Error Probabilities for Local Extrema in Gene expression Data

Perry Groot [a,*], Christian Gilissen [b], Michael Egmont-Petersen [b]

[a] *Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands*
[b] *Department of Human Genetics, UMC St. Radboud Nijmegen, the Netherlands*

Abstract

Current approaches for the prediction of functional relations from gene expression data often do not have a clear methodology for extracting features and are not accompanied by a clear characterisation of their performance in terms of the inherent noise present in such data sets. Without such a characterisation it is unclear how to focus on the most probable functional relations present. In this article, we start from the fundamental theory of scale-space for obtaining features (i.e., local extrema) from gene expression profiles. We show that under the assumption of Gaussian distributed noise, repeatedly measuring a local extrema behaves like a bivariate Gaussian distribution. Furthermore, the error of not re-observing local extrema is phrased in terms of the integral over the tails of this bivariate Gaussian distribution. Using integration techniques developed in the 50s, we demonstrate how to compute these error probabilities exactly.

Key words: Statistics, Scale-space theory, Bivariate Gaussian integration, Gene expression

1. Introduction

In the last few years, various international genome projects have yielded the near complete molecular sequences of a large number of species, including human. Novel high-throughput methodologies such as microarray-based gene expression profiling are now being used to generate genome-wide transcriptomic data sets at an ever-increasing rate to analyse and monitor the effects of intrinsic and exogenous variables on living cells, tissues, and organs. In general, the timing of mRNA expression for a given gene has been found to correlate well with the function of the resultant protein (Bähler, 2005; Bozdech *et al.*, 2003).

Identification of functional relations from gene expression data has remained difficult because of the inherent noise in such data sets. Methods that have been developed to determine such relations include clustering algorithms like hierarchical clustering, K-means clustering, and self-organising maps (Eisen *et al.*, 1998; Datta and Datta, 2003). The simplest approach to clustering is to select a gene and determine the nearest neighbouring genes according to a distance measure between gene expression profiles. This approach, called hierarchical clustering, allows the clustering of groups of genes that are co-regulated. As yet, however, it is unclear how well certain distance functions can deal with noise which is inherent to gene expression data sets. Model-based approaches like dynamic Bayesian networks offer more flexible techniques that can, in principle, deal with the inherent noise of gene expression data sets (Friedman *et al.*, 2000; Husmeier, 2003). However, such model-based approaches preferably use discretised expression data mapping expression levels to some discrete representation, which raises methodological questions. The interpretation of time series gene expression data sets is complex and is still regarded to be an open problem (Storey *et al.*, 2005).

In this article, we will focus on the representation proposed by Egmont-Petersen *et al.* (2004), which only registers the local extrema of the time course gene expression. This representation focusses on the most likely time points a gene changes from up regulated to becoming down regulated or vice versa and allows the prediction of functional relations without regarding the amplitude of the signal (cf. Section 2). We will start the analysis of gene expression profiles by using a fundamental approach developed in the computer vision community, called scale-space theory (Koenderink, 1984), for analysing images and signals at multiple scales (Section 3). This allows us to formulate criteria for detecting local extrema in noisy signals. By interpreting point measurements as a stochastic process we are able to derive its exact distribution and give a characterisation of

* Corresponding author. Tel.:+31 24 3652075, Fax:+31 24 3653366.
Email addresses: perry@cs.ru.nl (Perry Groot),
C.Gilissen@antrg.umcn.nl (Christian Gilissen),
M.EgmontPetersen@antrg.umcn.nl (Michael Egmont-Petersen).

not re-observing an extremum as the result of noise and/or smoothing. More specifically, the contributions of this paper are the following:

- Under the assumption of Gaussian distributed additive noise, the repeated measurement of two subsequent points in the scale-space representation of a one-dimensional discrete signal is shown to have a bivariate Gaussian distribution (Section 4).
- Using the criteria for detecting local extrema in the scale-space representation, the probability that an extremum is not re-observed because of noise and/or smoothing is phrased in terms of integrating the tails of a bivariate Gaussian distribution that cross the horizontal and vertical axis, respectively (Section 4).
- We apply the method of (Owen, 1956) for the integration of a bivariate Gaussian distribution (Section 5) and provide an algorithm for the procedure (Section 7).

In summary, the paper *uses a methodological approach for detecting local extrema in a signal with Gaussian distributed noise that is accompanied with a precise quanitification of the measurement quality of the local extrema that can be computed with the provided algorithm*.

The remaining sections discuss in more detail the motivations behind our work (Section 2), related work (Section 6), and conclusions and further work (Section 8).

2. Motivation

One of the major research directions in bioinformatics is the identification of functional relations between gene expression profiles based on extracted features. For example, Figure 1 shows that when the pol32-profile has a local extremum at time n, the rad51-profile has a similar extremum at time $n+1$, strongly suggesting a functional relation between the pol32 and rad51 gene. However, in contrast with the smooth profiles shown in Figure 1, gene expression profiles inherently contain a lot of noise making it difficult to clearly distinguish features in such profiles. The aim of our work is to have a clear underlying methodology for obtaining features (i.e., local extrema) from gene expression profiles for predicting functional relations between genes, which is accompanied with a precise characterisation of not being able to observe the feature due to noise present.

Here, we focus on locating local extrema in gene expression profiles. By discretising gene expression levels on the basis of local extrema we focus on the most likely time points at which a gene (and eventually its associated protein) is active (local maximum) and inactive (local minimum). As local extrema are invariant under scaling and vertical shifting (dosage effects (Goldenthal *et al.*, 2004)), this representation effectively captures the global dynamics of the time-dependent data, i.e., local extrema allow the prediction of functional relations between a transcription factor with small absolute changes in expression ratio, and a target gene, because the amplitude is disregarded (Egmont-Petersen *et al.*, 2004).

