Closure Relations

Adrian Baddeley

Joint work with G. Nair, A. Mira, R.S. Anderssen, F. DeHoog, P. Grabarnik
Statistical theory often says we “should” act in a particular way
Statistical theory often says we “should” act in a particular way

- minimum variance estimation
Statistical theory often says we “should” act in a particular way

- minimum variance estimation
- maximum likelihood
Statistical theory often says we “should” act in a particular way

- minimum variance estimation
- maximum likelihood
- minimum Bayes risk
Statistical theory often says we “should” act in a particular way

- minimum variance estimation
- maximum likelihood
- minimum Bayes risk

Markov chains do not have such a “compulsory” interpretation
In a nutshell

Suppose $\pi P = \pi$
In a nutshell

Suppose $\pi P = \pi$

where $\pi$ is a vector of probabilities and $P$ is a transition matrix.
Suppose $\pi P = \pi$

where $\pi$ is a vector of probabilities and $P$ is a transition matrix. [Assuming ...],

- the Markov chain with transition probabilities $P$ converges in law to $\pi$
Suppose $\pi P = \pi$

where $\pi$ is a vector of probabilities and $P$ is a transition matrix. [Assuming ...],

- the Markov chain with transition probabilities $P$ converges in law to $\pi$
  $\rightarrow$ MCMC methods for simulating from $\pi$
In a nutshell

Suppose $\pi P = \pi$
where $\pi$ is a vector of probabilities and $P$ is a transition matrix. [Assuming ...],

- the Markov chain with transition probabilities $P$ converges in law to $\pi$
  $\rightarrow$ MCMC methods for simulating from $\pi$
- for any vector $v$,
  $$\pi v = \pi [P v]$$
In a nutshell

Suppose $\pi P = \pi$

where $\pi$ is a vector of probabilities and $P$ is a transition matrix. [Assuming \ldots],

- the Markov chain with transition probabilities $P$ converges in law to $\pi$
  $\quad \rightarrow$ MCMC methods for simulating from $\pi$
- for any vector $v$,
  $$\pi v = \pi [Pv]$$
  $\quad \rightarrow$ Recurrence relation between moments of $\pi$
Suppose \( \pi \, P = \pi \)

where \( \pi \) is a vector of probabilities and \( P \) is a transition matrix. [Assuming \ldots],

- the Markov chain with transition probabilities \( P \) converges in law to \( \pi \)
  - \textit{MCMC methods} for simulating from \( \pi \)
- for any vector \( v \),
  \[
  \pi v = \pi \left[ P v \right]
  \]
  - \textit{Recurrence relation} between moments of \( \pi \)
- \( \pi \) is characterised by
  \[
  \pi \left[ (P - I) v \right] = 0 \quad \text{for all } v
  \]
In a nutshell

Suppose \( \pi P = \pi \)
where \( \pi \) is a vector of probabilities and \( P \) is a transition matrix. [Assuming …],

- the Markov chain with transition probabilities \( P \) converges in law to \( \pi \)
  \( \rightarrow \) MCMC methods for simulating from \( \pi \)
- for any vector \( v \),
  \[ \pi v = \pi [Pv] \]
  \( \rightarrow \) Recurrence relation between moments of \( \pi \)
- \( \pi \) is characterised by
  \[ \pi [(P - I)v] = 0 \quad \text{for all } v \]
  \( \rightarrow \) Stein’s method for approximation by \( \pi \)
In a nutshell

Suppose $\pi P = \pi$
where $\pi$ is a vector of probabilities and $P$ is a transition matrix. [Assuming ...],

- the Markov chain with transition probabilities $P$ converges in law to $\pi$
  $\Rightarrow$ MCMC methods for simulating from $\pi$
- for any vector $v$,
  \[ \pi v = \pi [Pv] \]
  $\Rightarrow$ Recurrence relation between moments of $\pi$
- $\pi$ is characterised by
  \[ \pi [(P - I)v] = 0 \quad \text{for all } v \]
  $\Rightarrow$ Stein’s method for approximation by $\pi$
  $\Rightarrow$ Estimating equations for fitting $\pi_\theta$ to data
Plan

1. Introduction
2. Spatial point process models
3. MCMC for spatial point processes
4. Recurrence relations for moments
5. Estimating equations
2. Spatial point process models
Point pattern dataset
Point pattern dataset

\[ x = \{x_1, \ldots, x_n\}, \quad x_i \in D, \quad n \geq 0 \]
Model $x$ as a realisation of a *random point process* $X$ whose possible outcomes are (finite) point patterns in $D$. 
Poisson point process model

The homogeneous *Poisson* point process $\Pi$ with intensity $\lambda > 0$ in $D$ is characterised by
The homogeneous *Poisson* point process $\Pi$ with intensity $\lambda > 0$ in $D$ is characterised by

1. total number of points $N = n(\Pi)$ has a Poisson distribution with mean $\mu = \lambda|D|$
The homogeneous Poisson point process $\Pi$ with intensity $\lambda > 0$ in $D$ is characterised by

1. total number of points $N = n(\Pi)$ has a Poisson distribution with mean $\mu = \lambda |D|
2. Given $N = n$, the $n$ points are stochastically independent and uniformly distributed in $D$. 
Poisson point process model

The homogeneous *Poisson* point process $\Pi$ with intensity $\lambda > 0$ in $D$ is characterised by

1. total number of points $N = n(\Pi)$ has a Poisson distribution with mean $\mu = \lambda |D|$  
2. Given $N = n$, the $n$ points are stochastically independent and uniformly distributed in $D$.

