Optimal Estimation

J. Rissanen

Helsinki Institute for Information Technology, Technical University of Tampere,
and University of London

11/1/2008

Overview

- Common theory of optimal estimation of both real-valued parameters and their number and structure
- Criterion which generalizes and replaces Fisher's maximum likelihood principle
- No "true" data generating distribution
- Same theory for confidence and interval estimation

Modeling problem

- Data $Y = \{y_t : t = 1, 2, \ldots, n\}$, or
  $Y|X = \{(y_t, x_{1,t}, x_{2,t}, \ldots)\}$, X *explanatory* variables
- Want to learn properties expressed by set of distributions

$$\mathcal{M}_s = \{f(Y|X_s; \theta, s)\}$$

  s structure parameter
  $\theta = \theta_1, \ldots, \theta_{k(s)}$ real-valued parameters
- Optimal selection of **model class** $\mathcal{M}_s$ non-computable, not considered

Examples

In absence of X two general ways:

- Span space of signals by basis vectors $\{w_i^n\}$,

$$\begin{aligned} \hat{y}^n &= \sum_{i \in s} \theta_i w_i^n \\ y^n &= \hat{y}^n + e^n \end{aligned}$$

  $f(e_t; \hat{y}_t, 1)$ gaussian, extend by independence

- Explanatory variables as states $s_t = F(y^t)$; *Markov* model for y^n defined by product of conditionals $f(y_t | s_{t-1}, \theta)$.

Models and estimators

- To simplify notations $Y = X$ (no explanatory variables); also only structures determined by number of real-valued parameters
- Classes of parametric models

$$\begin{aligned} \mathcal{M}_k &= \{f(x^n;\theta,k) : \theta \in \Omega^k\}\ k \leq n \\ \mathcal{M} &= \{\mathcal{M}_k : k = 1,2,\ldots,K, K \leq n\}, \end{aligned}$$

- For $\mathcal{M}_k$ a set $\mathcal{F}_k$ of *estimator* functions

$$\bar{\theta}(\cdot), k : x^n \mapsto \bar{\theta}, k$$

- For $\mathcal{M}$ a set $\mathcal{F}$

$$\bar{\theta}(\cdot), \bar{k}(\cdot) : x^n \mapsto \bar{\theta}, \bar{k}$$

Models or parameters

Important difference:

- θ, k determine model $f(y^n; \theta, k)$ but not conversely
- Both $\theta_1, \ldots, \theta_k$ and $\theta_1, \ldots, \theta_k, 0, \ldots, 0$ determine same model
- Properties of data still represented by model $f(y^n; \bar{\theta}(x^n), \bar{k}(x^n))$ defined by estimated parameters
- **criterion**: function $F(\bar{\theta}(x^n), \bar{k}(x^n))$ of parameters - not of models $f(y^n; \bar{\theta}(x^n), \bar{k}(x^n))$

Truth versus no truth

- Customary estimation based on "truth" $f(x^n; \theta^*, k^*)$
- Want theory without this assumption: two major differences
  1. Cannot measure quality of fit by distance of $\bar{\theta}(x^n), \bar{k}(x^n)$ to "true" parameters θ^*, k^*; different yardstick needed:
     - take $F(\bar{\theta}(x^n), \bar{k}(x^n))$ as probability or density function
       large $F(\bar{\theta}(x^n), \bar{k}(x^n))$ - good fit
       small $F(\bar{\theta}(x^n), \bar{k}(x^n))$ - bad fit
     - equivalently: code length $\log 1/F(\bar{\theta}(x^n), \bar{k}(x^n))$
  2. "truth" assumption superfluous - even harmful because of red herring effect; richer theory without it

Yardstick: probability assigned to data

- For $\mathcal{M}_k$

$$\begin{aligned} \bar{f}(x^n; k) &= \frac{f(x^n; \bar{\theta}(x^n), k)}{\bar{C}_{k,n}} \\ \bar{C}_{k,n} &= \int f(y^n; \bar{\theta}(y^n), k) dy^n \end{aligned}$$

