

Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks

Jonas Rothfuss ^{*†} Fabio Ferreira^{*†} Simon Walther[†] Maxim Ulrich[†]

ABSTRACT

Given a set of empirical observations, conditional density estimation aims to capture the statistical relationship between a conditional variable \mathbf{x} and a dependent variable \mathbf{y} by modeling their conditional probability $p(\mathbf{y}|\mathbf{x})$. The paper develops best practices for conditional density estimation for finance applications with neural networks, grounded on mathematical insights and empirical evaluations. In particular, we introduce a noise regularization and data normalization scheme, alleviating problems with over-fitting, initialization and hyper-parameter sensitivity of such estimators. We compare our proposed methodology with popular semi- and non-parametric density estimators, underpin its effectiveness in various benchmarks on simulated and Euro Stoxx 50 data and show its superior performance. Our methodology allows to obtain high-quality estimators for statistical expectations of higher moments, quantiles and non-linear return transformations, with very little assumptions about the return dynamic.

^{*}denotes equal contribution

[†]The authors are with the Computational Risk and Asset Management Research Group at the Karlsruhe Institute of Technology (KIT), Germany. Reference to jonas.rothfuss@gmail.com

I. Introduction

A wide range of problems in econometrics and finance are concerned with describing the statistical relationship between a vector of explanatory variables \mathbf{x} and a dependent variable or vector \mathbf{y} of interest. While regression analysis aims to describe the conditional mean $\mathbb{E}[\mathbf{y}|\mathbf{x}]$, many problems in risk and asset management require gaining insight about deviations from the mean and their associated likelihood. The stochastic dependency of \mathbf{y} on \mathbf{x} can be fully described by modeling the conditional probability density $p(\mathbf{y}|\mathbf{x})$. Inferring such a density function from a set of empirical observations $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ is typically referred to as conditional density estimation (CDE).

We propose to use neural networks for estimating conditional densities. In particular, we discuss two models in which a neural network controls the parameters of a Gaussian mixture. Namely, these are the Mixture Density Network (MDN) by Bishop (1994) and the Kernel Mixture Network (KMN) by Ambrogioni et al. (2017). When chosen expressive enough, such models can approximate arbitrary conditional densities.

However, when combined with maximum likelihood estimation, this flexibility can result in overfitting and poor generalization beyond the training data. Addressing this issue, we develop a noise regularization method for conditional density estimation. By adding small random perturbations to the data during training, the conditional density estimate is smoothed and generalizes better. In fact, we mathematically derive that adding noise during training is equivalent to penalizing the second derivatives of the conditional log-probability. Graphically, the penalization punishes very curved or even spiky density estimators in favor of smoother variants. Our experimental results demonstrate the efficacy and importance of the noise regularization for attaining good out-of-sample performance.

Moreover, we attend to further practical issues that arise due to different value ranges of the training data. In this context, we introduce a simple data normalization scheme, that fits the conditional density model on normalized data, and, after training, transforms the density estimate, so that it corresponds to the original data distribution. The normalization scheme makes the hyperparameters and initialization of the neural network based density estimator insensitive to differing value ranges. Our empirical evaluations suggest that this increases the consistency of the training results and significantly improves the estimator's performance.

Aiming to compare our proposed approach against well-established CDE methods, we report a comprehensive benchmark study on simulated densities as well as on EuroStoxx 50 returns. When trained with noise regularization, both MDNs and KMNs are able to outperform previous standard semi- and nonparametric conditional density estimators. Moreover, the results suggest that even for small sample sizes, neural network based conditional density estimators can be an equal or superior alternative to well established conditional kernel density estimators.

Our study adds to the econometric literature, which discusses two main approaches towards CDE. The majority of financial research assumes that the conditional distribution follows a standard parametric family (e.g. Gaussian) that captures the dependence of the distribution parameters on \mathbf{x} with a (partially) linear model. The widely used ARMA-GARCH time-series model (Engle, 1982;

Nelson and Cao, 1992) and many of its extensions (Glosten et al., 1993; Hansen et al., 1994; Sentana, 1995) fall into this category. However, inherent assumptions in many such models have been refuted empirically later on (Harvey and Siddique, 1999; Jondeau and Rockinger, 2003). Another example for this class of models are linear factor models (Fama and French, 1993; Carhart, 1997; Fama and French, 2015). Here, too, evidence for time variation in the betas of these factor models, as documented by Jagannathan and Wang (1996), Lewellen and Nagel (2006) or Gormsen and Jensen (2017), cast doubt about the actual existence of the stated linear relationships. Overall, it is unclear to which degree the modelling restrictions are consistent with the actual mechanisms that generate the empirical data and how much they bias the inference.

Another major strand of research approaches CDE from a nonparametric perspective, estimating the conditional density with kernel functions, centered in the data points (Hyndman et al., 1996; Li and Racine, 2007). While kernel methods make little assumptions about functional relationships and density shape, they typically suffer from poor generalization in the tail regions and from data sparseness when dimensionality is high.

In contrast, CDE based on high-capacity function approximators such as neural networks has received little attention in the econometric and finance community. Yet, they combine the global generalization capabilities of parametric models with little restrictive assumptions regarding the conditional density. Aiming to combine these two advantages, this work studies the use of neural networks for estimating conditional densities. Overall, this paper establishes a sound framework for fitting high-capacity conditional density models. Thanks to the presented noise regularization and data normalization scheme, we are able to overcome common issues with neural network based estimators and make the approach easy to use. The conditional density estimators are available as open-source python package.¹

II. Background

A. Density Estimation

Let X be a random variable with probability density function (PDF) $p(x)$ defined over the domain \mathcal{X} . When investigating phenomena in the real world, the distribution of an observable variable X is typically unknown. However, it is possible to observe realizations $\mathbf{x}_n \sim p(\mathbf{x})$ of X . Given a collection $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of such observations, it is our aim to find a good estimate $\hat{p}(x)$ of the true density function p . Typically, the goodness of a fitted distribution \hat{p} is measured by a statistical divergence between the estimated \hat{p} and the true density function p . Throughout the density estimation literature (Bishop, 2006; Li and Racine, 2007; Shalizi, 2011) the most common criteria are integrated mean squared error (IMSE) and the Kullback-Leibler divergence.

In its most general form, density estimation aims to find the best \hat{p} among all possible PDFs over the domain \mathcal{X} , while only given a finite number of observations. Even in the simple case $\mathcal{X} = \mathbb{R}^1$, this would require estimating infinitely many distribution parameters with a finite amount

¹https://github.com/freelunchtheorem/Conditional_Density_Estimation

of data, which is not feasible in practice. Hence, it is necessary to either restrict the space of possible PDFs or to embed other assumptions into the density estimation. The kind of imposed assumptions characterize the distinction between the sub-fields of parametric and non-parametric density estimation.

A.1. Parametric Density Estimation

In parametric estimation, the PDF \hat{p} is assumed to belong to a parametric family $\mathcal{F} = \{\hat{p}_\theta(\cdot) | \theta \in \Theta\}$ where the density function is described by a finite dimensional parameter $\theta \in \Theta$. A classical example of \mathcal{F} is the family of univariate normal distributions $\{\mathcal{N}(\cdot | \mu, \sigma) | (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$.

The standard method for estimating θ is *maximum likelihood estimation*, wherein θ^* is chosen so that the likelihood of the data \mathcal{D} is maximized:

$$\theta^* = \arg \max_{\theta} \prod_{n=1}^N \hat{p}_\theta(\mathbf{x}_n) = \arg \max_{\theta} \sum_{n=1}^N \log \hat{p}_\theta(\mathbf{x}_n) \quad (1)$$

In practice, the optimization problem is restated as maximizing the sum of log-probabilities which is equivalent to minimizing the Kullback-Leibler divergence between the empirical data distribution $p_{\mathcal{D}}(x) = \frac{1}{N} \sum_{n=1}^N \delta(\|\mathbf{x} - \mathbf{x}_n\|)$ (i.e. point mass in the observations \mathbf{x}_n) and the parametric distribution \hat{p}_θ :

$$\theta^* = \arg \min_{\theta} \mathcal{D}_{KL}(p_{\mathcal{D}} || \hat{p}_\theta) \quad (2)$$

For further details on parametric density estimation, we refer to Bishop (2006) page 57.

A.2. Nonparametric Density Estimation

In contrast to parametric methods, nonparametric density estimators do not explicitly restrict the space of considered PDFs. The most popular nonparametric method, kernel density estimation (KDE), places a symmetric density function $K(z)$, the so-called kernel, in each training data point \mathbf{x}_n (Rosenblatt, 1956; Parzen, 1962). The resulting density estimate for univariate distributions is formed as equally weighted mixture of the N densities centered in the data points. In the case of multivariate kernel density estimation, i.e. $\dim(\mathcal{X}) = l > 1$, the density can be estimated as product of marginal kernel density estimates. Such a kernel density estimate reads as follows:

$$\hat{p}(\mathbf{x}) = \prod_{j=1}^l \hat{p}(x^{(j)}) = \prod_{j=1}^l \frac{1}{Nh^{(j)}} \sum_{n=1}^N K\left(\frac{x^{(j)} - x_n^{(j)}}{h^{(j)}}\right) \quad (3)$$

In that, $x^{(j)}$ denotes the j -th element of the column vector $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^l$ and $h^{(j)}$ the bandwidth / smoothing parameter corresponding to the j -th dimension.

One popular choice of $K(\cdot)$ is the Gaussian kernel: $K(z) = (2\pi)^{-\frac{1}{2}} e^{-\frac{z^2}{2}}$. Other common choices of $K(\cdot)$ are the Epanechnikov and exponential kernels. Provided a continuous kernel function, the

estimated PDF in (3) is continuous. Beyond the appropriate choice of $K(\cdot)$, a central challenge is the selection of the bandwidth parameter h which controls the smoothing of the estimated PDF. For details on bandwidth selection, we refer the interested reader to Li and Racine (2007).

B. Conditional Density Estimation (CDE)

Let (X, Y) be a pair of random variables with respective domains $\mathcal{X} \subseteq \mathbb{R}^l$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ and realizations \mathbf{x} and \mathbf{y} . Let $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}, \mathbf{y})/p(\mathbf{x})$ denote the conditional probability density of \mathbf{y} given \mathbf{x} . Typically, Y is referred to as a dependent variable (i.e. explained variable) and X as conditional (explanatory) variable. Given a dataset of observations $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ drawn from the joint distribution $(\mathbf{x}_n, \mathbf{y}_n) \sim p(\mathbf{x}, \mathbf{y})$, the aim of conditional density estimation (CDE) is to find an estimate $\hat{p}(\mathbf{y}|\mathbf{x})$ of the true conditional density $p(\mathbf{y}|\mathbf{x})$.

In the context of conditional density estimation, the Kullback-Leibler divergence objective is expressed as expectation over $p(\mathbf{x})$:

$$\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [\mathcal{D}_{KL}(p(\mathbf{y}|\mathbf{x}) || \hat{p}(\mathbf{y}|\mathbf{x}))] = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})} [\log p(\mathbf{y}|\mathbf{x}) - \log \hat{p}(\mathbf{y}|\mathbf{x})] \quad (4)$$

Similar to the unconditional case in (1) - (2), parametric maximum likelihood estimation following from (4) can be expressed as

$$\theta^* = \arg \max_{\theta} \sum_{n=1}^N \log \hat{p}_{\theta}(\mathbf{y}_n | \mathbf{x}_n) \quad (5)$$

Given a dataset \mathcal{D} drawn i.i.d from $p(\mathbf{x}, \mathbf{y})$, (5) can be viewed as equivalent to minimizing a monte-carlo estimate of the expectation in (4). The nonparametric KDE approach, discussed in Section II.A.2 can be extended to the conditional case. Typically, unconditional KDE is used to estimate both the joint density $\hat{p}(\mathbf{x}, \mathbf{y})$ and the marginal density $\hat{p}(\mathbf{x})$. Then, the conditional density estimate follows as the density ratio

$$\hat{p}(\mathbf{y}|\mathbf{x}) = \frac{\hat{p}(\mathbf{x}, \mathbf{y})}{\hat{p}(\mathbf{x})} \quad (6)$$

where both the numerator and denominator are the sums of Kernel functions as in (3). For more details on CDE, we refer the interested reader to Li and Racine (2007).

III. Related Work

This chapter discusses related work in the areas of finance, econometrics and machine learning. In that, we use the the differentiation between parametric and non-parametric methods, as discussed in Section II. In particular, we organize the following review in three categories: 1) parametric conditional density and time-series models with narrowly defined parametric families 2) non-parametric density estimation and 3) parametric models based on high-capacity function approximators such as neural networks.