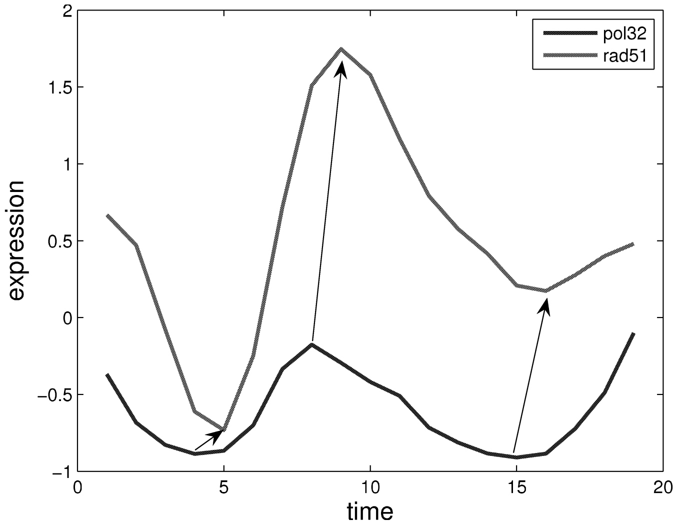


Fig. 1. Functional relations based on local extrema between gene expression profiles in yeast.

We start the analysis of local extrema from the theory of scale-space (Section 3), which allows one to analyse signals at different scales. At each scale a different amount of smoothing is applied, resulting in a simplification of the signal as spurious structures (i.e., local extrema) as the result of noise is removed. This is shown in Figure 2 for the gene expression profile of the swi4 gene along the scale-space dimension.

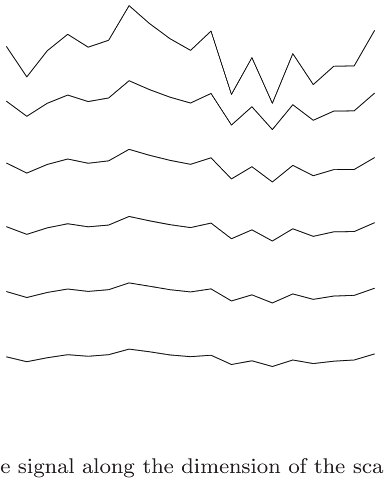


Fig. 2. The same signal along the dimension of the scale parameter. A larger scale parameter results in a smoother signal, i.e., with less structure.

After formulating criteria for detecting local extrema in the scale-space representation of a gene expression profile, the next step is to precisely characterise the probability of not detecting a local extremum due to noise. This allows one to focus on those profiles that have the most likely correct measurements when determining local extrema for the prediction of functional relations based on those local extrema.

3. Scale-space Theory

Objects in the real-world and details in images exist only at a limited range of resolution. For example, a branch of a tree only makes sense at a scale of a few centimetres up to a few meters (Lindeberg, 1990). The computer vision community has developed a multi-resolution representation, called *scale-space* theory, which allows one to analyse images at various levels of scale without choosing a scale a priori. This framework derives for some domain T, a one-dimensional continuous signal $f : T \to \mathbb{R}$, and continuous scale parameter $s \in \mathbb{R}^+$ a family of signals $L : T \times \mathbb{R}^+ \to \mathbb{R}$ that represent the original signal f such that [1]

- All representations are linear shift-invariant smoothings (i.e., generated by a *convolution* of the original signal with a kernel).
- An increasing scale parameter s corresponds with coarser levels of scale and signals with less structure (i.e., local extrema), and $s = 0$ represents the original signal, i.e., $L(t; 0) = f(t)$ and for $s \geq 0$ it holds that $L(t; s)$ has no more structure, i.e., local extrema, than $L(t; 0)$.
- Each signal in the scale-space representation is a real-valued function on the same domain as the original function, i.e., $L(\cdot; s) : T \to \mathbb{R}$.

In this article, we will regard the number of local extrema as the measure of the structure of the signals in the scale-space representation. This leads to the following definition of a scale-space kernel (Lindeberg, 1990):

Definition 3.1 *A one-dimensional kernel $K : T \to \mathbb{R}$ is denoted a scale-space kernel if for all signals $f : T \to \mathbb{R}$ the number of local extrema in the convolved signal $f' = K * f$ does not exceed the number of local extrema in the original signal.*

Many authors have shown (Iijima, 1962; Babaud *et al.*, 1986; Koenderink, 1984; Lindeberg, 1994; Witkin, 1983) that for continuous signals a unique solution exists that satisfies the scale-space axioms, i.e., the Gaussian function [2]

$$k(t; \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-(t-\mu)^2/2\sigma^2} \tag{1}$$

with mean $\mu = 0$ and standard deviation σ, the *width* of the kernel, which is some fixed value. By taking for the scale parameter $s = \sigma$ one gets a family of kernels $K_s : T \to \mathbb{R}$. For a signal f and scale parameter s, the scale-space representation $L : T \times \mathbb{R}^+ \to \mathbb{R}$ is then given by the convolution of the Gaussian kernels K_s with the signal f:

$$\begin{aligned} L(t; s) &= (K_s * f)(t) \\ &= \int_{-\infty}^{\infty} K_s(\tau) \cdot f(t - \tau)\, d\tau \\ &= \int_{-\infty}^{\infty} \frac{1}{s\sqrt{2\pi}} e^{-\tau^2/2s^2} \cdot f(t - \tau)\, d\tau \end{aligned} \tag{2}$$

Note that local extrema can equivalently be rephrased in terms of zero-crossings as a local extremum in a continuous function f is equivalent to a zero-crossing in its first difference $\mathcal{D}f$, which is known to commute with the convolution operator:

$$\mathcal{D}(K_s * f) = K_s * (\mathcal{D}f) = (\mathcal{D}K_s) * f \tag{3}$$

Hence, equivalently we can take a convolution of the original signal f with the differentiated Gaussian function to generate a scale-space representation of the first-order derivative

$$\begin{aligned} L(t; s) &= (\mathcal{D}K_s * f)(t) \\ &= \int_{-\infty}^{\infty} \frac{-\sqrt{2}}{2s^3\sqrt{\pi}} \tau\, e^{-\tau^2/2s^2} \cdot f(t - \tau)\, d\tau \end{aligned} \tag{4}$$

in which we can analyse the zero-crossings at each scale. Note that this regularises differentiation, as differentiation of discrete data without some smoothing is an ill-posed problem.

