Bayesian Regression of Piecewise Constant Functions

Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

Valencia ISBA, 1–6 June 2006

Table of Contents

Abstract

I derive an exact and efficient Bayesian regression algorithm for piecewise constant functions of unknown segment number, boundary location, and levels. It works for any noise and segment level prior, e.g. Cauchy which can handle outliers. I derive simple but good estimates for the in-segment variance. I also propose a Bayesian regression curve as a better way of smoothing data without blurring boundaries. The Bayesian approach also allows straightforward determination of the evidence, break probabilities and error estimates, useful for model selection and significance and robustness studies. I present an application to microarray-CGH data analysis. Many possible extensions will be discussed.

Keywords: Bayesian regression, exact polynomial algorithm, non-parametric inference, piecewise constant function, dynamic programming, application, microarray-CGH data.

Advantages of Bayesian Regression

- Very principled, hence involves less heuristic design choices.
- Important for estimating the number of segments.
- One can decide among competing models solely on evidence.
- Bayes often works well in theory and practice.
- Probability estimates and variances for quantities of interest.
- Bayesian regression curve (better than local smoothing which wiggles more and blurs jumps).

Setup / Likelihood

True function $f = (f_1, ..., f_n)$

has k segments with boundaries

$0 = t_0 < t_1 < ... < t_{k-1} < t_k = n$,

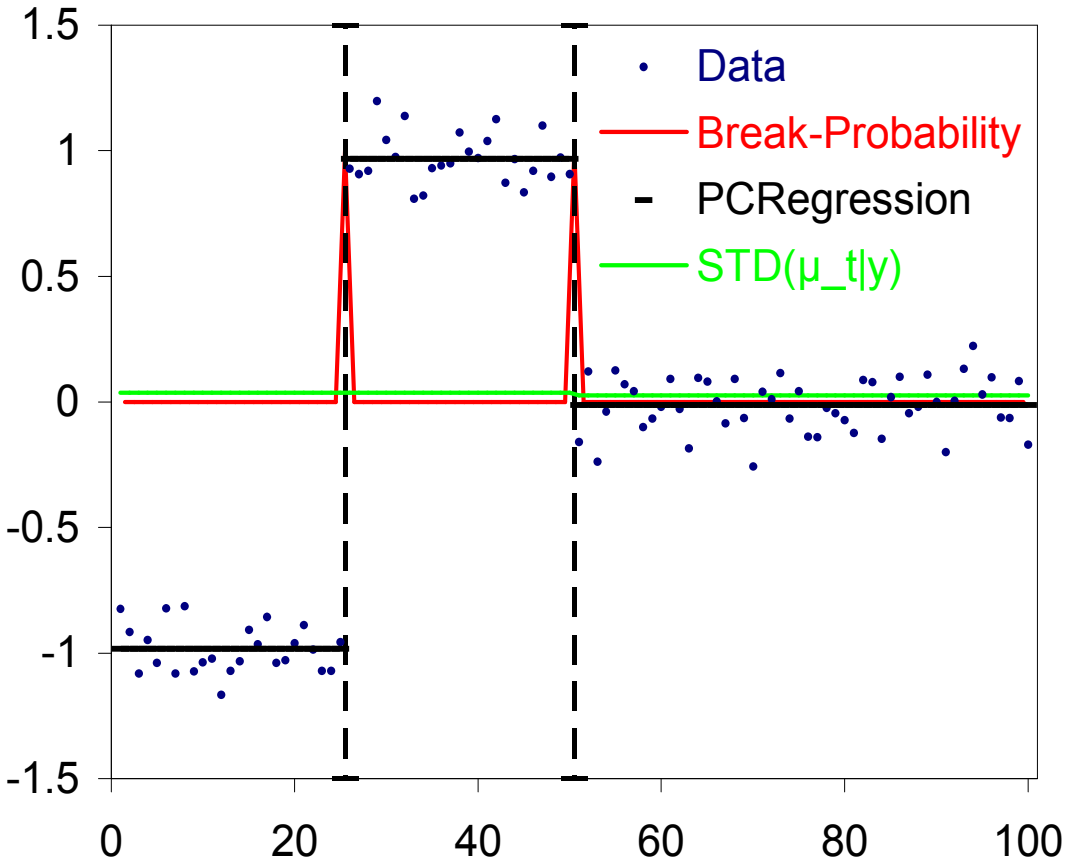
i.e. f is const. on $\{t_{q-1}+1, .., t_q\}$

for each $0 < q \leq k$.

Noisy observations $\boldsymbol{y} = (y_1, ..., y_n)$.

Any independent noise with

mean μ_q and variance σ^2.



$$\Longrightarrow \quad \text{Likelihood:} \quad P(\boldsymbol{y}|\boldsymbol{\mu}, \sigma) = \prod_{q=1}^{k} \prod_{i=t_{q-1}+1}^{t_q} P(y_i|\mu_q, \sigma)$$

Bayesian Regression

Goal: Estimate

– segment levels $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)$,

– boundaries $\boldsymbol{t} = (t_0, ..., t_k)$,

– and their number k.

Bayesian regression: Assume prior $P(\boldsymbol{\mu}, \boldsymbol{t}, k)$

Compute posterior: $P(\boldsymbol{\mu}, \boldsymbol{t}, k|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{t}, k)P(\boldsymbol{\mu}, \boldsymbol{t}, k)}{P(\boldsymbol{y})}$

Evidence: $P(\boldsymbol{y}) = \sum_{k,\boldsymbol{t}} \int P(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{t}, k)P(\boldsymbol{\mu}, \boldsymbol{t}, k)\, d\boldsymbol{\mu}$

Too complex: We need summaries like mean or MAP.

Prior

- We model the level of each segment by a broad (e.g. Gaussian) distribution $P(\mu_q|\nu,\rho)$
- Uniform distribution among all segmentations into k segments: $P(\boldsymbol{t}|k) = \binom{n-1}{k-1}^{-1}$
- Uniform prior over segment number k: $P(k) = 1/n$.

