

Bayesian Inference for Multivariate Medians and Multivariate Quantiles

Subhashis Ghoshal,
North Carolina State University

Aad's 60th Birthday Conference
Leiden, June 17–21, 2019

Based on collaborations with my graduate student Indrabati Bhattacharya



Bayesian inference using a Dirichlet Process prior

- Samples $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P$, P probability distribution on a domain $\mathfrak{X}$.
- To make Bayesian inference on P or some characteristic of P .
- A convenient and popular prior on P is a Dirichlet process DP_α with base measure α : for every finite measurable partition $\{A_1, \dots, A_k\}$,

$$(P(A_1), \dots, P(A_k)) \sim \text{Dirichlet}_k(\alpha(A_1), \dots, \alpha(A_k)).$$

- Such a process exists and has many remarkable properties:
 - $\mathbb{E}P(A) = \alpha(A)/M$, $M = \alpha(\mathfrak{X})$, the total mass.
 - **Posterior conjugacy** $P|(X_1, \dots, X_n) \sim \text{DP}_{\alpha+n\mathbb{P}_n}$, where $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ is the empirical measure.
 - P is a.s. discrete: great for clustering, unsuitable for density estimation.
- Allows Bayesian inference on any k -dimensional functional $\psi(P)$.

Objectives

- Bayesian inference on **multivariate median and quantiles**.
- To establish asymptotic frequentist coverage of Bayesian credible sets.

Univariate situation

- Analytical formula for the posterior distribution of the univariate u th quantile $Q(u)$, $0 < u < 1$: If F_n stands for posterior mean of the distribution function F of P , then for any $x \in \mathbb{R}$, $P(Q(u) \leq x)$ is

$$\int_u^1 \frac{\Gamma(N)}{\Gamma(NF_n(x))\Gamma(N\bar{F}_n(x))} v^{NF_n(x)-1} (1-v)^{N\bar{F}_n(x)-1} dv,$$

where $N = M + n$, $\bar{F}_n = 1 - F_n$.

- Analytical formula for the posterior distribution of the mean is a lot more fascinating and non-trivial.

The multivariate ℓ_1 -median

- For $p \geq 1$, the population ℓ_1 -median of P (with respect to the ℓ_p -norm $\|\cdot\|_p$ on $\mathbb{R}^k$) is given by

$$\theta(P) = \theta_p(P) = \arg \min_{\theta} P(\|X - \theta\|_p - \|X\|_p).$$

- Usually $p = 2$, called the spatial median.
- For a sample $X_1, \dots, X_n \in \mathbb{R}^k$, the sample ℓ_1 -median is

$$\hat{\theta}_n = \hat{\theta}_{n,p} = \arg \min_{\theta} \mathbb{P}_n \|X - \theta\|_p.$$

- Clearly, the sample ℓ_1 -median is an M-estimator.
- If $p = 1$, everything reduces to component-wise univariate medians, and hence is excluded from our study, i.e. $1 < p < \infty$, and is also fixed throughout and is dropped from the notation.

Multivariate quantiles

- More generally, for any $u \in B_q := \{u \in \mathbb{R}^k : \|u\|_q < 1\}$, where q is the conjugate index of p , the population u th multivariate quantile of P is given by

$$Q(u; P) = \arg \min_{\theta} P[\Phi_p(u; X - \theta) - \Phi_p(u; X)],$$

where $\Phi_p(u; t) = \|t\|_p + \langle u, t \rangle$.

- The multivariate quantile process is

$$\hat{Q}_n(u) = \arg \min_{\theta} \mathbb{P}_n \Phi_p(u; X - \theta).$$

- The case $u = 0$ corresponds to the multivariate median.

Bayesian inference on multivariate median and quantiles

- Multivariate median $\theta(P)$ and each multivariate quantile $Q(u; P)$ is a functional of P , and hence gets its induced posterior distribution from the posterior distribution $\text{DP}_{\alpha+n\mathbb{P}_n}$ of P .
- Unlike univariate quantiles and medians, there are no explicit formulas for the posterior (even the prior) distribution.
- These can be obtained by simulation from a Dirichlet process and evaluating these functionals through efficient optimization algorithm.
- We wish to study posterior contraction and frequentist coverage properties of resulting Bayesian credible sets — dual uncertainty quantification.
- This is achieved by a Bernstein-von Mises theorem.

