FBST

Full Bayesian Significance Test

Carlos Alberto de Bragança Pereira
cpereira@ime.usp.br

Julio Michael Stern
jstern@ime.usp.br

NOPEF , BIOINFO and IME-USP
University of Sao Paulo

VII School on Regression Models
São Carlos, 12-02-2001

III Continuos Optimization Workshop
Florianopolis, 02-03-2001

Density Function $\quad f(x, \theta)$

Data $x \in S$, the Sample Space,
Parameter Space $\Theta = \{\theta \in R^k \,|\, g(\theta) \leq 0\}$

Given (after we have observed)
the data x, sample size n,

Likelihood Function $f(\theta \,|\, x) = (1/c) f(x, \theta)$
$c = \int_\Theta f(\theta \,|\, x) \, d\theta$

1- Likelihood, Plausibility, Favorable Chances.
2- Wahrscheinlichkeit , Verossimilhança,
Possibility it is Truth, Give Proof, Veracity.

1- Should we consider $f(\theta \,|\, x)$
as a probability function of θ ?
Bayesians (yes) X Frequentists (no)

2- Posterior Function,
given a (Subjective) Prior $g(\theta)$,
$f(\theta \,|\, x) \propto l(\theta \,|\, x) g(\theta)$

Estimation:
What is (my, the) best guess for θ?

1- $\theta^* \in \arg\max f(\theta \,|\, x)$

2- $\theta^* \in \arg\min_{\theta' \in \Theta} \int_\Theta f(\theta' \,|\, x)\, p(\theta', \theta)\, d\theta$

$p(\theta', \theta)$ Loss, penalty or distance function.

ex: $p(\theta', \theta) = ||\theta' - \theta||_p$
$\theta^* =$ Median, for $p = 1$,
$\theta^* =$ Mean, for $p = 2$,
$\theta^* =$ Mode, for $p = \infty$

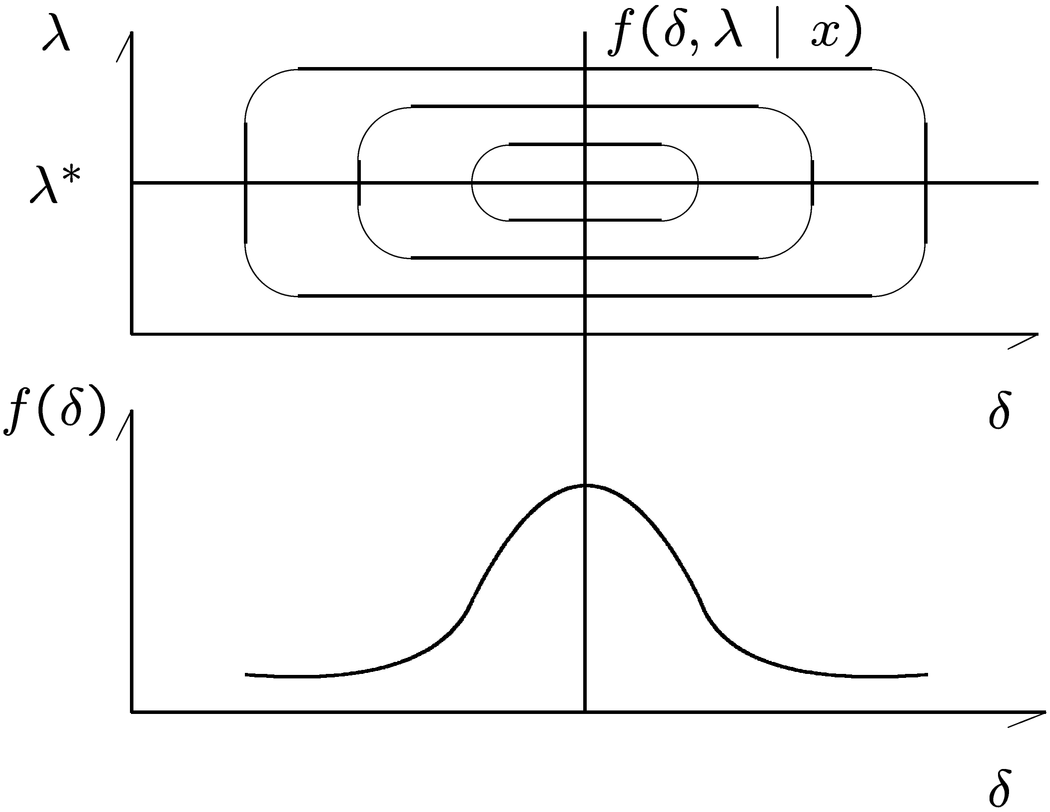
Operators: $\max_\Theta$, $\int_\Theta \, d\theta$

