Using Score-based Methods for Unnormalisable Probability Density Estimation

Truncated Density Estimation and Parameter Derivative Estimation

Daniel Williams
University of Bristol
12th July 2024

Contents

A brief literature review will accompany each project

(Brief) Introduction

Unnormalisable Models

The data PDF is given by $q(\boldsymbol{x})$, which we only access through samples $\{\boldsymbol{x}_i\} \sim q(\boldsymbol{x})$. We seek to model $q(\boldsymbol{x})$ with $p(\boldsymbol{x};\boldsymbol{\theta})$.

General form of a model PDF

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\bar{p}(\boldsymbol{x};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}, \qquad Z(\boldsymbol{\theta}) = \int_V \bar{p}(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}, \tag{1}$$

- We **do not have access** to $Z(\boldsymbol{\theta})$
- This appears in applications such as
  - Energy based modelling (EBM)
  - **Truncated Density Estimation**

(Brief) Introduction

Truncated Distributions

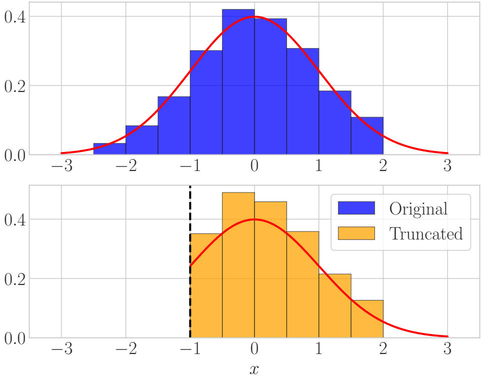


Figure 1. Example of a Normal distribution truncated at $x = -1$.

Chapter 3: **Estimating Density Models with Truncation Boundaries using Score Matching**

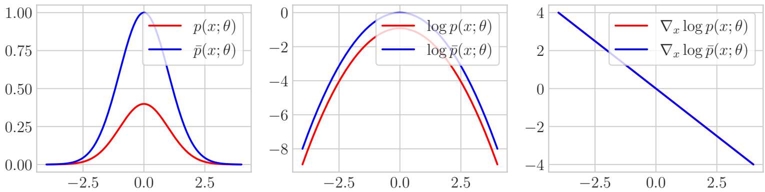
Publication

Liu, S., T. Kanamori, and D. J. Williams (2022). Estimating density models with truncation boundaries using score matching. Journal of Machine Learning Research 23(186), 1–38.

Note on contribution. This chapter is based on the paper authored by Song Liu, Takafumi Kanamori and myself. My primary contribution was in the development of the algorithm and the experiments.

Literature Review: Score Matching

Overview



The score function for $p(\boldsymbol{x};\boldsymbol{\theta})$ is given by

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x};\boldsymbol{\theta}) = \nabla_{\boldsymbol{x}} \log \bar{p}(\boldsymbol{x};\boldsymbol{\theta}) - \nabla_{\boldsymbol{x}} \log Z(\boldsymbol{\theta}) = \nabla_{\boldsymbol{x}} \log \bar{p}(\boldsymbol{x};\boldsymbol{\theta}).$$

Therefore computing the score function *does not require computing the normalising constant.*

Literature Review: Score Matching

Objective Functions

Score Matching (Hyvärinen, 2005)

Objective function

$$\mathcal{J}_{\mathrm{SM}}(\boldsymbol{\theta}) := \mathbb{E}_q \left[\left\| \boldsymbol{\psi}_q(\boldsymbol{x}) - \boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}) \right\|^2 \right], \tag{2}$$

where $\boldsymbol{\psi}_q(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x})$ and $\boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}; \boldsymbol{\theta})$.

Generalised Score Matching (Yu et al., 2018, 2019)

Objective function

$$\mathcal{J}_{\mathrm{GSM}}(\boldsymbol{\theta}) := \mathbb{E}_q \left[\left\| \boldsymbol{h}(\boldsymbol{x})^{1/2} \odot \left(\boldsymbol{\psi}_q(\boldsymbol{x}) - \boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}) \right) \right\|^2 \right]. \tag{3}$$

where $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Literature Review: Score Matching

Tractability

$\mathcal{J}_{\mathrm{SM}}(\boldsymbol{\theta})$ can be written as

$$\mathcal{J}_{\mathrm{SM}}(\boldsymbol{\theta}) = \mathbb{E}_q \left[\boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x})^\top \boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}) + 2\mathrm{tr}\left(\nabla_{\boldsymbol{x}} \boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}) \right) \right] + C_q$$

where $C_q = \int_{\mathbb{R}^d} q(\boldsymbol{x}) \boldsymbol{\psi}_q(\boldsymbol{x})^\top \boldsymbol{\psi}_q(\boldsymbol{x}) d\boldsymbol{x}$. This relies on the boundary condition:

Score Matching Boundary Condition

$$\lim_{\|\boldsymbol{x}\| \to \infty} q(\boldsymbol{x}) = 0.$$

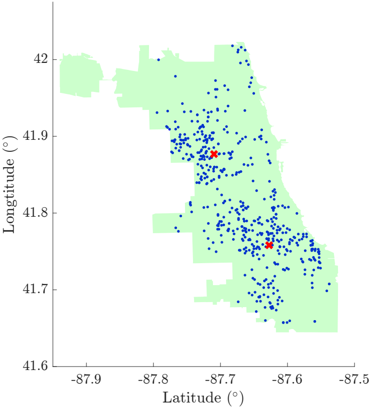
$\mathcal{J}_{\mathrm{GSM}}(\boldsymbol{\theta})$ can be written in a similar tractable way. This relies on the boundary condition

Generalised Score Matching Boundary Condition

$$\lim_{\substack{x_j \to a_j(\boldsymbol{x}_{-j})^+ \\ x_j \to b_j(\boldsymbol{x}_{-j})^-}} h_j(\boldsymbol{x}) q(\boldsymbol{x}) = 0, \forall j = 1, \ldots, d,$$

where $\boldsymbol{a}(\boldsymbol{x}_{-j})$ and $\boldsymbol{b}(\boldsymbol{x}_{-j})$ are the lower and upper endpoints of the domain, respectively, and $x_j \to 0^+$ denotes where x_j approaches zero from above, and $x_j \to \infty^-$ denotes where x_j approaches infinity from below

Truncated Domain



Homicides in Chicago.

