Sufficient observables for large scale structure in galaxy surveys

J. Carron* and I. Szapudi
Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI, 96822

24 October 2013

ABSTRACT
Beyond the linear regime, the power spectrum and higher order moments of the matter field no longer capture all cosmological information encoded in density fluctuations. While non-linear transforms have been proposed to extract this information lost to traditional methods, up to now, the way to generalize these techniques to discrete processes was unclear; ad hoc extensions had some success. We pointed out in Carron & Szapudi (2013) that the logarithmic transform approximates extremely well the optimal “sufficient statistics”, observables that extract all information from the (continuous) matter field. Building on these results, we generalize optimal transforms to discrete galaxy fields. We focus our calculations on the Poisson sampling of an underlying lognormal density field. We solve and test the one-point case in detail, and sketch out the sufficient observables for the multi-point case. Moreover, we present an accurate approximation to the sufficient observables in terms of the mean and spectrum of a non-linearly transformed field. We find that the corresponding optimal non-linear transformation is directly related to the maximum a posteriori Bayesian reconstruction of the underlying continuous field with a lognormal prior as put forward in Kitaura et al. (2010). Thus simple recipes for realizing the sufficient observables can be built on previously proposed algorithms that have been successfully implemented and tested in simulations.

Key words: large-scale structure of Universe, cosmology: theory, methods: statistical

1 INTRODUCTION
In the current inflationary paradigm, the small primordial density fluctuations are believed to be very close to a Gaussian field. The natural descriptors of such fields are two-point correlation functions, or (power) spectra in Fourier space. However, it was established first in $N$-body simulations (Rimes & Hamilton 2005, 2006), and subsequently with analytical calculations (Neyrinck et al. 2006) that the spectrum of the matter field loses its effectiveness as the fluctuations grow. Fourier modes of the density become strongly coupled (Meiksin & White 1999, Scoccimarro et al. 1999), resulting in large covariances effectively diminishing the available information. In this context, a variety of non-linear transformations of the field such as Gaussianization (Weinberg 1992, Nevrinck et al. 2011, Zhang et al. 2011, Yu et al. 2011), the logarithmic mapping (Nevrinck et al. 2009, Seo et al. 2011, 2013, Caron 2012), Box-Cox transformations (Joachimi et al. 2013) or clipping (Simpson et al. 2012, 2013), have been shown to increase the fidelity to linear theory and/or recapture information in the noise free fields. These results can be understood within the simple yet qualitatively and to some extent quantitatively accurate lognormal model (Coles & Jones 1991, Kayo et al. 2001) of the statistics of the matter field. It has been shown that the full set of $N$-point moments of fields of this type carry very little information in the high variance regime (see Carron & Nevrinck 2012 and references therein for an extensive discussion on these statistical issues).

Before these ideas can applied to extract more information from galaxy surveys, a key issue to be dealt with is discreteness. Galaxy fields correspond to a set of points rather than a continuous random field, and it is not entirely clear how and to what extent the methods and conclusions of these works apply. While studies such as Nevrinck et al. (2011) suggests that analogous methods should still bring some improvement, up to now the estimators relied on ad hoc generalizations of the logarithmic and similar transforms to discrete data sets.

In Carron & Szapudi (2013) we have shown rigorously that such non-linear transforms of the matter field typically work because they i) undo (some of) the non-linear evolution, thus ii) Gaussianize the distribution, and last but not least iii) they correspond to a good approximation to
sufficient statistics, observables that extract all available information from the matter field on a given cosmological parameter. The logarithmic transformation would correspond to exact sufficient statistics if the underlying continuous field were lognormal. While the evolved dark matter field is only approximately lognormal, apparently it is close enough that the amount of information extracted by the simple logarithmic transformation is virtually indistinguishable from that of the exact (and vastly more complicated) sufficient statistics. Thus a simple way presents itself to generalize these findings to the discrete galaxy fields in surveys: assume a lognormal field sampled in a Poisson fashion, and construct the corresponding sufficient statistics. Both of these assumptions are approximate, nevertheless, we expect these results to capture the essential properties of dark matter fields encoded by the distribution of galaxies. Also, it will be clear how our results generalize to more complex statistical models.

