2015 Workshop on High-Dimensional Statistical Analysis
DEC.11 (Friday) ~ DEC. 15 (Tuesday)
Humanities and Social Sciences Center, Academia Sinica, Taiwan

Information Geometry
and
Spontaneous Data Learning

Shinto Eguchi, ISM, Japan

江口真透, 統数研

Joint work with Osamu Komori and Akifumi Notsu

小森　理, 野津　昭文

Spontaneous learning?

We focus on a collapse of consistency for estimation

Model assumes unimodality;

the true distribution may be multimodal.

There is a wide class of consistent estimators

Select the optimal estimator under a fixed model

to get the information for the true distribution

Cf. Model selection

Power divergence

β-cross entropy $$C_\beta(g,f) = -\frac{1}{\beta}\int f^\beta g \mathrm{d}\mu + \frac{1}{\beta+1}\int f^{\beta+1}\mathrm{d}\mu$$

β-diagonal entropy $$H_\beta(f) = C_\beta(f,f)$$

β-divergence $$D_\beta(g,f) = C_\beta(g,f) - H_\beta(g)$$

Hill (1971), Tsallis (1988) Basu et al (1998) Minami-Eguchi (2002)

$$U(s) = \frac{1}{1+\beta}(1+\beta s)^{\frac{1+\beta}{\beta}},\ \phi(t) = \frac{t^\beta - 1}{\beta},\ C_U(f,g) = \int \{-\phi(g)f + U(\phi(g))\}\mathrm{d}\mu$$
$$= C_\beta(f,g)$$

Projective power divergence

γ-cross entropy $$C_\gamma(g,f) = -\frac{1}{\gamma(\gamma+1)}\int \left\{\frac{f(x)}{\| f \|}\right\}^\gamma g(x)\mathrm{d}\mu(x)$$

$$\| f \| = \left\{\int f(x)^{\gamma+1}\mathrm{d}\mu(x)\right\}^{\frac{1}{1+\gamma}} \quad (\text{Lebesgue } \mathrm{L}_p \text{ norm}, p = \gamma+1)$$

γ-diagonal entropy $$H_\gamma(f) = C_\gamma(f,f)$$

γ-divergence $$D_\gamma(g,f) = C_\gamma(g,f) - H_\gamma(g)$$

Fujisawa-Eguchi (2008), Eguchi-Kato (2010)

Three properties

$$C_\gamma(g,f) = -\int \left(\frac{f}{\|f\|}\right)^\gamma g \, \mathrm{d}\mu$$

1. linearity $\quad C_\gamma(\alpha g+\beta h, f) = \alpha C_\gamma(g,f) + \beta C_\gamma(h,f)$

2. scale invariance $\quad C_\gamma(g,\lambda f) = C_\gamma(g,f) \quad (\forall \lambda > 0)$

3. lower bound $\quad C_\gamma(g,f) \geq H_\gamma(g)$ or $D_\gamma(g,f) \geq 0$

Characterization

Thm: Let $F(g,f) = \Phi\left(\int \rho(f)\right) \int \psi(f) g$

If F satisfies

(i) $F(g, \lambda f) = F(g,f) \quad (\forall \lambda > 0)$

(ii) $F(g,f) \geq F(g,g)$

then there is a constant γ suct that $F(g,f) = C_\gamma(g,f)$

Remark: If $(\Phi(Y), \rho(f), \psi(f)) = (Y^{-\frac{\gamma}{1+\gamma}}, f^{1+\gamma}, f^\gamma)$,

then $F(g,f) = \left(\int f(x)^{1+\gamma} dx\right)^{-\frac{\gamma}{1+\gamma}} \int f^\gamma g = C_\gamma(g,f)$

γ-power divergence geometry

For a model $M = \{f_\theta(x) : \theta \in \Theta\}$

γ -**power metric**

$$G_{ij}^{(\gamma)}(\theta) = \frac{1}{\gamma}\int \frac{\partial f_\theta}{\partial \theta_i}\frac{\partial}{\partial \theta_j}\left(\frac{f_\theta}{\| f_\theta \|}\right)^\gamma \mathrm{d}\mu$$

γ -**power connection pair**

$$\Gamma_{ij,k}^{(\gamma)}(\theta) = \frac{1}{\gamma}\int \frac{\partial^2 f_\theta}{\partial \theta_i \partial \theta_j}\frac{\partial}{\partial \theta_k}\left(\frac{f_\theta}{\| f_\theta \|}\right)^\gamma \mathrm{d}\mu$$

$$^*\Gamma_{ij,k}^{(\gamma)}(\theta) = \frac{1}{\gamma}\int \frac{\partial f_\theta}{\partial \theta_k}\frac{\partial^2}{\partial \theta_j \partial \theta_j}\left(\frac{f_\theta}{\| f_\theta \|}\right)^\gamma \mathrm{d}\mu$$

β-**power metric**

$$G_{ij}^{(\beta)}(\theta) = \frac{1}{\beta}\int \frac{\partial f_\theta}{\partial \theta_i}\frac{\partial f_\theta^{\beta}}{\partial \theta_j} \mathrm{d}\mu$$

β -**power connection pair**

$$\Gamma_{ij,k}^{(\beta)}(\theta) = \frac{1}{\beta}\int \frac{\partial^2 f_\theta}{\partial \theta_i \partial \theta_j}\frac{\partial f_\theta^{\beta}}{\partial \theta} \mathrm{d}\mu$$

$$^*\Gamma_{ij,k}^{(\beta)}(\theta) = \frac{1}{\beta}\int \frac{\partial f_\theta}{\partial \theta_k}\frac{\partial^2 f_\theta^{\beta}}{\partial \theta_j \partial \theta_j} \mathrm{d}\mu$$

γ-estimation

Parametric model $M = \{ f_\theta(x) : \theta \in \Theta \}$

γ-power loss function $L_\gamma(\theta) = -\frac{1}{n}\frac{1}{\gamma(\gamma+1)}\sum_{i=1}^{n}\left(\frac{f_\theta(x_i)}{\| f_\theta \|}\right)^\gamma$

γ-estimator $\hat{\theta}_\gamma = \arg\min_{\theta \in \Theta} L_\gamma(\theta)$

β-power loss function $L_\beta(\theta) = -\frac{1}{n\beta}\sum_{i=1}^{n} f(x_i,\theta)^\beta + \frac{1}{\beta+1}\| f(\cdot,\theta) \|^{\beta+1}$

β-estimator $\hat{\theta}_\beta = \arg\min_{\theta \in \Theta} L_\beta(\theta)$

Consistency

$$\text{If } x_1,\ldots,x_n \underset{\text{iid}}{\sim} g(x), \text{ then } \hat{\theta}_\gamma \underset{\text{as}}{\to} \theta_g$$

$$\text{where } \theta_g = \arg\min_{\theta \in \Theta} C_\gamma(g, f(\cdot,\theta))$$

$$\text{If } g(x) = f(x,\theta), \text{ then } \quad \theta_g = \theta$$

$$\text{because } C_\gamma(f(\cdot,\theta), f(\cdot,\theta^*)) \geq H_\gamma(f(\cdot,\theta))$$

$$\boxed{\text{If } x_1,\ldots,x_n \underset{\text{iid}}{\sim} f(x,\theta), \text{ then } \hat{\theta}_\gamma \underset{\text{as}}{\to} \theta}$$

Cf. Wald (1947), White (1982)

γ-estimating function

γ-estimating function

$$\mathrm{e}_{\gamma}(x,\theta) = f_{\theta}(x)^{\gamma} S_{\theta}(x) - \boxed{\frac{f_{\theta}(x)^{\gamma}}{\| f_{\theta} \|^{\gamma+1}}} \mathrm{E}_{f_{\theta}}(f_{\theta}{}^{\gamma} S_{\theta})$$

β-estimating function

$$\mathrm{e}_{\beta}(x,\theta) = f_{\theta}(x)^{\beta} S_{\theta}(x) - \mathrm{E}(f_{\theta}{}^{\beta} S_{\theta})$$

Asymptotic variance

$$\mathrm{var}_{\mathrm{A}}(\hat{\theta}_{\gamma}) = \frac{1}{n}\left\{\mathrm{E}\left(\frac{\partial \mathrm{e}_{\gamma}}{\partial \theta^{\mathrm{T}}}\right)\right\}^{-1} \mathrm{Var}(\mathrm{e}_{\gamma}) \left\{\mathrm{E}\left(\frac{\partial \mathrm{e}_{\gamma}{}^{\mathrm{T}}}{\partial \theta}\right)\right\}^{-1}$$

Note: $\beta = 0 \quad \Rightarrow \quad \mathrm{e}_{\beta}(x,\theta) = S_{\theta}(x)$; $\gamma = 0 \quad \Rightarrow \quad \mathrm{e}_{\gamma}(x,\theta) = S_{\theta}(x)$

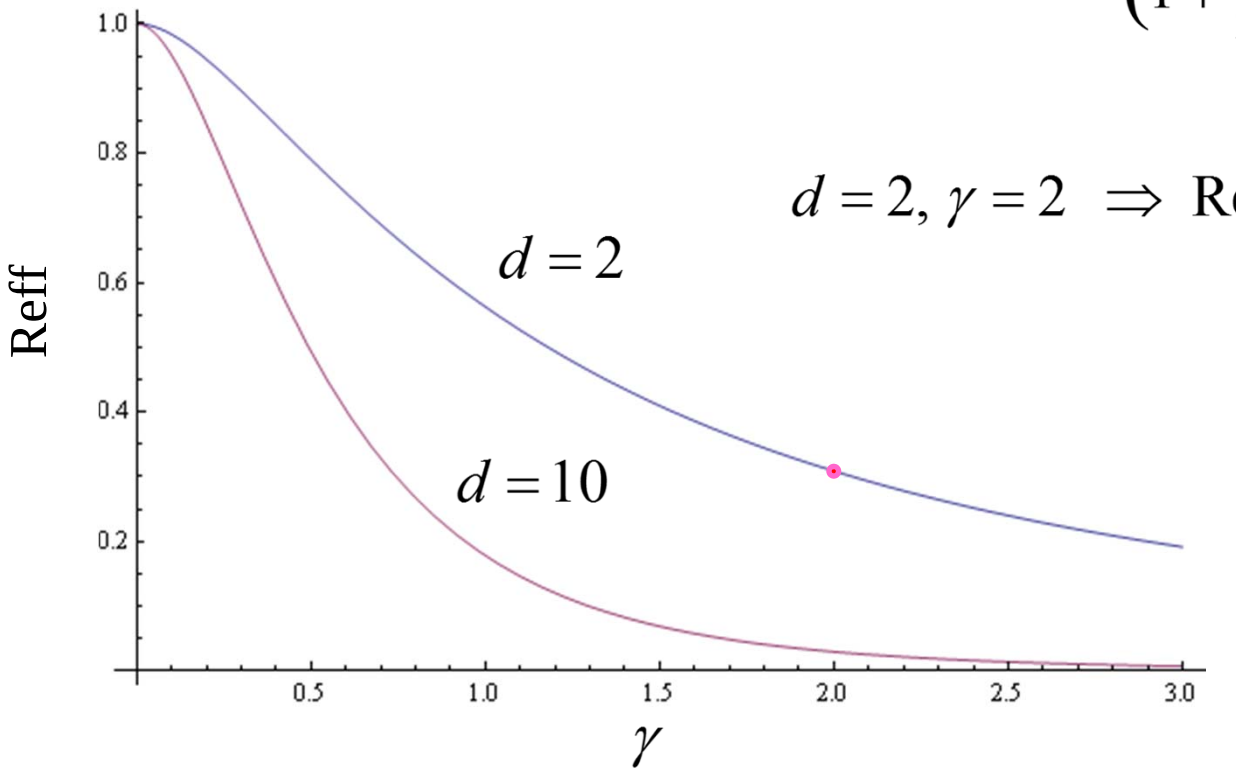
For any $\gamma \neq 0$, $\hat{\theta}_{\gamma}$ is ineffecient relative to MLE (Also $\hat{\theta}_{\beta}$)

Relative efficiency for γ-estimator

Normal mean model $f(\boldsymbol{x},\boldsymbol{\theta}) = (2\pi)^{-d/2} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\theta})\}, \quad \boldsymbol{x} \in \mathbb{R}^d$

Relative efficiency $$\mathrm{Reff} = \frac{\det(\mathrm{var}_{\mathrm{A}}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}))}{\det(\mathrm{var}_{\mathrm{A}}(\hat{\boldsymbol{\theta}}_{\gamma}))} = \frac{1}{\left(1+\frac{\gamma^2}{2\gamma+1}\right)^{-\frac{d+2}{2}}} < 1$$

$$d = 2,\ \gamma = 2 \ \Rightarrow\ \mathrm{Reff} = 0.309$$



γ-estimator for normal mean

γ-loss function $L_\gamma(\boldsymbol{\theta}) \propto -\sum_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{\theta})^\gamma$ for normal mean model

$$\frac{\partial}{\partial \boldsymbol{\theta}} L_\gamma(\boldsymbol{\theta}) \propto -\sum_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{\theta})^\gamma (\boldsymbol{x}_i - \boldsymbol{\theta}) = 0$$

$\hat{\boldsymbol{\theta}}_\gamma$ is a weighted mean: $\hat{\boldsymbol{\theta}}_\gamma = \dfrac{\sum \{f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_\gamma)\}^\gamma \boldsymbol{x}_i}{\sum \{f(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_\gamma)\}^\gamma}$

Iteratively reweighted mean

$$\boldsymbol{\theta}_{t+1} \leftarrow \frac{\sum \{f(\boldsymbol{x}_i, \boldsymbol{\theta}_t)\}^\gamma \boldsymbol{x}_i}{\sum \{f(\boldsymbol{x}_i, \boldsymbol{\theta}_t)\}^\gamma} \quad \text{starting from } \hat{\boldsymbol{\theta}}_0 = \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$$

Note: β-estimator $\hat{\boldsymbol{\theta}}_\beta$ is the same as γ-estimator $\hat{\boldsymbol{\theta}}_\gamma$

γ-estimator for normal mean and variance

Normal model $$M = \{ \; f(\boldsymbol{x},\theta,V) = \det(2\pi V)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^{\mathrm{T}} V^{-1}(\boldsymbol{x}-\boldsymbol{\theta})\}$$

γ-loss function $$L_\gamma(\boldsymbol{\theta},\boldsymbol{V}) = -\frac{1}{n}\frac{(2\pi)^d}{\gamma(\gamma+1)^{d+1}}\sum_{i=1}^{n}\det(V)^{-\frac{1}{2}\frac{\gamma}{\gamma+1}} f(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{V})^\gamma$$

γ-estimator $$\hat{\boldsymbol{\theta}}_\gamma = \frac{\sum\{f(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_\gamma,\hat{\boldsymbol{V}}_\gamma)\}^\gamma \boldsymbol{x}_i}{\sum\{f(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_\gamma,\hat{\boldsymbol{V}}_\gamma)\}^\gamma}$$

$$\hat{\boldsymbol{V}}_\gamma = \frac{(\gamma+1)\sum\{f(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_\gamma,\hat{\boldsymbol{V}}_\gamma)\}^\gamma(\boldsymbol{x}_i-\hat{\boldsymbol{\theta}}_\gamma)(\boldsymbol{x}_i-\hat{\boldsymbol{\theta}}_\gamma)^{\mathrm{T}}}{\sum\{f(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_\gamma,\hat{\boldsymbol{V}}_\gamma)\}^\gamma}$$

Note: β-estimator $\hat{\boldsymbol{V}}_\beta$ has not such a weighted expression

