Robust Higher Order Statistics

Max Welling
School of Information and Computer Science
University of California Irvine
Irvine CA 92697-3425 USA
welling@ics.uci.edu

Abstract

Sample estimates of moments and cumulants are known to be unstable in the presence of outliers. This problem is especially severe for higher order statistics, like kurtosis, which are used in algorithms for independent components analysis and projection pursuit. In this paper we propose robust generalizations of moments and cumulants that are more insensitive to outliers but at the same time retain many of their desirable properties. We show how they can be combined into series expansions to provide estimates of probability density functions. This in turn is directly relevant for the design of new robust algorithms for ICA. We study the improved statistical properties such as B-robustness, bias and variance while in experiments we demonstrate their improved behavior.

1 INTRODUCTION

Moments and cumulants are widely used in scientific disciplines that deal with data, random variables or stochastic processes. They are well known tools that can be used to quantify certain statistical properties of the probability distribution like location (first moment) and scale (second moment). Their definition is given by,

$$\mu_n = \mathbb{E}[x^n] \tag{1}$$

where $\mathbb{E}[\cdot]$ denotes the average over the probability distribution $p(x)$. In practise we have a set of samples from the probability distribution and compute sample estimates of these moments. However, for higher order moments these estimates become increasingly dominated by outliers, by which we will mean the samples which are far away from the mean. Especially for heavy tailed distributions this implies that these estimates have high variance and are generally unsuitable to measure properties of the distribution.

An undesirable property of moments is the fact that lower order moments can have a dominating influence on the value of higher order moments. For instance, when the mean is large it will have a dominating effect on the second order moment,

$$\mathbb{E}[x^2] = \mathbb{E}[x]^2 + \mathbb{E}[x - \mathbb{E}[x]]^2 \tag{2}$$

The second term which measures the variation around the mean, i.e. the variance, is a much more suitable statistic for scale than the second order moment. This process of subtracting lower order information can be continued to higher order statistics. The resulting estimators are called centralized moments or cumulants. Well known higher order cumulants are skewness (third order) measuring asymmetry and kurtosis (fourth order) measuring "peakiness" of the probability distribution . Explicit relations between cumulants and moments are given in appendix A (set $\mu_0 = 1$ for the classical case). Since cumulants are functions of moments up to the same order, they also suffer from high sensitivity to outliers.

Many statistical methods and techniques use moments and cumulants because of their convenient properties. For instance they follow easy transformation rules under affine transformations. Examples in the machine learning literature are certain algorithms for independent components analysis [3, 2, 1]. A well known downside of these algorithm is their sensitivity to outliers in the data. Thus, there is a need to define robust cumulants which are relatively insensitive to outliers but retain most of the convenient properties that moments and cumulants enjoy. This will be the topic of this paper.

2 MOMENTS AND CUMULANTS

A formal definition of the relation between moments and cumulants to all orders can be given in terms the characteristic function (or moment generating function) of a probability distribution,

$$\Psi(t) = \mathbb{E}[e^{ixt}] = \sum_{n=0}^{\infty} \frac{1}{n!}\mu_n (it)^n \tag{3}$$

where the last expression follows by Taylor expanding the exponential. The cumulants can now be defined by

$$\sum_{n=0}^{\infty} \frac{1}{n!} \kappa_n (it)^n = \ln \Psi(t) \tag{4}$$

where we expand the right hand side in powers of (it) and match terms at all orders.

The generalization of the above to the multivariate case is straightforward. Moments are defined as expectations of monomials,

$$\mu_{i_1,\dots,i_m} = \mathbb{E}[x_{i_1} \dots x_{i_m}] \tag{5}$$

and the cumulants are again defined through the characteristic function (see Eq.7), where in addition to the univariate cumulants we now also have cross-cumulants.

From the definition of the cumulants in terms of the moments we can derive a number of interesting properties, which we will state below. It will be our objective to conserve most of these properties when we define the robust cumulants.

Lemma 1 *The following properties are true for cumulants:*

- **I.** *For a Gaussian density, all cumulants higher than second order vanish.*
- **II.** *For independent random variables, all cross-cumulants vanish.*
- **III.** *All cumulants transform multi-linearly with respect to affine transformations.*
- **IV.** *All cumulants higher than first order are invariant with respect to translations.*

The proofs of these statements can for instance be found in [9] and are very similar to the proofs for the robust cumulants which we will present in the next section.

3 ROBUST MOMENTS AND CUMULANTS

In this section we will define robust moments and cumulants by introducing an isotropic decay factor which downweights outliers. With this decay factor we will have introduced a preferred location and scale. We therefore make the following important assumption: *The probability density function has zero-mean and unit-variance (or covariance equal to the identity in the multivariate case).* This can always be achieved by a linear transformation of the random variables. Analogously, data will need to be centered and sphered. One may worry that these preprocessing steps are non-robust operations. Fortunately, we can rely on an extensive body of literature [5][6] to compute robust estimates of location and scale.

As will become apparent in the following, a convenient choice for the robust moments is given by the following expression,

Definition 1 *The robust moments are given by:*

$$\mu^{(\alpha)}_{i_1 \dots i_n} = \mathbb{E}\left[(\alpha x_{i_1}) \dots (\alpha x_{i_n}) \frac{\phi(\alpha \mathbf{x})}{\phi(\mathbf{x})}\right] \qquad \alpha \geq 1 \tag{6}$$

where $\phi(\mathbf{x})$ is the multivariate standard normal density.

