Error Probabilities for Local Extrema in Gene Expression Data

Perry Groot,[1] Christian Gilissen,[2]
Michael Egmont-Petersen[2]

[1]Institute for Computing and Information Sciences, Radboud University Nijmegen
perry@cs.ru.nl
[2]Department of Human Genetics, UMC St. Radboud Nijmegen
{C.Gilissen,M.EgmontPetersen}@antrg.umcn.nl

LaQuSo Lunchcolloquium, 25 Jan 2007

Introduction

- All biological knowledge is encoded in the DNA
- DNA consists of A,C,T,G nucleotides
- Human DNA forms a helix of two strands (complementary), roughly 1.3×10^9 base pairs.
- Genes are regions in the DNA, e.g., encoding proteins
- Gene expression is translation of genes into proteins (transcription, splicing, translation)

Introduction

- Identification of functional relations among genes is an important challenge in bioinformatics

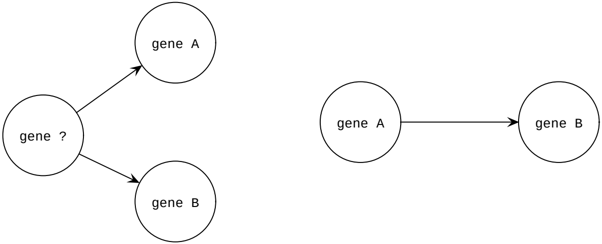
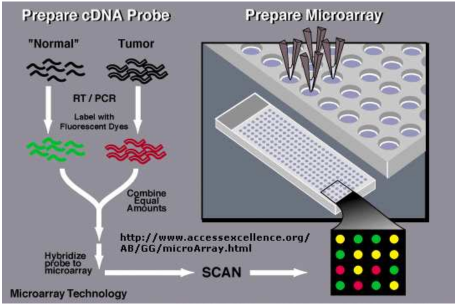


Figure: Left: co-regulation. Right: control regulation.

Introduction

- Microarray measures mRNA levels indicating gene expression levels, but with a lot of noise



Introduction

- Local extrema invariant under *scaling* and *vertical shifting*
- Local extrema significant feature [Gilissen et al.]

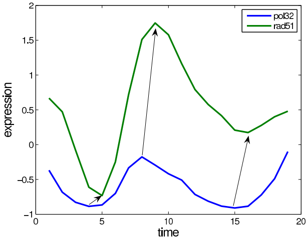


Figure: Functional relations between gene expression profiles.

Outline

- Scale-space theory
- Local Extrema in Scale-Space
- Bivariate Gaussian Integration
- Empirical Validation

Outline

- Scale-space theory
- Local Extrema in Scale-Space
- Bivariate Gaussian Integration
- Empirical Validation

Scale-space theory

- Multi-resolution representation from computer vision
- Given a one-dimensional continuous signal $f : T \rightarrow \mathbb{R}$ and continuous scale parameter $s \in \mathbb{R}^+$, derives a family of signals $L : T \times \mathbb{R}^+ \rightarrow \mathbb{R}$ such that
  - For each s a linear shift-invariant smoothing, i.e., a convolution with a kernel

$$K * f = \int_{-\infty}^{\infty} K(\tau) \cdot f(t - \tau)\, d\tau$$

  - $s = 0$ original signal: $L(t; 0) = f(t)$
  - Increasing s corresponds with less structure

$$s_1 \leq s_2 \text{ implies } \#_{extrema} L(t; s_1) \leq \#_{extrema} L(t; s_2)$$

  - Each s same domain as f: $L(\cdot; s) : T \rightarrow \mathbb{R}$

Scale-space theory

- For continuous signal, Gaussian kernel is *unique* solution
- Observing local extrema is equivalent to finding zero-crossings of the derivative
- Differentiation and convolution commute

$$\mathcal{D}(K_s * f) = K_s * (\mathcal{D}f) = (\mathcal{D}K_s) * f$$

- Continuous scale-space representation

$$\begin{aligned} L(t;s) &= (\mathcal{D}K_s * f)(t) \\ &= \int_{-\infty}^{\infty} \frac{-\sqrt{2}}{2s^3\sqrt{\pi}} \tau\, e^{-\tau^2/2s^2} \cdot f(t-\tau)\, d\tau \end{aligned}$$

Scale-space theory

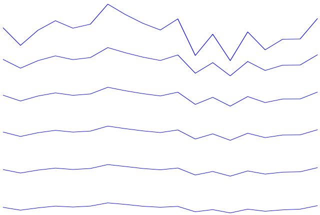


Figure: Signal along the dimension of the scale parameter. A larger scale parameter results in a smoother signal, i.e., with less structure.