4. Formal Derivation of Error Probabilities

The scale-space theory in the previous section is stated in terms of continuous signals. In practice, many signals and images are, however, discrete as is the case for example with gene expression data. Here, we take our time domain T to be discrete, i.e., $T = \{\ldots, t_{-1}, t_0, t_1, \ldots\}$ such that each $t_l \in T$ can be identified with $l \in \mathbb{Z}$. A commonly used approach for discrete signals is to apply the continuous scale-space theory and discretise the resulting equations giving good approximations (Canny, 1986; Aström and Heyden, 1999). Here, we follow that approach by sampling from the first-order derivative of the Gaussian kernel

$$K_s(t) = \frac{-\sqrt{2}}{2s^3\sqrt{\pi}} t\, e^{-t^2/2s^2} \tag{5}$$

to generate the following scale-space representation for a discrete signal f

$$L(t_l; s) = \sum_{n=-\infty}^{\infty} \frac{-\sqrt{2}}{2s^3\sqrt{\pi}} n\, e^{-n^2/2s^2} \cdot f(t_l - n) \tag{6}$$

which allows us to detect extrema in the signal f by looking for zero-crossings in-between two points t_l and t_{l+1}. Zero-crossings occur when the corresponding values in the scale-space representation changes sign

[1] $\mathbb{R}^+ = \{r \mid r \in \mathbb{R} \wedge r \geq 0\}$

[2] The Gaussian function follows uniquely from an axiomatisation that assumes no knowledge about the domain under study. Additional knowledge about the domain, may result in a different kernel.

$$
\begin{aligned}
&L(t_l;s) < 0 \text{ and } L(t_{l+1};s) > 0 \quad \text{(local minimum)} \\
&L(t_l;s) > 0 \text{ and } L(t_{l+1};s) < 0 \quad \text{(local maximum)}
\end{aligned} \tag{7}
$$

In the real world, signals and images are, however, often distorted by noise. Here, we assume that the underlying signal f is distorted by Gaussian distributed additive noise

$$g(t_l) = f(t_l) + \epsilon, \quad \epsilon \sim N(0, \sigma_g^2) \tag{8}$$

For readability, we attach the function g used in the construction of its scale-space representation $L(t_l; s)$ as a subscript, i.e., $L_g(t_l; s)$. Furthermore, to simplify notation, we keep s fixed for the remainder of this section and rewrite the differentiated Gaussian kernel K_s by a number of indexed variables $k_n = K_s(n)$. Note that because each point $t_l \in T$ is identified with $l \in \mathbb{Z}$ according to Equations (5) and (6) it holds that

$$
\begin{aligned}
L_g(t_l;s) &= \sum_{n=-\infty}^{\infty} k_n g(t_l - n) \\
&= \sum_{n=-\infty}^{\infty} k_n g(t_{l-n})
\end{aligned} \tag{9}
$$

As our criterion for deciding whether a local minimum or local maximum is present is based on the values $L_g(\cdot; s)$ in two subsequent points t_l and t_{l+1} we are interested in the bivariate probability density $p(L_g(t_l; s), L_g(t_{l+1}; s))$. We consider $g(t_l)$ and $g(t_{l+1})$ as observations of a stochastic process. The variance of $L_g(t_l; s)$ is then given by

$$\text{var}(L_g(t_l;s)) = \sum_{n=-\infty}^{\infty} k_n^2 \, \sigma_g^2 \tag{10}$$

because $\text{var}(aX) = a^2\text{var}(X)$ and, with X and Y independent variables, $\text{var}(X+Y) = \text{var}(X)+\text{var}(Y)$ (Shao, 1999).

Next, define the function $\delta_s(t_l, t_{l+1})$ as the difference between the two corresponding L_g values

$$
\begin{aligned}
\delta_s(t_l, t_{l+1}) &= L_g(t_l;s) - L_g(t_{l+1};s) \\
&= \sum_{n=-\infty}^{\infty} k_n g(t_{l-n}) - \sum_{n=-\infty}^{\infty} k_n g(t_{l+1-n}) \\
&= \sum_{n=-\infty}^{\infty} (k_n - k_{n+1}) g(t_{l-n})
\end{aligned} \tag{11}
$$

As δ_s is a linear function of *independently* sampled normally distributed variables for which holds that (Shao, 1999) $\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y)$ it follows from Equation (11) that

$$\text{var}(\delta_s(t_l, t_{l+1})) = \sum_{n=-\infty}^{\infty} (k_n - k_{n+1})^2 \sigma_g^2 \tag{12}$$

By definition, for two variables X and Y it holds that

$$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\,\text{cov}(X, Y) \tag{13}$$

Placing the covariance on the left hand side results in

$$\text{cov}(X, Y) = \frac{1}{2}(\text{var}(X) + \text{var}(Y) - \text{var}(X - Y)) \tag{14}$$

Hence, for variables $X = L_g(t_l; s)$ and $Y = L_g(t_{l+1}; s)$ the covariance $\text{cov}(L_g(t_l; s), L_g(t_{l+1}; s))$ is given by the formula

$$
\begin{aligned}
&\text{cov}(L_g(t_l;s), L_g(t_{l+1};s)) = \\
&\frac{1}{2}\left(\text{var}(L_g(t_l;s) + \text{var}(L_g(t_{l+1};s)) - \text{var}(\delta_s(t_l, t_{l+1}))\right)
\end{aligned} \tag{15}
$$

Combining Equations (10), (12), and (15) gives

$$
\begin{aligned}
&\text{cov}(L_g(t_l;s), L_g(t_{l+1};s)) = \\
&\sigma_g^2 \left(\sum_{n=-\infty}^{\infty} k_n^2 - \frac{1}{2} \sum_{n=-\infty}^{\infty} (k_n - k_{n+1})^2 \right)
\end{aligned} \tag{16}
$$

and the covariance matrix Σ_{L_g} therefore becomes

$$
\Sigma_{L_g} = \begin{bmatrix} \sigma_g^2 \sum_{n=-\infty}^{\infty} k_n^2 & cov(L_g(t_l;s), L_g(t_{l+1};s)) \\ cov(L_g(t_l;s), L_g(t_{l+1};s)) & \sigma_g^2 \sum_{n=-\infty}^{\infty} k_n^2 \end{bmatrix} \tag{17}
$$

From these results follows that the bivariate probability density function $p(L_g(t_l; s); L_g(t_{l+1}; s))$ we are interested in has a bivariate Gaussian distribution $G(\mu_{L_g}, \Sigma_{L_g})$ with mean vector $\mu_{L_g} = (L_g(t_l; s), L_g(t_{l+1}; s))^T$. The bivariate Gaussian distribution can be visualised as concentric ellipses with each ellipse being an iso-density curve.

