Riemannian Framework for Assessing Bayes Robustness

Karthik Bharath

School of Mathematical Sciences

University of Nottingham

Joint work with Sebastian Kurtek at The Ohio State University.

Motivation

Important to assess sensitivity of posterior inference to:

- Prior distribution;
- Likelihood;
- Data.

Motivation

Important to assess sensitivity of posterior inference to:

- Prior distribution;
- Likelihood;
- Data.

Two observations:

- 'Distance'-based measures are commonly used with inadequate considerations of the geometry of the space of models;
- Perturbations and sensitivity measures are usually developed independent of the inferential methodology.

Objective

Unify the perturbation mechanisms for prior, likelihood and data with inference under a Riemannian framework to develop sensitivity measures which are geometrically calibrated.

Why bother with the geometry?

- The space of probability densities is a nonlinear manifold.
- Divergence measures are not true distances(positive definiteness, symmetry and triangle inequality).
- Geodesic distances provide geometrically calibrated measures of disparity between densities. Under the Fisher-Rao metric they are also bounded.
- Geometry might lead to statistical insights.

Fisher–Rao metric

- Banach manifold of probability densities on $\mathbb{R}$:

$$\mathcal{P} = \left\{ p : \mathbb{R} \to \mathbb{R}^+ \cup \{0\} : \int_{\mathbb{R}} p(x) dx = 1 \right\}.$$

- For a point p in $\mathcal{P}$ define the tangent space as:

$$T_p(\mathcal{P}) = \left\{ \delta p : \mathbb{R} \to \mathbb{R} : \int_{\mathbb{R}} \delta p(x) p(x) dx = 0 \right\}.$$

- The nonparametric Fisher-Rao metric then is:

$$\langle\langle \delta p_1, \delta p_2 \rangle\rangle_p = \int_{\mathbb{R}} \delta p_1(x) \delta p_2(x) \frac{1}{p(x)} dx.$$

Fisher–Rao metric

- Banach manifold of probability densities on $\mathbb{R}$:

$$\mathcal{P} = \left\{ p : \mathbb{R} \to \mathbb{R}^{+} \cup \{0\} : \int_{\mathbb{R}} p(x)dx = 1 \right\}.$$

- For a point p in $\mathcal{P}$ define the tangent space as:

$$T_p(\mathcal{P}) = \left\{ \delta p : \mathbb{R} \to \mathbb{R} : \int_{\mathbb{R}} \delta p(x) p(x) dx = 0 \right\}.$$

- The nonparametric Fisher-Rao metric then is:

$$\langle\langle \delta p_1, \delta p_2 \rangle\rangle_p = \int_{\mathbb{R}} \delta p_1(x) \delta p_2(x) \frac{1}{p(x)} dx.$$

The metric is invariant to reparameterizations (Ćencov 1982).

Connection to Fisher Information matrix

- Consider the parametric family $\mathcal{F} = \{f(x,\theta) | \theta \in \Theta\}$.
- The tangent vectors at $f(x,\theta)$ are $\frac{\partial}{\partial\theta} f(x,\theta)$.
- Then, the norm on $\mathcal{F}$ is induced by the Fisher-Rao Riemannian metric

$$\int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta} f(x,\theta) \right)^2 \frac{1}{f(x,\theta)} dx = \int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta} log(f(x,\theta)) \right)^2 f(x,\theta) dx$$
$$= E_\theta \left[\frac{\partial}{\partial\theta} log(f(x,\theta)) \right]^2,$$

Fisher–Rao metric

Issue: difficult to use the metric directly as it changes from point to point on the manifold $\mathcal{P}$.

Fisher–Rao metric

Issue: difficult to use the metric directly as it changes from point to point on the manifold $\mathcal{P}$.

Solution: find a different representation, which simplifies the computations.

Different choices are available:

log representation; CDF representation; positive square-root.

The Square-Root Representation (SRT): Bhattacharya (1943)

- Define the map $\phi : \mathcal{P} \mapsto \Psi$ where the space Ψ is the space containing the positive square-root of all possible density functions.

- Using this mapping, define the square-root transform of probability density functions as $\phi(p) = \psi = +p^{1/2}$. The inverse mapping is simply $\phi^{-1}(\psi) = p = \psi^2$.

