Noise contrastive estimation

Poisson transform, asymptotics, comparison with MC-MLE

Nicolas Chopin (ENSAE)

(based on joint work with Simon Barthelmé, CNRS, Gipsa-LAB; and Lionel Riou-Durand, ENSAE)

Part I

Unnormalised statistical models

Unnormalised statistical models

- "Unnormalised" statistical models: models with an intractable normalisation constant in the likelihood.
- Popular in Machine Learning, Computer Vision (deep learning), neuroscience.
- Creates computational difficulties (*"doubly intractable problems"* in Bayesian context).

Unnormalised models: examples

Unnormalised models often correspond to some exponential family

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{\exp\left\{\boldsymbol{\theta}^T S(y)\right\}}{Z(\boldsymbol{\theta})}$$

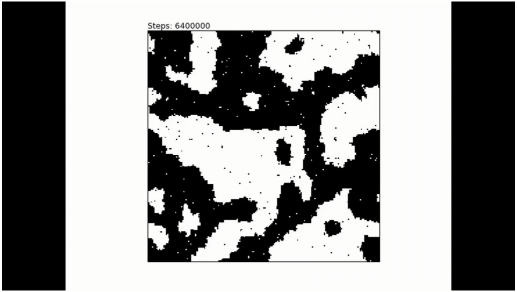
where $Z(\boldsymbol{\theta})$ is intractable. Examples follow.

Example 1: Ising models

$\boldsymbol{y} = (y_1, \ldots, y_n)$ where the y_i's are binary variables y_i observed on a lattice:

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{\exp\left\{\alpha \sum_i y_i + \beta \sum_{i \sim j} \mathbf{1}(y_i = y_j)\right\}}{Z(\boldsymbol{\theta})}$$

with $\boldsymbol{\theta} = (\alpha, \beta)$, $Z(\boldsymbol{\theta})$ is then a sum of 2^n terms.



Example 2: exponential random graphs (and networks)

$$p_{\boldsymbol{\theta}}(\mathbf{y}) = \frac{\exp\left\{\boldsymbol{\theta}^T S(\mathbf{y})\right\}}{Z(\boldsymbol{\theta})}$$

where $\mathbf{y} = (y_1, \ldots, y_p)$, and $y_i = 1$ if edge i in the network is active; $Z(\boldsymbol{\theta})$ is then a sum of 2^p terms, with $p = n(n-1)/2$.

Example 3: truncated Gaussian distribution

IID data-points $y_1, \ldots, y_n$ from a Gaussian distribution truncated to $]0, \infty[^d$:

$$f_{\mu,\Sigma}(y) = \frac{1}{Z(\mu, \Sigma)} \exp\left\{-(y-\mu)^T \Sigma^{-1}(y-\mu)\right\} \mathbf{1}_{]0,\infty[^d}(y)$$

then $Z(\mu, \Sigma)$ is a $d-$dimensional integral which is difficult to approximate when d gets large.

Unnormalised *sequential* models

- Markov sequence $\mathbf{y}_0, \ldots, \mathbf{y}_n$ where the transition kernel is defined up to a constant.
- Example: sequential Ising

$$p(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{a}, \mathbf{Q}, \mathbf{R}) \propto \exp\left(\mathbf{a}^t \mathbf{y}_t + \mathbf{y}_t \mathbf{Q} \mathbf{y}_t + \mathbf{y}_t \mathbf{R} \mathbf{y}_{t-1}\right)$$

- Nastier than IID version: n normalisation constants missing.

Current strategies for inference

- Classical estimation: MCMC-MLE (Geyer, 1994), contrastive divergence (Bengio and Delalleau, 2009), noise-contrastive divergence (Gutmann and Hyvärinen, 2012).
- Bayesian: exchange algorithm (Murray et al., 2012), ABC, russian roulette (Girolami et al., 2013).
- I do not know of methods for *sequential* unnormalised models.

Our contribution

- Poisson transform shows you can treat the missing normalisation constant as just another parameter. Gives you an alternative likelihood function.
- Applies to sequential problems as well.
- Noise-contrastive divergence is an approximation of the Poisson transform and we can now extend it to the sequential setting.
- Sequential estimation can be turned into a *semiparametric logistic regression* problem.
- Poisson transform simplifies asymptotic theory and formal comparison between methods.

Part II

The Poisson transform

Poisson point processes

- Poisson processes are distributions over countable subsets of a domain Ω (e.g., $\Omega = \mathbb{R}$ for a temporal point process).
- Let S be a realisation from a PP. For all (measurable) $\mathcal{A} \subseteq \Omega$, the number of points of S in $\mathcal{A}$ follows a Poisson distribution with parameter

$$\lambda_{\mathcal{A}} = \mathbb{E}\left(|S \cap \mathcal{A}|\right) = \int_{\mathcal{A}} \lambda\left(\mathbf{y}\right) \mathrm{d}\mathbf{y}$$

where $\lambda\left(\boldsymbol{y}\right)$ is the intensity function.

