# New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit

Aapo Hyvärinen

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science

**Report A47**

**Otaniemi 1997**

# New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit

**Aapo Hyvärinen**

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science
Rakentajanaukio 2 C, FIN-02150 Espoo, FINLAND

aapo.hyvarinen@hut.fi
http://www.cis.hut.fi/~aapo

Report A47
August 1997

ISBN 951-22-3721-0
ISSN 1455-0784

A shorter version of this paper will appear in:
Advances in Neural Information Processing Systems 10 (Proc. NIPS*97).

# New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit

Aapo Hyvärinen
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 2200, FIN-02015 HUT, Finland
Email: aapo.hyvarinen@hut.fi

**Abstract**

We derive a first-order approximation of the density of maximum entropy for a continuous 1-D random variable, given a number of simple constraints. This results in a density expansion which is somewhat similar to the classical polynomial density expansions by Gram-Charlier and Edgeworth. Using this approximation of density, an approximation of 1-D differential entropy is derived. The approximation of entropy is both more exact and more robust against outliers than the classical approximation based on the polynomial density expansions, without being computationally more expensive. The approximation has applications, for example, in independent component analysis and projection pursuit.

# 1 Introduction

The basic information-theoretic quantity for continuous one-dimensional random variables is differential entropy. The differential entropy $H$ of a scalar random variable $X$ with density $f(x)$ is defined as

$$H(X) = -\int f(x) \log f(x) dx. \tag{1}$$

The 1-D differential entropy, henceforth called simply entropy, has important applications in areas that have attracted considerable interest in the neural network community lately, such as independent component analysis (Comon 1994) and projection pursuit (Huber 1985).

Unfortunately, the estimation of entropy is quite difficult in practice. Using definition (1) requires estimation of the density of $X$, which is recognized to be both theoretically difficult and computationally demanding. Simpler approximations of entropy have been proposed both in the context of projection pursuit (Jones and Sibson 1987) and independent component analysis (Amari, Cichocki, and Yang 1996; Comon 1994). These approximations are usually based on approximating the density $f(x)$ using the polynomial expansions of Gram-Charlier or Edgeworth (Kendall and Stuart 1958). This construction leads to the use of higher-order cumulants, like kurtosis. However, such cumulant-based methods often provide a rather poor approximation of entropy. There are two main reasons for this. Firstly, finite-sample estimators of higher-order cumulants are highly sensitive to outliers: their values may depend on only a few, possibly erroneous, observations with large values (Huber 1985). This means that outliers may completely determine the estimates of cumulants, thus making them useless. Secondly, even if the cumulants were estimated perfectly, they measure mainly the tails of the distribution, and are largely unaffected by structure near the centre of the distribution (Friedman 1987).

Therefore, better approximations of entropy are needed. To this end, we introduce in this paper approximations of entropy that are both more exact in the expectation and have better finite-sample statistical properties, when compared to the cumulant-based approximations. Nevertheless, they retain the computational and conceptual simplicity of the cumulant-based approach. Our approximations are based on an approximative maximum entropy method. This means that we approximate the *maximum* entropy that is compatible with our measurements of the random variable $X$. This maximum entropy, or further approximations thereof, can then be used as a meaningful approximation of the entropy of $X$. To accomplish this, we derive a first-order approximation of the density that has the maximum entropy given a set of constraints, and then use it to derive approximations of the differential entropy of $X$.

This paper is organized as follows. First, some applications of 1-D differential entropy are reviewed in Section 2. The maximum entropy approach is motivated in Section 3. An approximation of the density of maximum entropy is derived in Section 4, and is then used to approximate the entropy in Section 5. The choice of the 'measuring' functions used is treated in Section 6, and some practical examples of the approximations are given in Section 7. Simulation results backing up our theoretical arguments are presented in Section 8, and a discussion is given in Section 9.