Parameter Space $\Theta = \Delta \times \Lambda$ best δ^* ?

λ are "nuisance" parameters
$\equiv$ we do not care about $\lambda^* \in \Lambda$

nuisance < O.F. nuire
< V.L. nocere (to harm), V.L. necare (to kill)
annoyance < anoye < O.F. enoyer , SP. enojar
< V.L. in odio (to hate)
nonsense < F. non sens
< L. sentire (to feel), G. sinn (meaning)

Eliminate λ
(get rid of the killer paramameters)
How? Ans: Projection Π



Projection Operator, Π:

1- Profile Posterior:
$f(\delta \,|\, x) = f(\delta, \lambda^* \,|\, x)$

2- Marginal Posterior:
$f(\delta \,|\, x) = \int f(\delta, \lambda \,|\, x)\, d\lambda$

Hypothesis Testing:

$\Theta = \Theta_0 \; + \; \Theta_1$

Null, $H_0 : \theta \in \Theta_0$ Alternative, $H_1 : \theta \in \Theta_1$

$$LR = \frac{\max_{\Theta_0} f(\theta \,|\, x)}{\max_{\Theta_1} f(\theta \,|\, x)} = \frac{f(\theta_0^* \,|\, x)}{f(\theta_1^* \,|\, x)}$$

Likelihood Ratio, Not a Probability

Neyman-Pearson Theorem $\Theta = \{\theta_0\} \cup \{\theta_1\}$

$LR \leq k$ criterion minimizes convex combination of errors: 1-reject $H_0 \,|\, \theta_0$, 2-accept $H_0 \,|\, \theta_1$

P-Value, Extreme events probability:

$T = \{x \in S \,|\, LR(x) \leq LR\}$, Tail points

$\alpha = Pr(T \,|\, H_0) \equiv Pr(T \,|\, \tilde{\theta})$, Nominal Level

$\tilde{\theta} = \arg\max_{\Theta_0} Pr(T \,|\, \theta)$

1- α is a probability in the sample space

2- $\forall \theta$, $n \to \infty \;\Rightarrow\; \alpha \to 0$

"Increase sample size to reject"

H_0 Posterior Probability
$\alpha = (1/c) \ \int_{\Theta_0} f(\theta \,|\, x)\, d\theta$

1- Is a probability in Θ
2- If $\theta \in \dot{\Theta}_0$, $n \to \infty \ \Rightarrow \ \alpha \to 1$
3- If $\theta \in \dot{\Theta}_1$, $n \to \infty \ \Rightarrow \ \alpha \to 0$

Odds Ratio

$$OR(H_0\,, x) = \frac{\int_{\Theta_0} f(\theta \,|\, x)\, d\theta \ /c}{\int_{\Theta_1} f(\theta \,|\, x)\, d\theta \ /c}$$

$O(p) = p/(1-p)$

Precise Hypothesis:

$\Theta_0 = \{\theta \in \Theta \mid \; h(\theta) = 0\}$

$\mathrm{dim}(\Theta_0) < \mathrm{dim}(\Theta) \;\; \Leftrightarrow \mathrm{dim}(h) \geq 1$

we may use $\delta = h(\theta)$

$\int_{\Theta_0} 1 d\theta = 0 \Rightarrow OR = 0 \;,\; \forall x$!

