2016 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 13–16, 2016, SALERNO, ITALY

SCORE-MATCHING ESTIMATORS FOR CONTINUOUS-TIME POINT-PROCESS REGRESSION MODELS

Maneesh Sahani, Gergő Bohner, Arne Meyer

University College London
Gatsby Computational Neuroscience Unit
25 Howland Street, London W1T 4JG.

ABSTRACT

We introduce a new class of efficient estimators based on score matching for probabilistic point process models. Unlike discretised likelihood-based estimators, score matching estimators operate on continuous-time data, with computational demands that grow with the number of events rather than with total observation time. Furthermore, estimators for many common regression models can be obtained in closed form, rather than by iteration. This new approach to estimation may thus expand the range of tractable models available for event-based data.

***Index Terms*—** point-process, score matching, estimation, spike train, neural data

1. INTRODUCTION

A point process is a probability law governing the distribution of a random subset of points drawn from a specified space [1]. In the most common application, the space is the real line, and the points define events in time. Such data arise in a wide range of applications: including neurophysiology, seismology, queuing theory and network traffic analysis. Many of the principles developed here apply to point processes over any continuous space; however, for simplicity and compactness we limit the exposition to point processes in time.

We often wish to fit the parameters of a point process model to one or more observed sets of events. The simplest point process is the Poisson process, which may be defined by a parametrised intensity function $\lambda^\theta(t)$. The log likelihood of a Poisson process for observed events $\{t_1, t_2, \dots\}$ is:

$$\log p(t_1, t_2, \dots|\theta) = \sum_i \log \lambda^\theta(t_i) - \int_{-\infty}^{\infty} dt\ \lambda^\theta(t)$$

For many choices of parametrisation, evaluation of the integral in this likelihood is intractable. Thus, exact maximum likelihood estimation of θ may be computationally challenging. One common approach is to discretise time, replacing the integral by a sum (e.g., as in [2]).

Score matching was introduced by Hyvärinen [3] as an alternative estimation approach based on matching the derivative in the data space of the log-density of the model to the log-density of the empirical (unknown) distribution. It was motivated as a way to estimate parameters in distributions for which the normalising constant is intractable (as the derivatives being matched do not depend on this constant). It might thus seem plausible that a score matching estimator would also help to avoid the complications associated with an intractable integral in the likelihood above. This hope will be borne out.

2. SCORE MATCHING FOR POINT PROCESSES

Consider a point process defined on an interval $[0, T]$. A single sample-path from the process may be represented as non-decreasing counting function $N : [0, T] \to \mathbb{Z}_+$, which makes unit transitions at the event times $\mathcal{T} = \{t_1, t_2, t_3, \dots, t_{N(T)}\}$ with $0 \le t_1 < t_2 < t_3 < \dots t_{N(T)} \le T$ (we have assumed that no two events occur at the same time). Note that the number of events $N(T)$ is also a random variable. We write $\mathbb{T}$ for the collection of feasible event sets and $\mathbb{T}_N$ for the collection conditioned on the count $N(T)$. Let $\mathcal{P}^*$ represent the true process that generated the sample, and assume that it has an associated density function $p^*(\mathcal{T})$. We have a parametric model process $\mathcal{P}^\theta$, with density p^θ depending on an unknown θ which we would like to estimate.

We define the point-process score-matching objective function to be

$$J(\theta) = \frac{1}{2}\left\langle \sum_{i=1}^{N} \left(\partial_{t_i} \log p^*(\mathcal{T}) - \partial_{t_i} \log p^\theta(\mathcal{T})\right)^2 \right\rangle_{\mathcal{T}\sim\mathcal{P}^*} \quad (1)$$

with the score-matching estimate of θ being the value at which this objective is minimised. This choice of objective may be related to the original score-matching objective of [3] in one of two ways. First, it can be seen as the difference in

This work was supported by the Gatsby Charitable Foundation and the Simons Foundation (SCGB 323228, MS).