# 2 Applications of Differential Entropy

First, we discuss some applications of the approximations introduced in this paper. Two important applications of differential entropy are independent component analysis (ICA) and projection pursuit. In the general formulation of ICA (Comon 1994), the purpose is

to transform an observed random vector $\mathbf{x} = (x_1, ..., x_m)^T$ linearly into a random vector $\mathbf{s} = (s_1, ..., s_m)^T$ whose components are statistically as independent from each other as possible. The mutual dependence of the $s_i$ is classically measured by mutual information. Assuming that the linear transformation is invertible, the mutual information $I(s_1, ..., s_m)$ can be expressed as

$$I(s_1, ..., s_m) = \sum_i H(s_i) - H(x_1, ..., x_m) - \log |\det \mathbf{M}| \tag{2}$$

where $\mathbf{M}$ is the matrix defining the transformation $\mathbf{s} = \mathbf{M}\mathbf{x}$. The second term on the right-hand side does not depend on $\mathbf{M}$, and the minimization of the last term is a simple matter of differential calculus. Therefore, the critical part is the estimation of the 1-D entropies $H(s_i)$: finding an efficient and reliable estimator or approximation of entropy enables an efficient and reliable estimation of the ICA decomposition.

In projection pursuit, the purpose is to search for projections of multivariate data which have 'interesting' distributions (Friedman 1987; Huber 1985; Jones and Sibson 1987). Typically, interestingness is considered equivalent with non-Gaussianity. A natural criterion of non-Gaussianity is entropy (Huber 1985; Jones and Sibson 1987), which attains its maximum (for constant variance) when the distribution is Gaussian, and all other distributions have smaller entropies. Because of the difficulties encountered in the estimation of entropy, many authors have considered other measures of non-Gaussianity (Cook, Buja, and Cabrera 1993; Friedman 1987; Hall 1989), but entropy remains, in our view, the best choice of a projection pursuit index, especially because it provides a simple connection to ICA. Indeed, Eq. (2) can be manipulated to show (Comon 1994) that in ICA as well as in projection pursuit, the basic problem is to find directions in which entropy is minimized for constant variance.

## 3 Why maximum entropy?

Assume that the information available on the density $f(x)$ of the scalar random variable $X$ is of the form

$$\int f(x) G_i(x) dx = c_i, \text{ for } i = 1, ..., n, \tag{3}$$

which means in practice that we have estimated the expectations $E\{G_i(X)\}$ of $n$ different functions of $X$. Since we are not assuming any model for the random variable $X$, the estimation of the entropy of $X$ using this information is not a well-defined problem: there exist an infinite number of distributions for which the constraints in (3) are fulfilled, but whose entropies are very different from each other. In particular, the differential entropy reaches $-\infty$ in the limit where $X$ takes only a finite number of values.

A simple solution to this dilemma is the maximum entropy method. This means that we compute the *maximum* entropy that is compatible with our constraints or measurements in (3), which is a well-defined problem. This maximum entropy, or further approximations thereof, can then be used as an approximation of the entropy of $X$.

Our approach thus is very different from the asymptotic approach often used in projection pursuit (Cook, Buja, and Cabrera 1993; Friedman 1987; Hall 1989). In the asymptotic approach, one establishes a sequence of functions $G_i$ so that when $n$ goes to infinity, the information in (3) gives an asymptotically convergent approximation of some theoretical projection pursuit index. We avoid in this paper any asymptotic considerations, and consider directly the case of finite information, i.e., finite $n$. This non-asymptotic approach is justified by the fact that often in practice, only a small number of measurements of the form (3) are used, for computational or other reasons.

