Comparison of Maximum Entropy and Higher-Order Entropy Estimators

Amos Golan* and **Jeffrey M. Perloff****

ABSTRACT

We show that the generalized maximum entropy (GME) is the only estimation method that is consistent with a set of five axioms. The GME estimator can be nested using a single parameter, α, into two more general classes of estimators: GME-α estimators. Each of these GME-α estimators violates one of the five basic axioms. However, small-sample simulations demonstrate that certain members of these GME-α classes of estimators may outperform the GME estimation rule.

Keywords: *generalized entropy, maximum entropy, axioms*

* American University
** University of California, Berkeley

Correspondence: Amos Golan, Dept. of Economics, Roper 200, American University, 4400 Mass. Ave. NW, Washington, DC 20016-8029; E-mail: agolan@american.edu; Fax: (202) 885-3790; Tel: (202) 885-3783.

Comparison of Maximum Entropy and Higher-Order Entropy Estimators

1. Introduction

The generalized maximum entropy (GME) estimation method (Golan, Judge, and Miller, 1996) has been widely used for linear and nonlinear estimation models. We derive two more general estimation methods by replacing the GME's entropy objective with higher-order entropy indexes. We then show that the GME is the only estimation technique that is consistent with five axioms (desirable properties of an estimator) and that each of the two new estimators violates one of these axioms. Nevertheless, linear model sampling experiments demonstrate that these new estimators may perform better than the GME for small or ill-behaved samples.

The GME estimator is based on the classic maximum entropy (ME) approach of Jaynes (1957a, 1957b), which uses Shannon's (1948) entropy-information measure to recover the unknown probability distribution of underdetermined problems. Shannon's entropy measure reflects the uncertainty (state of knowledge) that we have about the occurrence of a collection of events. To recover the unknown probabilities that characterize a given data set, Jaynes proposed maximizing the entropy, subject to the available sample-moment information and the requirement of proper probabilities. The GME approach generalizes the maximum entropy problem by taking into account individual noisy observations (rather than just the moments) while keeping the objective of minimizing the underlying distributional, or likelihood, assumptions.[1]

---

[1] The GME is a member of the class of information-theoretic estimators (empirical and generalized empirical likelihood, GMM when observations are Gaussian and BMOM).

We nest the GME estimator into two more general classes of estimators. To derive each of these new estimators, we replace the Shannon entropy measure with either the Renyi (1970) or Tsallis (1988) generalized entropy measures. Each of these entropy measures is indexed by a single parameter α (the Shannon measure is a special case or both of these measures where $\alpha = 1$). We call our generalized entropy estimators GME-α estimators.

In Section 2, we compare the Shannon, Renyi, and Tsallis entropy measures. We start Section 3 by briefly summarizing the GME approach for the linear model. Then we derive the two GME-α estimation methods. In Section 4, we show that GME can be derived from five axioms or properties that we would like an estimator to possess. Unfortunately, each of the GME-α estimators violates one of these axioms. Nonetheless, one might want to use these GME-α estimators for small samples as they outperform the GME in some sampling experiments presented in Section 5. We summarize our results in Section 6.

2. Properties of Entropy Measures

After formally defining the discrete versions of the three entropy measures, we note that they share many properties but are distinguished by their additivity properties. For a random vector x with K discrete values x_k, each with a probability $p_k = P(x_k)$ and $p = \{p_1,...,p_K)$ where p is a proper distribution, the Shannon entropy measure is

---

These estimators avoid using an explicit likelihood (e.g., Owen, 1991; Qin and Lawless, 1994; Kitamura and Stutzer, 1997; Imbens et al., 1998; Zellner, 1997).

$$H(\boldsymbol{x}) = -\sum_k p_k \log p_k, \qquad (2.1)$$

where $0 \log 0$ is defined as 0.

The two more general families of information measures are indexed by a single parameter α, which we restrict to be strictly positive: $\alpha > 0$. The Renyi (1970) entropy measure is

$$H_\alpha^R(\boldsymbol{x}) = \frac{1}{1-\alpha} \log \sum_k p_k^\alpha. \qquad (2.2)$$

The Tsallis (1988) measure is

$$H_\alpha^T(\boldsymbol{x}) = c \frac{\sum_k p_k^\alpha - 1}{1-\alpha}, \qquad (2.3)$$

where the value of c, a positive constant, depends on the particular units used. For simplicity, we set $c = 1$. (The Renyi functional form resembles the CES production function and the Tsallis function form is similar to the Box-Cox.) Both of these more general families include the Shannon measure as a special case: as $\alpha \to 1$,

$H_\alpha^R(\boldsymbol{x}) = H_\alpha^T(\boldsymbol{x}) = H(\boldsymbol{x})$.

With Shannon's entropy measure, events with high or low probability do not contribute much to the index's value. With the generalized entropy measures for $\alpha > 1$, higher probability events contribute more to the value than do lower probability events. Unlike the Shannon's measure (2.1), the average logarithm is replaced by an average of powers α. Thus, a change in α changes the relative contribution of event k to the total sum. The larger the α, the more weight the "larger" probabilities receive in the sum. For a detailed discussion on entropy and information see Retzer and Soofi (this volume).

These entropy measures have been compared in Renyi (1970), Tsallis (1988), Curado

and Tsallis (1991), and Holste et al. (1998).[2]   The Shannon, Renyi, and Tsallis measures

share three properties.   First, all three entropy measures are nonnegative for any arbitrary

p. These measures are strictly positive except when all probabilities but one equal zero

(perfect certainty). Second, these indexes reach a maximum value when all probabilities

are equal. Third, each measure is concave for arbitrary *p*. In addition, the two

generalized entropy measures share the property that they are monotonically decreasing

functions of α for any *p*.

The three entropy measures differ in terms of their additivity properties. Shannon

entropy of a composite event equals the sum of the marginal and conditional entropies:

$$H(x,y)=H(y)+H(x|y)=H(x)+H(y|x), \tag{2.4}$$

where *x* and *y* be two discrete and finite distributions.[3]   However, this property does not

hold for the other two measures (see Renyi, 1970). If *x* and *y* are independent, then Eq.

(2.4) reduces to

---

[2] Holste et al. (1998) show that H_α^R and $H_\alpha^T(x)$ are related:
$H_\alpha^R(x) = (1/1-\alpha)\log[1+(1-\alpha)\log H_\alpha^T]$.

[3] Let *x* and *y* be two discrete and finite distributions with possible realizations
$x_1, x_2,..., x_K$ and $y_1, y_2,..., y_J$ respectively. Let *p(x,y)* be a joint probability distribution.
Now, define $P(\mathbf{x} = x_k) = p_k$, $P(\mathbf{y} = y_j) = q_j$, $P(\mathbf{x} = x_k, \mathbf{y} = y_j) = w_{kj}$,
$P(\mathbf{x}|\mathbf{y}) = P(\mathbf{x} = x_k | \mathbf{y} = y_j) = p_{k|j}$ and $P(\mathbf{y}|\mathbf{x}) = P(\mathbf{y} = y_j | \mathbf{x} = x_k) = q_{j|k}$ where
$p_k = \sum_{j=1}^{J} w_{kj}$, $q_j = \sum_{k=1}^{K} w_{kj}$ and the conditional probabilities satisfy $w_{kj} = q_j p_{k|j} = p_k q_{j|k}$.
The conditional information (entropy) *H(x|y)* is the total information in *x* with the
condition that *y* has a certain value:
$$H(\mathbf{x}|\mathbf{y}) = \sum_j q_j \left[-\sum_k p_{k|j} \log p_{k|j}\right] = \sum_j q_j \left[-\sum_k \left(\frac{w_{kj}}{q_j}\right) \log \left(\frac{w_{kj}}{q_j}\right)\right] = \sum_{k,j} w_{kj} \log \left(\frac{q_j}{w_{kj}}\right).$$

$$H(x,y)=H(x)+H(y),\qquad\qquad(2.5)$$

which is the property of standard additivity, that holds for the Shannon and Renyi entropy

measures, but not for the Tsallis measure.[4]

Finally, only the Shannon and Tsallis measures have the property of Shannon

additivity. The total amount of information in the entire sample is a weighted average of

the information in two mutually exclusive subsamples, A and B. Let the probabilities for

subsample A be $\{p_1,...,p_L\}$ and those for B be $\{p_{L+1},...,p_K\}$, and define $p_A = \sum_{k=1}^{L} p_k$ and

$p_B = \sum_{k=L+1}^{K} p_k$. Then, for all α, (*including* $\alpha = 1$),

$$H_\alpha^T(p_1,...,p_K) = H_\alpha^T(p_A,p_B) + p_A^\alpha H_\alpha^T(p_1/p_A,...,p_L/p_A)$$
$$+ p_B^\alpha H_\alpha^T(p_{L+1}/p_B,...,p_K/p_B).$$

3. GME-α Estimators

We can derive the GME-α estimators using an approach similar to that used to derive the

classic maximum entropy (ME) and the GME estimators. The GME and GME-α

coefficient estimates converge in the limit as the sample size grows without bound.

