Hermite Polynomials and Measures of Non-gaussianity

Jouni Puuronen and Aapo Hyvärinen

Dept of Mathematics and Statistics, Dept of Computer Science and HIIT,
University of Helsinki, Finland
jouni.puuronen@helsinki.fi, aapo.hyvarinen@helsinki.fi

Abstract. We first review some rigorous properties of the Hermite polynomials, and demonstrate their usefulness in estimating probability distributions as series from data samples. We then proceed to explain how these series can be used to obtain precise and robust measures of non-Gaussianity. Our measures of non-Gaussianity detect all kinds of deviations from Gaussianity, and thus provide reliable objective functions for ICA. With a linear computational complexity with respect to the sample size, our method is also suitable for large data sets.

Keywords: ICA, Hermite polynomials, non-Gaussianity.

1 Introduction

Measuring the non-Gaussianity of a random variable is a central problem in the theory of independent component analysis (ICA) and related domains. Most approaches are based either on cumulants such as kurtosis and skewness, or differential entropy. Cumulants are computationally simple but they are very sensitive to outliers which seriously limits their practical utility. Differential entropy is statistically optimal in the context of ICA estimation, but its estimation and computation is very difficult. Various compromises between the two have been proposed, for example in [6].

Cumulant-based non-Gaussianity measures can be motivated by the theory of Hermite polynomials. When a suitable polynomial series is used to approximate the probability density function (pdf) near a Gaussian density, cumulants are obtained naturally as the coefficients in the series [1,2]. The advantage of such an approach could be that we can include an arbitrary number of terms in the expansion and thus it may be possible to obtain very general approximations.

In this paper, we first review some existing theory of Hermite polynomials which gives a more general way of approximating pdf's than what has been used so far in the signal processing or machine learning literature. In particular, we show how the Hermite polynomials can be used to obtain an expansion whose coefficients can be naturally estimated as expectations of dampened polynomials, which are robust against outliers. The dampening is done by multiplying the polynomials by a Gaussian kernel.

T. Honkela et al. (Eds.): ICANN 2011, Part II, LNCS 6792, pp. 205–212, 2011.

Based on the Hermite polynomial expansion, we propose a family of non-Gaussianity measures which is a) derived in a principled way from a polynomial expansion, b) zero only for the Gaussian distribution, c) robust against outliers, and d) easy to compute since it is essentially obtained by expectations of analytical functions of the data.

2 The Hermite Polynomial Series

2.1 Definition

We use a following definition for the Hermite polynomials:

$$H_n(x) = (-1)^n e^{\frac{1}{2}x^2} D_x^n e^{-\frac{1}{2}x^2}, \tag{1}$$

where D_x is the derivative operator. The orthogonality and formal completeness properties of these polynomials are given by

$$\int\limits_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} H_n(x) H_m(x) dx = \sqrt{2\pi} n!\, \delta_{nm}, \tag{2}$$

$$\sum_{n=0}^{\infty} \frac{1}{n!} H_n(x) H_n(x') = \sqrt{2\pi}\delta(x - x') e^{\frac{1}{2}x^2}. \tag{3}$$

This means that once a function $f(x)$ on a real axis is given, and once coefficients a_n are defined as follows

$$a_n = \frac{1}{\sqrt{2\pi} n!} \int\limits_{-\infty}^{\infty} e^{(\alpha - \frac{1}{2})x^2} H_n(x) f(x) dx, \tag{4}$$

for all $n = 0, 1, 2, \ldots$, we may attempt to recover the function $f(x)$ as a series

$$f(x) = e^{-\alpha x^2} \sum_{n=0}^{\infty} a_n H_n(x). \tag{5}$$

If one substitutes the formula given in (4) into the coefficients a_n in (5), and uses the delta function identity given in (3), the factors depending on α cancel out. Thus, at least formally, these series should work with arbitrary α.

If a choice $\alpha = \frac{1}{2}$ is used, the series may be called a Gram-Charlier series. With a choice $\alpha = \frac{1}{4}$ the series should be called a Gauss-Hermite series. There may be some ambiguity about the terminology in some contexts, but some authors are strict with these conventions [3]. Thus, we have generalized these two well-known series expansions by introducing a general parameter α. It is also possible to interpret this series representation as a Gram-Charlier series of the function $e^{(\alpha - \frac{1}{2})x^2} f(x)$ with any α.