Poisson point processes (II)

Let's assume that $\int_\Omega \lambda\left(\mathbf{y}\right) \mathrm{d}\mathbf{y} < \infty$, then

- The cardinal of S is Poisson, with parameter $\int_\Omega \lambda\left(\mathbf{y}\right) \mathrm{d}\mathbf{y} < \infty$;
- conditional on $|S| = k$, the elements of S are IID with density

$$\propto \exp\left\{\lambda(\boldsymbol{y})\right\}.$$

Poisson point processes (III)

The likelihood of a Poisson process is:

$$\log p\left(S|\lambda\right) = \sum_{\mathbf{y}_i \in S} \log \lambda\left(\mathbf{y}_i\right) - \int_{\Omega} \lambda\left(\mathbf{y}\right) \mathrm{d}\mathbf{y}$$

Poisson point processes (IV)

Consider $S_1 \sim \mathrm{PP}(\lambda_1)$, $S_2 \sim \mathrm{PP}(\lambda_2)$:

- $S_1 \cup S_2 \sim \mathrm{PP}(\lambda_1 + \lambda_2)$
- a point y in $S_1 \cup S_2$ originates from S_1 with probability

$$\frac{\lambda_1(y)}{\lambda_1(y) + \lambda_2(y)}$$

The Poisson transform

- Generalisation of the Poisson-Multinomial transform (Baker, 1994)
- For estimation purposes, you can treat IID data in just about any space as coming from a Poisson process.
- New likelihood function: no loss of information, one extra latent parameter.

Theorem statement (I)

Data: $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \Omega$, density $p(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\{f_{\boldsymbol{\theta}}(\mathbf{y})\}$, so log-likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} f_{\boldsymbol{\theta}}(\mathbf{y}_i) - n \log \int_{\Omega} \exp\{f_{\boldsymbol{\theta}}(\mathbf{y})\} \, d\mathbf{y}.$$

Theorem statement (I)

Data: $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \Omega$, density $p(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\{f_{\boldsymbol{\theta}}(\mathbf{y})\}$, so log-likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} f_{\boldsymbol{\theta}}(\mathbf{y}_i) - n \log \int_{\Omega} \exp\{f_{\boldsymbol{\theta}}(\mathbf{y})\} \mathrm{d}\mathbf{y}.$$

Poisson log-likelihood:

$$\mathcal{M}(\boldsymbol{\theta}, \nu) = \sum_{i=1}^{n} \{f_{\boldsymbol{\theta}}(\mathbf{y}_i) + \nu\} - n \int_{\Omega} \exp\{f_{\boldsymbol{\theta}}(\mathbf{y}) + \nu\} \mathrm{d}\mathbf{y}$$

i.e. log-likelihood of a PP with intensity $\lambda(\boldsymbol{y}) = f_{\boldsymbol{\theta}}(\boldsymbol{y}) + \nu$.

Theorem statement (II)

Theorem

Let $\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta})$ *and* $\left(\tilde{\boldsymbol{\theta}}, \nu^\star\right) = \underset{\boldsymbol{\theta}\in\Theta, \nu\in\mathbb{R}}{\operatorname{argmax}}\, \mathcal{M}(\boldsymbol{\theta}, \nu)$*. Then* $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star$ *and* $\nu^\star = -\log\left(\int \exp\{f_{\boldsymbol{\theta}^\star}(\mathbf{y})\}\, d\mathbf{y}\right)$.

In other words, the MLE can be computed by maximising $\mathcal{M}(\boldsymbol{\theta}, \nu)$ in both variables. There is no loss of information. Also, asymptotic confidence intervals for $\boldsymbol{\theta}$ are the same. The latent variable ν "estimates" the normalisation constant.

Proof

For fixed $\boldsymbol{\theta}$, maximise $\mathcal{M}(\boldsymbol{\theta}, \nu)$ wrt ν leads to:

$$\nu^\star(\boldsymbol{\theta}) = -\log\left(\int \exp\left\{f_{\boldsymbol{\theta}^\star}(\mathbf{y})\right\} \mathrm{d}\mathbf{y}\right)$$

and

$$\mathcal{M}(\boldsymbol{\theta}, \nu^\star(\boldsymbol{\theta})) = \mathcal{L}(\boldsymbol{\theta}) - n.$$

Extension to sequential models

The same logic can be applied to sequential models:

$$p_{\theta}(\mathbf{y}_t|\mathbf{y}_{t-1}) \propto \exp\left\{f_{\theta}(\mathbf{y}_t, \mathbf{y}_{t-1})\right\}$$

We will apply the Poisson transform to each conditional distribution.