- Neither boundary condition is satisfied when the support of $q(\boldsymbol{x})$ is a **truncated domain**.
- Truncated data can come in many forms - e.g. geographical applications

Truncated Score Matching

Methodology Overview

Let V be a truncated domain with boundary ∂V and the support of $q(\boldsymbol{x})$ is V. We propose

Truncated Score Matching Objective

$$\mathcal{J}_{\mathrm{TSM}}(\boldsymbol{\theta}) := \mathbb{E}_q \left[\left\| \boldsymbol{g}(\boldsymbol{x})^{1/2} \odot \left(\boldsymbol{\psi}_q(\boldsymbol{x}) - \boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}) \right) \right\|^2 \right], \tag{4}$$

where $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ and $\boldsymbol{g}(\boldsymbol{x}') = \boldsymbol{0}, \forall \boldsymbol{x}' \in \partial V$.

For which the following boundary condition is satisfied

Truncated Score Matching Boundary Condition

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}'} g_j(\boldsymbol{x}) q(\boldsymbol{x}) = 0, \ \forall j = 1, \ldots, d, \tag{5}$$

where $\boldsymbol{x}' \in V$ is denoted as a boundary point, and $\boldsymbol{x} \to \boldsymbol{x}'$ takes any point sequence converging to $\boldsymbol{x}'$ into account.

Truncated Score Matching Objective

$\mathcal{J}_{\text{TSM}}(\boldsymbol{\theta})$ can be written as

$$\mathcal{J}_{\text{TSM}}(\boldsymbol{\theta}) = \sum_{j=1}^{d} \mathbb{E}_q \Big[\mathrm{g}_j(\boldsymbol{x}) \psi_{p_{\boldsymbol{\theta}},j}(\boldsymbol{x})^2 + 2\mathrm{g}_j(\boldsymbol{x}) \partial_{x_j} \psi_{p_{\boldsymbol{\theta}},j}(\boldsymbol{x}) + 2\partial_{x_j} \mathrm{g}_j(\boldsymbol{x}) \psi_{p_{\boldsymbol{\theta}},j}(\boldsymbol{x}) \Big] + C_q \tag{6}$$

provided that $\mathbf{g}$ is *weakly differentiable* and V is a *Lipschitz domain.*

Choice of Weighting Function

$\mathcal{J}_{\text{TSM}}(\boldsymbol{\theta})$ depends on $\mathbf{g}$, a *weighting function*. We choose $\mathbf{g}$ as a *distance function*, given by

$$\mathbf{g} = \begin{bmatrix} g_0 \\ \vdots \\ g_0 \end{bmatrix}, \quad g_0(\boldsymbol{x}) = \min_{\boldsymbol{x}' \in \partial V} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{x}').$$

This naturally fulfills the boundary criteria ($g(\boldsymbol{x}') = 0, \forall \boldsymbol{x}' \in \partial V$).

Additionally, g_0 is justified via solving a Stein discrepancy estimator (Barp et al., 2019) and obtaining the truncated score matching objective with g_0.

Some Key Results

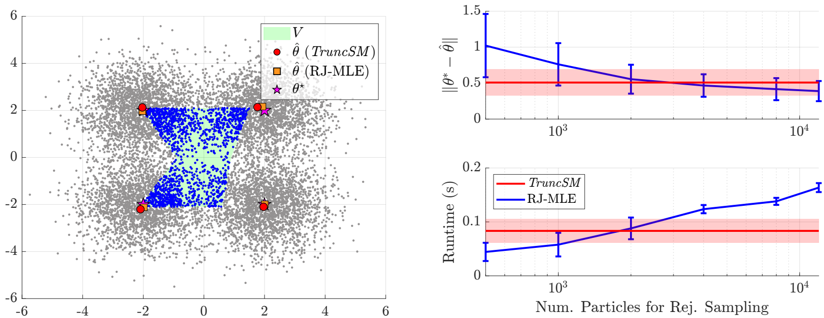


Figure 2. Mixed multivariate Normal example, with performance comparisons to rejection sampling MLE (RJ-MLE).

Some Key Results

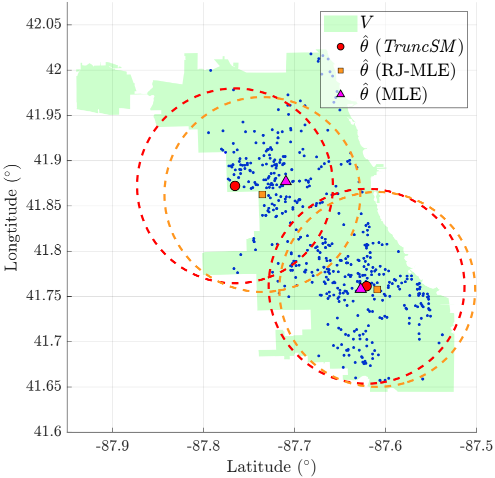


Figure 3. Estimating a mixed multivariate Normal distribution for homicide locations in the city of Chicago.

Some Key Results

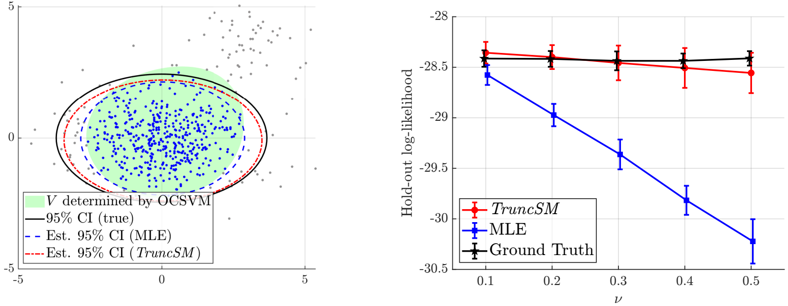


Figure 4. Outlier over-trimming compensation using *TruncSM*, in comparison with a naive MLE approach and the ground truth.

