Selection functions in astronomical data modeling, with the space density of white dwarfs as a worked example

Rix, H.W.; Hogg, D.W.; Boubert, D.; Brown, A.G.A.; Casey, A.; Drimmel, R.; ... ; Price-Whelan, A.M.

Draft version June 16, 2021
Typeset using LaTeX default style in AASTeX63

Selection Functions in Astronomical Data Modeling, with the Space Density of White Dwarfs as Worked Example.

Hans-Walter Rix,[1] David W. Hogg,[1,2,3] Douglas Boubert,[4] Anthony G.A. Brown,[5] Andrew Casey,[6] Ronald Drimmel,[7] Andrew Everall,[8] Morgan Fouesneau,[1] and Adrian M. Price-Whelan[2]

[1] *Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany*
[2] *Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Ave, New York, NY 10010, USA*
[3] *Center for Cosmology and Particle Physics, Department of Physics, New York University, 726 Broadway, New York, NY 10003, USA*
[4] *Magdalen College, Oxford University, Oxford OX1 4AU, United Kingdom*
[5] *Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, Netherlands*
[6] *School of Physics and Astronomy, Monash University, VIC 3800, Australia*
[7] *INAF - Osservatorio Astrofisico di Torino, Strada Osservatorio 20, Pino Torinese 10025 Torino Italy*
[8] *Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom*

ABSTRACT

Statistical studies of astronomical data sets, in particular of cataloged properties for discrete objects, are central to astrophysics. One cannot model those objects' population properties or incidences without a quantitative understanding of the conditions under which these objects ended up in a catalog or sample, the sample's *selection function*. As systematic and didactic introductions to this topic are scarce in the astrophysical literature, we aim to provide one, addressing generically the following questions: What is a selection function? What arguments $\mathbf{q}$ should a selection function depend on? Over what domain must a selection function be defined? What approximations and simplifications can be made? And, how is a selection function used in 'modelling'? We argue that volume-complete samples, with the volume drastically curtailed by the faintest objects, reflect a highly sub-optimal selection function that needlessly reduces the number of bright and usually rare objects in the sample. We illustrate these points by a worked example, deriving the space density of white dwarfs (WD) in the Galactic neighbourhood as a function of their luminosity and Gaia color, $\Phi_0(M_G, (B-R))$ in [mag^{-2} pc^{-3}]. We construct a sample $\mathcal{C}$ of 10^5 presumed WDs through straightforward selection cuts on the Gaia EDR3 catalog, in magnitude, color, parallax, and astrometric fidelity, $\mathbf{q} = (G, (B-R), \varpi, p_{\rm af})$. We then combine a simple model for Φ_0 with the effective survey volume derived from this selection function $S_{\mathcal{C}}(\mathbf{q})$ to derive a detailed estimate of $\Phi_0(M_G, (B-R))$ is robust against the detailed choices for $S_{\mathcal{C}}(\mathbf{q})$. This resulting white dwarf luminosity-color function $\Phi_0(M_G, (B-R))$ differs dramatically from the initial number density distribution in the luminosity-color plane: by orders of magnitude in density and by four magnitudes in density peak location.

Keywords: Stars: white dwarf stars; Galaxy: disk; Methods: data analysis, statistical; Catalogs

1. SELECTION FUNCTIONS IN ASTRONOMY

Statistical studies of astronomical data sets or catalogs are central to many, if not most, aspects of astrophysics. They usually entail *making a model* of some of the cataloged quantities that characterize (usually discrete) sets of objects, and constraining that model by asking quantitatively whether the data in the catalog match model expectations. This requires that one understands under which circumstances an object would have had a chance to be in the catalog, or in a sub-sample drawn from a catalog. This understanding can be captured in a *selection function* (or *selection probability*). The probability of an object to be in the catalog depends on, for example, the detection efficiency of the observational survey from which the catalog was derived as well as on choices made during the construction of the catalog, such as removing potential entries deemed to be of insufficient 'quality'.

Implicitly, the issue of selection functions in astronomy has been around for as long as there have been astronomical catalogues, with perhaps the first articulations given in classic textbooks such as Trumpler & Weaver (1953). The selection function of a sample is closely related to the effective or maximal survey volume $V_{\rm max}$ of a catalog (a spatial

or a generalized parameter-space volume), a concept introduced quantitatively by Schmidt (1968) in the context of seminal work on quasars. The concept of a selection function has also been linked to the concepts of *selection effect* or *selection bias*; we deem those concepts to be more nebulous, as they have been used both as a synonym for the selection function itself, and for biased results arising from ignoring important aspects of the selection function in an analysis.

The concept of a selection function is very widely used in contemporary astrophysics: ADS lists 700 instances of it in the refereed publications of the year 2020 alone. Yet, didactic expositions of selection functions – definitions, worked examples, best-practices guidance – are hard to find in the astrophysical literature (see Bovy & Rix 2013; Wojno et al. 2017; Bovy 2017; Boubert & Everall 2020; Gaia Collaboration et al. 2020b for notable recent exceptions). This paper aims to fill this gap by providing an exposition of the conceptual and practical issues that arise when devising and applying a selection function.

We will use the Gaia catalog (Gaia Collaboration & et al. 2016; Gaia Collaboration et al. 2020a) – one of the most extensive, multi-dimensional all-sky catalogs of discrete astronomical objects – as a backdrop and input to our worked example. However, we stress that the basic formulation and many aspects of the suggested best-practices should have far broader applicability. Framing these issues in the Gaia context is based on two considerations. First, with precision measurements of 10^6-10^9 objects, statistical analyses of Gaia data will rarely be limited by the sample size (and its Poisson variance), and often not by the individual measurement precision. Instead, modelling will be limited by the precision and the incorporation rigor of the selection function. Second, awareness of the central role of a selection function and of established techniques to implement it in modelling are perhaps not as widespread across all aspects of 'Gaia science' as they are in cosmological large scale structure (e.g. Cole et al. 2005), or in gravitational wave detection.

The remainder of the paper is structured as follows: we aim to summarize selection function basics in Section 2; we illustrate these with a *worked example* in Section 3, deriving the white dwarf (WD) *luminosity–color function* (LCF) from Gaia data, the solar neighbourhood space density of WDs, in pc^{-3}, as a function of their absolute magnitude *and* color. A python notebook for this worked example can be found here. We then lay out a number of further issues that should be considered when applying and deriving selection functions in Section 4; for most, their detailed resolution is beyond the scope of this paper.

Some readers may prefer to start by seeing a concrete example of how to model data including a selection function, before considering the broad guidance in Section 2. We encourage those readers to skip forward to Section 2.4 and then go to Section 3, before returning to the rest of Section 2.

2. SELECTION FUNCTION 'BASICS'

In this Section we give a general introduction to the concept and use of selection functions in astrophysics, addressing: What is a selection function, what are desirable properties for it, what is its role in modeling?

It is a common situation in astronomy that we have a model for the physical properties of discrete objects, say stars, and we want to test this model (or find its best-fit parameters) through a comparison with data. And it is quite likely that one of astronomy's vast catalogs lists observational constraints (fluxes, color, etc.) on such objects; one then selects a pertinent subset of such objects and fits a model to them.

What is a selection function? There are several ways to look at it: One may view the selection function, $S_{\mathcal{C}}(\mathbf{q})$, as the *probability* that an object with attributes $\mathbf{q}$ will be contained within a catalog or sample $\mathcal{C}$ under consideration; we use the subscript $\mathcal{C}$ for the selection function as a constant reminder that it is 'for a given catalog or sample $\mathcal{C}$'. Operationally, a catalog $\mathcal{C}$ in the current context is simply a list that specifies attributes $\mathbf{q}$ for a set of discrete objects. The full set of these catalog attributes can and often will be more extensive than the set of $\mathbf{q}$ that enter the selection function or are being modeled.

The most common use of selection functions is in 'modelling' data sets, based on some model family, $\mathcal{M}(\mathbf{q} \mid \boldsymbol{\Theta}_{\rm mod})$ parameterized by $\boldsymbol{\Theta}_{\rm mod}$. There the selection function may also be viewed as the multiplicative link between the probability density predictions of $\mathcal{M}(\mathbf{q} \mid \boldsymbol{\Theta}_{\rm mod})$ for the quantities $\mathbf{q}$, and the *expected* catalog-incidence (if the model were correct), $\mathrm{d}\Lambda_{\mathcal{C}}(\mathbf{q})$, of these quantities:

$$\mathrm{d}\Lambda_{\mathcal{C}}(\mathbf{q}) = \mathcal{M}(\mathbf{q} \mid \boldsymbol{\Theta}_{\rm mod})\ S_{\mathcal{C}}(\mathbf{q})\ \mathrm{d}\mathbf{q}. \tag{1}$$

Note that while the model prediction is a probability density with units 'per d$\mathbf{q}$', the selection function is simply a unitless probability (function), bounded between zero and unity[1]. We will lay out below the question of which arguments $\mathbf{q}$ of the selection function are suitable and necessary.

When does one need a selection function? Broadly, we need to determine and apply a selection function whenever we want to answer a question or constrain a model through data comparison, and when that model predicts densities, rates, or other incidences for objects with certain characteristics (that are reflected in 'observables'). Note that selection functions are not only needed when analyzing large sets of catalog entries, but just as much when an extensive search for elusive objects has yielded one specimen (or even none); after all, the predicted catalog incidences for a physical model in Eq.1 can well be $\Lambda_{\mathcal{C}}(\mathbf{q}) < 1$.

2.1. *How does one construct a selection function?*

It is easy to state that all one needs for stringent modelling of objects in $\mathcal{C}$ is a sensible model family $M(\mathbf{q} \mid \boldsymbol{\Theta}_{\rm mod})$ and a selection function $S_{\mathcal{C}}(\mathbf{q})$ in the sense of Eq. 1; but this does not address how to devise a good selection function and its resulting sample $\mathcal{C}$.

In the context of large, contemporary astronomical data catalogs (Gaia, PanSTARRS, 2MASS, WISE, GALEX, eROSITA, ...) it is rare that anyone aims to build a model that tries to constrain the physical properties of all objects in the entire *parent* catalog at once: usually such catalogs encompass objects of very different physical natures, from say AGN to White Dwarfs, for which it makes little sense to build a model simultaneously. Instead it is most common to model only subsets of objects from the parent catalog, constructed foremost by cuts or selections in properties $\mathbf{q}$ or in aspects of 'data quality'.

For most applications it therefore makes sense to think of the construction of a selection function as consisting of two parts: the first is to characterize the *detection efficiency* of the underlying experiment and the resulting *completeness* of the parent catalog; the second is the definition of the sub-sample to be modelled through selection on cataloged properties $\mathbf{q}$. Figure 1 presents a schematic of this multi-step process towards a 'final' or 'total' selection function for modelling, which we can use as a guide throughout. Any step in Figure 1 towards constructing a selection function reflects inevitably a number of choices: in the experimental design leading to the parent catalog, or scientific choices in selecting a sub-sample from it for the astrophysical problem at hand. Therefore, it may be a formidable task to construct $S_{\mathcal{C}}(\mathbf{q})$ and understand its fidelity.

Figure 1 makes explicit that the parent catalog is in practice derived from a survey through some form of processing to turn raw data into the catalog quantities $\mathbf{q}$. The survey also has a selection function which is determined by the sky coverage, the sensitivity of the telescope-instrument combination, wavelength range, etc. In practice most scientific analyses start from the catalog entries $\mathbf{q}$, and in the rest of this paper we only discuss the selection function of the parent catalog. We assume that this selection function implicitly accounts for the survey selection function, where for example the sensitivity limit, in combination with a minimum signal to noise ratio needed for the data processing, is translated to a magnitude limit for the parent catalog.

We start by discussing the inherent 'completeness' $S^{\rm parent}(\mathbf{q})$ of the overall parent catalog (e.g. Boubert & Everall 2020, for Gaia DR2), quantifying the probability that a source with observable characteristics $\mathbf{q}$ is included in the parent catalogue. There are two separate (but non-exclusive) paths to determine $S^{\rm parent}(\mathbf{q})$. The most straightforward path is through knowing some 'ground truth', a complete and sufficiently extensive set of objects whose properties we know from external empirical information; to be useful the $\mathbf{q}$ of the ground-truth sample must be such that some will and some won't end up in the parent catalog. Such information can come e.g. from deeper or higher resolution data over restricted survey areas. The parent catalog completeness is then constrained by asking which $S^{\rm parent}(\mathbf{q})$ makes the actual catalog entries among the ground-truth sample a likely outcome, as a function of $\mathbf{q}$. But very often sufficient ground truth is not known. Then $S^{\rm parent}(\mathbf{q})$ must be constructed from an understanding of the overall experiment and of the data processing that leads to catalog entries. In practice, implementing such an approach rigorously requires considerable effort. For example, hardly any survey is simply flux-limited. In the case of the Gaia mission, $S^{\rm parent}(\mathbf{q})$ depends in complex ways on sky position, both because of Gaia's scanning law and because of source crowding. But both aspects are known and can be accounted for, and indeed a parent catalog selection function for Gaia has been derived by Boubert et al. (2020, 2021) using Binomial statistics to infer the detection efficiency (Boubert & Everall

[1] The upper bound of 1 may in rare cases be exceeded, if a catalog construction leads to a finite probability that one object leads to multiple (unlinked) catalog entries.

