MTH707: Kernel Density Estimation using MCMC

Dwija Kakkad

Contents

1 Kernel Density Estimation

Density estimation is estimating the probability density function given some data. More formally, given $X_1, \ldots, X_n$ we want the probability density function from which this data was generated. The parametric approach assumes that the functional form of this density is known, and approaches like point estimation are used to estimate the parameters, and consequently the distribution. Non-parametric approaches do not keep as strict assumptions on the density function. A popular non-parametric method for density estimation is kernel density estimation (KDE). In this report, except Section 3.4, we shall consider univariate distributions only. Given data $X_1, \ldots, X_n \in \mathbb{R}$, if we want the value of the probability density function $\hat{f}$ at point y_0, the kernel density estimator is given as

$$\hat{f}(y_0) = \frac{1}{nh_n} \sum_{i=1}^{n} k\left(\frac{y_0 - X_i}{h_n}\right), \tag{1}$$

where k is a kernel function and h_n is the bandwidth, which we will later see turns out to be a function of n. The bandwidth controls the "smoothing" of the resultant density function, with a higher value of h_n giving a smoother function. We impose some assumptions on the kernel function k:

[0]The code for this project is publicly available at `https://github.com/dwija04/IS_KDE_MH`.

1. k is symmetric, i.e. $k(x) = -k(x)$, for all $x \in \mathbb{R}$.
2. $\int k(x)dx = 1$.
3. $\lim_{x\to-\infty} k(x) = \lim_{x\to\infty} k(x) = 0$.

We see now that $\hat{f}_n$ is a valid probability density, as $\hat{f}_n(x) > 0$ for all $x \in \mathscr{X}$, and $\int_{\mathscr{X}} \hat{f}_n(x)dx = 1$. The choice of kernel depends on the user, and some popular kernels include the Gaussian kernel, Uniform kernel, Epachenikov kernel etc.

1.1 Asymptotic properties of the kernel density estimator

We now analyse some properties of the kernel density estimator given in 1. Assume that $X_1, \dots X_n$ are i.i.d. samples from a distribution having an unknown density function $f : \mathscr{X} \subseteq \mathbb{R} \to R$, and assume that we want the quality of estimation of our estimator $\hat{f}$ at a point y_0. We assume that f is double differentiable, and that $f''(x)$ is bounded for all $x \in \mathscr{X}$. The bias of the KDE at y_0 is

$$\mathbb{E}[\hat{f}(y_0)] - f(y_0) = \frac{1}{2}h_n^2 f''(y_0)\mu_k + o(h_n^2), \tag{2}$$

where $\mu_k = \int t^2 k(t)dt$. Also, the variance of $\hat{f}$ is given by

$$\mathrm{Var}(\hat{f}(y_0)) = \frac{1}{nh_n} f(y)\rho_k + o\left(\frac{1}{nh_n}\right), \tag{3}$$

where $\rho_k = \int k^2(t)dt$. Consequently, we can see that the MSE of the estimator turns out to be

$$\begin{aligned}\mathrm{MSE}(\hat{f}(y_0)) &= \mathrm{bias}^2(\hat{f}(y_0)) + \mathrm{V}(\hat{f}(y_0)) \\ &= \frac{1}{4}h_n^4(f''(y_0))^2\mu_k^2 + \frac{1}{nh_n}f(y_0)\rho_k^2 + o(h_n^4) + o\left(\frac{1}{nh_n}\right).\end{aligned}$$

The bandwidth h minimizing the AMSE ($\frac{1}{4}h_n^4(f''(y_0))^2\mu_k^2 + \frac{1}{nh_n}f(y_0)\rho_k^2$) is given by

$$h_{\mathrm{opt}}(y_0) = \left(\frac{4}{n}\frac{f(y_0)\rho_k^2}{(f''(y_0))^2\mu_k^2}\right)^{1/5}.$$

Also, since $h_n \to 0$ as $n \to \infty$,

$$\sqrt{nh_n}(\hat{f}(y_0) - \mathbb{E}[\hat{f}(y_0)]) \xrightarrow{d} N(0, f(y_0)\rho_k).$$

Additionally, if $nh_n^5 \to 0$ as $n \to \infty$, then

$$\sqrt{nh_n}(\hat{f}(y_0) - f(y_0)) \xrightarrow{d} N(0, f(y_0)\rho_k). \tag{4}$$

2 Importance sampling kernel density estimation

Suppose we have i.i.d. samples $X_1, \dots X_n \in \mathbb{R}$ from a univariate distribution having density g, and f is our target density (univariate) of interest. An importance sampling kernel density estimator can be constructed. Nakayama (2011) proposes the following estimator

$$\hat{f}_{\mathrm{IS}}(y_0) = \frac{1}{nh_n}\sum_{i=1}^{n}\frac{f(X_i)}{g(X_i)}k\left(\frac{y_0 - X_i}{h_n}\right). \tag{5}$$

In this report, we will denote the term $f(s)/g(s)$ as $w(s)$. The same assumptions are imposed on the kernel as given in 1. Furthermore, assume that f is doubly differentiable, and f'' is bounded in a neighbourhood of y_0. Then, the bias of this estimator is given by

$$\mathbb{E}[\hat{f}_{\text{IS}}(y_0) - f(y_0)] = \frac{1}{2}h_n^2 f''(y_0)\mu_k + o(h_n^2), \tag{6}$$

where $\mu_k = \int t^2 k(t)dt$. Additionally, if we assume that the double derivative of g exists, and g'' is bounded in a neighbourhood of y, and $\sup_{x\in\mathscr{X}} L(x) = M' < \infty$, the variance of $\hat{f}_{\text{IS}}(y_0)$ is given by

$$\text{Var}(\hat{f}_{\text{IS}}(y_0)) = \frac{1}{nh_n}w(y_0)f(y_0)\rho_k + o\left(\frac{1}{nh_n}\right), \tag{7}$$

where $\rho_k = \int k^2(t)dt$. From 6 and 7, we can conclude that

$$\text{MSE}(\hat{f}_{\text{IS}}(y_0)) = \frac{1}{nh_n}w(y_0)f(y_0)\rho_k + o\left(\frac{1}{nh_n}\right) + h_n^4\left(\frac{f''(y_0)}{2}\mu_k\right)^2 + o(h_n^4) \tag{8}$$

as $n \to \infty$ when $h_n \to 0$ and $nh_n \to \infty$. The AMSE is defined as the higher order terms of the MSE, i.e.