Change measure over Θ_0,

$$OR = \quad \frac{\pi}{1-\pi} \quad \frac{\int_{\Theta_0} f(\delta{=}0,\lambda\,,x)g(\delta,\lambda)\;d\lambda}{\int_{\Theta_1} f(\theta\,,x)g(\theta)\;d\theta}$$

1- which mass π ?

2- which measure $d\lambda$ in Θ_0?

natural = easy parameterization of Θ_0 ?

3- $\forall\theta$, $\pi > 0$, $n \to \infty \;\Rightarrow\;$ OR $\to$ 1

"Increase sample size to accept"

Savage's allegory:
King Hiero suspects his goldsmith cheated, using impure gold for his new crown. Arquimedes must measure the crown density, $N(m, s)$, test $m = g$, and advise the king about the hanging.

Epistemological questions:

1- Do precise (sharp) hypothesis really exist?
Instrumental (model) versus
Essential (real) theory.
Galileo's "Epur si muove"

2- How can we "check" a precise hypothesis?

I.J.Good's positions on precise hypothesis:
1- Denial of real sharp tests,
2- Favoring Jeffrey's scheme,
3- Reduction to dichotomies.

1- “Let us consider a null hypothesis that is a (sharp) simple statistical hypothesis H. This is often done is statistical practice, although it would usually be more realistic to lump in with H a small neighborhood of close hypotheses. ... Of course if H is redefined to be a composite hypothesis by including within H a small neighborhood of the sharp null hypothesis, then it becomes possible to obtain much more evidence in favor of H, even without assuming a prior over the components of H. Similarly, if by the truth of Newtonian mechanics we mean that it is approximately true in some appropriate well defined sense we could obtain strong evidence that it is true; but if we mean by its truth that it is exactly true then it has already been refuted. ...

Very often the statistician doesn't bother to make it quite clear whether his null hypothesis is intended to be sharp or only approximately sharp. He also often has in mind an experiment of moderate size but does not state this explicitly. It is hardly surprising then that many Fisherians (and Popperians) say that - you can't get (much) evidence in favor of the null hypothesis but can only refute it."

"Since I regard refutation and corroboration as both valid criteria for this demarcation it is convenient to use another term, Checkability, to embrace both processes. I regard checkability as a measure to which a theory is scientific, where checking is to be taken in both its positive and negative senses, confirming and disconfirming."

2- “My own view on induction is close to that of Jeffreys (1939) in that I think that the initial probability is positive for every self-consistent scientific theory with consequences verifiable in a probabilistic sense. No contradiction can be inferred from this assumption since the number of statable theories is at most countably infinite (enumerable).”

3- “It is very difficult to decide on numerical values for the probabilities, but it is not quite so difficult to judge the ratio of the subjective initial probabilities of two theories by comparing their complexities. This is one reason why the history of science is scientificaly important.”

Text book history is a post-mortem analysis.

When do we have a Precise Hypothesis?

A Scientific Theory, ex:

1- Classical Mechanics, errors in measurements

2- Statistical Physics, too many variables

3- Old (Einstein) Quantum Mechanics,
Unobservable (complex) wave function

4- Modern (Heisemberg) Quantum Mechanics,
No hidden states, von Neuman theorem

The Physical law is exact (precise) in Θ
Measurements in S are inexact, incomplete, unattainable, or essentially uncertain.

