Some Model Selection Problems in Image Analysis

Mário A. T. Figueiredo
"Instituto Superior Técnico"
Lisboa
PORTUGAL

People involved in work here presented:

José M. N. Leitão, Instituto Superior Técnico, PORTUGAL
Anil K. Jain, Michigan State University, USA
Robert D. Nowak, Rice University, USA

Outline

- Image analysis as Bayesian inference: a brief review
- Model selection problems
- Approaches to model selection:

  Bayesian model selection (Bayes factors)

  Minimum description length (MDL)

- Example: discontinuity-preserving restoration
- Example: contour estimation
- Example: segmentation of Poisson imagery
- Concluding remarks

Image analysis as statistical inference

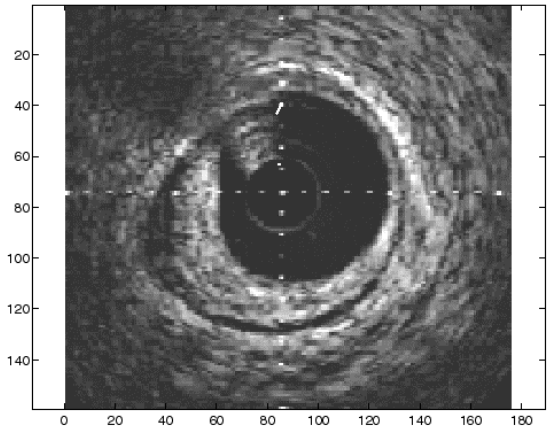
Most image analysis problems can/should be formulated as

Given observed data **g**, infer **f**

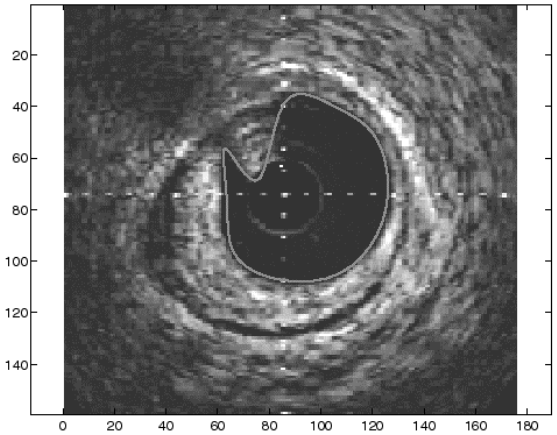
Seems a trivial statement.

My message: “start by writing down what f and g are”

- Examples



inference

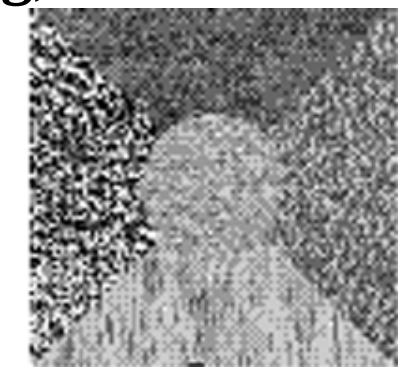


g, an observed image

f, a contour, e.g., represented by a sequence of points

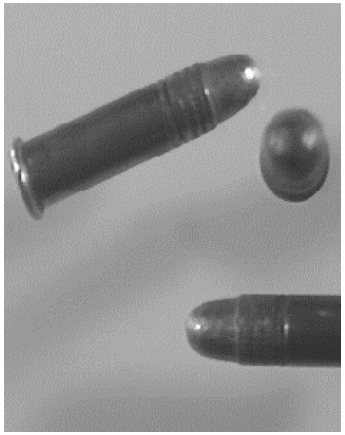
Image analysis as statistical inference: more examples

g, observed image



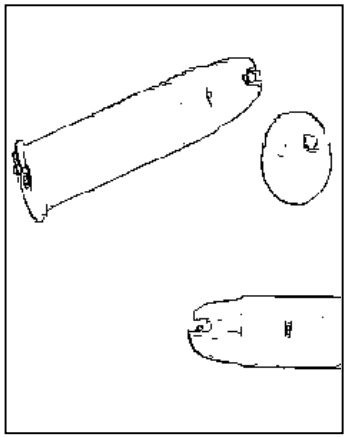
inference

f, segmentation



g, observed image

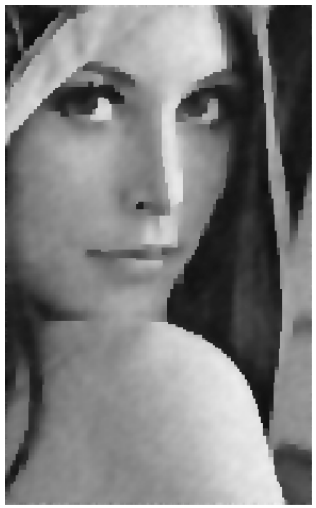
inference



f, edge map

g, noisy image

inference



f, restored image

Image analysis as statistical inference

- In all the cases here considered: **g** = “one observed image”
- Other possibilities:

  extracted features (e.g., local filters),

  tomographic projections,

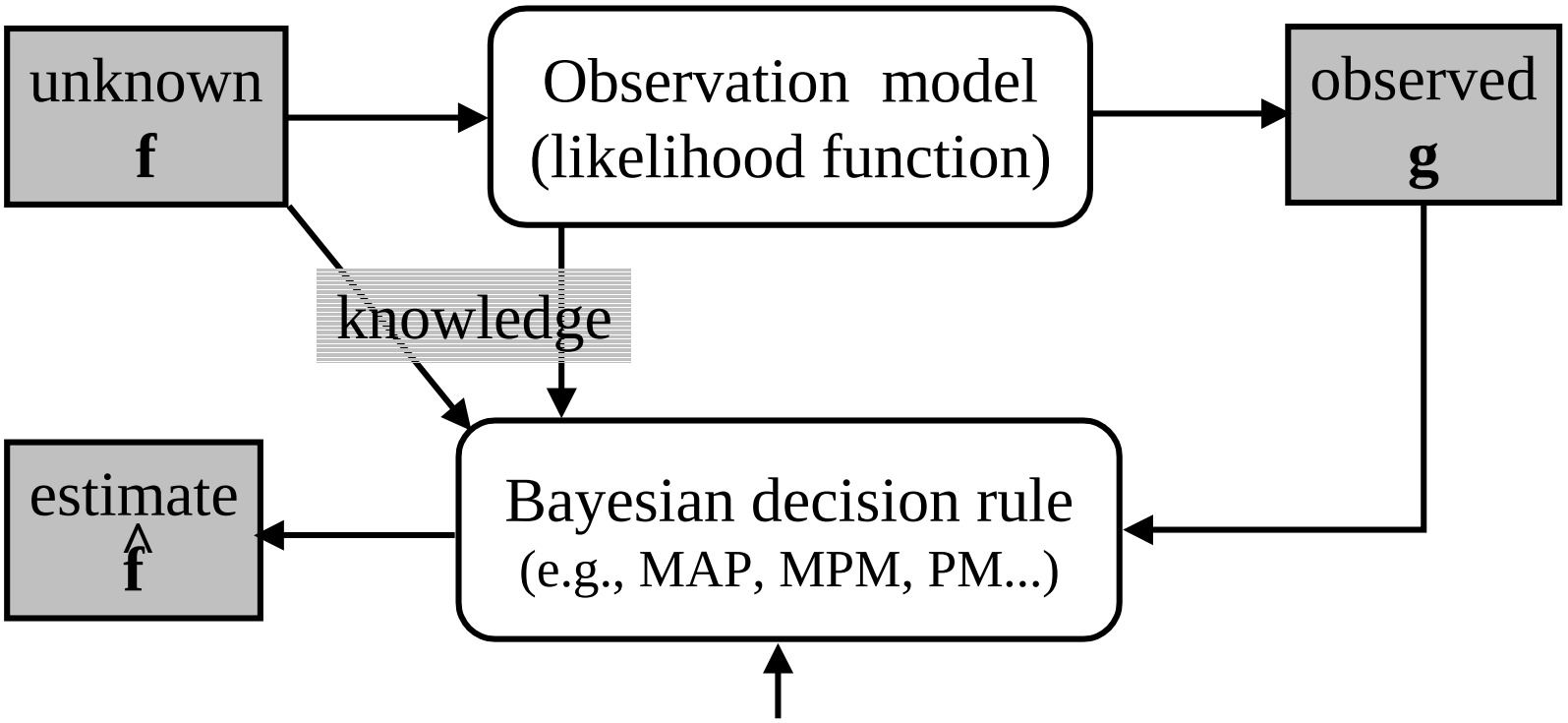
  sequences of images,

  ….

- **f** can be image-like (e.g., in image restoration or segmentation)
- **f** can be a parametric representation (e.g., a spline contour)

Many uncertainties
(noise, uncertainty about f,…) $\Longrightarrow$ Natural frameworks:
- statistical inference
- statistical decision making

Bayesian image analysis



The Bayesian approach is explicitly model-based

Bayesian image analysis: basic elements

- Observation model / likelihood function:

$$p(\mathbf{g} \mid \mathbf{f}, \phi)$$

$\mathbf{f}$ is the unknown
$\mathbf{g}$ is the observed data
ϕ are parameters (assumed known, by now)

- Prior knowledge:

$$p(\mathbf{f} \mid \psi)$$

$\mathbf{f}$ is the unknown
ψ are parameters (assumed known, by now)

- *A posteriori* knowledge (via Bayes theorem):

$$p(\mathbf{f} \mid \mathbf{g}, \phi, \psi) = \frac{p(\mathbf{g} \mid \mathbf{f}, \phi)\, p(\mathbf{f} \mid \psi)}{p(\mathbf{g} \mid \phi, \psi)}$$

Bayesian image analysis: loss function and Bayes risk

Given $p(\mathbf{f} \mid \mathbf{g}, \phi, \psi)$ (obtained via Bayes law)

…and a loss function $L(\mathbf{f}, \hat{\mathbf{f}})$

A posteriori expected loss:

$$E[L(\mathbf{f}, \hat{\mathbf{f}}) \mid \mathbf{g}, \phi, \psi] = \int L(\mathbf{f}, \hat{\mathbf{f}})\, p(\mathbf{f} \mid \mathbf{g}, \phi, \psi)\, d\mathbf{f}$$

…expected value of the loss function, given the observed data **g**.

Optimal Bayes rule

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}'} E[L(\mathbf{f}, \mathbf{f}') \mid \mathbf{g}, \phi, \psi]$$

…minimizes the *a posteriori* expected loss:

Bayesian image analysis: common decision rules

So, what to do with the posterior $\boxed{p(\mathbf{f} \mid \mathbf{g}, \phi, \psi)}$?

It all depends on the loss function. Examples:

$$L(\mathbf{f}, \hat{\mathbf{f}}) = \begin{cases} 1 \Leftarrow \mathbf{f} \neq \hat{\mathbf{f}} \\ 0 \Leftarrow \mathbf{f} \neq \hat{\mathbf{f}} \end{cases} \longrightarrow \hat{\mathbf{f}}_{\text{MAP}} = \arg\max_{\mathbf{f}} p(\mathbf{f} \mid \mathbf{g}, \phi, \psi)$$

$$L(\mathbf{f}, \hat{\mathbf{f}}) = \| \mathbf{f} - \hat{\mathbf{f}} \|^2 \longrightarrow \hat{\mathbf{f}}_{\text{PM}} = \text{E}[\mathbf{f} \mid \mathbf{g}, \phi, \psi]$$

$$L(\mathbf{f}, \hat{\mathbf{f}}) = \sum_i L_i(f_i, \hat{f}_i) \longrightarrow \hat{f}_i = \arg\min_{f_i} \text{E}[L_i(f_i - \hat{f}_i) \mid \mathbf{g}, \phi, \psi]$$

(additive loss)

Marginal criteria, e.g., *maximizer of posterior marginals* (MPM), Marroquin, Mitter, and Poggio, JASA 1987)

Unknown (hyper)parameters

What if ϕ and/or ψ (the hyper-parameters) are unknown ?

- Unknown: $\mathbf{f}, \phi, \psi$

- *A posteriori* probability function

$$p(\mathbf{f}, \phi, \psi \mid \mathbf{g}) = \frac{p(\mathbf{g} \mid \mathbf{f}, \phi)\, p(\mathbf{f} \mid \psi)\, p(\phi, \psi)}{p(\mathbf{g})}$$

$p(\phi, \psi)$
"hyper-prior"

- There is nothing fundamentally new

- This is called a hierarchical Bayes model

Rules involving (hyper)parameters

Again, it all depends on the loss function. Examples:

$$L\left((\mathbf{f},\phi,\psi),(\hat{\mathbf{f}},\hat{\phi},\hat{\psi})\right)=\begin{cases}1 \Leftarrow \mathbf{f}\neq\hat{\mathbf{f}}\\ 0 \Leftarrow \mathbf{f}=\hat{\mathbf{f}}\end{cases} \longrightarrow \hat{\mathbf{f}}_{\text{MMAP}}=\arg\max_{\mathbf{f}} p(\mathbf{f}\,|\,\mathbf{g})$$

(*marginal MAP*)

(does not care about ϕ and ψ)

$$L(\cdot,\cdot)=\begin{cases}1 \Leftarrow (\mathbf{f},\phi,\psi)\neq(\hat{\mathbf{f}},\hat{\phi},\hat{\psi})\\ 0 \Leftarrow (\mathbf{f},\phi,\psi)=(\hat{\mathbf{f}},\hat{\phi},\hat{\psi})\end{cases} \longrightarrow (\hat{\mathbf{f}},\hat{\phi},\hat{\psi})_{\text{GMAP}}=\arg\max_{\mathbf{f},\phi,\varphi} p(\mathbf{f},\phi,\psi\,|\,\mathbf{g})$$

(*generalized MAP*)

$$L(\cdot,\cdot)=\begin{cases}1 \Leftarrow (\phi,\psi)\neq(\hat{\phi},\hat{\psi})\\ 0 \Leftarrow (\phi,\psi)=(\hat{\phi},\hat{\psi})\end{cases} \longrightarrow (\hat{\phi},\hat{\psi})_{\text{PMMAP}}=\arg\max_{\phi,\psi} p(\phi,\psi\,|\,\mathbf{g})$$

(*parameter marginal MAP*, usually via EM)

If $p(\phi,\psi)\propto \text{const}$ $\quad (\hat{\phi},\hat{\psi})_{\text{PMMAP}}=(\hat{\phi},\hat{\psi})_{\text{MML}}=\arg\max_{\phi,\psi} p(\mathbf{g}\,|\,\phi,\psi)$

(maximum marginal likelihood)

Empirical Bayes

1. Estimate hyper-parameters, e.g., $(\hat{\phi},\hat{\psi})_{\mathrm{MML}} = \arg\max_{\phi,\psi} p(\mathbf{g}\,|\,\phi,\psi)$

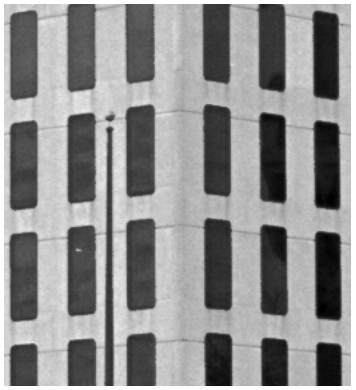
2. Plug them in as if they were known, e.g.

$$\hat{\mathbf{f}}_{\mathrm{MAP}} = \arg\max_{\mathbf{f}}\ p(\mathbf{f}\,|\,\mathbf{g},\hat{\phi},\hat{\psi})$$

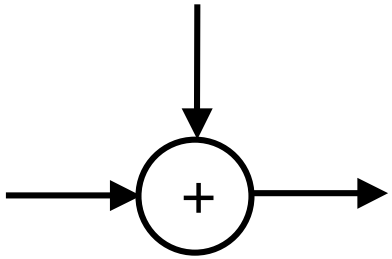
This is **not** a truly Bayesian procedure,

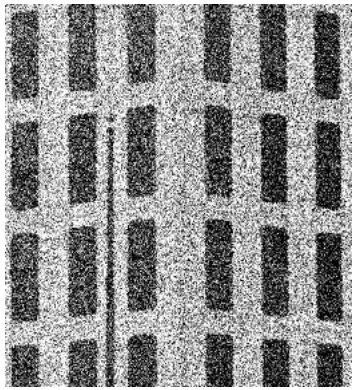
…but has nice asymptotic properties.