$$\text{AMSE}(\hat{f}_{\text{IS}}(y_0)) = \frac{1}{nh_n}w(y_0)f(y_0)\rho_k + h_n^4\left(\frac{f''(y_0)}{2}\mu_k\right)^2.$$

Here, the bandwidth minimizing the AMSE is given by

$$h_{\text{opt}}(y_0) = \left(\frac{w(y_0)f(y_0)\rho_k}{(f''(y_0)\eta_k)^2}\right)^{1/5} n^{-1/5}. \tag{9}$$

A similar central limit theorem type result holds for this IS-KDE. Since $h_n \to 0$ as $n \to \infty$,

$$\sqrt{nh_n}(\hat{f}_{\text{IS}}(y_0) - \mathbb{E}[\hat{f}_{\text{IS}}(y_0)]) \xrightarrow{d} N(0, f(y_0)w(y_0)\rho_k),$$

and additionally if $nh_n \to 0$ as $n \to \infty$,

$$\sqrt{nh_n}(\hat{f}_{\text{IS}}(y_0) - f(y_0)) \xrightarrow{d} N(0, f(y_0)w(y_0)\rho_k). \tag{10}$$

3 Metropolis Hastings

The Metropolis Hastings algorithm ((Metropolis et al., 1953); (Hastings, 1970)) is a standard algorithm which enables us to construct a Markov chain having stationary distribution f (say), using proposals from a distribution q (say). Let $x \in \mathscr{X}$, and $A \in \mathscr{B}(\mathscr{X})$, where $\mathscr{B}(\mathscr{X})$ is the Borel-σ field of $\mathscr{X}$. The Markov transition kernel for this algorithm is given by

$$P(x, A) = \Pr(X_n \in A | X_{n-1} = x) = \int_A q(x, y)\alpha(x, y)dy + \delta_x(A)\int_{\mathscr{X}}(1 - \alpha(y, x))q(y, x)dy \tag{11}$$

Note that this kernel is not absolutely continuous with respect to the Lebesgue measure. We will establish some notation for this Markov transition kernel. Rewriting the Markov transition kernel as

$$P(x, A) = \int_A \alpha(x, y)q(x, y)dy + r(x)\delta_x(A),$$

where

$$r(x) = \int [1 - \alpha(x, y)] q(x, y) dy = 1 - a(x), \tag{12}$$

we can denote the $i-$step transition kernel

$$P^{(i)}(x, A) = \Pr(X_{i+j} \in A | X_j = x) = \int_A \tilde{p}_x^{(i)}(y) dy + r(x)^i \delta_x(A),$$

where $\tilde{p}_x^{(i)}$ is the absolutely continuous part of the Markov transition kernel $P^{(i)}(x, \cdot)$. Denote the density of X_i by $p^{(i)}$ and the corresponding measure by $P^{(i)}$.

3.1 Kernel Density Estimation using Metropolis Hastings

So far, we have dealt with i.i.d. samples from either the target, or a proposal distribution g. If we introduce some dependence in the data we have, we need to re-analyze the expressions of bias and variance for the kernel density estimator. Hart (1996) explore kernel density estimation with time-series data. We focus our attention to the case when $X_1, \dots X_n$ is a Markov chain. Note that we are still dealing with the univariate setup. Hall et al. (1995) show that the same results of asymptotic variance and bias hold when we replace the i.i.d. samples from f used in 1 with samples obtained from an MCMC algorithm where the kernel is absolutely continuous with respect to the Lebesgue measure. However, this expression changes when $X_1, \dots X_n$ are obtained using an MH algorithm, because the resultant Markov transition kernel is absolutely continuous with respect to a mixing measure, and not absolutely continuous with respect to the Lebesgue measure.
Let $X_1, \dots X_n$ be samples obtained from a Metropolis Hastings algorithm with proposal q and target f. To estimate the probability density at a point y_0, Sköld, Martin and Roberts, G O (2003) consider the following estimator

$$\hat{f}^{\text{MH}}(y_0) = \frac{1}{nh_n} \sum_{i=1}^{n} k\left(\frac{X_i - y_0}{h_n}\right), \tag{13}$$

and under certain assumptions give the expressions for asymptotic variance and bias of this estimator. Here, we will try to replicate the proofs for an importance sampling version of this estimator.

3.2 Importance Sampling Kernel Density Estimation with Metropolis Hastings

Following the notation of the previous sections, let f denote the target density we want to estimate. Let g denote the importance density, i.e. we have samples from g and we wish to estimate the density of f at $y \in \mathbb{R}$. Let $X_1, \dots X_n$ be samples from a Metropolis Hastings algorithm run with a proposal density q and having g as the invariant distribution. We assume f, g and q to be univariate densities. Consider the estimator

$$\hat{f}_{\text{IS}}^{\text{MH}}(y_0) = \frac{1}{nh_n} \sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} k\left(\frac{X_i - y_0}{h_n}\right). \tag{14}$$

We give the asymptotic bias and variance for this estimator. Although, in this report, we will assume all our Markov chains to be start from stationarity, the following theorem holds even when this assumption is relaxed. Note that assuming that the chain starts from stationarity implies that $p^{(i)}(x) = g(x)$ for all i. To prove some theoretical properties of this estimator, some lemmas are needed. Throughout this report, we will denote

$$k_h(y_0 - s) = \frac{1}{h_n} k\left(\frac{y_0 - s}{h_n}\right).$$

A thing to note here is that since $\int k(t)dt = 1$, using a simple change of variables one can easily verify that $\int k_h(t)dt = 1$. We will use this multiple times in the proof. Also, from now, we shall denote $\hat{f}_{\text{IS}}^{\text{MH}}(y_0)$ by $\hat{f}_n(y_0)$ for notational simplicity.