Extension to sequential models

- Original log-likelihood of sequence:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{n} \left[f_{\boldsymbol{\theta}}(\mathbf{y}_t; \mathbf{y}_{t-1}) - \log \left(\int_{\Omega} \exp \left\{ f_{\boldsymbol{\theta}} \left(\mathbf{y}; \mathbf{y}_{t-1} \right) \right\} \mathrm{d}\mathbf{y} \right) \right]$$

Extension to sequential models

- Original log-likelihood of sequence:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{n} \left[f_{\boldsymbol{\theta}}(\mathbf{y}_t; \mathbf{y}_{t-1}) - \log \left(\int_{\Omega} \exp\left\{ f_{\boldsymbol{\theta}}\left(\mathbf{y}; \mathbf{y}_{t-1}\right)\right\} \mathrm{d}\mathbf{y} \right) \right]$$

- Poisson-transformed log-likelihood:

$$\mathcal{M}\left(\boldsymbol{\theta}, \boldsymbol{\nu}\right) = \sum_{t=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathbf{y}_t; \mathbf{y}_{t-1}) + \nu_{t-1} \right\} - \int_{\Omega} \sum_{t=1}^{n} \exp\left\{ f_{\boldsymbol{\theta}}\left(\mathbf{y}; \mathbf{y}_{t-1}\right) + \nu_{t-1} \right\} \mathrm{d}\mathbf{y}$$

We have introduced one latent variable ν_t per observation. Sum of integrals becomes integral of a *single* sum.

Extension to sequential models

Maximising the Poisson-transformed likelihood wrt $(\boldsymbol{\theta}, \boldsymbol{\nu})$, gives the MLE for $\boldsymbol{\theta}$, and

$$\nu^{\star}_{t-1}\left(\boldsymbol{\theta}^{\star}\right) = -\log\left(\int \exp\left\{f_{\boldsymbol{\theta}^{\star}}\left(\mathbf{y};\mathbf{y}_{t-1}\right)\right\} \mathrm{d}\mathbf{y}\right),$$

i.e. minus the log-marginalisation constant for the conditional

$$p(\mathbf{y}|\mathbf{y}_{t-1}, \boldsymbol{\theta}^{\star}) \propto \exp\left\{f_{\boldsymbol{\theta}^{\star}}(\mathbf{y};\mathbf{y}_{t-1})\right\}.$$

From parametric to semi-parametric inference

The value of the latent variables at the mode are a function of $\mathbf{y}_{t-1}$:

$$\nu^{\star}_{t-1}\left(\boldsymbol{\theta}^{\star}\right) = -\log\left(\int \exp\left\{f_{\boldsymbol{\theta}^{\star}}\left(\mathbf{y};\mathbf{y}_{t-1}\right)\right\} \mathrm{d}\mathbf{y}\right) = \chi(\mathbf{y}_{t-1}).$$

If $\mathbf{y}_t$, $\mathbf{y}_{t'}$ are close, ν_t, ν_{t-1} should be close as well, i.e., $\chi\left(\mathbf{y}\right)$ is (hopefully) smooth.

From parametric to semi-parametric inference

The value of the latent variables at the mode are a function of $\mathbf{y}_{t-1}$:

$$\nu_{t-1}^{\star}\left(\boldsymbol{\theta}^{\star}\right) = -\log\left(\int \exp\left\{f_{\boldsymbol{\theta}^{\star}}\left(\mathbf{y};\mathbf{y}_{t-1}\right)\right\} \mathrm{d}\mathbf{y}\right) = \chi(\mathbf{y}_{t-1}).$$

If $\mathbf{y}_t$, $\mathbf{y}_{t'}$ are close, ν_t, ν_{t-1} should be close as well, i.e., $\chi\left(\mathbf{y}\right)$ is (hopefully) smooth.

$\Rightarrow$ Do inference over χ: e.g. if you have n points but χ is well captured by a spline basis with $k \ll n$ components, use spline basis instead. Poisson likelihood becomes:

$$\begin{aligned}\mathcal{M}(\boldsymbol{\theta},\chi) = & \sum_{t=1}^{n}\left\{f_{\boldsymbol{\theta}}(\mathbf{y}_t;\mathbf{y}_{t-1}) + \chi(\mathbf{y}_{t-1})\right\} \\ & - \int_{\Omega}\sum_{t=1}^{n}\exp\left\{f_{\boldsymbol{\theta}}\left(\mathbf{y};\mathbf{y}_{t-1}\right) + \chi(\mathbf{y}_{t-1})\right\}\mathrm{d}\mathbf{y}\end{aligned}$$

Using the Poisson transform in practice

Back to the IID case: Poisson-transformed likelihood still involves an intractable integral

$$\mathcal{M}(\boldsymbol{\theta}, \nu) = \sum_{i=1}^{n} \{f_{\boldsymbol{\theta}}(\mathbf{y}_i) + \nu\} - n \int_{\Omega} \exp\{f_{\boldsymbol{\theta}}(\mathbf{y}) + \nu\} \, \mathrm{d}\mathbf{y}$$

which we need to approximate.