$\Longrightarrow$ Prior: $P(\boldsymbol{\mu},\boldsymbol{t},k) = \prod_{q=1}^{k} P(\mu_q|\nu,\rho) \times P(\boldsymbol{t}|k) \times P(k)$

- (ρ,ν,σ) are fixed hyper-parameters determined later.

Quantities of Interest

Segments: $\hat{k} = \arg\max_k P(k|\boldsymbol{y})$

Boundaries: $\hat{t}_q = \arg\max_{t_q} P(t_q|\boldsymbol{y},\hat{k})$

Segment level: $\hat{\mu}_q = \mathbf{E}[\mu_q|\boldsymbol{y},\hat{\boldsymbol{t}},\hat{k}] = \int P(\mu_q|\boldsymbol{y},\hat{\boldsymbol{t}},\hat{k})\mu_q d\mu_q$

The estimate $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{t}},\hat{k})$ defines a (single) piecewise constant (PC) function $\hat{f}$, which is our estimate of f.

A (very) different quantity is to Bayes-average over all piecewise constant functions and to ask for the mean at location i as an estimate for f_i:

Regression curve: $\hat{\mu}'_i = \mathbf{E}[\mu'_i|\boldsymbol{y}] = \int P(\mu'_i|\boldsymbol{y})\mu'_i d\mu'_i$

Dynamic Programming

- Dynamic programming: Fix a break, then data left and right of the break are independent.
- Evidence and moments of single segment from $i+1$ to j.

  $A^r_{ij} := \int P(\mu_m) \prod_{t=i+1}^{j} P(y_t|\mu_m) \mu_m^r d\mu_m$
- Analytical for exponential family with conjugate prior like Gauss and numerically for others like Cauchy.
- $L_{kj} :\propto P(y_1..y_j|k)$ of first j data, given k segments.
- $R_{ki} :\propto P(y_{i+1}..y_n|k)$ of last $n-i$ data, given k segments.
- Left recursion: Evidence of $y_1..y_j$ with $k+1$ segments $=$ evidence of y_{0h} with k segments $\times$ single-segment evidence of $y_{h+1}..y_j$, summed over all locations h of boundary k:
  $L_{k+1,j} = \sum_{h=k}^{j-1} L_{kh} A^0_{hj}$
- Similarly: Right recursion: $R_{k+1,i} = \sum_{h=i+1}^{n-k} A^0_{ih} R_{kh}$

Efficient Solutions for Quantities of Interest

Evidence $P(\boldsymbol{y}) = \sum_{k=1}^{n} P(\boldsymbol{y}|k)P(k) = \frac{1}{n}\sum_{k=1}^{n} L_{kn}/\binom{n-1}{k-1}$

The posterior of k and its MAP estimate are

$$P(k|\boldsymbol{y}) = \frac{L_{kn}}{\binom{n-1}{k-1}k_{max}E} \quad \text{and} \quad \hat{k} = \operatorname*{arg\,max}_{k=1..k_{max}} P(k|\boldsymbol{y})$$

Prob. that boundary p located at h is $P(t_p{=}h|\boldsymbol{y},\hat{k}) = L_{ph}R_{\hat{k}-p,h}/L_{\hat{k}n}$

MAP segment boundary p is $\hat{t}_p := \operatorname{arg\,max}_h P(t_p = h|\boldsymbol{y},\hat{k})$

Segment level moments are $\widehat{\mu_p^r} = A^r_{\hat{t}_{p-1}\hat{t}_p}/A^0_{\hat{t}_{p-1}\hat{t}_p}$

Regression curve: Fix single segment $t_{m-1} = i, .., t_m = j$ containing t, then $\mu'_t = \mu_m$. Now sum over all such segments:

$$\widehat{\mu'^r_t} = \sum_{i<t\leq j} \frac{1}{L_{\hat{k}n}} \sum_{m=1}^{\hat{k}} L_{m-1,i} A^r_{ij} R_{\hat{k}-m,j}$$

Determination of the Hyper-Parameters

- Global variance ρ and mean ν of $\boldsymbol{\mu}$, in-segment variance σ.
- (Empirical) Bayes: Averaging or maximizing $P(\boldsymbol{y}|\sigma,\nu,\rho)$ is expensive.
- Fast semi-principled estimation
  Global mean $\hat{\nu} \approx \frac{1}{n}\sum_{t=1}^{n} y_t$
  Global variance $\hat{\rho}^2 \approx \frac{1}{n-1}\sum_{t=1}^{n}(y_t-\hat{\nu})^2$
- In-segment variance σ more tricky without knowing segmentation:
  $\hat{\sigma}^2 \approx \frac{1}{2(n-1)}\sum_{t=1}^{n-1}(y_{t+1}-y_t)^2$
  – Good for large noise.
  – Crude estimate is enough if noise is low (regression easy).

Quartiles for Heavy-Tailed Robust Distributions

Let $[\boldsymbol{y}]$ be the data vector $\boldsymbol{y}$ sorted in ascending order.

Global median $\hat{\nu} \approx [\boldsymbol{y}]_{n/2}$

Global scale $\hat{\rho} \approx \frac{[\boldsymbol{y}]_{3n/4} - [\boldsymbol{y}]_{n/4}}{2\alpha}$ with $\alpha \approx 1$

Differences $\Delta_t := y_{t+1} - y_t$.

