Kernel Conditional Exponential Family

Michael Arbel   **Arthur Gretton**

Gatsby Computational Neuroscience Unit, University College London

{michael.n.arbel, arthur.gretton}@gmail.com

Abstract

A nonparametric family of conditional distributions is introduced, which generalizes conditional exponential families using functional parameters in a suitable RKHS. An algorithm is provided for learning the generalized natural parameter, and consistency of the estimator is established in the well specified case. In experiments, the new method generally outperforms a competing approach with consistency guarantees, and is competitive with a deep conditional density model on datasets that exhibit abrupt transitions and heteroscedasticity.

1 Introduction

Distribution estimation is one of the most general problems in machine learning. Once an estimator for a distribution is learned, in principle, it allows to solve a variety of problems such as classification, regression, matrix completion and other prediction tasks. With the increasing diversity and complexity of machine learning problems, regressing the conditional mean of y knowing x may not be sufficiently informative when the conditional density $p(y|x)$ is multimodal. In such cases, one would like to estimate the conditional distribution itself to get a richer characterization of the dependence between the two variables y and x. In this paper, we address the problem of estimating conditional densities when x and y are continuous and multi-dimensional.

Our conditional density model builds on a generalisation of the exponential family to infinite dimensions (Barron et al., 1991; Canu et al., 2006; Fukumizu, 2009; Gu et al., 1993; Pistone et al., 1995), where the natural parameter is a function in a reproducing kernel

Hilbert space (RKHS): in this sense, like the Gaussian and Dirichlet processes, the kernel exponential family (KEF) is an infinite dimensional analogue of the finite dimensional case, allowing to fit a much richer class of densities. While the maximum likelihood solution is ill-posed in infinite dimensions, Sriperumbudur et al. (2017) have demonstrated that it is possible to fit the KEF via score matching (Hyvärinen, 2005), which entails solving a linear system of size $n \times d$, where n is the number of samples and d is the problem dimension. It is trivial to draw samples from such models using Hamiltonian Monte Carlo (Neal, 2010), since they directly return the required potential energy (Rasmussen, 2003; Strathmann et al., 2015). In high dimensions, fitting a KEF model to samples becomes challenging, however: the computational cost rises as d^3, and complex interactions between dimensions can be difficult to model.

The complexity of the modelling task can be significantly reduced if a directed graphical model can be constructed over the variables, (Jordan, 1999; Pearl, 2001), where each variable depends on a subset of parent variables (ideally much smaller than the total, as in e.g. a Markov chain). In the present study, we extend the non-parametric family of Sriperumbudur et al., 2017 to fit conditional distributions. The natural parameter of the *conditional* infinite exponential family is now an operator mapping the conditioning variable to a function space of features of the conditioned variable: for this reason, the score matching framework must be generalised to the vector-valued kernel regression setting of Micchelli et al., 2005. We establish consistency in the well specified case by generalising the arguments of Sriperumbudur et al., 2017 to the vector-valued RKHS. While our proof allows for general vector-valued RKHSs, we provide a practical implementation for a specific case, which takes the form of a linear system of size $n \times d$ [1].

A number of alternative approaches have been proposed to the problem of conditional density estimation.

---

[1] The code can be found at:
https://github.com/MichaelArbel/KCEF

Sugiyama et al., 2010 introduced the Least-Square Conditional Density Estimation (LS-CDE) method, which provides an estimate of a conditional density function $p(y|x)$ as a non-negative linear combination of basis functions. The method is proven to be consistent, and works well on reasonably complicated learning problems, although the optimal choice of basis functions for the method is an open question (in their paper, the authors use Gaussians centered on the samples). Earlier non-parametric methods such as variants of Kernel Density Estimation (KDE) may also be used in conditional density estimation (Fan et al., 1996; Hall et al., 1999). These approaches also have consistency guarantees, however their performance degrades in high-dimensional settings (Nagler et al., 2016). Sugiyama et al., 2010 found that kernel density approaches performed less well in practice than LS-CDE.

Kernel Quantile Regression (KQR), introduced by (Takeuchi et al., 2006; Zhang et al., 2016), allows to predict a percentile of the conditional distribution when y is one-dimensional. KQR is formulated as a convex optimization problem with a unique global solution, and the entire solution path with respect to the percentile parameter can be computed efficiently (Takeuchi et al., 2009). However, KQR applies only to one-dimensional outputs and, according to Sugiyama et al., 2010, the solution path tracking algorithm tends to be numerically unstable in practice.

It is possible to represent and learn conditional probabilities without specifying probability densities, via conditional mean embeddings (Grunewalder et al., 2012; Song et al., 2010). These are conditional expectations of (potentially infinitely many) features in an RKHS, which can be used in obtaining conditional expectations of functions in this RKHS. Such expected features are complementary to the infinite dimensional exponential family, as they can be thought of as conditional expectations of an infinite dimensional sufficient statistic. This statistic can completely characterise the conditional distribution if the feature space is sufficiently rich (Sriperumbudur et al., 2010), and has consistency guarantees under appropriate smoothness assumptions. Drawing samples given a conditional mean embedding can be challenging: this is possible via the Herding procedure (Bach et al., 2012; Chen et al., 2010), as shown in (Kanagawa et al., 2016), but requires a non-convex optimisation procedure to be solved for each sample.

A powerful and recent deep learning approach to modelling conditional densities is the Neural Autoregressive Network (Uria et al., 2013, Raiko et al., 2014 and Uria et al., 2016). These networks can be thought of as a generalization of the Mixture Density Network introduced by Bishop, 2006. In brief, each variable is represented by a mixture of Gaussians, with means and variances depending on the parent variables through a deep neural network. The network is trained on observed data using stochastic gradient descent. Neural autoregressive networks have shown their effectiveness for a variety of practical cases and learning problems. Unlike the earlier methods cited, however, consistency is not guaranteed, and these methods require non-convex optimization, meaning that locally optimal solutions are found in practice.

We begin our presentation in Section 2, where we briefly present the Kernel Exponential Family. We generalise this model to the conditional case, in our first major contribution: this requires the introduction of vector-valued RKHSs and associated concepts. We then show that a generalisation of score matching may be used to fit the conditional density models for general vector valued RKHS, subject to appropriate conditions. We call this model the kernel conditional exponential family (KCEF).

