Leverage and influence diagnostics for Gibbs spatial point processes

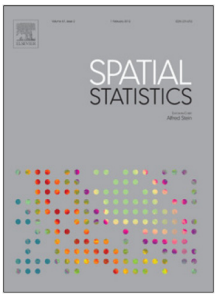
Baddeley, Adrian; Rubak, Ege; Turner, Rolf

Published in:
Spatial Statistics

DOI (link to publication from Publisher):
10.1016/j.spasta.2018.09.004

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version



Leverage and influence diagnostics for Gibbs spatial point processes

Adrian Baddeley, Ege Rubak, Rolf Turner

PII: S2211-6753(18)30118-0
DOI: https://doi.org/10.1016/j.spasta.2018.09.004
Reference: SPASTA 328

To appear in: *Spatial Statistics*

Received date : 7 June 2018
Accepted date : 28 September 2018

Please cite this article as: Baddeley A., et al., Leverage and influence diagnostics for Gibbs spatial point processes. *Spatial Statistics* (2018), https://doi.org/10.1016/j.spasta.2018.09.004

Leverage and influence diagnostics for Gibbs spatial point processes

Adrian Baddeley[a,b], Ege Rubak[c], Rolf Turner[d]

[a]*Department of Mathematics & Statistics, Curtin University, Perth, Australia*
[b]*Data61, CSIRO, Perth, Australia*
[c]*Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark*
[d]*Department of Statistics, University of Auckland, New Zealand*

Abstract

For point process models fitted to spatial point pattern data, we describe diagnostic quantities analogous to the classical regression diagnostics of leverage and influence. We develop a simple and accessible approach to these diagnostics, and use it to extend previous results for Poisson point process models to the vastly larger class of Gibbs point processes. Explicit expressions, and efficient calculation formulae, are obtained for models fitted by maximum pseudolikelihood, maximum logistic composite likelihood, and regularised composite likelihoods. For practical applications we introduce new graphical tools, and a new diagnostic analogous to the effect measure `DFFIT` in regression.

Keywords: composite likelihood, conditional intensity, DFBETA, DFFIT, likelihood influence, model diagnostics, model validation, pseudolikelihood.
2010 MSC: 62M30, 62J20

1. Introduction

This paper develops tools for model criticism in the analysis of spatial point pattern data [5, 26, 27, 33, 39]. Model criticism is good statistical practice in the analysis of any kind of data [20, 19]. The necessary tools — such as plots of residuals, leverage and influence diagnostics — are well-developed and widely used for linear models [1, 12] and generalized linear models [42, 36, 54, 25, 32,

Email address: `adrian.baddeley@curtin.edu.au` (Adrian Baddeley)

47, 30]. However, for some other kinds of data and models, these tools are not yet available.

For spatial point process models fitted to spatial point pattern data, it is only recently that residuals [9] and leverage and influence diagnostics [2] were developed by adapting the classical definitions of these quantities to the setting of spatial point processes.

Leverage and influence can be viewed as Taylor approximations of the effect of changes in the data on properties of the fitted model. They depend on the type of model, but also on the fitting method. In [2] we focused on Poisson point process models fitted by maximum likelihood, derived explicit formulae for the diagnostics, and demonstrated their utility on real data. For applications, it is important to extend these diagnostics to larger model families such as the Cox, Neyman-Scott, Gibbs, and determinantal point processes [5, Chapters 12 & 13]. Extension to these model classes is more complicated than envisaged in [2], and some of the formulae stated without proof in [2, Section 4.2] need to be corrected. Additionally we need to extend the original approach in [2] to deal with new model-fitting techniques, including logistic composite likelihood [3], quasilikelihood [29], hierarchical pseudolikelihood [31] and penalised pseudolikelihood [6].

Apart from these advances in methodology and technique, there is a great need to make these diagnostic tools more accessible to applied statisticians. The original definitions in [2] were intimidating abstract statements about derivatives in function spaces. In this paper we pursue a much simpler way to define and understand the diagnostics. They are here defined as the "obvious" Taylor approximations of the effect of changes in the data on properties of the fitted model. Using this approach, we derive explicit formulae for the diagnostics, including greatly expanded results for Gibbs models. This approach unifies and improves the understanding of the diagnostics, clarifies their role, and strengthens the connections with the classical leverage and influence diagnostics for generalized linear models [42, 36, 54, 25, 32, 47, 30].

The paper begins in Section 2 with an accessible overview of the diagnostics, including worked examples (with R code supplied), practical advice, and some new graphical tools. Section 3 defines notation and technical assumptions. The

diagnostics are defined and developed in Sections 4–6: leverage in Section 4, parameter influence in Section 5, and likelihood influence in Section 6. These sections include explicit formulae for the diagnostics in common cases, including details such as edge corrections, as well as general formulae for models with a flexible parametric form. Parameter influence and likelihood influence are defined using the spatial point process counterpart of "case deletion" diagnostics [54, 16, 43, 41]. In Section 7 we introduce a new point process model diagnostic analogous to the regression diagnostic `DFFIT`. Section 8 completes the analysis in the worked example that was started in Section 2. An efficient software implementation requires sparse-matrix calculations, for which we provide detailed formulae in Section 9. Implementation and timings are described in Section 10.

Appendix A recalls the classic definitions of leverage and influence in a generalized linear model, for reference. Appendix B discusses regularised composite likelihoods. Proofs of some results are relegated to Appendix C. Online supplements give a general guide to the software, code scripts, a detailed analysis of a real example dataset (gold deposits in Western Australia), and a re-analysis of the famous Chorley-Ribble cancer data correcting our earlier analysis in [2, 5].

2. Overview of diagnostics

2.1. Example data and model

A spatial point pattern dataset is a finite set $\mathbf{x} = \{x_1, \ldots, x_n\}$ of points observed in a survey region or "window" W in the two-dimensional plane. Figure 1 shows the classic Swedish Pines dataset of Strand [49] popularised by Ripley [46]. It gives the locations of 71 pine saplings in a 9.6×10 metre survey quadrat. These data are often used as an example of regularity or inhibition between points, which could be explained by the effects of plant competition [46, pp. 172–175], [52, p. 483], [6], [5, pp. 221, 265, 336, 405, 488, 503–520, 529, 536].

However, the pattern also appears to be spatially inhomogeneous [5, pp. 169–175 and 513]. The left panel of Figure 2 shows a nonparametric estimate of the intensity using a Gaussian kernel estimate [22] with bandwidth 1.4 metres selected by Scott's rule of thumb [48]. It suggests that a ridge of higher intensity runs from top left to bottom right in the plot.

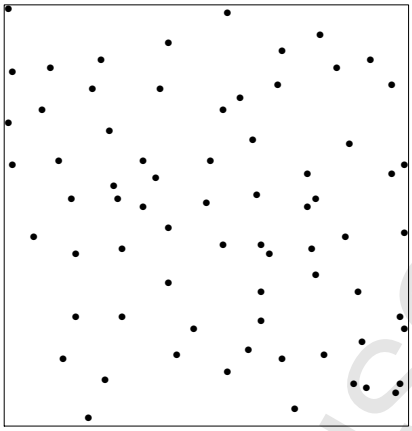


Figure 1: Swedish Pines data: 71 saplings in a 9.6×10 metre region.

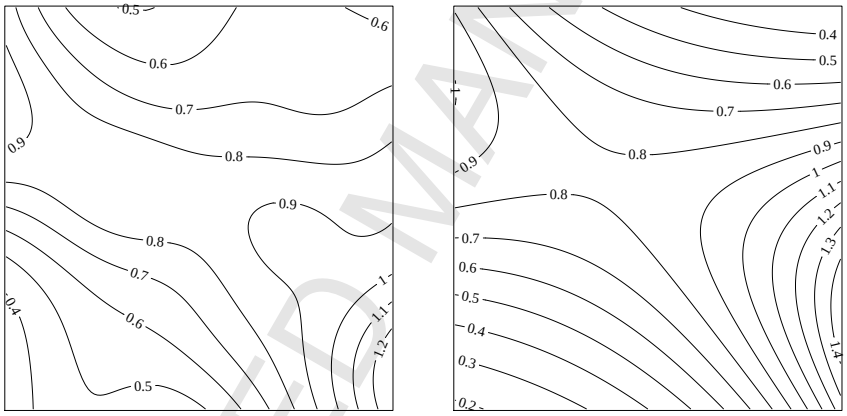


Figure 2: Estimates of intensity for the Swedish Pines data. *Left:* kernel estimate; *Right:* fitted log-quadratic function of coordinates.

Conflicting conclusions are obtained from different fitted models and hypothesis tests applied to these data [5, pp. 488, 512, 517] and there is no consensus in the literature. This motivates us to investigate the sensitivity of the analysis to individual data points.

For simplicity, we use Poisson point process models in this overview. We fitted a Poisson process to the Swedish Pines data in which the intensity or rate $\lambda(u)$ at spatial location u is a log-quadratic function of the Cartesian coordinates:

$$\lambda_{\boldsymbol{\theta}}((x,y)) = \exp(\theta_0 + \theta_1 x + \theta_2 y + \theta_3 x^2 + \theta_4 xy + \theta_5 y^2) \qquad (1)$$

where $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_5)^\top$ is the parameter vector. The model was fitted by maximum likelihood using the Berman-Turner [13] quadrature approximation. The fitted intensity is shown in the right hand panel of Figure 2.

The evidence for non-uniform intensity is equivocal. The likelihood ratio test of $H_0 : \theta_1 = \theta_2 = \cdots = \theta_5 = 0$ against $H_1 : \theta_i \neq 0$ for some i, is not significant at level 0.05, but the likelihood ratio tests of each of the hypotheses $H_{0i} : \theta_i = 0$ against $H_{1i} : \theta_i \neq 0$ for each $i = 1, \ldots, 5$ are all significant at the 0.05 level. Backward stepwise model selection using AIC retains all the terms in (1) except the x^2 term.

2.2. Diagnostics

Next we introduce and demonstrate the diagnostics that are defined in this paper. An online supplement provides R code to generate the figures shown in the paper.

For reference, Appendix A contains the classic definitions of leverage and influence diagnostics for a generalized linear model, which have been adapted here to the spatial point process setting.

2.2.1. Leverage

Figure 3 shows the *leverage function* defined in Section 4. Leverage is an index of the sensitivity of the fitted model to the addition of new data points. At any spatial location u, the leverage $h(u)$ is a Taylor approximation to the change in the fitted intensity that would occur at location u if a new data point were added at that location:

$$h(u) \approx \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\})}(u) - \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x})}(u), \tag{2}$$

where $\widehat{\boldsymbol{\theta}}(\mathbf{x})$ denotes the parameter estimate based on the pattern $\mathbf{x}$, and $\mathbf{x} \cup \{u\}$ is the result of augmenting $\mathbf{x}$ by adding a new point at location u. Values of leverage are intensities, expressed as the mean number of points per unit area, in this case, points per square metre.

Figure 3 indicates that the top left and bottom right corners of the survey region have the highest leverage, and are therefore the most sensitive to the presence of data points, for this model.

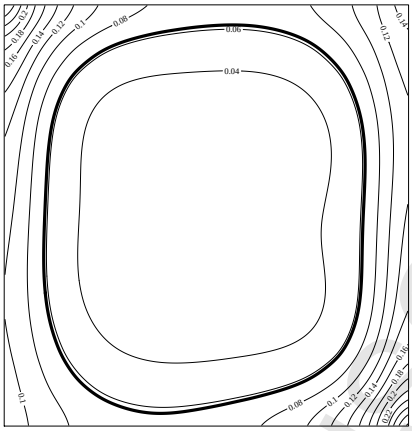


Figure 3: Contour map of the leverage function for a log-quadratic Poisson model (1) fitted to the Swedish Pines data by maximum likelihood. The thick line is the contour at the average value of leverage.

There is no "critical value" of leverage in the sense of statistical significance. In linear models with i.i.d. errors, the leverage is completely determined by the design matrix, which in turn is determined by the form of the model and by the covariate values. In generalized linear models and in spatial point process models, this statement is "almost" true: leverage depends *mainly* on the covariate functions and the form of the model, but also depends on the fitted parameter values (and hence on the observations, unlike the linear model setting). The dependence on the fitted parameters may place high or low importance on particular covariates.

Common practice in regression analysis is to treat any leverage value greater than the mean leverage as relatively "high". The thick black line in Figure 3 is the contour of leverage at a level equal to the average value of leverage over the survey region, namely 0.0625. High leverage is typically — but not always — associated with extreme values of the covariate, in this case, extreme values of the Cartesian coordinates.

Another possible benchmark in our case is the leverage value for the uniform Poisson process ("complete spatial randomness"). For such a process the effect (2) of adding one new data point would be to increase the estimated intensity $\hat{\lambda}$ by $1/96 = 0.01$ points per square metre. For the log-quadratic model, leverage

values in Figure 3 range from 0.03 to 0.3, indicating that this model is very sensitive to data in the corners of the study region.

2.2.2. *Influence*

Figure 4 depicts the (likelihood) *influence* function defined in Section 6. Influence is a case deletion diagnostic. The influence $s(x_i)$ of a data point $x_i \in \mathbf{x}$ is a Taylor approximation to the (negative) change in the log-likelihood of the data $\mathbf{x}$ that would occur if the parameter estimate $\widehat{\boldsymbol{\theta}}$ were based on the data excluding x_i:

$$s(x_i) \approx \frac{2}{p} \log \frac{L(\widehat{\boldsymbol{\theta}})}{L(\widehat{\boldsymbol{\theta}}_{-i})} \tag{3}$$

where p is the number of parameters, and $L(\boldsymbol{\theta})$ is the likelihood for parameter $\boldsymbol{\theta}$ based on the entire dataset $\mathbf{x}$, while $\widehat{\boldsymbol{\theta}}$ is the parameter estimate based on $\mathbf{x}$, and $\widehat{\boldsymbol{\theta}}_{-i} = \widehat{\boldsymbol{\theta}}(\mathbf{x} \setminus \{x_i\})$ is the parameter estimate obtained after deleting x_i from the point pattern.

The circles in Figure 4 are centred at the data points (locations of the observed trees), and the diameter of each circle is proportional to the influence value, as indicated on the scale at left. The influence values are dimensionless (log likelihood ratios multiplied by $2/p = 1/3$). Large influence values occur at some data points near the corners of the survey rectangle, as intuitively expected; Figure 4 shows that the fitted model is highly sensitive to the observed data in the corners of the survey region.

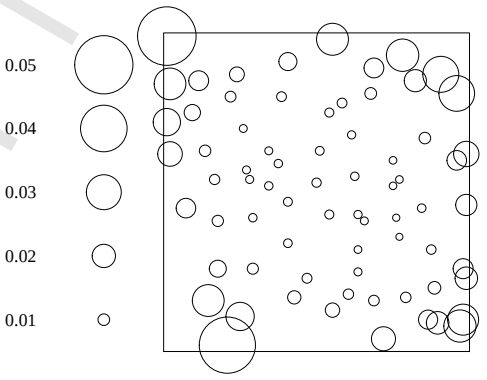


Figure 4: Influence for a log-quadratic Poisson model fitted to the Swedish Pines data by maximum likelihood.

The scale factor $2/p$ in (3) comes from the classical definition of likelihood influence [25] (see Appendix A) and corresponds to rescaling a likelihood ratio test statistic to have a null mean value equal to 1. Although a statistical significance interpretation is not really appropriate, influence values greater than 1 would be cause for concern. Another benchmark is the influence value for each data point in the uniform Poisson process: for this model the right-hand side of (3) is equal to $2(n\log(n/(n-1))-1) = 0.014$. Using this benchmark, many of the peripheral data points are highly influential for the log-quadratic model.

The influence reflects the change in the overall fit that would occur if a data point were omitted. It is therefore an index of both "anomaly" and "influence". As discussed in [30, Section 4.8, p. 74 ff.], observations may be overly influential because there are too few observations for the complexity of the model; because of data errors; because the covariate values are extreme; because the observations are genuinely anomalous; or for other reasons. Modelling strategies are discussed in [30, Section 4.10, p. 79 ff.].

2.2.3. Parameter influence `DFBETA`

The leverage plot in Figure 3 identified those parts of the survey region where the fitted model is most sensitive to new data. The influence plot in Figure 4 identified data points whose presence is most highly influential on the fitted model. Neither of these plots indicates *how* these data affect the fitted model.

Figure 5 depicts the *parameter influence measure* $\mathbb{D}\widehat{\boldsymbol{\theta}}$ defined in Section 5, which corresponds to the quantity commonly called `DFBETA` in regression analysis [30, p. 76]. The parameter influence describes the effect of data changes on the fitted parameter estimates, and is the most detailed diagnostic considered in this paper. Each panel in Figure 5 corresponds to one of the coefficients θ_i of the model and indicates the effect on the estimate of that coefficient.

The parameter influence is a *"spatial deletion"* diagnostic which describes the effect on the parameter estimates of deleting any specified subregion of the spatial domain W. "Deletion" of a subregion B signifies that any points of $\mathbf{x}$ which fall in B will be removed, but moreover that the covariate values $Z(u), u \in B$ associated with that region will also be removed from consideration. Formally, the likelihood or composite likelihood is redefined so that W is replaced by $W \backslash B$. Effectively we consider what would have changed if the survey region had not

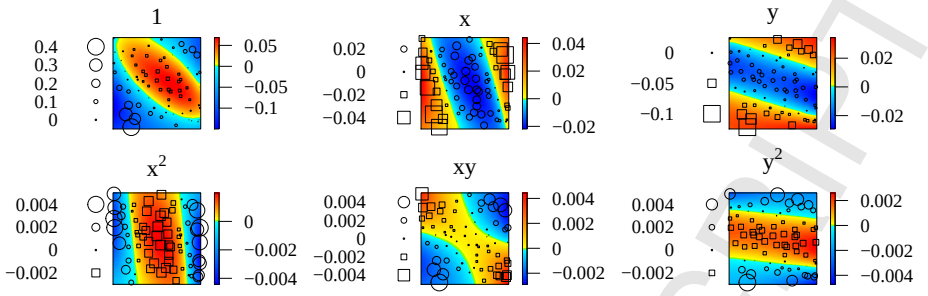


Figure 5: Parameter influence measure for log-quadratic Poisson model fitted to the Swedish Pines data by maximum likelihood.

included the subregion B.

