Generalized Cross-entropy Methods with Applications to Rare-event Simulation and Optimization

Z. I. Botev
D. P. Kroese
T. Taimre
Department of Mathematics
The University of Queensland
Brisbane 4072, Australia
botev@maths.uq.edu.au

The cross-entropy and minimum cross-entropy methods are well-known Monte Carlo simulation techniques for rare-event probability estimation and optimization. In this paper, we investigate how these methods can be extended to provide a general non-parametric cross-entropy framework based on ϕ-divergence distance measures. We show how the χ^2 distance, in particular, yields a viable alternative to the Kullback–Leibler distance. The theory is illustrated with various examples from density estimation, rare-event simulation and continuous multi-extremal optimization.

Keywords: generalized cross-entropy, maximum entropy method, cross-entropy method, rare-event simulation, stochastic optimization, Csisár's ϕ-divergence

1. Introduction

In the standard cross-entropy (CE) method [1], the importance sampling density (also called proposal or *instrumental* density) is restricted to some parametric family. The optimal instrumental density is the solution to a *parametric* CE minimization program and in special cases (in particular with multi-dimensional Bernoulli and Gaussian distributions) can be found explicitly, providing fast updating rules. Rubinstein [2] developed a non-parametric alternative referred to as minimum CE (MCE) which, like the standard CE method, aims to minimize the Kullback–Leibler (KL) CE distance. Instead of minimizing the distance within a parametric model, the MCE method minimizes the CE distance over all possible densities satisfying certain generalized moment-matching constraints. Thus, in contrast to the standard CE method, a *functional* optimization program is solved. The MCE method suggests the possibility of searching for an instrumental density in a non-parametric way, i.e. without the imposition of a fixed parametric model for the importance sampling pdf.

In both the CE and MCE methods the principal measure of interest is the KL CE. Kapur and Kesavan [3] argue that one could also use the more general Havrda–Charvát one-parameter family [4] as the distance measure between densities. Unfortunately, the solutions to the corresponding functional optimization program are not necessarily positive functions (probability densities), as is the case with the KL CE. Recently, however, it was shown [5] that if the equality constraints in the functional optimization program are replaced by *inequality* constraints, valid solutions (probability densities) to quite general functional CE minimization programs can be found. A particularly useful CE distance is the χ^2 distance which has several advantages, including an intuitive 'least-squares' distance interpretation and easy sampling from the resulting model.

In this paper we explore how the non-parametric framework can be used for rare-event estimation and optimization, paying particular attention to the χ^2 distance. For both rare-event simulation and optimization, the crucial point is to choose the instrumental density as close as possible to the target density (e.g. the minimum-variance Importance Sampling density). Our aim is to develop non-parametric methods for obtaining good instrumental densities that (1) are flexible enough to approximate the target

SIMULATION, Vol. 83, Issue 11, November 2007 785–806

DOI: 10.1177/0037549707087067
Figures 2–4 appear in color online: http://sim.sagepub.com

pdf well, and (2) are easy enough to sample from. In particular, we will focus on *kernel mixture densities*, which arise naturally from CE minimization using the χ^2 distance, and have a close connection with *density estimation*; see e.g. [6] and [7].

The rest of the paper is organized as follows. In Section 2 we present some background to the CE method. In Section 3 we provide the generalized CE (GCE) framework, and show how both CE and MCE are special cases. Moreover, we present a χ^2 GCE program as a convenient alternative to the conventional KL program. In Section 4 we indicate how the GCE method can be applied to density estimation, rare estimation and optimization. This is illustrated by numerical experiments in Section 5. Finally, in Section 6 we formulate our conclusions and give directions for future research.

2. The CE Method

The CE method is a well-known Monte Carlo technique for rare-event estimation and optimization [1]. In the estimation setting, the CE method provides an adaptive way of approximating the optimal importance sampling distribution for quite general problems. By formulating an optimization problem as an estimation problem, the CE method becomes a general and powerful stochastic search algorithm. The method is based on a simple iterative procedure where each iteration contains two phases: (a) generate a random data sample (trajectories, vectors, etc.) according to a specified mechanism; (b) update the parameters of the random mechanism on the basis of the data, in order to produce a 'better' sample in the next iteration. This last step involves minimizing the KL CE distance between the optimal importance sampling density and the instrumental density.

Specifically, consider the estimation of

$$\ell = \mathbb{P}_{\mathbf{u}}(S(\mathbf{X}) \geqslant \gamma) = \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geqslant \gamma\}}, \tag{1}$$

for some fixed level γ. Here $S(\mathbf{X}) \ \mathbb{R}^n \to \mathbb{R}$ is the sample performance and $\mathbf{X}$ is a random vector on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where the sample space $\Omega \equiv \mathbb{R}^n$, $\mathcal{F}$ is the $\sigma-$algebra of Borel subsets of $\mathbb{R}^n$ and $\mathbb{P}$ is a probability measure on $\mathcal{F}$. In addition, $\mathbf{X}$ has pdf $f(\cdot; \mathbf{u})$, belonging to some parametric family $\{f(\cdot; \mathbf{v}), \mathbf{v} \in \mathcal{V}\}$. We assume that $\{S(\mathbf{X}) \geqslant \gamma\}$ is a *rare event*, i.e. the probability ℓ is very small, e.g. $< 10^{-5}$ (how small ℓ needs to be for a rare event depends on the problem under consideration; in this sense the concept of a rare event is similar to the concept of ill-conditioning of a matrix as measured by the matrix condition number). We can estimate ℓ using the Importance Sampling (IS) estimator

$$\hat{\ell} = \frac{1}{N} \sum_{k=1}^{N} I_{\{S(\mathbf{X}_k) \geqslant \gamma\}} \, W(\mathbf{X}_k; \mathbf{u}, \mathbf{v}), \tag{2}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ is a random sample from $f(\mathbf{x}; \mathbf{v})$, and $W(\mathbf{X}_k; \mathbf{u}, \mathbf{v}) = f(\mathbf{X}_k; \mathbf{u}) / f(\mathbf{X}_k; \mathbf{v})$ is the likelihood ratio.

The challenging problem is how to select a vector $\mathbf{v}$ that gives the most accurate estimate of ℓ for a fixed simulation effort. The ideal (zero variance) IS density is given by

$$\pi(\mathbf{x}) = \frac{f(\mathbf{x}; \mathbf{u}) \, I_{\{S(\mathbf{x}) \geqslant \gamma\}}}{\ell}.$$

The idea behind the CE method is to choose $\mathbf{v}$ such that the KL CE distance between π and $f(\cdot; \mathbf{v})$ is minimized. That is, minimize

$$\mathcal{D}(\pi \to f(\cdot; \mathbf{v})) = \mathbb{E}_{\pi} \ln \frac{\pi(\mathbf{X})}{f(\mathbf{X}; \mathbf{v})}. \tag{3}$$

This implies that the optimal reference parameter $\mathbf{v}^*$ is given by

$$\mathbf{v}^* = \operatorname*{argmax}_{\mathbf{v} \in \mathcal{V}} \mathbb{E}_{\mathbf{u}} I_{\{S(\mathbf{X}) \geqslant \gamma\}} \ln f(\mathbf{X}; \mathbf{v}), \tag{4}$$

which can, in principle, be estimated using

$$\operatorname*{argmax}_{\mathbf{v} \in \mathcal{V}} \frac{1}{N} \sum_{k=1}^{N} I_{\{S(\mathbf{X}_k) \geqslant \gamma\}} \ln f(\mathbf{X}_k; \mathbf{v}), \tag{5}$$

with $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{iid} f(\cdot; \mathbf{u})$. However, this is void of meaning if $\{S(\mathbf{X}) \geqslant \gamma\}$ is a rare event under $f(\cdot; \mathbf{u})$, since it is likely that all indicators in the sum above are zero. To circumvent this problem a multi-level approach is used, where a sequence of reference parameters $\{\mathbf{v}_t, \ t \geqslant 0\}$ and a sequence of levels $\{\gamma_t, \ t \geqslant 1\}$ are generated, while iterating in both γ_t and $\mathbf{v}_t$.

In particular, starting from $\mathbf{v}_0 = \hat{\mathbf{v}}_0 = \mathbf{u}$, one proceeds as follows.

1. **Adaptive updating of γ_t.** For a fixed $\mathbf{v}_{t-1}$, let γ_t be the $(1-\rho)$-quantile of $S(\mathbf{X})$ under $\mathbf{v}_{t-1}$. Here ρ is a user-specified parameter supplied to the algorithm in advance. To estimate γ_t, draw a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from $f(\cdot; \hat{\mathbf{v}}_{t-1})$ and evaluate the sample $(1-\rho)$-quantile $\hat{\gamma}_t$.

2. **Adaptive updating of $\mathbf{v}_t$.** For fixed γ_t and $\mathbf{v}_{t-1}$, derive $\mathbf{v}_t$ as

$$\mathbf{v}_t = \operatorname*{argmax}_{\mathbf{v} \in \mathcal{V}} \mathbb{E}_{\mathbf{v}_{t-1}} I_{\{S(\mathbf{X}) \geqslant \gamma_t\}} W(\mathbf{X}; \mathbf{u}, \mathbf{v}_{t-1}) \times \ln f(\mathbf{X}; \mathbf{v}). \tag{6}$$

   The stochastic counterpart of equation (6) is as follows: for fixed $\hat{\gamma}_t$ and $\hat{\mathbf{v}}_{t-1}$, derive $\hat{\mathbf{v}}_t$ as the solution

$$\hat{\mathbf{v}}_t = \operatorname*{argmax}_{\mathbf{v} \in \mathcal{V}} \frac{1}{N} \sum_{\mathbf{X}_k \in \mathcal{E}_t} W(\mathbf{X}_k; \mathbf{u}, \hat{\mathbf{v}}_{t-1}) \times \ln f(\mathbf{X}_k; \mathbf{v}), \tag{7}$$

   where $\mathcal{E}_t$ is the *set of elite samples* on the tth iteration, that is, the samples $\mathbf{X}_k$ for which $S(\mathbf{X}_k) \geqslant \hat{\gamma}_t$.

The procedure is set to terminate only when, at some iteration T, a level $\hat{\gamma}_T$ is reached which is at least γ, at which point the original value of γ can be used without too few samples having non-zero indicators. We then reset $\hat{\gamma}_T$ to γ, reset the corresponding elite set, and deliver the final reference parameter $\hat{\mathbf{v}}^* = \hat{\mathbf{v}}_T^*$, using equation (7) as before. This $\hat{\mathbf{v}}^*$ is then used in equation (2) to estimate ℓ. Assuming that the probability ℓ does not vanish in a neighborhood of $\hat{\mathbf{v}}^*$ (which is the case, for example, if the distribution of $S(\mathbf{X})$ has infinite tail), then it can be shown [8] that $\hat{\mathbf{v}}^*$ converges in probability to the output of program (5). The asymptotic consistency of $\hat{\mathbf{v}}^*$ as an estimator of $\mathbf{v}^*$ under some technical conditions and the convergence of the algorithm above are also discussed in Homem-de-Mello and Rubinstein [8].

The following toy example explains the essence of the CE method and will be used later to motivate the MCE method. All of these methods and ideas will be unified in a single framework which we call the generalized CE (GCE) framework.

Example 1 (Exponential Distribution) Suppose we wish to estimate via simulation the probability $\ell = \mathbb{P}_u(X \geqslant \gamma)$, with X exponentially distributed with mean u; we write $X \sim \text{Exp}(u^{-1})$. Suppose further that γ is large in comparison to u, so that $\ell = \mathrm{e}^{-\gamma/u}$ is a rare-event probability. The updating formula for $\hat{v}_t$ in equation (7) follows from the solution v to

$$\frac{\partial}{\partial v} \sum_{X_k \in \mathcal{E}_t} W_k \ln\left(v^{-1}\mathrm{e}^{-X_k/v}\right)$$

$$= \quad -\sum_{X_k \in \mathcal{E}_t} W_k \frac{1}{v} + \sum_{X_k \in \mathcal{E}_t} W_k \frac{X_k}{v^2} = 0,$$

where $W_k = \mathrm{e}^{-X_k(u^{-1}-v^{-1})} v/u$, yielding

$$\widehat{v}_t = \frac{\sum_{X_k \in \mathcal{E}_t} W_k X_k}{\sum_{X_k \in \mathcal{E}_t} W_k}. \tag{8}$$

In other words, $\hat{v}_t$ is simply the sample mean of the elite samples, weighted by the likelihood ratios $\{W_k\}$.