However, these estimators produce different results for small samples. As the chief

application of maximum entropy estimation approaches is to extract information from

limited or ill-conditioned data, we concentrate on such cases.

---

[4] For two *independent* subsets A and B, H_α^T is "pseudo-additive" and satisfies
$H_\alpha^T(A,B) = H_\alpha^T(A) + H_\alpha^T(B) + (1-\alpha)H_\alpha^T(A)H_\alpha^T(B)$ for all α where
$H_\alpha^T(A,B) \equiv H_\alpha^T(x,y) = \left(\sum_{k,j} w_{kj}^\varepsilon - 1\right)/(1-\alpha)$.

We consider the classical linear model:

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\tag{3.1}$$

where $y \in \mathfrak{R}^N$, X is an $N \times K$ known linear operator, $\boldsymbol{\beta} \in \mathfrak{R}^K$ and $\varepsilon \in \mathfrak{R}^N$ is a noise vector. Our objective is to estimate the vector $\boldsymbol{\beta}$ from the noisy observations y. If the observation matrix X is irregular or ill-conditioned, if $K \gg N$, or if the underlying distribution is unknown, the problem is ill-posed. In such cases, one has to (i) incorporate some prior (distributional) knowledge, or constraints, on the solution, or (ii) specify a certain criterion to choose among the infinitely many solutions, or (iii) do both.

3.1. GME Estimation Method

We briefly summarize how we would obtain a GME estimate of Eq. (3.1). Instead of making an explicit likelihood or distributional assumptions we view the errors, $\boldsymbol{\varepsilon}$, in this equation as another set of unknown parameters to be estimated simultaneously with the coefficients, $\boldsymbol{\beta}$. Rather than estimate the unknowns directly, we estimate the probability distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ within bounded support spaces.

Let z_k be an M-dimensional vector $z_k = (z_{k1},...,z_{kM})'$ for all k. Let p_k be an M-dimensional proper probability distribution for each covariate k defined on the set z_k such that

$$\beta_k \equiv \sum_m p_{km} z_{km} \quad \text{and} \quad \sum_m p_{km} = 1.\tag{3.2}$$

Similarly, we redefine each error term as

$$\varepsilon_i \equiv \sum_j w_{ij} v_j \quad \text{and} \quad \sum_j w_{ij} = 1.\tag{3.3}$$

The support space for the coefficients is often determined by economic theory.

For example, according to an economic theory, the propensity to consume out of income is an element of (0, 1), therefore we would specify the support space to be $v_k = (0, 0.5, 1)'$ for $M = 3$. Lacking such theoretical knowledge, we usually assume the support space is symmetric around zero with large range.

Golan, Judge, and Miller (1996) recommend using the "three-sigma rule" of Pukelsheim (1994) to establish bounds on the error components: the lower bound is $\underline{v} = -3\sigma_y$ and the upper bound is $\bar{v} = -3\sigma_y$, where σ_y is the (empirical) standard deviation of the sample y. For example if $J = 5$, then $v = (-3\sigma_y, -1.5\sigma_y, 0, 1.5\sigma_y, 3\sigma_y)'$.

Imposing bounds on the support spaces in this manner is equivalent to making the following convexity assumptions:

Convexity Assumption C1. $\beta \in B$ *where B is a bounded convex set.*

Convexity Assumption C2. $\varepsilon \in V$ *where V is a bounded convex set that is symmetric around zero.*

Having reparameterized β and ε, we rewrite the linear model as

$$y_i = \sum_{k=1}^{K} x_{ik}\beta_k + \varepsilon_i = \sum_{k=1}^{K}\sum_{m=1}^{M} x_{ik} z_{km} p_{km} + \sum_j v_j w_{ij}, \quad i=1, \ldots, N \tag{3.4}$$

We obtain the GME estimator by maximizing the joint entropies of the distributions of the coefficients and the error terms subject to the data and the requirement for proper probabilities:

$$\text{GME} = \begin{cases} (\widetilde{\mathbf{p}}, \widetilde{\mathbf{w}}) = \underset{\mathbf{p},\mathbf{w}}{\arg\max} \left\{ -\sum_k \sum_m p_{km} \log p_{km} - \sum_i \sum_j w_{ij} \log w_{ij} \right\} \\ s.t. \\ y_i = \sum_{k,m} x_{ik} z_{km} p_{km} + \sum_j v_j w_{ij} \; ; \; \sum_m p_{km} = 1; \; \sum_j w_{ij} = 1 \end{cases} \tag{3.5}$$

Jaynes's traditional maximum entropy (ME) estimator is a special case of the GME where the last term in the objective – the entropy of the error terms – is dropped and (pure) moment conditions replace the individual observation restrictions (i.e., $y = X\boldsymbol{\beta} = Xp$ and $\boldsymbol{\varepsilon} = \boldsymbol{0}$).

The estimates (solutions to this maximization problem) are

$$\widetilde{p}_{km} = \frac{\exp(-\sum\limits_{i=1}^{N}\widetilde{\lambda}_i z_{km} x_{ik})}{\sum\limits_{m}\exp(-\sum\limits_{i-1}^{N}\widetilde{\lambda}_i z_{km} x_{ik})} \equiv \frac{\exp(-\sum\limits_{i}\widetilde{\lambda}_i z_{km} x_{ik})}{\Omega_k(\widetilde{\lambda})} \tag{3.6}$$

and

$$\tilde{w}_{ij} = \frac{\exp(-\tilde{\lambda}_i v_j)}{\sum\limits_{j}\exp(-\tilde{\lambda}_i v_j)} \equiv \frac{\exp(-\tilde{\lambda}_i v_j)}{\Psi_i(\tilde{\lambda})}, \tag{3.7}$$

The point estimates are $\widetilde{\beta}_k \equiv \sum_m z_{km} \widetilde{p}_{km}$ and $\widetilde{\varepsilon}_i \equiv \sum_j v_j \tilde{w}_{ij}$. Golan, Judge, and Miller (1996) and Golan (2001) provide detailed comparisons of the GME method with other regularization methods and information-theoretic methods such as the GMM and BMOM.

3.2. GME-α Estimator

We can extend the ME or GME approach by replacing the Shannon entropy measure with the Renyi or Tsallis entropy measure.

We now extend the GME approach[5]. For notational simplicity, we define H_α^e, $e=$

---

[5] In the appendix, we derive the generalization of the ME estimator.

R (Renyi) or T (Tsallis), to be the entropy of order α. The order of the signal and the noise entropies, α and α', may be different. We define $H_\alpha^e(\beta_k)$ and $H_{\alpha'}^e(\varepsilon_i)$ for each coordinate $k = 1, 2, \ldots, K$ and each observation $i = 1, 2, \ldots, N$ respectively. Then, the GME-α model is

$$
\text{GME-}\alpha = \begin{cases}
\underset{p,w}{\text{Max}} \quad \{H_\alpha^e(\beta) + H_{\alpha'}^e(\varepsilon)\} = \left\{ \sum_k H_\alpha^e(\beta_k) + \sum_i H_{\alpha'}^e(\varepsilon_i) \right\} \\
\text{s.t.} \\
y = XZp + Vw, \ p \geq 0, \ w \geq 0, \\
\sum_m p_{km} = 1, \ \sum_j p_{ij} = 1 \text{ for all } k \text{ and } i
\end{cases}
\tag{3.8}
$$

where Z is a matrix representation of the K M-dimensional support vectors z_k, and V is a matrix of the N J-dimensional support space v.