Our second contribution, in Section 3, is an empirical estimator for the *natural parameter* of the KCEF (Theorem 1), with convergence guarantees in the well-specified case (Theorem 2). In our experiments (Section 4), we empirically validate the consistency of the estimator and compare it to other methods of conditional density estimation. Our method generally outperforms the leading alternative with consistency guarantees (LS-CDE). Compared with the deep approach (RNADE) which lacks consistency guarantees, our method has a clear advantage at small training sample sizes while being competitive at larger training sizes.

2 Kernel Exponential Families

In this section we first present the kernel exponential family, which we then extend to a class of conditional exponential families. Finally, we provide a methodology for unnormalized density estimation within this class.

2.1 Kernel Conditional Exponential Family

We consider the task of estimating the density $p(y)$ of a random variable Y with support $\mathcal{Y} \subseteq \mathbb{R}^d$ from i.i.d samples $(Y_i)_{i=1}^n$. We propose to use a family of densities parametrized by functions belonging to a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\mathcal{Y}$ with positive semi-definite kernel k (Canu et al., 2006; Fukumizu, 2009; Sriperumbudur et al., 2017). This exponential family of density functions takes the form

$$\left\{ p(y) := q_0(y) \frac{\exp \langle f, k(y, \cdot) \rangle_{\mathcal{H}_\mathcal{Y}}}{Z(f)} \middle| f \in \mathcal{F} \right\}, \quad (1)$$

where q_0 is a base density function on $\mathcal{Y}$ and $\mathcal{F}$ is the set of functions in the RKHS space $\mathcal{H}_\mathcal{Y}$ such that $Z(f) := \int_\mathcal{Y} \exp \langle f, k(y,.)\rangle_{\mathcal{H}_\mathcal{Y}} q_0(y) dx < \infty$. In what follows, we call this family the *kernel exponential family* (KEF) by analogy to classical exponential family. f plays the role of the natural parameter while $k(y,.)$ is the sufficient statistic. Note that with an appropriate choice of the base distribution q_0 and a finite dimensional RKHS $\mathcal{H}_\mathcal{Y}$, one can recover any finite dimensional exponential family. When $\mathcal{H}_\mathcal{Y}$ is infinite-dimensional, however, the family can approximate a much broader class of densities on $\mathbb{R}^d$: under mild conditions, it is shown in Sriperumbudur et al., 2017 that the KEF approximates all densities of the form $\{q_0(y) \exp(f(y) - A)|f \in C_0(\mathcal{Y})\}$, A being the normalizing constant and $C_0(\mathcal{Y})$ the set of continuous functions with vanishing tails.

Given two subsets $\mathcal{Y}$ and $\mathcal{X}$ of $\mathbb{R}^d$ and $\mathbb{R}^p$ respectively, we now propose to extend the KEF to a family of conditional densities $p(y|x)$. We modify equation (1) by making the function f depend on the conditioning variable x. The parameter f is a function of two variables x and y, $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $y \mapsto f(x,y)$ belongs to the RKHS $\mathcal{H}_\mathcal{Y}$ for all x in $\mathcal{X}$. In all that follows, we will denote by T the mapping

$$T \,:\, \mathcal{X} \to \mathcal{H}_\mathcal{Y} \qquad x \mapsto T_x$$

such that $T_x(y) = f(x,y)$ for all y in $\mathcal{Y}$

We next consider how to enforce a smoothness requirement on T to make the conditional density estimation problem well-posed. To achieve this, we will require that the mapping T belongs to a vector valued RKHS $\mathcal{H}$: we now briefly review the associated theory, following (Micchelli et al., 2005). A Hilbert space $(\mathcal{H}, \langle ., .\rangle_\mathcal{H})$ of functions $T : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ taking values in a vector space $\mathcal{H}_\mathcal{Y}$ is said to be a vector valued RKHS if for all $x \in \mathcal{X}$ and $h \in \mathcal{H}_\mathcal{Y}$, the linear functional $T \mapsto \langle h, T_x\rangle_{\mathcal{H}_\mathcal{Y}}$ is continuous. The reproducing property for vector-valued RKHSs follows from this definition. By the Riesz representation theorem, for each $x \in \mathcal{X}$ and $h \in \mathcal{H}_\mathcal{Y}$, there exists a linear operator Γ_x from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}$ such that for all $T \in \mathcal{H}$,

$$\langle h, T_x\rangle_{\mathcal{H}_\mathcal{Y}} = \langle T, \Gamma_x h\rangle_\mathcal{H} \qquad (2)$$

Considering the dual operator Γ_x^* from $\mathcal{H}$ to $\mathcal{H}_\mathcal{Y}$, we also get

$$\Gamma_x^* T = T_x.$$

We can define a vector-valued reproducing kernel by composing the operator Γ_x with its dual,

$$\Gamma(x, x') = \Gamma_x^* \Gamma_{x'},$$

where for all x and x', $\Gamma(x, x')$ is a bounded linear operator from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}_\mathcal{Y}$, i.e., $\Gamma(x, x') \in \mathcal{L}(\mathcal{H}_\mathcal{Y})$. The space $\mathcal{H}$ is said to be generated by an operator valued reproducing kernel Γ. One practical choice for Γ is to define it as:

$$\Gamma(x, x') = k_\mathcal{X}(x, x') I_{\mathcal{H}_\mathcal{Y}} \qquad \forall x, x' \in \mathcal{X}, \quad (3)$$

where $I_{\mathcal{H}_\mathcal{Y}}$ the identity operator on $\mathcal{H}_\mathcal{Y}$ and $k_\mathcal{X}$ is now a real-valued kernel which generates a real valued RKHS $\mathcal{H}_\mathcal{X}$ on $\mathcal{X}$ (as in the conditional mean embedding; see Grunewalder et al., 2012). A simplified form of the estimator of T will be presented in Section 3 for this particular choice for Γ and will be used in the experimental setup in Section 4.