Similarly, the *deterministic* updating formula equation (6) gives

$$v_t = \frac{\mathbb{E}_{\mathbf{u}} I_{\{X \geqslant \gamma_t\}} X}{\mathbb{E}_{\mathbf{u}} I_{\{X \geqslant \gamma_t\}}} = \mathbb{E}_{\mathbf{u}}\left[X \mid X \geqslant \gamma_t\right] = \gamma_t + u,$$

where $\gamma_t = -v_{t-1} \ln \rho$ is the $(1-\rho)$-quantile of the $\text{Exp}(v_{t-1}^{-1})$ distribution. The CE optimal parameter is $v^* = \gamma + u$. The Relative Error (RE) of $\hat{\ell}$, that is, $\text{RE} = \sqrt{\mathbb{E}\left[(\hat{\ell}-\ell)^2\right]}/\ell$, under any $v > u/2$ is [1, pg. 77]:

$$\text{RE} = \frac{1}{\sqrt{N}} \sqrt{\frac{v^2 \mathrm{e}^{\gamma/v}}{u(2v-u)} - 1}.$$

Substituting $v = v^* = \gamma + u$ shows that the relative error grows in proportion to $\sqrt{\gamma}$. More specifically, for fixed u, $\text{RE} \sim \sqrt{\gamma\,\mathrm{e}/N2u}$ as $\gamma \to \infty$. Thus the estimator (2) under the CE reference parameter v^* is *polynomial*. This is in contrast to the crude Monte Carlo (CMC) estimator ($v = u$) which is *exponential*, i.e. $\text{RE} \sim \exp(\gamma/2u)/\sqrt{N}$ as $\gamma \to \infty$. For comparison, the Minimum Variance (VM) parameter [1, pg. 77] is ${}_*v = (\gamma + u + \sqrt{\gamma^2 + u^2})/2 \sim \gamma + u/2$, $\gamma \to \infty$, which gives asymptotically the same relative error as the CE case. □

Although in the example above both CE and VM yield substantial variance reduction compared with CMC, the relative errors still increase with γ in both cases. To contrast, the zero-variance IS pdf is a shifted exponential pdf, given by $\pi(x) = I_{\{x \geqslant \gamma\}}\, u^{-1}\, \mathrm{e}^{-(x-\gamma)u^{-1}}$. This suggests looking for g in a larger class of distributions. For instance, in Example 1 one could consider the class of shifted exponentials.

Example 2 (Cauchy Density, Example 1 Continued) Suppose we choose as instrumental density in Example 1 such as the Cauchy density, that is, $g(x; h, \mu) = (h/\pi)\left(h^2 + (x-\mu)^2\right)^{-1}$. We wish to find the parameters h and μ that give minimal variance for the corresponding IS estimator $\hat{\ell}$. By standard arguments [1] this means minimizing

$$\mathbb{E}_\pi \frac{\pi(X)}{g(X; h, \mu)} \tag{9}$$

with respect to the parameters h and μ. Here the target

$$\pi(x) = I_{\{x \geqslant \gamma\}} f(x; u)/\ell \propto I_{\{x \geqslant \gamma\}} \mathrm{e}^{-x/u}$$

is again the optimal IS pdf. Now expression (9) is proportional to

$$\int_\gamma^\infty \frac{1}{h} \mathrm{e}^{-2x/u}\left(h^2 + (x-\mu)^2\right) dx$$

$$= \quad \frac{u\mathrm{e}^{-2\gamma/u}}{4h}\left(2h^2 + (\gamma-\mu)^2 + (u+\gamma-\mu)^2\right),$$

which has extrema at $(\mu, h) = (\gamma + u/2, \pm u/2)$. We take the solution with $h > 0$, leaving $({}_*\mu, {}_*h) = (\gamma + u/2, u/2)$ as the parameter pair that minimizes the variance of $\hat{\ell}$. The corresponding minimum is

$$\int_{\mathbb{R}} \frac{\pi^2(x; u)}{g(x; {}_*\mu, {}_*h)} dx$$

$$= \int_\gamma^\infty (\ell u)^{-2} \mathrm{e}^{-2x/u} \frac{2\pi}{u}\left(\frac{u^2}{4} + \left(x - \left(\gamma + \frac{u}{2}\right)\right)^2\right) dx$$

$$= \frac{1}{\ell^2} N \mathbb{E}_{*g}\left[\hat{\ell}^2\right] = \frac{\pi}{2}$$

and the relative error of the minimum variance Cauchy estimator is given by

$$RE = \frac{\mathrm{Std}_{*g}(\hat{\ell})}{\ell} = \frac{\sqrt{(\frac{\pi}{2}\ell^2 - \ell^2)/N}}{\ell} = \sqrt{\frac{\pi - 2}{2N}}.$$

Thus, we have a relative error which does not depend on γ, so that an estimator with *bounded* relative error is obtained. □

3. The GCE Framework

The GCE framework [5, 9] is a natural generalization and unification of the ideas behind the MCE and CE methods [1, 2] and the maximum entropy principle [10]. Similar to the MCE method, the idea is to choose the instrumental pdf as close to some *prior* pdf as possible, while at the same time satisfying certain (generalized) moments constraints. These constraints enforce a moment-matching condition between the model (instrumental pdf) and the target (optimal importance sampling) pdf.

To explain the GCE framework, recall that the idea behind the CE method is to choose the instrumental density g such that the KL CE distance between the optimal (minimum variance) importance sampling density π and g is minimal. If we search for the optimal g over all densities, then g is the solution of the *functional* optimization program $\min_g \mathcal{D}(\pi \to g)$. The solution to this functional optimization program is $g = \pi$, which is not useful because π is unknown. In the CE method this problem is resolved by restricting the space of densities within which we search for g to be a parametric family of densities, say $\{f(\cdot; \mathbf{v}), \mathbf{v} \in \mathcal{V}\}$. Instead of solving a functional optimization problem, one therefore solves the *parametric* optimization problem $\min_{\mathbf{v}} \mathcal{D}(\pi \to f(\cdot; \mathbf{v}))$. In many cases the parametric family $\{f(\cdot; \mathbf{v}),\ \mathbf{v} \in \mathcal{V}\}$ can be quite a rigid and inflexible model for the target π.

Some important questions therefore arise. Is it possible to obtain an instrumental pdf in a *non-parametric way*, that is, not directly linked to a class of parameterized densities? Is it possible to modify the functional optimization problem $\min_g \mathcal{D}(\pi \to g)$ to obtain a useful instrumental which is close to but not immediately identical to π? Moreover, is the KL CE distance the best distance criterion for obtaining a good instrumental pdf? Is it possible to use more general distance measures, such as the Csiszár family of distances, to obtain useful instrumental densities? These questions motivate the following non-parametric procedure of getting close to the target π, where 'closeness' is measured by a generalized CE distance.

GCE Program

1. Given an *a priori* probability density p on the set $\mathcal{X} \subset \mathbb{R}^n$,
2. minimize the Csiszár ϕ-divergence (measure of CE):
$$\mathcal{D}(g \to p) = \int_{\mathcal{X}} p(\mathbf{x})\, \phi\left(\frac{g(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x} \qquad (10)$$
over all probability densities g, where ϕ is any continuous twice-differentiable function, with $\phi(1) = 0$ and $\phi''(x) > 0, x > 0$,
3. subject to the *generalized moment-matching* constraints (equalities or inequalities):
$$\mathbb{E}_g K_i(\mathbf{X}) = \int_{\mathcal{X}} g(\mathbf{x}) K_i(\mathbf{x}) d\mathbf{x} \underset{=}{\geqslant} \kappa_i,$$
$$i = 0, \ldots, m, \qquad (11)$$
where $\{K_i : \mathbb{R}^n \to \mathbb{R}\}_{i=1}^m$ is a set of linearly independent smooth functions called *kernels*, and $\kappa_i = \mathbb{E}_\pi K_i(\mathbf{X}),\ i = 1, \ldots, m$ (in practice κ_i is estimated via a Monte Carlo estimate $\hat{\kappa}_i$ discussed later). For notational simplicity we set $K_0 \equiv 1, \kappa_0 = 1$ and insist that the zeroth constraint (the constraint corresponding to $i = 0$) always be a strict equality constraint so that the function g integrates to unity.

To solve the above optimization problem we employ the Karush–Kuhn–Tucker (KKT) theory [11] of constrained optimization which is a generalization of the usual Lagrangian theory of constrained optimization. Traditional KKT theory applies to finite-dimensional vector spaces but it has been shown [12, 13, 14] that it can also be extended to infinite-dimensional functional spaces. Using the KKT theory and ignoring the non-negativity constraint on g, we have the following proposition [5, 9, 14].

Proposition 1 *The solution to the GCE program is given by:*

$$g(\mathbf{x}) = p(\mathbf{x})\Psi'\left(\sum_{i=0}^m \lambda_i K_i(\mathbf{x})\right), \quad \Psi' = \phi'^{-1}, \qquad (12)$$

where the $\{\lambda_i\}_{i=0}^m$ are Lagrange multipliers which solve the convex optimization program:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \sum_{i=0}^m \lambda_i \kappa_i - \mathbb{E}_p \Psi\left(\sum_{i=0}^m \lambda_i K_i(\mathbf{X})\right) \qquad (13)$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \qquad (14)$$

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^T$.

The inequality constraint $\boldsymbol{\lambda} \geqslant \mathbf{0}$ holds component-wise, i.e. $\lambda_i > 0\ \forall i$, and is enforced only when the moment-matching constraints equation (11) are inequality constraints. If the moment constraints are equalities, then the constraint (14) is omitted. Since the zeroth constraint is always a strict equality, λ_0 is not included in (14).

Remark 1 (Csiszár's measure) Csiszár's measure [15] of directed divergence, defined by the functional in equation (10), can be interpreted as a distance measure between the pdfs g and p over the probability space $(\Omega, \mathbb{P}, \mathcal{F})$. The definition of Csiszár's measure ensures that it has the following properties:

1. $\mathcal{D}(g \rightarrow p) \geqslant 0$, following Jensen's inequality: $\mathbb{E}_p \phi\left(\frac{g(\mathbf{X})}{p(\mathbf{X})}\right) \geqslant \phi\left(\mathbb{E}_p \frac{g(\mathbf{X})}{p(\mathbf{X})}\right)$;
2. $\mathcal{D}(g \rightarrow p) = 0$ if and only if $g \equiv p$; and
3. $\mathcal{D}(g \rightarrow p)$ is a convex function of g and p.

Notice, however, that $\mathcal{D}$ differs from the usual metric functions over a metric space in the following properties.

1. In general, $\mathcal{D}(g \rightarrow p) \neq \mathcal{D}(p \rightarrow g)$, i.e. $\mathcal{D}$ is not symmetric, hence the label *directed* divergence applied to it.
2. In general, $\mathcal{D}(g \rightarrow p) + \mathcal{D}(p \rightarrow s) \ngeq \mathcal{D}(g \rightarrow s)$ for any probability density s, i.e. the measure does not satisfy the triangle inequality which is characteristic of all Euclidean measures of distance, for example.

If we set $\phi(x) = \left(x^\alpha - x\right) / \left(\alpha(\alpha - 1)\right), \alpha \neq 0, 1$, then the resulting CE distance:

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

is indexed by the parameter α. Specific choices of α give rise to the most notable CE measures [15]. For example,

$$\mathcal{D}_2(g \rightarrow p) = \frac{1}{2} \int \frac{\left[g(\mathbf{x}) - p(\mathbf{x})\right]^2}{p(\mathbf{x})} d\mathbf{x}$$

yields the χ^2 distance measure and $\lim_{\alpha \rightarrow 1} \mathcal{D}_\alpha(g \rightarrow p)$ yields the KL CE distance. In subsequent sections, we will make extensive use of the χ^2 distance measure. Our usage of the χ^2 measure is in part motivated by the following relations with respect to other measures [16, pg. 224]:

$$\begin{aligned} 2\,\mathcal{D}_2(g \rightarrow p) &\geqslant \lim_{\alpha \rightarrow 1} \mathcal{D}_\alpha(g \rightarrow p) \\ &\geqslant \ln\left(1 + 2\mathcal{D}_2(g \rightarrow p)\right). \end{aligned}$$

Hence $\mathcal{D}_2$ dominates the KL CE distance. In addition, it is also easy to show that $\mathcal{D}_2$ dominates the L_1 metric distance:

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

Thus, if we minimize $\mathcal{D}_2$ with respect to g, then we also minimize an upper bound on two very fundamental distance measures: the KL CE and the L_1 metric distance. The importance of the L_1 metric is derived from the fact that it is the only L_p metric that is invariant to monotone transformations of $\mathbf{x}$ [16, Introduction].

Remark 2 (Non-negativity of g) Note that, since we have ignored the non-negativity constraint on g in the proposition, equation (12) is typically not a non-negative function. For some choices of ϕ, however, the non-negativity constraint $g(\mathbf{x}) \geqslant 0$ need not be imposed explicitly. In particular, if $\phi(x) = x\ln(x) - x + 1$, corresponding to minimization of the KL distance, then $\Psi'(x) = \exp(x)$ and the condition $g(\mathbf{x}) \geqslant 0$ is automatically satisfied. In this case the proposition above yields the unique optimal solution. For a general ϕ, however, the non-negativity constraint has to be enforced explicitly. We explain one practical way of achieving non-negativity for g later. For an interesting theoretical treatment of the non-negativity constraint see Ben-Tal and Teboulle [14].

Remark 3 In the GCE approach one always takes the uniform density on $\mathcal{X}$ to be the most uninformative prior. As in the Bayesian methodology, the prior could be an *improper* density, i.e. one for which the integral is not finite. For example, the most uninformative prior on $\mathbb{R}^n$ is $p(\mathbf{x}) \propto 1, \ \forall \mathbf{x} \in \mathbb{R}^n$. The prior p can be assumed to be an improper uniform density over the set $\mathcal{X}$ as long as the integration in equation (13) with weight p can be carried out. Note, however, that in the Bayesian approach the most uninformative priors are the so-called *Jeffrey's priors* [17].

Apart from the problem of choosing the distance measure $\mathcal{D}$, we also need to decide which features of the target density need to be modeled, i.e. which moment-matching constraints need to be enforced. We consider a number of features and argue that the most convenient 'closeness' measures are the KL and χ^2 distances.