For example if $\alpha = \alpha'$, the Tsallis GME-α Lagrangean is

$$
L = \sum_{k,m} \left(\frac{1}{1-\alpha} \right) \left(p_{km}^\alpha - 1 \right) + \sum_{i,j} \left(\frac{1}{1-\alpha} \right) \left(w_{ij}^\alpha - 1 \right)
$$

$$
+ \sum_i \lambda_i \left(y_i - \sum_{k,m} x_{ik} z_{km} p_{km} - \sum_j v_j w_{ij} \right) + \sum_k \rho_k \left(1 - \sum_m p_{km} \right)
\tag{3.9}
$$

$$
+ \sum_i \phi_i \left(1 - \sum_j w_{ij} \right) + \sum_{k,m} \theta_{km} (-p_{km}) + \sum_{i,j} \vartheta_{ijm} (-w_{ij})
$$

and the first and second order conditions are

$$
\frac{\partial L}{\partial p_{km}} = \frac{1}{1-\alpha} p_{km}^{\alpha-1} - \sum_i \lambda_i x_{ik} z_{km} - \rho_k - \theta_{km} \leq 0
\tag{3.10}
$$

$$\frac{\partial^2 L}{\partial p_{km}^2} = -\alpha p_{km}^{\alpha-2} < 0 \tag{3.11}$$

$$\frac{\partial^2 L}{\partial p_{km} \partial p_{jm}} = 0 = \frac{\partial^2 L}{\partial p_{jm} \partial p_{km}} \tag{3.12}$$

$$\frac{\partial L}{\partial w_{ij}} = \frac{1}{1-\alpha} w_{ij}^{\alpha-1} - \lambda_i v_j - \phi_i - \vartheta_{ij} \leq 0 \tag{3.13}$$

$$\frac{\partial^2 L}{\partial w_{ij}^2} = -\alpha w_{ij}^{\alpha-2} < 0 \tag{3.14}$$

$$\frac{\partial^2 L}{\partial w_{ij} \partial w_{tj}} = 0 = \frac{\partial^2 L}{\partial w_{tj} \partial w_{ij}} \tag{3.15}$$

Given Eqs. (3.11), (3.12), (3.14), and (3.15), the Hessian has negative values along the diagonal and zero values off the diagonal, so an interior solution would be unique, if one exists[6].

---

[6] We reformulated this problem making α endogenous. We normalized the α-entropy measures so that they were elements of [0, 1], thereby making the entropy measures comparable across various values of α. Unfortunately, two problems arose. First, any normalization involves constructing a new objective that is also a function of α, so that some of the entropy properties discussed in Section 2 do not hold. Second, the normalized objective functions had multiple local maximum points under plausible conditions. Thus, we henceforth treat α as exogenous.

4. Axiomatic Derivation

We next show that the GME approach is consistent with our convexity assumptions C1 and C2 and five additional axioms. Then we demonstrate that each of our more general estimation models violates one of these axioms.

Our axiomatic approach is an extension of the axiomatic approaches to the Classical ME by Shore and Johnson (1980), Skilling (1988, 1989), and Csiszar (1991). We start by characterizing the properties (axioms) that we want our method of inference to possess. Then, we determine which estimation approaches possess those properties.

4.1. Axioms

Our five axioms represent a minimum set of requirements for a logically consistent method of inference from a finite data set. Following Skilling, we start by defining a distribution $f(x)$ as a positive, additive distribution function (PADF). It is positive by construction: $f(x_i) = p_i \geq 0$ for each realization x_i, $i = 1, 2, ..., N$ and strictly positive for at least one x_i. It is additive in the sense that the probability in some well-defined domain (e.g., B and V) is the *sum* of all the probabilities in *any* decomposition of this domain into sub-domains. A PADF lack the property of proper distribution that is always normalized: $\sum_i p_i = 1$.[7] The inference question can be viewed as a search for those PADFs that best characterize the finite data set. Working with PADFs allows us to avoid the complexity of dealing with normalizations, which simplifies our analysis.

---

[7] We can work with improper probability distributions, *p*,* that sum up to some number $s<1$, by normalizing so that $p_i = \left(p_i^* / \sum_i p_i^*\right) = \left(p_i^* / s\right)$.

Following Shore and Johnson and Skilling, we want to identify an estimate that is the *best* according to some criterion. We want a transitive means of ranking estimates so that we can determine which estimate maximizes (or minimizes) a certain function. We use the following axioms to determine the exact form of that function, while requiring that that function be independent of the data. Let $f(I, q)$, or similarly $\hat{\beta}[I, q]$, be the estimates provided by maximizing some function H with respect to the available data I (y, X) given some prior model q.

The five axioms are:

A1. "Best" Posterior: Completeness, Transitivity, and Uniqueness. *All posteriors can be ranked, the rankings are transitive, and, for any given prior and data set, the "best" posterior (the one that maximizes H) is unique.*

A2. Permutation or Coordinate Invariance. *Let H be any unknown criterion and f(I, q) is the estimate that obtained by optimizing H given the information set I (data) and prior model q. For Δ, a coordinate transformation, Δf(I, q) = f(ΔI, Δq).* (This axiom states that if we solve a given problem in two different coordinate systems, both sets of estimates are related by the same coordinate transformation.)

A3. Scaling. *If no additional information is available, the posterior should equal the prior.*[8]

---

[8] Following Skilling, we use this axiom for convenience only. It guarantees that the posterior's units are equivalent (rather than proportional) to those of the priors. If we use proper probability distributions instead of PADFs, this axiom is not necessary, but the resulting proof is slightly more complicated.

A4. Subset Independence. *Let I_1 be a constraint on $f(x)$ in the domain $x \in B_1$.*

Let I_2 be another constraint in a different domain $x \in B_2$. Then, we require that

our estimation (inference) method yield

$$f\left[B_1|I_1\right] \cup f\left[B_2|I_2\right] = f\left[B_1 \cup B_2 | I_1 \cup I_2\right],$$

where $f\left[B|I\right]$ is the chosen PADF in the domain B, given the information I.

(Our estimation rule produces the same results whether we use the subsets

separately or their union. That is, the information contained in one subset of the

data, or a specific data set, should not affect the estimates based on another subset

if these two subsets are independent.)

A5. System Independence. *The same estimate should result from optimizing*

independent information (data) of independent systems separately using their

different densities or together using their joint density.

4.2. Theorems

The following theorem holds for the GME method of inference (and hence for the

classical ME, which is a special case of the GME).

Theorem 1. *For the linear model (3.1) satisfying C1 and C2 with a finite and limited*

data set, the set of $N \times K$ PADFs in (3.5) that satisfy (A1-A5), that are defined on the

convex sets B and V, and that result from an optimization procedure (with respect to the

N observed data points) contains only the GME.

Proof of Theorem 1. Given convexity assumptions C1 and C2, each coefficient and error point estimate can be represented as an expected value of the support B and V for some $N \times K$ PADFs. To simplify our notation, we present the results for the discrete case in terms of a $K \times M$ matrix P and a $NJ \times 1$ vector $\mathbf{w}$ where $\mathbf{p}_k$ is an M-dimensional PADF and $\mathbf{w}_i$ is a J-dimensional PADF. Similarly, we define the prior models as P^0 and $\mathbf{w}^0$ with dimensions equal to P and $\mathbf{w}$. For simplicity, we assume that both sets of priors are uniform (within their support spaces) so that we can ignore them for the rest of the proof. Given A2 and A4, we choose the PADFs P and $\mathbf{w}$ by maximizing over the pair $\{P, \mathbf{w}\}$ some function (the "sum rule") of the form

$$H*(P,\mathbf{w}) = \sum_{k,m} g_{km}(p_{km}) + \sum_{i,j} h_{ij}(w_{ij}) \tag{4.1}$$

for some unknown functions $g_{km}(p_{km})$ and $h_{ij}(w_{ij})$. By imposing these axioms, we eliminate all cross terms between the different domains. This result yields the required independence between $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$.