## 4 Approximating the maximum entropy density

In this section, we shall derive an approximation of the density of maximum entropy compatible with the measurements in (3). The basic results of the maximum entropy method tell us (Cover and Thomas 1991) that under some regularity conditions, the density $f_0(x)$ which satisfies the constraints (3) and has maximum entropy among all such densities, is of the form

$$f_0(x) = A \exp(\sum_i a_i G_i(x)), \tag{4}$$

where $A$ and $a_i$ are constants that are determined from the $c_i$, using the constraints in (3) (i.e., by substituting the right-hand side of (4) for $f$ in (3)), and the constraint $\int f_0(x)dx = 1$. This leads in general to a system of $n+1$ non-linear equations which is difficult to solve. Therefore, we decide to make a simple approximation of $f_0$. This is based on the assumption that the density $f(x)$ is *not very far from a Gaussian distribution* of the same mean and variance. In addition, we can make the technical assumption that $f(x)$ is near the standardized Gaussian density $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, since this amounts simply to making $X$ zero-mean and of unit variance. Therefore we put two additional constraints in (3), defined by $G_{n+1}(x) = x, c_{n+1} = 0$ and $G_{n+2}(x) = x^2, c_{n+2} = 1$. To further simplify the calculations, let us make another, purely technical assumption: The functions $G_i, i = 1, ..., n$, form an orthonormal system according to the metric defined by $\varphi$, and are orthogonal to all polynomials of second degree. In other words, for all $i, j = 1, ..., n$

$$\int \varphi(x) G_i(x) G_j(x) dx = \begin{cases} 1, \text{ if } i = j \\ 0, \text{ if } i \neq j \end{cases}, \quad \int \varphi(x) G_i(x) x^k dx = 0, k = 0, 1, 2. \tag{5}$$

For any linearly independent functions $G_i$, this assumption can always be made true by ordinary Gram-Schmidt orthonormalization.

Now, note that the assumption of near-Gaussianity implies that all the other $a_i$ in (4) are very small compared to $a_{n+2} \approx -1/2$, since the exponential in (4) is not far from $\exp(-x^2/2)$. Thus we can make a first-order approximation of the exponential function (detailed derivations can be found in the Appendix). This allows for simple solutions for the constants in (4), and we obtain the *approximative maximum entropy density*, which we denote by $\hat{f}(x)$:

$$\hat{f}(x) = \varphi(x)(1 + \sum_{i=1}^{n} c_i G_i(x)) \tag{6}$$

where $c_i = E\{G_i(X)\}$. To estimate this density in practice, the $c_i$ are estimated, for example, as the corresponding sample averages of the $G_i(X)$. The density expansion in (6) is somewhat similar to the Gram-Charlier and Edgeworth expansions (Kendall and Stuart 1958).

## 5 Approximating the differential entropy

An important application of the approximation of density shown in (6) is in approximation of entropy. A simple approximation of entropy can be found by approximating both occurences of $f$ in (1) by $\hat{f}$ as defined in Eq. (6), and using a Taylor approximation of the logarithmic function, which yields $(1+\epsilon)\log(1+\epsilon) \approx \epsilon + \epsilon^2/2$. Thus one obtains after some algebraic manipulations (see Appendix)

$$H(X) \approx -\int \hat{f}(x) \log \hat{f}(x) dx \approx H(\nu) - \frac{1}{2} \sum_{i=1}^{n} c_i^2 \tag{7}$$

where $H(\nu) = \frac{1}{2}(1 + \log(2\pi))$ means the entropy of a standardized Gaussian variable, and $c_i = E\{G_i(X)\}$ as above. Note that even in cases where this approximation is not very accurate, (7) can be used to construct a projection pursuit index (or a measure of non-Gaussianity) that is consistent in the sense that (7) obtains its maximum value, $H(\nu)$, when $X$ has a Gaussian distribution.

## 6 Choosing the measuring functions

Now it remains to choose the 'measuring' functions $G_i$ that define the information given in (3). As noted in Section 4, one can take practically any set of linearly independent functions, say $\bar{G}_i, i = 1, ..., n$, and then apply Gram-Schmidt orthonormalization on the set containing those functions and the monomials $x^k, k = 0, 1, 2$, so as to obtain the set $G_i$ that fulfills the orthogonality assumptions in (5). This can be done, in general, by numerical integration. In the practical choice of the functions $\bar{G}_i$, the following criteria must be emphasized:

1. The practical estimation of $E\{\bar{G}_i(x)\}$ should not be statistically difficult. In particular, this estimation should not be too sensitive to outliers.

