The Generalized Cross Entropy Method, with Applications to Probability Density Estimation

Zdravko I. Botev and Dirk P. Kroese*

Department of Mathematics
The University of Queensland
Brisbane 4072
AUSTRALIA
E-mail: {botev,kroese}@maths.uq.edu.au

Abstract: Nonparametric density estimation aims to determine the sparsest model that explains a given set of empirical data and which uses as few assumptions as possible. Many of the currently existing methods do not provide a sparse solution to the problem and rely on asymptotic approximations. In this paper we describe a framework for density estimation which uses information-theoretic measures of model complexity with the aim of constructing a sparse density estimator that does not rely on large sample approximations. The effectiveness of the approach is demonstrated through an application to some well-known density estimation test cases.

1. Introduction

The problem of *density estimation* is to find the sparsest probability model which fits a given set of empirical data with the introduction of as little extraneous information as possible.

Many integration problems can be solved efficiently using an appropriate density estimation technique. Examples include:

1. *Monte Carlo integration*, where the problem is to estimate integrals of the form $\int_{\mathscr{X}} H(\mathbf{x}) \, d\mathbf{x}$, for an arbitrary function H and set $\mathscr{X}$. These problems can be efficiently solved by sampling from a good estimate

---

of the density $f(\mathbf{x}) = c\,|H(\mathbf{x})|$, where c is an unknown normalizing constant.

2. *Rare-event simulation*, where a small probability $\ell = \mathbb{P}_h(S(\mathbf{X}) \geqslant \gamma)$ needs to be estimated, for some real-valued function S of a random variable $\mathbf{X}$ with probability density h. This problem is solved efficiently by sampling from an good estimate of the minimum variance importance sampling density [50] $f(\mathbf{x}) = c\,I_{\{S(\mathbf{X}) \geqslant \gamma\}}\,h(\mathbf{x})$, where I denotes the indicator function.

Both of the above problems can be solved efficiently provided one can estimate an optimal (in minimum variance sense) probability density from a given set of empirical data [65]. Thus density estimation is not only an important tool for data analysis, but is also crucial for the performance of many nonparametric population-based Monte Carlo simulation techniques for multidimensional integration [65].

The classical method of density estimation is the parametric approach advocated by Fisher. Here one specifies the model up to a small number of parameters and these are estimated optimally via the likelihood principle. A major problem with this classical paradigm is the problem of specification, in which one has to specify the probability density function. The specification is subjective in the sense espoused by Fisher when he wrote "As regards the problem of specification, these are entirely a matter for the practical statistician,..." [34]. Moreover, it is hard to verify the validity of the parametric model assumptions. For example, [53] argues that with large samples, goodness-of-fit tests almost always reject quite reasonable models. Bayesian statistics is usually also a parametric approach, because in most applications a functional form for the model is assumed.

The non-parametric approach to statistical modelling, initiated by Pearson, takes a more direct path, by trying to estimate the entire probability density, rather than a few parameters of a subjectively specified function. Currently the most popular non-parametric approach to density estimation is the *kernel approach* (for a general introduction see [53], [64], [57]) with its many different flavors (see, e.g., [14], [39], [37], [52], [35], [59], [1]). One of the disadvantages of the kernel approach is that it does not provide a sparse probability model for the data - in many cases the shape described by the kernel estimator can be mimicked by a mixture model with only a few components while the kernel density estimator is a mixture with as many components as the number of data points. This lack of sparsity makes any subsequent inference/analysis of the resulting model computationally intensive. Furthermore, the problem of bandwidth estimation is still a very

contentious issue. The most popular data driven bandwidth selection techniques are only justified asymptotically [27] and have been criticized for obscuring important features in the data [35]. More importantly, as pointed out in [36], to derive the asymptotic rates of convergence of the estimator one makes incompatible assumptions about the smoothness of the underlying unknown density. The most popular non-asymptotic (i.e., not derived using asymptotic arguments) technique is the *Least Squares Cross Validation* method, which has been criticized for its large sampling variability, slow rate of convergence and the concomitant necessity of computationally intensive multi-modal optimization [36], [27].

Mixture modelling is another popular approach which combines elements of the parametric and nonparametric models [43]. Mixtures behave like a parametric model when the number of components is specified in advance. If, however, the number of components is allowed to vary and depend on the data, then the mixture model behaves more like a nonparametric kernel density estimator in the sense that the number of components acts as a smoothing parameter similar to the bandwidth. Just as a large bandwidth reduces the variance of the estimator at the expense of greater bias, small number of components in the mixture model will result in smaller variance, but greater specification bias. Alternatively, a larger number of components reduces the specification bias, but leads to greater variance of the density estimator. Thus mixtures yield models that range on a spectrum with parametric models at one end and nonparametric models at the other end. A lot of research has focused on the maximum likelihood data-driven selection of the number of components in mixtures and similar to the bandwidth selection problem, the issue is still not resolved completely satisfactorily. Lack of necessary regularity conditions [60] precludes the use of likelihood ratio tests for the number of components. There exist bootstrap approaches, Bayesian approaches such as the BIC, and information-theoretic approaches such as the AIC for the optimal selection of the number of components [41, 43]. The performance of these methods, however, has been contentious for various theoretical and practical reasons [60, 12, 4].

The aim of this paper is to describe a different approach to density estimation, one which uses information-theoretic concepts and functional optimization. The proposed procedure can be viewed as a method for fitting a mixture model where the number of components is automatically selected. Alternatively, it can be viewed as a kernel density estimator with weights and data-driven bandwidth selection.

An advantage of the proposed approach is that it is *non-asymptotic* in the sense that we compute a smoothing bandwidth in a manner that does not

rely on asymptotic expansions. An additional advantage is that it provides a *sparse* model for the data - most of the weights for our density estimator are exactly zero, which is in contrast to standard kernel density estimators, where all weights (as many as there are data points) are strictly positive. Furthermore, if we view our estimator as a mixture model with common variance, the selection of the number of components is automatic.

Our approach was motivated by the work of Kapur and Kesavan [30] and the well-known *Cross Entropy* (CE) method [50, 49], and will be referred to as the *Generalized Cross Entropy* method (GCE) due to its emphasis on using generalized CE measures, as opposed to using the traditional Shannon and Kullback-Leibler information measures which have been extensively used.

Regarding the literature dealing with CE principles we refer to the classical works of Havrda-Charvát [24], Kullback-Leibler [33], Shannon [55], Jaynes [26], and, more recently, Kapur and Kesavan [32, 29, 30, 28, 31, 38]. The Generalized measures of distance between pdfs pioneered by Havrda and Charvat [24] and advocated by Kapur and Kesavan [30] have later been used successfully in statistical mechanics by Tsallis [61] and in neural networks [16, 44], but so far have not been used in nonparametric or semi-parametric density estimation. Although the GCE principles were laid out by Kapur and Kesavan [29], a successful application to nonparametric continuous density estimation is missing. This paper fills the gap by demonstrating how these generalized entropy principles can be applied to the problem of probability density estimation and serves as a prequel to [8], where the same ideas are applied to classifying binary data.

The rest of the paper is organized as follows. In Section 2 we formulate the general postulates on which the GCE method is based. A generic GCE algorithm is described in Section 3. Section 4 deals with an applications of the GCE method to density estimation; corresponding numerical results are given in Section 5. Finally, in Section 6 we formulate our conclusions.

2. The Cross Entropy Postulate

The GCE principles are described in [30]. Here we review the main ideas briefly. Similar to the Bayesian approach, the GCE approach starts with a prior probability density about the empirical data. This prior density is then updated in view of the empirical data by minimizing a loss (or risk) function criterion. This Bayesian like approach is summarized in the *CE Postulate* (see [30] and [28]).

The Cross Entropy Postulate:
Given any three of:

1. a prior probability density p,
2. a generalized Cross Entropy distance $\mathcal{D}$ (also known as relative/directed divergence) between two probability densities,
3. a finite set $\mathscr{C}$ of constraints connecting the probability model with the data,
4. a posterior density g,

then under suitable conditions the fourth entity can be found uniquely.

Here $\mathcal{D}$ in combination with $\mathscr{C}$ serves a similar purpose as the risk or loss function in Bayesian inference. The difference is that in the Bayesian setting we are usually estimating a parameter using a loss function, while here we are estimating a density function. Similar to the Bayesian paradigm, the approach provides a framework for updating a prior density function in view of newly available data. Note that [30] and [28] consider the problem of finding any one of the four entities from the other three, but here we will only be concerned with finding the posterior density g given the prior p, the distance measure $\mathcal{D}$, and the constraints $\mathscr{C}$. We now specify each one of these in order to determine g.

The Prior Probability Density

The GCE method assumes that the proposal probability density is updated iteratively. The prior density p at the current iteration is the posterior density from the previous iteration. The prior density which is used to initialize the iteration is the uniform density over the region of interest. In some cases the prior density is the improper uniform density and the normalizing constant over the region of interest is not strictly computable. Similar to the Bayesian methodology, the GCE takes in that case $p(\mathbf{x}) \propto 1$, $\forall \mathbf{x} \in \mathscr{X}$, without any reference to the value of the normalizing constant. The GCE always takes the uniform density, *improper* or otherwise, as the most unbiased and uninformative prior density, in accordance with Laplace's *Principle of Insufficient Reason* [32],[30], which argues that the uniform density is the most unbiased and objective density in the absence of any information about the analyzed probabilistic system. Note that this is different from the Bayesian approach, which uses so-called uninformative *Jeffrey's priors* — densities defined over the space Θ of a model parameter θ and often different from the uniform density over the set Θ. In the GCE method we deal directly with the most uninformative density over the space of the observables, i.e., the uniform density over $\mathscr{X}$.

The Cross Entropy distance $\mathcal{D}$

We use the notion of Cross Entropy distance (directed divergence) between two probability densities. We restrict our attention to the class of directed divergence measures first analyzed by Csiszár [15]. These measures constitute a direct generalization of the most widely used and computationally tractable information-theoretic measures. A distinguishing property of these measures is their convexity.

Definition 1 (Csiszár Measure). The Csiszár generalized measure of directed divergence between two continuous probability densities g and p is:

$$\mathcal{D}(g \to p) \;\; = \;\; \int p(\mathbf{x}) \, \psi\left(\frac{g(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} \;, \quad \mathbf{x} \in \mathbb{R}^d,$$

where

1. $\psi : \; \mathbb{R}^+ \to \mathbb{R}$ is a continuous twice-differentiable function;
2. $\psi(1) = 0$;
3. $\psi''(x) > 0$ for all $x \in \mathbb{R}^+$.