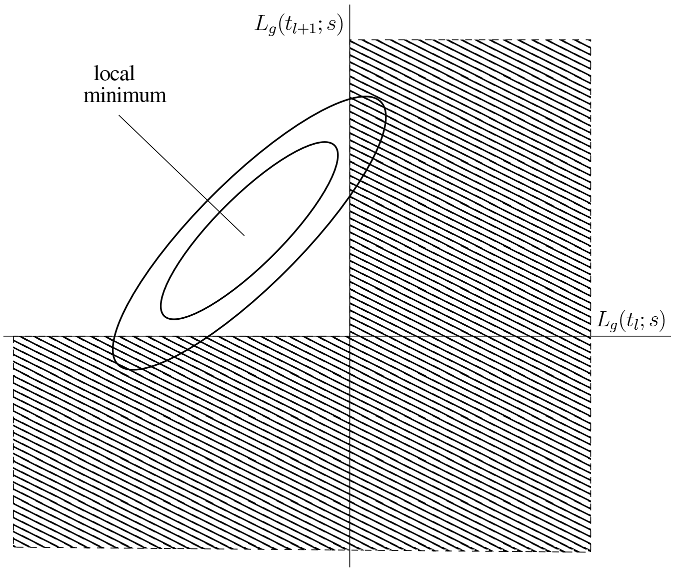


Fig. 3. Bivariate distribution of the two time-points through which the smooth first-order derivative changes sign.

This allows us to give an explicit formula for the error that an extremum will not be re-observed because of noise and/or smoothing of the scale-space kernel, i.e., that the

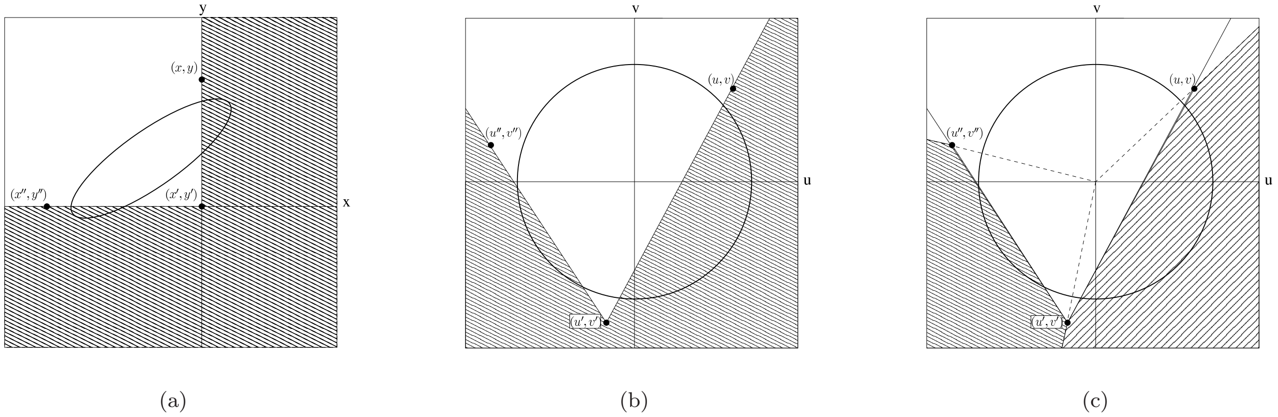


Fig. 4. Stepwise process of integrating a correlated bivariate normal distribution. (a) Correlated bivariate normal distribution with shaded area denoting area of integration. Transforming the distribution by rotating and stretching the axes results in (b) an uncorrelated bivariate normal distribution with zero means. The shaded area to integrate is also affected by this transformation. (c) The final step is to make a subdivision of the area to integrate into a number of polygons such that they can be integrated using the formulas derived by Owen (1956). In this case, the polygons in the uv-plane can be constructed by first taking three points in the xy-plane (x, y), (x', y'), and (x'', y''), which are the origin and two points on the x- and y-axis further from the origin than any data point and second, mapping these to points (u, v), (u', v'), and (u'', v'') in the uv-plane using the same transformation to normalise the bivariate normal distribution.

measurements $L_g(t_l; s)$ and $L_g(t_{l+1}; s)$ do not fulfil the requirements stated in Equation (7). Hence, for a local minimum the probability of not detecting it due to noise and/or smoothing becomes

$$P_{min}(L_g(t_l; s), L_g(t_{l+1}; s); \mu_{L_g}, \Sigma_{L_g}) = 1 - \iint\limits_{\substack{t_l,\, L_g(t_l; s) < 0, \\ t_{l+1},\, L_g(t_{l+1}; s) > 0}} G(\mu_{L_g}, \Sigma_{L_g})\, dt_l\, dt_{l+1} \tag{18}$$

The probability of missing a local maximum is defined analogous with t_l, t_{l+1} such that $L_g(\cdot; s)$ ranges over the lower right quadrant as shown in Figure 3, i.e., $L_g(t_l; s) > 0$ and $L_g(t_{l+1}; s) < 0$.

$$P_{max}(L_g(t_l; s), L_g(t_{l+1}; s); \mu_{L_g}, \Sigma_{L_g}) = 1 - \iint\limits_{\substack{t_l,\, L_g(t_l; s) > 0, \\ t_{l+1},\, L_g(t_{l+1}; s) < 0}} G(\mu_{L_g}, \Sigma_{L_g})\, dt_l\, dt_{l+1} \tag{19}$$

The derived formulas make it possible to calculate the part of the bivariate Gaussian distribution that covers the shaded area, in Figure 3. The probability that an extremum is not detected because of noise, for a given smoothing factor reduces to integrating the tails of a bivariate Gaussian distribution that cross the horizontal and vertical axis, respectively. In the next section we show how to compute this probability by integration.

5. Bivariate Gaussian Integration

The integration of a correlated bivariate normal distribution over arbitrary polygons can be performed in a number of steps as shown in Figure 4. Figure 4(a) shows the outline of a correlated bivariate normal distribution together with a shaded area denoting the area over which one wants to integrate. The first step is to transform the correlated bivariate normal distribution into an uncorrelated bivariate normal distribution with zero means as shown in Figure 4(b). This can be done by a rotation and a stretching of the axes. This transformation will also affect the shaded area over which we need to integrate as shown in Figure 4(b), which will therefore need to be re-computed. The final step of the process is to divide the shaded area into a number of polygons that are suitable for integration using the formulas derived by Owen (1956). These polygons are such that they are bounded by a finite line segment and two lines meeting in the origin as shown in Figure 4(c). Note, that the polygons drawn in Figure 4(c) do not cover the entire area shaded in Figure 4(b). However, the polygons can be chosen in such a way that the area not covered has a large distance from the origin and therefore has a volume that converges to zero. In the following subsections, we discuss this process in more detail.