Parametric CDE in finance and econometrics. The majority of work in finance and econo-

metrics uses a standard parametric family to model the conditional distribution of stock returns and other instruments. Typically, a Gaussian distribution is employed, whereby the parameters of the conditional distribution are predicted with time-series models. A popular instantiation of this category is the ARMA-GARCH method which models mean and variance of the conditional Gaussian through linear relationships (Engle, 1982; Hamilton, 1994). Various generalizations of GARCH attempt to model asymmetric return distributions and negative skewness (Nelson and Cao, 1992; Glosten et al., 1993; Sentana, 1995). Further work employs the student-t distribution as conditional probability model (Bollerslev et al., 1987; Hansen et al., 1994) and models the dependence of higher-order moments on the past (Gallant et al., 1991; Hansen et al., 1994).

Though the neural network based CDE approaches, which are presented in this paper, are also parametric models, they make very little assumptions about the underlying relationships and density family. Both the relationship between the conditional variables and distribution parameters, as well as the probability density itself are modelled with flexible function classes (i.e. neural network and GMM). In contrast, traditional financial models impose strong assumptions such as linear relationships and Gaussian conditional distributions. It is unclear to which degree such modelling restrictions are consistent with the empirical data and how much they bias the inference.

Non-parametric CDE. A distinctly different line of work in econometrics aims to estimate densities in a non-parametric manner. Originally introduced in Rosenblatt (1956); Parzen (1962), KDE uses kernel functions to estimate the probability density at a query point, based on the distance to all training points. In principle, kernel density estimators can approximate arbitrary probability distributions without parametric assumptions. However, in practice, data is finite and smoothing is required to achieve satisfactory generalization beyond the training data. The fundamental issue of KDE, commonly referred to as the bandwidth selection problem, is choosing the appropriate amount of smoothing (Park et al., 1990; Cao et al., 1994). Common bandwidth selection methods include rules-of-thumb (Silverman, 1982; Sheather and Jones, 1991; Botev et al., 2010) and selectors based on cross-validation (Rudemo, 1982; Bowman, 1984; Hall et al., 1992).

In order to estimate conditional probabilities, previous work proposes to estimate both the joint and marginal probability separately with KDE and then computing the conditional probability as their ratio (Hyndman et al., 1996; De Gooijer and Zerom, 2003; Li and Racine, 2007). Other approaches combine non-parametric elements with parametric elements (Tresp, 2001; Sugiyama and Takeuchi, 2010), forming semi-parametric conditional density estimators. Despite their theoretical appeal, non-parametric density estimators suffer from the following drawbacks: First, they tend to generalize poorly in regions where data is sparse which especially becomes evident in the tail regions of the distribution. Second, their performance deteriorates quickly as the dimensionality of the dependent variable increases. This phenomenon is commonly referred to as the "curse of dimensionality".

CDE with neural networks: MDN, KDE, Normalizing Flows. The third line work approaches conditional density estimation from a parametric perspective. However, in contrast to parametric modelling in finance and econometrics, such methods use high-capacity function

approximators instead of strongly constrained parametric families. Our work builds upon the work of Bishop (1994) and Ambrogioni et al. (2017), who propose to use a neural network to control the parameters of mixture density model. When both the neural network and the mixture of densities is chosen to be sufficiently expressive, any conditional probability distribution can be approximated (Hornik, 1991; Li and Andrew Barron, 2000). Sarajedini et al. (1999) propose neural networks that parameterize a generic exponential family distribution. However, this limits the overall expressiveness of the conditional density estimator.

A recent trend in machine learning is the use of neural network based latent density models (Mirza and Osindero, 2014; Sohn et al., 2015). Although such methods have been shown successful for estimating distributions of images, it is not possible to recover the PDF of such latent density models. More promising in this sense are normalizing flows which use a sequence of invertible maps to transform a simple latent distribution into more complex density functions (Rezende and Mohamed, 2015; Dinh et al., 2017; Trippe and Turner, 2018). Since the PDF of normalizing flows is tractable, this could be an interesting direction to supplement our work.

While neural network based density estimators make very little assumptions about the underlying density, they suffer from severe over-fitting when trained with the maximum likelihood objective. In order to counteract over-fitting, various regularization methods have been explored in the literature (Krogh and Hertz, 1992; Holmstrom and Koistinen, 1992; Webb, 1994; Srivastava et al., 2014). However, these methods were developed with emphasis on regression and classification problems. Our work focuses on the regularization of neural network based density estimators. In that, we make use of the noise regularization framework (Webb, 1994; Bishop, 1995), discussing its implications in the context of density estimation and empirically evaluating its efficacy.

IV. Conditional Density Estimation with Neural Networks

The following chapter introduces and discusses two neural network based approaches for estimating conditional densities. Both density estimators are, in their nature, parametric models, but exhibit substantially higher flexibility than traditional parametric methods. In the first part, we formally define the density models and explain their fitting process. The second part of this chapter attends to the challenges that arise from this flexibility, introducing a form of smoothness regularization to combat over-fitting and enable good generalization.

A. The Density Models

A.1. Mixture Density Networks

Mixture Density Networks (MDNs) combine conventional neural networks with a mixture density model for the purpose of estimating conditional distributions $p(\mathbf{y}|\mathbf{x})$ (Bishop, 1994). In particular, the parameters of the unconditional mixture distribution $p(\mathbf{y})$ are outputted by the neural network, which takes the conditional variable \mathbf{x} as input.

For our purpose, we employ a Gaussian Mixture Model (GMM) with diagonal covariance matrices as density model. The conditional density estimate $\hat{p}(\mathbf{y}|\mathbf{x})$ follows as weighted sum of K Gaussians

$$\hat{p}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^K w_k(\mathbf{x}; \theta) \mathcal{N}(\mathbf{y}|\mu_k(\mathbf{x}; \theta), \sigma_k^2(\mathbf{x}; \theta)) \quad (7)$$

wherein $w_k(\mathbf{x}; \theta)$ denote the weight, $\mu_k(\mathbf{x}; \theta)$ the mean and $\sigma_k^2(\mathbf{x}; \theta)$ the variance of the k -th Gaussian component. All the GMM parameters are governed by the neural network with parameters θ and input \mathbf{x} . It is possible to use a GMM with full covariance matrices Σ_k by having the neural network output the lower triangular entries of the respective Cholesky decompositions $\Sigma_k^{1/2}$ (Tansey et al., 2016). However, we choose diagonal covariance matrices in order to avoid the quadratic increase in the neural network’s output layer size as the dimensionality of \mathcal{Y} increases.

The mixing weights $w_k(\mathbf{x}; \theta)$ must resemble a categorical distribution, i.e. it must hold that $\sum_{k=1}^K w_k(\mathbf{x}; \theta) = 1$ and $w_k(\mathbf{x}; \theta) \geq 0 \forall k$. To satisfy the conditions, the softmax function is used.

$$w_k(\mathbf{x}) = \frac{\exp(a_k^w(\mathbf{x}))}{\sum_{i=1}^K \exp(a_i^w(\mathbf{x}))} \quad (8)$$

In that, $a_k^w(\mathbf{x}) \in \mathbb{R}$ denote the logit scores emitted by the neural network. Similarly, the standard deviations $\sigma_k(\mathbf{x})$ must be positive. To ensure that the respective neural network satisfy the non-negativity constraint, a softplus non-linearity is applied:

$$\sigma_k(\mathbf{x}) = \log(1 + \exp(a_k^\sigma(\mathbf{x}))) \quad (9)$$

Since the component means $\mu_k(\mathbf{x}; \theta)$ are not subject to such restrictions, we use a linear layer without non-linearity for the respective output neurons.

A.2. Kernel Mixture Networks

While MDNs resemble a purely parametric conditional density model, a closely related approach, the Kernel Mixture Network (KMN), combines both non-parametric and parametric elements (Ambrogioni et al., 2017). Similar to MDNs, a mixture density model of $\hat{p}(y)$ is combined with a neural network which takes the conditional variable \mathbf{x} as an input. However, the neural network only controls the weights of the mixture components while the component centers and scales are fixed w.r.t. to \mathbf{x} . Figuratively, one can imagine the neural network as choosing between a very large amount of pre-existing kernel functions to build up the final combined density function. As common for non-parametric density estimation, the components/kernels are placed in each of the training samples or a subset of the samples. For each of the kernel centers, one or multiple scale/bandwidth parameters σ_m are chosen. As for MDNs, we employ Gaussians as mixture components, wherein the scale parameter directly coincides with the standard deviation.

Let K be the number of kernel centers μ_k and M the number of different kernel scales σ_m . The KMN conditional density estimate reads as follows:

$$\hat{p}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^K \sum_{m=1}^M w_{k,m}(\mathbf{x}; \theta) \mathcal{N}(\mathbf{y}|\mu_k, \sigma_m^2) \quad (10)$$

As previously, the weights $w_{k,m}$ correspond to a softmax function. Ambrogioni et al. (2017) propose to choose the kernel centers μ_k by subsampling the training data by recursively removing each point y_n that is closer than a constant δ to any of its predecessor points. This can be seen as a naive form of clustering which depends on the ordering of the dataset. Instead, we suggest to use a well-established clustering method such as K-means for selecting the kernel centers. The scales of the Gaussian kernels can either be fixed or jointly trained with the neural network weights. In practice, considering the scales $\{\sigma_m\}_{m=1}^M$ as trainable parameters consistently improves the performance.

Overall, the KMN model is more restrictive than MDN as the locations and scales of the mixture components are fixed during inference and cannot be controlled by the neural network. However, due to the reduced flexibility of KMNs, they are less prone to over-fit than MDNs.

B. Fitting the Density Models

The parameters θ of the neural network are fitted through maximum likelihood estimation. In practice, we minimize the negative conditional log-likelihood of the training data $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$:

$$\theta^* = \arg \min_{\theta} - \sum_{n=1}^N \log p_{\theta}(\mathbf{y}_n|\mathbf{x}_n) \quad (11)$$

In that, the negative log-likelihood in (11) is minimized via numerical optimization. Due to its superior performance in non-convex optimization problems, we employ stochastic gradient descent in conjunction with the adaptive learning rate method [Adam](#) (Kingma and Ba, 2015).

B.1. Variable Noise as Smoothness Regularization

A central issue when training high capacity function approximators such as neural networks is determining the optimal degree of complexity of the model. Models with too limited capacity may not be able to sufficiently capture the structure of the data, inducing a strong restriction bias. On the other hand, if a model is too expressive it is prone to over-fit the training data, resulting in poor generalization. This problem can be regarded as finding the right balance when trading off variance against inductive bias. There exist many techniques allowing to control the trade-off between bias and variance, involving various forms of regularization and data augmentation. For an overview of regularization techniques, the interested reader is referred to Kukačka et al. (2017).

Adding noise to the data during training can be viewed as form of data augmentation and regularization that biases towards smooth functions (Webb, 1994; Bishop, 1994). In the domain of finance, assuming smooth return distribution is a reasonable assumption. Hence, it is desirable

to embed an inductive bias towards smoothness into the learning procedure in order to reduce the variance. Specifically, we add small perturbances in form of a random vector $\boldsymbol{\xi} \sim q(\boldsymbol{\xi})$ to the data $\tilde{\mathbf{x}}_n = \mathbf{x}_n + \boldsymbol{\xi}_x$ and $\tilde{\mathbf{y}}_n = \mathbf{y}_n + \boldsymbol{\xi}_y$. Further, we assume that the noise is zero centered as well as identically and independently distributed among the dimensions, with standard deviation η :

$$\mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}] = 0 \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi} \boldsymbol{\xi}^\top] = \eta^2 I \quad (12)$$

Before discussing the particular effects of randomly perturbing the data during conditional maximum likelihood estimation, we first analyze noise regularization in a more general case. Let $\mathcal{L}_{\mathcal{D}}(\mathcal{D})$ be a loss function over a set of data points $\mathcal{D} = \{x_1, \dots, x_N\}$, which can be partitioned into a sum of losses corresponding to each data point x_n :

$$\mathcal{L}_{\mathcal{D}}(\mathcal{D}) = \sum_{n=1}^N \mathcal{L}(\mathbf{x}_n) \quad (13)$$

The loss $\mathcal{L}(\mathbf{x}_n + \boldsymbol{\xi})$, resulting from adding random perturbations can be approximated by a second order Taylor expansion around x_n

$$\mathcal{L}(\mathbf{x}_n + \boldsymbol{\xi}) = \mathcal{L}(\mathbf{x}_n) + \boldsymbol{\xi}^\top \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n} + \frac{1}{2} \boldsymbol{\xi}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n} \boldsymbol{\xi} + \mathcal{O}(\boldsymbol{\xi}^3) \quad (14)$$

Assuming that the noise $\boldsymbol{\xi}$ is small in its magnitude, $\mathcal{O}(\boldsymbol{\xi}^3)$ is neglectable. Using the assumption about $\boldsymbol{\xi}$ in (12), the expected loss can be written as

$$\mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\mathcal{L}(\mathbf{x}_n + \boldsymbol{\xi})] \approx \mathcal{L}(\mathbf{x}_n) + \frac{1}{2} \mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}^\top \mathbf{H}^{(n)} \boldsymbol{\xi}] = \mathcal{L}(\mathbf{x}_n) + \frac{\eta^2}{2} \text{tr}(\mathbf{H}^{(n)}) \quad (15)$$

where $\mathcal{L}(\mathbf{x}_n)$ is the loss without noise and $\mathbf{H}^{(n)} = \frac{\partial^2 \mathcal{L}}{\partial x^2}(\mathbf{x})|_{\mathbf{x}_n}$ the Hessian of \mathcal{L} w.r.t x , evaluated at x_n . This result has been obtained earlier by Webb (1994) and Bishop (1994). See Appendix.A for derivations.