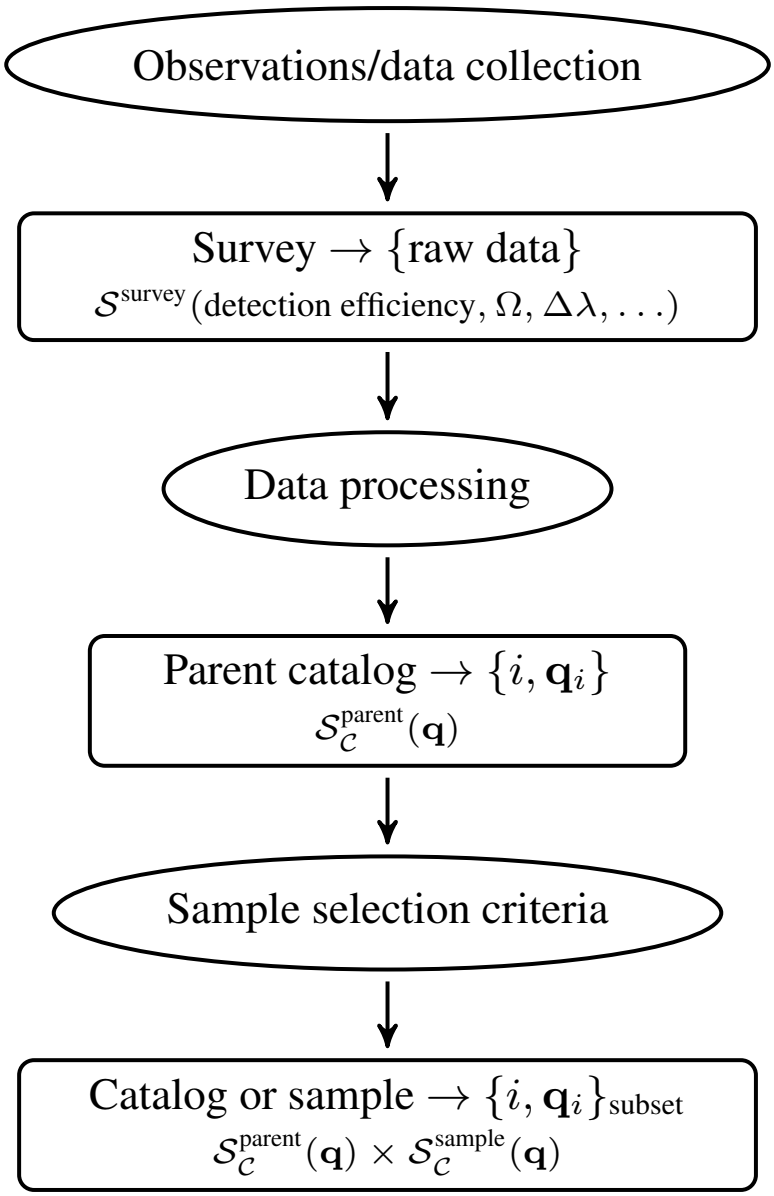

Figure 1. Schematic of all the factors that sequentially set the selection function $S_{\mathcal{C}}(\mathbf{q})$ of a sample of discrete astrophysical objects to be modelled. $S_{\mathcal{C}}(\mathbf{q})$ describes the probability that such an object with observable properties $\mathbf{q}$ will enter a sample. Viewed end-to-end, this starts with the overall experiment (say, the Gaia mission) and its *detection efficiency* of astrophysical sources, $S^{\rm survey}$. After data processing, this results in a parent catalog (say, the Gaia EDR3 catalog), whose *completeness*, $S_{\mathcal{C}}^{\rm parent}(\mathbf{q})$ must be characterized. Astrophysical models for some class of objects can then be constrained by comparison with catalog data. But in practice only a (often tiny) subset of objects in the entire parent catalog will be modelled. Commonly, these are objects that represent some particular class of astrophysical objects, say white dwarfs, or QSO, etc. Such sub-samples are typically defined through a set of selection 'cuts', $S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$. In the end, the individual selection function factors are multiplied and summarised in the overall *selection function*, $S_{\mathcal{C}}(\mathbf{q})$. In practice, the parent catalog completeness, $S_{\mathcal{C}}^{\rm parent}(\mathbf{q})$ is often the same 'given' across many modelling applications, while the sample selection $S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$ will always be tailored to the astrophysical case at hand.

2020). In that application, the resulting $S^{\rm parent}(\mathbf{q})$ made no assumption about the ground truth but solely relied on an understanding of the experiment.

While deriving $S^{\rm parent}(\mathbf{q})$ may be hard, it can then be used unchanged for basically all science analyses based on this parent catalog. And it is common that subsequent sample cuts $S^{\rm sample}_{\mathcal{C}}(\mathbf{q})$ keep only $\mathbf{q}$ that stay well away from the survey's detection limit, where the approximation $S^{\rm parent}(\mathbf{q}) = \text{const.} \approx 1$ is sensible.

We now turn our attention to the second step (see Fig. 1), present in almost all astrophysical modelling of catalog data: devising a sub-sample $\mathcal{C}$ from the parent catalog that encompasses only the objects to be compared to the model, through a suitable choice of $S^{\rm sample}_{\mathcal{C}}(\mathbf{q})$. Generally, $S^{\rm sample}_{\mathcal{C}}(\mathbf{q})$, will select $\mathcal{C}$ only on the basis of a subset of the parent catalog's attributes, such as sky position, $\mathbf{x}_{\rm sky}$, magnitude, color, parallax (ϖ), etc... . But it is also common and sensible to include *signal-to-noise* of attribute estimates (such as $\frac{\varpi}{\sigma_\varpi}$) or 'error flags' among those $\mathbf{q}$, as we will discuss below. In general, the final selection function of a selected catalogue $\mathcal{C}$ from such multi-step procedures can be captured as the result of multiplications:

$$S_{\mathcal{C}}(\mathbf{q}) = S^{\rm parent}_{\mathcal{C}}(\mathbf{q}) \;\cdot\; S^{\rm sample}_{\mathcal{C}}(\mathbf{q}), \tag{2}$$

as per Figure 1. In this context the terminology 'selection cut' does not imply that $S^{\rm sample}_{\mathcal{C}}(\mathbf{q})$ is either 1 or 0. A probabilistic but controlled sub-sampling of the full parent catalog within the domain where $S^{\rm sample}_{\mathcal{C}}(\mathbf{q})$ has support may well lead to a perfectly well-behaved selection function. If one chooses to model a random fraction f of objects that satisfy $S_{\mathcal{C}}(\mathbf{q})$ one simply has to use $S'_{\mathcal{C}}(\mathbf{q}) \equiv f \cdot S_{\mathcal{C}}(\mathbf{q})$

2.2. *What should be the arguments of a selection function?*

Generally, the selection function should be a function of the minimal set of object attributes $\mathbf{q}$ that describes any object's probability to enter $\mathcal{C}$ with sufficient precision and accuracy (which depends on the science question). This broad conjecture deserves some elaboration and qualification:

• For almost all modelling – involving integrals over 'parameter space' – it is important that the selection function describes such probabilities for both actual objects and for arbitrary, or counter-factual, objects that may have postulated attributes $\mathbf{q}$ not represented in the $\mathcal{C}$. For instance, the selection function of a magnitude-limited survey must return 0 for any counter-factual object fainter than the magnitude-limit, of which there are none in the catalogue. More generally, we must know the selection function beyond the domain in $\mathbf{q}$ where we actually have cataloged objects.

• Formally, $S_{\mathcal{C}}$ may be a function of more arguments, say of the full set of cataloged attributes $\mathbf{q}_C \equiv (\mathbf{q}, \mathbf{q}_{\rm irr})$, where the set of irrelevant attributes $\mathbf{q}_{\rm irr}$ is (quasi-)defined by $S^{\rm full}_{\mathcal{C}}(\mathbf{q}_C) = S_{\mathcal{C}}(\mathbf{q})\, S^{\rm irr}_{\mathcal{C}}(\mathbf{q}_{\rm irr})$ with $S^{\rm irr}_{\mathcal{C}}(\mathbf{q}_{\rm irr}) \equiv 1 \,\forall\, \mathbf{q}_{\rm irr}$. Note that $S_{\mathcal{C}}$ is and remains unitless, irrespective of the number of arguments. In photometric sample selection, magnitudes and colors are manifestly selection-relevant $\mathbf{q}$, while photometric selection may be 'blind' to radial velocities v_{rad} or proper motions $\boldsymbol{\mu}$ (if not too large), even if they were listed in the catalog; then v_{rad} and $\boldsymbol{\mu}$ would be part of $\mathbf{q}_{\rm irr}$.

• If possible, the $\mathbf{q}$ should represent simple 'observables'; that is, the $\mathbf{q}$ should be quantities such as sky positions, magnitudes, colors or parallaxes, rather than model-derived quantities such as *intrinsic colors*, *ages*, *bolometric luminosities*, *temperatures* or *distances*. However, if one looks closely, one realizes that the distinction between observables and model-dependent quantities is not well defined: technically, every catalog entry reflects some form of (data) model for a true observable.

So, the following guidance may be more practicable:

The selection function should depend on (a minimal set of) catalog attributes that can be predicted by the model and causally determine the catalog membership probability to sufficient precision.

Model predictability matters: for example, young stellar objects can be effectively selected by their optical variability (in conjunction with other criteria). Yet, only a subset of young stellar objects vary and there is no good model that describes which vary and by how much, making quantitative modelling (say, of the age-mass distribution) of variability-selected young stellar objects complex or even impossible.

There is a subset of common catalog attributes that deserve special attention in this context: *data quality flags* and *signal-to-noise cuts*, or cuts on the quoted measurement uncertainties (not the mean estimates).

2.3. *Data quality cuts as part of the selection function*

If one wants to model a set of *objects of interest*, say QSOs or WDs, it is often sensible to apply – beyond the initial selection – a number of cuts that eliminate spurious measurements, physical contaminants[2], and assure high

[2] By physical contaminants we mean astrophysical objects of a different nature than the objects of interest one seeks to mode, yet which project into nearly the same region of $\mathbf{q}$ used for initial sample cuts

data quality:

Selection function cuts to eliminate spurious measurements: It is often indispensable to eliminate objects from the parent catalog whose $\mathbf{q}$ are not to be trusted – commonly by means of data quality flags – for two reasons: first, almost all modelling assumes that one understands the precision of the $\mathbf{q}$ in the catalog, i.e. the $\sigma_{\mathbf{q}}$ must not be spurious. Second and related, the sample contamination from spurious objects, scattered to their seeming $\mathbf{q}$ as a result of poor attribute estimates, must be small; else it must be explicitly modelled. Such selection cuts, effectively terms in $S_{\mathcal{C}}(\mathbf{q})$ that depend on a catalog's data quality flags, should foremost 'clean' the sample; but one must check to which extent they affect the sample completeness, thereby altering $S_{\mathcal{C}}(\mathbf{q})$. Such a check must be done empirically by applying an analogous selection cut on data quality flags to a sample known externally to be *bona fide* within the intended $\mathbf{q}$ regime. To the extent that data quality cuts do not affect the sample's completeness, they need not be treated in $S_{\mathcal{C}}(\mathbf{q})$ in the subsequent modelling (Eq. 1).

Cuts to reduce 'physical contaminants': It is also often indispensable to apply additional cuts to the parent catalog, merely to separate the objects of interest from other classes of objects, 'physical contaminants', with similar $\mathbf{q}$ used in the selection so far; such additional selection cuts also aim to boost the *purity* of the sample. As above, such cuts ideally reduce only the number of contaminants, leaving the set of objects of interest untouched. In this limit, again the selection function would remain unchanged. As before, such cuts aim to make a simpler or better model for the remaining data; one may not need an elaborate model for the sample contaminants. The difference to the elimination of spurious objects is that the cuts to eliminate physical contaminants usually involve new and discriminating observables among the $\mathbf{q}$, not data-quality flags. In good but realistic cases, such cuts can dramatically reduce the contamination while only eliminating a small fraction of the objects of interest. As above, the fraction of removed objects of interest must be determined as a function of the $\mathbf{q}$, as they will lower $S_{\mathcal{C}}(\mathbf{q})$. We will give a specific example of such cuts in Section 3.2.

Signal-to-Noise Cuts: Signal-to-noise (S/N) cuts may be advisable for a number of reasons: often, the parent catalog has a vast set of catalog entries with marginal S/N in some attribute among the $\mathbf{q}$. From a purely mathematical perspective, there is no cogent reason to eliminate such entries from consideration. Yet, modelling them requires an increasing, and often problematic, reliance (towards small S/N) on the precise and accurate estimate of catalog uncertainties. And it requires much more careful and explicit differentiation between the 'true' $\mathbf{q}$ and the 'cataloged' $\mathbf{q}$. Given that $S_{\mathcal{C}}(\mathbf{q})$ is a function of cataloged attributes, one must model which objects are being scattered in and out of the sample by $\mathbf{q}_{obs} \neq \mathbf{q}_{true}$(see Frankel et al. 2018; Everall & Das 2020 and Section 4).

In light of this, how should S/N cuts, such as $\frac{\varpi}{\sigma_\varpi} > \overline{\mathrm{S/N}}_{\varpi,\mathrm{min}}$ be accounted for? In principle there are two options: We can include S/N attributes among the arguments of $S_{\mathcal{C}}(\mathbf{q})$; but this then requires that the model $\mathcal{M}(\mathbf{q} \mid \boldsymbol{\Theta}_{\mathrm{mod}})$ also predicts the uncertainties among the catalog attributes (e.g. predict both ϖ and σ_ϖ). Or, almost equivalently, one can express the *expected* S/N in term of other catalog 'observables'. We focus on this latter approach in our worked example, where we make a cut in the *expected* parallax S/N, $\overline{\mathrm{S/N}}_\varpi$, which we express as a function of G and ϖ; we do this because this straightforward approach may deserve wider use.

Taking a selection by $\frac{\varpi}{\sigma_\varpi}$ from the Gaia EDR3 catalog as an example, we now spell out the math of converting a S/N selection to one in terms of observables, as it does not seem to be documented in the literature. We start with the simple scaling that reflects how $\frac{\varpi}{\sigma_\varpi}$ varies with parallax and the ability to centroid a point source:

$$\overline{\mathrm{S/N}}_\varpi\,(\varpi, G) \sim \varpi \;\cdot\; \sqrt{\mathrm{Flux}(G)}.$$

One can then obtain a simple expression for the minimal parallax (maximal distance) at a given G where the expected parallax S/N should exceed a threshold $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}}$:

$$\varpi \geq \overline{\mathrm{S/N}}_{\varpi,\mathrm{min}} \;\cdot\; 10^{\frac{G-G^{\mathrm{r}}}{5}}. \tag{3}$$

The reference magnitude G^{r} in Eq. 3 can be derived from first principles, or scaled empirically: for Gaia EDR3 one finds $G^{\mathrm{r}} \approx 22$. Of course, it is likely, and in the case of Gaia known, that σ_ϖ, and by extension $\overline{S/N_\varpi}$ and G^{r}, vary distinctly with position on the sky (Lindegren et al. 2020; Everall et al. 2021). In that case, the condition of Eq. 3 can simply become position-dependent if the level of accuracy is to be boosted.

So, we can recapitulate the simple upshot of this example: when one aims to implement signal-to-noise selection criteria on one particular component of $\mathbf{q}_i$, one could make the selection function an explicit function of measurement

uncertainties (which would need to be modelled). However, it will often be preferable to keep the selection function simple (have fewer arguments) by instead applying selection cuts in observables among the $\mathbf{q}$ that amount to cuts in the *expected* S/N in that $\mathbf{q}_i$. But in general, sample cuts on S/N are perfectly legitimate, and often advantageous, if reflected correctly in $S_{\mathcal{C}}(\mathbf{q})$.

2.4. *Implementing the selection function in data - model comparisons*

In its most general form, the role of the selection function in modelling is summarized in Eq. 1, a simple multiplicative function of the $\mathbf{q}$. However, in many cases we want to compare only a few physical quantities between model and the data in the sub-sample, not the entire vector $\mathbf{q}$. Indeed, quite often the selection function may depend on components of $\mathbf{q}$ that we do not want to model; then we need to marginalize out these 'nuisance parameters'.