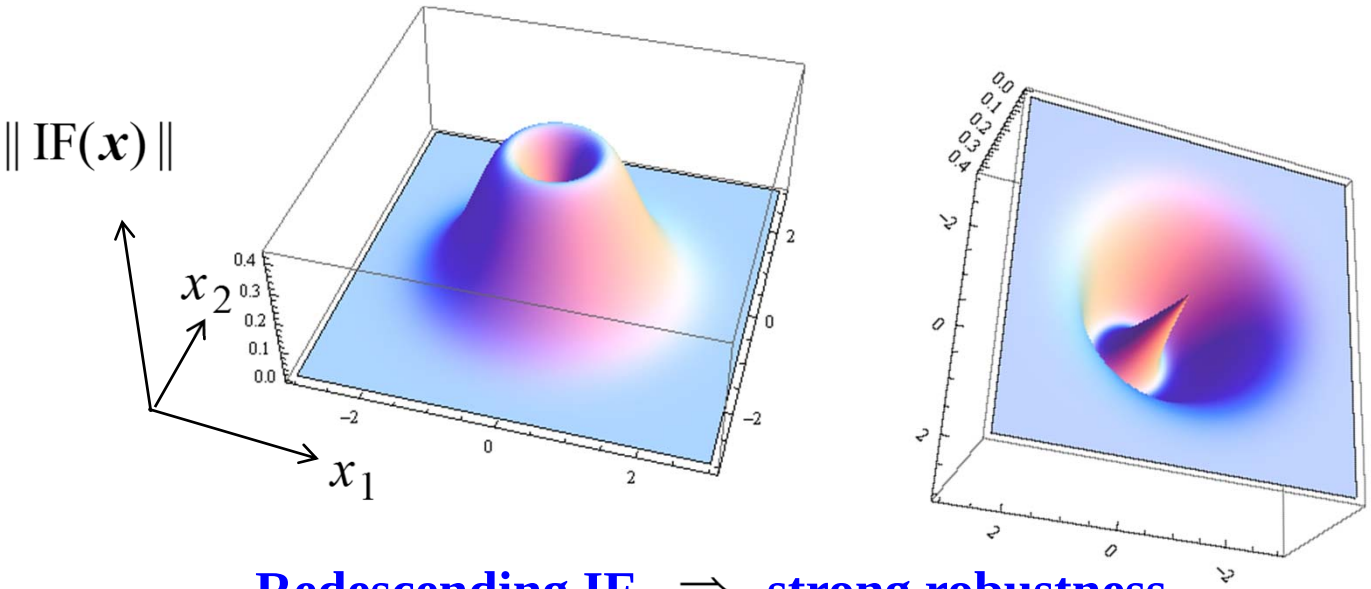
Influence function of γ-estimator

γ-estimator functional $$\hat{\boldsymbol{\theta}}_\gamma(g) = \arg\min_{\boldsymbol{\theta}\in\Theta} C_\gamma(g, f_{\boldsymbol{\theta}})$$

ε-contamination model $$g_\varepsilon(\boldsymbol{x}) = (1-\varepsilon) f(\boldsymbol{x},\boldsymbol{\theta}) + \varepsilon\delta(\boldsymbol{x})$$

$\delta(\boldsymbol{x})$ is Dirac delta function

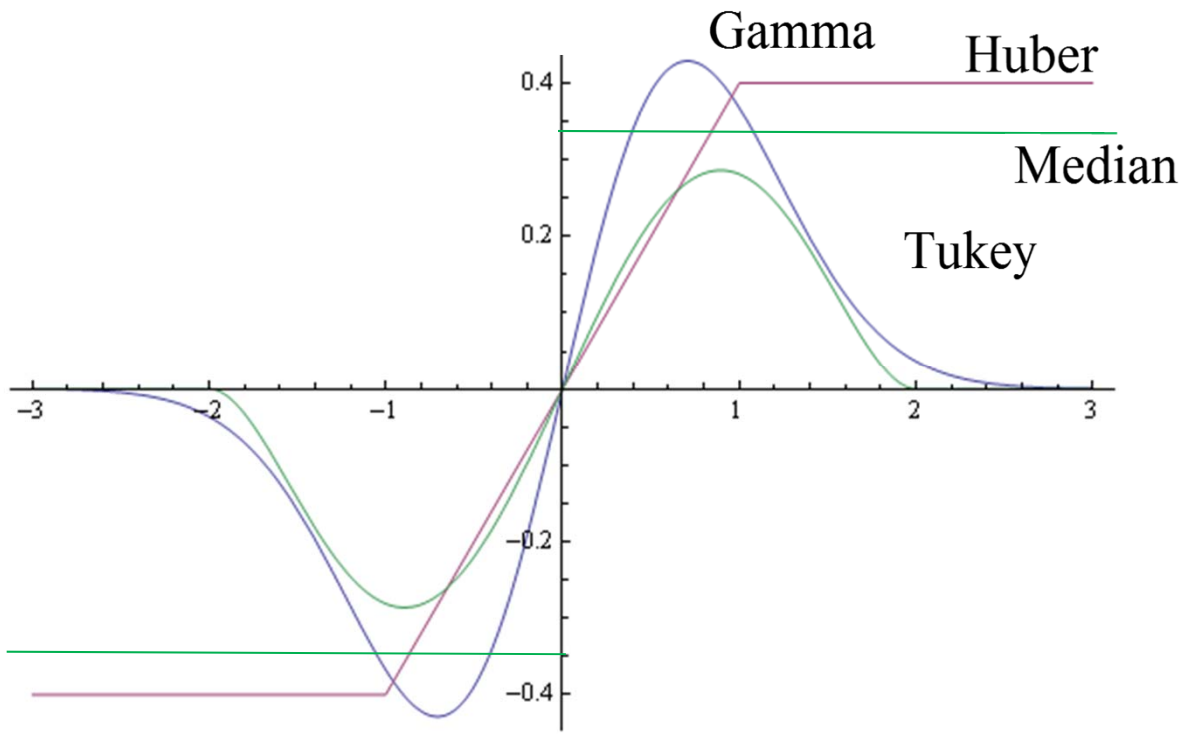
Influence function $$\mathrm{IF}(\boldsymbol{x}) = \frac{\partial}{\partial\varepsilon}\hat{\boldsymbol{\theta}}_\gamma(g_\varepsilon)|_{\varepsilon=0} = f(\boldsymbol{x},\boldsymbol{\theta})^\gamma(\boldsymbol{x}-\boldsymbol{\theta})$$



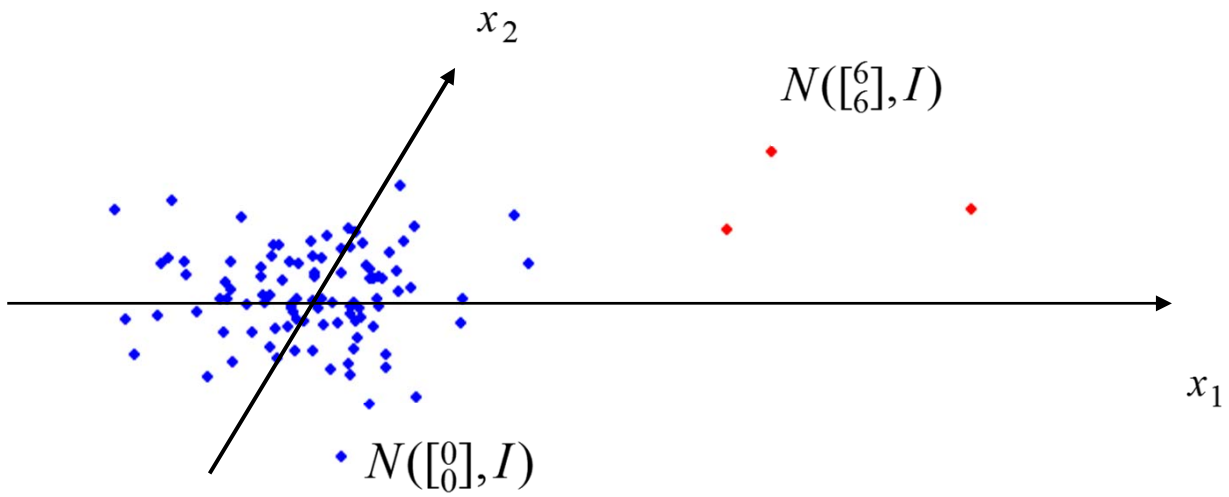
Redescending IF $\Rightarrow$ **strong robustness**

Influence functions

γ-weight	$xe^{-\frac{\gamma}{2}x^2}$
Median weight	$\mathrm{sgn}(x)$
Huber's weight	$xI_{[-c,c]}(x)+cI_{(c,\infty)}(x)-cI_{(-\infty,-c)}(x)$
Tukey's biweight	$x(1-x^2/c^2)^2 I_{[-c,c]}(x)$

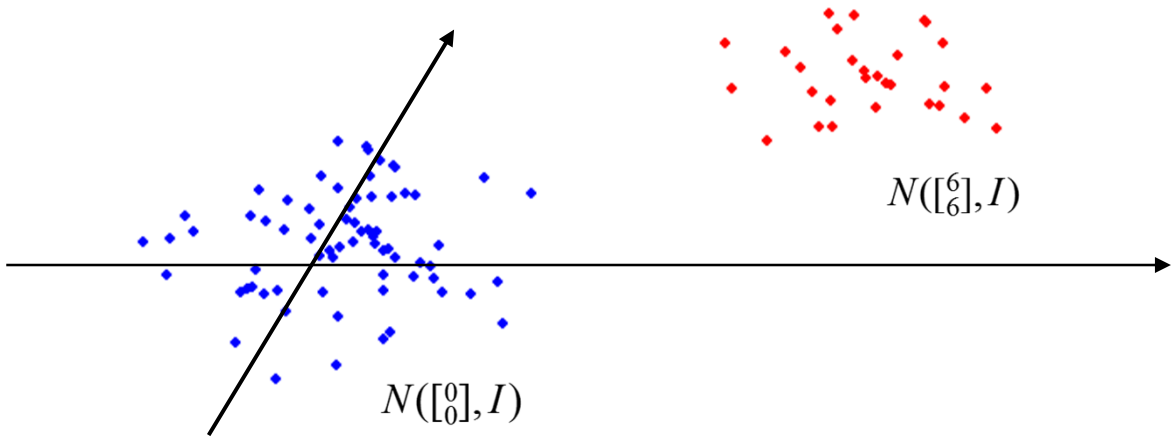


Robustness



true value	(0, 0)	(0, 0) + **outlying (6,6)**
maximum likelihood	(–0.0738, 0.0065)	(–0.1082, 0.1612)
median	(0.0260,-0.0606)	(0.0166,–0.0241)
γ–estimator (γ=1)	(–0.0192, 0.0279)	(–0.0192, 0.0279)

Super robustness



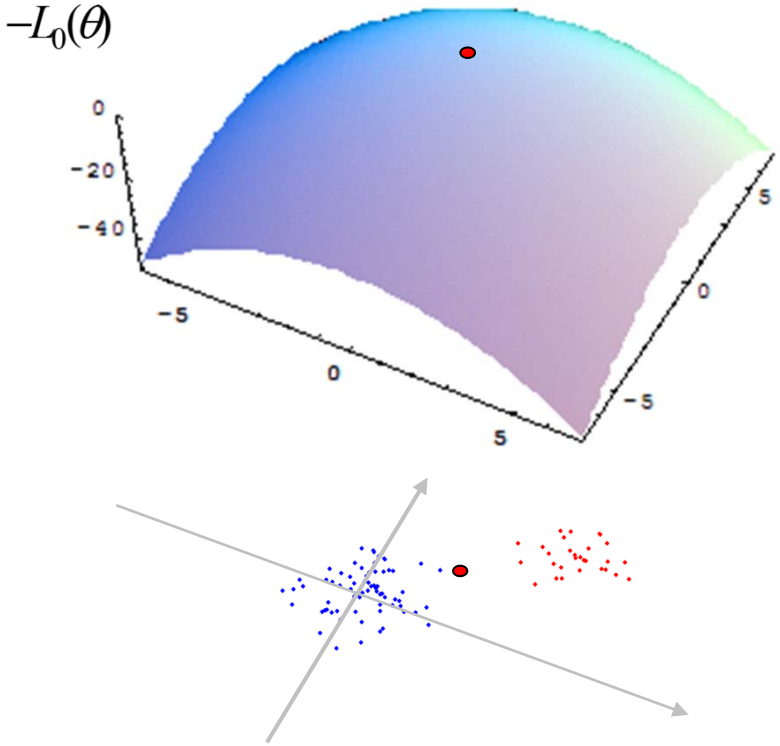
true value	(0, 0)	(0, 0) + **more outlying (6,6)**
maximum likelihood	(0.0067, –0.0196)	(1.93068, 1.76655)
median	(-0.06663, 0.0429)	(0.56484, 0.399286)
γ-estimator (γ=1)	(0.01069, 0.01595)	(0.01069, 0.01595)

Convex Learning

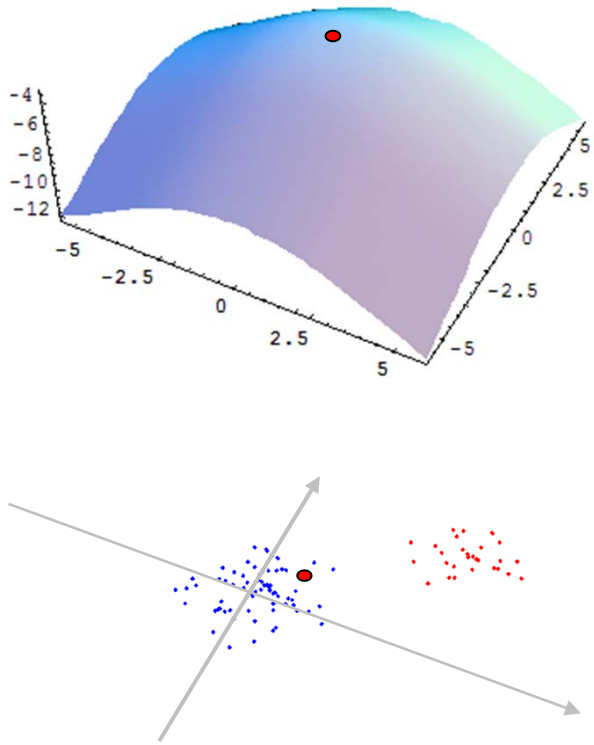
$$L_0(\mu) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^{\mathrm{T}}(x_i - \mu)$$