The Square-Root Representation (SRT): Bhattacharya (1943)

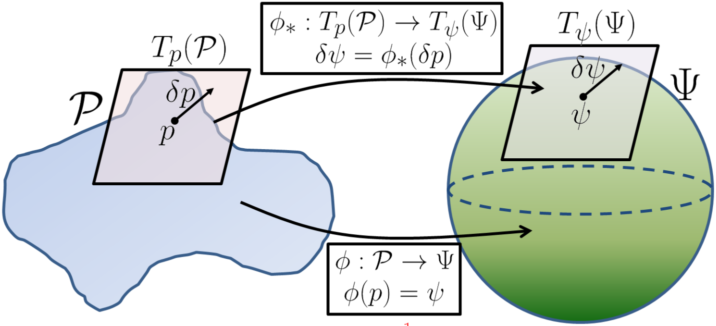
Fact 1: The space of all SRT representations of probability density functions is the positive orthant of the unit $\mathbb{L}^2$ sphere:

$$\Psi = \left\{\psi : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}; \int_{\mathbb{R}} |\psi(x)|^2 dx = 1\right\}.$$

Fact 2: Ψ is a Hilbert manifold with the unique global chart which is the identify map.

Fact 3: The nonparametric Fisher-Rao metric equips Ψ with a Riemannian structure and reduces to the standard $\mathbb{L}^2$ metric.

Fisher–Rao metric under SRT



$$d_{FR}(p_1, p_2) = \cos^{-1}(\langle \psi_1, \psi_2 \rangle)$$

Geometry of unit Hilbert sphere is well-known

- Tangent space at a point $\psi \in \Psi$: $T_\psi(\Psi) = \left\{ \delta\psi : \langle \delta\psi, \psi \rangle = 0 \right\}$.
- The $\mathbb{L}^2$ Riemannian metric is: $\langle \delta\psi_1, \delta\psi_2 \rangle = \int_{\mathbb{R}} \delta\psi_1(x)\delta\psi_2(x)dx$.
- Geodesic distance:

$$d_{FR}(p_1, p_2) = \theta = \cos^{-1}(\langle \psi_1, \psi_2 \rangle).$$

- $0 \leq d_{FR}(p_1, p_2) \leq \frac{\pi}{2}$.
- The geodesic path between ψ_1 and ψ_2, indexed by $\tau \in [0, 1]$, is $\eta(\tau) = (\sin(\theta))^{-1}[\sin(\theta - \tau\theta)\psi_1 + \sin(\tau\theta)\psi_2]$

Geometry of unit Hilbert sphere is well-known

- Exponential map $\exp : T_{\psi_1}(\Psi) \mapsto \Psi$, to map tangent vectors back to sphere:

$$\exp_{\psi_1}(\delta\psi) = \cos(\|\delta\psi\|)\psi_1 + \sin(\|\delta\psi\|)\delta\psi(\|\delta\psi\|)^{-1}.$$

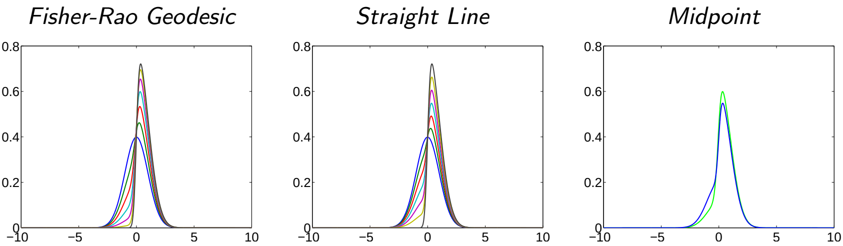
- Inverse Exponential map $\exp_{\psi_1}^{-1} : \Psi \mapsto T_{\psi_1}(\Psi)$:

$$\exp_{\psi_1}^{-1}(\psi_2) = [\theta(\sin(\theta))^{-1}(\psi_2 - \cos(\theta)\psi_1)].$$

These are tremendously useful!