There are no conceptual differences for the case in which g and h are discrete densities. The integral is simply replaced by the sum: $\sum_i p_i \, \psi\left(\frac{g_i}{p_i}\right)$.

Csiszár's family of measures subsumes all of the information-theoretic measures used in practice [30], [24]. For example, if $\psi(x) = \frac{x^\alpha - x}{\alpha(\alpha-1)}, \quad \alpha \neq 0, 1$, for some parameter α, then the family of divergences indexed by α

$$\mathcal{D}_\alpha(g \to p) = \frac{1}{\alpha(\alpha - 1)} \left(\int g^\alpha(\mathbf{x}) p^{1-\alpha}(\mathbf{x}) \, d\mathbf{x} - 1 \right) \tag{1}$$

includes the *Hellinger distance* for $\alpha = 1/2$, *Pearson's χ^2 discrepancy measure* for $\alpha = 2$, *Neymann's χ^2 measure* for $\alpha = -1$, the *Kullback-Leibler distance* in the limit as $\alpha \to 1$, and *Burg CE distance* as $\alpha \to 0$.

In this paper we will use the Pearson χ^2 measure for the construction of the density estimator as opposed to the more traditional Burg and Kullback-Leibler measures. The reasons for this are twofold. First, Burg and Kullback-Leibler measures lack resistance to outliers and robustness with respect to model misspecification. Scott [54] shows the unpredictable influence that outliers can have on fitting mixture models using Burg's measure (equivalent to likelihood maximization) and argues that L_2 measures such as the

χ^2 measure, while not the most efficient asymptotically, are robust. Another study which shows the deficiencies of the Kullback-Leibler measure is Hall's study [22]. It shows that, unlike the L_2 measure, the Kullback-Leibler measure is extraordinarily sensitive to the tails of the target density and thus minimization of the Kullback-Leibler distance can fail to provide a reliable density estimator in the presence of outliers. All studies indicate that while Burg and Kullback-Leibler measures lack robustness, whenever they work, they provide asympotically efficient estimators. Thus the index α in the divergence (1) provides a continuous range of measures on a scale from the very robust to the asymptotically efficient. At one end of the scale is K-L divergence and on the other end of the scale is the robust χ^2. Added to these considerations is the fact that Pearson's χ^2 measure dominates the Kullback-Leibler and L_1 measures:

$$2\,\mathcal{D}_2(g \to p) \geqslant \lim_{\alpha \to 1} \mathcal{D}_\alpha(g \to p),$$

$$2\,\mathcal{D}_2(g \to p) \geqslant \left(\int |g(\mathbf{x}) - p(\mathbf{x})|\, d\mathbf{x} \right)^2.$$

Thus, minimizing Pearson's χ^2 measure is guaranteed to reduce both the asymptotically efficient K-L measure and the robust and dimensionless L_1 metric [18].

Second, we favor the Pearson χ^2 measure due to its computational tractability over all other members in the family (1). L_2 measures have enjoyed a long tradition in nonparametric density estimation primarily due to their computational tractability [54, 40].

Now that we have a reasonable choice for the second ingredient of the CE postulate we comment briefly on the third ingredient.

The Constraint Set $\mathscr{C}$

For the purposes of the GCE method the posterior (proposal) density g is required to satisfy a finite number of linear integral constraints of the form:

$$\mathbb{E}_g K_i(\mathbf{X}) \geqq \mathbb{E}_f K_i(\mathbf{X}), \quad i = 1, \dots, n, \tag{2}$$

where $\{K_i\}_{i=1}^n$ is a set of suitably chosen functions and f is the target density, which solves a statistical or simulation problem. For example, each K_i can be a Gaussian density and f can be the optimal Importance Sampling density for rare-event simulation [50]. Note that the CE postulate gives us a consistent updating rule when the prior density, the constrains and the distance

metric have been specified. It does not, however, provide any guidance as to the choice of the constraints or CE distance in the first place. Our choice of $\mathscr{C}$ will be guided by the following considerations:

1. If the expectations $\mathbb{E}_f K_i(\mathbf{X})$ have to be estimated from empirical data, then the corresponding estimators $\hat{\kappa}_i$ should be asymptotically efficient. I.e., $\hat{\kappa}_i$ should preferably be the Maximum Likelihood Estimator of $\mathbb{E}_f K_i(\mathbf{X})$.
2. The computation of $\hat{\kappa}_i$ should be easy. For example, a computationally manageable and reliable estimate of $\mathbb{E}_f K_i(\mathbf{X})$ may be the Monte Carlo average $\hat{\kappa}_i = \frac{1}{J} \sum_{j=1}^{J} K_i(\mathbf{X}_j)$, where $\mathbf{X}_1, \ldots, \mathbf{X}_J \sim f$.

The constraints (2) connecting the probabilistic model g with the observed behavior of the system (as given by the target f) embody nothing more than a generalization of Pearson's *moment matching* method. We match the generalized moments of the proposed model $\mathbb{E}_g K_i(\mathbf{X})$ to the corresponding empirical moments $\hat{\kappa}_i$ (which approximate the true but unknown $\mathbb{E}_f K_i(\mathbf{X})$). Given the prior density, the CE distance, and the constraints, we now show how to obtain the posterior density g via the CE postulate.

3. The GCE Algorithm

In this section a quite general iterative algorithm for stochastic optimization and machine learning is presented. Suppose that at a given step of the iterative procedure we have a given prior sampling density p which we wish to update on the basis of empirical data, with the aim of obtaining a better probability model for the unknown process which generated the data. Furthermore, let the target density that solves the simulation, optimization or learning problem be denoted by f (e.g., f could be the optimal Importance Sampling density). Then the prior density p is updated to g using the CE postulate with the following ingredients:

1. Given the prior density p on the set $\mathscr{X} \subset \mathbb{R}^d$,
2. minimize the Csiszár measure of Cross Entropy :

$$\mathcal{D}(g \to p) = \int_{\mathscr{X}} p(\mathbf{x})\, \psi\left(\frac{g(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}$$

in terms of the density g, where $\mathbf{x} \in \mathbb{R}^d$ is a column vector.
In other words, we have to solve the functional optimization problem:

$$\min_{g \in \mathscr{G}} \mathcal{D}(g \to p), \tag{3}$$

where $\mathscr{G} = \left\{ g : \int g(\mathbf{x}) \, d\mathbf{x} = 1, \ g(\mathbf{x}) \geqslant 0, \forall \mathbf{x} \in \mathscr{X} \right\}$ is the set of all bona fide density functions on $\mathscr{X}$,

3. subject to the *generalized moment* constraints:

$$\mathbb{E}_g K_i(\mathbf{X}) = \int_{\mathscr{X}} g(\mathbf{x}) \, K_i(\mathbf{x}) \, d\mathbf{x} \ \geqslant \ \hat{\kappa}_i \, , \quad i = 1, \dots, n. \qquad (4)$$

Here

1. $\hat{\kappa}_i$ is a stochastic estimate (obtained from Monte Carlo simulation) or a deterministic estimate (obtained from, e.g., a quadrature procedure, if f is known) of $\mathbb{E}_f K_i(\mathbf{X})$,
2. each $K_i : \mathbb{R}^d \to \mathbb{R}$ is an absolutely continuous function. We refer to the $\{K_i\}$ as *kernel* functions.

Typically the GCE method assumes that each kernel K_i has the properties:

a) $\int_{\mathscr{X}} K_i(\mathbf{x}) \, d\mathbf{x} = 1, \quad K_i(\mathbf{x}) \geqslant 0$ for all $\mathbf{x} \in \mathbb{R}^d$,
b) $K_i(\mathbf{x}) = K(\mathbf{x}; \mathbf{x}_i, \sigma^2)$, so that each kernel K_i has a fixed functional form with a common scale parameter σ^2 and a variable location parameter $\mathbf{x}_i$. The location parameters $\mathcal{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are (usually independent and identically distributed) realizations from the prior density p or, if possible, from the target f. The parameter σ is referred to as the *bandwidth*. For example,

$$K_i(\mathbf{x}) = K(\mathbf{x}; \mathbf{x}_i, \sigma^2) = |\sigma|^{-d/2} \phi\left(\frac{(\mathbf{x} - \mathbf{x}_i)}{\sigma} \right), \quad \mathbf{x} \in \mathbb{R}^d,$$

where $\phi(\mathbf{x}) = (2\pi)^{-d/2} \exp(-\mathbf{x}^T \mathbf{x}/2)$, gives the popular Gaussian kernel with bandwidth σ.

In many cases we choose the kernels to be highly localized functions which are non-zero only in a small neighborhood of the observations at which they are anchored. This can save computational resources when evaluating the kernels and minimizing the CE distance.

Remark 1 (Choice of Constraints). Our choice of constraints is guided by the consistency properties of non-parametric estimators. The constraints $\mathscr{C}$ include the whole empirical sample $\mathcal{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ because it represents all of the available information about the unknown pdf.

One reason for choosing inequality constraints is to make sure that g dominates the unknown f by assigning probability mass in the neighborhood of each point $\mathbf{x}_i$ at least as large as the true (or the estimated) mass. This makes

g a good proposal density for an Acceptance Rejection algorithm designed to simulate from f. Another reason for choosing inequality constraints is that they allow us to handle the non-negativity restriction $g(\mathbf{x}) \geqslant 0$ in $\mathscr{G}$. Later in the paper it will become clear that choosing inequality constraints, as opposed to simple equality constraints, allow us to use the full power of the Karush-Kuhn-Tucker duality to deduce a novel bandwidth selection rule. Moreover, as demonstrated in the examples in the last section, with the inequality constraints the optimal model g exhibits model sparsity similar to that observed in *Support Vector Machines* [62].

Remark 2 (Non-negativity of Density)**.** Note that for some choices of ψ the non-negativity constraint $g(\mathbf{x}) \geqslant 0$ in $\mathscr{G}$ need not be imposed explicitly. If $\psi(x) = x \ln(x)$, corresponding to the Kullback–Leibler distance, the condition $g(\mathbf{x}) \geqslant 0$ is automatically satisfied. In general, however, the non-negativity constraint has to be enforced explicitly in the functional optimization program.