Lemma 1. *Let k be a compactly supported and bounded function, and function g be integrable in a neighbourhood of y. Then, Sköld, Martin and Roberts, G O (2003) mentions, as $h_n \to 0$,*

$$\int k_h(s)g(y-s)ds \to g(y)\int k(s)ds$$

for all continuity points y of g.

As a consequence of this lemma

$$\begin{aligned}\int k_h(s)g(y-s) &= \frac{1}{h}\int k\left(\frac{s}{h}\right)g(y-s)ds \\ &= \int k(t)g(y-ht)dt\end{aligned}$$

Therefore, the above lemma is the same as saying

$$\int k(t)g(y-ht)dt = g(u)\int k(t)dt \tag{15}$$

as $h \to 0$. We will use this consequence in all our proofs instead of the lemma. The compact support of k here is not needed if we assume g and k to be bounded.

Theorem 1. *Let $X_0, X_1, \ldots, X_{n-1}$ be samples from a Metropolis Hastings algorithm with target g. Fix $y_0 \in \mathbb{R}$, and suppose there are functions $V : \mathbb{R} \to \mathbb{R}^+$, $R : \mathbb{R} \to (0,1)$ and constants $\epsilon > 0$ and $M < \infty$ such that uniformly for $x \in [y-\epsilon, y+\epsilon]$, we have the following assumptions:*

1. *$|\tilde{p}_y^{(i)}(x) - g(x)| < V(y)R(i)$ and $\sum_{i=0}^{\infty} R(i) < M$.*
2. *$g(x)$, $a(x)^{-1}$, $p^{(i)}(x)$, $V(x)$, $\mathbb{E}[V(X_1)]$ are all bounded by M.*
3. *$a(x)$, $f(x)$ and $g(x)$ are uniformly continuous.*
4. *$\sup_{x\in\mathcal{X}} w(x) = M' < \infty$, where $w(x) = f(x)/g(x)$.*
5. *The Markov chain begins from stationarity, i.e. $X_1 \sim G$. This implies $p^{(i)}(x) = g(x)$, for all $x \in \mathcal{X}$ and for all $i = 1, \ldots n$.*
6. *The kernel k is compactly supported.*
7. *Lemma 1 holds.*

Note that some of these assumptions can be relaxed in a few special cases. We will discuss this in the proof. Under these assumptions,

$$\lim_{n\to\infty} nh_n\, Var\left[\hat{f}_n(y_0)\right] = A(y_0)\rho_k f(y_0)L(y_0). \tag{16}$$

and

$$Var\left[\hat{f}_n(y_0)\right] = A(y_0)\frac{\rho_k f(y_0)L(y_0)}{nh_n} + o\left(\frac{1}{nh_n}\right), \tag{17}$$

where

$$A(y_0) = \frac{2}{a(y_0)} - 1, \rho_k = \int k^2(t)dt. \tag{18}$$

Further, under the additional assumption

1. *$f(x)$ has bounded third derivative in $x \in [y_0 - \epsilon, y_0 + \epsilon]$,*

we get the asymptotic bias

$$\mathbb{E}\left[\hat{f}_n(y_0)\right] - f(y_0) = \frac{1}{2}h_n^2 \mu_K f''(y_0) + o(h_n^2), \tag{19}$$

as $n \to \infty$ and $h \to 0$, where $\mu_k = \int t^2 k(t)dt$.

Proof. We can write the variance of the estimator as

$$\mathrm{Var}\left[\hat{f}_n(y_0)\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[k_h(y_0 - X_i)w(X_i)] + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n-1} \mathrm{Cov}[k_h(y_0 - X_i)w(X_i), k_h(y_0 - X_j)w(X_j)].$$

Because the chain starts from stationarity,

$$\mathrm{Cov}[X_i, X_j] = \mathrm{Cov}[X_1, X_{j-i}], \quad \forall j > i$$

i.e. the double sum of covariances can be replaced by a sum over k of all lag $k-$ covariances. We can re-write the variance as

$$\mathrm{Var}\left[\hat{f}_n(y_0)\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(k_h(y_0 - X_i)w(X_i)) + \frac{2}{n^2}\sum_{j=1}^{n-1}(n-j)\mathrm{Cov}(k_h(y_0 - X_1)w(X_1), k_h(y_0 - X_{1+j})w(X_{1+j})). \tag{20}$$

First, we analyze the covariance term.

$$\begin{aligned}
&\mathrm{Cov}_G[k_h(y_0 - X_1)w(X_1), k_h(y_0 - X_{1+j})w(X_{1+j})] \\
&= \mathbb{E}_G[k_h(y_0 - X_1)k_h(y_0 - X_1)w(X_{1+j})w(X_{1+j})] - \mathbb{E}_G[k_h(y_0 - X_1)w(X_1)]\mathbb{E}_G[k_h(y_0 - X_{1+j})w(X_{1+j})] \\
&= \mathbb{E}_G[k_h(y_0 - X_1)k_h(y_0 - X_j)w(X_1)w(X_{1+j})\mathbb{I}(X_i \neq X_{1+j})] \\
&\quad + \mathbb{E}_G[k_h(y_0 - X_1)w(X_1)k_h(y_0 - X_{1+j})L(X_{1+j})\mathbb{I}(X_1 = X_{1+j})] \\
&\quad - \mathbb{E}_G[k_h(y_0 - X_1)w(X_1)]\mathbb{E}_G[k_h(y_0 - X_{1+j})L(X_{1+j})] \\
&= \int\int_{x \neq y} k_h(y_0 - x)k_h(y_0 - y)w(x)w(y)g(x)\tilde{p}_x^{(j)}(y)dxdy \\
&\quad + \mathbb{E}_G[k_h^2(y_0 - X_1)w^2(X_1)\mathbb{E}_G(\mathbb{I}(X_{1+j} = X_1)|X_1)] - \mathbb{E}_G[k_h(y_0 - X_1)w(X_1)]\mathbb{E}_G[k_h(y - X_{1+j})w(X_{1+j})] \\
&= \int\int \Big(k_h(y_0 - x)k_h(y_0 - y)w(x)w(y)g(x)\tilde{p}_x^{(j)}(y) - k_h(y_0 - x)k_h(y_0 - y)w(x)w(y)g(x)g(y)\Big)dxdy \\
&\quad + \mathbb{E}_G[k_h^2(y_0 - X_1)w^2(X_1)r(X_1)^{(j)}] \\
&= \int\int k_h(y_0 - x)k_h(y_0 - y)w(x)w(y)g(x)G_j(x,y)dxdy + \mathbb{E}_G[k_h^2(y_0 - X_1)w^2(X_1)r(X_1)^{(j)}] \\
&= \theta_j + \eta_j,
\end{aligned}$$

