An Alternative Method for Estimating and Simulating Maximum Entropy Densities

Jae-Young Kim
and
Joonhwan Lee
Seoul National University
May 25, 2008

Abstract

This paper proposes a method of estimating and simulating a maximum entropy distribution given moment conditions based on Monte-Carlo approach. The method provides an simple alternative to conventional calculation methods of maximum entropy densities which involve complex numerical integration and are subject to occasional failure. We first show that maximum entropy density on a random sample converges to minimum cross entropy solution to the sample density. Using this result, we show that the empirical maximum entropy density on a uniform sample supported within sufficiently large boundary converges to the true maximum entropy density. The performance of the proposed method is checked by several examples.
1 Introduction

A maximum entropy density is obtained by maximizing Shannon’s entropy measure. Jayens (1957) describes it to be “uniquely determined as the one which is maximally noncommittal with regard to missing information, and that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters.” Maximum entropy densities provide a convenient and flexible tool for approximating probability distributions known only to satisfy certain moment conditions. It is known that with various forms of moment conditions, we can derive many known probability distributions. For example, with first two arithmetic moments, the maximum entropy distribution is the normal distribution.

The maximum entropy (maxent) approach has been used widely in many literatures. In Bayesian econometrics, Bayesian method of moments (BMOM) employed the maxent approach to estimate the posterior density of parameters (Zellner, 1997). In the finance literature, there have been many applications of the maxent approach. Stutzer (1996) and Buchen and Kelly (1996) used the maxent approach to derive risk neutral measures (or state price densities) of asset returns to price derivatives. Rockinger and Jondeau (2002) and Bera and Park (2004) extended ARCH type models by applying maxent densities.

One big problem with calculating a maxent density is that the analytic derivation is impossible in general. Therefore, one should resort to numerical methods to calculate maxent densities. However, early studies noted that there are some difficulties with numerical procedures, which typically involve complex numerical integration and iterative nonlinear optimization (Zellner and Highfield, 1988; Ormoneit and White, 1999). Particularly, numerical optimization methods employed by them are highly sensitive to starting values because integrals need to be re-evaluated in each iteration. Wu (2003) partially resolved the difficulties by proposing a new algorithm to setting up appropriate starting value with sequential updating of moment conditions. However, it still has difficulty when calculating maxent densities given extreme moment conditions.

\footnote{Bera and Park (2004) provides a comprehensive table on characterizations of maximum entropy density.}
In this paper, we propose an alternative method of calculating maxent density. The method approximates maxent density by using discrete probability measure, empirical measure from uniform distribution for example. Discrete maxent problem does not involve numerical integration and standard numerical optimization procedures work well. In Section 2, we show that the cross-entropy minimizing distribution on empirical measure converges to the true cross-entropy minimizing distribution. In Section 3 and 4, the result directly leads to the approximation of maxent densities by maxent on uniform random sample or discrete uniform measure. In Section 5, the performance of the method is checked by numerical experiments.

2 Cross-Entropy Minimization on Random Sample

The maxent density is obtained by maximizing Shannon’s entropy measure, which is the Kullback-Leibler distance (or the cross-entropy) defined relative to uniform measure with the negative sign,

\[ W(p) = \int -p(x) \log p(x) dx \quad KL(p, q) = \int p(x) \log \frac{p(x)}{q(x)} dx \]

subject to some known moment conditions defined by,

\[ \int c_j(x)p(x)dx = d_j, \quad j = 1, \ldots, m \]

It is clear that the maxent problem is a special case of the cross-entropy minimization.

\[ p(x) = \arg \min_{p(x)} \int p(x) \log \frac{p(x)}{q(x)} dx \quad \text{s.t} \]

\[ \int p(x)dx = 1, \quad \int c_j(x)p(x)dx = d_j, \quad j = 1, \ldots, m \]

It is known that the problem can be simplified to an unconstrained optimization problem. (Golan et al., 1996 or Buchen and Kelly, 1996)

**Proposition 1.** The problem (1) is equivalent to:

\[ \min_{\lambda} \int \exp\left\{ \sum_{j=1}^{m} \lambda_j(c_j(x) - d_j) \right\} q(x) dx \quad (2) \]

\[ \text{It should be noted that for the remaining part, we assume that } \int \exp(\lambda'c(x))q(x)dx \text{ exists for } \lambda \in V, \text{ an open set} \]
and the solution of the problem is:

\[ p(x) = \frac{1}{\mu} \exp\{\lambda'c(x)\}q(x), \quad \text{where } \mu = \int \exp\{\lambda'c(x)\}q(x)dx \]