Using the Poisson transform in practice

Back to the IID case: Poisson-transformed likelihood still involves an intractable integral

$$\mathcal{M}(\boldsymbol{\theta}, \nu) = \sum_{i=1}^{n} \{f_{\boldsymbol{\theta}}(\mathbf{y}_i) + \nu\} - n \int_{\Omega} \exp\{f_{\boldsymbol{\theta}}(\mathbf{y}) + \nu\} \, \mathrm{d}\mathbf{y}$$

which we need to approximate.
Several ways, but an interesting one is to go through logistic regression.

Stochastic gradient descent

Before we go to logistic regression, note that another approach would be to use Monte Carlo (importance sampling) to obtain an *unbiased* estimate of the gradient:

$$
\begin{aligned}
\frac{1}{n}\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{M}(\boldsymbol{\theta},\nu) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{y}_i;\boldsymbol{\theta}) - \int_\Omega \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{y}_i;\boldsymbol{\theta}) \exp\left(f_{\boldsymbol{\theta}}(\mathbf{y})+\nu\right)\mathrm{d}\mathbf{y} \\
\frac{1}{n}\frac{\partial}{\partial \nu}\mathcal{M}(\boldsymbol{\theta},\nu) &= 1 - \int_\Omega \exp\left(f_{\boldsymbol{\theta}}(\mathbf{y})+\nu\right)\mathrm{d}\mathbf{y}
\end{aligned}
$$

The we could use SGD (stochastic gradient descent) to maximise $\mathcal{M}(\boldsymbol{\theta},\nu)$.

Part III

The logistic trick & noise-contrastive divergence

The logistic trick

- Idea: reduce an estimation problem to a classification problem.
- Several versions:
  - Logistic regression for density estimation: Hastie et al. (2003), intensity estimation: Baddeley et al. (2010).
  - Logistic regression for normalisation constants: Geyer (1994).
  - Logistic regression for estimation in unnormalised models: Gutmann and Hyvärinen (2012).
- The last one is called “noise-contrastive divergence” by the authors.

The logistic trick

We have n random points from distributions $p(y)$ and n points from $q(y)$. We note $z_i = 1$ if the i-th point is from p, $z_i = 0$ otherwise. Logistic regression models the log-odds ratio:

$$\eta(y) = \log \frac{p(z=1|y)}{p(z=0|y)}.$$

We have that:

$$\eta(y) = \log \frac{p(y)}{q(y)}$$

$\Rightarrow$ provided $q(y)$ is known, we can first estimate η (doing some form of logistic regression), and then recover $p(y)$ from $\eta(y)$.

From the logistic trick to noise-contrastive divergence

If we have a *normalised* model $p_{\boldsymbol{\theta}}(y)$ then we can run a logistic regression with the following model for the log-odds:

$$\eta\left(y;\boldsymbol{\theta}\right)=\log p_{\boldsymbol{\theta}}(y)-\log q\left(y\right).$$

From the logistic trick to noise-contrastive divergence

If we have a *normalised* model $p_{\boldsymbol{\theta}}(y)$ then we can run a logistic regression with the following model for the log-odds:

$$\eta\left(y;\boldsymbol{\theta}\right) = \log p_{\boldsymbol{\theta}}(y) - \log q\left(y\right).$$

If the model is *unnormalised,* $p_{\boldsymbol{\theta}}(y) \propto \exp\left\{f_{\boldsymbol{\theta}}(y)\right\}$, we introduce an intercept in the logistic regression

$$\eta\left(y;\boldsymbol{\theta}\right) = f_{\boldsymbol{\theta}}(y) + \nu - \log q\left(y\right).$$

This is the *noise-contrastive divergence* (NCD) technique of Gutmann and Hyvärinen (2012).

Simple interpretation in terms of PP

- We interpret $\boldsymbol{y} = (y_1, \ldots, y_n)$ as the realisation of a PP with intensity $\lambda_1(y) = n \exp\{f_{\boldsymbol{\theta}}(y) + \nu\}$;
- We interpret $\boldsymbol{x} = (x_1, \ldots, x_m)$ as the realisation of a PP with (known) intensity $\lambda_2(y) = mq(y)$;
- We mix $\boldsymbol{x}$ and $\boldsymbol{y}$, and then try to recover the probability that a given point comes from the original data as

$$\log \frac{p}{1-p} = \log \frac{\lambda_1(y)}{\lambda_2(y)} = \log \frac{n}{m} + f_{\boldsymbol{\theta}}(y) + \nu - \log q(y)$$

- If $f_{\boldsymbol{\theta}}(y) = \boldsymbol{\theta}^T S(y)$ (exponential model), this is equivalent to a basic logistic regression (with $S(y)$ as covariate).