5.1. *The Fundamental Formulas*

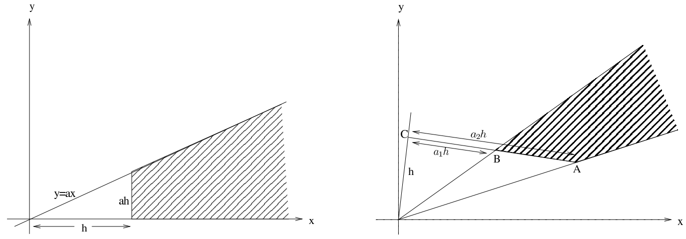
Following the derivation by Owen (1956), the T-function gives, for h and $a > 0$, the volume of an uncorrelated bivariate normal distribution with zero means and unit variances over the area between $y = ax$ and $y = 0$ and to the right of $x = h$ (cf. Fig 5(a)) and is given by

$$T(h,a) = \frac{\arctan a}{2\pi} - \frac{1}{2\pi}\sum_{j=0}^{\infty} c_j a^{2j+1} \tag{20}$$

where

$$c_j = (-1)^j \frac{1}{2j+1}\left[1 - e^{\left(-\frac{1}{2}h^2\right)}\sum_{i=0}^{j}\frac{h^{2i}}{2^i i!}\right] \tag{21}$$

which converges rapidly for small values of a or h. Values for negative a or h can be obtained by using $T(h,-a) = -T(h,a)$ and $T(-h,a) = T(h,a)$.



(a) (b)

Fig. 5. (a) area over which $T(h,a)$ gives the volume of a standardised bivariate normal with correlation zero. (b) a typical area for computing the bivariate normal integral over a polygon.

The area over which $T(h,a)$ gives the volume of a standardised bivariate normal distribution with correlation zero is shown in Figure 5(a). To compute the volume for an arbitrary polygon as shown in Figure 5(b), which is bounded by the side AB is given by $T(h,a_2) - T(h,a_1)$ for $a_2 > a_1$, where h is the length of the perpendicular line from the origin to the intersection C of the line through AB, a_1h is the distance from C to B, and a_2h is the distance from C to A. If C lies between A and B, the T-values are added instead of subtracted.

5.2. *Transformation*

For finding volumes of the general correlated bivariate normal over polygons, the first step is to make a rotation and stretching of the axes to reduce the function under the integral to the form of the T-function, i.e., transforming the general correlated bivariate normal distribution into an uncorrelated bivariate normal distribution. We will denote this transformation as a mapping from the xy-plane into the uv-plane, which can be done using the transformation

$$\begin{aligned} u(x,y;\mu_x,\mu_y,\sigma_x,\sigma_y,\rho) &= \\ &\frac{1}{\sqrt{2+2\rho}}\left[\frac{x-\mu_X}{\sigma_X} + \frac{y-\mu_Y}{\sigma_Y}\right] \\ v(x,y;\mu_x,\mu_y,\sigma_x,\sigma_y,\rho) &= \\ &\frac{-1}{\sqrt{2-2\rho}}\left[\frac{x-\mu_X}{\sigma_X} - \frac{y-\mu_Y}{\sigma_Y}\right] \end{aligned} \tag{22}$$

for the remainder shortened to $u(x,y), v(x,y)$ for readability, which maps a point (x,y) to $(u(x,y), v(x,y))$ for $\rho^2 < 1$, where ρ is the correlation of variables X and Y, μ_x and μ_y are the means of the X and Y variables, and σ_X and σ_Y the standard deviations of the X and Y variables respectively. Hence, in terms of Section 4, for two points t_l, t_{l+1} of interest, we have

$$\begin{aligned} \mu_x &= (K_s * f)(t_l) = L_f(t_l;s), \\ \mu_y &= (K_s * f)(t_{l+1}) = L_f(t_{l+1};s), \\ \sigma_X &= \sigma_Y = \sqrt{\left(\sum_i k_i^2\right)\cdot\sigma_g^2}, \\ \rho &= \frac{\left(\sum_i k_i^2 - \frac{1}{2}\sum_j (k_j - k_{j+1})^2\right)}{\left(\sum_i k_i^2\right)}. \end{aligned} \tag{23}$$

In the next subsection, an empirical validation will be given of the methods shown. As can already be seen in Figure 8, using the transformation described above, empirical data with a correlated bivariate distribution as shown in Figure 8(a) is transformed into an uncorrelated bivariate distribution as shown in Figure 8(b).

Using this transformation from the xy-plane into the uv-plane also transforms the volume of interest, i.e., the polygon over which we need to integrate. Any polygon in the xy-plane will be transformed into a polygon in the uv-plane. The vertices in the uv-plane of the transformed polygon need to be computed to construct the polygon over which we need to integrate in the uv-plane. In general, one can draw a graph of the polygon in the uv-plane and compute its volume using the following formulas (Owen, 1956)

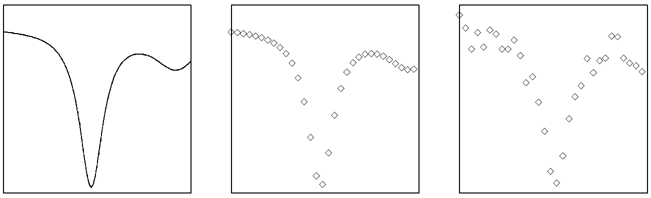
$$\begin{aligned} h &= \frac{|u_1v_2 - u_2v_1|}{\sqrt{(u_2-u_1)^2 + (v_2-v_1)^2}} \\ a_1 &= \frac{|u_1(u_2-u_1) + v_1(v_2-v_1)|}{|u_1v_2 - u_2v_1|} \\ a_2 &= \frac{|u_2(u_2-u_1) + v_2(v_2-v_1)|}{|u_1v_2 - u_2v_1|} \end{aligned} \tag{24}$$

where (u_1,v_1) and (u_2,v_2) are the coordinates of two adjacent vertices on the transformed polygon, i.e., the polygon in the uv-plane, and h, a_1, a_2 are used as described in Section 5.1 and Figure 5(b). With the aid of the graph, these volumes are then combined to give the volume over the outside (or inside) of the polygon.