2 METHODS

We build on our previous work (Carron & Szapudi 2013) to which we refer for details, but can be summarized as follows. Let \( p(\delta) \) be the one-point PDF of the fluctuation \( \delta \). It is well known that the statistic

\[
o(\delta) = \delta^2 \tag{1}
\]

(i.e. the variance) contains the entire Fisher information \( \langle (\partial_\alpha \ln p)^2 \rangle \) of \( p(\delta) \) whenever the probability density \( p(\delta) \) is Gaussian. This is a special case of a more general identity, valid for any \( p(\delta) \)-the observable

\[
o(\delta) = c_1 \frac{\partial \ln p(\delta)}{\partial \alpha} + c_2 \tag{2}
\]

carries the entire Fisher information content of \( p(\delta) \) on the parameter \( \alpha \) (in this equation \( c_1 \) and \( c_2 \) are arbitrary constants). These statistics can be therefore be considered ‘sufficient’, or optimal, with respect to their Fisher information content. They can be read directly from the shape of the PDF.

In this paper we construct the corresponding sufficient observables for the PDF of a discrete galaxy field, given schematically as

\[
o(N) = c_1 \frac{\partial \ln P(N)}{\partial \alpha} + c_2, \tag{3}
\]

where \( P(N) \) is the probability of observing \( N = 0, 1, \ldots \) galaxies in a given cell. Initially, we restrict our analysis to one-point probabilities and one-point optimal observables, to obtain local transformations exactly analogous to previous successful methods. Our focus on the one-point PDF makes the variance of the fluctuations the only model parameter of relevance. Nevertheless, generalizations to multipoint probabilities \( P(N_1, N_2, \ldots) \) will be obvious after the detailed calculation.

We will use the local Poisson model as a natural way for discrete sampling of an underlying continuous distribution. Less obvious is the choice of the underlying PDF \( p(\delta) \). In (Carron & Szapudi 2013) we used results from perturbation theory on the moments of the dark matter \( \delta \) field to show that

\[
o(\delta) = \ln^2(1 + \delta) \tag{4}
\]

was a very good approximation to the optimal observable when the index \( n \) of the power spectrum is reasonably close to \(-1\). This result was very successfully tested against the Millennium Simulation (Springel et al. 2005) density field. Since equation (4) is the optimal statistic (2) for the lognormal PDF, this model presents itself as a plausible choice for the underlying continuous PDF for the construction of the sufficient observable.

2.1 Local Poisson model with lognormal density

Using the local Poisson model with lognormal underlying matter density, the probability \( P(N) \) of observing \( N = 0, 1, \ldots \) galaxies in a given cell is given by

\[
P(N) = \int_{-1}^{\infty} d\delta \ p(\delta) p(N|\delta) \tag{5}
\]

where

\[
p(N|\delta) = \frac{1}{N!} e^{-\bar{N}(1+\delta)} [\bar{N} (1 + \delta)]^N, \tag{6}
\]

and \( \bar{N} = \langle N \rangle \) is the mean number of galaxy in the cell. More realistic local models for discreteness including deviations from the Poissonity or biasing schemes can be implemented analogously (Kitaura et al. 2013, e.g.). The lognormal PDF of \( \delta \) reads

\[
p(\delta) = \frac{1}{\sqrt{2\pi} \sigma^2_\delta} \exp \left[ -\frac{1}{2\sigma^2_\delta} \left( \ln(1 + \delta) - \bar{A} \right)^2 \right]. \tag{7}
\]

Writing \( A = \ln (1 + \delta) \), then the parameters of the above equation are set by

\[
\bar{A} = \langle A \rangle = -\frac{1}{2} \sigma^2_\delta
\]

\[
\sigma^2_\delta = \langle (A - \bar{A})^2 \rangle = \ln (1 + \sigma^2) , \tag{8}
\]

where \( \sigma^2 = \langle \delta^2 \rangle \). The first relation ensures that \( \delta \) has zero mean. The log-density \( A \) has a Gaussian PDF.

Saddle-point approximation

We found that \( P(N) \) can be obtained with great accuracy via saddle-point approximation. Specifically, we first write Eq. (5) using \( A = \ln (1 + \delta) \) as

\[
P(N) = \int_{-\infty}^{\infty} dA \exp [g_N(A)], \tag{9}
\]

with \( g_N(A) = \ln p(A) + \ln P(N|A) \). We then approximate the integrand as

\[
g_N(A) \simeq g_N(A^*) - \frac{1}{2} (A - A^*)^2 |g''_N(A^*)| \tag{10}
\]

where \( A^* = A^*(N) \) is the point where \( g_N(A) \) is maximal. Eq. (9) becomes a Gaussian integral resulting in

\[
P(N) \simeq \sqrt{\frac{2\pi}{|g''_N(A^*)|}} \exp (g_N(A^*)). \tag{11}
\]