**Proof.** The Lagrangian function of the problem is

\[ \mathcal{L}(p) = \int p(x) \log \frac{p(x)}{q(x)}dx + (1 + \lambda_0)(1 - \int p(x)dx) + \sum_{j=1}^{m} \lambda_j(d_j - \int c_j(x)p(x)dx) \]

\(^3\) \(\mathcal{L}(p)\) is maximized when its Frechet derivative equals zero and we have the first order condition

\[ \delta \mathcal{L} = \int [1 + \log \frac{p(x)}{q(x)} - (1 + \lambda_0) - \sum_{j=1}^{m} \lambda_jc_j(x)]\delta p(x)dx = 0 \]

This leads immediately to

\[ \log p(x) = \log q(x) + \lambda_0 + \sum_{j=1}^{m} \lambda_jc_j(x) \]

Letting \(\mu \equiv \exp -\lambda_0\) as defined above, we have

\[ p(x) = \frac{1}{\mu} \exp(\sum_{j=1}^{m} \lambda_jc_j(x))q(x) \]

The objective function now can be written in unconstrained form

\[ \int p(x) \log \frac{p(x)}{q(x)}dx = -\log \mu + \sum_{j=1}^{m} \lambda_jd_j \]

Therefore, the problem is equivalent to minimizing \(\int \exp\{\lambda'c(x) - d\}q(x)dx\) with respect to \(\lambda\)'s. Note that the Hessian matrix of the object function is the covariance matrix of moment conditions. That is,

\[ \frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \int \exp\{\lambda'c(x) - d\}q(x)dx = \int c_i(x)c_j(x)q(x)dx \]

Therefore, the object function is convex and the solution is unique. \(\Box\)

\(^3\)The \(\lambda\)'s have opposite signs to that of Wu (2003). This is due to different setup of Lagrangian function.
Now we consider the sample analogue of the problem (1). Let $X_1, X_2, \ldots, X_n$ be a i.i.d random sample and $X_1$ has the probability density $q(x)$. Then it is straightforward to construct the cross-entropy minimization on the empirical measure given the random sample.

$$p = \arg\min_p \sum_{i=1}^n p_i \log \frac{p_i}{1/n} \quad \text{s.t}$$

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n c_j(X_i)p_i = d_j, \quad j = 1, \ldots, m$$

Here $p = (p_1, \ldots, p_n)$, a vector of probabilities assigned on a given random sample. For this sample analogue version, we have a similar result to Proposition 1.

**Proposition 2.** The problem (3) is equivalent to:

$$\min_\lambda \sum_{i=1}^n \exp\left\{ \sum_{j=1}^m \lambda_j (c_j(X_i) - d_j) \right\} \frac{1}{n}$$

and the solution of the problem is:

$$p = \frac{1}{\hat{\mu}} \exp\{\lambda c(X_i)\} \frac{1}{n}, \quad \text{where} \quad \hat{\mu} = \sum_{i=1}^n \exp\{\lambda c(X_i)\} \frac{1}{n}$$

**Proof.** The proof is a special case of that of Proposition 1. \hfill $\blacksquare$

Let $\hat{P}_n$ be the probability measure induced by the empirical cross-entropy minimization with sample size $n$. That is,

$$\hat{P}_n(t) = \sum_{i=1}^n p_i \mathbf{1}_{t \geq X_i}(t)$$

It is natural to consider the relation between $\hat{P}_n(t)$ and $P(t)$, the true cross-entropy minimizing probability measure defined by

$$P(t) = \int_{-\infty}^t \frac{1}{\mu} \exp\{\lambda'c(x)\}q(x)dx$$

We first show that Lagrange multiplier $\lambda$ is consistently estimated by the empirical cross-entropy minimization.
Theorem 3. Let $\lambda^*$ be the unique solution of the problem (2) and $\hat{\lambda}_n$ be that of the problem (4). Then we have

$$
\hat{\lambda}_n \xrightarrow{p} \lambda^*
$$

Proof. Let $\rho(X_i, \lambda) \equiv \exp\{\lambda'(c(X_i) - d)\}$. The object function of (2) is $E_Q\rho(X, \lambda)$, where $Q$ is the probability measure corresponding to the density $q(x)$. Also, the object function of (4) is $\bar{\rho}_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \lambda)$. By the weak law of large numbers, we have

$$
\bar{\rho}_n(\lambda) \xrightarrow{p} E_Q\rho(X, \lambda)
$$

Then the standard argument used to prove the consistency of a MCE (Minimum Contrast Estimator) can be directly applied and that establishes the desired result. (See Van der Vaart, 1998 for example)

We now establish the uniform convergence of $\hat{P}_n(t)$ to $P(t)$, which is similar to the Glivenko-Cantelli theorem.