For any real-valued function $h(x)$ of the configuration $x$,

$$\mathbb{E} h(\Pi) = e^{-\lambda |D|} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \int_D \cdots \int_D h(x) \, dx_1 \, dx_2 \ldots \, dx_n$$
Poisson process
We can specify a finite point process model by its \textit{probability density} \( f(x) \).
We can specify a finite point process model by its probability density \( f(x) \) relative to the distribution of the Poisson point process with unit intensity in \( D \).
We can specify a finite point process model by its probability density $f(x)$ relative to the distribution of the Poisson point process with unit intensity in $D$.

\[
\mathbb{E} h(X) = \mathbb{E}[h(\Pi)f(\Pi)] \\
= e^{-|D|} \sum_{n=0}^{\infty} \frac{1}{n!} \int_{D} \cdots \int_{D} h(x)f(x) \, dx_1 \, dx_2 \cdots \, dx_n
\]
Gibbs model

A *Gibbs point process model* has a probability density of the form

\[ f(x) = \exp V(x) \]

where

\[ V(x) = V_0 + \sum_i V_1(x_i) + \sum_{i<j} V_2(x_i, x_j) + \ldots \]

is called the *potential*. 
A **Gibbs point process model** has a probability density of the form

\[ f(x) = \exp V(x) \]

where

\[ V(x) = V_0 + \sum_i V_1(x_i) + \sum_{i<j} V_2(x_i, x_j) + \ldots \]

is called the *potential*. \( V_0 \) is the normalising constant, and \( V_1, V_2, V_3, \ldots \) are functions with values in \([-\infty, \infty)\) called the *potentials of order* 1, 2, 3, \ldots.
A **Gibbs point process model** has a probability density of the form

\[ f(x) = \exp V(x) \]

where

\[ V(x) = V_0 + \sum_i V_1(x_i) + \sum_{i<j} V_2(x_i, x_j) + \ldots \]

is called the **potential**.

\( V_0 \) is the normalising constant, and \( V_1, V_2, V_3, \ldots \) are functions with values in \( (-\infty, \infty) \) called the **potentials of order** \( 1, 2, 3, \ldots \).

The potentials \( V_2, V_3, \ldots \) introduce **dependence/interaction between points**.
**Hard core process**

Assume \( \exp V_1(u) \equiv \beta > 0 \) and

\[
\exp V_2(u, v) = \begin{cases} 
0 & \text{if } \|u - v\| \leq R \\
1 & \text{otherwise}
\end{cases}
\]

where \( R > 0 \) is a threshold distance, and \( V_k \equiv 0 \) for \( k > 2 \). Then

\[
f(x) = \begin{cases} 
\alpha \beta^n(x) & \text{if } \|x_i - x_j\| > R \text{ for all } i \neq j \\
0 & \text{otherwise}
\end{cases}
\]

where \( \alpha = \exp(V_0) \). This is equivalent to a Poisson process of intensity \( \beta \), conditioned on the event that \( \|x_i - x_j\| > R \) for all \( i \neq j \). Known as a hard core process.
Hard core process

Assume $\exp V_1(u) \equiv \beta > 0$ and

$$
\exp V_2(u, v) = \begin{cases} 
0 & \text{if } \|u - v\| \leq R \\
1 & \text{otherwise}
\end{cases}
$$

where $R > 0$ is a threshold distance, and $V_k \equiv 0$ for $k > 2$. Then

$$
f(x) = \begin{cases} 
\alpha \beta^n(x) & \text{if } \|x_i - x_j\| > R \text{ for all } i \neq j \\
0 & \text{otherwise}
\end{cases}
$$

where $\alpha = \exp(V_0)$. This is equivalent to a Poisson process of intensity $\beta$, conditioned on the event that $\|x_i - x_j\| > R$ for all $i \neq j$. Known as a hard core process.

Note $\alpha = e^{(1-\beta)|D|}/p(\beta, R)$ where $p(\beta, R) = \mathbb{P} \{ \Pi_\beta \text{ has no } R\text{-close pairs} \}$ is not tractable.
Hard core process

Poisson

Hard core
Let $\exp V_1(u) \equiv \beta > 0$ and

$$\exp V_2(u, v) = \begin{cases} \gamma & \text{if } \|u - v\| \leq R \\ 1 & \text{otherwise} \end{cases}$$

where $0 \leq \gamma \leq 1$ is the “interaction strength”. Then

$$f(x) = \alpha \beta^n(x) \gamma^{s(x)}$$

where

$$s(x) = \sum_{i<j} 1\{\|x_i - x_j\| \leq R\}.$$
Plan

1. Introduction
2. Spatial point process models
3. MCMC for spatial point processes
4. Recurrence relations for moments
5. Estimating equations
3. MCMC for spatial point processes
Gibbs sampler for finite system