Write $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x}, W)$ for the parameter estimate obtained from the data in the spatial domain W. For a subregion $B \subset W$, consider

$$(\nabla\widehat{\boldsymbol{\theta}})(B) = \widehat{\boldsymbol{\theta}}(\mathbf{x}, W) - \widehat{\boldsymbol{\theta}}(\mathbf{x} \setminus B, W \setminus B),$$

the (negative) change in the parameter estimate that would occur if the data inside B were deleted. We follow standard practice in calculating the change with the sign reversed [12, p. 11 ff.], [32, pp. 149–170], so that a positive value of $(\nabla\widehat{\boldsymbol{\theta}})(B)$ signifies that the data in B tend to increase the fitted parameter value — that is, we get a larger value of $\widehat{\boldsymbol{\theta}}$ if we *keep* these data.

The spatial deletion diagnostic $\mathbb{D}\widehat{\boldsymbol{\theta}}$ is a set function defined so that $(\mathbb{D}\widehat{\boldsymbol{\theta}})(B)$ is a Taylor approximation to $(\nabla\widehat{\boldsymbol{\theta}})(B)$. It is most easily described using the language of infinitesimals. For a location $u \in W$, consider an infinitesimal region $\mathrm{d}u$ centred around u, with infinitesimal area $|\mathrm{d}u|$. Then $(\mathbb{D}\widehat{\boldsymbol{\theta}})(\mathrm{d}u)$ is a Taylor approximation to $(\nabla\widehat{\boldsymbol{\theta}})(\mathrm{d}u)$ which takes the form

$$(\mathbb{D}\widehat{\boldsymbol{\theta}})(\mathrm{d}u) = h(u)N(\mathrm{d}u) + g(u)\,|\mathrm{d}u| \tag{4}$$

where h and g are explicit functions derived from the parametric form of the model, and $N(\mathrm{d}u)$ is the number of data points falling in $\mathrm{d}u$, which will be either 0 or 1 under the assumptions made in the paper. Summing all the infinitesimal contributions (4) over any desired set $B \subset W$ gives

$$(\mathbb{D}\widehat{\boldsymbol{\theta}})(B) = \sum_{x_i \in \mathbf{x} \cap B} h(x_i) + \int_B g(u)\,\mathrm{d}u. \tag{5}$$

Equations (4) and (5) are equivalent descriptions of the same diagnostic. Although we will use the integrated form (5), some readers may prefer to think in

terms of infinitesimal contributions as in (4).

The diagnostic has two components (given on the right hand side of (4) or (5)) which describe, respectively, the effect of deleting data points and of deleting regions that do not contain data points. For data point $x_i \in \mathbf{x}$, the term $h(x_i)$ is a Taylor approximation to the (negative) change in $\widehat{\boldsymbol{\theta}}$ that would occur if x_i were deleted, that is, $\widehat{\boldsymbol{\theta}}(\mathbf{x}) - \widehat{\boldsymbol{\theta}}(\mathbf{x} \setminus \{x_i\})$. This "discrete" component is depicted in Figure 5 using symbols in the form of circles and squares, with circles representing positive values and squares representing negative values. These values are expressed in the same units as the corresponding coefficient of the model. For example, values for the x coordinate panel are expressed in metre^{-1}. For each panel the symbol scale is indicated in the legend to the left of the panel.

The estimate $\widehat{\boldsymbol{\theta}}$ would also change if we deleted a subset of the survey region where *no* data points were observed. At a spatial location u which is not a data point, if we delete the infinitesimal region $\mathrm{d}u$, the change in $\widehat{\boldsymbol{\theta}}$ is infinitesimal: $\widehat{\boldsymbol{\theta}}(\mathbf{x}, W) - \widehat{\boldsymbol{\theta}}(\mathbf{x}, W \setminus \mathrm{d}u) = g(u)\,|\mathrm{d}u|$. The function $g(u)$ is the "density component" of the parameter influence measure. Background colours or shades of grey in each panel of Figure 5 represent the density function $g(u)$ for the corresponding parameter (with values shown by the colour ribbon at the right of the panel). Density values must be integrated over a region to obtain values on the same scale as the corresponding coefficient. In this case the window area is about 100 square metres, so a density value $g(u) = 0.01$ (say) over the whole window would integrate to about 1. Another way to say this, roughly, is that $g(u)$ gives the effect of deleting a *unit* area.

In order to reveal important detail, each panel in Figure 5 is plotted using a different colour map and symbol map.

Interpretation of the plot is based mainly on the relative sizes of the symbols and the relative values of colours, within a given panel. Some benchmark values are available. For the Intercept panel in Figure 5, one may refer to the expected behaviour for a homogeneous Poisson model, for which the parameter influence measure would have discrete mass $1/n = 1/71 = 0.014$ at each data point and density $-1/A = -1/96 = -0.010$. The density values in Figure 5 are of the same order as this reference value, but the discrete masses in the Intercept

panel of Figure 5 are up to 30 times the reference value. Before jumping to conclusions, we note that the values of parameter influence for the Intercept component would be different if the origin of spatial coordinates was shifted. This phenomenon is familiar from linear regression. Currently the origin is the bottom left corner; shifting the origin to the centre of the region would yield different results. For this reason, the Intercept panel is often ignored, as we would ignore the Intercept component of `DFBETA` in regression.

Looking at the top right-hand corner of each panel of Figure 5 we can conclude that the presence of the data points in the top right-hand corner of Figure 1 causes a decrease in the fitted coefficients of x and y (associated with square symbols in the panels for x and y) and an increase in all the other fitted coefficients (associated with circular symbols in the other panels). The background colours indicate that the inclusion of the non-data locations from this corner of the survey region tends to increase (reddish colours) the fitted coefficients of x and y, and decrease (bluish colours) the other fitted coefficients.

2.2.4. Interpretation and visualisation of measures

In this paper we always use the term "measure" in the technical sense: a measure m is a set function, that is, it assigns a value $m(B)$ to any set B, and is additive with respect to set unions. For those unfamiliar with this concept we offer some explanation for its practical use in this application.

A measure can best be understood as a spatial distribution of "mass" or "electric charge", in which the mass or charge may either be concentrated at individual points ("atoms") or spread diffusely over the space, or may combine both of these kinds of behaviour. Given a measure m we can evaluate, for any spatial domain B, the total "mass" or "charge" $m(B)$ falling inside B. Conversely, knowledge of the values of $m(B)$ for all domains B is sufficient to determine the measure m.

The point process residuals introduced in [9] are a measure in this sense.

The parameter effect measure $\mathbb{D}\widehat{\boldsymbol{\theta}}$ is a measure. For any chosen subset B, the value $(\mathbb{D}\widehat{\boldsymbol{\theta}})(B)$ gives the approximate effect of deleting all the data in B, that is, deleting those data points x_j which fall in B and removing the non-data locations $u \in B$ from the survey region, as the sum of the two contributions on the right hand side of (5). This sum could be difficult to judge visually

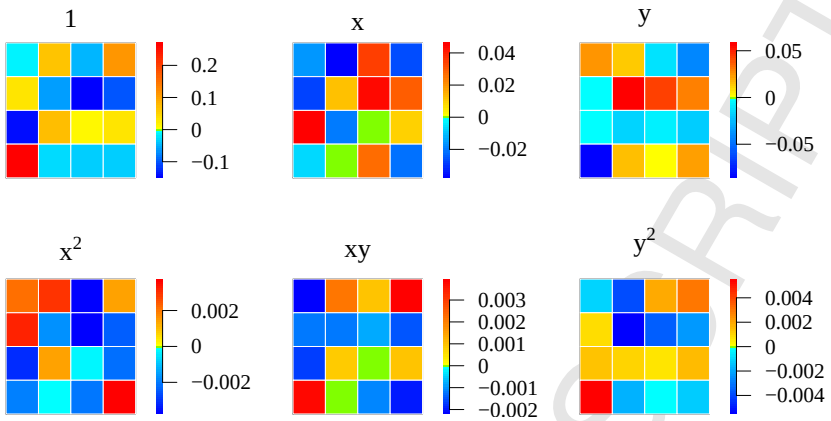


Figure 6: Values of the parameter influence measure for each tile in a 4×4 grid.

from Figure 5. We may prefer a plot like Figure 6 which shows the values $(\mathbb{D}\widehat{\boldsymbol{\theta}})(B)$ of the parameter effect measure for each tile B in a 4×4 grid of tiles across the survey region. The value for a given tile B is (5), the sum of point masses plus the integral of the density in this tile. The values in each panel are predicted changes to the corresponding coefficient $\widehat{\theta}_j$, expressed in the same units as θ_j. Instead of the rectangular tiles in Figure 6, one could use any regions of space which are meaningful in the application context, such as administrative subdivisions, or regions defined by distance from a reference point.

Another alternative for visualising a measure is to apply kernel smoothing, as shown in Figure 7. Using fixed-bandwidth smoothing with kernel ψ, the result of smoothing (5) is a density function

$$t(u) = \sum_{x_i \in \mathbf{x}} \psi(u - x_i) h(x_i) + \int_W \psi(u - u') g(u') \, \mathrm{d}u' \tag{6}$$

defined for all $u \in W$. Edge corrections may also be included. The smoothed function t can be interpreted as a density in the same way as g. Kernel smoothing avoids possible artefacts due to the sharp boundaries in a tessellation like Figure 6.

In summary, the parameter influence makes it possible to predict the sign and magnitude of the change in each fitted parameter that would occur if any chosen subset of the data were omitted. When scrutinising the data points $x_i \in \mathbf{x}$, one should first use the likelihood influence plot (Figure 4) to identify those data points which have a substantial influence on the fitted model, then

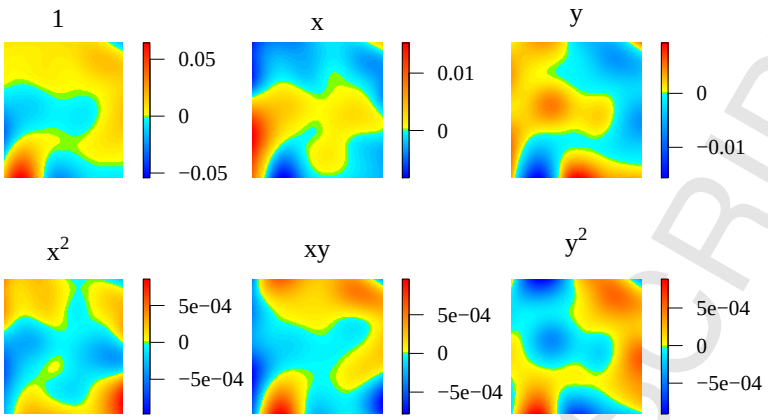


Figure 7: Kernel-smoothed density of parameter influence measure. Different colour scales are used in each panel.

use a parameter influence plot (Figure 5 or 6 or 7) to judge which of the fitted parameters is substantially changed, and in which direction, when that data point is omitted. When scrutinising the non-data locations, one may first use the leverage plot (Figure 3) to identify areas where the fitted intensity is highly sensitive to the presence of new data points, and then examine the parameter influence plots.

2.2.5. Effect change `DFFIT`

The parameter influence `DFBETA` (Figures 5–7) encapsulates the effect of data changes on the fitted parameters. To understand how these changes would affect the *predictions* of the fitted model, a strategy used for linear and generalized linear models is to multiply each component of `DFBETA` by the corresponding covariate value, to obtain the effect on the linear predictor. This is the *effect change* diagnostic, commonly known as `DFFIT` in linear regression [30, p. 76].

In this paper we define an analogue of `DFFIT` for point process models, the effect change measure, depicted in Figure 8. It was calculated by multiplying each numerical value (encoded as a symbol diameter or a colour) in Figure 5 by the value of the corresponding covariate at the same location. Each panel of Figure 8 represents the effect on a term in the linear predictor (i.e. the logarithm of the intensity λ) corresponding to one of the model parameters. The circle and square symbols are values on the scale of the linear predictor. Some of the circles have values as large as $+0.4$, which corresponds to increasing the fitted

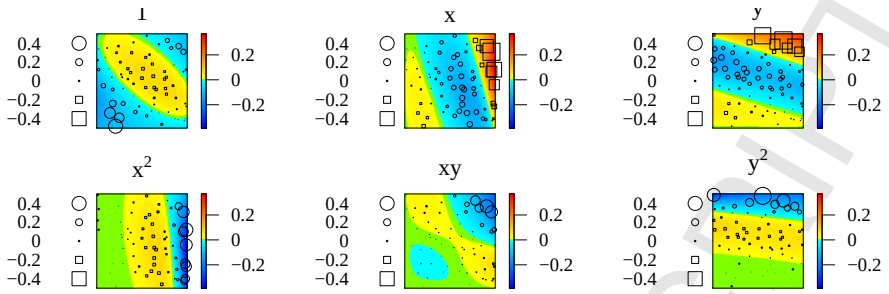


Figure 8: The DFFIT measure for a log-quadratic Poisson model fitted to the Swedish Pines data. Identical colour and symbol maps are used in all panels.

intensity by a factor of $\exp(0.4) \approx 1.5$. Some of the background colours reach values as low as -0.3, meaning that deletion of a unit area surrounding that location would reduce the fitted intensity by a factor $\exp(-0.3) \approx 0.75$.

The values of the discrete component in each panel of Figure 8 can be compared, since they are all on the scale of the linear predictor. The values of density in different panels can also be compared. Figure 8 is plotted using identical colour maps and identical symbol maps in each panel, in order to show the relative importance of each component. For example, the xy term is relatively unimportant, except in the top right corner of the window.

There is a data point near the middle of the upper boundary of Figure 1. From Figure 4 we see that this point has moderately large influence. Figure 8 shows that this data point has a negative effect on the y term in the linear predictor, a positive effect on the y^2 term, and negligible effect on other terms. A weakness of the effect measure is that it gives only the effect of deleting a data point on the predicted intensity at the *same* data point.

Figure 8 suggests that the strongest support for the non-stationary trend is provided by the data points in the top right corner.

Again it may be easier to interpret a plot similar to Figure 6 showing the DFFIT measure for each tile in a 4×4 grid of tiles covering the window, or a kernel-smoothed version similar to Figure 7.

2.3. Interaction between points

For simplicity, this overview has focused on a Poisson point process model for the Swedish Pines data. However this dataset shows strong evidence of a

"regular" arrangement or "inhibitory" interaction between points. One simple model for this interaction is the Strauss point process [50, 35], [5, pp. 497–500], an example of a pairwise interaction Gibbs point process [5, Chap. 13]. The Strauss model with constant intensity was first fitted to the Swedish Pines data in [45], [46, pp. 172–175].

If spatially-varying intensity is suspected as well, then in order to avoid Simpson's Paradox, the data should be analysed using a model that incorporates both spatially-varying intensity and interpoint interaction [5, pp. 503–506, 518, 529, 536]. The main goal of this paper is to extend the diagnostic tools presented above to this much larger class of Gibbs point process models. Accordingly we postpone further analysis of the Swedish Pines until Section 8, and begin the development of diagnostics for Gibbs models.

3. Notation and technical assumptions

This section begins the technical part of the paper. It defines the notation and records the assumptions we make about spatial point process models and the likelihoods used to fit them to data. It can mostly be skipped by readers who are not interested in technical details.

3.1. Data

The data consist of a spatial point pattern $\mathbf{x} = \{x_1, \ldots, x_n\}$ in a bounded region ('window') $W \subset \mathbb{R}^d$, where $x_i \in W$ for $i = 1, \ldots, n$, and the number of points $n \geq 0$ is not fixed in advance. We assume the points are distinct, $x_i \neq x_j$ for $i \neq j$.

There may also be covariate information of various kinds; we assume that this information is encoded into real-valued spatial covariate functions $Z_j(u), j = 1, \ldots, p$ defined at all spatial locations $u \in W$.

3.2. Models

The point pattern dataset $\mathbf{x}$ is assumed to be a realisation of a point process $\mathbf{X}$. We assume that, with probability 1, the total number of random points in $\mathbf{X}$ is finite. The sample space $\mathcal{X}$ of all possible realisations is the collection of all finite point patterns of distinct points, that is, finite subsets of W.

In order to define likelihoods and composite likelihoods, we assume the point process $\mathbf{X}$ has a probability density $f(\mathbf{x}), \mathbf{x} \in \mathcal{X}$. Although the details of the definition are not crucial, we formally define $f(\mathbf{x})$ as a density with respect to the Poisson process with unit intensity (rate) in W. This means that the expected value of any function $h(\mathbf{X})$ is given by $\mathbb{E}[h(\mathbf{X})] = \mathbb{E}[h(\mathbf{Y})f(\mathbf{Y})]$ where $\mathbf{Y}$ is the Poisson process with rate 1 in W. For further details, see [39, Section 6.1] or [5, Section 13.12].