Software Controlled Randomized Process

Randomization may be:
- External, exogenous, or

- Internally induced, endogenous
1- Efficient Algorithms
   Simulated Annealing, Genetic, Tabu Search
2- Desirable unpredictability

Ex. of internal randomization applications:
1- Cryptography
2- Automatic Transactions,
   Exchange, Clearing, etc
3- Games and Lotteries

Precise hypothesis may be:
1- Software specifications
2- Functional specifications
3- Normative or Legal Requirements

Test $h(\theta) = 0$

No access to source code
Confidentiality, Outsourcing, Privacy, Security

Data from Black Box Simulation
or Historical Observations

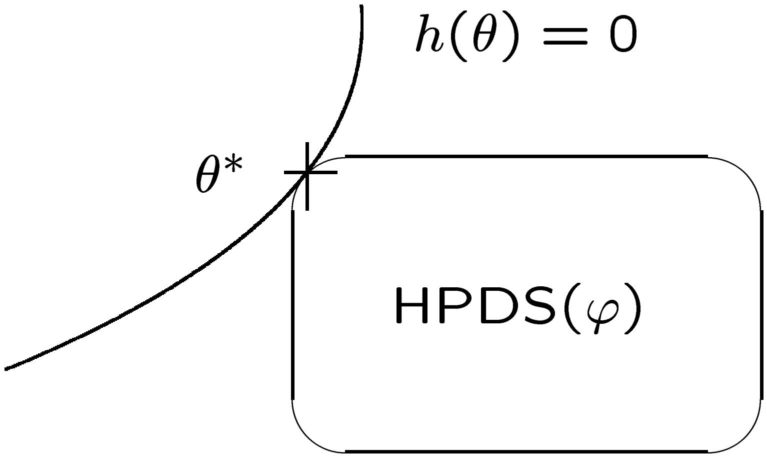
Benefit of the Doubt, Onus Probandi,
Presumption of Innocence, In Dubito Pro Reo
Safe Harbor Liability Rule:

This kind of principle establishes that there is no liability as long as there is a reasonable basis for belief, effectively placing the burden of proof on the plaintiff, who, in a lawsuit, must prove false a defendant's misstatement.

Such a rule also prevents the plaintiff of making any assumption not explicitly stated by the defendant, or tacitly implied by existing law or regulation.

Must consider the plaintif's claim, H_0,
in the most favorable way $\Rightarrow$
$\theta^* \in$ arg max$_{\Theta_0} f(\theta \,|\, x)$, $\varphi = f(\theta^* \,|\, x)$
Can not integrate on Θ_0,
(it is a classic, not a quantic plaintif)

Can only consider as evidence againts H_0
the more likelly alternatives
HPDS$(\varphi) = \{\theta \in \Theta \,|\, f(\theta \,|\, x) \geq \varphi\}$
Tangent Highest Probability Density Set



FBST Operational Definition

$\Theta = \{\theta \in R^n \,|\, g(\theta) \leq 0\}$

$\Theta_0 = \{\theta \in \Theta \,|\, h(\theta) = 0\}$

1- Numerical Optimization step:

$\theta^* \in arg\,\max_{\theta \in \Theta_0} f(\theta)$, $\quad \varphi = f^* = f(\theta^*)$

2- Numerical Integration step:

$\kappa^* = \int_\Theta f_\varphi(\theta \,|\, d) d\theta$

where $f_\varphi(x) = \begin{cases} f(x) \text{ if } f(x) \geq \varphi \\ 0 \text{ otherwise} \end{cases}$

HPDS$(\varphi) = \{x \in \Theta \,|\, f(x) \geq \varphi\}$
HPDS Tangent to Hypothesis (manifold)

FBST properties:

Clear and Simple definition

Ev(H) is a Probability in Θ

"Increase sample size
to get it right, accept / reject"

Testing $\sim$ "Checking" H,
as every scientist always wanted

... but also Compatible with the
Decision Theory framework, as shown by
Madruga, Esteves and Wechsler

Computationally Intensive:
Let Optimization, Num. Analysis
people earn a living :-)

Optimization Problems:

Size: Small–Medium
100 variables , 50 constraints
Large Scale Soft: Inappropriate
Modern Optim. Ideas: Essential

Optimization Techniques:
Variable Metric , Trust Region
Penalty , Merit functions , SQP
Sparsity: No , Structure: Yes
Stable Methods , Factorizations: Yes
Global Optimization: Yes.