Next, we impose the system independence (A5) and scaling (A3) axioms to obtain

$$H*(P,\mathbf{w}) = \sum g(\mathbf{p}_k) + \sum h(\mathbf{w}_i) = -\sum_{k,m} p_{km} \log p_{km} - \sum_{i,j} w_{ij} \log w_{ij} = H(P,\mathbf{w}), \tag{4.2}$$

which is the sum of the joint entropies of the signal and noise defined over B x V. This equation is of the same functional form as the objective function in the GME. Moreover, the axioms can lead to no other function. We can complete the proof by showing that (4.2) satisfies A1-A5, which we do by applying Theorem IV of Shore and Johnson (1980) within the support spaces Z and V (or use assumptions C1-C2).

Theorem 2. *For the linear model (3.1) satisfying* C1-C2 *with a finite and limited data,*

the set of $N \times K$ PADFs resulting from optimizing (with respect to the N observed data points) the Renyi GME-α on the convex set $B \times V$, satisfies axioms A1- A3 *and* A5.

Proof of Theorem 2. Transitivity and uniqueness (A1) follow directly from taking the second derivative of the Lagrangean for the Renyi GME-α [which is analogous to Eq. (3.9) for the Tsallis GME-α] and the first-order conditions. Axiom A2 holds trivially. We know that Axiom A5 holds from the additivity property discussed in Section 2, which is a necessary and sufficient condition for any function *H* to satisfy the system independence requirement. Finally, imposing A3 completes the proof.

Theorem 3. *For the linear model (3.1) satisfying* C1 *and* C2 *with a finite and limited data, the set of $N \times K$ PADFs resulting from optimizing (with respect to the N observed data points) the Tsallis-GME-α on the convex set $B \times V$, satisfies axioms* A1-A3 *and* A4.

Proof of Theorem 3. Transitivity and uniqueness (A1) follow directly from (3.11) and (3.12). Axiom A2 holds immediately. Axiom A4 follows from the property of Shannon additivity (see Section 2), where we use the relevant weights. Finally, imposing A3 completes the proof.[9]

4.4. Discussion

Given Theorem 1, if one wishes to choose a post-data distribution (PADF) for each coordinate *K* and *N*, that satisfies C1-C2 and A1-A5,[10] the appropriate rule is the GME.

---

[9] We know that H_α^T does not obey system independence, because it violates the additivity property, as we discussed in Section 2.

[10] We note here that Csiszar uses a more relaxed version of the *subset* and *system independent* axioms. For lack of space we do not provide here a full comparison of the three sets of axioms developed by Shore and Johnson (1980), Skilling (1988, 1989), and Csiszar (1991). This discussion is available upon request from the authors.

Equivalently, the GME is the appropriate method of assigning probability distributions within the set $B \times V$ given the available data and our axioms. If one is willing to give up either axiom A4 or A5, the Renyi or Tsallis GME-α estimator may be used respectively. However, one might want to use a GME-α estimator, rather than the GME rule, with a small and ill-conditioned data set due to the GME-α's faster rate of shrinkage.[11] In the following section we provide several examples.

As a closing remark, we note that an important extension of this work would be to develop an axiomatic framework that covers a larger class of information-related estimation rules, such as the (generalized) empirical likelihood, relevant GMM methods, and the BMOM (e.g., Owen, 1991; Qin and Lawless, 1994; Imbens, Johnson, and Spady, 1998; Kitamura and Stutzer, 1997; Zellner, 1997). By doing so, one could show how these estimation approaches differ in terms of desirable (axiomatic) properties, as we compared the GME and the GME-α methods.

5. Sampling Experiments

The following sampling experiments illustrate that the mean squared error (MSE) may be lower for a GME-α estimator than for the GME ($\alpha = 1$). Our objective is to provide examples showing that the GME-α may outperform the GME rather than to provide a full comparison of these estimation techniques relative to the GME and other information-theoretic methods.

---

[11] Because they are shrinkage estimators, all GME and GME-α estimators may be bias in small samples. In future work, it may be possible to correct for the bias in a small sample using a method similar to that in Holste et al. (1998).

We conduct four sets of experiments based on the linear model, Eq. (3.1). The first set uses a well-posed orthonormal experimental design with four covariates. The second set replaces the orthonormal covariates with four covariates, each of which was generated from a standard normal. The third set adds outliers to our second experimental design, as described in Ferretti et al. (1999). The fourth set of experiments uses an ill-conditioned design matrix with condition number of 90. For each of these designs, we vary the number of observations.

In all experiments and for all rules, we use the empirical three-standard-deviations rule (for each sample) to determine the errors' supports and with $J = 3$. For the first three sets of experiments we use the same support space, $z = (-100, 0, 100)'$, for each coefficient.

5.1. Example 1: Well-Conditioned, Orthonormal Design Matrix

We consider the orthonormal design (condition number of one), three different numbers of right-hand side variables ($K = 2$, 4, and 5) and two sample sizes ($N = 10$ and 30), and with 250 replications. Because we impose $\boldsymbol{\beta} = \mathbf{0}$, the linear model is $y_i = \sum_k \beta_k x_{ik} + \varepsilon_i = \varepsilon_i$. Figure 1 graphs the mean squared error (MSE) against α for a sample size of 30 and $K=4$. Here, because the model fits the data so well, the bias is practically zero, so the total variance equals the MSE. Both GME-α models have lower MSEs than does the GME ($\alpha=1$) for some values of α greater than 1. The Renyi rule dominates both the Tsallis and the GME rules for all examined values of α greater than 1. To save space, we do not report the figures for the other cases as they are qualitatively the same.

5.2. Example 2: Well-Conditioned Normal Design Matrix

The second experiment is a variant of the first one, where the orthonormal exogenous

variables are now generated from a N(0,1), $K = 4$, and $N = 10$ (Figure 2a) or 40 (Figure

2b). Again, for some values of $\alpha > 1$, both GME-α estimators dominate the GME.

Again, the bias is nearly zero, so that the total variance is virtually identical to the MSE.

The Renyi GME-α dominates the Tsallis when $N = 40$, however, the Tsallis GME-α

dominates the Renyi when $N = 10$.

5.3. Example 3: Outliers

We now add outliers to our linear model. We use the influential-outlier experimental

design of Ferretti et al. (1999), which uses a linear model (3.1) without an intercept. As

before, we consider three different values of $K (= 2, 4,$ and 5) and two sample sizes

(N=10 and 40). Here we used 100 replications. Each $\mathbf{x}_k$ is drawn from a N(0, 1) and

$$y_i = \sum_k \beta_k x_{ik} + \varepsilon_i = \varepsilon_i \text{ for } N(1\text{-}\delta) \text{ observations and } y_i = 6 \text{ and } x_{i1} = 10 \text{ for } N\delta$$

observations, separately for each $\delta = 0.5, 0.1, 0.2$ and 0.3. Thus, each sample has a

proportion δ of influential outliers, each with a value equal to six standard deviations

from the mean.

Because the results are qualitatively the same for all the parameter values we

examined, we illustrate our results with only two sets of results. Figure 3a presents the

MSEs and variances for $\alpha \in (0, 6.5)$ when $N = 40$, $K = 4$, and $\delta = 0.2$. Unlike in the

previous examples, the bias in this example is large due to the effects of the outliers.

Here, the Tsallis GME-α has virtually the same MSE for all values of α as does the GME,

though the GME-α slightly outperforms for very low and very high values of α. The

Renyi GME-α slightly dominates the GME for values of $\alpha < 1$, but is much worse for larger values of α.

Figure 3b reports the results of the experiment where $K = 4$, $N = 10$, $\delta = 0.3$ but instead of having $x_{i1} = 10$ for the outliers, we generate x_{i1} from a standard normal, so all exogenous variables are generated from a N(0,1). Both GME-α estimators have lower MSEs for some values of $\alpha > 1$. The Tsallis GME-α dominates the GME and the Renyi GME-α for all $\alpha > 1$. We also show the variance in this figure because there is a measurable bias.