2. The maximum entropy method assumes that the function $f_0$ in (4) is integrable. Therefore, to ensure that the maximum entropy distribution exists in the first place, the $\bar{G}_i(x)$ must not grow faster than quadratically as a function of $|x|$, because a function growing faster might lead to non-integrability of $f_0$ (Cover and Thomas 1991).

3. The $\bar{G}_i$ must capture aspects of the distribution of $X$ that are pertinent in the computation of entropy. In particular, if the density $f(x)$ were known, the optimal function $\bar{G}_{opt}$ would clearly be $-\log f(x)$, because $-E\{\log f(X)\}$ gives directly the entropy. Thus, one might use the log-densities of some known important densities as $\bar{G}_i$.

The first two criteria are met if the $\bar{G}_i(x)$ are functions that do not grow too fast (not faster than quadratically) when $|x|$ grows. This excludes, for example, the use of higher-order polynomials, as are used in the Gram-Charlier and Edgeworth expansions. One might then search, according to point 3, for log-densities of some well-known distributions that also fulfill the first two conditions. Examples will be given in the next section.

It should be noted, however, that the criteria above only delimit the space of function that can be used. Our framework enables the use of very different functions (or just one) as $\bar{G}_i$. The choice is not restricted to some well-known basis of a functional space, as in most approaches (Amari, Cichocki, and Yang 1996; Comon 1994; Jones and Sibson 1987). However, if prior knowledge is available on the distributions whose entropy is to estimated, point 3 above shows how to choose the optimal function.

## 7 A simple special case

A simple special case of (6) is obtained if one uses two functions $\bar{G}_1$ and $\bar{G}_2$, which are chosen so that $\bar{G}_1$ is *odd* and $\bar{G}_2$ is *even*. Such a system of two functions can measure the two most important features of non-Gaussian 1-D distributions. The odd function measures the asymmetry, and the even function measures the bimodality/sparsity dimension (called central hole/central mass concentration in (Cook, Buja, and Cabrera 1993)). After extensive experiments, Cook et al (1993) also came to the conclusion that two such measures (or two terms in their projection pursuit index) are enough for projection pursuit in most cases. Classically, these features have been measured by skewness and kurtosis, which correspond

to $\bar{G}_1(x) = x^3$ and $\bar{G}_2(x) = x^4$, but we do not use these functions for the reasons explained in Section 6.

In this special case, the approximation in (7) simplifies to

$$H(X) \approx H(\nu) - [k_1(E\{\bar{G}_1(X)\})^2 + k_2(E\{\bar{G}_2(X)\} - E\{\bar{G}_2(\nu)\})^2] \tag{8}$$

where $k_1$ and $k_2$ are positive constants (see Appendix), and $\nu$ is a Gaussian random variable of zero mean and unit variance. Practical examples of choices of $\bar{G}_i$ that are consistent with the requirements in Section 6 are the following. First, for measuring bimodality/sparsity, one might use, according to the recommendations of Section 6, the log-density of the double exponential (or Laplace) distribution:

$$\bar{G}_{2a}(x) = |x|. \tag{9}$$

For computational reasons, a smoother version of $\bar{G}_{2a}$ might also be used. Another choice would be the Gaussian function, which may be considered as the log-density of a distribution with infinitely heavy tails:

$$\bar{G}_{2b}(x) = \exp(-x^2/2). \tag{10}$$

For measuring asymmetry, one might use, on more heuristic grounds, the following function:

$$\bar{G}_1(x) = x \exp(-x^2/2). \tag{11}$$

which corresponds to the second term in the projection pursuit index of Cook et al (1993).

