_2 on G, as opposed to the standard Hellinger metric h for data density p_G

Existence of tests

Suppose that G_0 has k support points. Let $\mathcal{G} \subset \bar{\mathcal{G}}(\Theta)$.
The Hellinger information of W_2 metric for $\mathcal{G}$ is

$$C_k(\mathcal{G}, r) = \inf_{G \in \mathcal{G}: W_2(G_0, G) \geq r} h^2(p_{G_0}, p_G).$$



- both $\mathcal{G}$ and $C_k(\mathcal{G}, \cdot)$ may be non-convex.
- behavior near 0 of $C_k(\mathcal{G}, \cdot)$ depends on both $f(x|\theta)$ and $\mathcal{G}$

A test φ_n is an indicator function of the iid sample $X_1, \ldots, X_n$.

Lemma

Let $D(\epsilon)$ be the covering number in Wasserstein metric of a certain subset of $\bar{\mathcal{G}}(\Theta)$. There exist tests φ_n such that for any small $\epsilon > 0$,

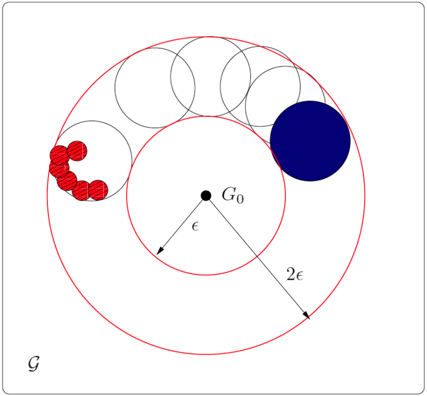
$$\begin{aligned} P_{G_0}\varphi_n &\leq D(\epsilon) \sum_{t=1}^{\lceil \mathrm{diam}(\Theta)/\epsilon \rceil} \exp[-nC_k(\mathcal{G}, t\epsilon)/8] \\ \sup_{G \in \mathcal{G}: W_2(G_0, G) > \epsilon} P_G(1-\varphi_n) &\leq \exp[-nC_k(\mathcal{G}, \epsilon)/8]. \end{aligned}$$

If $\mathcal{G}$ is convex, then $D(\epsilon)$ is the $\epsilon/2$-covering number of the "ring set":

$$\mathcal{S} := \{G : W_2(G_0, G) \in [\epsilon, 2\epsilon]\}$$

If $\mathcal{G}$ is non-convex, then $D(\epsilon)$ is the ϵ'-covering number of set $\mathcal{S}$, where

$$\epsilon' \asymp C_k^{1/4}(\mathcal{G}, \epsilon/2).$$



For convex $\mathcal{G}$, $D(\epsilon)$ is the number **blue** balls, of radius $\epsilon/2$, that cover ring set $\mathcal{S}$.

For non-convex $\mathcal{G}$, $D(\epsilon)$ is the number of red balls, of radius $C_k^{1/4}(\mathcal{G}, \epsilon/2)$.

- for finite mixtures, $C_k^{1/4}(\mathcal{G}, \epsilon/2) = O(\epsilon)$, but typically $C_k^{1/4}(\mathcal{G}, \epsilon/2) = o(\epsilon)$

Entropy bounds

Since Wasserstein metric inherits the geometry of the space of atoms Θ, it is simple to obtain bounds on the covering number in Wasserstein space:

Lemma

(a) $\log N(2\epsilon, \mathcal{G}_k(\Theta), d_\rho) \leq k(\log N(\epsilon, \Theta, \rho) + \log(e + e\,\mathrm{diam}(\Theta)/\epsilon))$.

(b) $\log N(2\epsilon, \bar{\mathcal{G}}(\Theta), d_\rho) \leq N(\epsilon, \Theta, \rho) \log(e + e\,\mathrm{diam}(\Theta)/\epsilon)$.

(c) *Let* $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$. *Assume that* $M = \max_{i=1}^k 1/p_i^* < \infty$ *and* $m = \min_{i,j \leq k} \rho(\theta_i^*, \theta_j^*) > 0$. *Then,*

$$\log N(\epsilon/2, \{G \in \mathcal{G}_k(\Theta) : d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \leq k(\sup_{\Theta'} \log N(\epsilon/4, \Theta', \rho) + \log(32k\,\mathrm{diam}(\Theta)/m)),$$

where the supremum in the right side is taken over all bounded subsets $\Theta' \subseteq \Theta$ *such that* $\mathrm{diam}(\Theta') \leq 4M\epsilon$.

Kullback-Leibler dense property

The Kullback-Leibler dense property, which provides a lower bound on the probability that the Kullback-Leibler distance to a given mixture density p_{G_0} is small can be derived from "small ball probability":

Lemma

Let $G \sim DP(\nu, P_0)$, where P_0 is a non-atomic base probability measure on a compact set Θ. For a small $\epsilon > 0$, let $D = D(\epsilon, \Theta, \rho)$ denote the packing number of Θ under ρ metric. Then, under the Dirichlet process distribution,

$$\Pi(G : W_2(G_0, G) \leq \sqrt{5}\epsilon) \geq \Gamma(\nu)[\epsilon^2(2D)^{-1}\operatorname{diam}(\Theta)^{-2}]^{D-1}\nu^D \prod_{i=1}^{D} P_0(S_i).$$

Here, $(S_1, \ldots, S_D)$ denotes the D disjoint $\epsilon/2$-balls that form a maximal packing of Θ. $\Gamma(\cdot)$ is the gamma function.

Summary

The question of posterior concentration of mixing measures is useful especially for clustering applications

Wasserstein metric provides a natural way to explore this question

- rates established for both finite and Dirichlet process mixtures
- minimax optimal rates?

For details, see:

- X. Nguyen, "Convergence of latent mixing measures in finite and infinite mixture models". Technical Report available at

www.stat.lsa.umich.edu/∼xuanlong