5.4. *Example 4: Ill-Conditioned Design Matrix*

Finally, we use an ill-conditioned design matrix experiment from Golan, Judge and Miller (1996) where $N = 10$, $K = 4$, $\delta = 0$, condition number is 90. Here, we use a tighter support space for each k: $z = (-10, 0, 10)'$. The results are summarized in Figure 4. Both GME-α models have lower MSE than does the GME for some values of α greater than 1. The Renyi rule dominates both the Tsallis and the GME rules for all examined values of α greater than 1. Because there are no outliers, the bias is practically zero for both rules over the entire range of α. Similar experiments with more observations yielded the same results and are not reported here.

6. Summary and Conclusions

Renyi (1970) and Tsallis (1988) independently generalized the Shannon entropy measure. Each of their indexes uses a single parameter, α, to nest many entropy measures, and each includes the Shannon measure as a special case when $\alpha = 1$. We showed that each of these GME-α entropy measures can be used as an objective in an estimation procedure in

much the same way as Golan, Judge, and Miller (1996) used the Shannon entropy measure to formulate the GME estimation method.

We then demonstrated that the GME estimation approach is the only one that is consistent with a set of five basic axioms: completeness*, transitivity, and uniqueness*, *permutation* or *coordinate invariance*, *scaling, subset independence*, and *system independence*. We showed that the Renyi GME-α models is consistent with all of the axioms except *subset independence*, and the Tsallis GME-α is consistent with all except *system independence*. Thus, to employ either of the GME-α estimators, one must be willing to give up one axiom.

We then noted that one might be interested in using the GME-α estimator despite the loss of a desirable property when dealing with an ill-posed, or a small-sample problem. We illustrated that the GME-α has lower mean squared error than does the GME for some values of α in a set of experiments involving small samples, possibly influential outliers and ill-conditioned data matrix.

Appendix: ME-α Estimator

We extend the classical ME approach (Jaynes, 1957a, 1957b; Levine, 1980) using the

Renyi and Tsallis general entropy measures to obtain ME-α estimators. Let $y = Xp$,

where p is a K-dimensional proper probability distribution. For $K >> N$, the number of

observations, the ME-α, is

$$ME - \alpha = \begin{cases} \tilde{p}^e = \arg\max\{H_\alpha^e\} \\ \\ \text{s.t. } y = Xp, \sum_k p_k = 1 \text{ and } p_k \geq 0, \end{cases} \tag{A1}$$

where $e = R$ (Renyi) or T (Tsallis).

For example, the ME-α estimator based on the H_α^R measure is

$$ME - H_\alpha = \begin{cases} \tilde{p}^R = \arg\max\left\{\dfrac{1}{1-\alpha}\log\sum_k p_k^\alpha\right\} \\ \\ \text{s.t. } y = Xp, \sum_k p_k = 1 \text{ and } p_k \geq 0 \end{cases} \tag{A2}$$

(We now omit the superscript R for notational simplicity.) The Lagrangean is

$$L = H_\alpha + \sum_m \lambda_m(y_m - \sum_k p_k x_{mk}) + \mu(1 - \sum_k p_k) + \sum_k \theta_k(-p_k). \tag{A3}$$

The optimal conditions are

$$\frac{\partial L}{\partial p_k} = \frac{\alpha}{1-\alpha}\frac{p_k^{\alpha-1}}{\sum_k p_k^\alpha} - \sum_m \lambda_m x_{mk} - \mu - \theta_k \leq 0. \tag{A4}$$

$$\sum_k p_k \frac{\partial L}{\partial p_k} = 0 = \sum_k p_k \left[\left(\frac{1}{1-\alpha} \right) \frac{\alpha p_k^{\alpha-1}}{\sum_k p_k^{\alpha}} - \sum_m \lambda_m x_{mk} - \mu - \theta_k \right] = 0. \qquad (A5)$$

Solving for μ, we find that $\mu = \dfrac{\alpha}{1-\alpha} - \sum_m \lambda_m y_m$. We assume that Eq. A5 holds with

equality, substitute for μ, and rearrange the equation to obtain

$$b_k^{*\alpha-1} \equiv \frac{p_k^{*\alpha-1}}{\sum_k p_k^{*\alpha}} = \left(\frac{1-\alpha}{\alpha} \right) \left[\sum_m \lambda_m^* x_{mk} + \frac{\alpha}{1-\alpha} - \sum_m \lambda_m^* y_m + \theta_k^* \right]. \qquad (A6)$$

If $\alpha = 2$, the exact solution is

$$p_k^* = \frac{b_k^*}{\Omega} \equiv \frac{b_k^*}{\sum_k b_k^*} = \left(\frac{\sum_m \lambda_m^*(y_m - x_{mk}) - \theta_k^* + 2}{\sum_k \left[\sum_m \lambda_m^*(y_m - x_{mk}) - \theta_k^* + 2 \right]} \right). \qquad (A7)$$

References

Curado, E. M. F. and C. Tsallis (1991), "Generalized Statistical Mechanics: Connection with Thermodynamics," *Journal Physics A: Math. Gen,* 24, L69-L72.

Csiszar, I. (1991), "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems," *The Annals of Statistics*, 19, 2032-2066.

Ferretti, N., D. Kelmansky, V. J. Yohai and R. H. Zamar (1999), A Class of Locally and Globally Robust Regression Estimates, *J. of the American Stat. Assoc*, 94, 174-188.

Golan, A., (2001), "A Simultaneous Estimation and Variable Selection Rule," *Journal of Econometrics*, 101, 165-193.

Golan, A., and G. Judge and D. Miller (1996), *Maximum Entropy Econometrics: Robust Estimation With Limited Data*, John Wiley & Sons, New York.

Holste, D., I Grobe and H. Herzel (1998), "Bayes' Estimators of Generalized Entropies," *J. Phys. A: Math. Gen.*, 31, 2551-2566.

Imbens, G.W., Johnson, P. and R.H. Spady, "Information-Theoretic Approaches to Inference in Moment Condition Models," Econometrica 66 (1998), 333-357.

Jaynes, E.T. (1957a), "Information Theory and Statistical Mechanics," *Physics Review*, 106, 620-630.

Jaynes, E.T. (1957b), "Information Theory and Statistical Mechanics II," *Physics Review*, 108, 171-190.

Kitamura, Y. and M. Stutzer (1997), "An information-theoretic alternative to generalized method of moment estimation," *Econometrica*, *66(4)*, 861-874.

Kullback, J. (1959), *Information Theory and Statistics,* New York: John Wiley & Sons.

Levine, R.D. (1980), "An Information Theoretical Approach to Inversion Problems," *Journal of Physics*, *A*, 13, 91-108.

Owen, A. (1991), "Empirical Likelihood for Linear Models," *The Annals of Statistics*, 19,

1725-1747.

Pukelsheim, F. (1994), "The Three-Sigma Rule," *The American Statistician,* 48(4), 88-91.

Qin, J. and J. Lawless (1994), Empirical Likelihood and General Estimating Equations, *The Annals of Statistics*, 22, 300–325.

Renyi, A. (1970), *Probability Theory*, North-Holland, Amsterdam.

Retzer, J. and E. Soofi (2001) "Information Indices: Unification and Applications," *Journal of Econometrics* (Forthcoming).

Shannon, C.E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379-423.

Shore, J.E., and R.W. Johnson (1980), "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transactions on Information Theory*, IT-26(1), 26-37.

Skilling, J. (1988), "The Axioms of Maximum Entropy," in G. J. Erickson and C. R. Smith (eds.) *Maximum Entropy and Bayesian Methods in Science and Engineering*, Kluwer Academic, Dordrecht, 173-187.

Skilling, J. (1989), "Classic Maximum Entropy," in J. Skilling (ed.) *Maximum Entropy and Bayesian Methods in Science and Engineering*, Kluwer Academic, Dordrecht, 45-52.