For a finite system of random variables $X = (X_1, \ldots, X_n)$, the Gibbs sampler is a Markov chain with states $x = (x_1, \ldots, x_n)$ in which each transition involves choosing a site $i$ and updating $x_i$ according to the one-site conditional probabilities

$$\mathbb{P} \{ X_i = x_i \mid X_j = x_j, j \neq i \}$$
Gibbs sampler for point process

For a finite spatial point process $X$, the analogue of the Gibbs sampler is a spatial birth-death process:
Gibbs sampler for point process

For a finite spatial point process $X$, the analogue of the Gibbs sampler is a spatial birth-death process: a continuous-time Markov process whose states are point patterns $x$. 
Gibbs sampler for point process

For a finite spatial point process $X$, the analogue of the Gibbs sampler is a spatial birth-death process: a continuous-time Markov process whose states are point patterns $x$, in which each transition is either instantaneous deletion ("death")

$$x \mapsto x \setminus \{x_i\}$$

or instantaneous addition ("birth")

$$x \mapsto x \cup \{u\}$$
birth

dead
Gibbs sampler for point process

For detailed balance,

\[
\frac{\text{birth}}{\text{death}} = \frac{b(x \mapsto x \cup \{u\})}{d(x \cup \{u\} \mapsto x)} = \frac{f(x \cup \{u\})}{f(x)}
\]
Gibbs sampler for point process

For detailed balance,

\[
\frac{\text{birth}}{\text{death}} = \frac{b(x \mapsto x \cup \{u\})}{d(x \cup \{u\} \mapsto x)} = \frac{f(x \cup \{u\})}{f(x)}
\]

Without loss of generality,

\[
d() \equiv 1
\]

\[
b(x \mapsto x \cup \{u\}) = \frac{f(x \cup \{u\})}{f(x)}
\]

so each point has an independent exponential(1) lifetime
Conditional intensity of point process

For a finite point process with probability density $f(x)$, the Papangelou \textit{conditional intensity} at a location $u \not\in x$ is

$$\Lambda(u, x) = \frac{f(x \cup \{u\})}{f(x)}$$
Roughly speaking, $\Lambda(u, x) \, du$ is the probability that $X$ has a point in an infinitesimal neighbourhood of $u$, given complete information about the realisation of $X$ outside this neighbourhood.
Conditional intensity of hard core process

For a hard core process,

\[
\Lambda(u, x) = \begin{cases} 
\beta & \text{if } ||u - x_i|| > R \text{ for all } i \\
0 & \text{otherwise}
\end{cases}
\]
Conditional intensity of hard core process

For a hard core process,

\[ \Lambda(u, x) = \begin{cases} \beta & \text{if } \| u - x_i \| > R \text{ for all } i \\ 0 & \text{otherwise} \end{cases} \]

In the Gibbs sampler, new points are added at a uniform rate \( \beta \) at all locations permitted by the hard core constraint.
For a Strauss process,

$$\Lambda(u, x) = \begin{cases} 
\beta \gamma^t(u, x) & \text{if } \|u - x_i\| > R \text{ for all } i \\
0 & \text{otherwise}
\end{cases}$$

where $t(u, x) = \sum_i 1\{\|x_i - u\| \leq R\}$ is the number of points of $x$ close to $u$. 
Conditional intensity \( \Lambda(u, x) \) as a function of \( u \) for fixed realisation \( x \)
Plan

1. Introduction
2. Spatial point process models
3. MCMC for spatial point processes
4. Recurrence relations for moments
5. Estimating equations
4. Recurrence relations for moments
Moments

The *intensity function* $\lambda(u)$ of a point process $X$ is the spatially-varying expected number of points per unit area, defined so that, for all $B \subseteq D$,

$$\mathbb{E}n(X \cap B) = \int_B \lambda(u) \, du.$$
Goal:

For a Gibbs point process $X$ with given potentials $V_k$, find the intensity
Moments

Goal:

For a Gibbs point process $X$ with given potentials $V_k$, find the intensity

This problem is analytically intractable
Moments

Intractability of the moments of Gibbs point processes is a major obstacle, because the moments are needed in
Moments

Intractability of the moments of Gibbs point processes is a major obstacle, because the moments are needed in

- statistical physics (“equation of state”)
Moments

Intractability of the moments of Gibbs point processes is a major obstacle, because the moments are needed in

- statistical physics ("equation of state")
- probability modelling/prediction
Moments

Intractability of the moments of Gibbs point processes is a major obstacle, because the moments are needed in

- statistical physics ("equation of state")
- probability modelling/prediction
- fitting models to point pattern data
Moments

Intractability of the moments of Gibbs point processes is a major obstacle, because the moments are needed in

- statistical physics ("equation of state")
- probability modelling/prediction
- fitting models to point pattern data
- model diagnostics
Intractability of the moments of Gibbs point processes is a major obstacle, because the moments are needed in

- statistical physics ("equation of state")
- probability modelling/prediction
- fitting models to point pattern data
- model diagnostics
- . . .
Moments

Intractability of the moments of Gibbs point processes was the original motivation for the invention of *Markov Chain Monte Carlo* methods

N. Metropolis et al (1953)
Equation of state calculations by fast computing machines.
*Journal of Chemical Physics* 21 1087–1092
“Homogeneous” case

Let $D$ be a rectangle, and identify opposite edges, making it a torus. Assume $V_k$ are invariant under translation on the torus, so the point process is “homogeneous” (“stationary”).

