# (Draft) On estimating the structure factor of a point process, with applications to hyperuniformity

Diala Hawat[*], Guillaume Gautier, Rémi Bardenet[†]

Université de Lille, CNRS, Centrale Lille

UMR 9189 – CRIStAL, F-59000 Lille, France

and

Raphaël Lachièze-Rey

Université de Paris cité, MAP5, Paris, France

March 4, 2022

## Abstract

Hyperuniformity is the study of stationary point processes with a sub-Poisson variance of the number of points in a large window. In other words, counting the points of a hyperuniform point process that fall in a given large region yields a small-variance Monte Carlo estimator of the volume of that region. Hyperuniform point processes have received a lot of attention in statistical physics, both for the investigation of natural organized structures and the synthesis of materials. Unfortunately, rigorously proving that a point process is hyperuniform is usually difficult. A common practice in statistical physics and chemistry is to use a few point process samples to estimate a spectral measure called the structure factor. Plotting the estimated structure factor, and evaluating its decay around zero, provides a graphical diagnostic of hyperuniformity. Different applied fields use however different estimators, and important algorithmic choices proceed from each field's lore. This paper is a systematic survey and derivation of known or otherwise natural estimators of the structure factor. In an effort to make investigations of the structure factor and hyperuniformity systematic and reproducible, we further provide a `Python` toolbox `structure_factor`, containing all the estimators and tools that we discuss.

*Keywords: stochastic geometry, Monte Carlo, spectrum estimation, multitapering, isotropy.*

[*]Corresponding author, diala.hawat@univ-lille.fr

# Contents

# 1 Introduction

Condensed matter physicists have observed that, for some random particle systems, the variance of the number of points in a large window scales slower than the volume of that window, a phenomenon called *hyperuniformity* (Torquato, 2018). This statistical property in turn implies desirable physical properties for materials (Torquato, 2018, Section 14). Outside physics, hyperuniform point processes have generated broad interest in statistics, machine learning, and probability. By definition, a hyperuniform point process leads to a fast-decaying Monte Carlo error when estimating volumes. Projection determinantal point processes, for instance, have thus been proposed for Monte Carlo integration (Bardenet and Hardy, 2020; Belhadji et al., 2019, 2020a), subsampling stochastic gradients (Bardenet et al., 2021), or feature selection (Belhadji et al., 2020b). The Ginibre ensemble, arguably one of the most famous determinantal point processes arising from random matrix theory (Anderson et al., 2010a), is another typical example of hyperuniform point processes. In stochastic geometry, and beyond determinantal point processes, hyperuniform point processes appear, e.g., in the zeros of Gaussian analytic functions (Hough et al., 2009), or in matching constructions (Klatt et al., 2020). Intriguingly, there is also empirical evidence that repeatedly applying an algorithm involving local repulsion of points to an initial point process leads to a hyperuniform point process (Klatt et al., 2019). Across all these scientific fields, there is a need for statistical diagnostics of hyperuniformity.

In theory, under mild assumptions, a point process in $\mathbb{R}^d$ is hyperuniform if and only if its structure factor $S(\mathbf{k})$ decays to zero as $\|\mathbf{k}\|_2$ goes to zero (Coste, 2021). The structure factor is the Fourier transform of a measure that encodes the pairwise correlations of a point process, and the behavior in zero of the structure factor reflects long-range correlations in the original point process. Given a realization of a point process, the standard empirical diagnostic of hyperuniformity thus involves estimating and plotting the structure factor of the underlying point process (Torquato, 2018; Klatt et al., 2019). This graphical assessment is however not standardized and often not described in full reproducible detail in the literature, with implementation choices and statistical properties often part of each field's folklore. There is also no standard software to estimate structure factors. In this

paper, we thus propose ($i$) a survey of existing estimators of the structure factor and their main properties, ($ii$) an empirical benchmark of the corresponding graphical diagnostics of hyperuniformity, and ($iii$) a modular, open-source Python package[1] that implements all these estimators and the corresponding hyperuniformity diagnostics.

In Section 2, we introduce hyperuniformity and the relevant background. In Section 3, we rederive two families of estimators of the structure factor. The first family assumes that the underlying point process is stationary, while the second further requires isotropy. These estimators are all present in our companion Python toolbox `structure_factor`, which we illustrate in Section 4 on four point processes exhibiting behaviors such as repulsiveness, spatial independence, and clustering. In Section 5, we numerically compare the accuracy of the estimators. We conclude with a few research leads in Section 5.4.

**Related work.**    The closest work to our paper is the preprint of Rajala et al. (2020). They introduce important new estimators based on the idea of tapering in time series analysis (Percival and Walden, 2020), and investigate central limit theorems for their estimators. On our end, we limit ourselves to a survey – including of the estimators of Rajala et al. (2020) – and simpler properties like asymptotic unbiasedness and its relation to implicit implementation choices in statistical physics papers. One reason for this is our motivation for the study of hyperuniformity: hyperuniform point processes are unlikely to satisfy the assumptions[2] behind the central limit theorems referenced by Rajala et al. (2020). More-over, our survey includes a broader choice of estimators, including numerical quadratures of Hankel transforms, and a companion Python package. Overall, we believe that our approach and that of Rajala et al. (2020) are complementary.

**Notation.**    Throughout this paper, bold lowercase letters like $\mathbf{x}, \mathbf{r}$ indicate vectors in $\mathbb{R}^d$, and the corresponding non-bold characters are scalars. In particular, $\mathbf{x} = (x_1, \ldots, x_d)$. Whenever not confusing, we use the same letter in different fonts for a vector and its Euclidean norm, i.e., $r = \|\mathbf{r}\|_2$ and $k = \|\mathbf{k}\|_2$.

---

[1] https://github.com/For-a-few-DPPs-more/structure-factor

[2] E.g., Hypothesis (H4) of Biscio and Waagepetersen (2019), when the linear statistic is the number of points, contradicts hyperuniformity.

# 2 Point processes and their structure factor

Investigating the hyperuniformity of a point process commonly goes through the visual inspection of its *structure factor*, also known as the *scaled spectral density function*. In this section, we introduce the key definitions to understand that procedure. For a general reference on point processes, we refer to Chiu et al. (2013). For hyperuniformity and structure factors, we primarily refer to Torquato (2018) for readers with a physics background, and to Coste (2021) for readers with a mathematics background.

## 2.1 Point processes and their correlation measures

Loosely speaking, a point process is a random set of points, such that we can count the number of points falling in any observation window. More formally, let a *configuration* of $\mathbb{R}^d$ be a locally finite set $\mathcal{X} \subset \mathbb{R}^d$, i.e., for any compact $K$ of $\mathbb{R}^d$, the cardinality $\mathcal{X}(K)$ of $\mathcal{X} \cap K$ is finite. The family $\mathcal{N}$ of configurations of $\mathbb{R}^d$ is endowed by the $\sigma$-algebra generated by the mappings $\mathcal{X} \mapsto \mathcal{X}(K)$, for $K$ compact. A *point process* is a random element of $\mathcal{N}$.

In this paper, we restrict ourselves to *stationary (or homogeneous)* point processes, i.e., the law of the point process $\mathcal{X}$ is identical to that of $\mathcal{X} + \mathbf{y} \triangleq \{\mathbf{x} + \mathbf{y}; \ \mathbf{x} \in \mathcal{X}\}$, for all $\mathbf{y} \in \mathbb{R}^d$. In that case, the *intensity measure* $\rho^{(1)}$ of $\mathcal{X}$, defined by

$$\rho^{(1)}(A) = \mathbb{E}\left[\mathcal{X}(A)\right],$$

is proportional to the Lebesgue measure. We then write $\rho^{(1)}(\mathrm{d}\mathbf{x}) = \rho \, \mathrm{d}\mathbf{x}$, and $\rho \geq 0$ is called the *intensity* of $\mathcal{X}$.

Assessing pairwise correlations in a point process, and – as we shall see – hyperuniformity, usually goes through estimating the two-point correlation measure $\rho^{(2)}$ of $\mathcal{X}$. It is formally defined by

$$\mathbb{E}\left[\sum_{\mathbf{x},\mathbf{y}\in\mathcal{X}}^{\neq} f(\mathbf{x},\mathbf{y})\right] = \int_{\mathbb{R}^d \times \mathbb{R}^d} f(\mathbf{x},\mathbf{y})\rho^{(2)}(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y})$$

for any non-negative measurable bounded function $f$ with compact support. If $\mathcal{X}$ is stationary, the previous expression can be factorized in

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} f(\mathbf{x}+\mathbf{y},\mathbf{y})\rho^2 g_2(\mathrm{d}\mathbf{x})\mathrm{d}\mathbf{y}, \tag{1}$$

where $g_2$ is called the *pair correlation measure*.

Intuitively, $g_2(\mathrm{d}\mathbf{x})$ is the probability that $\mathcal{X}$ has a point at location $\mathrm{d}\mathbf{x}$, given that $\mathcal{X}$ already contains the origin. If in addition $g_2$ has a density $g$ w.r.t. the Lebesgue measure, i.e., $g_2(\mathrm{d}\mathbf{x}) = g(\mathbf{x})\mathrm{d}\mathbf{x}$, then $g$ is called the *pair correlation function* of $\mathcal{X}$. Informally, $g(\mathbf{r})$ counts the pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{X}$ such that $\mathbf{x} - \mathbf{y} = \mathbf{r} \in \mathbb{R}^d$. Finally, when $\mathcal{X}$ is assumed both stationary and *isotropic*, i.e., that the distribution of $\mathcal{X}$ is both translation- and rotation-invariant; the pair correlation function $g$ is radial, i.e., it is only a function of the norm of its argument, and we shall denote, abusively, $g(\mathbf{r}) = g(r)$, where $r = \|\mathbf{r}\|_2$.

## 2.2 The Fourier transform

The Fourier transform $\mathcal{F}$ of a square integrable function $f : \mathbb{R}^d \to \mathbb{C}$ is the square integrable function

$$\mathcal{F}(f)(\mathbf{k}) = \int_{\mathbb{R}^d} f(\mathbf{x})e^{-\mathrm{i}\langle \mathbf{k}, \mathbf{x}\rangle}\mathrm{d}\mathbf{x},$$

where $\langle \mathbf{k}, \mathbf{x}\rangle$ is the dot product of the *wavevector* $\mathbf{k}$ by $\mathbf{x}$. If $f$ is furthermore a radial function, denoted abusively by $f(\mathbf{r}) = f(r)$, where $r = \|\mathbf{r}\|_2$, then the Fourier transform of $f$ is the corresponding *symmetric Fourier transform* $\mathcal{F}_s$, namely

$$\mathcal{F}_s(f)(k) = \mathcal{F}(f)(\mathbf{k}) = (2\pi)^{d/2}\int_0^\infty r^{d/2}f(r)\frac{J_{d/2-1}(kr)}{k^{d/2-1}}\mathrm{d}r, \tag{2}$$

where $k = \|\mathbf{k}\|_2$ and $J_\nu$ is the Bessel function of the first kind of order $\nu$ (Osgood, 2014). If we further define the *Hankel transform* of order $\nu \geq -1/2$ of a function $f : \mathbb{R} \to \mathbb{C}$ as

$$\mathcal{H}_\nu(f)(k) = \int_0^\infty f(r)J_\nu(kr)r\mathrm{d}r, \quad \text{for } k \geq 0, \tag{3}$$

then (2) rewrites as

$$\mathcal{F}_s(f)(k) = \frac{(2\pi)^{d/2}}{k^{d/2-1}}\mathcal{H}_{d/2-1}(\tilde{f})(k), \quad \text{where } \tilde{f} : x \mapsto f(x)x^{d/2-1}. \tag{4}$$

**Fourier transform of the scaled-intersection volume.** Let $W \subset \mathbb{R}^d$ be a convex set, later called *observation window*, and denote $|W|$ its Lebesgue measure, i.e., its volume. The Fourier transform of the corresponding *scaled-intersection volume* function

$$\alpha_0(\mathbf{r}, W) = \frac{1}{|W|}\int_{\mathbb{R}^d} \mathbb{1}_W(\mathbf{r} + \mathbf{y})\mathbb{1}_W(\mathbf{y})d\mathbf{y}, \tag{5}$$

6

is given by

$$\mathcal{F}(\alpha_0)(\mathbf{k}, W) = \frac{1}{|W|} \left( \mathcal{F}(\mathbb{1}_W)(\mathbf{k}) \right)^2 . \tag{6}$$

In particular, if $W = B^d(\mathbf{0}, R)$ is the Euclidean ball of radius $R$, the scaled-intersection volume function (5) is radial (Torquato, 2018, Section 3.1.1). In this case, we abusively denote $\alpha_0(\mathbf{r}, W) = \alpha_0(r, W)$ and

$$\mathcal{F}_s(\alpha_0)(k, W) = 2^d \pi^{d/2} \frac{\Gamma(1 + d/2)}{k^d} J_{d/2}^2(kR), \tag{7}$$

where $\Gamma$ is Euler's Gamma function. On the other hand, if $W = \prod_{j=1}^d [-L_j/2, L_j/2]$, Equation (6) simplifies to

$$\mathcal{F}(\alpha_0)(\mathbf{k}, W) = \left( \prod_{j=1}^d \frac{\sin(k_j L_j/2)}{k_j \sqrt{L_j}/2} \right)^2 . \tag{8}$$

## 2.3   The structure factor

The *structure factor measure* $\mathcal{S}$ of a stationary point process $\mathcal{X}$ in $\mathbb{R}^d$ with intensity $\rho > 0$ is the measure on $\mathbb{R}^d$, when it exists, defined by

$$\mathcal{S} = \mathcal{L}^d + \rho \mathcal{F}[g_2 - \mathcal{L}^d] = \mathcal{F}(\delta_0 + \rho(g_2 - \mathcal{L}^d)), \tag{9}$$

where $\mathcal{L}^d$ is the $d$-dimensional Lebesgue measure and $\delta_0$ is the Dirac mass in 0. When the measure $\mathcal{S}$ is absolutely continuous w.r.t. the Lebesgue measure, i.e., $\mathcal{S}(\mathrm{d}\mathbf{k}) = S(\mathbf{k})\mathrm{d}\mathbf{k}$, we call $S$ the *structure factor*.[3] When $g_2$ is absolutely continuous w.r.t. the Lebesgue measure and $g - 1$ is integrable, $S$ can be explicitly written as

$$S(\mathbf{k}) = 1 + \rho \mathcal{F}(g - 1)(\mathbf{k}). \tag{10}$$

If $\mathcal{X}$ is further assumed to be isotropic with intensity $\rho > 0$, then both the pair correlation function $g$ and the structure factor $S$ are radial functions. Abusively denoting $S(k) = S(\mathbf{k})$, Equation (10) reads

$$S(k) = 1 + \rho \mathcal{F}_s(g - 1)(k), \tag{11}$$

---

[3]The literature is inconsistent as to whether the structure factor is the measure $\mathcal{S}$ or its density $S$. We choose the density, which is also sometimes known as the *scaled spectral density function.*

see Torquato (2018, Section 2), which can be expressed analytically using (4) as

$$S(k) = 1 + \rho \frac{(2\pi)^{d/2}}{k^{d/2-1}} \int_0^\infty (g(r) - 1) r^{d/2} J_{d/2-1}(kr) \, \mathrm{d}r. \tag{12}$$

Finally, if the pair correlation function $g$ (when it exists) is smooth, then one can expect that

$$S(\mathbf{k}) \xrightarrow[\|\mathbf{k}\|_2 \to \infty]{} 1.$$

On the other hand, the behavior of $S$ in zero measures the fluctuations of $g$ around 1 at large scales $\|\mathbf{r}\|_2 \gg 1$, which can in turn help quantify properties like hyperuniformity.