978-1-5090-0746-2/16/$31.00 ©2016 IEEE

variational derivatives of the log-densities taken with respect to the counting function $N(t)$ defined on $[0, T]$ (see [1]) and subject to the constraints that it be non-decreasing and piecewise constant with unit steps. Alternatively, we can follow [4] and introduce hypothetical location parameters $\{\mu_1, \mu_2, \dots\}$ to define a translated point process based on $\mathcal{P}^*$ with density $p_{\boldsymbol{\mu}}(\{t_1 \dots t_N\}) = p^*(\{t_1 + \mu_1 \dots t_N + \mu_N\})$. Equation (1) then follows by matching Fisher score functions with respect to the parameters μ_i evaluated at $\mu_1 = \mu_2 = \dots = \mu_N = 0$, and with derivatives with respect to $\mu_{N+1}, \mu_{N+2}, \dots$ all set to 0.

As in the usual score matching development, this cost function cannot be minimised directly as it depends on derivatives of an unknown density p^*. The manipulations necessary to remove this dependence broadly follow the general derivation of [3], with some complications arising from different limits of integration. We first expand the square and drop the terms in $(\partial_{t_i} \log p^*(\mathcal{T}))^2$, as these do not depend on the parametric density and so do not affect the maximum with respect to the parameters. Examining the cross-terms, conditioned on the number of events N, we have:

$$\begin{aligned}
&\left\langle \partial_{t_i} \log p^*(\mathcal{T})\ \partial_{t_i} \log p^\theta(\mathcal{T}) \right\rangle_{\mathcal{P}^*|N} \\
&\quad = \int_{\mathbb{T}_N} d\mathcal{T}\ p^*(\mathcal{T}) \partial_{t_i} \log p^*(\mathcal{T})\ \partial_{t_i} \log p^\theta(\mathcal{T}) \\
&\quad = \int_{\mathbb{T}_N} d\mathcal{T}\ \partial_{t_i} p^*(\mathcal{T}) \partial_{t_i} \log p^\theta(\mathcal{T}) \\
&\quad = \int_{\mathbb{T}_N} d\mathcal{T}\ \Big(p^*(\mathcal{T}) \partial_{t_i} \log p^\theta(\mathcal{T}) \big(\delta(t_{i+1} - t_i) - \delta(t_i - t_{i-1})\big) \\
&\qquad\qquad - p^*(\mathcal{T}) \partial_{t_i}^2 \log p^\theta(\mathcal{T}) \Big)
\end{aligned}$$

where the final step required integration by parts, and we have used delta-function notation to evaluate the limits $t_i \in (t_{i-1}, t_{i+1})$, set by the order constraint on samples, for the complete portion of the integral. Now, provided that the parametric density satisfies the smoothness property

$$\partial_{t_i} \log p^\theta(\mathcal{T})|_{t_i = t_{i+1}} = \partial_{t_{i+1}} \log p^\theta(\mathcal{T})|_{t_{i+1} = t_i}\,,$$

the majority of delta function terms will cancel when summing over i, leaving

$$\begin{aligned}
&\left\langle \left\langle \sum_i \partial_{t_i} \log p^*(\mathcal{T})\ \partial_{t_i} \log p^\theta(\mathcal{T}) \right\rangle_{\mathcal{T} \sim \mathcal{P}^*|N} \right\rangle_{N \sim \mathcal{P}^*} \\
&\quad = \Big\langle \partial_{t_N} \log p^\theta(\mathcal{T}) \delta(t_N - T) - \partial_{t_1} \log p^\theta(\mathcal{T}) \delta(t_1) \\
&\qquad\qquad - \sum_i \partial_{t_i}^2 \log p^\theta(\mathcal{T}) \Big\rangle_{\mathcal{P}^*}
\end{aligned}$$

To construct the final empirical cost we recombine this term with the remaining model derivatives from (1) and replace expectations over $\mathcal{P}^*$ with evaluations at the observed event times. This will eliminate the remaining delta function values almost surely. Thus we arrive at the *empirical point process score matching* objective:

$$\widehat{J}(\theta) = \sum_i \frac{1}{2} \left(\partial_{t_i} \log p^\theta(\mathcal{T}) \right)^2 + \partial_{t_i}^2 \log p^\theta(\mathcal{T}) \qquad (2)$$

3. LOG-LINEAR POISSON-PROCESS REGRESSION

The point process score matching objective derived above applies quite generally to any parametric point process. In the remainder of this paper, we focus on a class of models in which the log-intensity function of a point process is taken to depend on an observed covariate function $\mathbf{x}(t)$. For example, the times of action potentials (or "spikes") generated by a sensory neuron may depend on a sensory stimulus being presented to the animal.