Remark 3 (Comparison with other Entropy methods)**.** Note that the GCE method solves a *functional* optimization problem to find the optimal posterior density $g(\mathbf{x})$. In contrast, population-based Monte Carlo methods like the CE method [50], and maximum likelihood methods (which essentially minimize Shannon's information measure) solve the *parametric* optimization problem

$$\min_{\boldsymbol{\theta}} \mathcal{D}\big(g(\cdot; \boldsymbol{\theta}) \to f\big)$$

to find the optimal density $g(\mathbf{x}; \boldsymbol{\theta})$ within a pre-specified family of densities $\big\{g(\cdot; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\big\}$ indexed by $\boldsymbol{\theta}$. Instead of specifying the functional form of the density in advance, in the GCE method we specify the generalized moments in (4). Note that given a density function f, we can determine an infinite number of moments $\{\mathbb{E}_f K_i(\mathbf{X})\}_{i=-\infty}^{i=\infty}$, but a finite number of moments are not enough to reconstruct the underlying density. Thus the GCE approach assumes less prior information about the unknown density f.

The problem (3) + (4) above is a constrained functional optimization problem. More specifically, without the *algebraic* constraint $g(\mathbf{x}) \geqslant 0$, it is an *isoperimetric* Calculus of Variations problem with integral equality and/or inequality constraints. Since ψ is strictly convex by assumption, the functional (2) is strictly convex and we can use the theory of duality to simplify the problem. There are a number of difficulties which have to be overcome, however. Since we have inequality constraints we cannot use the standard Lagrangian methods. Moreover, the standard duality techniques for convex

optimization, [6], [11], apply to finite dimensional optimization problems, i.e., problems in which instead of integrals one has sums with summation over a finite countable set. The optimization problem that we face is thus not a standard one. The application of the duality theory to infinite dimensional optimization problems is subtle and has been only recently established. We apply the formalism of this duality theory to our specific information-theoretic optimization problem omitting technical details. For a rigorous justification of this duality formalism with the necessary technical detail, the reader is referred to [5], [17] and [3].

3.1. The Dual Optimization Problem

The isoperimetric problem, (3) + (4), obtained in the previous section is convex and hence there is a dual formulation. In this case the dual problem is much easier to solve than the primal problem. This is essentially the reason why the strict convexity condition is imposed in the definition of the CE measures. In our case, let the **Primal Problem** be:

$$\min_{g} \ \mathcal{D}(g \to p)$$

$$\text{subject to:} \quad \int g(\mathbf{x}) \, K_i(\mathbf{x}) \, d\mathbf{x} \geqslant \hat{\kappa}_i , \quad i = 1, \dots, n \tag{5}$$

$$\int g(\mathbf{x}) \, d\mathbf{x} = 1$$

Note that the algebraic constraint $g(\mathbf{x}) \geqslant 0$, $\mathbf{x} \in \mathbb{R}^d$ is not included in the formulation of the primal problem. For the time being we assume that the non-negativity constraint is satisfied by the solution of the primal and need not be imposed. To derive the dual corresponding to the primal, define the Lagrangian:

$$\mathcal{L}(g; \boldsymbol{\lambda}, \lambda_0) =$$

$$= \int p(\mathbf{x}) \, \psi\left(\frac{g(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} - \lambda_0 \left(\int g(\mathbf{x}) \, d\mathbf{x} - 1\right) - \sum_{i=1}^{n} \lambda_i \left(\int g(\mathbf{x}) \, K_i(\mathbf{x}) \, d\mathbf{x} - \hat{\kappa}_i\right)$$

$$= \sum_{i=0}^{n} \lambda_i \, \hat{\kappa}_i + \int \left[p(\mathbf{x}) \, \psi\left(\frac{g(\mathbf{x})}{p(\mathbf{x})}\right) - g(\mathbf{x}) \sum_{i=0}^{n} \lambda_i \, K_i(\mathbf{x}) \right] d\mathbf{x},$$

where for convenience we define $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_n]^T$, $\hat{\kappa}_0 = 1$ and $K_0(\cdot) = 1$.

Then, (see, e.g., [5] and [17]) the **Dual Problem** is:

$$\max_{\lambda,\lambda_0} \left\{ \inf_g \mathcal{L}(g;\lambda,\lambda_0) \right\} \tag{6}$$

$$\text{subject to:} \qquad \lambda \geqslant \mathbf{0} \tag{7}$$

The dual can be simplified substantially. First $\inf_g \mathcal{L}(g;\lambda,\lambda_0)$ can be calculated explicitly using the Euler-Lagrange equation [63]. In this particular case the Euler-Lagrange equation yields:

$$\psi'\left(\frac{g(\mathbf{x})}{p(\mathbf{x})}\right) = \sum_{k=0}^{n} \lambda_k\, K_k(\mathbf{x})\ . \tag{8}$$

Since $\psi''(x) > 0$ for $x > 0$, the function $\psi'(x)$ has a unique inverse on the domain $x \in \mathbb{R}^+$. The functional form of the solution of the Euler-Lagrange equation can thus be written explicitly as:

$$g(\mathbf{x}) = p(\mathbf{x})\, \psi'^{-1}\left(\sum_{k=0}^{n} \lambda_k\, K_k(\mathbf{x})\right). \tag{9}$$

We can then substitute this $g(\mathbf{x})$ into the Lagrangian to obtain:

$$
\begin{aligned}
\mathcal{L}^*(\lambda,\lambda_0) \;=\;& \inf_g \mathcal{L}(g;\lambda,\lambda_0) \\[2mm]
=\;& \sum_{i=0}^{n} \lambda_i\, \hat{\kappa}_i + \mathbb{E}_p\, \psi\left(\psi'^{-1}\left(\sum_{k=0}^{n} \lambda_k\, K_k(\mathbf{X})\right)\right) \\[2mm]
& - \sum_{i=0}^{n} \lambda_i\, \mathbb{E}_p\, K_i(\mathbf{X})\, \psi'^{-1}\left(\sum_{k=0}^{n} \lambda_k\, K_k(\mathbf{X})\right).
\end{aligned}
$$

Thus, the dual becomes:

$$\max_{\lambda,\lambda_0} \quad \mathcal{L}^*(\lambda,\lambda_0)\ , \tag{10}$$

$$\text{subject to:} \qquad \lambda \geqslant \mathbf{0}\ . \tag{11}$$

Further simplification of $\mathcal{L}^*$ is possible if we set $\Psi' = \psi'^{-1}$ and observe that straightforward integration by parts yields:

$$\Psi(x) \;=\; x\, \Psi'(x) - \psi\left(\Psi'(x)\right) + \text{constant}\ .$$

Then $\mathcal{L}^*$ can be written compactly as:

$$\mathcal{L}^*(\boldsymbol{\lambda}, \lambda_0) = \sum_{i=0}^{n} \lambda_i \, \hat{\kappa}_i - \mathbb{E}_p \, \Psi\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right), \tag{12}$$

where the constant of integration is ignored as it is irrelevant to the optimization problem. We can finally state the simplest form of the **Dual Problem**:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \quad \sum_{i=0}^{n} \lambda_i \, \hat{\kappa}_i - \mathbb{E}_p \, \Psi\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right) \tag{13}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}. \tag{14}$$

The solution of the **Primal Problem** is obtained via the transformation:

$$g(\mathbf{x}) = p(\mathbf{x}) \, \Psi'\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right). \tag{15}$$

At this stage it is important to note that our primal problem is strictly convex. It is well known [63] that if the primal problem is (strictly) convex the dual problem is (strictly) concave and the solution of the primal, which is a (unique) minimizer, coincides exactly with the solution of the dual— a (unique) maximizer. Thus, the solution of our dual problem must be unique.

Important quantities for the optimization are the *gradient* and the *Hessian* of $\mathcal{L}^*$:

$$\frac{\partial \mathcal{L}^*}{\partial \lambda_i} = \hat{\kappa}_i - \mathbb{E}_p \, \Psi'\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right) K_i(\mathbf{X}) \tag{16}$$

$$\frac{\partial^2 \mathcal{L}^*}{\partial \lambda_i \partial \lambda_j} = -\mathbb{E}_p \, K_i(\mathbf{X}) \, \Psi''\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right) K_j(\mathbf{X}), \tag{17}$$

where $i, j \in \{0, 1, \ldots, n\}$. Since there are no constraints on λ_0, the gradient with respect to λ_0 has to be zero:

$$\frac{\partial \mathcal{L}^*}{\partial \lambda_0} = 1 - \mathbb{E}_p \, \Psi'\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right) = 0. \tag{18}$$

Note that if the *characterizing moment* constraints (4) are strict equalities instead of inequalities then the restriction $\boldsymbol{\lambda} \geqslant \mathbf{0}$ is omitted. Thus, with strict

equality constraints the dual optimization problem is:

$$\max_{\boldsymbol{\lambda},\lambda_0} \mathcal{L}^* = \max_{\boldsymbol{\lambda},\lambda_0} \quad \sum_{i=0}^{n} \lambda_i \, \hat{\kappa}_i - \mathbb{E}_p \, \Psi \left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X}) \right), \qquad (19)$$

though we may still have to enforce $g(\mathbf{x}) \geqslant 0, \forall \mathbf{x} \in \mathbb{R}^d$ explicitly. Special cases of (19) are given in the following examples.

Example 1 (The MCE method [49]). Choose $\psi(x) = x \ln(x) - x$, then $\psi'^{-1}(x) = \exp(x) = \Psi'(x) = \Psi(x)$, $g(\mathbf{x}) = p(\mathbf{x}) \exp \left(\sum_{k=0}^{n} \lambda_k K_k(\mathbf{x}) \right) \geqslant 0$ and $\mathcal{D}(g \to p) = \int g(\mathbf{x}) \ln \left(g(\mathbf{x})/p(\mathbf{x}) \right) d\mathbf{x} - 1$. The Lagrange multipliers are determined from the maximization of the dual (19). In this case, since there are no constraints on $\boldsymbol{\lambda}$ and λ_0, the unconstrained maximization of the strictly concave $\mathcal{L}^*$ leads to the set of non-linear equations for $\nabla_{\boldsymbol{\lambda}} \mathcal{L}^* = \mathbf{0}$:

$$\mathbb{E}_p \exp \left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X}) \right) K_i(\mathbf{X}) = \hat{\kappa}_i \,, \quad i = 0, \dots, n \qquad (20)$$

The solution gives the unique optimal $g(\mathbf{x})$ for the MCE method. In summary the MCE method chooses the proposal density (note that we have substituted for λ_0)

$$g(\mathbf{x}) = \frac{p(\mathbf{x}) \exp \left(\sum_{k=1}^{n} \lambda_k \, K_k(\mathbf{x}) \right)}{\mathbb{E}_p \exp \left(\sum_{k=1}^{n} \lambda_k \, K_k(\mathbf{X}) \right)} \qquad (21)$$

from the General Exponential Family [46] and then minimizes, without any constraints, what appears to be a distance measure:

$$\min_{\boldsymbol{\lambda}} \quad -\mathcal{L}^*(\boldsymbol{\lambda}) = \mathbb{E}_f \ln \frac{p(\mathbf{X})}{g(\mathbf{X})}.$$

An advantage of the MCE method is that $\hat{\kappa}_i = \frac{1}{J} \sum_{j=1}^{J} K_i(\mathbf{X}_j)$, with $\mathbf{X}_1, \dots, \mathbf{X}_J \sim f$, is the asymptotically efficient (i.e. Maximum likelihood) estimator of $\mathbb{E}_f K_i(\mathbf{X})$. This is a consequence of the fact that g in this case belongs to the General Exponential Family. The salient features of the MCE method can be summarized as follows:

1. In the MCE method the dual (19) of the primal functional optimization problem becomes a concave Geometric Programming Problem.