Tsallis, C. (1988), "Possible Generalization of Boltzmann-Gibbs Statistics," *J. Stat. Phys.*, 52, 479-487.

Zellner, A. (1997), "The Bayesian Method of Moments (BMOM): Theory and Applications," in, T. Fomby and R.C. Hill (eds.), *Advances in Econometrics*.
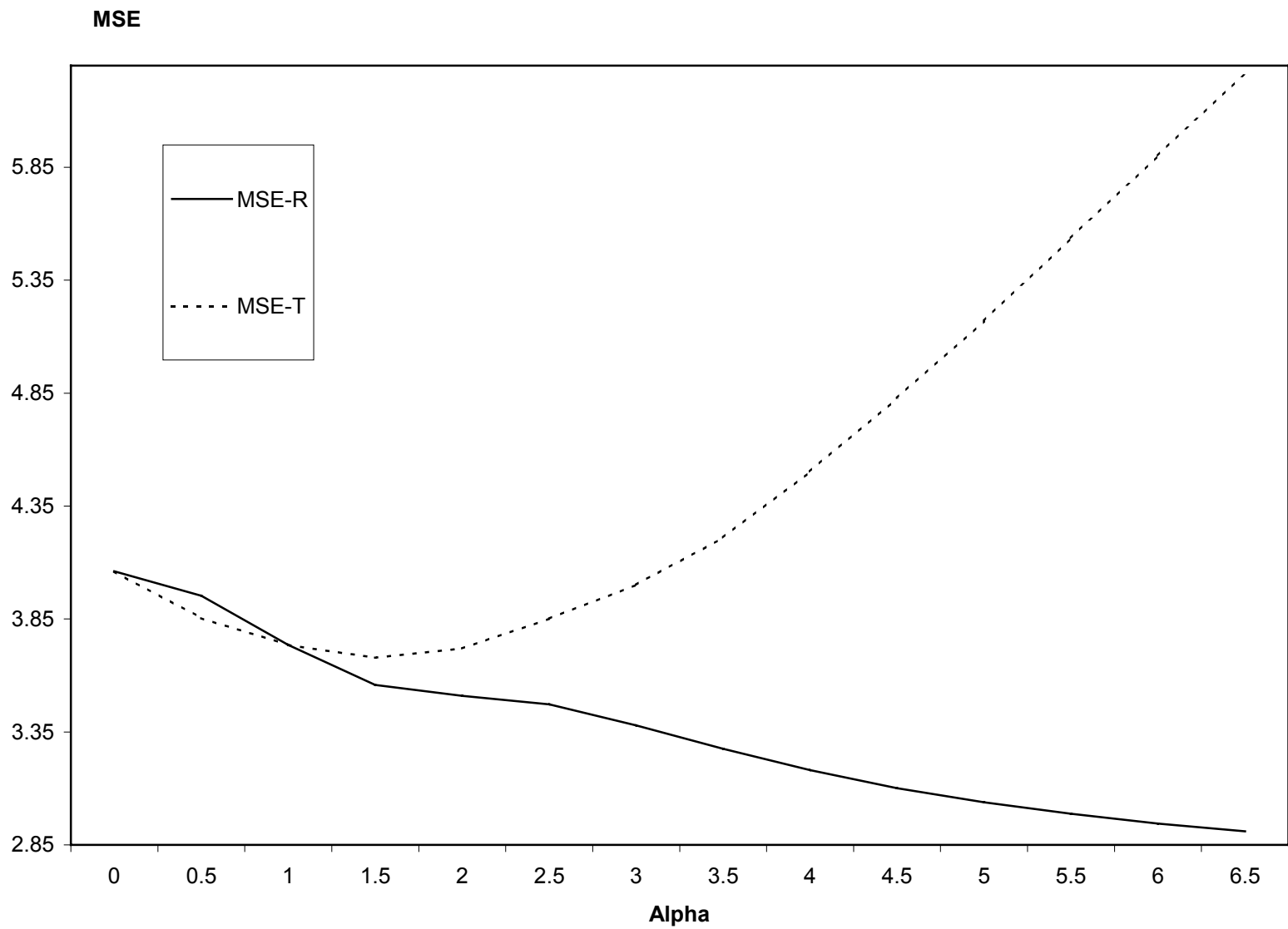
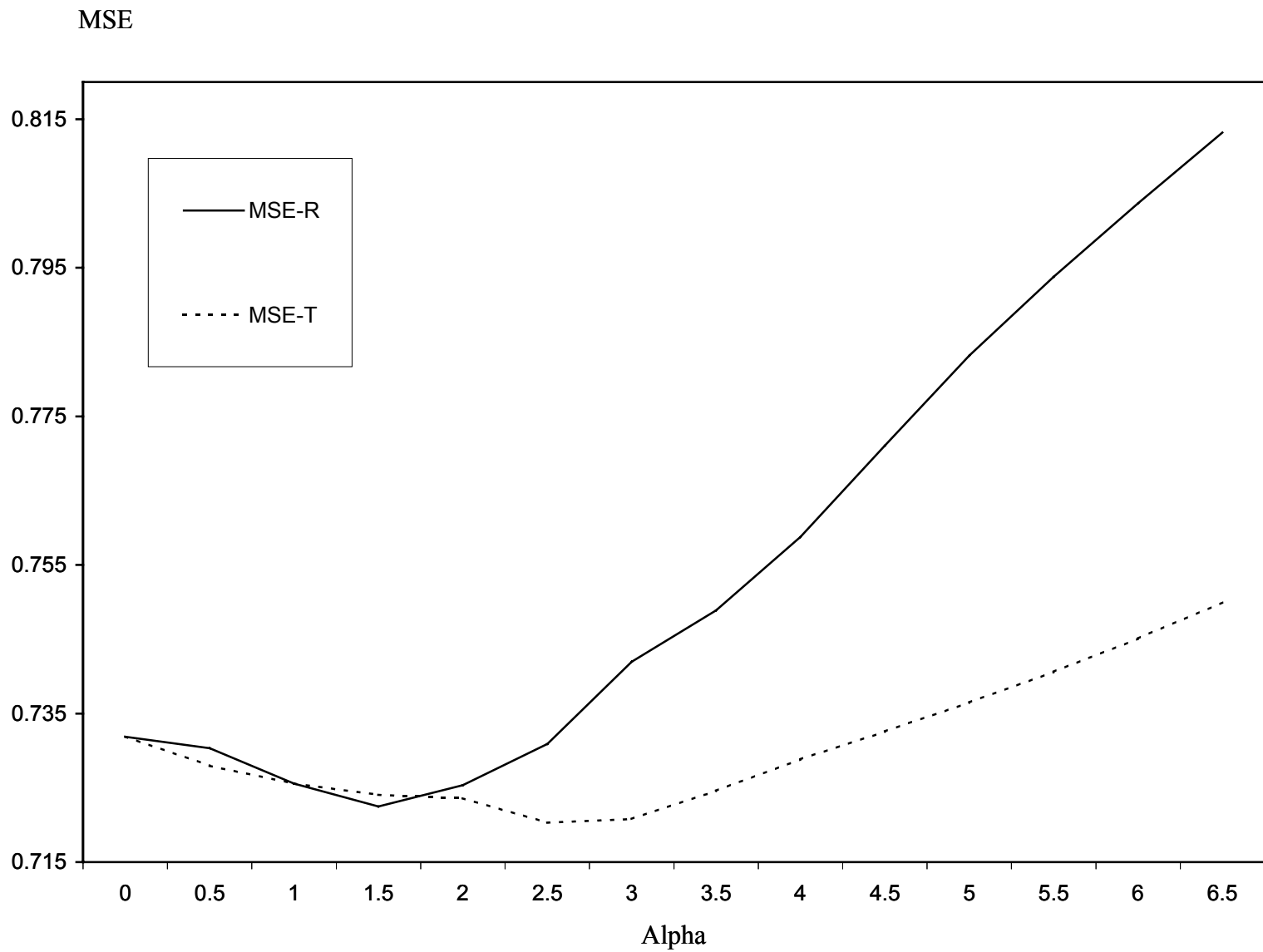Fig. 1. MSE for the two GME-α estimators for the orthonormal case with 30 observations.