Example: Normal and Skew-normal

$p_1 \sim N(0,1); \quad p_2 SN(5).$



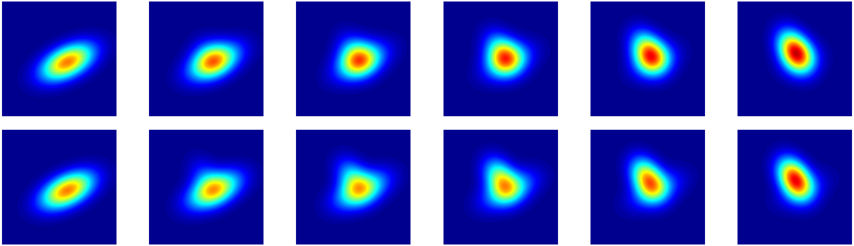
$$d_{FR}(p_1, p_2) = 0.6700;\ KL(p_1, p_2) = 6.6692;\ KL(p_2, p_1) = 0.5520$$

- Linear interpolation midpoint is shown in blue.
- Fisher-Rao geodesic midpoint is shown in green.

Example: Bivariate normals

$p_1 \sim N(\mu_1, \Sigma_1)$ and $p_2 \sim N(\mu_2, \Sigma_2)$ where

$$\mu_1 = \begin{bmatrix} .5 \\ .2 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1.2 & .4 \\ .4 & .6 \end{bmatrix} \text{ and } \mu_2 = \begin{bmatrix} 0 \\ .5 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} .5 & -.2 \\ -.2 & .7 \end{bmatrix}$$



$$d_{FR}(p_1, p_2) = 0.7157;\ KL(p_1, p_2) = 1.2522;\ KL(p_2, p_1) = 1.3653.$$

GEOMETRIC ϵ- PERTURBATION CLASS

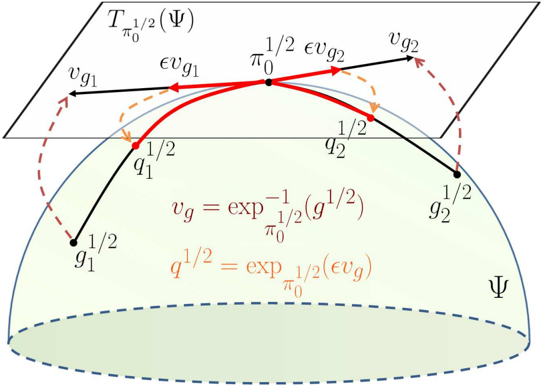
- Let $\mathcal{G} = \{g_1, \ldots, g_n\}$ denote a finite class of contaminants.
- We construct a set of tangent vectors $v_{g_1}, \ldots, v_{g_n} \in T_{\pi_0^{1/2}}(\Psi)$ using the inverse exponential map as $v_{g_i} = \exp^{-1}_{\pi_0^{1/2}}(g_i^{1/2}),\ i = 1, \ldots, n.$

Definition

For a class of densities $\mathcal{G} = \{g_1, \ldots, g_n\}$, the geometric ϵ-contamination class corresponding to the baseline prior π_0 is defined as

$$\Gamma = \left\{[\exp_{\pi_0^{1/2}}(\epsilon v_{g_i})]^2; 0 \leq \epsilon \leq 1,\ g_i \in \mathcal{G}, i = 1, \ldots, n\right\}. \tag{1}$$

Geometric ϵ-perturbation of prior



Two fundamental properties

Theorem

- *Any perturbation of the baseline prior should not have an effect on the sampling distribution.*
- *Given two perturbations of the baseline prior, the Riemannian metric on the space of joint densities show be independent of the sampling distribution.*

Theorem

- *The effects of simultaneous perturbations of the prior and likelihood on the joint density should be separable.*
- *Two separate perturbations of the prior and likelihood should be orthogonal to each other on the space of joint densities.*

GLOBAL SENSITIVITY MEASURES

Geodesic distance as sensitivity measure

- We assess global sensitivity to perturbations of the prior or likelihood using the Fisher-Rao geodesic distance between the baseline posterior and the perturbed posterior
- Upper bound of $\pi/2$ privdes a natural scale.
- Intrinsic distance captures the geometry of the space of densities.
- One can additionally assess sensitivity of functionalsof the posterior by computing them at the nearest and farthest perturbed posteriors.