## 2.4   Hyperuniformity

A point process $\mathcal{X}$ of $\mathbb{R}^d$ is often said to be *hyperuniform* (or *super-homogeneous*) if the variance of the number of points that fall in a Euclidean ball scales slower than the volume of that ball, i.e.

$$\lim_{R \to \infty} \frac{\mathrm{Var}\left[\mathcal{X}(B(\mathbf{0}, R))\right]}{|B(\mathbf{0}, R)|} = 0, \tag{13}$$

where Var stands for the variance. Some comments are in order. Although hyperuniformity *a priori* depends on the shape of the window, e.g., a ball in (13), mild technical assumptions allow considering any growing window (Coste, 2021, Section 2). Second, being hyperuniform is not a standard feature of point processes; a homogeneous Poisson process, for instance, is not hyperuniform, as the ratio in (13) is a positive constant. Third, the most general definition of hyperuniformity goes through the structure factor of a point process (9); under the mild assumption that $g_2 - 1$ is a signed measure it is equivalent to (13), see Coste (2021, Proposition 2.2). Moreover, Torquato (2018, Section 5.3.2) states that if the structure factor undergoes a power decay $|S(\mathbf{k})| \sim c\|\mathbf{k}\|_2^\alpha$ in the neighborhood of zero, the process can be classified into three categories, depending on the value of $\alpha$, as summarized in Table 1.

Since the variance cannot grow more slowly than $R^{d-1}$ for a spherical window (Beck, 1987), Class I represents the *strongest* form of hyperuniformity. It includes, for instance, the Ginibre point process, as we shall see in Section 2.1. Class II hyperuniform processes include, for instance, the sine process, a central object in random matrix theory (Anderson

| Class | $\alpha$ | $\mathrm{Var}\,[\mathcal{X}(B(0,R))]$ |
|-------|----------|---------------------------------------|
| I | $> 1$ | $\mathcal{O}(R^{d-1})$ |
| II | $= 1$ | $\mathcal{O}(R^{d-1}\log(R))$ |
| III | $\in\,]0,1[$ | $\mathcal{O}(R^{d-\alpha})$ |

Table 1: Classes of hyperuniformity for point processes satisfying $|S(\mathbf{k})| \underset{\mathbf{k}\to 0}{\sim} c\,\|\mathbf{k}\|_2^\alpha$.

et al., 2010b). By contrast, systems that fall in Class III present the *weakest* form of hyperuniformity. To the best of our knowledge, no point process has been rigorously proved to be Class III yet, although there are many candidates (Torquato, 2018, Table 1).

## 2.5 Empirical diagnostics of hyperuniformity

As suggested in Section 2.4, the behavior in zero of the structure factor quantifies hyperuniformity. Investigating whether one or several samples come from a hyperuniform point process is thus often carried out by estimating the structure factor, and then either visually inspecting the resulting plots around zero, or regressing the estimated values. Say for a stationary and isotropic point process, one option is to regress $\log S$ onto $\log k$ around zero, to assert a potential value for the decay rate $\alpha$ in Table 1.

Another criterion of *effective* hyperuniformity has been proposed; see Klatt et al. (2019, Supplementary material) and Torquato (2018, Section 11.1.6). For a stationary and isotropic hyperuniform point process, given a set of estimated values

$$\{(k_1, \widehat{S}(k_1)), \ldots, (k_n, \widehat{S}(k_n))\}$$

with $k_1, \ldots, k_n > 0$, the $H$-index is defined by

$$H = \frac{\widehat{S}(0)}{\widehat{S}(k_{\mathrm{peak}})}, \tag{14}$$

where $\widehat{S}(0)$ is a linear extrapolation of the structure factor in $k = 0$ based on the estimated values of $S$, and $k_{\mathrm{peak}}$ is the location of the first dominant peak value of the estimated

structure factor, defined here as

$$S(k_{\text{peak}}) = 1 \wedge S\left(\inf\left\{k_i \text{ such that } \widehat{S}(k_i) > 1, \widehat{S}(k_{i-1}) < \widehat{S}(k_i), \text{ and } \widehat{S}(k_{i+1}) < \widehat{S}(k_i)\right\}\right).$$

When $H < 10^{-3}$, the process is called *effectively hyperuniform* by Torquato (2018, Section 11.1.6). Note that the linear extrapolation is chosen for simplicity and not based on model selection. Like the threshold of $10^{-3}$, the definition of a dominant peak is also somewhat arbitrary.

Finally, both an estimate of the decay rate and the $H$-index (14) require estimators of the structure factor. Before we survey these estimators, we introduce a few benchmark point processes and discuss their structure factors.

## 2.6   Some point processes and their structure factor



(a) KLY process        (b) Ginibre ensemble        (c) Poisson process        (d) Thomas process
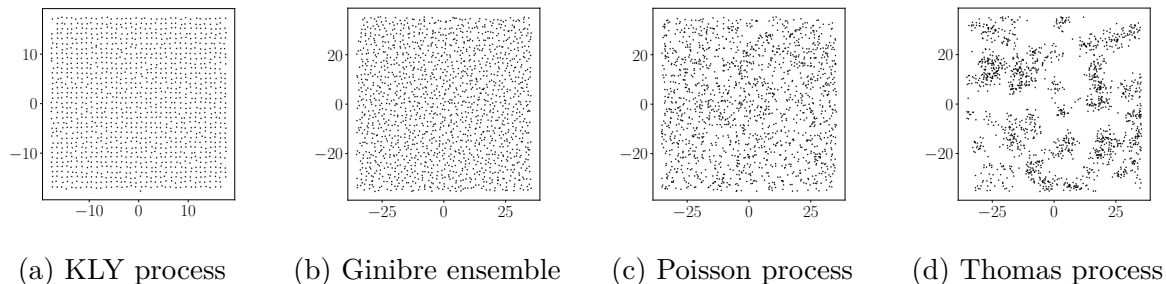
Figure 1: Samples of benchmark point processes.

### 2.6.1   The KLY process

The first point process is the result of a matching algorithm proposed by Klatt, Last, and Yogeshwaran (2020). It can be seen as a random perturbation of the square lattice. More specifically, each point of $\mathbb{Z}^d$ is matched with a close-by point of a user-defined point process in $\mathbb{R}^2$, like a homogeneous Poisson point process. The KLY process is an example of a point process that is known to be stationary, ergodic, and hyperuniform (but not isotropic). However the corresponding pair correlation function, structure factor, and hyperuniformity class are unknown. Klatt et al. (2020) use hyperuniformity diagnostics (Section 2.5) to assess the degree of hyperuniformity, and we shall reproduce their experiment with our

software. Figure 1a shows a sample from the KLY process generated by matching a subset of $\mathbb{Z}^2$ with a realization of a Poisson point process with intensity $\rho = 11$. The intensity of the resulting point process is equal to 1. The same method is applied to generate the samples from the KLY process used in Section 4.

### 2.6.2 The Ginibre ensemble

Our second hyperuniform point process is the Ginibre ensemble in $\mathbb{C} \approx \mathbb{R}^2$, both stationary and isotropic. The Ginibre ensemble can be defined (and approximately sampled) as the limit of the set of eigenvalues of matrices filled with i.i.d. standard complex Gaussian entries, as the size of the matrix goes to infinity (Hough et al., 2009, Theorem 4.3.10). A sample is shown in Figure 1b as well.

Its intensity is equal to $\rho_{\text{Ginibre}} = 1/\pi$, and its pair correlation function (Figure 2a) is

$$g_{\text{Ginibre}}(r) = 1 - \exp(-r^2). \tag{15}$$

The structure factor (Figure 2b) can be computed exactly,

$$S_{\text{Ginibre}}(k) = 1 - \exp(-k^2/4). \tag{16}$$

In particular, we have,

$$S_{\text{Ginibre}}(k) \sim k^2, \quad \text{for } k \to 0. \tag{17}$$

Thus, according to Table 1, the Ginibre ensemble is a hyperuniform point process of Class I, with $\alpha = 2$. Much is known about the Ginibre ensemble, and we use it to benchmark our toolbox and its different estimators, rather than to infer new results about it.

### 2.6.3 The homogeneous Poisson point process

The homogeneous Poisson point process on $\mathbb{R}^d$ is stationary and isotropic. It is often thought of as having as little structure as possible and can be defined as the limit of $N$ uniform i.i.d. points in a window of volume $N/\rho$, as $N$ goes to infinity. In particular, for any collection of mutually disjoint measurable subsets of $\mathbb{R}^d$, the number of points that fall in these subsets are independent; Figure 1c displays a realization from a Poisson process

11

with the same intensity $\rho_{\text{Poisson}} = 1/\pi$ as the Ginibre ensemble. In the same vein, the pair correlation function and the structure factor of the Poisson point process are both equal to one; see Figures 2a and 2b. The Poisson process is not hyperuniform.

### 2.6.4 The Thomas point process

The Thomas point process is a point process in $\mathbb{R}^d$ that is reminiscent of a mixture of Gaussians and typically exhibits clusters (Chiu et al., 2013, Section 5.3). Formally, consider a homogeneous Poisson point process $\mathcal{X}_{\text{parent}}$ of intensity $\rho_{\text{parent}}$, and let $\lambda, \sigma$ be positive. Conditionally to $\mathcal{X}_{\text{parent}}$, let $(N_{\mathbf{x}})_{\mathbf{x} \in \mathcal{X}_{\text{parent}}}$ be i.i.d. Poisson variables with mean $\lambda$. For any $\mathbf{x} \in \mathcal{X}_{\text{parent}}$, conditionally to $N_{\mathbf{x}}$, let $(Y_{\mathbf{x},i})_{i=1}^{N_{\mathbf{x}}}$ be i.i.d. centered two-dimensional isotropic Gaussian vectors with variance $\sigma$. The resulting Thomas process is given by

$$\mathcal{X}_{\text{Thomas}} = \cup_{\mathbf{x} \in \mathcal{X}_{\text{parent}}} \{\mathbf{x} + Y_{\mathbf{x},i}, i = 1, \ldots, N_{\mathbf{x}}\},$$

and its intensity is given by $\rho_{\text{Thomas}} = \rho_{\text{parent}} \times \lambda$. Figure 1d shows a realization of a Thomas point process of intensity $1/\pi$, generated from a parent Poisson point process of intensity $\rho_{\text{parent}} = 1/(20\pi)$, and a standard deviation $\sigma = 2$. Since $\sigma$ is small enough compared to $\lambda = 20$, clusters naturally appear. The pair correlation function and the structure factor of the Thomas process, see Figures 2a and 2b, are both radial functions (Chiu et al., 2013, Equation 5.52), given by

$$g_{\text{Thomas}}(r) = 1 + \frac{1}{\rho_{\text{parent}} \left(\sqrt{4\pi\sigma^2}\right)^d} \exp\left(-\frac{r^2}{4\sigma^2}\right), \tag{18}$$

and

$$S_{\text{Thomas}}(k) = 1 + \lambda \exp\left(-k^2\sigma^2\right). \tag{19}$$

The intuition that the Thomas point process is not hyperuniform is confirmed by the expression of the structure factor (19) which is even larger than 1.

## 3 Estimators of the structure factor

Starting from the theoretical definition of the structure factor (10), we derive all known estimators and add a few other natural candidates based on numerical quadratures of the

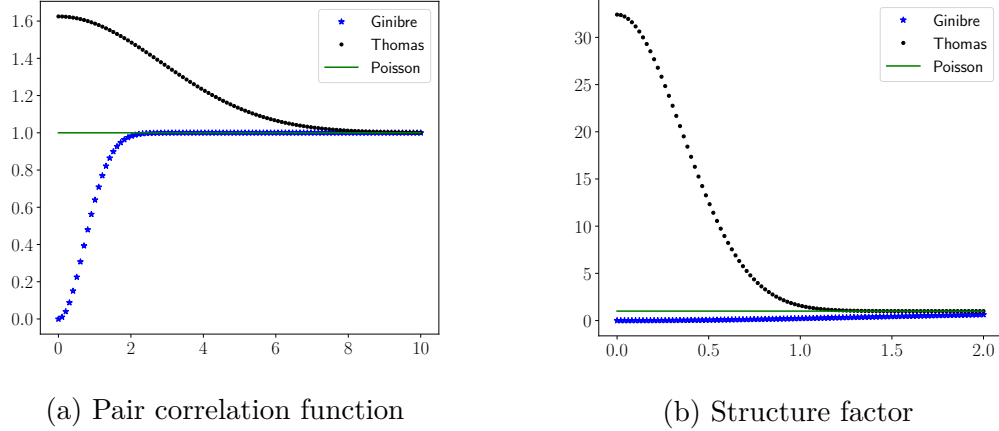(a) Pair correlation function      (b) Structure factor

Figure 2: Pair correlation functions and structure factors of some point processes.

symmetric Fourier transform. We pay particular attention to the sources of asymptotic (in the size of the window) bias for each estimator. We split the estimators according to whether they only assume stationarity or stationarity *and* isotropy.

## 3.1 Estimators assuming stationarity

The most common estimator of the structure factor is the so-called *scattering intensity* (Torquato, 2018; Klatt et al., 2019, 2020; Coste, 2021). Its name stems from its origins in the physics of diffraction, and it corresponds to a scaled version of Bartlett's periodogram in time series analysis. After introducing the scattering intensity, we follow Rajala et al. (2020) and generalize it to so-called *tapered* estimators. We group our assumptions and notation in Assumption A-1.

**Assumption A-1.** $\mathcal{X}$ *is a stationary point process of* $\mathbb{R}^d$ *with intensity* $\rho > 0$. *Its pair correlation function* $g$ *exists, and* $\mathbf{r} \mapsto g(\mathbf{r}) - 1$ *is integrable on* $\mathbb{R}^d$. *Moreover, we only observe a realization of* $\mathcal{X} \cap W = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ *of* $\mathcal{X}$ *in a centered, rectangular window* $W = \prod_{j=1}^{d} [-L_j/2, L_j/2]$. *We denote* $\mathbf{L} = (L_1, \ldots, L_d)$.

### 3.1.1 The scattering intensity

The following derivation is often assumed to be known to readers in the physics literature, and it seemed worthwhile to us to make it explicit, in particular, to understand the role

played by the so-called *allowed wavevectors.* Note that Rajala et al. (2020) provide similar derivations.

The basic idea is to introduce the scaled-intersection volume $\alpha_0$ (5) in the definition (10) of the structure factor. We obtain

$$S(\mathbf{k}) = 1 + \rho \int_{\mathbb{R}^d} (g(\mathbf{r}) - 1) e^{-i\langle \mathbf{k}, \mathbf{r}\rangle} \, d\mathbf{r} \tag{20}$$

$$= 1 + \rho \int_{\mathbb{R}^d} \lim_{\min_j L_j \to \infty} (g(\mathbf{r}) - 1) \alpha_0(\mathbf{r}, W) e^{-i\langle \mathbf{k}, \mathbf{r}\rangle} \, d\mathbf{r} \tag{21}$$

$$= 1 + \lim_{\min_j L_j \to \infty} \rho \int_{\mathbb{R}^d} (g(\mathbf{r}) - 1) \alpha_0(\mathbf{r}, W) e^{-i\langle \mathbf{k}, \mathbf{r}\rangle} \, d\mathbf{r} \tag{22}$$

where we used dominated convergence and the limit $\lim_{\min_j L_j \to \infty} \alpha_0(\mathbf{r}, W) = 1$. Splitting the integral from (22) into two, we shall recognize an expectation under our censored point process $\mathcal{X} \cap W$ and a bias term,

$$S(\mathbf{k}) = 1 + \lim_{\min_j L_j \to \infty} \left[ \frac{\rho}{|W|} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \underbrace{e^{-i\langle \mathbf{k}, \mathbf{r}\rangle} \mathbb{1}_W(\mathbf{r} + \mathbf{y}) \mathbb{1}_W(\mathbf{y})}_{f(\mathbf{r}+\mathbf{y}, \mathbf{y})} g(\mathbf{r}) \, d\mathbf{y} \, d\mathbf{r} - \underbrace{\rho \mathcal{F}(\alpha_0)(\mathbf{k}, W)}_{\triangleq \epsilon_0(\mathbf{k}, \mathbf{L})} \right].$$

Now, by definition (1) of the pair correlation measure, and still for any $\mathbf{k} \in \mathbb{R}^d$,

$$S(\mathbf{k}) = 1 + \lim_{\min_j L_j \to \infty} \left[ \frac{\rho}{\rho^2 |W|} \mathbb{E}\left[ \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}}^{\mathbf{x} \neq \mathbf{y}} \mathbb{1}_W(\mathbf{x}) \mathbb{1}_W(\mathbf{y}) e^{-i\langle \mathbf{k}, \mathbf{x} - \mathbf{y}\rangle} \right] - \epsilon_0(\mathbf{k}, \mathbf{L}) \right] \tag{23}$$

$$= \lim_{\min_j L_j \to \infty} \left[ \mathbb{E}\left[ \frac{1}{\rho |W|} \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X} \cap W} e^{-i\langle \mathbf{k}, \mathbf{x} - \mathbf{y}\rangle} \right] - \epsilon_0(\mathbf{k}, \mathbf{L}) \right] \tag{24}$$

$$= \lim_{\min_j L_j \to \infty} \left[ \mathbb{E}\left[ \frac{1}{\rho |W|} \left| \sum_{\mathbf{x} \in \mathcal{X} \cap W} e^{-i\langle \mathbf{k}, \mathbf{x}\rangle} \right|^2 \right] - \epsilon_0(\mathbf{k}, \mathbf{L}) \right]. \tag{25}$$

Note that the bias term satisfies

$$|\epsilon_0(\mathbf{k}, \mathbf{L})| = \rho \frac{\mathcal{F}^2(\mathbb{1}_W)(\mathbf{k})}{|W|} = \rho \left( \prod_{j=1}^d \frac{\sin(k_j L_j / 2)}{\sqrt{L_j} k_j / 2} \right)^2$$

$$\leq \begin{cases} 0 & \text{if } \mathbf{k} \in \mathbb{A}_{\mathbf{L}}, \\ \rho \prod_{j=1}^d L_j & \text{as } \|\mathbf{k}\|_2 \to 0, \\ 2^{2d} \prod_{j=1}^d \frac{1}{L_j k_j^2} & \text{as } \|\mathbf{k}\|_2 \to \infty, \end{cases}$$

where we defined

$$\mathbb{A}_{\mathbf{L}} = \left\{ \mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{R}^d \setminus \{\mathbf{0}\} \; ; \; \exists j \in \{1, \ldots, d\}, n \in \mathbb{Z}^*, \text{ s.t. } k_j = \frac{2\pi n}{L_j} \right\}. \tag{26}$$

We have thus proved the following proposition.