Some Key Results

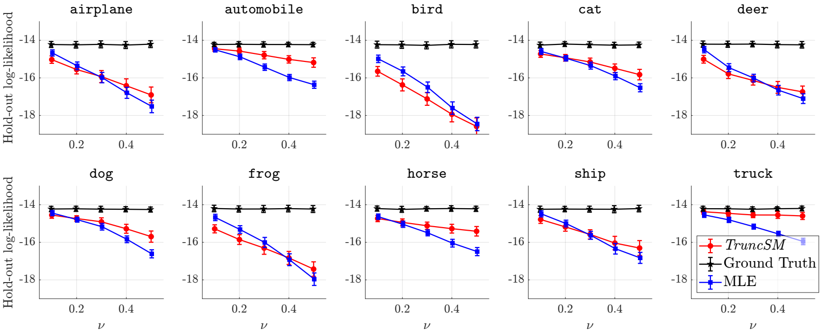


Figure 5. Validation log-likelihood as ν increases across the 10 image classes in the CIFAR-10 dataset.

Chapter 4: **Score Matching for Truncated Density Estimation on a Manifold**

Publication

Williams, D. J. and S. Liu (2022b, 25 Feb–22 Jul). Score matching for truncated density estimation on a manifold. In Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022, Volume 196 of Proceedings of Machine Learning Research, pp.312–321. PMLR.

Literature Review: Manifolds/Spheres

Let M be a Riemannian manifold, and M_C be a *compact* Riemannian manifold, that is *without boundary* or edges.

We focus on the sphere,

$$\mathbb{S}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_2 = 1\},$$

specifically the 2D sphere, $\mathbb{S}^2 \subset \mathbb{R}^3$, and two spherical distributions:

- Kent distribution
- von Mises-Fisher Distribution

both analogous to the multivariate Normal distribution.

Literature Review: Manifold Score Matching

Mardia et al. (2016) developed *manifold score matching* for M_C, whose objective is given by

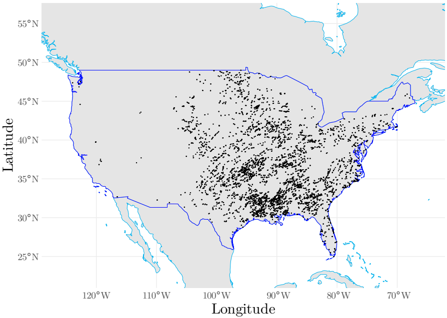
Manifold Score Matching

$$\mathcal{J}_{\text{MSM}}(\boldsymbol{\theta}) := \int_{M_C} q(\boldsymbol{z}) \left\| \boldsymbol{\psi}_q(\boldsymbol{z}) - \boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{z}) \right\|^2 d\boldsymbol{z}, \tag{7}$$

where $\boldsymbol{z} \in M_C$.

This objective can also be made tractable via the lack of boundary terms in *Stokes' theorem* as M_C is without boundary.

Truncated Manifold



Storms in the U.S.

- When the manifold does have a boundary (e.g. upper hemisphere), then manifold score matching is invalid.
- (Euclidean) truncated score matching will be bias when the manifold varies significantly from a Euclidean setting, e.g. the curvature of the Earth is more pronounced over a larger surface.

Truncated Manifold Score Matching

Let M_T be a Riemannian manifold *with boundary*, which is denoted as ∂M_T.

We propose *truncated manifold score matching* (TMSM), which follows the same process as truncated score matching, but defined on M_T.

TMSM objective

$$\mathcal{J}_{\text{TMSM}}(\boldsymbol{\theta}) := \int_{M_T} q(\boldsymbol{z}) \mathrm{g}(\boldsymbol{z}) \left\| \boldsymbol{\psi}_q(\boldsymbol{z}) - \boldsymbol{\psi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{z}) \right\|^2 d\boldsymbol{z}, \tag{8}$$

where $\boldsymbol{z} \in M_T$, $\mathrm{g}(\boldsymbol{z}') = 0, \forall \boldsymbol{z}' \in \partial M_T$.

With the boundary condition on g, Stokes' theorem allows the TMSM objective to be written tractably.

Truncated Manifold Score Matching

Following truncated score matching, g is a distance function. We propose two distance functions defined on the sphere $\mathbb{S}^2$.

- **Haversine distance**, the geodesic distance between two points along the surface of the sphere,
- **Projected Euclidean distance**, Euclidean distance calculated between two points projected from the sphere to a two dimensional plane.

Some Key Results

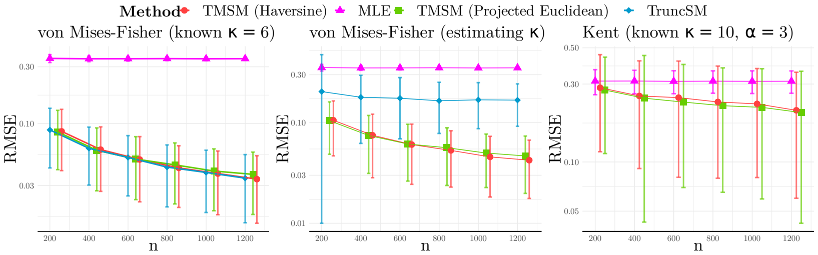


Figure 6. Mean estimation error and standard deviation against sample size n, on a log scale, from 512 trials for each method.

Some Key Results

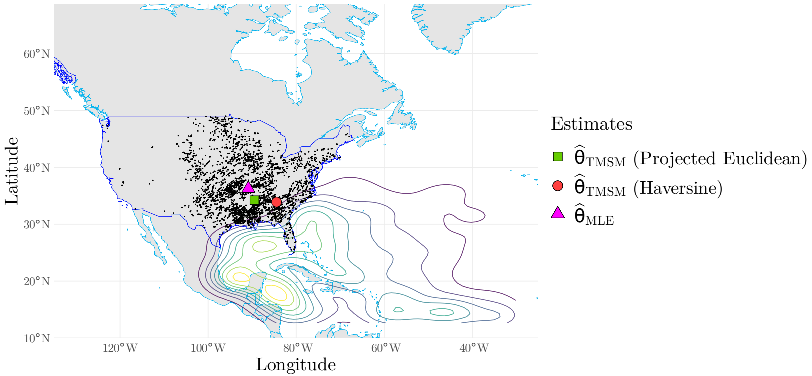


Figure 7. Observed storms as recorded by the U.S. from 2017 to 2020, with a kernel density estimate contour of storm origin locations over the Atlantic ocean.