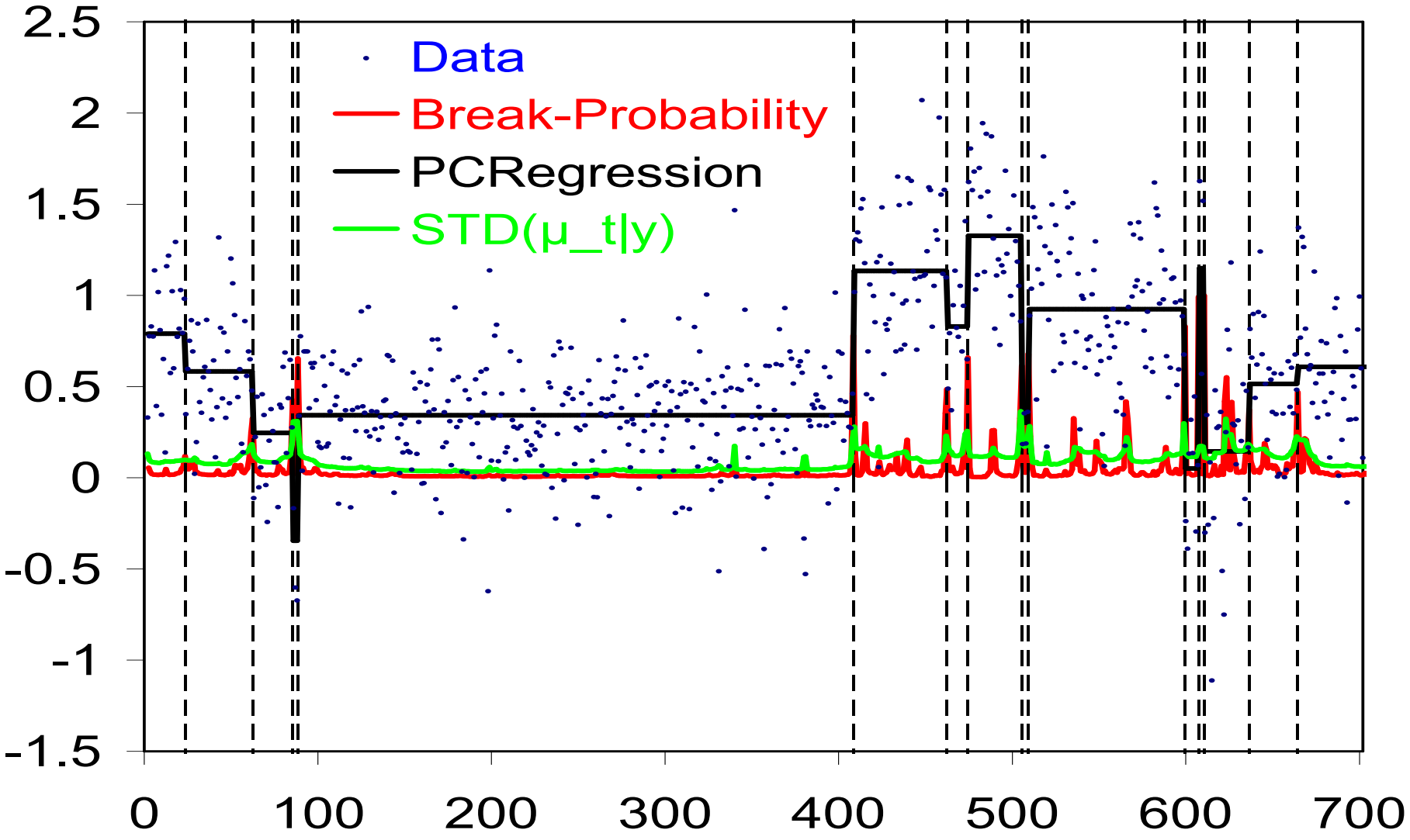
In-segment scale $\hat{\sigma} \approx \frac{[\boldsymbol{\Delta}]_{3n/4} - [\boldsymbol{\Delta}]_{n/4}}{2\beta}$ with $\beta \approx 12$

Iteratively improve them, if the estimates are really not sufficient.

Example: Gene Copy # Data

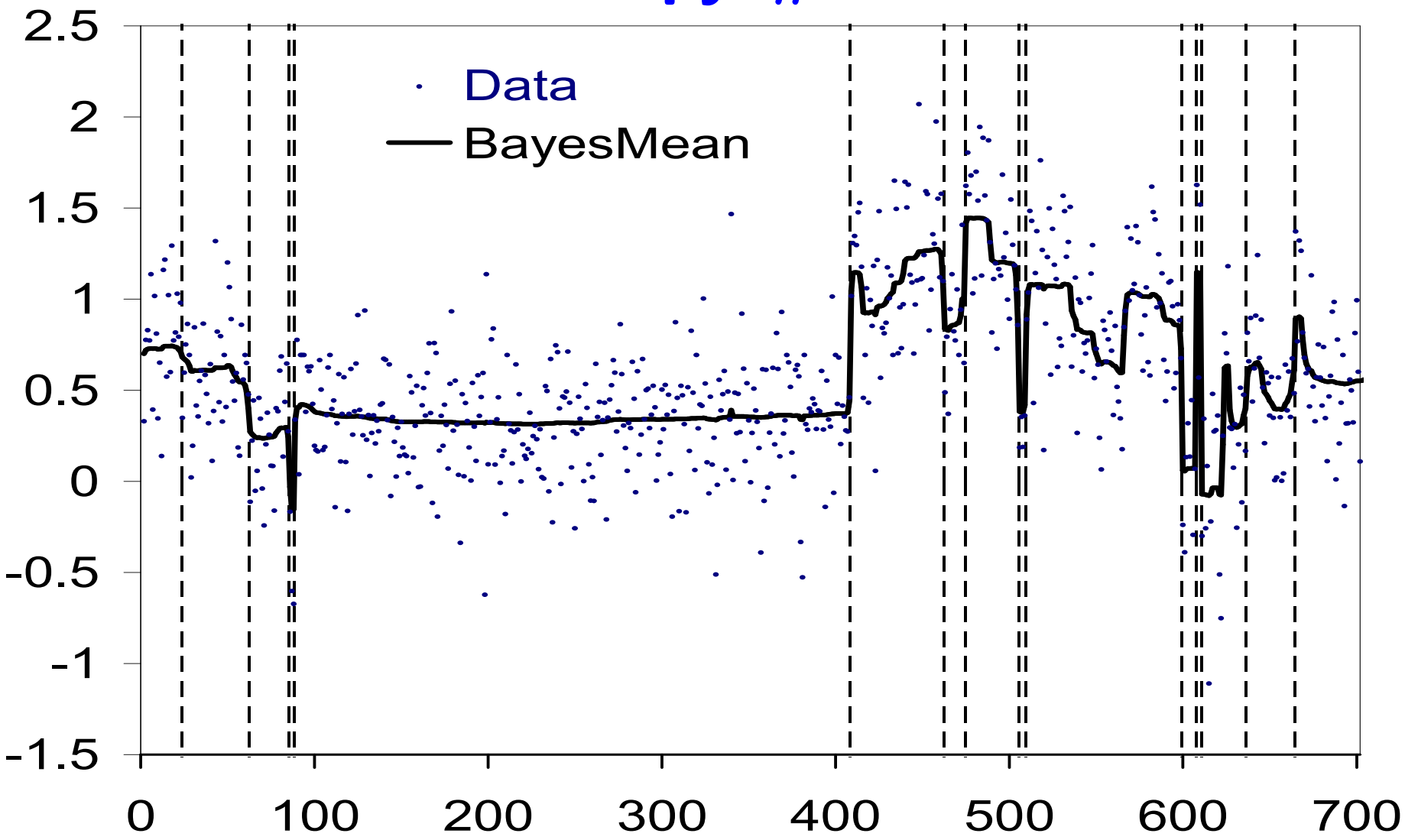
- All genes in a healthy human cell come in pairs, but can be lost or multiplied in tumor cells.
- With modern micro-arrays one can measure the copy-number of genes along a chromosome.
- It is important to determine the breaks, where copy-number chances.
- The measurements are very noisy [Pinkel'98].
- Hence this is a natural application for piecewise constant regression of noisy (one-dimensional) data.
- Regression results of one aberrant and one healthy chromosome (without biological interpretation) are shown ...