Simple example

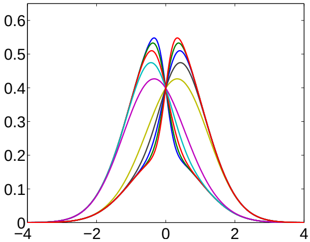
- Data: 50 data points simulated from the baseline model.

- Baseline model:

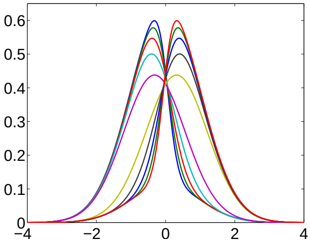
$$x_i|\theta \stackrel{i.i.d}{f} =\sim N(\theta, 1);$$

$$\theta \sim \pi_o = N(0, 1).$$

- Prior perturbation class: $SN(\alpha), -5 \leq \alpha \leq 5$.

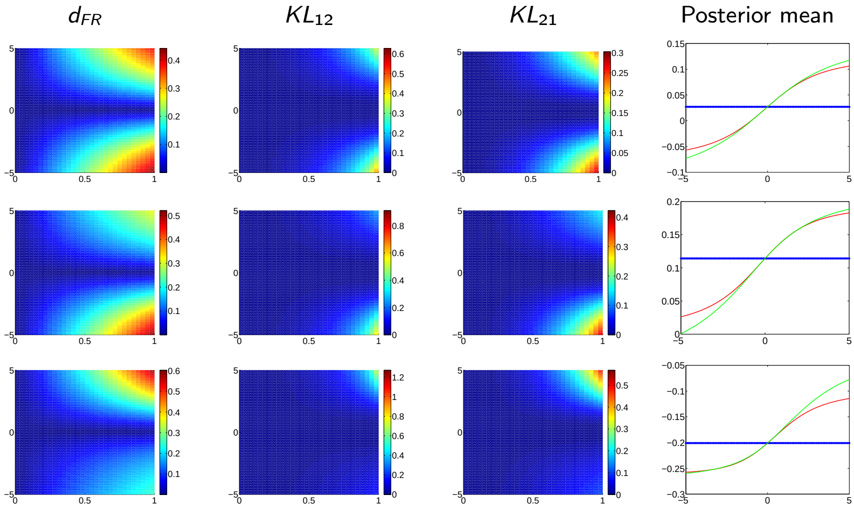


(a)



(b)

Simple example



Last columns is posterior mean for varying values of α and $\epsilon = 0.5$

(baseline=blue, geometric contamination=green, linear contamination=red).

Directional Data example

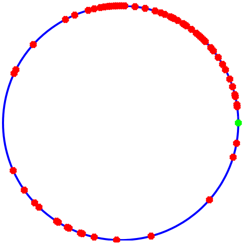
- Data: 76 directions of turtle movement after applying a treatment.
- Baseline model:

$$x_i|\theta \overset{i.i.d}{\sim} f = vM(\theta, \hat{\kappa}), \quad \hat{\kappa} = 1.14;$$

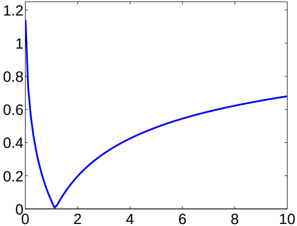
$$\theta \sim \pi_o = vM(0, 0.01).$$

Goal: assess global sensitivity to changing $\hat{\kappa}$. We will do this by varying the concentration parameter in the likelihood from 0.01 to 10.

Turtle Data



d_{FR}



Example: Generalized Mixed Effects Model

- Data: Binary response–presence or absence of bacteria; predictors–treatement (placebo, drug, drug+), week of test.
- Baseline model:

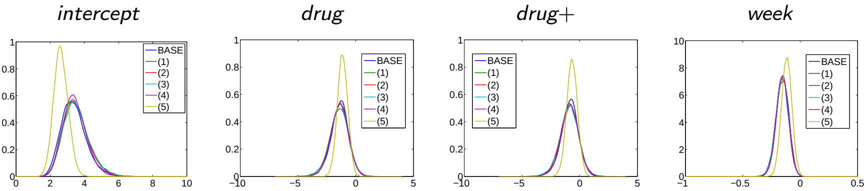
$$Y_{ij} \sim Bernoulli(p_{ij}); \quad logit(p_{ij}) = \mu + \sum_{k=1}^{3} x_{ij}^k \beta^k + V_i;$$

$$\mu \sim N(0, 100); \qquad \beta^k \stackrel{\text{i.i.d.}}{\sim} N(0, 100);$$

$$V_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2); \qquad \tau = \frac{1}{\sigma^2} \sim \Gamma(0.01, 0.01),$$