14

**Proposition 3.1.** *Under Assumption A-1, for* $\mathbf{k} \in \mathbb{A}_\mathbf{L}$*, the scattering intensity estimator*

$$\widehat{S}_{\mathrm{SI}}(\mathbf{k}) \triangleq \frac{1}{\rho|W|} \left| \sum_{j=1}^{N} e^{-\mathrm{i}\langle \mathbf{k}, \mathbf{x}_j \rangle} \right|^2, \tag{27}$$

*is asymptotically unbiased, i.e.*

$$\sup_{\mathbf{k} \in \mathbb{A}_\mathbf{L}} \left| \mathbb{E}[\widehat{S}_{\mathrm{SI}}(\mathbf{k})] - S(\mathbf{k}) \right| \xrightarrow[\min_j L_j \to \infty]{} 0.$$

This motivates restricting to $\mathbf{k} \in \mathbb{A}_\mathbf{L}$ if one is interested in estimating the behavior of $S$ in zero. In the literature, the scattering intensity is actually often evaluated on a subset of $\mathbb{A}_\mathbf{L}$, namely

$$\mathbb{A}_\mathbf{L}^{\mathrm{res}} = \left\{ \left( \frac{2\pi n_1}{L_1}, \ldots, \frac{2\pi n_d}{L_d} \right) \text{ with } \mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{Z}^d \setminus \{\mathbf{0}\} \right\}. \tag{28}$$

The set $\mathbb{A}_\mathbf{L}^{\mathrm{res}}$ is called the set of *allowed wavevectors* in physics (Klatt et al., 2020, Section 10), or part of the *dual lattice* of fundamental cell $W$ in sampling theory (Osgood, 2014, Chapter 5), or *Fourier grid* in time series analysis (Rajala et al., 2020). We add the superscript res to underline that it is actually a restriction of the set $\mathbb{A}_\mathbf{L}$ of wavevectors on which the scattering intensity is asymptotically unbiased. It is unclear to us why one should consider $\mathbb{A}_\mathbf{L}^{\mathrm{res}}$ instead of $\mathbb{A}_\mathbf{L}$.

The minimal *wavenumber* (norm of a wavevector) over $\mathbb{A}_\mathbf{L}^{\mathrm{res}}$ is

$$k_{\min}^{\mathrm{res}} = 2\pi \sqrt{\sum_{j=1}^{d} L_j^{-2}},$$

while working with $\mathbb{A}_\mathbf{L}$ (26) relaxes this threshold to a $k_{\min}$ satisfying

$$\frac{2\pi}{\max_j L_j} < k_{\min} < k_{\min}^{\mathrm{res}}.$$

We give evidence below as to why the scattering intensity should never be evaluated for $\|\mathbf{k}\|_2 \leqslant \frac{\pi}{\sqrt{d}\max_j L_j}$.

Finally, when $\rho$ is unknown, and $\mathcal{X}$ is further assumed to be ergodic, it is common to replace the denominator $\rho|W|$ by $N$ in (27), leading to the self-normalized scattering intensity estimator

$$\widehat{S}_{\mathrm{SI,s}}(\mathbf{k}) \triangleq \frac{1}{N} \left| \sum_{j=1}^{N} e^{-\mathrm{i}\langle \mathbf{k}, \mathbf{x}_j \rangle} \right|^2. \tag{29}$$

15

Indeed, by ergodicity,

$$\frac{\widehat{S}_{\text{SI,s}}}{\widehat{S}_{\text{SI}}} \xrightarrow[\min_j L_j \to \infty]{} 1, \quad \text{a.s..}$$

Many authors define the scattering intensity as its self-normalized version (Torquato, 2018; Klatt et al., 2019; Coste, 2021), but we argue that when $\rho$ is known, it is not clear whether this leads to an estimator with a smaller mean squared error than (27).

**A lower bound for $k_{\min}$.** It is clear that when $W$ is fixed, the scattering intensity (29) is not relevant when $\mathbf{k}$ is too close to 0, since for a fixed sample $\mathcal{X}_N = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$,

$$\lim_{\mathbf{k} \to 0} \widehat{S}_{\text{SI,s}}(\mathbf{k}) = N.$$

The theoretical convergence to 0 for hyperuniform processes is dictated by the compensation occurring between exponential terms $e^{-i\langle \mathbf{k}, \mathbf{x}_j \rangle}$ on large portions of the space. One can safely infer that if a large portion of the terms gives a positive contribution, then the compensation does not occur and the global result will not be accurate. Still working with the window $W = \prod_{j=1}^{d} [-L_j/2, L_j/2]$ from Assumption A-1, if $\|\mathbf{k}\|_2 < \frac{\pi}{\sqrt{d} \max_j L_j}$, then for $\mathbf{x} \in W$,

$$|\langle \mathbf{x}, \mathbf{k}_j \rangle| \leqslant k_{\min} \|\mathbf{x}\|_2 \leqslant \frac{\pi}{2}.$$

So there exists $\epsilon > 0$ s.t.,

$$\frac{1}{N} \left| \sum_{j=1}^{N} e^{-i\langle \mathbf{k}, \mathbf{x}_j \rangle} \right|^2 \geq \frac{(N\epsilon)^2}{N}.$$

Thus for any $\|\mathbf{k}\|_2 < \frac{\pi}{\sqrt{d} \max_j L_j}$ we have a lower bound of $\widehat{S}_{\text{SI,s}}(\mathbf{k})$ independent of the point process and which diverges as the number of points goes to infinity. Consequently, $\pi/\sqrt{d} \max_j L_j$ seems like a reasonable lower bound for the wavenumbers, which one might improve with a finer study of the estimator bias.

### 3.1.2 Tapered variants of the scattering intensity

The derivation of the scattering intensity can be generalized to the *tapered estimator*

$$\widehat{S}_{\text{T}}(t, \mathbf{k}) \triangleq \frac{1}{\rho} \left| \sum_{j=1}^{N} t(\mathbf{x}_j, W) e^{-i\langle \mathbf{k}, \mathbf{x}_j \rangle} \right|^2, \quad \mathbf{k} \in \mathbb{R}^d, \tag{30}$$

where $t(\cdot, W)$ is a uniformly (in $W$) square-integrable function supported on the observation window $W$, called a *taper*, subject to the renormalization condition (33). The tapered estimator (30) corresponds to a scaled version of what is called a tapered periodogram in the signal processing literature (Rajala et al., 2020, Section 3). The vocabulary is adapted from the spectral analysis of time series, where tapers are now well established (Percival and Walden, 2020).

In particular, one recovers the scattering intensity (29) from the tapered formulation (30) by plugging the taper

$$t_0(x, W) \triangleq \frac{1}{\sqrt{|W|}} \mathbb{1}_W(x).$$ (31)

To follow the derivation of (25), we further let

$$\alpha_t(\mathbf{r}, W) \triangleq \int_{\mathbb{R}^d} t(\mathbf{r} + \mathbf{y}, W) t(\mathbf{y}, W) \mathrm{d}\mathbf{y},$$ (32)

and require that

$$\lim_{W \to \mathbb{R}^d} \alpha_t(\mathbf{r}, W) = 1,$$ (33)

where the limit is taken so that the window progressively covers the whole space at roughly equal speed in all directions, i.e., $\min_j L_j \to \infty$ under Assumption A-1. Note first that our requirement that (33) holds differs from the treatment of Rajala et al. (2020). We find (33) to be a more natural generalization of the scattering intensity arguments of Torquato (2018). Note also that, for simplicity, we denoted $\alpha_{t_0}$ by $\alpha_0$ in Section 2.2, and we shall stick to this simplified notation.

Now, Cauchy-Schwarz and the uniform integrability of $t$ guarantee that $\alpha_t(\cdot, W)$ is uniformly (in $W$) bounded, so that by dominated convergence,

$$S(\mathbf{k}) = 1 + \rho \lim_{\min_j L_j \to \infty} \int_{\mathbb{R}^d} (g(\mathbf{r}) - 1) \alpha_t(\mathbf{r}, W) e^{-\mathrm{i}\langle \mathbf{k}, \mathbf{r} \rangle} \, \mathrm{d}\mathbf{r}.$$

Following the lines of (25), we obtain

$$S(\mathbf{k}) = \lim_{\min_j L_j \to \infty} \left( \mathbb{E}\left[ \widehat{S}_\mathrm{T}(t, \mathbf{k}) \right] - \underbrace{\rho \left| \mathcal{F}(t)(\mathbf{k}, W) \right|^2}_{\triangleq \epsilon_t(\mathbf{k}, \mathbf{L})} \right).$$ (34)

where the tapered estimator $\widehat{S}_\mathrm{T}$ is defined in (30). To eliminate the asymptotic bias $\epsilon_t(\mathbf{k}, \mathbf{L})$, one can restrict oneself again to a set of allowed wavevectors as we did in (26), i.e., the

17

zeros of $\epsilon_t(\cdot, \mathbf{L})$. For general tapers, however, finding these zeros is not straightforward, and an alternative way to escape the bias is to correct it, as in

$$\widehat{S}_{\mathrm{UDT}}(t, \mathbf{k}) \triangleq \frac{1}{\rho} \left| \sum_{j=1}^{N} t(\mathbf{x}_j, W) e^{-\mathrm{i}\langle \mathbf{k}, \mathbf{x}_j \rangle} \right|^2 - \rho \left| \mathcal{F}(t)(\mathbf{k}, W) \right|^2. \tag{35}$$

We refer to $\widehat{S}_{\mathrm{UDT}}$ as the *undirectly debiased tapered estimator*, which is a scaled version of what Rajala et al. (2020) define. The major issue of Estimator (35) is that it may give negative values. To remedy this, Rajala et al. (2020) propose to remove the bias inside the summation before taking the squared modulus, namely, to define the *directly debiased tapered estimator*

$$\widehat{S}_{\mathrm{DDT}}(t, \mathbf{k}) \triangleq \frac{1}{\rho} \left| \sum_{j=1}^{N} t(\mathbf{x}_j, W) e^{-\mathrm{i}\langle \mathbf{k}, \mathbf{x}_j \rangle} - \rho \mathcal{F}(t)(\mathbf{k}, W) \right|^2. \tag{36}$$

Finally, note that these debiasing techniques naturally apply to the special case of the scattering intensity, and thus offer an alternative to allowed values (26).

### 3.1.3   The multitapered estimator

In the spectral analysis of time series, *multitapering* (MT) was first introduced by Thomson (1982); see also Percival and Walden (2020) for a modern reference. The idea is to average a periodogram over many tapers, in the hope to reduce the variance of the resulting estimator. Rajala et al. (2020) propose to adapt the method to point processes, and we follow their lines.

For any $\mathbf{k} \in \mathbb{R}^d$ and $P \in \mathbb{N}^*$, and under Assumption A-1, Rajala et al. (2020) define the *multitapered estimator* $\widehat{S}_{\mathrm{MT}}$ by

$$\widehat{S}_{\mathrm{MT}}((t_q)_{q=1}^P, \mathbf{k}) = \frac{1}{P} \sum_{q=1}^{P} \widehat{S}(t_q, \mathbf{k}), \tag{37}$$

where the $P$ tapers $(t_q)_{q=1}^P$ are typically taken to be pairwise orthogonal, and $\widehat{S}(t_q, \cdot)$ is any of the tapered estimators, whether undebiased (30), undirectly debiased (35), or directly debiased (36). The directly and undirectly debiased versions of $\widehat{S}_{\mathrm{MT}}$ will be respectively denoted by $\widehat{S}_{\mathrm{DDMT}}$ and $\widehat{S}_{\mathrm{UDMT}}$.

### 3.1.4 On the choice of tapers

Common taper choices in time series analysis are Slepian tapers, sinusoidal tapers, and minimum bias tapers (Riedel and Sidorenko, 1995). For instance, still assuming a centered rectangular window $W = \prod_{j=1}^{d}[-L_j/2, L_j/2]$, the family of sinusoidal tapers $(t_q)_{q \geq 1}$ supported on $W$ is defined by

$$t_q(\mathbf{x}, W) = t(\mathbf{x}, \mathbf{p}^q, W) \triangleq \frac{\mathbb{1}_W(\mathbf{x})}{\sqrt{|W|}} \prod_{j=1}^{d} \sqrt{2} \sin\left(\frac{\pi p_j^q}{L_j}(x_j + L_j/2)\right), \tag{38}$$

where $\mathbf{p}^q = (p_1^q, \ldots, p_d^q) \in (\mathbb{N}^d)^*$ and $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The sinusoidal tapers are pairwise orthogonal and satisfy (33). For $\mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{R}^d$ the Fourier transform of $t_q$ for any $q$ is defined by,

$$\mathcal{F}(t_q)(\mathbf{k}, W) = \frac{1}{\sqrt{|W|}} \prod_{j=1}^{d} \sqrt{2} i^{(p_j^q+1)} \left[ \frac{\sin\left(\left(k_j - \frac{\pi p_j^q}{L_j}\right)\frac{L_j}{2}\right)}{k_j - \frac{\pi p_j^q}{L_j}} - (-1)^{p_j^q} \frac{\sin\left(\left(k_j + \frac{\pi p_j^q}{L_j}\right)\frac{L_j}{2}\right)}{k_j + \frac{\pi p_j^q}{L_j}} \right],$$

and can thus be used in any debiasing scheme.

## 3.2 Estimators assuming stationarity and isotropy

For isotropic point processes, a common approach is simply to numerically rotation-average the structure factor estimators presented in Section 3.1. Alternatively, one could start from the analytical expression (11) of the structure factor as symmetric Fourier transform – a univariate integral – involving the pair correlation function. Then again, two approaches have been identified. First, identifying an expectation under the point process as in Section 3.1 leads to a natural estimator originally derived by Bartlett (1964). Second, estimation of the pair correlation followed by numerical quadrature leads to two natural estimators, depending on whether the quadrature is that of Ogata (2005), or that of Baddour and Chouinard (2015). We review all these estimators in turn. All point processes in Section 3.2 satisfy Assumption A-2.