Collecting the terms from the lognormal and Poisson PDFs, we have in our case

\[
g_N(A) = -\frac{1}{2\sigma^2_\delta} (A - \bar{A})^2 - \frac{1}{2} \ln (2\pi\sigma^2_\delta) - \bar{N} e^A + \bar{N} A + N \ln \bar{N} - N! \tag{12}
\]
The saddle-point $A^*$ is obtained by setting $g_N'(A^*) = 0$, giving the following non-linear equation

$$e^{A^*} + \frac{A^*}{N\sigma_A^2} = \frac{N - 1/2}{N}.\quad (13)$$

Equivalently

$$\delta^* = \delta_y - \frac{\ln (1 + \delta^*)}{N\sigma_A^2} - \frac{1}{2N},\quad (14)$$

where $\delta_y = N/N - 1$. The curvature term $g_N''(A^*)$ is

$$g_N''(A^*) = -\frac{1}{\sigma_A^2} - \bar{N}e^{A^*} = -\frac{1}{\sigma_A^2} [1 + \bar{N}\sigma_A^2 (1 + \delta^*)].\quad (15)$$

Equation (13) is always well behaved with a unique solution.

### 2.2 The sufficient observable

With the above results we proceed to construct the sufficient observable. Let us start with some general observations. To get the optimal observable we need $\partial_\alpha \ln P(N)$. From Eq. (11),

$$\ln P(N) = g_N(A^*) - \frac{1}{2} \ln g''(A^*) + \frac{1}{2} \ln (2\pi).\quad (16)$$

When performing the derivatives, some care must be taken, as $\alpha$ enters in two different ways: in the density PDF $p(A)$ and in the solution to the saddle-point equation $A^*$. The first term in (16) can be dealt with simply. We note that in general $g_N(A) = \ln p(A) + \ln P(N)$. Thus, we can write

$$\frac{dg_N(A^*)}{d\alpha} = \frac{\partial g_N(A^*)}{\partial A} \frac{\partial A^*}{\partial \alpha} + \frac{\partial g_N(A^*)}{\partial \alpha} = \partial_\alpha \ln p(A^*)\quad (17)$$

The first term in the upper equation vanishes by definition of the saddle point. The second term reduces to the right hand side of the lower equation since $\ln P(N|A)$ carries no dependence on $\alpha$. Thus (17) is simply the optimal statistics of the underlying continuous density field evaluated at the point $A^*(N)$. To obtain the optimal statistics of the galaxy field we only need to add a correction from the curvature term. To write explicitly the curvature term, we need $\partial_\alpha A^*$. This is obtained taking the derivative of the saddle point equation $g_N'(A^*) = 0$, with the result

$$\left( \frac{\partial A^*}{\partial \alpha} \right) g''(A^*) = \frac{\partial \ln p(A^*)}{\partial \alpha}\quad (18)$$

We have all the ingredients to write down the optimal observable $o(N)$ for our Poisson lognormal model. Collecting the relevant terms we get after some algebra

$$o(N) = \ln^2(1 + \delta^*) - \frac{\ln (1 + \sigma^2) \bar{N} \sigma_A^2 (1 + \delta^*)}{1 + \bar{N} \sigma_A^2 (1 + \delta^*)} \left( 1 + \frac{\ln (1 + \delta^*)}{1 + \bar{N} \sigma_A^2 (1 + \delta^*)} \right).\quad (19)$$

We present the interpretation of $\delta^*$ and of the second term in the above observable later on.

### 3 TESTS AND RESULTS

To illustrate the behavior of $P(N)$ and test our statistics in the parameter space spanned by $\sigma^2$ and $\bar{N}$, we proceed as follows. Letting $R$ be the radius in $h^{-1}$Mpc of a (spherical) cell, we set the expected power law behaviors

$$\sigma^2 = \sigma^2_N \left( \frac{R}{8 \, h^{-1} \text{Mpc}} \right)^{-(n+3)}, \quad \bar{N} = \bar{N}_N \left( \frac{R}{8 \, h^{-1} \text{Mpc}} \right)^3,\quad (20)$$

where we use standard values $n = -1$, and $\sigma^2_N = 0.8$. We will use for the purposes of this paper two different values of $\bar{N}_N$. The approximate SDSS LRGs density $P_{\text{Percival et al.}}$ gives $\bar{N}_N = 0.2$, and a larger sampling rate $\bar{N}_N = 1$ for comparison.