Integration Techniques:
Monte Carlo , Importance Sampling

Parallel Computing Techniques:
PVM: Parallel Virtual Machine
Coarse Granularity
Heterogeneous (archit./use) environment

The two parameter Weibull density, reliability (survival) and hazard functions for a failure time $t \geq 0$ shape, and characteristic life (or scale) parameters $\beta > 0$, and $\gamma > 0$, are:

$$\begin{aligned} w(t \,|\, \beta, \gamma) &= (\beta\, t^{\beta-1}/\gamma^{\beta})\, exp(-(t/\gamma)^{\beta}) \\ r(t \,|\, \beta, \gamma) &= exp(-(t/\gamma)^{\beta}) \\ z(t \,|\, \beta, \gamma) &\equiv w(\ )/r(\ ) = \beta\, t^{\beta-1}/\gamma^{\beta} \\ \mu(W) &= \gamma\, \Gamma(1 + 1/\beta) \\ \sigma^2(W) &= \gamma^2(\Gamma(1 + 2/\beta) + \Gamma^2(1 + 1/\beta)) \end{aligned}$$

$\beta = 1$, W is the exponential distribution
$\beta = 2$, W is the Rayleigh distribution
$\beta = 2.5$, W approximates the lognormal
$\beta = 3.6$, W approximates the normal
$\beta = 5.0$, W approximates the peaked normal
Regions $\beta < 1$, $\beta = 1$, $\beta > 1$
decreasing, constant, increasing hazard rates
infant mortality, memoryless, wearout failures

$$X_i \sim w(t\,|\,\beta,\gamma) \;\Rightarrow\; X_{[1,n]} \sim w(t\,|\,\beta,\gamma/n^{1/\beta})$$

Afine transformation $t = t' + \alpha$
three parameter truncated Weibull distribution
location (or threshold) parameter, $\alpha > 0$

$$w(t\,|\,\alpha,\beta,\gamma) = (\beta\,(t+\alpha)^{\beta-1}/\gamma^{\beta})$$
$$exp(-((t+\alpha)/\gamma)^{\beta})/r(\alpha\,|\,\beta,\gamma)$$

$$r(t\,|\,\alpha,\beta,\gamma) = exp(-((t+\alpha)/\gamma)^{\beta})/r(\alpha\,|\,\beta,\gamma)$$

A panel displays 12 to 18 characters. Each character is displayed as a 5×8 matrix of pixels, and each pixel is made of 2 (RG) or 3 (RGB) individual color elements, (like a light emitting diode or gas plasma device). A panel fails when the first individual color element fails. The color elements are "burned in" so they are not at the infant mortality region, i.e. we assume $\beta > 1$.

A lot of panels were purchased as used components, taken from surplus machines. The dealer informed the machines had been operated for a given time, and also informed the mean life of the panels at those machines. Only working panels were acquired. The acquired panels were installed as components on machines of a different type. The use intensity of the panels at each type of machine corresponds to a different time scale, so mean lives are not directly comparable. The shape parameter however is an intrinsic characteristic of the panel. The used time over mean life ratio, $\rho = \alpha/\mu$, is adimensional, and can therefore be used as an intrinsic measure of wearout. We have recorded the time to failure, or times of withdrawal with no failure, of the panels at the new machines, and want to use this data to corroborate (or not) the wearout information provided by the surplus equipment dealer.

Model:

$\Theta = \{(\alpha, \beta, \gamma) \ \in \ ]0, \infty] \times [1, \infty] \times [0, \infty[\ \}$
$\Theta_0 = \{(\alpha, \beta, \gamma) \in \Theta \,|\, \alpha = \rho\mu(\beta, \gamma)\,\}$
$f(\alpha, \beta, \gamma \,|\, D) \propto \prod_{i=1}^{n} w(t_i \,|\, \alpha, \beta, \gamma)$
$\qquad\qquad\qquad\quad \prod_{j=1}^{m} r(t_j \,|\, \alpha, \beta, \gamma)$
data D are failure times, $t_i > 0$,
and the times of withdrawal $t_j > 0$.
Prior $\quad \beta \in [3.0,\ 4.0]$