Then the intensity $\lambda$ is constant, and $\mathbb{E}_n(X \cap B) = \lambda |B|$ for $B \subseteq D$.

Let’s try to determine the constant $\lambda$. 
Georgii-Nguyen-Zessin identity

For any functional $h(x)$ and any fixed location $u \in D$

$$\mathbb{E} h(X) = \mathbb{E} \left[ \frac{\Lambda(u, X \setminus \{u\})}{\lambda} h(X \setminus \{u\}) \right| u \in X$$

where $\lambda$ is the (constant) intensity, and the right hand side is formally defined as an expectation with respect to the Palm distribution of $X$ at $u$. 
For any functional $h(x)$ and any fixed location $u \in D$

$$\mathbb{E}h(X) = \mathbb{E}\left[ \frac{\Lambda(u, X \setminus \{u\})}{\lambda} h(X \setminus \{u\}) \bigg| u \in X \right]$$

where $\lambda$ is the (constant) intensity, and the right hand side is formally defined as an expectation with respect to the Palm distribution of $X$ at $u$.

Taking $h(x) \equiv 1$ gives

$$\lambda = \mathbb{E} [\Lambda(u, X \setminus \{u\}) \mid u \in X].$$
GNZ identity

$$\lambda = \mathbb{E} [\Lambda(u, X \setminus \{u\}) \mid u \in X]$$
\[ \lambda = \mathbb{E} \left[ \Lambda(u, X \setminus \{u\}) \mid u \in X \right] \]

For example, for the Strauss process,

\[ \Lambda(u, x) = \beta \gamma^{t(u, x)} \]

where \( t(u, x) = \sum_i 1\{\|x_i - u\| \leq R\} \).
GNZ identity

\[ \lambda = \mathbb{E} \left[ \Lambda(u, X \setminus \{u\}) \mid u \in X \right] \]

For example, for the Strauss process,

\[ \Lambda(u, x) = \beta \gamma^{t(u, x)} \]

where \( t(u, x) = \sum_i 1\{\|x_i - u\| \leq R\} \). So the intensity of the Strauss process satisfies

\[ \lambda = \beta \mathbb{E} \left[ \gamma^{t(u, x) \setminus \{u\}} \mid u \in X \right]. \]
\[ \lambda = \mathbb{E} \left[ \Lambda(u, X \setminus \{u\}) \mid u \in X \right] \]

For example, for the Strauss process,

\[ \Lambda(u, x) = \beta \gamma^{t(u, x)} \]

where \( t(u, x) = \sum_i 1\{\|x_i - u\| \leq R\} \). So the intensity of the Strauss process satisfies

\[ \lambda = \beta \mathbb{E} \left[ \gamma^{t(u, x \setminus \{u\})} \mid u \in X \right]. \]

Here the conditional intensity \( \Lambda(u, x) \) is known explicitly.
GNZ identity

\[ \lambda = \mathbb{E} \left[ \Lambda(u, X \setminus \{u\}) \mid u \in X \right] \]

For example, for the Strauss process,

\[ \Lambda(u, x) = \beta \gamma^{t(u,x)} \]

where \( t(u, x) = \sum_i 1\{\|x_i - u\| \leq R\} \). So the intensity of the Strauss process satisfies

\[ \lambda = \beta \mathbb{E} \left[ \gamma^{t(u,x \setminus \{u\})} \mid u \in X \right] . \]

Here the conditional intensity \( \Lambda(u, x) \) is known explicitly. But the right hand side is an expectation with respect to the unknown (Palm) distribution of the point process, so both sides are intractable.
Idea for approximation

Idea:

Approximate the RHS by an expectation with respect to a Poisson process with the same unknown intensity $\lambda$
Idea for approximation

Idea:

Approximate the RHS by an expectation with respect to a Poisson process with the same unknown intensity $\lambda$

i.e.

$$\lambda \approx \mathbb{E}_{\text{Pois}(\lambda)} \left[ \Lambda(u, \mathbf{X} \setminus \{u\}) \mid u \in \mathbf{X} \right]$$
Idea for approximation

Idea:

Approximate the RHS by an expectation with respect to a Poisson process with the same unknown intensity $\lambda$

i.e.

$$\lambda \approx \mathbb{E}_{\text{Pois}(\lambda)} \left[ \Lambda(u, X \setminus \{u\}) \mid u \in X \right]$$

by Slivnyak’s Theorem this is equivalent to

$$\lambda \approx \mathbb{E}_{\text{Pois}(\lambda)} \left[ \Lambda(u, X) \right]$$

Solve for $\lambda$ to obtain the “Poisson-saddlepoint approximation” $\lambda^{PS}$
For the Strauss process

\[ \mathbb{E}_{\text{Pois}}(\lambda) \left[ \Lambda(u, X) \right] = \mathbb{E}_{\text{Pois}}(\lambda) \left[ \beta \gamma^{t(u, X)} \right] = \beta \exp(-\lambda(1 - \gamma)\pi R^2). \]

Thus we solve

\[ \lambda = \exp(-\lambda G) \]

where \( G = (1 - \gamma)\pi R^2. \) The solution is

\[ \lambda^{PS} = \frac{W(\beta G)}{G} \]

where \( W \) is the inverse function of \( x \mapsto xe^x \), known as Lambert's \( W \)-function.
\[ \beta = 100, \ r = 0.05 \]

- exact values
- Poisson-saddlepoint
- mean field
Similarity to mean field approximation

Possion-saddlepoint:

$$\lambda \approx \mathbb{E}_{\text{Pois}(\lambda)} [\Lambda(u, X)]$$

Mean field:

$$\lambda \approx \exp \mathbb{E}_{\text{Pois}(\lambda)} [\log \Lambda(u, X)]$$
Inhomogeneous case

Remove the “homogeneity” assumptions.