Fig. 2a. MSE for the two GME-α estimators for the well-posed case (all X's are from standard normal) with 10 observations.
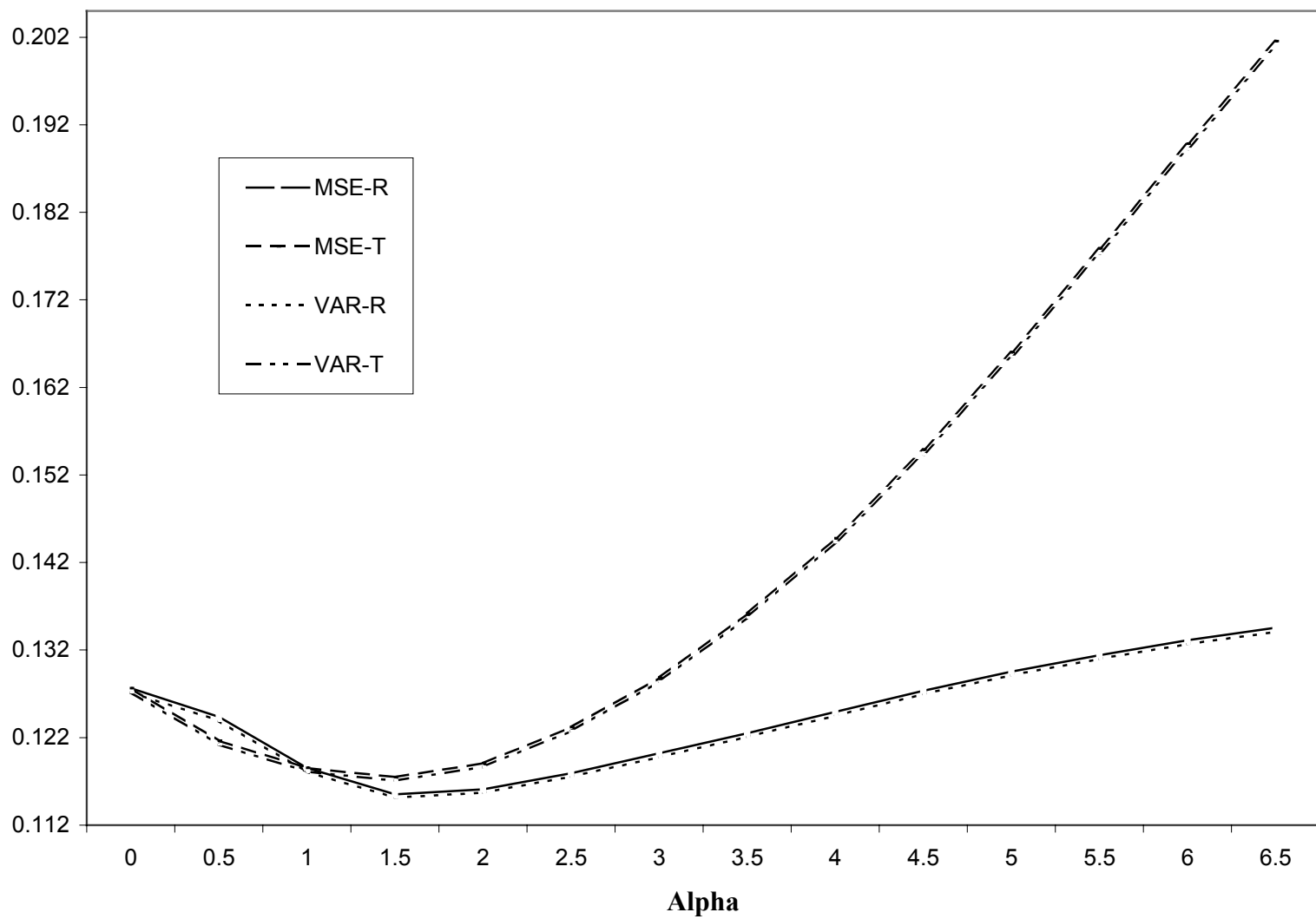
MSE, VAR

Fig. 2b. MSE for the two GME-α estimators for the well-posed case (all X's are from standard normal) with 40 observations.

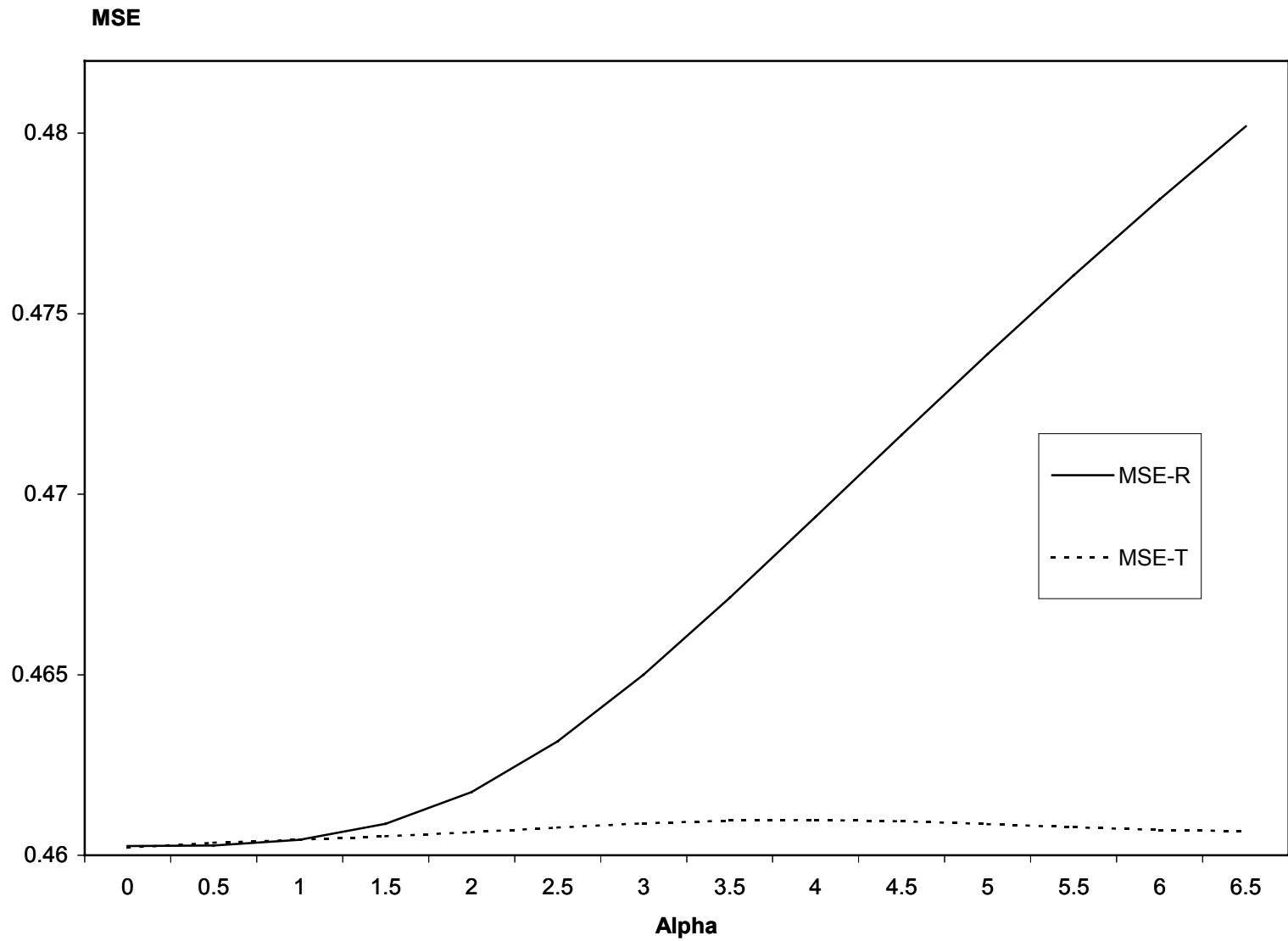Fig. 3a. MSE for the two GME-α estimators for experiment 2 with 20% outliers and 40 observations.