Chapter 5: **Approximate Stein Classes for Truncated Density Estimation**

Publication

Williams, D. J. and S. Liu (2023, 23–29 Jul). Approximate stein classes for truncated density estimation. In Proceedings of the 40th International Conference on Machine Learning, Volume 202 of Proceedings of Machine Learning Research, pp. 37066–37090. PMLR.

Literature Review: Stein Discrepancies

Stein discrepancies are based around the Stein identity, given below.

Stein Identity (Gorham and Mackey, 2015)

Let $q(\boldsymbol{x})$ be any smooth probability density supported on $\mathbb{R}^d$ and let $\mathcal{S}_q : \mathcal{F}^d \to \mathbb{R}$ be a map. $\mathcal{F}^d$ is a Stein class of functions, if for any $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d, \boldsymbol{f} \in \mathcal{F}^d$,

$$\mathbb{E}_q \left[\mathcal{S}_q \boldsymbol{f}(\boldsymbol{x}) \right] = 0, \tag{9}$$

Literature Review: Stein Discrepancies

One often uses $\mathcal{T}_q$, the Langevin-Stein operator, given by

$$\mathcal{T}_q := \sum_{j=1}^{d} \psi_{q,j}(\boldsymbol{x}) f_j(\boldsymbol{x}) + \partial_{x_j} f_j(\boldsymbol{x}),$$

for which the Stein identity holds under the boundary condition

Stein Boundary Condition

$$\lim_{\|\boldsymbol{x}\| \to \infty} q(\boldsymbol{x}) = 0.$$

Note: this is the same boundary condition as regular score matching.

Literature Review: Stein Discrepancies

The Stein identity can be used to construct a Stein discrepancy via

$$\mathcal{J}_{\mathrm{SD}}(\boldsymbol{\theta}) := \sup_{\boldsymbol{f} \in \mathcal{F}^d} \mathbb{E}_q[\mathcal{T}_{p_{\boldsymbol{\theta}}} \boldsymbol{f}(\boldsymbol{x})], \tag{10}$$

which can be interpreted as the *maximum violation of the Stein identity* between $q(\boldsymbol{x})$ and $p(\boldsymbol{x}; \boldsymbol{\theta})$.

When $\mathcal{F}^d = \mathcal{G}^d$, an RKHS, we obtain the kernelised stein discrepancy (KSD), which analytically solves the supremum and is closed form.

KSD (Chwialkowski et al., 2016; Liu et al., 2016)

$$\mathcal{J}_{\mathrm{KSD}}(\boldsymbol{\theta}) := \|\mathbb{E}_q[\mathcal{T}_{p_{\boldsymbol{\theta}}} k(\boldsymbol{x}, \cdot)]\|_{\mathcal{G}^d} \tag{11}$$

Approximate Stein Classes

We propose *Approximate Stein Classes*, for which the Stein identity holds in an *approximate* sense, i.e.

Approximate Stein Class

$\widetilde{\mathcal{F}}_m^d$ is an *approximate Stein class* of $q(\boldsymbol{x})$ if for all $\tilde{\boldsymbol{f}} : \mathbb{R}^d \to \mathbb{R}^d$, $\tilde{\boldsymbol{f}} \in \widetilde{\mathcal{F}}_m^d$,

$$\mathbb{E}_q[\mathcal{S}_{q,m}\tilde{\boldsymbol{f}}(\boldsymbol{x})] = O_P(\varepsilon_m), \tag{12}$$

where ε_m is a monotonically decreasing function of m and $\mathcal{S}_{q,m}$ depends on m. $O_P(\varepsilon_m)$ denotes a sequence indexed by m which is bounded in probability by ε_m.

(12) is the approximate Stein identity. We show as an example that the method introduced by Kanagawa et al. (2019) falls under this framework.

KSD for Truncated Density Estimation

- When the support of $q(\boldsymbol{x})$ is V, a truncated domain with boundary ∂V, the Stein identity no longer holds as the boundary condition does not hold.
- Conditions on $\mathcal{F}^d$, the Stein class, can be imposed such that the boundary condition holds, similar to the last projects.
- **However**, additionally assume we can *only* access ∂V through samples given by
$$\{\boldsymbol{x}'_i\}_{i=1}^m \sim \partial V.$$
We cannot just design a weighting function (e.g. Xu (2022)), as we don't have access to the true ∂V.

KSD for Truncated Density Estimation

Aims

- Derive a truncated density estimator that does not depend on a pre-defined distance function
  - More flexible
  - Applicable to only *samples* on boundary
- Define an *approximate Stein class* for truncated densities.
  - Hence a truncated Stein discrepancy

First, we explore the relationship between ∂V and $\widetilde{\partial V}$, via

$$\mathcal{G}_0^d = \{\mathbf{g} \in \mathcal{G}^d \mid \mathbf{g}(\boldsymbol{x}') = \mathbf{0} \, \forall \boldsymbol{x}' \in \partial V, \|\mathbf{g}\|^2_{\mathcal{G}^d} \leq 1\}, \tag{13}$$

and

$$\mathcal{G}_{0,m}^d = \{\tilde{\mathbf{g}} \in \mathcal{G}^d \mid \mathbf{g}(\boldsymbol{x}') = \mathbf{0} \, \forall \boldsymbol{x}' \in \widetilde{\partial V}, \|\mathbf{g}\|^2_{\mathcal{G}^d} \leq 1\}. \tag{14}$$

KSD for Truncated Density Estimation

- **Lemma 5.3** shows the difference between $\tilde{\mathbf{g}}(\boldsymbol{x}')$ and $\mathbf{g}(\boldsymbol{x})$ decreases with m, under the assumption that $\widetilde{\partial V}$ is ε_m-dense in ∂V.
- **Proposition 5.4** shows the denseness assumption holds with high probability with a lower bound on ε_m

Combined, we have

Theorem 5.6 (Approximate Stein Class for Truncated Densities)

Let $q(\boldsymbol{x})$ be a smooth density supported on V. For any $\tilde{\mathbf{g}} \in \mathcal{G}^d_{0,m}$, then

$$\mathbb{E}_q[\mathcal{T}_q \tilde{\mathbf{g}}(\boldsymbol{x})] = O_P(\varepsilon_m). \tag{15}$$

As m (number of samples of ∂V) increases, the Stein identity becomes more ‘accurate’.