Outline

- Scale-space theory
- Local Extrema in Scale-Space
- Bivariate Gaussian Integration
- Empirical Validation

Local Extrema in Scale-Space

- Discretise resulting equations

$$L(t_l; s) = \sum_{n=-\infty}^{\infty} \frac{-\sqrt{2}}{2s^3\sqrt{\pi}} n\, e^{-n^2/2s^2} \cdot f(t_l - n)$$

- Conditions for zero-crossings

$$\begin{array}{ll} L(t_l; s) < 0 \text{ and } L(t_{l+1}; s) > 0 & \text{(local minimum)} \\ L(t_l; s) > 0 \text{ and } L(t_{l+1}; s) < 0 & \text{(local maximum)} \end{array}$$

Local Extrema in Scale-Space

- Add Gaussian distributed, additive noise

$$g(t_l) = f(t_l) + \epsilon, \quad \epsilon \sim N(0, \sigma_g^2)$$

- Derive distribution of $p(L(t_l; s), L(t_{l+1}; s))$,
- Variance:

$$\sigma^2(L(t_l; s)) = \sum_{n=-\infty}^{\infty} k_n^2 \, \sigma_g^2$$

- Covariance:

$$cov(L(t_l; s), L(t_{l+1}; s)) = \sigma_g^2 \left(\sum_{n=-\infty}^{\infty} k_n^2 - \frac{1}{2} \sum_{n=-\infty}^{\infty} (k_n - k_{n+1})^2 \right)$$

Local Extrema in Scale-Space

- Covariance matrix

$$\Sigma_L = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix}$$

$$\sigma_{1,1} = \sigma_{2,2} = \sigma_g^2 \sum_{n=-\infty}^{\infty} k_n^2$$

$$\sigma_{1,2} = \sigma_{2,1} = \sigma_g^2 \left(\sum_{n=-\infty}^{\infty} k_n^2 - \frac{1}{2} \sum_{n=-\infty}^{\infty} (k_n - k_{n+1})^2 \right)$$

- $p(L(t_l; s), L(t_{l+1}; s)) \sim G(\mu_L, \Sigma_L)$ with $\mu_L = (L(t_l; s), L(t_{l+1}; s))^T$.

Local Extrema in Scale-Space

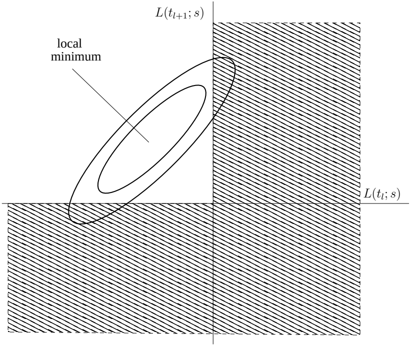


Figure: Bivariate distribution of the two time-points through which the smooth first-order derivative changes sign.

Local Extrema in Scale-Space

- Error probability for missing a local minimum

$$P_{min}(L(t_l;s),L(t_{l+1};s);\mu_L,\Sigma_L) = 1 - \int_{t_l,L(t_l;s)<0,} \int_{t_{l+1},L(t_{l+1};s)>0} G(\mu_L,\Sigma_L)\,dt_l\,dt_{l+1}$$

- Error probability for missing a local maximum

$$P_{max}(L(t_l;s),L(t_{l+1};s);\mu_L,\Sigma_L) = 1 - \int_{t_l,L(t_l;s)>0,} \int_{t_{l+1},L(t_{l+1};s)<0} G(\mu_L,\Sigma_L)\,dt_l\,dt_{l+1}$$

Outline

- Scale-space theory
- Local Extrema in Scale-Space
- Bivariate Gaussian Integration
- Empirical Results

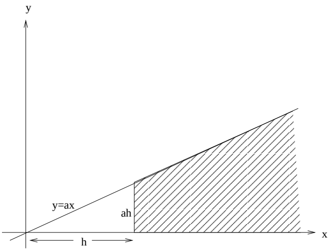
Bivariate Gaussian integration

- Fundamental formulas

$$T(h,a) = \frac{\arctan a}{2\pi} - \frac{1}{2\pi}\sum_{j=0}^{\infty} c_j a^{2j+1}$$

$$c_j = (-1)^j \frac{1}{2j+1}\left[1 - e^{\left(-\frac{1}{2}h^2\right)}\sum_{i=0}^{j}\frac{h^{2i}}{2^i i!}\right]$$

Bivariate Gaussian integration



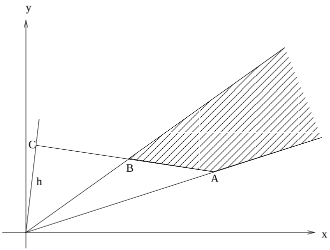


Figure: Left: area over which $T(h,a)$ gives the volume of a standardised bivariate normal with correlation zero. Right: a typical area for computing the bivariate normal integral over a polygon $(T(h,a_2) \pm T(h,a_1))$ depending on C).