3.1 The MCE method [2]

The MCE method is obtained as a special case of the GCE framework by choosing $\phi(x) = x\ln x - x + 1$ (so that $\Psi'(x) = \exp(x)$), corresponding to the minimization of the KL CE distance, and by taking *equality* constraints in equation (11). It follows from equation (12) that

$$g(\mathbf{x}) = p(\mathbf{x}) \exp\left(\sum_{i=0}^{m} \lambda_i K_i(\mathbf{x}) \right) \tag{15}$$

where the Lagrange multipliers are determined from the unconstrained maximization of the convex programming problem equation (13). This convex optimization problem can be solved by equating the corresponding gradient to zero, which leads to the following set of nonlinear equations:

$$\mathbb{E}_p \exp\left(\sum_{i=0}^{m} \lambda_i K_i(\mathbf{X}) \right) K_i(\mathbf{X}) = \kappa_i, \quad i = 0, \ldots, m. \tag{16}$$

The solution gives the unique optimal $g(\mathbf{x})$ for the MCE method. Note that λ_0 and the equation for $i = 0$ correspond to the normalization constraint $\int_{\mathcal{X}} g(\mathbf{x})d\mathbf{x} = 1$.

The expectations on the left-hand side of equation (16) typically have to be estimated via an empirical average to give the *stochastic counterpart* of equation (16):

$$\frac{1}{N}\sum_{j=1}^{N}\exp\left(\sum_{k=0}^{m}\lambda_k K_k\left(\mathbf{X}_j\right)\right)K_i\left(\mathbf{X}_j\right) = \hat{\kappa}_i,$$

$$\left\{\mathbf{X}_j\right\}_{j=1}^{N} \sim_{iid} p, \quad i = 0,\ldots,m,$$

where $\hat{\kappa}_i,\ i \geqslant 1$ is, for example, the IS estimate

$$\frac{\sum_{j=1}^{N}\frac{\pi(\mathbf{X}_j)}{p(\mathbf{X}_j)}K_i(\mathbf{X}_j)}{\sum_{j=1}^{N}\frac{\pi(\mathbf{X}_j)}{p(\mathbf{X}_j)}}$$

of $\mathbb{E}_\pi K_i(\mathbf{X})$. Simulation from equation (15) is in general feasible only via an Accept-Reject or a Markov Chain Monte Carlo algorithm. As mentioned in Remark 2, the non-negativity of $g(\mathbf{x})$ is ensured by its exponential functional form.

Remark 4 Note that the above MCE updating formulae can be obtained by choosing the following parametric pdf as the instrumental density:

$$g(\mathbf{x};\boldsymbol{\lambda}) = \frac{p(\mathbf{x})\exp\left(\sum_{i=1}^{m}\lambda_i K_i(\mathbf{x})\right)}{\mathbb{E}_p\exp\left(\sum_{i=1}^{m}\lambda_i\ K_i(\mathbf{X})\right)} \tag{17}$$

and then solving the parametric minimization program

$$\min_{\boldsymbol{\lambda}}\ \mathbb{E}_\pi \ln\frac{p(\mathbf{X})}{g(\mathbf{X};\boldsymbol{\lambda})}, \tag{18}$$

without any constraints on the parameters $\boldsymbol{\lambda}$. Thus equations (15) and (18) give identical results and correspond to choosing a model pdf from the general exponential family [18]. The minimization program (18) and the instrumental (17) are reminiscent of the CE minimization programs discussed previously. An important difference, however, is that here we can much more easily incorporate prior information via p.

Example 3 (Exponential Distribution, Example 1 Continued) We now consider the MCE approach to Example 1, using $K_1(x) = x$ ($K_0 \equiv 1$) and an improper prior $p(x) = 1$ on $[0,\infty)$. Equation (15) then becomes the exponential density $g(x) = \mathrm{e}^{\lambda_0+\lambda_1 x}$. Matching the zeroth equation in (16) normalizes g to give $g(x) = I_{\{x\geqslant 0\}}(-\lambda_1)\mathrm{e}^{\lambda_1 x}$ and solving the $i = 1$ equation yields:

$$-\lambda_1\int_0^\infty \mathrm{e}^{\lambda_1 x}\times dx = \kappa_1 = \mathbb{E}_\pi X.$$

Thus the optimal value of the Lagrange multiplier in this case is

$$\lambda_1^* = \frac{-1}{\mathbb{E}_\pi X} = \frac{-1}{\gamma+u},$$

which gives the same optimal density as in Example 1 (only the parameterization is different). It seems that by using a single moment constraint in equation (16) we have not gone beyond what the CE method can already do. In the MCE method, however, one can more easily incorporate prior information via p with the objective of getting closer to the target π. In particular, for this example we can take $p(x) = I_{\{x\geqslant\gamma\}}$, that is, 1 on $[\gamma,\infty)$ and 0 otherwise, indicating complete lack of prior information other than that the instrumental should be 0 for $x < \gamma$ (because we know that the minimal variance pdf π, regardless of its functional form, is 0 for $x < \gamma$).

From equation (17) and program (18), the form of the MCE instrumental density is

$$g(x) = \frac{I_{\{x\geqslant\gamma\}}\mathrm{e}^{\lambda_1 x}}{\int I_{\{x\geqslant\gamma\}}\mathrm{e}^{\lambda_1 x}dx} = I_{\{x\geqslant\gamma\}}(-\lambda_1)\mathrm{e}^{\lambda_1(x-\gamma)}, \tag{19}$$

where $\lambda_1 < 0$ is obtained from minimizing $\mathbb{E}_\pi \ln\left[p(X)/g(X)\right]$, that is,

$$\begin{aligned}\lambda_1 &= \operatorname*{argmax}_{\lambda<0}\int_\gamma^\infty \pi(x)\ln g(x)dx\\ &= \operatorname*{argmax}_{\lambda<0}\left[\ln(-\lambda)+u\lambda\right] = -u^{-1}.\end{aligned}$$

In other words, $g(x) = I_{\{x\geqslant\gamma\}}u^{-1}\mathrm{e}^{-(x-\gamma)u^{-1}}$, which happens to be the zero-variance IS density. □

3.2 The CE method

We now demonstrate that the CE method is a special case of the GCE framework when in the CE method we consider instrumentals from an exponential family [18], that is, densities of the form

$$g(\mathbf{x};\boldsymbol{\lambda}) = \frac{\exp\left(\sum_{k=1}^{m}\lambda_k K_k(\mathbf{x})\right)}{\int\exp\left(\sum_{k=1}^{m}\lambda_k K_k(\mathbf{x})\right)d\mathbf{x}},$$

where $\{K_i\}_{i=1}^m$ and $\{\lambda_i\}_{i=1}^m$ are called the *natural statistics* and *natural parameters*. In this case the parametric minimization program

$$\min_{\boldsymbol{\lambda}}\mathbb{E}_f\ln(\pi(\mathbf{X})/g(\mathbf{X};\boldsymbol{\lambda}))$$

is equivalent to the maximization program

$$\max_{\boldsymbol{\lambda}}\mathbb{E}_\pi\ln g(\mathbf{X};\boldsymbol{\lambda}).$$

Since g is in the exponential family, the maximization problem is concave. Setting the gradient equal to zero and

assuming that there exists a Lebesgue integrable function h such that

$$\left| \frac{\partial}{\partial \lambda_i} \ln g(\mathbf{x}; \boldsymbol{\lambda}) \right| < h(\mathbf{x}),$$

for all $\mathbf{x}$ and all $\boldsymbol{\lambda}$ with g continuously differentiable with respect to $\boldsymbol{\lambda}$, then by the Lebesgue dominated convergence theorem the expectation (integration) and differential operator can be interchanged, i.e. we have

$$\mathbb{E}_\pi \nabla \ln g(\mathbf{X}; \boldsymbol{\lambda}) = \mathbf{0}.$$

Thus we can write, for $i = 1, \ldots, m$:

$$\frac{\int \exp\left(\sum_{k=1}^m \lambda_k K_k(\mathbf{x})\right) K_i(\mathbf{x}) d\mathbf{x}}{\int \exp\left(\sum_{k=1}^m \lambda_k K_k(\mathbf{x})\right) d\mathbf{x}} = \mathbb{E}_\pi K_i(\mathbf{X}). \quad (20)$$

It can easily be verified that exactly the same equations can be obtained from the GCE program if we use: (1) $\phi(x) = x\ln(x) - x + 1$; (2) equality constraints in equation (11); and (3) proper or improper uniform prior $p(\mathbf{x}) \propto 1$ on $\mathcal{X}$. The updating equations between the GCE program and the CE method do not agree under any other conditions. We emphasize again that the single greatest advantage of the CE method is that for many exponential models equations (20) can be solved analytically to give simple and fast CE updating rules for the parameters $\{\lambda_i\}_{i=1}^m$ of the instrumental pdf.

Example 4 Consider the univariate case with the constraint $K_1(x) = x$ and $\mathcal{X} = [0, \infty)$; then equation (20) gives the exponential model $g(x) = -\lambda_1 e^{\lambda_1 x}$ with Lagrange multiplier (or CE parameter) $-1/\lambda_1 = \mathbb{E}_\pi X$. Similarly, if $K_1(x) = x$, $K_2(x) = x^2$ and $\mathcal{X} = \mathbb{R}$, then $g(x) \propto e^{\lambda_1 x + \lambda_2 x^2}$ which is simply a different parameterization of a Gaussian $g(x) = c \exp[-(x-\mu)^2/(2\sigma^2)]$ with optimal CE parameters given by $\mu = \mathbb{E}_\pi X$ and $\sigma^2 = \mathbb{E}_\pi X^2 - \mu^2$. □

We now deviate from using the KL CE distance and use a particular GCE measure [10] instead.

3.3 χ^2 GCE program

One can ask: apart from the KL distance measure, what other measures within the Csiszár's family yield 'useful' instrumentals? More specifically, we would like to choose the function ϕ in Csiszár's family of measures such that:

- it may be possible to complete the integration in equation (13) analytically;
- maximizing equation (13) (possibly with the constraints (14)) and hence finding the set of Lagrange multipliers $\{\lambda_k\}_{k=0}^m$ is relatively easy (e.g. only a linear $\phi'^{-1} = \Psi'$ will make the Hessian matrix of equation (13) constant and this will simplify the maximization of equation (13)); and
- generating random variables from the model g in equation (12) is relatively easy (e.g. if Ψ' is linear then g is a discrete mixture and the *composition method* for random variate generation applies).

We now show that it is possible to satisfy all these requirements by choosing the most general quadratic function $\phi(x) = a_2(x^2-1) + a_1(x-1)$, $a_2 > 0, a_1 \in \mathbb{R}$, which still satisfies Csiszár's requirement for a ϕ-divergence. It is easy to verify that the value of a_1 is not relevant to the primal optimization problem and any positive a_2 will yield an equivalent primal problem. For simplicity, we set $a_1 = 0$ and $a_2 = 1/2$, thus obtaining $\phi(x) = (x^2-1)/2$, in which case $\Psi'(x) = x$ and one obtains the χ^2 generalized CE distance

$$\mathcal{D}(g \to p) = \frac{1}{2} \int \frac{g^2(x)}{p(x)} dx - \frac{1}{2}.$$

For this GCE distance, the points above are satisfied if we

- choose some prior $p(\mathbf{x})$ over the set $\mathcal{X}$;
- let $\{K_i\}$ be *Gaussian kernel functions* with *bandwidth* σ, i.e.

$$K_i(\mathbf{x}) = K(\mathbf{x}|\mathbf{X}_i, \sigma) = c \exp\left(\frac{-||\mathbf{x} - \mathbf{X}_i||^2}{2\sigma^2}\right),$$

$$i = 1, \ldots, m,$$

  with $\mathbf{X}_1, \ldots, \mathbf{X}_m \sim_{approx.} \pi$ in (11) (c is a normalization constant); and
- select inequality constraints in (11).

Without the inequality constraints the non-negativity constraint on the pdf g will be very difficult to impose [14]. Next substitute these ingredients in Proposition 1. In this case, equation (12) becomes the *particle filter*-type density:

$$g(\mathbf{x}) = p(\mathbf{x}) \sum_{j=0}^m \lambda_j K(\mathbf{x}|\mathbf{X}_j, \sigma), \quad (21)$$

where the Lagrange multipliers $\boldsymbol{\lambda}, \lambda_0$ are obtained from equations (13) and (14). Specifically, equations (13) and (14) yield the convex *Quadratic Programming Problem* (QPP):

$$\min_{\boldsymbol{\lambda}, \lambda_0} \frac{1}{2} \left[\lambda_0, \boldsymbol{\lambda}^T\right] \begin{pmatrix} 1 & \mathbf{c}^T \\ \mathbf{c} & C \end{pmatrix} \begin{bmatrix} \lambda_0 \\ \boldsymbol{\lambda} \end{bmatrix} - \left[\lambda_0, \boldsymbol{\lambda}^T\right] \begin{bmatrix} 1 \\ \boldsymbol{\kappa} \end{bmatrix} \quad (22)$$

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

Here C is the $m \times m$ matrix with entries:

$$C_{ij} = \mathbb{E}_p K_i(\mathbf{X}) K_j(\mathbf{X}), \quad \text{for } i, j = 1, \ldots, m,$$

$\mathbf{c} = [\mathbb{E}_p K_1(\mathbf{X}), \ldots, \mathbb{E}_p K_m(\mathbf{X})]^T$ and $\boldsymbol{\kappa}(\sigma) = [\kappa_1(\sigma), \ldots, \kappa_m(\sigma)]^T$. Note that both the generalized moments $\boldsymbol{\kappa}(\sigma)$ and the matrix $C(\sigma)$ are functions of the bandwidth σ and the empirical data $\mathbf{X}_1, \ldots, \mathbf{X}_m$, because the kernels $\{K_i\}_{i=1}^m$ depend on σ and the data. In practice, $\boldsymbol{\kappa}(\sigma)$ is also estimated via Monte Carlo simulation.