Computation of induced posterior distribution

To sample P from $DP_{\alpha+n\mathbb{P}_n}$:

- Generate $U \sim \text{Be}(M, n)$
- (infinite dimensional component) generate $P_I \sim DP_\alpha$ as follows — for some N , get $V_1, \dots, V_{N-1} \sim \text{Be}(1, M)$ independently, $V_N = 1$, $\eta_1, \dots, \eta_N \sim G := \alpha/M$ independently, $W_i = \prod_{j=1}^{i-1} (1 - V_j) V_i$, $i = 1, \dots, N$, $P = \sum_{i=1}^N W_i \delta_{\eta_i}$ — N is the first time $\prod_{j=1}^{N-1} (1 - V_j) < \epsilon$, predetermined;
- (finite dimensional component) else let $E_1, \dots, E_n \sim \text{Exp}(M)$ independently, $P_F = \frac{\sum_{i=1}^n E_i \delta_{x_i}}{\sum_{i=1}^n E_i}$.
- Obtain $P = UP_I + (1 - U)P_F$.
- Compute all multivariate quantiles of interest for the generated probability measure P .
- Repeat many many times, obtain histograms, density estimates etc. to represent the induced posterior distributions.

The Bayesian bootstrap

- The non-informative limit of $\text{DP}_{\alpha+n\mathbb{P}_n}$ as $M = \alpha(\mathfrak{X}) \rightarrow 0$ is called the Bayesian bootstrap
- It is an exchangeable bootstrap $\mathbb{B}_n = \sum_{i=1}^n W_{in} \delta_{X_i}$, where $(W_{1n}, \dots, W_{nn}) \sim \text{Dirichlet}_n(n; 1, \dots, 1)$, the distribution uniform over the unit n -simplex.
- The most useful property for us is $\|P - \mathbb{B}_n\|_{TV} = O_p(n^{-1})$, where $P \sim \text{DP}_{\alpha+n\mathbb{P}_n}$, allowing importing results from the bootstrap theory.

Basics of Bernstein-von Mises theorems

- Let θ be a parameter with the true value θ_0 in some model based on n i.i.d. observations $X_1, \dots, X_n$.
- Let $\hat{\theta}_n$ be a “well-respected” (read, efficient) estimator.
- Let Π be a prior distribution and $\Pi(\cdot|X_1, \dots, X_n)$.
- If we are in the $\sqrt{n}$ -regime and asymptotic normality domain (we are in this problem), then the Bernstein-von Mises theorem says that conditionally on the data, $\sqrt{n}(\theta - \hat{\theta}_n)$ is distributed asymptotically $N(0, \Sigma)$, where Σ is the asymptotic dispersion matrix in the limiting distribution of $\hat{\theta}_n$, i.e. $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, \Sigma)$.
- The main implication is that the Bayesian and frequentist distributions of $\sqrt{n}(\theta - \hat{\theta}_n)$ are approximately the same, so are the Bayesian and frequentist uncertainty quantifications, meaning that Bayesian credible sets will have asymptotic frequentist validity.

Sample multivariate median: The Z-estimator view

- The estimator $\hat{\theta}_n$ is viewed as a Z-estimator, that is, it satisfies a system of equations $\Psi_n(\theta) = \mathbb{P}_n \psi(\cdot, \theta) = 0$.

$$\psi_j(x, \theta) = \frac{|x_j - \theta_j|^{p-1}}{\|x - \theta\|_p^{p-1}} \text{sign}(\theta_j - x_j), \quad j = 1, \dots, k.$$

- The higher dimension is helpful here: for $k \geq 2$, the criterion function in the original M-estimator form is differentiable, while if one dimension, that is not true.

- The true value θ_0 satisfies the system of equations $\Psi_0(\theta) = P_0\psi(\cdot, \theta) = 0$.
- The population multivariate median is unique unless P_0 is completely supported on a straight line (cf. Kemperman).

- Notations:

$$\dot{\Psi}_0 = [\partial \Psi_0 / \partial \theta]_{\theta=\theta_0} \quad \Sigma_0 = P_0 \psi(X, \theta_0) \psi^T(X, \theta_0).$$

- Hence

$$\dot{\Psi}_0 = P_0 \left(\frac{p-1}{\|X - \theta_0\|_p} \left[\text{diag} \left(\frac{|X_1 - \theta_{01}|^{p-2}}{\|X - \theta_0\|_p^{p-2}}, \dots, \frac{|X_k - \theta_{0k}|^{p-2}}{\|X - \theta_0\|_p^{p-2}} \right) - \frac{YY^T}{\|X - \theta_0\|_p^{2(p-1)}} \right] \right).$$

with Y given by

$$Y = [|X_1 - \theta_{01}|^{p-1} \text{sign}(X_1 - \theta_{01}), \dots, |X_k - \theta_{0k}|^{p-1} \text{sign}(X_k - \theta_{0k})]^T.$$

BVM Theorem for multivariate ℓ_1 -median

Theorem

Let $p \geq 2$ be a fixed integer. For $k \geq 2$, suppose that the following conditions hold.