Image analysis examples: image restoration



Gaussian
white noise





f (unknown)

g (observed)

Likelihood function

$$p(\mathbf{g}|\mathbf{f},\sigma^2) \equiv \mathcal{N}(\mathbf{g}|\mathbf{f},\sigma^2\mathbf{I}_N)$$

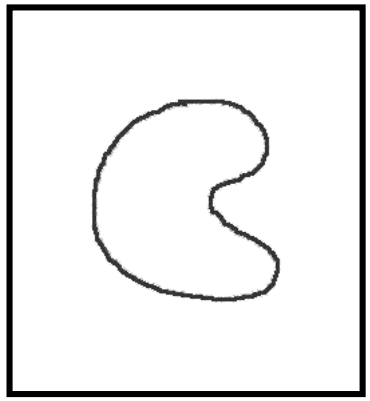
σ^2 → noise variance

N → number of pixels

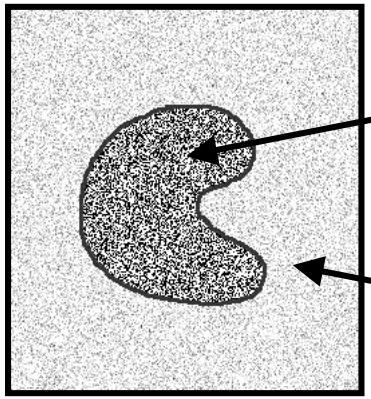
$\mathbf{I}_N$ → N x N identity matrix

See: M. Figueiredo and J. Leitão, "Unsupervised image restoration and Edge location using Gauss-Markov random fields and the MDL principle", *IEEE Transactions on Image Processing*, vol. 6, pp. 1089-1102, 1997.

Image analysis examples: contour estimation



observation mechanism



a statistical model for the inside,

...another one for the outside

f, contour description
(e.g., spline, Fourier,…)

g, observed image

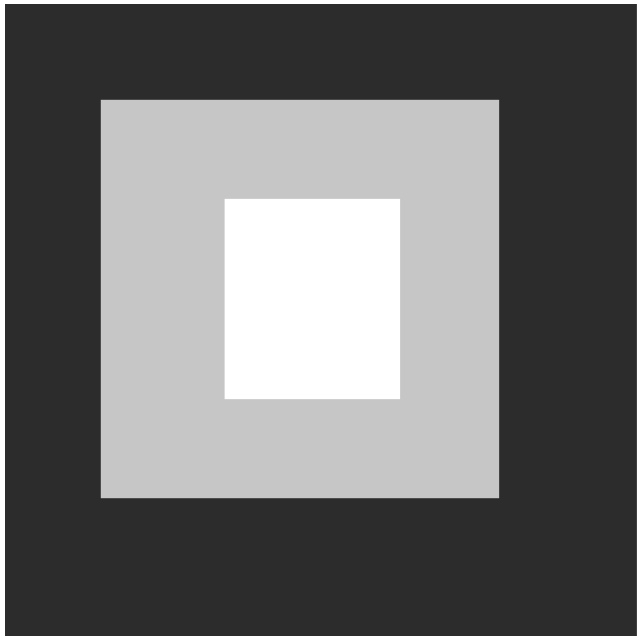
Under inside/outside independence assumption:

$$p(\mathbf{g}\,|\,\mathbf{f},\phi_{in},\phi_{out}) = p(\{g_i : i \in in(\mathbf{f})\}\,|\,\phi_{in})\;\; p(\{g_i : i \in out(\mathbf{f})\}\,|\,\phi_{out})$$

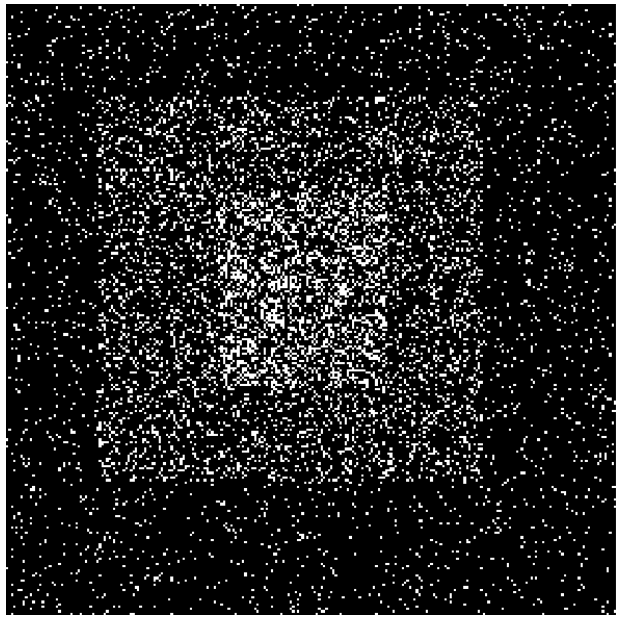
Examples: Gaussians of different means and variances;
Rayleigh of different variances (ultrasound images); Different textures.

See: M. Figueiredo, J. Leitão, and A. Jain, "Adaptive B-splines and boundary estimation",in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-97), pp. 724-729, San Juan, Puerto Rico, 1997.

Image analysis examples: Poisson image segmentation



observation mechanism



f, intensity image

g, observed counts

Under conditional independence assumption:

$$p(\mathbf{g}\,|\,\mathbf{f}) = \prod_i p(g_i\,|\,f_i) = \prod_i \frac{(f_i)^{g_i}\exp(-f_i)}{g_i!}$$

(independent Poisson observations)

Priors

What prior information to use ?

In image restoration → Images should be smooth, except at edges/discontinuities

In contour estimation → Contours should be smooth

In image segmentation → Nearby pixels tend to belong to the same region

Questions: How many edges should we allow ?

How smooth should the contour be ?

Into how many regions should we segment the image ?

These are all model selection questions

Model selection

Suppose there are K possible models, $m \in \{m_1, ..., m_K\}$

…with a priori probabilities $p(m_1),\ldots,p(m_K)$

Given model m:

- a likelihood function $p(\mathbf{g} \mid \mathbf{f}, \phi_{(m)}, m)$

- a prior $p(\mathbf{f} \mid \psi_{(m)}, m)$

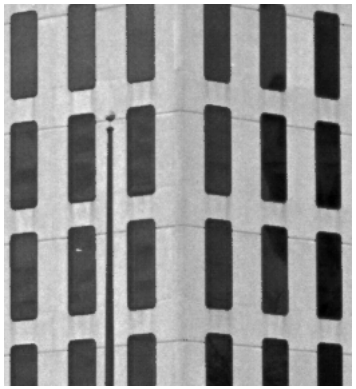
- a hyper-prior $p(\phi_{(m)}, \psi_{(m)} \mid m)$

$\phi_{(m)}, \psi_{(m)}$ are indexed by m: they may have different structures

There is nothing new: unknowns are now $m, \mathbf{f}, \phi_{(m)}, \psi_{(m)}$

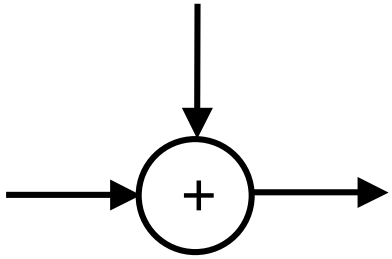
standard Bayes approach can be followed

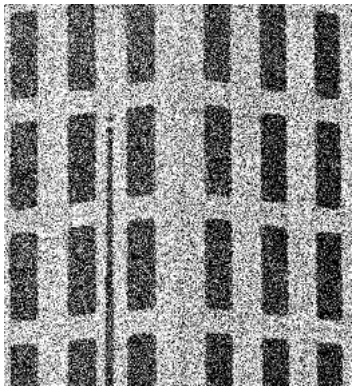
Model selection: image restoration



f (unknown)

Gaussian white noise





g (observed)

Likelihood function

$$p(\mathbf{g}|\mathbf{f},\sigma^2)=\mathcal{N}(\mathbf{g}|\mathbf{f},\sigma^2\mathbf{I})$$

Prior (smoothness almost everywhere)

$$p(\mathbf{f}\,|\,\varphi_{(m)},m)=\frac{1}{Z(\varphi_{(m)},m)}\exp\Big[-\frac{\mu}{2}\sum_{i\sim j}(1-d_{i,j})\left(f_i-f_j\right)^2\Big]$$

number of $d_{i,j}=1$

partition function

$d_{i,j}\in\{0,1\}$

$\varphi_{(m)}=\left(\mu,\{(i,j):d_{i,j}=1\}\right)$

edge locations

sum over all nearest neighbors

Model selection: image restoration

Compound Gauss-Markov random field (CGMRF)

$$p(\mathbf{f}\,|\,\boldsymbol{\psi}_{(m)}, m) = \frac{\sqrt{\det \mathbf{A}(\boldsymbol{\psi}_{(m)})}}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}\mathbf{f}^{T}\mathbf{A}(\boldsymbol{\psi}_{(m)})\mathbf{f}\right]$$

$\mathbf{A}(\boldsymbol{\psi}_{(m)})$ → "potential matrix" (inverse of the covariance)

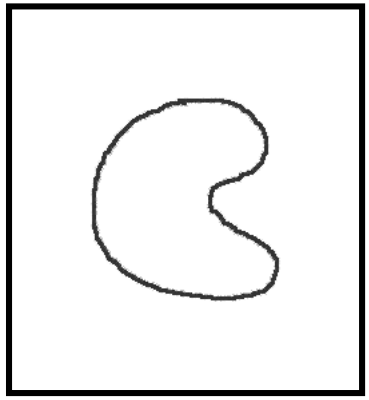
Given $\boldsymbol{\psi}_{(m)}$, the image estimate is given by a simple Wiener-type filter:

$$\hat{\mathbf{f}} = \left[\sigma^2 \mathbf{A}(\boldsymbol{\psi}_{(m)}) + \mathbf{I}\right]^{-1}\mathbf{g}$$

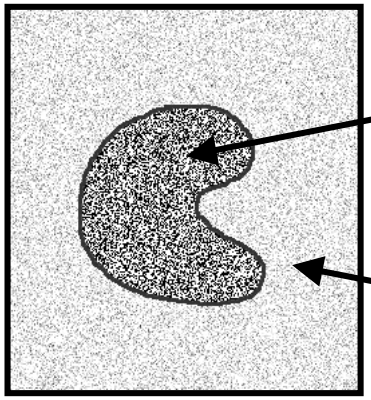
But m = ? How many edges ? → model selection problem

See: M. Figueiredo and J. Leitão, "Unsupervised image restoration and Edge location using Gauss-Markov random fields and the MDL principle", *IEEE Transactions on Image Processing*, vol. 6, pp. 1089-1102, 1997.

Model selection: contour estimation



observation mechanism



a statistical model for the inside,

...another one for the outside

f, contour description

g, observed image

A (discretized) B-spline contour description

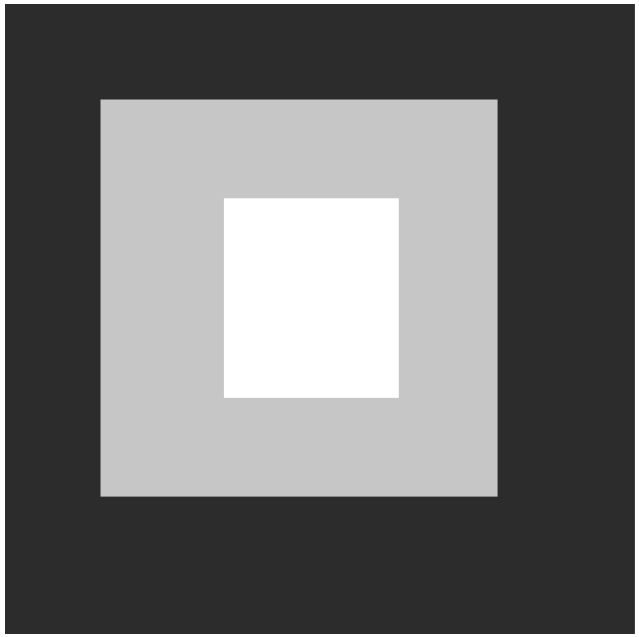
$\mathbf{f}_{(m)} = \mathbf{B}_{(m)}\theta_{(m)}$ → a 2-D spline curve with m control points

matrix with discretized B-spline basis

control points coordinates

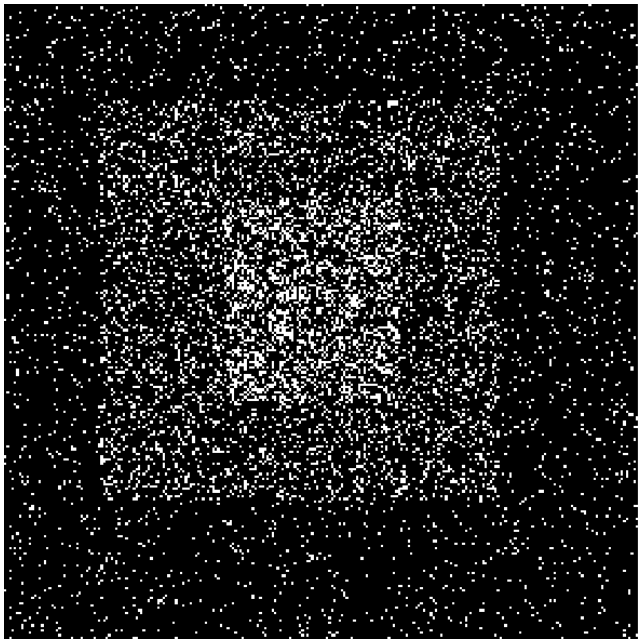
m = ? How complex should the contour be ? → model selection