$$L(\mu) = \frac{1}{n}\sum_{i=1}^{n}\| x_i - \mu \|_{L_1}$$

MLE $\begin{bmatrix} 1.93 \\ 1.76 \end{bmatrix}$

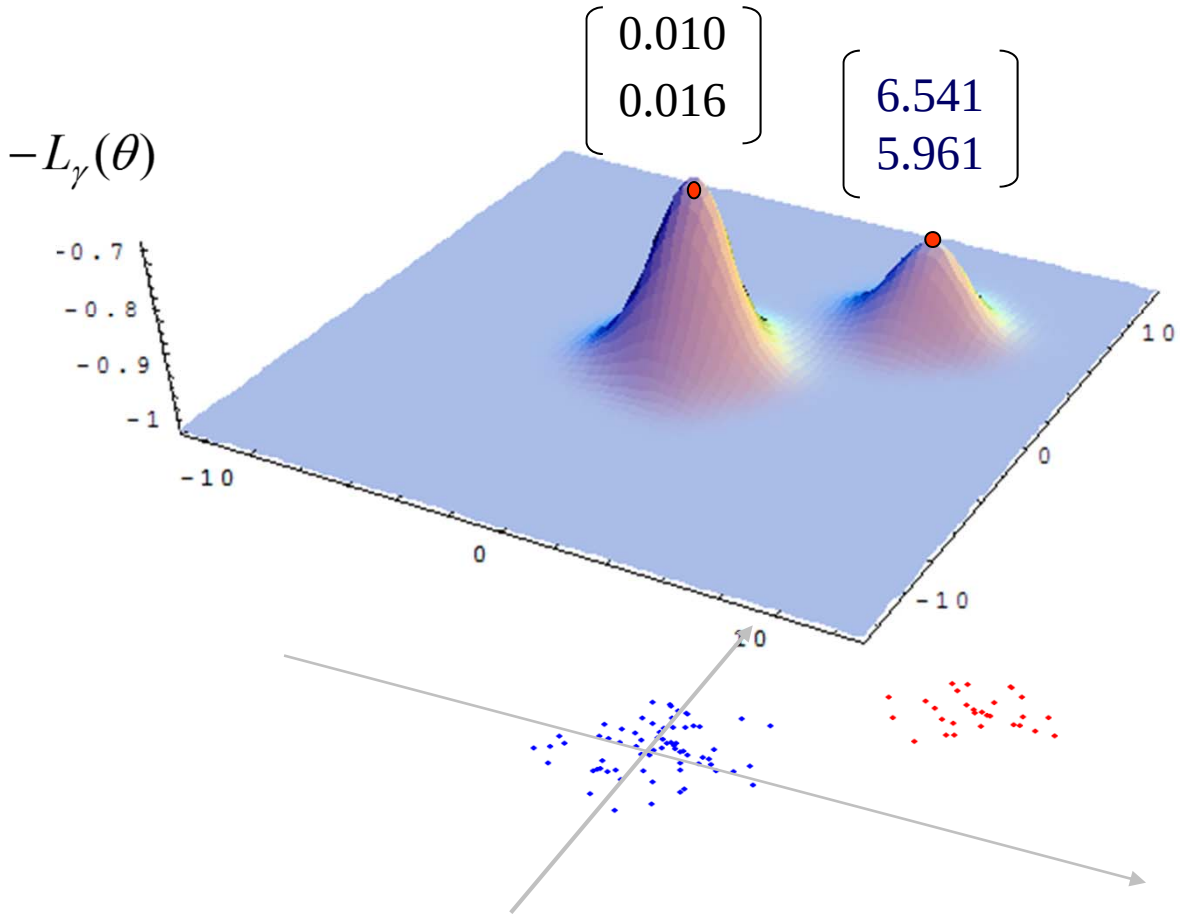


Median $\begin{bmatrix} 0.565 \\ 0.399 \end{bmatrix}$

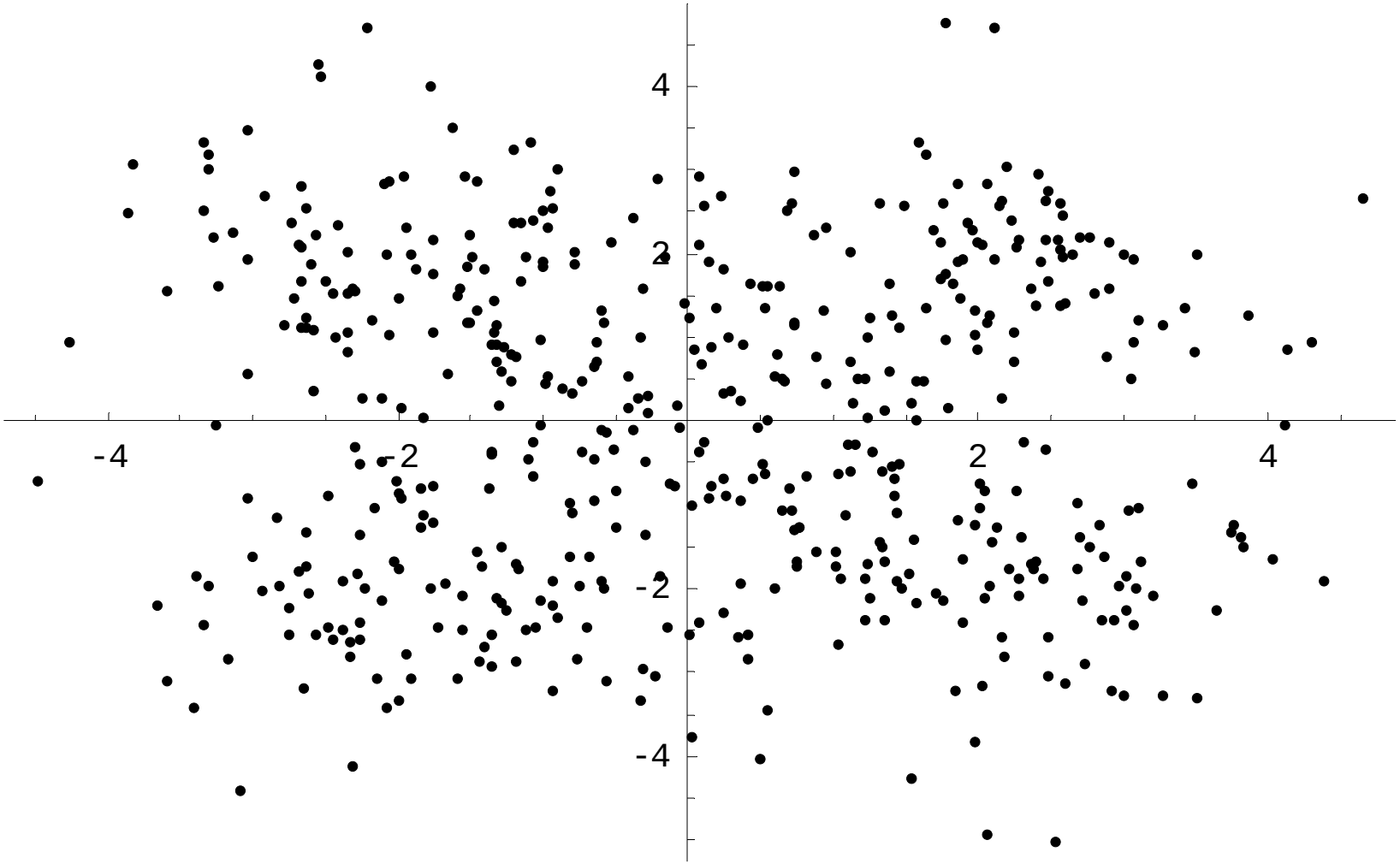


Non-convex learning

$$L_\gamma(\mu) = -\frac{1}{\gamma(\gamma+1)}\frac{1}{n}\sum_{i=1}^{n} e^{-\frac{\gamma}{2}(x_i-\mu)^{\mathrm{T}}(x_i-\mu)}$$



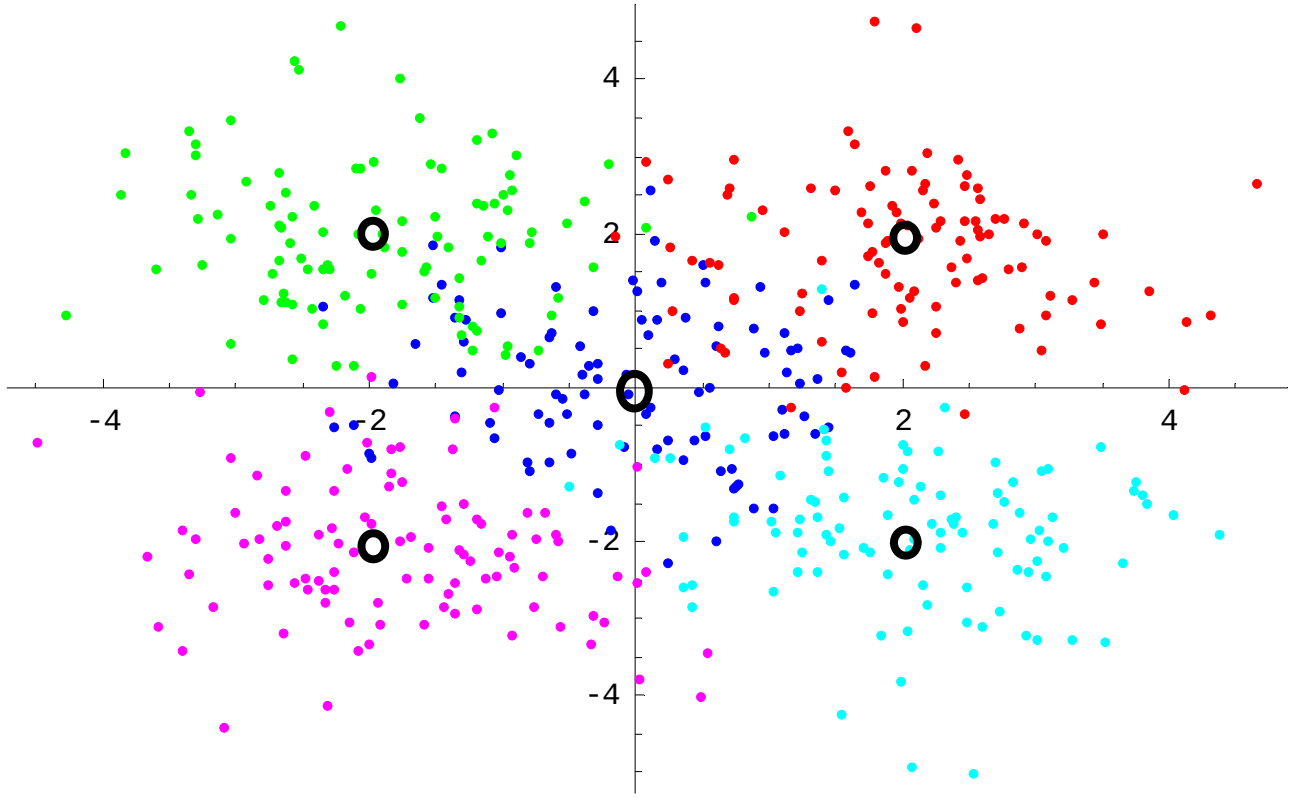
A data plot



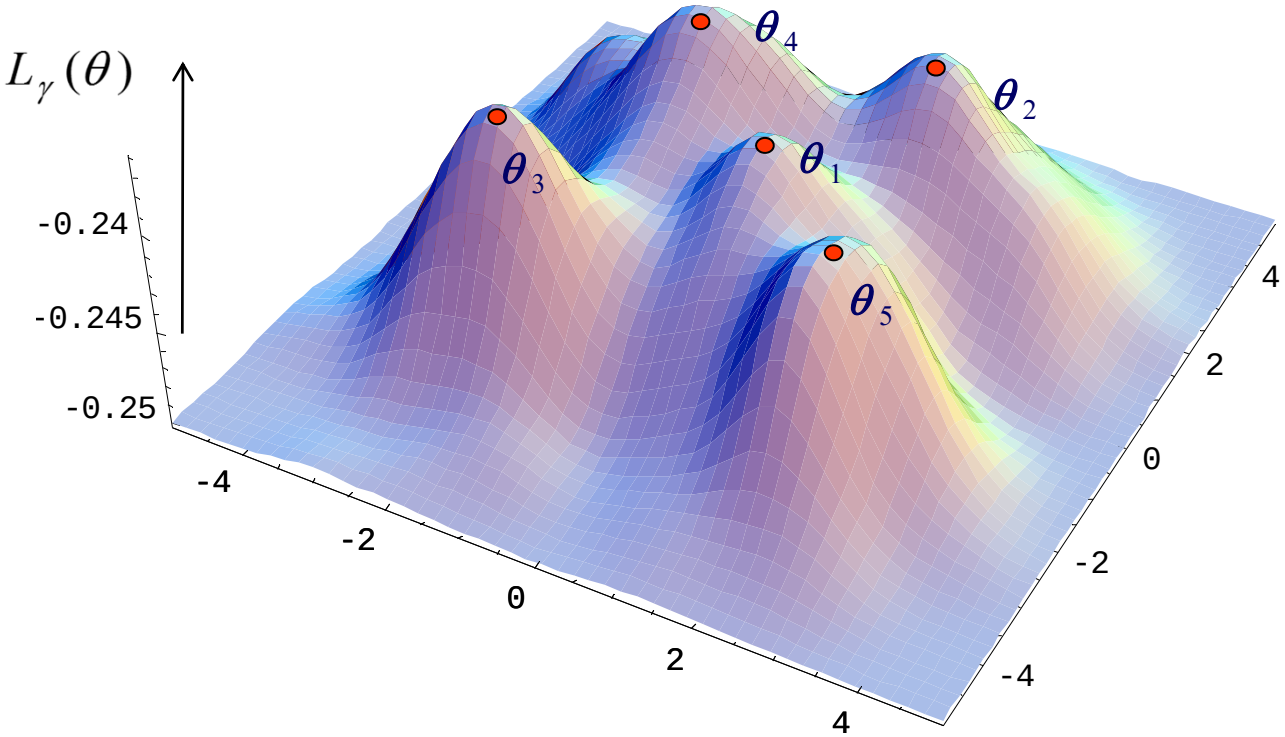
Five-mixture

Deta set $\{\boldsymbol{x}_i\}_{i=1}^n$ follows a data density

$$p(\boldsymbol{x}) = \frac{1}{5}\phi(\boldsymbol{x}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}) + \frac{1}{5}\phi(\boldsymbol{x}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}) + \frac{1}{5}\phi(\boldsymbol{x}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}) + \frac{1}{5}\phi(\boldsymbol{x}, \begin{bmatrix} -2 \\ 2 \end{bmatrix}) + \frac{1}{5}\phi(\boldsymbol{x}, \begin{bmatrix} -2 \\ -2 \end{bmatrix})$$



Plot of γ-loss



$\theta_1 = \{-0.6040, 0.4878\}$, Loss = -0.4576

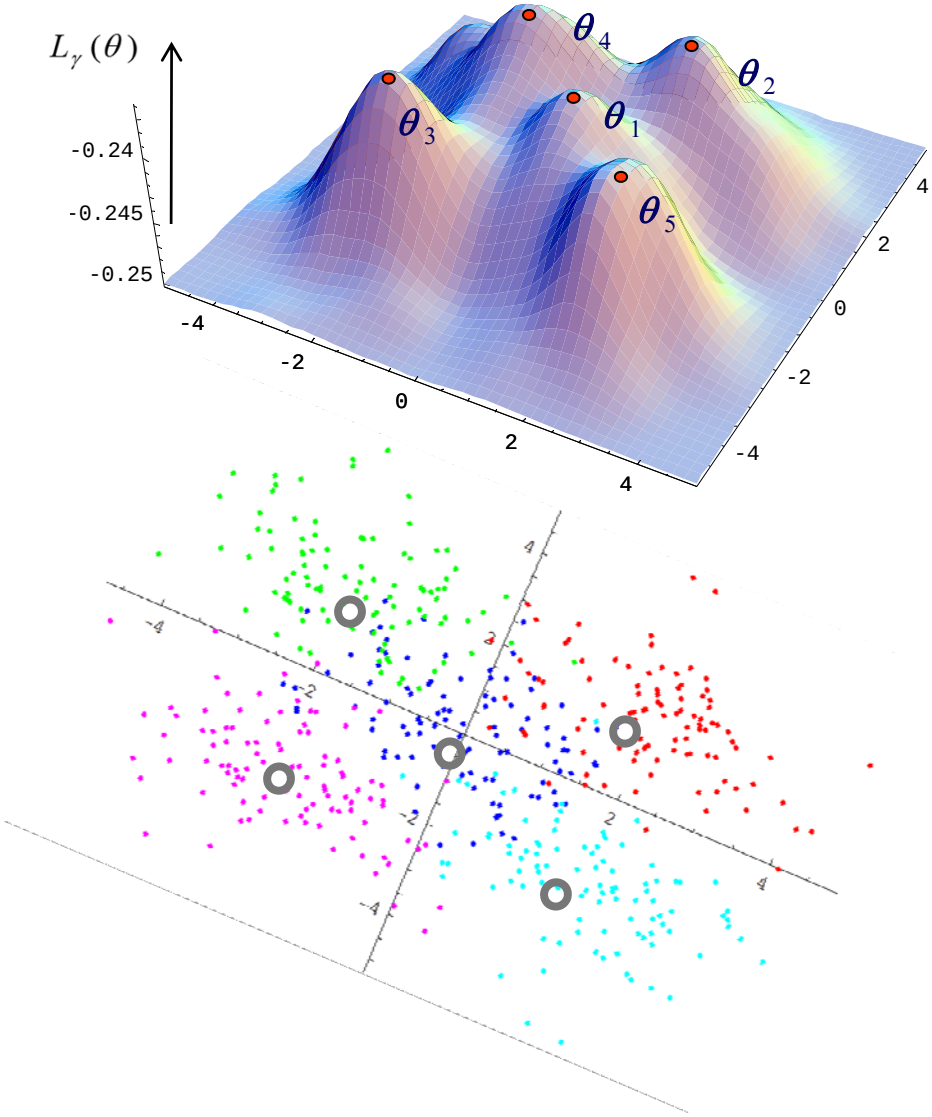
$\theta_2 = \{2.095, 1.846\}$, Loss = -0.4584