- C1 The true distribution of $X \in \mathbb{R}^k$, P_0 has a probability density that is bounded on compact subsets of $\mathbb{R}^k$.
- C2 The ℓ_1 -median of P_0 is unique.

Then

- (i) $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N_k(0, \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1})$,
 - (ii) given the data, $\sqrt{n}(\theta - \hat{\theta}_n) \rightsquigarrow N_k(0, \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1})$ in P_0 -probability.
- Further if $k = 2$, the assertions hold for any $1 < p < \infty$.

Brief history of Z-estimator

- Huber (1967, Berkeley Symp) proved the asymptotic normality of Z-estimators for finite-dimensional parameter space.
- Van der Vaart (1995, Stat Neerlandica) extended Huber's theorem to the infinite dimensional case.
- Wellner and Zhan (1996, U. Wash. Tech Report) extended van der Vaart's theorem to bootstrap.

Bootstrapping a Z-estimator

Assumptions on the weights:

- The bootstrap empirical measure is defined as $\hat{\mathbb{P}}_n = n^{-1} \sum_{i=1}^n W_{ni} \delta_{X_i}$.
- Let $W_n = (W_{n1}, W_{n2}, \dots, W_{nn})$ be a triangular array of bootstrap weights (on some probability space) satisfying the following conditions:
 - (i) The vectors $W_n = (W_{n1}, W_{n2}, \dots, W_{nn})^T$ are invariant under permutation for every n .
 - (ii) The weights $W_{ni} \geq 0$ for every n, i and $\sum_{i=1}^n W_{ni} = n$ for all n .
 - (iii) The $L_{2,1}$ norm $\|W_{n1}\|_{2,1} = \int_0^\infty \sqrt{P(W_{n1} \geq u)} du$ of W_{n1} is uniformly bounded.
 - (iv) $\lim_{\lambda \rightarrow \infty} \limsup_{n \rightarrow \infty} \sup_{t \geq \lambda} (t^2 P\{W_{n1} \geq t\}) = 0$.
 - (v) $n^{-1} \sum_{i=1}^n (W_{ni} - 1)^2 \rightarrow c^2 > 0$ in $\hat{P}$ -probability for some constant $c > 0$.
- The Bayesian bootstrap weights $W_{ni} = V_i / \bar{V}_n$, $i = 1, \dots, n$, $V_1, \dots, V_n$ i.i.d. exponential, satisfy the conditions with $c = 1$.

Assumptions on the function class:

- The class $\mathcal{F}_R = \{\psi_j(\cdot, \theta) : \|\theta - \theta_0\| \leq R, j = 1, 2, \dots, k\}$ has *enough measurability* to allow for randomization with independently and identically distributed multipliers to be possible.
- Typical conditions are $\mathcal{F}$ is countable, or if the empirical processes $\mathbb{G}_n$ is stochastically separable, or $\mathcal{F}$ is image admissible Suslin.

Notation: Let $\hat{\theta}_n$ be the corresponding bootstrap estimator: i.e., satisfies the system of equations $\hat{\mathbb{P}}_n \psi(\cdot, \theta) = 0$ (or more generally, $o_p(n^{-1/2})$).

Theorem (Wellner and Zhan)

Assume that

❶ $\Psi_0(\theta) = P_0\psi(X, \theta)$ is differentiable at θ_0 , and $\dot{\Psi}_0 = \left[\frac{\partial \Psi_0}{\partial \theta} \right]_{\theta=\theta_0}$ is nonsingular.

❷ $\sup \left\{ \frac{\|\mathbb{G}_n(\psi(\cdot, \theta) - \psi(\cdot, \theta_0))\|}{1 + \sqrt{n}\|\theta - \theta_0\|} : \|\theta - \theta_0\| \leq \delta_n \right\} \rightarrow_p 0$ if $\delta_n \rightarrow 0$.

❸ $\Sigma_0 = P_0\psi(X, \theta_0)\psi^T(X, \theta_0) < \infty$.