Using the above examples one obtains two practical examples of (8):

$$H_a(X) = H(\nu) - [k_1(E\{X \exp(-X^2/2)\})^2 + k_2^a(E\{|X|\} - \sqrt{2/\pi})^2], \tag{12}$$

$$H_b(X) = H(\nu) - [k_1(E\{X \exp(-X^2/2)\})^2 + k_2^b(E\{\exp(-X^2/2)\} - \sqrt{1/2})^2], \tag{13}$$

with $k_1 = 36/(8\sqrt{3} - 9)$, $k_2^a = 1/(2 - 6/\pi)$, and $k_2^b = 24/(16\sqrt{3} - 27)$. As above, $H(\nu) = \frac{1}{2}(1 + \log(2\pi))$ means the entropy of a standardized Gaussian variable. These approximations $H_a(X)$ and $H_b(X)$ can be considered more robust and accurate generalizations of the approximation derived using the Gram-Charlier expansion in (Jones and Sibson 1987). Indeed, using the polynomials $\bar{G}_1(x) = x^3$ and $\bar{G}_2(x) = x^4$ one obtains the approximation of entropy in (Jones and Sibson 1987), which is in practice almost identical to the one proposed in (Comon 1994). Finally, note that the approximation in (13) is very similar to the first two terms of the projection pursuit index by (Cook, Buja, and Cabrera 1993).

## 8 Simulation Results

To show the validity of our approximations of differential entropy we compared the approximations $H_a$ and $H_b$ in Eqs (12) and (13) in Section 7, with the one offered by higher-order cumulants as given in (Jones and Sibson 1987). The expectations were here evaluated exactly, ignoring finite-sample effects.

First, we used a family of Gaussian mixture densities, defined by

$$f(x) = \mu\varphi(x) + (1 - \mu)2\varphi(2(x - 1)) \tag{14}$$

where $\mu$ is a parameter that takes all the values in the interval $0 \leq \mu \leq 1$. This family includes asymmetric densities of both negative and positive kurtosis. The results are depicted in Fig. 1. Note that the plots show approximations of negentropies: the negentropy

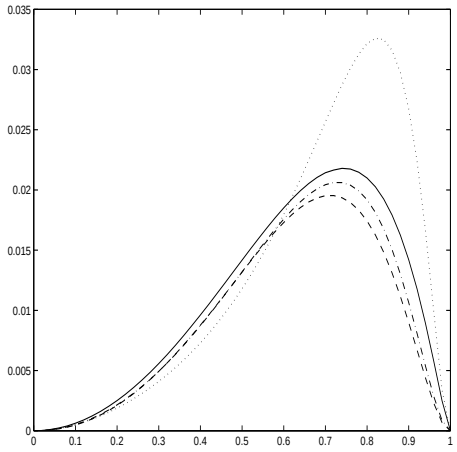


Figure 1: Comparison of different approximations of negentropy, for the family of mixture densities in (14) parametrized by $\mu$ ranging from 0 to 1. Solid line: true negentropy. Dotted line: cumulant-based approximation. Dashed line: approximation $H_a$ in (12). Dot-dashed line: approximation $H_b$ in (13). Our two approximations were much better than the cumulant-based one.

of $X$ equals $H(\nu) - H(X)$, where $\nu$ is again a standardized Gaussian variable. One can see that both of the approximations $H_a$ and $H_b$ introduced in Section 7 were considerably more accurate than the cumulant-based approximation.