- For $\mathcal{M}$ different construct (clarified later)

$$\begin{aligned} \bar{f}(x^n) &= \frac{\bar{f}(x^n; \bar{k}(x^n))}{\bar{C}_n} \\ \bar{C}_n &= \sum_k \int_{\bar{k}(y^n)=k} f(y^n; \bar{\theta}(y^n), k) dy^n \end{aligned}$$

Traditional theory

- Traditional measure of goodness of estimator of "true" θ
  covariance $V_{\bar{\theta}} = E_\theta(\bar{\theta} - \theta)(\bar{\theta} - \theta)'$
- Cramer-Rao inequality $V_{\bar{\theta}} \geq \min_{\bar{\theta}} V_{\bar{\theta}}$, equality for small subset of $\mathcal{M}_k$ only
- Solution to

$$\min_{\bar{\theta}} \max_{\theta \in \Omega^k} V_{\bar{\theta}}$$

  not *ML* estimator $\hat{\theta}$, except asymptotically.
- Yet *ML* estimator claimed to capture all information in data that can be done by $\mathcal{M}_k$! (true but not because of C-R!)
- **Conclusion:** cannot define optimal estimation by covariance - even of θ let alone also the structure

Partial optimality

- Partial fix in terms of mean KL distance

$$\min_{\bar{\theta}(\cdot)} \max_{\theta} \int f(x^n;\theta,k)\bar{D}(x^n)dx^n$$

$$\bar{D}(x^n) = D(f(Y^n;\theta,k)\|f(Y^n;\bar{\theta}(x^n),k))$$

$$D(p\|q) = \int p(x)\log\frac{p(x)}{q(x)}dx$$

- Solution ML estimator $\hat{\theta}(\cdot)$ - at least asymptotically:

$$\int f(x^n;\theta,k)\hat{D}(x^n)dx^n \to k/2$$

- Does not generalize to estimation of structure k!

Structure estimation by ad hoc criteria

- Multitude of criteria for k; Cross-validation, hypothesis testing etc; a few examples:

$$\min_k \log 1/f(x^n; \hat{\theta}(x^n), k) + k, \;\; AIC,$$

$$\min_k \log 1/f(x^n; \hat{\theta}(x^n), k) + \frac{k}{2} \log n, \;\; BIC,$$

$$\min_k \log 1/ \int f(x^n; \theta, k) w(\theta) d\theta$$

- Which to prefer and why?

Algorithmic theory

- For all strings, length of shortest program with prefix property in universal computer U, *Kolmogorov complexity*,

$$K(x^n) = \min\{|p(x^n)| : U(p) = x^n\}$$

- defines unique best code

$$P(x^n) = \frac{2^{-K(x^n)}}{C}$$

- Kolmogorov complexity **non computable**: cannot tell how close any approximation is
- Cannot apply directly; can imitate idea however

General MDL principle

- "*Find a model with which each data set and the model can be encoded with shortest code length*":

$$\min_{\theta,k} \log 1/f(x^n;\theta,k) + L(\theta,k)$$

- I left selection of $L(\theta,k)$ vague: anything decodable OK
- For decoding prefix requirement, generalized Kraft inequality

$$\sum_k \int 2^{-L(\theta,k)} d\theta \leq 1$$

- Even then the code

$$P_L(x^n) = f(x^n;\bar{\theta}(x^n),\bar{k}(x^n))2^{-L(\bar{\theta}(x^n),\bar{k}(x^n))}$$

incomplete: does not sum to unity.