In particular, in our case we need three vertices (u,v), (u',v'), and (u'',v'') to construct the two polygons as illustrated in Figure 4. To calculate their volume, we will need to compute the T-values using the parameters h, a_1, a_2

as computed above, once using the pair of vertices (u, v), (u', v') and once using the pair of vertices (u', v'), (u'', v'').

5.3. *Empirical Validation*

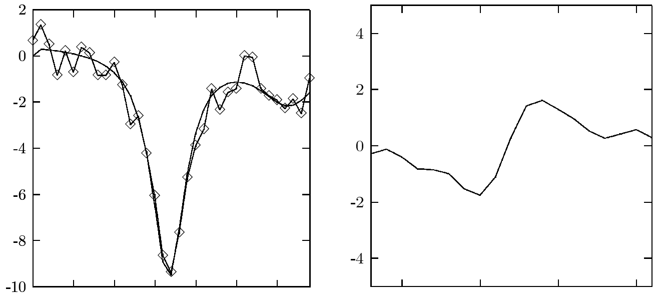


(a) (b) (c)

Fig. 6. Illustrating the acquisition of the model signal. (a) Represents the original, continuous signal f. In (b), the continuous signal is sampled to form a discrete signal f_i. Finally, in (c) distributed, additive noise is added to the discrete signal to obtain signal $\tilde{f}_i$.

The method presented in this article has been empirically validated. A time series was constructed using the process as described in Figure 6. Firstly, a continuous signal f was created (cf. Figure 6(a)). Secondly, the continuous signal f was discretised by sampling data points, denoted by f_i (cf. Figure 6(b)). Thirdly, noise was added to the discrete samples, denoted by $\tilde{f}_i$. (cf. Figure 6(c)).

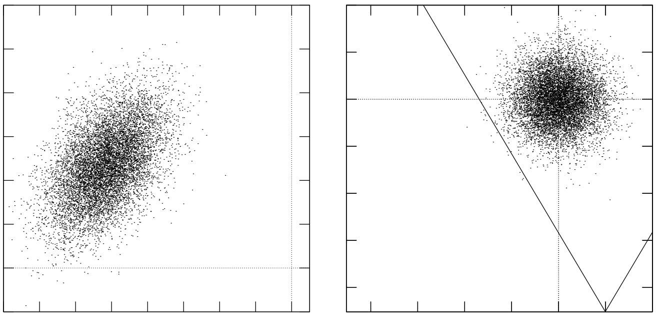
The discretised signal f_i has a length of 34 samples and there is a corresponding local minimum in the continuous signal f between points 15 and 16. By repeating the process in Figure 6, we obtained several time series $g_e = \{\tilde{f}_{1,e}, \ldots, \tilde{f}_{34,e}\}$, $e = 1, \ldots, 10^4$ with added Gaussian noise $\epsilon \sim N(0, \frac{1}{2})$. Each time series g_e can be considered a stochastic experiment in which we are interested in the detection of a local extrema between points $g_{e,15}$ and $g_{e,16}$.



(a) (b)

Fig. 7. (a) Original continuous signal f together with one noisy realisation $g = f_i + \epsilon$, where $\epsilon \sim N(0, \frac{1}{2})$. (b) The convolved signal of a noisy time series with a discretely sampled differentiated Gaussian.

Therefore, for each time series g_e, its convolution $v_e = K_s * g_e$ was calculated with K_s a discretely sampled differentiated Gaussian as in Equation (5), with $\mu = 0$ and scale $s = 1.1$ (cf. Figure 7(b)). The value of σ was arbitrarily chosen and the number of samples used for the kernel was 16 as all infinite sums need to be bounded for implementation purposes.[3] By plotting the value of $v_{e,15}$ against $v_{e,16}$, we obtained the empirical data shown in Figure 8(a). Note that the top left quadrant corresponds with a true minimum. All data points that lie outside of the top left quadrant are experiments in which a local minimum could not be re-observed. The volume of the data points outside of the top left quadrant represents the probability of not re-observing a local extremum (cf. Equation (18)).



(a) (b)

Fig. 8. (a) Empirical data of a general correlated bivariate normal. (b) The same data after transformation, corresponding to an uncorrelated bivariate normal distribution with zero means and unit variances on the right. The lines in (b) correspond with the x- and y-axis in (a) after transformation.

To compute the volume of interest, we first transform the empirical data to obtain data with an uncorrelated bivariate normal distribution using Equation (22), which results in the empirical data of Figure 8(b). As the area for integration also transforms when transforming the empirical data, we need to re-compute the area for integration in the uv-plane (cf. Figure 8(b)).

The final step in our method is shown in detail in Figure 9. As explained in Figure 4, we need to divide the area of integration in the uv-plane in a number of polygons, such that these polygons can be integrated using the method of Owen (1956). These polygons are such that given two vertices, the polygon is bounded by the line segment between those two vertices, and the two lines through the origin O and a vertex. For example, in Figure 9 the area of integration (cf. the shaded area in Figure 4(b)) is subdivided into two polygons. One polygon is bounded by the line between (u, v) and (u', v'), the line through O and (u, v), and the line through O and (u', v'). The other polygon is bounded by the line between (u', v') and (u'', v''), the line through O and (u', v'), and the line through O and (u'', v''). The volume of the area not covered by these polygons rapidly

[3] Infinite summations can be bounded as their tails rapidly converge to zero. If one wants to obtain a certain error bound in the truncation one can also derive a value for the upper bound (Lindeberg, 1990).

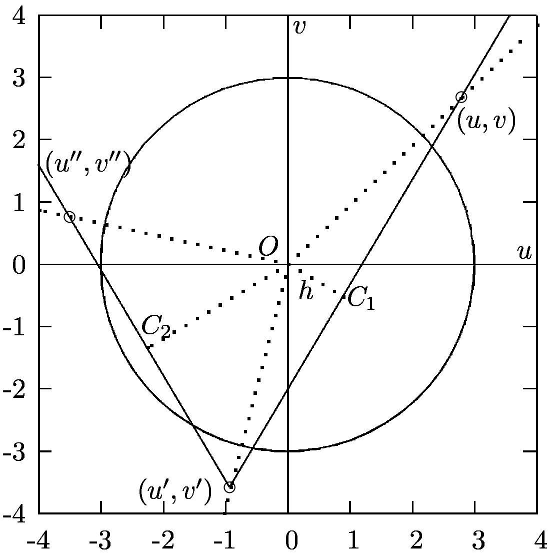


Fig. 9. Detailed construction to compute the volume bounded by the solid lines according to the method of Owen (1956) for an uncorrelated bivariate normal distribution, for which almost all data points are bounded by the given circle.

converges to zero when points (u,v) and (u'',v'') are taken further away from the origin.