$\theta_3 = \{-1.595, -2.028\}$, Loss = -0.4575

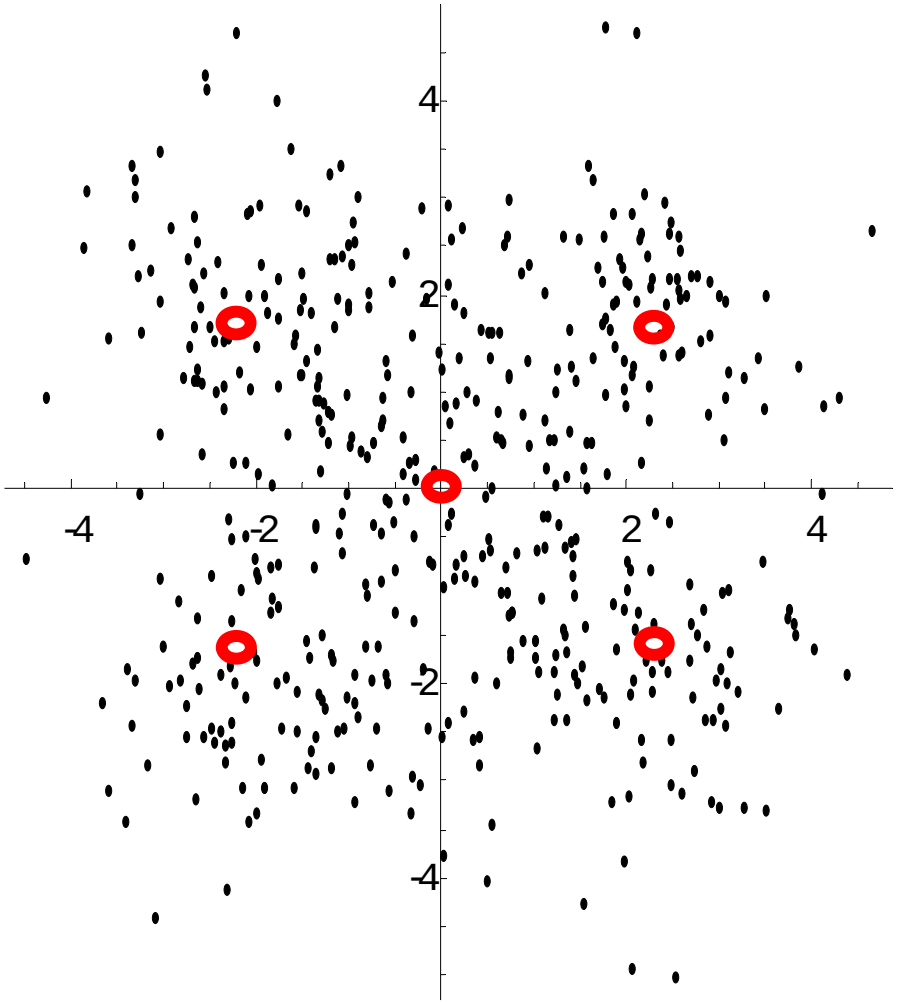
$\theta_4 = \{-1.437, 1.287\}$, Loss = -0.4576

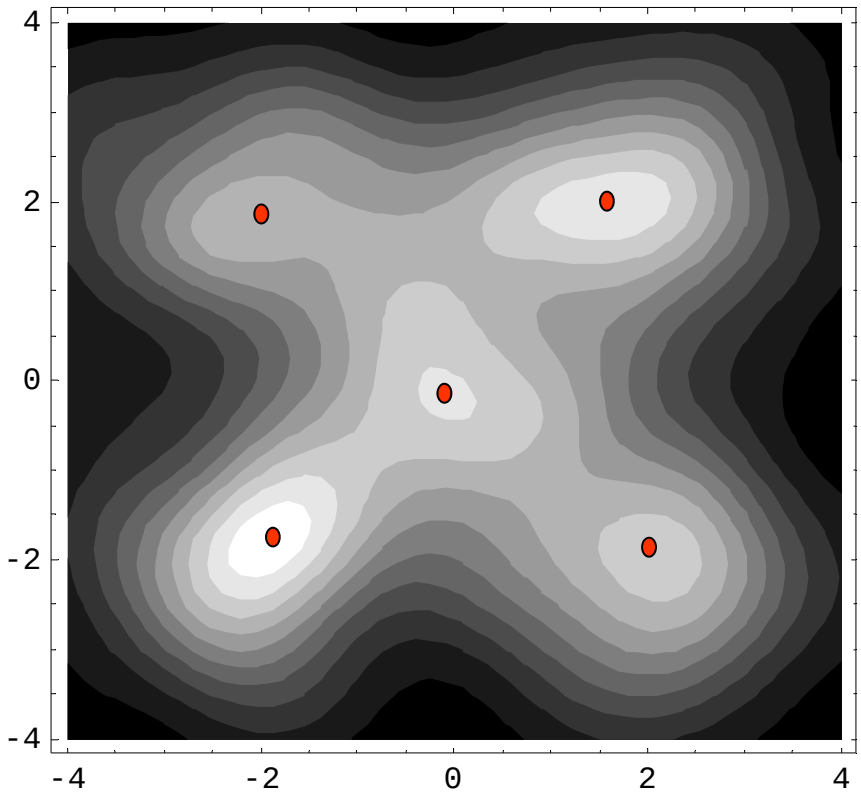
$\theta_5 = \{1.268, -1.277\}$, Loss = -0.4584

Detecting hidden structure!



Cluster Number = 5





Clustering algorithm

$$\text{IRM}: \quad \boldsymbol{\mu} \leftarrow \sum_{i=1}^{n} w_i(\boldsymbol{\mu})\boldsymbol{x}_i, \qquad w_i(\boldsymbol{\mu}) = \frac{\{f(\boldsymbol{x}_i, \boldsymbol{\mu})\}^{\gamma}}{\sum \{f(\boldsymbol{x}_i, \boldsymbol{\mu})\}^{\gamma}}$$

Step 1. Find $\boldsymbol{\mu}_1 \xleftarrow{\text{IRM}} \boldsymbol{\mu}_{\text{ML}}$

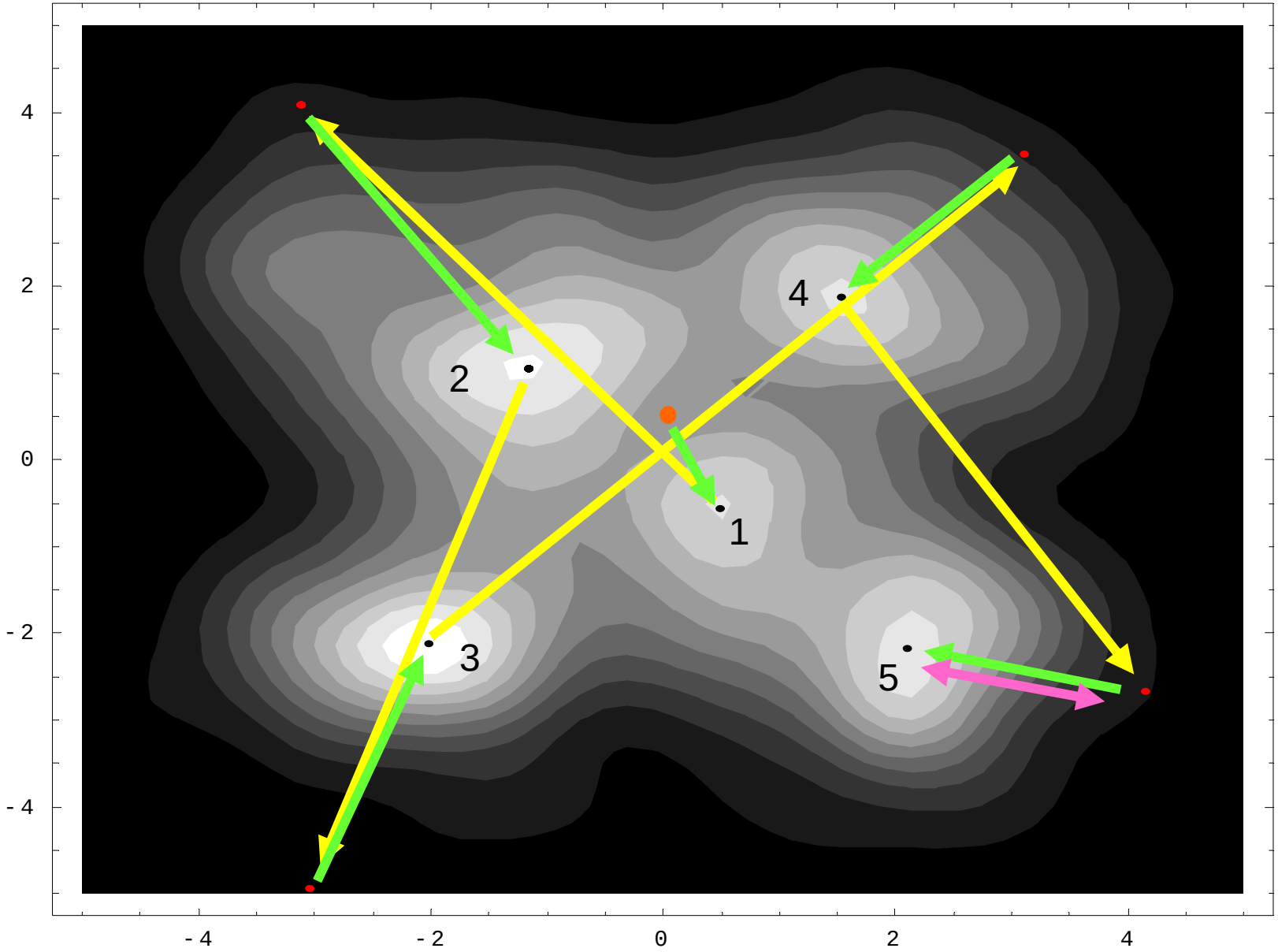
Step 2. Find $\boldsymbol{\mu}_{t+1} \xleftarrow{\text{IRM}} \boldsymbol{\mu}_{0t}$ for $t \geq 1$

where $\boldsymbol{\mu}_{0t} = \arg\max \{ \min_{1 \leq s \leq t} \| \boldsymbol{\mu} - \boldsymbol{\mu}_s \| : \boldsymbol{\mu} \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \}$

Repeat until $\boldsymbol{\mu}_{k+1} \in \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k\}$.

Cf. k-means, model-based

Starting from MLE



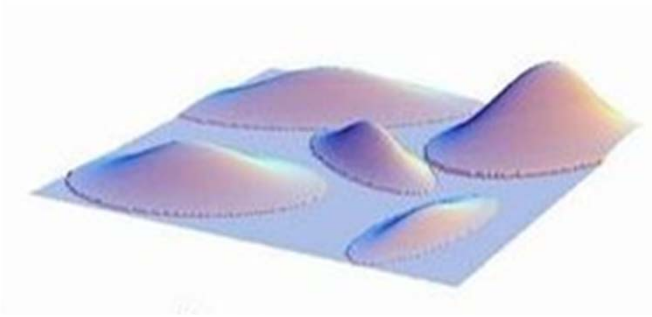
Variance adapt

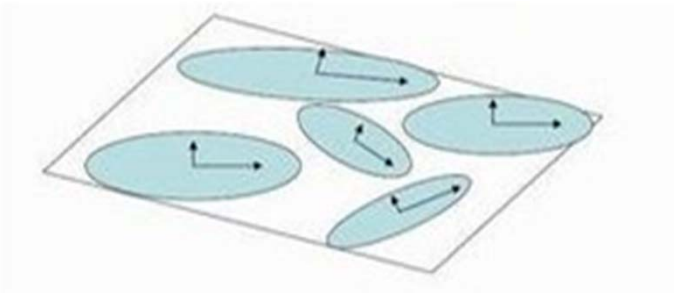
γ-loss function

$$L_\gamma(\mu\ ,\Sigma) = -\frac{1}{n}\det(\Sigma)^{-\frac{\gamma}{1+\gamma}}\sum_{i=1}^{n}\exp\{-\tfrac{\gamma}{2}(x_i-\mu)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu)\}$$

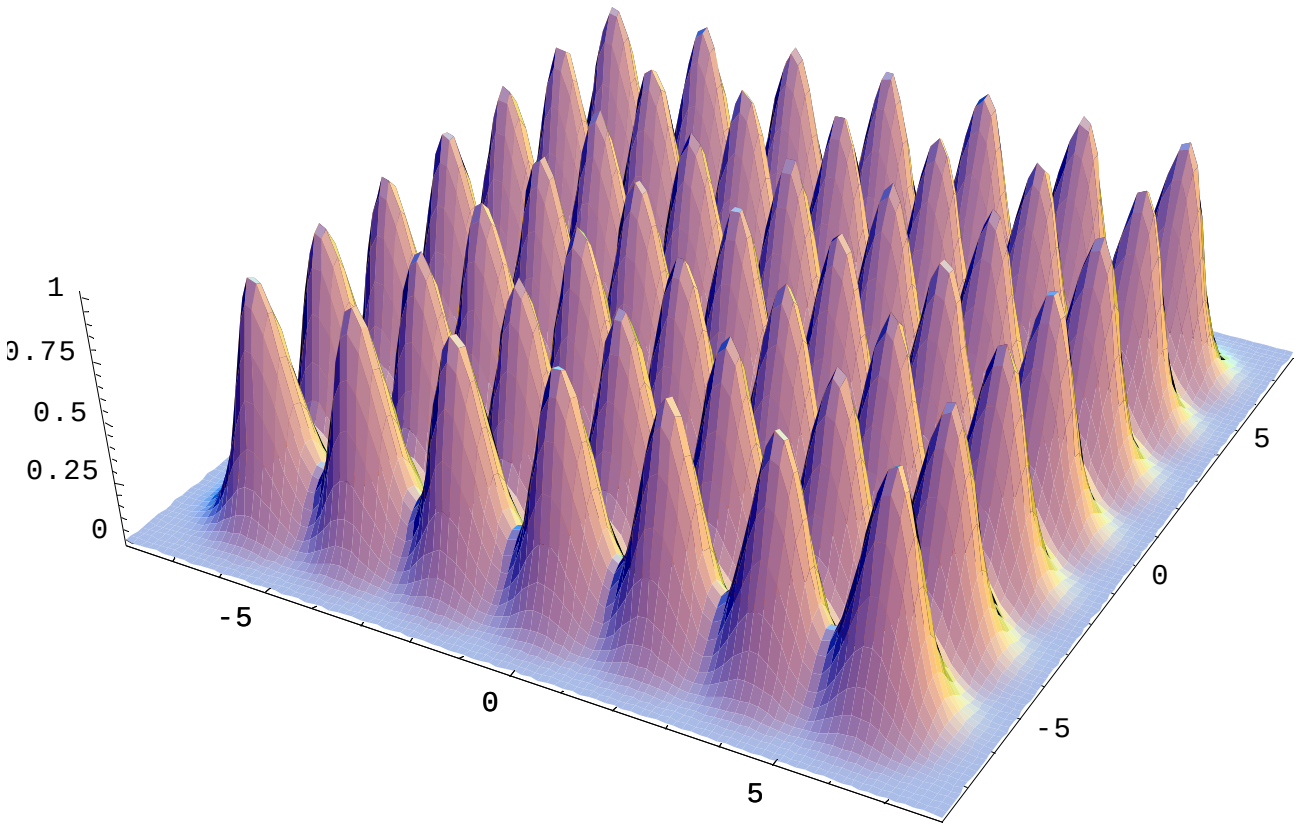
$$\text{IRV} \quad (**) \quad \boldsymbol{\Sigma} \quad \leftarrow \quad (1+\gamma)\sum_{i=1}^{n} w_i(\boldsymbol{\mu},\boldsymbol{\Sigma})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

$$\text{with weights} \quad w_i(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{f(\boldsymbol{x}_i,\boldsymbol{\mu},\boldsymbol{\Sigma})^\gamma}{\sum f(\boldsymbol{x}_i,\boldsymbol{\mu},\boldsymbol{\Sigma})^\gamma}$$