#### Saddle-point approximation

The exact evaluation of $P(N)$ with numerical methods poses no particular difficulties. We used a basic Newton-Raphson algorithm to solve Eq. (13) for the saddle point. We show...
$P(N) \ versus \ A^*(N)$ as the crosses on the lower panel of Figure 1 for $N_8 = 1$, for different values of $\sigma$ as indicated. The solid lines show the saddle-point approximation. For convenience these lines omit the discreteness of the PDF, we continued $N$ to non-integer values by replacing the factorial function $N!$ by the Gamma function $\Gamma(N + 1)$. We found the relative deviation to be always subpercent even for large values of $\sigma$, except at the void probability $P(N = 0)$, where the accuracy worsens as $\sigma$ increases. The accuracy is similar or better for $N_8 = 0.2$. It turns out that for all the purposes of this paper, the saddle-point approximation is as good as the exact result. The lower panel shows the exact sufficient observable $\partial_\sigma \ln P(N)$ and its approximation with $o(N)$, Eq. (19). It is clear from the upper panel that the transformation from $N$ to $A^*(N)$ no longer Gaussianizes the PDF on small scales. Nevertheless, the shape of $\partial_\sigma \ln P$ remains close to a parabola on all scales. This suggests that the entire information is contained the first two moments, regardless of the non-Gaussianity of $P(A^*)$. This observation will be fully exploited next.

**Efficiencies**

We define the efficiency of the statistics as the ratio $\epsilon$ of the information of the statistics to the total information. This is shown in figure 2 for two different values of $N_8$. Each panel shows several curves. The lower solid line is the optimal statistic $o(N)$ according to our calculations, Eq. (19). For the most part, it is indistinguishable from unity, confirming our expectations. A slight deviation is observed on the smallest scales, due to the the saddle-point approximation. The dotted curves show the efficiency of the second moment $\ln^2(1 + \delta^*)$ alone, neglecting the curvature term in the optimal statistic; the latter still accounts for up to 20% of the total information in the intermediate regime. The dash-dotted lines displays the efficiency of the naive $\delta^2_{\sigma}$. This performs poorly for large variances, as expected. For comparison, we also show with dashes the statistic

$$\log^2_{+}(\delta) = \begin{cases} \ln^2(1 + \delta) &\delta > 0 \\ \delta^2 &\delta \leq 0 \end{cases} \quad (21)$$

where the $\log_{+}$ mapping was introduced in Nevrinck et al. (2013) as local transform alternative to the logarithmic mapping that is well defined when $\delta \approx -1$. While useful on large scales when $N$ is still sufficiently large, it performs poorly on smaller scales. In fact, its sensitivity to $\alpha$ vanishes around $\sigma \approx 1$.

Finally, motivated by the parabolic shape of $\partial_\sigma \ln P$ on the lower panel of figure 1 the upper solid lines present the efficiency of the first two moments (jointly) of $\ln(1 + \delta^*)$. Strikingly, these curves are indistinguishable from unity, showing perfect extraction of information. This suggests that the curvature term in $o(N)$ glean the information content (perhaps in a complex way) of the first moment that is negligible for small, but carries information for larger $\sigma$. Note that this simple procedure appears to perform slightly better that $o(N)$ for the largest variances only due to the slight inaccuracy of the saddle point integration used to realise the optimal sufficient statistic.

![Figure 2. Upper panel: the efficiency of various statistics to capture the information within the count-in-cell PDF, as a function of $\sigma$. $N$ scales according to Eq. (20), with $N_8 = 0.2$. The dotted line shows $\ln^2(1 + \delta^*)$, the dashed line $\log^2_{+}(1 + \delta)$, as defined in the text. The dash-dotted line shows the second moment of the galaxy density fluctuations $\delta^2_{\sigma}$. Two other statistics are shown as solid lines, both almost indistinguishable from unity: the optimal statistics $o(N)$ according to the saddle-point approximation, and the combined information content of the first two moments of $\ln(1 + \delta^*)$, accounting for their covariance. The former shows a very slight deviation from optimality for the largest variances. The lower panel is the same for $N_8 = 1$. In both cases, the optimal statistic is equivalent to the extraction of the first two moments of $\ln(1 + \delta^*)$.](https://example.com/figure2.png)
of the matter field needs to be evaluated at this point. We now reinterpret these results in terms of reconstructing the underlying $A$ field from the observation of $N$. In a Bayesian setting, the posterior for $A$ is

$$p(A|N) = \frac{P(N|A)p(A)}{P(N)}.$$  

By definition $A^*(N)$ maximizes $\ln P(N|A) + \ln p(A)$, thus the right hand side of the above equation with respect to $A$. Therefore we can now interpret $A^*$ the maximum a posteriori (MAP) solution in a Bayesian reconstruction of the $A$ field. The generalization of these one-dimensional considerations to the multipoint case using the full joint probability $P(N_1, \ldots, N_{N_{\text{cell}}})$ is clear. Statistics optimal with respect to $A$ are given now by