We will now express $T_x(y)$ in a convenient form that will allow to extend the KEF. For a given x, recalling that T_x belongs to $\mathcal{H}_\mathcal{Y}$, one can write $T_x(y) = \langle T_x, k(y,.)\rangle_{\mathcal{H}_\mathcal{Y}}$ for all y in $\mathcal{Y}$. Using the reproducing property in (2), one further gets $T_x(y) = \langle T, \Gamma_x k(y,.)\rangle_\mathcal{H}$. By considering the subset $\mathcal{T}$ of $\mathcal{H}$ such that for all x in $\mathcal{X}$ the integral

$$Z(T_x) := \int_\mathcal{Y} q_0(y) \exp(\langle T, \Gamma_x k(y,.)\rangle_\mathcal{H}) dy < \infty$$

is finite, we define the *kernel conditional exponential family* (KCEF) as the set of conditional densities

$$\left\{ p_T(y|x) := q_0(y) \frac{\exp \langle T, \Gamma_x k(y,.)\rangle_\mathcal{H}}{Z(T_x)} \bigg| T \in \mathcal{T} \right\}. \quad (4)$$

Here T plays the role of the natural parameter while $\Gamma_x k(y,.)$ is the sufficient statistic. When T is restricted to be constant with respect to x, we recover the *kernel exponential family* (KEF). The KCEF is therefore an extension of the KEF introduced in Sriperumbudur et al., 2017. It is also a special case of the family introduced in Canu et al., 2006. In the latter, the inner product is given by $\langle T, \phi(x,y)\rangle_\mathcal{H}$ where ϕ is a general feature of x and y. In the present work, ϕ has the particular form $\phi(x,y) = \Gamma_x k(y,.)$. This allows to further express $p_T(y|x)$ for a given x as an element in a KEF with sufficient statistic $k(y,.)$, by using the identity $\langle T, \Gamma_x k(y,.)\rangle_\mathcal{H} = \langle T_x, k(y,.)\rangle_{\mathcal{H}_\mathcal{Y}}$. This is desirable since $p_T(y|x)$ remains in the same KEF family as x varies and only the natural parameter T_x changes.

2.2 Unnormalized density estimation

Given i.i.d samples $(X_i, Y_i)_{i=1}^n$ in $\mathcal{X} \times \mathcal{Y}$ following a joint distribution $\pi(x) p_0(y|x)$, where π defines a marginal distribution over $\mathcal{X}$ and $p_0(y|x)$ is a conditional density function, we are interested in estimating p_0 from the samples $(X_i, Y_i)_{i=1}^n$. Our goal is to find the optimal conditional density p_T in the KCEF that best

approximates p_0. The intractability of the normalizing constant $Z(T_x)$ makes maximum likelihood estimation difficult. Sriperumbudur et al., 2017 used a score-matching approach (see Hyvärinen, 2005) to avoid this normalizing constant; in the case of the KCEF, however, the score function between $\pi(x)p_0(y|x)$ and $\pi(x)p_T(y|x)$ contains additional terms that involve the derivatives of the log-partition function $\log Z(T_x)$ with respect to x. Instead, we now propose a different approach with a modified version of the score-matching objective.

We define the expected conditional score between two conditional densities $p_0(y|x)$ and $q(y|x)$ under a marginal density π on x to be:

$$J(p_0|q) := \int_{\mathcal{X}} \pi(x)\mathcal{J}(p_0(.|x)\|q(.|x))dx$$

where:

$$\mathcal{J}(p_0(.|x)\|q(.|x)) = \frac{1}{2}\int_{\mathcal{Y}} p_0(y|x)\left\|\nabla_y \log \frac{p_0(y|x)}{q(y|x)}\right\|^2 dy.$$

For a fixed value x in $\mathcal{X}$, $\mathcal{J}(p_0(.|x)\|q(.|x))$ is the score-matching function between $p_0(.|x)$ and $q(.|x)$ as defined by Hyvärinen, 2005. We further take the expectation over x to define a divergence over conditional densities. The normalizing constant of $q(y|x)$, which is a function of x, is never involved in this formulation, as we take the gradient of the log-densities over y only. For a conditional density $p_0(y|x)$ that is supported on the whole domain $\mathcal{Y}$ for all x in $\mathcal{X}$, the expected conditional score is well behaved in the sense that $J(p_0|q)$ is always non-negative, and reaches 0 if and only if the two conditional distributions $p_0(y|x)$ and $q(y|x)$ are equal for π-almost all x. More discussion can be found in Appendix D, for the case when this condition fails to hold. The goal is then to find a conditional distribution p_T in the KCEF for a given $T \in \mathcal{T}$ that minimizes this score over the whole family.

Under mild regularity conditions on the densities (see Hyvärinen, 2005; Sriperumbudur et al., 2017, and below), the score can be rewritten

$$J(p_0\|p_T) = \mathbb{E}\left[\sum_{i=1}^d \partial_i^2 T_x(y) + \frac{1}{2}(\partial_i T_x(y))^2\right]$$

$$+ \mathbb{E}\left[\sum_{i=1}^d \partial_i T_x(y)\partial_i \log q_0(y)\right] + J(p_0\|q_0)$$

where $J(p_0\|q_0)$ is a constant term for the optimization problem and the expectation is taken over $\pi(x)p_0(y|x)$. All derivatives are with respect to y, and we used the notation $\partial_i f(y) = \frac{\partial}{\partial y_i}f(y)$. In the case of KCEF,

conditions to obtain this expression are satisfied under assumptions in Appendix A.4, as proved in Theorem 3 of Appendix B.1 . The expression is further simplified using the reproducing property for the derivatives of functions in an RKHS (Lemma 3 of Appendix C),

$$\partial_i T_x(y) = \langle T, \Gamma_x \partial_i k(y,.)\rangle_{\mathcal{H}}$$
$$\partial_i^2 T_x(y) = \langle T, \Gamma_x \partial_i^2 k(y,.)\rangle_{\mathcal{H}}$$

which leads to:

$$J(T) = \mathbb{E}\left[\sum_{i=1}^d \frac{1}{2}\langle T, \Gamma_x \partial_i k(y,.)\rangle_{\mathcal{H}}^2 + \langle T, \xi_i(x,y)\rangle_{\mathcal{H}}\right]$$

with:

$$\xi_i(x,y) = \Gamma_x(\partial_i^2 k(y,.) + \partial_i \log q_0(y)\partial_i k(y,.)). \quad (5)$$

We introduced the notation $J(T) := J(p_0\|p_T) - J(p_0\|q_0)$ for convenience. This formulation depends on $p_0(y|x)$ only through an expectation, therefore a Monte Carlo estimator of the score can be derived as a quadratic functional of T in the RKHS $\mathcal{H}$,

$$\hat{J}(T) = \frac{1}{n}\sum_{\substack{b\in[n]\\i\in[d]}} \frac{1}{2}\langle T, \Gamma_{X_b}\partial_i k(Y_b,.)\rangle_{\mathcal{H}}^2 + \langle T, \xi_i(X_b, Y_b)\rangle_{\mathcal{H}}.$$

Note that the objective functions $J(T)$ and $\hat{J}(T)$ can be defined over the whole space $\mathcal{H}$, whereas $J(p_0|p_T)$ is meaningful only if T belongs to $\mathcal{T}$.