where

$$\theta_j = \int\int k_h(x)k_h(y)w(x)w(y)g(x)G_j(x,y)dxdy,$$

$$\eta_j = \mathbb{E}_G[k_h^2(X_1)w^2(X_1)r(X_1)^{(j)}],$$

and, from Assumption 1,

$$G_j(x,y) = \tilde{p}_x^{(j)}(y) - g(y) \leq V(x)R(j). \tag{21}$$

We, now analyze the two terms θ_j and η_j one by one. For θ_j,

$$\begin{aligned}
&\frac{2}{n^2}\sum_{j=1}^{n-1}(n-j)\int\int k_h(y_0-x)k_h(y_0-y)w(x)w(y)g(x)G_j(x,y)dxdy \\
&\leq \frac{2}{n^2}\int\int\sum_{j=1}^{n-1}(n-j)k_h(y_0-x)k_h(y_0-y)w(x)w(y)g(x)V(x)R(j)dxdy \\
&= \frac{2}{n}\int\int\sum_{j=1}^{n-1}k_h(y_0-x)k_h(y_0-y)w(x)w(y)g(x)V(x)R(j)dxdy \\
&\quad - \frac{2}{n^2}\int\int\sum_{j=1}^{n-1}jk_h(y_0-x)k_h(y_0-y)w(x)w(y)g(x)V(x)R(j)dxdy \\
&\leq \frac{2}{n}\int\int k_h(y_0-x)k_h(y_0-y)w(x)w(y)g(x)V(x)\sum_{j=1}^{n-1}R(j)dxdy \\
&\quad - \frac{2}{n^2}\int\int k_h(y_0-x)k_h(y_0-y)w(x)w(y)g(x)V(x)\sum_{j=1}^{n-1}jR(j)dxdy
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{n}\int\int k_h(y_0-x)k_h(y_0-y)w(x)w(y)g(x)V(x)Mdxdy && \text{Using Assumption 2} \\
&\quad + \frac{2}{n^2}\int\int k_h(y_0-x)k_h(y_0-y)w(x)w(y)g(x)V(x)\sum_{j=1}^{n-1}nR(j)dxdy && \text{(Second term is } \geq 0,\ j \leq n)
\end{aligned}$$

Since k is compactly supported, for a sufficiently small h, we can bring the support of k_h to an ϵ neighbourhood where Assumption 2 holds. So, since $V(x) < M$ for all x in an ϵ neighbourhood of y, after shrinking the support, we can take uniformly put a bound on $V(x)$ as the integral will be 0 outside $[y_0-\epsilon, y_0+\epsilon]$, and so the limits of the integral wil also change to the compact support. Note that if we do not use a compact kernel, we need the integral

$$\int k_h(y_0-x)w(x)g(x)V(x)dx$$

to be finite over the entire support of g. One way to ensure this is assuming the Markov chain to be uniformly ergodic so that $V(x)$ can uniformly bounded on the support. Here, we take a bound on $V(x)$ and proceed, however, if on assumes the above integral to be finite, the first term can be directly written as $O(n^{-1})$. Using the uniform bound on the weights, g and a bound on V

(Assumption 2 and Assumption 4),

$$
\begin{aligned}
&\frac{2}{n^2}\sum_{j=1}^{n-1}(n-j)\theta_j \\
&\leq \frac{2M^2M'^2}{n}\int\int k_h(y_0-x)k_h(y_0-y)dxdy \\
&\quad + \frac{2M^2M'^2}{n^2}\int\int k_h(y_0-x)k_h(y_0-y)\sum_{j=1}^{n-1} nR(j)dxdy \\
&= \frac{2M^2M'^2}{n} + \frac{2M^2M'^2}{n}\int\int k_h(y_0-x)k_h(y_0-y)\sum_{j=1}^{n-1} R(j)dxdy \qquad \text{(kernels integrate to 1)} \\
&\leq \frac{2M^2M'^2}{n} + \frac{2M^3M'^2}{n}\int\int k_h(y_0-x)k_h(y_0-y)dxdy \\
&= \frac{2M^2M'^2}{n} + \frac{2M^3M'^2}{n} \qquad \text{(kernels integrate to 1)} \\
&= O(n^{-1})
\end{aligned}
$$

For the second term η_j,

$$
\begin{aligned}
&\frac{2}{n^2}\sum_{j=1}^{n-1}(n-j)\mathbb{E}\left[k_h^2(y_0-X_1)w^2(X_1)r(X_1)^{(j)}\right] \\
&= \frac{2}{n^2}\mathbb{E}\left[k_h^2(y_0-X_1)w^2(X_1)\sum_{j=1}^{n-1}(n-j)r(X_1)^{(j)}\right] \\
&= \frac{2}{n^2}\mathbb{E}\left[k_h^2(y_0-X_1)w^2(X_1)\left(\frac{r^2(X_1)(1-r^{n-1}(X_1))}{a^2(X_1)} + \frac{(n-1)r(X_1)}{a(X_1)}\right)\right] \\
&= \frac{2}{n^2}\int k_h^2(y_0-s)w^2(s)\frac{r^2(s)(1-r^{n-1}(s))}{a^2(s)}g(s)ds \\
&\quad + \frac{2}{n^2}\int k_h^2(y_0-s)w^2(s)\frac{(n-1)r(s)}{a(s)}g(s)ds
\end{aligned}
$$