The decaying factor is thus given by $\frac{\phi(\alpha\mathbf{x})}{\phi(\mathbf{x})} = \alpha^d \exp(-\frac{1}{2}(\alpha^2 - 1)\mathbf{x}^T\mathbf{x})$, where d is the dimension of the space. In the limit $\alpha \to 1$ we obtain the usual definition of moments.

In order to preserve most of the desirable properties that cumulants obey, we will use the same definition to relate moments to cumulants as in the classical case,

Definition 2 *The robust cumulants are defined by:*

$$\sum_{n=0}^{\infty} \sum_{i_1=1}^{M} \dots \sum_{i_n=1}^{M} \frac{1}{n!} \kappa^{(\alpha)}_{i_1 \dots i_n} (it_{i_1}) \dots (it_{i_n}) = \\ \ln\left(\sum_{m=0}^{\infty} \sum_{j_1=1}^{M} \dots \sum_{j_m=1}^{M} \frac{1}{m!} \mu^{(\alpha)}_{j_1 \dots j_m} (it_{j_1}) \dots (it_{j_m})\right) \tag{7}$$

The right hand side can again be defined as the logarithm of the moment generating function for robust moments,

$$\Psi^{(\alpha)}(\mathbf{t}) = \mathbb{E}\left[\exp(i\alpha \mathbf{x}^T \mathbf{t}) \frac{\phi(\alpha\mathbf{x})}{\phi(\mathbf{x})}\right] \tag{8}$$

The explicit relation between robust moments and cumulants up to fourth order is given in appendix A.

With the above definitions we can now state some important properties for the robust cumulants. Since we assume zero-mean and unit-variance we cannot expect the cumulants to be invariant with respect to translation and scalings. However, we will prove that the following properties are still valid,

Theorem 1 *The following properties are true for robust cumulants:*

- **I.** *For a standard Gaussian density, all robust cumulants higher than second order vanish.*
- **II.** *For independent random variables, robust cross-cumulants vanish.*
- **III.** *All robust cumulants transform multi-linearly with respect to rotations.*

Proof: I: For a standard Gaussian we can compute the moment generating function analytically giving $\Psi^{(\alpha)}(\mathbf{t}) = -\frac{1}{2}\mathbf{t}^T\mathbf{t}$, implying that $\kappa^{(\alpha)}_{i_1 i_2} = \delta_{i_1 i_2}$ and all other cumulants vanish.
II: We note that if the variables $\{x_i\}$ are independent, $\Psi^{(\alpha)}(\mathbf{t})$ factorizes into a product of expectations which the logarithm turns into a sum, each term only depending on one t_i. Since cross cumulants on the left hand side of Eq.7 are precisely those terms which contain distinct t_i, they must be zero.
III: From Eq.6 we see that since the decay factor is isotropic, robust moments still transform multi-linearly with respect to rotations. If we rotate both the moments and $\mathbf{t}$ in the right-hand side of Eq.7, it remains invariant. To ensure that the left-hand side of Eq.7 remains invariant we infer that the robust cumulants must also transform multi-linearly with respect to rotations,

$$\kappa^{(\alpha)}_{i_1\ldots i_n} \rightarrow O_{i_1 j_1}\ldots O_{i_n j_n}\kappa^{(\alpha)}_{j_1\ldots j_n}, \qquad \mathbf{O}\mathbf{O}^T = \mathbf{O}^T\mathbf{O} = \mathbf{I} \tag{9}$$

This concludes the proof. ⊠

4 ROBUST GRAM-CHARLIER AND EDGEWORTH EXPANSIONS

Assuming we have computed robust cumulants (or equivalently robust moments) up to a given order, can we combine them to provide us with an estimate of the probability density function? For the classical case it is long known that the Gram-Charlier and Edgeworth expansions are two possibilities [8]. In this section we will show that these expansions can be generalized to the robust case as well. To keep things simple, we will discuss the univariate case here. Multivariate generalizations are relatively straightforward.

Both robust Gram-Charlier and Edgeworth expansions will be defined as series expansions in the scaled Hermite polynomials $H_n(\alpha x)$.

$$p(x) = \sum_{n=0}^{\infty} c^{(\alpha)}_n H_n(\alpha x)\phi(x) \quad \text{with} \tag{10}$$

$$c^{(\alpha)}_n = \frac{1}{n!}\int_{-\infty}^{\infty} p(x) H_n(\alpha x)\phi^{-1}(x)\, d\nu_\alpha \tag{11}$$

where we have defined the measure $d\nu_\alpha = \phi(\alpha x)\, dx$ and used the following generalized orthogonality relation,

$$\int_{-\infty}^{\infty} H_n(\alpha x) H_m(\alpha x)\, d\nu_\alpha = n!\,\delta_{nm} \tag{12}$$

When $c^{(\alpha)}_n$ is estimated by averaging over samples (Eq.25), we see that the decay factor $\frac{\phi(\alpha x)}{\phi(x)}$ will again render them robust against outliers.

We may also express the above series expansion directly in terms of the robust cumulants. The explicit expression is given by the following theorem[1],

Theorem 2 *The series expansion of a density $p(x)$ in terms of its robust cumulants is given by*

$$p(x) = \frac{\phi(x)}{\phi(\alpha x)} e^{\left(\sum_{n=0}^{\infty} \frac{1}{n!}\tilde{\kappa}^{(\alpha)}_n (-1)^n \frac{d^n}{d(\alpha x)^n}\right)} \phi(\alpha x) \tag{13}$$

$$\textit{with} \qquad \tilde{\kappa}^{(\alpha)}_n = \kappa^{(\alpha)}_n - \delta_{n,2} \tag{14}$$

Proof: see appendix B.