To make this concrete, we consider the example that we will work out in detail in the next Section: we want to learn about the luminosity-color function (LCF) of white dwarfs at the Sun's location in the Galaxy, $\Phi_0(M, c)$, which is the number density of WDs that have absolute magnitude and color (M, c), a quantity with units [mag^{-2} pc^{-3}]. This density can of course vary with position, $\mathbf{x}$. We can write such a model, specified by parameters $\boldsymbol{\Theta}_{\rm mod}$, as

$$\Phi(M, c, \mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod}) = \Phi_0(M, c \mid \boldsymbol{\Theta}_{\rm mod})\, \hat{n}(\mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod}), \tag{4}$$

where we have assumed that the spatial variation $\hat{n}(\mathbf{x})$ separates multiplicatively. This may be convenient if our science interest is focussed on $\Phi_0(M, c)$ not on the dimensionless $\hat{n}(\mathbf{x})$, which may be approximately known or be just less interesting; in almost all analogous modelling cases the actual positions of sources such as WDs $\{\mathbf{x}_i\}$ are of little astrophysical interest. Yet, they are crucial arguments in the selection function that reflects limits in sky coverage and distance (and implicitly the apparent magnitude). Therefore constraining $\Phi_0(M, c)$ requires us to marginalize over $\mathbf{x}$. The need for such marginalization is common, and must be accounted for in the selection function.

Given a model that is phrased in terms of *physical* quantities, here M, c and 3D-position $\mathbf{x}$, Eq. 1 becomes

$$\mathrm{d}\Lambda_s(M, c, \mathbf{x}) = \Phi(M, c, \mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod})\, S_{\mathcal{C}}(\mathbf{q}(M, c, \mathbf{x}))\, \mathrm{d}M\mathrm{d}c\, \mathrm{d}V\ . \tag{5}$$

We now want to link this to the actual numbers of WDs with (M, c) in our chosen sub-sample (e.g. illustrated in Fig. 5 below). For this we need to predict the expected number of entries in the *whole* sample (per $\mathrm{d}M \cdot \mathrm{d}c$) through volume-integration in Eq. 5:

$$\Lambda_{\mathcal{C}}(M, c) = \Phi_0(M, c \mid \boldsymbol{\Theta}_{\rm mod}) \int \hat{n}(\mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod})\ S_{\mathcal{C}}(\mathbf{q}(M, c, \mathbf{x}))\ \mathrm{d}^3\mathbf{x}\ , \tag{6}$$

with $\Phi_0(M, c \mid \boldsymbol{\Theta}_{\rm mod})$ and $\hat{n}(\mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod})$ from Eq. 4; for compactness of notation we have dropped the explicit differentials $\mathrm{d}\Lambda$ and $\mathrm{d}M, \mathrm{d}c$ for the remainder of the paper.

Note that if one wanted to constrain the spatial distribution of the tracers from the data at hand, rather than incorporate the presumed-to-be-known information about it in the modelling, we would of course not marginalize over the three spatial dimensions, but retain and model $\Lambda_s(M, c, \mathbf{x})$.

So, while the selection function appears initially as a multiplicative factor in the model prediction (Eq. 1), in practice some or all of its dimensions are subject to marginalization integrals. This integration is over the quantities $\mathbf{q}$ that matter for the selection function but are mere nuisance parameters for the model.

2.5. *Is Sample Completeness Important?*

It is often deemed the holy grail of sample design to be *complete* with respect to some properties: 'our parent sample contains basically all point sources in the sky with G brighter than X magnitudes'; 'our sample contains all stars of type X within Y parsec of the Sun', etc.. Sample completeness in this sense has an immediate visceral and intellectual appeal, with completeness to a magnitude limit or volume-completeness being perhaps the most common desiderata.

But the merit of completeness in constraining models is much more nuanced: in many circumstances completeness is nice, but not necessary; in many other circumstances striving for completeness forces compromises in the sample design that lead to highly sub-optimal answers for the science questions of interest. We illustrate this crucial point here briefly and qualitatively, with a more quantitative underpinning in the worked WD example.

Completeness, nice but not necessary: We know that the Gaia EDR3 catalog – when averaged across the sky – is nearly complete ($S \geq 0.95$) for magnitudes $12 < G < 19$ (Boubert & Everall 2020; Gaia Collaboration et al.

2020b). This is of course a fundamental piece of information to determine the incidence of astronomical phenomena and sources. But let us imagine we model a random 2/3 of the sources in the Gaia catalog, instead of the whole catalog. Then the sample is plenty large enough in most of the cases to do the model fitting; and if this 2/3 sub-sampling is properly reflected in the selection function and the modelling, we will get an identical modelling result. *Knowing the level of completeness is far more important than being 'complete'!* If there is no difference in the effort or resources needed to analyse the fully complete sample, there is no harm in it. But often – think of spectroscopic follow-up of a photometric catalog – striving for $S \equiv 1$ completeness implies an enormous additional resource effort; then the need for completeness must be justified by the necessity for answering the science question of interest.

'Complete' samples are often very sub-optimal: Let us presume we want to estimate the space density of different objects (as in Section 3) that span a wide range in luminosities; and let's presume the common case that intrinsically faint objects are more numerous than more luminous objects. If we then want to construct a volume limited sample from some flux-limited parent catalog[3] the maximal volume is set by the distance at which the least luminous objects fall below the basic flux limit of the parent catalog (eg. Hollands et al. 2018; McCleery et al. 2020). The more luminous and more rare objects can be found in the parent catalog across much larger volumes. Yet, they get discarded for the sake of volume-completeness: 10-fold more luminous objects can be (naively) seen across $10^{3/2} \approx 30$-fold larger volumes, meaning that 97% of them in the parent catalog get discarded from the volume complete sample. If luminous objects are very rare, such stringent cuts to achieve volume-completeness may even leave them without any representation in the sample. As we will show in the next Section, there are selection function choices that are mathematically just as simple, provide unbiased model estimates, but can draw on much larger samples. And this situation is quite generic: *Striving for (e.g. volume-) completeness just for its conceptual appeal may greatly increase the effort needed to get suitable data, or – at a given data quality – gravely limit the quality of the subsequent modelling!*

3. A WORKED EXAMPLE: THE COLOR-LUMINOSITY FUNCTION OF WHITE DWARFS

Following on these general considerations, we now turn to a worked example to flesh-out and illustrate the points above. For this we choose the *luminosity-color function* (LCF) of White Dwarfs in the Galactic disk at $R_\odot$: their space density $\Phi_0(M, c)$ as a function of absolute magnitude M and color c. This is for a number of reasons: the example is of astrophysical interest; it can be implemented in a highly simplified form, with results differing (instructively) by orders of magnitude from the 'naive' plotting of the face-value color-absolute magnitude diagram (CAMD); it allows us to illustrate almost all aspects from the above Section, and it can illustrate how much or how little impact sensible, but in detail arbitrary, choices in the sample selection make.

The distribution of WDs in mass-age space has long been recognized as a powerful diagnostic of both stellar physics and Galactic archaeology (Wood 1992; Fontaine et al. 2001). Their mass distribution reflects the distribution of predecessor masses, in combination with the initial-to-final mass ratio (Weidemann 2000; El-Badry et al. 2018). Their distribution in luminosity at a given mass reflects both the birth-rate of such objects and their cooling histories. This is reflected in the *observable* luminosity-color function, as those two observable quantities reflect a combination of WD mass and cooling age.

The spectacular data from Gaia DR2 (Gaia Collaboration et al. 2018; Gentile Fusillo et al. 2019) immediately revealed how intricate the CAMD of WDs is in Gaia's M_G vs. $(B-R)$ space:[4] the WD distribution shows two branches in the CAMD around $(B-R) \approx 0.2$, and the distribution shows a ridge across most colors at low luminosities, presumably related to the energy release following crystallization (Cheng et al. 2019; Tremblay et al. 2019). This distribution can also be seen in Figure 10 at the end of this paper. The detailed physical interpretation of this distribution's morphology is not yet settled (Brown 2021) and is beyond the scope of this paper.

Between the most luminous and dimmest WDs in typical Gaia-derived samples (e.g. Gentile Fusillo et al. 2019) there are about 10 magnitudes, or a factor of 10,000 in luminosity. This implies that the volume across which WDs remain bright enough to enter a magnitude-limited sample varies by *many* orders of magnitude[5].

This vast range in effective WD survey volume must be accounted for. And, indeed has recently been done in determining the WD 'luminosity function' (Gaia Collaboration et al. 2020b), which is the LCF integrated across the full color range at a given luminosity. Yet, the intricacy and information-richness of the M_G *vs.* $(B-R)$ of WDs suggest that such an exercise should be generalized to retain the color-structure. We work this out here as an

[3] 'Volume-limited' in the sense of: nearly all objects of interest across all luminosities in this volume.

[4] Throughout this text we refer to the $(G_{\rm BP} - G_{\rm RP})$ color defined from the Gaia BP and RP bands as $(B-R)$.

[5] This situation is perfectly analogous to any galaxy survey, where luminous galaxies can be seen across vastly larger volumes than dwarf galaxies (e.g. Blanton et al. 2003)

example: determining the WDs LCF, the space-density of WDs as a function of M_G and of color (where we take $(B-R)$); subsequently we take M and c as shorthand for its two arguments. We will remain very cursory on the many astrophysical implications of this analysis in order to retain this paper's focus on the basics of how to devise and apply selection functions.

3.1. *A Model for the Luminosity-Color Function of WDs*

The general LCF model has already been spelled out in Eq. 4. But to actually make predictions, we need to specify functional forms for both the density normalization $\Phi_0(M, c \mid \boldsymbol{\Theta}_{\rm mod})$ and for the dimensionless spatial density distribution $\hat{n}(\mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod})$. For $\Phi_0(M, c \mid \boldsymbol{\Theta}_{\rm mod})$ we face the issue that there is no simple parameterized model that captures the white dwarfs CAMD patterns seen e.g. in Figure 10. Therefore, we adopt a model where $\Phi_0(M, c \mid \boldsymbol{\Theta}_{\rm mod})$ is described by independent top-hat functions within any small (M, c) patch. If we then choose a 120×120 grid in (M, c), we have $14,400$ parameters $\boldsymbol{\Theta}_{\rm mod}$ for $\Phi_0(M, c \mid \boldsymbol{\Theta}_{\rm mod})$. For $\hat{n}(\mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod})$, which is essentially a nuisance parameter in the current context, we will adopt two very simplified functional forms, either a homogenous distribution, or a plane-parallel slab with a vertically Gaussian density profile of scale h_z. Note that we adopt a $\hat{n}(\mathbf{x})$ here, and hence do not fit for parameters in $\hat{n}(\mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod})$; this is just one of the many 'astrophysical choices' faced in model-building.

3.2. *WD Sample Selection*

To constrain this model by confronting it with data, we need to choose a suitable sub-sample of WDs, which we now do. We start with an initial query to the Gaia EDR3 as our parent catalog, designed to yield an initial set of possible WD *candidates* within a few hundred pc around the Sun with reasonably well-measured photometry and astrometry; we then refine this initial selection in a number of steps, resulting in an $S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$ that identifies WDs across their full parameter range with high purity.

Fundamentally, WDs can be selected by their exceptional position in the CAMD, far below the main sequence at colors bluer than $(B-R) \approx 2$. We want to capture WDs at all relevant colors (equivalently, surface temperatures), and in the end we want to eliminate most contaminants (i.e. sources that lie below the main sequence but are manifestly not single WDs). We also want to keep the sample as large as sensible because our model for $\Phi_0(M, c)$ has many parameters.

Following the considerations of the previous Section, we describe the sequence of selection cuts, which form multiplicative terms of the eventual sub-sample selection function $S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$ that can be expressed solely as a function of $\mathbf{q} = (G, (B-R), \varpi)$.

The initial Gaia EDR3 query encapsulates the following aspects:

- We select on Gaia G band apparent magnitude $G < G^{\rm lim}$ mag with a fiducial magnitude limit of $G^{\rm lim} = 20$, so that the approximation that Gaia EDR3 is 'largely complete to that magnitude' is sensible.
- We choose the color range $-0.8 < (B-R) < 2.5$ mag, which entails basically the full color range expected for WDs. We take observed colors and ignore the issue of dust reddening for the time being.
- We apply a selection cut of parallax $\varpi > \varpi_{\rm lim} = 3$ mas, corresponding to a maximum sample extent of 333 pc. While luminous WDs can be seen to greater distances, this choice eliminates the need for sophisticated models of the spatial density distribution of WDs, $\hat{n}(\mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod})$, and of sophisticated treatment of 3D dust extinction. To limit the size of the initial candidate WD sample, we also require `parallax_over_error` > 5, a cut that will be superseeded by more stringent requirements in the subsequent analysis.
- We select objects that lie at least two magnitudes below the main sequence at that color through $M_G(c) > M_{\rm MS}(c) + 2$, with the absolute magnitude estimated as $M_G \equiv G + 5 \log_{10} \frac{\varpi}{100 \text{ mas}}$ (we design a sample where the difference between the true and the estimated absolute magnitude is negligible).

This translates into the following query to the Gaia EDR3 catalog:

```
SELECT *
FROM gaiaedr3.gaia_source
WHERE
    phot_g_mean_mag < 20.0
and bp_rp between -0.8 and 2.5
```

```
and parallax > 3.
and parallax_over_error > 5.
and phot_g_mean_mag +5*log10(parallax/100.) > 4.+ (13/3.3)*(bp_rp+0.8),
```

where the mean main sequence slope $\Delta M_G/\Delta(B-R)$ is adopted to be 13/3.3.

This query results in 737,899 returned Gaia EDR3 entries, whose CAMD is shown in Figure 2, where we again equate $M_G = G + 5\ \log_{10} \frac{\varpi}{100\ \mathrm{mas}}$. This Figure shows both the sequence of presumed WDs, around $((B-R), M_G) \approx (0.2, 12)$, and a dominant set of other sources centered near $((B-R), M_G) \approx (1.4, 13)$. The diagonal right edge of the distribution reflects our well-below-the-main-sequence sample cut. The latter group of sources, $((B-R), M_G) \approx (1.4, 13)$, turns out to be almost entirely spurious in their CAMD position, illustrating the need for sample cleaning.