To construct the points (u,v), (u',v'), and (u'',v'') in the uv-plane, we take the following three points in the xy-plane $(x,y) = (0,n)$, $(x',y') = (0,0)$, and $(x'',y'') = (-n,0)$ for sufficiently large n (e.g., $n = 1000$) and use the transformation in Equation (22), which we used to normalise the bivariate distribution, to map those points to the corresponding points in the uv-plane (u,v), (u',v'), and (u'',v'').

The volume of both polygons can now be computed using Equations (24) and (20) derived by Owen (1956). Firstly, given a polygon, one computes the values for h, a_1, and a_2 using Equation (24) and the vertices of the polygon (e.g., (u,v) and (u',v')). Secondly, one computes the volume using the values of h, a_1, and a_2 in Equation (24). When the intersection point C_i of the line segment between the two vertices used and the line perpendicular through the origin lies between the two vertices, the two T-values are added. When C_i does not lie between the two vertices the T-values are subtracted (cf. Figure 9).

For the empirical data of Figure 8 we obtained an error probability of 0.0018 using the method presented. This value has been empirically verified by counting the number of data points (out of 10^4 experiments) outside the top left quadrant, which gave an empirical estimate of 0.0016.

6. Related Work

Scale-space theory was first pioneered in the work of Iijima (1962), but was inaccessible for Western researchers as it was published in Japanese. In the western world, the concept was first introduced by Witkin (1983) and independently developed by Koenderink to a complete multi-scale theory (Koenderink, 1984). Since then, several researchers have derived the unique solution of the Gaussian kernel for a scale-space representation for continuous signals. Lindeberg introduced the notion of semi-group structure (Lindeberg, 1990), i.e., the convolution of two scale-space kernels results in a scale-space kernel

$$K(\cdot; s_1) * K(\cdot; s_2) = K(\cdot; s_1 + s_2) \tag{25}$$

which implies that the structure decreases between any two levels when the scale parameter increases

$$s_1 \leq s_2 \text{ implies } \#_{extrema} L(t; s_1) \leq \#_{extrema} L(t; s_2) \tag{26}$$

Lindeberg showed that a sampled Gaussian kernel does not fulfil the semi-group property and developed a complete scale-space theory for discrete signals (Lindeberg, 1990) and showed that for a one-dimensional discrete signal, the discrete analogue of the Gaussian kernel

$$T(n; s) = e^{-t} I_n(t) \tag{27}$$

where I_n are the modified Bessel functions of integer order is the unique solution that satisfies the scale-space axioms and semi-group property. Here, we have taken the common approach (Lindeberg, 1990) of a sampled Gaussian kernel as these kernels are almost similar for $\sigma > 1$ (ter Haar Romeny, 2002). The derived results also hold for the discrete analogue of the Gaussian kernel as this merely means that the indexed variables k_n are replaced by different values.

By interpreting two subsequent points in a signal with Gaussian distributed noise as a stochastic process we derived the bivariate Gaussian distribution $G(\mu_L, \Sigma_L)$ and showed that the error for not detecting local extrema between two subsequent points to be equal to an integration over $G(\mu_L, \Sigma_L)$. This is one of many examples of bivariate Gaussian integration for practical problem solving that abound the literature (e.g., (Smith, 1953)). Tables for integrating the bivariate Gaussian over a rectangle with sides parallel to the axes were already known a century ago (e.g., (Pearson, 1931)). Cadwell extended these results by presenting a method for computing the volume of an uncorrelated bivariate Gaussian over *any polygon* (Cadwell, 1953). In the work of Owen (1956) formulas were derived for computing volumes of the general correlated bivariate normal, which we use in this paper. Basically, the method of Owen (1956) consists of two steps. First, transform the correlated bivariate Gaussian into an uncorrelated Gaussian. Second, compute the polygon of interest after transformation and compute the volume over the transformed polygon for the uncorrelated Gaussian. For the probability of not detecting local extrema due to noise and/or smoothing we are able to formulate this procedure into an algorithm.

Closely related to our approach of investigating deterministic and stochastic aspects of re-observing (features of) a continuous signal given noisy, discretely sampled data is the work of Aström and Heyden (1999). However, Aström and Heyden (1999) investigates the problem of noisy edge-detection, i.e., where the first-order derivative of the signal is maximal, whereas we investigate the problem of noisy extrema detection, i.e., where the first-order derivative of the signal is zero. Furthermore, Aström and Heyden (1999) derives results for the *estimated variance* of the detected edge, in terms of, among others, parameters of camera blurring and intensity jump. In this article, the approach taken is less complex, resulting in *precise probabilities* for the error of not re-observing a local extremum for each value of the scale-space parameter.

7. Practical Considerations

In practice, it is of course not possible to compute the bivariate Gaussian distribution, which is needed to compute the error probabilities, by running a large number of experiments, and therefore needs to be estimated. The covariance matrix Σ_{L_g} is completely specified in terms of σ_g^2 and k_n (cf. Equation (17)). By considering a large number of different gene profiles from the same microarray experiment, one can select those genes that show the least amount of change in expression values and use their variance to get an upper bound for the value of σ_g^2. The underlying assumption is that these genes are dormant and have no role in the samples under study, i.e., these genes should have a constant expression profile without noise present. The values for k_n are easily computed when a scale s has been selected. The vector μ_{L_g} can be estimated by evaluating the scale-space representation of the signal under study in the two points of interest for the detection of an extremum. Both estimates for μ_{L_g} and Σ_{L_g} can be combined to get an estimate for the bivariate Gaussian distribution $G(\mu_{L_g}, \Sigma_{L_g})$. Using this distribution, the rest of the method presented can be used without change to compute the error probabilities.

Summarising, to compute the error probability of not detecting a local minimum, algorithm E, which is shown below, can be used:

Algorithm E (*Error probability local minimum*). This algorithm computes the error probability of not measuring a local minimum of a signal g, given values for the scale parameter s, variance σ_g^2, and points of interest t_l, t_{l+1}.