- Goal: assess global sensitivity of marginal posteriors of μ and β to the following choices of priors on σ^2:
  - Half-normal with variance 100 on σ;
  - Half-Cauchy with scale 100 on σ;
  - Uniform(0,100) on σ;
  - $\Gamma(1, 2)$ on τ;
  - $\Gamma(1, 2)$ on τ.

Example: Generalized Mixed Effects Model



Fixed Effect	*Model*				
	(1)	(2)	(3)	(4)	(5)
intercept	0.1054	0.0864	0.0982	0.0740	0.6716
drug	0.0716	0.0499	0.0590	0.0435	0.3835
drug+	0.0666	0.0580	0.0683	0.0445	0.3432
week	0.0524	0.0572	0.0630	0.0311	0.3670

LOCAL SENSITIVITY MEASURES

Local perturbation measures based on ϵ-perturbation

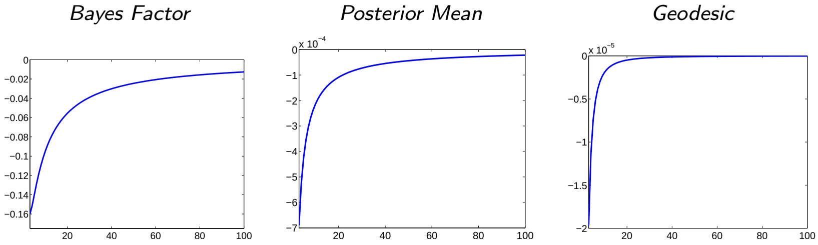
- Use directional derivatives to derive local sensitivity measures under the geometric ϵ-perturbation class.
- Utilize the underlying geometry of the space to develop sensitivity measures for posterior functionals.
- Second-order analysis on the geodesic distance itself can be used obtain finer measures.

Toy example

- Data: 50 data points simulated from the baseline model.
- Baseline model:

$$x_i|\theta \overset{i.i.d}{f} =\sim N(\theta,1); \theta \sim \pi_o = N(0,1).$$

- Prior perturbation class: $t_\nu, \nu = 3, 4, \ldots, 100$.
- Candidate prior for Bayes factor: $\pi_1 = N(0,5)$.

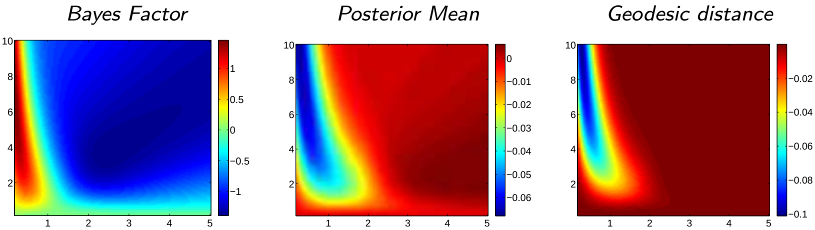


Local analysis for Turtle data

- Data: 76 directions of turtle movement after applying a treatment.
- Baseline model:

$$x_i|\theta \overset{i.i.d}{\sim} f = vM(\theta, \hat{\kappa}), \hat{\kappa} = 1.14; \quad \theta \sim \pi_0 = vM(0, 0.01).$$

- Prior perturbations: wrapped Laplace with skewness parameter $0.2 \leq \eta \leq 5$, and concentration parameter $0.2 \leq \lambda \leq 10$.
- $\eta < 1$ is skewed anti-clockwise; and, $\eta > 1$ is skewed clockwise; $\eta = 1$ is symmetric.



Detecting Influential Observations

Influential Observations

- If p_0 is the baseline posterior obtained with all observations, denote p_k to be posterior obtained having deleted the *kth* observation.