Theorem 4.

$$
\lim_{n \to \infty} Q_n \left( \sup_{t \in \mathbb{R}} |\hat{P}_n(t) - P(t)| > \epsilon \right) = 0
$$

Proof. Let $\mu_n = \frac{1}{n} \sum \exp\{\hat{\lambda}'_n c(X_i)\}$. Then the triangle inequality, the continuous mapping theorem and WLLN yields

$$
\hat{\mu}_n \xrightarrow{p} \mu = \int \exp\{\lambda'^* c(x)\} q(x) dx
$$

$$
\hat{P}_n(t) = \sum_{i=1}^{n} p_i 1_{(t \geq X_i)}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\hat{\mu}_n} \exp\{\hat{\lambda}'_n c(X_i)\} 1_{(t \geq X_i)}(t)
$$

By the Slutsky lemma and WLLN, we have

$$
\hat{P}_n(t) \xrightarrow{p} P(t) \quad \forall t \in \mathbb{R}
$$

Since $\hat{P}_n(t)$ and $P(t)$ are both increasing and bounded, the uniform convergence can be proved using the same argument for the Glivenko-Cantelli theorem.
3 The Monte-Carlo Method of Calculating Maxent Density

This section proposes an alternative method of calculating and simulating from a maxent density based on the result of the previous section. Conventional methods (Zellner and Highfield, 1988; Ormoneit and White, 1999; Wu, 2003 for example) involve complex numerical integrations, which lead to difficulties associated with obtaining the numerical solution. These methods employ standard Newton-Rhapson algorithm which uses explicit gradient values. However, in most of the search region, Jacobian matrices are near-singular and the search algorithm often fails unless the starting value is set close to the solution. (Ormoneit and White, 1999) Once we get an estimated maxent density, we need to run another algorithm to draw random numbers from the maxent density.

The method we propose here uses Monte-Carlo draws to estimate maxent densities. We first draw random numbers from uniform distribution supported within appropriate bound. Then we calculate the empirical cross-entropy minimization solution with given moment conditions. The numerical minimization problem is solved by Nelder-Mead simplex search algorithm, a direct search algorithm which does not involve explicit or implicit gradient values. With the calculated discrete probabilities, we can directly draw a random number from the maxent density by weighted resampling on the original uniform draws.

**Proposition 5.** The maxent density \( p(x) \) can be approximated with an arbitrary precision by the empirical cross-entropy minimization on random draws from \( U(-M, M) \) with \( M \) sufficiently large.

**Proof.** Let \( \hat{\lambda}_M \) be the solution to

\[
\min_{\lambda} \int \exp\left\{ \sum_{j=1}^{m} \lambda_j (c_j(x) - d_j) \right\} q_M(x) dx,
\]

where \( q_M(x) = \frac{1}{2M} 1_{(-M,M)}(x) \). As \( M \to \infty \), converges to the maxent solution

\[
\lambda^* = \argmax_{\lambda} \int -\exp\left\{ \sum_{j=1}^{m} \lambda_j (c_j(x) - d_j) \right\} dx
\]

by MCT and Theorem 3. Also, by Theorem 3,

\[
\hat{\lambda}_{n,M} \xrightarrow{p} \hat{\lambda}_M
\]
Therefore, as \( n \to \infty \) and \( M \to \infty \), \( \hat{\lambda}_{n,M} \overset{p}{\to} \lambda^* \). The proposition follows from the continuous mapping theorem.

In practice, with normalization to zero-mean and unit-variance, \( M \) of 15 or less is a reasonable choice for most of cases. For the numerical minimization, we employ the Nelder-Mead simplex search algorithm, which is widely used for unconstrained minimization problem. The algorithm may seem inefficient at the first glance, because it does not use any gradient value, which is easily available for the maxent problem. However, Jacobian matrices on most of the search region are near-singular and therefore Newton-like algorithms often fail even for small number of moment conditions. The Nelder-Mead algorithm first sets a simplex of \( n + 1 \) vertices for \( n \)-dimensional problem around the initial values. Then the algorithm transforms the simplex along the surface by reflection, expansion, contraction or shrinkage for each iteration step. Only several additional function evaluations are needed for an iteration step and therefore computational load is relatively small. The algorithm continues until the diameter of the simplex become smaller than specified threshold. For the detailed description of the algorithm and convergence results, see Lagarias et al., 1997. Although the computation is inefficient compared to Newton-like algorithms, the direct search algorithm successfully obtains numerical solutions for the maxent problem in cases that conventional algorithms fail.