Model selection: segmentation



f, intensity image

observation mechanism



g, observed counts

f is a piece-wise constant image

How many regions of constant intensity ?

model selection problem

Model selection: the full posterior

$$p(m,\mathbf{f},\phi_{(m)},\psi_{(m)}\,|\,\mathbf{g}) = \frac{p(\mathbf{g}\,|\,\mathbf{f},\phi_{(m)},m)\;p(\mathbf{f}\,|\,\psi_{(m)},m)\;p(\phi_{(m)},\psi_{(m)}\,|\,m)\;p(m)}{p(\mathbf{g})}$$

Again, what to do with this, depends on the loss function

("true Bayesians" just report it, or sample from it; e.g., RJ-MCMC)

Seen strictly as a model selection problem, the natural choice is

$$L(\cdot,\cdot) = \begin{cases} 1 \Leftarrow m \neq \hat{m} \\ 0 \Leftarrow m \neq \hat{m} \end{cases} \longrightarrow \hat{m} = \arg\max_{m}\; p(m\,|\,\mathbf{g})$$

$$\hat{m} = \arg\max_{m}\left\{p(m\,|\,\mathbf{g})\right\} = \arg\max_{m}\left\{\underbrace{p(\mathbf{g}\,|\,m)}\;p(m)\right\}$$

sometimes called the "evidence for model m"

Model selection and Bayes factors

$$\hat{m} = \arg\max_m \{ p(m \mid \mathbf{g}) \} = \arg\max_m \{ p(\mathbf{g} \mid m)\, p(m) \}$$

When comparing two models

$$\frac{p(m_1 \mid \mathbf{g})}{p(m_2 \mid \mathbf{g})} = \underbrace{\frac{p(\mathbf{g} \mid m_1)}{p(\mathbf{g} \mid m_2)}}_{\text{Bayes factor}} \underbrace{\frac{p(m_1)}{p(m_2)}}_{\text{Prior odds ratio}}$$

Common approach (we may call it the “model selection approach”):

1. Take $\hat{m}$, e.g. $\hat{m} = \arg\max_m p(m \mid \mathbf{g})$
2. Use it as if it were “the true model”

This is an empirical Bayes approach

Alternative: model averaging (not focused here).

Evidence and Schwarz's BIC

$$\hat{m} = \arg\max_{m} \{ \log p(\mathbf{g} \mid m) + \log p(m) \}$$

Main difficulty: computing the evidence

To simplify the notation: $\mathbf{f}_{(m)} \equiv (\mathbf{f}, \phi_{(m)}, \psi_{(m)})$

Common approach: a Laplace approximation (valid asymptotically)

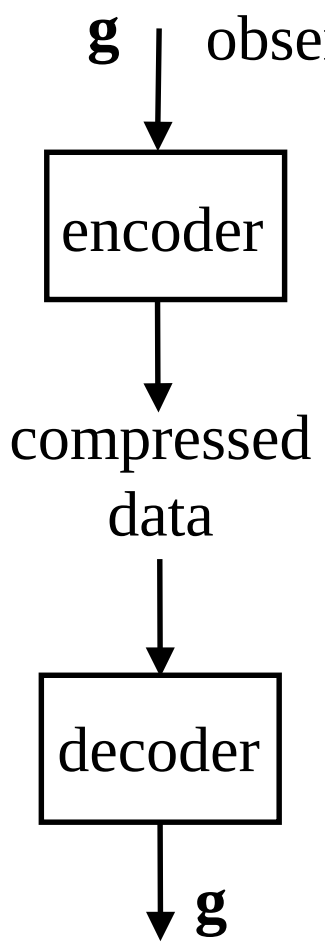
$$\log p(\mathbf{g} \mid m) \approx \log p(\mathbf{g} \mid \hat{\mathbf{f}}_{(m)}^{ML}, m) - \frac{\dim(\mathbf{f}_{(m)})}{2} \log n$$

Schwarz's Bayesian inference criterion (BIC)

Can be seen as an "order-penalized" log-likelihood.

The minimum description length (MDL) criterion

Introduction to MDL



Rationale: short code $\Leftrightarrow$ good model

long code $\Leftrightarrow$ bad model

code length $\Leftrightarrow$ model adequacy

Several flavors: Rissanen 78, JRSS 87
Rissanen 96,
Wallace and Freeman, JRSS 87

Formalizing the MDL criterion

Scenario: we are given a set of models for the data

each model m is characterized by (unknown) "parameters" $\mathbf{f}_{(m)}$

$$\{p(\mathbf{g} \mid \mathbf{f}_{(m)}, m),\ m = m_1, m_2, ..., m_K\}$$

no (or vague) prior information about $\mathbf{f}_{(m)}$

Goal: given data **g**,

build the shortest possible code for **g**

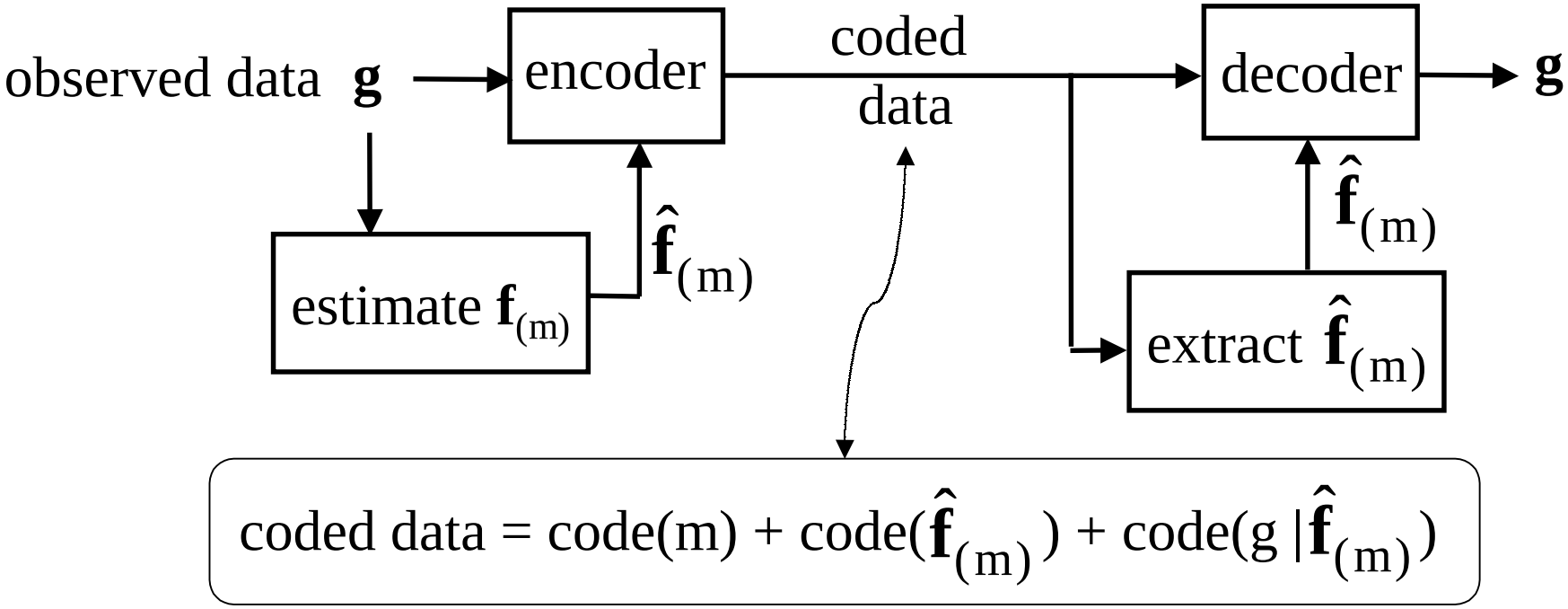
With $\mathbf{f}_{(m)}$ known, the shortest code-length for **g** is (Shannon's)

$$\boxed{L(\mathbf{g} \mid \mathbf{f}_{(m)}) = -\log p(\mathbf{g} \mid \mathbf{f}_{(m)}, m)}$$

However, $\mathbf{f}_{(m)}$ is, a priori, unknown; it has to be estimated.

Formalizing the MDL criterion

Assumption: given $\mathbf{f}_{(m)}$, both encoder and decoder know how to build the same code



MDL principle:
choose m and $\hat{\mathbf{f}}_{(m)}$ so that length(coded data) is shortest

Formalizing the MDL criterion

coded data = code(m) + code($\hat{\mathbf{f}}_{(m)}$) + code(g | $\hat{\mathbf{f}}_{(m)}$)

$$L(m, \mathbf{f}_{(m)}, \mathbf{g}) = \underbrace{L(m)}_{\text{Usually constant}} + L(\mathbf{f}_{(m)} \mid m) + L(\mathbf{g} \mid \mathbf{f}_{(m)})$$

MDL criterion

$$(\hat{m}, \hat{\mathbf{f}}_{(\hat{m})})_{\text{MDL}} = \arg \min_{m, \mathbf{f}_{(m)}} \{L(\mathbf{f}_{(m)}) + L(\mathbf{g} \mid \mathbf{f}_{(m)})\}$$

$$= \arg \min_{m, \mathbf{f}_{(m)}} \{L(\mathbf{f}_{(m)}) - \log p(\mathbf{g} \mid \mathbf{f}_{(m)})\}$$

Formalizing the MDL criterion

$$(\hat{m}, \hat{\mathbf{f}}_{(\hat{m})})_{\mathrm{MDL}} = \arg \min_{m, \mathbf{f}_{(m)}} \{ L(\mathbf{f}_{(m)}) - \log p(\mathbf{g} \mid \mathbf{f}_{(m)}) \}$$

$L(\mathbf{f}_{(m)})$? Finite $L(\mathbf{f}_{(m)}) \Rightarrow$ truncate to finite precision: $\tilde{\mathbf{f}}_{(m)}$

High precision

$$-\log f(\mathbf{g} \mid \tilde{\mathbf{f}}_{(m)}) \approx -\log f(\mathbf{g} \mid \hat{\mathbf{f}}^{\mathrm{ML}}_{(m)}) \quad \text{but} \quad L(\tilde{\mathbf{f}}_{(m)}) \nearrow$$

Low precision

$L(\tilde{\mathbf{f}}_{(m)}) \searrow$ but

$-\log f(g \mid \tilde{\mathbf{f}}_{(m)})$ may be >> $-\log p(\mathbf{g} \mid \hat{\mathbf{f}}^{\mathrm{ML}}_{(m)})$

Optimal compromise (under regularity conditions, and asymptotic)

$$L(\text{each component of } \mathbf{f}_{(m)}) = \frac{1}{2}\log(n)$$

n is the sample size from with the parameter is estimated (growth rate of Fisher info.)

Formalizing the MDL criterion

Under regularity conditions

$$(\hat{m}, \hat{\mathbf{f}}_{(\hat{m})})_{\mathrm{MDL}} = \arg \min_{m, \mathbf{f}_{(m)}} \{ -\log p(\mathbf{g} \mid \mathbf{f}_{(m)}) + \frac{\dim(\mathbf{f}_{(m)})}{2} \log(n) \}$$

Coincides with Schwarz's BIC.

the dimension of $\mathbf{f}_{(m)}$

This is the "classical" (1978) MDL criterion.

It is not valid in our problems.

Basically: each "parameter" is not estimated from all the "data".

Quick and dirty engineering solution:

use the "natural" description length

Back to the image restoration problem

Likelihood function $p(\mathbf{g}|\mathbf{f},\sigma^2)=\mathcal{N}(\mathbf{g}|\mathbf{f},\sigma^2\mathbf{I})$

Prior $p(\mathbf{f}|\psi_{(m)},m)=\mathcal{N}\left(\mathbf{f}|0,\left[A(\psi_{(m)})\right]^{-1}\right)$

$$p(\mathbf{g},\mathbf{f}|\sigma^2,\psi_{(m)},m)\propto p(\mathbf{g}|\mathbf{f},\sigma^2)\,p(\mathbf{f}|\psi_{(m)},m)$$

$$\left(\hat{m},\hat{\psi}_{(\hat{m})},\hat{\sigma}^2\right)_{MDL}=\arg\min_{m,\psi_{(m)}}\{-\log p(\mathbf{g},\mathbf{f}|\sigma^2,\psi_{(m)},m)+L(\psi_{(m)})\}$$

Can be seen as an MDL criterion with missing data (**f** is unknown).
Implemented via an EM-type algorithm

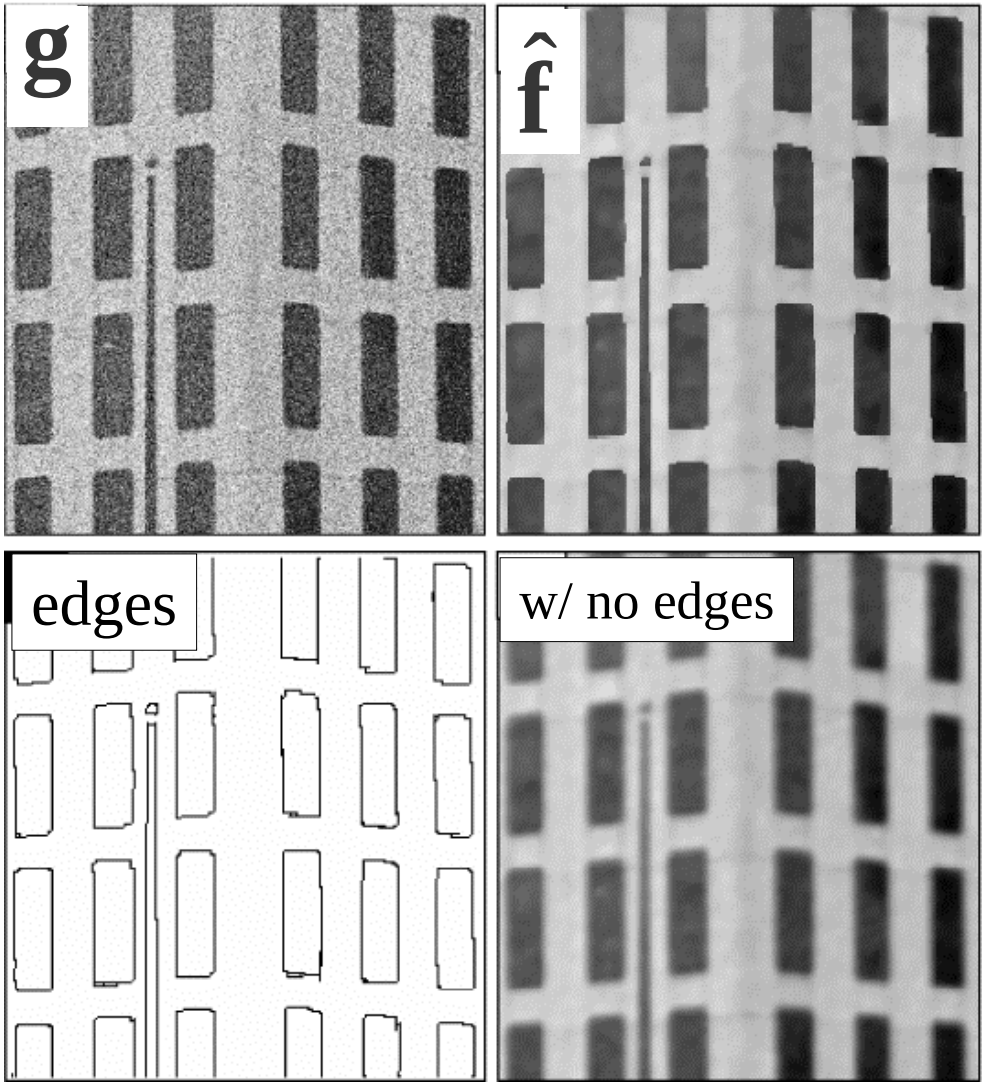
$$L(\psi_{(m)})=m\log(N)$$

Natural code-length for m edge locations on an N-pixels image

Given $\hat{\sigma}^2$ and $\hat{\psi}_{(m)}$

$$\hat{\mathbf{f}}=\left[\hat{\sigma}^2\mathbf{A}(\hat{\psi}_{(m)})+\mathbf{I}\right]^{-1}\mathbf{g}$$