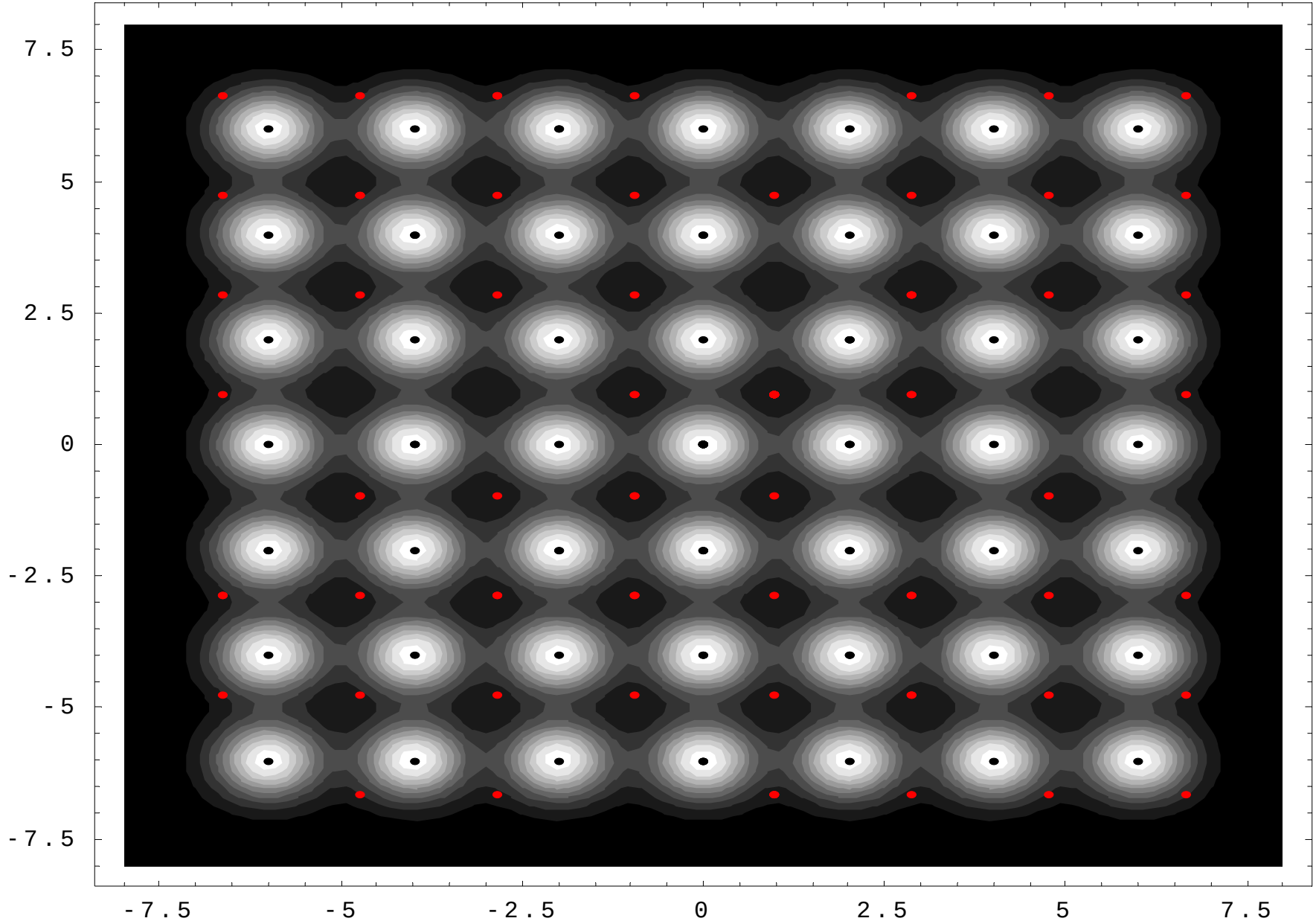


7 x 7 modes

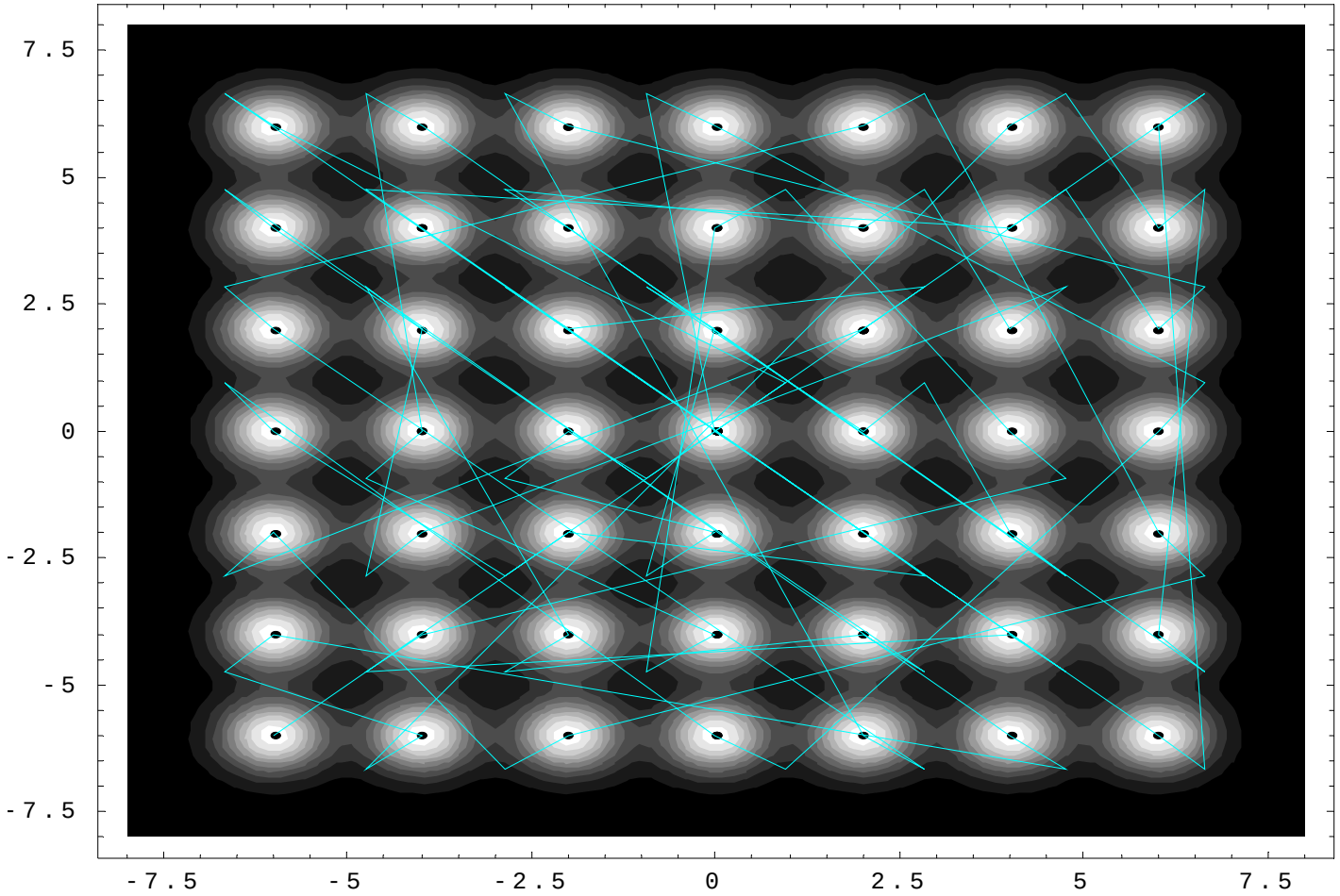


$$g(\boldsymbol{x}) = \sum_{j=1}^{7}\sum_{i=1}^{7}\frac{1}{2\pi}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{ij})^{\mathrm{T}}((\boldsymbol{x}-\boldsymbol{\mu}_{ij})\}$$

Expected case



Trace of IRM



Scalable !

Why spontaneous data learning?

- We just estimate a normal mean adopting γ-estimator
- The γ-loss function becomes nonconvex if data distribution is multimodal, and γ is selected between 1 and 3.
- Local minima correspond to cluster centers if γ is adaptively selected, for example K-fold cross validation, information criterion.

Difference of convex functions

γ–loss function

$$\log\{-L_\gamma(\boldsymbol{\mu})\} = \log\sum_{i=1}^{n}\exp\{-\tfrac{\gamma}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{\mu})\} + \mathrm{const}$$

$$= \log\Big(\sum_{i=1}^{n} e^{\gamma\,\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\mu}-\frac{\gamma}{2}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i}\Big) - \tfrac{\gamma}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$$

Hessian

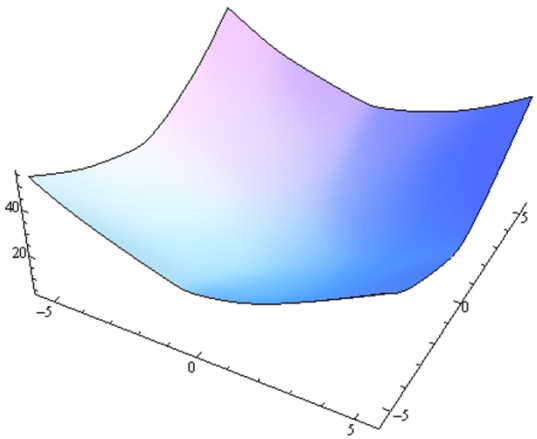
$$\frac{\partial}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^{\mathrm{T}}}\log\Big\{\sum_{i=1}^{n} e^{\gamma\,\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\mu}-\frac{\gamma}{2}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i}\Big\} = \sum_{i=1}^{n} w_i(\boldsymbol{\mu})(\boldsymbol{x}_i-\overline{\boldsymbol{\mu}})(\boldsymbol{x}_i-\overline{\boldsymbol{\mu}})^{\mathrm{T}}$$

$$\text{where}\quad w_i(\boldsymbol{\mu}) = \frac{f(\boldsymbol{x}_i,\boldsymbol{\mu})^{\gamma}}{\sum f(\boldsymbol{x}_j,\boldsymbol{\mu})^{\gamma}},\ \overline{\boldsymbol{\mu}} = \sum w_i(\boldsymbol{\mu})\boldsymbol{x}_i$$

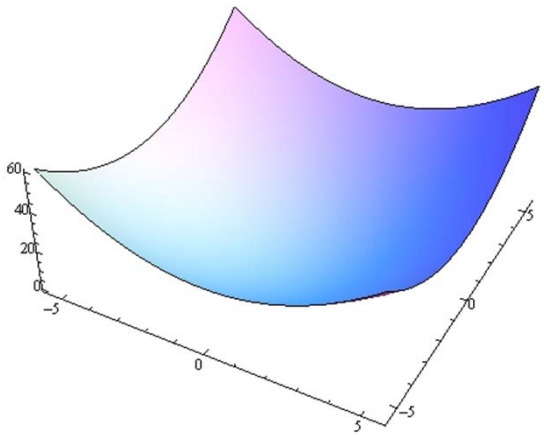
Cf. CCCP, Yuille & Rangarajan (2002), An & Tao (2005)

Difference of convex functions

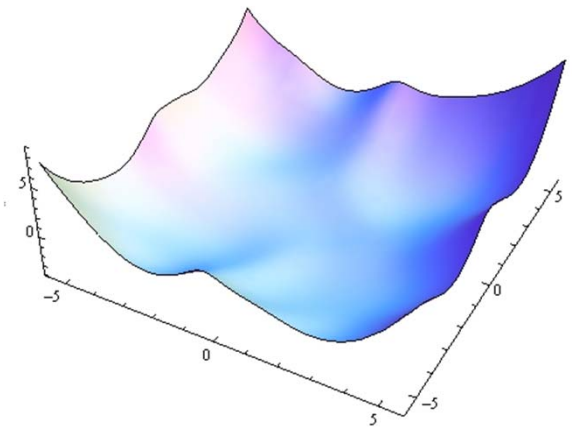
$$\log\{\sum_{i=1}^{n} e^{\gamma \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta} - \frac{\gamma}{2}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i}\}$$



$$\frac{\gamma}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$$



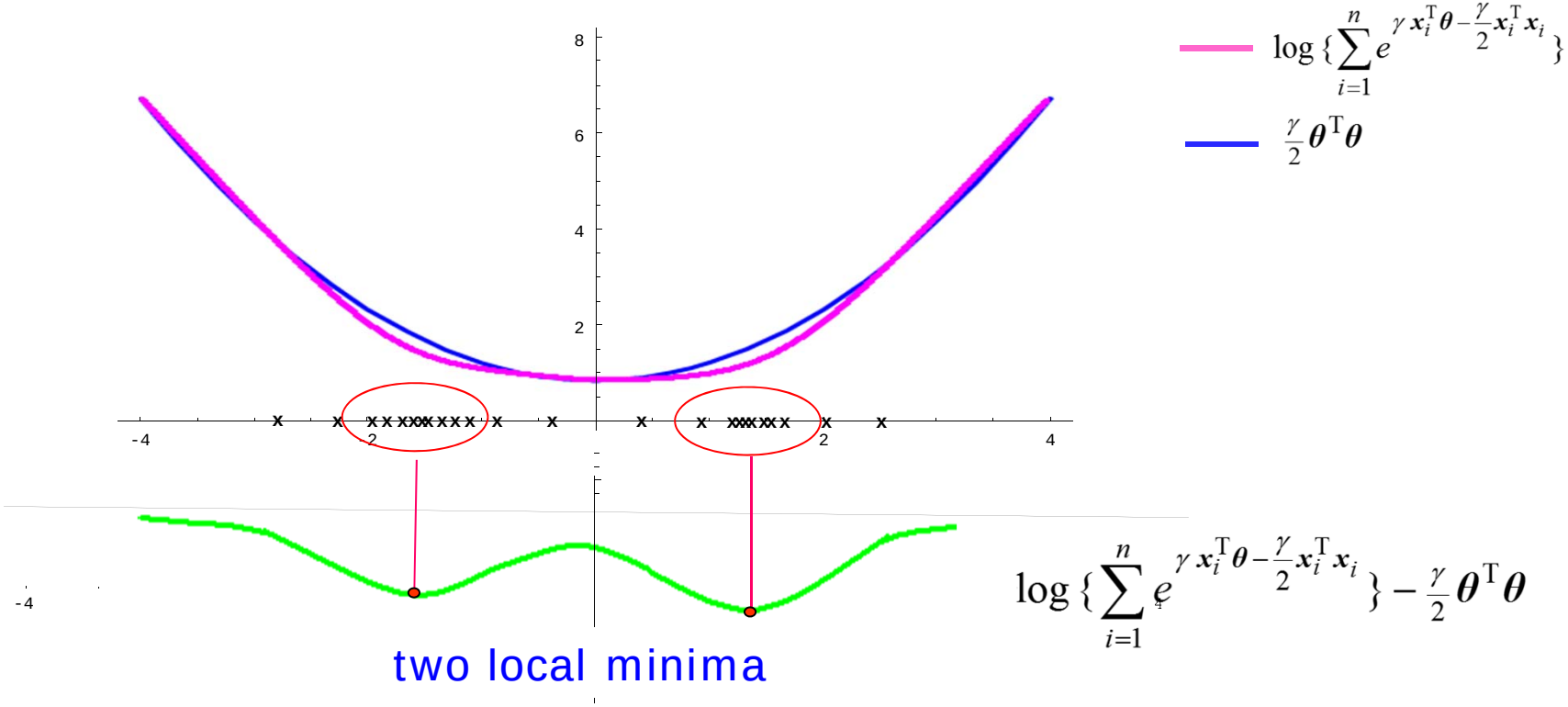
difference



Rem If $\{\boldsymbol{x}_i\} \sim \sum_{k=1}^{K} \pi_k N(\boldsymbol{\theta}_k, I)$, then $L_\gamma(\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} \sum_{k=1}^{K} \pi_k e^{-\frac{1}{2}\frac{\gamma}{\gamma+1}(\boldsymbol{\theta}-\boldsymbol{\theta}_k)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_k)}$

If $\{\boldsymbol{x}_i\} \sim N(\boldsymbol{\theta}_0, I)$, then $L_\gamma(\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} e^{-\frac{1}{2}\frac{\gamma}{\gamma+1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}$

Convexity depending on data



γ–loss function suggests local minima as centers of clusters

Why spontaneous data learning?