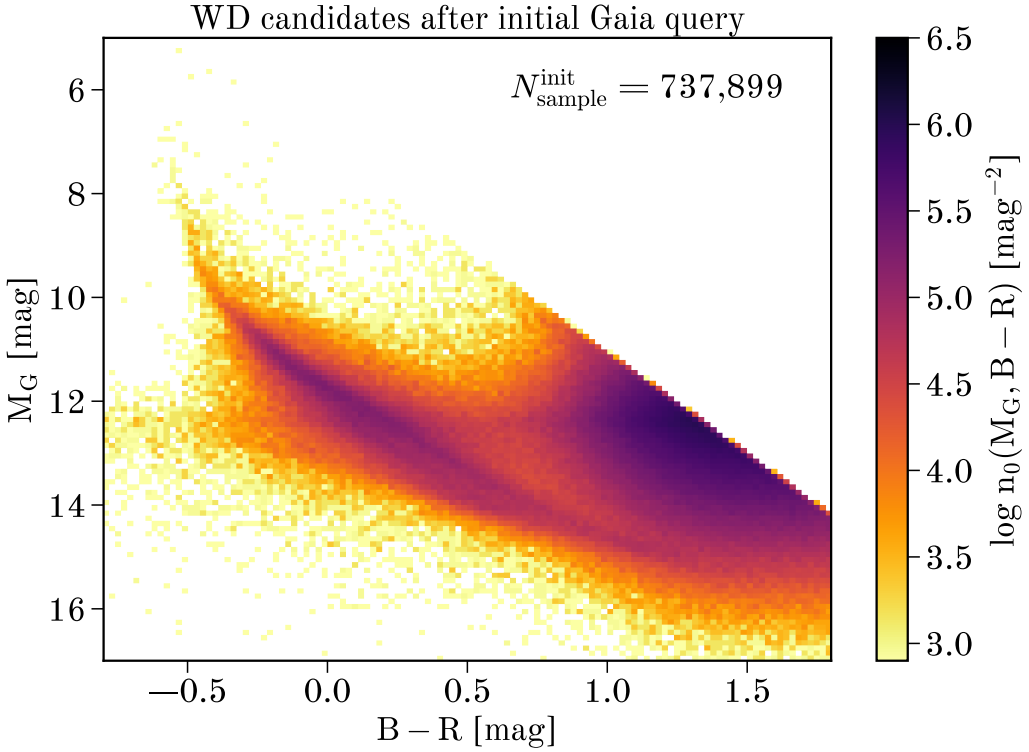

Figure 2. Distribution of objects in the color – absolute magnitude diagram (CAMD, here $((B-R),M_G)$) resulting from the initial sample query. This query selects WDs as it should – the sequence of objects centered on $((B-R), M_G) = (0.2, 12)$ – but the initial candidate sample is overall dominated by objects seemingly near $((B-R), M_G) = (1.4, 13)$; it turns out that almost all of these sources (seem to) lie in this portion of the CAMD because they have spurious astrometry or photometry in EDR3 (see Fig.3). The units of this distribution is number of objects / mag^2.

We now turn to additional cleaning cuts in the sample selection, made either to assure 'data quality', or to achieve sufficient 'astrophysical purity' (see Section 2.3).

• We start with a cut on the fidelity of the astrometric solution of the initial catalog entries. As Figure 2 illustrates, objects that seemingly lie below the main sequence in a CAMD do so for chiefly two reasons: their physical nature indeed places them there; or spurious astrometric (or color) measurements scatter them there in rare instances. These rare instances, however, matter in this part of the CAMD, as it is extremely sparsely filled with *bona fide* objects: even rare spurious measurements may dominate the face-value population, as shown by Gaia Collaboration et al. (2020b) in this context. Therefore, a cut to remove spurious sources is indispensible here. We select on the astrometric fidelity parameter p_{af} recently proposed by Rybizki et al. (2021a), designed and verified to eliminate objects with spurious parallax measurements, specifically $p_{\mathrm{af}} > 0.9$. This removes the dominant fraction (83%!) of all initial candidate

sources as presumably spurious sources contaminating the sample. The impact of this data quality cut is remarkable, as Figure 3 shows: it removes almost all objects seemingly around $((B-R), M_G) \approx (1.4, 13)$.

If this cut removed nothing but spurious objects, it would have *no* impact on the selection function. To check its impact on the objects of interest, we applied this cut to a sample of *bona fide* WDs (spectroscopically verified by Kleinman et al. 2013), and found that this quality cut only removed $\leq 5\%$ of the WDs with $G < 19.5$. In a more stringent analysis than executed here, one could and should empirically calibrate how this cut affects the selection function (as a function of G and $(B-R)$); below we simply neglect this few-percent effect. Selection cuts on data-quality parameters such as these are generic to all analyses of large catalogs.

• We now proceed to apply a cut that is designed to eliminate 'physical' contaminants in this portion of the CAMD very effectively, yet leave the objects of interest (here WDs) essentially unaffected. The most common physical contaminants are presumably binary stars involving a WD and a low-mass main sequence star: these can be detached binaries, or in mass-transfer systems such as CVs, and are expected to be found in the vicinity of around $((B-R), M_G) \approx (1, 12)$, where objects can be seen in Figure 3. Single WDs (or two WDs of the similar $T_{\rm eff}$) form an extremely tight sequence in $(B-G)$ *vs.* $(G-R)$ space, as shown in Figure 4. Yet, almost all contaminating WD-MS binaries have an SED that is the combination of photospheres (or accretion disks) of very different $T_{\rm eff}$, scattering these objects over a wide area in the color-color locus (see left panel of Fig.4). To eliminate these objects, we apply a cut in the color plane, $(G-R) < (G-R)_{\rm lim} = f\,((B-G))$, as indicated in the right panel of Figure 4:

$$(G-R) < 0.48 + 1.15x - x^2 + 0.70x^3 - 0.2x^4 \qquad x \equiv (B-G) - 0.2. \tag{7}$$

This cut eliminates 13,255 sources (about 10%). Most of them are presumed physical contaminants (for WDs as objects of interest) as the resulting CAMD distibution shows (see Fig.5). This selection cut should not eliminate any of the objects of interest, as it encloses the full range $(B-G)-(G-R)$ expected for them. Only WDs with exceptionally poorly measured colors may get eliminated, and we checked that the impact of this cut on the spectroscopically verified WD sample from Kleinman et al. (2013) was at the few percent level.

While we will not model WD colors beyond $((B-R))$, any SED model for single WDs will predict zero probability for colors off the *stellar locus* in the color-color plane; so the right-hand side of Eq.1 would stay unchanged for $\mathbf{q}$ away from this locus. An explicit treatment of such sample-purifying cuts may therefore not be necessary.

• Finally, we turn to the issue of sub-selecting sources with 'sufficiently precise parallax estimates'. As discussed in Section 2.3, we do this through the condition $\overline{\rm S/N}_{\varpi}(\varpi, G) > \overline{\rm S/N}_{\varpi,\rm min}$, as per Eq. 3. For estimating the WD LCF, we choose thresholds of $\overline{\rm S/N}_{\varpi,\rm min} = 5, 10, \mathbf{20}, 40, 80$. Expressing this S/N threshold through $\overline{\rm S/N}_{\varpi,\rm min}(G, \varpi)$ leaves us mostly (but not exclusively) with sample members whose directly determined $\frac{\varpi}{\sigma_\varpi}$ is above the desired threshold; and we gain through this approach that we do not need to expand the arguments of the sub-sample selection function $S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$ beyond $\mathbf{q} = \{G, (B-R), \varpi\}$.

Again, the value of any $\overline{\rm S/N}_{\varpi,\rm min}$ cut is an astrophysical *choice* that can make uncertainties in M_G less important, at the expense of reduced sample size.

To summarize, we create the sub-sample to be modelled from the Gaia EDR3 parent catalog through the following steps:

•**Initial Sample Cuts:** G< $G^{\rm lim} \approx 20$, $\varpi > \varpi_{\rm min} = 3$ mas, $\varpi/\sigma_\varpi > 5$, and sources 'below the main sequence': $G + 5\log_{10}(\varpi/100) > 4 + 3.94\,((B-R) + 0.8)$.

• **Data Quality Cut**: $p_{\rm af} > 0.9$ from Rybizki et al. (2021b) to eliminate spurious astrometry contaminants in this intrinsically sparse part of the CAMD.

• **Physical Contaminant Elimination**: select only sources with SEDs resembling single WDs through a cut in the $(B-G)$ *vs.* $(G-R)$ plane (Figure 4 and Eq.7).

• **Parallax S/N Cuts**: This is implemented through $\overline{\rm S/N}_{\varpi}(\varpi, G) > \overline{\rm S/N}_{\varpi,\rm min}$, following Eq.3, where we take $\overline{\rm S/N}_{\varpi,\rm min} = 20$ as fiducial, but explore other choices.

The full selection function, as used in Eq. 2, consists of the following terms:

$$\begin{aligned} S_{\mathcal{C}}(\mathbf{q}) = S_{\mathcal{C}}^{\rm parent}(\mathbf{q}) \quad & \times S_{\mathcal{C}}^{\rm init.\ query}(G, \varpi, \sigma_\varpi, (B-R)) \times S_{\mathcal{C}}^{\rm data\ quality}(p_{\rm af} \mid \mathbf{q}) \\ & \times S_{\mathcal{C}}^{\rm contam.}((B-G), (G-R)) \times S_{\mathcal{C}}^{\varpi SN}(\overline{\rm S/N}_{\varpi} \mid G, \varpi)\,. \end{aligned} \tag{8}$$

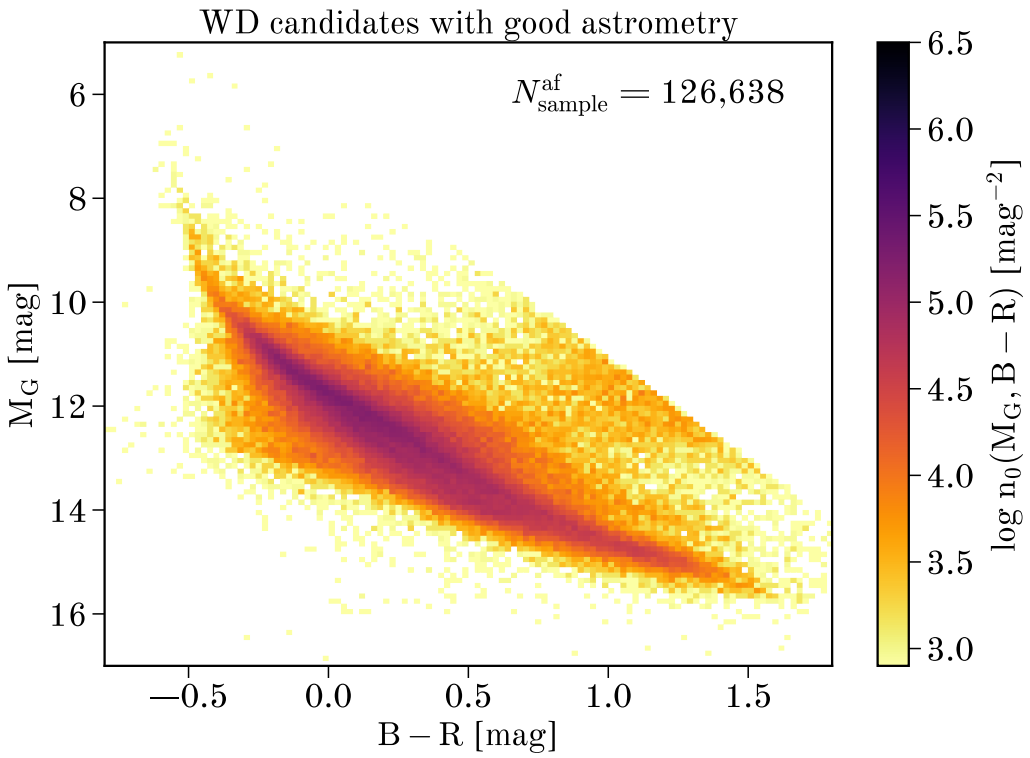

Figure 3. Distribution of objects in the CAMD after removing the (dominant subset of) presumably spurious objects, by requiring an astrometric fidelity of $p_{\rm af} > 0.9$ from Rybizki et al. (2021a). This cut, cast as $S_{\mathcal{C}}^{\rm data\ quality}(p_{\rm af} \mid \mathbf{q})$, eliminates most spurious sources, but leaves $\geq 95\%$ of the spectroscopically confirmed WDs with $G < 19.5$ unaffected.

We refer to each selection function term as a 'cut', as we have implemented all terms in Eq.2 indeed as [0,1] step function reflecting Boolean conditions.

These selection cuts lead to a catalog whose face-value number density distribution is shown in Figure 5, as a function of the key quantities $(M, c) = (M_G, (B-R))$. It is worth iterating that all selection function cuts in Eq. 8 constitute examples of the types of cuts that many other analyses will also perform, whether they spell them out explicitly or not.

For ease in implementation of the subsequent modelling, we will make a number of further approximations whose astrophysical impact should be small. First, we assume that the parent sample is complete across the sky, $S_{\mathcal{C}}^{\rm parent} \approx 1 \,\forall\, \mathbf{x}_{\rm sky}$. This is manifestly not true (e.g. Boubert & Everall 2020), but at magnitudes brighter than $G \leq 20$ it is a sky-averaged approximation that is good at the 5% level. Next, we assume that we do not need to treat the terms $S_{\mathcal{C}}^{\rm data\ quality}(p_{\rm af} \mid \mathbf{q})$ and $S_{\mathcal{C}}^{\rm contam.}((B-G),(G-R))$ explicitly, as these terms cut out almost exclusively spurious measurements and physical contaminants, but not our objects of interest, single WDs. Therefore, these terms are ≈ 1 for all $\mathbf{q}$ where the model makes non-zero predictions. Again, this holds more broadly: selection cuts that only eliminate $\mathbf{q}$-space where $\mathcal{M}(\boldsymbol{\Theta}_{\rm mod}) = 0$ leave the selection function unaffected.

Finally, we will assume that the remaining terms of the selection function are only functions of three variables, $\mathbf{q} = (G, (B-R), \varpi)$. Formally, the initial cut $S_{\mathcal{C}}^{\rm init.\ query}(G, \varpi, \sigma_\varpi, (B-R))$ is a function of σ_ϖ, since it included the criteria $\frac{\varpi}{\sigma_\varpi} > 5$, but subsequent cuts in $\overline{\rm S/N}_{\varpi,\rm min}$ are more stringent, and expressed via $\mathbf{q} = (G, (B-R), \varpi)$.