To find an explicit expression up to a certain order in the robust cumulants, one expands the exponential and uses $(-1)^n \frac{d^n}{dx^n}\phi(x) = H_n(x)\phi(x)$ to convert derivatives into Hermite polynomials.

Analogous to the classical literature we will talk about a Gram-Charlier expansion when we expand in $c^{(\alpha)}_n$ and an Edgeworth expansion when we expand in $\kappa^{(\alpha)}_n$. Their only difference is therefore in their convention to break the series off after a finite number of terms.

When $\alpha = 1$ the Hermite expansions discussed in this section will be normalized, even when only a finite number of terms is taken into account. This holds since $H_0 = 1$ and $c_0 = 1/N\sum_n 1 = 1$, while all higher order polynomials are orthogonal to "1". When generalizing to robust cumulants this however no longer holds true. To correct this we will add an extra term to the expansion,

$$p_R(x) = \{\sum_{n=0}^{R} c^{(\alpha)}_n H_n(\alpha x) + \psi(x)\}\phi(x), \tag{15}$$

The correction factor can be computed by a Gram-Schmidt procedure resulting in,

$$\psi(x) = (1 - \sum_{n=0}^{R} n!\, a_n c^{(\alpha)}_n)\left(\frac{\phi(x)}{\phi(\alpha x)} - \sum_{n=0}^{R} a_n H_n(\alpha x)\right). \tag{16}$$

with $a_n = \frac{(n-1)!!}{n!}(\alpha^2 - 1)^{\frac{n}{2}}\ \delta_{n,2k}$ for $k \in \{0,1,2,3,...\}$ and $(n-1)!!$ denotes the double factorial of $(n-1)$ defined by $1\cdot 3\cdot 5 ...(n-1)$. The correction factor is thus orthogonal to all Hermite polynomials $H_n(\alpha x)$ with $n = 1..R$ under the new measure $d\nu_\alpha$. We can also show that $p_R(x)$ always integrates to 1 and that when $\alpha \rightarrow 1$ the correction term will reduce to $\psi(x) \rightarrow c_{R+K} H_{R+K}(x)$ with $K = 1$ when R is odd and $K = 2$ when R is even. Finally we note that since $\int_{-\infty}^{\infty}\phi^2(x)/\phi(\alpha x)dx = 1/(\alpha\sqrt{2-\alpha^2})$ the

---

[1] The equivalent result in the multivariate case is,

$$p(\mathbf{x}) = \frac{\phi(\mathbf{x})}{\phi(\alpha\mathbf{x})} \times$$
$$e^{\left(\sum_{n=0}^{\infty}\sum_{i_1=1}^{M}\cdots\sum_{i_n=1}^{M} \frac{1}{n!}\tilde{\kappa}^{(\alpha)}_{i_1\ldots i_n}(-1)^n \frac{d}{d(\alpha x)_{i_1}}\cdots\frac{d}{d(\alpha x)_{i_n}}\right)}\phi(\alpha\mathbf{x})$$
$$\text{with} \qquad \tilde{\kappa}^{(\alpha)}_{i_1 i_2} = \kappa^{(\alpha)}_{i_1 i_2} - \delta_{i_1 i_2}$$

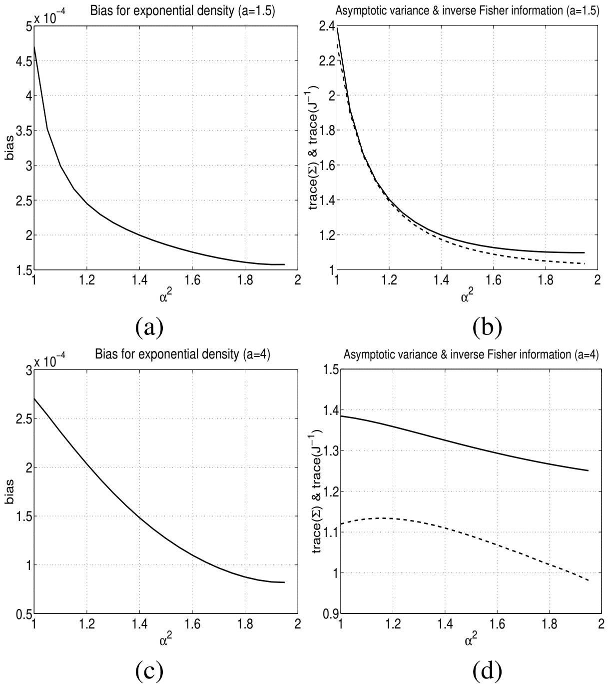


Figure 1: (a)-Bias as a function of α^2 for a generalized Laplacian with $a = 1.5$ (super-Gaussian). (b)-Asymptotic variance (solid line) and inverse Fisher information (dashed line) as a function of α^2 for $a = 1.5$. (c)-(d) Similar plots for $a = 4$ (sub-Gaussian)

correction is only normalizable for $\alpha^2 < 2$, which is what we will assume in the following.