Encoding of parameters

- MDL method very general: almost anything can be encoded
- Quantization of components

$$\hat{\theta} \mapsto \underline{\theta}$$
$$|\hat{\theta}_i - \underline{\theta}_i| \leq \delta$$

- Optimization with Taylor expansion ($\log 1/\delta$ bits/parameter)

$$\min_{\delta,k} [\log 1/f(x^n; \underline{\theta}, k) + k \log 1/\delta]$$

- Asymptotic criterion same as BIC

$$\min_{k} [\log 1/f(x^n; \hat{\theta}(x^n), k) + \frac{k}{2} \log n]$$

Predictive MDL

- Example: predictor function of past data

$$\hat{y}_t = P(y^{t-1}; \hat{\theta}(y^{t-1}), k)$$

- Cumulative squared (honest) prediction errors

$$S_n = \sum_{t \le n} (y_t - \hat{y}_t)^2$$

- Errors define Gaussian model

$$f(y^n; k) = (2\pi)^{-n/2} e^{-S_n/2}$$

- Criterion: $\min_k \log 1/f(y^n; k) \Leftrightarrow \min_k S_n$
- Asymptotically as good as any criterion for Gaussian models; large initial errors

Estimation capacity

- Estimation **capacity** for real-valued parameters (k fixed)

$$\hat{C}_{k,n} = \max_{\bar{\theta}(\cdot)} \bar{C}_{k,n} = \max_{\bar{\theta}(\cdot)} \int f(y^n; \bar{\theta}(y^n), k) dy^n < \infty$$

- Estimation **capacity** for structure (number of parameters)

$$\begin{aligned} \hat{C}_n &= \max_{\bar{\theta}(\cdot), \bar{k}(\cdot)} \bar{C}_n \\ \bar{C}_n &= \sum_k \int_{\bar{k}(y^n)=k} f(y^n; \bar{\theta}(y^n), k) dy^n / \bar{C}_{k,n} \end{aligned}$$

- Names suggested by Shannon's channel capacity

Shtarkov's *NML* code

- $\log \hat{C}_k \Leftrightarrow \hat{\theta}(\cdot)$ and *NML* code $f(\mathbf{x}; \hat{\theta}(\mathbf{x}), k)/\hat{C}_k$. Why good?
- Justification: closest code to "ideal" target as solution to

$$\min_q \max_{\mathbf{x}} \log \frac{f(\mathbf{x}; \hat{\theta}(\mathbf{x}), k)}{q(\mathbf{x})} = \log \hat{C}_k$$

- Good intuition but formalization fails; minmax argument applies to **all** estimators:

$$\min_q \max_{\mathbf{x}} \log \frac{f(\mathbf{x}; \bar{\theta}(\mathbf{x}), k)}{q(\mathbf{x})} = \log \bar{C}_k$$

  and $\bar{\theta}(\cdot) \neq \hat{\theta}(\cdot) \Rightarrow \log \bar{C}_k < \log \hat{C}_k$
- Why being closest to ideal target better than being closer to some non-ideal one?

Better justification

- $\hat{f}(x^n; k)$ only code in family $\mathcal{F}_k = \{\bar{f}(x^n; k)\}$ satisfying

$$\ln 1/\bar{f}(x^n; k) = \min_{\theta} [\ln 1/f(x^n; \theta, k) + \ln \bar{C}_{k,n}], \ \textit{all } x^n$$

- Let $\bar{\theta} \neq \hat{\theta}$, some x^n (discrete). Then for some $\tilde{\theta}$ and $\Delta > 0$, $f(x^n; \tilde{\theta}, k) = f(x^n; \bar{\theta}, k) + \Delta$ defines better code

$$\bar{f}(x^n; k) < \tilde{f}(x^n; k) = \frac{f(x^n; \bar{\theta}, k) + \Delta}{\bar{C}_{k,n} + \Delta}$$