3.2.1. Poisson models

A Poisson model postulates that $\mathbf{x}$ is a realisation of a Poisson point process $\mathbf{X}$ in W with intensity (rate) function $\lambda_{\boldsymbol{\theta}}(u)$, $u \in W$ where $\boldsymbol{\theta}$ is a p-dimensional parameter vector. In order that the expected number of points is finite, the integral $\int_W \lambda_{\boldsymbol{\theta}}(u)\,\mathrm{d}u$ must be finite. The likelihood is

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \left[\prod_{v\in\mathbf{x}} \lambda_{\boldsymbol{\theta}}(v)\right] \exp \int_W (1 - \lambda_{\boldsymbol{\theta}}(u))\,\mathrm{d}u, \qquad \mathbf{x} \in \mathcal{X}. \tag{7}$$

Likelihoods are defined only up to a constant factor, and the definition in (7) is calibrated so that the homogeneous Poisson process with unit rate $\lambda(u) \equiv 1$ has log-likelihood equal to zero.

In principle the intensity $\lambda_{\boldsymbol{\theta}}(u)$ could have any functional form, provided it is integrable. Regularity conditions are imposed in Section 3.4. A *loglinear* Poisson model postulates that

$$\lambda_{\boldsymbol{\theta}}(u) = \exp(\boldsymbol{\theta}^\top Z(u)) \tag{8}$$

where the "canonical covariate" $Z(u), u \in W$ is a p-dimensional vector valued function of spatial location. In this case, the likelihood (7) takes the form

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = c(\boldsymbol{\theta}) \exp(\boldsymbol{\theta}^\top \sum_{v\in\mathbf{x}} Z(v))$$

where $c(\boldsymbol{\theta})$ is the normalising constant; this likelihood is an exponential family model.

3.3. Gibbs models

A finite point process is a Gibbs process if its probability density $f(\mathbf{x})$ exists and has *hereditary positivity*, meaning that $f(\mathbf{x}) > 0$ implies $f(\mathbf{y}) > 0$ for all

sub-patterns $\mathbf{y} \subset \mathbf{x}$. See [15, 21] or [39, Chapter 6] for details. In particular, any Poisson process in W with an integrable intensity function is also a Gibbs process.

The class of Gibbs *processes* is so large that it embraces almost all useful models for finite point patterns. In that sense, the general results obtained below are very widely applicable. However, the explicit formulae for diagnostics depend on the feasibility of calculating particular terms in the model, and this may be a severe restriction on their scope.

We consider a Gibbs finite point process model with probability density $f_{\boldsymbol{\theta}}(\mathbf{x})$. For the vast majority of applications, we can assume an exponential family model

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = c(\boldsymbol{\theta}) \exp(\boldsymbol{\theta}^\top V(\mathbf{x})) \tag{9}$$

where the canonical sufficient statistic $V(\mathbf{x})$ is a p-dimensional vector valued function of the point pattern $\mathbf{x}$, and $c(\boldsymbol{\theta})$ is the (usually intractable) normalising constant. More generally we assume

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = c(\boldsymbol{\theta}) b(\mathbf{x}) \exp(\boldsymbol{\theta}^\top V(\mathbf{x})) \tag{10}$$

where $b(\mathbf{x}) \geq 0$ must have hereditary positivity. Numerous models are presented in [5, Chap. 13].

The main tool for modelling and model-fitting is the (Papangelou) *conditional intensity*

$$\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) = \frac{f_{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\})}{f_{\boldsymbol{\theta}}(\mathbf{x} \setminus \{u\})}, \quad u \in W. \tag{11}$$

This can be interpreted as the intensity at a location u given the existing configuration $\mathbf{x}$ at all other locations [5, Chapter 13]. The definition (11) is a convenient way to embrace the two cases

$$\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) = \begin{cases} f_{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\}) / f_{\boldsymbol{\theta}}(\mathbf{x}) & \text{if } u \notin \mathbf{x}, \\ f_{\boldsymbol{\theta}}(\mathbf{x}) / f_{\boldsymbol{\theta}}(\mathbf{x} \setminus \{u\}) & \text{if } u \in \mathbf{x}. \end{cases}$$

In the special case of a Poisson point process with intensity function $\lambda_{\boldsymbol{\theta}}(u)$, the conditional intensity $\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})$ is equivalent to the intensity $\lambda_{\boldsymbol{\theta}}(u)$.

Result 1. *The conditional intensity of a finite Gibbs process is* exvisible *[51], that is,* $\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) = \lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x} \setminus \{u\})$*, and has* hereditary positivity *in the sense that* $\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) > 0$ *implies* $\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{y}) > 0$ *for all* $\mathbf{y} \subset \mathbf{x}$.

In the exponential family Gibbs model (9), the conditional intensity is log-linear in $\boldsymbol{\theta}$,

$$\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = \exp(\boldsymbol{\theta}^\top Z(u\,|\,\mathbf{x})), \tag{12}$$

where $Z(u\,|\,\mathbf{x}) = V(\mathbf{x}\cup\{u\}) - V(\mathbf{x}\setminus\{u\})$ is exvisible by construction. The conditional intensity is convenient for modelling because the often intractable normalising constant $c(\boldsymbol{\theta})$ cancels out in the ratio (11). Similarly for (10) we have

$$\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = m(u\,|\,\mathbf{x})\ \exp(\boldsymbol{\theta}^\top Z(u\,|\,\mathbf{x})) \tag{13}$$

where $m(u\,|\,\mathbf{x}) = b(\mathbf{x}\cup\{u\})/b(\mathbf{x}\setminus\{u\})$ is exvisible and we use the convention $0/0 = 0$.

Gibbs models used for practical data analysis usually have finite interaction range which we denote by R. Having this property means that

$$\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = \lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}\cap D(u,R)) \tag{14}$$

for all configurations $\mathbf{x}$ and all $u\in W$, where $D(u,R)$ denotes the disc of radius $R>0$ centred at u.

The Poisson intensity $\lambda_{\boldsymbol{\theta}}(u)$ or Gibbs conditional intensity $\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ may also depend on nuisance parameters. For the moment we assume that these are held fixed.

3.4. Regularity conditions in the general case

In the most general case, we need the following assumptions.

(A1) the conditional intensity $\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ is twice differentiable with respect to $\boldsymbol{\theta}\in\Theta$ for all fixed u and $\mathbf{x}$, where Θ is an open subset of $\mathbb{R}^p$.

(A2) for all $\boldsymbol{\theta}\in\Theta$, either $\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})>0$ everywhere, or more generally

$$\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = m(u\,|\,\mathbf{x})\lambda^+_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) \tag{15}$$

where $m(u\,|\,\mathbf{x})$ takes only the values 0 and 1, and $\lambda^+_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ is positive everywhere and is twice differentiable with respect to $\boldsymbol{\theta}$;

(A3) the first and second derivatives of $\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ with respect to $\boldsymbol{\theta}$ are absolutely integrable with respect to u over W, for each fixed $\mathbf{x}$ and $\boldsymbol{\theta}\in\Theta$.

Assumption (A2) is needed because many popular Gibbs models include a "hard core" interaction term causing the conditional intensity to take the value zero at some locations. Equation (15) states that this hard core term does not depend on the parameter $\boldsymbol{\theta}$. The practical implication is that any parameters governing the hard core interaction are held fixed.

The form of (15) and Result 1 imply that both $m(u\,|\,\mathbf{x})$ and $\lambda^+_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ must be exvisible, that is, $m(u\,|\,\mathbf{x}) = m(u\,|\,\mathbf{x}\setminus\{u\})$ and $\lambda^+_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = \lambda^+_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}\setminus\{u\})$, and also that $m(u|\mathbf{x})$ has hereditary positivity, $m(u|\mathbf{x}) > 0$ implies $m(u|\mathbf{y}) > 0$ for all $\mathbf{y} \subset \mathbf{x}$.

Assuming (A1)–(A3) define

$$\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) \;=\; \frac{\partial}{\partial\boldsymbol{\theta}} \log \lambda^+_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}), \tag{16}$$

$$\kappa_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) \;=\; \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log \lambda^+_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}). \tag{17}$$

It follows that

$$\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) \;=\; \zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\,\lambda_{\boldsymbol{\theta}}(u|\mathbf{x}), \tag{18}$$

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) \;=\; \left[\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})^\top + \kappa_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\right]\,\lambda_{\boldsymbol{\theta}}(u|\mathbf{x}). \tag{19}$$

The functions $\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ and $\kappa_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ are exvisible. In an exponential family model (9) or (10) which respectively imply loglinear conditional intensity (12) or (13), we have $\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = Z(u\,|\,\mathbf{x})$ and $\kappa_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = \mathbf{0}$. (Note that $\mathbf{0}$ denotes a matrix of zeroes here. Elsewhere it may denote a zero vector.)

3.5. Likelihoods and composite likelihoods

The explicit form of the leverage and influence diagnostics will depend on the method used to fit the model, because these diagnostics are based on Taylor approximations to the composite likelihood and its derivatives. Accordingly, here we list the main choices of composite likelihood for Poisson and Gibbs point process models.

For a Poisson point process with intensity $\lambda_{\boldsymbol{\theta}}(u), u \in W$, the loglikelihood is, from (7) up to an additive constant,

$$\log \mathsf{L}(\boldsymbol{\theta}, \mathbf{x}) = \sum_{v\in\mathbf{x}} \log \lambda_{\boldsymbol{\theta}}(v) - \int_W \lambda_{\boldsymbol{\theta}}(u)\,\mathrm{d}u. \tag{20}$$

Throughout this paper we shall use the letter v to represent a data point while u represents any spatial location.

An alternative choice is the logistic conditional likelihood [3] constructed by generating a Poisson process D of sample points ("dummy" or non-data points), with known intensity function $\rho(u) > 0$, then conditioning on the locations of the superimposed data and dummy points, and forming the conditional loglikelihood of the data:

$$\begin{aligned} \log \mathsf{LL}(\boldsymbol{\theta}; \mathbf{x}, D) &= \sum_{v \in \mathbf{x}} \log \left(\frac{\lambda_{\boldsymbol{\theta}}(v)}{\lambda_{\boldsymbol{\theta}}(v) + \rho(v)} \right) + \sum_{u \in D} \log \left(\frac{\rho(u)}{\lambda_{\boldsymbol{\theta}}(u) + \rho(u)} \right) \\ &= \sum_{v \in \mathbf{x}} \log p_{\boldsymbol{\theta}}(v) + \sum_{u \in D} \log(1 - p_{\boldsymbol{\theta}}(u)) \end{aligned} \tag{21}$$

where $\rho(u)$ is the intensity of the dummy process and

$$p_{\boldsymbol{\theta}}(u) = \frac{\lambda_{\boldsymbol{\theta}}(u)}{\lambda_{\boldsymbol{\theta}}(u) + \rho(u)} \tag{22}$$

is the conditional probability that a point of $\mathbf{x} \cup D$ at location u belongs to $\mathbf{x}$. Thus $\mathsf{LL}(\boldsymbol{\theta}; \mathbf{x}, D)$ is the conditional likelihood, given the locations of the combined pattern of data and dummy points, of the data/dummy status of each point. The logistic conditional likelihood for Poisson models is used frequently in Geographical Information Systems for computational efficiency purposes, and used occasionally in spatial statistics because of inferential advantages [24, 3], [5, Section 9.10, pp. 355–359].

For a Gibbs point process with conditional intensity $\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})$, Besag's [14] log *pseudolikelihood* is

$$\log \mathsf{PL}\,(\boldsymbol{\theta}; \mathbf{x}) = \sum_{v \in \mathbf{x}_\ominus} \log \lambda_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) - \int_{W_\ominus} \lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})\, \mathrm{d}u, \tag{23}$$

where $W_\ominus \subseteq W$ is a designated subset of the observation window W, and $\mathbf{x}_\ominus = \mathbf{x} \cap W_\ominus$. There are several versions of the pseudolikelihood, all taking the same common form (23), which use different corrections for edge effects. A common example is the *border correction* in which

$$W_\ominus = \{u \in W : \, d(u, W^c) \geq R\} \tag{24}$$

is the subset of W lying at least R units away from the complement of W, where R is a threshold distance; setting R equal to the interaction range of the

model defined in (14) implies that the pseudolikelihood (23) is computable from information observable inside W. In most other types of edge correction, $W_\ominus$ is simply equal to W, and $\lambda_{\boldsymbol{\theta}}(u \mid \mathbf{x})$ is replaced by a modified version of the conditional intensity including an edge effect weighting factor. Details of these edge corrections are given in [6]. Here it suffices to assume the general form (23).

The log pseudolikelihood has the same algebraic form as the Poisson loglikelihood (20), and reduces to the loglikelihood (up to a constant) if the model is Poisson. Practical methods for fitting point process models by maximum likelihood and maximum pseudolikelihood were developed in [13] and [6] respectively. Edge corrections are described in detail in [6]. Likewise the logistic conditional likelihood (21) can be extended to Gibbs point process models, as developed in [17, 3]:

$$\begin{aligned} \log \mathsf{LL}(\boldsymbol{\theta}; \mathbf{x}, D) &= \sum_{v \in \mathbf{x}_\ominus} \log\left(\frac{\lambda_{\boldsymbol{\theta}}(v \,|\, \mathbf{x})}{\lambda_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) + \rho(v)}\right) + \sum_{u \in D_\ominus} \log\left(\frac{\rho(u)}{\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) + \rho(u)}\right) \\ &= \sum_{v \in \mathbf{x}_\ominus} \log p_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) + \sum_{u \in D_\ominus} \log(1 - p_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})), \end{aligned} \tag{25}$$

where $D_\ominus = D \cap W_\ominus$, while $\rho(u)$ is again the intensity of the dummy process, and

$$p_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) = \frac{\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})}{\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) + \rho(u)} \tag{26}$$

is the analogue of the mean in logistic regression, generalising (22). The properties of this composite likelihood, its statistical advantages, and fitting algorithms are discussed in [3], [5, Section 13.13.7, pp. 556–557].

Other composite likelihoods are sometimes used, including regularised versions of the composite likelihoods above, and products of composite likelihoods for hierarchical interaction point process models [31, 28, 34]. Our results can be extended to these "composite composite likelihoods". Regularised composite likelihoods are discussed in Appendix B. Results for hierarchical composite likelihoods can be deduced from our results below, but are omitted for brevity.

3.6. Composite score and sensitivity

The Gibbs point process model with conditional intensity $\lambda_{\boldsymbol{\theta}}(u \mid \mathbf{x})$ is assumed to have been fitted to the data $\mathbf{x}$ by maximising a composite likelihood

$\mathsf{CL}(\boldsymbol{\theta}, \mathbf{x})$ yielding parameter estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$. We make the following regularity assumptions:

(C1) The composite loglikelihood $\log \mathsf{CL}(\boldsymbol{\theta}, \mathbf{x})$ is twice differentiable with respect to $\boldsymbol{\theta}$ in a neighbourhood of $\widehat{\boldsymbol{\theta}}$, for the given data $\mathbf{x}$. We define the composite score

$$U(\boldsymbol{\theta}, \mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathsf{CL}(\boldsymbol{\theta}, \mathbf{x}) \tag{27}$$

and the negative Hessian

$$H(\boldsymbol{\theta}, \mathbf{x}) = -\frac{\partial}{\partial \boldsymbol{\theta}} U(\boldsymbol{\theta}, \mathbf{x})^\top. \tag{28}$$

(C2) The maximum composite likelihood is achieved at a stationary point, that is, $\widehat{\boldsymbol{\theta}}$ is a solution of the composite score equation

$$U(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{0}. \tag{29}$$

(C3) The negative Hessian $H(\widehat{\boldsymbol{\theta}}, \mathbf{x})$ is positive definite.

It can be verified directly that each of the composite likelihoods listed in Section 3.5 satisfies (C1) for any $\mathbf{x}$ if the regularity conditions (A1)–(A2) hold.

Note that the conditions (C2)–(C3) only need to hold for the dataset $\mathbf{x}$, not for all possible realisations. In effect this excludes trivial cases (such as an empty point pattern) where the parameters are unidentifiable.

The following two results are straightforwardly obtained by first principles.

Result 2. *For a Poisson process with intensity $\lambda_{\boldsymbol{\theta}}(u)$ fitted by maximum likelihood, under regularity conditions (A1)–(A3) the likelihood score is*

$$U(\boldsymbol{\theta}, \mathbf{x}) = \sum_{v \in \mathbf{x}} \zeta_{\boldsymbol{\theta}}(v) - \int_W \zeta_{\boldsymbol{\theta}}(u) \lambda_{\boldsymbol{\theta}}(u) \, \mathrm{d}u \tag{30}$$

and the negative Hessian is

$$H(\boldsymbol{\theta}, \mathbf{x}) = -\sum_{v \in \mathbf{x}} \kappa_{\boldsymbol{\theta}}(v) + \int_W \left[\zeta_{\boldsymbol{\theta}}(u) \zeta_{\boldsymbol{\theta}}(u)^\top + \kappa_{\boldsymbol{\theta}}(u)\right] \lambda_{\boldsymbol{\theta}}(u) \, \mathrm{d}u. \tag{31}$$

In an exponential family model (9) *or* (10) *we have $\zeta_{\boldsymbol{\theta}}(u) = Z(u)$ and $\kappa_{\boldsymbol{\theta}}(u) \equiv \mathbf{0}$ so that*

$$H(\boldsymbol{\theta}, \mathbf{x}) = H(\boldsymbol{\theta}) = \int_W Z(u) Z(u)^\top \lambda_{\boldsymbol{\theta}}(u) \, \mathrm{d}u$$

coincides with the Fisher information.