Truncated Kernelised Stein Discrepancy (TKSD)

Using Theorem 5.6, we construct a Truncated Kernelised Stein Discrepancy (TKSD), given by

$$\mathcal{J}_{\mathrm{TKSD}}(\boldsymbol{\theta})^2 := \sup_{\mathbf{g} \in \mathcal{G}^d_{0,m}} \mathbb{E}_q[\mathcal{T}_{p_{\boldsymbol{\theta}}} \mathbf{g}(\boldsymbol{x})]. \tag{16}$$

and show it can be written as

Theorem 5.7 (Closed Form TKSD)

$$\mathcal{J}_{\mathrm{TKSD}}(\boldsymbol{\theta}) = \sum_{l=1}^{d} \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})} \mathbb{E}_{\boldsymbol{y}\sim q(\boldsymbol{x})} \left[u_l(\boldsymbol{x}, \boldsymbol{y}) - \mathbf{v}_l(\boldsymbol{x})^\top (\mathbf{K}')^{-1} \mathbf{v}_l(\boldsymbol{y}) \right] \tag{17}$$

where $u_l(\boldsymbol{x}, \boldsymbol{y})$ is the same quantity inside the sum as KSD, $\mathbf{v}_l(\boldsymbol{z}) = \psi_{p,l}(\boldsymbol{z}) \boldsymbol{\phi}^\top_{\boldsymbol{z},\mathbf{x}'} + (\partial_{z_l} \boldsymbol{\phi}_{\boldsymbol{z},\mathbf{x}'})^\top$, $\boldsymbol{\phi}_{\boldsymbol{z},\mathbf{x}'} = [k(\boldsymbol{z}, \boldsymbol{x}'_1), \ldots, k(\boldsymbol{z}, \boldsymbol{x}'_m)]$ is a vector of kernel evaluations on boundary points, $\boldsymbol{\phi}_{\mathbf{x}'} = [k(\boldsymbol{x}'_1, \cdot), \ldots, k(\boldsymbol{x}'_m, \cdot)]^\top$ and $\mathbf{K}' = \boldsymbol{\phi}_{\mathbf{x}'} \boldsymbol{\phi}^\top_{\mathbf{x}'}$.

Consistency Analysis

We study $\hat{L}(\boldsymbol{\theta}) := \lim_{m\to\infty} \hat{\mathcal{J}}_{\text{TKSD}}(\boldsymbol{\theta})^2$ and under some assumptions show consistency with the following theorem.

Theorem 5.11 (TKSD Consistency)

Suppose there exists a unique $\boldsymbol{\theta}^\star$ such that $q = p_{\boldsymbol{\theta}^\star}$, Assumption 5.9 and Assumption 5.10 hold, then

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

where

$$\hat{\boldsymbol{\theta}}_n := \arg\min_{\boldsymbol{\theta}} \hat{L}(\boldsymbol{\theta}).$$

Some Key Results

Empirical Consistency

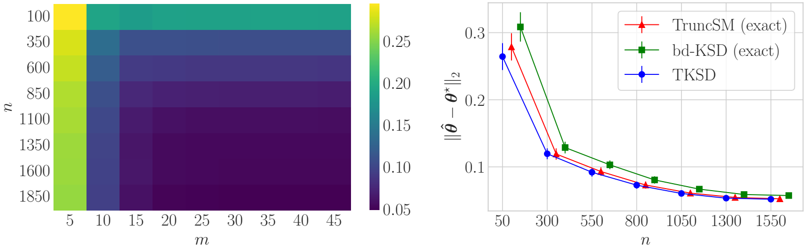


Figure 8. Estimation error as either m or n increase, fixed $d = 2$, estimating the mean of a truncated multivariate Normal.

Some Key Results

Boundary Distribution Effect

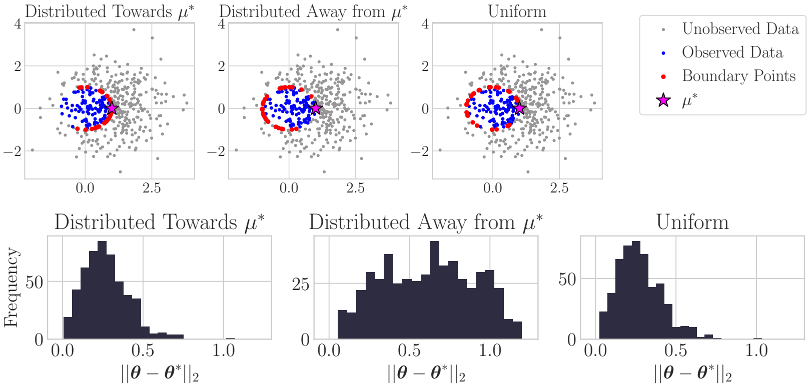


Figure 9. Quantification of effect of distribution of boundary samples, for three types of skewed boundary samples: towards the centre, away from the centre and uniformly distributed.