5 CONSISTENCY, ROBUSTNESS, BIAS AND VARIANCE

In this section we will examine the robustness, bias and efficiency of our generalized expansion. Many definitions in this section are taken from [5]. Our analysis will assume that the data arrive centered and sphered, which allows us to focus on the analysis of the higher order statistics. For a thorough study of the robustness properties of first and second order statistics see [5].

First we mention that the estimators $\hat{c}_n^{(\alpha)}[p_R]$ for the truncated series expansion (Eq.15) are Fisher consistent. This can be shown by replacing $p(x)$ in Eq.11 with $p_R(x)$ and using orthogonality between $\psi(x)$ and the Hermite polynomials $H_n(\alpha x)$ $n = 1..R$ w.r.t. the measure $d\nu_\alpha$.

To prove B-robustness we need to define and calculate the influence function IF for the estimators $\hat{c}_n^{(\alpha)}$. Intuitively, the influence function measures the sensitivity of the estimators to adding one more observation at location x,

$$IF(x) = \lim_{t \to 0} \frac{c_n^{(\alpha)}[(1-t)p_R + t\delta_x] - c_n^{(\alpha)}[p_R]}{t}. \quad (17)$$

An estimator is called B-robust if its influence function is finite everywhere. We will now state the following result.

Theorem 3 *The estimates $\hat{c}_n^{(\alpha)}[p_R]$ are B-robust for $\alpha > 1$.*

Proof: It is straightforward to compute the influence function defined in Eq.17,

$$IF(x) = \frac{1}{n!} H_n(\alpha x) \frac{\phi(\alpha x)}{\phi(x)} - c_n^{(\alpha)} \quad (18)$$

Since for $\alpha > 1$ this IF is finite everywhere, the result follows. ⊠

Since cumulants are simple functions of the $c_n^{(\alpha)}$ up to the same order, we conclude that cumulants are also B-robust. It is important to notice that in the classical case ($\alpha = 1$) the theorem does not hold, confirming that classical cumulants are not robust. Analogously one can show that the sensitivity to shifting data-points is also bounded for $\alpha > 1$.

We now turn to the analysis of bias and variance. It is well known that the point-wise mean square error can be decomposed into a bias and a variance term,

$$\mathbf{MSE}_x(p_R^{(N)}(x)) = \mathbb{E}\left[(p_R^{(N)}(x) - p(x))^2\right] =$$
$$\mathbb{E}\left[(p_R^{(N)}(x) - p_R(x))^2\right] + (p_R(x) - p(x))^2 \quad (19)$$

where $p_R^{(N)}$ is the estimate of p_R using a sample of size N. The expectation $\mathbb{E}$ is taken over an infinite number of those samples. Clearly, the first term represents the variance and the second the bias which is independent of N. The variance term (V) may be rewritten in terms of the influence function,

$$V = \frac{1}{N} \sum_{n,m=0}^{R} \Sigma(c_n^{(\alpha)}, c_m^{(\alpha)}) H_n(\alpha x) H_m(\alpha x) \phi^2(x) \quad (20)$$

$$\Sigma(c_n^{(\alpha)}, c_m^{(\alpha)}) = \int_{-\infty}^{\infty} p(x) IF(x, c_n^{(\alpha)}) IF(x, c_m^{(\alpha)}) dx \quad (21)$$

So the variance decreases as $1/N$ with sample size while the data independent part is completely determined by the asymptotic covariance matrix Σ which is expressed in terms of the influence function.

Finally, by defining the Fisher information as,

$$J(c_n^{(\alpha)}, c_m^{(\alpha)}) = \mathbb{E}\left[\frac{1}{p(x)} \frac{\partial}{\partial c_n^{(\alpha)}} p_R(x) \frac{1}{p(x)} \frac{\partial}{\partial c_m^{(\alpha)}} p_R(x)\right]_p$$
$$= \int_{-\infty}^{\infty} \frac{H_n(\alpha x) H_m(\alpha x) \phi(x)^2}{p(x)}\, dx \quad (22)$$

the well known Cramer-Rao bound follows: $\Sigma(c_n^{(\alpha)}, c_m^{(\alpha)}) \geq J^{-1}(c_n^{(\alpha)}, c_m^{(\alpha)})$.

In figure 1 we plot the bias and the total variation (trace of the covariance) as a function of α^2 for a super-Gaussian and a sub-Gaussian density (generalized Laplace density $p \propto \exp(-b|x|^a)$ with unit variance and $a = 1.5$ and $a = 4$ respectively) . The trace of the inverse Fisher information was also plotted (dashed line). The model included 10 orders in the expansion $n = 0, ..., 9$ plus the normalization term $\psi(x)$. All quantities were computed using numerical integration. We conclude that *both* bias and efficiency improve when α moves away from the classical case $\alpha = 1$.

6 INDEPENDENT COMPONENTS ANALYSIS

Although robust moments and cumulants can potentially find applications in a broad range of scientific disciplines, we will illustrate their usefulness by showing how they can be employed to improve algorithms for independent components analysis (ICA). The objective in ICA is to find a new basis for which the data distribution factorizes into a product of independent one-dimensional marginal distributions. To achieve this, one first removes first and second order statistics from the data by shifting the sample mean to the origin and sphering the sample covariance to be the identity matrix. These operations render the data *de-correlated* but higher order dependencies may still remain. It can be shown [2] that if an independent basis exists, it must be a rotation away from the basis in which the data is de-correlated, i.e. $\mathbf{x}_{ica} = \mathbf{O}\mathbf{x}_{decor}$ where $\mathbf{O}$ is a rotation. One approach to find $\mathbf{O}$ is to propose a contrast function that, when maximized, returns a basis onto which the data distribution is a product of independent marginal distributions. Various contrast functions have been proposed, e.g. the neg-entropy [4] and the mutual information [1]. All contrast functions share the property that they depend on the marginal distributions which need to be estimated from the data. Naturally, the Edgeworth expansion [4, 3] and the Gram-Charlier expansion [1] have been proposed for this purpose. This turns these contrast functions into functions of moments or cumulants. However, to obtain reliable estimates one needs to include cumulants of up to fourth order. It has been observed frequently that in the presence of outliers these cumulants often become unreliable (e.g. [7]).