At the optimization step it is better, for numerical stability, to maximize the log-likelihood:

$wl_i = \log(\beta) + (\beta - 1)\log(t_i + \alpha)$
$\qquad\quad -\beta\log(\gamma) - ((t_i + \alpha)/\gamma)^{\beta} + (\alpha/\gamma)^{\beta}$
$rl_j = -((t_j + \alpha)/\gamma)^{\beta} + (\alpha/\gamma)^{\beta}$
$fl = \sum_{i=1}^{n} wl_i + \sum_{j=1}^{m} rl_j$
$h(\alpha, \beta, \gamma) = \rho\,\gamma\,\Gamma(1 + 1/\beta) - \alpha = 0$

$n = 45$ failure times, in years,

0.01	0.19	0.51	0.57	0.70
0.73	0.75	0.75	1.11	1.16
1.21	1.22	1.24	1.48	1.54
1.59	1.61	1.61	1.62	1.62
1.71	1.75	1.77	1.79	1.88
1.90	1.93	2.01	2.16	2.18
2.30	2.30	2.41	2.44	2.57
2.61	2.62	2.72	2.76	2.84
2.96	2.98	3.19	3.25	3.31

and $m = 5$ withdrawals

1.19	3.50	3.50	3.50	3.50

Evidence for some values of ρ:

ρ	0.05	0.10	0.20	0.30	0.40
e	0.04	0.14	0.46	0.98	1.00

ρ	0.50	0.60	0.70	0.80	0.90
e	0.98	0.84	0.47	0.21	0.01

Multiple Linear Regression Model
$y = X\beta + u \quad , \quad X \; n \times k$
n observations, k independent variables,
β parameters, u Gaussian white noise
$E(u) = 0 \; , \quad \mathrm{Cov}(u) = \sigma^2 I$
diffuse prior $p(\beta, \sigma) = 1/\sigma$
Posterior:

$$
\begin{aligned}
f(\beta, \sigma \,|\, y, X) &= \frac{1}{\sigma^{n+1}} \exp(-\frac{1}{2\sigma^2}(\,(n-k)s^2 \\
&\quad +(\beta - \widehat{\beta})' X' X (\beta - \widehat{\beta})\,)\,) \\
\widehat{\beta} &= (X'X)^{-1} X' y \\
\widehat{y} &= X\widehat{\beta}\,, \\
s^2 &= (y - \widehat{y})'(y - \widehat{y})/(n-k) \\
fl(\beta, \sigma) &= +(\beta - \widehat{\beta})' X' X (\beta - \widehat{\beta})\,) \\
&\quad -(n+1)\log(\sigma) - \frac{1}{2\sigma^2}(\,(n-k)s^2 \\
\frac{\partial\, fl}{\partial\, \beta}(\beta, \sigma) &= -\frac{1}{\sigma^2}(\beta - \widehat{\beta})'\, X' X \\
\frac{\partial\, fl}{\partial\, \sigma}(\beta, \sigma) &= -\frac{n+1}{\sigma} + \frac{1}{\sigma^3}(\,(n-k)s^2 \\
&\quad +(\beta - \widehat{\beta})' X' X (\beta - \widehat{\beta})\,)
\end{aligned}
$$

Polynomial Regression of Order k
$X_{i,j} = (x_i)^j$, $i = 1 \ldots n$, $j = 0 \ldots k$
dimension $d = k + 2$
parameters $\beta_0, \beta_1, \ldots \beta_k$, and σ

Model Selection and Regularization
$R_{pen} = r(d, n) * R_{emp}$
Regularization (or penalization) function
$r(d, n)$, using $p = (d/n)$