Some Key Results

Density Truncated by the U.S. border

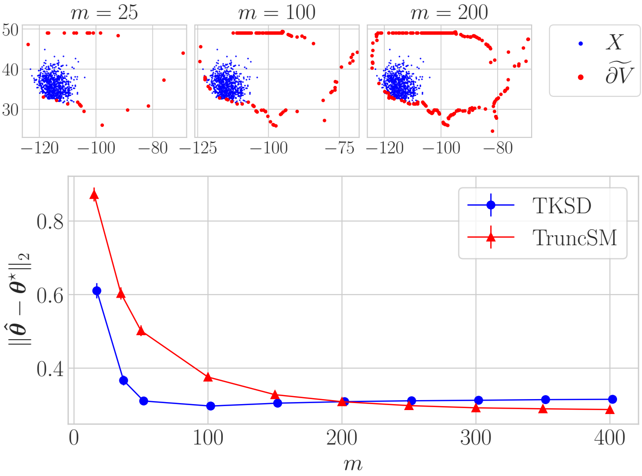


Figure 10. Density estimation when the truncation boundary is the border of the U.S.

Some Key Results

Mixed Multivariate Normal Comparison

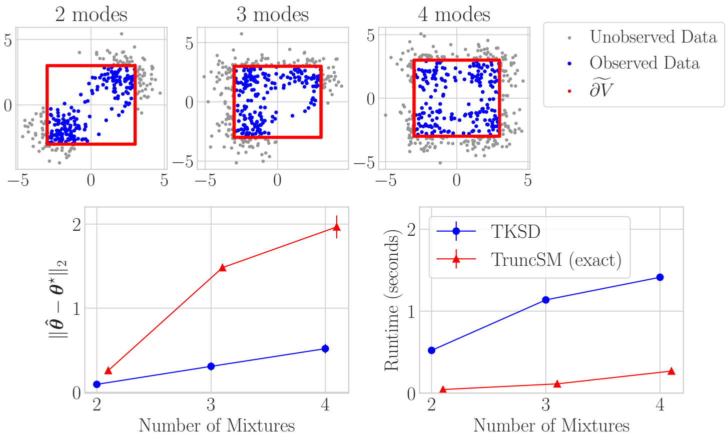


Figure 11. Estimation of Gaussian mixture modes using TKSD and *TruncSM*.

Some Key Results

Regression

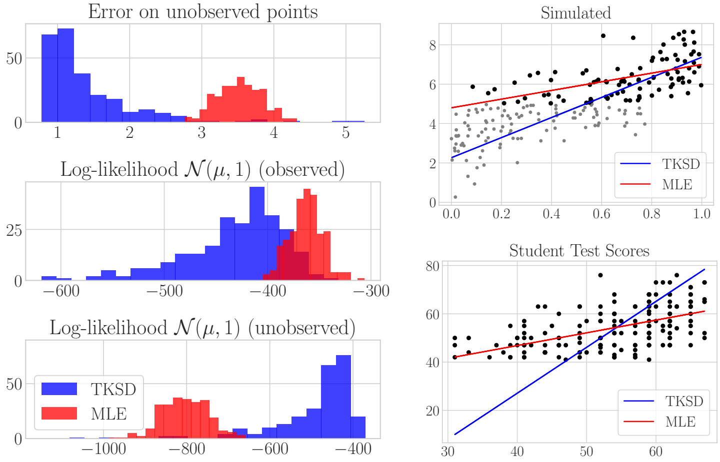


Figure 12. Estimation of regression parameters using TKSD.

Chapter 6: **Time Score Matching for Parameter Change Estimation and Changepoint Detection**

As yet unpublished.

Literature Review: Changepoint Detection

Classical changepoint detection:

$$x_t = \mu_t + \epsilon_t, \qquad \mu_t = \begin{cases} \mu_1, & \text{if } 0 < t < \tau \\ \mu_2, & \text{if } \tau \leq t \leq t_{\text{end}} \end{cases}$$

- Assumes abrupt shift in parameter at a given time τ
- Popular methods are moving window methods e.g. MOSUM (Eichinger and Kirch, 2018)
- For more complex tasks it is common to extend existing methods for specific circumstances
  - E.g. piecewise linear change MOSUM (Kim et al., 2024)

Literature Review: Time Score Matching

Let $q_t(\boldsymbol{x})$ be a density function that changes over time $t \in [0, 1]$. The time score matching (Choi et al., 2022) objective is given by

$$\mathcal{J}_{\text{time}}(\boldsymbol{\theta}) = \mathbb{E}_{q(t)}\mathbb{E}_{q_t(\boldsymbol{x})}\left[h(t)\|\dot{\psi}_{q_t}(\boldsymbol{x}) - s_0(\boldsymbol{x}, t)\|^2\right], \tag{18}$$

where $\dot{\psi}_{q_t}(\boldsymbol{x}) = \partial_t \log q_t(\boldsymbol{x})$ is the *time score* function, $s_0(\boldsymbol{x}, t)$ is a model of $\dot{\psi}_{q_t}(\boldsymbol{x})$, $h(t)$ is a weighting function and $q(t)$ is a uniform distribution over timescales.

Boundary conditions enable $\mathcal{J}_{\text{time}}(\boldsymbol{\theta})$ to be written tractably.

Estimating Parameter Change in the Exponential Family using Score Matching

Given an exponential family parameterisation

$$q_t(\boldsymbol{x}) = q(\boldsymbol{x}; \boldsymbol{\theta}^\star(t), \boldsymbol{\varphi}^\star) = \frac{\exp\left(\langle \boldsymbol{\theta}^\star(t), \boldsymbol{f}(\boldsymbol{x})\rangle + \langle \boldsymbol{\varphi}^\star, \boldsymbol{h}(\boldsymbol{x})\rangle\right)}{Z(\boldsymbol{\theta}^\star(t), \boldsymbol{\varphi}^\star)}, \quad (19)$$

the work was motivated by the following theorem

Theorem 6.1

Let q_t be defined as in equation (19), then the time score function for q_t can be written as

$$\dot{\psi}_{q_t}(\boldsymbol{x}) = \partial_t \log q_t = \langle \partial_t \boldsymbol{\theta}_t^\star, \boldsymbol{f}(\boldsymbol{x})\rangle - \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}_t^\star, \boldsymbol{f}(\boldsymbol{x})\rangle\right]. \quad (20)$$

Hence $\dot{\psi}_{q_t}(\boldsymbol{x})$ is *unnormalised*, and only depends on the parameters through $\partial_t \boldsymbol{\theta}_t^\star$.