Now we want to find the intensity function $\lambda(u), u \in D$ defined by

$$\mathbb{E} n(X \cap B) = \int_B \lambda(u) \, du.$$
Inhomogeneous case

The GNZ identity gives

\[ \lambda(u) = \mathbb{E} \left[ \Lambda(u, X \setminus \{u\}) \mid u \in X \right] \]

Again we approximate this by

\[ \lambda(u) = \mathbb{E}_{\text{Pois}(\lambda(\cdot))} \left[ \Lambda(u, X) \right] \]

where the expectation is with respect to a Poisson process with intensity function \( \lambda(\cdot) \).
Consider a general pairwise interaction process (where the only condition is $V_k \equiv 0$ for $k > 2$).

Then the Poisson-saddlepoint approximation $\lambda^{\text{PS}}(\cdot)$ is the solution of the integral equation

$$
\lambda(u) = \beta(u) \exp\left(- \int_D (1 - \gamma(u, v)) \lambda(v) \, dv\right)
$$

where $\beta(u) = \exp V_1(u)$ and $\gamma(u, v) = \exp V_2(u, v)$. 


Consider a general pairwise interaction process (where the only condition is $V_k \equiv 0$ for $k > 2$).

Then the Poisson-saddlepoint approximation $\lambda_{PS}(\cdot)$ is the solution of the integral equation

$$
\lambda(u) = \beta(u) \exp\left(- \int_D (1 - \gamma(u, v)) \lambda(v) \, dv\right)
$$

where $\beta(u) = \exp V_1(u)$ and $\gamma(u, v) = \exp V_2(u, v)$.

Solve by (under-relaxed) Picard iteration
Example 1: Simulation in bounded domain

Simulation of the stationary Strauss process in a bounded domain $D$ does not yield a stationary process, due to “edge effects”. How bad are the edge effects?
Example 1: Simulation in bounded domain

Simulation of the stationary Strauss process in a bounded domain $D$ does not yield a stationary process, due to “edge effects”. How bad are the edge effects?

Intensity of finite Strauss process in rectangle
Example 2: data analysis

Japanese Pines data
Example 2: data analysis

data

exp $V_1$

exp $V_2$
Intensity estimated by MCMC and by Poisson-saddlepoint approximation
Second moments

$K$-function of the Strauss process: true value (grey shading), Poisson-saddlepoint approximation (dashed lines), sparse approximation (dotted lines).
Summary

- Accurate approximations to the first two moments of a Gibbs process
Summary

- Accurate approximations to the first two moments of a Gibbs process
- Many potential applications
Summary

- Accurate approximations to the first two moments of a Gibbs process
- Many potential applications
  - model fitting, model validation, prediction, conditional prediction,
  - Bayesian inference, improvements to MCMC.
Summary

- Accurate approximations to the first two moments of a Gibbs process
- Many potential applications
  - model fitting, model validation, prediction, conditional prediction,
    Bayesian inference, improvements to MCMC.
- Potential for fast maximum likelihood estimation
Summary

- Accurate approximations to the first two moments of a Gibbs process
- Many potential applications
  - model fitting, model validation, prediction, conditional prediction,
    Bayesian inference, improvements to MCMC.
- Potential for fast maximum likelihood estimation
- Available in open source software spatstat
A. Baddeley and G. Nair.
**Fast approximation of the intensity of Gibbs point processes.**

A. Baddeley and G. Nair.
**Approximating the moments of a spatial point process.**

R.S. Anderssen *et al.*
**Solution of an integral equation arising in spatial point process theory.**
Plan

1. Introduction
2. Spatial point process models
3. MCMC for spatial point processes
4. Recurrence relations for moments
5. Estimating equations
5. Estimating equations
Goal:

find *computationally tractable* parameter estimators for fitting Gibbs point process models to point pattern data.
Maximum likelihood

Assume parametric model, with probability density

\[ f(x, \theta) \]

where \( \theta \) is the parameter, whose value we want to estimate. The **maximum likelihood estimate** (MLE) of \( \theta \) from data \( x \) is

\[ \hat{\theta} = \arg\max_{\theta} f(x, \theta). \]
Under regularity conditions, the MLE is the solution of

\[ U(\hat{\theta}, x) = 0 \]

where \( U(\theta, x) = \frac{\partial}{\partial \theta} \log f(x, \theta) \) is the score function.
Likelihood score equations