Akaike's f.pred.er.	$\text{fpe} = (1 + p)/(1 - p)$
Schartz' Bay. crt.	$\text{sbc} = 1 + \ln(n) p/(2 - 2p)$
Gen. cross valid.	$\text{gcv} = (1 - p)^{-2}$
Shibata mdl. slct.	$\text{sms} = 1 + 2p$

FBST "Occham razor" selection criterion:
Pluralitas non est ponenda sine necessitas:
The model shall not include a parameter unless there is strong evidence that it is not null

FBST selection criterion has a intrinsic regularization mechanism

example: d-dim. $x \,|\, \beta \ \sim \ N(\beta, I)$
FBST for $H: \beta_1 = 0$
Prior $\beta \ \sim \ N(0, I)$
Posterior for $\beta \ \sim \ N(x/2, (1/2)I)$
Weight of HPDS tangent to $H: \beta_1 = 0$ is

$$
\begin{aligned}
\kappa^* &= Pr\{\chi^2_d \leq x_1^2/2\} \\
&= \mathrm{Pr}\left\{ \frac{\chi^2_d}{\sqrt{2d}} - \sqrt{\frac{d}{2}} \ \leq \ t \ \right\} \\
t &\equiv \frac{x_1^2}{2\sqrt{2d}} - \sqrt{\frac{d}{2}} \\
d &\rightarrow \infty \ \ \text{Central Limit Theorem} \\
\kappa^* &\approx \Phi\,(t) \rightarrow 0 \ , \quad \text{evid}(H) \rightarrow 1
\end{aligned}
$$

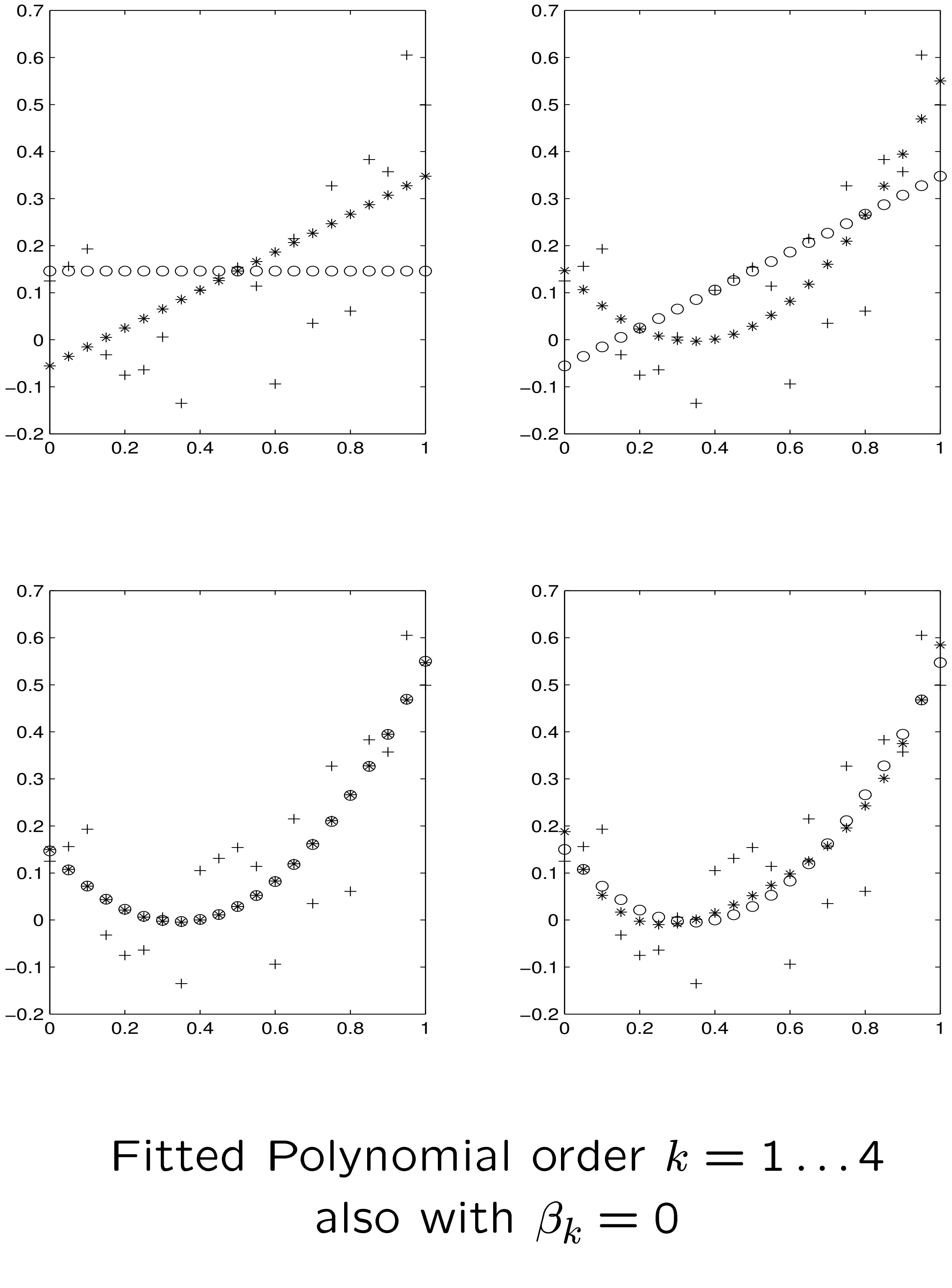
Intrinsic Regularization is explained
by Simple Geometry Related to Symmetry
N is Spherically (Elliptically) Symmetric
(scaled) Invariance for Orthogonal Group
whose invariant metric is $|| \; ||_2$
whereas the unitary volume in R^d is
defined by a cube, a sphere $|| \; ||_\infty$
Model's invariant sphere has comparatively
"small volume in high dimension"