**Assumption A-2.** *$\mathcal{X}$ is a stationary isotropic point process of $\mathbb{R}^d$, of intensity $\rho$. Its pair correlation function $g$ exists, and $r \mapsto g(r) - 1$ is integrable. Moreover, we only observe a realization $\mathcal{X} \cap W = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of $\mathcal{X}$ in the centered ball $W = B^d(\mathbf{0}, R)$.*

### 3.2.1 Bartlett's isotropic estimator

From the observation that the scaled-intersection volume $\alpha_0$ (5) is a radial function, $\alpha_0(\mathbf{r}, W) = \alpha_0(r, W)$, and following the lines of Section 3.1, dominated convergence yields

$$S(k) = 1 + \lim_{R \to \infty} \left[ \rho \frac{(2\pi)^{d/2}}{k^{d/2-1}} \int_0^\infty \alpha_0(r, W) \, g(r) r^{d/2} J_{d/2-1}(kr) \, \mathrm{d}r - \underbrace{\rho \mathcal{F}(\alpha_0)(k, W)}_{\triangleq \epsilon_1(k,R)} \right] \quad (39)$$

Now, precisely because $\alpha_0$ is radial, we have

$$\alpha_0(r, W) = \frac{1}{\omega_{d-1}} \int_{S^{d-1}} \alpha_0(r\mathbf{u}, W) \mathrm{d}\mathbf{u}, \quad (40)$$

where $\mathrm{d}\mathbf{u}$ is the $(d-1)$-dimensional Hausdorff measure and $S^{d-1}$ is the unit sphere of $\mathbb{R}^d$, with surface area $\omega_{d-1}$. Plugging (40) into (39) yields

$S(k) - 1$

$$= \lim_{R \to \infty} \frac{\rho(2\pi)^{d/2} k^{1-d/2}}{|W|\omega_{d-1}} \int_0^\infty \int_{S^{d-1}} \int_{\mathbb{R}^d} \mathbb{1}_W(r\mathbf{u}+\mathbf{y})\mathbb{1}_W(\mathbf{y})g(r)r^{d/2}J_{d/2-1}(kr) \, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{u}\mathrm{d}r - \epsilon_1(k, R)$$

$$= \lim_{R \to \infty} \frac{\rho(2\pi)^{d/2}}{|W|\omega_{d-1}} \int_0^\infty \int_{S^{d-1}} \int_{\mathbb{R}^d} \mathbb{1}_W(r\mathbf{u}+\mathbf{y})\mathbb{1}_W(\mathbf{y})g(r)\frac{J_{d/2-1}(kr)}{(kr)^{d/2-1}} \, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{u} \, r^{d-1}\mathrm{d}r - \epsilon_1(k, R)$$

$$= \lim_{R \to \infty} \frac{\rho(2\pi)^{d/2}}{|W|\omega_{d-1}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}_W(\mathbf{r}+\mathbf{y})\mathbb{1}_W(\mathbf{y}) \frac{J_{d/2-1}(k\|\mathbf{r}\|_2)}{(k\|\mathbf{r}\|_2)^{d/2-1}}g(\|\mathbf{r}\|_2) \, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{r} - \epsilon_1(k, R).$$

We now recognize an expectation using (1), namely

$$S(k) = 1 + \lim_{R \to \infty} \frac{(2\pi)^{d/2}}{\rho|W|\omega_{d-1}} \mathbb{E}\left[ \sum_{\mathbf{x},\mathbf{y} \in \mathcal{X} \cap W}^{\neq} \frac{J_{d/2-1}(k\|\mathbf{x}-\mathbf{y}\|_2)}{(k\|\mathbf{x}-\mathbf{y}\|_2)^{d/2-1}} \right] - \epsilon_1(k, R). \quad (41)$$

We thus define a new estimator

$$\widehat{S}_{\mathrm{BI}}(k) = 1 + \frac{(2\pi)^{d/2}}{\rho|W|\omega_{d-1}} \sum_{\substack{i,j=1 \\ i \neq j}}^N \frac{J_{d/2-1}(k\|\mathbf{x}_i-\mathbf{x}_j\|_2)}{(k\|\mathbf{x}_i-\mathbf{x}_j\|_2)^{d/2-1}}, \quad (42)$$

along with its self-normalized version

$$\widehat{S}_{\mathrm{BI,s}}(k) = 1 + \frac{(2\pi)^{d/2}}{N\omega_{d-1}} \sum_{\substack{i,j=1 \\ i \neq j}}^N \frac{J_{d/2-1}(k\|\mathbf{x}_i-\mathbf{x}_j\|_2)}{(k\|\mathbf{x}_i-\mathbf{x}_j\|_2)^{d/2-1}}, \quad (43)$$

as in the case of the self-normalized scattering intensity (29). When $d = 2$, $\widehat{S}_{\mathrm{BI,s}}$ corresponds to Bartlett's isotropic estimator (Bartlett, 1964).

20

Here also, there are two sources of bias in the estimator (42). The first one is due to the restriction of the point process to a bounded observation window, which shall disappear as the window grows. The second source of bias $\epsilon_1(k, R)$ is again related to the Fourier transform of the scaled-intersection volume $\alpha_0(\cdot, W)$. Diggle et al. (1987) observed that $|\epsilon_1(k, R)|$ is larger when $k > 0$ is small, and thus proposed to artificially set the value of the estimator to some constant when $k$ is smaller than a certain minimal wavenumber (Diggle et al., 1987, Equation 3.4). Obviously, this correction is inadequate to study hyperuniformity, i.e., the behavior of $S$ near zero.

An alternative to Diggle's clipping procedure is to proceed as done for the scattering intensity (27) and estimate the structure factor only at a set of allowed wavenumbers, defined as the zeros of $\epsilon_1(\cdot, R)$. Using (7), it comes, for fixed $d$,

$$\epsilon_1(k, R) = \mathcal{F}(\alpha_0)(k) = 2^d \pi^{d/2} \frac{\Gamma(d/2 + 1)}{k^d} J_{d/2}^2(kR)$$

$$\leq \begin{cases} 0 & \text{if } k \in \{\frac{x}{R}; J_{d/2}(x) = 0\}, \\ \mathcal{O}(R^d) & \text{as } k \to 0, \\ \frac{1}{k^d(Rk)^{2/3}} & \text{as } k \to \infty. \end{cases}$$

The two bounds respectively come from the fact that $J_\nu(x) \sim \frac{1}{\Gamma(\nu+1)}(\frac{x}{2})^\nu$ in the neighborhood of zero, and that for all $\nu > 0$ and $x \in \mathbb{R}$, $|J_\nu(x)| \leq c|x|^{-1/3}$ (with $c \approx 0.8$) (Landau, 2000). Thus, for the estimator (42), we let the set of allowed wavenumbers associated to the window $W = B^d(\mathbf{0}, R)$ be

$$\mathbb{A}_R = \left\{ \frac{x}{R} \in \mathbb{R}, \text{ s.t. } J_{d/2}(x) = 0 \right\}. \tag{44}$$

**Proposition 3.2.** *Under Assumption A-2, for $k \in \mathbb{A}_R$, the estimator $\widehat{S}_{\mathrm{BI}}$ is asymptotically unbiased, i.e.,*

$$\sup_{k \in \mathbb{A}_R} \left| \mathbb{E}\left[ \widehat{S}_{\mathrm{BI}}(k) \right] - S(k) \right| \xrightarrow[R \to \infty]{} 0.$$

Note that we can also define the debiased tapered and multitapered estimators assuming stationarity and isotropy, as done in Section 3.1, but the choice of the taper(s) requires more attention, as they must be separable and isotropic.

### 3.2.2 Estimating the structure factor using Ogata quadrature

Still working under Assumption A-2, we can define an estimator of the structure factor (11) by first approximating the pair correlation function from a realization of $\mathcal{X}$, and then approximating the Hankel transform (4).

Estimators of the pair correlation function have been thoroughly investigated; see (Baddeley et al., 2016). In a nutshell, they divide in, on one side, numerical derivatives of estimates of Ripley's $K$ function, and on the other side, direct kernel density estimators based on the collection of pairwise distances in the sample. Both families come with sophisticated edge correction techniques, and, at least for small sample sizes, it seems reasonable to build on this previous work. Henceforth, we assume that an estimator is available, and defer the discussion of which estimator to use to Section 4.

It remains to perform a numerical quadrature on a Hankel transform. Ogata (2005) defines a quadrature for integrals of the form

$$\mathcal{I}_\nu(f) = \int_0^\infty f(x) J_\nu(x) \mathrm{d}x \tag{45}$$

$$\approx \pi \sum_{j=1}^\infty w_{\nu,j} f\left(\frac{\pi}{h}\psi(h\xi_{\nu,j})\right) J_\nu\left(\frac{\pi}{h}\psi(h\xi_{\nu,j})\right) \psi'(h\xi_{\nu,j}), \tag{46}$$

with

$$w_{\nu,j} = \frac{Y_\nu(\pi\xi_{\nu,j})}{J_{\nu+1}(\pi\xi_{\nu,j})}, \quad \text{and} \quad \psi(t) = t \times \tanh\left(\frac{\pi}{2}\sinh(t)\right),$$

$Y_\nu$ is the Bessel function of the second kind of order $\nu$, $h$ is a positive constant called the *stepsize*, and $(\xi_{\nu,j})_{j\geq 1}$ are the positive zeros of the Bessel function $J_\nu(\pi x)$ of the first kind of order $\nu$, arranged in increasing order. In practice, the infinite sum on the right-hand side of (46) can be truncated at a small number of function evaluations since the quadrature nodes approach the zeros of $J_\nu(x)$, that is $\frac{\pi}{h}\psi(h\xi_{\nu,j}) \sim \pi\xi_{\nu,j}$, very fast as $j \to \infty$.

Ogata's quadrature applies to the Hankel transform (3) of an integrable function $f$, since

$$\mathcal{H}_\nu(f)(k) = \mathcal{I}_\nu\left(x \mapsto \frac{x}{k^2} f\left(x/k\right)\right). \tag{47}$$

22

In particular, it applies to the computation of structure factors since, by (11) and (4),

$$S(k) = 1 + \rho \mathcal{F}_s(g-1)(k)$$

$$= 1 + \rho \frac{(2\pi)^{d/2}}{k^{d/2-1}} \mathcal{I}_{d/2-1} \left( x \mapsto \frac{x^{d/2}}{k^{d/2+1}} g(x/k) - 1 \right).$$

We thus define the Hankel-Ogata estimator of the structure factor as

$$\widehat{S}_{\mathrm{HO}}(k) = 1 + \rho \frac{(2\pi)^{d/2}\pi}{k^\nu} \sum_{j=1}^{N} w_{\nu,j} \tilde{h}_k \left( \frac{\pi}{h} \psi(h\xi_{\nu,j}) \right) J_\nu \left( \frac{\pi}{h} \psi(h\xi_{\nu,j}) \right) \psi'(h\xi_{\nu,j}), \qquad (48)$$

with $\nu = d/2 - 1$, $N \in \mathbb{N}$, $\tilde{h}_k(x) = x \mapsto \frac{x^{d/2}}{k^{d/2+1}}(\hat{g}(x/k) - 1)$, and $\hat{g}$ an estimator of the pair correlation function. Finally, note that Ogata's quadrature is also implemented in the Python package `hankel` of Murray and Poulin (2019).

**Relation between $k_{\min}$ and $r_{\max}$.** There exists a hidden inverse proportional relation in Equation (48), between the minimal wavenumber $k_{\min}$ and the maximal radius $r_{\max}$ at which the pair correlation function has been estimated. Truncating the infinite sum after $N$ terms in Equation (46) has been informally justified by

$$\psi(h\xi_{d/2-1,N}) \approx h\xi_{d/2-1,N}. \qquad \text{(Condition 1)}$$

The maximum radius $r_{\max}$ at which $\hat{g}$ is available should in turn satisfy

$$r_{\max} = \max \left\{ \frac{\pi}{hk} \psi(h\xi_{d/2-1,j}), \ k \in \mathbb{R}^*, \ j \in \{1, \ldots, N\} \right\}. \qquad \text{(Condition 2)}$$

(Condition 1) and (Condition 2) entail that

$$k_{\min} \approx \frac{\pi \xi_{d/2-1,N}}{r_{\max}}. \qquad (49)$$

Thus $k_{\min}$ is not only proportional to $1/r_{\max}$ but also the maximum of the considered zero of the Bessel function $J_{d/2-1}(x)$.

### 3.2.3 Estimating the structure factor using Baddour-Chouinard quadrature

Instead of using the quadrature of Ogata (2005), one can estimate Hankel transforms more directly, similarly to how the discrete Fourier transform is used to approximate Fourier

transforms. Intuitively, assuming that either $f$ or its Hankel transform $\mathcal{H}_\nu(f)$ (3), has bounded support allows rewriting it as a Fourier-Bessel series, with coefficients involving evaluations of either $\mathcal{H}_\nu(f)$ or $f$, respectively. Truncating the resulting Fourier-Bessel series yields approximate direct and inverse Hankel transforms. This discrete Hankel transform (DHT) was derived by Baddour and Chouinard (2015), and is implemented as a `Matlab` code. Moreover, Guizar-Sicairos and Gutiérrez-Vega (2004) developed a `Python` package, `pyhank`, based on the same idea.

In detail, let $N > 0$ and $f : \mathbb{R}^+ \to \mathbb{R}$ be a continuous function, Baddour and Chouinard (2015) approximate

$$\mathcal{H}_\nu(f)(k_m) \approx \frac{R^2}{\eta_{\nu,N}} \sum_{j=1}^{N-1} \frac{2}{\eta_{\nu,N} J_{\nu+1}^2(\eta_{\nu,j})} J_\nu\left(\frac{\eta_{\nu,m}\eta_{\nu,j}}{\eta_{\nu,N}}\right) f(r_j), \quad 1 \leq j, m \leq N-1, \qquad (50)$$

where $\eta_{\nu,m} = \pi\xi_{\nu,m}$ is the $m^{th}$ positive zero of the Bessel function $J_\nu(x)$ of the first kind,

$$r_j = \frac{\eta_{\nu,j}}{\eta_{\nu,N}}R, \quad k_m = \frac{\eta_{\nu,m}}{\eta_{\nu,N}}K, \quad K = \frac{\eta_{\nu,N}}{R}, \quad \text{with} \quad R > 0. \qquad (51)$$

The user thus needs to specify both $N$ and $R$. Intuitively, the choice of $R$ is governed by how far on the positive axis one has been able to evaluate $f$. Once $R$ is fixed, $N$ decides how large $K$ is, that is, how high in frequency one wishes to estimate the Hankel transform.

To conclude, given an estimator $\hat{g}$ of the pair correlation function, we define yet another estimator of the structure factor, called the Hankel-Baddour-Chouinard estimator,

$$\widehat{S}_{\text{HBC}}(k_m) = 1 + \rho(2\pi)^{d/2} \frac{R^2}{\eta_{\nu,N}} \sum_{j=1}^{N-1} \frac{2}{\eta_{\nu,N} J_{d/2}^2(\eta_{\nu,j}) k_m^\nu} J_\nu\left(\frac{\eta_{\nu,m}\eta_{\nu,j}}{\eta_{\nu,N}}\right) \tilde{h}(r_j), \qquad (52)$$

where $\nu = d/2 - 1$, $\tilde{h}(x) = x^\nu(\hat{g}(x) - 1)$, and the set of wavenumbers $\{k_m\}_m$ is fixed by (51). Finally, we can deduce from (51) that the minimal wavenumber of $\widehat{S}_{\text{HBC}}$ (52) is $k_1 = \frac{\eta_{d/2-1,1}}{R}$. Comparing $k_1$ with the minimal wavenumber $k_{\min}$ (49) of $\widehat{S}_{\text{HO}}$ (48) for the same number of points $N$ and $r_{\max} = R$, we observe that $k_1 < k_{\min}$, as $k_1$ is proportional to the first zero of the Bessel function $J_{d/2-1}(x)$ while $k_{\min}$ is proportional to the $N^{th}$ zero of $J_{d/2-1}(x)$.

# 4 Demonstrating all estimators using our toolbox

We have implemented all estimators of Section 3, as well as the regression diagnostics of Section 2.5, in an open-source `Python` toolbox called `structure_factor`[4]. The package is licensed under the MIT license, and is available on Github and PyPI (under the project name `structure-factor`).