**Algorithm 1** (Nelder-Mead Direct Search Algorithm). *First, we need to specify four parameters, namely reflection, expansion, contraction and shrinkage parameters denoted by \( \rho, \chi, \gamma \) and \( \sigma \). Set an arbitrary simplex with \( n + 1 \) vertices in \( \mathbb{R}^n \) and continue the following iteration of transforming simplex until the diameter of the simplex become smaller than specified value.*

1. **Order the \( n + 1 \) vertices to satisfy** \( f(x_1) \leq f(x_2) \leq \cdots \leq f(x_{n+1}) \).

2. **Reflection**: Calculate \( x_r = (1 + \rho)\bar{x} - \rho x_{n+1} \), where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \). If \( f(x_1) \leq f(x_r) \leq f(x_n) \), then accept the point \( x_r \) to replace \( x_{n+1} \) and terminate the iteration; otherwise, go to step 3 or 4.

3. **Expansion**: If \( f(x_r) < f(x_1) \), calculate \( x_e = (1 + \rho\chi)\bar{x} - \rho\chi x_{n+1} \) and evaluate \( f(x_e) \). If \( f(x_e) < f(x_r) \), then accept \( x_e \) and terminate the iteration; otherwise, accept \( x_r \) and terminate.

4. **Contraction**: If \( f(x_r) \geq f_n \), perform either outside or inside contraction.
(a) Outside Contraction : If $f(x_r) < f(x_{n+1})$, calculate $x_c = (1 + \rho \gamma)\bar{x} - \rho \gamma x_{n+1}$ and evaluate $f(x_c)$. If $f(x_c) \leq f(x_r)$, accept $x_c$ and terminate the iteration; otherwise go to step 5.

(b) Inside Contraction : If $f(x_r) \geq f(x_{n+1})$, calculate $x_{cc} = (1 - \gamma)\bar{x} + \gamma x_{n+1}$ and evaluate $f(x_{cc})$. If $f(x_{cc}) < f(x_{n+1})$, accept $x_{cc}$ and terminate the iteration; otherwise go to step 5.

5. Shrinkage : Set $n$ new points $v_i = x_1 + \sigma(x_i - x_1)$, $i = 2, \ldots, n + 1$. Replace $x_2, \ldots, x_{n+1}$ by $v_2, \ldots, v_{n+1}$ and terminate the iteration.

The precision of the method is somewhat disappointing like other Monte-Carlo methods. With very large random draws (500,000 or more), the estimation of $\lambda$ is accurate to at least 4 decimal places. However, the Monte-Carlo method is relatively simple and stable so that there is less concern of failure. It is later shown by numerical experiments that the proposed method is able to obtain estimated maxent densities given large number of moment conditions or moment conditions that is close to legitimate bounds.

Moreover, through the maxent probabilities on the uniform random draws, $p_i = \frac{1}{n} \exp\{\hat{\lambda}_{n,M}c(X_i)\} \frac{1}{n}$, $i = 1, 2, \ldots, n$, we can generate a random number that follows the estimated maxent density. First, we can immediately obtain the empirical maxent measure $\hat{Q}_{n,M}(t) = \sum_{i=1}^n p_i 1(t \geq X_i)(t)$. By the Theorem 4, $\hat{Q}(t)$ converges uniformly to the true maxent measure. Thus, with $n$ and $M$ sufficiently large, Monte-Carlo simulation of the maxent density can be approximated by a random draw from $\hat{Q}_{n,M}$. Note that we can perform the Monte-Carlo simulation of the maxent density directly unlike the conventional methods. The following algorithm can be used to draw a random number from $\hat{Q}_{n,M}$

**Algorithm 2.** We can draw a random number from the maxent density by the following algorithm.

1. Draw $V \sim U(0, 1)$.

2. Let $X^{(1)}, \ldots, X^{(n)}$ be the order statistic of the random sample.

3. Let

$$k = \inf \{ k \mid \sum_{i=1}^p p_i \geq V, k \in \mathbb{N} \}$$

3. Let $Y = X^{(k)}$. Then $Y \sim p(x)$
The weighted empirical measure $\hat{P}_n$ can be used to evaluate any functional of the maxent density. For example, quantiles, trimmed mean and so forth. Stutzer(1996)’s canonical valuation can be viewed as a calculation of trimmed(truncated) mean from the empirical maxent measure.