Previous work (Webb, 1994; Bishop, 1994; An, 1996) has introduced noise regularization for regression and classification problems. However, to our best knowledge, noise regularization has not been used in the context of parametric density estimation. In the following, we derive and analyze the effect of noise regularization w.r.t. maximum likelihood estimation of conditional densities.

When concerned with maximum likelihood estimation of a conditional density $p_\theta(y|x)$, the loss function coincides with the negative conditional log-likelihood $\mathcal{L}(y_n, x_n) = -\log p(y_n|x_n)$. Let the standard deviation of the additive data noise $\boldsymbol{\xi}_x, \boldsymbol{\xi}_y$ be η_x and η_y respectively. Maximum likelihood

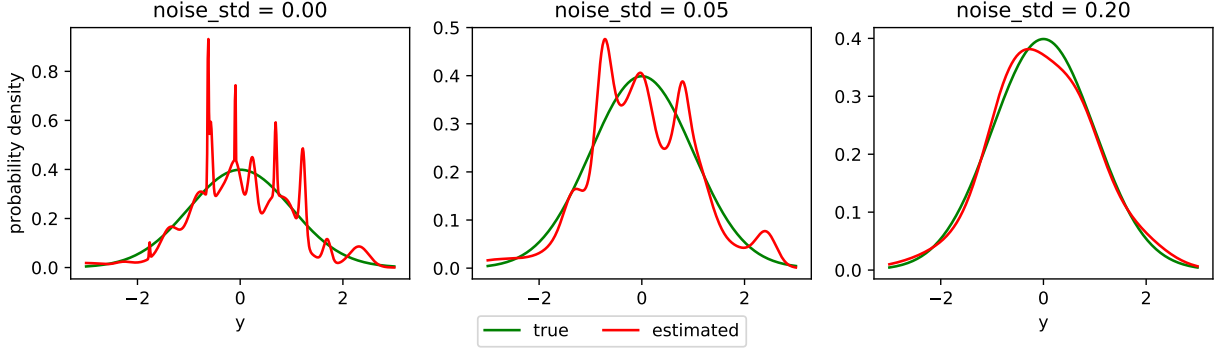


Figure 1. Effect of noise regularization on density estimate. Conditional MDN density estimate (red) and true conditional density (green) for different noise regularization intensities η_x, η_y . The MDN has been fitted with 3000 samples.

estimation (MLE) with data noise is equivalent to minimizing the loss

$$\mathcal{L}(\mathcal{D}) \approx - \sum_{n=1}^N \log p_{\theta}(\mathbf{y}_n | \mathbf{x}_n) + \sum_{n=1}^N \frac{\eta_y^2}{2} \text{tr}(\mathbf{H}_{\mathbf{y}}^{(n)}) + \sum_{n=1}^N \frac{\eta_x^2}{2} \text{tr}(\mathbf{H}_{\mathbf{x}}^{(n)}) \quad (16)$$

$$= - \sum_{n=1}^N \log p_{\theta}(\mathbf{y}_n | \mathbf{x}_n) - \frac{\eta_y^2}{2} \sum_{n=1}^N \sum_{j=1}^m \frac{\partial^2 \log p_{\theta}(\mathbf{y} | \mathbf{x})}{\partial y^{(j)} \partial y^{(j)}} \Big|_{\mathbf{y}=\mathbf{y}_n} - \frac{\eta_x^2}{2} \sum_{n=1}^N \sum_{j=1}^l \frac{\partial^2 \log p_{\theta}(\mathbf{y} | \mathbf{x})}{\partial x^{(j)} \partial x^{(j)}} \Big|_{\mathbf{x}=\mathbf{x}_n} \quad (17)$$

In that, the first term corresponds to the standard MLE objective while the other two terms constitute a smoothness regularization. The second term in (16) penalizes large negative second derivatives of the conditional log density estimate $\log p_{\theta}(\mathbf{y} | \mathbf{x})$ w.r.t. \mathbf{y} . As the MLE objective pushes the density estimate towards high densities and strong concavity in the data points \mathbf{y}_n , the regularization term counteracts this tendency to over-fit and overall smoothes the fitted distribution. The third term penalizes large negative second derivatives w.r.t. the conditional variable \mathbf{x} , thereby regularizing the sensitivity of the density estimate on changes in the conditional variable. This smoothes the functional dependency of $p_{\theta}(\mathbf{y} | \mathbf{x})$ on \mathbf{x} . As stated previously, the intensity of the smoothness regularization can be controlled through the standard deviation (η_x and η_y) of the perturbations.

Figure 1 illustrates the effect of the introduced noise regularization scheme on MDN density estimates. Plain maximum likelihood estimation (left) leads to strong over-fitting, resulting in a spiky distribution that generalizes poorly beyond the training data. In contrast, training with noise regularization (center and right) results in smoother density estimates that are closer to the true conditional density. In Section V.B, a comprehensive empirical evaluation demonstrates the efficacy and importance of noise regularization.

B.2. Data Normalization

In many applications of machine learning and econometrics, the value range of raw data varies widely. Significant differences in scale and range among features can lead to poor performance of many learning algorithms. When the initial distribution during training is statistically too far away from the actual data distribution, the training converges only slowly or might fail entirely. Moreover, the effect of many hyperparameters is often influenced by the value range of learning features and targets. For instance, the efficacy of noise regularization, introduced in the previous section, is susceptible to varying data ranges. In order to circumvent these and many more issues that arise due to different value ranges of the data, a common practice in machine learning is to normalize the data so that it exhibits zero mean and unit variance (Sola and Sevilla, 1997; Grus, 2015). While this practice is straightforward for classification and regression problems, such a transformation requires further consideration in the context of density estimation. The remainder of this section, elaborates on how to properly perform data normalization for estimating conditional densities. In that, we view the data normalization as change of variable and derive the respective density transformations that are necessary to recover an estimate of the original data distribution.

In order to normalize the training data \mathcal{D} which originates from $p(\mathbf{x}, \mathbf{y})$, we estimate mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ along each data dimension followed by subtracting the mean from the data points and dividing by the standard deviation.

$$\tilde{\mathbf{x}} = \text{diag}(\hat{\sigma}_x)^{-1}(\mathbf{x} - \hat{\mu}_x) \quad \text{and} \quad \tilde{\mathbf{y}} = \text{diag}(\hat{\sigma}_y)^{-1}(\mathbf{y} - \hat{\mu}_y) \quad (18)$$

The normalization operations in (18) are linear transformations of the data. Subsequently, the conditional density model is fitted on the normalized data, resulting in the estimated PDF $\hat{q}_\theta(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$.

However, when performing inference, one is interested in an unnormalized density estimate $\hat{p}_\theta(\mathbf{y}|\mathbf{x})$, corresponding to the conditional data distribution $p(\mathbf{y}|\mathbf{x})$. Thus, we have to transform the learned distribution $\hat{q}_\theta(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$ so that it coincides with $p(\mathbf{y}|\mathbf{x})$. In that, both the transformations $\mathbf{x} \rightarrow \tilde{\mathbf{x}}$ and $\mathbf{y} \rightarrow \tilde{\mathbf{y}}$ must be accounted for.

The former is straightforward: Since the neural network is trained to receive normalized inputs $\tilde{\mathbf{x}}$, it is sufficient to transform the original inputs \mathbf{x} to $\tilde{\mathbf{x}} = \text{diag}(\hat{\sigma}_x)^{-1}(\mathbf{x} - \hat{\mu}_x)$ before feeding them into the network at inference time. In order to account for the linear transformation of \mathbf{y} , we have to use the change of variable formula since the volume of the probability density is not preserved if $\sigma_y \neq 1$. The change of variable formula can be stated as follows.

THEOREM 1: *Let \tilde{Y} be a continuous random variable with probability density function $q(\tilde{y})$, and let $Y = v(\tilde{Y})$ be an invertible function of \tilde{Y} with inverse $\tilde{Y} = v^{-1}(Y)$. The probability density function $p(y)$ of Y is:*

$$p(y) = q(v^{-1}(y)) * \left| \frac{d}{dy} (v^{-1}(y)) \right| \quad (19)$$

In that, $\left| \frac{d}{dy} (v(y)) \right|$ is the determinant of the Jacobian of v which is vital for adjusting the volume of $q(v^{-1}(y))$, so that $\int p(y)dy = 1$. In case of the proposed data normalization scheme, v is

a linear function

$$v^{-1}(\mathbf{y}) = \text{diag}(\hat{\sigma}_{\mathbf{y}})^{-1}(\mathbf{y} - \hat{\mu}_{\mathbf{y}}) \quad (20)$$

and, together with (19), \hat{p}_{θ} follows as

$$\hat{p}_{\theta}(\mathbf{y}|\mathbf{x}) = |\text{diag}(\hat{\sigma}_{\mathbf{y}})^{-1}| \hat{q}_{\theta}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}) = \frac{1}{\prod_{j=1}^l \hat{\sigma}_{\mathbf{y}}^{(j)}} \hat{q}_{\theta}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}) \quad (21)$$

The above equation provides a simple method for recovering the unnormalized density estimate from the normalized mixture density $\hat{q}_{\theta}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$.

Alternatively, we can directly recover the conditional mixture parameters corresponding to $p_{\theta}(\mathbf{y}|\mathbf{x})$. Let $(\tilde{w}_k, \tilde{\mu}_k, \text{diag}(\tilde{\sigma}_k))$ be the conditional parameters of the GMM corresponding to $q(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$. Based on the change of variable formula, Theorem 2 provides a simple recipe for re-parameterizing the GMM so that it reflects the unnormalized conditional density. As special case of Theorem 2, with $\Sigma = \text{diag}(\tilde{\sigma})$ and $B = \text{diag}(\hat{\sigma}_{\mathbf{y}})$, the transformed GMM corresponding to $\hat{p}_{\theta}(\mathbf{y}|\mathbf{x})$ has the following parameters:

$$w_k = \tilde{w}_k \quad (22)$$

$$\mu_k = \hat{\mu}_{\mathbf{y}} + \text{diag}(\hat{\sigma}_{\mathbf{y}})\tilde{\mu}_k \quad (23)$$

$$\sigma_k = \text{diag}(\hat{\sigma}_{\mathbf{y}}) \tilde{\sigma}_k \quad (24)$$

THEOREM 2: Let $\mathbf{x} \in \mathbb{R}^n$ be a continuous random variable following a Gaussian Mixture Model (GMM), this is $\mathbf{x} \sim p(\mathbf{x})$ with

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mu_k, \Sigma_k) \quad (25)$$

Any linear transformation $\mathbf{z} = a + B\mathbf{x}$ of $\mathbf{x} \sim p(\mathbf{x})$ with $a \in \mathbb{R}^n$ and B being an invertible $n \times n$ matrix follows a Gaussian Mixture Model with density function

$$p(\mathbf{z}) = \sum_{k=1}^K w_k \mathcal{N}(a + B\mu_k, B\Sigma_k B^{\top}) \quad (26)$$

Proof. See Appendix.B

Overall, the training process with data normalization includes the following steps:

1. Estimate empirical unconditional mean $\hat{\mu}_{\mathbf{x}}$, $\hat{\mu}_{\mathbf{y}}$ and standard deviation $\hat{\sigma}_{\mathbf{x}}$, $\hat{\sigma}_{\mathbf{y}}$ of training data
2. Normalize the training data: $\{(\mathbf{x}_n, \mathbf{y}_n)\} \rightarrow \{(\tilde{\mathbf{x}}_n, \tilde{\mathbf{y}}_n)\}$

$$\tilde{\mathbf{x}}_n = \text{diag}(\hat{\sigma}_{\mathbf{x}})^{-1}(\mathbf{x}_n - \hat{\mu}_{\mathbf{x}}), \quad \tilde{\mathbf{y}}_n = \text{diag}(\hat{\sigma}_{\mathbf{y}})^{-1}(\mathbf{y}_n - \hat{\mu}_{\mathbf{y}}), \quad n = 1, \dots, N$$

3. Fit the conditional density model $\hat{q}_{\theta}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$ using the normalized data
4. Transform the estimated density back into the original data space to obtain $\hat{p}_{\theta}(\mathbf{y}|\mathbf{x})$. This

can be done by either

- (a) directly transforming the mixture density \hat{q}_θ with the change of variable formula in (21)
- or**
- (b) transforming the mixture density parameters outputted by the neural network according to (22)-(24)

V. Empirical Evaluation with Simulated Densities

This chapter comprises an extensive experimental study based on simulated densities. It is organized as follows: In the first part, we explain the methodology of the experimental evaluation, including the employed conditional density simulations and evaluation metric. The following sections include an evaluation of the noise regularization and the data normalization scheme, which have been introduced in section IV.B. Finally, we present a benchmark study, comparing the neural network based approaches with state-of-the art CDE methods.