$$\frac{\partial \ln p(A_1^*, \ldots, A_{N_{\text{cell}}}^*)}{\partial \alpha} = \frac{1}{2} \frac{\partial \ln \det H}{\partial \alpha},$$

where $H_{ij}$ is the curvature matrix $-\partial_{ij}^2 \ln p(A_1, \ldots | N_1, \ldots)$ evaluated at $A^*$, the saddle-point solution for the $A$ field. These coupled equations result in general in a more complicated, non-local transformation of the galaxy field. The Bayesian reconstruction of the matter field with a Poisson lognormal assumption is the exact approach used in Kitaura et al. (2010). According to the above, their Bayesian MAP equations are always identical to our saddle-point equations. They implemented and solved these equations in three dimensions, which proves the feasibility of our approach even in the more involved multipoint case. On the other hand, their good overall success in reconstructing the density field and the direct connection we revealed between the MAP solution and sufficient statistics suggest that the logarithmic transformation of this reconstructed field does capture most information in the data.

Our initial theoretical results suggest recipes for efficient observables from galaxy fields. The simplest is the following: map in each cell $N$ to $A^*(N)$ and extract the spectrum and mean of the new field. The mapping requires only an estimate of $\sigma_\alpha^2$ and $\bar{N}$ that can be obtained from count-in-cells. There are some obvious generalizations and refinements that are left for future work. The multipoint case, including the curvature term, can be implemented with straightforward extensions of the algorithms of Kitaura et al. (2010). Further, these methods should be tested in simulations to see to what extent the conclusions from the lognormal model hold in real dark matter distributions. Refinements of the statistical models can include the use of the exact sufficient observable of the matter field and/or using a more accurate PDFs. Deviations from Poisson sampling and diverse biasing schemes can be modeled as well.

ACKNOWLEDGMENTS

We thank Mark C. Neyrinck for his careful reading of the manuscript, triggering improvements to the paper. We acknowledge NASA grants NNX12AF83G and NNX10AD53G for support.

REFERENCES

Carron J., 2012, Physical Review Letters, 108, 071301
Carron J., Neyrinck M. C., 2012, ApJ, 750, 28
Carron J., Szapudi I., 2013, MNRAS, 434, 2961
Coles P., Jones B., 1991, MNRAS, 248, 1
Joachimi B., Taylor A. N., Kiessling A., 2011, MNRAS, 418, 145
Kayo I., Taruya A., Suto Y., 2001, ApJ, 561, 22
Kitaura F.-S., Jasche J., Metcalf R. B., 2010, MNRAS, 403, 589
Kitaura F.-S., Yepes G., Prada F., 2013, ArXiv e-prints
Meiksin A., White M., 1999, MNRAS, 308, 1179
Neyrinck M. C., Szapudi I., Rimes C. D., 2006, MNRAS, 370, L66
Neyrinck M. C., Szapudi I., Szalay A. S., 2009, ApJ, 698, L90
Neyrinck M. C., Szapudi I., Szalay A. S., 2011, ApJ, 731, 116
Percival W. J., Nichol R. C., Eisenstein D. J., Frieman J. A., Fukugita M., Loveday J., Pope A. C., Schneider D. P., Szalay A. S., Tegmark M., Vogeley M. S., Weinberg D. H., Zehavi I., Bahcall N. A., Brinkmann J., Connolly A. J., Meiksin A., 2007, ApJ, 657, 645
Rimes C. D., Hamilton A. J. S., 2005, MNRAS, 360, L82
Rimes C. D., Hamilton A. J. S., 2006, MNRAS, 371, 1205
Scoccimarro R., Zaldarriaga M., Hui L., 1999, ApJ, 527, 1
Seo H.-J., Sato M., Jain B., Takada M., 2011, ApJ, 729, L11+
Seo H.-J., Sato M., Takada M., Dodelson S., 2012, ApJ, 748, 57
Simpson F., Heavens A. F., Heymans C., 2013, ArXiv e-prints
Simpson F., James J. B., Heavens A. F., Heymans C., 2011, Physical Review Letters, 107, A261301
Springel V., White S. D. M., Jenkins A., Frenk C. S., Yoshida N., Gao L., Navarro J., Thacker R., Croton D., Helly J., Peacock J. A., Cole S., Thomas P., Couchman H., Evrard A., Colberg J., Pearce F., 2005, Nature, 435, 629
Weinberg D. H., 1992, MNRAS, 254, 315
Yu Y., Zhang P., Lin W., Cui W., Fry J. N., 2011, Phys. Rev. D, 84, 023523
Zhang T.-J., Yu H.-R., Harnois-Déraps J., MacDonald I., Pen U.-L., 2011, ApJ, 728, 35