4 The Performance and Numerical Experiments

In this section, we provide some performance checks for the method proposed in the previous section. First, we follow the numerical experiment done by Ormoneit and White(1999) and Wu(2003). Similar to them, the moment conditions\(^4\) imposed are

\[
\begin{align*}
EX &= 0 \quad & EX^2 &= 1 \\
EX^3 &\equiv \mu_3 \in [0, 3] \quad & EX^4 &\equiv \mu_4 \in [(EX^3)^3 + 1, 1.1, 10]
\end{align*}
\]

Figure 1 and 2 shows the estimated $\lambda_i$, $i = 1, 2, 3, 4$, plotted against the value of $\mu_3$ and $\mu_4$. We can find that the patterns of them closely resemble those reported by Ormoneit and White (1999) and Wu (2003). Also, Figure 3 shows some examples of calculated maxent densities with some extreme moment conditions that caused failure with the algorithm proposed by Omoneit and White (1999). For example, they reported that their algorithm failed when $\mu_3 = 0$ and $\mu_4 > 3$. Also, they encountered numerical error when $\mu_4 > 10$.

The discretized maxent method is able to get solutions with extreme moment conditions, such as $\mu_3 = 5, \mu_4 = 26.1$ or $\mu_3 = 0, \mu_4 = 12$. The plotted densities are obtained by the weighted kernel smoothing where the weights are given as $p_i$’s.

\[
\hat{p}(x) = \sum_{i=1}^{n} p_i \frac{1}{h} K\left( \frac{x - X_i}{h} \right)
\]

The performance when there are large number of moment conditions, is also checked. We imposed ten moment conditions which are slightly modified from those of the standard normal distribution. Figure 4 shows various maxent densities with ten moment conditions imposed and compare them with the standard normal density. $M1$ imposes $2k$th moments larger than

\(^4\)It is noted in previous studies that $\mu_3^2 + 1 < \mu_4$ must hold to ensure the covariance matrix to be positive definite. (Wu, 2003)
the standard normal distribution, while $M_2$ imposes those smaller than the standard normal. We also imposed empirical (standardized) moments of KOSPI 200 daily returns. The proposed method combined with sequential updating of initial values resulted in fast calculation and stable numerical procedures.

5 Conclusion

The maximum entropy approach provides a rich set of probability distributions given some moment conditions. Large classes of distributions used frequently are special cases of maxent density. In this paper, we first discuss the sample analogue of the cross-entropy minimization problem, which is a generalization of the maxent problem. Various convergence results are obtained in a straightforward manner. The results can be directly applied to the calculation and simulation of maxent density.

Traditional approaches of calculating maxent density involved complex numerical integration procedure and are subject to occasional failure. The numerical optimization is highly sensitive to the starting value when there is numerous moment conditions. The paper proposes an attractive alternative to the calculation of maxent density. The method is justified by the convergence of empirical maxent measure to the true maxent measure. This Monte-Carlo based method is simple and stable; a few lines of codes is enough to implement the method. Moreover, the result can be extended to discretized uniform density, which can be viewed as a refinement to the Monte-Carlo method. The refinement yielded significant improvement in precision of calculation. Several numerical experiments which were used to test traditional methods, are applied to the proposed method. The result suggests that the proposed method is an attractive alternative to conventional methods for calculating and simulating maxent densities.

The 2$^{\text{nd}}$ moment of standard normal distribution is $\frac{(2k)!}{2^k k!}$.

The standardized moments of KOSPI 200 daily returns are $\mu_3 = -0.34$, $\mu_4 = 6.17$, $\mu_5 = -14.42$, $\mu_6 = 139.47$, $\mu_7 = -733.67$, $\mu_8 = 5763.1$, $\mu_9 = -37815$, $\mu_{10} = 279256$. 

11
Figure 1: $\hat{\lambda}_1$ (top) and $\hat{\lambda}_2$ (bottom) for given $\mu_3$ and $\mu_4$, $(n = 10,000, \ M = 8)$
Figure 2: $\hat{\lambda}_3$ (top) and $\hat{\lambda}_4$ (bottom) for given $\mu_3$ and $\mu_4$, ($n = 10,000$, $M = 8$)
Figure 3: Estimated Maxent Density with Extreme Moment Conditions ($n = 30000$, $M = 8$)
Figure 4: Various Maxent Densities with Ten Moment Conditions ($n = 30000, M = 8$)