Here again, we can either use the compact kernel assumption to get a bound on a^{-1} after shrinking the support of the kernel, as a^{-1} is locally bounded (Assumption 2). Or we can assmue a^{-1} to be uniformly bounded over the support of g. This is a fairly reasonable assumption as this implies that a does not go to 0 at any point. This means that the Markov chain moves at some point in

time. Using the bound on a, and using the uniform bound on weights (Assumption 4),

$$
\begin{aligned}
&\leq \frac{2M'^2M^3}{n^2}\int k_h^2(y_0-s)(r^2(s)(1-r^{n-1}(s)))ds \\
&\quad + \frac{2}{n^2}\int k_h^2(y_0-s)w^2(s)\frac{nr(s)}{a(s)}g(s)ds - \frac{2}{n^2}\int k_h^2(y_0-s)w^2(s)\frac{r(s)}{a(s)}g(s)ds \\
&\leq \frac{2M'^2M^3(1-M^{-1})^2}{n^2h_n^2}\int k^2\left(\frac{y_0-s}{h_n}\right)ds \\
&\quad + \frac{2}{n}\int k_h^2(y_0-s)w^2(s)\frac{r(s)}{a(s)}g(s)ds \\
&\quad + \frac{2M'^2M^2(1-M^{-1})}{n^2h_n^2}\int k^2\left(\frac{y_0-s}{h_n}\right)ds \\
&= \frac{2M'^2M^3(1-M^{-1})^2}{n^2h_n}\int k^2(t)dt \\
&\quad + \frac{2}{n}\int k_h^2(y_0-s)w^2(s)\frac{r(s)}{a(s)}g(s)ds \\
&\quad + \frac{2M'^2M^2(1-M^{-1})}{n^2h_n}\int k^2(t)dt \\
&= \frac{2}{n}\int k_h^2(y_0-s)w^2(s)\frac{r(s)}{a(s)}g(s)ds + O\left(\frac{1}{n^2h_n}\right) \\
&= \frac{2}{nh_n^2}\int k^2\left(\frac{y_0-s}{h_n}\right)w^2(s)\frac{r(s)}{a(s)}g(s)ds + O\left(\frac{1}{n^2h_n}\right) \\
&= \frac{2}{nh_n}\int k^2(t)w^2(y_0-h_nt)\frac{r(y_0-h_nt)}{a(y_0-h_nt)}g(y_0-h_nt)ds + O\left(\frac{1}{n^2h_n}\right).
\end{aligned}
$$

Therefore, the covariance term

$$
\begin{aligned}
&\frac{2}{n^2}\sum_{j=1}^{n-1}\mathrm{Cov}[k_h(y_0-X_1)w(X_1), k_h(y_0-X_{1+j})w(X_{1+j})] \\
&= \frac{2}{nh_n}\int k^2(t)w^2(y_0-h_nt)\frac{r(y_0-h_nt)}{a(y_0-h_nt)}g(y_0-h_nt)ds + O\left(\frac{1}{n^2h_n}\right) \qquad (22)
\end{aligned}
$$

Taking the variance term,

$$
\begin{aligned}
\frac{1}{n^2}\sum_{i=0}^{n-1}\mathrm{Var}(k_h(y_0-X_i)w(X_i)) &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\int k_h^2(y_0-s)w^2(s)g(s)ds - \left(\int k_h(y_0-s)g(s)ds\right)^2\right) \\
&= \frac{1}{n^2h_n^2}\sum_{i=0}^{n-1}\int k^2\left(\frac{y_0-s}{h_n}\right)w^2(s)g(s)ds + O(n^{-1}) \\
&= \frac{1}{nh_n^2}\int k^2\left(\frac{y_0-s}{h_n}\right)w^2(s)g(s)ds + O(n^{-1}) \\
&= \frac{1}{nh_n}\int k^2(t)w^2(y_0-h_nt)g(y_0-h_nt)dt + O(n^{-1})
\end{aligned}
$$

We can write the second term as $O(n^{-1})$ as $(\int k_h(y_0 - s)g(s)ds))^2$ is bounded (Assumptions 2 and 5). Adding the variance and covariance terms, we get

$$\begin{aligned}
&\text{Var}\left[\hat{f}_n(y_0)\right] \\
&\qquad = \frac{2}{nh_n}\int k^2(t)w^2(y_0 - h_n t)\frac{r(y_0 - h_n t)}{a(y_0 - h_n t)}g(y_0 - h_n t)dt \\
&\qquad\quad + \frac{1}{nh_n}\int k^2(t)w^2(y_0 - h_n t)g(y_0 - h_n t)dt + O(n^{-1}) + O\left(\frac{1}{n^2 h_n}\right).
\end{aligned}$$

Using Lemma 1, for

$$\begin{aligned}
&\text{Var}\left[\hat{f}_n(y_0)\right] \\
&\qquad = \frac{2}{nh_n}w^2(y_0)\frac{r(y_0)}{a(y_0)}g(y_0)\int k^2(t)dt \\
&\qquad\quad + \frac{1}{nh_n}w^2(y_0)g(y_0)\int k^2(t)dt + O(n^{-1}) + O\left(\frac{1}{n^2 h_n}\right) \\
&\qquad = \frac{1}{nh_n}\frac{2w(y_0)f(y_0)r(y_0)\rho_k}{a(y_0)} + \frac{1}{nh_n}w(y_0)f(y_0)\rho_k + o\left(\frac{1}{nh_n}\right) \\
&\qquad = \frac{1}{nh_n}w(y_0)f(y_0)\rho_k\left(\frac{2r(y_0)}{a(y_0)} + 1\right) + o\left(\frac{1}{nh_n}\right) \\
&\qquad = \frac{1}{nh_n}w(y_0)f(y_0)\rho_k\left(\frac{2(1 - a(y_0))}{a(y_0)} + 1\right) + o\left(\frac{1}{nh_n}\right) \\
&\qquad = \frac{1}{nh_n}w(y_0)f(y_0)\rho_k\left(\frac{2}{a(y_0)} - 1\right) + o\left(\frac{1}{nh_n}\right) \\
&\qquad = \frac{1}{nh_n}w(y_0)f(y_0)\rho_k\left(\frac{2}{a(y_0)} - 1\right) + o\left(\frac{1}{nh_n}\right) \\
&\qquad = \frac{1}{nh_n}w(y_0)f(y_0)\rho_k A(y_0) + o\left(\frac{1}{nh_n}\right)
\end{aligned}$$

$$\begin{aligned}
&nh_n\text{Var}\left[\hat{f}_n(y_0)\right] \\
&\qquad = \frac{2}{nh_n}\int k^2(t)w^2(y_0 - h_n t)\frac{r(y_0 - h_n t)}{a(y_0 - h_n t)}g(y - h_n t)dt \\
&\qquad\quad + \frac{1}{nh_n}\int k^2(t)w^2(y_0 - h_n t)g(y_0 - h_n t)dt + O(h_n) + O(n^{-1}).
\end{aligned}$$