3 Empirical KCEF and consistency

In this section, we will first estimate the optimal $T^* = \text{argmin}_{T\in\mathcal{H}} J(T)$ over the whole space $\mathcal{H}$ by minimizing a regularized version of the quadratic form in equation $\hat{J}(T)$, then we will state conditions under which all of the obtained solutions belong to $\mathcal{T}$ defining therefore conditional densities in the KCEF.

Following Sriperumbudur et al., 2017, we define the kernel ridge estimator to be $T_{n,\lambda} = \text{argmin}_{T\in\mathcal{H}} \hat{J}(T) + \frac{\lambda}{2}\|T\|_{\mathcal{H}}^2$ where $\|T\|_{\mathcal{H}}$ is the RKHS norm of T. $T_{n,\lambda}$ is then obtained by solving a linear system of nd variables as shown in the next theorem:

Theorem 1. *Under assumptions listed in Appendix A.4, and in particular if $\|\Gamma(x,x)\|_{Op}$ is uniformly bounded on $\mathcal{X}$ for the operator norm, then the minimizer $T_{n,\lambda}$ exists, is unique, and is given by*

$$T_{n,\lambda} = -\frac{1}{\lambda}\hat{\xi} + \sum_{b\in[n];i\in[d]} \beta_{(b,i)}\Gamma_{X_b}\partial_i k(Y_b,\cdot),$$

where

$$\hat{\xi} = \frac{1}{n}\sum_{b\in[n];i\in[d]} \xi_i(X_b, Y_b),$$

and ξ_i are given by (5). $\beta_{(b,i)}$ denotes the $(b-1)d+i$ entry of a vector β in $\mathbb{R}^{nd}$, obtained by solving the linear system

$$(G + n\lambda I)\beta = \frac{h}{\lambda},$$

where G is an nd by nd Gram matrix, and h is a vector in $\mathbb{R}^{nd}$,

$$(G)_{(a,i),(b,j)} = \langle \Gamma_{X_a} \partial_i k(Y_a, \cdot), \Gamma_{X_b} \partial_j k(Y_b, \cdot) \rangle_{\mathcal{H}}$$
$$(h)_{(b,i)} = \langle \hat{\xi}, \Gamma_{X_b} \partial_i k(Y_a, \cdot) \rangle_{\mathcal{H}}.$$

The result is proved in Theorem 4 of Appendix B.2.

For the particular choice of Γ in (3), the estimator takes a simplified form

$$T_{n,\lambda}(x,y) = -\frac{1}{\lambda}\hat{\xi}(x,y) + \sum_{\substack{b \in [n] \\ i \in [d]}} \beta_{(b,i)} k_{\mathcal{X}}(X_b, x) \partial_i k(Y_b, y),$$

with

$$\hat{\xi}(x,y) = \frac{1}{n} \sum_{b \in [n]; i \in [d]} k_{\mathcal{X}}(X_b, x) \partial_i^2 k(Y_b, y)$$
$$+ \frac{1}{n} \sum_{b \in [n]; i \in [d]} k_{\mathcal{X}}(X_b, x) \partial_i \log q_0(Y_b) \partial_i k(Y_b, y)$$

The coefficients β are obtained by solving the same system $(G + n\lambda I)\beta = \frac{h}{\lambda}$, where G and h reduce to

$$(G)_{(a,i),(b,j)} = k_{\mathcal{X}}(X_a, X_b) \partial_i \partial_{j+d} k(Y_a, Y_b),$$
$$(h)_{(b,i)} = \partial_i \hat{\xi}(X_b, Y_b),$$

and all derivatives are taken with respect to y.

The above estimator generalizes the estimator in Sriperumbudur et al., 2017 to conditional densities. In fact, if one choses the kernel $k_{\mathcal{X}}$ to be a constant kernel, then one exactly recovers the setting of the KEF.

This linear system has a complexity of $\mathcal{O}(n^3 d^3)$ in time and $\mathcal{O}(n^2 d^2)$ in memory, which can be problematic for higher dimensions d as n grows. However, in practice, if the goal is to estimate a density of the form $p(x_1, ..., x_d)$, one can use the general chain rule for distributions, $p(x_1)p(x_2|x_1)....p(x_d|x_1, ..., x_{d-1})$, and estimate each conditional density $p(x_i|x_1, ..., x_{i-1})$ using the KCEF in (4). This reduces the complexity of the algorithm to $\mathcal{O}(n^3 d)$. A reduction to the cubic complexity in the number of data points n could be managed via a Nyström-like approximation (Sutherland et al., 2017).

In the well-specified case where the true conditional density $p_0(y|x)$ is assumed to be in (4) (i.e. $p_0(y|x) = p_{T_0}(y|x)$), we analyze the parameter convergence of

the estimator $T_{n,\lambda}$ to T_0 and the convergence of the corresponding density $p_{T_{n,\lambda}}(y|x)$ to the true density $p_0(y|x)$. First, we consider the covariance operator C of the joint feature $\Gamma_x k(y, \cdot)$ under the joint distribution of x and y, as introduced in Theorem 3 of Appendix B.1, and we denote by $\mathcal{R}(C^\gamma)$ the range space of the operator C^γ. We then have the following consistency result:

Theorem 2. *Let $\gamma > 0$ be a positive parameter and define $\alpha = \max(\frac{1}{2(\gamma+1)}, \frac{1}{4}) \in (\frac{1}{4}, \frac{1}{2})$. Under the conditions in Appendix A.4, for $\lambda = n^{-\alpha}$, and if $T_0 \in \mathcal{R}(C^\gamma)$, then*

$$\|T_{n,\lambda} - T_0\| = \mathcal{O}_{p_0}(n^{-\frac{1}{2}+\alpha}).$$

Furthermore, if $\sup_{y \in \mathcal{Y}} k(y,y) < \infty$, then

$$KL(p_0 || p_{T_{n,\lambda}}) = \mathcal{O}_{p_0}(n^{-1+2\alpha}).$$

These asymptotic rates match those obtained for the unconditional density estimator in Sriperumbudur et al., 2017. The smoother the parameter T_0, the closer α gets to $\frac{1}{4}$, which in turns leads to a convergence rate in KL divergence of the order of $\frac{1}{\sqrt{n}}$. The worst case scenario happens when the range-space parameter γ gets closer to 0, in which case convergence in KL divergence happens at a rate close to $\frac{1}{n^\gamma}$. A more technical formulation of this theorem along with a proof is presented in Appendix B.3 (see Theorems 5 and 6).