❹ For any $\delta_n \rightarrow 0$, $\lim_{\lambda \rightarrow \infty} \limsup_{n \rightarrow \infty} \sup_{t \geq \lambda} t^2 P_0(D_n(X_1) > t) = 0$, where

$$D_n(x) = \sup \left\{ \frac{|\psi_j(x, \theta) - \psi_j(x, \theta_0)|}{1 + \sqrt{n}\|\theta - \theta_0\|} : \|\theta - \theta_0\| \leq \delta_n, j = 1, 2, \dots, k \right\}.$$

❺ $\|\hat{\theta}_n - \theta_0\| \xrightarrow{P_0} 0$ and $\|\hat{\hat{\theta}}_n - \hat{\theta}_n\| \xrightarrow{P} 0$ in P_0 -probability.

Then

(i) $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N_k(0, \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1})$;

(ii) $\sqrt{n}(\hat{\hat{\theta}}_n - \hat{\theta}_n) \rightsquigarrow N_k(0, c^2 \dot{\Psi}_0^{-1} \Sigma_0 \dot{\Psi}_0^{-1})$ in P_0 -probability.

Main challenge in the proof

To show that $\mathcal{F}_R$ is P_0 -Donsker. Let

$$\mathcal{F}_{1R} = \left\{ |x_j - \theta_j|^{p-1} / \|\mathbf{x} - \boldsymbol{\theta}\|_p^{p-1} : j = 1, 2, \dots, k, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq R \right\},$$
$$\mathcal{F}_{2R} = \left\{ \text{sign}(\theta_j - x_j) : j = 1, 2, \dots, k, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq R \right\}.$$

- To show that $\mathcal{F}_{1R}$ and $\mathcal{F}_{2R}$ satisfy the uniform entropy condition.
- $\mathcal{F}_{2R}$ is easy.
- $\mathcal{F}_{1R}$ is a bounded VC-major class of functions, that is, the sets $\{x : f(x) > t\}$, $f \in \mathcal{F}_{1R}$, $t \in \mathbb{R}$, form a VC class of sets.
- Boils down to verifying VC property of $S_{\boldsymbol{\theta}, t}^1 = \left\{ x : |x_1 - \theta_1|^p > \frac{t^{p/(p-1)}}{1 - t^{p/(p-1)}} \sum_{j=2}^k |x_j - \theta_j|^p \right\}$.
- Split into pieces for each possible sign of the absolute value, and reduce to a collection spanned by finitely many functions.
- Important to have p an integer in the analysis, unless $k = 2$.

One more point

- The posterior Dirichlet process multivariate median is not the multivariate median for the Bayesian bootstrap distribution.
- But we can treat it through the Bayesian bootstrap result since it is an approximate multivariate median, within $O_p(n^{-1}) = o_p(n^{-1/2})$ error.
- If a prior input is not essential, then one can work with the noninformative limit $M \rightarrow 0$, i.e., the Bayesian bootstrap, which is (much) simpler for computation, since the infinite dimensional component disappears completely.

Affine Equivariant ℓ_1 -median

- The multivariate ℓ_1 -median lacks affine equivariance.
- The basic idea [Chakraborty, Chaudhuri and Oja, 1998, Statistica Sinica] is to use a part of the data to create a co-ordinate system, so that each linear transformation affects the data and the co-ordinate system in the same way.
- Compute the multivariate median in the transformed co-ordinate system, and finally transform back.
- Mathematically, with $S = \{i_0, i_1, \dots, i_k\} \subset \{1, 2, \dots, n\}$, define $Z_j^{(S)} = \{X(i) - X(i_0) : i = i_1, \dots, i_k\}^{-1} X_j$, $j \notin S$. Compute $\hat{\phi}_n^{(S)} = \arg \min_{\phi} \sum_{j \notin S} \|Z_j^{(S)} - \phi\|$, and transform back to $\hat{\theta}_n^{(S)} = X(S) \hat{\phi}_n^{(S)}$.

Bayesian Affine Equivariant ℓ_1 -median

- As before, we use a part of the data to create a co-ordinate system, so that each linear transformation affects the data and the co-ordinate system in the same way.
- Put a Dirichlet process prior on the transformed P_Z — effectively a data dependent prior.
- By a basic property of the Dirichlet, the original $P = P_X$ also gets a Dirichlet process prior, with data dependent parameters.
- The resulting multivariate median is affine-equivariant.
- We also have a Bernstein-von Mises theorem for this affine equivariant median.