A few problems should now be apparent. Firstly, Proposition 1 provides the optimal solution of the GCE program ignoring the fact that g has to be a non-negative function. Although we require (see equation (23)) the mixture weights $\boldsymbol{\lambda}$ in equation (21) to be non-negative, λ_0 is not constrained and could still be negative rendering the model (21) an invalid mixture pdf. Secondly, there seems to be no objective method of choosing an appropriate value for the bandwidth parameter σ. It turns out, however, that it is possible to choose the bandwidth parameter σ such that at the optimal solution of the QPP we have $\lambda_0 = 0$. Since λ_0 is not constrained, setting the gradient of equation (22) with respect to λ_0 to zero gives the relationship $\lambda_0 = 1 - \mathbf{c}^T\boldsymbol{\lambda}$. Thus, if $\mathbf{c}^T\boldsymbol{\lambda} = 1$ at the optimal solution of the QPP, then $\lambda_0 = 0$. This suggests that we can find the value of σ which gives $\mathbf{c}^T\boldsymbol{\lambda} = 1$ via the following implicit root-finding program:

$$(\sigma^*, \boldsymbol{\lambda}^*) = \left\{ \begin{matrix} (\sigma, \boldsymbol{\lambda}) \;\; \mathbf{c}^T\boldsymbol{\lambda}(\sigma) = 1, \\ \boldsymbol{\lambda}(\sigma) = \operatorname{argmin}_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \left(\frac{1}{2}\boldsymbol{\lambda}^T C(\sigma)\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \boldsymbol{\kappa}(\sigma) \right) \end{matrix} \right\}. \quad (23)$$

This program finds the value for σ, denoted σ^*, which will force the solution $\boldsymbol{\lambda}^*$ of the sum of equations (22) and (23) to satisfy $\mathbf{c}^T\boldsymbol{\lambda}^* = 1$, rendering $\lambda_0 = 0$. Note that with $\sigma = \sigma^*$, the solution of (22) and (23) is given by $[\lambda_0; \boldsymbol{\lambda}] = [0; \boldsymbol{\lambda}^*]$ where $\boldsymbol{\lambda}^*$ is the output of equation (23). Thus, to compute the optimal Lagrange multipliers and bandwidth we solve the program (23), the output of which also solves (22) and (23). This time, however, the solution of (22) and (23), given by equation (21), is a proper mixture pdf with non-negative mixture weights $\boldsymbol{\lambda}^*$.

Explicit calculation of the entries of matrix C is possible if, for example, we have $\mathcal{X} = \mathbb{R}^n$ and $p \propto 1$, giving $C_{ij} = \int_{\mathbb{R}^n} K_i(\mathbf{x}) K_j(\mathbf{x})\, d\mathbf{x} = K(\mathbf{X}_i \,|\, \mathbf{X}_j, \sqrt{2}\sigma)$, where K_i is a Gaussian pdf with bandwidth σ anchored at the point $\mathbf{X}_i$. In this case $\mathbf{c} = [1, \ldots, 1]^T$ in equation (23).

To summarize, a valid solution of the GCE program is the mixture pdf equation (21), with $\lambda_0 = 0$ and non-negative mixture components $\boldsymbol{\lambda}^*$ calculated from the program (23) with $\mathbf{c} = \mathbf{1}$, obtained by using $\phi(x) = (x^2 - 1)/2$ in $\mathcal{D}(g \to p)$, inequality moment constraint in equation (11), $p \propto 1$ and $K_i(\mathbf{x}) = K(\mathbf{x} \,|\, \mathbf{X}_i, \sigma) = c \exp\left(-||\mathbf{x} - \mathbf{X}_i||^2/(2\sigma^2)\right)$, $i = 1, \ldots, m$.

Remark 5 (Estimating $\kappa_i = \mathbb{E}_\pi K_i(\mathbf{X})$) Assume we use the same set $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ as both location parameters for the Gaussians $K_i(\mathbf{x}) = K(\mathbf{x} \,|\, \mathbf{X}_i, \sigma)$ and as a sample for the estimation of each $\mathbb{E}_\pi K_i(\mathbf{X})$. Note that $\mathbb{E}_\pi K_i(\mathbf{X})$ is a function of $\mathbf{X}_i$ and hence is a random variable. Under the assumption that $\mathbf{X}_1, \ldots, \mathbf{X}_m \sim_{iid} \pi$, 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:

$$\hat{\kappa}_i = \frac{1}{m-1} \sum_{j \neq i}^{m} K_i(\mathbf{X}_j). \quad (24)$$

This is the *cross-validatory*, also known as *leave-one-out* estimator and its consistency properties are established in Bowman [19].

Example 5 (Exponential Distribution, Examples 1 & 3 continued) Suppose once more that we wish to estimate $\ell = \mathbb{P}_u(X \geqslant \gamma)$, with $X \sim \text{Exp}(u^{-1})$, as in Examples 1 and 3. Further, suppose that we have a random sample $\mu_1, \mu_2, \ldots, \mu_N$ from $p \sim \text{Exp}(\nu)$. Finally, suppose that we wish to build an approximation to the target $\pi(x) = I_{\{x \geqslant \gamma\}} u^{-1} \mathrm{e}^{-u^{-1}(x-\gamma)}$ using Gaussian kernels, of the form

$$K_k(x) = K(x \,|\, \mu_k, \sigma) = (\sqrt{2\pi}\sigma)^{-1} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}.$$

For the χ^2 GCE quadratic programming problem, we calculate

$$c_i = \mathbb{E}_p K_i(X) = \left(1 - erf\left(\frac{\nu\sigma^2 - \mu_i}{\sqrt{2}\sigma}\right)\right) \frac{\nu \mathrm{e}^{\frac{\sigma^2\nu^2}{2} - \nu\mu_i}}{2},$$

$$\kappa_i = \mathbb{E}_\pi K_i(X) = \left(1 - erf\left(\frac{(\gamma - \mu_i)}{\sqrt{2}\sigma} + \frac{\sigma}{\sqrt{2}u}\right)\right) \frac{\mathrm{e}^{\frac{(\gamma-\mu_i)}{u} + \frac{\sigma^2}{2u^2}}}{2u},$$

and

$$C_{ij} = \left(1 - erf\left(\frac{\sigma^2\nu - (\mu_i + \mu_j)}{2\sigma}\right)\right) \times \frac{\nu \mathrm{e}^{-\frac{1}{2}\left(\frac{(\mu_i - \mu_j)^2 - \sigma^4\nu^2 + 2\sigma^2\nu(\mu_j + \mu_j)}{2\sigma^2}\right)}}{2}.$$

As a concrete example, suppose $u = 1$, $\gamma = 10$ and $\nu^{-1} = (11 + \sqrt{101})/2 \approx 10.5249$, corresponding to the VM optimal parameter ${}_*\upsilon$ in Example 1. A typical outcome of the GCE procedure is depicted in Figure 1, using a sample of size $m = 30$ from p.

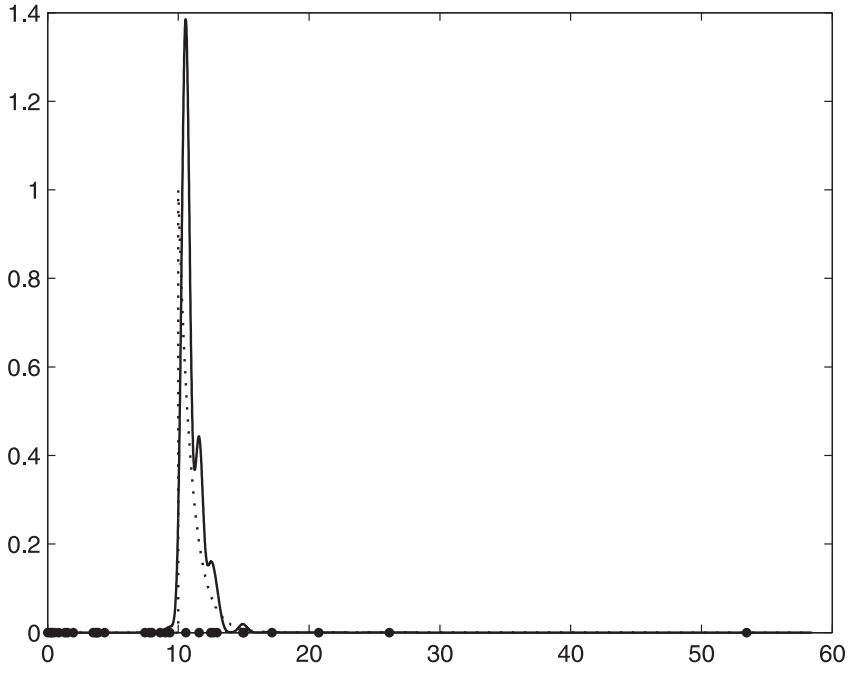

Figure 1. Target $\pi(x)$ (dotted) and kernel mixture $g(x)$ (solid)

Table 1. Non-zero weights and points of equation (21)

Weight	0.25	31.13	10.77	3.56	1.78	0.65	0.087 4	0.003 45	2.60×10^{-5}
Point	9.36	10.57	11.60	12.48	12.96	14.95	17.16	20.74	26.14

Recall that the instrumental density is a mixture of the form equation (21). The mixture has bandwidth $\sigma \approx 0.31$, with non-zero weights and points as listed in Table 1. □

4. Applications

Areas to which the GCE methodology can be applied are density estimation (both continuous and discrete), rare event probability estimation and optimization.

4.1 Density Estimation

Consider the one-dimensional density estimation problem where we are given the sample $\mathcal{X}_m \equiv \{X_1, \ldots, X_m\}$ 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 $\mathcal{X}_m$ with few stringent assumptions is the *kernel method* [6, 20, 21]. The method assumes that the true, but unknown, underlying density function π can be approximated by a pdf of the form:

$$\hat{\pi}(x \mid \sigma, \mathcal{X}_m) = \frac{1}{m\sigma} \sum_{i=1}^{m} K\left(\frac{x - X_i}{\sigma}\right),$$

where $\sigma \in \mathbb{R}_+ \backslash \{0\}$ is the bandwidth parameter, which controls the smoothness or 'resolution' of $\hat{\pi}$, and K is a positive, symmetric (around 0) and unimodal kernel. For our purposes we choose to use the Gaussian kernel $K(x) = (1/\sqrt{2\pi}) \times \exp(-x^2/2)$. Everything in the kernel estimator is fixed and known except the bandwidth σ. There are various methods [6] for tuning σ so that the approximation of π is as good as possible. Currently, the prevailing method for bandwidth selection is the Sheather–Jones (SJ) method [22]. An alternative is to use the χ^2 GCE method.

We now summarize the GCE program for this problem, as given in equation (21), the sum of equations (22) and (23) and equation (23).

4.1.1 χ^2 GCE Program for Density Estimation

For $i = 1, \ldots, m$ choose

$$K_i(x) = K(x \mid X_i, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - X_i)^2}{2\sigma^2}\right)$$

and

1. solve the program

$$(\sigma^*, \boldsymbol{\lambda}^*) = \left\{ \begin{array}{l} (\sigma, \boldsymbol{\lambda}) \;\; \mathbf{1}^T \boldsymbol{\lambda}(\sigma) = 1, \\ \boldsymbol{\lambda}(\sigma) = \operatorname{argmin}_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \left(\frac{1}{2} \boldsymbol{\lambda}^T C(\sigma) \boldsymbol{\lambda} - \boldsymbol{\lambda}^T \hat{\kappa}(\sigma) \right) \end{array} \right\}, \quad (25)$$

where the $m \times m$ matrix C has entries

$$C_{ij} = \int_{\mathbb{R}} K_i(x) K_j(x) dx = \frac{\exp\left(\frac{(X_i - X_j)^2}{(4\sigma^2)} \right)}{\sqrt{4\pi\sigma^2}},$$

$\hat{\kappa}(\sigma)$ is the cross-validatory estimator in equation (24), and

2. present the weighted Gaussian mixture density

$$g(x) = \sum_{j=1}^{m} \lambda_j^* K(x | X_j, \sigma^*) \quad (26)$$

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

Note that we have obtained a standard kernel density estimator in equation (26) with weights. Hall and Turlach [23] have studied the asymptotic properties of estimators of the form equation (26).

4.2 Discrete Density Estimation

Assume that we are given the binary data $\mathcal{X}_m \equiv \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$, where $\{\mathbf{X}_i\}$ are n-dimensional binary vectors. Let X_{il} denote the lth component of $\mathbf{X}_i$. We model the data using a kernel estimator with the discrete kernel given by:

$$K_i(\mathbf{x}) = K(\mathbf{x} \mid \mathbf{X}_i, \sigma) = \prod_{l=1}^{n} \sigma^{I_{\{x_l = X_{il}\}}} (1-\sigma)^{1 - I_{\{x_l = X_{il}\}}}$$

$$= \sigma^{d(\mathbf{x}, \mathbf{X}_i)} (1-\sigma)^{n - d(\mathbf{x}, \mathbf{X}_i)},$$

where $n - d(\mathbf{x}, \mathbf{y}) = n - \sum_{l=1}^{n} I_{\{x_l = y_l\}}$ measures the number of mismatches between the vectors $\mathbf{x}$ and $\mathbf{y}$ (i.e. the Hamming distance). The kernel is therefore a multivariate Bernoulli pmf. Note that:

$$\lim_{\sigma \uparrow 1} K(\mathbf{x} \mid \mathbf{X}_i, \sigma) = \left\{ \begin{array}{lll} 1, & \text{if} & \mathbf{x} = \mathbf{X}_i \\ 0, & \text{if} & \mathbf{x} \neq \mathbf{X}_i \end{array} \right., \quad (27)$$

$$K\left(\mathbf{x} \mid \mathbf{X}_i, \tfrac{1}{2}\right) = \frac{1}{2^n}. \quad (28)$$

The end points of the interval $\left[1/2, 1\right]$ represent two extremes of modeling the data: for $\sigma = 1$, g is simply estimated from the corresponding relative frequencies; for $\sigma = 1/2$, K is not unimodal and g is the uniform pmf on $\mathcal{X}$. We find the non-trivial value of $\sigma \in (1/2, 1)$ by solving the following χ^2 GCE program.