A. Methodology

A.1. Density Simulation

In order to benchmark the proposed conditional density estimators and run experiments that aim to answer different sets of questions, several data generating models (simulators) are employed. The density simulations allow us to generate unlimited amounts of data, and, more importantly, compute the statistical distance between the true conditional data distribution and the density estimate. The density simulations, briefly introduced below, are inspired by financial models and exhibit properties of empirical return distributions, such as negative skewness and excess kurtosis.

- **EconDensity** (Fig. 6a): Conditional Gaussian distribution that exhibits heteroscedasticity and a mean with quadratic dependence on the conditional variable.
- **ArmaJump** (Fig. 6b): AR(1) process with a jump component, thus exhibiting negative skewness and excess kurtosis.
- **SkewNormal** (Fig. 6c): Conditional skew normal distribution (Anděl et al., 1984). The skewness, volatility and mean parameters of the conditional distribution are functions of the conditional variable.
- **GaussianMixture** (Fig. 6d): The joint $p(\mathbf{x}, \mathbf{y})$ distribution follows a GMM that can be factorized into a x - and y -component, i.e. $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})p(\mathbf{x})$. This ensures that the conditional probability $p(\mathbf{y}|\mathbf{x})$ is tractable.

A detailed description of the conditional density simulations alongside an illustration (Figure 6) of the respective probability density can be found in Appendix.C.

A.2. Evaluation Metrics

In order to assess the goodness of the estimated conditional densities, we measure the statistical distance between the estimate and the true conditional probability density. In particular, the Hellinger distance is used as evaluation metric. We choose the Hellinger Distance over other popular statistical divergences, because it is symmetric and constrained to $[0, 1]$, making it easier to interpret. Since the training data is simulated from a joint distribution $p(\mathbf{x}, \mathbf{y})$, but the density estimates $\hat{p}(\mathbf{y}|\mathbf{x})$ are conditional, we evaluate the statistical distance across different conditional values \mathbf{x} . For that, we uniformly sample 10 values for \mathbf{x} between the 10%- and 90%-percentile of $p(\mathbf{x})$, compute the respective Hellinger distance and finally average the conditional statistical distances. In all experiments, 5 different random seeds are used for the data simulation and density estimation. The reported Hellinger distances are averages over the random seeds while the translucent areas in the plots depict the standard deviation across the seeds.

B. Evaluation of Noise Regularization

This section empirically examines the noise regularization, presented in Section IV.B.1. As has been mathematically derived, the standard deviation of the variable noise controls the intensity of the smoothness regularization. Accordingly, the standard deviation parameters η_x and η_y , can be used to trade-off between bias and variance of the conditional density estimator.

In the following, we aim to empirically determine which values of η_x and η_y lead to the best density fit. For that, a grid-search on the noise regularization intensities is performed. Figure 2 holds the results in form of a heatmap. We observe that no regularization or noise regularization with low intensity leads to poor generalization of the density estimates, resulting in a high statistical distance w.r.t. the true density. In contrast, strong regularization may result in too much inductive bias, as it is the case for the ArmaJump simulation. The configuration $\eta_x = 0.2, \eta_y = 0.1$ leads to low statistical distances in all density simulations. Hence, we choose it as default configuration for further experiments.

Figure 3 compares the goodness of KMN/MDN density estimates with and without noise regularization across different sizes of the training set. Consistent with Figure 2, the noise regularization reduces the Hellinger distance by a significant margin. As the number of training samples increases, the estimator is less prone to over-fit and the positive effect of noise regularization tends to become smaller. Overall, the experimental results underpin the efficacy of noise regularization and its importance for achieving good generalization beyond the training samples.

C. Evaluation of Data Normalization

The data normalization scheme, introduced in section IV.B.2, aims to make the hyperparameters of the density estimator invariant to the distribution of the training data. If no data normalization is used, the goodness of the hyperparameters is sensitive to the level and volatility of the training data. The most sensitive hyperparameter is the initialization of the parametric density estimator. If

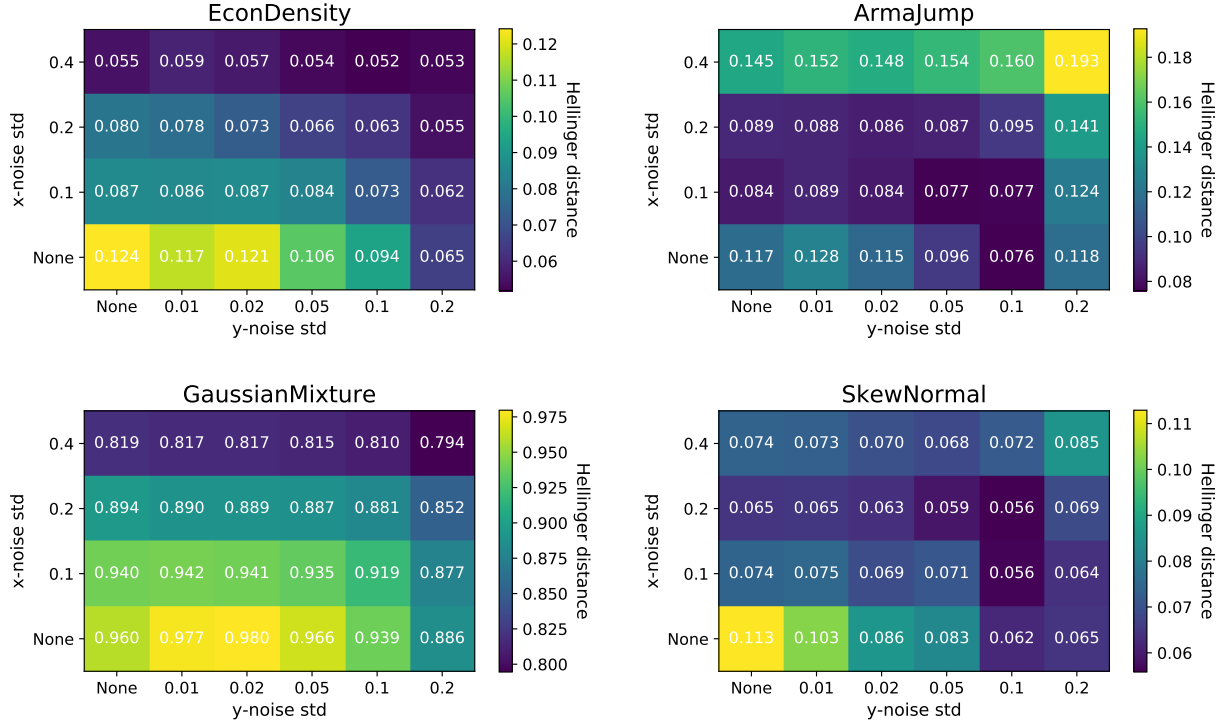


Figure 2. Effect of various noise regularization intensities η_x and η_y Hellinger distance between MDN estimate, fitted with 1600 samples, and the true density across various noise regularization intensities. The displayed values are averages over 5 seeds.

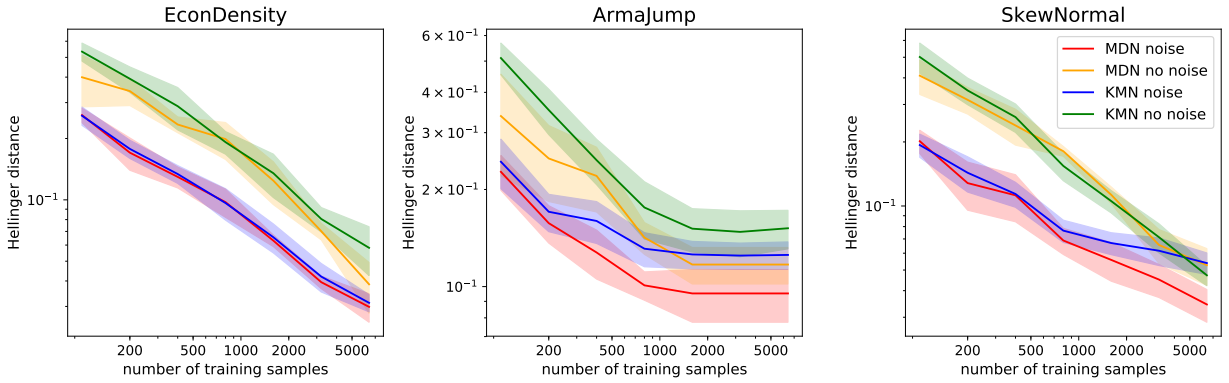


Figure 3. Effect of noise regularization on goodness of estimated density. Goodness of MDN/KMN density estimate, fitted with ($\eta_x = 0.2, \eta_y = 0.1$) and without noise regularization. The colored graphs display the Hellinger distance between estimated and true density, averaged over 5 seeds, and the translucent areas the respective standard deviation across varying samples sizes.

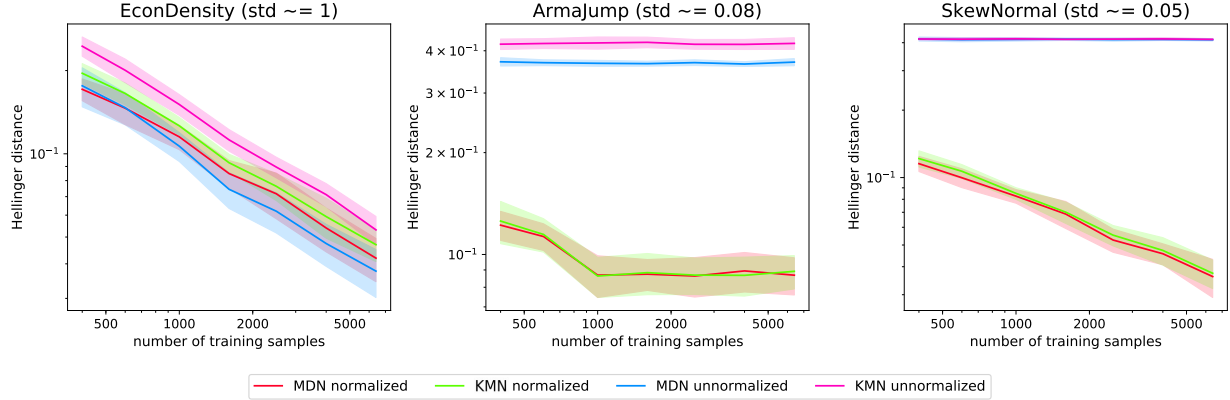


Figure 4. Effect of data normalization. Goodness of MDN / KMN density estimate, fitted with and without data normalization. The colored graphs display the Hellinger distance between estimated and true density, averaged over 5 seeds, and the translucent areas the respective standard deviation across varying samples sizes. While the EconDensity has a unconditional standard deviation of 1, the other two density simulations have a substantially lower unconditional volatility, ca. 0.08 and 0.05.

the initial density estimate is statistically too far away from the true density, numerical optimization might be slow or fail entirely to find a good fit. Figure 4 illustrates this phenomenon and emphasizes the practical importance of proper data normalization.