Image restoration examples



Totally unsupervised

μ, σ^2, number of edges, their locations

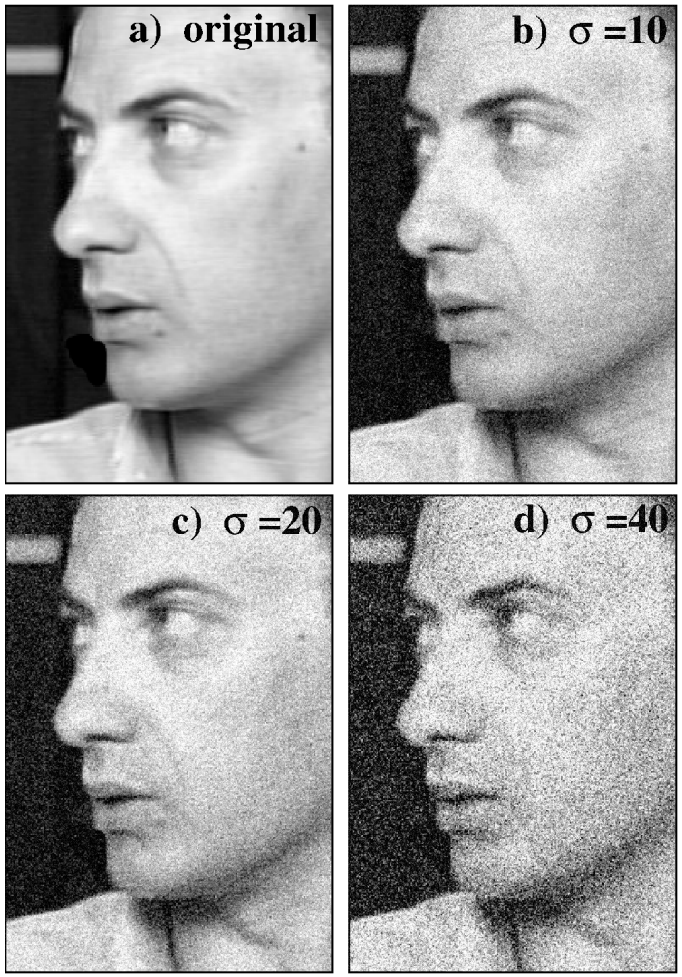
↓

all unknown

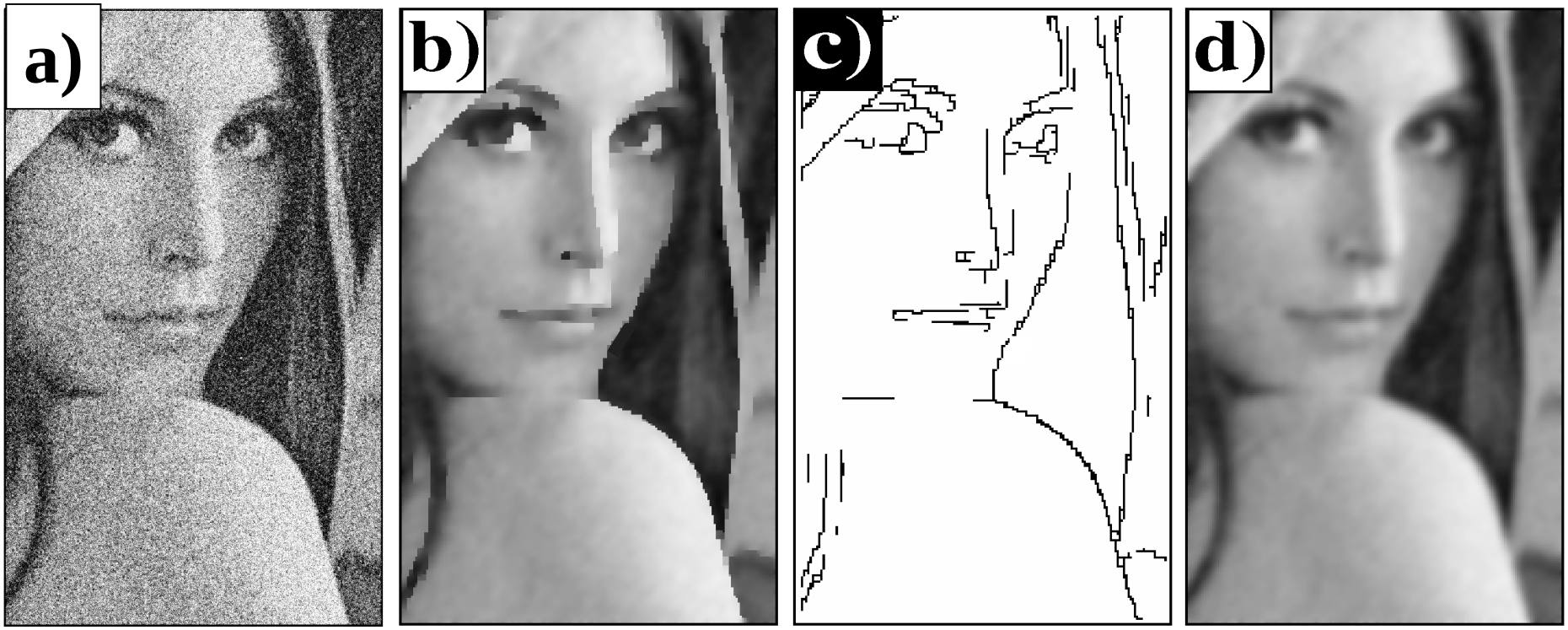
Implemented via an EM-type algorithm, together with a continuation method, and a pseudo-likelihood approximation

M. Figueiredo and J. Leitão, "Unsupervised image restoration and Edge location using Gauss-Markov random fields and the MDL principle", *IEEETransactions on Image Processing*, vol. 6, pp. 1089-1102, 1997.

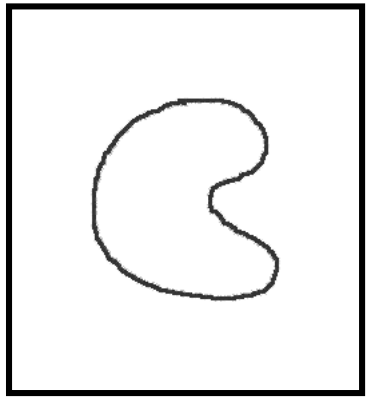
Further image restoration examples



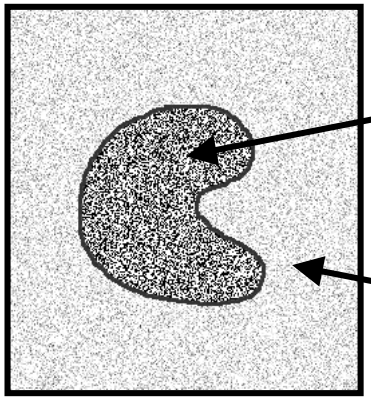
Further image restoration examples



Back to the contour estimation problem



observation mechanism



f, contour description

g, observed image

$\mathbf{f}_{(m)} = \mathbf{B}_{(m)}\theta_{(m)}$ → a 2-D spline curve with m control points

$$(\hat{m}, \hat{\theta}_{(\hat{m})})_{MDL} = \arg \min_{m,\boldsymbol{\theta}_{(m)}} \{-\log p(\mathbf{g} \mid \theta_{(m)}) + L(\theta_{(m)})\}$$

$$p(\mathbf{g} \mid \theta_{(m)}, \phi_{in}, \phi_{out}) = \prod_{i \in \text{inside}(\mathbf{f}_{(m)})} p(g_i \mid \phi_{in}) \prod_{i \in \text{outside}(\mathbf{f}_{(m)})} p(g_i \mid \phi_{out})$$

Contour estimation problem

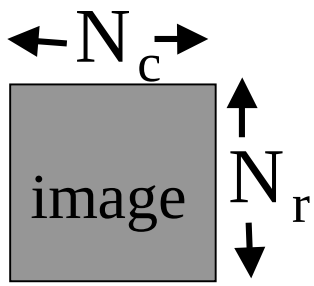
Regarding ϕ_{in}, ϕ_{out}

- minimize also w.r.t. them
- adopt a (vague, conjugate) prior, and integrate them out

What about $L(\theta_{(m)})$?

Natural code length for m control points

$$L(\theta_{(m)}) = m[\log(N_r) + \log(N_c)] = m \log(N)$$



$\left\| \mathbf{B}_{(m)} \right\|_\infty = 1$ because of the partition of unity of B-splines

quantizing control points at pixel resolution guarantees also pixel resolution for the contour

Contour estimation problem

For each m, $\hat{\theta}_{(m)}^{ML}$ is obtained by a gradient-projection-type algorithm

Then,

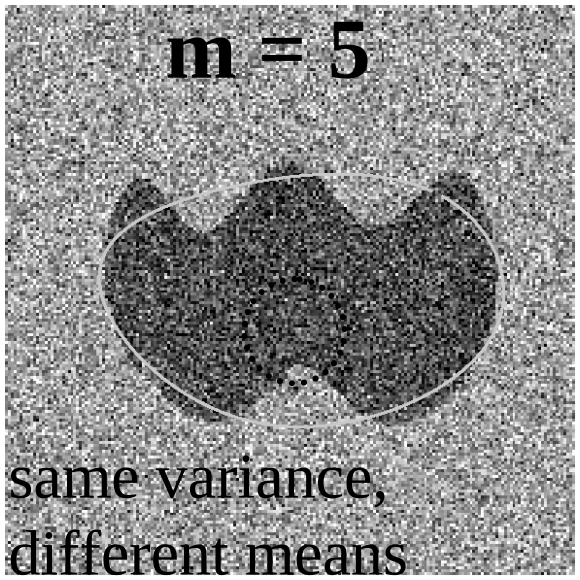
$$\hat{m}_{MDL} = \arg \min_{m} \{ -\log p(\mathbf{g} \mid \hat{\theta}_{(m)}^{ML}) + m \log(N) \}$$

Examples:

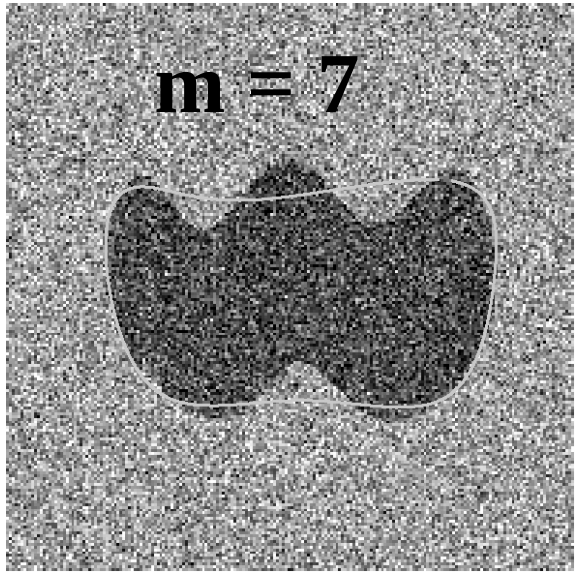
- synthetic images: Gaussian regions of different mean and variance
- ultrasound images: Rayleigh regions of different variance
- NMR: Riccian regions of different mean and variance

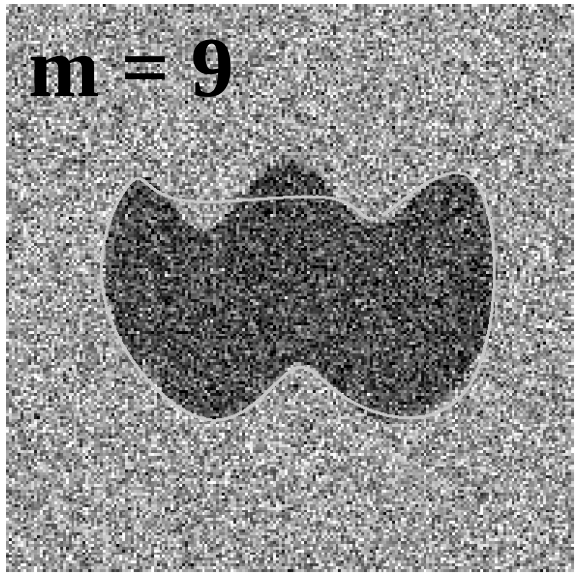
See: M. Figueiredo, J. Leitão, and A. Jain, "Adaptive B-splines and boundary estimation",in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-97), pp. 724-729, San Juan, Puerto Rico, 1997.