MSE, VAR

Alpha
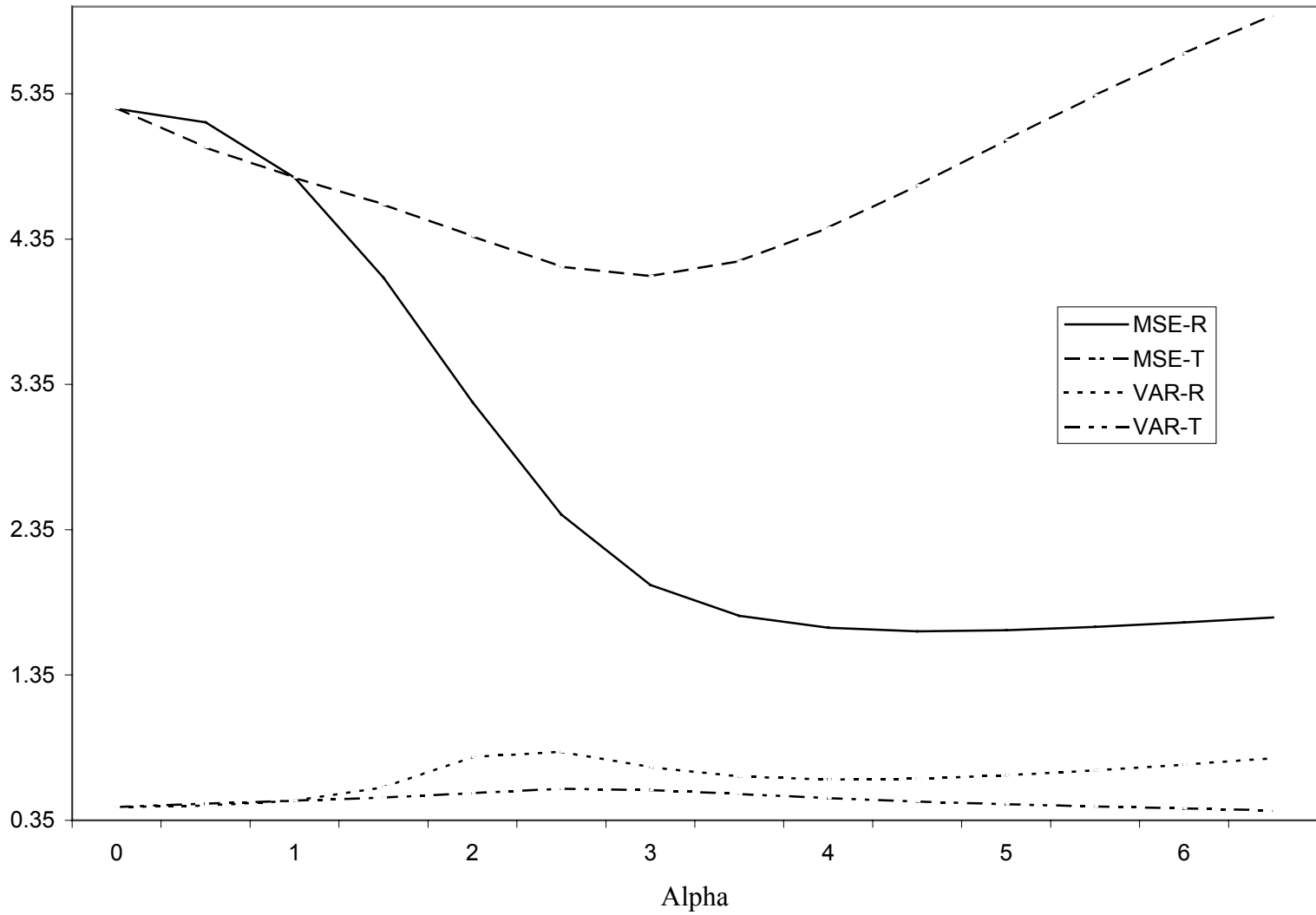
Fig. 3b. MSE for the two GME-α estimators for experiment 2 with 30% outliers and 10 observations.
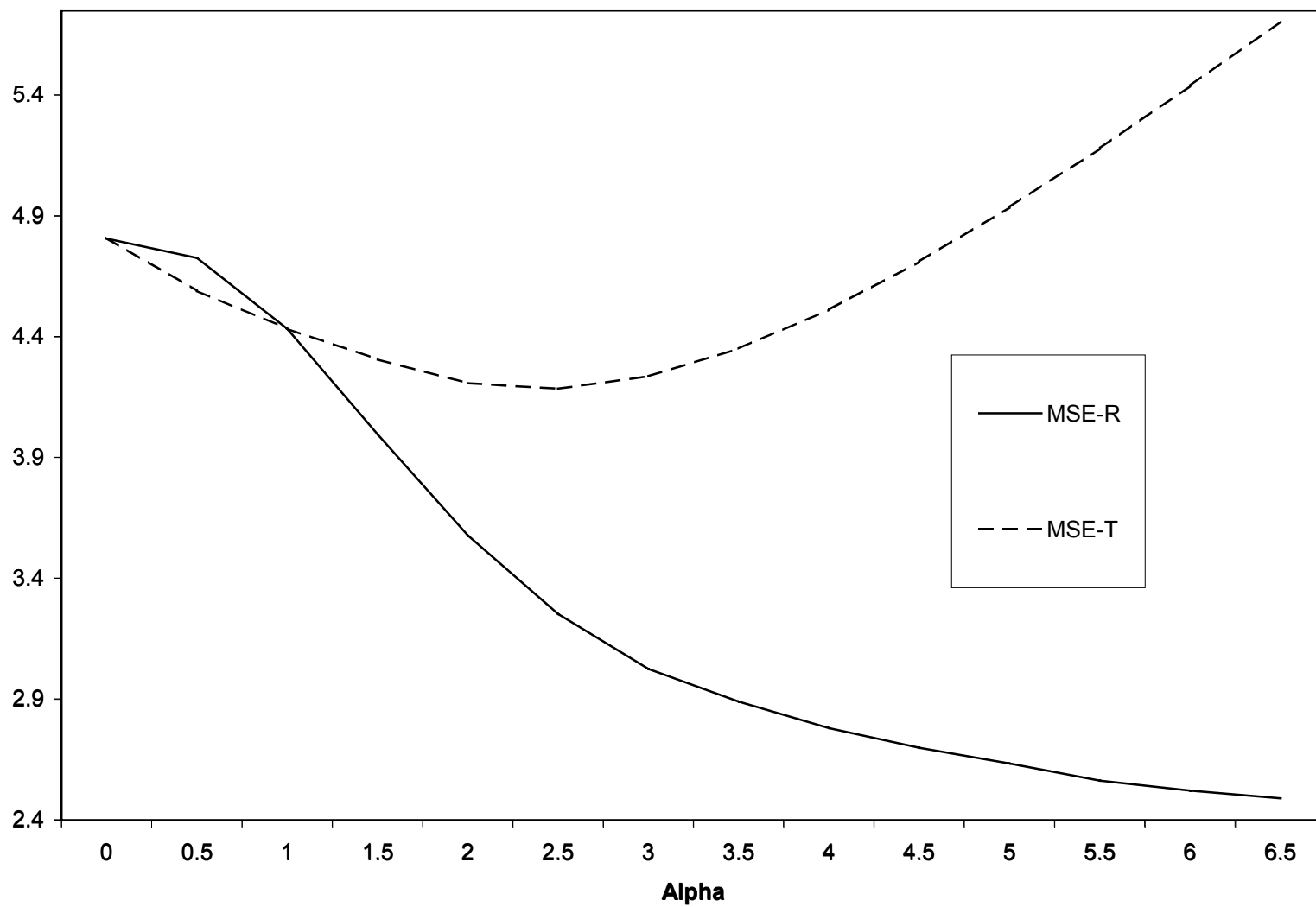
Fig. 4. MSE for the two GME-α estimators for the ill-conditioned case (condition number 90).