Robertson, John C.; Tallman, Ellis W.; Whiteman, Charles H.

Working Paper

Forecasting using relative entropy

Working Paper, No. 2002-22

Provided in Cooperation with:
Federal Reserve Bank of Atlanta

Suggested Citation: Robertson, John C.; Tallman, Ellis W.; Whiteman, Charles H. (2002) : Forecasting using relative entropy, Working Paper, No. 2002-22, Federal Reserve Bank of Atlanta, Atlanta, GA

This Version is available at:
https://hdl.handle.net/10419/100850

FEDERAL
RESERVE
BANK
of ATLANTA

Forecasting Using Relative Entropy

John C. Robertson, Ellis W. Tallman and Charles H. Whiteman

Working Paper 2002-22
November 2002

Working Paper Series

Federal Reserve Bank of Atlanta
Working Paper 2002-22
November 2002

Forecasting Using Relative Entropy

John C. Robertson, Federal Reserve Bank of Atlanta
Ellis W. Tallman, Federal Reserve Bank of Atlanta
Charles H. Whiteman, University of Iowa

Abstract: The paper describes a relative entropy procedure for imposing moment restrictions on simulated forecast distributions from a variety of models. Starting from an empirical forecast distribution for some variables of interest, the technique generates a new empirical distribution that satisfies a set of moment restrictions. The new distribution is chosen to be as close as possible to the original in the sense of minimizing the associated Kullback-Leibler Information Criterion, or relative entropy. The authors illustrate the technique by using several examples that show how restrictions from other forecasts and from economic theory may be introduced into a model's forecasts.

JEL classification: E44, C53

Key words: approximate prior information, Kullback-Leibler Information Criterion, relative numerical efficiency

---

The authors thank David Aadland, William Roberds, Frank Schorfheide, and Tao Zha for helpful discussions. They also received helpful comments from the participants in the Atlanta Fed brown bag lunch series, the Western Economics Association Meetings in Seattle 2002, the NBER Summer Workshop on Forecasting in July 2002, and seminars at the Economics Departments of the University of Georgia, Vanderbilt University, and the University of Virginia. The views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Please address questions regarding content to John C. Robertson, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street, N.E., Atlanta, Georgia 30309-4470, 404-498-8782, 404-498-8956 (fax), john.c.robertson@atl.frb.org; Ellis W. Tallman, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street, N.E., Atlanta, Georgia 30309-4470, 404-498-8915, 404-498-8956 (fax), ellis.tallman@atl.frb.org; or Charles H. Whiteman, W380 PBB, University of Iowa, Iowa City, Iowa 52242-1000, whiteman@uiowa.edu.

Forecasting Using Relative Entropy

INTRODUCTION

One of the frustrations of macroeconometric modeling and policy analysis is that empirical models that forecast well are typically nonstructural, yet making the kinds of theoretically coherent forecasts policymakers wish to see requires imposing structure that may be difficult to implement and that in turn often makes the model empirically irrelevant. In this paper, we describe the application of a procedure that can, in principle, be used to produce forecasts that are consistent with a set of moment restrictions without imposing them directly on the model. Even when it is desirable to impose the restrictions directly on the forecasting model, the technique in this paper can be used to examine the likely validity of a range of restrictions without the need to re-fit the model each time, and thereby provides the modeler with considerable flexibility to experiment with various types of restrictions.

Our procedure, inspired by Stutzer (1996) and Kitamura and Stutzer (1997), involves changing the initial predictive distribution to a new one that satisfies specified moment conditions, but that changes the other properties of the new distribution the least. That is, we minimize the relative entropy between the two distributions, subject to the restriction that the new distribution satisfies the specified moment conditions. Stutzer (1996) used this idea to modify a nonparametric predictive distribution for the price of an asset to satisfy the martingale condition associated with risk-neutral pricing. Foster and Whiteman (2002) build on this idea to price soybean options using a predictive model reflecting weather, market conditions, etc. Kitamura and Stutzer (1997) used the idea to provide an alternative to generalized method of moments estimation in which the moment conditions hold exactly relative to a new measure (but not necessarily in the data); likewise, our procedure imposes the moment conditions exactly on a

new predictive distribution that is as close (in the information-theoretic sense) as possible to the original.

The need to incorporate conditioning information into a forecast arises routinely. This is particularly true in the context of handling data release lags. In circumstances when observations on some variables are released before others, a forecaster would like to make predictions for the *unknown* post-sample values conditional on all the available data. In these circumstances, the *known* post-sample data could be thought of as a mean restriction on the forecast.