Result 3. *For a Gibbs point process model, if the composite likelihood is Besag's pseudolikelihood* (23)*, then under regularity conditions (A1)–(A3) the composite score is*

$$U(\boldsymbol{\theta};\mathbf{x}) = \sum_{v\in\mathbf{x}_\ominus} \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) - \int_{W_\ominus} \zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\,\mathrm{d}u \tag{32}$$

and the negative Hessian is

$$H(\boldsymbol{\theta};\mathbf{x}) = -\sum_{v\in\mathbf{x}_\ominus} \kappa_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) + \int_{W_\ominus} \left(\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})^\top + \kappa_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\right)\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\,\mathrm{d}u. \tag{33}$$

For the exponential family model (9) *with loglinear conditional intensity* (12)*, these reduce to*

$$U(\boldsymbol{\theta},\mathbf{x}) = \sum_{v\in\mathbf{x}_\ominus} Z(v\,|\,\mathbf{x}) - \int_{W_\ominus} Z(u\,|\,\mathbf{x})\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\,\mathrm{d}u \tag{34}$$

$$H(\boldsymbol{\theta},\mathbf{x}) = \int_{W_\ominus} Z(u\,|\,\mathbf{x})Z(u\,|\,\mathbf{x})^\top\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\,\mathrm{d}u. \tag{35}$$

The logistic composite likelihood (25) depends on the randomly-generated dummy points as well as the observed data points. We shall analyse the composite likelihood conditionally on both data and dummy points, that is, we treat the dummy points as fixed when defining diagnostics.

Result 4. *For a Gibbs point process model using the logistic composite likelihood* (25)*, under regularity conditions (A1)–(A3) the composite score is*

$$U(\boldsymbol{\theta};\mathbf{x},D) = \sum_{v\in\mathbf{x}_\ominus} \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) - \sum_{u\in\mathbf{x}_\ominus\cup D_\ominus} \zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) \tag{36}$$

and the negative Hessian is

$$\begin{aligned} H(\boldsymbol{\theta};\mathbf{x},D) = &\sum_{u\in\mathbf{x}_\ominus\cup D_\ominus} p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})(1-p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}))\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})^\top \\ &+ \sum_{u\in\mathbf{x}_\ominus\cup D_\ominus} p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})\kappa_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) - \sum_{v\in\mathbf{x}_\ominus} \kappa_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}). \end{aligned} \tag{37}$$

In the exponential family model (9) *with loglinear conditional intensity* (12)*, these reduce to*

$$U_W(\boldsymbol{\theta},\mathbf{x},D) = \sum_{v\in\mathbf{x}_\ominus} Z(v\,|\,\mathbf{x}) - \sum_{v\in\mathbf{x}_\ominus\cup D_\ominus} Z(v\,|\,\mathbf{x})p_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) \tag{38}$$

$$H_W(\boldsymbol{\theta},\mathbf{x},D) = \sum_{v\in\mathbf{x}_\ominus\cup D_\ominus} Z(v\,|\,\mathbf{x})Z(v\,|\,\mathbf{x})^\top p_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\,(1-p_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})). \tag{39}$$

In the special case of a Poisson process with intensity $\lambda_{\boldsymbol{\theta}}(u)$, these results take the simpler form in which the conditioning on $\mathbf{x}$ is dropped: that is, $\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ is replaced by $\zeta_{\boldsymbol{\theta}}(u)$, $p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ is replaced by $p_{\boldsymbol{\theta}}(u)$, $\kappa_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ is replaced by $\kappa_{\boldsymbol{\theta}}(u)$ and $Z(u|\mathbf{x})$ is replaced by $Z(u)$.

To prove this result, we simply use elementary calculus which gives

$$\tfrac{\partial}{\partial\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = p_{\boldsymbol{\theta}}(u|\mathbf{x})(1 - p_{\boldsymbol{\theta}}(u|\mathbf{x}))\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$$

so that $(\partial/\partial\boldsymbol{\theta}) \log p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = (1 - p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}))\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ and $(\partial/\partial\boldsymbol{\theta}) \log(1 - p_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})) = -p_{\boldsymbol{\theta}}(u|\mathbf{x})\zeta_{\boldsymbol{\theta}}(u|\mathbf{x})$. Differentiating (25) gives

$$\begin{aligned} U(\boldsymbol{\theta}; \mathbf{x}, D) &= \sum_{v \in \mathbf{x}_\ominus} \frac{\partial}{\partial\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(u|\mathbf{x}) + \sum_{u \in D_\ominus} \frac{\partial}{\partial\boldsymbol{\theta}} \log(1 - p_{\boldsymbol{\theta}}(u|\mathbf{x})) \\ &= \sum_{v \in \mathbf{x}_\ominus} \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\,(1 - p_{\boldsymbol{\theta}}(u|\mathbf{x})) - \sum_{u \in D_\ominus} \zeta_{\boldsymbol{\theta}}(u|\mathbf{x}) p_{\boldsymbol{\theta}}(u|\mathbf{x}) \\ &= \sum_{v \in \mathbf{x}_\ominus} \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) - \sum_{u \in \mathbf{x}_\ominus \cup D_\ominus} p_{\boldsymbol{\theta}}(u|\mathbf{x})\zeta_{\boldsymbol{\theta}}(u|\mathbf{x}) \end{aligned}$$

i.e. gives (36). Differentiation of (36) then yields (37).

4. Leverage in a point process model

This section begins the core material of the paper in which we define the diagnostics and give explicit expressions for them.

The earlier paper [2] gave detailed derivations of the leverage and influence diagnostics for the case of a Poisson process with loglinear intensity $\lambda_{\boldsymbol{\theta}}(u) = \exp(\boldsymbol{\theta}^\top Z(u))$, where $Z(u)$ is a fixed, known vector-valued function, and where the model is fitted by maximum likelihood.

The extension of these results to a *Gibbs* point process model fitted by maximum pseudolikelihood was discussed briefly in [2, Section 6.4] and [5, page 544] where explicit formulae for the diagnostics were presented, without proof. However, these formulae were partially incorrect, as were Figures 7 and 8 in [2] and Figures 13.36–13.38 in [5, p. 545–547]. The present paper gives the corrected results. Corrected figures are provided in an online supplement.

Leverage and influence for point process models were formally defined in [2] as derivatives, with respect to the data, of properties of the fitted model. In this paper we shall give a more accessible but less rigorous derivation of the diagnostics as Taylor approximations to properties of the model.

4.1. General definition of leverage

Consider any Gibbs point process model with conditional intensity $\lambda(u\,|\,\mathbf{x}) = \lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ at location u for configuration $\mathbf{x}$. When a model is fitted to the data $\mathbf{x}$, we obtain parameter estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$, and the fitted conditional intensity is $\widehat{\lambda}(u\,|\,\mathbf{x}) = \lambda_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x})$.

If the point pattern $\mathbf{x}$ is changed by adding a new point at location u, the fitted conditional intensity at the same location u is changed by an amount

$$\lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x}\cup\{u\})}(u\,|\,\mathbf{x}\cup\{u\}) - \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x})}(u\,|\,\mathbf{x}) = \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x}\cup\{u\})}(u\,|\,\mathbf{x}) - \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x})}(u\,|\,\mathbf{x}), \qquad (40)$$

where the right hand side follows because of the exvisibility of the conditional intensity.

The *leverage* is a function $h(u)$ giving a Taylor approximation to (40). To define it we introduce some notation from spatial statistics [4, 3, 18].

Definition 1. *For a real-valued or vector-valued function g defined for point patterns $\mathbf{x} \in \mathcal{X}$, the* difference operator Δ_u *is defined for each location $u \in W$ by*

$$\Delta_u g(\mathbf{x}) = g(\mathbf{x}\cup\{u\}) - g(\mathbf{x}\setminus\{u\}) = \begin{cases} g(\mathbf{x}\cup\{u\}) - g(\mathbf{x}), & \text{for } u \notin \mathbf{x} \\ g(\mathbf{x}) - g(\mathbf{x}\setminus\{u\}) & \text{for } u \in \mathbf{x}. \end{cases} \qquad (41)$$

Similarly if $g(u, \mathbf{x})$ is a function defined for locations $u \in W$ and point patterns $\mathbf{x} \in \mathcal{X}$, we define for $u' \in W$

$$\Delta_u g(u', \mathbf{x}) = g(u', \mathbf{x}\cup\{u\}) - g(u', \mathbf{x}\setminus\{u\}). \qquad (42)$$

Note that Δ_u is the effect of *adding* a data point; it is conceptually different from the effect of "case deletion". Case deletion is often denoted by Δ in the literature on generalized linear models [32] but in this paper we shall use the symbol ∇ (defined in Section 5).

The leverage $h(u)$ will be defined as a Taylor approximation to $\Delta_u \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x})}(u\,|\,\mathbf{x})$ for each $u \in W$. The relation (40) greatly simplifies this calculation. Applying the chain rule to the right-hand side of (40), we obtain the first order Taylor approximation which is

$$\Delta_u \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x})}(u\,|\,\mathbf{x}) \approx \left[\frac{\partial}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) \right] \Delta_u \widehat{\boldsymbol{\theta}}(\mathbf{x}) = \lambda_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x}) \zeta_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x}) \Delta_u \widehat{\boldsymbol{\theta}}(\mathbf{x}). \qquad (43)$$

To approximate $\Delta_u \widehat{\boldsymbol{\theta}}(\mathbf{x})$ we expand the composite score $U(\boldsymbol{\theta}, \mathbf{x})$ about $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}(\mathbf{x})$ giving

$$U(\widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\}), \mathbf{x}) - U(\widehat{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{x}) \approx -H(\widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\}), \mathbf{x})\ \Delta_u \widehat{\boldsymbol{\theta}}(\mathbf{x}). \tag{44}$$

Since $\widehat{\boldsymbol{\theta}}$ is the solution of the composite score equation (29), the left side of (44) can be rewritten

$$\begin{aligned} U(\widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\}), \mathbf{x}) - U(\widehat{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{x}) &= U(\widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\}), \mathbf{x}) - 0 \\ &= U(\widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\}), \mathbf{x}) - U(\widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\}), \mathbf{x} \cup \{u\}) \\ &= -\Delta_u U(\tilde{\boldsymbol{\theta}}, \mathbf{x}) \end{aligned}$$

where $\tilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\})$ is held fixed. This gives the approximation $\Delta_u \widehat{\boldsymbol{\theta}}(\mathbf{x}) \approx H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1} \Delta_u U(\tilde{\boldsymbol{\theta}}, \mathbf{x})$, where $\Delta_u U(\tilde{\boldsymbol{\theta}}, \mathbf{x})$ denotes the value of $\Delta_u U(\boldsymbol{\theta}, \mathbf{x})$ when $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ is held fixed. Alternatively, exchanging the roles of $\widehat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ yields the approximation

$$\Delta_u \widehat{\boldsymbol{\theta}}(\mathbf{x}) \approx H(\tilde{\boldsymbol{\theta}}, \mathbf{x})^{-1}\ \Delta_u U(\widehat{\boldsymbol{\theta}}, \mathbf{x}). \tag{45}$$

Further approximating $H(\tilde{\boldsymbol{\theta}}, \mathbf{x}) \approx H(\widehat{\boldsymbol{\theta}}, \mathbf{x})$ motivates the following definition.

Definition 2. *Consider a Gibbs point process model with conditional intensity $\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})$, fitted using the estimating function $U(\boldsymbol{\theta}, \mathbf{x})$, and satisfying regularity conditions (A1)–(A3) and (C1)–(C3). The (standardised)* leverage *value at location u is the first order approximation to $\Delta_u \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x})}(u \,|\, \mathbf{x})$ given by*

$$h(u) = \lambda_{\widehat{\boldsymbol{\theta}}}(u \,|\, \mathbf{x}) \zeta_{\widehat{\boldsymbol{\theta}}}(u \,|\, \mathbf{x})^{\top}\ H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}\ \Delta_u U(\widehat{\boldsymbol{\theta}}, \mathbf{x}), \tag{46}$$

where $H(\boldsymbol{\theta}, \mathbf{x})$ is the negative Hessian defined in (28)*.*

In theory, Definition 2 is very general, since almost all interesting point process models satisfy the Gibbs property, and the estimating function U is very general. In practice, the scope of application is a much narrower range of "tractable" Gibbs models for which we can find computable expressions for $H(\boldsymbol{\theta}, \mathbf{x})$ and $\Delta_u U(\boldsymbol{\theta}, \mathbf{x})$. This scope is still quite broad, as it includes all the Poisson and Gibbs models presented in [5, Chapters 9 and 13].

4.2. Leverage for Poisson likelihood

For a Poisson process model with intensity $\lambda_{\boldsymbol{\theta}}(u)$ fitted by maximum likelihood, the likelihood score and negative Hessian are given by (30) and (31) above.

The effect on the score of adding a new point at u is trivially $\Delta_u U(\boldsymbol{\theta}, \mathbf{x}) = \zeta_{\boldsymbol{\theta}}(u)$, so we get the following result, stated in [2]:

Result 5. *For a Poisson process model with intensity $\lambda_{\boldsymbol{\theta}}(u)$, fitted by maximum likelihood, if regularity conditions (A1)–(A3) and (C1)–(C3) hold then the leverage function* (46) *reduces to*

$$h(u) = \lambda_{\widehat{\boldsymbol{\theta}}}(u) \zeta_{\widehat{\boldsymbol{\theta}}}(u)^\top \, H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1} \, \zeta_{\widehat{\boldsymbol{\theta}}}(u). \tag{47}$$

Since $H(\widehat{\boldsymbol{\theta}}, \mathbf{x})$ is positive definite, $h(u) \geq 0$ for all u. In the loglinear model $\lambda_{\boldsymbol{\theta}}(u) = \exp(\boldsymbol{\theta}^\top Z(u))$, the leverage is

$$h(u) = \lambda_{\widehat{\boldsymbol{\theta}}}(u) Z(u)^\top \, H(\widehat{\boldsymbol{\theta}})^{-1} \, Z(u). \tag{48}$$

4.3. Interpretation of leverage

Interpretation of the leverage function for a Poisson model was discussed in Section 2.2.1. Those comments remain true for Gibbs models.

Leverage values are expressed in the same units as the conditional intensity, namely length^{-d} (number of points per unit volume). The values $h(u)$ can be interpreted as approximations to the change in intensity $\Delta_u \lambda_{\widehat{\boldsymbol{\theta}}}(u \mid \mathbf{x}) = \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x} \cup \{u\})}(u | \mathbf{x} \cup \{u\}) - \lambda_{\widehat{\boldsymbol{\theta}}(\mathbf{x})}(u | \mathbf{x})$.

Common practice in linear models is to declare a leverage value to be "large" if it exceeds the average leverage of all observations. In the loglinear Poisson case the average leverage is $(1/|W|) \int_W h(u) \, \mathrm{d}u = p/|W|$ where $|W|$ is the volume of W, using a matrix trace formula [2, eq. (23)]. We have not been able to obtain a simple expression for the average leverage in the Gibbs case.

In simple linear regression, the leverage is highest when the explanatory variable is most extreme. However, this is not necessarily true for generalized linear models: the claim in [42] was refuted in [37, p. 117]; see [32, p. 153 ff.]. Likewise, it is not true for point process models. Consider the special case where there is a single scalar explanatory variable $z(u)$ and we fit the loglinear Poisson model $\lambda_{\boldsymbol{\theta}}(u) = \exp(\theta_0 + \theta_1 z(u))$. Then the leverage (48) is a function of the covariate $z(u)$, of the form $h(z) = (a + bz + cz^2) \exp(\hat{\theta}_0 + \hat{\theta}_1 z)$. Depending on the values of the fitted coefficients, on the variance terms a, b, c and on the range of z values, the maximum value of $h(z)$ may occur at one or both of the extremes of z, or at a stationary point at some intermediate value of z. A detailed example is given in the online supplementary material.

4.4. Leverage for pseudolikelihood

For a Gibbs model fitted by maximum pseudolikelihood, the composite score $U(\boldsymbol{\theta}, \mathbf{x})$ is given by (32). The following result is obtained from first principles, using exvisibility.

Result 6. *Consider a Gibbs model with conditional intensity $\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ fitted by maximum pseudolikelihood. Under regularity conditions (A1)–(A3) and (C2)–(C3), the leverage is* (46) *with*

$$\Delta_u U(\widehat{\boldsymbol{\theta}}, \mathbf{x}) = \mathbf{1}\{u \in W_\ominus\}\, \zeta_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x}) + \sum_{v \in \mathbf{x}_\ominus} \Delta_u \zeta_{\widehat{\boldsymbol{\theta}}}(v\,|\,\mathbf{x}) - \int_{W_\ominus} \Delta_u \xi_{\widehat{\boldsymbol{\theta}}}(v\,|\,\mathbf{x})\, \mathrm{d}v, \quad (49)$$

where $\xi_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) = \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) = (\partial/\partial\boldsymbol{\theta})\lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})$.

It is instructive to prove this result from first principles:

$$\begin{aligned}
\Delta_u U(\boldsymbol{\theta}, \mathbf{x}) &= U(\boldsymbol{\theta}, \mathbf{x} \cup \{u\}) - U(\boldsymbol{\theta}, \mathbf{x} \setminus \{u\}) \\
&= \sum_{v \in \mathbf{x}_\ominus} \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\}) + \mathbf{1}\{u \in W_\ominus\}\, \zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x} \cup \{u\}) \\
&\quad - \int_{W_\ominus} \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\}) \lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\})\, \mathrm{d}v \\
&\quad - \sum_{v \in \mathbf{x}_\ominus} \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) + \int_{W_\ominus} \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \setminus \{u\}) \lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \setminus \{u\})\, \mathrm{d}v \\
&= \mathbf{1}\{u \in W_\ominus\}\, \zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) + \sum_{v \in \mathbf{x}_\ominus} [\zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\}) - \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \setminus \{u\})] \\
&\quad - \int_{W_\ominus} [\zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\}) \lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\}) - \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \setminus \{u\}) \lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \setminus \{u\})]\, \mathrm{d}v.
\end{aligned}$$

This is equivalent to (49), proving the result.