The simplest such model is conditionally Poisson, with a log-intensity function that is a linear function of $\mathbf{x}(t)$:

$$\log \lambda^\theta(t) = \boldsymbol{\theta}^\mathsf{T} \mathbf{x}(t)\,. \qquad (3)$$

This scheme resembles a generalised linear model (GLM) for a Poisson *count* observation, and indeed the point-process likelihood may be obtained as a limit of the Poisson-count GLM [2]. Thus, in practice, such models are often fit by discretising time, counting events that fall in each discrete bin, and using the iterative GLM framework. The score-matching estimator is far simpler.

Recalling that the log model density is given by $p^\theta(\mathcal{T}) = \sum_i \log \lambda^\theta(t_i) - \int dt\ \lambda^\theta(t) = \sum_i \boldsymbol{\theta}^\mathsf{T} \mathbf{x}(t_i) + \textit{constant}$, we have:

$$\widehat{J}(\theta) = \sum_i \frac{1}{2} \left(\boldsymbol{\theta}^\mathsf{T} \mathbf{x}'(t_i) \right)^2 + \boldsymbol{\theta}^\mathsf{T} \mathbf{x}''(t_i) \qquad (4)$$

where primes represent temporal derivatives. Solving for the minimum in $\boldsymbol{\theta}$ (and writing $\mathbf{x}_i'$ for $\mathbf{x}'(t_i)$ etc.) we obtain:

$$\hat{\boldsymbol{\theta}} = -\Big(\sum_i \mathbf{x}_i' \mathbf{x}_i'^\mathsf{T} \Big)^{-1} \sum_i \mathbf{x}_i'' \qquad (5)$$

This is a simple estimator that depends only on derivatives of the regression covariate evaluated at the times of events. It does not depend on the value of $\lambda^\theta(t)$ at other times, and thus its computational burden scales with the number of events rather than with total observation time T. As will be seen in the experiments below, it can compare favourably to maximum-likelihood estimators in accuracy, at a small fraction of the computational cost.

If the continuous-time covariate function $\mathbf{x}(t)$ and its derivatives are not known exactly, and instead must be sampled, potentially quantised and possibly corrupted by noise, then the process by which the smooth function is reconstructed from these samples will affect the quality of the

estimate in (5). The exact nature of any bias will depend on the properties of the estimates of $\mathbf{x}_i'$ and $\mathbf{x}_i''$. However, two general points are worth noting: First, the separate sums in the numerator and denominator of (5) will reduce the impact of correlation between estimates of the first and second derivatives; and second, while zero-mean perturbations in estimates of $\mathbf{x}''$ will average away in the numerator, noise in $\mathbf{x}'$ will contribute a positive-definite bias to the squared term in the denominator. In effect, such estimation noise contributes a term very similar to that encountered in ridge regression.

4. LNP REGRESSION

The log-linear assumption of (3) is common, but not always appropriate. More general linear-nonlinear-Poisson (LNP) models have attracted interest, particularly from the neuroscience community. These models assume an intensity function of the form $\lambda(t) = f(K\mathbf{x}(t))$, where K is a vector or matrix and $f()$ is an unknown nonlinear function mapping the column space of K to $\mathbb{R}_+$. Note that for arbitrary $f()$, only the row space of K is identifiable.

If the regressor covariate values $\mathbf{x}(t)$ are normally distributed, then spectral methods offer efficient—and generally unbiased and consistent—estimators for the row space of K [5, 6]. However, these approaches suffer from considerable bias when the distribution is non-normal. The alternative is to assume a basis of nonlinear functions $\phi_i()$, with $f()$ then estimated within the space spanned by this basis. This approach may be formulated in information-theoretic terms [7], although the resulting cost function is identical to that obtained by the conventional likelihood-based treatment [8].