Conditioning information has been incorporated into forecasting models in a variety of settings (see for example Theil, 1971). In the VAR literature, Doan, Litterman and Sims (1984) exploit the contemporaneous and inter-temporal variance-covariance matrix structure in a VAR to account for the impact of conditioning a forecast on post-sample values for some variables in the model. Waggoner and Zha (1999) extended the Doan, Litterman and Sims analysis to accommodate uncertainty in model parameters in a fully Bayesian setting. Our procedure can be viewed as an alternative to the Waggoner-Zha technique**,** where we incorporate the conditioning information directly into the prior.

In what follows, we first sketch the theory underlying the application of relative entropy to forecasting. We then turn to three examples that illustrate the technique. The first example involves incorporation of conditioning information implicit within financial market forecasts into the predictive distribution of a vector autoregressive (VAR) model. We then turn to two examples that involve the incorporation of moment conditions implied by economic theory into VAR model forecasts.

I. UPDATING PREDICTIONS USING RELATIVE ENTROPY

I.1 *Relative Entropy and Moment Conditions.* Our interest is in the predictive distribution of an *M*-dimensional random variable *y*. In practice, it is usually difficult to derive this distribution analytically, but it is often straightforward to sample from the distribution using computer simulation techniques. Specifically, we have a sample of *N* draws $\{y_i, i = 1, \cdots, N\}$ on *y*, together with weights $\{\pi_i, i = 1, \cdots, N\}$, which ensure that each observation receives weight in the sample dictated by the predictive distribution. For a random sample from the predictive density itself, the weights are $\pi_i = 1/N$ for all *i*.

Further, we assume that we have other information about functions of *y* not used in the creation of the draws from the predictive distribution. This information takes the form of moments of a function *g*(*y*) representing quantitites such as the mean, median, standard deviation, or quantiles of the predictive distribution. The question is how to use this "new" information.

Suppose that the expectation of *g*(*y*) is equal to a known quantity, $\bar{g}$. In general,

$$(1) \qquad \sum_{i=1}^{N} \pi_i g(y_i) \neq \bar{g};$$

that is, the mean computed under the original weights will not satisfy the moment condition associated with the new information. This, of course, is what makes the information "new". Accommodating the new information requires modifying the beliefs embodied in the original weights $\{\pi_i, i = 1, \cdots, N\}$. Following Stutzer (1996) and Kitamura and Stutzer (1997), we find a new set of weights $\{\pi_i^*, i = 1, \cdots, N\}$ representing a new predictive density that is as close as possible to the original, in the information-theoretic sense, but that satisfy the specified moment

restriction. Following the notation of Soofi and Retzer (2002), the Kullback-Leibler Information Criterion (KLIC), or relative entropy of π^* to π is

$$(2) \qquad K\left(\pi^* : \pi\right) = \sum_{i=1}^{N} \pi_i^* \log\left(\frac{\pi_i^*}{\pi_i}\right) \quad .$$

This function is one convenient way to measure the new information introduced in moving from π to π^*[1]. Thus we seek new weights that minimize $K\left(\pi^* : \pi\right)$, subject to the following constraints:

$$(3) \qquad \pi_i^* \geq 0, \ \sum_{i=1}^{N} \pi_i^* = 1, \ \sum_{i=1}^{N} \pi_i^* g(y_i) = \overline{g} \ .$$

There is a substantial literature in science and statistics motivating KLIC and demonstrating its successful application (see volume 107, spring 2002, of *The Journal of Econometrics* for examples). Solution of this problem is straightforward using the method of Lagrange (see Csiszar, 1975 for the solution in the case of general probability distributions); the solution can be written as

$$(4) \qquad \pi_i^* = \frac{\pi_i \exp\left(\gamma' g\left(y_i\right)\right)}{\sum_{i=1}^{N} \pi_i \exp\left(\gamma' g\left(y_i\right)\right)}$$

where γ is the vector of Lagrange multipliers associated with the moment constraints. Thus the initial weights π have been modified, or "exponentially tilted", via (4) to generate the new weights π^* in much the same way that the state-price density modifies objective probabilities of payoffs to risk-neutral probabilities in contingent-claims asset pricing. Moreover, using the fact

[1] The KLIC is a "directed divergence" between two probability distributions. Reversing the roles of π and π^* in the objective function would yield a different set of weights. In the estimation context, the formulation we have adopted leads to the "information-theoretic" estimator of Kitamur and Stutzer (1997); the alternative leads to the "empirical likelihood" estimator of Qin and Lawless (1994).

that $\sum_{i=1}^{N}\pi_i^* = 1$, and $\sum_{i=1}^{N}\pi_i^* g(y_i) = \bar{g}$, the vector of "tilting parameters" γ can be computed as the solution to a minimization problem:

(5) $$\gamma = \arg\min_{\tilde{\gamma}} \sum_{i=1}^{N}\pi_i \exp\left(\tilde{\gamma}'[g(y_i) - \bar{g}]\right).$$

Then, with the weights in hand, one can compute the updated expectation of any other function of interest *h(y)* as $\sum_{i=1}^{N}\pi_i^* h(y_i)$.