Second, we considered the following family of density functions:

$$f_\alpha(x) = C_1 \exp(C_2 |x|^\alpha) \tag{15}$$

where $\alpha$ is a positive constant, and $C_1, C_2$ are normalization constants that make $f_\alpha$ a probability density of unit variance. For different values of $\alpha$, the densities in this family exhibit different shapes. For $.5 \leq \alpha < 2$, one obtains (sparse) densities of positive kurtosis. For $\alpha = 2$, one obtains the Gaussian density, and for $\alpha > 2$, a density of negative kurtosis. Thus the densities in this family can be used as examples of different symmetric non-Gaussian densities. In Figure 2, the different approximations are plotted for this family, using parameter values $.5 \leq \alpha \leq 3$. Since the densities used are all symmetric, the first terms in the approximations were neglected. Again, it is clear that both of the approximations $H_a$ and $H_b$ introduced in Section 7 were considerably more accurate than the cumulant-based approximation in (Comon 1994; Jones and Sibson 1987). (In the case of symmetric densities, these two cumulant-based approximations are identical). Especially in the case of sparse densities (or densities of positive kurtosis), the cumulant-based approximations performed very poorly; this is probably because it gives too much weight to the tails of the distribution.

# 9 Discussion

Novel methods of approximating the density and the differential entropy of a 1-D random variable were introduced. The approximations were based on the maximum entropy principle, i.e., we approximated the (density of) maximum entropy compatible with given measurements. The approximations used a linear approximation of the maximum entropy equations to obtain a computationally simple approximation of the density of maximum entropy for given constraints. Using the Taylor expansion of the logarithmic function, a simple approximation of differential entropy was then obtained. The resulting approximation of differential entropy was shown to be considerably more accurate and statistically well-behaving than existing methods of low computational complexity. The methods can be viewed as generalizations of the polynomial density approximations, based on truncations of the Gram-Charlier and Edgeworth expansions. These novel approximations can be used, for

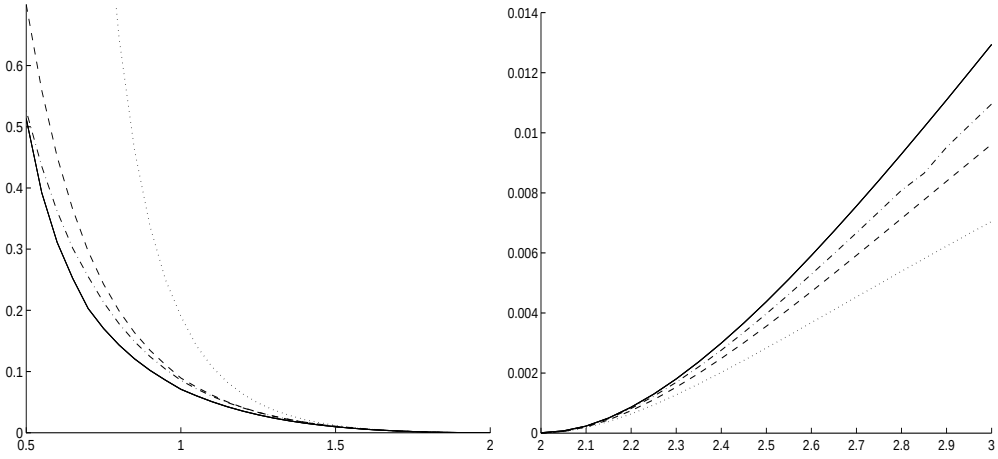


Figure 2: Comparison of different approximations of negentropy, for the family of densities (15) parametrized by $\alpha$. On the left, approximations for densities of positive kurtosis ($.5 \leq \alpha < 2$) are depicted, and on the right, approximations for densities of negative kurtosis ($2 < \alpha \leq 3$). Solid line: true negentropy. Dotted line: cumulant-based approximation. Dashed line: approximation $H_a$ in (12). Dot-dashed line: approximation $H_b$ in (13). Clearly, our two approximations were much better than the cumulant-based one, especially in the case of densities of positive kurtosis.

example, in the estimation of mutual information, as in indepedendent component analysis, and as measures of non-normality in projection pursuit.