We propose to use the robust Edgeworth and Gram-Charlier expansions discussed in this paper instead of the classical ones. As we will show in the experiments below, it is safe to include robust cumulants to very high order in these expansions (we have gone up to order 20), which at a moderate computational cost will have a significant impact on the accuracy of our estimates of the marginal distributions. We note that the derivation of the contrast function in e.g. [4] crucially depends on properties I,II and III from theorem 1. This makes our robust cumulants the ideal candidates to replace the classical ones. Instead of going through this derivation we will argue for a novel contrast function that represents a slight generalization of the one proposed in [4],

$$I(\mathbf{O}) = \sum_{n=1}^{R}\sum_{i=1}^{M} w_n (\tilde{\kappa}_{i...i}^{(\alpha)})^2 \qquad w_n \geq 0, \tag{23}$$

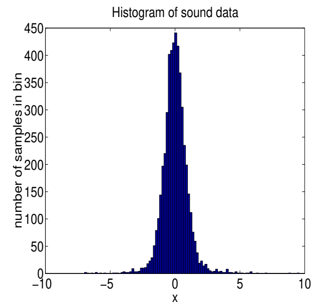


Figure 2: Histogram of sound-data (5000 samples).

where $\tilde{\kappa}_{i...i}^{(\alpha)}$ only differ from the usual $\kappa_{i...i}^{(\alpha)}$ in second order, $\tilde{\kappa}_{ii}^{(\alpha)} = \kappa_{ii}^{(\alpha)} - 1$. These cumulants are defined on the rotated axis $\mathbf{e}'_i = \mathbf{O}^T\mathbf{e}_i$.

We will now state a number of properties that show the validity of $I(\mathbf{O})$ as a contrast function for ICA,

Theorem 4 *The following properties are true for* $I(\mathbf{O})$*:*

i. $I(\mathbf{O})$ *is maximal if the probability distribution on the corresponding axis factors into an independent product of marginal distributions.*

ii. $I(\mathbf{O})$ *is minimal (i.e.* 0*) if the marginal distributions on the corresponding axis are Gaussian.*

Proof: To prove (i) we note that the following expression is scalar (i.e. invariant) w.r.t. rotations[2],

$$\sum_{i_1...i_n} (\tilde{\kappa}_{i_1...i_n}^{(\alpha)})^2 = constant \qquad \forall n \tag{24}$$

We now note that this expression can be split into two terms: a sum over the "diagonal terms" where $i_1 = i_2 = ... = i_n$ and a sum over all the remaining cross-cumulant terms. When all directions are independent all cross-cumulants must vanish by property II of theorem 1. This minimizes the second term (since it's non-negative). Hence, by the fact the sum of these terms is constant, the first term, which equals $I(\mathbf{O})$, must be maximal for independent directions.
To prove (ii) we invoke property I of theorem 1 that for Gaussian random variables all cumulants $\tilde{\kappa}$ must vanish. ⊠

By the above theorem we see that $I(\mathbf{O})$ simultaneously searches for independent directions and non-Gaussian directions. Observe however, that for practical reasons we have ignored cumulants of order higher than R. Hence, there will certainly be more than one distribution which

[2]For vectors this reduces to the statement that an inner product is scalar. To prove the general case we use $\mathbf{O}^T\mathbf{O} = \mathbf{I}$ for every index separately.

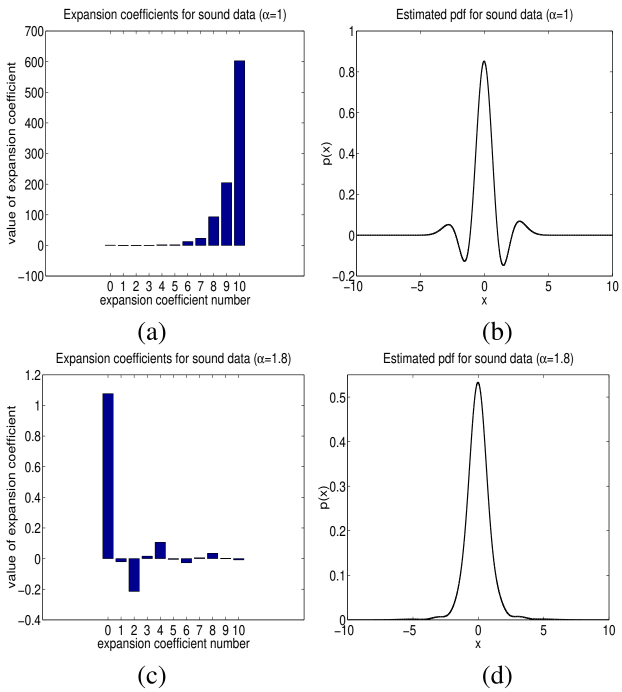


Figure 3: (a)-Expansion coefficients for classical Gram-Charlier expansion ($\alpha = 1$). (b)-Density estimate for $\alpha = 1$ after four orders. The negative tails signal the onset of a diverging series. (c)-Decreasing expansion coefficients for $\alpha = 1.8$. (f)-Density estimate after 10 orders for $\alpha = 1.8$.