Under regularity conditions, the MLE is the solution of

\[ U(\hat{\theta}, x) = 0 \]

where \( U(\theta, x) = \frac{\partial}{\partial \theta} \log f(x, \theta) \) is the score function. The score satisfies

\[ \mathbb{E}_\theta [U(\theta, X)] = 0 \]
MLE for exponential family

The likelihoods are typically in regular exponential family form

\[ f(x, \theta) = \frac{\exp\{\theta^t V(x)\}}{Z(\theta)} \]

where
- \(x\) is the data,
- \(\theta\) the parameter,
- \(V(x)\) a statistic (easily computable),
- \(Z(\theta)\) the normalising constant.
MLE for exponential family

The score is

\[ U(x, \theta) = V(x) - \mathbb{E}_\theta[V(X)] \]

so the MLE is the solution \( \hat{\theta} \) of

\[ \mathbb{E}_\theta[V(X)] = V(x) \]
MLE for exponential family

Problem: $Z(\theta)$ is intractable

$\Rightarrow$ moments $E_\theta[V(X)]$ are intractable
$\Rightarrow$ maximum likelihood is analytically intractable
$\Rightarrow$ finite sample behaviour is poorly understood
Suppose \( X = (X_1, \ldots, X_n) \) is a vector of discrete random variables. Define the pseudolikelihood

\[
\text{PL}(\theta, x) = \prod_{i=1}^{n} P_{\theta} \{ X_i = x_i \mid (X_j = x_j, \ j \neq i) \}
\]

the product of the conditional likelihoods of each \( X_i \) given the other variables. Estimate \( \theta \) by

\[
\tilde{\theta} = \arg\max_{\theta \in \Theta} \text{PL}(\theta, x).
\]
Suppose $X$ is a spatial point process with Papangelou conditional intensity $\Lambda_\theta(u, X)$. Define the **pseudolikelihood**

$$PL(\theta, x) = \left[ \prod_{i=1}^{n} \Lambda_\theta(x_i, x) \right] \exp \left( - \int_{D} \Lambda(u, x) \, du \right).$$

Again estimate $\theta$ by

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} PL(\theta, x).$$
Maximum pseudolikelihood

The MPLE for a Gibbs model is typically

- easy to compute
- consistent and asymptotically normal
- biased and inefficient in small samples
MCMC methods

MCMC maximum likelihood inference is elegant, statistically efficient, but requires large computing resources.

Pseudolikelihood is cheap to compute, but is statistically inefficient.

We need something in between. . .
Pseudolikelihood estimating equations

Suppose \( \mathbf{X} = (X_1, \ldots, X_N) \) has regular exponential family distribution with density

\[
f(\mathbf{x}; \theta) = \frac{\exp\{\theta^\top V(\mathbf{x})\}}{Z(\theta)}.
\]

Then, writing \( X_{-i} \) for \( (X_j, j \neq i) \),

\[
\mathbb{P}_\theta \left\{ X_i = x_i \mid X_{-i} = \mathbf{x}_{-i} \right\} = \frac{\exp\{\theta^\top V(\mathbf{x})\}}{\sum_{\mathbf{x}' \prime} \exp\{\theta^\top V(\mathbf{x}' \prime)\}}
\]

where \( \mathbf{x}' \) runs over all states satisfying \( \mathbf{x}'_{-i} = \mathbf{x}_{-i} \).
Pseudolikelihood

We find easily

\[
\frac{\partial}{\partial \theta} \log PL(\theta, x) = \sum_{i=1}^{n} (V(x) - E_{\theta}[V(X) \mid X_{-i} = x_{-i}])
\]

so \( \tilde{\theta} \) is the solution of

\[
\frac{1}{n} \sum_{i=1}^{n} E_{\theta}[V(X) \mid X_{-i} = x_{-i}] = V(x)
\]

This is an unbiased estimating equation.
Consider a parameter estimator $\hat{\theta}$ which is the solution of an estimating equation

$$g(x, \theta) = 0$$

for a given function $g$, where $x$ is the data.
Consider a parameter estimator $\hat{\theta}$ which is the solution of an estimating equation

$$g(x, \theta) = 0$$

for a given function $g$, where $x$ is the data. We call $g$ unbiased if, for all $\theta$,

$$\mathbb{E}_\theta [g(X, \theta)] = 0$$
Consider a parameter estimator \( \hat{\theta} \) which is the solution of an *estimating equation*

\[
g(x, \theta) = 0
\]

for a given function \( g \), where \( x \) is the data. We call \( g \) *unbiased* if, for all \( \theta \),

\[
\mathbb{E}_\theta [g(X, \theta)] = 0
\]

This does not mean the estimator \( \hat{\theta} \) is unbiased, but it will usually be consistent.
Example 1: The MLE is the solution of the estimating equation

\[ U(x; \theta) = 0 \]

where \( U \) is the score.
Estimating equations (reminder)

*Example 1:* The MLE is the solution of the estimating equation

\[ U(x; \theta) = 0 \]

where \( U \) is the score.