Toy example: truncated exponential

Recall the truncated exponential model:

$$p(y|\theta) \propto \exp(\theta y)$$

We produce reference samples from $U(0,1)$, so that the logistic model for NCD is just:

$$\eta(y;\theta) = \theta y + \nu$$

Fitting in R:

```
m <- glm(z~y+offset(logratio),data=df,family=binomial)
```

Summary

- Logistic trick: get a logistic classifier to discriminate true data from random reference data (from a known distribution). It implicitly learns a model for the true data
- NCD: in unnormalised models, introduce an intercept for the missing normalisation constant
- Our interpretation: NCD is an approximation of the Poisson-transformed likelihood

NCD approximates the Poisson transform

- In NCD, you can introduce as many reference points (points simulated from q) as you like.
- Parametrise the log-odds by

$$\eta(\mathbf{y}) = f_{\boldsymbol{\theta}}(\mathbf{y}) + \nu + \log \frac{n}{m} - \log q(\mathbf{y})$$

  where m is the number of reference points.
- Theorem: as $m \to +\infty$, the logistic log-likelihood $\mathcal{R}^m(\boldsymbol{\theta}, \nu)$ tends to the Poisson log-likelihood $\mathcal{M}(\boldsymbol{\theta}, \nu)$ (pointwise).

NCD approximates the Poisson transform

To sum up: take your true n datapoints, add m random reference datapoints, and estimate the model

$$p_{\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left\{f_{\boldsymbol{\theta}}(\mathbf{y})\right\}$$

using a logistic regression with log-odds

$$\eta\left(\mathbf{y}\right) = f_{\boldsymbol{\theta}}(\mathbf{y}) + \nu + \log\frac{n}{m} - \log q(\mathbf{y})$$

The intercept will be used to estimate the missing normalisation constant. The technique is effectively a practical way of approximating a Poisson-transformed likelihood.

NCD for sequential models

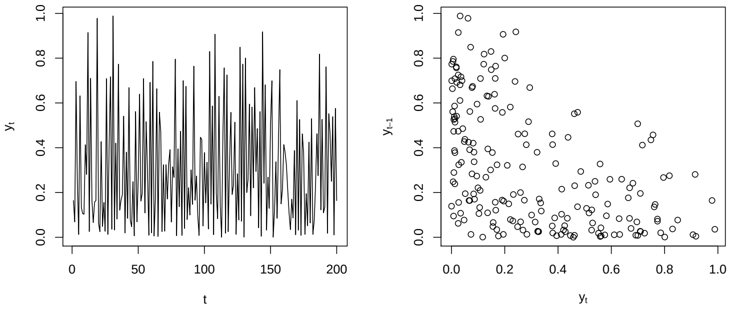
The relationship between the Poisson transform and NCD shows directly how to adapt NCD to sequential models: apply NCD to each conditional distribution (the transition kernels)

- Reference density $q(\mathbf{y})$ becomes a reference kernel $q\left(\mathbf{y}_t|\mathbf{y}_{t-1}\right)$
- Include an intercept ν_t per conditional distribution $p(\mathbf{y}_t|\mathbf{y}_{t-1},\boldsymbol{\theta})$

Truncated exponential, revisited

We turn our previous example into a Markov chain:

$$p(y_t | y_{t-1}, \theta) \propto \exp\left(\theta y_t y_{t-1}\right)$$



Truncated exponential, revisited

Consider the NCD approximation for *fixed* y_{t-1}. The model for the log-odds will take the form:

$$\eta\left(y_t\right) = \theta y_t y_{t-1} + \nu_{t-1} + \log\frac{n}{m} - \log q(y_t|y_{t-1})$$

This leads to a linear logistic regression with $y_t y_{t-1}$ as a covariate.

Parametric vs. semi-parametric model

It is wasteful to fit a separate intercept per time-point. As in the semi-parametric version of the Poisson transform, we can use:

$$\eta\left(y_t\right) = \theta y_t y_{t-1} + \chi\left(y_{t-1}\right) + \log\frac{n}{n_r} - \log q(y_t|y_{t-1})$$

where $\chi\left(y_{t-1}\right)$ will be fitted using splines.