Aberrant Gene Copy # of Chromosome 1



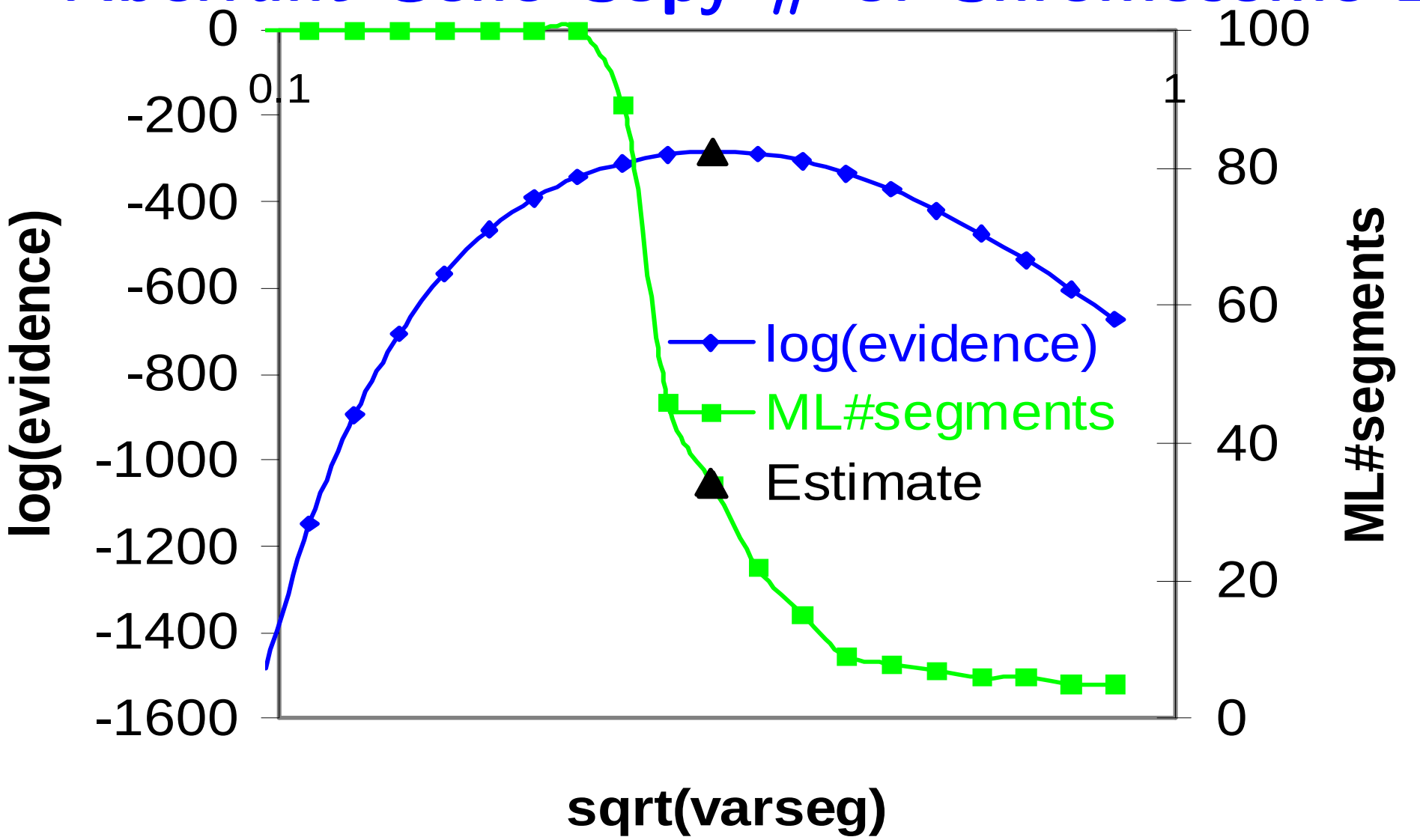
Data (blue), PCR (black), BP (red), and variance$^{1/2}$ (green).