E1. [Initialise.] Construct for signal g its scale-space representation $L_g(\cdot; s) = K_s * g$ (cf. Equation 6)), with K_s the first-order derivative of the Gaussian kernel (cf. Equation (5)). Furthermore, define functions $u(x, y)$, $v(x, y)$ as in Equation (22) using the values given (cf. Equation 23)), with $\mu_x = (K_s * g)(t_l)$, $\mu_y = (K_s * g)(t_{l+1})$.

E2. [Vertices.] Define the three vertices $A = (0, 1000)$, $B = (0, 0)$, and $C = (-1000, 0)$. Next, compute the vertices A', B', and C' using functions $u(x, y)$ and $v(x, y)$, i.e., $A' = (u(A), v(A)) = (u(0, 1000), v(0, 1000))$.

E3. [Volume 1.] Compute h, a_1, a_2 using Equation (24) with vertices A', B'. Set $volume_1 := T(h, a_2) + T(h, a_1)$.

E4. [Volume 2.] Compute h, a_1, a_2 using Equation (24) with vertices B', C'. Set $volume_2 := T(h, a_2) + T(h, a_1)$.

E5. [Error probability.] Return $volume_1 + volume_2$. □

Although, similar results can also be obtained by sampling from the bivariate Gaussian distribution given the values as provided to Algorithm E, the algorithm gives a computationally cheap method to compute (nearly) precise probability values. The infinite sums that are used within the calculations of Algorithm E can be implemented using small upper bounds to produce good approximations.

8. Conclusions and Further Work

This study began with the premise that an approach for predicting functional relations from gene expression profiles should have a clear underlying methodology for feature extraction and should be accompanied with a characterisation of its performance w.r.t. the inherent noise present in gene expression data sets. Here, we started the analysis from the fundamental theory of scale-space for formulating criteria for detecting local extrema. It was shown that interpreting the measurement of a local extrema in scale-space as a stochastic process behaves like a bivariate Gaussian distribution. The error of not re-observing the extremum due to noise could be rephrased in terms on an integral over the tails of this distribution. Finally, we demonstrated how to use integration techniques developed in the 50s for an exact computation of these error probabilities. This resulted in an exact value that characterises the quality of the measurements that are used in predicting functional relations between genes.

The current study has laid a fundamental basis for predicting functional relations between gene expression profiles based on local extrema that allows one to focus on the most probable predictions for further analysis.

Some issues fell outside the scope of this article and will be dealt with in future research. Firstly, the current approach allows one to compute the error probability for each value of the scale-space parameter, but does not give any indications how to choose the 'best' scale for practical purposes (Lindeberg, 2004). Secondly, no comparison was made between results obtained for different scale-space levels. (The analysis of structures through scale-space is also referred to as 'deep structure' in literature (Lindeberg, 1994).) Often, different possibly conflicting criteria need to be fulfilled in order to obtain the best trade-off between uncertainty (variance) and location accuracy (Janssen *et al.*, 2002).

References

Aström, K. and Heyden, A. (1999). Stochastic analysis of image acquisition, interpolation and scale-space smoothing. *Advances in Applied Probability*, **31**(4), 855–894.

Babaud, J., Witkin, A., Baudin, M., and Duda, R. (1986). Uniqueness of the gaussian kernel for scale-space filtering. *IEEE Transactions on Pattern Analysis of Machine Intelligence*, **PAMI-8**(1), 26–33.

Bähler, J. (2005). Cell cycle control of gene expression in budding and fission yeast. *Annual Review of Genetics*, **39**, 69–94.

Bozdech, Z., Llins, M., Pulliam, B., Wong, E., Zhu, J., and DeRisi, J. (2003). The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS Biol*, **1**(1), 85–100.

Cadwell, J. (1953). The bivariate normal integral. *Biometrika*, **38**, 475–481.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intell.*, **8**(6), 679–698.

Datta, S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**(4), 459–466.

Egmont-Petersen, M., de Jonge, W., and Siebes, A. (2004). Discovery of regulatory connections in microarray data. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 149–160.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. In *Proc. Natl. Acad. Sci. Unit. States Am.*, volume 95, pages 14863–14867.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Goldenthal, M., Vanoni, M., Buchferer, B., and Marmur, J. (2004). Regulation of mal gene expression in yeast: Gene dosage effects. *Molecular and General Genetics*, pages 508–517.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.

Iijima, T. (1962). Basic theory on the normalization of pattern. *Bulletin of Electrical Laboratory*, **26**, 368–388. In Japanese.

Janssen, J., Egmont-Petersen, M., Hendriks, E., Reinders, M., van der Geest, R., Hogendoorn, P., and Reiber, J. (2002). Scale-invariant segmentation of dynamic contrast-enhanced perfusion MR-images with inherent scale selection. *Journal of Visualization and Computer Animation*, **13**(1), 1–19.

Koenderink, J. (1984). The structure of images. *Biological Cybernetics*, **50**, 363–396.

Lindeberg, T. (1990). Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**(3), 234–54.

Lindeberg, T. (1994). *Scale-Space Theory in Computer Visison*. Kluwer, Dordrecht.

Lindeberg, T. (2004). Feature detection with automatic scale selection. *International Journal of Computer Vision*, **30**(2), 79–116.

Owen, D. (1956). Tables for computing bivariate normal probabilities. *Ann Math Stat*, **27**, 1075–1090.

Pearson, K. (1931). Tables for statisticians and biometricians. *Biometrika Office*, **2**.

Shao, J. (1999). *Mathematical Statistics*. Springer-Texts in Statistics. Springer-Verlag, New York-Berlin-Heidelberg.

Smith, R. (1953). Conduction of heat in the semi-finite solid, with a short table of an important integral. *Australian Journal of Physics*, **6**, 127–130.

Storey, J., Xiao, W., Leek, J., Tompkins, R., and Davis, R. (2005). Significance analysis of time course microarray experiments. In *Proceedings of the National Academy of Sciences*, volume 102, pages 12837–12842.

ter Haar Romeny, B. (2002). *Front-End Vision and Multiscale Image Analysis*. Kluwer Academic Publishers.

Witkin, A. (1983). Scale-space filtering. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 1019–1022, Karlsruhe, Germany.