In practice (I)

Positive examples are given by:

Value at time $t-1$	Value at time t	Label
y_1	y_2	1
y_2	y_3	1
$\vdots$	$\vdots$	$\vdots$
y_{n-1}	y_n	1

While negative examples are given by:

Value at time $t-1$	Value at time t	Label
y_1	r_2	0
y_2	r_3	0
$\vdots$	$\vdots$	$\vdots$
y_{n-1}	r_n	0

In practice (II)

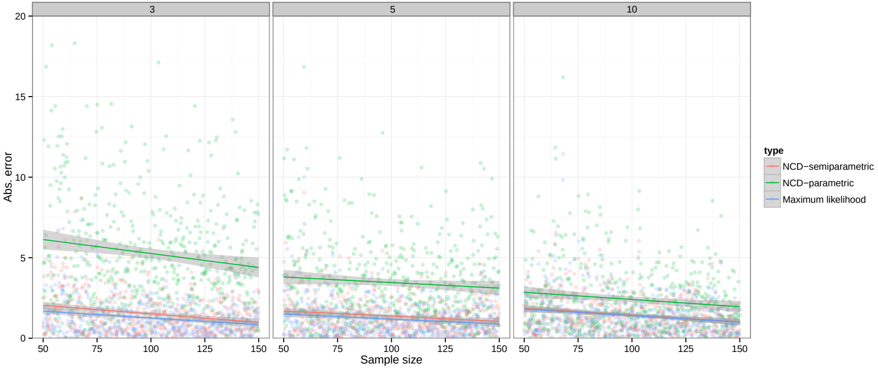
We can fit the (semi-parametric) model via:

```
m <- gam(label ~ I(y_t*y_tminusone)+s(y_tminusone),data=
```

The fully parametric model corresponds to:

```
m <- gam(label ~ I(y_t*y_tminusone)+as.factor(y_tminuso
```

Parametric vs. semi-parametric model



Part IV

Application: LATKES

Data

Figure: A sequence of eye movements extracted from the dataset of Kienzle et al. (2009). Fixation locations are in red and successive locations are linked by a straight line.

Eye movements recorded while 14 subjects where exploring a set of photographs (Fig. 1); each contributing between 600 and 2,000 datapoints.

LATKES

LATKES: Log-Additive Transition Kernels. A class of spatial Markov chain models, with applications to eye movement data:

$$p(y_t|y_{t-1},\ldots,y_{t-k}) \propto \exp\left\{\sum \beta_i v_i(y_t) + g(y_t, y_{t-1},\ldots,y_{t-k})\right\}$$

where $y_1 \ldots y_t$ are spatial locations (e.g. on a screen), $v_i\,(y)$ are spatial covariates, $g(...)$ is an interaction kernel.

Fitting LATKES using logistic regression

- Transition kernel only specified up to normalisation constant.
- Can use sequential version of NCD to turn the problem into (semiparametric) logistic regression.
- Standard packages can be used (mgcv, INLA).

Example

We fit the model:

$$p(y_t|y_{t-1}) \propto \exp\left\{b(||y_t||) + r_{\mathrm{dist}}\left(||\mathbf{y}_t - \mathbf{y}_{t-1}||\right) + r_{\mathrm{ang}}\left(\angle\left(\mathbf{y}_t - \mathbf{y}_{t-1}\right)\right)\right\}$$

where:

- $b(||y_t||)$ should reflect a centrality bias;
- $r_{\mathrm{dist}}\left(||\mathbf{y}_t - \mathbf{y}_{t-1}||\right)$ should reflect the fact that successive fixations are close together;
- $r_{\mathrm{ang}}\left(\angle\left(\mathbf{y}_t - \mathbf{y}_{t-1}\right)\right)$ should reflect a tendency for making movements along the cardinal axes (vertical and horizontal).

Note on NCD implementation

- We fitted functions b, r_{dist} and r_{ang} (plus the log-normalising constant χ, as already explained) using smoothing splines. (Extension of NCD to smoothing splines is direct: simply add appropriate penalty to log-likelihood).
- We used R package *mgcv* (Wood, 2006).
- Reference points were sampled from an Uniform distribution (over the screen); 20 reference datapoints per datapoint.
- Requires one line of code of R, took about 5 minutes.

Results

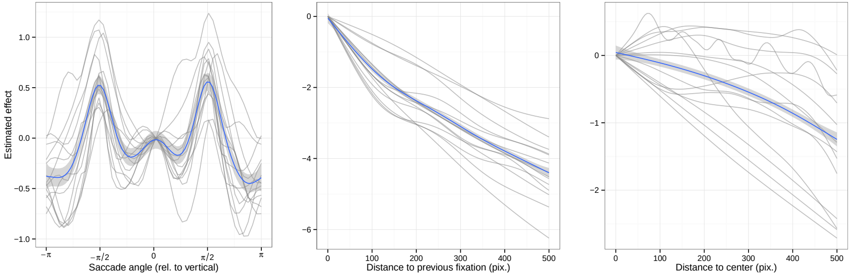


Figure: The different panels display the estimated effects of saccade angle (r_{ang}), distance to previous fixation (r_{dist}) and centrality bias (s). Individual subjects are in gray, and the group average is in blue.