For a heuristic interpretation of (49), we note that the addition of a new point u to the dataset $\mathbf{x}$ gives rise to an extra term $\zeta_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ in the sum in (32) provided $u \in W_\ominus$. It also changes the values of the existing summands in (32) from $\zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})$ to $\zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\})$, giving rise to the second term on the right of (49). The third term on the right of (49) is the effect on the integral in (32). Note that $\Delta_u \xi_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})$ expands to

$$\Delta_u \xi_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) = \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\}) \lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \cup \{u\}) - \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \setminus \{u\}) \lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x} \setminus \{u\}).$$

It is not easy to characterise the locations where the leverage (46) will take a large value, except to say that, by definition, these are the locations where

the addition of a new data point would have substantially altered the fitted conditional intensity. Note that, whereas the leverage (47) of a Poisson model is always positive, the leverage (46) of a Gibbs process can include negative values; we have encountered this in practice.

4.5. Leverage for logistic composite likelihood

The logistic composite likelihood (21) or (25) involves randomly-generated dummy locations; we shall treat the dummy points as fixed when computing diagnostics. The composite score is (36).

Result 7. *Consider a Poisson model with intensity $\lambda_{\boldsymbol{\theta}}(u)$ fitted by maximum logistic composite likelihood* (21)*. Under regularity conditions (A1)–(A3) and (C2)–(C3) the leverage is*

$$h(u) = \lambda_{\widehat{\boldsymbol{\theta}}}(u)\, \zeta_{\widehat{\boldsymbol{\theta}}}(u)^{\top}\, H_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}, D)^{-1}\, \zeta_{\widehat{\boldsymbol{\theta}}}(u) \tag{50}$$

where $H_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}, D)$ is given in (39)*.*

This result is very similar to (47), and again the leverage must be nonnegative.

Result 8. *Consider a Gibbs model with conditional intensity $\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ fitted by maximum logistic composite likelihood. Under regularity conditions (A1)–(A3) and (C2)–(C3), the leverage is*

$$h(u) = \lambda_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x})\, \zeta_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x})\, H_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}, D)^{-1}\, \Delta_u U_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}, D) \tag{51}$$

where $H_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}, D)$ is given in (39) *and*

$$\Delta_u U_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}, D) = \mathbf{1}\{u \in W_{\ominus}\}\, \zeta_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x}) + \sum_{v \in \mathbf{x}_{\ominus}} \Delta_u \zeta_{\widehat{\boldsymbol{\theta}}}(v\,|\,\mathbf{x}) - \sum_{v \in \mathbf{x}_{\ominus} \cup D_{\ominus}} \Delta_u \pi_{\widehat{\boldsymbol{\theta}}}(v\,|\,\mathbf{x}), \tag{52}$$

where $\pi_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) = p_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\, \zeta_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})$.

This is derived by first principles in a similar fashion to Result 6. In the Gibbs case, the leverage (51) can take negative values.

5. Parameter influence in a point process model

The remaining diagnostics discussed in this paper are defined in terms of the effect of *deleting* observations, and are thus fundamentally different from the leverage.

Case deletion diagnostics are well developed for generalized linear models [54, 30], [12, p. 11 ff.], [32, pp. 149–170], [47, pp. 227–235] and for mixed models and generalised estimating equations, at least in theory [16, 43, 41]. The challenge is to develop them for spatial data.

5.1. General definition of parameter influence

In a generalized linear model, the parameter influence (in software parlance the "`DFBETA`" [12, Equation (2.1) p. 13], [47, p. 228 ff.]) of the ith observation is a Taylor approximation to the negative change in the fitted parameters $\widehat{\boldsymbol{\theta}}$ which would occur if the ith observation were deleted from the dataset. See (A.3) in Appendix A.

In a spatial point process model, the analogue of a single case deletion is to delete a small region of space B, along with its contents.

Definition 3. *For any real-valued or vector-valued function $f(A)$ of a set argument $A \subseteq W$, define*

$$(\nabla f)(B) = f(W) - f(W \setminus B). \tag{53}$$

That is, $(\nabla f)(B)$ determines the (negative) effect on $f(W)$ of deleting *the subset B from W.*

In order to define deletion diagnostics for spatial models, suppose that for any subset $A \subseteq W$, we can define a (composite) likelihood $\mathsf{CL}_A(\boldsymbol{\theta}, \mathbf{x}) = \mathsf{CL}_A(\boldsymbol{\theta}, \mathbf{x} \cap A)$ obtained by restricting the original (composite) likelihood to the data in A. Since the objective is to study the influence of different subsets of the data on the final model, we shall assume that *the edge correction is not changed* when the composite likelihood is restricted to A. For the likelihoods and composite likelihoods defined in Section 3.5, this is achieved by replacing $\mathbf{x}$, W, $\mathbf{x}_\ominus$, and $W_\ominus$ by $\mathbf{x} \cap A$, A, $\mathbf{x}_\ominus \cap A$ and $W_\ominus \cap A$ respectively in the definitions. Correspondingly the composite score is $U_A(\boldsymbol{\theta}, \mathbf{x}) = U_A(\boldsymbol{\theta}, \mathbf{x} \cap A)$ and the negative Hessian is $H_A(\boldsymbol{\theta}, \mathbf{x}) = H_A(\boldsymbol{\theta}, \mathbf{x} \cap A)$. Let $\widehat{\boldsymbol{\theta}}(A)$ be the maximiser of

$\mathsf{CL}_A(\boldsymbol{\theta}, \mathbf{x})$, that is, the maximum composite likelihood estimate based on the data inside A. Applying Definition 3 to $\widehat{\boldsymbol{\theta}}$, we consider

$$(\nabla\widehat{\boldsymbol{\theta}})(B) = \widehat{\boldsymbol{\theta}}(W) - \widehat{\boldsymbol{\theta}}(W \setminus B), \tag{54}$$

the negative change in the parameter estimate that would occur if we omitted all the data in the subregion B. We seek a Taylor approximation to this change. The standard first-order approximation is

$$(\nabla\widehat{\boldsymbol{\theta}})(B) \approx H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1} \, (\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(B), \tag{55}$$

where

$$(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(B) = U_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}) - U_{W\setminus B}(\widehat{\boldsymbol{\theta}}, \mathbf{x} \setminus B) \tag{56}$$

is the function $(\nabla U(\boldsymbol{\theta}, \mathbf{x}))(B)$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}(W)$.

Following the approach of [2, Definition 2] we require that approximation $(\mathbb{D}\widehat{\boldsymbol{\theta}})(B)$ to $(\nabla\widehat{\boldsymbol{\theta}})(B)$ should be *additive* as a function of the set argument B, that is, $(\mathbb{D}\widehat{\boldsymbol{\theta}})(A \cup B) = (\mathbb{D}\widehat{\boldsymbol{\theta}})(A) + (\mathbb{D}\widehat{\boldsymbol{\theta}})(B)$ for disjoint sets $A, B \subset W$.

For example, consider the homogeneous Poisson point process model in two dimensions, fitted to a pattern of n data points in a window W of area $|W|$. The maximum likelihood estimate of intensity is $\hat{\lambda} = n/|W|$. The canonical parameter is $\theta = \log\lambda$. Suppose we delete a subregion B of very small area $|B|$. If B contains a data point, the negative change in $\widehat{\boldsymbol{\theta}}$ is $(\nabla\widehat{\boldsymbol{\theta}})(B) = \log(n/|W|) - \log((n-1)/(|W| - |B|))$ or approximately $1/n$. If B does not contain data points, the negative change is $(\nabla\widehat{\boldsymbol{\theta}})(B) = \log((|W|-|B|)/|W|)$ or approximately $-|B|/|W|$. The effect of deleting any subregion $B \subset W$ can be approximated by two components: a positive change of $+1/n$ associated with each data point $v \in \mathbf{x} \cap B$; and a negative change of $-|B|/|W| = -\int_B (1/|W|)\,\mathrm{d}u$ associated with the "background" locations.

Definition 4. *The* parameter influence measure $\mathbb{D}\widehat{\boldsymbol{\theta}}$ *is the vector-valued measure on* W *with increments*

$$(\mathbb{D}\widehat{\boldsymbol{\theta}})(\mathrm{d}u) = H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(\mathrm{d}u) = H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1} \left[U_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}) - U_{W\setminus \mathrm{d}u}(\widehat{\boldsymbol{\theta}}, \mathbf{x})\right]. \tag{57}$$

For experts in measure theory, Definition 4 means that $\mathbb{D}\widehat{\boldsymbol{\theta}}$ is the countably-additive measure on Borel subsets of W obtained by applying Method II of Monroe [40] to the set function $f(A) = H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(A)$.

The parameter influence measure $\mathbb{D}\widehat{\boldsymbol{\theta}}$ is defined so that $\mathbb{D}\widehat{\boldsymbol{\theta}}(B)$ gives the approximate effect, on the fitted parameter $\widehat{\boldsymbol{\theta}}$, of removing all the data in the region B (i.e. both the observed data points and the background locations without data points). This makes it possible to predict the *sign and magnitude* of the change in parameter estimates that would occur if any chosen subset of the data were omitted from the fitting. One should look at this plot to understand why a particular data point has high influence (identified from the likelihood influence plot described in Section 6) and to ascertain the sign and magnitude of its effect. Each component of the parameter influence (corresponding to one of the canonical coefficients) is a real-valued measure; the values of the measure are expressed in the same units as those of the corresponding coefficient.

5.2. Parameter influence for Poisson likelihood

Result 9. *If the model is a Poisson process fitted by maximum likelihood, then*

$$(\mathbb{D}\widehat{\boldsymbol{\theta}})(B) = H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1} \left(\sum_{v \in \mathbf{x} \cap B} \zeta_{\widehat{\boldsymbol{\theta}}}(v) - \int_B \zeta_{\widehat{\boldsymbol{\theta}}}(u) \lambda_{\widehat{\boldsymbol{\theta}}}(u) \, \mathrm{d}u \right). \tag{58}$$

That is, the parameter influence measure consists of a diffuse component with density $-H^{-1}\zeta_{\widehat{\boldsymbol{\theta}}}(u)\lambda_{\widehat{\boldsymbol{\theta}}}(u)$ and a discrete component concentrated on the data points $v \in \mathbf{x}$ with masses $H^{-1}\zeta_{\widehat{\boldsymbol{\theta}}}(v)$, where $H = H(\widehat{\boldsymbol{\theta}}, \mathbf{x})$.

To prove this we note that

$$(\nabla U(\boldsymbol{\theta}, \mathbf{x}))(A) = \sum_{v \in \mathbf{x} \cap A} \zeta_{\boldsymbol{\theta}}(v) - \int_A \zeta_{\boldsymbol{\theta}}(u) \lambda_{\boldsymbol{\theta}}(u) \, \mathrm{d}u. \tag{59}$$

This expression is countably additive as a function of the set argument A, that is, it is a measure. Consequently the measure with increments (57) is (58), proving the result.

The form of $\mathbb{D}\widehat{\boldsymbol{\theta}}$ in (58) shows that it is a "weighted residual measure" in the sense of [9].

In the special case of a homogeneous Poisson process, the assertion established in Result 9 agrees with the rough calculation in Section 5.1.

5.3. Parameter influence for pseudolikelihood

Result 10. *For a Gibbs model fitted by maximum pseudolikelihood, the parameter influence measure* $\mathbb{D}\widehat{\boldsymbol{\theta}}$ *is*

$$(\mathbb{D}\widehat{\boldsymbol{\theta}})(B) = H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1} \left(\sum_{v \in \mathbf{x} \cap B} g^{\#}_{\widehat{\boldsymbol{\theta}}}(v \,|\, \mathbf{x}) - \int_B g_{\widehat{\boldsymbol{\theta}}}(u \,|\, \mathbf{x}) \lambda_{\widehat{\boldsymbol{\theta}}}(u \,|\, \mathbf{x}) \, \mathrm{d}u \right), \tag{60}$$

where

$$g_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) = \mathbf{1}\{u \in W_{\ominus}\} \, \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) \tag{61}$$

$$g^{\#}_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) = \mathbf{1}\{v \in W_{\ominus}\} \, \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) + \sum_{u \in \mathbf{x}_{\ominus}} \Delta_v \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) - \int_{W_{\ominus}} \Delta_v \xi_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) \, \mathrm{d}u \tag{62}$$

where $\xi_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) = \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) \, \lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})$*. That is, the parameter influence measure consists of a diffuse component with density* $-H^{-1} g_{\widehat{\boldsymbol{\theta}}}(u \mid \mathbf{x}) \lambda_{\widehat{\boldsymbol{\theta}}}(u \mid \mathbf{x})$ *and a discrete component concentrated on the data points* $v \in \mathbf{x}$ *with masses* $H^{-1} g^{\#}_{\widehat{\boldsymbol{\theta}}}(v \,|\, \mathbf{x})$*, where* $H = H(\widehat{\boldsymbol{\theta}}, \mathbf{x})$*.*

The proof is given in Appendix C. Note that, in the Gibbs case, $(\mathbb{D}\widehat{\boldsymbol{\theta}})(B)$ includes all contributions to the composite score from the region B, but additionally includes interaction terms between data points inside and outside this region. This is no longer a weighted residual measure in the sense of [9].

5.4. Parameter influence for logistic likelihood

Result 11. *For a Gibbs model fitted by maximum logistic composite likelihood, the parameter influence measure is a discrete measure concentrated on the data and dummy points, with masses* $H(\widehat{\boldsymbol{\theta}}, \mathbf{x}, D)^{-1} g^{\ddagger}_{\widehat{\boldsymbol{\theta}}}(u \,|\, \mathbf{x})$ *for* $u \in \mathbf{x} \cup D$*, where*

$$g^{\ddagger}_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) = -\mathbf{1}\{u \in W_{\ominus}\} \, \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) p_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) + \mathbf{1}\{u \in \mathbf{x}\} \, \Delta_u U_W(\boldsymbol{\theta}, \mathbf{x}, D), \tag{63}$$

and where $\Delta_u U_W(\boldsymbol{\theta}, \mathbf{x}, D)$ *is given in* (52)*.*

The proof is given in Appendix C.

6. Likelihood influence

6.1. General definition of influence

Next we define the likelihood influence measure for a point process, corresponding to the classical definition (equation (A.4) in Appendix A). Let

$$s(B) = \frac{2}{p} \log \frac{\mathsf{CL}(\widehat{\boldsymbol{\theta}}(W), \mathbf{x})}{\mathsf{CL}(\widehat{\boldsymbol{\theta}}(W \setminus B), \mathbf{x})}, \tag{64}$$

essentially the change in the value of log composite likelihood caused by omitting the data inside B when estimating $\boldsymbol{\theta}$ (but retaining these data when evaluating the composite likelihoods, in conformity with the classical definition).

The (composite likelihood) *influence measure* S will be a measure (countably-additive set function) such that $S(B)$ is a second-order Taylor approximation to (64) for each $B \subseteq W$. Note that, just as in the classical case [25], the *first order* Taylor approximation is

$$s(B) \approx \frac{2}{p}\left((\nabla\widehat{\boldsymbol{\theta}})(B)\right)^{\top} \frac{\partial}{\partial\boldsymbol{\theta}}\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \log \mathsf{CL}(\boldsymbol{\theta}) = \frac{2}{p}\left((\nabla\widehat{\boldsymbol{\theta}})(B)\right)^{\top} U(\widehat{\boldsymbol{\theta}}, \mathbf{x}) = 0,$$

because the derivative of $\log \mathsf{CL}$ is the composite score $U(\boldsymbol{\theta}, \mathbf{x})$, which is equal to zero at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ by assumption (C2). Hence the first order Taylor approximation is zero. This explains the need for the second order Taylor approximation,

$$s(B) \approx \frac{2}{p}\,\frac{1}{2}((\nabla\widehat{\boldsymbol{\theta}})(B))^{\top} H(\widehat{\boldsymbol{\theta}}, \mathbf{x})\,((\nabla\widehat{\boldsymbol{\theta}})(B)). \tag{65}$$

Applying (55) gives $s(B) \approx (1/p)(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(B)^{\top} H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(B)$. See [36, 54, 25].

Definition 5. *The (composite likelihood)* influence measure *is the measure S with increments*

$$S(\mathrm{d}u) = \frac{1}{p}(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(\mathrm{d}u)^{\top} H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(\mathrm{d}u). \tag{66}$$

In all the cases we consider, we find that S is a discrete measure putting weight only on the data points (and on the dummy points in the logistic composite likelihood). That is, for a data point or dummy point v, we will have

$$S(\{v\}) = \frac{1}{p}((\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(v))^{\top} H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(v)$$

where $(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(v) = U(\widehat{\boldsymbol{\theta}}, \mathbf{x}) - U(\widehat{\boldsymbol{\theta}}, \mathbf{x} \setminus v)$. For an infinitesimal region $\mathrm{d}u$ that does not contain any data points, $(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(\mathrm{d}u)$ is of order $|\mathrm{d}u|$, so that (66) is of order $|\mathrm{d}u|^2$ which is negligible.

In practical terms, the likelihood influence reflects the change in overall fit that would occur if a data point were omitted. One should look at this plot to identify highly influential and anomalous data points.