4.2.1 χ^2 GCE Program for Discrete Density Estimation

Choose the binary kernel $K_i(\mathbf{x}) = K(\mathbf{x} \mid \mathbf{X}_i, \sigma) = \sigma^{d(\mathbf{x}, \mathbf{X}_i)} (1-\sigma)^{n - d(\mathbf{x}, \mathbf{X}_i)}$ and

1. solve the program

$$(\sigma^*, \boldsymbol{\lambda}^*) = \left\{ \begin{array}{l} (\sigma, \boldsymbol{\lambda}) \;\; \mathbf{1}^T \boldsymbol{\lambda}(\sigma) = 1, \\ \boldsymbol{\lambda}(\sigma) = \operatorname{argmin}_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \left(\frac{1}{2} \boldsymbol{\lambda}^T C(\sigma) \boldsymbol{\lambda} - \boldsymbol{\lambda}^T \hat{\kappa}(\sigma) \right) \end{array} \right\}, \quad (29)$$

with the matrix C given by

$$C_{ij} = \sum_{\mathbf{x} \in \mathcal{X}} K_i(\mathbf{x}) K_j(\mathbf{x}) = K(\mathbf{X}_i \mid \mathbf{X}_j, \varsigma),$$

$\varsigma = \sigma^2 + (1-\sigma)^2$, $i, j = 1, \ldots, m$, and $\hat{\kappa}(\sigma)$ is again the cross-validatory estimator in equation (24), and

2. present the weighted Bernoulli mixture density

$$g(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j^* K(\mathbf{x} \mid \mathbf{X}_j, \sigma^*) \quad (30)$$

as the optimal GCE probability mass function that models the discrete data $\mathcal{X}_m$.

4.3 Rare-event Simulation and Optimization

We now explain how the parametric estimation procedure for the instrumental via KL CE minimization at each step of the CE method can be substituted with a non-parametric estimation procedure such as the χ^2 GCE program, the MCE method or the kernel method of Jones et al. [22].

Recall that for each level γ_t, $t = 1, 2, \ldots$ in the CE method we estimate an optimal parameter $\mathbf{v}_t^*$ associated with the parametric instrumental $g(\mathbf{x}) \equiv f(\cdot; \mathbf{v})$, i.e. the functional form of the instrumental is kept fixed, and only its parameter $\mathbf{v}$ is updated at each step t. In the MCE and GCE framework, however, for each level γ_t we update the instrumental $g(\mathbf{x})$ non-parametrically. Thus for each level γ_t, instead of a sequence of finite dimensional parameters $\{\mathbf{v}_t, \; t = 1, 2, \ldots\}$, we estimate a sequence of instrumentals $\{g_t(\mathbf{x}), \; t = 1, 2, \ldots\}$ which approximates the optimal sequence of IS densities, $\{\pi_t(\mathbf{x}) \propto I_{\{S(\mathbf{x}) > \gamma_t\}} f(\mathbf{x}; \mathbf{u}), \; t = 1, 2, \ldots\}$. For each t the CE program is reformulated in the following way. Begin with a uniform instrumental $g_0(\mathbf{x})$ over $\mathcal{X}$ or $g_0(\mathbf{x}) \equiv f(\mathbf{x}; \mathbf{u})$.

1. **Adaptive updating of γ_t.** For a given $g_{t-1}(\mathbf{x})$, let γ_t be the $(1-\rho)$-quantile of $S(\mathbf{X})$ under $g_{t-1}(\mathbf{x})$. We can estimate γ_t by drawing a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from $g_{t-1}(\mathbf{x})$ and evaluating the sample $(1-\rho)$-quantile $\hat{\gamma}_t$.

2. **Updating of $g_t(\mathbf{x})$.** For a given $\hat{\gamma}_t$ and $g_{t-1}(\mathbf{x})$, estimate a new instrumental $g_t(\mathbf{x})$ using either the χ^2 GCE program, the MCE program or the Sheather–Jones kernel method [22].

For example, in the MCE method, $g_t(\mathbf{x})$ is obtained from the program:

$$\begin{aligned} & \min_{g_t} \mathcal{D}(g_t \;\rightarrow\; g_{t-1}) \\ = & \int_{\mathcal{X}} g_t(\mathbf{x}) \ln(g_t(\mathbf{x})/g_{t-1}(\mathbf{x}))\, d\mathbf{x} \\ & \text{subject to:} \quad \mathbb{E}_{g_t} K_i(\mathbf{X}) \\ = & \mathbb{E}_{\pi_t} K_i(\mathbf{X}), \quad i = 0, \ldots, m. \end{aligned}$$

Ideally, in the GCE framework, at iteration t we would like to take $p_t = g_{t-1} \approx \pi_{t-1}$, as is done with MCE. In the χ^2 GCE program, however, we choose $p_t \propto 1, \;\; \forall t$ in order to obtain simple closed-form entries $C_{ij} = \mathbb{E}_p K_i(\mathbf{X}) K_j(\mathbf{X})$ for the Hessian matrix in the QPP (22)+(23). Thus, with χ^2 GCE, g_t is the output of the program:

$$\begin{aligned} & \min_{g_t} -\frac{1}{2} + \frac{1}{2} \int_{\mathbb{R}^n} \frac{g_t^2(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} \\ \equiv & \min_{g_t} \int_{\mathbb{R}^n} g_t^2(\mathbf{x})\, d\mathbf{x} \\ & \text{subject to:} \quad \mathbb{E}_{g_t} K_i(\mathbf{X}) \\ \geqslant & \kappa_i^{(t)} = \mathbb{E}_{\pi_t} K_i(\mathbf{X}), \quad i = 0, \ldots, m = N, \end{aligned}$$

where, just as with density estimation, $\{K_i(\mathbf{x})\}_{i=1}^m$ is a Gaussian kernel with mean at $\mathbf{X}_i$ and with common bandwidth σ^*, given by the output of the root-finding program (23). Note that we have as many constraints as number of samples, i.e. $m = N$, and so the QPP involves N constraints. This is not a major problem if the QPP solver exploits the convexity of the problem.

Remark 6 In practice, at each step t, each $\kappa_i^{(t)}$ needs to be estimated. We suggest the following procedure. Given a sample from g_{t-1}, we estimate $\kappa_i^{(t)}$ via the IS estimator

$$\hat{\kappa}_i^{(t)} = \sum_{j=1, j\neq i}^{m} \frac{\pi_t(\mathbf{X}_j)}{g_{t-1}(\mathbf{X}_j)} K_i(\mathbf{X}_j) \Big/ \sum_{j=1, j\neq i}^{m} \frac{\pi_t(\mathbf{X}_j)}{g_{t-1}(\mathbf{X}_j)},$$

$$\mathbf{X}_1, \ldots, \mathbf{X}_m \;\sim_{iid}\; g_{t-1}.$$

Here the sums do not include the ith component, in keeping with the cross-validatory approach explained in Remark 5. An alternative, which is stochastically equivalent to the IS estimator and which is easier to implement in a practical simulation, is to employ the *sample importance resampling* (SIR) method (e.g. [24]) to obtain a sample $\mathbf{X}_1^*, \ldots, \mathbf{X}_m^* \sim_{approx.} \pi_t$ and then use the estimator

$$\hat{\kappa}_i^{(t)} = \frac{1}{m-1} \sum_{\{j: \mathbf{X}_j^* \neq \mathbf{X}_i\}} K_i(\mathbf{X}_j^*).$$

Since the solution of the χ^2 GCE has the functional form of a non-parametric kernel density estimator, it is quite natural to consider constructing an instrumental via a standard kernel density estimation method such as the Sheather–Jones (SJ) method [24]. In particular, step 2 in the above algorithm can be substituted with:

2. **Updating of $g_t(\mathbf{x})$.** For a given $\hat{\gamma}_t$ and $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{iid} g_{t-1}(\mathbf{x})$, compute the normalized IS weights

$$\begin{aligned} w_i^{(t)} \;\propto\; & \frac{\pi_t(\mathbf{X}_i)}{g_{t-1}(\mathbf{X}_i)} \propto \frac{f(\mathbf{X}_i; \mathbf{u}) I_{\{S(\mathbf{X}_i) > \gamma_t\}}}{g_{t-1}(\mathbf{X}_i)}, \\ i \;=\; & 1, \ldots, N, \end{aligned}$$

then apply SIR to $\mathbf{X}_1, \ldots, \mathbf{X}_N$ to obtain the new sample $\mathbf{X}_1^*, \ldots, \mathbf{X}_N^*$. Based on the elite sample $\mathbf{X}_1^*, \ldots, \mathbf{X}_N^*$ construct the SJ non-parametric kernel density estimator.

Determining which approach (GCE, MCE, CE, SJ, etc.) is most appropriate in certain situations remains a matter for further research. MCE allows us to use prior information easily but yields model pdfs that are difficult to sample from; χ^2 GCE does not provide a practical way to incorporate prior information, but yields solutions that are easy to sample from. In addition, solving the QPP can be quite slow. Using SJ-type methods side-steps the need to solve a QPP, but provides no objective method for estimating the bandwidth in higher dimensions.

4.3.1 Optimization

The rare-event estimation procedure can be easily modified to an optimization procedure. Suppose we wish to maximize a function $S(\mathbf{x})$ over $\mathbf{x} \in \mathcal{X}$. The idea, exactly as in the CE method, is to associate with this optimization problem a rare-event estimation problem, namely the estimation of $\mathbb{E}_f I_{\{S(\mathbf{X}) \geqslant \gamma\}}$, where γ is left unspecified and f is some pdf on $\mathcal{X}$. By using a multi-level approach, and choosing the target pdf at each iteration t to be the uniform distribution on the level set $\{S(\mathbf{x}) \geqslant \gamma_{t-1}\}$, a sequence of levels $\{\gamma_t\}$ and instrumentals $\{g_t\}$ is generated such that the former increases towards the maximum γ^* and the latter are steered towards the degenerate measure at $\mathbf{x}^* \in \text{argmax}\, S(\mathbf{x})$.

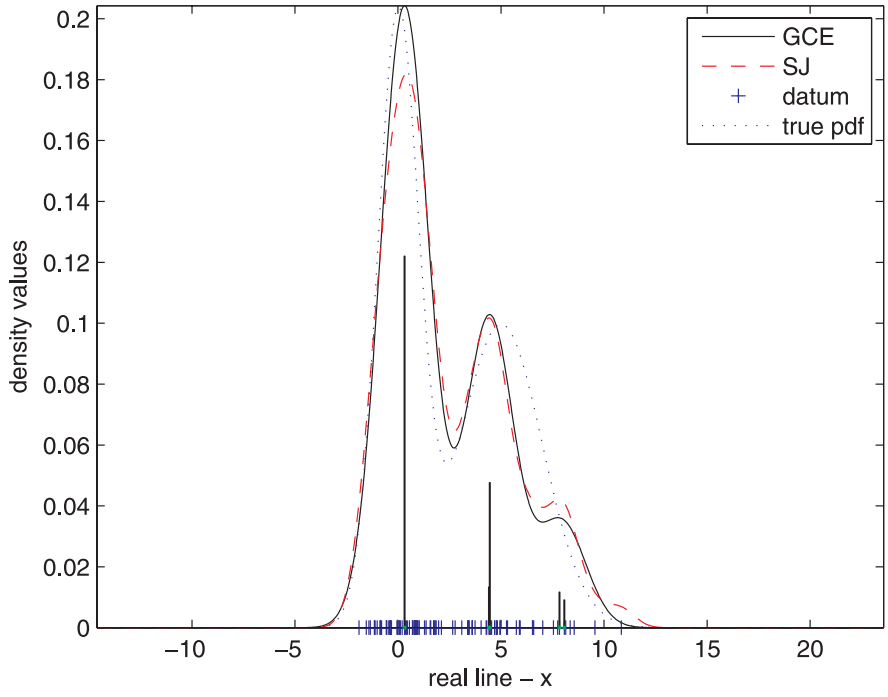

Figure 2. Data modeling via the GCE and kernel methods

5. Numerical Experiments

We used MatLab software in all of the numerical experiments. Although we solved the QPP using the Optimization toolbox of MatLab, we recommend the use of a quadratic programing solver which can exploit the convexity of the QPP problem to achieve greater efficiency.