Part V

Asymptotics

Inspiration: MCMC-MLE

From now on, we focus on the IID case: $y_1, \dots, y_n$ have density $\propto h_{\boldsymbol{\theta}}(y) \stackrel{\Delta}{=} \exp\{f_{\boldsymbol{\theta}}(y)\}$. Up to a linear transformation, the log-likelihood is:

$$\frac{1}{n}\sum_{i=1}^{n}\{f_{\boldsymbol{\theta}}(y_i) - f_{\psi}(y_i)\} - \log\frac{Z(\boldsymbol{\theta})}{Z(\psi)}$$

where ψ is arbitrary. MCMC-MLE (Geyer, 1994): runs MCMC to sample M points $\boldsymbol{x}_m$ from $\propto \exp\{f_{\psi}(y)\}$, use importance sampling to approximate second term

$$\frac{1}{m}\sum_{j=1}^{m}\frac{h_{\boldsymbol{\theta}}(\boldsymbol{x}_m)}{h_{\psi}(\boldsymbol{x}_m)} \approx \frac{Z(\boldsymbol{\theta})}{Z(\psi)}$$

(since $Z(\boldsymbol{\theta}) = \int h_{\boldsymbol{\theta}}(y)\,dy$) then maximise:

$$\ell^{\mathrm{IS}}_{n,m}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\{f_{\boldsymbol{\theta}}(y_i) - f_{\psi}(y_i)\} - \log\left\{\frac{1}{m}\sum_{i=1}^{m}\frac{h_{\boldsymbol{\theta}}(\boldsymbol{x}_m)}{h_{\psi}(\boldsymbol{x}_m)}\right\}$$

Objective

Clearly NCE and MCMC-MLE have a lot in common: both generate M “fake” data-points from the model with $\boldsymbol{\theta} = \boldsymbol{\psi}$, and both maximise some pseudo-likelihood. How to compare them formally?

Comparing the maximisers of three (pseudo-)likelihoods

$$\hat{\theta}_n = \arg\max \ell_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\{f_{\boldsymbol{\theta}}(y_i) - f_{\psi}(y_i)\} - \log\frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\psi})}$$

$$\hat{\theta}^{\mathrm{IS}}_{n,m} = \arg\max \ell^{\mathrm{IS}}_{n,m}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\{f_{\boldsymbol{\theta}}(y_i) - f_{\psi}(y_i)\} - \log\left\{\frac{1}{m}\sum_{j=1}^{m}\frac{h_{\boldsymbol{\theta}}(}{h_{\psi}(}\right.$$

$$(\hat{\theta}^{\mathrm{NCE}}_{n,m}, \hat{\nu}) = \arg\max \ell^{\mathrm{NCE}}_{n,m}(\boldsymbol{\theta}, \nu) = \sum_{i=1}^{n}\log q_{\boldsymbol{\theta},\nu}(y_i) + \sum_{i=1}^{n}\log\{1 - q_{\boldsymbol{\theta},\nu}(y_i)$$

with

$$\log\left\{\frac{q_{\boldsymbol{\theta},\eta}(y)}{1 - q_{\boldsymbol{\theta},\eta}(y)}\right\} = f_{\boldsymbol{\theta}}(y) - f_{\psi}(y) + \eta + \log\left(\frac{n}{m}\right)$$

s

Dealing with different parameter spaces

It is convenient to define all these functions with respect to the extended parameter space $\boldsymbol{\xi} = (\boldsymbol{\theta}, \nu) \in \Theta \times \mathbb{R}^+$. To do so:

1. We use the Poisson transform for the true log-likelihood:

$$\ell_n(\boldsymbol{\theta}, \eta) = \frac{1}{n} \sum_{i=1}^{n} \{f_{\boldsymbol{\theta}}(y_i) - f_{\psi}(y_i)\} + \eta - e^{\nu} \times \frac{Z(\boldsymbol{\theta})}{Z(\psi)}$$

2. We define a similar Poisson transform for the MCMC-MLE likelihood.

All the derivations are done in terms of $\boldsymbol{\xi}$, but for your convenience, I will state the results in terms of $\boldsymbol{\theta}$.

Asymptotic regime?