And so, since $h_n \to 0$ as $n \to \infty$,

$$\begin{aligned}
\lim_{n\to\infty} nh_n\text{Var}\left[\hat{f}_n(y_0)\right] &= w(y_0)f(y_0)\rho_k\left(\frac{2}{a(y_0)} - 1\right) \\
&= w(y_0)f(y_0)\rho_k A(y_0)
\end{aligned}$$

To see the bias of the estimator,

$$\begin{aligned}
&\mathbb{E}\left[\hat{f}_n(y_0)\right] - f(y_0) \\
&= \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{E}[k_h(y_0 - X_i)w(X_i)] - f(y_0) \\
&= \frac{1}{n}\sum_{i=0}^{n-1} \int k_h(y_0 - s)w(s)g(s)ds - f(y_0) \\
&= \int \frac{1}{h_n} k\left(\frac{y_0 - s}{h_n}\right) f(s)ds - f(y_0)
\end{aligned}$$

Substituting $(y-s)/h_n = t$, and using second order Taylor series expansion around t for the second term (we have assumed f to have a bounded third derivative around y),

$$\begin{aligned}
&\int k(t)f(y_0 - h_n t)dt - f(y_0) \\
&= \int k(t)f(y_0)dt - \int t h_n k(t) f'(y_0) + \frac{f''(y_0)}{2}\int t^2 k(t)dt + o(h_n^2) - f(y_0) \\
&= \frac{1}{2}h_n^2 f''(y_0)\mu_k + o(h_n^2).
\end{aligned}$$

Hence,

$$\mathbb{E}\left[\hat{f}_n(y_0)\right] - f(y_0) = \frac{1}{2}h_n^2 f''(y_0)\mu_k + o(h_n^2). \tag{23}$$

□

3.3 A self-normalized importance sampling kernel density estimator

We will now discuss a self normalized version of this IS kernel density estimator. That is, consider

$$\hat{f}^{\text{MH}}_{\text{sn-IS}} = \frac{\hat{f}_n}{\hat{W}_n}, \tag{24}$$

where

$$W_n = \frac{1}{n}\sum_{i=1}^{n} w(X_i).$$

Note that here, we are still taking $X_1, \dots X_n$ to be MH samples, following the same assumptions as discussed in Theorem 2. Note that this estimator is especially useful in kernel density estimation for Bayesian posteriors, because in case the target has an intractable normalizing constant, it gets cancelled in the numerator and denominator. We can establish the same theory as we did in Theorem 2 for this new estimator. Before that, we need some lemmas.

Lemma 2. *Suppose a Markov transition kernel P is $G-$ Harris ergodic, and the chain starts from stationarity, $||P^n(x,\cdot) - G(\cdot)|| \leq M(x)\psi(n)$, and $\mathbb{E}_G[M(x)] < \infty$. Let $w : \mathscr{X} \to \mathbb{R}$, $\bar{W}_n = \sum_{i=1}^n w(X_i)/n$, and suppose that atleast one of the following holds:*

1. $\sup_x |w(x)| < \infty$ *and* $\sum_{n=1}^{\infty} \psi(n) < \infty$, *or*
2. $\mathbb{E}_G|w(x)|^{2+\delta} < \infty$ *for some* $\delta > 0$ *and* $\sum_{n=1}^{\infty} \psi(n)^{\delta/(2+\delta)} < \infty$.

Then,

$$\sigma_g^2 := \lim_{n\to\infty} n \operatorname{Var}_G(\bar{W}_n) = \operatorname{Var}_G(w(X_1)) + 2\sum_{j=2}^{\infty} \operatorname{Cov}_G(w(X_1), w(X_j)) < \infty.$$

Lemma 3. *Let $X_1, \dots X_n$ be samples obtained from an MH algorithm, and $\bar{W}_n$ be as defined above. Here, $w(X_i) = f(X_i)/g(X_i)$. Then,*

$$\mathbb{E}_G[\bar{W}_n] = \mathbb{E}_G\left[\frac{1}{n}\sum_{i=1}^{n} w(X_i)\right] = 1 \tag{25}$$

Proof.

$$\begin{aligned}
\mathbb{E}_G\left[\frac{1}{n}\sum_{i=1}^{n} w(X_i)\right] &= \mathbb{E}_G[w(X_1)] && \text{The marginal of each sample is } g \\
&= \int w(s)g(s)ds \\
&= \int f(s)ds \\
&= 1.
\end{aligned}$$

□

Lemma 4. *Let $\bar{W}_n$ be as defined above. Let $X_1, \dots X_n$ be the samples defined above, and let Assumptions 1, 2, 4 hold. Then,*

$$Var_G[\bar{W}_n] = O(n^{-1}) \tag{26}$$

Proof. Note that since the samples $X_1, \dots X_n$ are from a Metropolis Hastings transition kernel, the chain is Harris-ergodic. We have assumed the chain to start from stationarity, i.e. $X_0 \sim G$. Also, Assumption 1 implies $||P^i(x,\cdot) - G(\cdot)|| \leq V(x)R(i)$. From Assumption 4, we have $\sup_{\mathscr{X}} |w(x)| = M' < \infty$, and from Assumption 1 we have $\sum_{i=1}^{\infty} R(i) \leq M < \infty$. So, using Lemma 2, we have

$$\lim_{n\to\infty} n \operatorname{Var}_G(\bar{W}_n) = \operatorname{Var}_G(w(X_1)) + 2\sum_{j=2}^{\infty} \operatorname{Cov}_G(w(X_1), w(X_j)) < \infty.$$