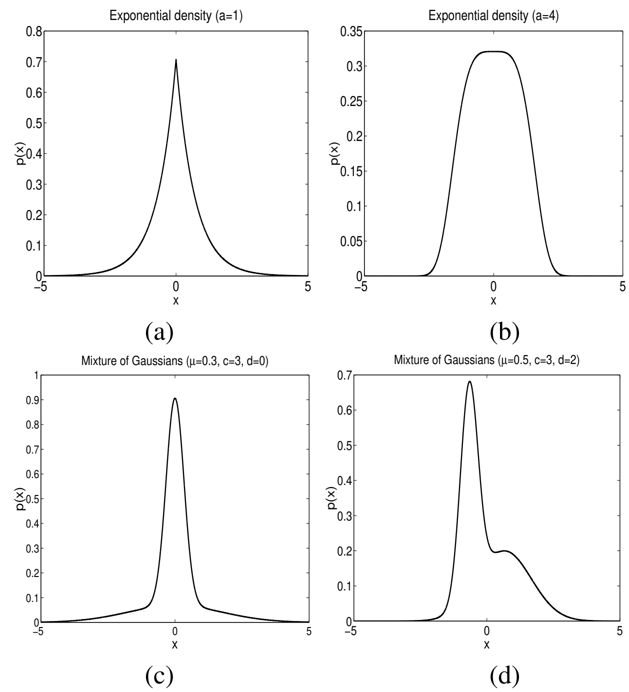


Figure 4: Top row: Generalized Laplace distributions with (a) $a = 1$, (b) $a = 4$. Bottom row: Mixture of Gaussians with (c) $\mu = 0.3$, $c = 3$, $d = 0$ and (d) $\mu = 0.5$, $c = 3$, $d = 2$.

maximizes $I(\mathbf{O})$ (for instance distributions which only differ in the statistics of order higher than R). Good objective functions are discriminative in the sense that there are only few (relevant) densities that maximize it. We can influence the ability of $I(\mathbf{O})$ to discriminate by changing the weighting factors w_n. Doing this allows for a more directed search towards predefined qualities, e.g. a search for high kurtosis directions would imply a large w_4.

A straightforward strategy to maximize $I(\mathbf{O})$ is gradient ascent while at every iteration projecting the solution back onto the manifold of rotations (e.g. see [10]). A more efficient technique which exploits the tensorial property of cumulants (i.e. property III of theorem 1) was proposed in [3]. This technique, called Jacobi-optimization, iteratively solves two dimensional sub-problems analytically.

7 EXPERIMENTS

The following set of experiments focus on density estimates based on the Gram-Charlier expansion (Eq.10) where we replace Eq.11 with a sample estimate,

$$\hat{c}_n^{(\alpha)} = \frac{1}{N}\frac{1}{n!}\sum_{A=1}^{N}\frac{\phi(\alpha x_A)}{\phi(x_A)}H_n(\alpha x_A) \tag{25}$$

The reason we focus on this task is that we can demonstrate robustness by showing that low order robust statistics are always dominant over higher order robust statistics, even for heavy tailed distributions. Yet at the same time they carry the relevant information of the probability density, i.e. they combine into an accurate estimate of it. This exercise is also relevant for cumulant based algorithms for independent components analysis because they rely on the fact that the Gram-Charlier or Edgeworth expansions describe the source distributions well.

Sound Data

We downloaded recordings from music CD's [3] and extracted 5000 samples from it. The histogram is shown in figure 2. Due to the presence of outliers we expect the classical expansion to break down. This can be observed from figure (3a) where the coefficients *increase* with the order of the expansion. In figure (3b) we see that the density estimate has become negative in the tails after 4 orders, which is an indication that the series has become unstable. In figures (3c,d) we see that for the robust expansion at $\alpha = 1.8$ the coefficients decrease with order and the estimate of the density is very accurate after 10 orders.

Synthetic Data

In this experiment we sampled 5000 data-points from two generalized Laplace densities $p \propto \exp(-b|x|^a)$ (figures 4a,b) and from two mixtures of two Gaussians parameterized as
$p_{\mathbf{mog}}(x) = \mu a\phi(ax + b) + c(1 - \mu)\phi(cx + d)$ (figures 4c,d). These include super-Gaussian distributions (figures