Why difference of convex functions?

I like to elucidate the reason why SDL occur

in information geometric understandings

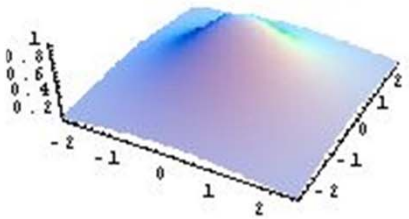
Max γ-entropy

γ-diagonal entropy $$H_\gamma(g) = \left(\int g(\boldsymbol{x})^{\gamma+1} \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{\gamma+1}}$$

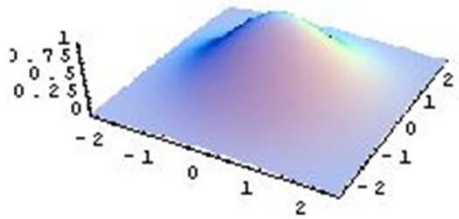
Equal mean space $$F(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{ f(\boldsymbol{x}) : \mathrm{E}_f(\boldsymbol{X}) = \boldsymbol{\mu}, \mathrm{V}_f(\boldsymbol{X}) = \boldsymbol{\Sigma} \}$$

Max γ-entropy $$f_\gamma(\cdot, \mu, \Sigma) = \underset{f \in F(\mu,\Sigma)}{\arg\max} \; H_\gamma(f)$$

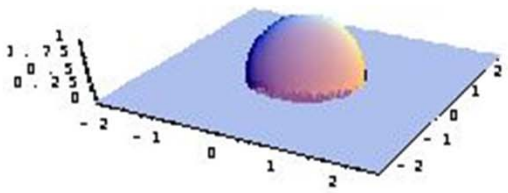
γ-model $$f_\gamma(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c_\gamma \det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} \{ 1 - \tfrac{1}{2}\gamma(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \}_+^{\frac{1}{\gamma}}$$



$\gamma = -0.3$ (t-distribution)



$\gamma = 0$ (normal)



$\gamma = 2$ (Wigner)

γ-estimator on γ-model

Let $(x_1,\ldots,x_n)$ be iid from $f_\gamma(\cdot,\mu,\Sigma)$

γ-loss function $$L_\gamma(\mu,\Sigma)=\frac{1}{n}\sum_{i=1}^{n}\det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{\gamma+1}}\{1-\frac{1}{2}\gamma(x_i-\mu)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu)\}$$

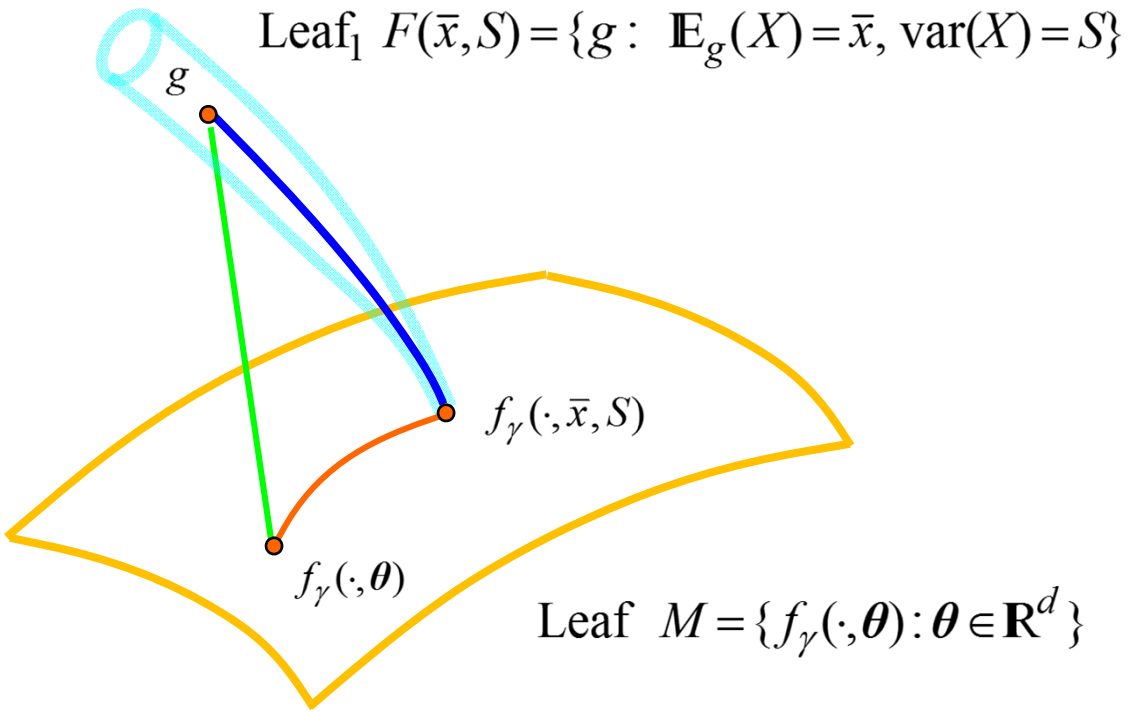
$$(\bar{x},S)=\arg\min_{(\mu,\Sigma)} L_\gamma(\mu,\Sigma) \quad \forall\gamma>\frac{2}{d+2}$$

$$L_\gamma(\mu,\Sigma)-L_\gamma(\bar{x},S)=D_\gamma(f_{\bar{x},S},f_{\mu,\Sigma})\geq 0$$

Let $g_{\mu,\Sigma}$ be a location-scale family.
The γ-estimator for (μ,Σ) is the sample mean and sample variance if and only if $g_{\mu,\Sigma}$ is the γ-model.

Cf. Teicher (1961)

Pythagoras foliation



Loss decomposition $L_\gamma(\boldsymbol{\theta})-L_\gamma(\bar{\boldsymbol{x}},\boldsymbol{S})=D_\gamma(f_\gamma(\cdot,\bar{\boldsymbol{x}},\boldsymbol{S}),f_\gamma(\cdot,\boldsymbol{\theta}))$

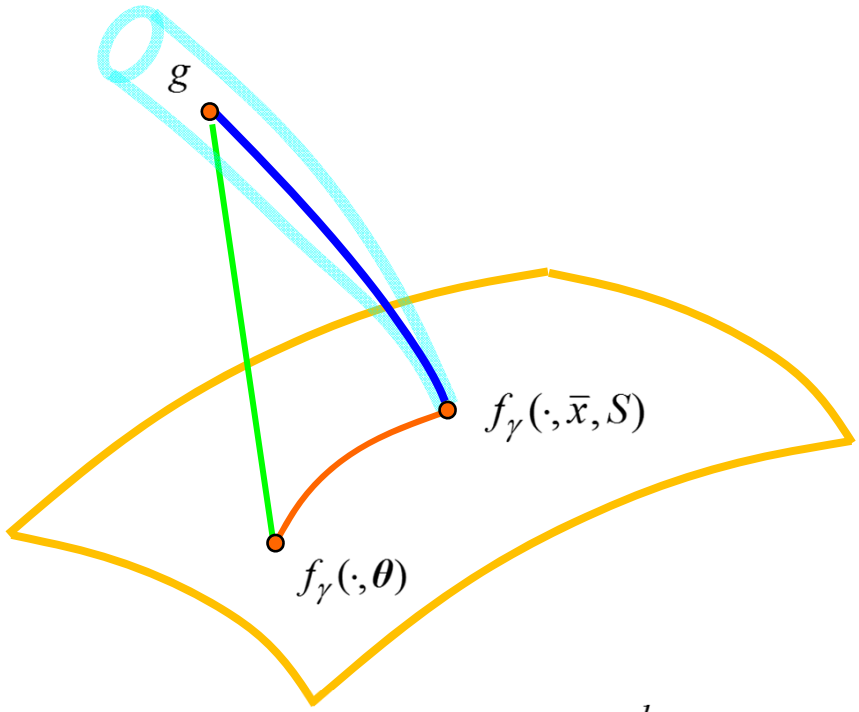
{Pythagorian triangle} $\bigcup_{\boldsymbol{\theta}\in\mathbf{R}^d} F(\boldsymbol{\theta})$

(γ-estimator , γ'-model)

model / loss function	0-model	γ-model
0-loss	$(\bar{x}, S)$ MLE	M-estimator
γ-loss	**Spontaneous Data learning**	$(\bar{x}, S)$ γ-estimator

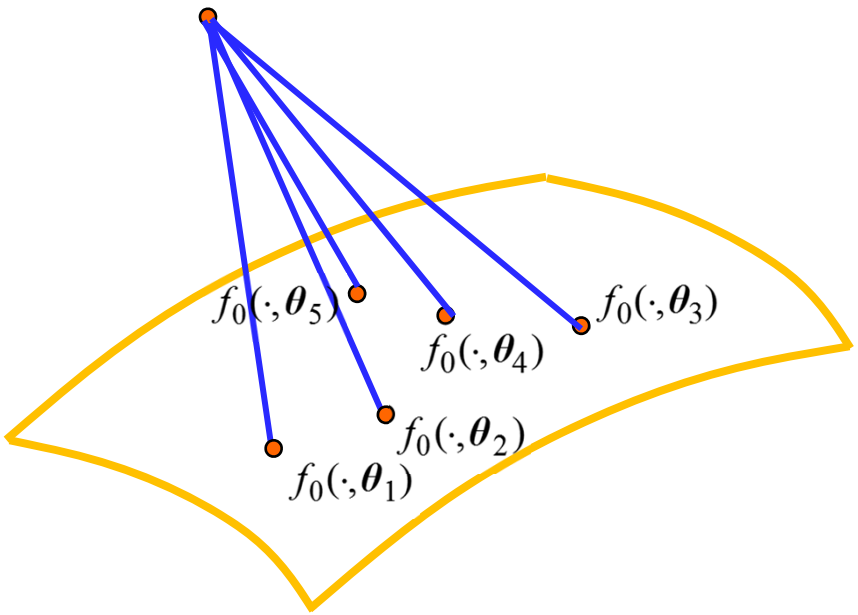
Break for Pythagoras foliation arises SDL

$\text{Leaf}_1 \;\; F(\bar{x}, S) = \{g : \; \mathbb{E}_g(X) = \bar{x}, \; \text{var}(X) = S\}$



$\text{Leaf} \;\; M = \{f_\gamma(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbf{R}^d\}$

$g = \sum_{k=1}^{K} \pi_k f(\cdot, \theta_k)$



$M_0 = \{f_0(\boldsymbol{x}, \boldsymbol{\theta}) := (2\pi)^{-d/2} \exp\{-(\boldsymbol{x} - \boldsymbol{\theta})^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{\theta})/2\} : \boldsymbol{\theta} \in \mathbb{R}^d\}$

Five local minima of $L_\gamma(\boldsymbol{\theta})$

What justification for SDL?

- **SDL is a result from nonconvex learning,**

  cf. CCCP, Yuille & Rangarajan (2002), An & Tao (2005)

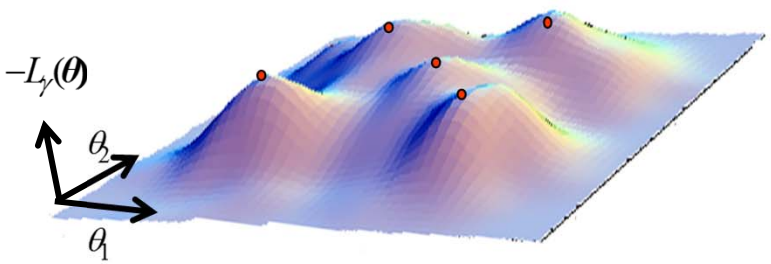
- **SDL is led to by a break of canonical duality between statistical model and estimation**

I like to elucidate a statistical justification for SDL

The break reveals the true distribution?

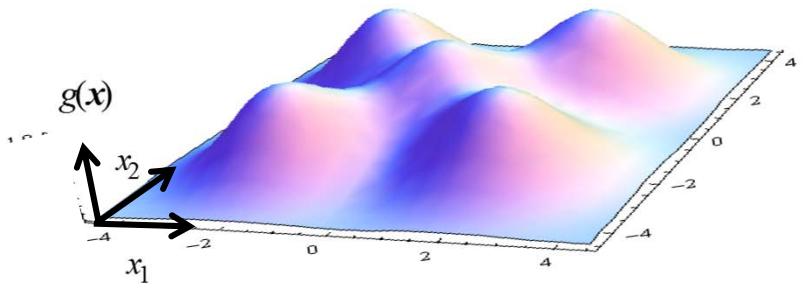
Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_1 \sim g(\boldsymbol{x}) = \sum_{j=1}^{K} \pi_j \phi(\boldsymbol{x}, \boldsymbol{\theta}_j, I)$ (normal mixture)

Then $$-L_\gamma(\boldsymbol{\theta}) \xrightarrow[n\to\infty]{} \mathrm{E}\{-L_\gamma(\boldsymbol{\theta})\} = \sum_{j=1}^{K} \pi_j \phi(\boldsymbol{\theta}, \boldsymbol{\theta}_j, \frac{\gamma+1}{\gamma} I)$$



$n = 500,\ \gamma = 2$

$\cong$



In general $n \to \infty, \gamma \to \infty \Rightarrow -L_\gamma(\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} g(\boldsymbol{\theta})$?