The regularity of the conditional density $p(y|x)$ with respect to x is captured by the boundedness assumption on the operator valued kernel Γ; i.e., $\|\Gamma(x,x)\|_{op} \leq \kappa$ for all $x \in \mathcal{X}$ in Assumption (**E**). This assumption allows to control the variations of the conditional distribution $p(y|x)$ as x changes. Roughly speaking, we may estimate the conditional density $p(y|x_0)$ at a given point x_0 from samples (Y_i, X_i) whenever there are X_i sufficiently close to x_0. The uniformly bounded kernel Γ allows to express the objective function $J(T)$ as a quadratic form $J(T) = \frac{1}{2}\langle T, CT \rangle_{\mathcal{H}} + \langle T, \xi \rangle_{\mathcal{H}} + c_0$ for constant c_0, where C is the covariance operator introduced in Theorem 3. Furthermore, this boundedness assumption ensures that C is a "well-behaved" operator, namely a positive semi-definite trace-class operator. The population solution of the regularized score objective is then given by $T_\lambda = (C + \lambda I)^{-1} CT_0$ while the estimator is given by: $\hat{T}_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\xi}$ where $\hat{C}$ and $\hat{\xi}$ are empirical estimators for C and ξ.

The proof of consistency makes use of ideas from Caponnetto et al., 2007; Sriperumbudur et al., 2017, exploiting the properties of trace-class operators. The main idea is to first control the error $\|T_0 - \hat{T}_{\lambda,n}\|_{\mathcal{H}}$ by introducing the population solution T_λ,

$$\|T_0 - \hat{T}_{\lambda,n}\|_{\mathcal{H}} \leq \|T_0 - T_\lambda\|_{\mathcal{H}} + \|T_\lambda - \hat{T}_{\lambda,n}\|_{\mathcal{H}}$$

The first term $\|T_0 - T_\lambda\|_{\mathcal{H}}$ represents the regularization error which is introduced by adding a regularization term λI to the operator C. This term doesn't depend on n, and can be shown to decrease as the amount of regularization goes to 0 with a rate $\lambda^{min(1,\gamma)}$. The second term represents the estimation error due to the finite number of samples n. This term decreases as $n \to 0$ but also increases when $\lambda \to 0$, therefore a trade-off needs to be made between decreasing the first term $\|T_0 - T_\lambda\|_{\mathcal{H}}$ by setting $\lambda \to 0$ and keeping the term $\|T_\lambda - \hat{T}_{\lambda,n}\|_{\mathcal{H}}$ small enough. Using decompositions similar to those of Caponnetto et al., 2007; Sriperumbudur et al., 2017, we apply concentration inequalities on the general Hilbert space $\mathcal{H}$ to get a probabilistic bound on the estimation error of order $\mathcal{O}(\frac{1}{\lambda\sqrt{n}})$.

Concerning the convergence in KL divergence, the requirement that the real-valued kernel k is bounded implies that $\mathcal{T}$ is in fact equal to $\mathcal{H}$. Therefore, minimizing the expected score $J(p_{T_0}||p_T)$ is equivalent to minimizing the quadratic form $J(T)$ over the whole RKHS $\mathcal{H}$. Finally, the rates in KL divergence are obtained from the error rate of $\hat{T}_{\lambda,n}$.

4 Experiments

We perform a diverse set of experiments, on both synthetic and real data, in order to validate our model empirically. In all experiments, the data are centered and rescaled such that the standard deviation for every dimension is equal to 1. Given $(X_1^{(n)}, ..., X_d^{(n)})_{n=1}^N$ i.i.d. samples of dimension d we are interested in approximating the joint distribution $p_0(X_1, ..., X_d)$ of data using different methods:

• The **KEF** model from Sriperumbudur et al., 2017 approximates p by a distribution p_f that belongs to the KEF (1) by minimizing the score loss between p and p_f to find the optimal parameter f.

• The **KCEF** model of Theorem 1 approximates p by a distribution $\hat{p}$ that is assumed to factorize according to some Directed Acyclic Graphical model (DAG): $\hat{p}(X_1, ..., X_d) = \Pi_{i=1}^d \hat{p}(x_i|x_{\pi(i)})$ where $\pi(i)$ are the parent nodes of i. Note that we do not necessarily make independence assumptions, as the graph can be fully connected. We will consider in particular two graphs, the Full graph (**F**) of the form $\hat{p}(X_1, ..., X_d) = \hat{p}(X_1)\Pi_{i=2}^d \hat{p}(X_i|X_1, ..., X_{i-1})$ and the Markov graph (**M**) of the form $\hat{p}(X_1, ..., X_d) = \hat{p}(X_1)\Pi_{i=2}^d \hat{p}(X_i|X_{i-1})$. Each of the factors is assumed to belong to the KCEF in (4), and is estimated independently from the others by minimizing the empirical loss $\hat{J}(T)$ to find the optimal operator T_i such that $\hat{p}(X_i|X_{\pi(i)}) = p_{T_i}(X_i|X_{\pi(i)})$.

• The **Orderless RNADE** model in Uria et al., 2016, where we train a 2 Layer Neural Autoregres-sive model with 100 units per layer. The model consists of a product of conditional densities of the form $\Pi_{i=1}^d p(X_{o_i}|X_{o_{<i}}, \theta, o)$, where o is a permutation of the dimensions $[1, ..., d]$ and θ is a set of parameters that are shared across the factors regardless of the chosen permutation o. RNADE is trained by minimizing the empirical expected negative log-likelihood, where the expectation is taken over all possible permutations and data,

$$\mathcal{L}(\theta) = \mathbb{E}_{o \in D} \, \mathbb{E}_{X \in \mathbb{R}^d} \Big[-\log p(X_{o_i}|X_{o_{<i}}, \theta, o) \Big].$$

• The **LSCDE** model in Sugiyama et al., 2010 where we also used the 2 factorizations of the joint distribution (**F**, **M**) and solve a least-squares problem to estimate each of the conditional densities. The approximate densities are of the form $\alpha^T \phi(X_i, X_{\pi(i)})$ where ϕ is a vector of m known non-negative functions and α is obtained by minimizing the squared error between $p(X_i, X_{\pi(i)})$ and $\alpha^T \phi(X_i, X_{\pi(i)})$. Only the non-negative component of the solution α is used.

For all variants of our model, we take the base density q_0 to be a centered gaussian with a standard deviation of 2. The kernel function used for both predicted variable y and conditioning variable x is the anisotropic radial basis function (RBF) with per-dimension bandwidths. The bandwidths and the regularization parameter λ are tuned by gradient descent on the cross validated score.