Here, we parametrise the *log* intensity in a similar way. Let $\boldsymbol{\phi}()$ be a fixed vector-valued function mapping the column space of K to $\mathbb{R}^m$ (that is, it collects the outputs of the m basis nonlinear functions into a single vector-valued output). Then the model intensity is $\log \lambda^\theta(t) = \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\phi}(K\mathbf{x}(t))$. [We continue to use the generic symbols $\mathcal{P}^\theta$, p^θ and λ^θ for the model, even though the parameters now form a tuple $(\boldsymbol{\theta}, K)$.]

The log-likelihood for this model is

$$\log p^\theta(\mathcal{T}|K,\boldsymbol{\theta}) = \sum_i \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\phi}(K\mathbf{x}(t_i)) - \int dt\, \lambda(t)$$

Writing $\mathbf{x}_i = \mathbf{x}(t_i)$, $\boldsymbol{\phi}_i = \boldsymbol{\phi}(K\mathbf{x}(t_i))$, and similar forms for derivatives as above, we have:

$$\begin{aligned}\partial_{t_i}\boldsymbol{\phi}(K\mathbf{x}(t_i)) &= \nabla\boldsymbol{\phi}^\mathsf{T}(K\mathbf{x}(t_i))K\mathbf{x}'(t_i) = \nabla\boldsymbol{\phi}_i \cdot K\mathbf{x}_i' \\ \partial_{t_i}^2\boldsymbol{\phi}(K\mathbf{x}(t_i)) &= (\nabla\nabla\boldsymbol{\phi}\cdot K\mathbf{x}_i')\cdot K\mathbf{x}_i' + \nabla\boldsymbol{\phi}_i \cdot K\mathbf{x}_i''\end{aligned}$$

where $\nabla\nabla\boldsymbol{\phi}$ is a 3-tensor of the form $[\nabla\nabla\boldsymbol{\phi}(\mathbf{z})]_{ijk} = \frac{\partial^2\phi_i}{\partial z_j \partial z_k}$. And so the LNP score matching objective is

$$\widehat{J}(K,\boldsymbol{\theta}) = \Big\langle \boldsymbol{\theta}^\mathsf{T}(\nabla\nabla\boldsymbol{\phi}\cdot K\mathbf{x}_i')\cdot K\mathbf{x}_i' + \boldsymbol{\theta}^\mathsf{T}\nabla\boldsymbol{\phi}_i\cdot K\mathbf{x}_i'' + \frac{1}{2}\|\boldsymbol{\theta}^\mathsf{T}\nabla\boldsymbol{\phi}_i\cdot K\mathbf{x}_i'\|^2\Big\rangle$$

It is straightforward to see that by setting $K = I$ and $\boldsymbol{\phi}(\mathbf{z}) = \mathbf{z}$ (that is $\boldsymbol{\phi}(K\mathbf{x}) = \mathbf{x}$), we obtain $\nabla\boldsymbol{\phi} = 1$ and $\nabla\nabla\boldsymbol{\phi} = 0$, and so recover the log-linear model score matching cost function of (4).

A similar closed-form solution is also available for the "generalised quadratic model" (GQM) case, where $K = I$ and $\boldsymbol{\phi}(\mathbf{z}) = \mathrm{vec}(\mathbf{z}\mathbf{z}^\mathsf{T})$ (the 'vec' operator unrolls a matrix argument into a vector). In this case we have, writing $\mathbf{e}_i$ for the cartesian basis vector along coordinate i:

$$\begin{aligned}&\nabla_j\boldsymbol{\phi}(\mathbf{x}) = \mathrm{vec}(\mathbf{e}_j\mathbf{x}^\mathsf{T} + \mathbf{x}\mathbf{e}_j^\mathsf{T}) \\ \Rightarrow\quad &\nabla\boldsymbol{\phi}_i\cdot\mathbf{x}_i' = \sum_j \mathrm{vec}(\mathbf{e}_j\mathbf{x}^\mathsf{T} + \mathbf{x}\mathbf{e}_j^\mathsf{T})x_{ij}' = \mathrm{vec}(\mathbf{x}'\mathbf{x}^\mathsf{T} + \mathbf{x}\mathbf{x}'^\mathsf{T}) \\ &\text{and}\quad \nabla\boldsymbol{\phi}_i\cdot\mathbf{x}_i'' = \mathrm{vec}(\mathbf{x}''\mathbf{x}^\mathsf{T} + \mathbf{x}\mathbf{x}''^\mathsf{T}) \\ &\text{also}\quad \nabla_j\nabla_k\boldsymbol{\phi}(\mathbf{x}) = \mathrm{vec}(\mathbf{e}_j\mathbf{e}_k^\mathsf{T} + \mathbf{e}_k\mathbf{e}_j^\mathsf{T}) \\ \Rightarrow\quad &(\nabla\nabla\boldsymbol{\phi}_i\cdot\mathbf{x}_i')\cdot\mathbf{x}_i' = 2\,\mathrm{vec}(\mathbf{x}_i'\mathbf{x}_i'^\mathsf{T})\end{aligned}$$

Collecting these expressions together we find the closed form estimator:

$$\hat{\boldsymbol{\theta}}_{GQM} = -\left(\sum_i \mathrm{vec}(\mathbf{x}_i'\mathbf{x}_i^\mathsf{T} + \mathbf{x}_i\mathbf{x}_i'^\mathsf{T})\mathrm{vec}(\mathbf{x}_i'\mathbf{x}_i^\mathsf{T} + \mathbf{x}_i\mathbf{x}_i'^\mathsf{T})^\mathsf{T}\right)^{-1} \sum_i \left(2\,\mathrm{vec}(\mathbf{x}_i'\mathbf{x}_i'^\mathsf{T}) + \mathrm{vec}(\mathbf{x}_i''\mathbf{x}_i^\mathsf{T} + \mathbf{x}_i\mathbf{x}_i''^\mathsf{T})\right)$$

"Maximum Expected Likelihood" methods have also been proposed for GQM estimation [9], but like other spectral methods depend on a known and tractable distribution of regression inputs. The score-matching estimator is efficient, and free of such assumptions — a point we investigate in experiments below.

5. EXPERIMENTS

We investigated the properties of the proposed score-matching-based estimators in numerical experiments, in which we fit model parameters to simulated data where the true parameters θ and K were known. Two aspects of the were of particular interest: the consistency and the computational costs of the estimators. We also compared these estimators to maximum likelihood (ML) estimates obtained by iterative procedures.

5.1. Generalised Linear Model

We began by evaluating the simple log-linear Poisson model estimator (5). Responses were generated according to (3). A 10-dimensional covariate function $\mathbf{x}(t)$ was obtained by filtering Gaussian white noise sampled at 1000 Hz with 10 Gammatone filters. A "true" weight vector θ was chosen randomly. Event times were generated by an inhomogeneous Poisson process with log-intensity given by the weighted filter outputs, offset to achieve a total event event rate of 10, 20 or 40 Hz. The agreement between recovered and true model

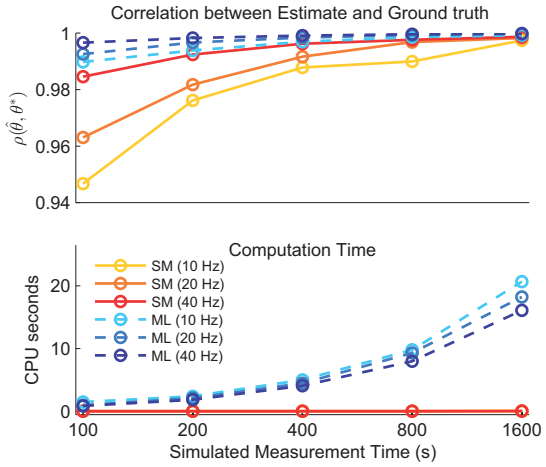

Fig. 1. Comparison between score matching (SM) and maximum likelihood (ML) estimation for the generalised linear Poisson model. SM achieves high correlation to ground truth (note the y-axis on the upper figure) and approaches the performance of ML with considerably lower computational cost.