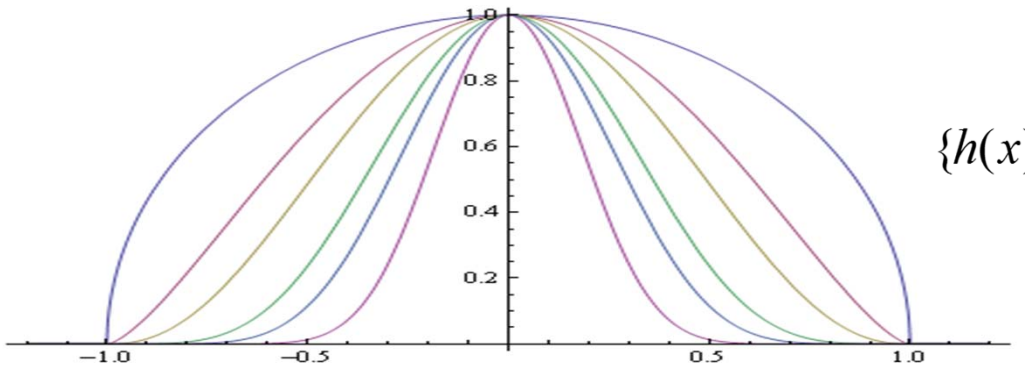
γ-loss for location model

A location model $M = \{ h(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^d \}$

γ-loss function $$\widetilde{L}_\gamma(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{h(\boldsymbol{x}_i - \boldsymbol{\theta})^\gamma}{\int h(\boldsymbol{y})^\gamma \, d\boldsymbol{y}}$$

Note that $-\widetilde{L}_\gamma(\boldsymbol{\theta})$ can be viewed as a density since $\int -\widetilde{L}_\gamma(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = 1$

Semi-circle distribution $h(x) = I(x^2 < 1)\,(1 - x^2)$



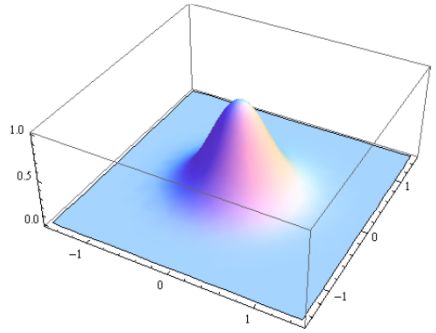
$\{ h(x)^\gamma : \ \gamma = 1, 3, 10, 15, 30 \}$

power

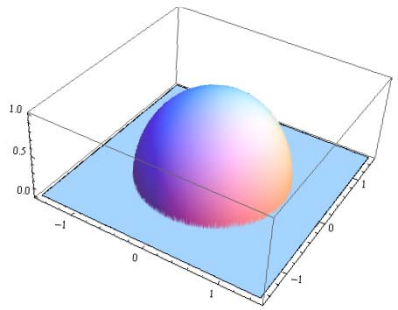
γ-loss as a density estimator

$h(\boldsymbol{x})$

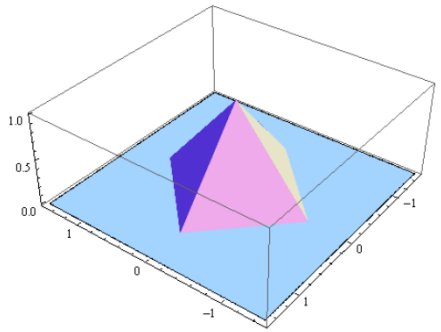
Normal



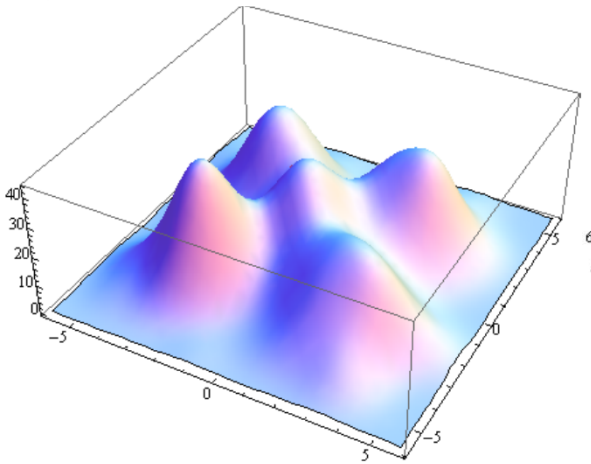
Semi-L_2circle

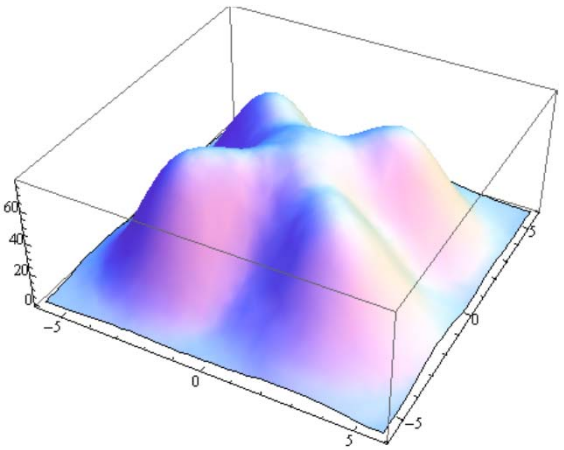


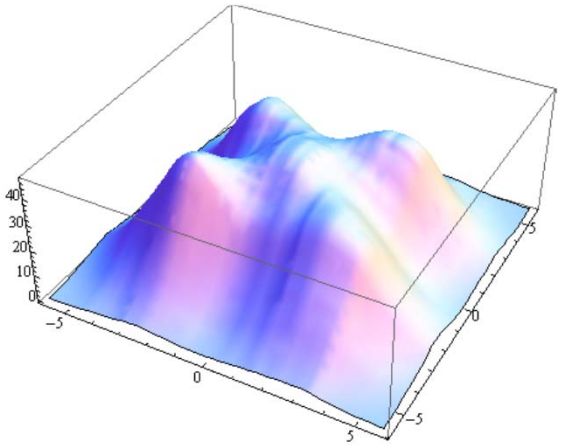
Semi-L_1circle



$-\widetilde{L}_{\gamma}(\boldsymbol{\theta})$







Nonparametric consistency

Theorem Let $g(\boldsymbol{x})$ be a true density function.

We assume : $h(\boldsymbol{x}) < h(0)$ if $x \neq 0$

Then $\lim_{\gamma \to \infty} \mathbb{E}\{\tilde{L}_\gamma(\boldsymbol{\theta})\} = -g(\boldsymbol{\theta})$

Proof.

$$\mathbb{E}\{L_\gamma(\boldsymbol{\theta})\} = -\int \psi_\gamma(\boldsymbol{\theta} - \boldsymbol{y}) g(\boldsymbol{y}) d\boldsymbol{y}, \text{ where } \psi_\gamma(\boldsymbol{x}) = \frac{h(\boldsymbol{x})^\gamma}{\int h(\boldsymbol{y})^\gamma d\boldsymbol{y}}$$

$$\int \psi_\gamma(\boldsymbol{\theta} - \boldsymbol{y}) g(\boldsymbol{y}) d\boldsymbol{y} \xrightarrow[\gamma \to \infty]{} \int \delta(\boldsymbol{\theta} - \boldsymbol{y}) g(\boldsymbol{y}) d\boldsymbol{y} = g(\boldsymbol{\theta})$$

(Dirac's delta function δ).

Mean Integrated Square Error

Theorem. We assume: $\exists$ a strictly decreasing function $\eta(s)$ for $s > 0$ such that $h(\boldsymbol{x}) = \eta(\| \boldsymbol{x} \|)$.

Then
$$\mathrm{MISE}_\gamma = \frac{1}{4}\int \left\{V_\gamma \mathrm{tr}\left(\frac{\partial^2 g(\boldsymbol{y})}{\partial \boldsymbol{y}\partial \boldsymbol{y}^{\mathrm{T}}}\right)\right\}^2 + \frac{1}{n}\frac{\int h(y)^{2\gamma}\,dy}{\left(\int h(y)^{\gamma}\,dy\right)^2}$$

where
$$V_\gamma = \int \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \frac{h(\boldsymbol{y})^{\gamma}}{\int h(\boldsymbol{z})^{\gamma}\,d\boldsymbol{z}}\,d\boldsymbol{y}$$

Kernel density estimator
$$\tilde{g}_h(\boldsymbol{x}) = \frac{1}{nh}\sum_{i=1}^{n} K((\boldsymbol{x}-\boldsymbol{x}_i)/h)$$

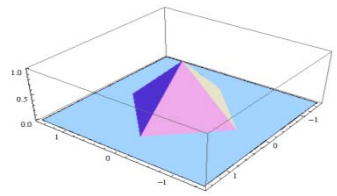
If $h(\boldsymbol{x})$ and $K(\boldsymbol{x})$ are normal, then $L_\gamma(\boldsymbol{x}) = -\tilde{g}_h(\boldsymbol{x})$ with $h = \gamma^{-\frac{1}{2}}$

In general $h_{\mathrm{opt}} = n^{-1/(d+4)}$ cf. Parzen (1962), Li - Racine (2004)

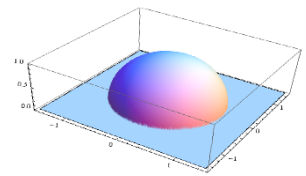
Semicircle distributions

A location model $M = \{ h(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^d \}$

L_1 semicircle distribution $h^{(1)}(\boldsymbol{x}) = \begin{cases} c_d^{(1)} \{1 - \| \boldsymbol{x} \|_{L_1}\} & \text{if} \quad \| \boldsymbol{x} \|_{L_1} \le 1 \\ 0 & \text{otherwise} \end{cases}$



L_2 semicircle distribution $h^{(2)}(\boldsymbol{x}) = \begin{cases} c_d^{(2)} (1 - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}) & \text{if} \quad \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x} \le 1, \\ 0 & \text{otherwise.} \end{cases}$



where $c_d^{(1)}$ and $c_d^{(2)}$ are normalizing constants.

Rem.

The support of $h^{(2)}$ is a d - dimesional hypershere S_d;

the support of $h^{(1)}$ is a d - dimensional hypercube C_d.

$$\mathrm{S}_d \subseteq \mathrm{C}_d$$

$$r_d = \frac{\mathrm{vol}(\mathrm{C}_d)}{\mathrm{vol}(\mathrm{S}_d)} = \frac{d 2^{d-1} \Gamma(d/2)}{\pi^{d/2}} \xrightarrow{d \to \infty} \infty$$

$d = 10, 15 \Rightarrow r_d = 401, 8.5905$ Cf. curse of dimensionality

Mean Integrated Square Error

Theorem.

Let $\mathrm{MISE}_\gamma(h)$ be the mean integrated square error of $L_\gamma(\boldsymbol{x})$ with $h(\boldsymbol{x})$.

Then

$$\mathrm{MISE}_\gamma(h^{(2)}) = \frac{d^2 e^2}{4\gamma^2} \int \left\{ \mathrm{tr}\left(\frac{\partial^2 g(\boldsymbol{y})}{\partial \boldsymbol{y} \partial \boldsymbol{y}^{\mathrm{T}}} \right) \right\}^2 d\boldsymbol{y} + \frac{1}{n} \left(\frac{\gamma}{4e\pi} \right)^{\frac{d}{2}}$$

$$\mathrm{MISE}_\gamma(h^{(1)}) = \frac{1}{4\gamma^4} \left(\sum_{k=1}^{d} \frac{1}{k^2} \right) \int \left\{ \mathrm{tr}\left(\frac{\partial^2 g(\boldsymbol{y})}{\partial \boldsymbol{y} \partial \boldsymbol{y}^{\mathrm{T}}} \right) \right\}^2 d\boldsymbol{y} + \frac{1}{n} \left(\frac{\gamma}{4} \right)^d$$

Kernel density estimator

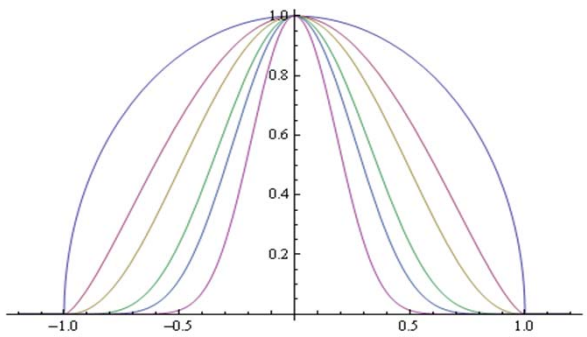
A location model $M = \{ h(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^d \}$

γ-loss function $$\widetilde{L}_\gamma(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{h(\boldsymbol{x}_i - \boldsymbol{\theta})^\gamma}{\int h(\boldsymbol{y})^\gamma \, d\boldsymbol{y}}$$

kernel density estimator $$\hat{g}_b(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b^d} h\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{b} \right)$$

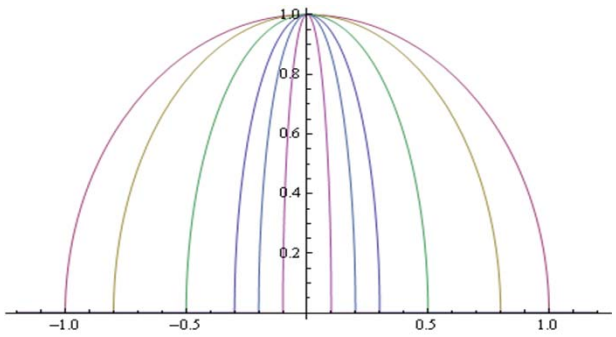
Semi-circle distribution $h(x) = I(x^2 < 1)\,(1 - x^2)$

$\{ h(x)^\gamma : \ \gamma = 1, 3, 10, 15, 30 \}$



power

$\{ h(x/b) : \ b = 1, 0.8, 0.6, 0.3, 0.1 \}$



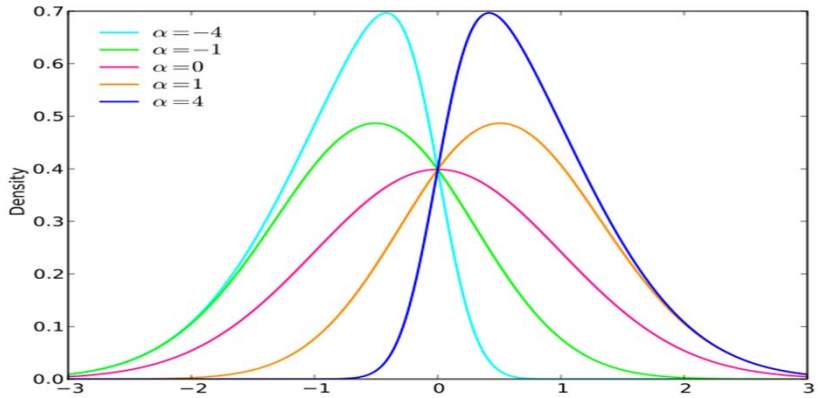
bandwidth

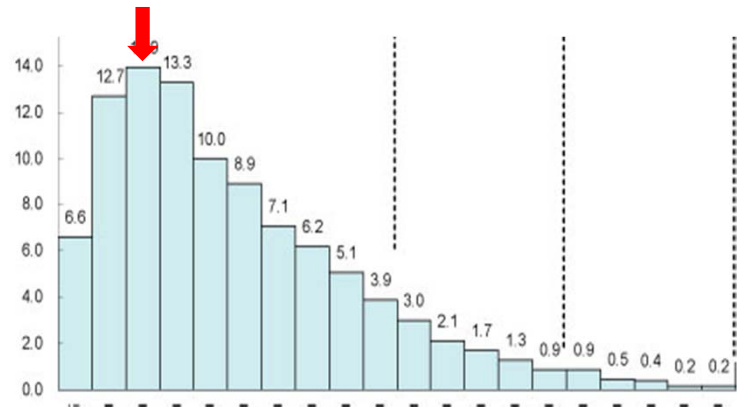
Skew normal distribution

Skew normal distribution $$f(x,\tau,\omega,\alpha)=\frac{2}{\omega}\phi(\frac{x-\tau}{\omega})\Phi(\alpha\frac{x-\tau}{\omega})$$