Synthetic data: We consider the 'grid' dataset, which is a d-dimensional distribution with a tractable density that factorizes in the form

$$p(x_i|x_{i-1}) = C_i(1 + \sin(2\pi w_i^a x_i)\sin(2\pi w_i^b x_{i-1}))$$

for all $i \in [d]$. C_i is a tractable normalizing constant. Samples are generated using rejection sampling for each dimension. To study the effect of sample size on the estimator, we generate n training points with n varying from 200 to 2000 and $d = 3$, and estimate the log-likelihood on 2000 newly generated points. To compare the effect of dimension, we generate 2000 datapoints of dimension d varying from 2 to 20, and estimate the log-likelihood on 2000 test points. Unlike in (Sriperumbudur et al., 2017; Sutherland et al., 2017), the score function $\hat{J}(T)$ cannot be used as a metric to compare different factorizations of the estimated distribution, as it is dependent on the specific factorization of the joint distribution. Instead, we estimated the log-likelihood for our proposed model **KCEF**, where the normalizing constants are computed using importance sampling. We discarded the **KEF** in this experiment, since estimating the normalizing constant in high dimensions becomes impractical.

In Figure 1(left), we plot the log-likelihood as the number of samples increases. Both variants of **KCEF** (**F**,

M) performed slightly better than the other methods in terms of speed of convergence as sample size increases. The variants that exploit the Markov structure of data **M** lead to the best performance for both **KCEF** and **LSCDE** as expected. The **NADE** method has comparable performance for large sample sizes, but the performance drops significantly for small sample sizes. This behaviour will also be observed in subsequent experiments on real data. The figure on the right shows the evolution of the log-likelihood per dimension as the dimension increases. In the **F** case, our approach is comparable to **LSCDE** with an advantage in small dimensions. The **F** approaches both use an anisotropic RBF kernel with tuned per-dimension bandwidth which end up performing a kind of automatic relevance determination. This helps getting comparable performance to the **M** methods. A drastic drop in performance can happen when an isotropic kernel is used instead as confirmed by Figure 5 of Appendix E. Finally, **NADE**, which is also agnostic to the Markov structure of data, seems to achieve comparable performance to the **F** methods with a slight disadvantage in higher dimensions.

Real data: We applied the proposed and existing methods to the R package benchmark datasets (Team, 2008) (see Table 1) as well as three UCI datasets previously used to study the performance of other density estimators (see Silva et al., 2011; Tang et al., 2012; Uria et al., 2013). In all cases data are centered and normalized.

First, the R benchmark datasets are low dimensional with few samples, but with a relatively complex conditional dependence between the variables. This setting allows to compare the methods in terms of data efficiency and overfitting. Each dataset was randomly split into a training and a test set of equal size. The models are trained to estimate the conditional density of a one dimensional variable y knowing x using samples $(x_i, y_i)_{i=1}^n$ form the training set. The accuracy is measured by the negative log-likelihood for the test samples $(\tilde{x}_i, \tilde{y}_i)_{i=1}^n$ averaged over 20 random splits of data. We compared the proposed method with **NADE** and **LSCDE** on 14 datasets. For **NADE** we used CV over the number of units per layer $\{2, 10, 100\}$ and number of mixture components $\{1, 2, 5, 10\}$ for a 2 layer network. We also used CV to chose the hyperparameters for **LSCDE** and the proposed method on a 20×20 grid (for λ and σ).

The experimental results are summmarized in Table 1. **LSCDE** worked well in general as claimed in the original paper, however the proposed method substantially improves the results. On the other hand, **NADE** performed rather poorly due to the small sample size of the training set, despite our attempts to improve its

	KCEF	NADE	LSCDE
caution	**0.99 ± 0.01**	4.12 ± 0.02	1.19 ± 0.02
ftcollinssnow	**1.46 ± 0.0**	3.09 ± 0.02	1.56 ± 0.01
highway	**1.17 ± 0.01**	11.02 ± 1.05	1.98 ± 0.04
heights	**1.27 ± 0.0**	2.71 ± 0.0	1.3 ± 0.0
sniffer	**0.33 ± 0.01**	1.51 ± 0.04	0.48 ± 0.01
snowgeese	**0.72 ± 0.02**	2.9 ± 0.15	1.39 ± 0.05
GAGurine	**0.46 ± 0.0**	1.66 ± 0.02	0.7 ± 0.01
geyser	1.21 ± 0.04	1.43 ± 0.07	**0.7 ± 0.01**
topo	**0.67 ± 0.01**	4.26 ± 0.02	0.83 ± 0.0
BostonHousing	**0.3 ± 0.0**	3.46 ± 0.1	1.13 ± 0.01
CobarOre	3.42 ± 0.03	4.7 ± 0.02	**1.61 ± 0.02**
engel	**0.18 ± 0.0**	1.46 ± 0.02	0.76 ± 0.01
mcycle	**0.56 ± 0.01**	2.24 ± 0.01	0.93 ± 0.01
BigMac2003	**0.59 ± 0.01**	13.8 ± 0.13	1.63 ± 0.03

Table 1: Mean and std. deviation of the negative log-likelihood on benchmark data over 20 runs, with different random splits. In all cases $d_y = 1$. Best method in boldface (two-sided paired *t-test* at 5%).

	white -wine	parkinsons	red wine
KCEF-F	13.05 ± 0.36	**2.86 ± 0.77**	11.8 ± 0.93
KCEF-M	14.36 ± 0.37	5.53 ± 0.79	13.31 ± 0.88
LSCDE-F	13.59 ± 0.6	15.89 ± 1.48	14.43 ± 1.5
LSCDE-M	14.42 ± 0.66	10.22 ± 1.45	14.06 ± 1.36
NADE	**10.55 ± 0.0**	3.63 ± 0.0	**9.98 ± 0.0**

Table 2: UCI results: average and standard deviation of the negative log-likelihood over 5 runs with different random splits. Best method in boldface (two-sided paired *t-test* at 5%).

performance by reducing the number of parameters to train and by introducing early stopping.

The UCI datasets (Red Wine, White Wine and Parkinsons) represent challenging datasets with non-linear dependencies and abrupt transitions between high and low density regions. This makes the densities difficult to model using standard tools such as mixtures of Gaussians or factor analysis. They also contain enough training sample points to allow a stronger performance by **NADE**. All discrete-valued variables were eliminated as well as one variable from every pair of variables that are highly correlated (Pearson correlation greater than 0.98). Following Uria et al., 2013, 90% of the data were used for training while 10% were held-out for testing. Two different graph factorizations (**F**, **M**) were used for the proposed method and for **LSCDE**.