Bivariate Gaussian integration

- Transformation to standard bivariate Gaussian distribution

$$u(x,y) = \frac{1}{\sqrt{2+2\rho}}\left[\frac{x-\mu_X}{\sigma_X} + \frac{y-\mu_Y}{\sigma_Y}\right]$$

$$v(x,y) = \frac{-1}{\sqrt{2-2\rho}}\left[\frac{x-\mu_X}{\sigma_X} - \frac{y-\mu_Y}{\sigma_Y}\right]$$

with

$$\mu_X = (f * K_s)(t_l), \qquad \mu_Y = (f * K_s)(t_{l+1}),$$

$$\sigma_X = \sigma_Y = \sqrt{\left(\sum_i k_i^2\right) \cdot \sigma_g^2},$$

$$\rho = \frac{\left(\sum_i k_i^2 - \frac{1}{2}\sum_j (k_j - k_{j+1})^2\right) \cdot \sigma_g^2}{\left(\sum_i k_i^2\right) \cdot \sigma_g^2}.$$

Bivariate Gaussian integration

- Formulas for obtaining parameters given two polygon vertices (u_1, v_1) and (u_2, v_2)

$$h = \frac{|u_1 v_2 - u_2 v_1|}{\sqrt{(u_2 - u_1)^2 + (v_2 - v_1)^2}}$$

$$a_1 = \frac{|u_1(u_2 - u_1) + v_1(v_2 - v_1)|}{|u_1 v_2 - u_2 v_1|}$$

$$a_2 = \frac{|u_2(u_2 - u_1) + v_2(v_2 - v_1)|}{|u_1 v_2 - u_2 v_1|}$$

Bivariate Gaussian integration

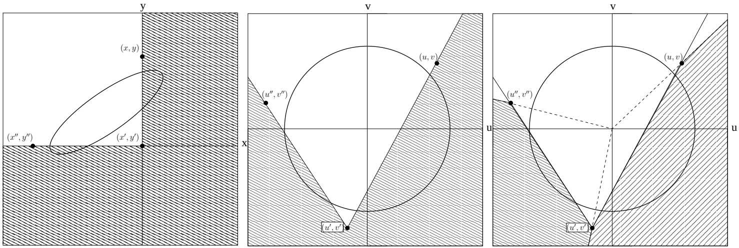


Figure: Steps for obtaining error probabilities.

Outline

- Scale-space theory
- Local Extrema in Scale-Space
- Bivariate Gaussian Integration
- Empirical Validation

Empirical Validation

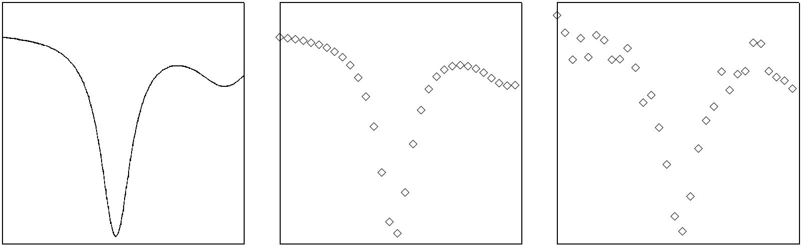


Figure: Acquisition of the model signal.

Empirical Validation

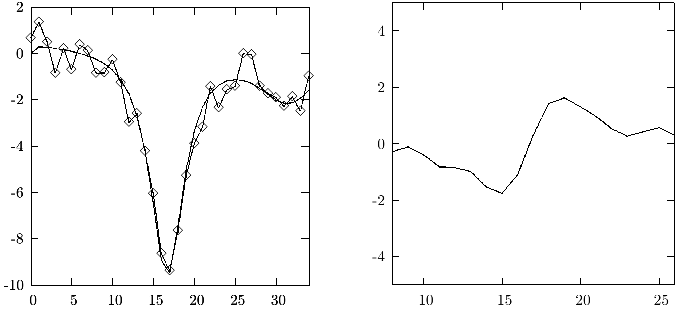


Figure: Left: Original time series f_i together with one noisy realization $g = f_i + \epsilon$, where $\epsilon \sim N(0, \frac{1}{2})$. Right: The convolved signal of a noisy time series with a discretely sampled differentiated Gaussian.

Empirical Validation

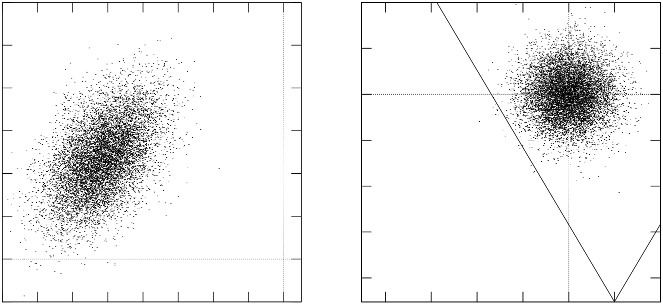


Figure: Left: Empirical data of a general correlated bivariate normal. Right: The same data after transformation, corresponding to an uncorrelated bivariate normal distribution with zero means and unit variances on the right. The lines in the right figure correspond with the x- and y-axis in the left figure after transformation.

Empirical Validation

- Number of experiments: 10.000
- Estimated error of probability: 0.0016
- Computed exact probability: 0.0018

Conclusions

- Starting point fundamental theory of scale-space
- Derived conditions for local extrema in scale space
- Derived bivariate Gaussian distribution of not re-observing local extrema
- Derived error probabilities in terms of integrating tails of a bivariate Gaussian distribution
- Applied techniques for bivariate Gaussian integration to obtain precise error probabilities