Normal distribution $$f(x,\tau,\omega,0)=\frac{1}{\omega}\phi(\frac{x-\tau}{\omega})$$

Parameter of interests $$\theta(\tau,\omega,\alpha)=\arg\max_{x\in\mathbb{R}} f(x,\tau,\omega,\alpha)$$





Japanese income data

Estimation for a mode θ

MLE under skew normal

$$\hat{\theta}_{\mathrm{ML}} = \theta(\hat{\tau}, \hat{\omega}, \hat{\alpha}), \text{where } \hat{\tau}, \hat{\omega}, \hat{\alpha} \text{ are MLEs for } \tau, \omega, \alpha.$$

γ-estimator under normal

$$\hat{\theta}_{\gamma} = \arg\min_{\theta \in \Theta} L_{\gamma}(\theta) \quad \text{where} \quad L_{\gamma}(\theta) = -\sum_{i=1}^{n} \exp\{-\frac{\gamma}{2}(x_i - \theta)^2\}$$

MSE

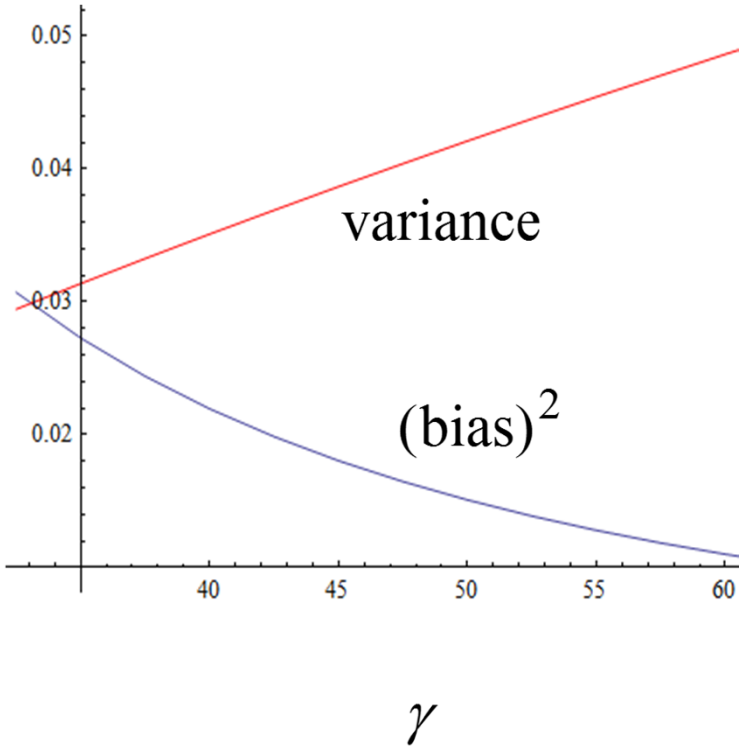
$$\mathrm{MSE}(\theta_{\mathrm{ML}}, \theta) = \frac{1}{n} I(\theta)^{-1} + o(n^{-1}) \quad \text{where} \quad I(\theta) = (\frac{\partial\theta}{\partial\tau}, \frac{\partial\theta}{\partial\omega}, \frac{\partial\theta}{\partial\alpha}) I(\tau, \omega, \alpha) (\frac{\partial\theta}{\partial\tau}, \frac{\partial\theta}{\partial\omega}, \frac{\partial\theta}{\partial\alpha})^{\mathrm{T}}$$

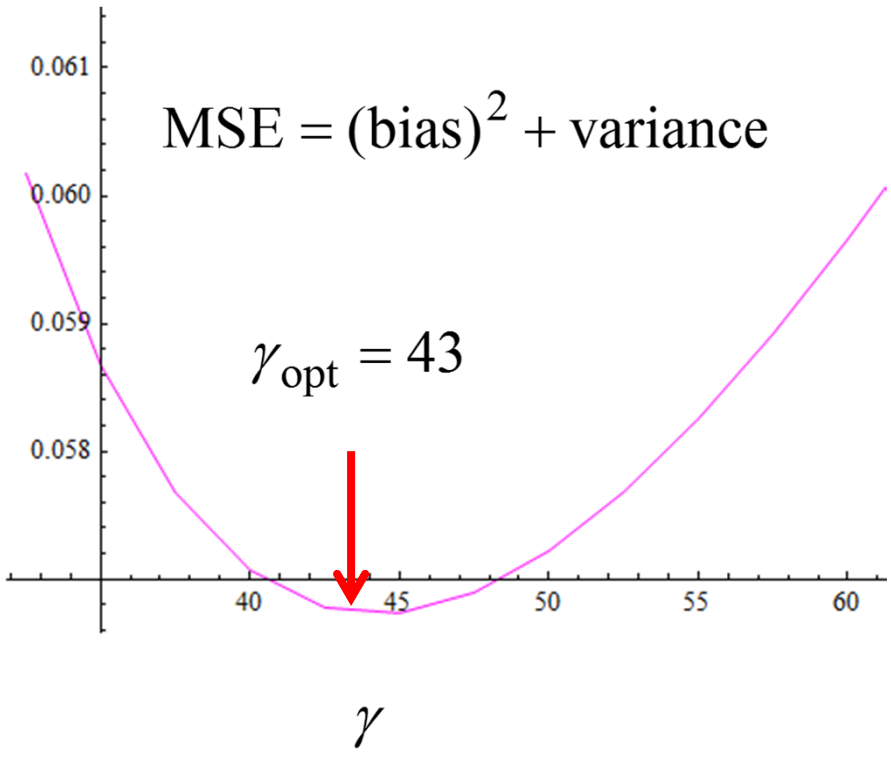
$$\mathrm{MSE}(\theta_{\gamma}, \theta) = \left\{\mathrm{E}(\phi(\theta - X)^{\gamma}) - f(\theta, \tau, \omega, \alpha)\right\}^2 + \frac{1}{n} \mathrm{Var}\left(\frac{\phi(\theta - X)^{\gamma}}{\mathrm{E}\{\phi(\theta - X)^{\gamma}\}}\right)$$

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{ML}}, \theta) < \mathrm{MSE}(\hat{\theta}_{\gamma}, \theta) \text{ for sufficiently large } n$$

Trade between bias and variance

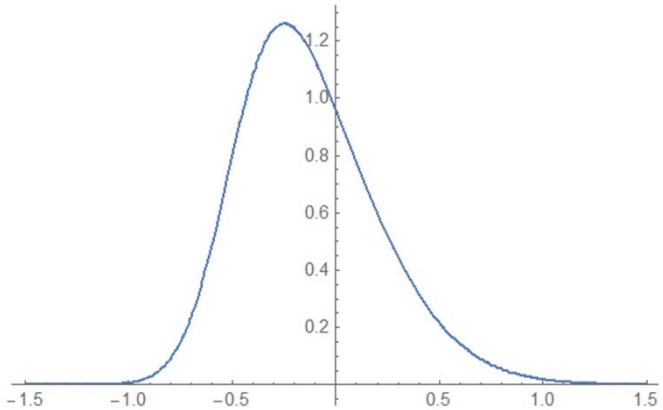
True values $(\tau, \omega, \alpha) = (0.5,\ 0.5,\ 2.5)$ $\theta = -0.248$, $n = 20$





Simulation study

True distribution



$$f(x,\tau_0,\omega_0,\alpha_0) = \frac{2}{\omega_0}\phi\left(\frac{x-\tau_0}{\omega_0}\right)\Phi\left(\alpha_0\frac{x-\tau_0}{\omega_0}\right)$$

$$(\tau_0,\omega_0,\alpha_0) = (0.5,\ 0.5,\ 2.5) \quad \theta_0 = -0.248$$

sample size	$\hat{\theta}_{\mathrm{ML}}$	$\hat{\theta}_{\gamma}$	$\bar{x}$
$n = 20$	2.35514	1.32099	2.02219
$n = 50$	1.41589	0.896704	1.67111
$n = 200$	0.248097	0.459803	1.48877

ML has also SDL?

loss \ model	0-model	γ-model
0-loss	$(\bar{x}, S)$	M-estimator
γ-loss	SDL	$(\bar{x}, S)$

γ-model

$$f_\gamma(\boldsymbol{x},\boldsymbol{\theta},I) = c_\gamma \{ 1 - \tfrac{1}{2}\gamma(\boldsymbol{x}-\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\theta}) \}_+^{\frac{1}{\gamma}}$$

log likelihood

$$L_0(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_\gamma(\boldsymbol{x}_i,\boldsymbol{\theta},I)$$

$$\frac{\partial}{\partial\boldsymbol{\theta}} L_0(\boldsymbol{\theta}) = \sum_{i=1}^{n} f_\gamma(\boldsymbol{x}_i,\boldsymbol{\theta},I)^{-\gamma}(\boldsymbol{x}_i-\boldsymbol{\theta})$$

Rem **If we assume normality, take $\gamma < 0$, then the log likelihood for γ–model has (weak) SDL.**

MLE for Cauchy location model

Cauchy location model

$$f(x,\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2} \qquad (-\infty < x < \infty)$$

Log-likelihood function

$$L(\theta) = -\sum_{i=1}^{n} \log\{1+(x_i-\theta)^2\} \text{ if } \{x_i\}_{i=1}^{n} \underset{\text{iid}}{\sim} f(x,\theta)$$

Likelihood equation

$$\frac{\partial}{\partial\theta} L(\theta) = \sum_{i=1}^{n} f(x_i,\theta)(\theta - x_i) = 0$$

Maximum likelihood

$$\hat{\theta}_{\mathrm{ML}} = \frac{\sum f(x_i,\hat{\theta}_{\mathrm{ML}})x_i}{\sum f(x_i,\hat{\theta}_{\mathrm{ML}})}$$

Fixed-point iteration

$$\hat{\theta}_{t+1} = \frac{\sum f(x_i,\hat{\theta}_t)x_i}{\sum f(x_i,\hat{\theta}_t)}$$

Reeds, Asymptotic number of roots of Cauchy location likelihood equations. Ann. Statist. (1985), 775-784.

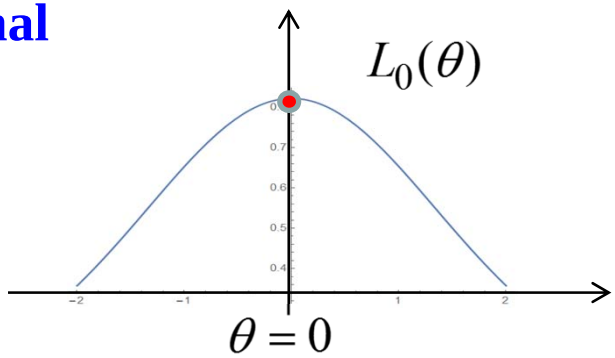
Weak SDL

Expected log-likelihood function

$$L_0(\theta) = -\mathbb{E}(\log\{1+(X-\theta)^2\}) = -\int_{-\infty}^{\infty} \log\{1+(x-\theta)^2\} g(x)\mathrm{d}x \quad \text{if} \quad X \sim g(x)$$

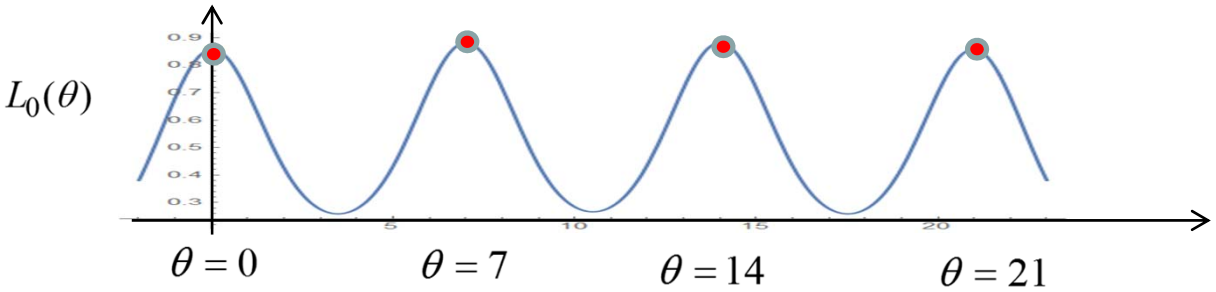
Case 1: true distribution is a normal

$$g(x) \sim N(0,1)$$



Case 2: true distribution is a normal mixture

$$g(x) \sim \tfrac{1}{4}N(0,1) + \tfrac{1}{4}N(7,1) + \tfrac{1}{4}N(14,1) + \tfrac{1}{4}N(21,1)$$



SDL is from the break of max-entropy

loss \ model	0-model	γ-model
0-loss	$(\bar{x}, S)$	M-estimator
γ-loss	γ-estimator	$(\bar{x}, S)$

(γ–model, γ–estimator) leads to a canonical statistics

(0 –model, γ–estimator) leads to strong SDL if $\gamma > 0$

(γ–model, MLE) leads to weak SDL if $\gamma < 0$

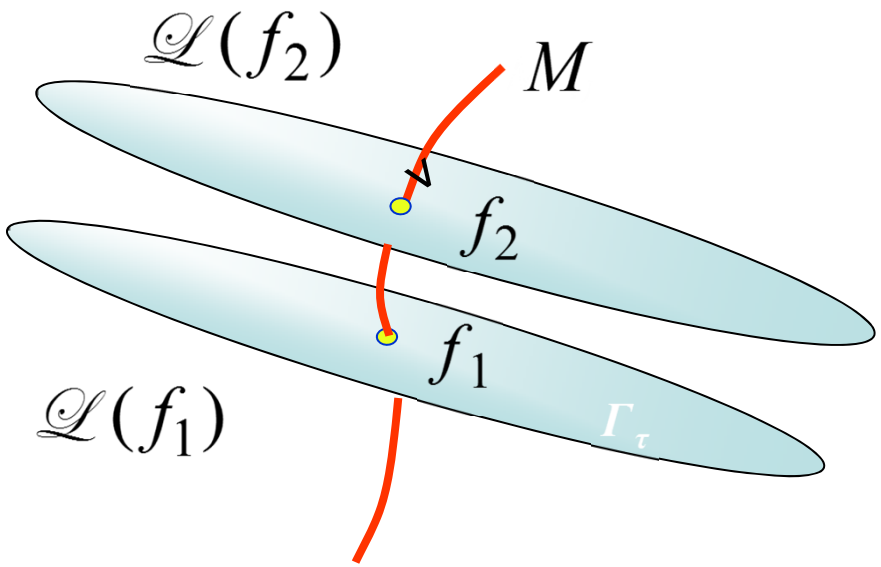
(γ–model, MLE) leads to non-robustness if $\gamma > 0$

(0 –model, γ–estimator) leads to unstable if $\gamma < 0$

Estimator selection vs model selection

Fix a model $M = \{f_\theta : \theta \in \Theta\}$, find a good estimator in $\{\hat{\theta}_\omega : \omega \in \Omega\}$

Fix a loglikelihood $L(M) = \sum \log f_\theta(x_i)$, find a good model in $\{M_k\}_k$



$$\mathscr{L}(f) = \{\, g \in \mathscr{F}_\mu : \hat{\theta}(g) = \hat{\theta}(f) \,\}$$

Future work

- Kernel density estimator vs. power entropy density estimator

  The curse of dimensionality? cf. Huber (1985)

- Informative incomplete data

  The MLE has a fragile property under selective samples.

  What about the γ-estimator ? Any robustness selection bias?

Τηανκ ψου!