- Geyer (1994) takes $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ fixed, and show that $\hat{\boldsymbol{\theta}}^{\mathrm{IS}}_{n,m} \to \hat{\boldsymbol{\theta}}_n$ as $m \to +\infty$.
- Gutmann and Hyvärinen (2012) take $m = \tau n$, and show that $\hat{\boldsymbol{\theta}}^{\mathrm{NCE}}_{n,m} \to \boldsymbol{\theta}^\star$ as m, $n \to +\infty$.

Monte Carlo error

Th 1

For $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ fixed, and some constant c, under conditions similar to Geyer (1994)

$$m(\hat{\theta}^{\mathrm{NCE}}_{n,m} - \hat{\theta}^{\mathrm{IS}}_{n,m}) \underset{m\to\infty}{\to} c.$$

Implications

Geyer (1994) showed that, under very mild assumptions (e.g. allowing the $\boldsymbol{x}_m$'s to be generated by MCMC), $\hat{\theta}^{\mathrm{IS}}_{n,m}$ is a consistent, asymptotically normal estimator of the true MLE $\hat{\theta}_n$. Our result implies that $\hat{\theta}^{\mathrm{NCE}}_{n,m}$ has exactly the same properties, since it is at m^{-1} distance from $\hat{\theta}^{\mathrm{IS}}_{n,m}$.
Note: Geyer (1994)'s proof is based on concepts such as hypoconvergence (the weakest form of convergence of functions that ensures that maximisers also converge).

Overall error

Th 2

Let $m = \tau n$, and take, $m, n \to +\infty$ (for τ fixed). Under mild assumptions (MCMC sampling),

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{IS}}_{n,m} - \boldsymbol{\theta}^\star\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0_d, V_\tau^{IS})$$

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{NCE}}_{n,m} - \boldsymbol{\theta}^\star\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0_d, V_\tau^{NCE})$$

Asymptotic variance comparison

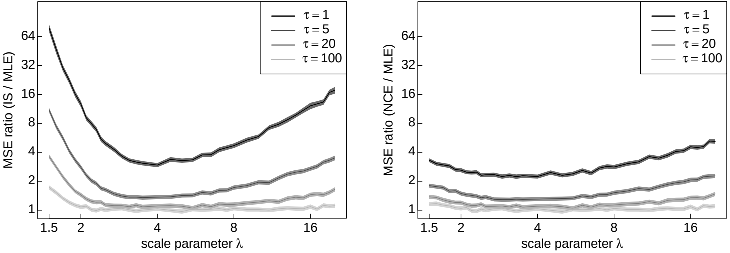
Th 3

Under IID sampling (for the $\boldsymbol{x}_j'$s), for any $\tau > 0$, and any ψ,

$$V_\tau^{NCE} \leq V_\tau^{IS}.$$

Numerical comparison

Model: $y_1, \ldots, y_n \sim \mathcal{N}_{>0}(\mu, \Sigma)$ (truncated to $]0, +\infty[^d$), reference data simulated from $\mathcal{N}_{>0}(0, \lambda I_d)$.



Conclusions

- Poisson transform: you can treat any data as coming from a Poisson point process in the appropriate space, and infer the *intensity* rather than the density.
    - *It is OK to treat the normalisation constant as a free parameter!*
- NCE effectively approximates the Poisson transform via *logistic regression*.
- Inference for unnormalised sequential models can be turned into semi-parametric logistic regression
    - True as well for unnormalised models with covariates
- For the same CPU budget (m simulations from reference parameter), NCE should be more robust and and more accurate that MCMC-MLE (and more convenient to use).

Relevant papers

- Barthelmé, S., and NC. The Poisson transform for unnormalised statistical models. Statistics and Computing 25, 4 (2015), 767--780.
- Riou-Durand, L. and NC. Noise contrastive estimation: asymptotics, comparison with MC-MLE, arXiv:1801.10381.

References I

Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A., Milne, R. K., Schuhmacher, D., Shah, R., and Turner, R. (2010). Spatial logistic regression and change-of-support in poisson point processes. *Electronic Journal of Statistics*, 4(0):1151–1201.

Baker, S. G. (1994). The Multinomial-Poisson transformation. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 43(4):495–504.

Bengio, Y. and Delalleau, O. (2009). Justifying and generalizing contrastive divergence. *Neural computation*, 21(6):1601–1621.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in markov chain monte carlo. Technical Report 568, School of Statistics, University of Minnesota.

Girolami, M., Lyne, A.-M., Strathmann, H., Simpson, D., and Atchade, Y. (2013). Playing russian roulette with intractable likelihoods. *arxiv 1306.4032*.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13(1):307–361.

References II

Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The Elements of Statistical Learning*. Springer, corrected edition.

Kienzle, W., Franz, M. O., Schölkopf, B., and Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of vision*, 9(5).

Murray, I., Ghahramani, Z., and MacKay, D. (2012). MCMC for doubly-intractable distributions.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 1 edition.