In case of the EconDensity simulation, the conditional standard deviation of the simulation density and the initial density estimate are similar. Both density estimation with and without data normalization yield quite similar results. Yet, the data normalization consistently reduces the Hellinger distance. The ArmaJump and SkewNormal density simulators have substantially smaller conditional standard deviations, i.e. 12 - 20 times smaller than the EconDensity. Without the data normalization scheme, the initial KMN/MDN density estimates exhibit a large statistical distance to the true conditional density. As a result, the numerical optimization is not able to sufficiently fit the density within 1000 training epochs. As can be seen in Figure 4, the resulting density estimates are substantially offset compared to the estimates with data normalization.

D. Conditional Density Estimator Benchmark Study

After empirically evaluating the noise regularization and data normalization scheme, we benchmark the neural network based density estimators against state-of-the art conditional density estimation approaches. Specifically, the benchmark study comprises the following conditional density estimators:

- **Mixture Density Network (MDN):** As introduced in Section IV.A.1. The MDN is trained with data normalization and noise regularization ($\eta_x = 0.2, \eta_y = 0.1$).
- **Kernel Mixture Network (KMN):** As introduced in Section IV.A.2. The KMN is trained with data normalization and noise regularization ($\eta_x = 0.2, \eta_y = 0.1$).

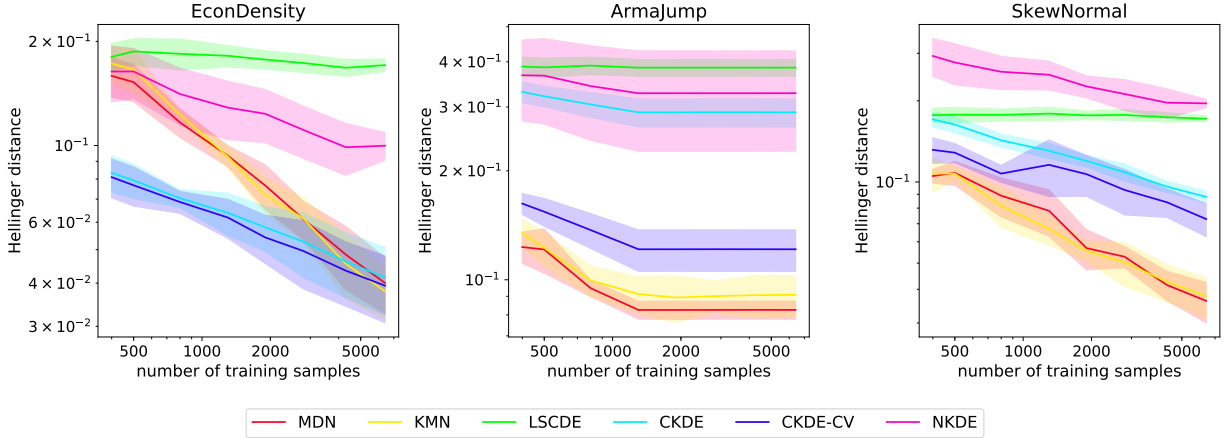


Figure 5. Conditional Density Estimator Benchmark. The illustrated benchmark study compares 6 density estimators across in 3 density simulations. To assess the goodness of fit, we report the Hellinger distance between the true density and the density estimate fitted with different sample sizes. The colored graphs display the Hellinger distance averaged over 5 seeds and the translucent areas the respective standard deviation.

- **Conditional Kernel Density Estimation (CKDE)**: Non-parametric approach introduced in II.B using bandwidth selection via the rule-of-thumb (Silverman, 1982).
- **CKDE with bandwidth selection via cross-validation (CKDE-CV)**: Similar to CKDE using bandwidth selection via maximum likelihood cross-validation (Li and Racine, 2007)
- **ϵ -Neighborhood kernel density estimation (NKDE)**: Non-parametric method that considers only a local subset of training points to form a density estimate.
- **Least-Squares Conditional Density Estimation (LSCDE)**: Semi-parametric estimator that computes the conditional density as linear combination of kernels (Sugiyama and Takeuchi, 2010).

Figure 5 depicts the evaluation results for the described estimators across different density simulations and number of training samples. Due to its limited modelling capacity, LSCDE yields poor estimates in all three evaluation cases and shows only minor improvements as the number of samples increases. CKDE consistently outperforms NKDE. This may be ascribed to the locality of the considered data neighborhoods of the training points that NKDE exhibits, whereas CKDE is able to fully use the available data. Unsurprisingly, the version of CKDE with bandwidth selection through cross-validation always improves upon CKDE with the rule-of-thumb.

In the EconDensity evaluation, CKDE achieves lower statistical distances for small sample sizes. However, the neural network based estimators KMN and MDN gain upon CKDE as the sample size increases and achieve similar results in case of 6000 samples. In the other two evaluation cases, KMN and MDN consistently outperform the other estimators. This demonstrates that even for small sample sizes, noise-regularized neural network estimators can be an equipollent or even superior alternative to well established CKDE.

VI. Empirical Evaluation on Euro Stoxx 50 Data

While the previous chapter is based on simulations, this chapter provides an empirical evaluation and benchmark study on real-world stock market data. In particular, we are concerned with estimating the conditional probability density of Euro Stoxx 50 returns. After a describing the data and density estimation task in detail, we report and discuss benchmark results.

A. The Euro Stoxx 50 data

The following empirical evaluation is based the Euro Stoxx 50 stock market index. The data comprises 3169 trading days, dated from January 2003 until June 2015. We define the task as predicting the conditional probability density of 1-day log returns, conditioned on 14 explanatory variables. These conditional variables include last period returns, risk free rate, realized volatility, option-implied moments and return factors. For a detailed description of the conditional variables, we refer to Appendix.F. Overall, the target variable is one-dimensional, i.e. $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}$, whereas the conditional variable \mathbf{x} constitutes a 14-dimensional vector, i.e. $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{14}$.

B. Evaluation Methodology

In order to assess the goodness of the different density estimators, out-of-sample validation is used. In particular, the available data is split in a training set which has a proportion of 80 % and a validation set consisting of the remaining 20 % of data points. The validation set \mathcal{D}_{val} is a consecutive series of data, corresponding to the 633 most recent trading-day and is only used for computing the following goodness-of-fit measures:

- *Avg. log-likelihood*: Average conditional log likelihood of validation data

$$\frac{1}{|\mathcal{D}_{val}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{val}} \log \hat{p}(\mathbf{y}|\mathbf{x}) \quad (27)$$

- *RMSE mean*: Root-Mean-Square-Error (RMSE) between the realized log-return and the mean of the estimated conditional distribution. The estimated conditional mean is defined as the expectation of \mathbf{y} under the distribution $\hat{p}(\mathbf{y}|\mathbf{x})$:

$$\hat{\mu}(\mathbf{x}) = \int_{\mathcal{Y}} \mathbf{y} \hat{p}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \quad (28)$$

Based on that, the RMSE w.r.t. $\hat{\mu}$ is calculated as

$$RMSE_{\mu} = \sqrt{\frac{1}{|\mathcal{D}_{val}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{val}} (\mathbf{y} - \hat{\mu}(\mathbf{x}))^2} \quad (29)$$

- *RMSE Std*: RMSE between the realized deviation from the predicted mean $\hat{\mu}(\mathbf{x})$ and the standard deviation of the conditional density estimate. The estimated conditional standard

	Avg. log-likelihood	RMSE mean (10^{-2})	RMSE std (10^{-2})
CKDE	3.3368 ± 0.0000	0.6924 ± 0.0000	0.8086 ± 0.0000
NKDE	3.1171 ± 0.0000	1.0681 ± 0.0000	0.5570 ± 0.0000
LSCDE	3.5079 ± 0.0087	0.7057 ± 0.0061	0.5442 ± 0.0028
MDN w/o noise	3.1386 ± 0.1501	0.5339 ± 0.0084	0.3222 ± 0.0064
KMN w/o noise	3.3130 ± 0.0743	0.6070 ± 0.0417	0.4072 ± 0.0372
MDN w/ noise	3.7539 ± 0.0324	0.5273 ± 0.0082	0.3188 ± 0.0016
KMN w/ noise	3.7969 ± 0.0250	0.5375 ± 0.0079	0.3254 ± 0.0063

Table I Out-of-sample validation on EuroStoxx 1-day returns.

	Avg. log-likelihood	RMSE mean (10^{-2})	RMSE std (10^{-2})
CKDE LOO-CV	3.8142 ± 0.0000	0.5344 ± 0.0000	0.3672 ± 0.0000
NKDE LOO-CV	3.3435 ± 0.0000	0.7943 ± 0.0000	0.4831 ± 0.0000
LSCDE 10-fold CV	3.5250 ± 0.0040	0.6836 ± 0.0056	0.5538 ± 0.0059
MDN 10-fold CV	3.8351 ± 0.0114	0.5271 ± 0.0060	0.3269 ± 0.0026
KMN 10-fold CV	3.8299 ± 0.0141	0.5315 ± 0.0075	0.3247 ± 0.0047

Table II Out-of-sample validation on EuroStoxx 1-day returns - Hyper-Parameter selection via cross-validation.

deviation is defined as

$$\hat{\sigma}(\mathbf{x}) = \sqrt{\int_{\mathcal{Y}} (\mathbf{y} - \hat{\mu}(\mathbf{x}))^2 \hat{p}(\mathbf{y}|\mathbf{x}) d\mathbf{y}} \quad (30)$$

The respective RSME is calculated as follows:

$$RMSE_{\sigma} = \sqrt{\frac{1}{|\mathcal{D}_{val}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{val}} (|\mathbf{y} - \hat{\mu}(\mathbf{x})| - \hat{\sigma}(\mathbf{x}))^2} \quad (31)$$

For details on the estimated conditional moments and the approximation of the associated integrals, we refer the interested reader to Appendix.G.

Calculating the average log-likelihood is a common way of evaluating the goodness of a density estimate (Rezende and Mohamed, 2015; Tansey et al., 2016; Trippe and Turner, 2018). The better the estimated conditional density approximates the true distribution, the higher the out-of-sample likelihood on expectation. Only if the estimator generalizes well beyond the training data, it can assign high conditional probabilities to the left-out validation data.

In finance, return distributions are often characterized by their centered moments. The RMSEs w.r.t. mean and standard deviation provide a quantitative measure for the predictive accuracy and consistency w.r.t. the predictive uncertainty. Overall, the training of the estimators and calculation of the goodness measures is performed with 5 different seeds. The reported results are averages over the 5 seeds, alongside the respective standard deviation.

C. Empirical Density Estimator Benchmark

The Euro Stoxx 50 estimator benchmark consists of two categories. First, we compare the estimators in their default hyper-parameter configuration. The default configurations have been selected with hyperparameter sweeps on the simulated densities in the previous chapters. The respective results are reported in Table I. Consistent with the simulation studies in Section V, noise regularization for MDNs and KMNs improves the estimate’s generalization (i.e. validation log-likelihood) by a significant margin. Moreover, the regularization improves the predictive accuracy (RMSE mean) and uncertainty estimates (RMSE std).

In the second part of the benchmark, the estimator parameters are selected through cross-validation on the Euro Stoxx training set. As goodness criterion, the log-likelihood is used. Following previous work (Duin, 1976; Pfeiffer, 1985; Li and Racine, 2007), we use leave-one-out cross-validation in conjunction with the downhill simplex method of Nelder and Mead (1965) for selecting the parameters of the kernel density estimators (CKDE and NKDE). For the remaining methods, hyper-parameter grid search with 10-fold cross-validation is employed. For details on the determined hyper-parameter settings, we refer to Appendix.E. Table II depicts the evaluation results with hyper-parameter search. Similar to the results without hyper-parameter selection, MDNs and KMNs with noise regularization consistently outperform previous methods in all three evaluation measures. This strengthens the results from the simulation study and demonstrates that, when regularized properly, neural network based methods are able to generate superior conditional density estimates.

Moreover, it is interesting to observe that both kernel density estimators, improve substantially, when cross-validation is used. This is a strong indication, that the bandwidth which is selected through the Gaussian rule-of-thumb is inferior and the underlying return data is non-Gaussian.

VII. Conclusion

This paper studies the use of neural networks for conditional density estimation. Addressing the problem of over-fitting, we introduce a noise regularization method that leads to smooth density estimates and improved generalization. Moreover, a normalization scheme which makes the model’s hyper-parameters insensitive to differing value ranges is proposed. Corresponding experiments showcase the effectiveness and practical importance of the presented approaches. In a benchmark study, we demonstrate that our training methodology endows neural network based CDE with a better out-of-sample performance than previous semi- and non-parametric methods. Overall, this work establishes a practical framework for the successful application of neural network based CDE in areas such as econometrics. Based on the promising results, we are convinced that the proposed method enhances the econometric toolkit and thus advocate further research in this direction. While this paper focuses on CDE with mixture densities, a promising avenue for future research could be the use of normalizing flows as parametric density representation.