The score is an unbiased estimating function:

\[ \mathbb{E}_\theta[U(X; \theta)] = 0. \]
Example 2: The method-of-moments estimator is the value of $\theta$ satisfying $g(x, \theta) = 0$ where

$$g(x, \theta) = S(x) - \mathbb{E}_\theta[S(X)]$$

where $S(x)$ is the observed value of a chosen statistic $S$, and $\mathbb{E}_\theta[S(X)]$ is the population moment.
Estimating equations (reminder)

Example 2: The method-of-moments estimator is the value of $\theta$ satisfying $g(x, \theta) = 0$ where

$$g(x, \theta) = S(x) - \mathbb{E}_\theta[S(X)]$$

where $S(x)$ is the observed value of a chosen statistic $S$, and $\mathbb{E}_\theta[S(X)]$ is the population moment.

The method-of-moments estimating function is an unbiased estimating function:

$$\mathbb{E}_\theta G(x, \theta) = \mathbb{E}_\theta[S(X) - \mathbb{E}_\theta[S(X)]] = 0.$$
Godambe-Heyde theory of estimating functions generalises Fisher-Cramér asymptotic theory of MLE, making it possible to:

- calculate asymptotic variance of estimators
- choose the optimal estimating function within a class
- construct an associated 'quasilikelihood'
Estimating equations

Godambe-Heyde theory of estimating functions generalises Fisher-Cramér asymptotic theory of MLE, making it possible to

- calculate asymptotic variance of estimators
- choose the optimal estimating function within a class
- construct an associated ‘quasilikelihood’

It doesn’t tell us how to

- find a class of unbiased estimating functions
- trade off statistical efficiency against computational cost.
Equilibrium principle

*Standard fact:* if $\mathbf{Y}$ is a stationary Markov process with equilibrium distribution $\pi$, and if $X$ has distribution $\pi$, then

$$
\mathbb{E} [(\mathcal{A} S)(X)] = 0
$$

for essentially all statistics $S = S(X)$, where $\mathcal{A}$ is the *generator* of $\mathbf{Y}$. 
Equilibrium principle

Discrete case

Define generator $\mathcal{A}$ by

$$(\mathcal{A}S)(x) = \mathbb{E}[S(Y_{n+1}) - S(Y_n) \mid Y_n = x]$$
Equilibrium principle

Discrete case
Define generator \( \mathcal{A} \) by

\[
(\mathcal{A} S)(x) = \mathbb{E}[S(Y_{n+1}) - S(Y_n) \mid Y_n = x] \\
= \mathbb{E}[S(Y_{n+1}) \mid Y_n = x] - S(x)
\]
Equilibrium principle

Discrete case
Define generator $\mathcal{A}$ by

$$
(\mathcal{A} S)(x) = \mathbb{E}[S(Y_{n+1}) - S(Y_n) \mid Y_n = x] \\
= \mathbb{E}[S(Y_{n+1}) \mid Y_n = x] - S(x) \\
= \sum_{y \in \mathcal{X}} (S(y) - S(x)) P(x, y)
$$
Equilibrium principle

Then

\[\mathbb{E}_\pi[(A S)(X)] = \mathbb{E}_\pi[S(Y_{n+1}) - S(Y_n)]\]
\[= \mathbb{E}_\pi[S(Y_{n+1})] - \mathbb{E}_\pi[S(Y_n)]\]
\[= 0.\]
Equilibrium principle

Time-invariance estimating equations


Find a Markov chain sampler $Y^{(\theta)} = (Y_t^{(\theta)}, \ t \geq 0)$ for the distribution of $X$ under $\theta$.

Let $A_\theta$ be the generator of $Y^{(\theta)}$. Then

$$\mathbb{E}_{\theta} \left[ (A_\theta S)(X) \right] = 0$$

for “any” statistic $S$. 
Choose a statistic $S'$. 

Given the observed data $x$, solve for $\theta$ in 

$$(\mathcal{A}_\theta S') (x) = 0.$$ 

This is an unbiased estimating equation for $\theta$. 

Solution $\hat{\theta}_T = \text{equilibrium estimator}$.
Example 1
Take any parametric model for $X$.
Let $Y_1^{(\theta)}, Y_2^{(\theta)}, \ldots$ be \textbf{i.i.d. realisations} of $X$ under $\theta$. Then $Y$ has discrete time generator

$$(\mathcal{A}_\theta S)(x) = \mathbb{E}_{\theta} S(X) - S(x)$$
Equilibrium principle

Example 1
Take any parametric model for $X$.
Let $Y_1^{(\theta)}$, $Y_2^{(\theta)}$, ... be i.i.d. realisations of $X$ under $\theta$. Then $Y$ has
discrete time generator

$$(\mathcal{A}_\theta S)(x) = \mathbb{E}_\theta S(X) - S(x)$$

Then $\hat{\theta}$ is the solution of

$$\mathbb{E}_\theta S(X) = S(x)$$

i.e. the method-of-moments estimator.
(In regular exponential families this gives the MLE)
Example 2
Finite system of r.v.'s \(X = (X_1, \ldots, X_M)\).
Let \(Y\) be the single-site **Gibbs sampler**.
Generator:

\[
(A_\theta S)(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_\theta[S(X) \mid X_{-i} = x_{-i}] - S(x)
\]
Equilibrium principle

Example 2

Finite system of r.v.’s \( \mathbf{X} = (X_1, \ldots, X_M) \).

Let \( Y \) be the single-site Gibbs sampler.