$$
\begin{aligned}
\mathrm{Vol}(S_d) &= (2/d)\pi^{d/2}/\Gamma(d/2) \\
\mathrm{Vol}(C_d) &= 2^d \\
\frac{\mathrm{Vol}(S_d)}{\mathrm{Vol}(C_d)} &= \frac{2}{d}\left(\frac{\pi}{4}\right)^{d/2} \Big/ \Gamma\left(\frac{d}{2}\right)
\end{aligned}
$$

Sakamoto's $n = 21$ points simulation of
$y_i = \text{exp}((x_i - 0.3)^2) - 1 + 0.1 * N(0, 1)$
$y = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots + \beta_k x^k + N(0, \sigma I)$
EMP $= ||y - \widehat{y}||_2^2 / n$ and regularizations
for models of order $k = 0, \ldots 5$
+AIC, +FBST $\quad H:\ \beta^k = 0$

k	EMP	FPE	SBC	GCV	SMS	-AIC	FBST
0	3712	4494	4307	4535	4419	0725	0.00
1	2223	2964	2787	3025	2858	2035	0.00
2	1130	1661	1534	1724	1560	3213	0.00
3	1129	1835	1667	1946	1667	3080	1.00
4	1088	1959	1751	2133	1710	2979	0.99
5	1087	2173	1913	2445	1811	2786	1.00

$*1E??$



Fitted Polynomial order $k = 1 \ldots 4$
also with $\beta_k = 0$

Sakamoto's model for 20 or 39 cities in Japan
y, average daily minimum January temper.
$x_{\bullet,j}$, latitude, altitude and longitude

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + N(0, \sigma I)$$

k	EMP	FPE	SBC	GCV	SMS	AIC	FBST
0	3266	3992	3810	4032	3919	1305	0.00
1	0873	1182	1104	1209	1135	1065	0.00
2	0247	0370	0339	0385	0345	0828	0.05
3	0246	0410	0369	0438	0369	0848	1.00

Sakamoto gives a fine analysis of the example, "detecting" a problem with $k = 2$, due to insufficient (20) data at high altitude.
With more (39) data, $FBST = 0.00$.