5.1 Density Estimation

Suppose we are given 100 data points from the Gaussian mixture model $\pi \sim \mathrm{N}(0,1)/2 + \mathrm{N}(5,4)/2$, in obvious notation. Given the data only, we wish to estimate the underlying density. Figure 2 shows the result of a typical density estimation experiment using the SJ [22] and the GCE density estimation method. The long thin bars above the data points represent the relative values of the Lagrange multipliers $\boldsymbol{\lambda}$ (i.e. mixture weights of equation (30)) associated with each point. Figure 3 shows the ratio of the exact integrated squared error (ISE) $\int (g_{\mathrm{GCE}}(x) - \pi(x))^2 dx / \int (g_{\mathrm{SJ}}(x) - \pi(x))^2 dx$ over 200 Monte Carlo experiments. The integration was performed numerically over the range $x \in [-10, 15]$ using 4000 regularly spaced points. Figure 3 shows that the error of the GCE estimator in estimating the pdf π from data is comparable to the error of the SJ kernel method.

From Figure 2 and other typical simulation experiments we can conclude the following. 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 [25], the model obtained via the GCE method is more sparse than that obtained via the traditional SJ 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 equation (30). The main disadvantage of the GCE method is that the computational cost of calculating the Lagrange multipliers via the associated QPP increases dramatically with the number of points m (complexity $O(5\,m^3)$). This makes the approach currently impractical for large sample sizes. Another problem, similar to a major problem with the support vector machine methodology, is that the number of non-zero Lagrange multipliers decreases as the dimension of the problem increases, and so the number of 'support vectors' increases with the increased dimensionality of the problem.

As a second example, we consider density estimation for a log-normal density. Figure 4 shows 800 points generated from a log-normal density with location 0 and scale 1, as well as the SJ and GCE estimates.

It is interesting to note that out of the 800 points only 21 points have non-zero Lagrange multipliers. Thus the GCE model for the 800 points is a Gaussian mixture with only 21 components. In contrast, the SJ estimator is an equally weighted mixture with 800 components. The spar-

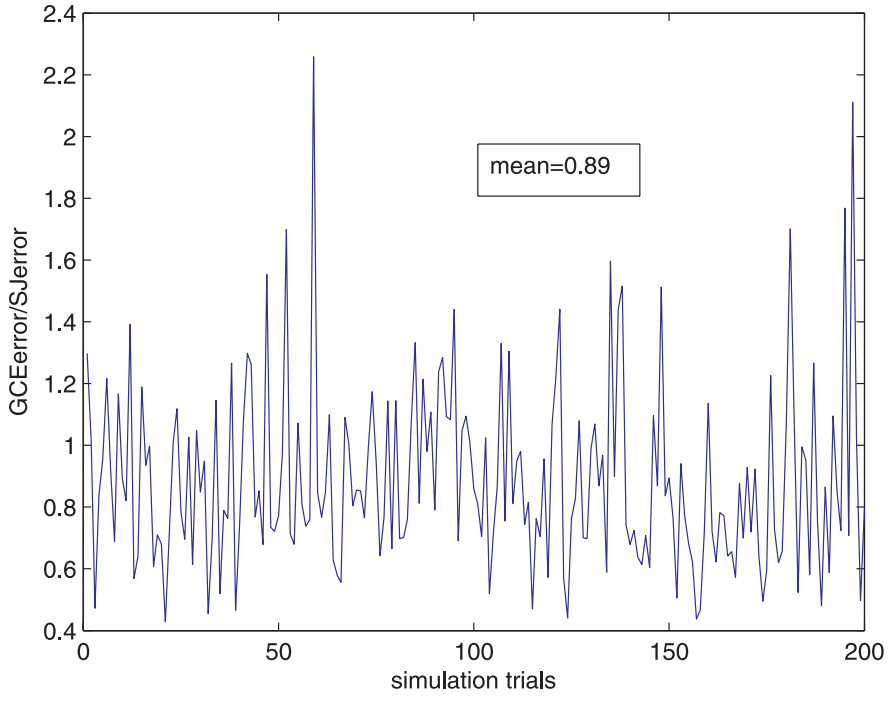

Figure 3. Behavior of the ratio of the GCE error to SJ error

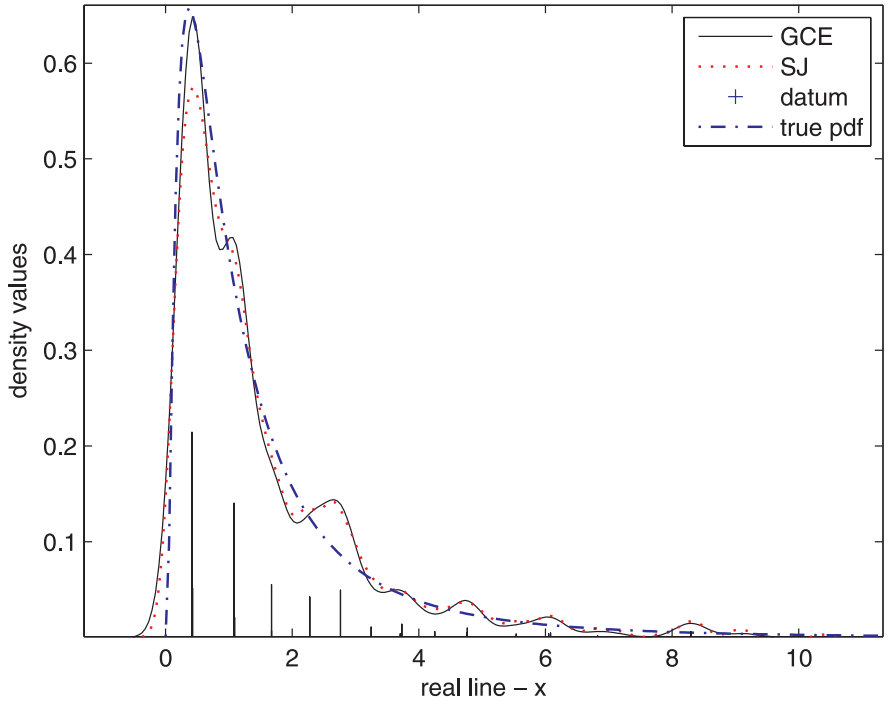

Figure 4. 800 points from the log-normal density with location 0 and scale 1

Table 2. A well-known medical dataset used as a test dataset for binary discrimination with kernel models

	data											data									
obs.	1	2	3	4	5	6	7	8	9	10	obs.	1	2	3	4	5	6	7	8	9	10
1	1	1	1	0	1	0	1	0	0	1	21	0	0	0	0	1	0	0	0	0	0
2	1	1	1	1	1	0	0	1	0	0	22	1	1	1	1	1	1	1	0	0	0
3	1	1	0	1	1	1	0	0	1	0	23	1	1	1	0	1	0	0	0	0	1
4	1	1	0	1	1	0	0	1	1	0	24	1	1	0	1	0	0	1	1	1	0
5	1	1	1	1	0	0	1	0	0	1	25	1	1	1	1	0	0	0	1	0	0
6	1	1	0	0	1	0	0	0	0	1	26	0	0	0	1	0	0	1	0	0	0
7	1	1	1	0	0	1	0	1	0	0	27	1	1	0	1	1	0	0	1	1	1
8	1	1	0	1	0	1	0	0	1	0	28	1	1	1	1	0	0	0	0	0	1
9	1	1	1	1	1	0	1	1	0	0	29	1	0	1	0	1	0	0	0	1	0
10	1	0	0	0	0	0	0	0	0	0	30	1	1	0	1	0	1	0	0	0	1
11	1	1	1	1	0	1	0	1	0	0	31	0	1	1	1	0	0	0	0	0	1
12	1	1	0	0	0	0	1	1	1	0	32	1	1	1	1	1	1	1	0	0	1
13	1	1	1	1	1	1	1	0	0	1	33	0	0	1	1	1	0	1	0	1	0
14	1	1	1	1	1	0	1	1	0	1	34	1	1	1	1	0	1	1	0	0	1
15	0	0	1	1	0	0	1	1	0	0	35	1	0	1	0	1	0	0	1	0	0
16	1	1	0	1	0	0	0	0	1	0	36	1	1	1	1	0	0	0	1	0	0
17	0	1	1	0	0	1	0	1	0	1	37	1	1	1	1	1	0	0	0	0	0
18	1	0	1	1	1	0	0	1	0	0	38	1	1	0	0	0	0	0	0	0	0
19	1	0	1	1	1	0	1	0	0	0	39	0	0	0	0	0	0	0	0	0	0
20	1	1	1	1	0	0	1	0	0	1	40	0	1	1	1	0	0	1	0	0	1

sity of the GCE estimator makes it computationally easier to evaluate at each point and to visualize the pdf.

5.2 Discrete Density Estimation

Table 2 represents a well-known medical binary data set described in Anderson et al. [26] and used throughout the statistical literature to test non-parametric statistical models [27, 28, 29]. The data describe 40 patients suffering from a certain disease. Each patient may or may not have any of 10 possible symptoms. The presence of the symptoms is represented as binary row vectors of length 10. A 1 means that the symptom is present and a 0 represents the lack of that symptom. The aim is then to build a model for the data showing which combination of symptoms are most likely to indicate the presence of the disease in patients.

For the data in Table 2, we obtained the Bernoulli mixture model given in Table 3. Note that we can read off the most weighty pattern in Table 1 from the mixture model in Table 3. More specifically, observations 20, 32 and 36 are representatives of the most predominant binary pattern in Table 2. Patients exhibiting these patterns of symptoms would therefore most likely be considered stricken with the disease. Also note that the GCE mixture pmf that models the data is sparse in the sense that the number of mixture components is usually much smaller than the number of observations. This is in sharp contrast to the traditional discrete kernel density estimation techniques where the model pmf is an equally weighted mixture with as many components as the number of observations.

5.3 Estimation Examples

In this section we give examples of estimating quantities of the form $\ell = \mathbb{E}_f H(\mathbf{X}; \gamma)$, for some function H possibly dependent on a level parameter γ. The estimation is achieved by building a sequence of instrumentals $\{g_t,\ t = 1, 2, \dots\}$ approximating the sequence of optimal IS densities $\{\pi_t \propto |H(\mathbf{x}; \gamma_t)| f(\mathbf{x}),\ t = 1, 2, \dots\}$, and subsequently estimating the quantity of interest using the IS estimator with g_T as the instrumental.

More specifically, for each of the estimation examples, we start with a uniform sample over an appropriate rectangular region. We proceed by iteratively resampling the points with weights π_t / g_{t-1}, where g_{t-1} is the approximation for level γ_{t-1}. We then build a new IS density g_t as a Gaussian kernel mixture. Next, we sample from this new density, and repeat the resampling-approximation-sampling process, until the final level γ has been hit. We perform a further P iterations with the level fixed to be γ, and then estimate ℓ using the IS estimator with $N_1 = 10^5$ samples from the final density g.

Table 3. The mixture pmf with $\sigma^* = 0.792\,75$. The table presents the mixture weight and location for each of the twenty kernels constituting the mixture model for the medical dataset in Table 2

ith binary vector	$K(x \mid \sigma^*, \mathbf{X}_i)$ with $\mathbf{X}_i$ given by										ith weight λ_i
20	1	1	1	1	0	0	1	0	0	1	0.22707
32	1	1	1	1	1	1	1	0	0	1	0.18974
36	1	1	1	1	0	0	0	1	0	0	0.18358
2	1	1	1	1	1	0	0	1	0	0	0.095159
8	1	1	0	1	0	1	0	0	1	0	0.05691
39	0	0	0	0	0	0	0	0	0	0	0.055622
10	1	0	0	0	0	0	0	0	0	0	0.039071
23	1	1	1	0	1	0	0	0	0	1	0.037531
4	1	1	0	1	1	0	0	1	1	0	0.025143
18	1	0	1	1	1	0	0	1	0	0	0.019046
12	1	1	0	0	0	0	1	1	1	0	0.012942
6	1	1	0	0	1	0	0	0	0	1	0.011055
31	0	1	1	1	0	0	0	0	0	1	0.010266
21	0	0	0	0	1	0	0	0	0	0	0.0093176
16	1	1	0	1	0	0	0	0	1	0	0.0072441
24	1	1	0	1	0	0	1	1	1	0	0.0066307
35	1	0	1	0	1	0	0	1	0	0	0.0057993
3	1	1	0	1	1	1	0	0	1	0	0.0053602
9	1	1	1	1	1	0	1	1	0	0	0.0024979
27	1	1	0	1	1	0	0	1	1	1	2.3899×10^{-5}

5.3.1 Fused Gaussians

The following example is a well-known test case for Monte Carlo algorithms. The problem is to estimate $\ell = \mathbb{E}_f X$, where f is given by the following 'fusion' of two Gaussian densities:

$$f(x, y) = c^{-1} \mathrm{e}^{-\frac{1}{2}\left(x^2y^2+x^2+y^2-8(x+y)\right)} \equiv c^{-1} f_0(x, y),$$

where c is assumed to be unknown (in fact, $c \approx 20\,216.335\,877\,352$). Note that in this example we have $H(x, y; \gamma) = x$, and so γ is irrelevant.

Figure 5 depicts a contour plot of the optimal IS density. By direct numerical integration, one can obtain the approximation $\ell = \mathbb{E}_f X \approx 1.859\,97$.

Since we take the normalization constant c to be unknown, it must be estimated. This is done with a random sample of size N_1 from g via

$$\hat{c} = \frac{1}{N_1} \sum_{i=1}^{N_1} \frac{f_0(X_i, Y_i)}{g(X_i, Y_i)}.$$

The quantity of interest is estimated using

$$\hat{\ell} = \frac{1}{N_1} \sum_{i=1}^{N_1} \frac{X_i \, f_0(X_i, Y_i)}{\hat{c}\, g(X_i, Y_i)}.$$

In order to illustrate the importance of good instrumental densities g, we run three experiments in which g is built up non-parametrically using different numbers of samples and iterations.