Theorem

Given a subset $S = \{i_0, i_1, \dots, i_k\}$ of $\{1, 2, \dots, n\}$ with size $k + 1$, suppose that the following conditions hold.

- C1 The true distribution of $Z^{(S)}$, P_{Z_0} has a probability density that is bounded on compact subsets of $\mathbb{R}^k$.
- C2 The ℓ_1 -median of P_{Z_0} is unique.

Then

- (i) $\sqrt{n}(\hat{\theta}_n^{(S)} - \theta_0^{(S)})| \{X_i : i \in S\} \rightsquigarrow N_k(0, X(S)\{\Psi_{Z_0}^{(S)}\}^{-1}\Sigma_{Z_0}^{(S)}\{\Psi_{Z_0}^{(S)}\}^{-1}\{X(S)\}^T)$;
- (ii) given the data,
 $\sqrt{n}(\theta^{(S)} - \hat{\theta}_n^{(S)}) \rightsquigarrow N_k(0, X(S)\{\Psi_{Z_0}^{(S)}\}^{-1}\Sigma_{Z_0}^{(S)}\{\Psi_{Z_0}^{(S)}\}^{-1}\{X(S)\}^T)$ in P_0 -probability.

BVM Theorem for multivariate quantiles

- For each u , the sample u -quantile is viewed as a Z -estimator solving $\Psi_n^{(u)}(\xi) = \mathbb{P}_n \psi^{(u)}(\cdot, \xi) = 0$.
- $\Psi^{(u)}(\xi) = P \psi^{(u)}(\cdot, \xi)$.
- The true value of $Q_0(u) \equiv Q_{P_0}(u)$ solves $\Psi_0^{(u)}(\cdot, \xi) = P_0 \psi^{(u)}(\cdot, \xi) = 0$.
- To study the joint asymptotic posterior distribution of $Q(u_1), \dots, Q(u_m)$.
- The Bayesian bootstrapped analogs of $\hat{Q}_n(u_1), \dots, \hat{Q}_n(u_m)$ are given by $Q_{\mathbb{B}_n}(u_1), \dots, Q_{\mathbb{B}_n}(u_m)$ respectively.

- The real-valued components of $\psi^{(u)}(\cdot, \xi)$ are then given by

$$\psi_j^{(u)}(x, \xi) = \frac{|x_j - \xi_j|^{p-1}}{\|x - \xi\|_p^{p-1}} \text{sign}(\xi_j - x_j) + u_j, \quad j = 1, \dots, k.$$

Define $\dot{\Psi}_0^{(u)} = \int \dot{\psi}_{x,0}^{(u)} dP_0$, where

$$\dot{\psi}_{x,0}^{(u)} = \left[\frac{\partial \psi^{(u)}(x, \xi)}{\partial \xi} \right]_{\xi=Q_0(u)}.$$

The matrix $\dot{\psi}_{x,0}^{(u)}$ is given by

$$\frac{p-1}{\|x - Q_0(u)\|_p} \left[\text{diag} \left(\frac{|x_1 - Q_{01}(u)|^{p-2}}{\|x - Q_0(u)\|_p^{p-2}}, \dots, \frac{|x_k - Q_{0k}(u)|^{p-2}}{\|x - Q_0(u)\|_p^{p-2}} \right) - \frac{YY^T}{\|x - Q_0(u)\|_p^{2(p-1)}} \right],$$

Let

- $Y = [|x_j - Q_{0j}(u)|^{p-1} \text{sign}(x_j - Q_{0j}(u)) : j = 1, \dots, k]^T$;
- $\Sigma_{0;u,v} = P_0 \psi^{(u)}(x, Q_0(u)) \{ \psi^{(v)}(x, Q_0(v)) \}^T$.

Let $p \geq 2$ be a fixed integer. For $k \geq 2$, suppose that the following conditions hold.

- C1 The true distribution of X , P_0 has a probability density that is bounded on compact subsets of $\mathbb{R}^k$.
- C2 For every $u_1, \dots, u_m \in B_q^{(k)}$, the $u_1, \dots, u_m$ -quantiles of P_0 are unique.