In this section, we quickly demonstrate the toolbox on the four point processes described in Section 2.6, that is, the KLY process of intensity $\rho_{\mathrm{KLY}} = 1$, the Ginibre ensemble of intensity $\rho_{\mathrm{Ginibre}} = 1/\pi$, the Poisson process of intensity $\rho_{\mathrm{Poisson}} = 1/\pi$, and the Thomas process with intensity $\rho_{\mathrm{Thomas}} = 1/\pi$, $\rho_{\mathrm{parent}} = 1/(20\pi)$ and $\sigma = 2$. This leads to samples of size 5800 points approximately.

The dimension is always $d = 2$. All figures in this section can be reproduced by following our demonstration notebooks.

## 4.1 Basic software objects

While we refer to our online documentation[5] for details, we believe that a quick overview of the main objects of our package is useful. All estimators from Section 3 are methods of the class `StructureFactor`. The class constructor takes as input an object of type `PointPattern`. In a nutshell, for a stationary point process $\mathcal{X}$ of intensity $\rho$, a `PointPattern` encapsulates a sample of $\mathcal{X} \cap W = \{x_1, \ldots, x_N\}$, the observation window $W$, and the intensity $\rho$ (optional). If the intensity of $\mathcal{X}$ is not provided by the user, it is automatically approximated by the asymptotically unbiased estimator $\widehat{\rho} = \frac{N}{|W|}$. Finally, to comply with the requirements of estimators on specific windows, we provide a `restrict_to_window()` method for the class `PointPattern`.

## 4.2 Demonstrating estimators that only assume stationarity

**The scattering intensity.** Figure 3 illustrates the scattering intensity estimator of Section 3.1. Columns respectively correspond to the KLY, Ginibre, Poisson, and Thomas

---

[4]https://github.com/For-a-few-DPPs-more/structure-factor
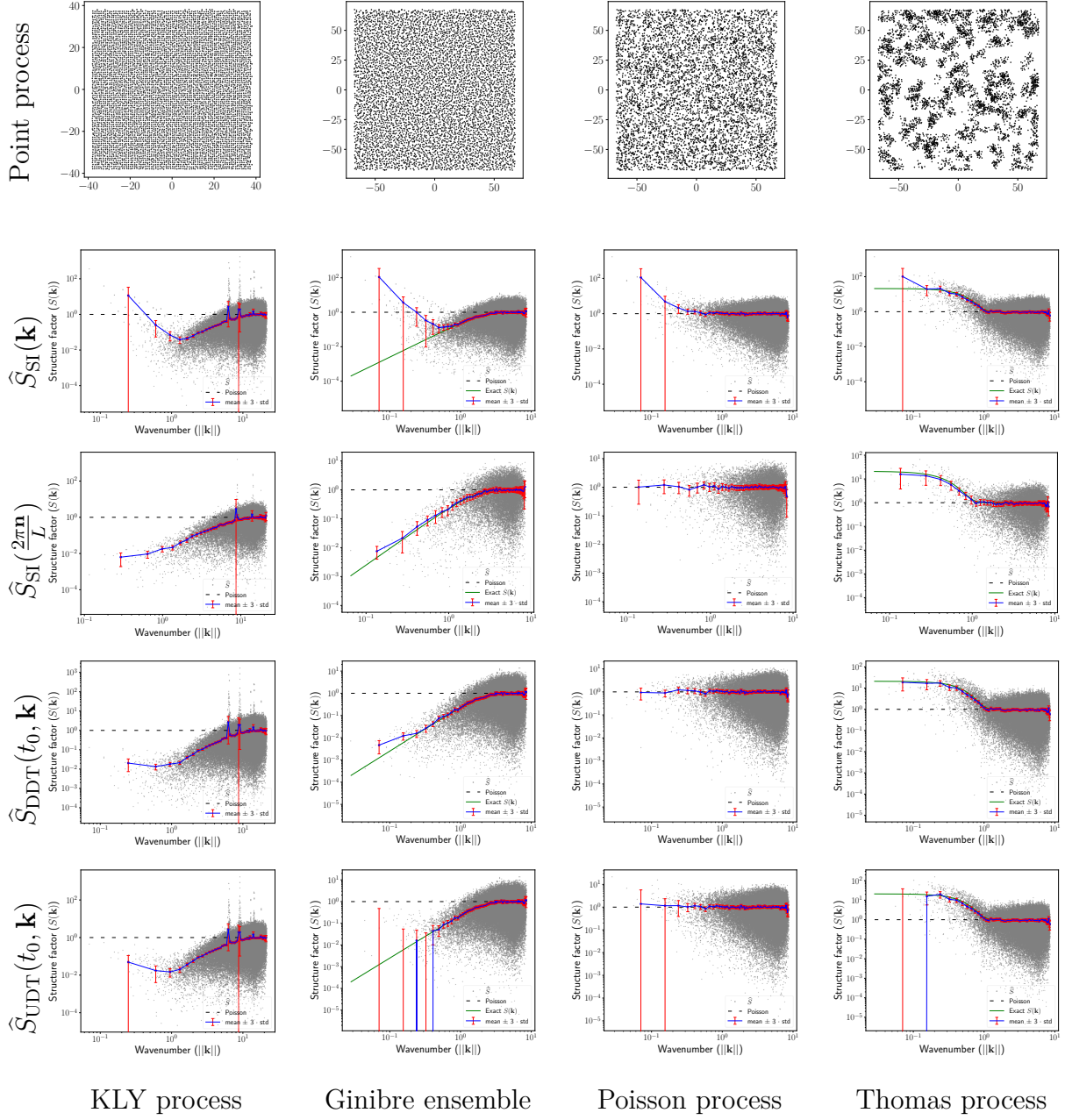[5]https://for-a-few-dpps-more.github.io/structure-factor/

Figure 3: Variants of the scattering intensity estimator applied to four point processes. The computation and visualization are done using `structure_factor`. See Section 4.2 for details.

point processes. The first row contains a sample of each point process, observed in square windows. The second row shows the scattering intensity $\widehat{S}_{\mathrm{SI}}$ in (27) on arbitrary wavevectors $\mathbf{k}$, while in the third row, the estimators are only evaluated on a subset of the allowed

wavevectors (26). The fourth and fifth rows illustrate the debiasing techniques, respectively the directly debiased scattering intensity $\widehat{S}_{\mathrm{DDT}}(t_0, \mathbf{k})$ from (36), and the undirectly debiased scattering intensity $\widehat{S}_{\mathrm{UDT}}(t_0, \mathbf{k})$ from (35).

The clouds of grey points are the approximated structure factors. Note that since all benchmark point processes are (at least approximately) isotropic, the structure factor $S$ is a radial function, so we plot $k \mapsto \hat{S}(k)$, $k \in \mathbb{R}$, and not $\mathbf{k} \mapsto \hat{S}(\mathbf{k})$, $\mathbf{k} \in \mathbb{R}^2$. To regularize the obtained estimator, we bin the norm of the wavevectors regularly and provide the empirical mean (in blue) and the empirical standard deviation of the mean (red bars indicate $\pm 3$ such standard deviations). Note that the binning can be specified by the user in our library. On each plot, the exact structure factor is represented by a green line when it is known. Finally, the dashed black lines are the structure factor of the homogeneous Poisson process, for reference.

While we refer to Section 5 for a more detailed comparison, one can already observe from Figure 3 that the most accurate estimators are the scattering intensity $\widehat{S}_{\mathrm{SI}}$ (27) evaluated on the set of allowed wavevectors (26) and the debiased scattering intensity $\widehat{S}_{\mathrm{DDT}}(t_0, \mathbf{k})$ (36). The bias at small, non-allowed wavenumbers of the scattering intensity is visible in the second row. As for the undirectly debiased variant, it produces a few negative values, visible as large error bars on our log-log plot.

**Using an alternate taper.** Now, as mentioned in Section 3, the scattering intensity $\widehat{S}_{\mathrm{SI}}$ is a particular case of the tapered periodogram $\widehat{S}_{\mathrm{T}}$ of Rajala et al. (2020), with the specific taper $t_0$. We are free to use other tapers verifying (33).

Figure 4 shows the estimated structure factors of the same four benchmark point processes (first row), using the tapered estimator $\widehat{S}_{\mathrm{T}}$ in (30) (second row), the corresponding directly debiased version $\widehat{S}_{\mathrm{DDT}}$ (third row), and the undirectly debiased version $\widehat{S}_{\mathrm{UDT}}$ (last row). The taper used is the first sinusoidal taper $t_1(\mathbf{x}, W) = t(\mathbf{x}, \mathbf{p}^1, W)$ with $\mathbf{p}^1 = (1, 1)$ in (38). The same legend applies as for Figure 3.

First, the asymptotic bias of $\widehat{S}_{\mathrm{T}}$ at small wavenumber $k$ is visible in the second row. Second, for the KLY process (first column), the Ginibre ensemble (second column), and the Poisson process (third column) the estimator $\widehat{S}_{\mathrm{UDT}}$ (last row) returned a few negative values

again, resulting in large inaccuracies in our log-log scale. The directly debiased estimator $\widehat{S}_{\mathrm{DDT}}$ yields the most accurate approximation of known structure factors, consistently across point processes.
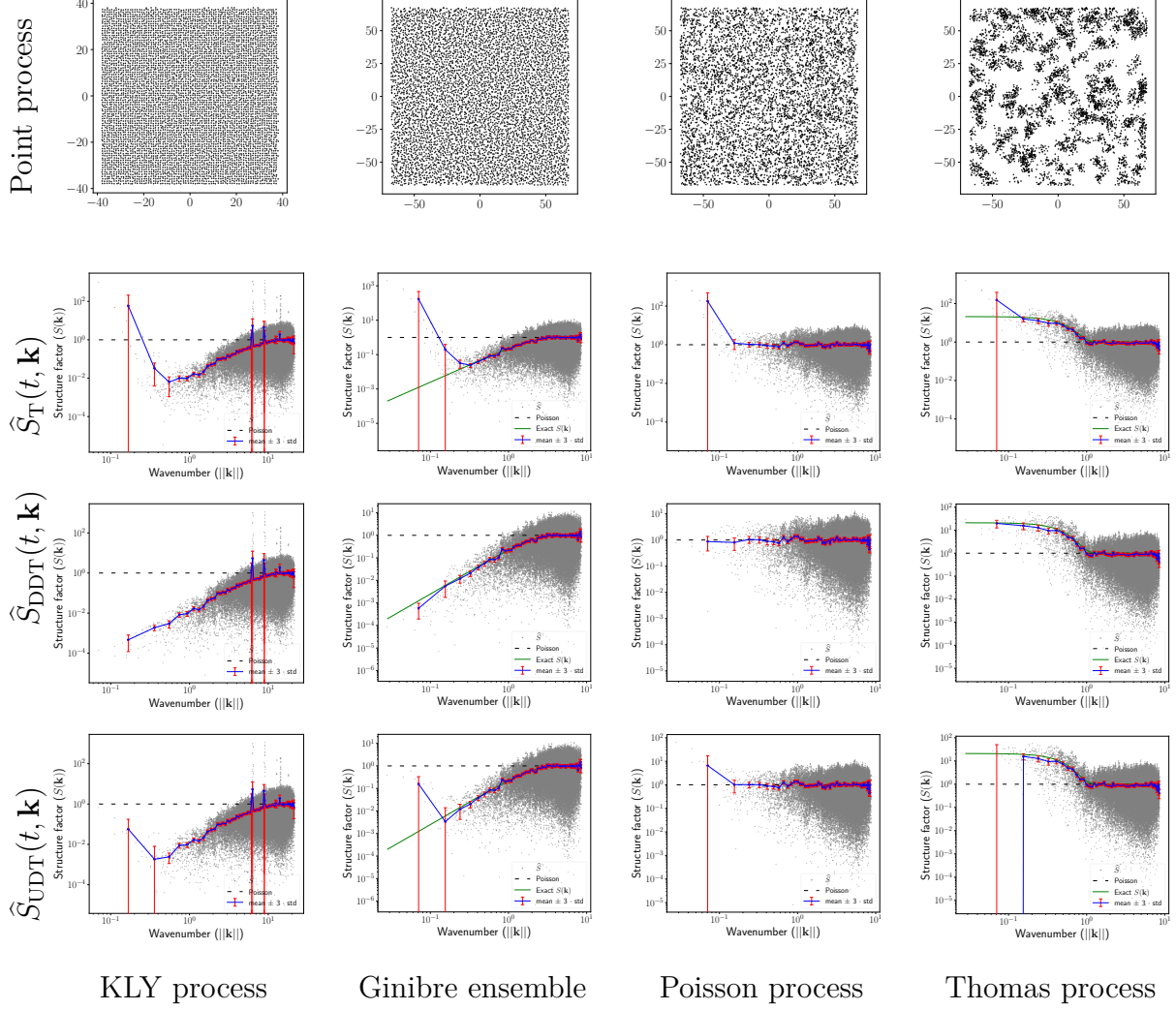


Figure 4: Scaled tapered periodogram and the corresponding debiased versions: KLY process (first column), Ginibre ensemble (second column), Poisson process (third column), and Thomas process (last column). The computation and visualization are done using structure_factor.

**Averaging over multiple tapers.** The multitapered estimator $\widehat{S}_{\mathrm{MT}}$ of (37) is now investigated in Figure 5, using the first four sinusoidal tapers, i. e., $(t_q)_{q=1}^4$ with $t_q(\mathbf{x}, W) =$

$t(\mathbf{x}, \mathbf{p}^q, W)$ and $\mathbf{p}^q \in \{1, 2\}^2$ in (38). We also show the results of the corresponding directly and undirectly debiased versions, $\widehat{S}_{\text{DDMT}}$ and $\widehat{S}_{\text{UDMT}}$.
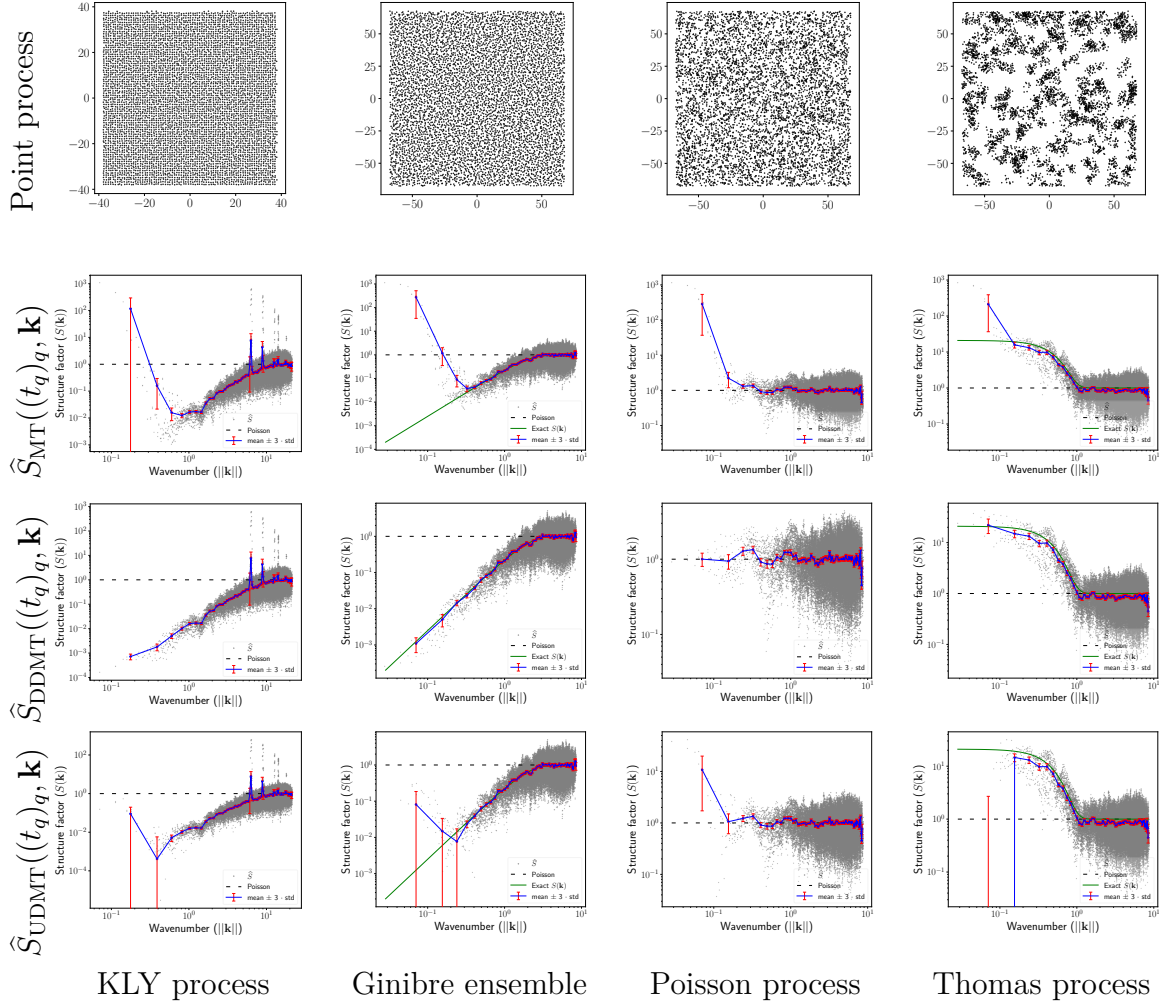


Figure 5: Scaled multitapered periodogram and the debiased version: KLY process (first column), Ginibre ensemble (second column), Poisson process (third column), and Thomas process (last column). The computation and visualization are done using structure_factor.