Appendix

Appendix A. Noise Regularization

Let $\mathcal{L}_{\mathcal{D}}(\mathcal{D})$ be a loss function over a set of data points $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, which can be partitioned into a sum of losses corresponding to each data point x_n :

$$\mathcal{L}_{\mathcal{D}}(\mathcal{D}) = \sum_{i=1}^N \mathcal{L}(\mathbf{x}_n) \quad (1)$$

Also, let each x_n be perturbed by a random noise vector $\boldsymbol{\xi} \sim q(\boldsymbol{\xi})$ with zero mean and i.i.d. elements, i.e.

$$\mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}] = 0 \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}_n \boldsymbol{\xi}_j^\top] = \eta^2 I \quad (2)$$

The resulting loss $\mathcal{L}(\mathbf{x}_n + \boldsymbol{\xi})$ can be approximated by a second order Taylor expansion around x_n

$$\mathcal{L}(\mathbf{x}_n + \boldsymbol{\xi}) = \mathcal{L}(\mathbf{x}_n) + \boldsymbol{\xi}^\top \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n} + \frac{1}{2} \boldsymbol{\xi}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n} \boldsymbol{\xi} + \mathcal{O}(\boldsymbol{\xi}^3) \quad (3)$$

Assuming that the noise $\boldsymbol{\xi}$ is small in its magnitude $\mathcal{O}(\boldsymbol{\xi}^3)$ may be neglected. The expected loss under $q(\boldsymbol{\xi})$ follows directly from (3):

$$\mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\mathcal{L}(\mathbf{x}_n + \boldsymbol{\xi})] = \mathcal{L}(\mathbf{x}_n) + \mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}^\top \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n}] + \frac{1}{2} \mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n} \boldsymbol{\xi}] \quad (4)$$

Using the assumption about $\boldsymbol{\xi}$ in (2) we can simplify (4) as follows:

$$\mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\mathcal{L}(\mathbf{x}_n + \boldsymbol{\xi})] = \mathcal{L}(\mathbf{x}_n) + \mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}^\top \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n}] + \frac{1}{2} \mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n} \boldsymbol{\xi}] \quad (5)$$

$$= \mathcal{L}(\mathbf{x}_n) + \frac{1}{2} \mathbb{E}_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} [\boldsymbol{\xi}^\top \mathbf{H}^{(n)} \boldsymbol{\xi}] \quad (6)$$

$$= \mathcal{L}(\mathbf{x}_n) + \frac{1}{2} E_{\boldsymbol{\xi} \sim q(\boldsymbol{\xi})} \left[\sum_j \sum_k \xi_j \xi_k \frac{\partial^2 \mathcal{L}(\mathbf{x})}{\partial x^{(j)} \partial x^{(k)}} \Big|_{\mathbf{x}_n} \right] \quad (7)$$

$$= \mathcal{L}(\mathbf{x}_n) + \frac{1}{2} \sum_j E_{\boldsymbol{\xi}} [\xi_j^2] \frac{\partial^2 \mathcal{L}(\mathbf{x})}{\partial x^{(j)} \partial x^{(j)}} \Big|_{\mathbf{x}_n} + \frac{1}{2} \sum_j \sum_{k \neq j} E_{\boldsymbol{\xi}} [\xi_j \xi_k] \frac{\partial^2 \mathcal{L}(\mathbf{x})}{\partial x^{(j)} \partial x^{(k)}} \Big|_{\mathbf{x}_n} \quad (8)$$

$$= \mathcal{L}(\mathbf{x}_n) + \frac{\eta^2}{2} \sum_j \frac{\partial^2 \mathcal{L}(\mathbf{x})}{\partial x^{(j)} \partial x^{(j)}} \Big|_{\mathbf{x}_n} \quad (9)$$

$$= \mathcal{L}(\mathbf{x}_n) + \frac{\eta^2}{2} \text{tr}(\mathbf{H}^{(n)}) \quad (10)$$

In that, $\mathcal{L}(\mathbf{x}_n)$ is the loss without noise and $\mathbf{H}^{(n)} = \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{x})|_{\mathbf{x}_n}$ the Hessian of \mathcal{L} at \mathbf{x}_n . With ξ_j we denote the elements of the column vector $\boldsymbol{\xi}$.

Appendix B. Data Normalization and Change of Variable

LEMMA 1: Let $\mathbf{x} \in S \subseteq \mathbb{R}^n$ be a continuous random variable with probability density function $p(\mathbf{x})$. Any linear transformation $\mathbf{z} = \mathbf{a} + B\mathbf{x}$ of $\mathbf{x} \sim p(\mathbf{x})$ with $\mathbf{a} \in \mathbb{R}^n$ and B being an invertible $n \times n$ matrix follows the probability density function

$$q(\mathbf{z}) = \frac{1}{|B|} p(B^{-1}(\mathbf{z} - \mathbf{a})) \quad , \quad \mathbf{z} \in \{\mathbf{a} + B\mathbf{x} \mid \mathbf{x} \in S\} \quad . \quad (11)$$

PROOF 1 (Proof of Lemma 1): The Lemma directly follows from the change of variable theorem (see Bishop (2006) page 18)

THEOREM 3: Let $\mathbf{x} \in \mathbb{R}^n$ be a continuous random variable following a Gaussian Mixture Model (GMM), this is $\mathbf{x} \sim p(\mathbf{x})$ with

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mu_k, \Sigma_k) \quad . \quad (12)$$

Any linear transformation $\mathbf{z} = \mathbf{a} + B\mathbf{x}$ of $\mathbf{x} \sim p(\mathbf{x})$ with $\mathbf{a} \in \mathbb{R}^n$ and B being an invertible $n \times n$ matrix follows a Gaussian Mixture Model with density function

$$p(\mathbf{z}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{a} + B\mu_k, B\Sigma_k B^\top) \quad . \quad (13)$$

PROOF 2 (Proof of Theorem 3): With $\mathbf{x} \in \mathbb{R}^n$ following a Gaussian Mixture Model, its probability density function can be written as

$$p(\mathbf{x}) = \sum_{k=1}^K w_k \mathcal{N}(\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{1}{2}}} \sum_{k=1}^K w_k \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)^\top \Sigma_k^{-1}(\mathbf{x} - \mu_k)\right)}{|\Sigma_k|^{\frac{1}{2}}} \quad (14)$$

Let $\mathbf{z} \sim q(\mathbf{z})$ be a linear transformation $\mathbf{z} = \mathbf{a} + B\mathbf{x}$ of $\mathbf{x} \sim p(\mathbf{x})$ with $\mathbf{a} \in \mathbb{R}^n$ and B being an invertible $n \times n$ matrix. From Lemma 1 follows that

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{\frac{1}{2}}} \sum_{k=1}^K w_k \frac{\exp\left(-\frac{1}{2}(B^{-1}\mathbf{z} - B^{-1}\mathbf{a} - \mu_k)^\top \Sigma_k^{-1}(B^{-1}\mathbf{z} - B^{-1}\mathbf{a} - \mu_k)\right)}{|B| |\Sigma_k|^{\frac{1}{2}}} \quad (15)$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}} \sum_{k=1}^K w_k \frac{\exp\left(-\frac{1}{2}(\mathbf{z} - (\mathbf{a} + B\mu_k))^\top (B^{-1})^\top \Sigma_k^{-1} B^{-1}(\mathbf{z} - (\mathbf{a} + B\mu_k))\right)}{|B| |\Sigma_k|^{\frac{1}{2}}} \quad (16)$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}} \sum_{k=1}^K w_k \frac{\exp\left(-\frac{1}{2}(\mathbf{z} - (\mathbf{a} + B\mu_k))^\top (B\Sigma_k B^\top)^{-1}(\mathbf{z} - (\mathbf{a} + B\mu_k))\right)}{|B\Sigma_k B^\top|^{\frac{1}{2}}} \quad (17)$$

$$= \sum_{k=1}^K w_k \mathcal{N}(\mathbf{a} + B\mu_k, B\Sigma_k B^\top) \quad \square$$

Appendix C. Density Simulation

This section holds detailed descriptions of the simulated conditional densities used for the experiments in chapter V. The respective conditional densities are illustrated in Figure 6.

Appendix C.1. EconDensity

This simple, economically inspired, distribution has the following data generating process $(x, y) \sim p(x, y)$:

$$x = |\epsilon_x|, \quad \epsilon_x \sim N(0, 1) \quad (18)$$

$$\sigma_y = 1 + x \quad (19)$$

$$y = x^2 + \epsilon_y, \quad \epsilon_y \sim N(0, \sigma_y) \quad (20)$$

The conditional density follows as

$$p(y|x) = \mathcal{N}(y|\mu = x^2, \sigma = 1 + x) \quad (21)$$

and is illustrated in Figure 6a. One can imagine x to represent financial market volatility, which is always positive with rare large realizations. y can be an arbitrary variable that is explained by volatility. We choose a non-linear relationship between x and y to check how the estimators can cope with that. To make things more difficult, the relationship between x and y becomes more blurry at high x realizations, as expressed in a heteroscedastic σ_y , that is rising with x . This reflects the common behaviour of higher noise in the estimators in times of high volatility.

Appendix C.2. ARMAJump

The underlying data generating process for this simulator is an AR(1) model with a jump component. A new realization x_t of the time-series can be described as follows:

$$x_t = [c(1 - \alpha) + \alpha x_{t-1}] + (1 - z_t)\sigma\epsilon_t + z_t[-c + 3\sigma\epsilon_t] \quad \epsilon_t \sim N(0, 1), z_t \sim B(1, p) \quad (22)$$

In that, $c \in \mathbb{R}$ is the long run mean of the AR(1) process and $\alpha \in \mathbb{R}$ constitutes the autoregressive factor, describing how fast the AR(1) time series returns to its long run mean c . Typically an ARMA process is perturbed by Gaussian White Noise $\sigma\epsilon_t$ with standard deviation $\sigma \in \mathbb{R}^+$. We add a jump component, that occurs with probability p and is indicated by the Bernoulli distributed binary variable z_t . If a jump occurs, a negative shock of the same magnitude as c is accompanied by Gaussian noise with three times higher standard deviation than normal. The dynamic is a discrete version of the class of affine jump diffusion models, which are heavily used in bond and option pricing. Here, for each time period t , the conditional density $p(x_t|x_{t-1})$ shall be predicted. Note

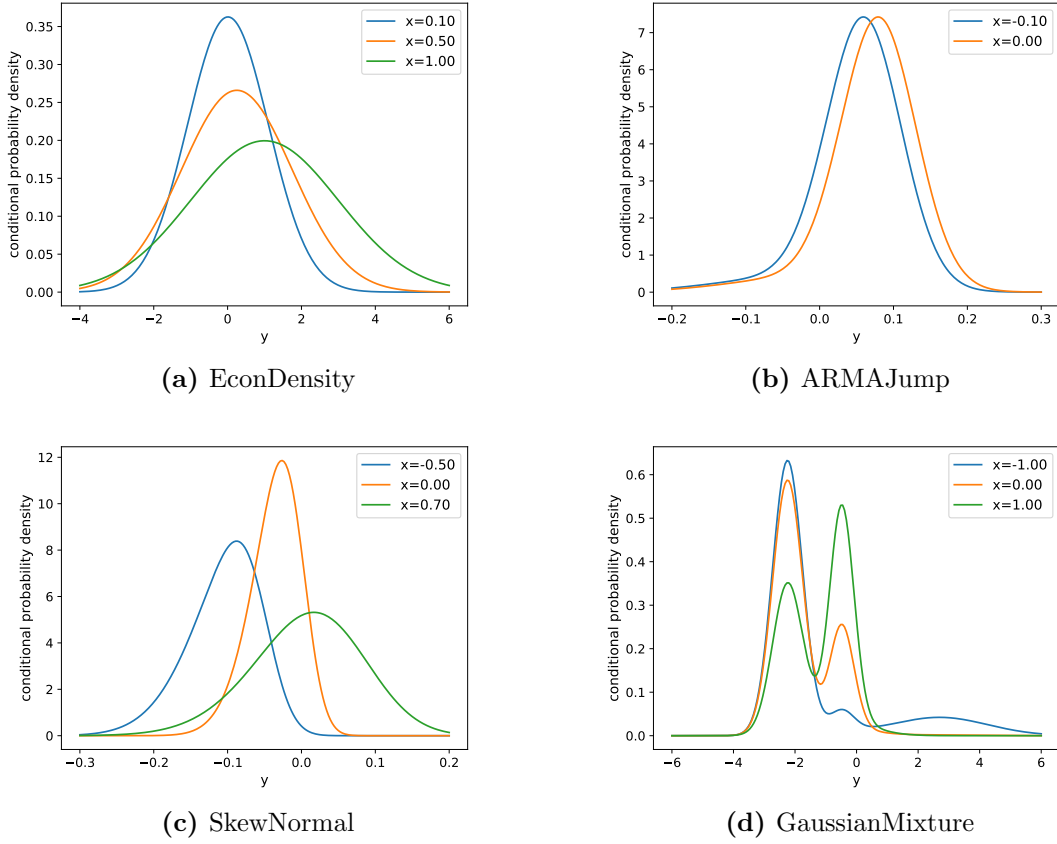


Figure 6. Conditional density simulation models. Conditional probability densities corresponding to the different simulation models. The coloured graphs represent the probability densities $p(y|x)$, conditioned on different values of x .

that in this case y corresponds to x_t . The conditional density follows as mixture of two Gaussians:

$$p(x_t|x_{t-1}) = (1 - p)\mathcal{N}(x_t|\mu = c(1 - \alpha) + \alpha x_{t-1}, \sigma) + p\mathcal{N}(x_t|\mu = \alpha(x_{t-1} - c), 3\sigma) \quad (23)$$

Figure 6b depicts the ARMAJump conditional probability density for the time-series parameters $c = 0.1, \alpha = 0.2, p = 0.1, \sigma = 0.05$. As can be seen in the depiction, the conditional distribution has a negative skewness, resulting from the jump component.