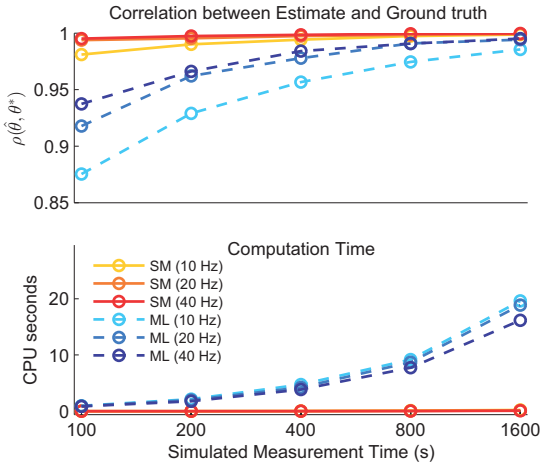

Fig. 2. Comparison between SM and binned ML estimation on the GLM. Binning high-frequency input adversely affects ML estimates. The continuous time SM estimate does not require any binning.

parameters was quantified using the Pearson correlation coefficient ρ.

As expected, the ML estimator performed slightly better than the score-matching estimator for small data sizes, particularly for low event rates (Fig. 1). However, correlation coefficients in both cases were high. With increasing data size both estimators converged to the true solution. The computational cost of the closed-form score matching estimator was over two orders of magnitude lower than the cost of the iterative ML estimator, which increases considerably with data size.

ML parameter estimation in practice is often based on binning events at a pre-determined timescale to make estimation computationally feasible. If the true intensity function varies more rapidly than the bin-width, such discretisation may mask the true underlying function. We constructed simulated data as before, but generated the input covariate and events at a sample rate of 5000 Hz, which we binned to 1000 Hz for ML estimation. In our simulations, this mismatch in bin width had a detrimental effect on the fitted parameters (Fig. 2). An advantage of the score matching estimator is that it is evaluated only at the exact event times and does not require any binning. In neural experiments, $\mathbf{x}(t)$ is often an experimental stimulus which varies at a fixed rate (for example, the frame rate of a monitor). Nonetheless, the physiological response may be more rapid — for example, with the neurons spiking immediately after a frame refresh rather than uniformly throughout the frame presentation time. Thus, an appropriate choice of bin width may often be unclear.

5.2. Generalised Quadratic Model

We then examined the performance of our score-matching estimator for the log-quadratic (GQM) case, where the log-intensity depended linearly on the outer product of the covariate function – $\mathbf{x}(t)\mathbf{x}(t)^\mathsf{T}$. Linear weights θ, now forming a matrix, were chosen randomly; event times simulated; and the resulting estimates of θ evaluated. Maximum likelihood estimation for the GQM is partiularly compuationally burdensome and thus not extensively employed. Instead, we compared the score-matching estimator to another closed-form solution that maximises the expected likelihood (MEL) of GQM [9].

While MEL methods provide a fast and robust way to fit GQM parameters, they rely on the assumption that expected likelihoods can be computed for the given input distribution. If this is the case, as for Gaussian inputs, the MEL estimator's performance exceeds that of the score matching estimator with only a small increase in computational cost (Fig. 3). However for covariates that follow a more natural non-Gaussian distribution, the MEL estimator is not appropriate. When events are driven by a covariate based on filtered speech waveforms, with the same Gammatone profiles as before, the MEL estimator collapses. The score-matching derivation does not depend on the stimulus distribution, and so is not affected by the change in statistics (Fig. 4). In fact, the estimated parameters are slightly more accurate than with the Gaussian noise covariate, probably because the speech-driven model generates events with greater temporal precision.