Aberrant Gene Copy # of Chromosome 1



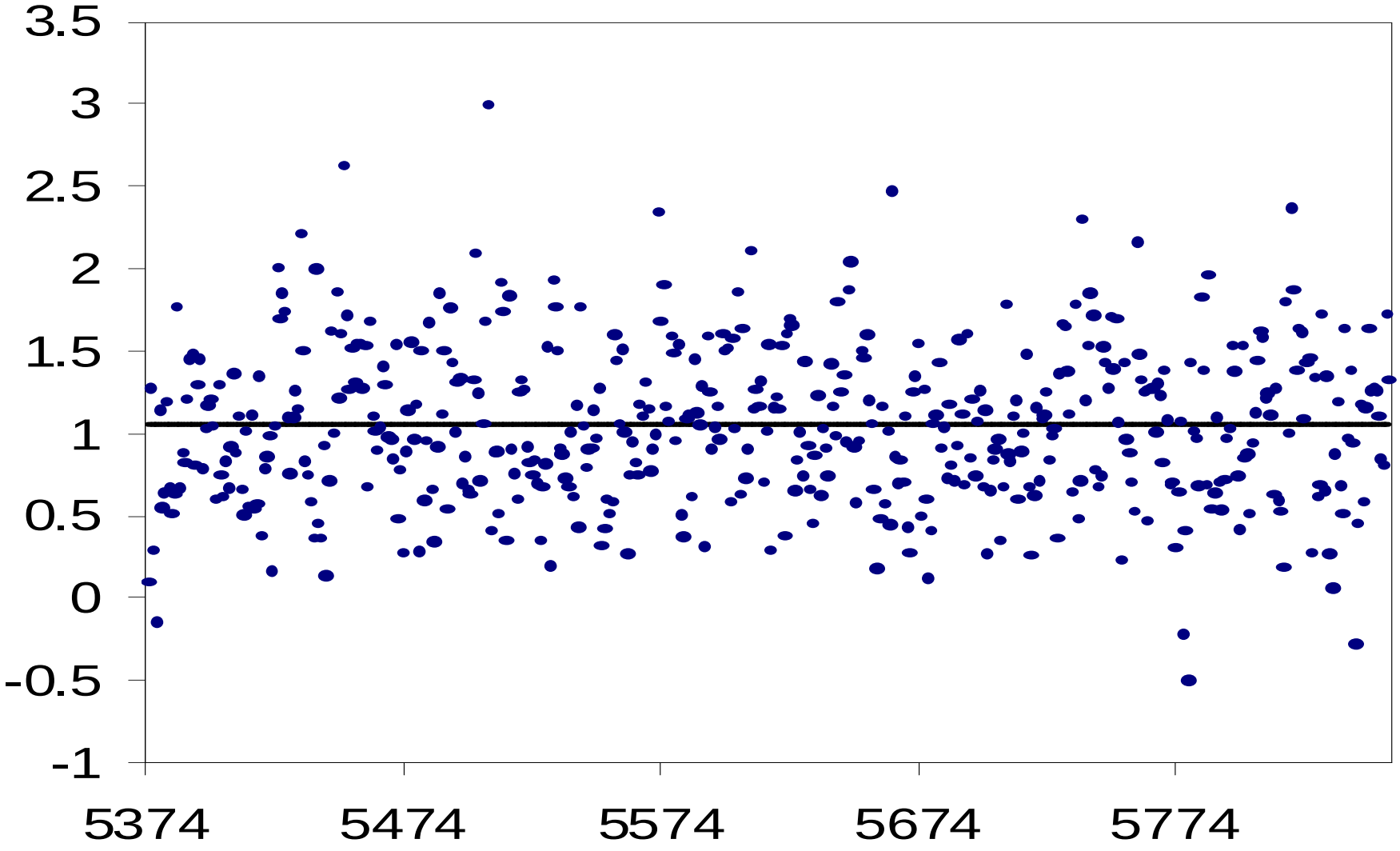
Data with Bayesian regression curve $\pm$ 1 std.-deviation.

Aberrant Gene Copy # of Chromosome 1



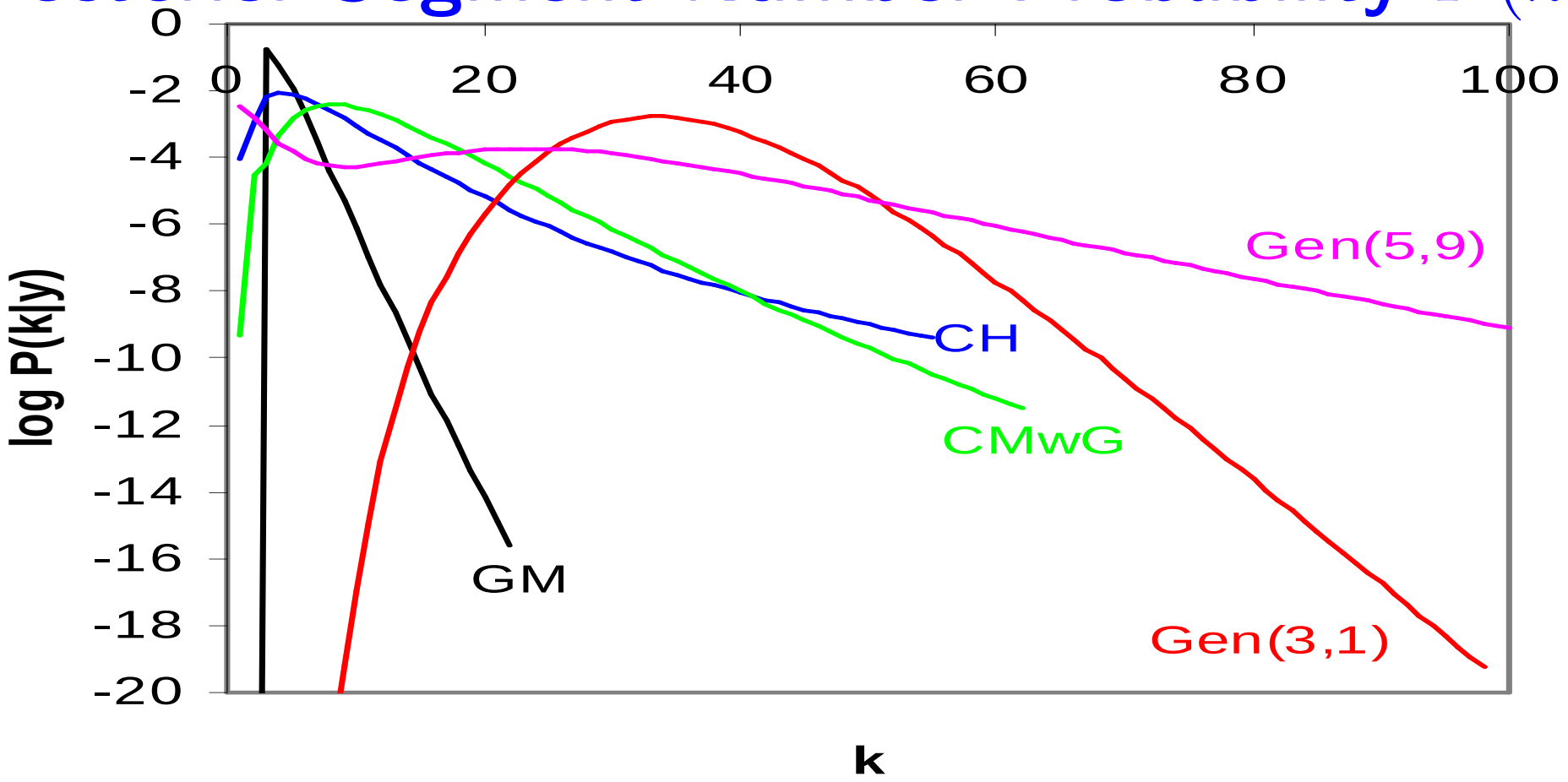
$\log P(\boldsymbol{y})$ (blue) and $\hat{k}$ (green) as function of σ and our estimate $\hat{\sigma}$ of $(\arg)\max_\sigma P(\boldsymbol{y})$ and $\hat{k}(\hat{\sigma})$ (black triangles).