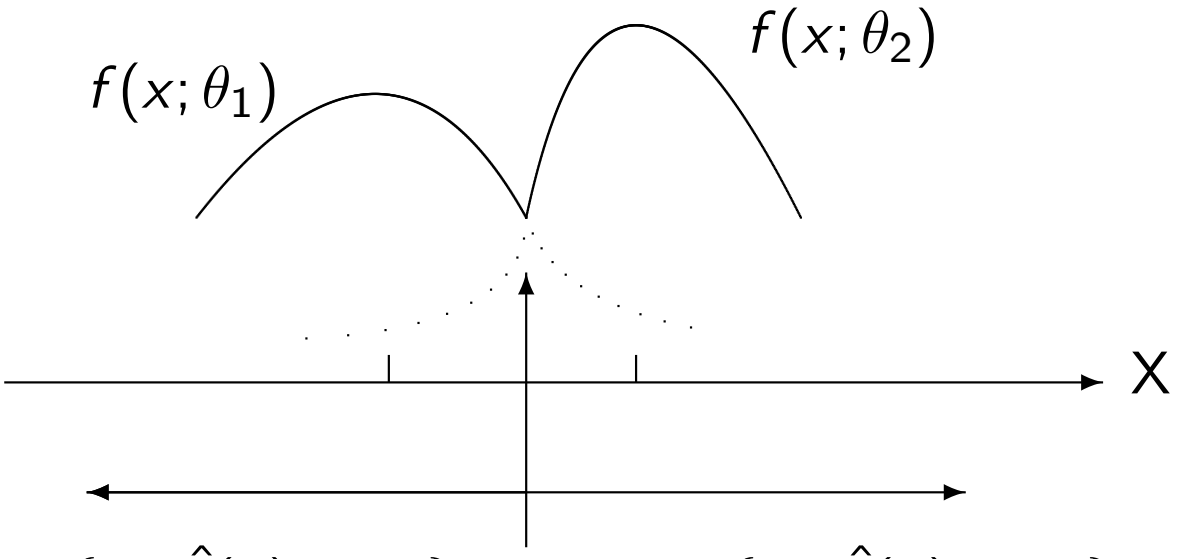
- Conclusion: $\hat{\theta}(x^n)$ and $\hat{f}(x^n; k)$ only code satisfying necessary conditions for optimality for all data

MDL optimality for $\mathcal{M}_k$

- *MDL* optimality: *NML* code $\hat{f}(x^n; k)$ is the only complete code determined by model class such that its codewords cannot be shortened except by making code dependent on some data (tailoring); for optimality at x^n without tailoring one must use $\hat{\theta}(x^n)$
- True even though $\ln 1/\hat{f}(x^n; k)$ not minimized for all x^n over estimators

$$\ln 1/\hat{f}(x^n; k) \neq \min_{\bar{\theta}(\cdot)} \ln 1/\bar{f}(x^n; k)$$

Example: Two models



$B_1 = \{x : \hat{\theta}(x) = \theta_1\}$ $B_2 = \{x : \hat{\theta}(x) = \theta_2\}$

$$\hat{C} = \int f(x;\hat{\theta}(x))dx = \sum_i \int_{B_i} f(x;\theta_i)dx$$

ML estimates maximize capacity and separation $\frac{1}{2}\hat{C}$

complete code $\hat{f}(x) = f(x;\hat{\theta}(x))/\hat{C}$

MDL optimality for $\mathcal{M}$

- Capacity $\hat{C}_n$ iff $\hat{k}(x^n)$, *ML* $\hat{\theta}(x^n)$, and **complete NML** code

$$\hat{f}(x^n;\mathcal{M}) = \frac{f(x^n;\hat{\theta}(x^n),\hat{k}(x^n))/\hat{C}_{\hat{k}(x^n),n}}{\hat{C}_n}$$

$$\hat{k}(x^n) = \mathit{argmax}\ \max_k f(x^n;\hat{\theta}(x^n),k)/\hat{C}_{k,n}$$