We again observe the bias of $\widehat{S}_{\text{MT}}$ at small wavenumbers (second row), and that the negative values output by $\widehat{S}_{\text{UDMT}}$ at small wavenumbers (last row) make visual assessments of hyperuniformity less straightforward. Like with single tapers, the directly debiased estimator $\widehat{S}_{\text{DDMT}}$ gives a consistently accurate approximation. Compared to Figure 4, however, it is not obvious whether multitapering yields a smaller mean square error than

single tapers, and a more quantitative study will investigate this in Section 5.

## 4.3   Demonstrating estimators that assume isotropy

**Bartlett's isotropic estimator.**   Figure 6 illustrates Bartlett's isotropic estimator of Section 3.2. Columns respectively correspond to the KLY, Ginibre, Poisson, and Thomas point processes. The first row contains a sample of each point process, observed in ball windows. The second row shows $\widehat{S}_{\mathrm{BI}}$ on arbitrary wavenumbers $k$, while in the last row, the estimator is only evaluated on a subset of the Bessel-specific allowed wavenumbers (44).

First, we note that, unlike scattering intensity variants, plotting Bartlett's isotropic estimator $k \mapsto \widehat{S}_{\mathrm{BI}}(k)$ in (42) does not require binning. On the other hand, Bartlett's estimator is significantly costlier than its scattering intensity counterpart; See Section 5.1. Now, we comment on the accuracy of the estimator in Figure 6. Here again, small, non-allowed wavenumbers give rise to large biases for $\widehat{S}_{\mathrm{BI}}(k)$, especially for the two hyperuniform point processes (KLY and Ginibre). When applied to allowed wavenumbers only, the estimator shows accuracy across all point processes, similarly to the directly debiased periodograms.

**Estimating the pair correlation function.**   The last two estimators of the structure factor are $\widehat{S}_{\mathrm{HO}}$ in (48) and $\widehat{S}_{\mathrm{HBC}}$ in (52). These estimators require an approximation of the pair correlation function (pcf) of the point process. We thus quickly investigate standard estimators of the pcf on our benchmark point processes.

There are two types of estimators of the pcf for stationary isotropic point processes (Baddeley et al., 2016): kernel density estimators applied to pairwise distances, and numerical derivatives of Ripley's $K$ function. The R library `spatstat` implements both, respectively as `pcf.ppp`, which uses an Epanechnikov kernel and Stoyan's rule of thumb for bandwidth selection (Baddeley et al., 2016, Section 7.6.2), and `pcf.fv`, which computes the derivative of a polynomial estimator of Ripley's $K$ function.

The kernel density estimator behaves badly for small values of $r$: for many point processes, its variance becomes infinite when $r$ goes to 0. The derivative estimator is recommended for large datasets, where direct estimation of the pcf can be time-consuming
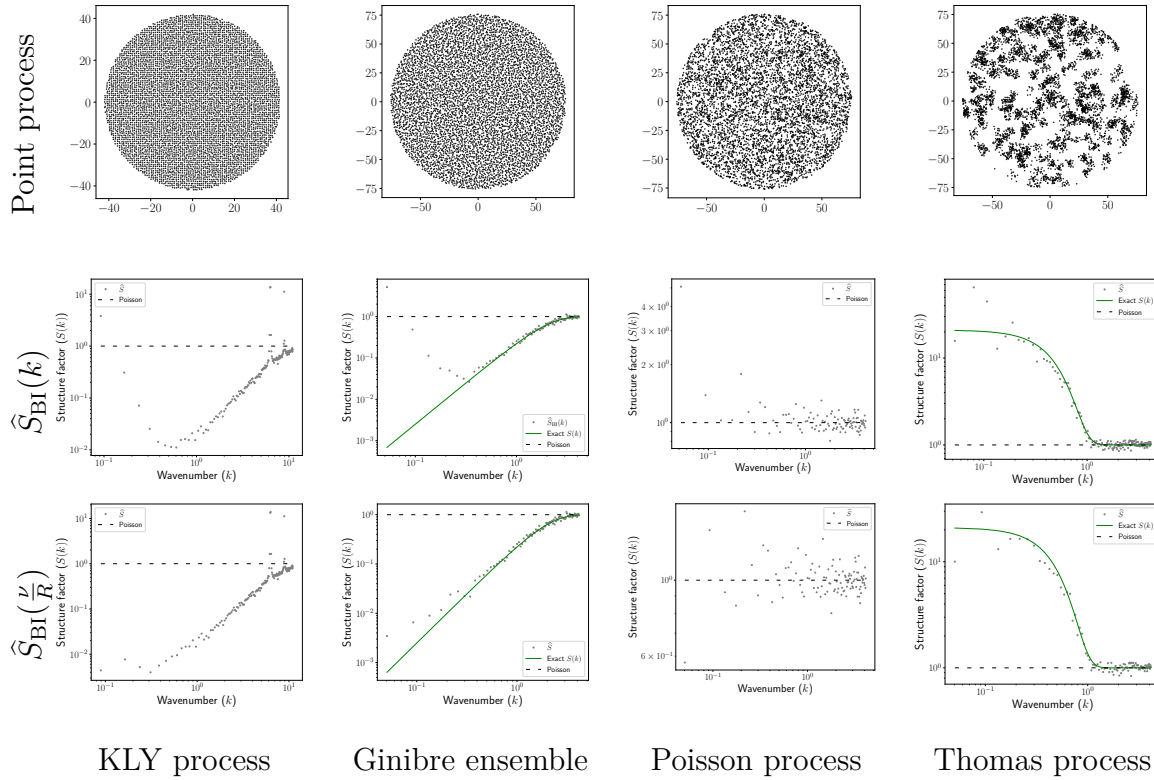
Figure 6: Bartlett isotropic estimator: KLY process (first column), Ginibre ensemble (second column), Poisson process (third column), and Thomas process (last column). The computation and visualization are done using `structure_factor`.

(Baddeley et al., 2016, Section 7.6.2).

Figure 7 shows the two estimators of the pair correlation function of the benchmark point processes. Note that we provide an independent, open-source Python interface[6] to the R library spatstat. The second row shows the estimation of the pair correlation function using `pcf.ppp`. This method provides a choice of boundary corrections, like `"trans"`, `"iso"`, or none (`"un"`). For more details see Baddeley et al. (2016, Sections 7.4.4 and 7.4.5). The last row of Figure 7 shows the estimation of the pair correlation function using `pcf.fv`.

We observe that, for the cardinalities considered here, the choice of edge correction method is irrelevant, as long as there is one. We also observe that the two methods for
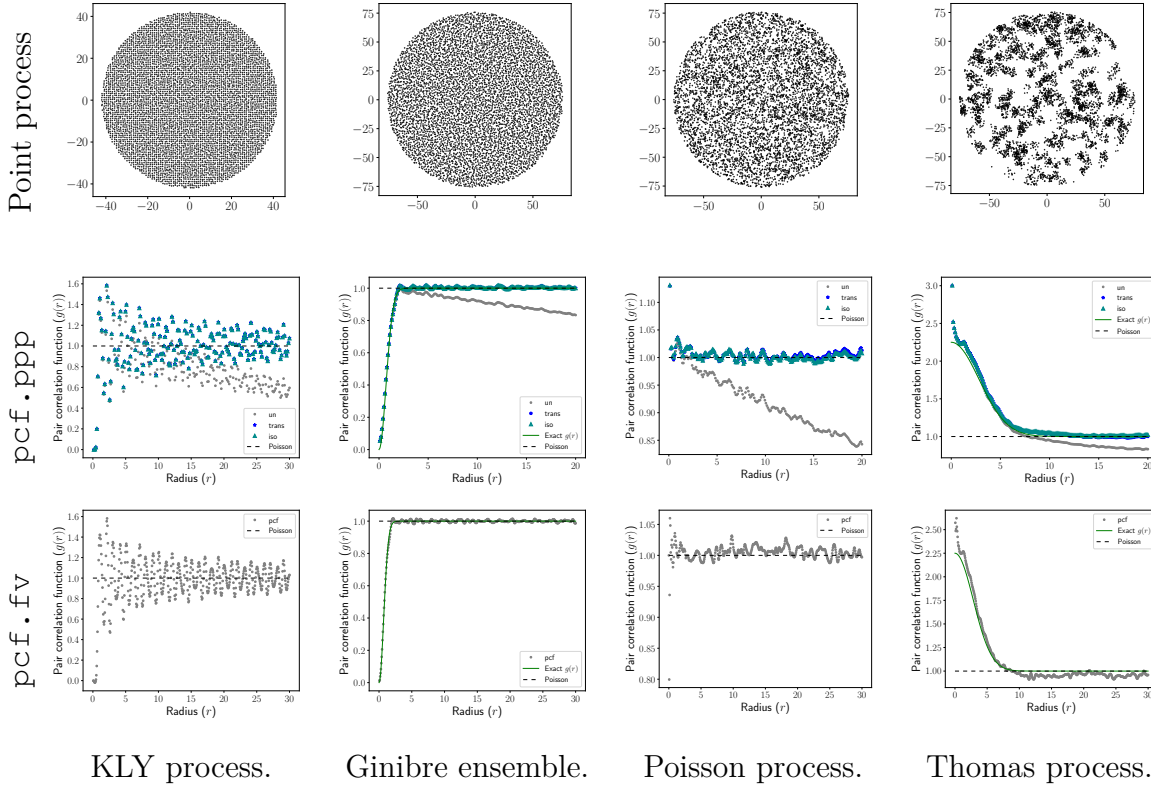
---

Figure 7: Approximated pair correlation function: KLY process (first column), Ginibre ensemble (second column), Poisson process (third column), and Thomas process (last column). The computation and visualization are done using `structure_factor`.

estimating the pcf perform similarly, and we pick `pcf.fv` for the rest of this section. We manually remove undefined values (`NaN`, `-Inf`, or `Inf`), and we interpolate the obtained discrete approximation of $g$, in order to evaluate it at any point required by the quadratures of Section 3.2.2, and 3.2.3. Finally, note that the maximum radius $r_{max}$ at which `spatstat` provides an approximation of $g$ is limited by the size of the observation window. Typically, it should be less than half the window diameter for a ball window, and less than $1/4$ of the smaller side length of the window for a rectangular window; see the documentation of `spatstat`. For larger values than the $r_{max}$ provided by `spatstat`, we manually set $g$ to be identically 1, which has the effect of automatically truncating quadratures that evaluate $g - 1$, like Ogata's quadrature (48).
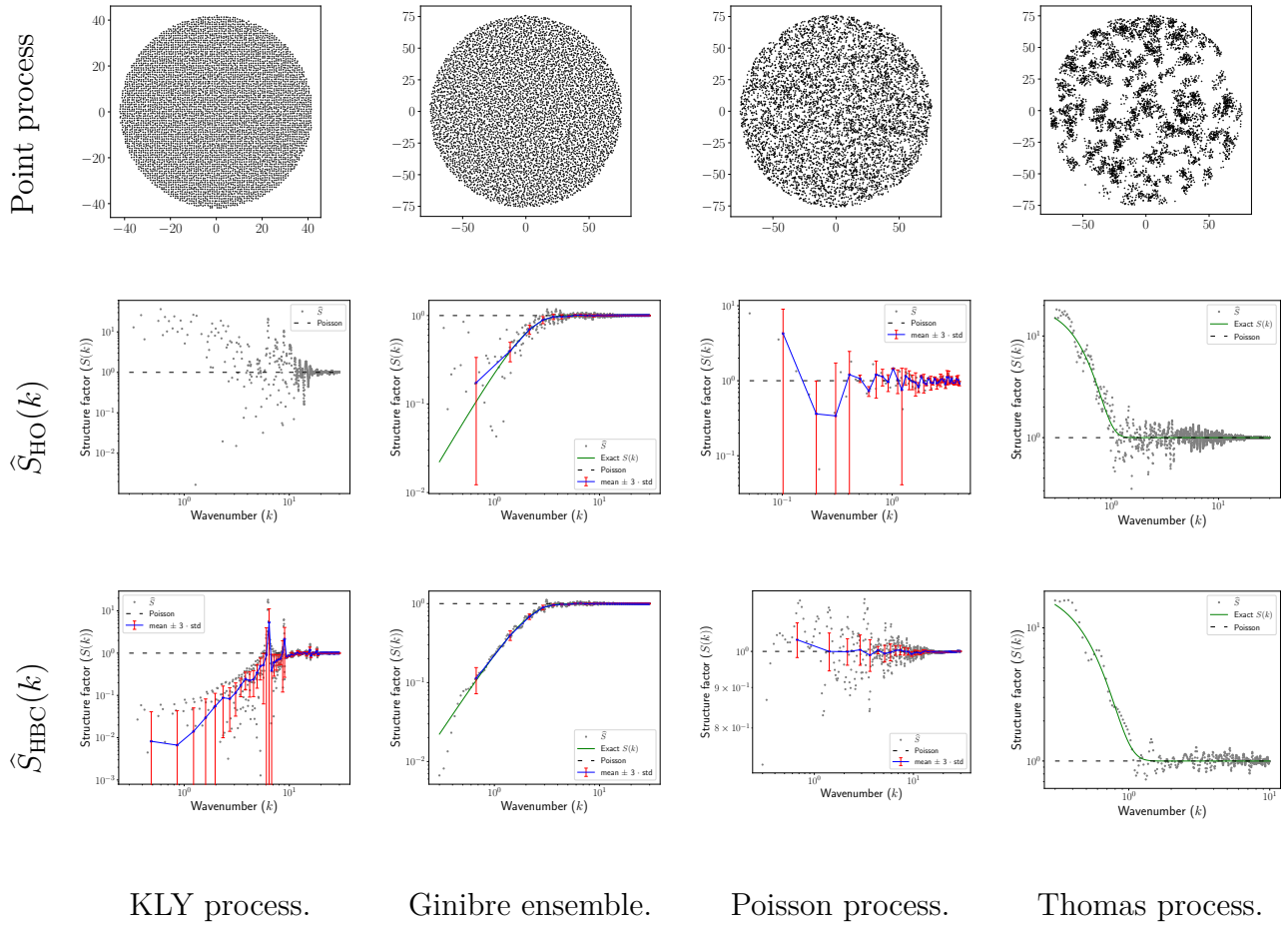
Figure 8: Estimation using Hankel transform quadratures: KLY process (first column), Ginibre ensemble (second column), Poisson process (third column), and Thomas process (last column). The computation and visualization are done using `structure_factor`.

**Hankel transform quadratures.** Figure 8 shows the results of Ogata's $\widehat{S}_{\mathrm{HO}}$ (second row) and Baddour-Chouinard's $\widehat{S}_{\mathrm{HBC}}$ (last row) on our four benchmark point processes from Section 2.1. The legend is the same as for Figure 3; see Section 4.2. For the accuracy of the estimators, we can see that $\widehat{S}_{\mathrm{HO}}$ failed to approximate the structure factor of the KLY process. Even the results of $\widehat{S}_{\mathrm{BC}}$ seem to be unreliable. The reason may be that the KLY process is not an isotropic point process, and thus that the fluctuations of its approximated pair correlation function in Figure 7 cause the quadrature to be inaccurate. For the remaining point processes, $\widehat{S}_{\mathrm{BC}}$ seems to give more accurate results than $\widehat{S}_{\mathrm{HO}}$.