In Table 2, we report the performance of the different models. Our method was among the statistically significant group of best models on Parkinsons dataset according to the two-sided paired $t-test$ at significance level of 5%. On the remaining datasets, it achieved the second best performance after **NADE**.

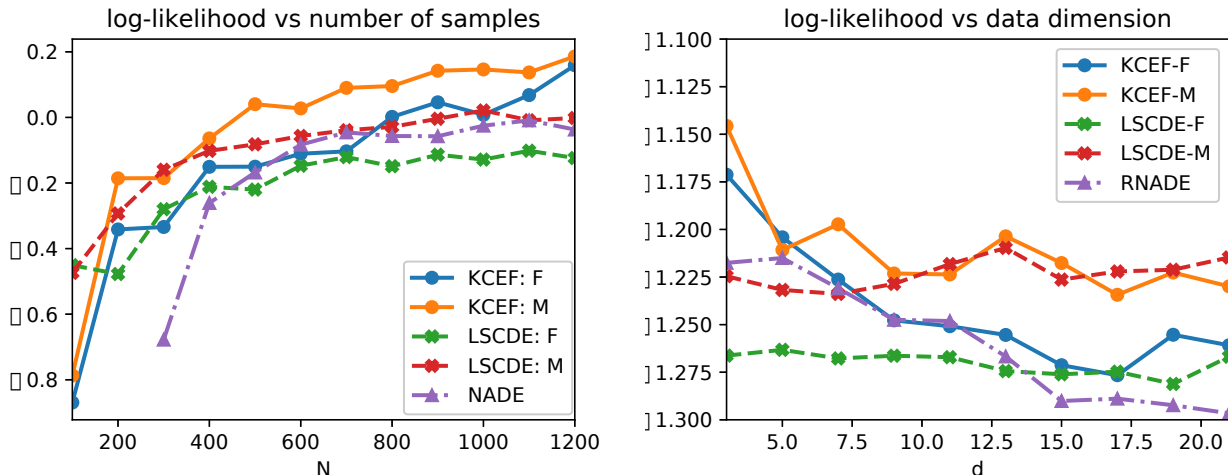**Sampling:** We compare samples generated from the approximate distribution obtained using differ-

Figure 1: Experimental comparison of proposed method KCEF and other methods (LSCDE and NADE) on synthetic *grid* dataset. **LEFT**: log-likelihood vs training samples size, $(d = 3)$. **RIGHT**: log-likelihood per dimension vs dimension, $N = 2000$. The log-likelihood is evaluated on a separate test set of size 2000.
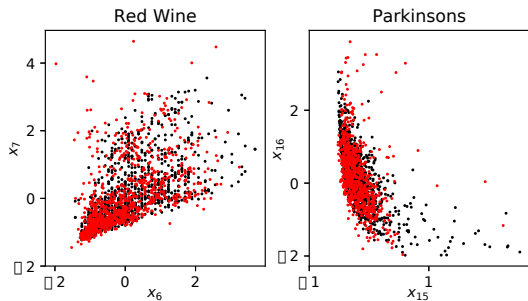

Figure 2: Scatter plot of 2-d slices of *red wine* and *parkinsons* data sets. Black points are real data, red are samples from the KCEF.

	$\mathcal{H}_{KEF<KCEF}$	$\mathcal{H}_{NADE<KCEF}$
parkinsons	0.523506	0.011467
red-wine	0.000791	0.326109

Table 3: *p*-values for the relative similarity test. Columns represents the p-values for testing whether samples from KEF (resp. KCEF) model are closer to the data than samples from the KCEF (resp. NADE).

We also performed a test of relative similarity between the generated samples and the ground truth data following the methodology and code of Bounliphone et al., 2015. Given samples from data X_m and generated samples Y_n and Z_r from two different methods, we test the hypothesis that P_x is closer to P_z than P_y according to the MMD metric. The null hypothesis

$$\mathcal{H}_{y<z} : MMD(P_x, P_y) \leq MMD(P_x, P_z)$$

is tested against the alternative at a significance level $\alpha = 5\%$ (see Bounliphone et al., 2015 for details). Table 3 shows the p-value for testing **KCEF** vs **KEF** and **NADE** vs **KCEF**. We see that **KCEF** significantly outperforms **NADE** with high confidence for the *parkinsons* dataset, consistent with Table 2. Performance of the two methods is not statistically distinguishable for the *red-wine* data. See the scatter plots in Figure 4 of Appendix E, which visually confirm the result. **KCEF** gives significantly better samples than **KEF** on *red-wine*: indeed, **KCEF** generally outperforms **KEF** on distributions where the densities exhibit abrupt transitions, as is clear by inspection of the plots in Figure 4 of Appendix E.

ent methods (**KEF**, **KCEF**, **NADE**). To get samples $(X_1, ..., X_d)$ from the joint distribution of **KCEF** we performed ancestral sampling, where a sample from the parents $\pi(i)$ of node i is first generated, and then X_i is sampled according to $p(X_i|X_{\pi(i)})$. We used the methodology and code in Strathmann et al., 2015 to sample from each conditional distribution $p(X_i|X_{\pi(i)})$ using an HMC proposal, since we have access to the gradient of the conditional densities and their un-normalized values. We trained the 3 models on Red Wine and Parkinsons datasets as described previously, and generated joint samples from two-dimensional slices of data (see Figure 2). Since each conditional distribution is low-dimensional, we assumed an idealized scenario where the burn-in is completed after 100 iterations of the HMC sampler. We then run 20 samplers for 1000 and thin by a factor 10, which results in 2000 samples. As shown in Figure 2, **KCEF** is able to capture challenging properties of the target distribution, such as heteroscedasticity and sharp thresholds.

References

Bach, F., S. Lacoste-Julien, and G. Obozinski (2012). "On the equivalence between herding and conditional gradient algorithms." In: *ICML*.

Barron, A. and C-H. Sheu (1991). "Approximation of density functions by sequences of exponential families." In: *Annals of Statistics* 19.3, pp. 1347–1369.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Bounliphone, Wacha, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou, and Arthur Gretton (2015). "A Test of Relative Similarity For Model Selection in Generative Models." In: *CoRR*. eprint: 1511.04581. URL: https://arxiv.org/abs/1511.04581.

Canu, Stephane and Alex J. Smola (2006). "Kernel methods and the exponential family." In: *Neurocomputing* 69.7, pp. 714–720.