- $\hat{f}(x^n;\mathcal{M})$ unique code in family $\{\bar{f}(x^n;\mathcal{M})\}$ satisfying *MDL* requirement: shortest code length for all data without tailoring (necessary conditions for optimality)
- $\ln 1/\hat{f}(x^n;\mathcal{M}) =$ **stochastic complexity**

Properties of capacities

- Capacities $\log \hat{C}_{k,n}$ and $\log \hat{C}_n$: all the information in data about parameters - no parameters excluded or possible properties data could have
- Capacities tend to equalize probabilities defining them: without constraints

$$\max_{\{P_i \leq 1\}} \sum_1^m P_i = m$$

- Each subset $\mathcal{M}_k$ (structure) consists of maximally similar models: having common property
- Can find features in data which we cannot even specify?

Two minmax problems

- Notations

$$\begin{aligned} f_{\theta,k}(x^n) &= f(x^n;\theta,k) \\ \bar{f}_k(x^n) &= f(x^n;\bar{\theta}(x^n),k)/\bar{C}_{k,n} \\ \bar{f}(x^n) &= \bar{f}_{\bar{k}(x^n)}/\bar{C}_n \end{aligned}$$

1. For all $\mathcal{M}_k$ (k not estimated)

$$\min_{\bar{\theta}(\cdot)} \max_{\theta} D(f_{\theta,k} \| \bar{f}_k)$$

2. For $\mathcal{M}$ and all k

$$\min_{\bar{\theta}(\cdot),\bar{k}(\cdot)} \max_{\theta,k} D(f_{\theta,k} \| \bar{f})$$

First optimality

Theorem

1. *Solution to minmax problem for $\mathcal{M}_k$ is $\hat{\theta}(\cdot)$ with NML model $\hat{f}_k(x^n)$. Minmax value is the capacity*
$$\ln \hat{C}_{k,n} = \min_{\bar{\theta}(\cdot)} \max_{\theta} D(f_{\theta,k} \| \bar{f}_k).$$
2. *For all θ and k*
$$\min_{\bar{\theta}(\cdot)} D(f_{\theta,k} \| \bar{f}_k) = \ln \hat{C}_{k,n} - a_{k,n}$$
$$a_{k,n} = E_{\theta,k} \ln \frac{f_{\hat{\theta},k}(X^n)}{f_{\theta,k}(X^n)}.$$

Regret

- Earlier related minmax problem in universal coding for $\mathcal{M}_k$

$$\min_q \max_\theta D(f_{\theta,k} \| q).$$

- Due to its difficulty problem changed into (prior w)

$$\max_w \min_q \int w(\theta) D(f_{\theta,k} \| q) d\theta$$

- For all w, minimizing q is mixture universal model

$$q_w(x^n) = \int f(x^n; \theta, k) w(\theta) d\theta.$$

With maximizing prior: *channel capacity*, the 'regret';
justification: asymptotic code length optimality

Optimality of *MDL* estimators $\hat{\theta}(\cdot), \hat{k}(\cdot)$

Theorem

For all k and n, if $\bar{C}_k \geq 1$, the solution to

$$\min_{\bar{\theta}(\cdot),\bar{k}(\cdot)} \max_{\theta,k} D(f_{\theta,k} \| \bar{f})$$

is $\hat{\theta}(\cdot), \hat{k}(\cdot)$ and $\hat{f}$. The minmax value is $\log \hat{C}_k + \log \hat{C}$

Optimal convergence rate (1984):

Theorem

For all $\bar{\theta}(\cdot), \bar{k}(\cdot)$

$$D(f_{\theta,k} \| \bar{f})/n \geq D(f_{\theta,k} \| \hat{f})/n \to k \log n/(2n)$$

as $n \to \infty$. Inequality holds for all k and θ except some in a set $A_{\theta,k}$ whose volume goes to zero as n grows.