3.3. *Estimating $\Phi_0(M, c)$ from the Sub-Sample Defined by $S_{\mathcal{C}}(\mathbf{q})$*

We now turn to working out the specifics of constraining our parameters, the 120×120 elements of $\Phi_0(M_G, (B-R))$, through the comparison with the data of this sub-sample, $\{M_G(G, \varpi), (B-R)\}_{i=1,N_{\rm sample}}$. We start with the rate prediction for the catalog entries, $\Lambda_{\mathcal{C}}$ from Eq. 6. This prediction of $\Lambda_{\mathcal{C}}(M_G, (B-R))$ entails a marginalization over 3D space, as the detailed objects positions are not of interest here. As often in astronomy, it is useful to separate such

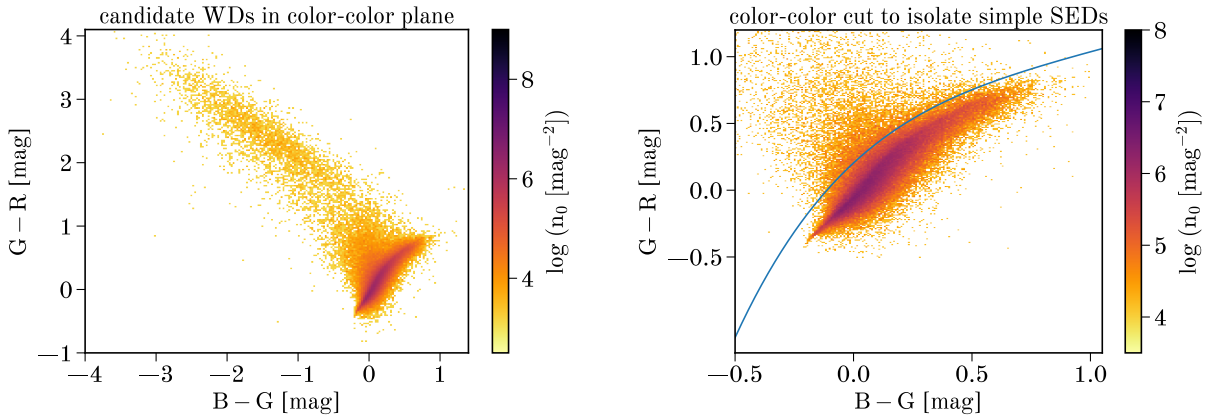

Figure 4. Distribution of the candidate WD sample (after the astrometric fidelity cut) in the $(B-G)$ *vs.* $(G-R)$ color-color plane. The density distribution shows a sharp ridge where our objects of interest are located, objects with SEDs (or colors) of single WDs. The left panel, shows the full color distribution, which exhibits a subset with an enormous spread in colors: most are presumably binaries, involving a WD (possibly with an accretion disk) and a low-mass star. The color-cut, shown in the right panel as the blue line, eliminates most of those, while preserving $\geq 95\%$ of spectroscopically confirmed WDs. This is an example of a sample selection cut that leaves the selection function for the objects of interest essentially unaffected. It simply makes the sample purer, lessening the need to explicitly model contamination.

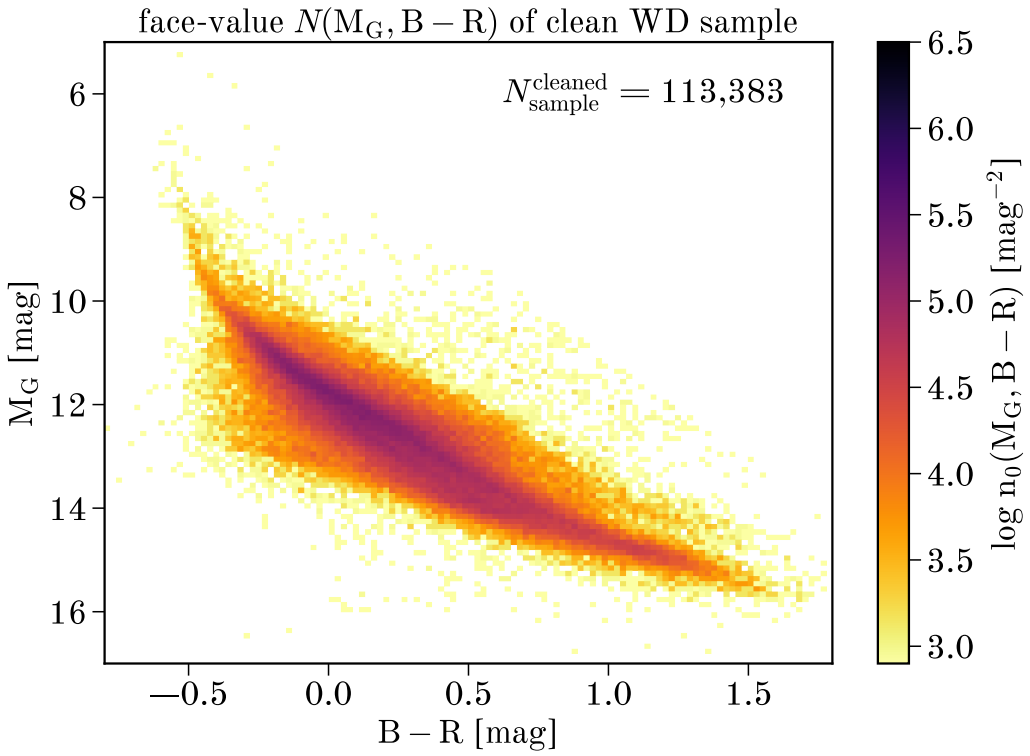

Figure 5. Face-value distribution of the WD sample members in the $N_C(M_G, (B-R))$-plane, after all the selection function cuts summarized in Eq.8 have been applied. The units of this distribution is number of objects / mag^2.

a volume integral in the angular components and the line-of-sight direction:

$$\Lambda_{\mathcal{C}}(M_G,(B-R)) = \Phi_0(M_G,(B-R)) \int_{4\pi} \mathrm{d}\Omega \int_0^\infty d^2\ \mathrm{d}d\ \ \hat{n}(\mathbf{x}_{\rm sky},d)\ S_{\mathcal{C}}\Big(\mathbf{q}(M_G,\mathbf{x}_{\rm sky},d)\Big). \tag{9}$$

From this we obtain with $\mathrm{d}d/\mathrm{d}\varpi = -\varpi^{-2}$

$$\Lambda_{\mathcal{C}}(M_G,(B-R)) = \Phi_0(M_G,(B-R)) \int_{4\pi} \mathrm{d}\Omega \int_{\varpi_{\rm min}(M_G|S_{\mathcal{C}}(\mathbf{q}))}^\infty \varpi^{-4}\mathrm{d}\varpi\ \hat{n}(\mathbf{x}_{\rm sky},d(\varpi)). \tag{10}$$

Note that in this case the impact of the selection function can be entirely subsumed in the lower bound of the last integral, $\varpi_{\rm min}(M_G|S_{\mathcal{C}}(\mathbf{q}))$; this is because we assumed that $S_{\mathcal{C}}(\mathbf{q}) \approx 1$ within these bounds, and zero beyond. Also, we can drop $(B-R)$ as an explicit argument of the selection function, as $S_{\mathcal{C}}(\mathbf{q})$ does not vary significantly with color within the chosen range.

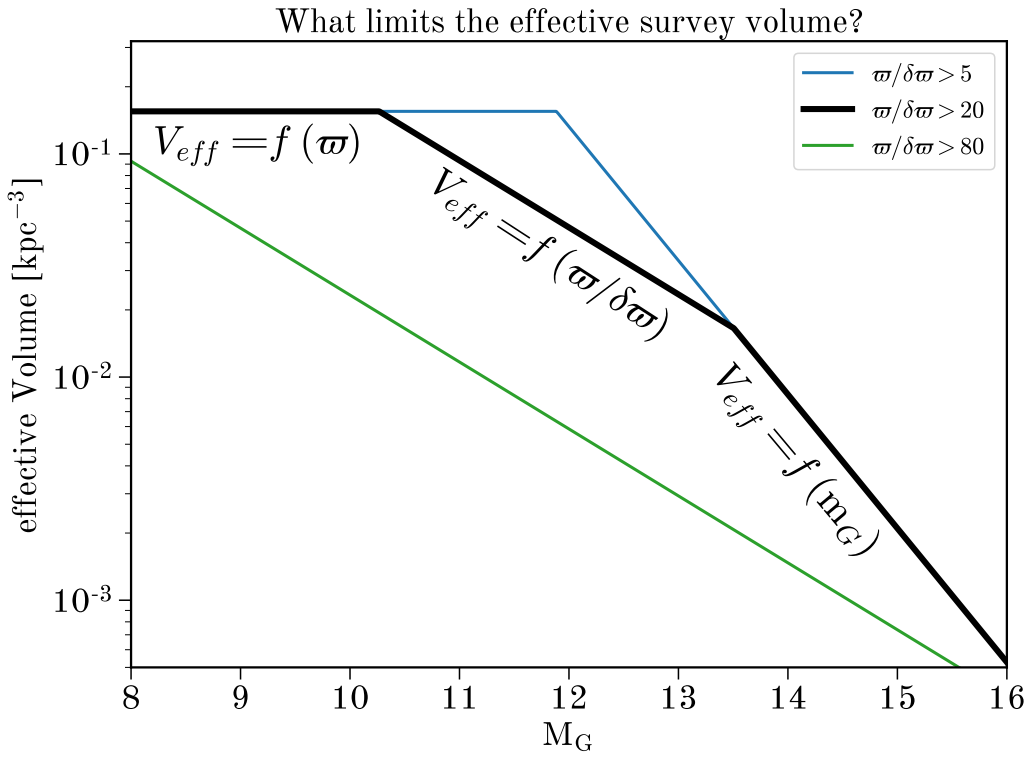

Figure 6. The effective survey volume, $V_{\rm eff}(M_G)$, which for our worked example is only a function of each object's estimated absolute magnitude M_G. The Figure shows three regimes (thick black line) where different terms in $S_{\mathcal{C}}(\mathbf{q})$ limit $V_{\rm eff}(M_G)$: for the most luminous objects ($M_G < 11$) $V_{\rm eff}$ is simply limited by the initial selection $\varpi > 3$ mas; for the least luminous objects in the volume it is limited by the initial cut apparent magnitude, $G < 20$. In the intermediate regime, the volume is limited by the (subsequently) chosen cut in *expected* parallax S/N, $\left\langle \frac{\varpi}{\sigma_\varpi} \right\rangle_{\rm min}$ (Eq.11). For very demanding choices in $\left\langle \frac{\varpi}{\sigma_\varpi} \right\rangle_{\rm min}$ this cut may dominate for all M_G (green line); if such a cut is omitted or very lenient (blue line), this regime may disappear. This Figure can also serve to illustrate why volume-limited samples are generally very sub-optimal: if we wanted to construct a volume-limited sample of WDs covering $7 < M_G < 15$, it would have a volume of only $V_{\rm eff} = 10^{-3}$ kpc^{-3}, and we would have to discard 90% (99%) of the accessible sample-members at $M_G = 13$ (10).

The formulation of Eq. 10 allows of course for angular variations in the selection function $S_{\mathcal{C}}(G(M_G,\varpi),\mathbf{x}_{\rm sky},\varpi)$ and for an arbitrary (but presumed known) density distribution $\hat{n}(\mathbf{x}_{\rm sky},d(\varpi))$. Here we approximate the entire selection function as isotropic, as its angular variation is orders of magnitude less important than its radial one (distance or parallax).

For this example and similar ones[6] the integral of Eq. 10 boils down to calculating a properly weighted effective volume, $V_{\rm eff}$, because one can write the result as $\Lambda_{\mathcal{C}}(M_G,(B-R)) = \Phi_0(M_G,(B-R)) \cdot V_{\rm eff}(M_G)$.

To evaluate this integral, we need to know how the non-zero domain of the selection function translates into the choice of the lower parallax integral boundary, $\varpi_{\rm min}$, in Eq. 10. This limit varies of course among the sample members, as its depends on their G and ϖ. The explicit selection function dependence is here only through $\varpi_{\rm min} = f(M_G(G,\varpi))$. Considering the case at hand and neglecting dust extinction, the condition of assuring a minimal $\overline{\rm S/N}_{\varpi,\rm min}$ (Eq. 3) translates into a minimal parallax of

$$\varpi_{\rm min}(M_G \mid \overline{\rm S/N}_{\varpi,\rm min}) \equiv \sqrt{\overline{\rm S/N}_{\varpi,\rm min}} \cdot 10^{\frac{M_G - G^{\rm r}+10}{10}}. \tag{11}$$

The selection function specified in Eq.8 implies that the maximal distance $\varpi_{\rm min}$ can be set by three different aspects:

$$\varpi_{\rm min}(M_G \mid S_{\mathcal{C}}(\mathbf{q})) = \max \begin{cases} \varpi^{\rm lim}, & \text{initial } \varpi \text{ cut} \\ 10^{(M_G - G^{\rm lim}+10)/5}, & \text{initial } G \text{ cut} \\ \sqrt{\overline{\rm S/N}_{\varpi,\rm min}}\, 10^{(M_G - G^{\rm r}+10)/10}, & \overline{\rm S/N}_{\varpi,\rm min}(\text{Eq. 11}), \end{cases} \tag{12}$$

where $\varpi_{\rm min}$ is in units of mas, the absolute magnitude $M_G(G,\varpi)$ is an implicit function of G and ϖ, and $G^{\rm r} \approx 22$ (for Gaia EDR3) is a reference magnitude for the sky-averaged parallax S/N scaling (see Eq. 3). The simple form of Eq. 12 is for the dust free case, but can get generalized to include dust exctinction (see Section 4).

This formulation also links this 'forward modelling with a selection function' to classical $V_{\rm max}$-analyses (Schmidt 1968; Avni & Bahcall 1980; Paczynski 1990; Lilly et al. 1995). The important difference here is that $V_{\rm eff}$ is an effective volume that equals the true survey volume $V_{\rm max}$ for objects of a given M_G only in the isotropic, homogeneous and dust-free limit. Note that this sample has a well-defined $V_{\rm eff}$ for each M_G, but as an ensemble is *not* volume limited in any global sense. As discussed in Section 2.5, a volume-limited sample with $\varpi_{\rm min}(M_G \mid S_{\mathcal{C}}(\mathbf{q})) = $ const. for all M_G would be sub-optimal and no simpler to model, as Eq. 10 with a variable $\varpi_{\rm min}(M_G \mid S_{\mathcal{C}}(\mathbf{q}))$ illustrates.

To proceed and actually evaluate Eq. 10 we now spell out two possible assumptions for the spatial density distribution of WDs: the first is that $\hat{n}(\mathbf{x}_{\rm sky}, d(\varpi)) \approx$ const. This is clearly a poor approximation as soon as the sample reach becomes comparable or larger than the vertical scale height, h_z, of Galactic disk WDs. But it can prove instructive as the simplest limiting case. The second case, is to view the WD distribution near the Sun as a plane parallel slab of a single, and known (Gaussian) scale height h_z:

$$\hat{n}\,(l,b,\varpi) = \exp\left(-\frac{z^2}{2h_z^2}\right) = \exp\left(-\frac{\sin^2 b}{2\varpi^2 h_z^2}\right) \tag{13}$$

Note that this neglects WDs density variations with Galactocentric radius and the likely age-dependence of the scale height, which has impact on $V_{\rm eff}$ at the level of only a few percent.

As stated before, we assume a) that, averaged over 4π, dust extinction to the typical distance of the sample members is negligible; b) that the selection cuts in G and ϖ are sharp and uniform across the sky, allowing us to bypass the marginalization of the uncertainties in calculating $V_{\rm eff}$; and c) that the initial parent sample (and consequently the sub-samples) is approximately complete within these cuts. All these approximations may lead to systematic errors in estimating $V_{\rm eff}$ that are, however, very small compared to the $V_{\rm eff}$-range, $V_{\rm eff}(M_G)$, with M_G ranging from 6 to 16 mag.