[3] http://sweat.cs.unm.edu/bap/demos.html

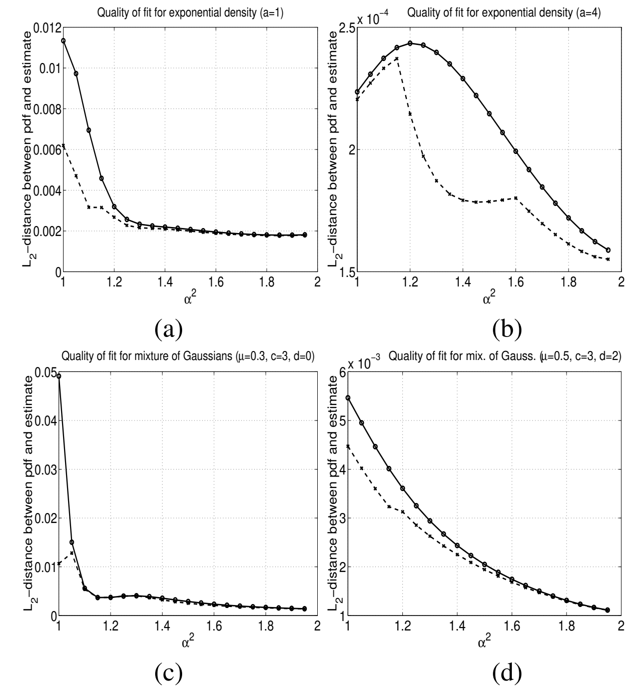


Figure 5: Top row: total L_2 distance between true and estimated densities as function of α^2 for generalized Laplace density with (a) $a = 1$, (b) $a = 4$. Bottom row: same as top row for the mixture of Gaussians distributions with (c) $\mu = 0.3$, $c = 3$, $d = 0$ and (d) $\mu = 0.5$, $c = 3$. The corresponding densities are shown in figure 4. Dashed line indicates the best estimate over all orders.

4a,c), a sub-Gaussian density (figures 4b) and an asymmetric density (figures 4d). We plot the total L_2 distance between the estimate and the true density as we vary α (figures 5a,b,c,d). Shown is the best estimate over all orders orders (dashed line) and the final estimate after 20 orders. In both cases it is observed that the best estimates are obtained around $\alpha^2 \approx 2$ (but recall that $\alpha^2 < 2$, see section 4. We also plot the L_2 distance between true and estimated density as a function of the order of the expansion for $\alpha^2 = 1$ and $\alpha^2 = 1.9$ ($a = 1$) in figures (6a,b). Clearly, the robust expansion converges while the classical expansion is unstable. Finally, in figure 7 we compare the best estimated PDFs for the general Laplace density at $a = 1$ with $\alpha^2 = 1$ (a) and $\alpha^2 = 1.9$ (b).

The general conclusion from these experiments is that in all cases (super- or sub-Gaussian PDF, symmetric or asymmetric PDF) we find that the quality (in L_2-norm) of the estimated densities improves considerably when we use the robust series expansion with a setting of α^2 close to (but smaller than) 2. This effect is more pronounced for super-Gaussian densities than for sub-Gaussian densities.

8 DISCUSSION

In this paper we have proposed robust alternatives to higher order moments and cumulants. In order to arrive at robust

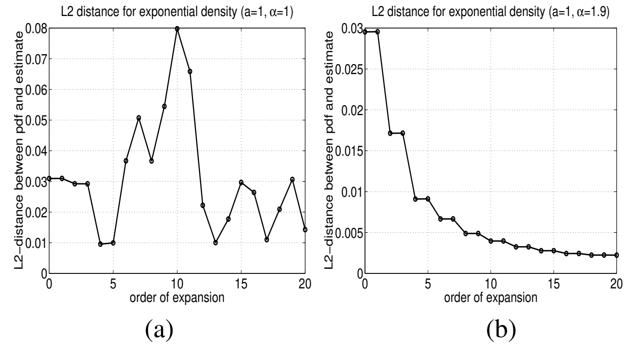


Figure 6: L_2-distance as a function of the order of the expansion for (a) $\alpha^2 = 1$ and (b) $\alpha^2 = 1.9$ for the generalized Laplace PDF with $a = 1$.

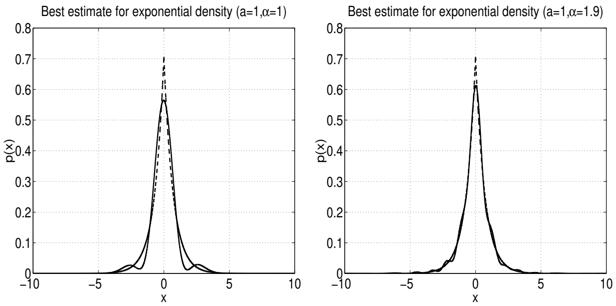
cumulants invariance w.r.t. translations was lost and the class of transformations under which they transform multilinearly was reduced from affine to orthogonal (i.e. rotations). However, all other cumulant properties were conveniently preserved. We argue that by first centering and sphering the data (using robust techniques described in the literature [5]), multi-linearity w.r.t. orthogonal transformations is all we need, which could make the trade-off with improved robustness properties worthwhile.

There is two well-known limitations of cumulants that one needs to be aware of. Firstly, they are less useful as statistics characterizing the PDF if the mass is located far away from the mean. Secondly, the number of cumulants grows exponentially fast with the dimensionality of the problem. With these reservations in mind, many interesting problems remain, even in high dimensions, that are well described by cumulants of low dimensional marginal distributions, as the ICA example has illustrated.

The sensitivity to outliers can be tuned with the parameter $\alpha^2 \in [1, 2)$. Our experiments have shown that if one includes many orders in the expansion, optimal performance was obtained when α^2 was close to (but smaller than) 2. Although unmistakeably some information is ignored by weighting down the impact of outliers, the experiments indicated that the relevant information to estimate the PDF was mostly preserved. In future experiments we hope to show that this phenomenon is also reflected in improved performance of ICA algorithms based on robust cumulants.