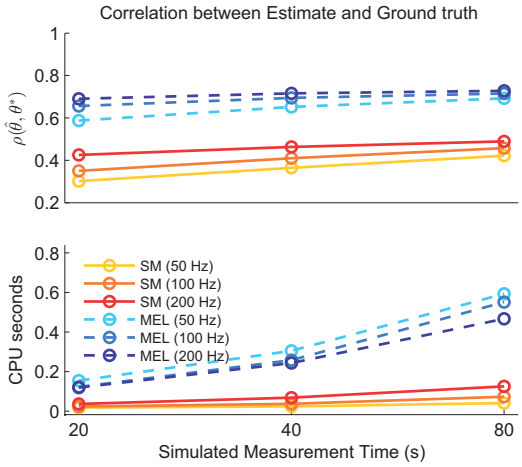

Fig. 3. Comparison between SM and maximum expected likelihood (MEL) estimation for a generalised quadratic model using Gaussian inputs. MEL offers clear advantages when expected likelihoods are computable with only modest computation burden.

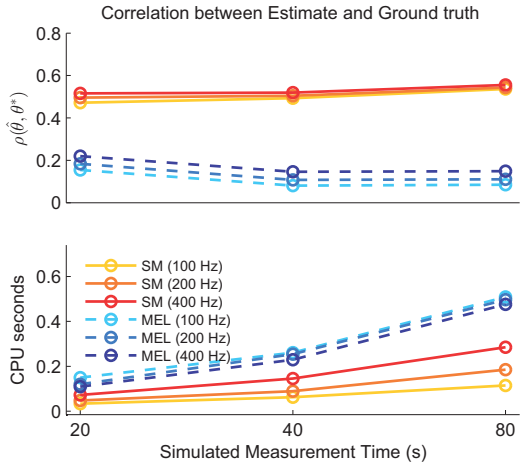

Fig. 4. Comparison between SM and MEL estimation for a generalised quadratic model using a speech signal input. Incorrectly computing expectations based on second order input statistics defeats the MEL approach as expected. SM remains robust (and, indeed, performs slightly better than with Gaussian noise).

6. CONCLUSION

We have introduced a new class of estimators for point-process models based on measurements in continuous time, and derived specific estimators for "regression" settings where the point-process intensity depends on a known external covariate. The estimators are frequently closed-form, and computation scales with the number of events rather than the total interval length. This approach to estimation may help to expand the range of point-process models that can be tractably fit to measured data.

7. REFERENCES

[1] Donald L. Snyder and Michael I. Miller, *Random Point Processes in Time and Space*, Springer Verlag, New York, second edition, 1991.

[2] Mark Berman and T Rolf Turner, "Approximating point process likelihoods with GLIM," *Appl. Stats.*, pp. 31–38, 1992.

[3] Aapo Hyvärinen, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, pp. 695–709, 2005.

[4] Aapo Hyvärinen, "Some extensions of score matching," *Comput. Stats. & Data Anal.*, vol. 51, no. 5, pp. 2499–2512, 2007.

[5] Odelia Schwartz, Jonathan W. Pillow, Nicole C. Rust, and Eero P. Simoncelli, "Spike-triggered neural characterization," *J. Vis.*, vol. 6, no. 4, pp. 484–507, 2006.

[6] Jonathan W. Pillow and Eero P. Simoncelli, "Dimensionality reduction in neural models: An information-theoretic generalization of spike-triggered average and covariance analysis," *J. Vis.*, vol. 6, no. 4, pp. 414–428, 2006.

[7] Tatyana Sharpee, Nicole C. Rust, and William Bialek, "Analyzing neural responses to natural signals: maximally informative dimensions.," *Neural Comput.*, vol. 16, no. 2, pp. 223–250, 2004.

[8] Ross S. Williamson, Maneesh Sahani, and Jonathan W. Pillow, "The equivalence of information-theoretic and likelihood-based methods for neural dimensionality reduction," *PLoS Comput. Biol.*, vol. 11, no. 4, pp. e1004141, 2015.

[9] Il Memming Park, Evan Archer, Nicholas J. Priebe, and Jonathan W. Pillow, "Spectral methods for neural characterization using generalized quadratic models.," in *Advances in Neural Information Processing Systems*, vol. 26, pp. 2454–2462. 2013.