In general, the result of marginalizing over space in Eq. 10 can be expressed as

$$\Lambda_{\mathcal{C}}(M_G,(B-R)) = \Phi_0(M_G,(B-R)) \cdot V_{\rm eff}(M_G)\,. \tag{14}$$

For the simplest, $\hat{n}(l,b,\varpi) =$ const., the effective volume $V_{\rm eff}$ from Eq. 10 is analytic and intuitive, $V_{\rm eff}(M_G) = \frac{4\pi}{3}\varpi_{\rm min}^{-3}(M_G)$, where $\varpi_{\rm min}$ reflects the most stringent among three basic selection function choices as per Eq. 12: an initial parallax cut ($\varpi_{\rm lim}$), an initial apparent magnitude cut (G), and a subsequently chosen parallax S/N cut $\overline{\rm S/N}_{\varpi,\rm min}$.

For the more realistic case of $\hat{n}(l,b,\varpi)$ specified by Eq.13, the rate prediction still can be written compactly as in Eq. 14, but $V_{\rm eff}(M_G)$ is now generalized to

$$V_{\rm eff}(M_G) = 2\pi \int_{-\pi/2}^{\pi/2} \cos b\; {\rm d}b \int_{\varpi_{\rm min}(M_G)}^{\infty} \varpi^{-4}\; \exp\left(-\frac{\sin^2 b}{2\varpi^2 h_z^2}\right) {\rm d}\varpi\,. \tag{15}$$

[6] The same holds whenever the model constrains 'number densities' of objects of interest.

For population-averaged scale heights of $h_z \approx$ 300 pc (e.g. Bovy et al. 2012) and for $\varpi_{\rm min}$=3 mas the resulting $V_{\rm eff}$ differ among the two approximations at the 15% level for luminous WDs ($M_G < 10$), and only $\sim 1\%$ for the faintest WDs. We will nonetheless use the more accurate calculation of $V_{\rm eff}$ from Eq. 15 throughout the remainder of the analysis. In all density plots shown in the paper, this difference would not be discernable.

The resulting $V_{\rm eff}(M_G)$ for our fiducial sample selection choices are shown in Figure 6, which indeed show that all three regimes of $\varpi_{\rm min}$ in Eq. 12 come to bear in this regime. For the most luminous objects $V_{\rm eff}$ is the same, set by the initial parallax cut $\varpi^{\rm lim}$; i.e. most luminous objects are volume complete. For the least luminous objects $V_{\rm eff}$ is set by the maximal distance, or $\varpi_{\rm min}$, at which they become fainter than $G^{\rm lim}$. For objects of intermediate luminosity, $V_{\rm eff}$ is set by $\overline{\rm S/N}_{\varpi,\rm min}$. For a low threshold of $\overline{\rm S/N}_{\varpi,\rm min}$, say > 5, this regime may be irrelevant; but for very demanding choices of $\overline{\rm S/N}_{\varpi,\rm min}$, say > 50, this regime may dominate, as illustrated by the thin blue and green lines in Figure 6.

We can now work out how to constrain the LCF at any one given color-pixel, $(M_G, (B-R))$. This is eqivalent to asking what the probability of the model parameter $\Phi_0(M_G, (B-R))$ is, given the number of sub-sample members in that pixel, $N_{\mathcal{C}}(M_G, (B-R))$ and $V_{\rm eff}(M_G)$? For flat priors on $\Phi_0(M_G, (B-R))$ and $N_{\mathcal{C}}(M_G, (B-R))$, this is just the Poisson probability

$$P\Big(\Phi_0 \mid N\Big) = P\Big(N \mid \Phi_0\Big) \cdot \frac{P(\Phi_0)}{P(N)} \sim \frac{\Lambda^N\ e^{-\Lambda}}{N!}, \tag{16}$$

where we use the shorthand of Φ_0 for $\Phi_0(M_G, (B-R))$, N for $N_{\mathcal{C}}(M_G, (B-R))$, and Λ for $\Lambda_{\mathcal{C}}(M_G, (B-R)) \equiv \Phi_0(M_G, (B-R))\ V_{\rm eff}(M_G)$. If one is only interested in a simple point estimate for $\Phi_0(M, c)$ one can adopt

$$\Phi_0(M_G, (B-R)) = \frac{N_{\mathcal{C}}(M_G, (B-R))}{V_{\rm eff}(M_G)}, \tag{17}$$

which is what we do for the plots in Section 3.4.

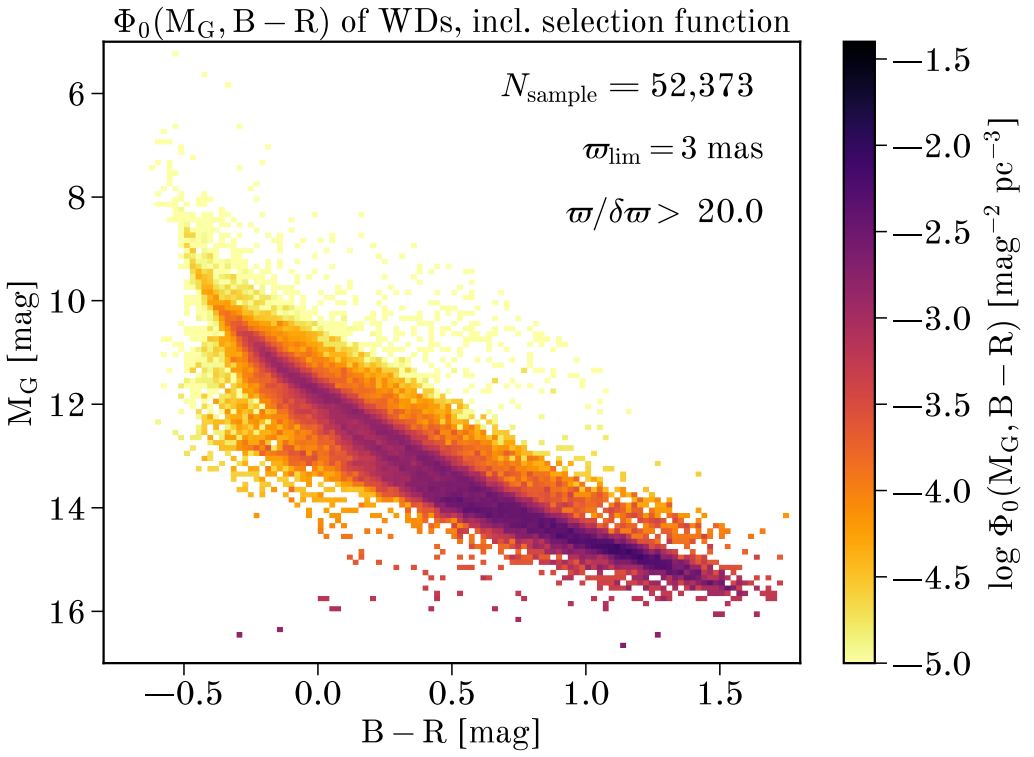

Figure 7. Estimate of $\Phi_0(M_G, (B-R))$, the WD luminosity-color function (LCF), derived from Eq. 14 for the particular sample selection function choices listed within the Figure. Note that $\Phi_0(M_G, (B-R))$ has units of [mag^{-2} pc^{-3}].

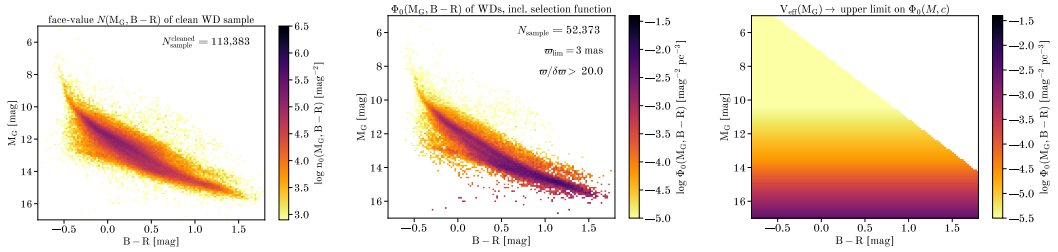

Figure 8. Comparison of the cleaned sample of likely WDs, $N_{\mathcal{C}}(M_G, (B-R))$ (left, with units [mag^{-2}]) with the modelled white dwarf LCF, $\Phi_0(M_G, (B-R))$ (center, with units [mag^{-2} pc^{-3}]). The $(M_G,(B-R))$ domain covered by both distribution is of course the same. Yet, the density peaks of $N_{\mathcal{C}}(M_G, (B-R))$ and of $\Phi_0(M_G, (B-R))$ are in dramatically different places in the $(M_G, (B-R))$-plane. The reason is that the two distributions differ by the selection function integral (expressed here via $V_{\rm eff}$), which is shown in the right panel; in the case at hand the selection function ends up being only a function of M_G, not $(M_G,(B-R))$; the survey volume in the top right corner of the $(M_G, (B-R))$-plane is zero, as such objects were excluded by the initial Gaia parent catalog query.

At first glance, the outcome of Eqs. 10-15 looks like a classic 'volume correction approach'. But it is important to keep in mind that in the formulation of Eq. 10, a number of the more subtle selection effects can be easily implemented, just requiring a numerical evaluation of the integral, as in the difference between Eq. 17 and Eq. 15; and this formulation will never apply[7] any noise-amplifying "upward correction" of the data by division where $S_{\mathcal{C}} \ll 1$.

3.4. *The Resulting Estimate of the White Dwarf LCF*

The basic point estimate of the WDs LCF is now simply an evaluation of Equation 17, after choosing an $(M_G, (B-R))$ grid on which our model parameters $\Phi_0(M, c)$ are to be evaluated. We know that the LCF has some fine-scale structure, and we therefore choose a fine grid of 120×120 points in M_G and $(B-R)$, covering $5 < M_G < 17$ and $-0.8 < (B-R) < 1.8$. Evaluation of Eq.17 results in the distribution shown in Figure 7: the white dwarf LCF in the Galactic solar neighbourhood. Again, it is crucial to note that this density has units of [mag^{-2} pc^{-3}], the number of white dwarfs per magnitude-color interval and per volume.

Figure 8 contrasts the face-value sample distribution of WDs $N_{\mathcal{C}}(M_G, (B-R))$ in the magnitude-color plane (left), with the estimate of the WDs LCF, $\Phi_0(M_G, (B-R))$ (center). The two distributions cover the same domain in $(M_G, (B-R))$, as per Eq.17: the two panels differ only in being re-weighted line-by-line by the selection function through the corresponding $V_{\rm eff}$ (right). But these quantitative differences are dramatic: while the sample-member density peaks near $(M_G, (B-R)) \approx (11, -0.1)$, the 'true' LCF density peaks at $(M_G, (B-R)) \approx (15, 1.2)$, where it is orders of magnitude higher than at $(11, -0.1)$. Of course, $N_{\mathcal{C}}(M_G, (B-R))$ and $\Phi_0(M_G, (B-R))$ also differ by their units.

We believe that this quantitative LCF distribution of WDs, $\Phi_0(M_G, (B-R))$, deserves much astrophysical follow-up: e.g. testing WD evolutionary models, as it properly reflects the density along cooling tracks for different WD masses; or an estimate of the density of WDs along the crystalization line. Such analyses are beyond the scope of this paper, especially as they would benefit from the inclusion of spectroscopic WD classification information.

We now only turn to a few more technical points regarding the selection function. In particular, we want to focus on the impact of different $\overline{\rm S/N}_{\varpi,\rm min}$ choices on the analysis. Figure 9 shows the same simple estimate of $\Phi_0(M_G, (B-R))$ (Eq.17) but for four alternate choices of $\overline{\rm S/N}_{\varpi,\rm min}$, namely $> 5, 10, 40, 80$; these different cuts lead to sample sizes that differ by nearly a factor of 10 in sample size. The first thing to note in Figure 9 is that the resulting density estimates of $\Phi_0(M, c)$ are mutually consistent, as they should be. An inclusive choice of $\overline{\rm S/N}_{\varpi,\rm min} = 5$ leads of course to a better sampled estimate of $\Phi_0(M, c)$. A far more stringent choice of $\overline{\rm S/N}_{\varpi,\rm min} > 80$ leads to a far smaller sample, but one with very precise distances (and luminosities): this clarifies the bifurcation of the LCF at intermediate colors and luminosities, yielding a sharper image of the LCF, but at the expense of sparser sampling.

This comparison also illustrates that the choices of sample cuts such as $\Delta{\rm S/N}(\varpi)$ are not universal, but should depend on the science goals. Of course, it is possible to combine the $\Phi_0(M, c)$ estimates resulting from different choices

[7] For convenience, we have phrased our estimate here as a volume correction in Eq. 17; but of course, a probabilistic constraint on $\Phi_0(M, c)$ from $p\,(N(M, c) \mid \Phi_0(M, c), V_{\rm eff})$ would also work well.