Normal Gene Copy # of Chromosome 9



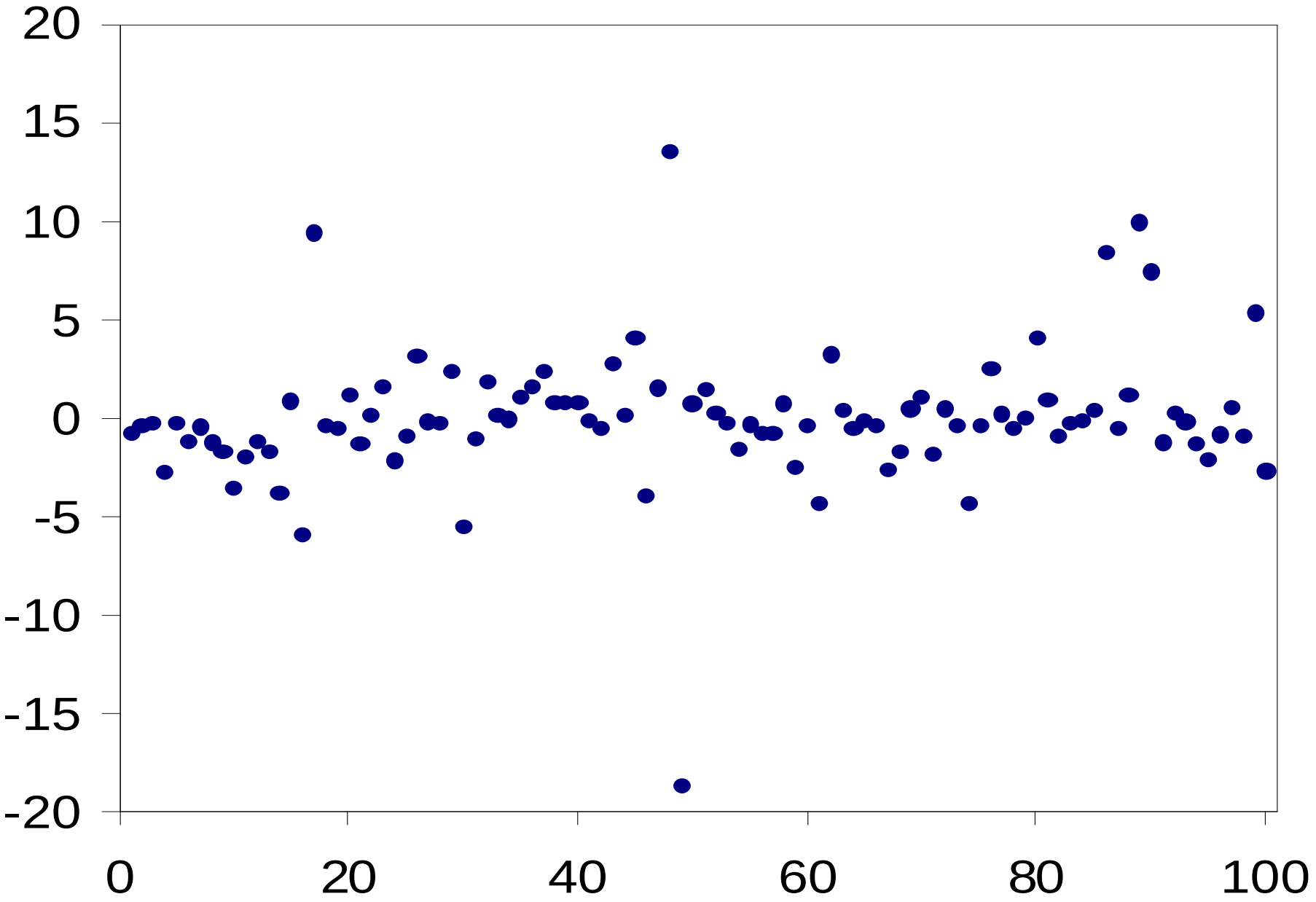
Data with Bayesian regression curve.