Aims

Therefore, by modelling a time-evolving density through its parameter *derivatives*, $\partial_t \boldsymbol{\theta}_t^\star$, one can have full information on how the parameters change over time.

- Includes any form of parameter change (e.g. abruptly changing parameters - classical changepoint detection)
- Not limited in scope to one *type* of change
- Since $\dot{\psi}_{q_t}(\boldsymbol{x})$ is unnormalised, free to have an extremely flexible parameterisation of density function (e.g. neural network)

Adapted Time Score Matching

We propose an adaptation of time score matching using Theorem 6.1,

$$\mathcal{J}_{\mathrm{C}}(\boldsymbol{\theta}_t) \coloneqq \int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g(t)\|\dot{\psi}_{q_t}(\boldsymbol{x}) - s(\boldsymbol{x};t)\|^2\right] dt \tag{21}$$

where $s(\boldsymbol{x};t) = \langle \partial_t \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\rangle - \mathbb{E}_{q_t}\left[\langle \partial_t \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\rangle\right]$ is our model, and $g(t)$ is a weighting function.

Theorem 6.2

Let $g(t) = 0$ at $t = 0$ and $t = 1$. $\mathcal{J}_{\mathrm{C}}(\boldsymbol{\theta}_t)$ can be written as

$$\int_{t=0}^{t=1} \mathbb{E}_{q_t}\left[g_t s_t(\boldsymbol{x})^2 + 2\partial_t g_t \langle \partial_t \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\rangle + 2g_t \left\langle \partial_t^2 \boldsymbol{\theta}_t, \boldsymbol{f}(\boldsymbol{x})\right\rangle\right] dt + C, \tag{22}$$

where C is a constant independent of $s_t(\boldsymbol{x})$.

Asymptotic Distribution of $\partial_t\widehat{\boldsymbol{\theta}}_t$

Declare a change if the derivative, $\partial_t\boldsymbol{\theta}_t$ is significantly different from zero.

Therefore consider a null hypothesis, $\partial_t\boldsymbol{\theta}_t^\star = \mathbf{0}$. Any violation of this null hypothesis for a time t means there is a changepoint at time t.

Theorem 6.4

Provided there exists a true parameter $\boldsymbol{\alpha}^\star$ such that $\boldsymbol{\theta}_t^\star = {\boldsymbol{\alpha}^\star}^\top \boldsymbol{\phi}(t)$, and $\mathcal{J}_{\mathrm{C}}(\boldsymbol{\theta}_t)$ is minimised at $\boldsymbol{\alpha}^\star$, $\widehat{\boldsymbol{\alpha}}$ is a minimiser of the sample version of the objective, and $\partial_t\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\alpha}}^\top(\partial_t\boldsymbol{\phi}(t))$. As $n \to \infty$ and $T \to \infty$, the asymptotic distribution of $\partial_t\widehat{\boldsymbol{\theta}}_t$, under the null hypothesis, is given by

$$\partial_t\widehat{\boldsymbol{\theta}}_t \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_{\partial_t\widehat{\boldsymbol{\theta}}_t}\right),$$

where $\boldsymbol{\Sigma}_{\partial_t\widehat{\boldsymbol{\theta}}_t}$ contains no unknown quantities.

Asymptotic Distribution of $\partial_t \widehat{\boldsymbol{\theta}}_t$

Use the asymptotic distribution of $\partial_t \widehat{\boldsymbol{\theta}}_t$ for changepoint detection. More convenient to use the following result

Corollary 6.5

Define

$$D(t) := (\partial_t \widehat{\boldsymbol{\theta}}_t)^\top \boldsymbol{\Sigma}^{-1}_{\partial_t \widehat{\boldsymbol{\theta}}_t} (\partial_t \widehat{\boldsymbol{\theta}}_t). \tag{23}$$

then for any t, as $n \to \infty$ and $T \to \infty$, $D(t) \xrightarrow{d} \chi^2_k$.

Then we can use $D(t)$ as a univariate measure to detect changes across all dimensions of $\partial_t \boldsymbol{\theta}_t$.

Modelling $\partial_t \boldsymbol{\theta}_t$

We let

$$\boldsymbol{\theta}_t = \boldsymbol{\alpha}^\top \boldsymbol{\phi}(t),$$

where $\boldsymbol{\alpha}$ is a $b \times d$ matrix, and $\boldsymbol{\phi}$ is the *time feature* function. Consequently,

$$\partial_t^{(l)} \boldsymbol{\theta}_t = \boldsymbol{\alpha}^\top (\partial_t^{(l)} \boldsymbol{\phi}(t)).$$

- Inspired by non-parameteric modelling methods
- $\boldsymbol{\phi}$ can be complicated and as flexible as needed

This formulation is required for asymptotic theory, but no prior derivations. Other models could be chosen based on application if needed.

Some Key Results

Synthetic Examples

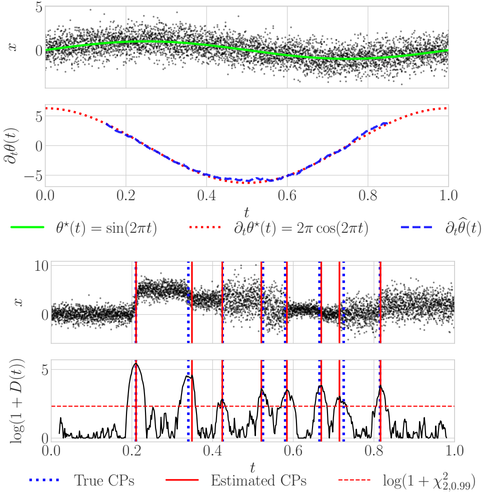


Figure 13. Two illustrative examples of using TSM-DE for parameter derivative estimation and changepoint detection.