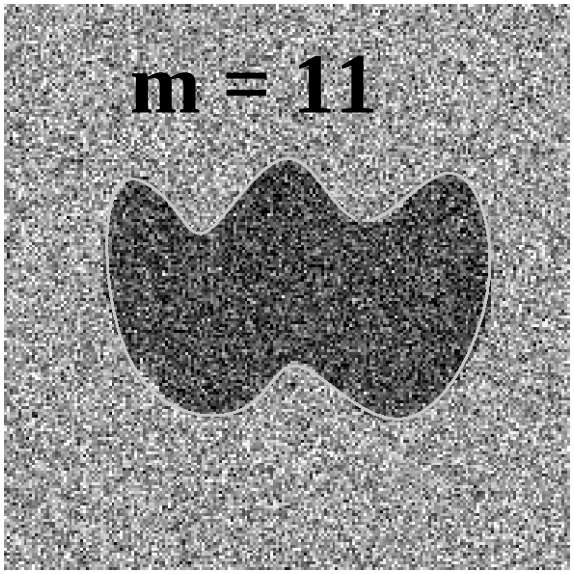
Contour estimation examples

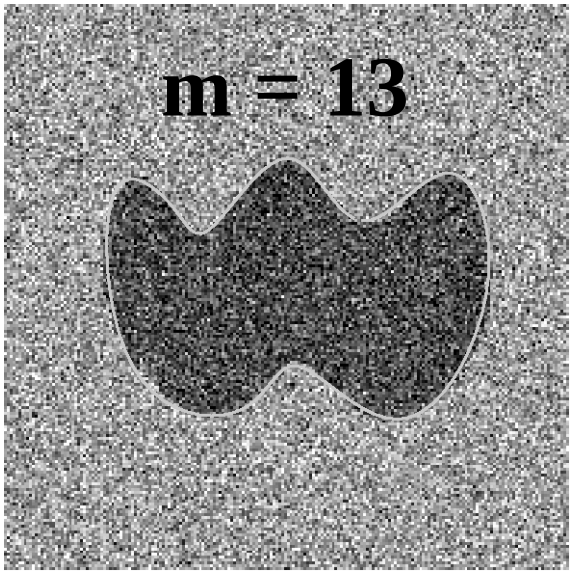


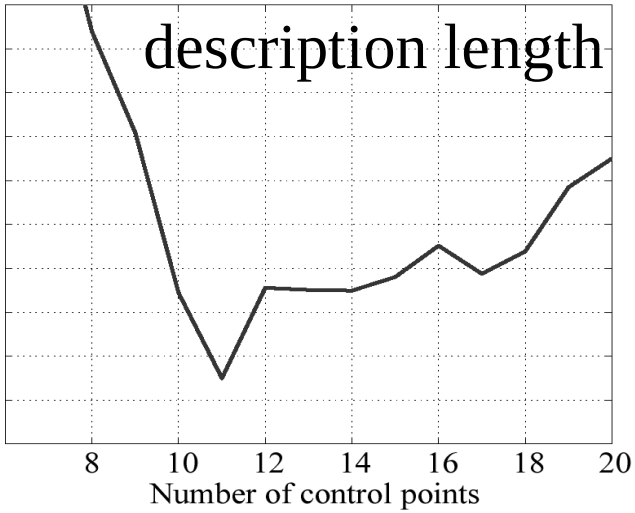
same variance,
different means





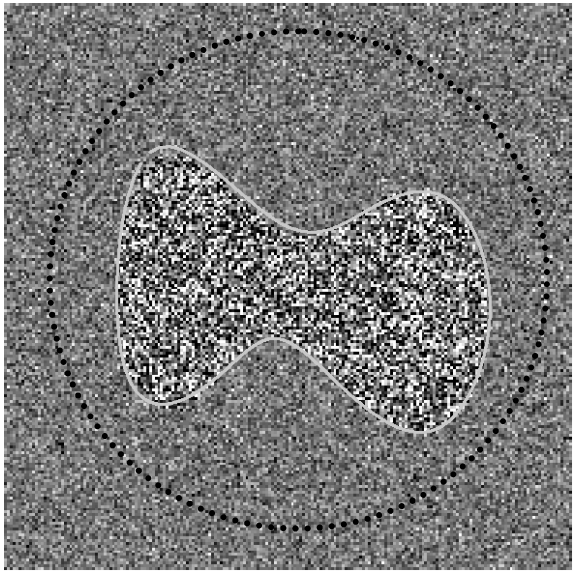




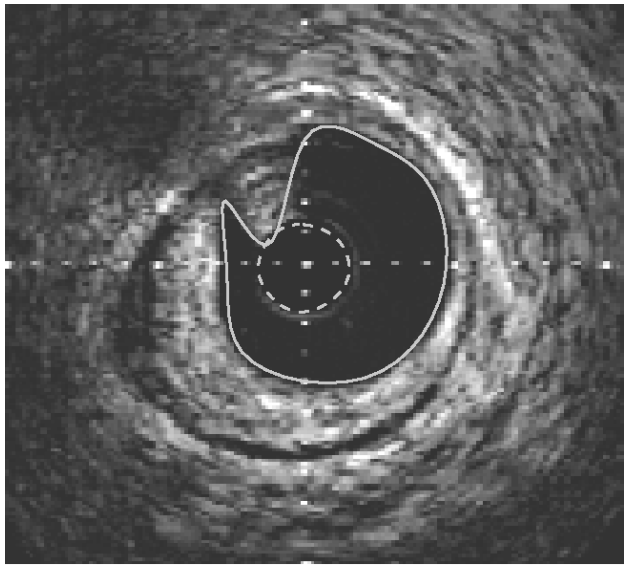


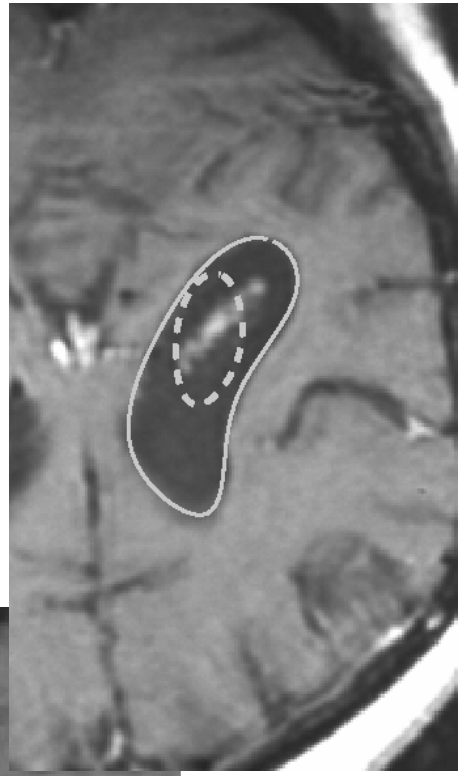
Contour estimation examples

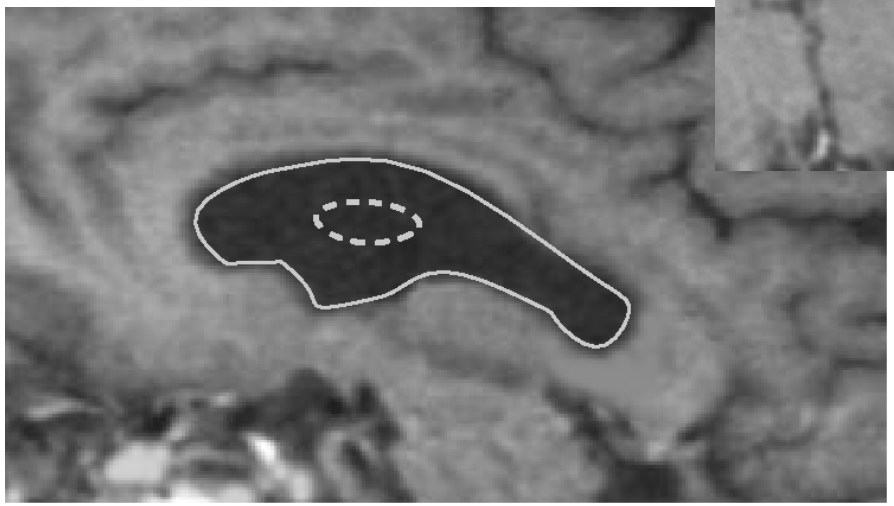
same mean,
different variances



initializations







The incompleteness issue in MDL

Classical MDL (Rissanen, 1978)

$$(\hat{m}, \hat{\mathbf{f}}_{(\hat{m})})_{\text{MDL}} = \arg \min_{m, \mathbf{f}_{(m)}} \{ \frac{m}{2} \log(n) - \log p(\mathbf{g} \mid \mathbf{f}_{(m)}) \}$$

Problems: not valid for small samples (asymptotic criterion);
assumes a redundant (called *incomplete*) code (Rissanen, 1996).

Incompleteness: in code-length $-\log p(\mathbf{g} \mid \mathbf{f}_{(m)})$

1. Transmitter estimates $\mathbf{f}_{(m)}$ from $\mathbf{g}$
2. Transmitter sends $\hat{\mathbf{f}}_{(m)}$ to receiver
3. Now, receiver knows that the only possible **g**'s are those that can lead to that $\hat{\mathbf{f}}_{(m)}$
   Code only needed for those **g**'s

The incompleteness issue in MDL

Example: $\mathbf{g} = \{g_1, g_2, \ldots, g_{10}\}$ $\quad g_i \in \{0,1\} \Rightarrow \mathbf{g} \in \{0,1\}^{10}$

10 i.i.d. observations of a Bernoulli source

$\text{Prob}(g_i=1) = f$

1. Transmitter computes $\hat{f} = \frac{1}{10}\sum_{i=1}^{10} g_i$

2. Transmitter sends $\hat{f}$ to receiver (rational number)

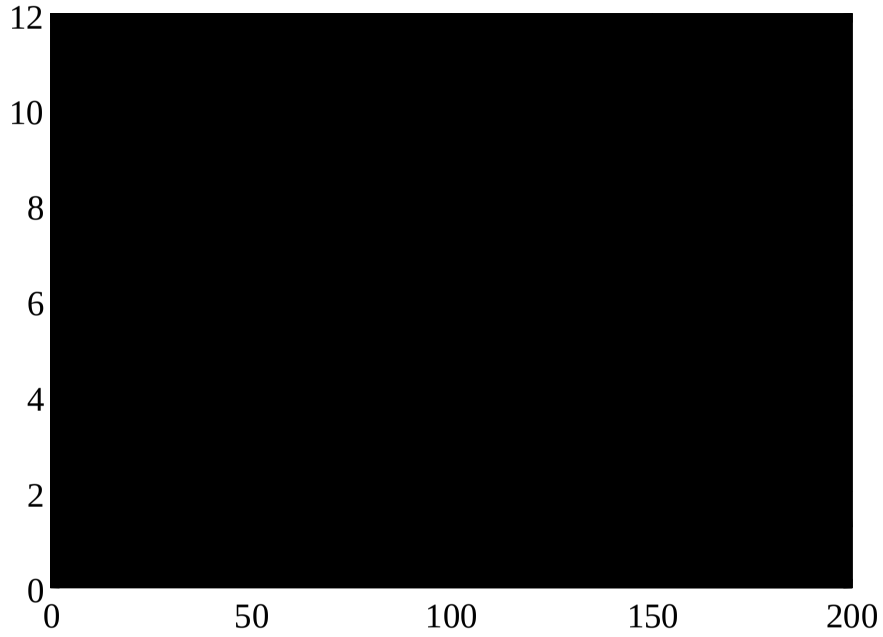
3. Not all **g**'s need be coded.
   Only those that have $10\,\hat{f}$ ones: $L(\mathbf{g}) = \log\binom{10}{10\,\hat{f}}$

Concrete example: $\mathbf{g} = \{1,0,0,0,1,0,1,0,0,0\}$, 3 ones $\longrightarrow \hat{f} = \frac{3}{10}$

$\log_2\binom{10}{3} = 6.9$ bits versus $-\log_2 p(\mathbf{g}|\hat{f}) = -\log_2[0.3^3(1-0.3)^7] \approx 8.8$ bits

An MDL criterion to segment Poisson data

Basic question: given a sequence of independent Poisson counts



$$\mathbf{g} = \{g_1, \ldots, g_N\}$$

Is there a change in the parameter ?
If so, where ?

We have N+1 models:

$$m_0 : p(\mathbf{g} \mid \lambda) = \prod_{i=1}^{N} \text{Poisson}(g_i \mid \lambda)$$

$$m_1, \ldots, m_N : p(\mathbf{g} \mid \lambda_a, \lambda_b, m_k) = \prod_{i=1}^{k} \text{Poisson}(g_i \mid \lambda_a) \prod_{i=k+1}^{N} \text{Poisson}(g_i \mid \lambda_b)$$

An MDL criterion to segment Poisson data

Fact: $g_1,\ldots,g_N$ be N i.i.d. Poisson counts (arbitrary parameter).

Let $t = \sum g_i$ be the total count

Conditioned on t, the g_i's follow a multinomial distribution

$$p(g_1,...,g_N \mid t) = \text{Multi}(g_1,...,g_N \mid t; \frac{1}{N},...,\frac{1}{N})$$

"each g_i is expected to get a 1/N fraction of the total count t"

What if $g_1,...,g_i \sim$ i.i.d. Poisson(λ_a) $g_{i+1},...,g_N \sim$ i.i.d. Poisson(λ_b)

$$p(g_1,...,g_N \mid t) = \text{Multi}(g_1,...,g_N \mid t; \underbrace{\frac{2\rho}{N},...,\frac{2\rho}{N}}_{1 \text{ to } k}, \underbrace{\frac{2(1-\rho)}{N},...,\frac{2(1-\rho)}{N}}_{k+1 \text{ to } N})$$

$$\boxed{\rho = \frac{\lambda_a}{\lambda_a + \lambda_b}}$$

An MDL criterion to segment Poisson data

Approach: rather than base the test on the Poisson model,
condition on the total count, and use the multinomial

The N+1 models:

$$m_0 : \; p(g_1,\ldots,g_N \mid t) = \binom{t}{g_1\; g_1 \cdots g_N} \left(\frac{1}{N}\right)^t$$

$$m_k : \; p(g_1,\ldots,g_N \mid t) = \binom{t}{g_1\; g_1 \cdots g_N} \left(\frac{2\rho_k}{N}\right)^{\sum_{i=1}^{k} g_i} \left(\frac{2(1-\rho_k)}{N}\right)^{\sum_{i=k+1}^{N} g_i}$$

Note that m_0 is nested into all the m_k's

$$\lambda_a = \lambda_b \;\Rightarrow\; \rho = \frac{\lambda_a}{\lambda_a + \lambda_b} = \frac{1}{2} \;\Rightarrow\; \frac{2\rho}{N} = \frac{2(1-\rho)}{N} = \frac{1}{N}$$

An MDL criterion to segment Poisson data

Our MDL approach:

1. given $\mathbf{g} = \{g_1, ..., g_N\}$ compute $t = \sum g_i$
2. transmit t (e.g., with an arbitrary technique, e.g. Elias code)
3. compute the description length under model m_0

$$m_0 : \; L(\mathbf{g} \mid t, m_0) = -\log \binom{t}{g_1 \; g_1 \cdots g_N} + t \log N$$

This is the total description length, because there is no parameter

4. compute the description length under models $m_1, \ldots, m_N$

An MDL criterion to segment Poisson data

4. compute the description length under models $m_1,\ldots,m_N$

Under m_k: -log(Multinomial with two parameters: $\frac{2\rho_k}{N}$ and $\frac{2(1-\rho_k)}{N}$)

First thing: estimate ρ_k : $\hat{\rho}_k = \frac{1}{t}\sum_{i=1}^{k} g_i = \frac{s_k}{t}$

Code length for $\hat{\rho}_k$?