6.2. Influence for Poisson likelihood

Result 12. *If the model is a Poisson process fitted by maximum likelihood, the influence is a discrete measure with atoms, at the data points $v \in \mathbf{x}$, having mass*

$$S(\{v\}) = \frac{1}{p}\zeta_{\widehat{\boldsymbol{\theta}}}(v)^\top H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}\zeta_{\widehat{\boldsymbol{\theta}}}(v). \tag{67}$$

The proof is given in Appendix C. The right hand side of (67) is a non-negative-definite quadratic form, so it is always the case that $S(\{v\}) \geq 0$. In the loglinear setting, where $\zeta_{\boldsymbol{\theta}}(v) = Z(v)$ is a vector valued covariate, the influence value $S(\{v\}) = (1/p)Z(v)^\top H(\widehat{\boldsymbol{\theta}})^{-1}Z(v)$ is the squared Mahalanobis distance (defined by covariance matrix $H = H(\widehat{\boldsymbol{\theta}})$) between the origin and the point $Z(v)$, and can be interpreted as quantifying the "extremeness" of the covariate value.

6.3. Influence for pseudolikelihood

Result 13. *Suppose the model is a Gibbs process fitted by maximum pseudo-likelihood. The influence is a discrete measure with atoms, at the data points $v \in \mathbf{x}$, having mass*

$$S(\{v\}) = \frac{1}{p}g^{\#}_{\widehat{\boldsymbol{\theta}}}(v\,|\,\mathbf{x})^\top H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}g^{\#}_{\widehat{\boldsymbol{\theta}}}(v\,|\,\mathbf{x}) \tag{68}$$

where $g^{\#}_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})$ is given in (62).

The proof is similar to that of the preceding result. The right hand side of (68) is a non-negative-definite quadratic form, so it is always true that $S(\{v\}) \geq 0$.

6.4. Influence for logistic composite likelihood

Result 14. *If the model is fitted by maximising the logistic composite likelihood, the influence is a discrete measure with atoms at the* ***data and dummy*** *points $u \in \mathbf{x} \cup D$ with masses*

$$S(\{u\}) = \frac{1}{p}g^{\ddagger}_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x})^\top H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}g^{\ddagger}_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x}) \tag{69}$$

where $g^{\ddagger}_{\boldsymbol{\theta}}(u\,|\,\mathbf{x})$ is given in (63).

The proof is a slight modification of the preceding proofs. Again we always have $S(\{u\}) \geq 0$.

7. Effect change diagnostic DFFIT

The vector-valued parameter influence measure $\mathbb{D}\widehat{\boldsymbol{\theta}}$ gives the (negative) effect, on each parameter estimate $\widehat{\boldsymbol{\theta}}_j$, of deleting a part of the spatial domain and the associated data. In practical terms it can be difficult to interpret these values in terms of the predictions of the model. This problem is familiar in generalized linear modelling, where the usual remedy is to multiply each component of the parameter influence vector DFBETA by the corresponding covariate; the resulting diagnostic DFFIT gives the effect on each term in the linear predictor *at the same location* [12, eq. (2.10), p. 15], [53, p. 125], [47, p. 228 ff.], [30, pp. 76–77].

Definition 6. *The effect change diagnostic DFFIT is the vector-valued measure* $e(\cdot \,|\, \mathbf{x})$ *on* W *defined by*

$$e(A \mid \mathbf{x}) = \int_A \zeta_{\widehat{\boldsymbol{\theta}}}(u \,|\, \mathbf{x})^\top (\mathbb{D}\widehat{\boldsymbol{\theta}})(\,\mathrm{d}u) \tag{70}$$

for each $A \subseteq W$.

For a Poisson point process model with loglinear intensity fitted by maximum likelihood, the DFFIT measure has an atom at each data point $v \in \mathbf{x}$ of mass $e^{\#}(\{v\}) = Z(v)^\top H(\widehat{\boldsymbol{\theta}})^{-1} Z(v)$ and a diffuse component with density $e(u) = -\lambda_{\widehat{\boldsymbol{\theta}}}(u) Z(u)^\top H(\widehat{\boldsymbol{\theta}})^{-1} Z(u) = -h(u)$. This parallels the familiar connection between case-deletion residuals and leverage in generalised linear models.

For a Gibbs model fitted by maximum pseudolikelihood, the DFFIT measure has an atom at each data point $v \in \mathbf{x}$ of mass $e^{\#}(\{v\} \,|\, \mathbf{x}) = \zeta_{\widehat{\boldsymbol{\theta}}}(v \,|\, \mathbf{x})^\top H^{-1} g^{\#}(v \,|\, \mathbf{x})$ and a diffuse density $e(\,\mathrm{d}u \mid \mathbf{x}) = \zeta_{\widehat{\boldsymbol{\theta}}}(u \mid \mathbf{x})^\top H^{-1} g(u \mid \mathbf{x}) \lambda_{\widehat{\boldsymbol{\theta}}}(u \mid \mathbf{x})\,\mathrm{d}u$ over locations $u \in W$, where $H = H(\widehat{\boldsymbol{\theta}}, \mathbf{x})$, and $g^{\#}(v \,|\, \mathbf{x})$ and $g(u \,|\, \mathbf{x})$ are given by (62) and (61).

For a Gibbs model fitted by maximum logistic composite likelihood, the DFFIT measure e is discrete, with atoms at data points and dummy points of mass $Z(u \,|\, \mathbf{x})^\top H^{-1} g^{\ddagger}_{\widehat{\boldsymbol{\theta}}}(u \,|\, \mathbf{x})$, where $H = H_W(\widehat{\boldsymbol{\theta}}, \mathbf{x}, D)$, and $g^{\ddagger}_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})$ is given by (63).

In practical terms, the effect change DFFIT is the approximate effect *on the fitted model predictions* of deleting the data at a particular location. Its main limitation is that it only gives the effect at the same location. Values are *on the*

scale of the linear predictor, i.e. the **logarithm** of the intensity or conditional intensity.

8. Analysis of Swedish Pines including interaction

Using the diagnostics for Gibbs models developed in Sections 4–7 above, we now return to the analysis of the Swedish Pines data commenced in Section 2.

8.1. Inhomogeneous Strauss model

The Swedish Pines data are believed to exhibit regularity or inhibition between points, which could be explained by plant competition. Accordingly we modify the model used in Section 2 by introducing a new term in the likelihood which causes inhibition between points.

Fix a threshold distance $r > 0$. For any finite point pattern $\mathbf{x}$, let $s(\mathbf{x})$ be the number of unordered pairs of points in $\mathbf{x}$ that are closer than r units apart. The likelihood of the inhomogeneous Poisson process model fitted in Section 2 will now be multiplied by the *Strauss interaction term* $\gamma^{s(\mathbf{x})}$ where $0 \le \gamma \le 1$ is the interaction strength parameter, with $\gamma = 1$ yielding a Poisson process, and $\gamma = 0$ a hard core process in which random points never lie closer than r units apart [50, 35], [5, pp. 497–500]. The likelihood of this *inhomogeneous Strauss process* is

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = c_{\boldsymbol{\theta}} \left[\prod_{v \in \mathbf{x}} \beta_{\boldsymbol{\theta}}(v) \right] \gamma^{s(\mathbf{x})}, \tag{71}$$

where $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_5, \log \gamma)$ is the augmented parameter vector, $c_{\boldsymbol{\theta}}$ is the normalising constant, and $\beta_{\boldsymbol{\theta}}(u), u \in W$, is the log-quadratic function of the Cartesian coordinates (1) that previously served as the intensity of the Poisson model. The conditional intensity is

$$\lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) = \beta_{\boldsymbol{\theta}}(u) \gamma^{t(u|\mathbf{x})}, \tag{72}$$

where

$$t(u \,|\, \mathbf{x}) = \Delta_u s(\mathbf{x}) = s(\mathbf{x} \cup \{u\}) - s(\mathbf{x}) = \sum_{v \in \mathbf{x}} \mathbf{1}\{\|u - v\| \le r\} \tag{73}$$

is the number of points of $\mathbf{x}$ that lie closer than r units away from the location u.

For the Swedish Pines we take the interaction distance to be $r = 0.7$ metres selected in [5, p. 518]. The model was fitted by maximising Besag's pseudolikelihood using the border correction [6] with border width $R = r = 0.7$ metres, meaning that the domain $W_\ominus$ in (23) is obtained by trimming off a border of width R from the study region W, yielding an 8.2×8.6 metre rectangle. The fitted model has interaction strength $\widehat{\gamma} = 0.14$ corresponding to strong inhibition. The Strauss interaction term is significant at the 0.001 level, according to the adjusted composite likelihood ratio test [10]. An important detail is that, in order to perform this test, both models must have been fitted using the same composite likelihood; in this case the Poisson null model was re-fitted using the border correction, thus ignoring data close to the border. The non-stationary trend terms are marginally non-significant.

The left panel of Figure 9 shows the leverage function for the Strauss model, computed using (46) and (49). Some values of leverage are now 5 to 10 times higher than they were for the Poisson case (Figure 3). This occurs because the fitted trend $\beta_{\widehat{\boldsymbol{\theta}}}(u)$ in the Strauss model is approximately $1/\widehat{\gamma} \approx 7$ times larger than the fitted intensity in the Poisson model, in order to compensate for the strong inhibition in the Strauss model [5, pp. 496–7, 510]. Leverage is relatively low in the border region, as expected, because of the form of (49). Relatively high leverage occurs at several locations in the interior of the window. The leverage is now clearly very dependent on the spatial pattern of data points.

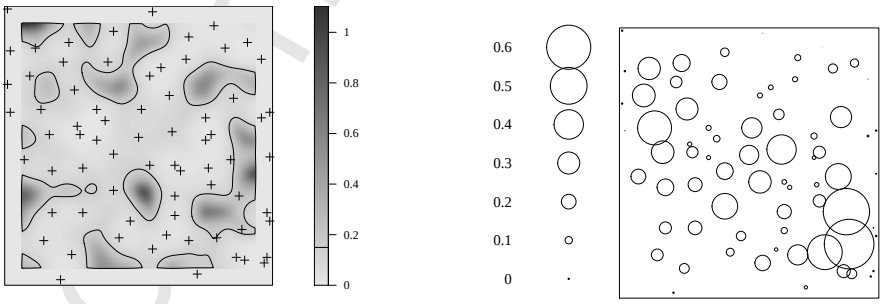


Figure 9: Leverage function (*left*) and influence measure (*right*) for a log-quadratic Strauss model fitted to the Swedish Pines data by maximum pseudolikelihood with border correction. In left panel, contour shows average value of leverage (also shown on the greyscale ribbon) and crosses are the original data.

The right panel of Figure 9 shows the (pseudolikelihood) influence for the Strauss model, computed from (68). The influence values are about 10 times larger than in the Poisson case, although these are log pseudolikelihood ratios which are not directly comparable with log likelihood ratios. Influence is again relatively low in the border region; influence is higher along a diagonal swath through the survey region. The highest influence occurs at data points near the bottom right corner of the survey region; these are the most isolated points, tending to make the regularity look strong.

Figure 10 shows the parameter influence measure for the Strauss model. For the trend parameters this shows a clear pattern in the way that the data tend to raise or lower each of the parameter values. For the canonical interaction parameter $\log \gamma$, the measure is highly dependent on the spatial configuration of the data points. Extremely large values indicate hypersensitivity to the pattern. The panel for the Intercept parameter is omitted for the reasons explained in Section 2.2.3, and to improve the layout.

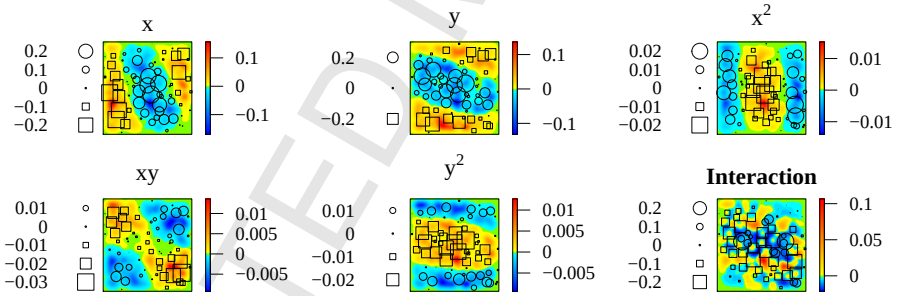


Figure 10: The parameter influence measure (omitting the Intercept panel) for a log-quadratic Strauss model fitted to the Swedish Pines data by maximum pseudolikelihood with border correction.

Figure 11 shows the corresponding `DFFIT` measure, using a common colour map and symbol map for each panel. The largest symbols, and the most extreme colour values, occur in the panels for the x and y coefficients. The Interaction panel in Figure 11 shows that changes in the interaction term are relatively small.

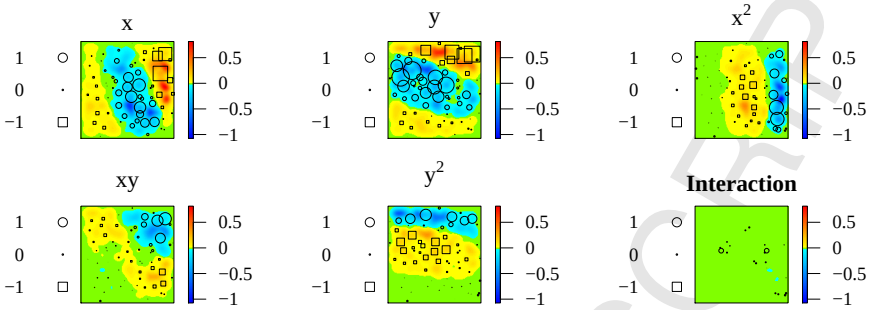


Figure 11: The DFFIT measure (omitting the Intercept panel) for a log-quadratic Strauss model fitted to the Swedish Pines data by maximum pseudolikelihood with border correction. Plotted using identical colour map and symbol map in all panels.

Figure 12 shows the sum of all the components of the DFFIT measure, that is, effectively the sum of all the panels in Figure 11. This "total DFFIT measure" expresses the effect of spatial deletions on the linear predictor (logarithm of the conditional intensity) of the fitted model. It reinforces the interpretations given above.

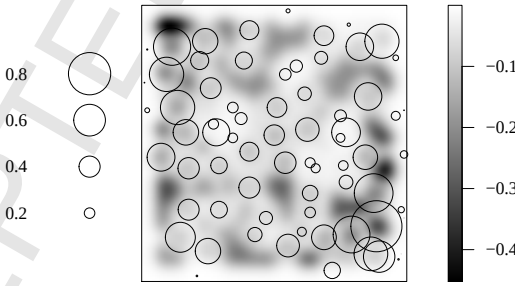


Figure 12: The total DFFIT measure for a log-quadratic Strauss model fitted to the Swedish Pines data by maximum pseudolikelihood with border correction.

d

8.2. Strauss model with isotropic edge correction

Our original motivation for studying the model diagnostics for the Swedish Pines was to resolve inconsistencies between the findings from different models. The diagnostics for the fitted Poisson model showed that locations close to the boundary of the survey region had high leverage, and there were some highly influential data points close to the boundary. In the diagnostics for the fitted Strauss model, fitted using the border correction, these data points were far less important. Indeed these peripheral data points do not contribute directly to the pseudolikelihood, because they fall in the border region $W \setminus W_{\ominus}$. This may account for inconsistencies between the Poisson and Strauss models fitted above. To assess this possible explanation, we re-fitted the inhomogeneous Strauss model using the "isotropic" edge correction [23, 6] which does not discard any data. That is, the domain of the pseudolikelihood or composite likelihood is $W_{\ominus} = W$. With this modification to the fitting procedure, the nonstationary trend is now significant at the 0.05 level according to the adjusted likelihood ratio test [10].

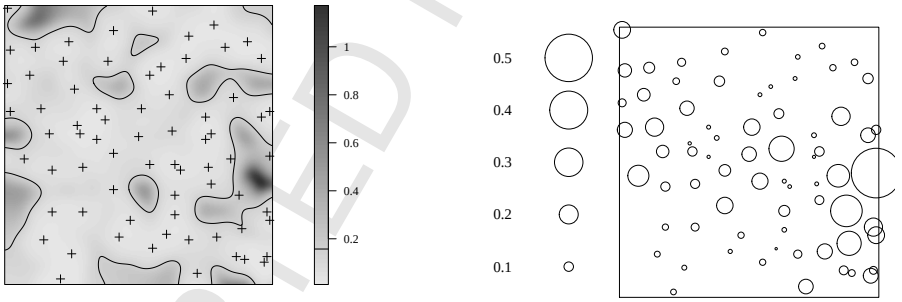


Figure 13: Leverage (*left*) and influence (*right*) for a log-quadratic Strauss model fitted to the Swedish Pines data with isotropic edge correction. Crosses show original data.

Figure 13 shows the leverage and influence computed for the log-quadratic Strauss model fitted using the modified procedure. The isotropic correction allows data near the edge of the survey region to contribute more strongly to the fit. The most influential data point is now a point near the right-hand edge of the survey region.

8.3. Conclusion for Swedish Pines

We can now draw a conclusion in the analysis of the Swedish Pines. First assuming no interaction between points, there was significant evidence of non-constant intensity, but the diagnostics (Section 2) showed that peripheral points were influential. To investigate interpoint interaction we fitted a Strauss model with the customary border correction, which suppresses contributions from data points near the border. For this model the non-constant trend terms were marginally non-significant, and diagnostics suggested that the peripheral points were much less important. This inconsistency motivated us to try an alternative choice of edge correction. In the Strauss model fitted with isotropic edge correction, the non-constant trend terms are significant. Figure 13 shows that data points in the border region have high leverage and influence for the Strauss model fitted with the isotropic correction. The end result is greater consistency between findings, and the conclusion that there is evidence for *both* trend *and* interaction, but that the fit is sensitive to data near the border.