2.2 Convergence of the Series

Many authors mention some conditions that may be used to guarantee the convergence, but the ones given by Szegö are among the most rigorous. [1] Applying Szegö's results, we obtain the following convergence conditions. If a function $f(x)$ is integrable over all finite intervals, and satisfies the condition

$$\int_{n}^{\infty} e^{(\alpha-\frac{1}{4})x^2} x^{-5/3}\big(|f(x)|+|f(-x)|\big)dx = o\Big(\frac{1}{n}\Big), \quad \text{as } n\to\infty \tag{6}$$

then the series in (5) converges towards the same number as the limit

$$\lim_{n\to\infty}\frac{1}{\pi}\int_{x-\delta}^{x+\delta} f(y)\frac{\sin(\sqrt{2n}(x-y))}{x-y}dy \tag{7}$$

with some fixed $\delta>0$ (or the series may diverge as this limit, whichever the limit does). If f further has the bounded variation property in some environment of x, then

$$\frac{\sin(\sqrt{2n}(x-y))}{x-y} \underset{n\to\infty}{\to} \pi\delta(x-y). \tag{8}$$

A formal calculation that verifies this delta identity can be computed with a change of variable. A rigorous proof can be carried out using techniques explained in literature of Fourier analysis [5]. In our context, pathological local oscillations are usually assumed to not be present. Thus, if the condition (6) holds, if f has the bounded variation property in some environment of x, and if f is continuous at x, then we know that the series (5) will converge towards the $f(x)$ at the point x.

We now have a prescription for a series representation of a given probability density function p_X. First fix a constant $0<\alpha\leq\frac{1}{2}$. In order for the series (5) to work optimally for $f(x)=p_X(x)$, we should preprocess a data sample $(x_1,x_2,\ldots,x_T)$ so that $E(X)=0$ and $E(X^2)=\frac{1}{2\alpha}$. Then $p_X(x)$ can be approximated with the series (5), once the integral of (4) is replaced with a sample average, giving us estimates $\hat{a}_n$. This scaling is not absolutely necessary, but the density function $\sqrt{\alpha/\pi}\exp(-\alpha x^2)$, which is proportional to the first term in the series (5), has the same variance $E(X^2)=\frac{1}{2\alpha}$. Other scalings would make the convergence slower.

However, standardized random variables are preferred for some applications, so we should note that if we define a variable $Y=\sqrt{2\alpha}X$, then $E(Y^2)=1$ and its pdf will be

$$p_Y(y)\approx\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\sum_{n=0}^{N}\hat{b}_nH_n\Big(\frac{y}{\sqrt{2\alpha}}\Big), \tag{9}$$

[1] Theorem 9.1.6 [4]. Szegö uses different convention where the Hermite polynomials are $2^{n/2}H_n(\sqrt{2}x)$ (here written using the definition (1)).

where $\hat{b}_n = \sqrt{\frac{\pi}{\alpha}}\hat{a}_n$. The coefficients b_n may equivalently be estimated by the formula

$$\hat{b}_n = \frac{1}{\sqrt{2\alpha}n!T}\sum_{t=1}^{T} e^{\frac{1}{2}(1-\frac{1}{2\alpha})y_t^2} H_n\Big(\frac{y_t}{\sqrt{2\alpha}}\Big). \tag{10}$$

Now, we propose to choose the parameter α so that the estimation of the coefficients b_n is robust against outliers. The choice $\alpha = \frac{1}{2}$ would give no robustness, because the exponential dampening term in (10) would vanish, leaving us with nothing more than expectation values of polynomials. The choices $0 < \alpha < \frac{1}{2}$ instead do give more robust formulas for coefficients $\hat{b}_n$, since the exponential dampening term will take weight away from the outliers.

For sake of example, consider the samples, generated out of a peculiarly shaped distribution, shown in the figure 1. This pdf consists of three pyramids close to the origin, and exponential tails.

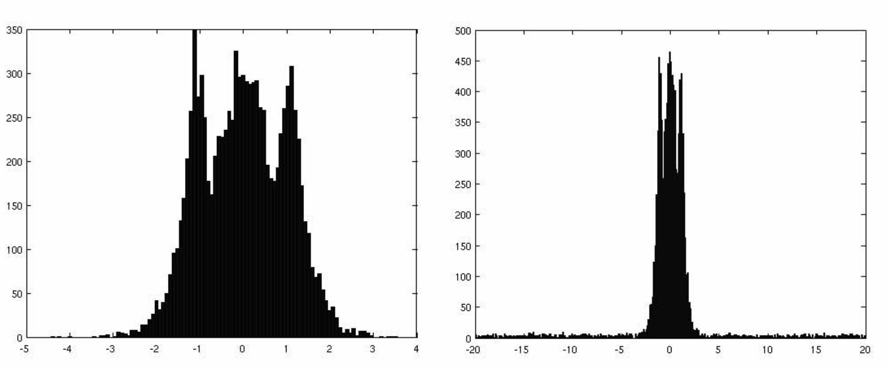

Fig. 1. On the left is a clean sample of 10000 points. On the right is otherwise the same sample but now with 100 outliers.