Since t was already sent, we only need to encode s_k

$s_k \in \{0,1,2,\ldots,t\} \longrightarrow \boxed{L(\rho_k) = L(s_k) = \log(t+1)}$

Should we now plug $\hat{\rho}_k$ in the multinomial above ?

Attention ! Incompleteness. Receiver already knows s_k

An MDL criterion to segment Poisson data

Best code for $\mathbf{g} = \{g_1,...,g_N\}$

…knowing that $\sum_{i=1}^{k} g_i = s_k$ ..and consequently that $\sum_{i=k+1}^{N} g_i = t - s_k$

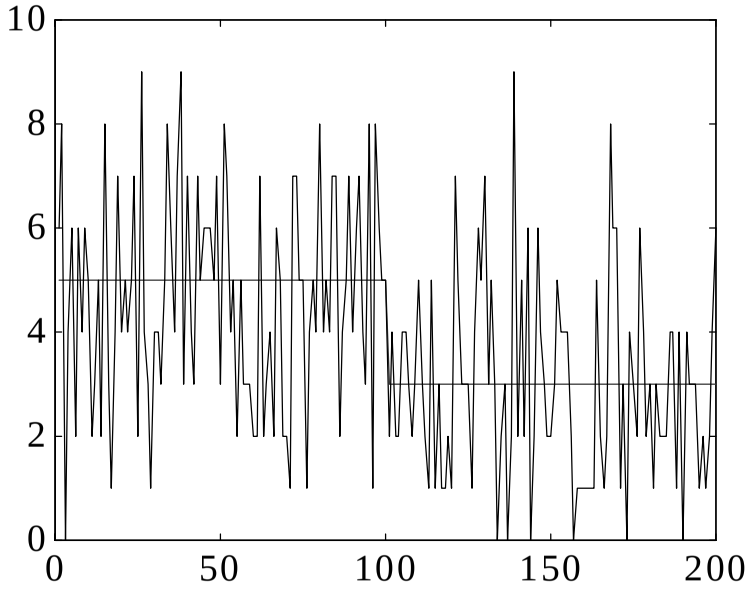
$$m_k:\ L(\mathbf{g}\,|\,t, m_k, s_k) = -\log \text{Multi}(g_1,...,g_k\,|\,s_k, \frac{1}{k},...,\frac{1}{k}) - \log \text{Multi}(g_{k+1},...,g_N\,|\,(t-s_k), \frac{1}{N-k},...,\frac{1}{N-k})$$

Finally, we choose the shortest of the N+1 total description lengths

$$L_0 = L(\mathbf{g}\,|\,t, m_0) \qquad L_k = \log(t+1) + L(\mathbf{g}\,|\,t, m_k, s_k), \quad k = 1,..,N$$

Surprisingly (or maybe not): this MDL criterion coincides with the Bayes model selection criterion with a uniform prior on $\rho \in [0,1]$

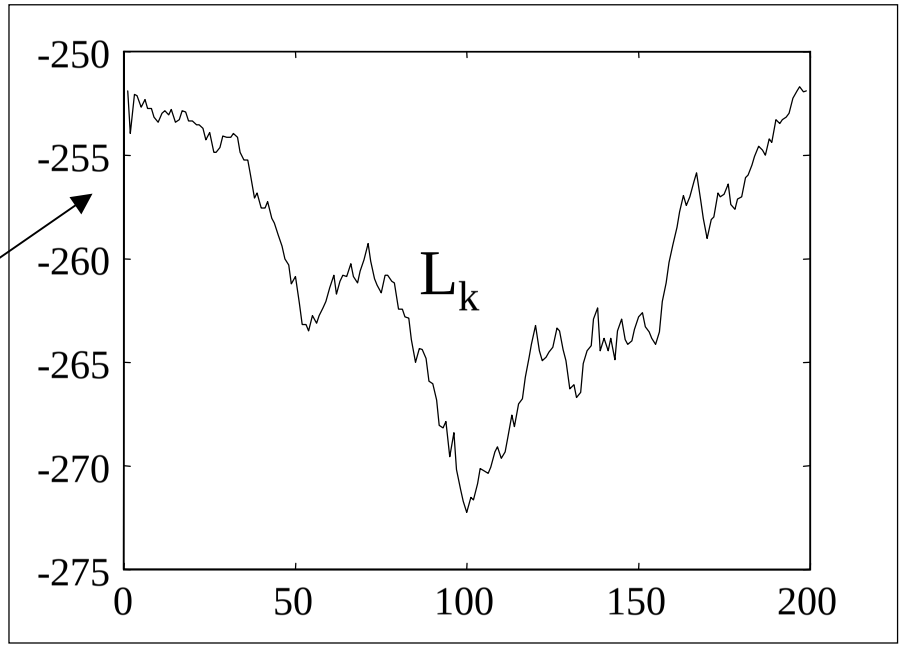
An MDL criterion to segment Poisson data



Minimum at k = 100
the correct vale (in this case)

(descr. Length - const)

L_0= -256



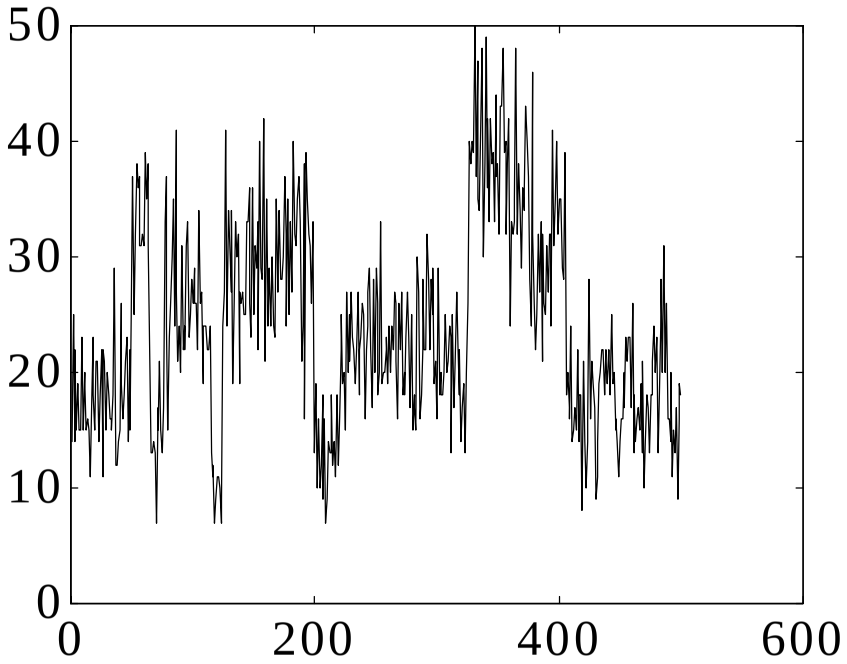
An MDL criterion to segment Poisson data

What about multiple change points ?

Simply take each segment and apply the criterion again (recursively)

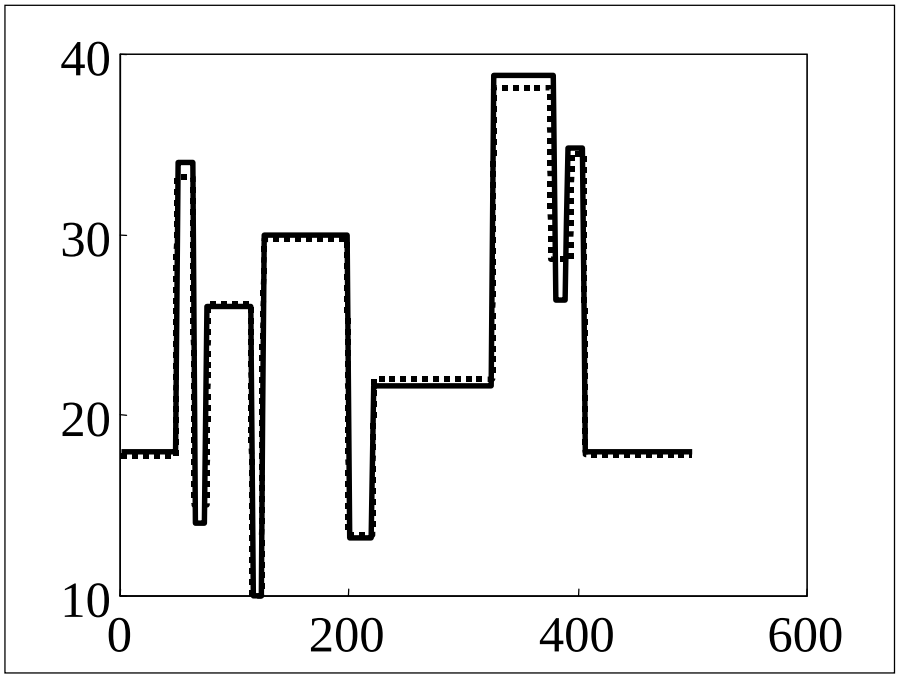
Example:

Observed counts



true intensity function

estimate



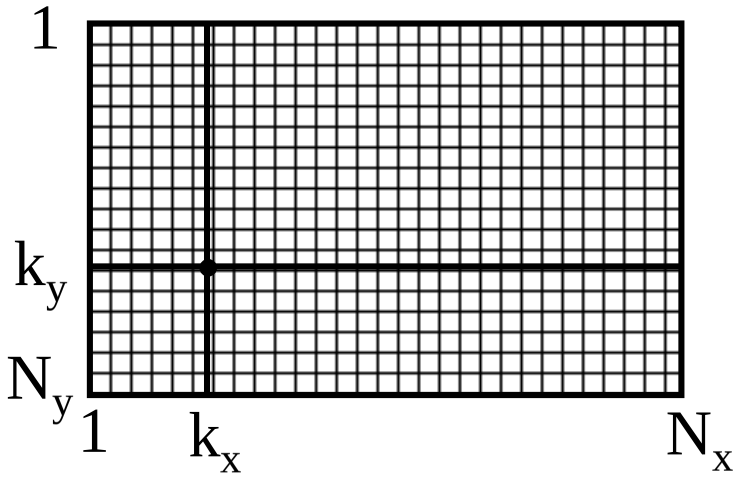
See: R. Nowak and M. Figueiredo, "Unsupervised progressive parsing of Poisson fields using minimum description length criteria",in Proceedings of the IEEE International Conference on Image Processing (ICIP-99), Kobe, Japan, October 1999 (to appear).

Segmenting Poisson images

Basic building block:

In 1D, we looked for the best (if any) change point

In 2D, we look for the best (if any) segmentation into two/four rectangles of the following type:



Same multinomial-based criterion:

- condition on the total count
- number of competing models:

$(1+(N_x+1)(N_y+1))$

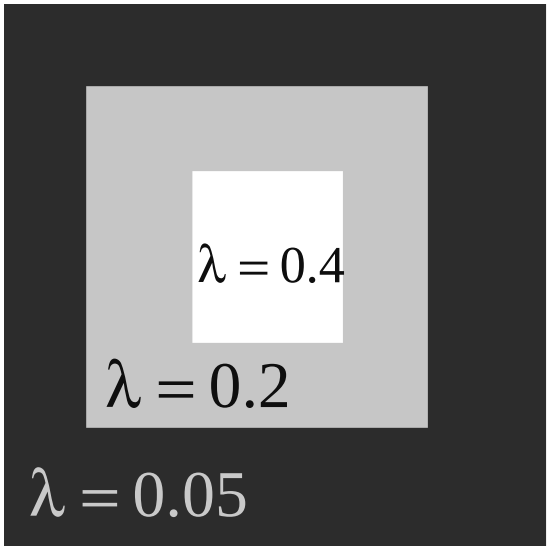
no segmentation,
all possible 4-segmentations,
and all possible 2-segmentations

To fully segment an image: apply criterion recursively.

Segmenting Poisson images: an example

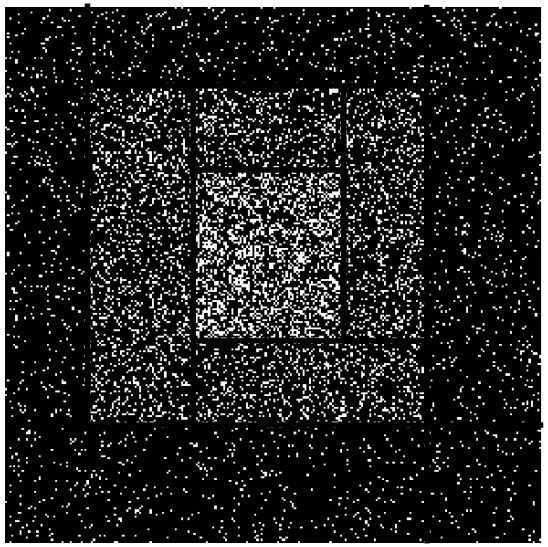
Example:

250*250 pixels/bins



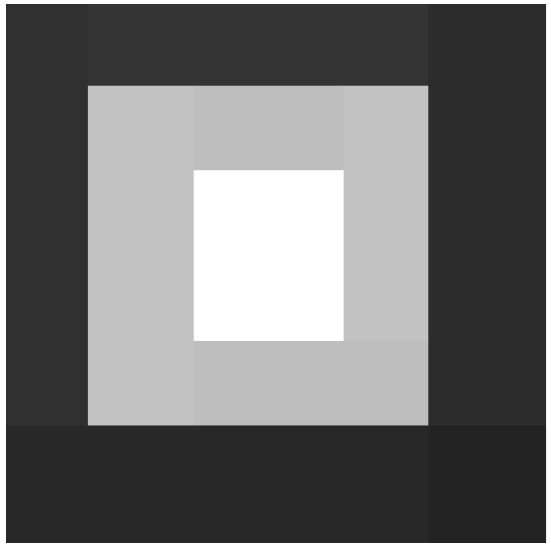
True intensity λ

counts



segmentation

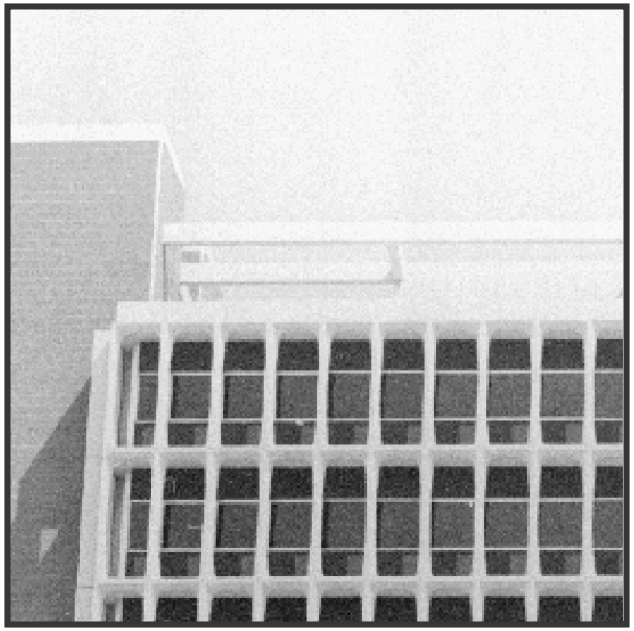
intensity estimate $\hat{\lambda}$



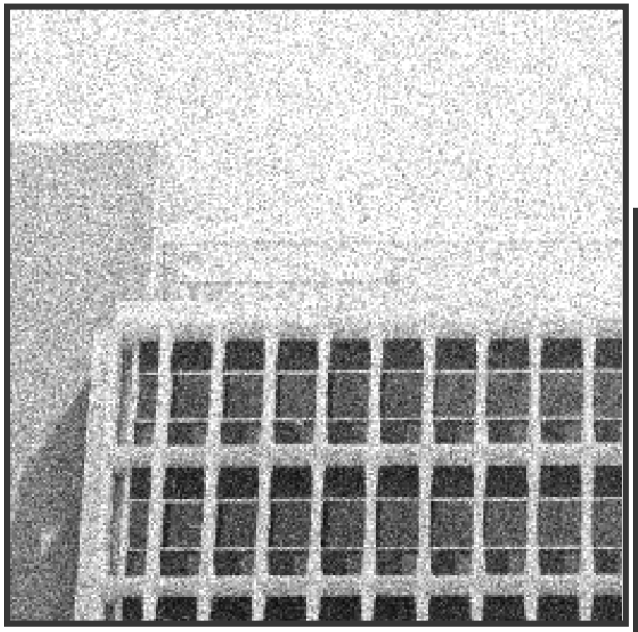
See: R. Nowak and M. Figueiredo, "Unsupervised progressive parsing of Poisson fields using minimum description length criteria",in Proceedings of the IEEE International Conference on Image Processing (ICIP-99), Kobe, Japan, October 1999 (to appear).

Segmenting Poisson images: another example

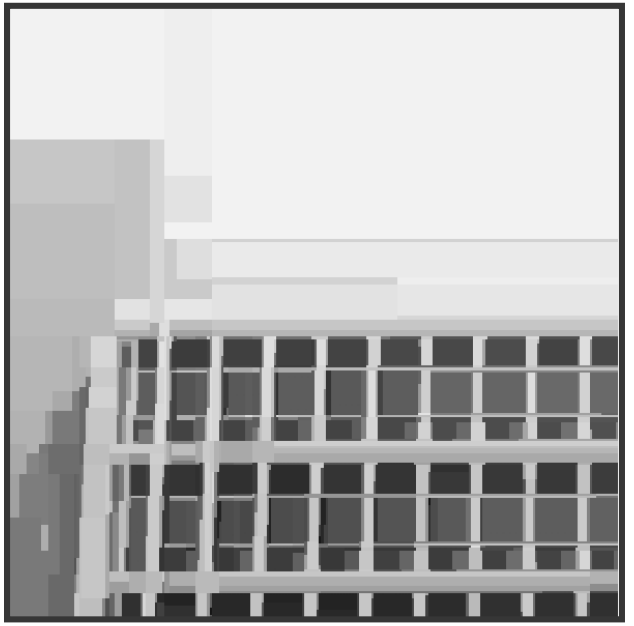
True intensity λ



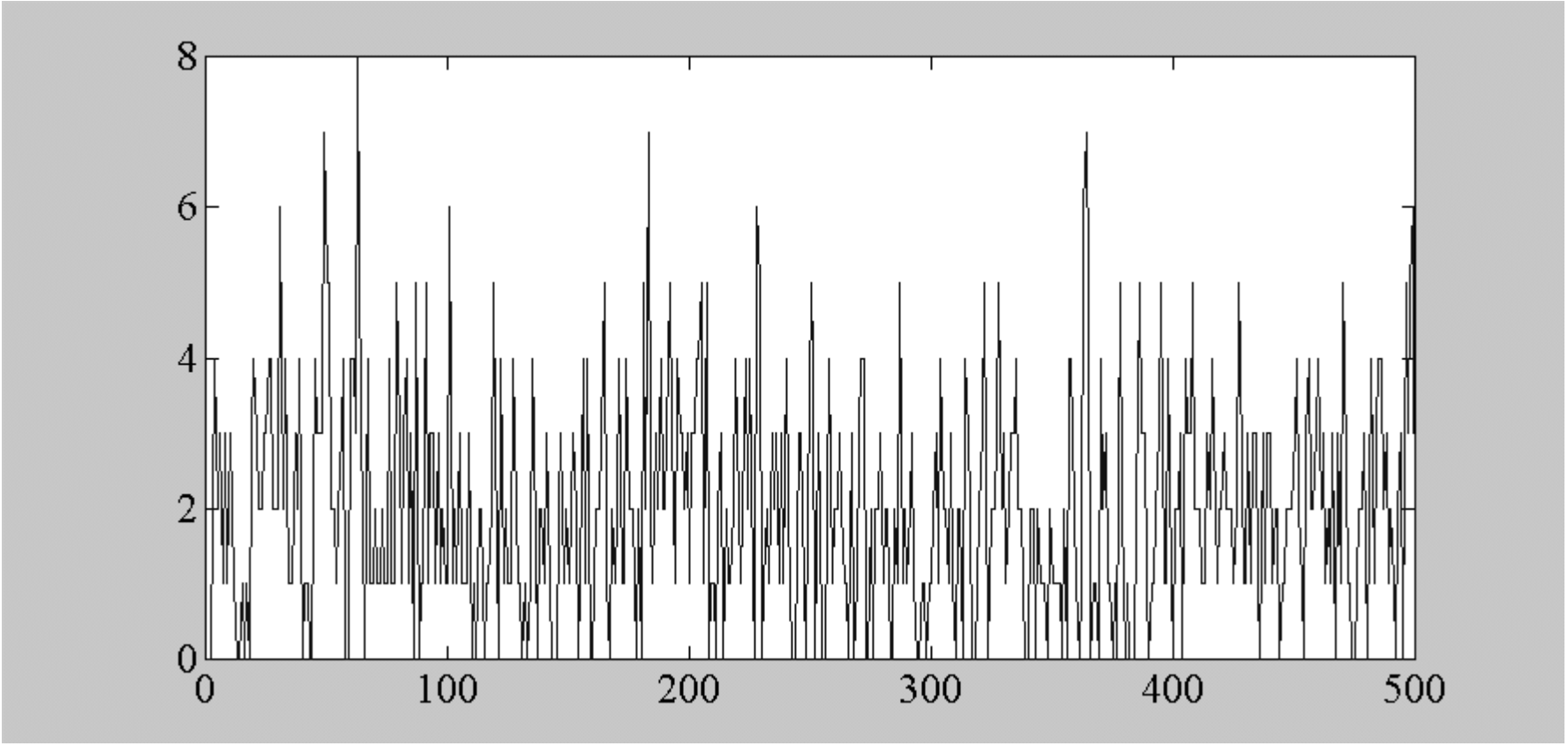
counts



intensity estimate $\hat{\lambda}$



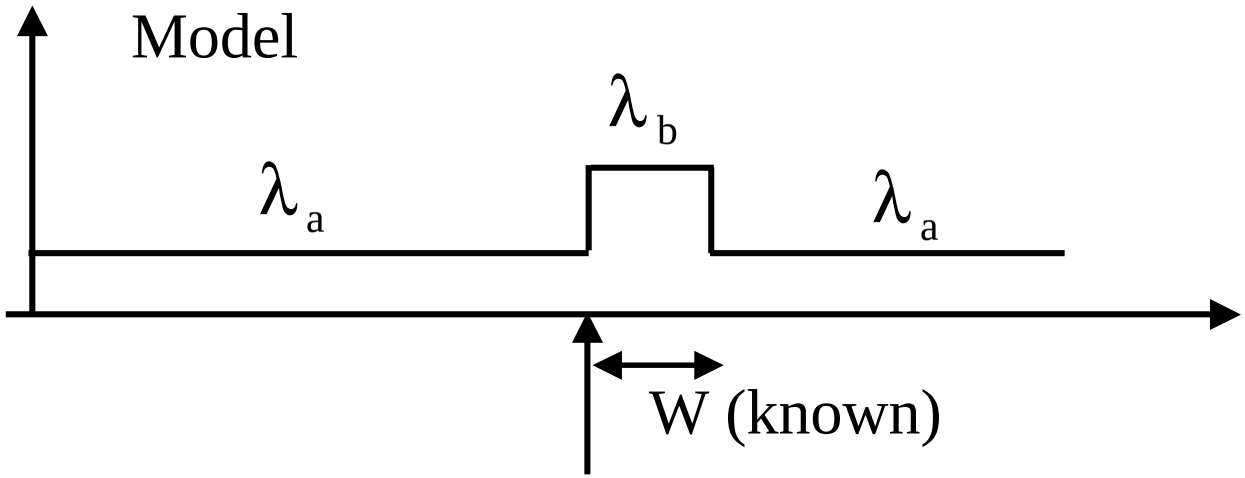
Singularities in Poisson data (“bursts”)



Question: given a sequence of Poisson observations,

...is there a “small” region of higher intensity than the background ?

Singularities in Poisson data (“bursts”)



λ_a, λ_b unknown (of course)

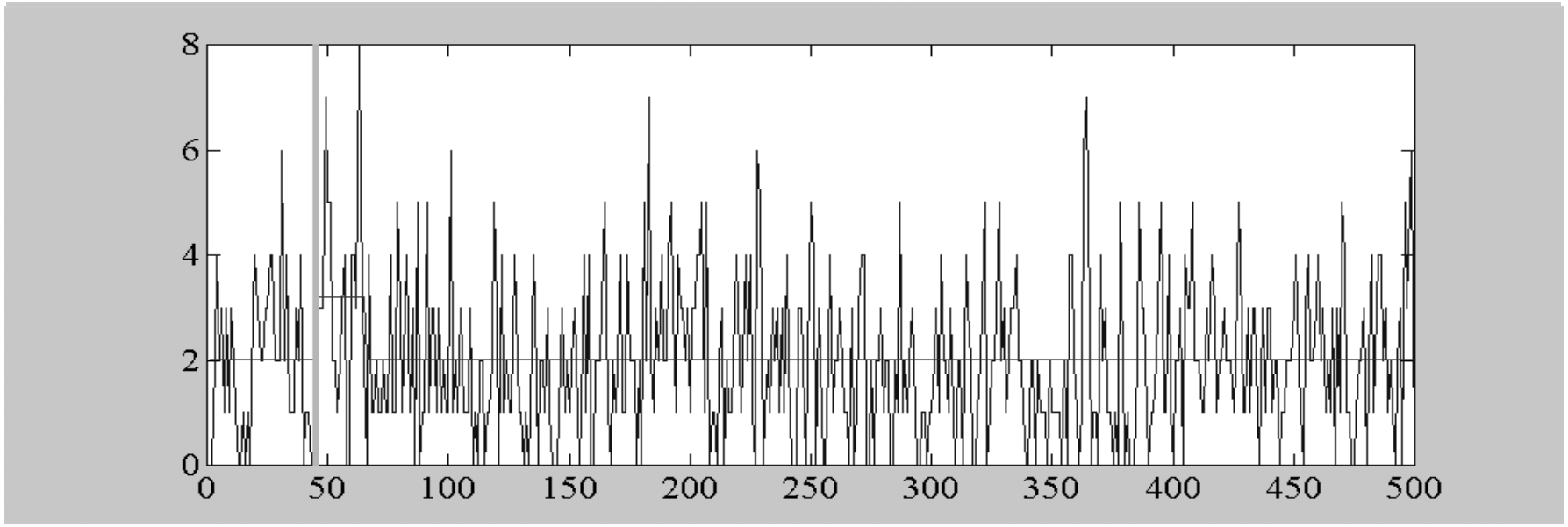
Approach as above: condition on the total count,
work with the multinomials

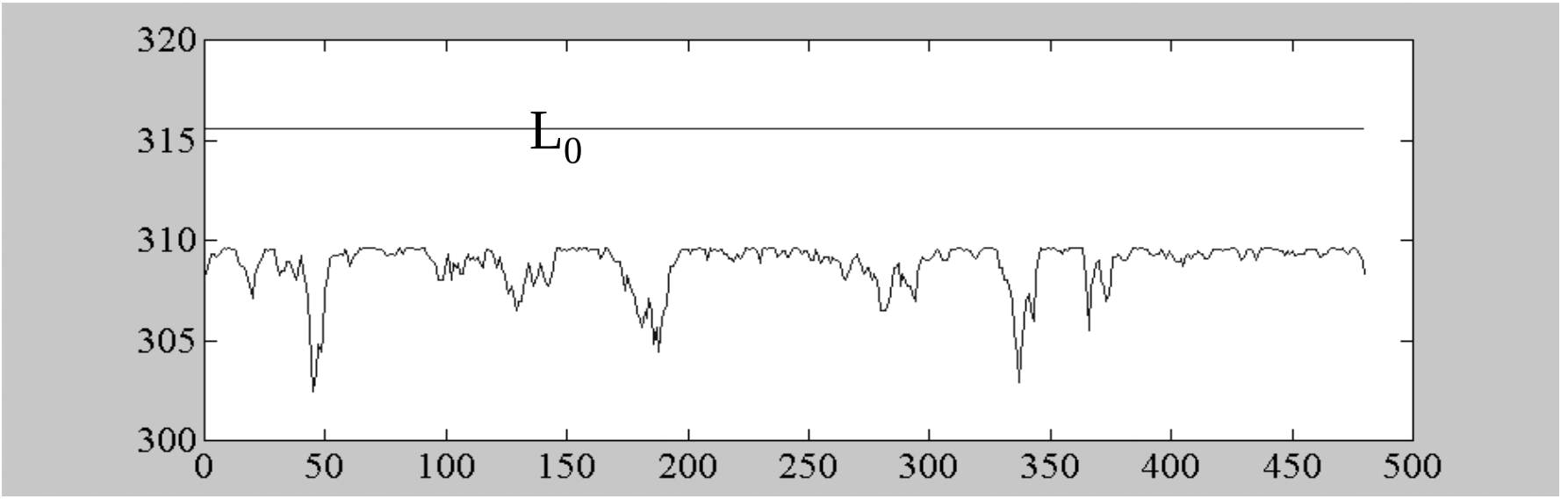
Again, equivalent to Bayes criterion with flat prior on $\rho = \dfrac{\lambda_a}{\lambda_a + \lambda_b}$