Appendix C.3. SkewNormal

The data generating process $(x, y) \sim p(x, y)$ resembles a bivariate joint-distribution, wherein $x \in \mathbb{R}$ follows a normal distribution and $y \in \mathbb{R}$ a conditional skew-normal distribution (And  l et al., 1984). The parameters (ξ, ω, α) of the skew normal distribution are functionally dependent on x .

Specifically, the functional dependencies are the following:

$$x \sim \mathcal{N}\left(\cdot \mid \mu = 0, \sigma = \frac{1}{2}\right) \quad (24)$$

$$\xi(x) = a * x + b \quad a, b \in \mathbb{R} \quad (25)$$

$$\omega(x) = c * x^2 + d \quad c, d \in \mathbb{R} \quad (26)$$

$$\alpha(x) = \alpha_{low} + \frac{1}{1 + e^{-x}} * (\alpha_{high} - \alpha_{low}) \quad (27)$$

$$y \sim \text{SkewNormal}(\xi(x), \omega(x), \alpha(x)) \quad (28)$$

Accordingly, the conditional probability density $p(y|x)$ corresponds to the skew normal density function:

$$p(y|x) = \frac{2}{\omega(x)} \mathcal{N}\left(\frac{y - \xi(x)}{\omega(x)}\right) \Phi\left(\alpha(x) \frac{y - \xi(x)}{\omega(x)}\right) \quad (29)$$

In that, $\mathcal{N}(\cdot)$ denotes the density, and $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution. The shape parameter $\alpha(x)$ controls the skewness and kurtosis of the distribution. We set $\alpha_{low} = -4$ and $\alpha_{high} = 0$, giving $p(y|x)$ a negative skewness that decreases as x increases. This distribution will allow us to evaluate the performance of the density estimators in presence of skewness, a phenomenon that we often observe in financial market variables. Figure 6c illustrates the conditional skew normal distribution.

Appendix C.4. GaussianMixture

The joint distribution $p(\mathbf{x}, \mathbf{y})$ follows a GMM. We assume that $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^l$ can be factorized, i.e.

$$p(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y},k}, \Sigma_{\mathbf{y},k}) \mathcal{N}(\mathbf{x} | \mu_{\mathbf{x},k}, \Sigma_{\mathbf{x},k}) \quad (30)$$

When \mathbf{x} and \mathbf{y} can be factorized as in (30), the conditional density $p(\mathbf{y}|\mathbf{x})$ can be expressed as:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^K W_k(x) \mathcal{N}(\mathbf{y} | \mu_{\mathbf{y},k}, \Sigma_{\mathbf{y},k}) \quad (31)$$

wherein the mixture weights are a function of \mathbf{x} :

$$W_k(x) = \frac{w_k \mathcal{N}(\mathbf{x} | \mu_{\mathbf{x},k}, \Sigma_{\mathbf{x},k})}{\sum_{j=1}^K w_k \mathcal{N}(\mathbf{x} | \mu_{\mathbf{x},j}, \Sigma_{\mathbf{x},j})} \quad (32)$$

For details and derivations we refer the interested reader to Guang Sung (2004) and Gilardi et al. (2002). Figure 6d depicts the conditional density of a GMM with 5 components (i.e. $K = 5$ and 1-dimensional \mathbf{x} and \mathbf{y} (i.e. $l = m = 1$)).

Appendix D. Conditional Density Estimators in the Benchmark Study

This section holds a detailed description of the CDE methods, compared to, in the benchmark studies:

- **Conditional Kernel Density Estimation (CKDE):** This non-parametric conditional density approach estimates both the joint probability $\hat{p}(\mathbf{x}, \mathbf{y})$ and the marginal probability $\hat{p}(\mathbf{x})$ with KDE (see Section II.A.2). The conditional density estimate follows as density ratio $\hat{p}(\mathbf{y}|\mathbf{x}) = \frac{\hat{p}(\mathbf{x}, \mathbf{y})}{\hat{p}(\mathbf{x})}$. For selecting the bandwidths $h_{\mathbf{x}}$ and $h_{\mathbf{y}}$ of the kernels, the rule-of-thumb of Silverman (1982) is employed:

$$h = 1.06\hat{\sigma}N^{-\frac{1}{4+d}} \quad (33)$$

In that, N denotes the number of samples, $\hat{\sigma}$ the empirical standard deviation and d the dimensionality of the data. The rule-of-thumb assumes that the data follows a normal distribution. If this assumption holds, the selected bandwidth h is proven to be optimal w.r.t. the *IMSE* criterion.

- **CKDE with bandwidth selection via cross-validation (CKDE-CV):** Similar to the CKDE above but the bandwidth parameters $h_{\mathbf{x}}$ and $h_{\mathbf{y}}$ are determined with leave-one-out maximum likelihood cross-validation. See Li and Racine (2007) for further details about the cross-validation-based bandwidth selection.
- **ϵ -Neighborhood kernel density estimation (NKDE):** For estimating the conditional density $p(\mathbf{y}|\mathbf{x})$, ϵ -neighbor kernel density estimation employs standard kernel density estimation in a local ϵ -neighborhood around a query point (\mathbf{x}, \mathbf{y}) (Sugiyama and Takeuchi, 2010). NKDE is similar to CKDE, as it uses kernels, placed in the training data points, to estimate the conditional probability density. However, rather than estimating both the joint probability $p(\mathbf{x}, \mathbf{y})$ and marginal probability $p(\mathbf{x})$, NKDE forms a density estimate by only considering a local subset of the training samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in \mathcal{I}_{\mathbf{x}, \epsilon}}$, where $\mathcal{I}_{\mathbf{x}, \epsilon}$ is the set of sample indices such that $\|\mathbf{x}_i - \mathbf{x}\|_2 \leq \epsilon$. The estimated density can be expressed as

$$p(\mathbf{y}|\mathbf{x}) = \sum_{j \in \mathcal{I}_{\mathbf{x}, \epsilon}} w_j \prod_{i=1}^l \frac{1}{h^{(i)}} K\left(\frac{y^{(i)} - y_j^{(i)}}{h^{(i)}}\right) \quad (34)$$

wherein w_j is the weighting of the j -th kernel and $K(z)$ a kernel function. In our implementation K is the density function of a standard normal distribution. The weights w_j can either be uniform, i.e. $w_j = \frac{1}{|\mathcal{I}_{\mathbf{x}, \epsilon}|}$ or proportional to the distance $\|\mathbf{x}_j - \mathbf{x}\|$. The vector of bandwidths $\mathbf{h} = (h^{(1)}, \dots, h^{(l)})^T$ is determined with the rule-of-thumb (see Equation 33), where the the number of samples N corresponds to the average number of neighbors in the training data:

$$N = \frac{1}{N} \sum_{n=1}^N |\mathcal{I}_{\mathbf{x}_n, \epsilon}| - 1 \quad (35)$$

- **Least-Squares Conditional Density Estimation (LSCDE):** A semi-parametric estimator that computes the conditional density as linear combination of kernels (Sugiyama and

Takeuchi, 2010).

$$\hat{p}_\alpha(\mathbf{y}|\mathbf{x}) \propto \alpha^T \phi(\mathbf{x}, \mathbf{y}) \quad (36)$$

Due to its restriction to linear combinations of Gaussian kernel functions ϕ , the optimal parameters α w.r.t. the integrated mean squared objective

$$J(\alpha) = \int \int (\hat{p}_\alpha(\mathbf{x}, \mathbf{y}) - p(\mathbf{x}, \mathbf{y}))^2 p(\mathbf{x}) d\mathbf{x} d\mathbf{y} \quad (37)$$

objective can be computed in closed form. However, at the same time, the linearity assumption makes the estimator less expressive than the KMN or MDN. See Appendix Sugiyama and Takeuchi (2010) for details.

Appendix E. Hyper-Parameter Settings and Selection

This section lists and describes the hyper-parameter configurations used for the empirical evaluations. In that, we first attend to the default configuration used in the benchmarks in Fig. 5 and Table I.

The neural network has two hidden layers with 16 neurons each, tanh non-linearities and weight normalization (Salimans and Kingma, 2016). For the KMN, we use $K = 50$ Gaussian mixture components and, for the MDN, $K = 20$ components. The neural network is trained for 1000 epochs with the Adam optimizer (Kingma and Ba, 2015). In that, the Adam learning rate is set to $\alpha = 0.001$ and the mini-batch size is 200. In order to select the kernel centers, for the KMN, we employ K-means clustering on the \mathbf{y}_i data points. In each of the selected centers, we place two Gaussians with initial standard deviation $\sigma_1 = 0.7$ and $\sigma_2 = 0.3$. While the locations of the Gaussians are fixed during training, the scale / standard deviation parameters are trainable and thus adjusted by the optimizer. Table III provides an overview of the default hyper-parameters.

	MDN	KMN
hidden layer sizes	(16,16)	(16,16)
hidden non-linearity	tanh	tanh
training epochs	1000	1000
Adam learning rate	0.001	0.001
K : number of components	20	50
η_x : noise std x	0.2	0.2
η_y : noise std y	0.1	0.1
weight normalization	True	True
data normalization	True	True
initialization of scales	-	[0.7, 0.3]
trainable scales	-	True

Table III Default hyper-parameter configuration for MDN and KMN

In the benchmark study with hyper-parameter selection (see Fig. II), a subset of the parameters has been optimized via 10-fold cross-validation grid search. Since the search space grows

exponentially with the number of parameters to optimize over, we constrained the grid search to the hyper-parameters which we found to have the greatest influence on the estimators performance. In the case of MDN and KMN, these are the number of training epochs, the number of mixture components and the noise regularization intensities. For LSCDE, the search comprised the number of kernels, bandwidth and damping parameter λ . Table IV holds the hyper-parameters that were determined through the parameter search and subsequently used to fit and evaluate the respective estimator.

	MDN-CV	KMN-CV	LSCDE-CV
training epochs	500	500	-
K : number of components	10	200	1000
η_x : noise std x	0.3	0.2	-
η_y : noise std y	0.15	0.15	-
bandwidth	-	-	0.5
λ : LSCDE damping parameter	-	-	0.1

Table IV Hyper-parameter configuration determined with 10-fold cross-validation on the EuroStoxx 50 data set

Appendix F. The Euro Stoxx 50 Data

The following section holds a detailed description of the EuroStoxx 50 dataset employed in chapter VI. The data comprises 3169 trading days, dated from January 2003 until June 2015. We define the task as predicting the conditional probability density of 1-day log returns, conditioned on 14 explanatory variables. These conditional variables are listed below:

- **log_ret_last_period:** realized log-return of previous trading day
- **log_ret_risk_free_1d:** risk-free 1-day log return, computed based on the overnight index swap rate (OIS) with 1 day maturity. The OIS rate r_f is transformed as $\log(\frac{r_f}{365} + 1)$.
- **RealizedVariation:** estimate of realized variance of previous day, computed as sum of squared 10 minute returns over the previous trading day
- **SVIX:** 30-day option implied volatility³ (Whaley, 1993)
- **bakshiKurt:** 30-day option implied kurtosis³ (Bakshi et al., 2003)
- **bakshiSkew:** 30-day option implied skewness³ (Bakshi et al., 2003)
- **Mkt-RF:** Fama-French market return factor (Fama and French, 1993)

- **SMB:** Fama-French Small-Minus-Big factor (Fama and French, 1993)
- **HML:** Fama-French High-Minus-Low factor (Fama and French, 1993)
- **WML:** Winner-Minus-Loser (momentum) factor (Carhart, 1997)
- **Mkt-RF 10-day risk:** risk of market return factor; sum of squared market returns over the last 10 trading days
- **SMB 10-day risk:** SMB factor risk; sum of squared factor returns over the last 10 days
- **HML 10-day risk:** HML factor risk; sum of squared factor returns over the last 10 days
- **WML 10-day risk:** WML factor risk; sum of squared factor returns over the last 10 days

Appendix G. Estimation of the Conditional Moments

This section briefly describes how the moments of the conditional density estimates $\hat{p}(\mathbf{y}|\mathbf{x})$ are computed. In particular, we will focus on the mean, covariance, skewness and kurtosis.