2. The expectations on the left-hand of (20) side can rarely be calculated analytically and thus have to be estimated via an empirical average to give the *stochastic counterpart* of (20):

$$\frac{1}{n} \sum_{j=1}^{n} \exp\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X}_j)\right) K_i(\mathbf{X}_j) = \hat{\kappa}_i, \quad \{\mathbf{X}_j\}_{j=1}^{n} \sim p, \quad i = 0, \ldots, n.$$

3. Simulation from (21) and any other member of the General Exponential Family is in general a difficult problem. Usually the Accept-Reject method or Markov Chain Monte Carlo methods are used to sample from (21).
4. The non-negativity of (21) is ensured by its exponential functional form. This makes the optimization easier.
5. If f does not belong to the General Exponential Family, then the MCE optimal density (21) may not converge to f as $n \to \infty$ (see [7]). Thus the functional form of (21) is not optimal.
6. While the functional form of (21) is not asymptotically optimal, the estimation of the characterizing moments $\mathbb{E}_g K_i(\mathbf{X})$ through $\hat{\kappa}_i = \sum_{j=1}^{J} K_i(\mathbf{X}_j)$, $\mathbf{X}_1, \ldots, \mathbf{X}_J \sim f$, is asymptotically optimal.

Example 2 (The CE method [50]). If in the CE method a proposal density is chosen from the General Exponential Family $g(\mathbf{x}) \propto \exp\left(\sum_{k=1}^{n} \lambda_k \, K_k(\mathbf{x})\right)$, then Maximizing the Likelihood $\sum_{j=1}^{J} \ln g(\mathbf{X}_j)$, where $\mathbf{X}_1, \ldots, \mathbf{X}_J \sim f$, gives the CE updating equations ($i = 0, \ldots, n$):

$$\frac{\int \exp\left(\sum_{k=1}^{n} \lambda_k \, K_k(\mathbf{x})\right) K_i(\mathbf{x}) \, d\mathbf{x}}{\int \exp\left(\sum_{k=1}^{n} \lambda_k \, K_k(\mathbf{x})\right) d\mathbf{x}} = \frac{1}{J} \sum_{j=1}^{J} K_i(\mathbf{X}_j) = \hat{\kappa}_i, \quad \{\mathbf{X}_j\}_{j=1}^{J} \sim f$$

for the parameters $\{\lambda_i\}_{i=0}^{n}$. Maximizing the Likelihood is approximately the same as minimizing Kullback–Leibler CE distance $\mathbb{E}_f \ln(f(\mathbf{X})/g(\mathbf{X}))$ between f and g. Minimization the Kullback–Leibler CE distance is the highlighting feature of the CE method. We conclude that the updating rules of the CE method (see [50] pages 68, 69 and Example 3.5) coincide with the updating rules of the GCE method in cases where

1. the CE method chooses a sampling/proposal density g from the General Exponential Family with *natural parameters* $\{\lambda_k\}_{k=1}^{n}$ and *natural statistics* $\{K_k(\mathbf{x})\}_{k=1}^{n}$ (see [46] page 95) and
2. the GCE method uses the convex $\Psi(x) = \exp(x)$ in (19).

The updating rules between the two methods do not agree under any other conditions. Note that the Maximum Likelihood estimators of parameters of densities in the General Exponential Family achieve the Cramer-Rao lower bound (see [46] page 223). This makes the simple estimator $\hat{\kappa}_i = \frac{1}{J} \sum_{j=1}^{J} K_i(\mathbf{X}_j)$, $\{\mathbf{X}_j\}_{j=1}^{J} \sim f$ the Minimum Variance Unbiased Estimator of $\mathbb{E}_f K_i(\mathbf{X})$. This is the advantage of using a proposal density from the General Exponential Family. Note, however, that typically one has random variables from the prior p and not from the target f. In this case the CE method uses the Likelihood Ratio (LR) estimator

$$\hat{\kappa}_i = \frac{\sum_{j=1}^{J} W(\mathbf{X}_j) K_i(\mathbf{X}_j)}{\sum_{j=1}^{J} W(\mathbf{X}_j)}, \quad W(\mathbf{X}_j) = \frac{f(\mathbf{X}_j)}{p(\mathbf{X}_j)}, \quad \{\mathbf{X}_j\}_{j=1}^{J} \sim p. \quad (22)$$

Since (22) no longer follows from the Maximum Likelihood Principle [46], the optimality of the LR estimator (22) is dubious and still an unresolved problem.

3.2. The choice for ψ

Our present aim is to choose the function ψ in Csiszár's measure such that:

1. The expectation in (13) can be evaluated analytically or at least without too much trouble.
2. Maximizing (13)+(14), and hence finding the set of Lagrange multipliers $\{\lambda_k\}_{k=0}^{n}$, is relatively easy. E.g., if ${\psi'}^{-1} = \Psi'$ are linear then (16) is linear in the Lagrange multipliers, and the Hessian matrix (17) is constant. This can greatly simplify the optimization.
3. Generating random variables from the optimal density g in (9) is relatively easy. E.g., if Ψ' is linear then g is a discrete mixture and the *composition method* (also known as the *convolution method*) for random variate generation applies [51].

Satisfying these requirements simultaneously is only possible for few specific choices of ψ. In particular we can choose Ψ' to be linear. Then ψ' is linear and the definition of Csiszár's measure requires:

$$\begin{aligned} \psi'(x) &= ax + b \\ \psi''(x) &> 0 \\ \psi(1) &= 0, \end{aligned}$$

hence :

$$\psi(x) = \frac{a}{2}(x^2 - 1) + b(x - 1), \quad a > 0$$

for arbitrary constants $a > 0$ and b. Note that the linear term $b(x - 1)$ is irrelevant and hence is omitted. Thus Csiszár's measure can be written as:

$$
\begin{aligned}
\mathcal{D}(g \to p) &= \frac{a}{2} \int p(\mathbf{x}) \left(\frac{g^2(\mathbf{x})}{p^2(\mathbf{x})} - 1 \right) d\mathbf{x} \\
&= -\frac{a}{2} + \frac{a}{2} \int \frac{g^2(\mathbf{x})}{p(\mathbf{x})} \, d\mathbf{x} \\
&= \frac{a}{2} \int \frac{\left(g(\mathbf{x}) - p(\mathbf{x}) \right)^2}{p(\mathbf{x})} \, d\mathbf{x}.
\end{aligned}
$$

Note that for optimization purposes the value of a is irrelevant as long as $a > 0$. We will thus choose $a = \frac{1}{2}$ to obtain:

$$\mathcal{D}(g \to p) = \frac{1}{2} \int \left(\frac{g^2(\mathbf{x})}{p(\mathbf{x})} - p(\mathbf{x}) \right) d\mathbf{x},$$

which is Pearson's χ^2 CE distance [24]. The choice $\psi(x) = \frac{1}{2}(x^2 - 1)$ ensures that:

1. $\psi'^{-1}(x) = x = \Psi'(x)$ allowing us to write (13) as a linear combination of integrals/expectations each of which, for various kernel functions K_i, can be evaluated analytically.
2. The Hessian matrix (17) of (13) is independent of the Lagrange multipliers.
3. The resulting density function (9) can be simulated using the *composition method*.

In fact (9) becomes the *smoothed bootstrap filter* density:

$$g(\mathbf{x}) = p(\mathbf{x}) \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}) . \tag{23}$$

In the particle filter context (see [19] page 296) the set $\{\lambda_i\}_{i=0}^{n}$ is the set of Sampling Importance Resampling (SIR) weights. The dual problem (13)+(14) becomes:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \quad -\frac{1}{2} + \sum_{i=0}^{n} \lambda_i \, \hat{\kappa}_i - \frac{1}{2} \sum_{i=0}^{n} \sum_{j=0}^{n} \lambda_i \, \lambda_j \, \mathbb{E}_p K_i(\mathbf{X}) K_j(\mathbf{X}), \tag{24}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}. \tag{25}$$

It is easy to verify that optimization is equivalent to :

$$\min_{\lambda,\lambda_0} \quad \frac{1}{2} \int \frac{\left(g(\mathbf{x}) - f(\mathbf{x})\right)^2}{p(\mathbf{x})} \, d\mathbf{x}, \tag{26}$$

$$\text{subject to:} \quad \lambda \geqslant \mathbf{0}, \tag{27}$$

with $g(\mathbf{x}) = p(\mathbf{x}) \sum_{k=0}^{n} \lambda_k K_k(\mathbf{x})$. Thus this approach is equivalent to choosing a discrete mixture of kernel functions as the sampling density and then minimizing the *projection pursuit index* (26) (see [57], page 129) between the sampling and the target density f. We now proceed to rewrite the dual problem in a form which is easier to interpret. First, since there are no constraints on λ_0 we can solve $\frac{\partial \mathcal{L}^*}{\partial \lambda_0} = 0$ in (16) and determine λ_0 as a function of λ:

$$\lambda_0 = 1 - \mathbb{E}_p \sum_{k=1}^{n} \lambda_k K_k(\mathbf{X}) = 1 - \sum_{k=1}^{n} \lambda_k \mathbb{E}_p K_k(\mathbf{X}).$$

We then substitute for λ_0 to obtain

$$g(\mathbf{x}) = p(\mathbf{x}) + p(\mathbf{x}) \sum_{k=1}^{n} \lambda_k \left(K_k(\mathbf{x}) - \mathbb{E}_p K_k(\mathbf{X}) \right) . \tag{28}$$