Generator:

\[
(A_{\theta} S) (\mathbf{x}) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{\theta} [S(\mathbf{X}) \mid X_{-i} = \mathbf{x}_{-i}] - S(\mathbf{x})
\]

Equilibrium equation \((A_{\theta} S)(\mathbf{x}) = 0\) is equivalent to maximum pseudolikelihood normal equations if we choose \( S \equiv V \) in exponential family \( f(\mathbf{x}; \theta) \propto \exp(\theta^T V(\mathbf{x})) \).
Equilibrium principle

Example 3
Point process $X$ with conditional intensity $\lambda(u, X)$.
Let $Y$ be the associated spatial birth-and-death process with unit death rate. Generator is

$$(A_\theta S)(x) = \int_W [S(x \cup \{u\}) - S(x)]\lambda(u, x) \, du$$

$$+ \sum_i [S(x \setminus \{x_i\}) - S(x)]$$
Equilibrium principle

Example 3
Point process $\mathbf{X}$ with conditional intensity $\lambda(u, \mathbf{X})$.
Let $\mathbf{Y}$ be the associated spatial birth-and-death process with unit death rate.
Generator is

$$\begin{align*}
(A_\theta S)(\mathbf{x}) &= \int_W \left[ S(\mathbf{x} \cup \{u\}) - S(\mathbf{x}) \right] \lambda(u, \mathbf{x}) \, du \\
&\quad + \sum_i \left[ S(\mathbf{x} \setminus \{x_i\}) - S(\mathbf{x}) \right]
\end{align*}$$

Time-invariance equation $(A_\theta S)(\mathbf{x}) = 0$ is equivalent to maximum pseudolikelihood normal equations if we choose $S \equiv V$ in exponential family $f(\mathbf{x}; \theta) \propto \exp(\theta^T V(\mathbf{x}))$. 
## Equilibrium principle

### Familiar examples

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$S(x)$</th>
<th>$\hat{\theta}_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>any</td>
<td>i.i.d.</td>
<td>any</td>
<td>method of moments</td>
</tr>
<tr>
<td>any</td>
<td>i.i.d.</td>
<td>$V(x)$</td>
<td>MLE</td>
</tr>
<tr>
<td>finite system of r.v.'s</td>
<td>1-site Gibbs sampler</td>
<td>$V(x)$</td>
<td>max. pseudolikelihood</td>
</tr>
<tr>
<td>Markov random field</td>
<td>1-site Gibbs sampler</td>
<td>$V(x)$</td>
<td>max. pseudolikelihood</td>
</tr>
<tr>
<td>spatial point process</td>
<td>spatial birth-death</td>
<td>any</td>
<td>Takacs-Fiksel estimator</td>
</tr>
<tr>
<td>spatial point process</td>
<td>spatial birth-death</td>
<td>$V(x)$</td>
<td>spatial MPLE</td>
</tr>
<tr>
<td>random censoring</td>
<td>i.i.d. (lifetimes)</td>
<td>e.d.f.</td>
<td>reduced sample estimator</td>
</tr>
</tbody>
</table>
## Equilibrium principle

### Unfamiliar/new estimators

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$S(x)$</th>
<th>$\hat{\theta}_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>continuous r.v.’s</td>
<td>Ornstein-Uhlenbeck diffusion</td>
<td>any</td>
<td>variational estimator</td>
</tr>
<tr>
<td>Markov random field</td>
<td>block-update Gibbs sampler</td>
<td>$V(x)$</td>
<td>generalised pseudolik.</td>
</tr>
<tr>
<td>spatial point process</td>
<td>multiple birth-death</td>
<td>$V(x)$</td>
<td>generalised pseudolik.</td>
</tr>
<tr>
<td>spatial point process</td>
<td>Barker dynamics</td>
<td>$V(x)$</td>
<td>logistic composite lik.</td>
</tr>
<tr>
<td>Markov random field</td>
<td>Metropolis-Hastings</td>
<td>$V(x)$</td>
<td>weird new estimator</td>
</tr>
</tbody>
</table>
Questions

What does it mean?
Questions

What does it mean?

- recipe for deriving estimators
Questions

What does it mean?

- recipe for deriving estimators
- common structure for different techniques of parameter estimation
  (MLE, method of moments, MPLE, Takacs-Fiksel, minimum contrast, variational, ... )
Questions

What does it mean?

- recipe for deriving estimators
- common structure for different techniques of parameter estimation (MLE, method of moments, MPLE, Takacs-Fiksel, minimum contrast, variational, ...)
- algebraic form of estimating equations corresponds to transition structure of Y
Questions

*What does it mean?*

- Recipe for deriving estimators
- Common structure for different techniques of parameter estimation (MLE, method of moments, MPLE, Takacs-Fiksel, minimum contrast, variational,…)
- Algebraic form of estimating equations corresponds to transition structure of $Y$
- When recipe is applied to ‘simple’ MCMC samplers $Y$ it yields ‘simple’ ad hoc estimators
A. Baddeley.

Time-invariance estimating equations.

A. Baddeley and D. Dereudre.

Variational estimators for the parameters of Gibbs point process models.

A. Baddeley, J.-F. Coeurjolly, E. Rubak and R. Waagepetersen

Logistic Regression for Spatial Gibbs Point Processes.