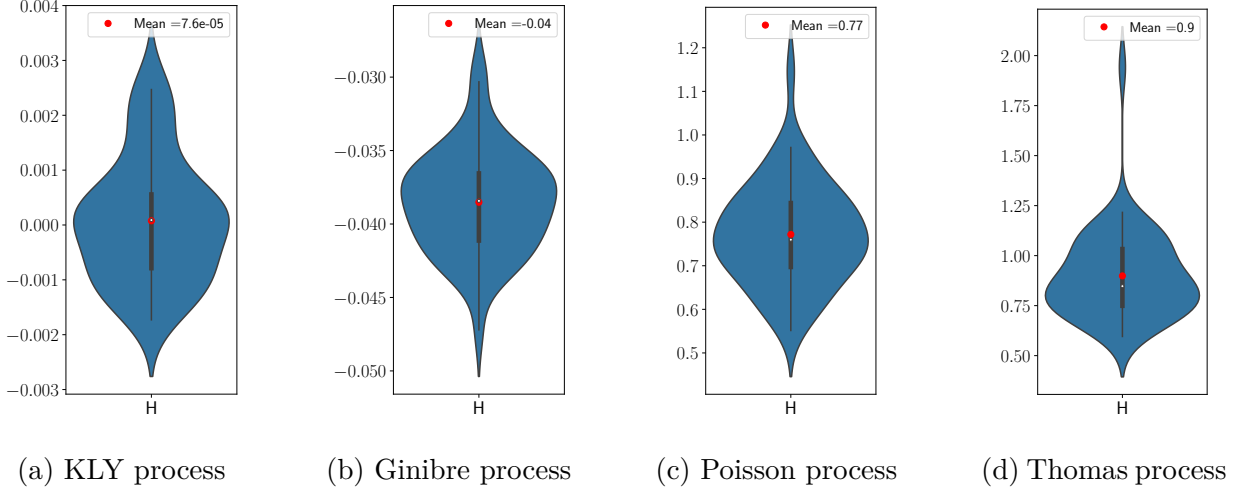(a) KLY process     (b) Ginibre process     (c) Poisson process     (d) Thomas process

Figure 9: Violin plot of the $H$-index across 50 samples from the KLY, Ginibre, Poisson, and Thomas point process. The computation is done using `structure_factor`.

## 4.4 Hyperuniformity diagnostics

We now demonstrate the estimation of the decay rate $\alpha$ from Section 2.4 and the hyperuniformity index $H$ described in (14).

Figure 9 illustrates the violin plots of $H$ across 50 samples from the benchmarks point processes, and Figure 10 shows the violin plots of the estimated power decay $\alpha$ across the samples of the KLY and the Ginibre point processes. We used the results of $\widehat{S}_{\mathrm{BI}}$, across 50 samples from the benchmark point processes of roughly $10^4$ points each (see Section 4.3). To fit the line required to compute $H$, we restricted to wavenumbers up to 0.6 for the Thomas process, and 1 for the remaining point processes. These values were chosen manually: the trade-off is to remain close to zero, while including enough data points to fit a line. The violin plots of Figure 9 indicate that, consistently across realizations of the Poisson and Thomas point processes, the $H$-index is larger than, say, 0.5. This is a strong hint that these point processes are not hyperuniform. On the contrary, for Ginibre, $H$ is even slighlty negative, hinting at hyperuniformity. For the KLY process, although the $H$-index is close to zero, we note that a threshold of $10^{-3}$ would not lead to the same answer across all 50 realizations.

Concerning the decay rate $\alpha$ of the structure factor, the maximum wavenumber used to
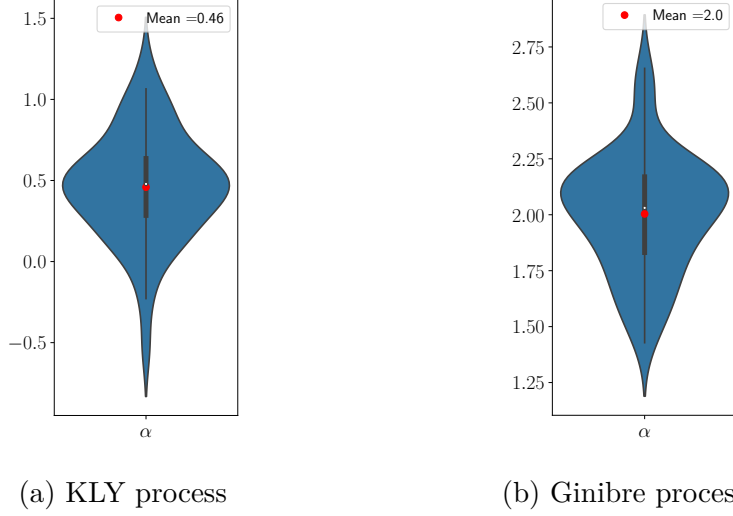
(a) KLY process  (b) Ginibre process

Figure 10: Violin plots of the structure factor power decay $\alpha$ across 50 samples of the KLY and the Ginibre point process. The computation is done using `structure_factor`.

fit the polynomial is equal to 0.45 for the Ginibre ensemble, and to 0.6 for the KLY process. Again, these thresholds are fit manually, and represent a trade-off between having enough points to fit our monomial and being close to zero. There is limited evidence that the KLY process is indeed hyperuniform: while the support of the distribution is large and includes 0, most estimated values of $\alpha$ are positive, and concentrate around 0.5. This is a hint that KLY might belong to Class III in Table 1. For the Ginibre ensemble, the concentration of values of $\alpha$ around 2 successfully reflects the known power decay (Section 2.1).

# 5  A quantitative comparison of the estimators

We now compare the cost and accuracy of all estimators more quantitatively.

## 5.1  Computational cost

For a given wavevector, evaluating any of the tapered variants of the scattering intensity requires a sum of $N$ terms. Multitapering naturally multiplies the cost by the number of tapers, but it can be trivially parallelized, especially since the number of tapers remains low in practice (Rajala et al., 2020).

On the other hand, for a sample of $N$ points, Bartlett's estimator $\widehat{S}_{\mathrm{BI}}$ is a sum of $\mathcal{O}((N^2 - N)/2)$ evaluations of a Bessel function. This makes Bartlett's estimator significantly costlier than its scattering intensity counterpart. As an example, for a realization of $N = 10^4$ points from a point process in $\mathbb{R}^2$, using a modern laptop, the evaluation of $\widehat{S}_{\mathrm{SI}}$ at a single wavevector took about one millisecond, compared to 2 seconds for $\widehat{S}_{\mathrm{BI}}$. Note however that for a similar accuracy on an isotropic point process, $\widehat{S}_{\mathrm{BI}}$ needs to be evaluated at fewer wavevectors than, say, the scattering intensity. Indeed, the latter is averaged over annuli to obtain binwise averages.

Now, for Hankel quadrature estimators, the main bottleneck is the approximation of the pcf. A kernel density estimator based on $N$ points is again a quadratic computation. In our experience, combining the pcf estimator `pcf.fv`, based on numerically differentiating an estimated Ripley's K function, and a Hankel quadrature to build $\widehat{S}_{\mathrm{HO}}$ and $\widehat{S}_{\mathrm{HBC}}$ is the least expensive pipeline.

## 5.2 Measuring the accuracy of the estimators near zero

To confirm the intuitions gained from Section 4, in particular that the multitapered variant of the scattering intensity and Bartlett's estimator dominate the others when it comes to visual diagnostics of hyperuniformity, we now compare the mean square error of the different estimators of the structure factor, integrated near 0.

For a subdivision $[k_1, k_2) \cup \cdots \cup [k_{J-1}, k_J)$ of some interval $[k_1, k_J)$, the integrated MSE reads

$$
\begin{aligned}
\mathrm{iMSE}(\widehat{S}) &= \int_{k_1}^{k_J} \mathrm{MSE}(\widehat{S})(k)\mathrm{d}k \\
&= \int_{k_1}^{k_J} \mathbb{E}[\widehat{S}(k) - S(k)]^2 \mathrm{d}k \\
&= \mathbb{E}\left[\int_{k_1}^{k_J} [\widehat{S}(k) - S(k)]^2 \mathrm{d}k\right] \\
&\approx \frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{J-1} (k_{j+1} - k_j)\left\{[\widehat{S}(k_{j+1}) - S(k_{j+1})]^2 + [\widehat{S}(k_j) - S(k_j)]^2\right\}\right], \quad (53)
\end{aligned}
$$

where the last line is a crude numerical integration using the trapezoidal rule. The expectation in (53) is under the investigated point process. Assuming it is easy to sample

36

from the point process, we can run a Student test to compare (53) for two estimators of the structure factor. Moreover, we can run a paired Student test by using the same point process samples for the two estimators. In other words, we can test whether the difference of the approximated integrated MSE (53) of two estimators is 0.

## 5.3 Comparing variants of the scattering intensity

In Section 4, we derived the intuition that $\widehat{S}_{\mathrm{SI}}$ on its allowed values, $\widehat{S}_{\mathrm{DDT}}(t_0, \mathbf{k})$, $\widehat{S}_{\mathrm{DDT}}(t_1, \mathbf{k})$, and $\widehat{S}_{\mathrm{DDMT}}((t_q)_{q=1}^4, \mathbf{k})$ gave the most accurate approximations among the estimators of Section **??**. We further expect $\widehat{S}_{\mathrm{DDMT}}((t_q)_{q=1}^4, \mathbf{k})$ to have the smallest integrated MSE among them. To test this, we run three paired, one-sided Student tests, comparing $\widehat{S}_{\mathrm{DDMT}}((t_q)_{q=1}^4, \mathbf{k})$ to each of the three former estimators.

Specifically, we use $M = 50$ independent realizations from each of the Ginibre, Poisson, and Thomas point processes, for which we know the exact structure factor. The observation window $W$ is always is the same two-dimensional cubic window of side $L = 70$, and the intensity of all processes is $\rho = 1/\pi$. The additional parameters of the Thomas process are, as used throughout the paper, $\rho_{\mathrm{parent}} = 1/(20\pi)$, and $\sigma = 2$. We evaluated $\widehat{S}_{\mathrm{SI}}$ (27) on its allowed wavevectors (26), between $k_1 = 0.1$ and $k_J = 2.8$. For the other estimators, $\widehat{S}_{\mathrm{DDT}}(t_0, \mathbf{k})$, $\widehat{S}_{\mathrm{DDT}}(t_1, \mathbf{k})$ (36), and $\widehat{S}_{\mathrm{DDMT}}((t_q)_{q=1}^4, \mathbf{k})$ (37), we used arbitrary wavevectors keeping the same range of the wavenumbers used for $\widehat{S}_{\mathrm{SI}}$. The tapers used are $t_0$ (31), and the first 4 tapers $(t_q)_{q=1}^4$ of the family of sinusoidal tapers (38). As we got different approximated values of $S$ for different wavevectors of the same wavenumber, we set the estimated structure factor to that sample mean.

Table 2 summarizes the results of the paired one-sided Student tests. For each point process, applying a Bonferroni correction, we can simultaneously reject at the level 0.01 the three hypotheses that there is no difference in mean between the estimated iMSEs. This confirms our intuition that $\widehat{S}_{\mathrm{DDMT}}((t_q)_1^4)$ yields the smallest integrated MSE among the considered variants of the scattering intensity. In particular, multitapering helps.

For information, the estimated MSEs are given in Table 3, in the form of a confidence interval (CI) of a sample mean plus or minus 3 empirical standard deviations of the mean.

| Estimators | $T$-score | $p$-value | $T$-score | $p$-value | $T$-score | $p$-value |
|---|---|---|---|---|---|---|
| $\widehat{S}_{\mathrm{DDMT}}((t_q)_1^4)$, $\widehat{S}_{\mathrm{SI}}(2\pi\mathbf{n}/L)$ | $-29.53$ | $3 \times 10^{-33}$ | $-41.59$ | $3 \times 10^{-40}$ | $-9.24$ | $10^{-12}$ |
| $\widehat{S}_{\mathrm{DDMT}}((t_q)_1^4)$, $\widehat{S}_{\mathrm{DDT}}(t_0)$ | $-22.40$ | $10^{-27}$ | $-30.42$ | $8 \times 10^{-34}$ | $-6.38$ | $2 \times 10^{-8}$ |
| $\widehat{S}_{\mathrm{DDMT}}((t_q)_1^4)$, $\widehat{S}_{\mathrm{DDT}}(t_1)$ | $-12.18$ | $9 \times 10^{-17}$ | $-25.39$ | $3 \times 10^{-30}$ | $-7.16$ | $10^{-9}$ |
| | Ginibre ensemble | | Poisson process | | Thomas process | |

Table 2: Paired $t$-tests for the variants of the scattering intensity.

There is roughly a factor 4 between the directly debiased multitapered estimator and each of the other three, which confirms a clear domination. For reference, we have also indicated $(\widehat{\mathrm{iVar}})$ the contribution of the sample integrated variance to the sample integrated MSE. Clearly, the variance is the biggest contributor, and the squared bias is at least one order of magnitude smaller, for all point processes.

Finally, Figure 11 shows the pointwise average of each estimator, across the $M = 50$ point process realizations. A reduction of bias at small $k$ for $\widehat{S}_{\mathrm{DDT}}(t_1, \mathbf{k})$ and $\widehat{S}_{\mathrm{DDMT}}((t_q)_{q=1}^4, \mathbf{k})$ is clearly visible for Ginibre.

| Estimators | $\widehat{\mathrm{iVar}}$ | $\mathrm{CI}[\widehat{\mathrm{iMSE}}]$ | $\widehat{\mathrm{iVar}}$ | $\mathrm{CI}[\widehat{\mathrm{iMSE}}]$ | $\widehat{\mathrm{iVar}}$ | $\mathrm{CI}[\widehat{\mathrm{iMSE}}]$ |
|---|---|---|---|---|---|---|
| $\widehat{S}_{\mathrm{SI}}(2\pi\mathbf{n}/L)$ | 0.32 | $0.32 \pm 0.02$ | 1.31 | $1.34 \pm 0.06$ | 69.51 | $70.71 \pm 17.95$ |
| $\widehat{S}_{\mathrm{DDT}}(t_0)$ | 0.32 | $0.33 \pm 0.03$ | 1.44 | $1.47 \pm 0.1$ | 72.15 | $73.63 \pm 26.12$ |
| $\widehat{S}_{\mathrm{DDT}}(t_1)$ | 0.34 | $0.35 \pm 0.06$ | 1.47 | $1.50 \pm 0.14$ | 79.29 | $80.51 \pm 27.20$ |
| $\widehat{S}_{\mathrm{DDMT}}((t_q)_1^4)$ | 0.08 | $\mathbf{0.08} \pm 0.007$ | 0.37 | $\mathbf{0.38} \pm 0.02$ | 17.90 | $\mathbf{18.19} \pm 4.19$ |
| | Ginibre ensemble | | Poisson process | | Thomas process | |

Table 3: Sample integrated variance and MSE for the variants of the scattering intensity across 50 samples from the Ginibre ensemble, Poisson process, and Thomas process. The computation is done using `structure_factor`.
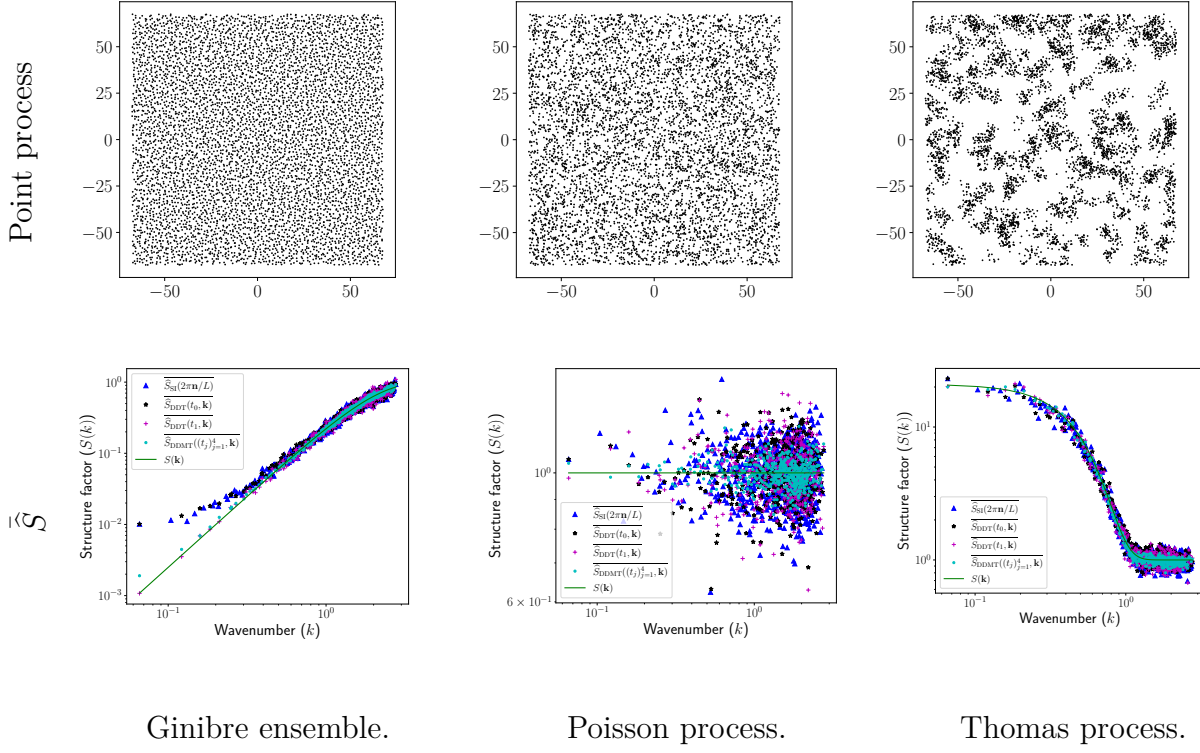
Figure 11: Pointwise averages of the variants of the scattering intensity, numerically rotation-averaged, across 50 independent realizations of the Ginibre, Poisson, and Thomas point processes. The computation and visualization are done using `structure_factor`.