The Lagrange multipliers are determined from optimization of the dual:

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i \left(\hat{\kappa}_i - \mathbb{E}_p K_i(\mathbf{X}) \right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \, \lambda_j \, \text{Cov}_p \left(K_i(\mathbf{X}); K_j(\mathbf{X}) \right) ,$$

subject to $\lambda \geqslant \mathbf{0}$. This can be written in matrix notation:

$$\min_{\lambda} \quad \frac{1}{2} \lambda^T C \lambda - \mathbf{c}^T \lambda \tag{29}$$

$$\text{subject to:} \quad \lambda \geqslant \mathbf{0}, \tag{30}$$

where

$$\mathbf{c} \;=\; \hat{\boldsymbol{\kappa}} - \mathbb{E}_p \mathbf{K}(\mathbf{X})$$

$$\text{with} \quad C \;=\; \mathbb{E}_p \left(\mathbf{K}(\mathbf{X}) - \mathbb{E}_p \mathbf{K}(\mathbf{X}) \right) \left(\mathbf{K}(\mathbf{X}) - \mathbb{E}_p \mathbf{K}(\mathbf{X}) \right)^T$$

$$\mathbf{K}(\mathbf{x}) \;=\; \left[K_1(\mathbf{x}) \; K_2(\mathbf{x}) \; \cdots \; K_n(\mathbf{x}) \right]^T$$

$$\hat{\boldsymbol{\kappa}} \;=\; \left[\; \hat{\kappa}_1 \;\; \hat{\kappa}_2 \;\; \hat{\kappa}_3 \; \ldots \; \hat{\kappa}_{n-1} \;\; \hat{\kappa}_n \; \right]^T.$$

Choosing $\psi(x) = \frac{1}{2}(x^2 - 1)$ thus makes the optimization problem (13)+(14) a Quadratic Programming Problem (QPP) for the Lagrange multipliers. Although C may be numerically ill-conditioned, with probability one C is a positive definite symmetric covariance matrix. Therefore the QPP (29) and (24) are strictly convex and the KKT conditions guarantee a unique global extremum for any concave constraints. In particular (29) and (24) have a unique global extremum under the concave constraints (30). The solution (c.f. (23)) in matrix form is:

$$g(\mathbf{x}) = p(\mathbf{x}) \left(\lambda_0 + \boldsymbol{\lambda}^T \mathbf{K}(\mathbf{x}) \right), \tag{31}$$

$$\text{where} \quad \lambda_0 = 1 - \boldsymbol{\kappa}^T \boldsymbol{\lambda}. \tag{32}$$

However, the solution of the QPP is not a pdf because λ_0 can go negative, and (31) will take negative values for some $\mathbf{x}$. This is unacceptable for a probability density function. To resolve this problem we find a value for the bandwidth parameter σ of the kernels $\{K_k\}_{k=1}^n$ so that $\lambda_0 \geqslant 0$. This will ensure that (31) is indeed a proper mixture pdf. In other words, we can find a value of the bandwidth parameter σ such that $\sigma \in \mathscr{S}$, where

$$\mathscr{S} = \left\{ \sigma : \ \mathbb{E}_p \mathbf{K}(\mathbf{X})^T \boldsymbol{\lambda}^* \leqslant 1, \ \boldsymbol{\lambda}^* = \underset{\boldsymbol{\lambda} \geqslant 0}{\operatorname{argmin}} \left[\boldsymbol{\lambda}^T C \boldsymbol{\lambda} - 2 \, \boldsymbol{\lambda}^T \mathbf{c} \right] \right\}$$

and C, $\mathbf{c}$ and $\mathbb{E}_p \mathbf{K}(\mathbf{X})$ depend implicitly on σ through the kernels $\{K(\mathbf{x}; \mathbf{x}_i, \sigma)\}_{i=1}^n$. The set $\mathscr{S}$ is the set of *admissible* bandwidth values in the sense that

$$\sigma \in \mathscr{S} \Leftrightarrow g \in \mathscr{G}.$$

We thus find a solution of the dual QPP such that $\sigma \in \mathscr{S}$, and this ensures that the solution of the primal $g \in \mathscr{G}$.

Remark 4. There are two extreme values for σ. We may either choose σ such that $\lambda_0 = 1$, in which case we assign maximum weight to the prior p; or we can choose σ such that $\lambda_0 = 0$, in which case we eliminate the prior as a mixture component in (31). Values for σ in between these two extremes represent a trade-off between the prior density and the observed empirical data.

In summary $\sigma \in \mathscr{S}$ implies that the solution (31) belongs to $\mathscr{G}$, i.e., is a proper pdf function which is non-negative and integrates to one.

Choosing K

For many choices of the kernel functions $\mathbf{K}$ we can find $\mathrm{Cov}_p\left(K_i(\mathbf{X}); K_j(\mathbf{X})\right) = C_{ij}$ analytically, provided p itself is another linear combination of kernels or a uniform prior. In practice the choice for $\mathbf{K}$ is dictated by the assumptions we make about the smoothness of the target density f. If f is known to be smooth then $\mathbf{K}$ should also be smooth. Naturally the more smooth f is, the easier it is to estimate. Examples of different possible kernels over continuous and discrete spaces for which the matrix C can be calculated analytically are given in [7]. Thus we emphasize that only $\mathbb{E}_f \mathbf{K}(\mathbf{X})$ needs to be estimated, either via a Monte Carlo sample or via standard quadrature methods, and all the other elements of the QPP can be calculated analytically for a wide variety of kernels.

Estimating $\mathbb{E}_f\mathbf{K(X)}$

Assume we use the same set $\mathcal{X}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ as both kernel location parameters and as a sample for the estimation of $\mathbb{E}_f\mathbf{K}(\mathbf{X})$. Note that $\mathbb{E}_f K_i(\mathbf{X})$ is a function of $\mathbf{X}_i$ and hence is a random variable. Under the assumption that $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d}{\sim} f$, each $\mathbf{X}_i$ is independent of all the other $\{\mathbf{X}_j\}_{j\neq i}$ and a simple unbiased estimator of $\mathbb{E}_f K_i(\mathbf{X})$ is:

$$\hat{\kappa}_i = \frac{1}{n-1} \sum_{j\neq i}^{n} K_i(\mathbf{X}_j).$$

This is the *cross-validatory*, also known as *leave-one-out*, estimator and its consistency properties are established in [9]. If f is known but we can not easily generate a sample from it then we can estimate $\mathbb{E}_f K_i(\mathbf{X})$ via the unbiased Importance Sampling estimator:

$$\hat{\kappa}_i = \frac{1}{n-1} \sum_{j\neq i} \frac{f(\mathbf{X}_j)}{p(\mathbf{X}_j)} K_i(\mathbf{X}_j), \quad \mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d}{\sim} p.$$

Here the GCE prior pdf p acts as a proposal density for the Importance Sampling estimation of $\mathbb{E}_f K_i(\mathbf{X})$. In an iterative GCE algorithm the current prior p will be the posterior pdf from the previous iteration.

3.3. Choosing σ

Recall that σ is the single common bandwidth parameter for each of the kernels. The main novelty of the proposed approach lies in the choice of the

bandwidth parameter σ. As discussed previously, we solve the functional optimization program (5) and obtain (31) as the solution. For an arbitrary choice of σ, the solution (31) of the optimization program is *not* a proper non-negative function, because we have not explicitly included the constraint $g(\mathbf{x}) \geqslant 0$ in (5). One can, however, choose a value for the bandwidth σ so that the solution of the program (31) is a bona-fide density, i.e., it integrates to one and is non-negative. We thus exploit the Kuhn-Tucker duality theory to arrive at a novel method for choosing the bandwidth parameter σ. We emphasize that σ is not a parameter which we optimize after solving the QPP (29)+(30), instead the QPP is solved as many times for various σ until we find a value for σ which makes the output of the QPP (31) a proper density. In other words, for a given σ, the QPP yields a function $g(\mathbf{x})$ which depends on σ. We thus obtain a family of solutions of the QPP indexed by σ, with most of the solutions failing to be non-negative functions. Out of this set of solutions we select one which is a proper pdf.

4. Application to Statistical Density Estimation

In this section we apply the GCE method to the problem of probability density estimation and compare it with standard nonparametric methods [56]. Consider the one-dimensional case where we are given the sample $\mathcal{X}_n \equiv \{X_1, \ldots, X_n\}$ on $\mathbb{R}$ and wish to visualize any patterns present in it, compress it or draw inferences based on statistical analysis. One of the most popular approaches to modeling the data $\mathcal{X}_n$ with few stringent assumptions is the *kernel method*. For a comprehensive treatment of the method see [56], [57] and [53]. The method assumes that the true, but unknown, underlying density function f can be approximated well by a probability density function of the form:

$$\widehat{f}(x \mid h, \mathcal{X}_n) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad X_1, \ldots, X_n \overset{i.i.d}{\sim} f(x) \tag{33}$$

where:

1. $h \in \mathbb{R}_+ \backslash \{0\}$ is a *bandwidth* parameter which controls the smoothness or "resolution" of $\widehat{f}$.
2. $K : \mathbb{R} \to \mathbb{R}_+$, $\int_{\mathbb{R}} K(x)\, dx = 1$, $K(-x) = K(x)$, i.e., K is a symmetric unimodal kernel. For our purposes we choose to use the Gaussian kernel $K(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

Everything in (33) is fixed except the bandwidth h. This is the only parameter over which one has control. There are various methods [53] for tuning h so that the approximation of f is as good as possible. Currently the prevailing method for bandwidth selection is the Sheather-Jones (SJ) plug-in estimate [27]. In the next section we will compare the Sheather-Jones estimator with the GCE estimator.