Posterior Segment Number Probability $P(k|y)$

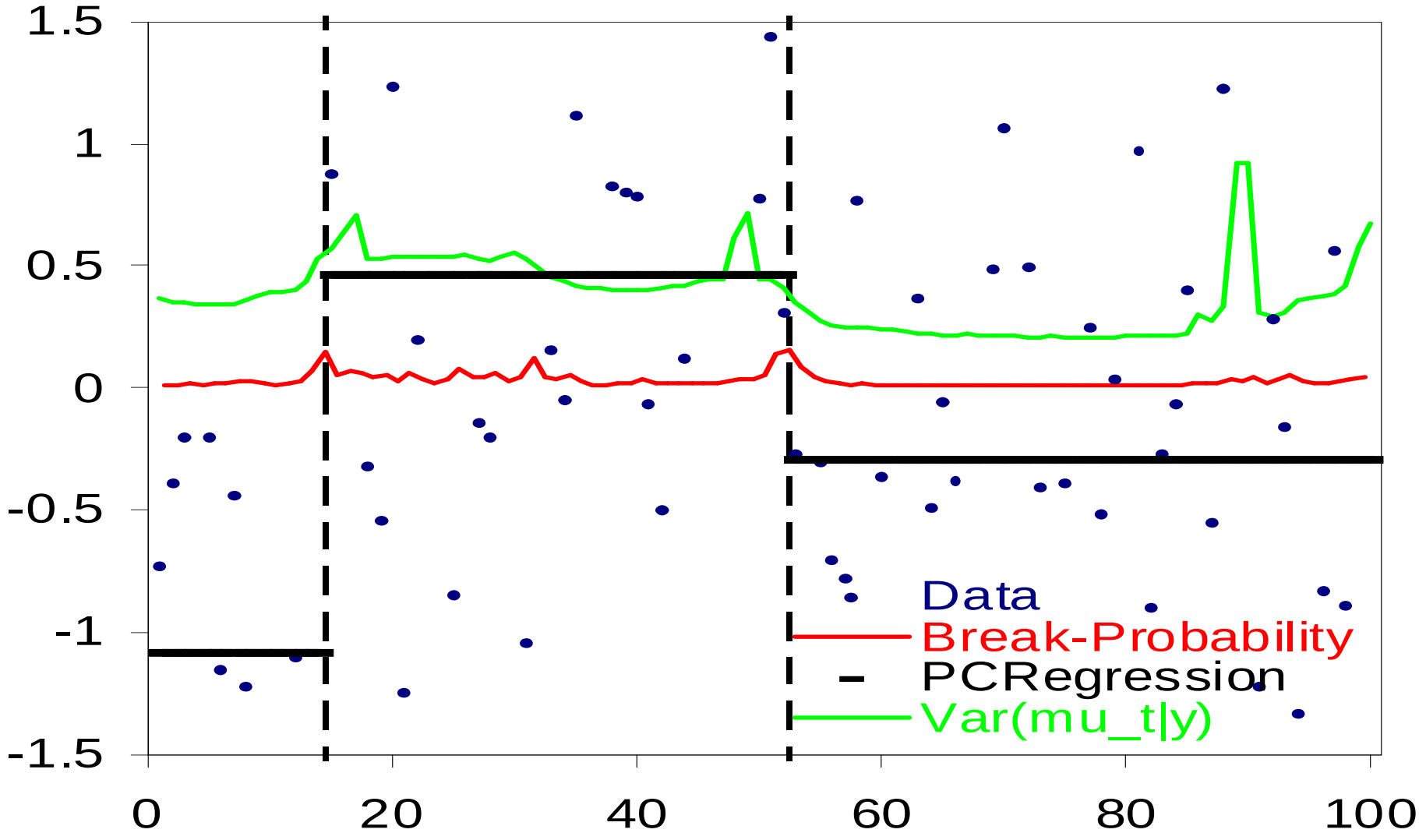


For medium Gaussian noise (GM, black), high Cauchy noise (CH, blue), medium Cauchy noise with Gaussian regression (CMwG, green), aberrant gene expression of chromosome 1 (Gen(3,1), red), normal gene expression of chromosome 9 (Gen(5,9), pink).

Synthetic Example: What do you see?