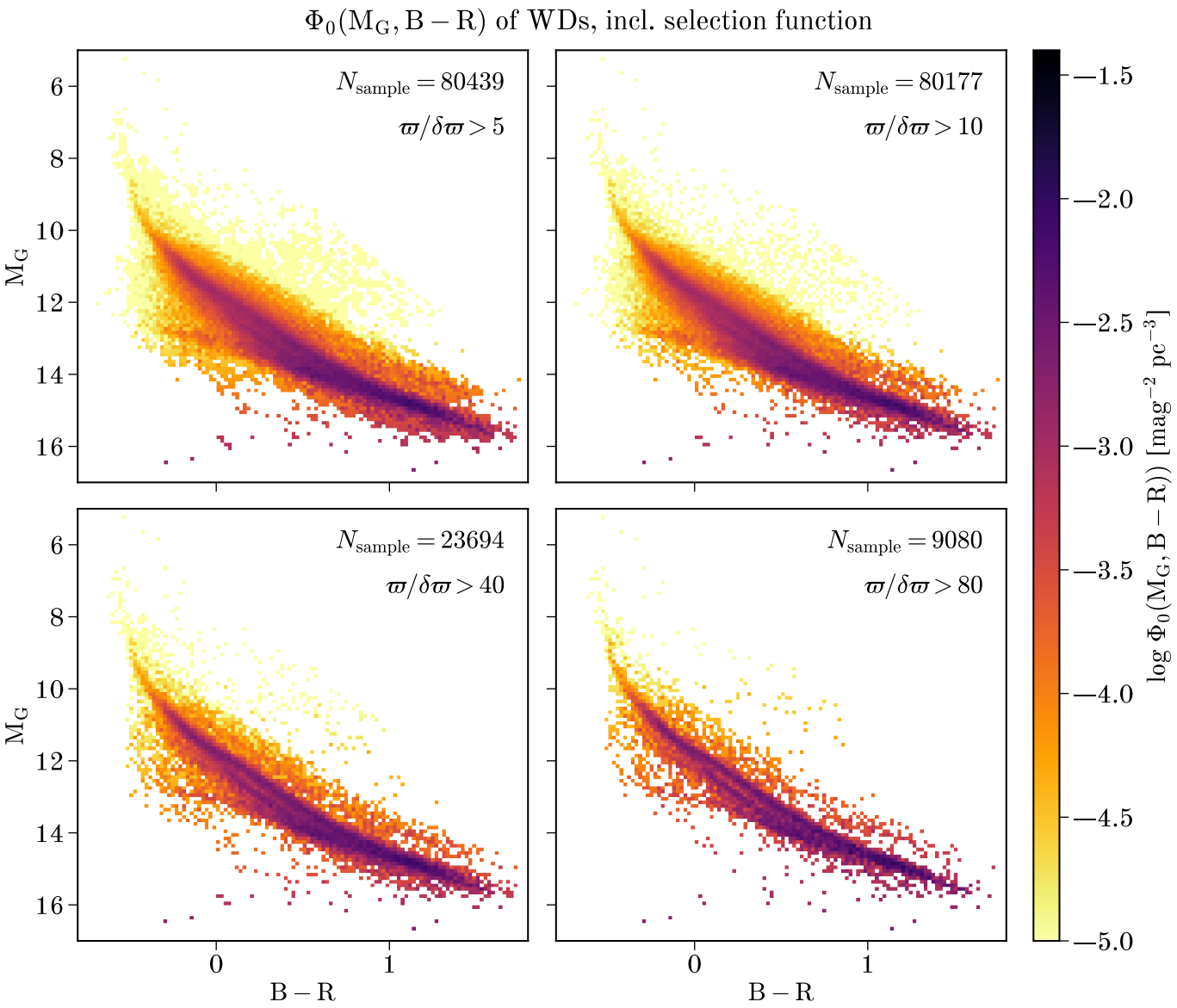

Figure 9. Estimate of $\Phi_0(M_G, (B-R))$, analogous to that in Figure 7, but for four different choices of $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}}$: 5,10,40,80. Despite the sample size differences of a factor of five, the inferred $\Phi_0(M_G, (B-R))$ are mutually consistent. High values of $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}}$ lead to small samples but better detail in the high-density structure of $\Phi_0(M_G, (B-R))$; low values of $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}}$ lead to larger samples and a better sampling of $(M_G,(B-R))$-plane, but at reduced resolution, as the uncertainties in M_G become noticable.

of $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}}$ and possible different grid in (M_G, c). We show this in Figure 10: we started with the estimate resulting from $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}} > 80$, which shows the sharpest high-density features but suffers from sparse sampling in the low-density regions. We retained the 20% highest density pixels in that density map, replacing the rest with the values from the $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}} > 20$ estimate; we repeated this exercise, retaining the 50% highest density pixels in this $\Phi_0(M, c)$ estimate, and replaced the rest with the values from a LCD density map that had been constructed from a 2× coarser (M_G, c) grid and a sample with $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}} > 5$. The resulting LCF map (Fig. 10) combines sharp features in the LCF, including the bifurcation, with better S/N and coverage in the low-density regions. We did this in part to stress that the LCF, determined via Eq. 17 is a model estimate of a function defined across the full portion of the (M_G, c).

This leads us to comment on the empty, white areas of the LCF distribution in the $(M_G, (B-R))$-plane of, say, Figure 10. It is not that there are no constraints on the LCF for these $(M_G, (B-R))$, as the selection function is defined also for counter-factual objects. Indeed, the Poisson estimate for $\Phi_0(M_G, (B-R))$ from Eq. 16 holds of course

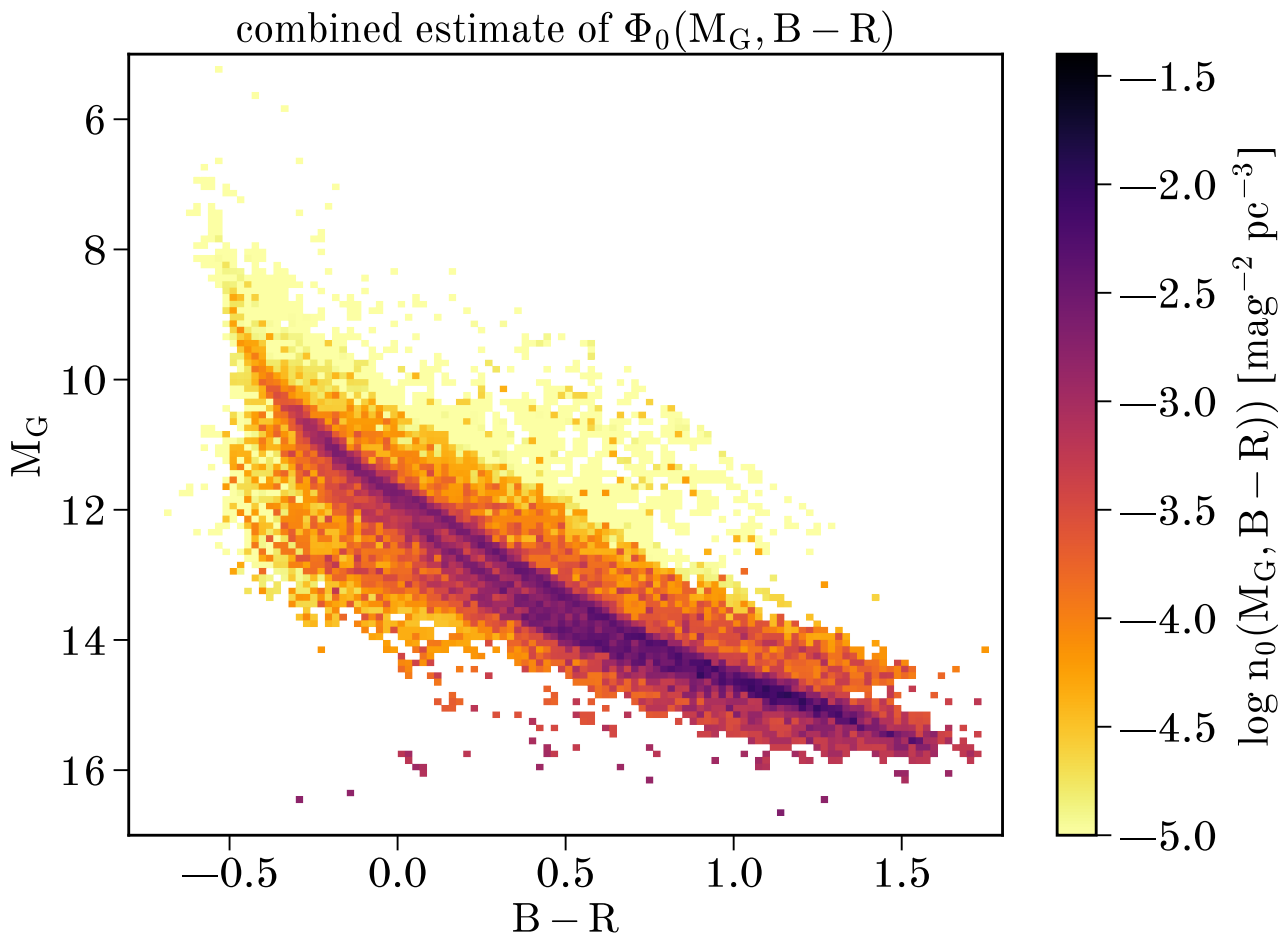

Figure 10. Estimate of the WD luminosity-color density (LCF) from a combination of the previous estimates. The high density regions were estimated using the $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}} > 80$ sample, to take full advantage of the precise parallaxes for the sharp features in the diagram. The other parts were estimated from the $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}} > 20$ sample, and finally from the $\overline{\mathrm{S/N}}_{\varpi,\mathrm{min}} > 5$ sample that had been calculated on a coarser grid; these latter two steps reduce sampling noise in the low-density parts of the distribution.

also for the empty pixels in the $(M_G, (B-R))$-plane, where $N_{\mathcal{C}}(M_G, (B-R)) = 0$. There, Eq. 16 implies e.g. a 1σ upper limit of $\Phi_0(M_G, (B-R)) = 1/V_{\mathrm{eff}}(M_G)$. So, this modelling implies estimates or upper limits across the entire $(M_G, (B-R))$ plane, as illustrated in Figure 11.

These considerations lead to a compact way to present modelling results, as the ones presented here. The most immediate result is simply the 2D-array of $N_{\mathcal{C}}(M_G, (B-R))$, the most likely $\Phi_0(M_G, (B-R))$ (Eq. 16) along with the $(M_G, (B-R))$ grid on which it is sampled. This allows to reconstruct the probability distribution for $\Phi_0(M_G, (B-R))$.

4. SUMMARY AND DISCUSSION

We begin the final part of our selection function exposition by summarizing very briefly the selection function fundamentals. We then touch on a number of practical subtleties that need to be considered, especially in cases going beyond the worked example above and its simplifying assumptions.

• *What's a selection function? When do we need it?* In modelling 'catalog data', a selection function is always needed when we want to ask questions about how frequently we are expected to find objects with certain physical attributes, their densities or property distribution; this type of modelling covers a very broad swath of astrophysical inquiry. The selection function, $S_{\mathcal{C}}(\mathbf{q})$, can be thought of as the multiplicative factor relating a model prediction, $\mathcal{M}(\mathbf{q} \mid \boldsymbol{\Theta}_{\mathrm{mod}})$ to

a catalog incidence $d\Lambda_{\mathcal{C}}(\mathbf{q})$; or it can be thought of as the (dimensionless) probability that an object of properties $\mathbf{q}$ will be in a parent catalog, or a sub-sample drawn from it.

• *How to Construct a Selection Function?* In practice, a selection function is constructed through a set of probabilistic conditions (often Boolean conditions) that describe the probability that an object (real or counter-factual) enters a (sub-)sample to be modelled. These conditions are intended to isolate near-optimal sets of objects that are informative about a physical question at hand. In general, there is no need or even benefit for samples to be 'complete' with respect to any simple quantity such as flux or volume, just their selection function must be sufficiently well known. Often, such sub-samples are drawn from a parent catalog (with its intrinsic selection function $S_{\mathcal{C}}^{\rm parent}(\mathbf{q})$) and pared down to a suitable sub-sample by subsequent user-defined selection cuts, $S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$. And it usually makes sense to describe the overall selection function as the product of these two terms. Ideally, the arguments of the selection function should be the minimal, or simplest, set of cataloged attributes $\mathbf{q}$ that isolates a suitable sample, and that can be modelled (i.e. are arguments of $\mathcal{M}(\mathbf{q} \mid \boldsymbol{\Theta}_{\rm mod})$). In general that means that the $\mathbf{q}$ should be "observables" (positions, fluxes, etc..), but under many circumstances they can also be data-quality flags.

• *How to Use Selection Functions in Modelling?* In most circumstances, there is no simple way to define or find an actually optimal selection function. The selection function $S_{\mathcal{C}}(\mathbf{q})$ inevitably reflects astrophysically informed choices and judgements. However, it is crucial that any chosen selection function is properly applied in the subsequent modelling inference, where it is indispensable. In the simplest form, $S_{\mathcal{C}}(\mathbf{q})$ appears in the modelling of its chosen sub-sample just as a multiplicative term. But in practice, $S_{\mathcal{C}}(\mathbf{q})$ often depends – for good reasons – on quantities within $\mathbf{q}$ not used in the data-model comparison; these quantities are then best marginalized out (see Eq. 6)

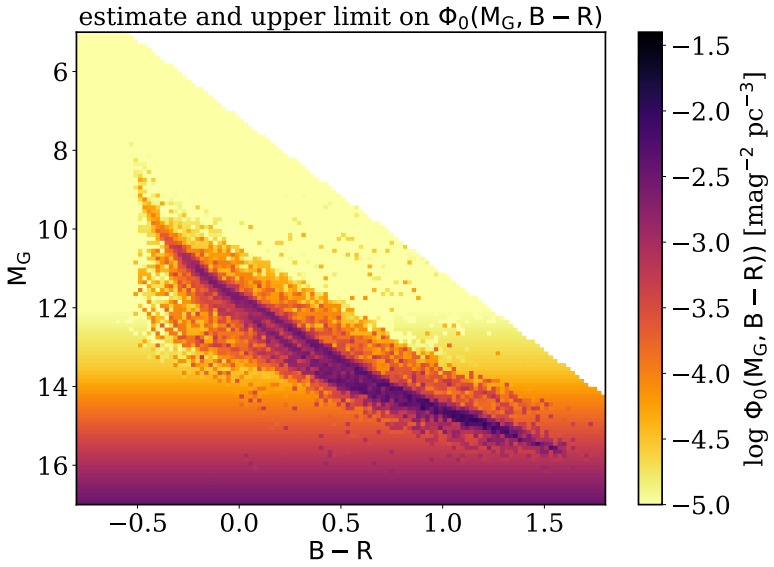

Figure 11. Representation of the $\Phi_0(M_G, (B-R))$, as in Fig. 10, but aigmented by 1σ upper limits across the entire (M_G,$(B-R)$) plane, where the sample does not contain any entries. Those limits were derived by calculating the density that even one single object in this (M_G,$(B-R)$)-pixel would imply. Knowing the selection function for counter-factual properties $\mathbf{q}$ permits to place constraints in models where we do not see any data in the sample.

Finally, we go through a number of complexities regarding the selection function that we have omitted so far in the interest of a simpler logical flow of the paper, as as they were not critical for our worked example.

• *Spatial complexity of selection functions:* The selection function of most large parent catalogs, $S_{\mathcal{C}}^{\rm parent}$ has very complex spatial structure for a variety of reasons. For catalogs based on space-based all-sky surveys, the spacecraft's orbit and ensuing scanning pattern gives $S_{\mathcal{C}}^{\rm parent}(m, \alpha, \delta)$ complex structure on the sky (e.g. Boubert & Everall 2020). In addition crowding of sources can affect (lower) $S_{\mathcal{C}}^{\rm parent}(m, \alpha, \delta)$ either because the overall source density is high, or if sources are spatially correlated, as for stellar clusters or binaries (e.g. El-Badry & Rix 2018). In particular, the lower completeness to faint sources near bright sources is a selection function aspect that afflicts most survey catalogs.