So, $\exists C_0 < \infty$ such that

$$\lim_{n\to\infty} \operatorname{Var}_G[\bar{W}_n] = C_0 < \infty$$

So, $\operatorname{Var}_G[\bar{W}_n] = C_0/n$ and hence $\operatorname{Var}_G[\bar{W}_n] = O(n^{-1})$. □

Theorem 2. *Fix $y \in \mathbb{R}$, and suppose there are functions $V : \mathbb{R} \to \mathbb{R}^+$, $R : \mathbb{R} \to (0,1)$ and constants $\epsilon > 0$ and $M < \infty$ such that uniformly for $x \in [y_0 - \epsilon, y_0 + \epsilon]$. Under the assumptions given in Theorem 2,*

$$Var\left[\hat{f}^{MH}_{sn\text{-}IS}(y_0)\right] = A(y_0)\frac{\rho_k f(y_0)w(y_0)}{nh_n} + o\left(\frac{1}{nh_n}\right), \tag{27}$$

where

$$A(y_0) = \frac{2}{a(y_0)} - 1, \rho_k = \int k^2(t)dt.$$

Further, under the additional assumption

1. *$f(x)$ has bounded third derivative in $x \in [y_0 - \epsilon, y_0 + \epsilon]$,*

we get the asymptotic bias

$$\mathbb{E}\left[\hat{f}^{MH}_{sn\text{-}IS}(y_0)\right] - f(y_0) = \frac{1}{2}h_n^2 \mu_K f''(y_0) + o(h_n^2), \tag{28}$$

as $n \to \infty$ and $h_n \to 0$, where $\mu_k = \int t^2 k(t)dt$. Again, some of these assumptions could be relaxed, and we will discuss the same in the proof of this.

Proof. We first define some notation. To study the bias and variance of the self normalized IS estimator (24) , we employ a Taylor series expansion. Define $p : \mathbb{R}^2 \to \mathbb{R}$, such that $p(a, b) = a/b$. Consequently,

$$\nabla p(a, b) = \left(\frac{1}{b} \ : \ -\frac{a}{b^2}\right)^T . \tag{29}$$

We can see that

$$\hat{f}^{\text{MH}}_{\text{sn-IS}} = \frac{\hat{f}_n(y)}{\bar{W}_n} = p(\hat{f}_n(y), \bar{W}_n).$$

Using a multivariate Taylor series expansion of $p(\hat{f}_n(y), \bar{W}_n)$ around $(f(y), 1)$,

$$\begin{aligned}\hat{f}^{\text{MH}}_{\text{sn-IS}}(y) = p(f(y), 1) &+ (\hat{f}_n(y) \ : \ \bar{W}_n)^\top \nabla p(f(y), 1) \\ &+ \frac{1}{2}(\hat{f}_n(y) - f(y) \ : \ \bar{W}_n - 1)^\top H(c)(\hat{f}_n(y) - f(y) \ : \ \bar{W}_n - 1)\end{aligned}$$

where $c \in [x, a]$ or $[a, x]$. We denote the last term by $R(x, a)$. i.e.

$$\hat{f}^{\text{MH}}_{\text{sn-IS}}(y) = p(f(y), 1) + (\hat{f}_n(y) \ : \ \bar{W}_n)^\top \nabla p(f(y), 1) + O(n^{-2})$$

Replacing the value ∇p from (29), we obtain

$$\hat{f}^{\text{MH}}_{\text{sn-IS}}(y) = \frac{f(y)}{1} + \frac{1}{1}(\hat{f}_n(y) - f(y)) - \frac{f(y)}{1}(\bar{W}_n - 1) + R(x, a).$$

$$\begin{aligned}\mathbb{E}\left[\hat{f}^{\text{MH}}_{\text{sn-IS}}(y)\right] &= f(y) + \mathbb{E}[\hat{f}_n(y)] - f(y) - (\mathbb{E}[\bar{W}_n] - 1) + O(n^{-2}) \\ &= \mathbb{E}[\hat{f}_n(y)] + O(n^{-2}) \\ &= f(y) + \frac{1}{2}h_n^2 \mu_k f''(y) + o(h_n^2) + O(n^{-1}) \qquad \text{From Lemma 2.}\end{aligned}$$

Now evaluating the variance,

$$\begin{aligned}\text{Var}\left[\hat{f}^{\text{MH}}_{\text{sn-IS}}(y)\right] &= \text{Var}\left[\hat{f}_n(y)\right] - f^2(y)\text{Var}[\bar{W}_n] \\ &= A(y)\frac{\rho_k f(y) L(y)}{nh_n} + o\left(\frac{1}{nh_n}\right) - f^2(y)O(n^{-1}) \\ &= A(y)\frac{\rho_k f(y) L(y)}{nh_n} + o\left(\frac{1}{nh_n}\right)\end{aligned}$$

At the point where we are evaluating the density, y , f is finite. Hence,

$$\text{Var}\left[\hat{f}^{\text{MH}}_{\text{sn-IS}}(y)\right] = A(y)\frac{\rho_k f(y) L(y)}{nh_n} + o\left(\frac{1}{nh_n}\right) + O(n^{-1}).$$

□

3.4 Estimating the marginal density of a multivariate distribution

So far, we have dealt with univariate densities. Now, assume f and g to be multivariate densities, with the support of f as $X \subseteq \mathbb{R}^d$. Let $X_1, \dots X_n \in \mathbb{R}^d$ be samples obtained from a Metropolis Hastings algorithm, with the kernel satisfying Assumptions 1, 2, 3. Let X_{ij} denote the j^{th} component of the i^{th} sample, and say our goal is to evaluate the marginal density of the j^{th} component at a point $y \in \mathbb{R}$. We can use the estimator

$$\hat{f}_j(y) = \frac{\sum_{i=1}^n w(X_i) k_h(y - X_{ij})}{\sum_{i=1}^n w(X_i)}. \tag{30}$$

Here, k is a univariate kernel satisfying the same properties discussed above. Similar variance and bias results can be established for this estimator, but we might need additional assumptions on the functions and kernel. In the example presented below, we use this estimator to estimate marginal densities in a Poisson random effects model.