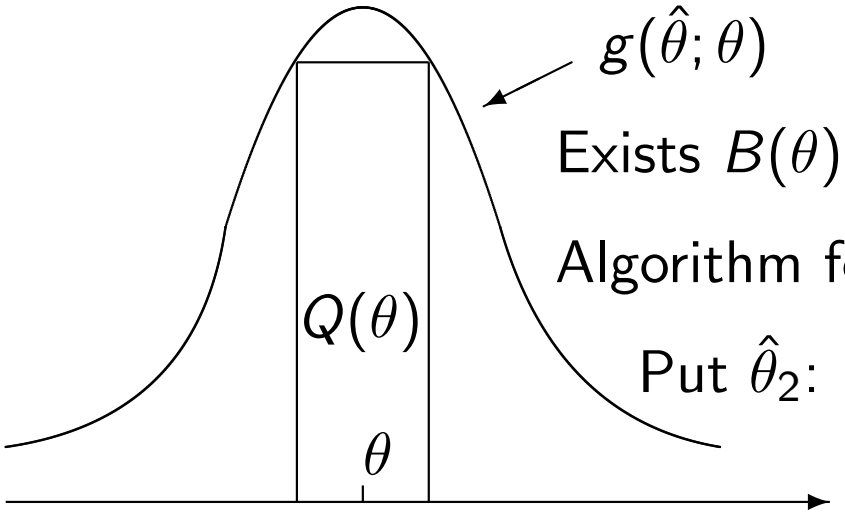
Optimal intervals

- Model classes too rich: do not quite correspond to properties intuitively desired
- For binary strings of length n cannot have more than 2^n properties - far less than size of Bernoulli class
- Dealt with traditionally by intuitive idea of *interval* estimation. Intent to cut off effect of too much "randomness"
- Formally: by estimation capacity and index of separation

Algorithm

Bernoulli family; normal approximations

partitions of Ω, $\Lambda = \{B(\theta_1), \ldots, B(\theta_m)\}$, $f(\mathbf{x}, \theta_i) \Rightarrow g(\hat{\theta}; \theta_i)$



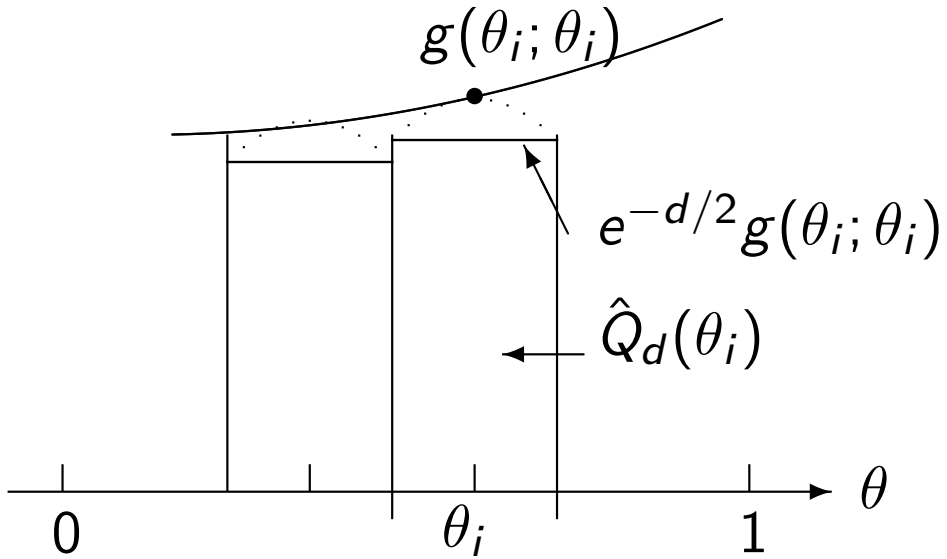
Exists $B(\theta)$ with max area $\hat{Q}(\theta)$, all θ

Algorithm for $\hat{\Lambda}$: Put $\hat{\theta}_1 = 1/2$, $B(\hat{\theta}_1), \hat{Q}(\hat{\theta}_1)$