In particular, this framework took explicitly into account the fact that we have only a finite number of information of $f(.)$. Thus our approach was very different from the one usually taken in projection pursuit (Friedman 1987; Cook, Buja, and Cabrera 1993; Hall 1989), where one tries to obtain a sequence of indexes that converges to a theoretical measure when the number $n$ of measurements of type (3) goes to infinity. Such asymptotic considerations may have limited validity in practical situations where one uses only a very small number of information in addition to the mean and the covariance. Indeed, in many applications, it is necessary for computational reasons to use only a minimum number of measurements of the form (3). Their number $n$ may be only one or two.

The principle introduced in this paper is a very general one, and thus possible extensions are numerous. For example, the first-order approximations made in Sections 4 and 5 might be extended to include higher-order terms, though this may lead to computationally more complex approximations. Furthermore, extensions to random variables of more than one dimension seem straightforward.

## References

Amari, S., A. Cichocki, and H. Yang (1996). A new learning algorithm for blind source separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Eds.), *Advances in Neural Information Processing 8 (Proc. NIPS'95)*, pp. 757–763. Cambridge, MA: MIT Press.

Comon, P. (1994). Independent component analysis – a new concept? *Signal Processing 36*, 287–314.

Cook, D., A. Buja, and J. Cabrera (1993). Projection pursuit indexes based on orthonormal function expansions. *J. of Computational and Graphical Statistics 2*(3), 225–250.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. John Wiley & Sons.

Friedman, J. (1987). Exploratory projection pursuit. *J. of the American Statistical Association 82*(397), 249–266.

Hall, P. (1989). Polynomial projection pursuit. *The Annals of Statistics 17*, 589–605.

Huber, P. (1985). Projection pursuit. *The Annals of Statistics 13*(2), 435–475.

Jones, M. and R. Sibson (1987). What is projection pursuit ? *J. of the Royal Statistical Society, ser. A 150*, 1–36.

Kendall, M. and A. Stuart (1958). *The Advanced Theory of Statistics*. Charles Griffin & Company.

# A Derivation of (6)

Due to the assumption of near-Gaussianity, we can write $f_0(x)$ as

$$f_0(x) = A \exp(-x^2/2 + a_{n+1}x + (a_{n+2} + 1/2)x^2 + \sum_{i=1}^{n} a_i G_i(x)), \tag{16}$$

where in the exponential, all other terms are very small with respect to the first one. Thus, using the first-order approximation $\exp(\epsilon) \approx 1 + \epsilon$, we obtain

$$f_0(x) \approx \tilde{A}\varphi(x)(1 + a_{n+1}x + (a_{n+2} + 1/2)x^2 + \sum_{i=1}^{n} a_i G_i(x)), \tag{17}$$

where $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ is the standardized Gaussian density, and $\tilde{A} = \sqrt{2\pi}A$. Due to the orthogonality constraints in (5), the equations for solving $\tilde{A}$ and $a_i$ become linear and almost diagonal:

$$\int f_0(x)dx = \tilde{A}(1 + (a_{n+2} + 1/2)) = 1 \tag{18}$$

$$\int f_0(x)xdx = \tilde{A}a_{n+1} = 0 \tag{19}$$

$$\int f_0(x)x^2dx = \tilde{A}(1 + 3(a_{n+2} + 1/2)) = 1 \tag{20}$$

$$\int f_0(x)G_i(x)dx = \tilde{A}a_i = c_i, \text{ for } i = 1, ..., n \tag{21}$$

and can be easily solved to yield $\tilde{A} = 1, a_{n+1} = 0$, $a_{n+2} = -1/2$ and $a_i = c_i, i = 1, .., n$. This gives (6).