Density Estimation via GCE

For clarity we now restate the crux of the GCE method in the context of one-dimensional density estimation ($d = 1$). Again assume that all we have is the empirical data $\mathcal{X}_n = \{X_1, \dots, X_n\}$. Then apply the GCE postulate with the following elements:

1. Given the uniform/uninformative prior $p \propto 1$ on $\mathbb{R}$,
2. Minimize the χ^2 CE distance $\mathcal{D}(g \rightarrow p) = -\frac{1}{2} + \frac{1}{2} \int \frac{g^2(x)}{p(x)}\,dx$ with respect to the density g, i.e.:

$$\min_{g \in \mathscr{G}} \mathcal{D}(g \rightarrow p) \equiv \min_{g \in \mathscr{G}} \int_{\mathbb{R}} g^2(x)\,dx, \tag{34}$$

3. subject to the constraint set $\mathscr{C}$:

$$\int_{\mathbb{R}} g(x)\,K_i(x)\,dx = \mathbb{E}_g\,K_i(X) \geqslant \hat{\kappa}_i = \frac{1}{n-1} \sum_{j \neq i} K_i(X_j), \quad i = 1, \dots, n. \tag{35}$$

Again $\mathscr{G} = \left\{ g : \int_{\mathbb{R}} g(x)\,dx = 1,\ g(x) \geqslant 0,\ x \in \mathbb{R} \right\}$ denotes the set of all probability density functions on $\mathbb{R}$ and we choose a Gaussian kernel $K_i(x) = K(x; X_i, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-X_i)^2}{2\sigma^2}\right) = \frac{1}{\sigma} \phi\left(\frac{x-X_i}{\sigma}\right)$. We can interpret the program $\min_{g \in \mathscr{G}} \mathcal{D}$ as minimization of the complexity of the proposed probabilistic model g and the imposition of the constraint set $\mathscr{C}$ as a means of ensuring that the model is consistent with the empirical data. The above problem is equivalent to the dual formulation:

1. Solve the program:

$$(\sigma^*, \lambda^*) = \left\{ (\sigma, \lambda) : \mathbf{1}^T \lambda(\sigma) = 1,\ \lambda(\sigma) = \underset{\lambda \geqslant 0}{\operatorname{argmin}} \left(\frac{1}{2} \lambda^T C(\sigma) \lambda - \lambda^T \hat{\boldsymbol{\kappa}}(\sigma) \right) \right\}, \tag{36}$$

where the matrix $C_{n \times n}$ has entries

$$C_{ij} = \int_{\mathbb{R}} K_i(x; X_i, \sigma)\,K_j(x; X_j, \sigma)\,dx = \frac{1}{\sqrt{2}\sigma} \phi\left(\frac{X_i - X_j}{\sqrt{2}\sigma}\right)$$

$$= \frac{1}{\sqrt{2\pi}(\sqrt{2}\sigma)} \exp\left(-\frac{(X_i - X_j)^2}{4\sigma^2}\right).$$

2. Present the Gaussian mixture density

$$g(x) = \sum_{j=1}^{n} \lambda_j^* \, K(x; X_j, \sigma^*) \tag{37}$$

as the optimal GCE density that models the data $\mathcal{X}_n$.

Note that the weighted kernel mixture (37) has been suggested in a different context. Hall & Turlach [23] analyzed the use of weights in density estimation and have successfully applied density estimators of the form (37). Later Girolami & He [20], [21] have applied the same idea to other statistical problems. These papers, however, do not exploit the fact that a weighted kernel mixture can be deduced from a functional optimization program using the χ^2 distance, where the weights can be interpreted as the Lagrange multipliers associated with functional optimization. It is this novel interpretation that justifies the bandwidth selection rule presented here.

5. Numerical Experiments

To illustrate the effectiveness of the proposed method we perform several simulation studies with synthetic data and compare the accuracy of the GCE estimator and the standard kernel density estimator [27]. In addition, we consider density estimation on a number of well-known real world datasets.

Table 1 describes the simulation experiments with synthetic data over 10 independent repetitions. The second column gives the analytical pdf from which the random samples were drawn. Some of the Gaussian mixture models are borrowed from [40]. The third column gives the number of data points generated from the analytical pdf. In order to asses the relative performance of the GCE estimator we use the ratio of the exact integrated squared errors:

$$\int \left(g(x) - f(x) \right)^2 dx \Big/ \int \left(\widehat{f}(x) - f(x) \right)^2 dx,$$

where $f(x)$ is the analytical pdf, $g(x)$ is the GCE density estimator, and $\widehat{f}(x)$ is the SJ kernel density estimator [27]. The last column of Table 1 gives the numerical value of the ratio of integrated squared errors.

From Table 1 we can conclude that the proposed estimator performs favorably compared to the standard kernel estimation method [27]. For

case study	analytical pdf $f(x)$	sample size	ISE ratio
1	$\frac{1}{2}N(0,1) + \frac{1}{2}N(5,4)$	100	0.87
2	$\frac{1}{3}\left(N(0;0.7) + N(3;0.7) + N(7;2.7)\right)$	2400	0.61
3	$\frac{1}{3}\left(N(-2;1) + N(2;1) + N(0;2)\right)$	150	0.52
4	$\frac{1}{2}N(0,1) + \sum_{k=0}^{4} \frac{1}{10}N\left(\frac{k}{2}-1, \left(\frac{1}{10}\right)^2\right)$	1200	0.59
5	$\sum_{k=0}^{2} \frac{2}{7}N\left(\frac{12k-15}{7}; \left(\frac{2}{7}\right)^2\right) + \sum_{k=8}^{10} \frac{1}{21}N\left(\frac{2k}{7}, \left(\frac{1}{21}\right)^2\right)$	150	0.9
6	$\frac{1}{10}N(0,1) + \frac{9}{10}N\left(0, \left(\frac{1}{10}\right)^2\right)$	200	0.78
7	$\frac{1}{2}N\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$	200	0.73
8	$\frac{9}{20}N\left(-\frac{6}{5}; \left(\frac{3}{5}\right)^2\right) + \frac{9}{20}N\left(\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{1}{10}N\left(0, \left(\frac{1}{4}\right)^2\right)$	600	0.96
9	$\sum_{k=0}^{1} \frac{46}{100}N\left(2k-1; \left(\frac{2}{3}\right)^2\right) + \sum_{k=1}^{3} \frac{1}{300}N\left(-\frac{k}{2}; \left(\frac{1}{100}\right)^2\right) + \frac{7}{300}N\left(\frac{k}{2}; \left(\frac{7}{100}\right)^2\right)$	1000	0.97
10	$\frac{1}{2x\sqrt{2\pi}}\exp\left(-\frac{(\ln(x))^2}{2\times 2^2}\right)$	800	0.51
11	$\frac{1}{3.5}\exp\left(\frac{x}{3.5}\right)\exp\left(-\exp\left(\frac{x}{3.5}\right)\right)$	140	0.77

example, Figure 1 shows the result of a typical run of case study 1. The long thin bars above the data points represent the relative values of the Lagrange multipliers λ (i.e., mixture weights of (37)) associated with each point. From the figure we can conclude the advantage of the GCE approach is that out of the 100 points there are only 5 "support vectors" (i.e., only 5 of the 100 points have non-zero Lagrange multipliers associated with them). Thus, as with the support vector models [62], the model obtained via the GCE method is much sparser than the one obtained via the traditional kernel density estimator (which is an equally weighted Gaussian mixture with 100 components). Note, however, that the support vector machine theory does not provide an optimal value for the smoothing parameter σ in (37).

Figure 2, which depicts the claw density of case study 4, shows that the proposed estimator estimates the peaks and troughs of the density much better.

Mixtures of Gaussian densities result in light-tailed distributions. Heavy-tailed distributions have always been notoriously difficult to estimate via kernel methods. The theoretical asymptotic analysis that justifies the band-width selection procedures in kernel methods does not hold when the target f is heavy-tailed like the log-normal density. Figure 3 is a typical outcome of case study 10. It shows 800 points generated from a log-normal density with location 0 and scale 2, along with the SJ and GCE estimates. It is interesting to note that out of the 800 points only 57 points have non-zero Lagrange multipliers. Thus the GCE model for the 800 points is a Gaussian mixture with only 57 components. In contrast, the kernel estimator is an equally weighted mixture with 800 components. The sparsity of the GCE estimator makes it computationally easier to evaluate the density at each point.
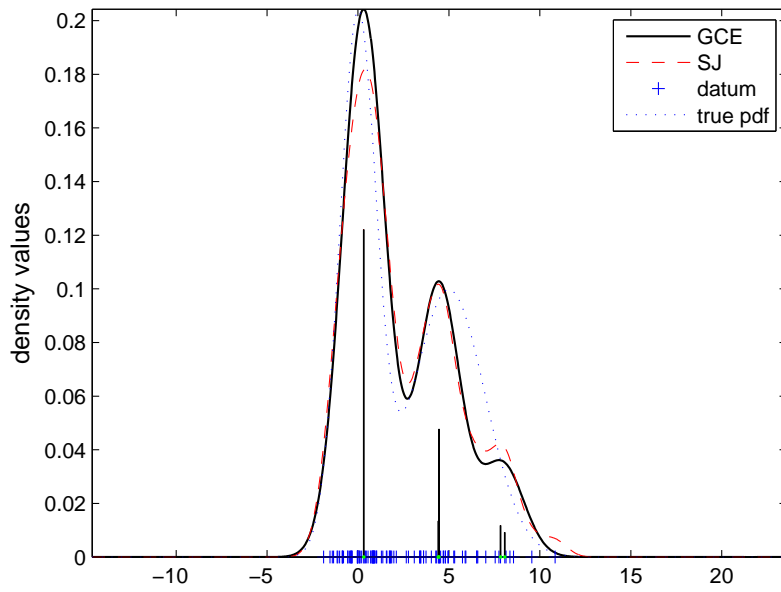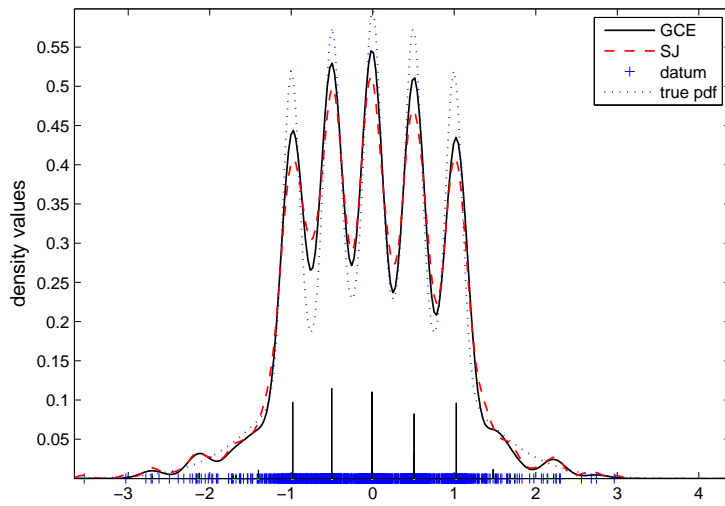
FIG 1. *Density estimation via the GCE and SJ methods*