Theorem

The following assertions hold:

- (i) *the joint distribution of $(\sqrt{n}(\hat{Q}_n(u_1) - Q_0(u_1)), \dots, \sqrt{n}(\hat{Q}_n(u_m) - Q_0(u_m)))$ converges to a km -dimensional normal distribution with mean zero, and the (j, l) th block of the covariance matrix is given by $\{\dot{\Psi}_0^{(u_j)}\}^{-1} \Sigma_{0; u_j, u_l} \{\dot{\Psi}_0^{(u_l)}\}^{-1}$, $1 \leq j, l \leq m$,*
- (ii) *given $X_1, \dots, X_n$, the joint distribution of $\{\sqrt{n}(Q(u_1) - \hat{Q}_n(u_1)), \dots, \sqrt{n}(Q(u_m) - \hat{Q}_n(u_m))\}$ converges to km -dimensional normal distribution with mean zero, and the (j, l) th block of the covariance matrix is given by $\{\dot{\Psi}_0^{(u_j)}\}^{-1} \Sigma_{0; u_j, u_l} \{\dot{\Psi}_0^{(u_l)}\}^{-1}$, $1 \leq j, l \leq m$.*

Further if $k = 2$, the assertions hold for any $1 < p < \infty$.

Simulation study

Table: Size and coverage of 95% Bayesian credible ellipsoids of ℓ_1 -median with $k = 2$ (non-affine equivariant)

Radii of credible ellipsoids				
Data		$n = 100$	$n = 1000$	$n = 10000$
Normal	$p = 2$	5.941	5.920	5.901
		0.950	0.953	0.960
	$p = 3$	6.181	6.005	5.992
		0.942	0.954	0.956
Laplace	$p = 2$	6.113	5.995	5.889
		0.949	0.950	0.956
	$p = 3$	6.062	5.997	5.816
		0.938	0.946	0.951

Fisher's Iris data

The data consists sepal length, sepal width, petal length and petal width measured for each sample of three plant species setosa, virginica and versicolor.

Table: Principal axes of 95% credible ellipsoid of spatial median

1st axis	2nd axis	3rd axis	4th axis
0.0580	-0.3129	-0.6747	-0.6629
-0.1461	0.2193	-0.6143	-0.7437
-0.2965	0.8626	0.4089	-0.0252
0.9420	0.3252	-0.0081	-0.0824

Figure: 95% Credible ellipsoids for the species Setosa

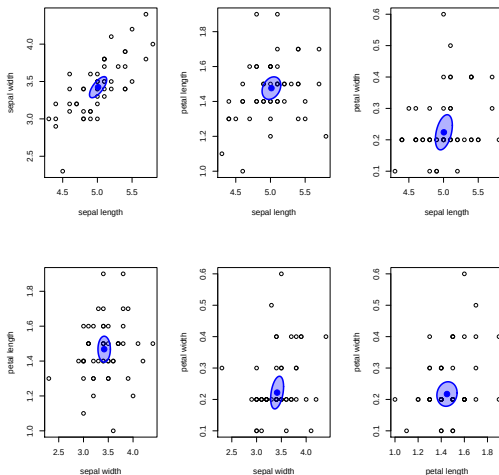


Figure: 95% Credible ellipsoids for the species *Virginica*

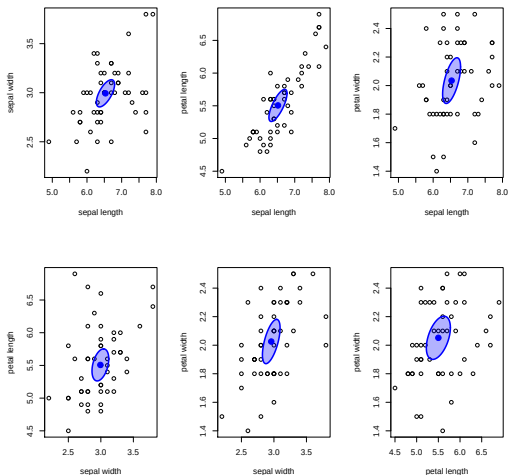
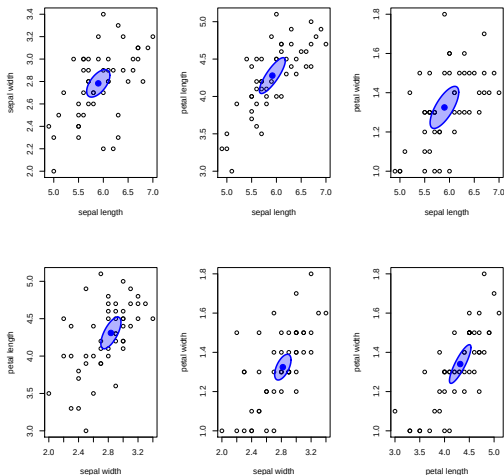


Figure: 95% Credible ellipsoids for the species Versicolor





Aad

Thank you