4 Example

4.1 Bayesian Poisson Random Effects Model

We will use the multivariate version of the estimator (30) to estimate a few marginal densities of the posterior of the Bayesian Poisson random effects model in Livingstone and Zanella (2020). The samples and weights have been taken from the implementation given in Shukla et al. (2025). The following hierarchical model is given for count data, where y_{ij} denotes the observed count for the j^{th} subject in the i^{th} class, and n_i is the number of subjects in class i. The model is defined as:

$$\begin{aligned} y_{ij} \mid \eta_i &\sim \text{Poisson}(e^{\eta_i}), \quad j = 1, 2, \dots, n_i, \\ \eta_i \mid \mu &\sim \mathcal{N}(\mu, \sigma_\eta^2), \quad i = 1, 2, \dots, I, \\ \mu &\sim \mathcal{N}(0, c^2). \end{aligned}$$

In line with the setup of Livingstone and Zanella (2020), we set:

$$I = 50, \quad \sigma_\eta = 3, \quad c = 10, \quad \lambda = 0.001.$$

Using the code given in Shukla et al. (2025), we get the samples from the importance density π^λ using Barker's proposal, where π^λ is the Moreau-Yosida envelope. We also get samples from the true target using Barker's proposal, and the importance weights. Using these, we compute the marginal kernel density estimators using the estimator discussed in (30). Note that here, we cannot use the unnormalized estimator. This is because the target f has an intractable normalizing constant. Using (30) will enable the normalizing constants in the numerator and denominator to cancel out. The marginal density plots for μ, η_5 and η_{20} are given here, and compared with the `density` function in `R`.For kernel density estimation, Sköld, Martin and Roberts, G O (2003) gives the optimal bandwidth for the estimator (13). However, these require the double derivative of the target, and the acceptance ratio of the MH algorithm. They also give the relation of the optimal bandwidth to the i.i.d. optimal bandwidth as follows:

$$h_{\text{MH}} = A^{1/5} h_{\text{i.i.d.}}.$$

where is as given in (18). In the implementation, however the bandwidth is fixed to be 0.05.

We can see that the marginal density plots differ quite a bit for components of η. This is primarily because the samples from the posterior are 'bad' samples.

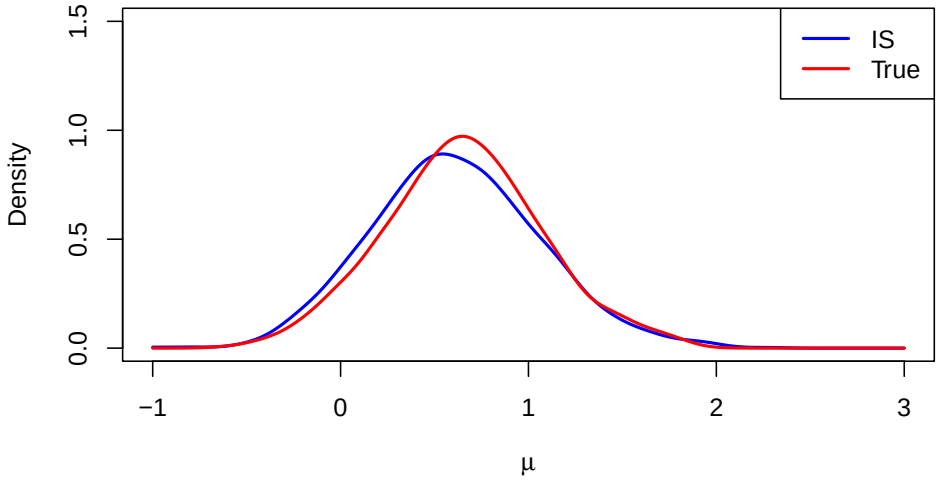
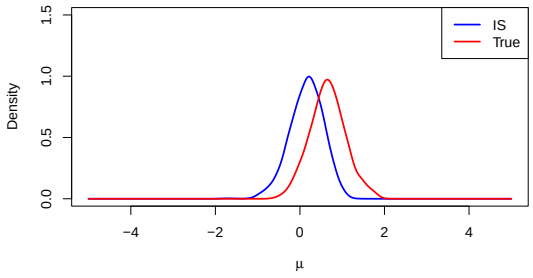
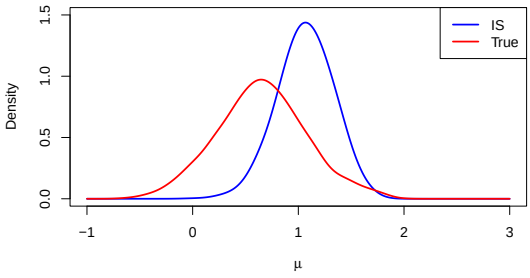


Figure 1: Plot of the marginal density of μ



(a) Plot of the marginal density of η_5



(b) Plot of the marginal density of η_{20}

Figure 2: Density plots of 2 marginals of η

5 Discussion

We see that the rejection step in the MH algorithm contributes to the asymptotic variance of the estimator. A self-normalized IS-KDE provides a practically useful estimator, especially in Bayesian scenarios involving intractable normalizing constants. We saw that the bias and variance of the self-normalized estimator remains asymptotically controlled, with the normalization only introducing a higher-order correction. The next step is to establish theory for the marginal density estimator presented, and also develop Markov chain CLT results for the estimators, similar to what we have in the i.i.d. case.

References

Hall, P., Lahiri, S., and Truong, Y. (1995). On bandwidth choice for density estimation with dependent data. the annals of statistics, 23(6), 2241-2263. *Ann. Statist.*, 23.

Hart, J. (1996). Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics - J NONPARAMETR STAT*, 6:115–142.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109.

Livingstone, S. and Zanella, G. (2020). The barker proposal: combining robustness and efficiency in gradient-based mcmc.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Nakayama, M. K. (2011). Asymptotic properties of kernel density estimators when applying importance sampling. In *Proceedings of the 2011 Winter Simulation Conference (WSC)*, pages 556–568. IEEE.

Shukla, A., Vats, D., and Chi, E. C. (2025). Mcmc importance sampling via moreau-yosida envelopes.

Sköld, Martin and Roberts, G O (2003). Density estimation for the Metropolis-Hastings algorithm. *Scandinavian Journal of Statistics*, 30(4):699–718.