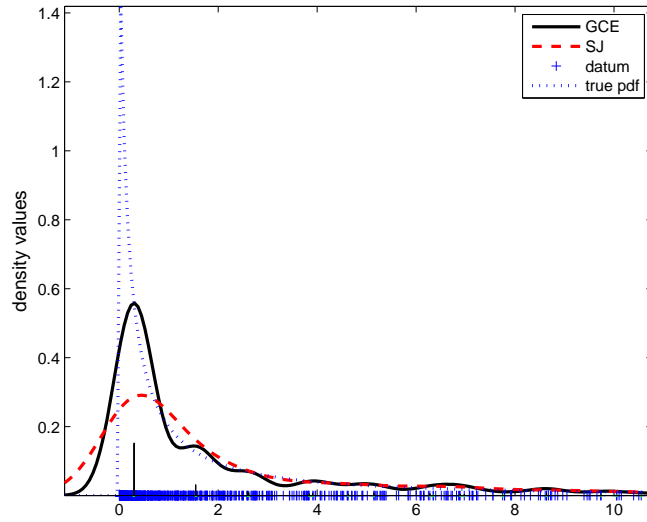FIG 2. *Case Study 4 - estimation of the claw density from 1200 points.*

FIG 3. *800 points from the log-normal density with location 0 and scale 2*
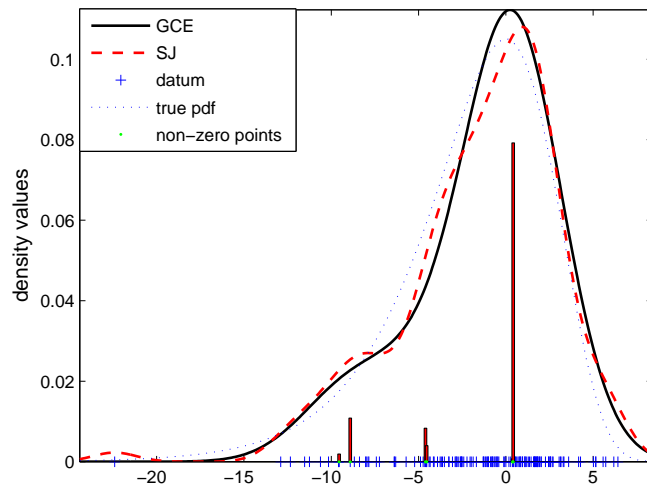

FIG 4. *140 points from the extreme value distribution with location=0 and scale=3.5*

Another example involving outliers is given in Figure 4, where we consider the Extreme Value density of case study 11. Again note the sparsity of the GCE density (only 5 points have non-zero Lagrange multipliers and thus our model is a mixture of 5 components) and the scarcity of any spurious modes which do not exist in the true underlying pdf. Extreme values, like the one at −25 do not affect the quality of the estimate.

Finally, we consider the application of the method on some widely analyzed real-world datasets. Figure 5 shows the estimated density from the *1872 Hidalgo stamp issue* data [25]. Our density estimate has 7 modes. This result is consistent with the findings in [25] and [2], where, based on historical information and physical considerations of the stamp production process, it is argued that an estimate with 7 modes is a sensible description of the data. Figure 6 shows the estimated density of the *acidity dataset*, measuring the acidity index in a sample of 155 lakes in Wisconsin [47]. The three modes are consistent with the analysis of the data in [42], where mixture models with two and three components are shown to be consistent with the data at the 5% and 10% level respectively using a Likelihood Ratio Test for the number of components. Finally, Figure 7 shows the density estimate of the widely analyzed *galaxy dataset* [48], which depicts the velocities of 82 galaxies diverging away from our own Milky Way. The density estimate exhibits 4 modes, which is a compromise between the 6 component mixture model with 6 modes favored by the Likelihood Ratio Test [42] and the 3 component mixture model favored by a full Bayesian analysis of the galaxy data [13].
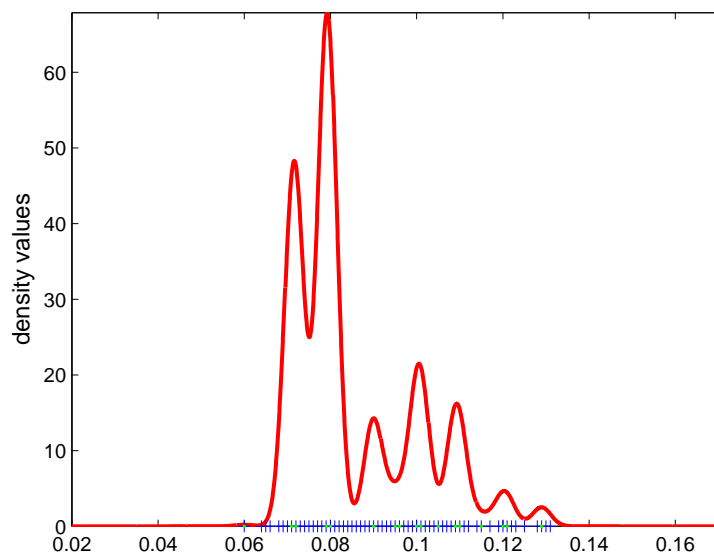

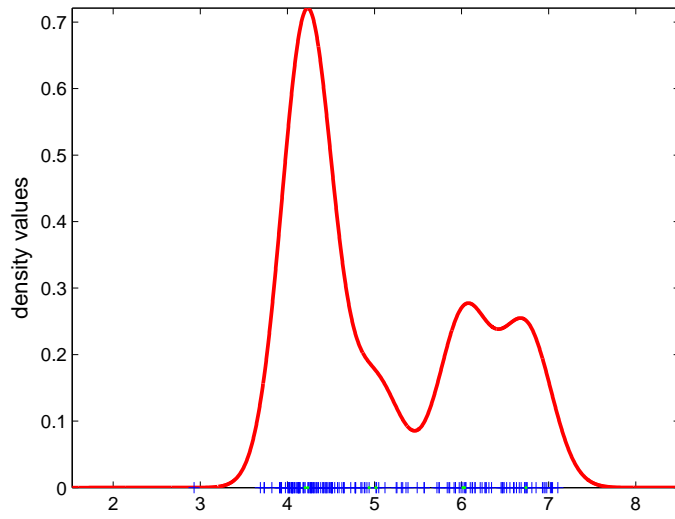Fɪɢ 5. *Plot of the mixture fit using the Hidalgo stamp issue data.*

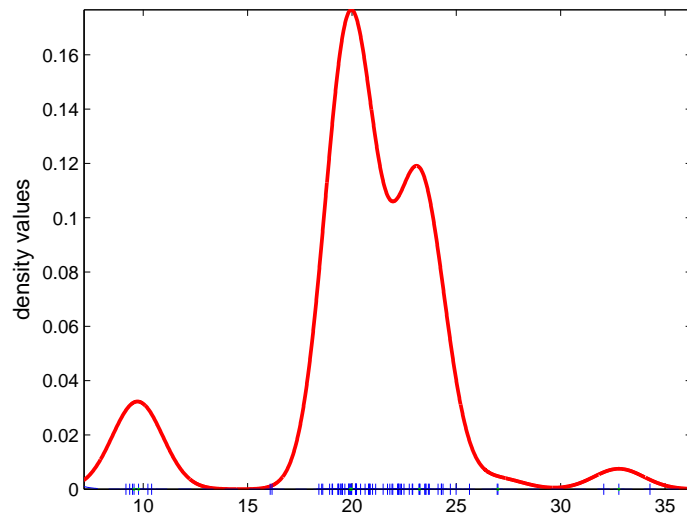Fɪɢ 6. *Density estimate of the acidity data [47].*

Fɪɢ 7. *Galaxy dataset and the corresponding density estimate.*

Remark 5. In all our simulation experiments we found a unique value for σ such that g integrates to unity, although we have no formal proof that this must always be so. As an illustration, Figure 8 shows $\mathbf{1}^T\lambda(\sigma) = \sum_{j=1}^{n} \lambda_j$ in (36) as a function of $\sigma \in [0, 10]$ for the galaxy dataset. It can be seen that in the range $[0, 10]$ there is a unique solution σ^* to program (36). In other words, there is a unique σ for which $\sum_{j=1}^{n} \lambda_j = 1$. The function $\mathbf{1}^T\lambda(\sigma)$ does not always appear monotonic in σ, but as seen from Figure 8, for values of $\sigma < \sigma^*$ the integral of g is smaller than unity, and for values of $\sigma > \sigma^*$ the integral of g is larger than unity. The same phenomenon was observed for the other test cases.
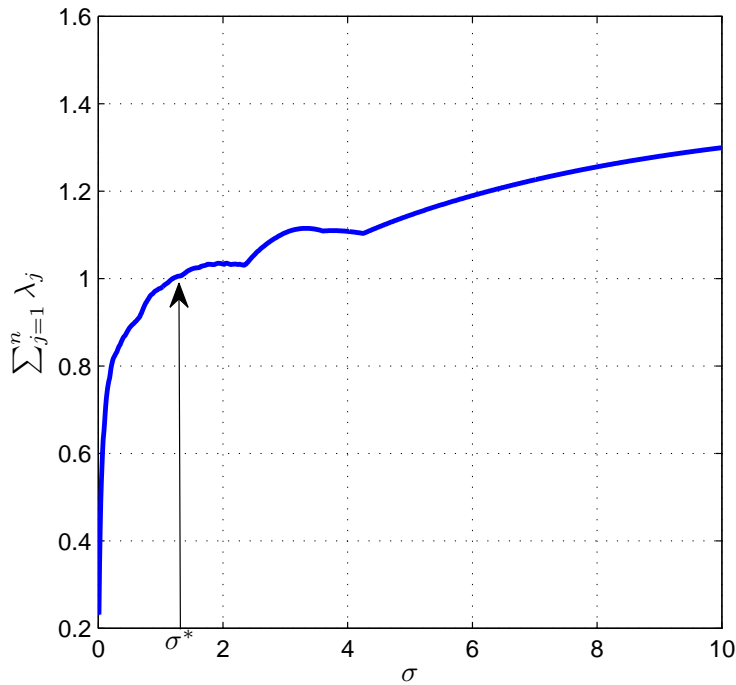

FIG 8. *A plot of* $\mathbf{1}^T\lambda(\sigma) = \sum_{j=1}^{n} \lambda_j$ *in (36) as a function of* $\sigma \in [0, 10]$ *for the galaxy dataset.*

In summary, we observe:

1. The standard kernel estimators rely on the availability of large samples to justify the asymptotic bandwidth plug-in selection procedures.
2. The GCE solves the problem directly without using asymptotic expan-

sions. The standard methods which do not explicitly rely on asymptotic approximations are the (least squares) cross validation methods ([10], [58]), which have been observed to give "rough" and "spiky" density estimates reflecting high variance of the estimator [64],[35].