• *Dust Extinction: Part of the Selection Function or of the Model?* Large catalogs that are based on X-ray, UV, optical or near-IR observations have observables that are affected by dust extinction. For catalogs of distant objects ($D \gtrsim 10$ kpc) dust is a foreground screen; for objects with 100 pc $\lesssim D \lesssim 10$ kpc the 3D distribution of the dust matters (Drimmel & Spergel 2001; Green et al. 2019; Lallement et al. 2019). This situation leaves several choices (e.g. Bovy et al. 2016): To model observed or dereddened fluxes? To make extinction part of the selection function, $S_{\mathcal{C}}(\mathbf{q})$, or part of the model, $\mathcal{M}(\mathbf{q} \mid \boldsymbol{\Theta}_{\rm mod})$? If the 2D/3D dust map was perfectly known, all these approaches are viable; they just differ in where the complexity is increased, either in the modelling or in the selection function. Yet two aspects argue for making dust extinction part of the modelling: It keeps the arguments $\mathbf{q}$ of the catalog incidence $\Lambda_{\mathcal{C}}(\mathbf{q})$ and the selection function $S_{\mathcal{C}}(\mathbf{q})$ better described as 'observables'. And making the dust extinction part of the model allows us to marginalize out any uncertainties in it.

How dust extinction, say a 3D dust map $A_\lambda(l, b, \varpi)$, is implemented in the modelling depends on the details, though it always alters the model-predicted apparent magnitude $m(M, l, b, \varpi)$ and color, say, $(B-R)(l, b, \varpi)$. In our worked example above, it just makes the maximal distance, $\varpi_{\rm min}$ to which an object is still in the sample a function of sky-position. The terms $M_G - m_G^{lim/r} + 10$ in the latter two cases of Eq. 12 must be replaced by $M_G - m_G^{lim/r} + 10 + A_G(l, b, \varpi)$; this makes Eq. 12 an (easily solved) implicit equation, with a sky-position dependence through $A_G(l, b, \varpi)$. The resulting $\varpi_{\rm min}(l, b)$ then becomes an integration boundary that explicitly depends on sky position, $\varpi_{\rm min}(\mathbf{x}_{\rm sky}, M_G | S_{\mathcal{C}}(\mathbf{q}))$ in the last integral of Eq. 10:

$$\Lambda_{\mathcal{C}}(M_G, (B-R)) = \Phi_0(M_G, (B-R)) \int_{4\pi} \mathrm{d}\Omega \int_{\varpi_{\rm min}(\mathbf{x}_{\rm sky}, M_G | S_{\mathcal{C}}(\mathbf{q}))}^{\infty} \varpi^{-4} \mathrm{d}\varpi \; \hat{n}(\mathbf{x}_{\rm sky}, d(\varpi)). \tag{18}$$

For our worked example, the inclusion of 3D dust extinction would simply reduce the effective survey volume, $V_{\rm eff}(M_G)$, which could be pre-computed for any given M_G. This integration again illustrates why the extinction model must exist also for counter-factual objects, i.e. all directions and distances that might be accessible in the dust-free case.

• *Sub-sample selection based on noisy catalog entries:* In the discussion so far, and in the worked example, we have assumed that the $\mathbf{q}$ in the catalog are precisely known at the values where $S_{\mathcal{C}}(\mathbf{q})$ varies strongly (or, where $S_{\mathcal{C}}(\mathbf{q})$ makes a cut). In that regime it is not necessary to differentiate explicitly between the 'observed' $\mathbf{q}_{obs}$ and the model-predicted $\mathbf{q}_{true}$. However, in the majority of pertinent cases the uncertainties of the $\mathbf{q}$ at the selection function boundaries will matter: if one makes a sharp sample-selection cut at a certain $\mathbf{q}_{obs}$, some sample members will have $\mathbf{q}_{true}$ that lie beyond that selection boundary, while other objects that have $\mathbf{q}_{true}$ within these boundaries will be absent from the chosen sub-sample. In that case one needs to have a model that covers a larger domain in $\mathbf{q}$ and needs to integrate over all $\mathbf{q}_{true}$, accounting for the $\mathbf{q}$ uncertainties, $\sigma_{\mathbf{q}}$. In our notation, Eq. 6 would generalize, for the homoscedastic regime of a uniform, typical uncertainty, $\sigma_{\mathbf{q}}$, to

$$\mathrm{d}\Lambda_s(M, c, \mathbf{x}) = \Phi(M, c, \mathbf{x} \mid \boldsymbol{\Theta}_{\rm mod}) \int \mathrm{d}\mathbf{q}_{obs} \; S_{\mathcal{C}}(\mathbf{q}_{obs})) \; p\Big(\mathbf{q}_{obs} \mid \mathbf{q}_{true}(M, c, \mathbf{x}), \sigma_{\mathbf{q}}\Big) \mathrm{d}M \mathrm{d}c \; \mathrm{d}V \; , \tag{19}$$

with corresponding changes in the equations that flow from Eq. 6. Such an approach is applied explicitly e.g. in Foreman-Mackey et al. (2014); Frankel et al. (2018).

• *What if the selection arises from the combination of two or more catalogs?* There are many instances, where the sub-sample selection arises from a combination of catalogs. The modelling of spectroscopic surveys, whose targets are almost inevitably drawn from a pre-existing photometric survey is a prime example. Modelling such spectroscopic surveys, say multi-object surveys such as SDSS, is based itself on a number of object selection steps: the selection of a sub-sample from photometric catalogs to yield objects eligible for spectroscopic targetting. This may be followed by the selection of objects among them that actually get spectroscopic observations, which then result in a new set of observables $\mathbf{q}_{\rm spec}$. And ultimately one will select yet another sub-sample from among those, based on their $\mathbf{q}_{\rm spec}$: e.g. selecting galaxies within a certain redshift range, stars of a certain metallicity, or stars spectroscopically classified as WDs. In most cases, these subsequent steps can still be treated as a sequence of multiplicative terms yielding $S_{\mathcal{C}}(\mathbf{q})$, as in Eqs. 2 & 8. Worked examples of this approach can be found e.g. in Bovy & Rix (2013) and Wojno et al. (2017).

• *The precision of the selection function?* In the discussion so far, we have presumed that the selection function has been determined with sufficient precision for whatever astrophysical problem is at hand. But we have not elaborated how one could assess the precision of $S_{\mathcal{C}}(\mathbf{q}) = S_{\mathcal{C}}^{\rm parent}(\mathbf{q}) \cdot S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$, which will vary case-by-case. An example of how to assess the precision of $S_{\mathcal{C}}^{\rm parent}(\mathbf{q})$ is given in Boubert & Everall (2020). In most cases, the precision of $S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$

should be high, as the selection functions are being designed in the same context as the modelling. However, there are cases where it is hard or even impossible to determine $S_{\mathcal{C}}^{\rm sample}(\mathbf{q})$ in sufficient approximation, even if $S_{\mathcal{C}}^{\rm parent}(\mathbf{q})$ is well-determined. The literature contains many samples of spectroscopic observations of objects contained in a catalog of well-defined $S_{\mathcal{C}}^{\rm parent}(\mathbf{q})$. Let's consider the case where a selection function is not constructable in a quantitative way with respect to some quantity a from among the $\mathbf{q}$, while it is well understood with respect to another element of $\mathbf{q}$, say b. As an example of such a situation we can take the radial metallicity distribution of stars in the Galactic disk from a spectroscopic survey, such as APOGEE (Majewski et al. 2017), $p([\mathrm{Fe/H}], R)$. In this case, the selection function for the Galactocentric radii, $R(l, b, \varpi)$ may be hard to construct, but the spectroscopic targetting is insensitive to different [Fe/H] among disk stars (see Frankel et al. 2018). It would be difficult to model $p([\mathrm{Fe/H}], R)$ for such a sample, but it is far easier to model the metalicity distribution *conditioned* on Galactocentric radius, $p([\mathrm{Fe/H}] \mid R)$. This can be generalized: if the selection function cannot be well determined for some components ($\mathbf{q}_{\rm noSF}$ of $\mathbf{q}$) that one wants to model, one should build a model for the incidence of the other components of $\mathbf{q}$, conditioned on $\mathbf{q}_{\rm noSF}$.

None of the individual aspects of this selection function formalism is without precedence in the literature. But we hope that this paper can help to clarify when, and when not, all these aspects have to come together to do justice to the information content of large astronomical catalogs.

For this paper, we have restricted ourselves to applying this selection function formalism to the estimate of the luminosity-color function of white dwarfs. There are a number of other applications where exactly the same approach should be pursued, such as the LCF of cataclysmic variables, the LCF of the lowest main sequence into the brown dwarf regime, the LCF of hot subdwarfs, etc., as they all fall into the regime of nearby objects where the detailed spatial distribution is a nuisance parameter to be integrated out, but only under inclusion of an appropriate selection function.

The input files and the computations for all plots in this paper can be found in this notebook: https://github.com/gaia-unlimited/WD-selection-function.

ACKNOWLEDGMENTS

The authors would like to thank Neige Frankel and Kareem El-Badry for thoughtful comments on the manuscript.

DB thanks Magdalen College for his fellowship and the Rudolf Peierls Centre for Theoretical Physics for providing office space and travel funds. AE thanks the Science and Technology Facilities Council of the United Kingdom for financial support.

This work is a result from the GaiaUnlimited project which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101004110. The GaiaUnlimited project was started at the 2019 Santa Barbara Gaia Sprint, hosted by the Kavli Institute for Theoretical Physics at the University of California, Santa Barbara.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

REFERENCES

Avni, Y., & Bahcall, J. N. 1980, ApJ, 235, 694, doi: 10.1086/157673

Blanton, M. R., Hogg, D. W., Bahcall, N. A., et al. 2003, ApJ, 592, 819, doi: 10.1086/375776

Boubert, D., & Everall, A. 2020, MNRAS, 497, 4246, doi: 10.1093/mnras/staa2305

Boubert, D., Everall, A., Fraser, J., Gration, A., & Holl, B. 2021, MNRAS, 501, 2954, doi: 10.1093/mnras/staa3791

Boubert, D., Everall, A., & Holl, B. 2020, MNRAS, 497, 1826, doi: 10.1093/mnras/staa2050

Bovy, J. 2017, MNRAS, 470, 1360, doi: 10.1093/mnras/stx1277

Bovy, J., & Rix, H.-W. 2013, ApJ, 779, 115, doi: 10.1088/0004-637X/779/2/115

Bovy, J., Rix, H.-W., Green, G. M., Schlafly, E. F., & Finkbeiner, D. P. 2016, ApJ, 818, 130, doi: 10.3847/0004-637X/818/2/130

Bovy, J., Rix, H.-W., Liu, C., et al. 2012, ApJ, 753, 148, doi: 10.1088/0004-637X/753/2/148

Brown, A. G. A. 2021, arXiv e-prints, arXiv:2102.11712. https://arxiv.org/abs/2102.11712

Cheng, S., Cummings, J. D., & Ménard, B. 2019, ApJ, 886, 100, doi: 10.3847/1538-4357/ab4989

Cole, S., Percival, W. J., Peacock, J. A., et al. 2005, MNRAS, 362, 505, doi: 10.1111/j.1365-2966.2005.09318.x

Drimmel, R., & Spergel, D. N. 2001, ApJ, 556, 181, doi: 10.1086/321556

El-Badry, K., & Rix, H.-W. 2018, MNRAS, 480, 4884, doi: 10.1093/mnras/sty2186

El-Badry, K., Rix, H.-W., & Weisz, D. R. 2018, ApJL, 860, L17, doi: 10.3847/2041-8213/aaca9c

Everall, A., Boubert, D., Koposov, S. E., Smith, L., & Holl, B. 2021, MNRAS, 502, 1908, doi: 10.1093/mnras/stab041

Everall, A., & Das, P. 2020, MNRAS, 493, 2042, doi: 10.1093/mnras/staa283

Fontaine, G., Brassard, P., & Bergeron, P. 2001, PASP, 113, 409, doi: 10.1086/319535

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, ApJ, 795, 64, doi: 10.1088/0004-637X/795/1/64

Frankel, N., Rix, H.-W., Ting, Y.-S., Ness, M., & Hogg, D. W. 2018, ApJ, 865, 96, doi: 10.3847/1538-4357/aadba5

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2020a, arXiv e-prints, arXiv:2012.01533. https://arxiv.org/abs/2012.01533

Gaia Collaboration, & et al. 2016, A&A, 595, A1, doi: 10.1051/0004-6361/201629272

Gaia Collaboration, Babusiaux, C., van Leeuwen, F., et al. 2018, A&A, 616, A10, doi: 10.1051/0004-6361/201832843

Gaia Collaboration, Smart, R. L., Sarro, L. M., et al. 2020b, arXiv e-prints, arXiv:2012.02061. https://arxiv.org/abs/2012.02061

Gentile Fusillo, N. P., Tremblay, P.-E., Gänsicke, B. T., et al. 2019, MNRAS, 482, 4570, doi: 10.1093/mnras/sty3016

Green, G. M., Schlafly, E., Zucker, C., Speagle, J. S., & Finkbeiner, D. 2019, ApJ, 887, 93, doi: 10.3847/1538-4357/ab5362

Hollands, M. A., Tremblay, P. E., Gänsicke, B. T., Gentile-Fusillo, N. P., & Toonen, S. 2018, MNRAS, 480, 3942, doi: 10.1093/mnras/sty2057

Kleinman, S. J., Kepler, S. O., Koester, D., et al. 2013, ApJS, 204, 5, doi: 10.1088/0067-0049/204/1/5

Lallement, R., Babusiaux, C., Vergely, J. L., et al. 2019, A&A, 625, A135, doi: 10.1051/0004-6361/201834695

Lilly, S. J., Tresse, L., Hammer, F., Crampton, D., & Le Fevre, O. 1995, ApJ, 455, 108, doi: 10.1086/176560

Lindegren, L., Klioner, S. A., Hernández, J., et al. 2020, arXiv e-prints, arXiv:2012.03380. https://arxiv.org/abs/2012.03380

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94, doi: 10.3847/1538-3881/aa784d

McCleery, J., Tremblay, P.-E., Gentile Fusillo, N. P., et al. 2020, MNRAS, 499, 1890, doi: 10.1093/mnras/staa2030

Paczynski, B. 1990, ApJ, 348, 485, doi: 10.1086/168257

Rybizki, J., Green, G., Rix, H.-W., et al. 2021a, arXiv e-prints, arXiv:2101.11641. https://arxiv.org/abs/2101.11641

—. 2021b, arXiv e-prints, arXiv:2101.11641. https://arxiv.org/abs/2101.11641

Schmidt, M. 1968, ApJ, 151, 393, doi: 10.1086/149446

Tremblay, P.-E., Fontaine, G., Fusillo, N. P. G., et al. 2019, Nature, 565, 202, doi: 10.1038/s41586-018-0791-x

Trumpler, R. J., & Weaver, H. F. 1953, Statistical astronomy

Weidemann, V. 2000, A&A, 363, 647

Wojno, J., Kordopatis, G., Piffl, T., et al. 2017, MNRAS, 468, 3368, doi: 10.1093/mnras/stx606

Wood, M. A. 1992, ApJ, 386, 539, doi: 10.1086/171038