# B Derivation of (7)

Using the Taylor expansion $(1 + \epsilon)\log(1 + \epsilon) = \epsilon + \epsilon^2/2 + o(\epsilon^2)$, one obtains

$$-\int \hat{f}(x) \log \hat{f}(x)dx = \tag{22}$$

$$-\int \varphi(x)(1 + \sum c_i G_i(x))(\log(1 + \sum c_i G_i(x)) + \log \varphi(x))dx = \tag{23}$$

$$-\int \varphi(x)\log\varphi(x) - \int \varphi(x)\sum c_i G_i(x)\log\varphi(x) \tag{24}$$

$$-\int \varphi(x)[\sum c_i G_i(x) + \frac{1}{2}(\sum c_i G_i(x))^2 + o((\sum c_i G_i(x))^2)] \tag{25}$$

$$= H(\nu) - 0 - 0 - \frac{1}{2}\sum c_i^2 + o((\sum c_i)^2) \tag{26}$$

due to the orthogonality relationships in (5).

# C Derivation of (8), (12) and (13)

First, we must orthonormalize the two functions $\bar{G}_1$ and $\bar{G}_2$ according to (5). To do this, it is enough to determine constants $\beta_1, \delta_1, \alpha_2, \gamma_2, \delta_2$ so that the functions $G_1(x) = (\bar{G}_1(x) + \beta_1 x)/\delta_1$ and $G_2(x) = (\bar{G}_2(x) + \alpha_2 x^2 + \gamma_2)/\delta_2$ are orthogonal to any second degree polynomials as in (5), and have unit norm in the metric defined by $\varphi$. In fact, as will be seen below, this modification gives a $G_1$ that is odd and a $G_2$ that is even, and therefore the $G_i$ are automatically orthogonal with respect to each other. Thus, first we solve the following equations:

$$\int \varphi(x)x(\bar{G}_1(x) + \beta_1 x)dx = 0 \tag{27}$$

$$\int \varphi(x)x^k(\bar{G}_2(x) + \alpha_2 x^2 + \gamma_2)dx = 0, \text{for } k = 0, 2 \tag{28}$$

A straight-forward solution gives:

$$\beta_1 = -\int \varphi(x)\bar{G}_1(x)x dx, \tag{29}$$

$$\alpha_2 = \frac{1}{2}(\int \varphi(x)\bar{G}_2(x)dx - \int \varphi(x)\bar{G}_2(x)x^2 dx), \tag{30}$$

$$\gamma_2 = \frac{1}{2}(\int \varphi(x)\bar{G}_2(x)x^2 dx - 3\int \varphi(x)\bar{G}_2(x)dx). \tag{31}$$

Next note that $\int \varphi(x)(\bar{G}_2(x) + \alpha_2 x^2 + \gamma_2)dx = 0$ implies together with the standardization

$$c_i = E\{G_i(X)\} = [E\{\bar{G}_i(X)\} - E\{\bar{G}_i(\nu)\}]/\delta_i. \tag{32}$$

This implies (8), with $k_i^2 = 1/(2\delta_i^2)$. Thus we only need to determine explicitly the $\delta_i$ for each function. We solve the two equations

$$\int \varphi(x)(\bar{G}_1(x) + \beta_1 x)^2/\delta_1 dx = 1, \tag{33}$$

$$\int \varphi(x)(\bar{G}_2(x) + \alpha_2 x^2 + \gamma_2)^2/\delta_2 dx = 1, \tag{34}$$

which yield after some tedious manipulations:

$$\delta_1^2 = \int \varphi(x)\bar{G}_1(x)^2 dx - (\int \varphi(x)\bar{G}_1(x)x\, dx)^2 \tag{35}$$

$$\delta_2^2 = \int \varphi(x)\bar{G}_2(x)^2 dx - (\int \varphi(x)\bar{G}_2(x)dx)^2$$
$$-\frac{1}{2}(\int \varphi(x)\bar{G}_2(x)dx - \int \varphi(x)\bar{G}_2(x)x^2 dx)^2. \tag{36}$$

Evaluating the $\delta_i$ for the given functions $\bar{G}_i$, one obtains (12) and (13) by the relation $k_i^2 = 1/(2\delta_i^2)$.