For case (1), we take $N = 5000$ samples per iteration for $P = 1$ iteration; for case (2), we take $N = 100$ samples per iteration for $P = 20$ iterations; and for case (3), we take $N = 1000$ samples per iteration for $P = 2$ iterations.

For this example, we start with a uniform sample over $[-2, 7]^2$ and the target density on iteration t is $\pi_t(x, y) \propto |x| f_0(x, y)$.

In Table 4, we give the minimum, maximum and average estimated relative errors over ten trials, as well as the smallest, largest and average point estimates for each of the three experimental setups.

We find that the trial that gave the smallest estimated relative error was (1); Figure 6 depicts the contour plot. This plot shows reasonable similarity to the contour plot of the true density, suggesting that we have built a reasonable importance sampling density.

On the other hand, we find that the contour plot of the most biased trial in (2) is visibly misshapen, as can be readily seen from Figure 7. This leads to the poor average performance of this experiment. We can imagine that this particular trial has a very biased g after the first resampling step, and it has too few iterations for π to be recovered.

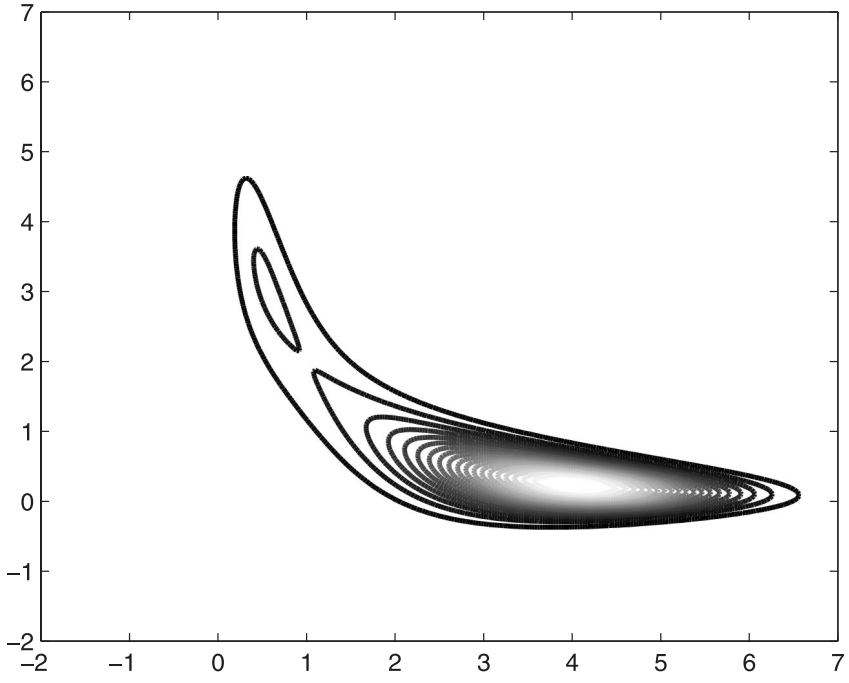

Figure 5. Contour plot of $\pi \propto |x| f_0(x, y)$

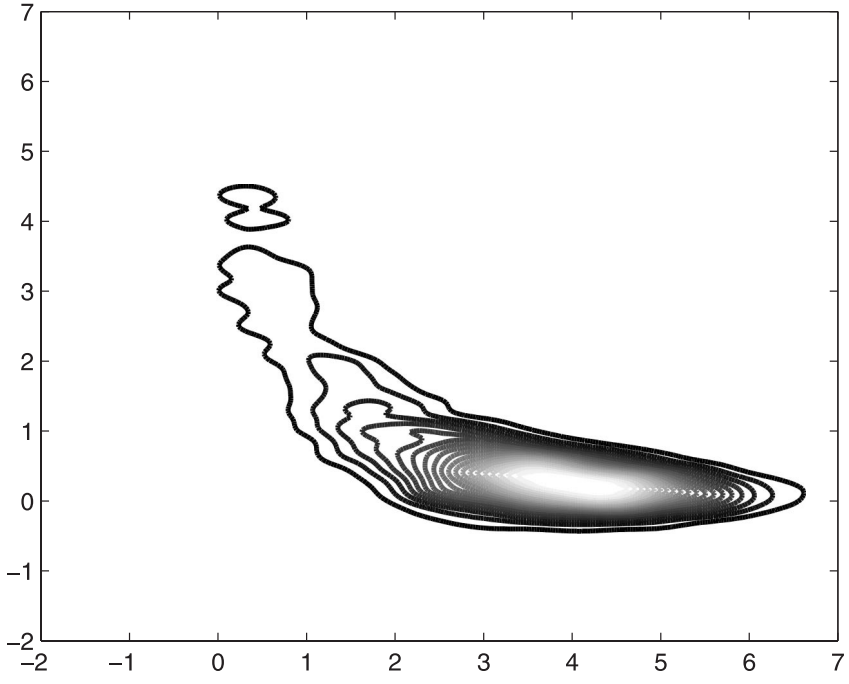

Figure 6. Contour plot of $g(x, y)$

Table 4. Fused Gaussians example

	min re	max re	average re	min est	max est	average est
1	0.00701	0.00802	0.00759	1.84503	1.89344	1.86620
2	0.00744	0.19385	0.05622	1.91516	3.31041	2.76723
3	0.00655	0.02544	0.01150	1.85009	1.95802	1.89823

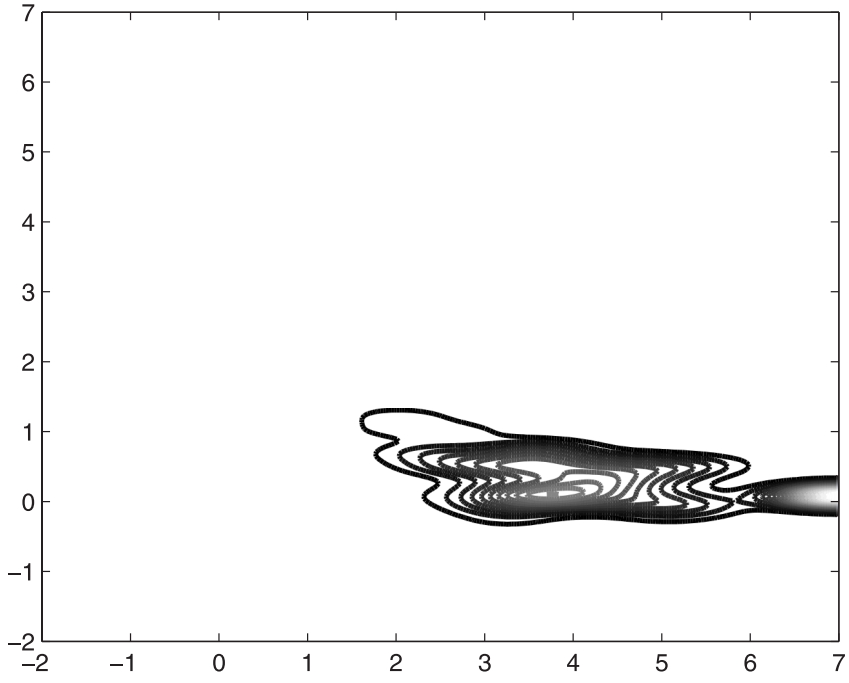

Figure 7. Contour plot of $g(x, y)$

5.3.2 Estimating $\mathbb{P}_f(X + Y > \gamma)$

The problem of interest is to estimate

$$\ell = \mathbb{P}_f\,(X + Y > \gamma)\,,$$

where X and Y are independent random variables with density

$$h(z) = I_{\{z \geqslant 0\}} b \mathrm{e}^{-bz},$$

and so $f(x, y) = h(x)h(y)$. In this example, $H(x, y; \gamma) = I_{\{x+y>\gamma\}}$ and we can calculate

$$\ell = \mathbb{P}_f\,(X + Y > \gamma) = (1 + b\gamma)\mathrm{e}^{-b\gamma}\,.$$

The target distribution π with parameters $b = 1$ and $\gamma = 10$ is depicted in Figure 8.

Suppose that $b = 5$, $\gamma = 10$. In this case, we can calculate the true probability, which is $\ell = 51\,\mathrm{e}^{-50} \approx 9.836\,62 \times 10^{-21}$. A typical contour plot of the final g for this example is given in Figure 9.

In a similar vein to the previous estimation example, we run three experiments in which g is is built up nonparametrically using different numbers of samples and iterations.

For case (1), we take $N = 1000$, $N_e = 500$ and $P = 20$; for case (2), we take $N = 300$, $N_e = 150$ and $P = 20$; and for case (3), we take $N = 500$, $N_e = 300$ and $P = 10$. Note that, unlike the previous example, we do have changing levels γ_t so that $\pi_t(x, y) \propto I_{\{x+y>\gamma_t\}} f(x, y)$. For this example, we start with a uniform sample over $[-10, 30]^2$. The results of these experiments are listed in Table 5.

As with the previous example, observe that taking too few samples per iteration ultimately gives rise to poor instrumental densities. Further, typical instrumental densities, such as those depicted in Figure 9, place a nonnegligible amount of mass in the lower triangular region. This prevents the relative error of the estimate from shrinking quickly as the sample size increases.

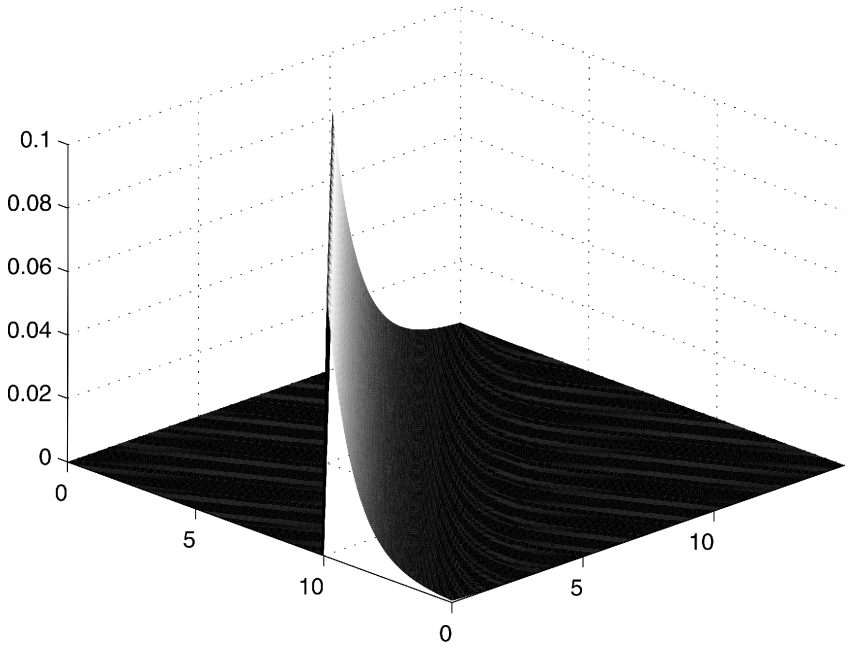

Figure 8. Plot of π for $a = b = 1$ and $\gamma = 10$

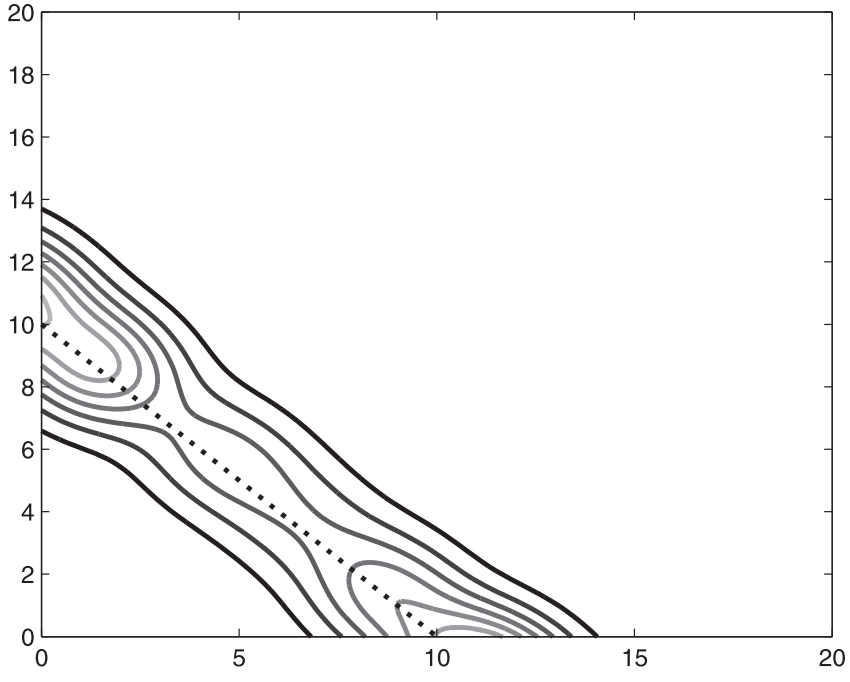

Figure 9. Contour plot of $g(x, y)$

Table 5. $\mathbb{P}_f(X + Y > \gamma)$ estimation

	min re	max re	average re	min est	max est	average est
1	0.018421	0.023962	0.020797	9.60048×10^{-21}	1.00973×10^{-20}	9.83595×10^{-21}
2	0.014821	0.231103	0.055028	8.17072×10^{-21}	1.03726×10^{-20}	9.68979×10^{-21}
3	0.015449	0.028022	0.022819	9.63451×10^{-21}	1.01999×10^{-20}	9.93677×10^{-21}