The figure 2 shows how well the approximation (9) works. We learn two things. Firstly, it would be a mistake to assume that the series approximation could be used only for distributions in which a reasonably small perturbation has been added to the Gaussian distribution. By computing sufficiently large number of terms in the series, the series can be made to converge to all kinds of distributions. Secondly, it would be a mistake to assume that outliers would make the computation of higher order terms impossible. With $\alpha = \frac{1}{4}$, the exponential dampening term provides a very good protection against outliers.

There are many interesting phenomena related to the convergence of these series, but we don't have space for a complete discussion on this topic. We merely note briefly some problems related to the use of too large $\alpha \approx \frac{1}{2}$ and too small $\alpha \approx 0$. The use of too large parameter can result in a complete failure in convergence, possibly due to the failure in satifying the condition (6), and possibly due to the lack of robustness. For example, if $\alpha > \frac{1}{4}$, the condition (6) is not satisfied

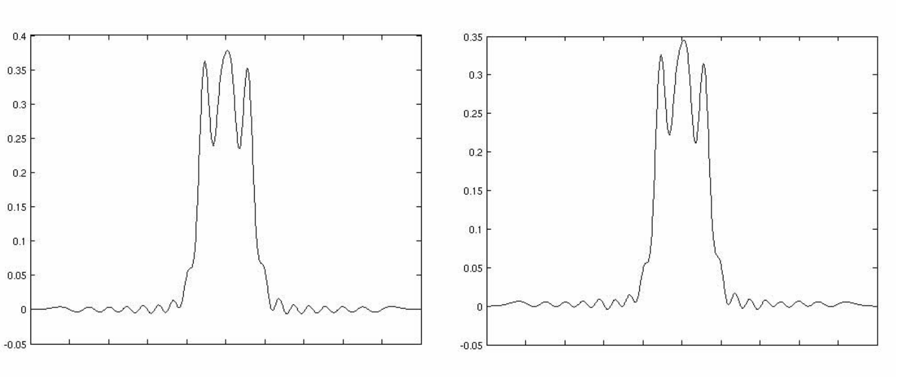

Fig. 2. Samples shown in figure 1 are estimated with approximation (9), with parameters $\alpha = \frac{1}{4}$ and $N = 40$

for the well-known Laplacian pdf, and in this case the series indeed turns out to be divergent. On the other hand, while using a very small parameter does guarantee robustness and pointwise convergence, the convergence can become highly non-uniform, which implies other difficulties for approximation purposes.

3 Measures of Non-gaussianity

3.1 Definition

Once a series representation (9) is obtained for a standardized distribution $p_Y(y)$, the coefficients b_n can be used to obtain a natural measure of non-Gaussianity,

$$J_\alpha = (1 - b_0)^2 + \sum_{n=1}^{N} n!\, b_n^2. \tag{11}$$

It is clear that J_α is zero only if Y is Gaussian. Other measures with the same property could be defined from these coefficients too, but the expression (11) is particularly well justified, as can be seen by examining its relationship to the L_2-norm and negentropy.

In fact, if $\alpha = \frac{1}{4}$ and $N = \infty$, then J_α is proportional to the L_2-distance between p_Y and the Gaussian distribution. If $\alpha = \frac{1}{2}$, then by using the approximation $\log(1 + t) \approx t$, it can be shown that J_α is approximately the negentropy of Y as in [6]. Both of these claims can be verified by using the orthogonality properties of Hermite polynomials. Thus the quantity J_α can be interpreted as a generalization of these two previously known measures of non-Gaussianity.

The special case $\alpha = \frac{1}{2}$ is nothing new for ICA, since it merely gives the skewness and kurtosis for b_3 and b_4. Hence it should be emphasized that the quantity J_α should not be used with this parameter value $\alpha = \frac{1}{2}$ if the robustness is an issue.