- The influence measure for the *kth* observation then is

$$I(k) = d_{FR}(p_0, p_k).$$

Influential Observations

- If p_0 is the baseline posterior obtained with all observations, denote p_k to be posterior obtained having deleted the *kth* observation.
- The influence measure for the *kth* observation then is

$$I(k) = d_{FR}(p_0, p_k).$$

- Issue: posterior may not be available in closed form and numerical computation of the marginal likelihood may not be possible.
- Solution: Estimate distance using Monte-Carlo based on MCMC samples or importance sampling. This estimate is consistent.

Example: Linear regression

- *Data*: response–natural log of survival time; predictors–blood clotting score, prognostic index, enzyme test, liver test, age, gender (binary), moderate alcohol use (binary), heavy alcohol use (binary); n=54.

- *Baseline model*:

$$y|\theta, X \sim f = N(X\theta, \sigma^2 \mathbf{I}_{54});$$

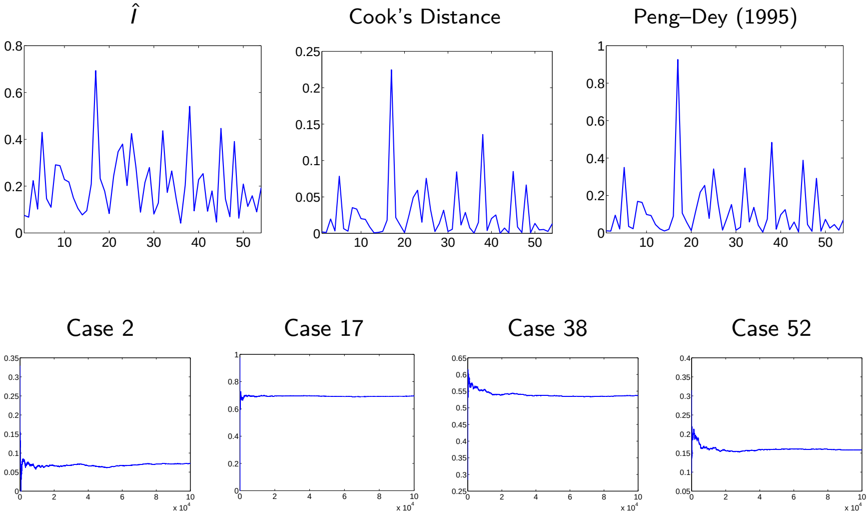
$$\theta \sim \pi = N(\mathbf{0}, 1000\mathbf{I}_9).$$

- Easy to evaluate baseline and case-deletion posterior. But $d_F R$ is a high-dimensional intergral.

- If $\{\theta_i\}$ is a sample from the baseline posterior, then

$$\hat{I}(k) = \hat{d}_{FR}(p_0, p_k) = \cos^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{p_k(\theta_i|y, X)}{p_0(\theta_i|y, X)}}\right].$$

Example: Linear regression



Points to ponder over

- There appears to be value in incorporating geometric information into the robustness assessment.

Points to ponder over

- There appears to be value in incorporating geometric information into the robustness assessment.
- Several measures in literature might perform better and might even e easy to implement. But, the objective of our work is to unify the robustness assessment endeavour under a geometric framework.

Points to ponder over

- There appears to be value in incorporating geometric information into the robustness assessment.
- Several measures in literature might perform better and might even e easy to implement. But, the objective of our work is to unify the robustness assessment endeavour under a geometric framework.
- Our claim is that the measures provided are 'geometrically calibrated' with a natural scale.

Points to ponder over

- There appears to be value in incorporating geometric information into the robustness assessment.
- Several measures in literature might perform better and might even e easy to implement. But, the objective of our work is to unify the robustness assessment endeavour under a geometric framework.
- Our claim is that the measures provided are ‘geometrically calibrated’ with a natural scale.
- The framework is really easy to implement in practice!

Future and Current work (plenty!)

- Develop good estimators for FR distance when posteriors are unavailable analytically.
- Geometric Variational Bayes—we have some preliminary results which appear promising. The nonparametric manifold should make a seamless transition to the nonparametric Bayesian framework. (Nonparamteric Invariant prior mimicing the Jeffreys prior).
- Investigate posterior consistency in topological neighbourhoods induced by the FR metric.
- ......

If you can't convince them, confuse them.

–*Harry Truman*