By default the centered moments are estimated via numerical or monte-carlo integration, using their respective definitions:

$$\hat{\mu}(\mathbf{x}) = \int_{\mathcal{Y}} \mathbf{y} \hat{p}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \quad (38)$$

$$\hat{\sigma}(\mathbf{x}) = \sqrt{\int_{\mathcal{Y}} (\mathbf{y} - \hat{\mu}(\mathbf{x}))^2 \hat{p}(\mathbf{y}|\mathbf{x}) d\mathbf{y}} \quad (39)$$

$$\widehat{Cov}(\mathbf{x}) = \int_{\mathcal{Y}} (\mathbf{y} - \hat{\mu}(\mathbf{x}))(\mathbf{y} - \hat{\mu}(\mathbf{x}))^T \hat{p}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \quad (40)$$

$$\widehat{Skew}(\mathbf{x}) = \int_{\mathcal{Y}} \left(\frac{\mathbf{y} - \hat{\mu}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})} \right)^3 \hat{p}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \quad (41)$$

$$\widehat{Kurt}(\mathbf{x}) = \int_{\mathcal{Y}} \left(\frac{\mathbf{y} - \hat{\mu}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})} \right)^4 \hat{p}(\mathbf{y}|\mathbf{x}) d\mathbf{y} - 3 \quad (42)$$

Our implementation only supports estimating skewness and kurtosis for univariate target variables, i.e. $\dim(\mathcal{Y}) = 1$. If $\dim(\mathcal{Y}) = 1$, the integral is approximated with numerical integration, using the Gaussian quadrature with 10000 reference points, for which the density values are calculated. If $\dim(\mathcal{Y}) > 1$, we use Monte-Carlo integration with 100,000 samples (see Appendix??).

³The option implied moments are computed based on a options with maturity in 30 days. Since the days to maturity vary, linear interpolation of the option implied moments, corresponding to different numbers of days till maturity, is used to compute an estimate for maturity in 30 days.

In case of the KMN and MDN, the conditional distribution is a GMM. Thus, we can directly calculate mean and covariance from the GMM parameters, outputted by the neural network. The mean follows straightforward as weighted sum of the Gaussian component centers: $\mu_k(\mathbf{x}; \theta)$

$$\hat{\mu}(\mathbf{x}) = \sum_{k=1}^K w_k(\mathbf{x}; \theta) \mu_k(\mathbf{x}; \theta) \quad (43)$$

The covariance matrix can be computed as

$$\hat{Cov}(\mathbf{x}) = \sum_{k=1}^K w_k(\mathbf{x}; \theta) ((\mu_k(\mathbf{x}; \theta) - \hat{\mu}(\mathbf{x}))(\mu_k(\mathbf{x}; \theta) - \hat{\mu}(\mathbf{x}))^T + \text{diag}(\sigma_k(\mathbf{x}; \theta)^2)) \quad (44)$$

wherein the outer product accounts for the covariance that arises from the different locations of the components and the diagonal matrix for the inherent variance of each Gaussian component.

REFERENCES

- Ambrogioni, Luca, Umut Güçlü, Marcel A. J. van Gerven, and Eric Maris, 2017, The Kernel Mixture Network: A Nonparametric Method for Conditional Density Estimation of Continuous Random Variables.
- An, Guozhong, 1996, The Effects of Adding Noise During Backpropagation Training on a Generalization Performance, *Neural Computation* 8, 643–674.
- Anděl, Ji, Ivan Netuka, and Karel Zvára, 1984, On Threshold Autoregressive Processes, *Kybernetika* 20, 89–106.
- Bakshi, Gurdip S., Nikunj Kapadia, and Dilip B. Madan, 2003, Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options, *Review of Financial Studies* .
- Bishop, Chris M., 1995, Training with Noise is Equivalent to Tikhonov Regularization, *Neural Computation* 7, 108–116.
- Bishop, Christopher M, 1994, Mixture Density Networks.
- Bishop, Christopher M, 2006, *Pattern Recognition and Machine Learning* (Springer).
- Bollerslev, Tim, Bollerslev, and Tim, 1987, A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return, *The Review of Economics and Statistics* 69, 542–47.
- Botev, Z. I., J. F. Grotowski, and D. P. Kroese, 2010, Kernel density estimation via diffusion, *The Annals of Statistics* 38, 2916–2957.
- Bowman, Adrian W., 1984, An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* 71, 353–360.
- Cao, Ricardo, Antonio Cuevas, and Wensceslao González Manteiga, 1994, A comparative study of several smoothing methods in density estimation, *Computational Statistics & Data Analysis* 17, 153–176.
- Carhart, Mark M., 1997, On Persistence in Mutual Fund Performance, *The Journal of Finance* 52, 57–82.

- De Gooijer, Jan G, and Dawit Zerom, 2003, On Conditional Density Estimation, Technical report.
- Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio, 2017, Density estimation using Real NVP, in *Proceedings of the International Conference on Learning Representations*.
- Duin, R. P. W, 1976, On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions, *IEEE Transactions on Computers* C-25, 1175–1179.
- Engle, Robert F., 1982, Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica* 50, 987.
- Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.
- Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.
- Gallant, A. Ronald, D. Hsieh, George Tauchen, A. Gallant, D. Hsieh, and George Tauchen, 1991, ON FITTING A RECALCITRANT SERIES: THE POUND/DOLLAR EXCHANGE RATE, *Nonparametric and Semiparametric Methods in Econometrics and Statistics* .
- Gilardi, Nicolas, Samy Bengio, and Mikhail Kanevski, 2002, Conditional Gaussian Mixture Models for Environmental Risk Mapping, in *NNSP*.
- Glosten, Lawrence R, Ravi Jagannathan, David E Runkle, Lawrence R Glosten, Ravi Jagannathan, and David E Runkle, 1993, On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks, *Journal of Finance* 48, 1779–1801.
- Gormsen, Niels Joachim, and Christian Skov Jensen, 2017, Conditional Risk, *SSRN Electronic Journal* .
- Grus, Joel, 2015, *Data Science from Scratch*.
- Guang Sung, Hsi, 2004, *Gaussian Mixture Regression and Classification*, Ph.D. thesis.
- Hall, Peter, JS Marron, and Byeong U Park, 1992, Smoothed cross-validation, *Probability Theory* 92, 1–20.

- Hamilton, James D. (James Douglas), 1994, *Time series analysis* (Princeton University Press).
- Hansen, Bruce E., Hansen, and Bruce, 1994, Autoregressive Conditional Density Estimation, *International Economic Review* 35, 705–30.
- Harvey, Campbell R, and Akhtar Siddique, 1999, Autoregressive Conditional Skewness, *The Journal of Financial and Quantitative Analysis* .
- Holmstrom, L., and P. Koistinen, 1992, Using additive noise in back-propagation training, *IEEE Transactions on Neural Networks* 3, 24–38.
- Hornik, Kurt, 1991, Approximation capabilities of multilayer feedforward networks, *Neural Networks* 4, 251–257.
- Hyndman, Rob J., David M. Bashtannyk, and Gary K. Grunwald, 1996, Estimating and Visualizing Conditional Densities, *Journal of Computational and Graphical Statistics* 5, 315.
- Jagannathan, Ravi, and Zhenyu Wang, 1996, The Conditional CAPM and the Cross-Section of Expected Returns, *The Journal of Finance* 51, 3.
- Jondeau, Eric, and Michael Rockinger, 2003, Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements, Technical report.
- Kingma, Diederik P., and Jimmy Ba, 2015, Adam: A Method for Stochastic Optimization, in *ICLR*.
- Krogh, Anders, and John A Hertz, 1992, A Simple Weight Decay Can Improve Generalization, Technical report.
- Kukačka, Jan, Vladimir Golkov, and Daniel Cremers, 2017, Regularization for Deep Learning: A Taxonomy, Technical report.
- Lewellen, Jonathan, and Stefan Nagel, 2006, The conditional CAPM does not explain asset-pricing anomalies, *Journal of Financial Economics* 82, 289–314.
- Li, Jonathan Q, and Abstract R Andrew Barron, 2000, Mixture Density Estimation, in *NIPS*.

- Li, Qi, and Jeffrey S. Racine, 2007, *Nonparametric econometrics : theory and practice* (Princeton University Press).
- Mirza, Mehdi, and Simon Osindero, 2014, Conditional Generative Adversarial Nets, Technical report.
- Nelder, J A, and R Mead, 1965, A simplex method for function minimization, *The Computer Journal* 7, 308–313.
- Nelson, Daniel B., and Charles Q. Cao, 1992, Inequality Constraints in the Univariate GARCH Model, *Journal of Business & Economic Statistics* 10, 229.
- Park, Byeong, Byeong Park, and J. S. Marron, 1990, Comparison of data driven bandwidth selectors, *J. STATIST. ASSOC* 66–72.
- Parzen, Emanuel, 1962, On Estimation of a Probability Density Function and Mode, *The Annals of Mathematical Statistics* 33, 1065–1076.
- Pfeiffer, K.P., 1985, Stepwise variable selection and maximum likelihood estimation of smoothing factors of kernel functions for nonparametric discriminant functions evaluated by different criteria, *Computers and Biomedical Research* 18, 46–61.
- Rezende, Danilo Jimenez, and Shakir Mohamed, 2015, Variational Inference with Normalizing Flows, in *Proceedings of the 32nd International Conference on Machine Learning*.
- Rosenblatt, Murray, 1956, Remarks on Some Nonparametric Estimates of a Density Function, *The Annals of Mathematical Statistics* 27, 832–837.
- Rudemo, Mats, 1982, Empirical Choice of Histograms and Kernel Density Estimators.
- Salimans, Tim, and Diederik P. Kingma, 2016, Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks, in *NIPS*.
- Sarajedini, A., R. Hecht-Nielsen, and P.M. Chau, 1999, Conditional probability density function estimation with sigmoidal neural networks, *IEEE Transactions on Neural Networks* 10, 231–238.
- Sentana, E., 1995, Quadratic ARCH Models, *The Review of Economic Studies* 62, 639–661.

- Shalizi, Cosma, 2011, Estimating Distributions and Densities, Technical report.
- Sheather, S. J., and M. C. Jones, 1991, A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation, *Journal of the Royal Statistical Society* 53, 683–690.
- Silverman, B., 1982, On the estimation of a probability density function by the maximum penalized likelihood method, *The Annals of Statistics* 10, 795–810.
- Sohn, Kihyuk, Honglak Lee, and Xinchun Yan, 2015, Learning Structured Output Representation using Deep Conditional Generative Models, in *Advances in Neural Information Processing Systems*, 3483–3491.
- Sola, J., and J. Sevilla, 1997, Importance of input data normalization for the application of neural networks to complex industrial problems, *IEEE Transactions on Nuclear Science* 44, 1464–1468.
- Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, 2014, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research* 15, 1929–1958.
- Sugiyama, Masashi, and Ichiro Takeuchi, 2010, Conditional density estimation via Least-Squares Density Ratio Estimation, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, 781–788.
- Tansey, Wesley, Karl Pichotta, and James G. Scott, 2016, Better Conditional Density Estimation for Neural Networks.
- Tresp, Volker, 2001, Mixtures of Gaussian Processes, in *NIPS*.
- Trippe, Brian L, and Richard E Turner, 2018, Conditional Density Estimation with Bayesian Normalising Flows, Technical report.
- Webb, A.R., 1994, Functional approximation by feed-forward networks: a least-squares approach to generalization, *IEEE Transactions on Neural Networks* 5, 363–371.
- Whaley, Robert E., 1993, Derivatives on Market Volatility: Hedging Is Long Overdue, *The Journal of Derivatives* .