Some Key Results

Real Data: Temperature Anomalies

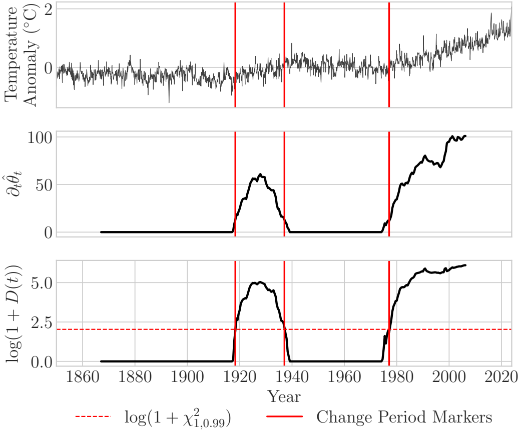


Figure 14. Changepoint detection (finding the beginning and end of the change periods) on temperature anomalies from 1850 to 2023.

Some Key Results

Real Data: Global Financial Crisis

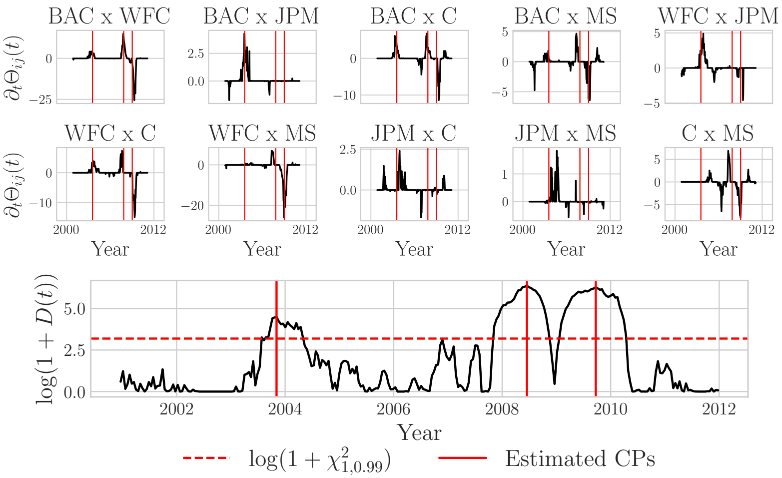


Figure 15. Estimates of $\partial_t \boldsymbol{\theta}_t$ and changepoint detection for correlation changes across the global financial crisis time period.

Some Key Results

Incremental Concept Drift

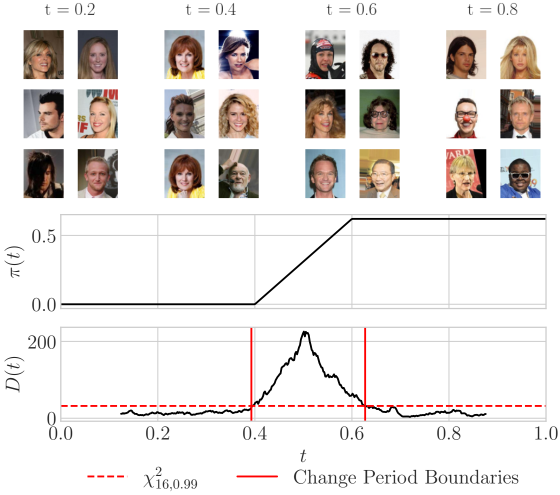


Figure 16. Detecting when out-of-sample images (glasses) are gradually introduced to a neural network representation.

References I

Barp, A., F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey (2019). Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, Volume 32.

Choi, K., C. Meng, Y. Song, and S. Ermon (2022). Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 2552–2573. PMLR.

Chwialkowski, K., H. Strathmann, and A. Gretton (2016, 20–22 Jun). A kernel test of goodness of fit. In M. F. Balcan and K. Q. Weinberger (Eds.), *Proceedings of The 33rd International Conference on Machine Learning*, Volume 48 of *Proceedings of Machine Learning Research*, New York, New York, USA, pp. 2606–2615. PMLR.

Eichinger, B. and C. Kirch (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli 24*(1), 526 – 564.

References II

Gorham, J. and L. Mackey (2015). Measuring sample quality with stein's method. *Advances in Neural Information Processing Systems 28*.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research 6*(Apr), 695–709.

Kanagawa, H., W. Jitkrittum, L. Mackey, K. Fukumizu, and A. Gretton (2019). A kernel stein test for comparing latent variable models. *arXiv preprint arXiv:1907.00586*.

Kim, J., H.-S. Oh, and H. Cho (2024). Moving sum procedure for change point detection under piecewise linearity. *Technometrics* (just-accepted), 1–19.

Liu, Q., J. Lee, and M. Jordan (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR.

References III

Liu, S., T. Kanamori, and D. J. Williams (2022). Estimating density models with truncation boundaries using score matching. *Journal of Machine Learning Research 23*(186), 1–38.

Mardia, K. V., J. T. Kent, and A. K. Laha (2016). Score matching estimators for directional distributions. *arXiv preprint arXiv:1604.08470*.

Williams, D. J. and S. Liu (2022). Score matching for truncated density estimation on a manifold. In *Topological, Algebraic and Geometric Learning Workshops 2022*, pp. 312–321. PMLR.

Williams, D. J. and S. Liu (2023, 23–29 Jul). Approximate stein classes for truncated density estimation. In *Proceedings of the 40th International Conference on Machine Learning*, Volume 202 of *Proceedings of Machine Learning Research*, pp. 37066–37090. PMLR.

Xu, W. (2022, 28–30 Mar). Standardisation-function kernel stein discrepancy: A unifying view on kernel stein discrepancy tests for goodness-of-fit. Volume 151 of *Proceedings of Machine Learning Research*, pp. 1575–1597. PMLR.

References IV

Yu, S., M. Drton, and A. Shojaie (2018, 09–11 Apr). Graphical models for non-negative data using generalized score matching. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Volume 84 of *Proceedings of Machine Learning Research*, pp. 1781–1790. PMLR.

Yu, S., M. Drton, and A. Shojaie (2019). Generalized score matching for non-negative data. *Journal of Machine Learning Research 20*(76), 1–70.