3.2 Simulations

In order to examine the performance of the quantity J_α, let us compare it with an ad hoc objective function

$$J_{\text{ad hoc}}(Y) \;=\; \frac{36}{8\sqrt{3}-9}\Big(E\big(Ye^{-\frac{1}{2}Y^2}\big)\Big)^2 + \frac{1}{2-\frac{6}{\pi}}\Big(E(|Y|) - \sqrt{\frac{2}{\pi}}\Big)^2, \qquad (12)$$

which is proposed in [6], and with a mean nearest neighbor (meanNN) objective function [7]

$$J_{\text{meanNN}}(Y) = -\sum_{t=1}^{T-1}\sum_{t'=t+1}^{T}\log(|Y_t - Y_{t'}|). \qquad (13)$$

This meanNN objective is also similar to Renyi-entropy estimators [9]. We used two-dimensional random variables $X = (X_1, X_2)$ in which X_2 is a Gaussian, and X_1 is given different distributions X^a, X^b, and X^c as shown in the figure 3. We then measured $J_{\text{ad hoc}}$, J_{meanNN} and J_α on a one dimensional random variable $Y = w \cdot X$, where $w = (\cos(\theta), \sin(\theta))$, and plotted $J_{\text{ad hoc}}$, J_{meanNN} and J_α as functions of θ. For all choices $X_1 = X^a, X^b, X^c$, we generated samples of 10000 points, and results are shown in figures 4, 5 and 6. The number of coefficients b_n in J_α was $N = 10$.

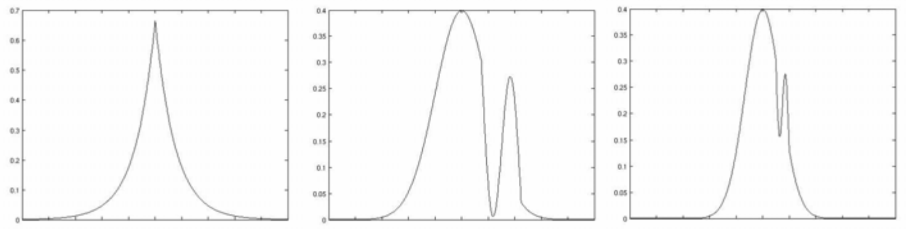

Fig. 3. From left to right the distributions X^a, X^b and X^c, which are used for X_1, while X_2 remains as a Gaussian. X^a is the Laplacian random variable, X^b is a Gaussian with a notable perturbation, and X^c a Gaussian with a smaller perturbation.

We see that sometimes, like with X_1 as Laplacian and X_2 as Gaussian, our objective J_α merely reproduces the same results that could have been obtained with some ad hoc objective function. On the other hand, if X_1 is almost a Gaussian with some perturbation, J_α can produce better results than some rivals, hence making it a serious tool for ICA.

It should be noted that it is a straightforward exercise to compute an explicit expression for the partial derivatives $\partial J_\alpha / \partial w_k$ (in terms of the derivatives of the Hermite polynomials.) Thus this objective function allows a simple application of the gradient method.

Another relevant remark is that the amount of time consumed for the computation of the coefficients b_n (and the gradient too) grows as $O(T)$ where T is the number of sample points. With meanNN algorithm the time grows as $O(T^2)$.

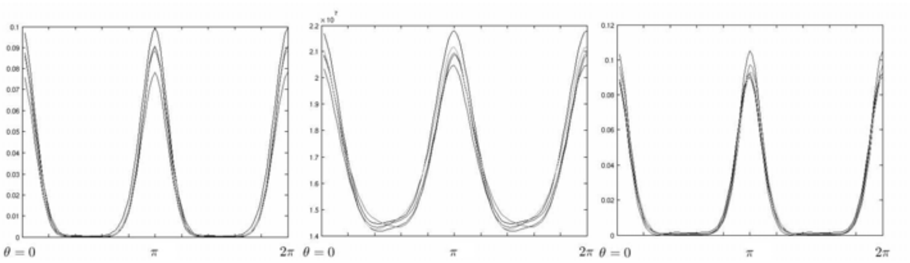

Fig. 4. The quantities $J_{\text{ad hoc}}$, J_{meanNN} and $J_{\alpha=\frac{1}{4}}$ (from left to right) plotted as a function of θ, when a choice $X = (X^a, X_2)$ was used. Samples were generated 5 times and results for all samplings are plotted here simultaneously. All objective functions succeed in finding the Laplacian in directions $\theta = 0$ and $\theta = \pi$.