Caponnetto, A. and E. De Vito (2007). "Optimal Rates for the Regularized Least-Squares Algorithm." In: *Foundations of Computational Mathematics* 7.3, pp. 331–368. URL: http://dx.doi.org/10.1007/s10208-006-0196-8.

Chen, Y., M. Welling, and A. Smola (2010). "Super-samples from Kernel-Herding." In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Fan, Jianqing, Qiwei Yao, and Howell Tong (1996). "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems." In: *Biometrika* 6704. URL: https://ideas.repec.org/p/ehl/lserod/6704.html.

Fukumizu, Kenji (2009). "Exponential manifold by reproducing kernel Hilbert spaces." In: *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, pp. 291–306.

Grunewalder, S., G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil (2012). "Conditional mean embeddings as regressors." In: *International Conference on Machine Learning (ICML)*.

Gu, C. and C. Qiu (1993). "Smoothing spline density estimation: Theory." In: *Annals of Statistics* 21.1, pp. 217–234.

Hall, Peter, Rodney C. L. Wolff, and Qiwei Yao (1999). "Methods for Estimating a Conditional Distribution Function." In: *Journal of the American Statistical Association* 94.445, pp. 154–163. URL: http://www.jstor.org/stable/2669691.

Hyvärinen, Aapo (2005). "Estimation of non-normalized statistical models by score matching." In: *Journal of Machine Learning Research* 6, pp. 695–709.

Jordan, Michael I. (1999). *Learning in Graphical Models*. Ed. by Michael I. Jordan. Cambridge, MA, USA: MIT Press.

Kanagawa, M., Y. Nishiyama, A. Gretton, and K. Fukumizu (2016). "Filtering with State-Observation Examples via Kernel Monte Carlo Filter." In: *Neural Computation* 28.2, pp. 382–444.

Micchelli, Charles A. and Massimiliano A. Pontil (2005). "On Learning Vector-Valued Functions." In: *Neural Comput.* 17.1, pp. 177–204. URL: http://dx.doi.org/10.1162/0899766052530802.

Nagler, Thomas and Claudia Czado (2016). "Evading the Curse of Dimensionality in Nonparametric Density Estimation with Simplified Vine Copulas." In: *J. Multivar. Anal.* 151.C, pp. 69–89. URL: https://doi.org/10.1016/j.jmva.2016.07.003.

Neal, Radford M. (2010). "MCMC using Hamiltonian dynamics." In: *Handbook of Markov Chain Monte Carlo*. eprint: 1206.1901. URL: https://arxiv.org/abs/1206.1901.

Pearl, J. (2001). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pistone, G. and C. Sempi (1995). "An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one." In: *Annals of Statistics* 23.5, pp. 1543–1561.

Raiko, Tapani, Li Yao, Kyunghyun Cho, and Yoshua Bengio (2014). "Iterative Neural Autoregressive Distribution Estimator (NADE-k)." In: *NIPS*. eprint: 1406.1485. URL: https://arxiv.org/abs/1406.1485.

Rasmussen, C.E. (2003). "Gaussian Processes to Speed up Hybrid Monte Carlo for Expensive Bayesian Integrals." In: *Bayesian Statistics 7*, pp. 651–659.

Mixed Cumulative Distribution Networks (2011). eprint: 1008.5386. URL: https://arxiv.org/abs/1008.5386.

Song, L., B. Boots, S. Siddiqi, G. Gordon, and A. Smola (2010). "Hilbert Space Embeddings of Hilbert Markov Models." In: *International Conference on Machine Learning (ICML)*.

Sriperumbudur, Bharath, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar (2017). "Density Estimation in Infinite Dimensional Exponential Families." In: *Journal of Machine Learning Research* 18.57, pp. 1–59. URL: http://jmlr.org/papers/v18/16-011.html.

Sriperumbudur, Bharath, Arthur Gretton, Kenji Fukumizu, Bernhard Schoelkopf, and Gert Lanckriet (2010). "Hilbert Space Embeddings and Metrics on Probability Measures." In: *Journal of Machine Learning Research* 11, pp. 1517–1561.

Strathmann, Heiko, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, and Arthur Gretton (2015). "Gradient-free Hamiltonian Monte Carlo with Effi-

cient Kernel Exponential Families." In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., pp. 955–963. URL: http://papers.nips.cc/paper/5890-gradient-free-hamiltonian-monte-carlo-with-efficient-kernel-exponential-families.pdf.

Sugiyama, Masashi, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara (2010). "Least-Squares Conditional Density Estimation." In: *IEICE Transactions* 93-D.3, pp. 583–594. URL: http://search.ieice.org/bin/summary.php?id=e93-d_3_583.

Sutherland, Dougal J., Heiko Strathmann, Michael Arbel, and Arthur Gretton (2017). "Efficient and principled score estimation." In: eprint: 1705.08360. URL: https://arxiv.org/abs/1705.08360.

Takeuchi, Ichiro, Quoc V. Le, Timothy D. Sears, and Alexander J. Smola (2006). "Nonparametric Quantile Estimation." In: *J. Mach. Learn. Res.* 7, pp. 1231–1264. URL: http://dl.acm.org/citation.cfm?id=1248547.1248592.

Takeuchi, Ichiro, Kaname Nomura, and Takafumi Kanamori (2009). "Nonparametric Conditional Density Estimation Using Piecewise-linear Solution Path of Kernel Quantile Regression." In: *Neural Comput.* 21.2, pp. 533–559. URL: http://dx.doi.org/10.1162/neco.2008.10-07-628.

Tang, Yichuan, Ruslan Salakhutdinov, and Geoffrey Hinton (2012). "Deep Mixtures of Factor Analysers." In: *Proceedings of the 29th International Conference on Machine Learning, 2012, Edinburgh, Scotland.* URL: https://arxiv.org/abs/1206.4635.

Team, R Development Core (2008). *The R Manuals.*

Uria, Benigno, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle (2016). "Neural Autoregressive Distribution Estimation." In: *Journal of Machine Learning Research* 17. eprint: 1605.02226. URL: https://arxiv.org/abs/1605.02226.

Uria, Benigno, Iain Murray, and Hugo Larochelle (2013). "RNADE: The real-valued neural autoregressive density-estimator." In: eprint: 1306.0186. URL: https://arxiv.org/abs/1306.0186.

Zhang, Chong, Yufeng Liu, and Yichao Wu (2016). "On Quantile Regression in Reproducing Kernel Hilbert Spaces with the Data Sparsity Constraint." In: *Journal of Machine Learning Research* 17.40, pp. 1–45. URL: http://jmlr.org/papers/v17/zhang16a.html.