9. Efficient computation formulae

The remainder of the paper presents algorithmic details which are vitally important in making the techniques feasible, but may be skipped by most users.

Computation of the diagnostics can be extremely demanding of time and computer memory. Most of the cost is incurred by computing (49) and (52). In principle, the quantity $\Delta_u \xi_{\boldsymbol{\theta}}(v \,|\, \mathbf{x})$ or $\Delta_u \pi_{\boldsymbol{\theta}}(v \,|\, \mathbf{x})$ must be evaluated for all ordered pairs of locations $u, v \in W$. In the case of (49), if the spatial domain is approximated by an $N \times N$ grid, then there are N^4 ordered pairs of locations to visit.

Efficient computational strategies are available in the cases of an exponential family model (9) corresponding to the loglinear conditional intensity (12), and of the exponential family with zeroes (10) corresponding to (13). The key quantities (49) and (52) reduce to

$$\Delta_u U(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{1}\{u \in W_\ominus\}\, Z(u \,|\, \mathbf{x}) + \sum_{v \in \mathbf{x}_\ominus} \Delta_u Z(v \,|\, \mathbf{x}) - \int_{W_\ominus} \Delta_u \xi_{\boldsymbol{\theta}}(v \,|\, \mathbf{x})\, \mathrm{d}v, \quad (74)$$

$$\Delta_u U_W(\boldsymbol{\theta}, \mathbf{x}, D) = \mathbf{1}\{u \in W_\ominus\}\, Z(u \,|\, \mathbf{x}) + \sum_{v \in \mathbf{x}_\ominus} \Delta_u Z(v \,|\, \mathbf{x}) - \sum_{v \in \mathbf{x}_\ominus \cup D_\ominus} \Delta_u \pi_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}). \quad (75)$$

Efficient computation is possible for models with a finite interaction range R as defined in (14). In the loglinear model (12) or loglinear-with-zeroes model (13) it follows from (14) that $Z(u \mid \mathbf{x}) = Z(u \mid \mathbf{x} \cap D(u,R))$ and $m(u \mid \mathbf{x}) = m(u \mid \mathbf{x} \cap D(u,R))$. Hence, if $\|u - v\| > R$ then $\Delta_u Z(v \mid \mathbf{x}) = \mathbf{0}$, $\Delta_u \xi_{\boldsymbol{\theta}}(v \mid \mathbf{x}) = \mathbf{0}$ and $\Delta_u \pi_{\boldsymbol{\theta}}(v \mid \mathbf{x}) = \mathbf{0}$. Computation can therefore be restricted to *close pairs* of locations u, v satisfying $\|u - v\| \le R$, with large savings in time and memory.

Implementation of this strategy involves developing fast algorithms to determine which pairs of locations u, v are closer than R units apart, and to evaluate $\Delta_u Z(v|\mathbf{x})$ and $\Delta_u m(v|\mathbf{x})$ for such pairs. We then require an expression for the integrand $\Delta_u \xi_{\boldsymbol{\theta}}(v|\mathbf{x})$ or $\Delta_u p_{\boldsymbol{\theta}}(v|\mathbf{x})$ using only the available values $\widehat{\boldsymbol{\theta}}$, $\lambda_{\widehat{\boldsymbol{\theta}}}(u|\mathbf{x})$, $\lambda_{\widehat{\boldsymbol{\theta}}}(v|\mathbf{x})$, $Z(u|\mathbf{x})$, $Z(v|\mathbf{x})$, $\Delta_u Z(v|\mathbf{x})$, $m(u|\mathbf{x})$, $m(v|\mathbf{x})$ and $\Delta_u m(v|\mathbf{x})$.

Define

$$s(u|\mathbf{x}) = (-1)^{\mathbf{1}\{u \in \mathbf{x}\}} \tag{76}$$

and for any function $g(u|\mathbf{x})$ define the signed difference

$$\Delta_u^{\#} g(v|\mathbf{x}) = s(u|\mathbf{x})\, \Delta_u g(v|\mathbf{x}) = \begin{cases} g(v|\mathbf{x} \cup \{u\}) - g(v|\mathbf{x}), & \text{for } u \notin \mathbf{x} \\ g(v|\mathbf{x} \setminus \{u\}) - g(v|\mathbf{x}), & \text{for } u \in \mathbf{x} \end{cases} \tag{77}$$

and the "effect"

$$\mathsf{R}_u g(v|\mathbf{x}) = g(v|\mathbf{x}) + s(u|\mathbf{x})\, \Delta_u g(v|\mathbf{x}) = \begin{cases} g(v|\mathbf{x} \cup \{u\}), & \text{for } u \notin \mathbf{x} \\ g(v|\mathbf{x} \setminus \{u\}), & \text{for } u \in \mathbf{x}. \end{cases} \tag{78}$$

Result 15. *In the loglinear model* (12) *fitted by maximum pseudolikelihood,*

$$\Delta_u \xi_{\boldsymbol{\theta}}(v|\mathbf{x}) = s(u|\mathbf{x}) \lambda_{\boldsymbol{\theta}}(v|\mathbf{x}) A_{\boldsymbol{\theta}}(u, v|\mathbf{x}) \tag{79}$$

where

$$A_{\boldsymbol{\theta}}(u, v|\mathbf{x}) = \mathsf{R}_u Z(v|\mathbf{x}) \exp(\boldsymbol{\theta}^\top \Delta_u^{\#} Z(v|\mathbf{x})) - Z(v|\mathbf{x}).$$

In the loglinear model with zeroes (13) *fitted by maximum pseudolikelihood,*

$$\Delta_u \xi_{\boldsymbol{\theta}}(v|\mathbf{x}) = s(u|\mathbf{x})\, \lambda_{\boldsymbol{\theta}}^{+}(v|\mathbf{x}) A_{\boldsymbol{\theta}}(u, v|\mathbf{x}) \tag{80}$$

where

$$A_{\boldsymbol{\theta}}(u, v|\mathbf{x}) = \mathsf{R}_u m(v|\mathbf{x}) \mathsf{R}_u Z(v|\mathbf{x}) \exp(\boldsymbol{\theta}^\top \Delta_u^{\#} Z(v|\mathbf{x})) - m(v|\mathbf{x}) Z(v|\mathbf{x}). \tag{81}$$

Proof. It suffices to prove (80) assuming (13).

If $u \notin \mathbf{x}$, we have

$$\begin{aligned}
\Delta_u \xi_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) &= Z(v\,|\,\mathbf{x}\cup\{u\})\lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}\cup\{u\}) - Z(v\,|\,\mathbf{x})\lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) \\
&= m(v\,|\,\mathbf{x}\cup\{u\})Z(v\,|\,\mathbf{x}\cup\{u\})\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}\cup\{u\}) - m(v\,|\,\mathbf{x})Z(v\,|\,\mathbf{x})\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) \\
&= \lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\left[m(v\,|\,\mathbf{x}\cup\{u\})Z(v\,|\,\mathbf{x}\cup\{u\})\frac{\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}\cup\{u\})}{\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})} - m(v\,|\,\mathbf{x})Z(v\,|\,\mathbf{x})\right] \\
&= \lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\left[m(v\,|\,\mathbf{x}\cup\{u\})Z(v\,|\,\mathbf{x}\cup\{u\})\exp(\boldsymbol{\theta}^\top\Delta_u Z(v\,|\,\mathbf{x})) - m(v\,|\,\mathbf{x})Z(v\,|\,\mathbf{x})\right] \\
&= \lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\Big[(m(v\,|\,\mathbf{x}) + \Delta_u m(v\,|\,\mathbf{x}))(Z(v\,|\,\mathbf{x}) + \Delta_u Z(v\,|\,\mathbf{x}))\exp(\boldsymbol{\theta}^\top\Delta_u Z(v\,|\,\mathbf{x})) \\
&\qquad\qquad -m(v\,|\,\mathbf{x})Z(v\,|\,\mathbf{x})\Big]. \qquad (82)
\end{aligned}$$

Alternatively if $u \in \mathbf{x}$,

$$\begin{aligned}
\Delta_u \xi_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) &= Z(v\,|\,\mathbf{x})\lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) - Z(v\,|\,\mathbf{x}\setminus\{u\})\lambda_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}\cup\{u\}) \\
&= m(v\,|\,\mathbf{x})Z(v\,|\,\mathbf{x})\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) - m(v\,|\,\mathbf{x}\setminus\{u\})Z(v\,|\,\mathbf{x}\setminus\{u\})\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}\cup\{u\}) \\
&= \lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\left[m(v\,|\,\mathbf{x})Z(v\,|\,\mathbf{x}) - m(v\,|\,\mathbf{x}\setminus\{u\})Z(v\,|\,\mathbf{x}\setminus\{u\})\frac{\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}\cup\{u\})}{\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})}\right] \\
&= \lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\,[m(v\,|\,\mathbf{x})Z(v\,|\,\mathbf{x}) - \\
&\qquad (m(v\,|\,\mathbf{x}) - \Delta_u m(v\,|\,\mathbf{x}))(Z(v\,|\,\mathbf{x}) - \Delta_u Z(v\,|\,\mathbf{x}))\exp(-\boldsymbol{\theta}^\top\Delta_u Z(v\,|\,\mathbf{x}))\Big]. \qquad (83)
\end{aligned}$$

Equations (82) and (83) can be rewritten in the common form (80). □

Result 16. *In the loglinear model* (12) *or loglinear-with-zeroes model* (13) *fitted by logistic composite likelihood,*

$$\Delta_u \pi_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}) = s(u\,|\,\mathbf{x})B(u, v\,|\,\mathbf{x}) \qquad (84)$$

where

$$B(u, v\,|\,\mathbf{x}) = R_u m(v\,|\,\mathbf{x})R_u Z(v\,|\,\mathbf{x})\frac{\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\exp(\boldsymbol{\theta}^\top\Delta^{\#}_u Z(v\,|\,\mathbf{x}))}{\lambda^+_{\boldsymbol{\theta}}(v\,|\,\mathbf{x})\exp(\boldsymbol{\theta}^\top\Delta^{\#}_u Z(v\,|\,\mathbf{x})) + \rho(u)} - Z(v\,|\,\mathbf{x})p_{\boldsymbol{\theta}}(v\,|\,\mathbf{x}). \qquad (85)$$

The results in this section can be extended with minor modification to the case where the first-order spatial trend is a general nonlinear function of the parameters:

$$\lambda_{\boldsymbol{\theta}}(u\,|\,\mathbf{x}) = m(u\,|\,\mathbf{x})\,\exp(O_{\boldsymbol{\theta}}(u) + \boldsymbol{\theta}^\top Z(u\,|\,\mathbf{x})), \qquad (86)$$

where $O_{\boldsymbol{\theta}}(u)$ is a real-valued function, twice differentiable with respect to $\boldsymbol{\theta}$ for all fixed u. This model is the hybrid [8] of a Gibbs process with loglinear

conditional intensity (13) and a Poisson process with very general form of the intensity. It often serves as the alternative hypothesis in a parametric test for interaction between points, where the null hypothesis is a very general Poisson process.

10. Software Implementation

The methods were implemented in the R language [44] using the spatial statistics package `spatstat` [7, 5]. The finished code is now released as part of `spatstat`.

Point process models are assumed to be fitted by Berman-Turner quadrature [13, 6] or logistic composite likelihood [17, 3] so that values of $Z(u\,|\,\mathbf{x})$ or $\zeta_{\widehat{\boldsymbol{\theta}}}(u\,|\,\mathbf{x})$ are available at a finite set of quadrature points $u_1, \ldots, u_m$ which include all the data points. The leverage function values, influence masses, and parameter influence contributions are then evaluated at these quadrature locations.

The software development cost was equivalent to 12 months full time work. Initially we developed simple algorithms which enumerate all triples or quadruples of locations, and store the results in three-dimensional arrays. We then extended the sparse matrix package `Matrix` [11] to three-dimensional sparse arrays with additional R and C code. We then implemented the sparse algorithms described in Section 9. A bottleneck is the computation of $\Delta_u Z(v\,|\,\mathbf{x})$ for all relevant pairs (u, v), and we developed special-purpose algorithms for computing this in various models. The fast code is extremely complex, and was checked and corrected by comparing results with the simple code.

Table 1 shows the computation times for evaluating the full set of leverage and influence diagnostics for models fitted to the Swedish Pines data. "Grid" indicates the spacing of dummy points used to fit the model. "Sparse" refers to the efficient sparse array methods described in Section 9 while "Non-Sparse" is the simple algorithm using full arrays and complete enumeration. The entry "NA" indicates that the non-sparse algorithm requires more memory than is available. Computational cost and memory usage are proportional to r^4 where r is the interaction range of the Strauss model. The relative efficiency of the sparse to non-sparse algorithms is also proportional to r^4, which justifies the effort expended on deriving and implementing the sparse algorithm.

Model	Grid	Sparse	Non-Sparse
Poisson	32×32	0.53	0.53
	64×64	0.56	0.56
	128×128	0.68	0.68
Strauss	32×32	0.69	1.15
	64×64	1.44	12.57
	128×128	11.38	NA

Table 1: Computation times (seconds) to evaluate the full set of leverage and influence diagnostics for models fitted to the Swedish Pines data, using different spacings of dummy points ("Grid") and different algorithms ("Sparse", "Non-Sparse"). Linux laptop, 2.6 GHz, quad core, 20 Gb RAM.

Acknowledgements

Adrian Baddeley was partially supported by the Australian Research Council, Discovery Outstanding Researcher Award DP130104470. Ege Rubak was supported by Curtin University as a visiting professor for six months; by The Danish Council for Independent Research | Natural Sciences, grant DFF – 7014–00074 "Statistics for point processes in space and beyond"; and by the Centre for Stochastic Geometry and Advanced Bioimaging, funded by grant 8721 from the Villum Foundation. We thank Robin Milne, Kassel Hingee and Gopalan Nair for helpful corrections.

References

[1] A.C. Atkinson. *Plots, Transformations and Regression*. Number 1 in Oxford Statistical Science Series. Oxford University Press/ Clarendon, 1985.

[2] A. Baddeley, Y.M. Chang, and Y. Song. Leverage and influence diagnostics for spatial point processes. *Scandinavian Journal of Statistics*, 40:86–104, 2013.

[3] A. Baddeley, J.-F. Coeurjolly, E. Rubak, and R. Waagepetersen. Logistic regression for spatial Gibbs point processes. *Biometrika*, 101(2):377–392, 2014.

[4] A. Baddeley, J. Møller, and A.G. Pakes. Properties of residuals for spatial point processes. *Annals of the Institute of Statistical Mathematics*, 60:627–649, 2008.

[5] A. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC, London, 2015.

[6] A. Baddeley and R. Turner. Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Australian and New Zealand Journal of Statistics*, 42(3):283–322, 2000.

[7] A. Baddeley and R. Turner. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005. URL: `www.jstatsoft.org`, ISSN: 1548-7660.

[8] A. Baddeley, R. Turner, J. Mateu, and A. Bevan. Hybrids of Gibbs point process models and their implementation. *Journal of Statistical Software*, 55(11):1–43, 2013.

[9] A. Baddeley, R. Turner, J. Møller, and M. Hazelton. Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society, Series B*, 67(5):617–666, 2005.

[10] A. Baddeley, R. Turner, and E. Rubak. Adjusted composite likelihood ratio test for spatial Gibbs point processes. *Journal of Statistical Computation and Simulation*, 86(5):922–941, 2016.

[11] Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2017. R package version 1.2-12.

[12] D.A. Belsley, E. Kuh, and R. Welsh. *Regression Diagnostics: Identifying Individual Data and Sources of Collinearity*. John Wiley and Sons, New York, 1980.

[13] M. Berman and T.R. Turner. Approximating point process likelihoods with GLIM. *Applied Statistics*, 41:31–38, 1992.

[14] J. Besag. Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute*, 47:77–91, 1977.

[15] D.S. Carter and P.M. Prenter. Exponential spaces and counting processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 21:1–19, 1972.

[16] R. Christensen, L.M. Pearson, and W. Johnson. Case-deletion diagnostics for mixed models. *Technometrics*, 34:38–45, 1992.

[17] M. Clyde and D. Strauss. Logistic regression for spatial pair-potential models. In A. Possolo, editor, *Spatial Statistics and Imaging*, volume 20 of *Lecture Notes - Monograph series*, chapter II, pages 14–30. Institute of Mathematical Statistics, Hayward, CA, 1991. ISBN 0-940600-27-7.

[18] J.F. Coeurjolly and E. Rubak. Fast covariance estimation for innovations computed from a spatial Gibbs point process. *Scandinavian Journal of Statistics*, 40:669–684, 2013.

[19] D.R. Cox and C.A. Donnelly. *Principles of Applied Statistics*. Cambridge University Press, Cambridge, UK, 2011.

[20] D.R. Cox and E.J. Snell. *Applied Statistics: Principles and Examples*. Chapman and Hall, 1981.

[21] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York, 1988.

[22] P.J. Diggle. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 34:138–147, 1985.

[23] P.J. Diggle, T. Fiksel, P. Grabarnik, Y. Ogata, D. Stoyan, and M. Tanemura. On parameter estimation for pairwise interaction processes. *International Statistical Review*, 62:99–117, 1994.

[24] P.J. Diggle and B. Rowlingson. A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A*, 157(3):433–440, 1994.

[25] E.L. Frome. The analysis of rates using Poisson regression models. *Biometrics*, 39:665–674, 1983.