3. The only approximation is in the estimation of the characterizing moments $\mathbb{E}_f K_i(\mathbf{X})$ through $\hat{\kappa}$. Apart from this approximation, the GCE solves a functional optimization problem exactly to find the optimal density function.

4. The GCE gives a sparse mixture model.

6. Discussion and Conclusions

We have formulated the GCE method and applied it to the problem of density estimation. The results of the numerical experiments are most encouraging and show that the GCE method has the potential of becoming a useful tool for addressing problems in probability density estimation. In our concluding discussion we note that:

1. The GCE ideas put forward by [30] and elaborated here seem to be related to the Support Vector Machines [62] and there is a possibility that the underlying principles of the Support Vector Machines can be derived via an information-theoretic approach similar to the one presented in [30]. An obvious similarity between the two approaches is that both of them require the solution of a QPP. Comparing the GCE approach to density estimation with the support vector machine approach [45], we find differences in the QPP constraints and the manner in which the regularization parameter σ is determined. More precisely, the QPP in the GCE method does not have the constraint $\sum_{i=1}^{N} \lambda_i = 1$ on the Lagrange multipliers. In the SVM approach [45] the Lagrange multipliers are constrained to sum to unity so that the model is a proper pdf. In the SVM approach the selection of σ is determined through experimentation and the only guidance for the selection of σ is provided by bounds of the form $\sigma \leqslant const.N^{-1/2}$. In the GCE approach, we select the (unique) σ for which the Lagrange multipliers sum to unity. It is interesting that the SVM and Generalized Entropy paradigms are developed from completely different principles and considerations, yet they both arrive at solutions which share many common characteristics. We point out, however, that the Generalized Cross Entropy principles [28, 30, 31] were used before the Support Vector Machines [62] gained popularity.

2. Similar to the Support Vector models, the GCE model is sparse in the sense that the number of non-zero mixture components is usually much smaller than the number of observations. This is in sharp contrast to the traditional kernel density estimation techniques where the model pdf is an equally weighted mixture with as many components as the number of observations. The sparsity of the model pdf makes it computationally less expensive to evaluate and manipulate it.

3. The Maximum Entropy Method of Jaynes [26] has for a long time been the only method to provide a pdf derived from the solution of a functional optimization problem. The proposed generalized model is thus the second density estimator to be derived from the solution of a functional optimization problem. In contrast to the popular Maximum Entropy Method, we use Pearson's divergence measure (as opposed to Kullback-Leibler's measure) in combination with generalized moment constraints to construct the model. To the best of our knowledge, Pearson's divergence measure has not been used in a functional optimization context to produce a probability model.

4. Whilst the popular maximum entropy programs [28] in probability and statistical mechanics utilize the Euler-Lagrange techniques of optimization, for the first time we exploit the more general Kuhn-Tucker duality to provide a bandwidth selection rule for the resulting estimator. The proposed approach gives rise to a bandwidth selection rule which is justified without appealing to asymptotic arguments.

References

[1] I.S. Abramson. On bandwidth variation in kernel estimates—a square root law. *Ann. Stat.*, 10:1217–1223, 1982.

[2] K. E. Basford, G. J. McLachlan, and M. G. York. Modelling the distribution of stamp paper thickness via finite normal mixtures: the 1872 stamp issue of mexico revisited. *Journal of Applied Statistics*, 24:169–179, 1997.

[3] A. Ben-Tal and M. Teboulle. Penalty functions and duality in stochastic programming via ϕ divergence functionals. *Mathematics of Operations Research*, 12:224–240, 1987.

[4] C. Biernacki, C. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated classification likelihood. Technical Report No. 3521. Rhône-Alpes:INRIA, 1998.

[5] J. M. Borwein and A. S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM J. Control and Optimization*, 29:325–338, 1991.

[6] J. M. Borwein and A. S. Lewis. *Convex analysis and Nonlinear Optimization: Theory and Examples*. Springer-Verlag, 2000.

[7] Z. I. Botev. *Stochastic Methods for Optimization and Machine Learning*. ePrintsUQ, http://eprint.uq.edu.au/archive/00003377/, BSc (Hons) Thesis, Department of Mathematics, School of Physical Sciences, The University of Queensland, 2005.

[8] Z. I. Botev and D. P. Kroese. Non-asymptotic bandwidth selection for density estimation of discrete data. *Methodology and Computing in Applied Probability*, 10:435–451, 2008.

[9] A. W. Bowman. A comparative study of some kernel-based nonparametric density estimators. *J. Statist. Comput. Simul.*, 21:313–327, 1985.

[10] A. W. Bowman, P. Hall, and D. M. Titterington. Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, 71:341–351, 1984.

[11] S. P. Boyd. *Convex Optimization*. New York: Cambridge, 2004.

[12] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal*, 13:195–212, 1996.

[13] S. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1982.

[14] S. T. Chiu. Bandwidth selection for kernel density estimation. *The Annals of Statistics*, 1991.

[15] I. Csiszár. A class of measures of informativity of observation channels. *Periodic Math. Hungarica*, 2:191–213, 1972.

[16] J. C. Principe D. Erdogmus. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Transactions on Signal Processing*, 50, 2002.

[17] A. Decarreau, D. Hilhorst, C. Lemarechal, and J. Navaza. Dual methods in entropy maximization. applications to some problems in crystalography. *SIAM Journal of Optimization*, 1992.

[18] L. Devroye and L. Gyofri. *Nonparametric Density Estimation: The L₁ View*. Wiley Series In Probability And Mathematical Statistics, 1985.

[19] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001.

[20] M. Girolami and C. He. Probability density estimation from optimally condensed data samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1253–1264, 2003.

[21] M. Girolami and C. He. Novelty detection employing an l_2 optimal non-parametric density estimator. *Pattern Recognition Letters*, 25:1389–1397, 2004.

[22] P. Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15:1491–1519, 1987.

[23] P. Hall and B. A. Turlach. Reducing bias in curve estimation by use of weights. *Computational Statistics and Data Analysis*, 30:67–86, 1999.

[24] J. H. Havrda and F. Charvat. Quantification methods of classification processes:concepts of structural α entropy. *Kybernatica*, 3:30–35, 1967.

[25] A. J. Izenman and C. J. Sommer. Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, 83:941–953, 1988.

[26] E. T. Jaynes. Information theory and statistical mechanics. *Physical Reviews*, 106:621–630, 1957.

[27] M. C. Jones, J. S. Marron, and S. J. Sheather. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, 11:337–381, 1996.

[28] J. N. Kapur. *Maximum Entropy Models in Science and Engineering*. Wiley Eastern, New Delhi, India, 1989.

[29] J. N. Kapur. *Measures of Information and Their Applications*. John Wiley & Sons, New Delhi, India, 1994.

[30] J. N. Kapur and H. K. Kesavan. *Generalized Maximum Entropy Principle (With applications)*. Standford Educational Press, University of Waterloo, Waterloo, Ontario, Canada, 1987.

[31] J. N. Kapur and H. K. Kesavan. The generalized maximum entropy principle. *IEEE Transactions on Syst., Man., and Cybernetics*, 19:1042–1052, 1989.

[32] J. N. Kapur and H. K. Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, New York, 1992.

[33] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann.Math.Stat.*, 22:79–86, 1951.

[34] E. L. Lehmann. Model specification: The views of fisher and neyman, and later developments. *Statistical Science*, 5:160–168, 1990.

[35] C. R. Loader. Bandwidth selection: Classical or plug-in. *The Annals of Statistics*, 27:415–438, 1999.

[36] C. R. Loader. *Local Regression and Likelihood*. Springer, 1999.

[37] J. S. Marron M. C. Jones and B. U. Park. A simple root n bandwidth selector. *The Annals of Statistics*, 19 4:1919–1932, 1991.

[38] H. K. Kesavan M. Srikanth and P. H. Roe. Probability density function estimation using the minmax measure. *IEEE Transactions on systems, Man. and Cybernetics—Part C: Applications and Reviews*, 30(1):77–83, 2000.

[39] J. S. Marron. An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *The Annals of Statistics*, 13:1011–1023, 1985.

[40] J. S. Marron and M. P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20:712–736, 1992.

[41] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, New York, 1997.

[42] G. J. Mclachlan and D. Peel. Contribution to the discussion of paper by s. richardson and p.j.green. *Journal of the Royal Statistical Society B*, 59:779–780, 1997.

[43] G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[44] R. A. Morejon and J. C. Principe. Advanced search algorithms for information-theoretic learning with kernel-based estimators. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 15, 2004.

[45] S. Mukherjee and V. Vapnik. Multivariate density estimation: a support vector machine approach. *Massachusetts Institute of Technology, ftp://publications.ai.mit.edu/ai-publications/1500-1999/AIM-1653.ps,* 1999.

[46] Y. Pawitan. *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Carendon Press Oxford, 2001.

[47] S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B*, 59:731–792, 1997.

[48] K. Roeder. Density estimation with confidence sets exemplified by super-clusters and voids in the galaxies. *Journal of the American Statistical Association*, 85:617–624, 1990.

[49] R. Y. Rubinstein. The stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation. *Methodology and Computing in Applied Probability*, 7:5–50, 2005.

[50] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method*. Springer-Verlag, 2004.

[51] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo Method, Second Edition*. John Wiley & Sons, New York, 2007.

[52] D. Ruppert and DBH Cline. Bias reduction in kernel density estimation by smoothed empirical transformations. *The Annals of Statistics*, 22:185–210, 1994.

[53] D. W. Scott. *Multivariate Density Estimation. Theory, Practice and Visualization*. John Wiley & Sons, 1992.

[54] D. W. Scott. Parametric statistical modeling by minimum integrated square error. *Technimetrics*, 43:274–285, 2001.

[55] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423;623–659, 1948.

[56] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[57] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.

[58] C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12, 1984.

[59] G. R. Terrell and David W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20:1236–1265, 1992.

[60] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

[61] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.*, 52:479, 1988.

[62] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[63] F. Y. M. Wan. *Introduction To The Calculus of Variations and Its Applications*. Chapman and Hall, 1995.

[64] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

[65] P. Zhang. Nonparametric importance sampling. *Journal of the American Statistical Association*, 91(435):1245–1253, 1996.