A ROBUST MOMENTS AND CUMULANTS TO 4'TH ORDER

This appendix contains the definition of the cumulants in terms of the moments and vice versa for general α. We have not denoted α explicitely in the following for nota-



(a) (b)

Figure 7: Best estimates for the generalized Laplace density at $a = 1$. In (a) we plot the best classical estimate which is found after four orders of Hermite polynomials are taken into account (i.e. $H_0(x), ..., H_4(x)$). For higher orders, the series becomes unstable and the calculation of the expansion coefficients is too sensitive to sample fluctuations. The best estimate from the robust expansion is depicted in (b). In that case the best estimate is found when all orders are taken into account, i.e. 20.

tional convenience.

$$\kappa_0 = \ln \mu_0 \qquad \kappa_1 = \frac{\mu_1}{\mu_0} \qquad \kappa_2 = \frac{\mu_2}{\mu_0} - (\frac{\mu_1}{\mu_0})^2$$
$$\kappa_3 = \frac{\mu_3}{\mu_0} - 3\frac{\mu_1\mu_2}{\mu_0^2} + 2(\frac{\mu_1}{\mu_0})^3$$
$$\kappa_4 = \frac{\mu_4}{\mu_0} - 3(\frac{\mu_2}{\mu_0})^2 - 4\frac{\mu_1\mu_3}{\mu_0^2} + 12\frac{\mu_1^2\mu_2}{\mu_0^3} - 6(\frac{\mu_1}{\mu_0})^4$$
$$\mu_0 = e^{\kappa_0} \qquad \frac{\mu_1}{\mu_0} = \kappa_1 \qquad \frac{\mu_2}{\mu_0} = \kappa_2 + \kappa_1^2$$
$$\frac{\mu_3}{\mu_0} = \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3$$
$$\frac{\mu_4}{\mu_0} = \kappa_4 + 4\kappa_3\kappa_1 + 3\kappa_2^2 + 6\kappa_2\kappa_1^2 + \kappa_1^4$$

B PROOF OF THEOREM 2

The characteristic function or moment generating function of a PDF is defined by:

$$\Psi(t) = \int_{-\infty}^{\infty} e^{ixt} p(x)\, dx = \sum_{n=0}^{\infty} \frac{1}{n!}\mu_n (it)^n = \mathcal{F}[p(x)] \tag{26}$$

where the last term follows from Taylor expanding the exponential and $\mathcal{F}$ denotes the Fourier transform. For arbitrary α we have,

$$\Psi^{(\alpha)}(t) = \int_{-\infty}^{\infty} e^{i\alpha xt} p(x) \frac{\phi(\alpha x)}{\phi(x)} dx$$
$$= \sum_{n=0}^{\infty} \frac{1}{n!} \mu^{(\alpha)}{}_n (it)^n dx = \mathcal{F}[p(x)\frac{\phi(\alpha x)}{\alpha\phi(x)}]. \tag{27}$$

Where in the last equality the definition of the generalized moments (Eq.6) was used. $\Psi^{(\alpha)}$ is the *(robust) moment generating function* of $p(x)$. We can find an expression for $p(x)$ if we invert the Fourier transform,

$$p(x) = \frac{\alpha\phi(x)}{\phi(\alpha x)} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\alpha xt} \Psi^{(\alpha)}(t)\, dt. \tag{28}$$

Next, we use the relation between the cumulants and the moments (Eq.7) to write,

$$p(x) = \frac{\alpha\phi(x)}{\phi(\alpha x)} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\alpha xt}\, e^{\sum_{n=0}^{\infty} \frac{1}{n!}\kappa_n^{(\alpha)} (it)^n}\, dt. \tag{29}$$

By defining $\tilde{\kappa}_n^{(\alpha)} = \kappa_n^{(\alpha)} - \delta_{n,2}$ we can separate a factor $\phi(t)$ (Gaussian) inside the integral,

$$p(x) = \frac{\alpha\phi(x)}{\phi(\alpha x)} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\alpha xt}\, e^{\sum_{n=0}^{\infty} \frac{1}{n!}\tilde{\kappa}_n^{(\alpha)} (it)^n} \phi(t)\, dt. \tag{30}$$

Finally, we will need the result

$$\mathcal{F}^{-1}[(it)^n \phi(t)] = \frac{\sqrt{2\pi}}{\alpha}(-1)^n \frac{d^n}{d(\alpha x)^n}\phi(\alpha x). \tag{31}$$

If we expand the exponential containing the cumulants in a Taylor series, and do the inverse Fourier transform on every term separately, after which we combine the terms again in an exponential, we find the desired result (Eq.14).

References

[1] S. Amari, A. Cichocki, and H.H. Yang. A new algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8:757–763, 1996.

[2] A.J. Bell and T.J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[3] J.F. Cardoso. High-order constrast for independent component analysis. *Neural Computation*, 11:157–192, 1999.

[4] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[5] F.R. Hampel, E.M. Ronchetti, P.J. Rousseuw, and W.A. Stahel. *Robust statistics*. Wiley, 1986.

[6] P.J. Huber. *Robust statistics*. Wiley, 1981.

[7] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, volume 10, pages 273–279, 1998.

[8] M.G. Kendall and A. Stuart. *The advanced theory of statistics Vol. 1*. Griffin, 1963.

[9] P. McCullagh. *Tensor Methods in Statistics*. Chapman and Hall, 1987.

[10] M. Welling and M. Weber. A constrained EM algorithm for independent component analysis. *Neural Computation*, 13:677–689, 2001.