Singularities in Poisson data (“bursts”): example

$\lambda_a = 2$

$\lambda_b = 3$

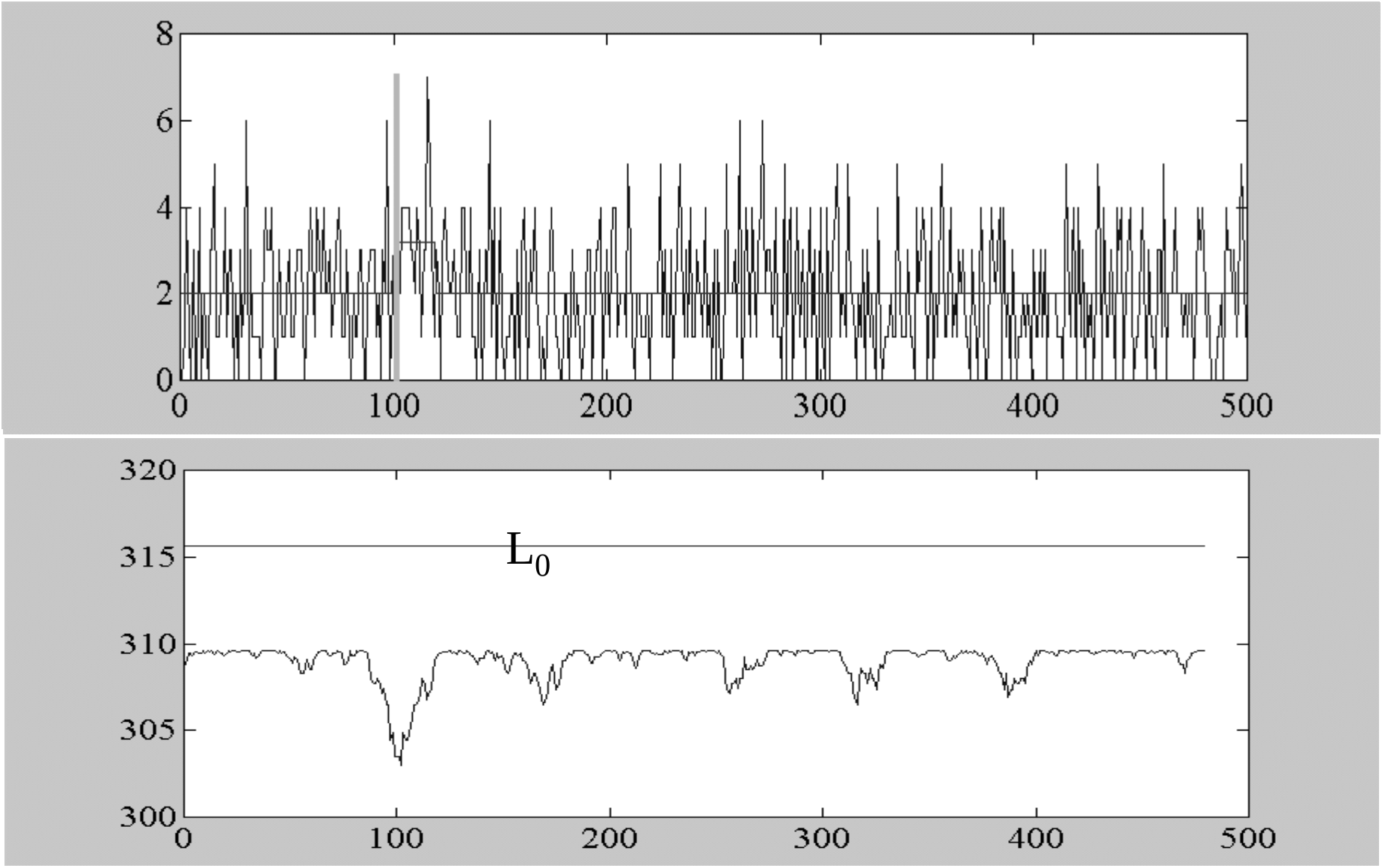




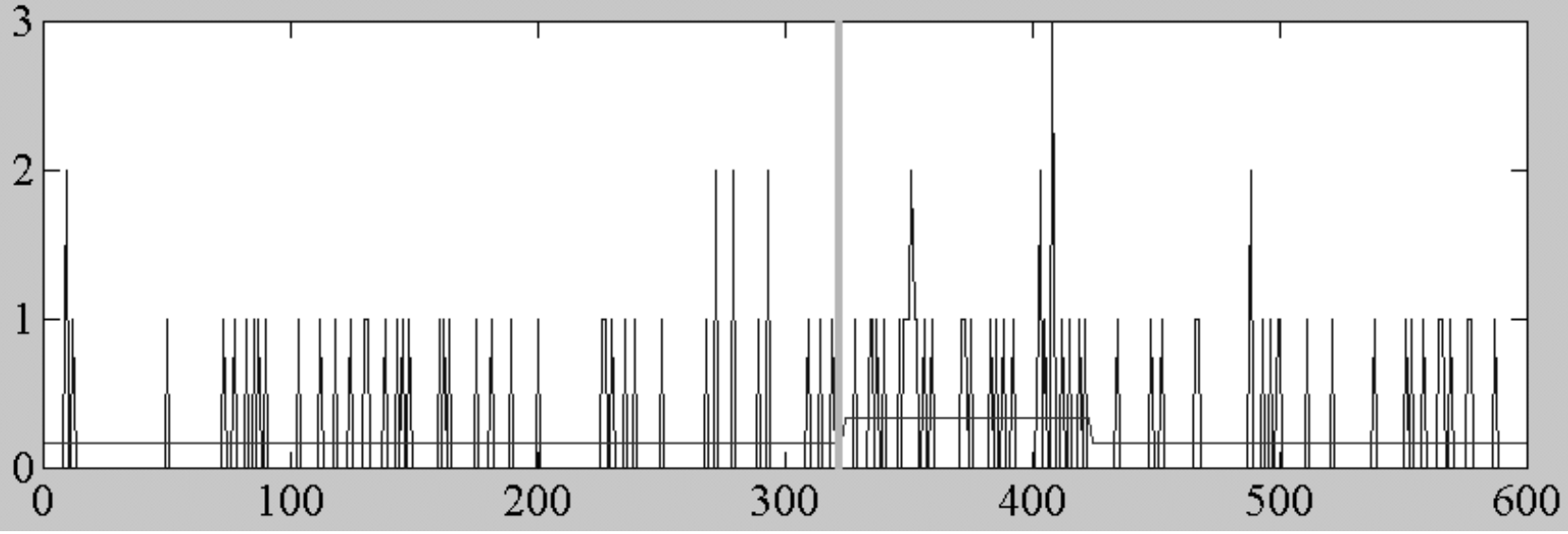
Singularities in Poisson data (“bursts”): example

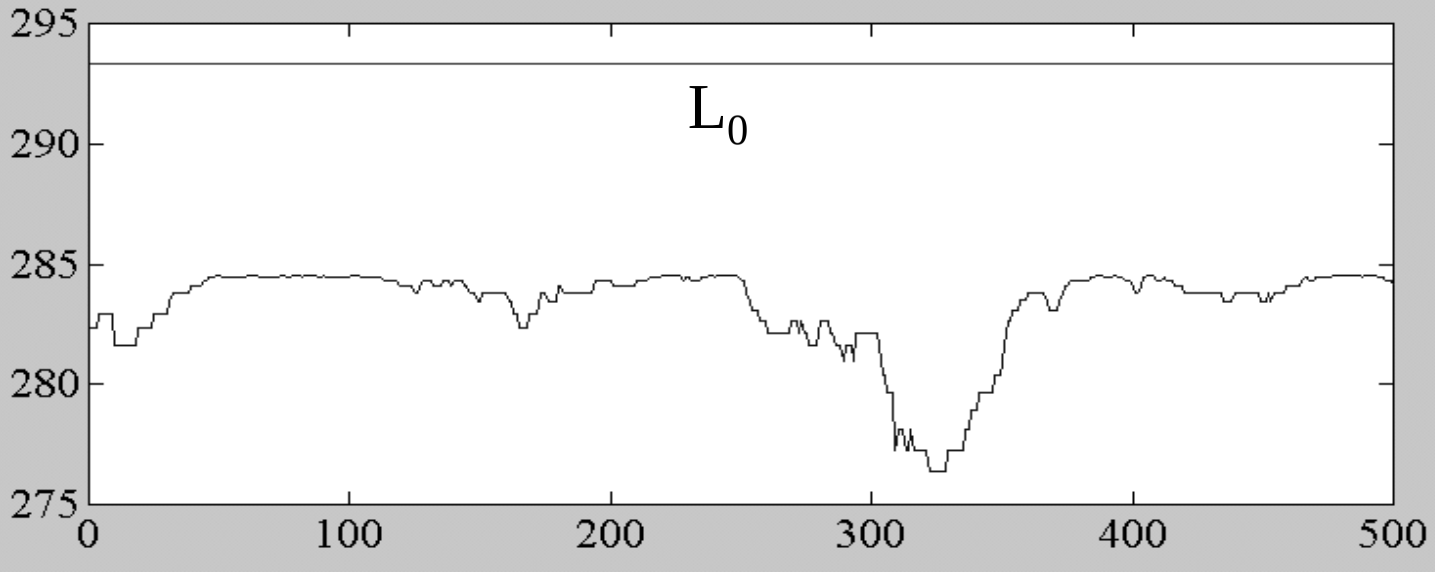
$\lambda_a = 2$

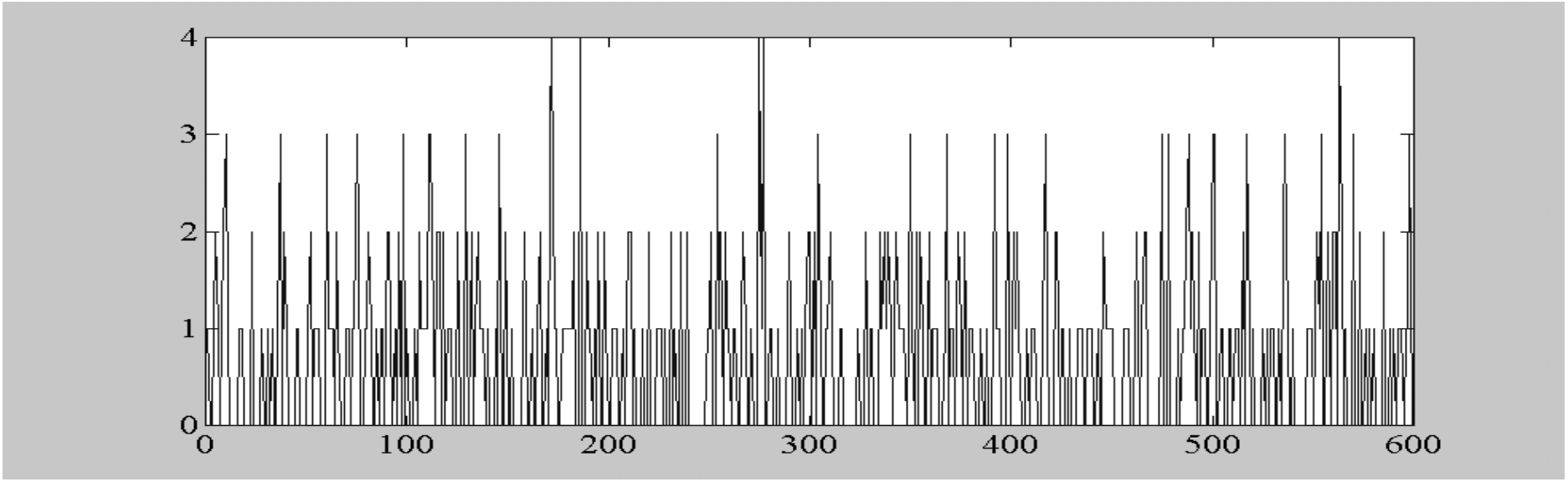
$\lambda_b = 3$

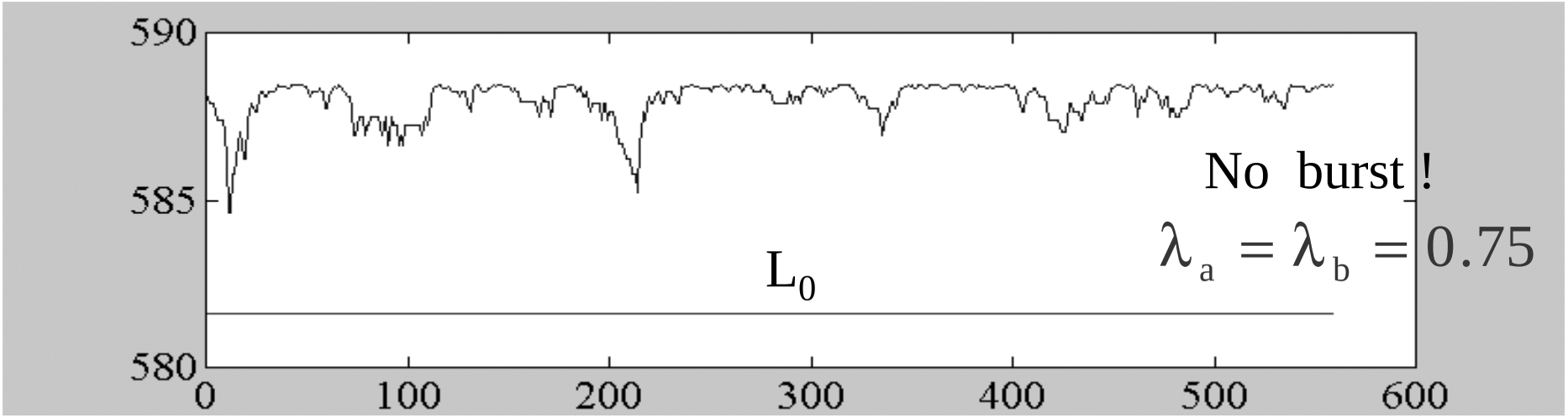


Singularities in Poisson data ("bursts"): example, low counts






4
3
2
1
0
0
100
200
300
400
500
600


590
585
580
0
100
200
300
400
500
600
No burst !
$\lambda_a = \lambda_b = 0.75$
L_0

Concluding remarks

- Most image analysis problems can/should addressed as statistical (namely Bayesian) inference.

- Many such problems involve model selection issues. Available approaches:

  Bayesian model selection
  Minimum description length (MDL) model selection
  Cross-validation / bootstrapping methods are computationally too heavy

- We addressed 3 examples of model selection problems in image analysis using MDL: image restoration, contour estimation, Poisson image segmentation

- In many applications, a natural code (rather than an optimal one) yields excellent results.

- If correctly formulated, MDL coincides with Bayesian model selection using reference priors.