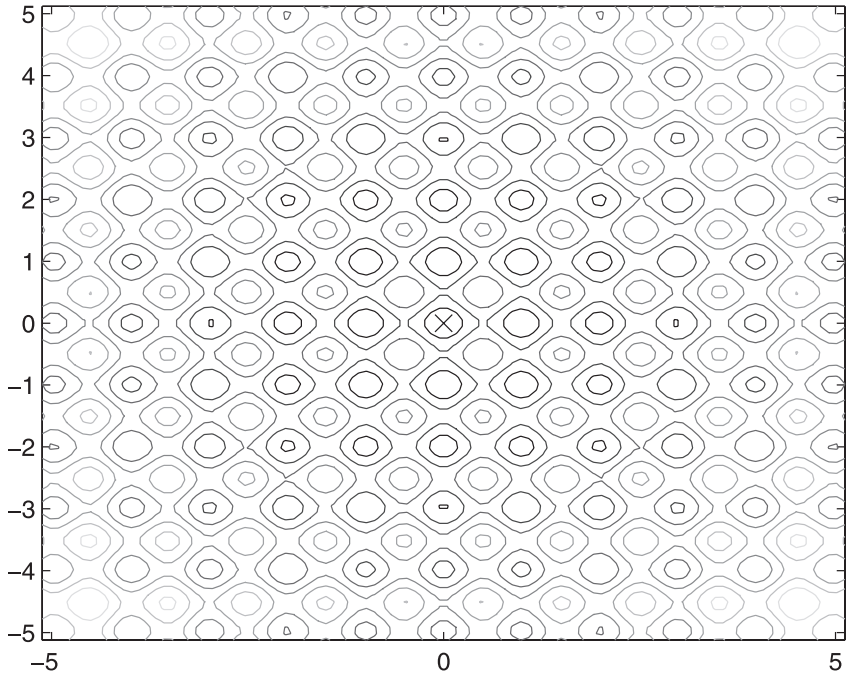

Figure 10. Contour plot of Rastrigin's function

5.4 Optimization Examples

The procedure followed for the optimization examples is almost the same as for the estimation examples. The only differences are: we do not estimate ℓ at the final step; the target density on iteration t is $\pi_t(\mathbf{x}) \propto I_{\{S(\mathbf{x}) \leqslant \gamma_t\}}$, where $\gamma_t = S_{(N_e)}$ is the largest of the elite sample scores from iteration $t-1$; and we stop the algorithm once 100 iterations have passed or $|\gamma_t - \gamma_{t-1}|, \ldots, |\gamma_t - \gamma_{t-5}| \leqslant 10^{-5}$.

Regarding the estimation examples, we run the algorithms 10 times and collect summary statistics; for these examples we collect statistics on the final function value and the number of function evaluations required by the algorithm. The statistics regarding the final function value are, for each trial, the average over 10 samples from the final pdf.

5.4.1 Rastrigin's Function

In this case, the problem is to minimize Rastrigin's function, depicted in Figure 10 and defined as

$$S(x, y) = 20 + x^2 + y^2 - 10\cos(2\pi x) - 10\cos(2\pi y),$$

which has known analytical solution $S(0, 0) = 0$ at $(x, y) = (0, 0)$.

We present the results of three numerical experiments that differ only in the settings of N and N_e. The settings for (1) were $N = 200$, $N_e = 50$; for (2) we have $N = 1000$, $N_e = 200$; and for (3) we have $N = 500$, $N_e = 100$. The initial region for this problem was $[-7.68, 7.68]^2$.

Note from the results in Table 6 that two of the ten trials in (1) gave poor results and that both runs exited by using the maximum number of allowed iterations. However, with a sample size large enough, the procedure consistently finds the global optimum.

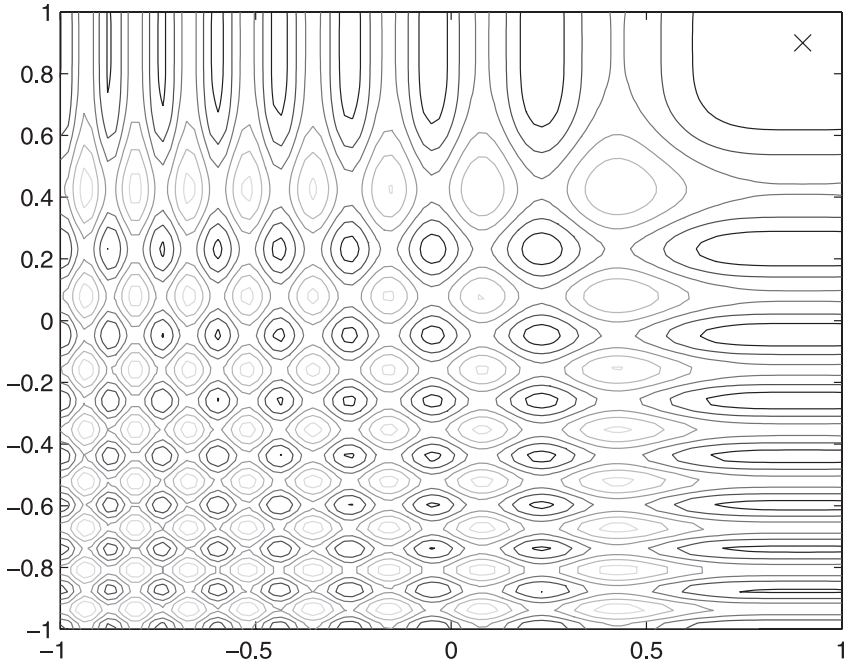

Figure 11. Contour plot of the trigonometric function with $\eta = 7$, $\mu = 1$, $x^* = 0.9$ and $y^* = 0.9$

Table 6. Rastrigin problem

	min score	max score	average score	min fevals	max fevals	average fevals
1	5.164×10^{-10}	1.679	0.275	4600	20200	7860
2	1.375×10^{-9}	6.914×10^{-9}	3.310×10^{-9}	19000	19000	19000
3	1.041×10^{-9}	1.089×10^{-4}	1.089×10^{-5}	9500	12000	9800

5.4.2 Trigonometric Function

Let $\eta = 7$, $\mu = 1$, $x^* = 0.9$ and $y^* = 0.9$. The problem is to minimize

$$\begin{aligned} S(x, y) &= 1 + 14 \left\{ \sin^2 \left[\eta (x - x^*)^2 \right] \right. \\ &+ \left. \sin^2 \left[\eta (y - y^*)^2 \right] \right\} \\ &+ \mu \left[(x - x^*)^2 + (y - y^*)^2 \right], \end{aligned}$$

depicted in Figure 11, with $-1 \leqslant x, y \leqslant 1$. The analytical solution is known to be $S(x^*, y^*) = 1$ at $(x, y) = (x^*, y^*)$.

We present the results of three numerical experiments with the same parameter settings as in the previous optimization example. The initial region for this problem was $[-1.5, 1.5]^2$, with any points falling outside $[-1, 1]^2$ set to have infinite score. The results are listed in Table 7.

Unlike the previous optimization example, the scores given in Table 7 indicate that the final density is consistently concentrated near the global minimum. Here, the effect of increasing the sample size is to improve an already adequate final density, whereas in the previous example the smallest sample size used gave inadequate results.

6. Conclusions

The GCE program provides a natural and general framework for constructing good instrumental densities g, such as those used in importance sampling, rare-event simulation, stochastic search and probability density estimation. The GCE procedure is specified by the prior density p (which conveys the available information on the target π), the function ϕ defining the CE distance and the kernels $\{K_i\}$ whose expectations under g are matched with the expectations under π. The instrumental g is derived from fundamental KKT optimization principles. By choosing p, ϕ and $\{K_i\}$ appropriately, both the classical parametric CE method (with densities in an exponential family) and the non-parametric MCE method can be recovered from the GCE method. Two choices for the distance-defining

Table 7. Trigonometric problem

	min score	max score	average score	min fevals	max fevals	average fevals
1	$1 + 2.190 \times 10^{-9}$	$1 + 1.348 \times 10^{-7}$	$1 + 1.922 \times 10^{-8}$	2800	3200	3080
2	$1 + 2.341 \times 10^{-9}$	$1 + 1.330 \times 10^{-8}$	$1 + 6.006 \times 10^{-9}$	13000	13000	13000
3	$1 + 2.396 \times 10^{-9}$	$1 + 1.086 \times 10^{-5}$	$1 + 1.172 \times 10^{-6}$	6500	7000	6550

function ϕ stand out. One gives the KL distance, used in both the CE and MCE method, and the other the χ^2 distance. The latter choice provides densities g that are weighted kernel mixtures, whose weights are found by solving a QPP. These are easy to sample from. If, in addition, the kernels are chosen to be Gaussian with a fixed bandwidth, the optimal bandwidth can be found from a simple root-finding problem.

The GCE method can be readily applied to density estimation, and compares well with the state of the art (SJ) in this area. The same holds for discrete (binary) density estimation. Notable differences between GCE and SJ are: (1) the SJ estimator relies on the availability of large samples and essentially solves an asymptotic approximation approximately, whereas GCE solves the problem without using any asymptotic approximations, and (2) the GCE gives a sparse mixture model (most weights are 0), whereas in SJ all weights (as many as there are data points) are nonzero.

Another application of the GCE method is IS estimation. Here, the crucial issue is to choose the instrumental g (the IS density) as close as possible to the target (the optimal IS pdf). As with the classical CE method, GCE can be implemented as a multi-level approach, thus steering the instrumental more gradually towards the target distribution. The method is then readily modified as a procedure to optimize general functions, as in the CE method. The advantage of using non-parametric CE approaches, as opposed to parametric approaches, is that the former are more flexible and can approximate the target better. A disadvantage is that the non-parametric algorithms tend to be significantly slower than their parametric counterparts. A future challenge is to devise non-parametric densities that are not only easy to simulate from (as in the χ^2 GCE approach) but are also fast to evaluate, especially in higher dimensions. Another direction to be explored is the use of non-Gaussian kernels, such as Cauchy kernels.

7. Acknowledgements

This project was supported by the Australian Research Council, under grant DP0558957. Thomas Taimre acknowledges the financial support of the ARC Centre of Excellence for Mathematics and Statistics of Complex Systems.

8. References

[1] Rubinstein, R. Y. and D. P. Kroese. 2004. *The Cross-Entropy Method*, New York: Springer.

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

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

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

[5] Botev, Z. I. and D. P. Kroese. 2006, http://espace.library.uq.edu.au/view.php?pid=UQ:12759. The generalized cross entropy method, with applications to probability density estimation. *Electronic Preprint*.

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

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

[8] Homem-de-Mello, T. and R. Y. Rubinstein. Rare event probability estimation for static models via cross-entropy and importance sampling. Technical Report 2002, Ohio State University: http://www.iwse.eng.ohio-state.edu/ISEFaculty/tito/pubs/rarevents.pdf.

[9] Botev, Z. I. 2005. *Stochastic Methods for Optimization and Machine Learning*. Technical Report: ePrintsUQ, http://eprint.uq.edu.au/archive/00003377/.

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

[11] Kuhn, H. W. and A. W. Tucker. 1951. Nonlinear programming. *In Proceedings of the 2nd Berkeley Symposium*, pp. 481–492. Berkeley: University of California Press.

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

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

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

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

[16] Devroye, L. and L. Gyofri. 1985. *Nonparametric Density Estimation: The L_1 View*. Wiley Series In Probability And Mathematical Statistics, New York: Wiley.

[17] Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian Data Analysis*. Chapman & Hall, 2nd edition.

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

[19] Bowman, A. W. 1985. A comparative study of some kernel-based non-parametric density estimators. *Journal of Statistical Computation and Simulation* 21, 313–327.

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

[21] Simonoff, J. S. 1996. *Smoothing Methods in Statistics*, New York: Springer.
[22] Jones, M. C., J. S. Marron, and S. J. Sheather. 1996. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* 11, 337–381.
[23] Hall, P. and B. A. Turlach. 1999. Reducing bias in curve estimation by use of weights. *Computational Statistics and Data Analysis* 30, 67–86.
[24] Robert, C. P. and G. Casella. 2004. *Monte Carlo Statistical Methods*. New York: Springer, 2nd edition.
[25] Vapnik, V. 1998. *Statistical Learning Theory*, New York: John Wiley & Sons.
[26] Anderson, J. A., K. Whale, J. Williamson, and W. W. Buchanan. 1972. A statistical aid to the diagnosis of keratoconjunctivitis sicca. *Quarterly Journal of Medicine* 41, 175–189.
[27] Aitchison, J. and C. Aitken. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* 63, 413–420.
[28] Hall, P. 1981. On nonparametric multivariate binary discrimination. *Biometrika* 68, 287–294.
[29] Titterington, D. 1980. A comparative study of kernel-based density estimates for categorical data. *Technometrics* 22, 259–268.

Zdravko I. Botev *is currently lecturing and pursuing a doctoral degree at the University of Queensland. He has won a university medal and several competitive scholarships including the Australian Bureau of Statistics scholarship.*

Dirk P. Kroese *has a wide range of publications in applied probability and simulation. He is a pioneer of the well-known Cross-Entropy method and coauthor (with R.Y. Rubinstein) of the first monograph on this method. He is associate editor of Methodology and Computing in Applied Probability and guest editor of Annals of Operations Research. He has held research and teaching positions at Princeton University and the University of Melbourne, and is currently working at the Department of Mathematics of the University of Queensland.*

Thomas Taimre *is currently pursuing a doctoral degree at the University of Queensland, Australia. He is the holder of an Australian Postgraduate Award and a scholarship from the Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex Systems.*