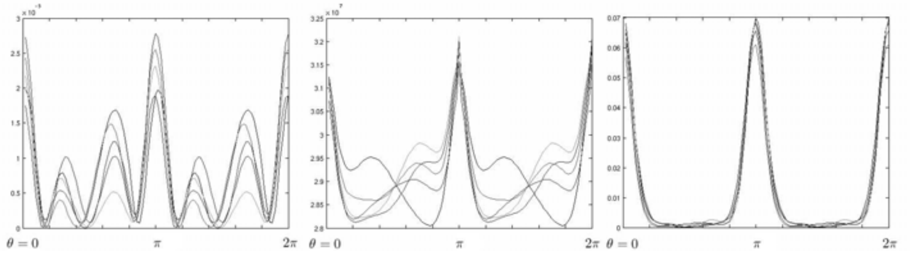

Fig. 5. The quantities $J_{\text{ad hoc}}$, J_{meanNN} and $J_{\alpha=\frac{1}{4}}$ plotted as a function of θ, when a choice $X = (X^b, X_2)$ was used. Samples were generated 5 times and results for all samplings are plotted here simultaneously. All functions have the correct global maxima, but $J_{\text{ad hoc}}$ has a severe local maxima problem.

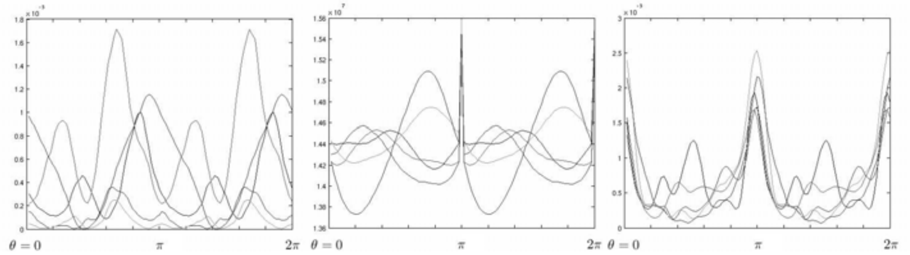

Fig. 6. The quantities $J_{\text{ad hoc}}$, J_{meanNN} and $J_{\alpha=\frac{1}{4}}$ plotted as a function of θ, when a choice $X = (X^c, X_2)$ was used. Samples were generated 5 times and results for all samplings are plotted here simultaneously. The J_α suffers from local maxima, but its overall performance is still better than those of $J_{\text{ad hoc}}$ and J_{meanNN}, whose results seem quite random.

4 Conclusions

We have proposed an ICA objective function, which is in a sense a very trivial application of the properties of the Hermite polynomials. Despite the triviality, with right choice of parameters, our objective function can be a very robust and precise measure of non-Gaussianity. Its advantage over ad hoc objective functions is that while ad hoc objective functions may work fine for certain distributions, our objective function measures all kinds of deviations from Gaussianity. Our objective function is also suitable for large data sizes, with only $O(T)$ asymptotic time consumption with respect to the sample size.

There also exists other strategies for ICA. For example, instead of measuring the non-Gaussianity, one can also directly attempt to measure the mutual information. Hulle [8] proposed a method for mutual information approximation using polynomial expansions. His method is not robust against outliers because it uses plain polynomials. We believe that the ideas we propose here for the measurement of non-Gaussianity could also be used to modify the methods of measuring the mutual information into a more robust form.

References

1. M. C. Jones, R. Sibson: What is Projection Pursuit? J. of Royal Statistical Society, Ser. A Vol 150, pages 1-36 (1987)
2. P. Comon: Independent component analysis—a new concept? Signal Processing Vol 36, pages 287-314 (1994)
3. S. Blinnikov, R. Moessner: Expansions for nearly Gaussian distributions. Astron. Astrophys. Suppl. Ser. 130, 193-205 (1998)
4. G. Szego: Orthogonal Polynomials. AMS (1939)
5. J. Duoandikoetxea: Fourier Analysis. AMS (2001)
6. A. Hyvarinen: New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit. Advances in Neural Information Processing Systems Vol 10, pages 273-279. (1998)
7. L. Faivishevsky, J. Goldberger: ICA based on a Smooth Estimation of the Differential Entropy. Advances in Neural Information Processing Systems Vol 21. (2009)
8. M. M. Van Hulle: Multivariate Edgeworth-based Entropy Estimation. Neural Computation Vol. 17, No. 9, pages 1903-1910 (2005)
9. D. Pal, P. Poczos, C. Szepesvari: Estimation of Renyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs. Advances in Neural Information Processing Systems (2010)