[26] C. Gaetan and X. Guyon. *Spatial Statistics and Modeling*. Springer, 2009. Translated by Kevin Bleakley.

[27] A.E. Gelfand, P.J. Diggle, M. Fuentes, and P. Guttorp, editors. *Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL, 2010.

[28] P. Grabarnik and A. Särkkä. Modelling the spatial structure of forest stands by multivariate point processes with hierarchical interactions. *Ecological Modelling*, 220:1232–1240, 2009.

[29] Y. Guan, A. Jalilian, and R. Waagepetersen. Quasi-likelihood for spatial point processes. *Journal of the Royal Statistical Society, Series B*, 77:677–697, 2015.

[30] F. Harrell. *Regression Modeling Strategies*. Springer, New York, 2001.

[31] H. Högmander and A. Särkkä. Multitype spatial point patterns with hierarchical interactions. *Biometrics*, 55:1051–1058, 1999.

[32] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, first edition, 1989.

[33] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley and Sons, Chichester, 2008.

[34] J.B. Illian, J. Møller, and R.P. Waagepetersen. Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics*, 16:389–405, 2009.

[35] F.P. Kelly and B.D. Ripley. A note on Strauss's model for clustering. *Biometrika*, 63:357–360, 1976.

[36] J.M. Landwehr, D. Pregibon, and A.C. Shoemaker. Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79:61–83, 1984.

[37] E. Lesaffre. *Logistic Discriminant Analysis with Applications to Echocardiography.* DSc thesis, University of Leuven, 1986.

[38] P. McCullagh and J.A. Nelder. *Generalized Linear Models.* Chapman and Hall, second edition, 1989.

[39] J. Møller and R.P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton, FL, 2004.

[40] M.E. Monroe. *Measure and Integration.* Addison-Wesley, Reading, MA, second edition, 1971.

[41] J. Pan, Y. Fei, and P. Foster. Case-deletion diagnostics for linear mixed models. *Technometrics*, 56:269–281, 2014.

[42] D. Pregibon. Logistic regression diagnostics. *Annals of Statistics*, 9:705–724, 1981.

[43] J.S. Preisser and B.F. Qaqish. Deletion diagnostics for generalised estimation equations. *Biometrika*, 83:551–562, 1996.

[44] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018. ISBN 3-900051-07-0.

[45] B.D. Ripley. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:172–212, 1977.

[46] B.D. Ripley. *Spatial Statistics.* John Wiley and Sons, New York, 1981.

[47] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection.* John Wiley and Sons, New York, 1987.

[48] D.W. Scott. *Multivariate Density Estimation. Theory, Practice and Visualization.* John Wiley and Sons, New York, 1992.

[49] L. Strand. A model for stand growth. In *IUFRO Third Conference Advisory Group of Forest Statisticians*, pages 207–216, Paris, 1972. INRA, Institut National de la Recherche Agronomique.

[50] D.J. Strauss. A model for clustering. *Biometrika*, 62:467–475, 1975.

[51] P.C.T. van der Hoeven. Une projection de processus ponctuels. *Zeitschrift fuer Wahscheinlichkeitstheorie und verwandte Gebiete*, 61:483–499, 1982.

[52] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-Plus.* Springer, second edition, 1997.

[53] S. Weisberg. *Applied Linear Regression.* John Wiley and Sons, New York, second edition, 1985.

[54] D.A. Williams. Generalised linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, 36:181–191, 1987.

Appendices

Appendix A. Leverage and influence in Poisson regression

For reference and comparison we recall the definitions [42, 36] of leverage and influence for a generalized linear model [38, 30] in the case of Poisson loglinear regression. Suppose there are observations from n experimental units with integer responses $y_1, \ldots, y_n$ and vector-valued covariate values $z_1, \ldots, z_n$. The model is based on the assumption that $y_1, \ldots, y_n$ are realisations of independent Poisson random variables $Y_1, \ldots, Y_n$ with means $\mu_i = \exp(\boldsymbol{\theta}^\top z_i)$ where $\boldsymbol{\theta}$ is the parameter vector.

The *leverage* of observation i is the (i, i) diagonal entry of the standardised leverage matrix

$$H^* = V^{1/2} Z (Z^\top V Z)^{-1} Z^\top V^{1/2} \tag{A.1}$$

where $Z = (z_1 z_2 \ldots z_n)^\top$ is the design matrix and V is the estimated variance matrix of the responses. The standardised leverage matrix satisfies the "leverage equation"

$$V^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \approx H^* V^{-1/2}(Y - \boldsymbol{\mu}), \tag{A.2}$$

where $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}})$ is the vector of fitted means. In the words of McCullagh and Nelder [38, p. 397] the leverage matrix "measures the influence, in Studentized units, of changes in Y on $\hat{\boldsymbol{\mu}}$."

The *parameter influence* ("`DFBETA`") of observation i is (a Taylor approximation of) the vector

$$\mathbf{b}_i = \widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{-i}, \tag{A.3}$$

where $\widehat{\boldsymbol{\theta}}_{-i}$ is the estimate of $\boldsymbol{\theta}$ obtained from the data after deleting the ith observation [12, eq. (2.1), p. 13], [47, p. 228 ff.], [30, p. 76]. The *(likelihood) influence* of observation i is (a Taylor approximation of) the scalar

$$s_i = \frac{2}{p} \log \frac{L(\widehat{\boldsymbol{\theta}})}{L(\widehat{\boldsymbol{\theta}}_{-i})} \tag{A.4}$$

where p is the dimension of $\boldsymbol{\theta}$ and $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, (\mathbf{y}, \mathbf{z}))$ is the likelihood function *evaluated for the full dataset* [42, 36, 54, 25]. The effect change ("`DFFIT`") [12, eq. (2.10), p. 15], [53, p. 125], [47, p. 228 ff.], [30, pp. 76–77] of observation i is

the vector $\mathbf{e}_i$ given by the entrywise product of the parameter influence b_i and the ith column of the design matrix, i.e. with components

$$e_{ij} = b_{ij} z_{ij}, \quad j = 1, \ldots, p. \tag{A.5}$$

Appendix B. Regularised composite likelihoods

The leverage and influence diagnostics for a fitted model depend on the method that was used to fit the model. Each new choice of fitting method requires, in principle, a new mathematical derivation of the form of the diagnostics.

For "regularised" versions of the pseudolikelihood and logistic likelihood, the diagnostics defined in Sections 4–7 are only slightly modified. Suppose the model is fitted by maximising the penalised composite likelihood

$$\log \mathsf{CL}^*(\boldsymbol{\theta}; \mathbf{x}) = \log \mathsf{CL}(\boldsymbol{\theta}; \mathbf{x}) - b(\boldsymbol{\theta}) \tag{B.1}$$

where $\mathsf{CL}(\boldsymbol{\theta}; \mathbf{x})$ is the pseudolikelihood or logistic composite likelihood, and $b(\boldsymbol{\theta})$ is a penalty term which is twice differentiable with respect to $\boldsymbol{\theta}$. In what follows we will write $b'(\boldsymbol{\theta})$ and $b''(\boldsymbol{\theta})$ respectively for the vector of derivatives and the matrix of second derivatives of $b(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. A common choice of penalty is the sum of squared parameter values, $b(\boldsymbol{\theta}) = \epsilon \boldsymbol{\theta}^\top \boldsymbol{\theta}$, where ϵ is a tuning constant. This would yield $b''(\boldsymbol{\theta}) = 2\epsilon I_p$.

In this setting, using the foregoing notation, the estimating function is

$$U^*(\boldsymbol{\theta}; \mathbf{x}) := \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathsf{CL}^*(\boldsymbol{\theta}; \mathbf{x}) = U(\boldsymbol{\theta}; \mathbf{x}) - b'(\boldsymbol{\theta}) \tag{B.2}$$

with negative Hessian

$$H^*(\boldsymbol{\theta}; \mathbf{x}) := -\frac{\partial}{\partial \boldsymbol{\theta}} U^*(\boldsymbol{\theta}; \mathbf{x}) = H(\boldsymbol{\theta}; \mathbf{x}) + b''(\boldsymbol{\theta}). \tag{B.3}$$

It follows from (B.2) that

$$\Delta_v U^*(\boldsymbol{\theta}; \mathbf{x}) = \Delta_v U(\boldsymbol{\theta}; \mathbf{x}).$$

Thus the diagnostics for the regularised and un-regularised fits will be the same, except that the negative Hessian is modified by adding the second derivative of the penalty:

Result 17. *Suppose that $b''(\widehat{\boldsymbol{\theta}})$ is positive definite. Then Results 5, 6, 9, 10, 12 and 13 remain true when the composite log likelihood is replaced by the penalised composite log likelihood* (B.1) *and the negative Hessian $H(\boldsymbol{\theta}, \mathbf{x})$ is replaced by* (B.3)*.*

Similarly, Results 7, 8, 11 and 14 remain true when the logistic log likelihood is replaced by the penalised logistic likelihood, and the negative Hessian $H_W(\boldsymbol{\theta}, \mathbf{x}, D)$ is replaced by $H^*_W(\boldsymbol{\theta}, \mathbf{x}, D) = H_W(\boldsymbol{\theta}, \mathbf{x}, D) + b''(\boldsymbol{\theta})$.

For example, for a Gibbs model fitted by maximum penalised pseudolikelihood, the leverage is

$$h(u) = \lambda_{\widehat{\boldsymbol{\theta}}}(u \,|\, \mathbf{x})\; Z(u \,|\, \mathbf{x})\; \Big(H(\widehat{\boldsymbol{\theta}}, \mathbf{x}) + b''(\boldsymbol{\theta})\Big)^{-1} \Delta_u U(\widehat{\boldsymbol{\theta}}, \mathbf{x}).$$

Appendix C. Proofs

This Appendix sketches proofs of Results 10, 11 and 12. We shall try to strike a balance between measure-theoretic rigor and intuitive clarity.

A completely rigorous approach would involve applying Method II of Monroe [40, pp. 47–49, 60–62, 80] to the set function $f(A) = (\nabla U)(A)$, where U is the score associated with the composite likelihood.

For the cases considered in this paper, a simplified approach can be used. We consider a disc Q of small radius ϵ, and study the asymptotic behaviour of $f(Q)$ as $\epsilon \to 0$. In order to prove (4) or (5), it is sufficient to show that when Q is centred on a data point x_i, we have $f(Q) \to h(x_i)$, while if Q is centred on a non-data location $u \notin \mathbf{x}$, we have $f(Q) \sim g(u)\,|Q|$ where $|Q| = \pi\epsilon^2$ is the area of Q.

Proof of Result 10

For any set Q we have from (32)

$$\begin{aligned}
(\nabla U(\boldsymbol{\theta}, \mathbf{x}))(Q) \;=\; & \sum_{v \in \mathbf{x}_\ominus \cap Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) - \int_Q \mathbf{1}\{u \in W_\ominus\}\, \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) \lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})\, \mathrm{d}u \\
& + \sum_{v \in \mathbf{x}_\ominus \setminus Q} [\zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) - \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x} \setminus Q)] \\
& - \int_{W_\ominus \setminus Q} [\zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) \lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) - \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x} \setminus Q) \lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x} \setminus Q)]\, \mathrm{d}u
\end{aligned} \tag{C.1}$$

Suppose $Q = D(u, \epsilon)$ does not contain any points of $\mathbf{x}$. Then the only non-zero term on the right-hand side of (C.1) is the second, integral term

$$(\nabla U(\boldsymbol{\theta}, \mathbf{x}))(Q) = -\int_{Q \cap W_\ominus} \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) \lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) \, \mathrm{d}u. \tag{C.2}$$

In the limit as the radius $\epsilon \to 0$, we obtain

$$(\nabla U(\boldsymbol{\theta}, \mathbf{x}))(Q) \sim |Q| \mathbf{1}\left\{u \in W_\ominus\right\} \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) \lambda_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}). \tag{C.3}$$

This proves (60) for discs $Q \subseteq W \setminus \mathbf{x}$, and hence for all regions A which do not contain data points.

Next suppose that Q is centred on a data point $v \in \mathbf{x}$. As $\epsilon \to 0$, the integral on the right-hand side of (C.1) tends to zero, so that $(\nabla U(\boldsymbol{\theta}, \mathbf{x}))(Q) \to g^{\#}(v \,|\, \mathbf{x})$. Thus proves (60) for subsets $A = \{v\}$ where $v \in \mathbf{x}$ and hence gives the result.

Proof of Result 11

Using (36), for a region $Q \subset W$,

$$\begin{aligned}
(\nabla U(\boldsymbol{\theta}, \mathbf{x}, D))(Q) = & \sum_{v \in \mathbf{x}_\ominus \cap Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) + \sum_{v \in \mathbf{x}_\ominus \setminus Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) \\
& - \sum_{v \in (\mathbf{x}_\ominus \cup D_\ominus) \cap Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) p_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) - \sum_{v \in (\mathbf{x}_\ominus \cup D_\ominus) \setminus Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) p_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) \\
& - \sum_{v \in \mathbf{x}_\ominus \setminus Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x} \setminus Q) + \sum_{v \in (\mathbf{x}_\ominus \cup D_\ominus) \setminus Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x} \setminus Q) p_{\boldsymbol{\theta}}(v \,|\, \mathbf{x} \setminus Q) \\
= & \sum_{v \in \mathbf{x}_\ominus \cap Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) - \sum_{v \in (\mathbf{x}_\ominus \cup D_\ominus) \cap Q} \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) p_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) \\
& + \sum_{v \in \mathbf{x}_\ominus \setminus Q} \left[\zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) - \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x} \setminus Q)\right] \\
& - \sum_{v \in (\mathbf{x}_\ominus \cup D_\ominus) \setminus Q} \left[\zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) p_{\boldsymbol{\theta}}(v \,|\, \mathbf{x}) - \zeta_{\boldsymbol{\theta}}(v \,|\, \mathbf{x} \setminus Q) p_{\boldsymbol{\theta}}(v \,|\, \mathbf{x} \setminus Q)\right].
\end{aligned} \tag{C.4}$$

If Q contains no points of $\mathbf{x} \cup D$, then $(\nabla U(\boldsymbol{\theta}, \mathbf{x}, D))(Q) = 0$. When the diameter of Q is sufficiently small, it may only contain a single point $u \in \mathbf{x} \cup D$. If $u \in D$ we have $\mathbf{x} \cap Q = \emptyset$ and $\mathbf{x} \setminus Q = \mathbf{x}$, so there is only one non-zero term $(\nabla U(\boldsymbol{\theta}, \mathbf{x}, D))(Q) = -\mathbf{1}\left\{u \in W_\ominus\right\} \zeta_{\boldsymbol{\theta}}(u \,|\, \mathbf{x}) p_{\boldsymbol{\theta}}(u \,|\, \mathbf{x})$. Alternatively if $u \in \mathbf{x}$ we

get

$$(\nabla U(\boldsymbol{\theta}, \mathbf{x}, D))(Q) = \mathbf{1}\{u \in W_{\ominus}\}[\zeta_{\boldsymbol{\theta}}(u|\mathbf{x}) - \zeta_{\boldsymbol{\theta}}(u|\mathbf{x})p_{\boldsymbol{\theta}}(u|\mathbf{x})] + \sum_{v \in \mathbf{x}_{\ominus} \setminus \{u\}} [\zeta_{\boldsymbol{\theta}}(v|\mathbf{x}) - \zeta_{\boldsymbol{\theta}}(v|\mathbf{x} \setminus \{u\})]$$
$$- \sum_{v \in (\mathbf{x}_{\ominus} \cup D_{\ominus}) \setminus \{u\}} [\zeta_{\boldsymbol{\theta}}(v|\mathbf{x})p_{\boldsymbol{\theta}}(v|\mathbf{x}) - \zeta_{\boldsymbol{\theta}}(v|\mathbf{x} \setminus \{u\})p_{\boldsymbol{\theta}}(v|\mathbf{x} \setminus \{u\})]$$
$$= \Delta_u U_W(\boldsymbol{\theta}, \mathbf{x}, D) - \mathbf{1}\{u \in W_{\ominus}\}\zeta_{\boldsymbol{\theta}}(u|\mathbf{x})p_{\boldsymbol{\theta}}(u|\mathbf{x}),$$

where $\Delta_u U_W(\boldsymbol{\theta}, \mathbf{x}, D)$ is given in (52). This proves the result.

Proof of Result 12

Consider a disc Q of radius ϵ and centre u. If Q contains no points of $\mathbf{x}$, then recalling (59), we have $(\nabla U(\boldsymbol{\theta}, \mathbf{x}))(Q) = \int_Q \zeta_{\boldsymbol{\theta}}(u')\lambda_{\boldsymbol{\theta}}(u')\,\mathrm{d}u' \sim \zeta_{\boldsymbol{\theta}}(u)\lambda_{\boldsymbol{\theta}}(u)|Q|$ as $\epsilon \to 0$. Consequently

$$(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(Q)^{\top} H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}(\nabla U(\widehat{\boldsymbol{\theta}}, \mathbf{x}))(Q) = O(|Q|^2) \to 0.$$

Alternatively if Q contains a single point of $\mathbf{x}$, say $Q \cap \mathbf{x} = \{v\}$, then $(\nabla U(\boldsymbol{\theta}, \mathbf{x}))(Q) = \zeta_{\widehat{\boldsymbol{\theta}}}(v) + O(|Q|)$ so that

$$s(Q) \to \frac{1}{p}\zeta_{\widehat{\boldsymbol{\theta}}}(v)^{\top} H(\widehat{\boldsymbol{\theta}}, \mathbf{x})^{-1}\zeta_{\widehat{\boldsymbol{\theta}}}(v)$$

as $\epsilon \to 0$. This yields (67).