Put $\hat{\theta}_2$: $B(\hat{\theta}_2), \hat{Q}(\hat{\theta}_2)$, repeat until $B(\hat{\theta}_{\hat{m}})$

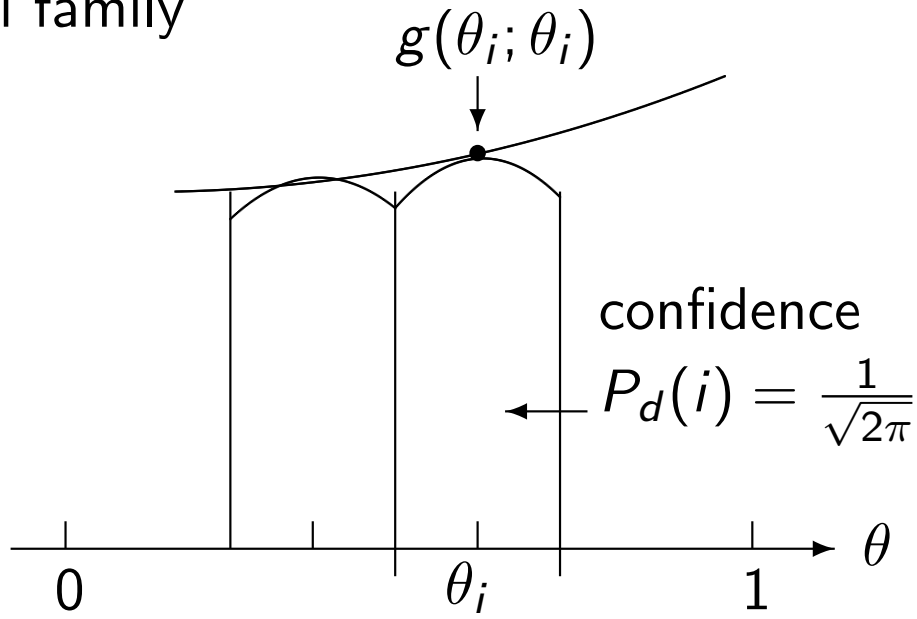
Continue

$|B_d(\theta_i)| = [4d\theta_i(1-\theta_i)/n]^{1/2}$, $g(\theta_i;\theta_i) = [n/(2\pi\theta_i(1-\theta_i))]^{1/2}$
$\hat{Q}_d(\theta_i) = g(\theta_i;\theta_i)|B_d(\theta_i)| = e^{-d/2}\sqrt{2d/\pi}$
capacity $\hat{C}_d = \sum_i \hat{Q}_d(\theta_i) = e^{-d/2}\hat{C}$
separation index $\hat{C}_d/|\Lambda_d| = \hat{Q}_d(\theta_i)$, maximized at $\hat{d} = 1$



Optimal distinguishability

Bernoulli family



$$|\Lambda| = m_d = \sqrt{2/(\pi d)}\hat{C}$$

for $\hat{d} = 1$ family $\{\theta_i\}_1^{m_{\hat{d}}}$ optimally distinguishable

Confidence and capacity

Theorem

1. *The number m_{d_n} of models $f(\mathbf{x};\theta_i,k)$ for $d=d_n$ that can be consistently estimated such that $P_{d_n}(i)\to 1$ cannot exceed $\hat{C}=O(\sqrt{n})$*
2. *If $m_{d_n}\leq \hat{C}^{1-\epsilon}$ (intervals shrink as $O(1/\sqrt{n-\epsilon})$) they can be so estimated*

Conclusions

- A common theory of estimation of both real-valued parameters and their structure
- First time optimal non asymptotic behavior of estimators with criterion defined and established
- No "true" data generating distribution assumed
- Results require basic information and coding theory
- Exact meaning of information in estimation
- Challenge to statisticians: do something comparable without it, or learn information theory