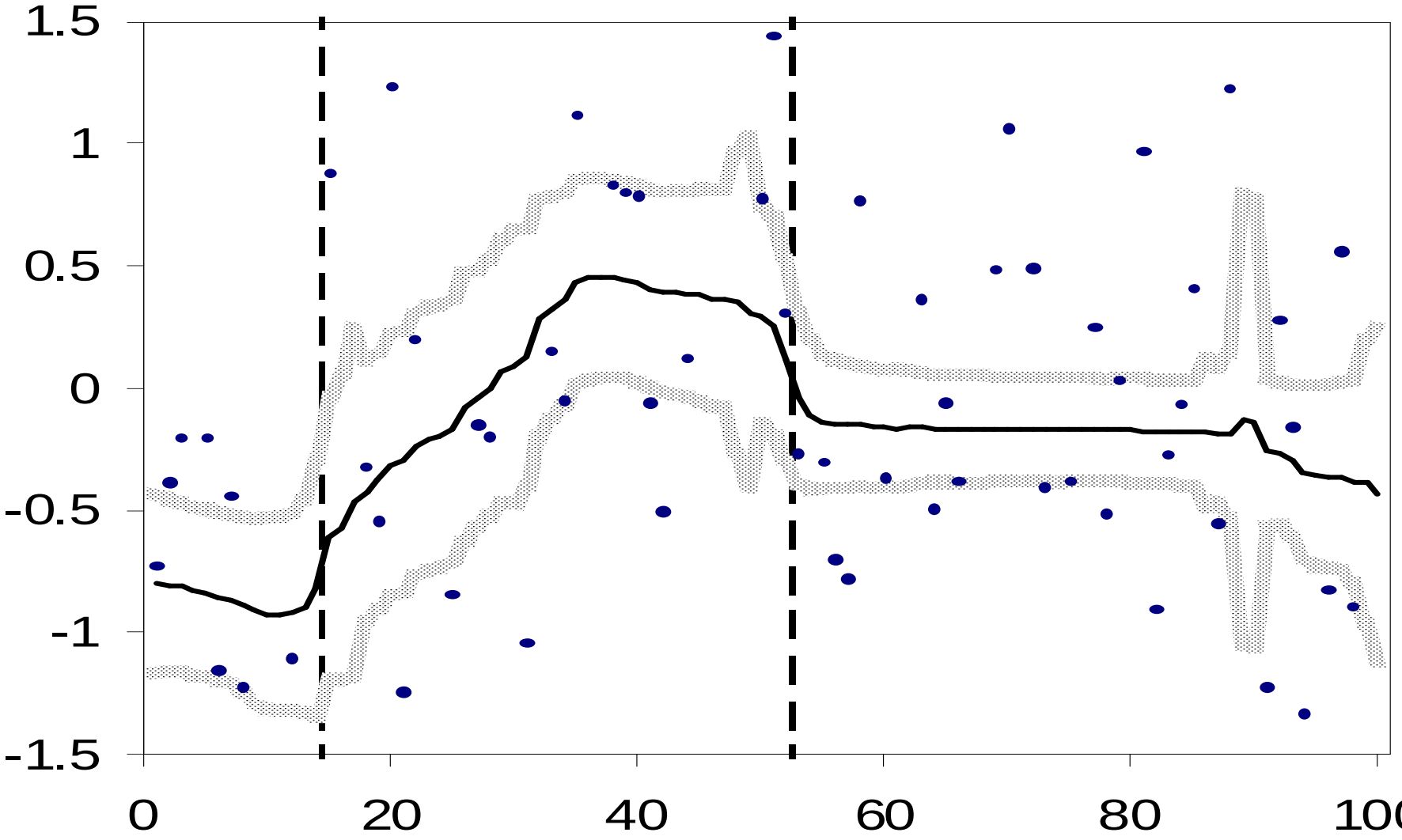


Synthetic Example: PC-Regression



Data was indeed sampled from a three segment function with high Cauchy noise. Data (blue), PCR (black), BP (red), and $\sqrt{\text{Var}}$ (green).

Synthetic Example: Bayesian Regression Curve



Data with Bayesian regression curve $\pm$ 1 std.-deviation.

Regression Summary of Gene and other Examples (↑Setup)

Gauss, Cauchy, Low, Medium, High noise, Gene	true noise scale	data size	method	global mean estimate	global deviation estimate	in-segment deviation est.	log-evidence $\log P(\boldsymbol{y})$	rel. log-likelihood $\frac{ll-\mathbf{E}[ll\|\hat{f}]}{\mathrm{Var}[ll\|\hat{f}]^{1/2}}$	Opt.#segm.	Confidence $P(\hat{k}(-1,+1)\|\boldsymbol{y})$
Name	σ	n	P	$\hat{\nu}$	$\hat{\rho}$	$\hat{\sigma}$	$\log E$	$\frac{ll-\mathbf{E}}{\sigma_{ll}}$	$\hat{k}$	$C_{k(-1,+1)}$
GL	0.10	100	G	-0.01	0.69	0.18	39	4.9	$3\|3$	74%(0|20)
GM	0.32	100	G	-0.03	0.73	0.35	-48	1.2	$3\|3$	44%(0|29)
GH	1.00	100	G	-0.10	1.15	1.03	-156	0.3	$3\|4$	13%(10|12)
CL	0.10	100	C	-0.02	0.58	0.09	-17	1.0	$3\|3$	69%(0|21)
CM	0.32	100	C	-0.09	0.70	0.27	-127	0.8	$3\|3$	38%(0|27)
CH	1.00	100	C	-0.20	0.99	0.86	-234	0.9	$3\|4$	12%(11|11)
GMwC	0.32	100	C	0.00	0.49	0.17	-70	1.5	$3\|3$	27%(0|26)
CMwG	0.32	100	G	0.01	1.24	1.22	-160	2.9	$5\|8$	8%(8|8)
Gen31	–	769	G	0.55	0.45	0.30	-283	-1.5	$15\|34$	6%(6|6)
Gen59	–	483	G	1.05	0.47	0.44	-336	-2.3	$1\|1$	8%(0|6)

Extensions

- Any generalized one-segment evidence (no problem)
- Known segment levels (even easier)
- (Non)constant regressors (easy)
- Piecewise linear regression (easy)
- Continuous regression (harder, approximate)
- Non-parametric prior and noise (easy)
- Very large n (break into overlapping pieces, heuristic)

Related work

[Sen&Srivastava'75] Frequentist solution for detecting a single break.

[Olshen&al'04] Generalization to pair of breaks.

Heuristic recursion for further remaining breaks.

[Jon'03,Lavielle'05] Penalized Maximum Likelihood.

[Endres&Földiák'05] Piecewise constant (PC) Bayesian density estimation.

[Lav'05,EF'05] Dynamic programming.

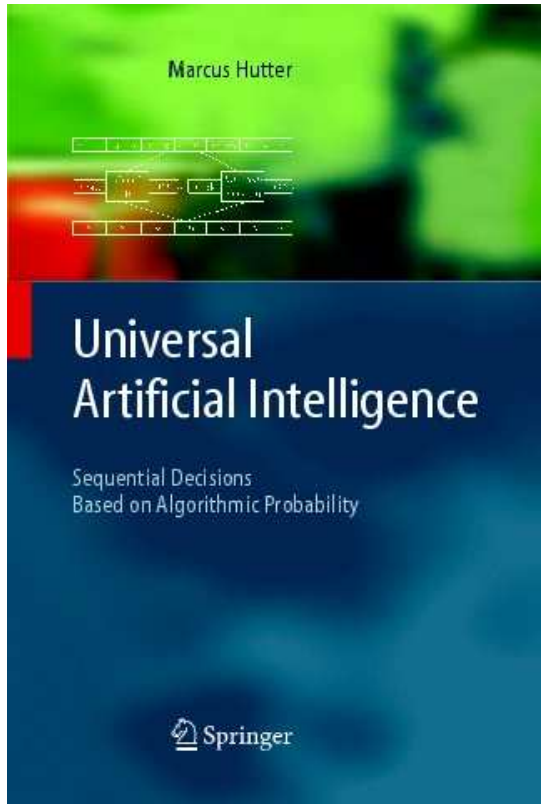
Summary

- Full Bayesian PC-regression
- (Non)Gaussian noise and prior
- Handling of outliers
- Analytic estimate for in-segment variance
- Bayesian regression curve
- Break probabilities and variances
- Global evidence for model comparison
- Principled, little parameters to choose (important for det. of k).

Thanks! Questions? Details:

Papers at http://www.idsia.ch/~marcus

Book intends to excite a broader AI audience about abstract Algorithmic Information Theory –and– inform theorists about exciting applications to AI.



$$\begin{array}{ccc} \text{Decision Theory} & = & \text{Probability + Utility Theory} \\ + & & + \\ \text{Universal Induction} & = & \text{Ockham + Bayes + Turing} \\ = & & = \\ \multicolumn{3}{c}{\text{A Unified View of Artificial Intelligence}} \end{array}$$