## 5.4 Comparing estimators that assume isotropy

We now run the same comparison on $\widehat{S}_{\mathrm{BI}}$ (42), $\widehat{S}_{\mathrm{HO}}$ (48), and $\widehat{S}_{\mathrm{HBC}}$ (52). For this study, we have sampled $M = 50$ independent realizations from the Ginibre, Poisson, and Thomas process, with the same parameters as in Section 5.3, except that the observation window for all point processes in now a 2-dimensional ball window centered at the origin with radius $R = 100$. For $\widehat{S}_{\mathrm{HO}}$, and $\widehat{S}_{\mathrm{HBC}}$ we used the method `pcf.fv` from `spatstat` to approximate the pair correlation function with maximal approximation radius $r_{\max} = 30$.

We have noted in Section 4 that $\widehat{S}_{\mathrm{HBC}}$ is more robust to non-isotropy than $\widehat{S}_{\mathrm{HO}}$, and that $\widehat{S}_{\mathrm{BI}}$ gave the tightest approximations in this family of estimators.

Table 4, summarizes the results of two one-sided paired $t$-tests per point process. For

| Estimators | $T$-score | $p$-value | $T$-score | $p$-value | $T$-score | $p$-value |
|---|---|---|---|---|---|---|
| $\widehat{S}_{\mathrm{BI}}$ vs. $\widehat{S}_{\mathrm{HO}}$ | $-12.24$ | $7 \times 10^{-17}$ | $-6.60$ | $10^{-8}$ | $-5.32$ | $10^{-6}$ |
| $\widehat{S}_{\mathrm{BI}}$ vs. $\widehat{S}_{\mathrm{HBC}}$ | $-25.51$ | $2 \times 10^{-30}$ | $-5.32$ | $10^{-6}$ | $-5.16$ | $2 \times 10^{-6}$ |
| | Ginibre ensemble | | Poisson process | | Thomas process | |

Table 4: Paired $t$-tests for the estimators that assume isotropy: $T$-scores of the $\widehat{\mathrm{iMSE}}$ and the associated $p$-values. The computation is done using `structure_factor`.

each point process, we can again simultaneously reject at level 0.01 the hypotheses that there is no difference in mean between the estimates iMSEs. This confirms the claim that Bartlett's estimator is the most accurate near 0, among estimators that assume isotropy. For reference, Table 5 gives the same summary statistics as Table 3 did for scattering intensity variants. Bartlett's estimator yields MSEs two orders of magnitude smaller than both Hankel transform quadratures, for both Ginibre and Poisson, and a factor 3 for Thomas. Again, the integrated MSE is mostly variance.

| $\widehat{S}$ | $\widehat{\mathrm{iVar}}$ | $\mathrm{CI}[\widehat{\mathrm{iMSE}}]$ | $\widehat{\mathrm{iVar}}$ | $\mathrm{CI}[\widehat{\mathrm{iMSE}}]$ | $\widehat{\mathrm{iVar}}$ | $\mathrm{CI}[\widehat{\mathrm{iMSE}}]$ |
|---|---|---|---|---|---|---|
| $\widehat{S}_{\mathrm{BI}}$ | $3.9 \times 10^{-3}$ | $\mathbf{4.0 \times 10^{-3}} \pm 3 \times 10^{-4}$ | $0.057$ | $\mathbf{0.058} \pm 9 \times 10^{-3}$ | $11.25$ | $\mathbf{11.65} \pm 4.71$ |
| $\widehat{S}_{\mathrm{HO}}$ | $0.37$ | $0.38 \pm 0.09$ | $2.12$ | $2.14 \pm 0.93$ | $43.63$ | $46.70 \pm 18.40$ |
| $\widehat{S}_{\mathrm{HBC}}$ | $0.03$ | $0.03 \pm 0.01$ | $2.44$ | $2.45 \pm 1.33$ | $57.62$ | $63.02 \pm 28.62$ |
| | Ginibre ensemble | | Poisson process | | Thomas process | |

Table 5: Sample integrated variance and MSE of all three estimators that assume isotropy, across 50 independent samples from the Ginibre, Poisson, and Thomas point processes. The computation is done using `structure_factor`.

Finally, Figure 12 shows the pointwise average of each estimator, across the $M = 50$ point process realizations. The accuracy of Ogata's quadrature at small wavenumbers is poor. The bias of the two Hankel quadrature estimators is well visible at small $k$ for all point processes. This is likely due to poor estimation of the pair correlation function at
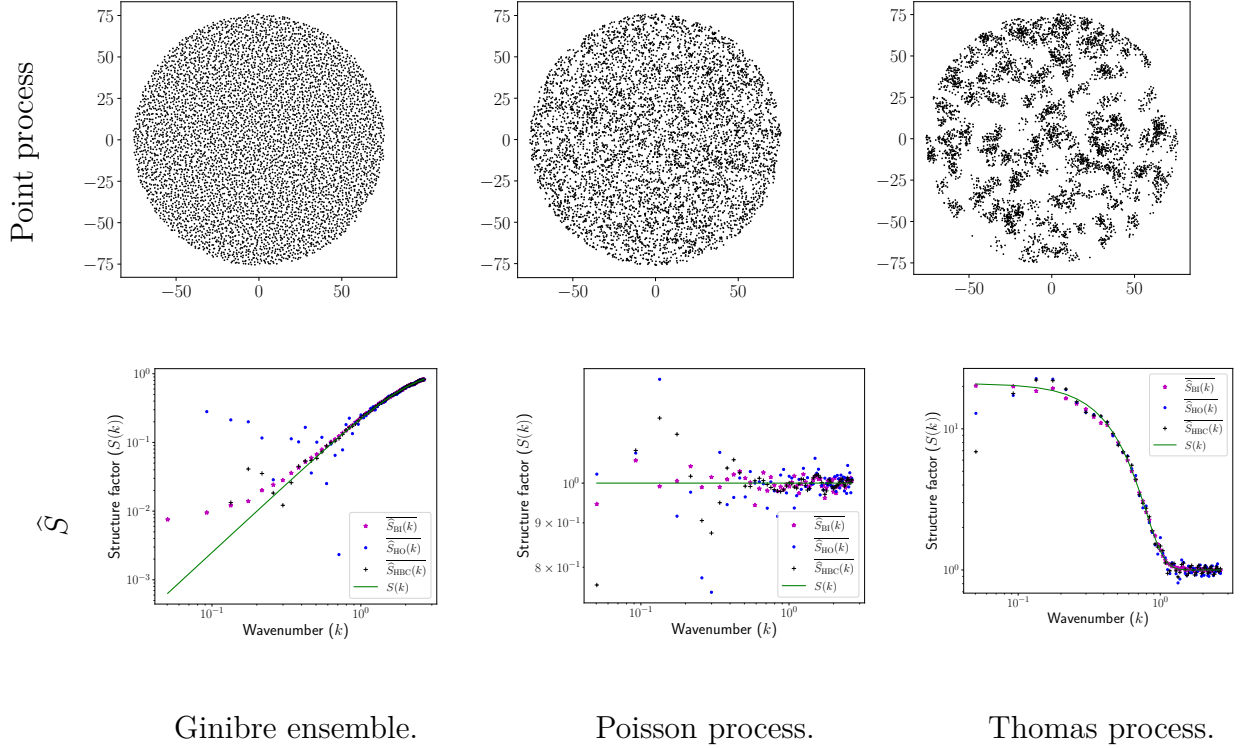
Figure 12: Pointwise averages of the three estimators that assume isotropy, across 50 independent samples of the Ginibre, Poisson, and Thomas point processes. The computation and visualization are done using `structure_factor`.

large scales.

# Summary and discussion

We surveyed estimators of the structure factor of a stationary point process, along with numerical diagnostics of hyperuniformity. We provided an open-source Python toolbox `structure_factor` containing the proposed estimators. The two estimators that fared best in our benchmark are the multitapered estimator of Rajala et al. (2020) and, in the case of isotropic point processes, Bartlett's isotropic estimator. Further comparing both estimators is a natural avenue for further work, but needs careful thinking. For starters, the two estimators are not defined on similar windows, and they do not require the same

number of evaluations for a similar accuracy. It is not clear how to best choose the tapers in the multitaper estimator, not even mentioning that a multitapered version of Bartlett's estimator can also in principle be derived. On the other hand, the computational cost of Bartlett's estimator should be lowered, e.g. by subsampling pairs of points.

Estimators based on Hankel transform quadratures, which rely on first estimating the pair correlation function, comparatively showed poor performance. But it is still possible that estimating the pcf can improve over non-quadrature structure factor estimators if, say, edge correction in pcf estimators plays an important role. If the user only has data collected on a non-rectangular, non-ball window, we would then recommend trying the Hankel-Baddour-Chouinard estimator. For our benchmarks, though, the large cardinalities and the regular windows involved do not build upon this strong point of quadrature-based estimators. It would also be interesting to investigate edge correction, or regularization schemes for the pcf, that are tailored to structure factor estimation.

Graphical hyperuniformity diagnostics allow recovering known information, and can be a useful tool, although summary statistics like the $H$-index of the estimated power decay of the structure factor require manual choices and show large variance. More automated and low-variance estimators would be desireable.

Finally, on the point process side, we hope that our survey and software can contribute to the standardization and reproducibility of empirical investigations involving structure factors, including the study of hyperuniform point processes. For starters, it would be interesting to empirically assess how the decay of the structure factor of the KLY process depends on the initial point process that is matched with $\mathbb{Z}^2$.

# References

G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118. Cambridge university press, 2010a.

G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118. Cambridge university press, 2010b.

A. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns Methodology and Applications with R.* Chapman and Hakk/CRC, 1nd edition, 2016.

N. Baddour and U. Chouinard. Theory and operational rules for the discrete hankel transform. *J. Opt. Soc. Am. A*, 32(4):611–622, Apr 2015. doi: 10.1364/JOSAA.32.000611. URL http://josaa.osa.org/abstract.cfm?URI=josaa-32-4-611.

R. Bardenet and A. Hardy. Monte Carlo with determinantal point processes. *Annals of Applied Probability*, 2020.

R. Bardenet, S. Ghosh, and M. Lin. Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

M. S. Bartlett. The spectral analysis of two-dimensional point processes. *Biometrika*, 51 (3-4):299–311, 12 1964. ISSN 0006-3444. doi: 10.1093/biomet/51.3-4.299. URL https://doi.org/10.1093/biomet/51.3-4.299.

J. Beck. Irregularities of distribution. I. *Acta Mathematica*, 159(none):1 – 49, 1987. doi: 10.1007/BF02392553. URL https://doi.org/10.1007/BF02392553.

A. Belhadji, R. Bardenet, and P. Chainais. Kernel quadrature with determinantal point processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

A. Belhadji, R. Bardenet, and P. Chainais. Kernel interpolation with continuous volume sampling. In *International Conference on Machine Learning (ICML)*, 2020a.

A. Belhadji, R. Bardenet, and P. Chainais. A determinantal point process for column subset selection. *Journal of Machine Learning Research (JMLR)*, 2020b.

C. Biscio and R. Waagepetersen. A general central limit theorem and a subsampling variance estimator for $\alpha$ mixing point processes. *Scandinavian Journal of Statistics*, 46 (4):1168–1190, 2019. ISSN 0303-6898. doi: 10.1111/sjos.12389.

S. Chiu, D. Stoyan, W. Kendall, and J. Mecke. *Stochastic Geometry and Its Applications.* 09 2013. ISBN 9780470664810. doi: 10.1002/9781118658222.

S. Coste. Order, fluctuations, rigidities. *arXiv preprint*, 2021. URL https://scoste.fr/assets/survey_hyperuniformity.pdf.

P. J. Diggle, D. J. Gates, and A. Stibbard. Nonparametric estimator for pairwise-interaction point. *Biometrika*, 74(4), 1987.

M. Guizar-Sicairos and J. C. Gutiérrez-Vega. Computation of quasi-discrete hankel transforms of integer order for propagating optical wave fields. *JOSA A*, 21(1):53–58, 2004.

J. B. Hough, M. Krishnapur, Y. Peres, and B. Virag. *Zeros of Gaussian Analytic Functions and Determinantal Point Processes.* 2009.

M. Klatt, J. Lovrić, D. Chen, C. K. Sebastian, M. S. Fabian, W. A. S. Philipp, S. G. Bruce, A. Smith, E. Gerd, and S. Torquato. Universal hidden order in amorphous cellular geometries. *ature Communications*, 10, 2019.

M. Klatt, G. Last, and D. Yogeshwaran. Hyperuniform and rigid stable matchings. *ArXiv:1810.00265v3*, 2020. URL https://arxiv.org/abs/1810.00265.

J. L. Landau. Bessel functions: Monotonicity and bounds. *Journal of the London Mathematical Society*, 61(1):197–215, 2000. doi: 10.1112/S0024610799008352.

S. Murray and F. Poulin. hankel: A python library for performing simple and accurate hankel transformations. *Journal of Open Source Software*, 4:1397, 05 2019. doi: 10.21105/joss.01397.

H. Ogata. A numerical integration formula based on the bessel functions. *Research Institute for Mathematical Sciences*, 41(4), 2005. doi: 10.2977/prims/1145474602.

P. Osgood. *Lecture Notes for EE 261 the Fourier Transform and Its Applications.* CreateSpace Independent Publishing Platform, 2014. ISBN 9781505614497. URL https://books.google.fr/books?id=omT6rQEACAAJ.

D. B. Percival and A. T. Walden. *Spectral analysis for univariate time series*, volume 51. Cambridge University Press, 2020.

T. Rajala, S. C. Olhede, and D. J. Murrell. Spectral estimation for spatial point patterns. *arXiv: Methodology, Version 1*, 2020.

K. S. Riedel and A. Sidorenko. Minimum bias multiple taper spectral estimation. *Signal Processing, IEEE Transactions on*, 43:188 – 195, 02 1995. doi: 10.1109/78.365298.

D. J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70 (9):1055–1096, 1982.

S. Torquato. Hyperuniform states of matter. *Physics Reports*, 745, 2018. doi: 10.1016/j. physrep.2018.03.001.