.

I.2 *A Gaussian Example.* To illustrate the tilting procedure in an analytical context, consider the problem of finding the KLIC-closest density f^* to a bivariate normal $f(y) = N(\theta, \Sigma)$ subject to the restriction that the second variable y_2, has mean equal to μ_2 and variance equal to Ω_{22}. Letting γ_1 denote the Lagrange multiplier associated with the mean restriction and γ_2 the multiplier associated with the variance restriction, the first order conditions lead to

(6) $$f^*(y) = c \cdot f(y) \cdot \exp\left\{\gamma_1 y_2 + \gamma_2 y_2^2\right\}$$

where *c* is the normalizing constant. The exponential tilt simply adds a linear and a quadratic term to the quadratic form in the exponent of the Gaussian kernel. Upon completing the square, we find that $f^*(y) = N(\mu, \Omega)$ where μ_2 and Ω_{22} are as given, and

$$\mu_1 = \theta_1 + \Sigma_{22}^{-1}\Sigma_{12}\left(\mu_2 - \theta_2\right)$$

$$\Omega_{12} = \Sigma_{12}\Sigma_{22}^{-1}\Omega_{22}$$

$$\Omega_{11} = \Sigma_{22}^{-1}\left[\Sigma_{11}\Sigma_{22} - \Sigma_{21}\Sigma_{12}\right] + \Omega_{22}\left[\Sigma_{22}^{-1}\Sigma_{21}\right]^2$$

Thus the moment conditions lead to the usual formula for the conditional mean. If, in addition, the variance condition is $\Omega_{22} = 0$, we obtain the usual formula for the conditional variance-covariance matrix as well.

The example illustrates the general principle, apparent from (4), that for a random vector *y* with density *f*, the probability density f^* closest to *f* in the KLIC sense, such that the mean of *g*(*y*) equals $\bar{g}$ has density given by

(7) $$f^*(y) \propto f(y) \cdot \exp\{\gamma' g(y)\}$$

where γ is set to ensure that the mean restriction holds. This relationship also suggests a convenient way to sample from the density f^*, a subject we take up next.

I.3 *Relation to Importance Sampling.* Expression (7) suggests how to generate a sample from the density f^* using "importance sampling" (Geweke, 1989). Heuristically, importance sampling involves re-weighting a sample drawn conveniently from one density f so that the sample corresponds to one drawn from the "target" density f^*. Those values in the support having lower density under f^* than *f* are down-weighted; values in the support having greater density under f^* are up-weighted. Specifically, given a sample $\{y_i, i = 1, \cdots, N\}$ from the density f with weights $\{\pi_i, i = 1, \cdots, N\}$ the sample from f^* is given by the same $\{y_i, i = 1, \cdots, N\}$, but with weights $\{\pi_i^*, i = 1, \cdots, N\}$ from equation (4). Note that the *drawings* are those from the *f* density and the *weights* are adjusted to make these a set of drawings from the

f^* density. Of course, for this procedure to make sense, the support of *f* and f^* must be the same.[2]

More generally, other conditions are needed to ensure that *f* is a "good" importance density for f^*. In essence, what is required is that the weights, $\{\pi_i^*, i = 1, \cdots, N\}$, must be well-behaved. For example, the new weights should not be "too far" from the original weights $\{\pi_i, i = 1, \cdots, N\}$: that is, the new density f^* should not be too far from *f* in the KLIC sense. To monitor this, Geweke (1989) suggests keeping track of the fraction of total weight assigned to the drawing receiving highest weight. A largest weight many times larger than 1/*N*, for example, is a clear signal of an inadequate importance density. Another monitoring device advocated by Geweke is more sensitive to unequal weighting: this device is the ratio of the average sum of the squares of the highest *m* weights to the average sum of the squares of all of the weights from the importance sample. Values much larger than unity indicate unwanted variation in the weights. In our applications, we tracked *m* = 1 and *m* = 10, and these are denoted as ω_1 and ω_{10}, respectively.

Geweke suggests still another indicator to assess the quality of an importance sampler, a concept referred to as "relative numerical efficiency" (RNE). To understand the RNE consider a function *h*(*y*) having mean μ_h, and define the Monte Carlo estimator of μ_h as

$$\overline{h}_N = \frac{\sum_{i=1}^{N} w(y_i) h(y_i)}{\sum_{i=1}^{N} w(y_i)}.$$

[2] Some restrictions may be inconsistent with the predictive distribution, i.e., there may be no observations to support the moment restrictions. In this case, there is no solution to the constrained KLIC-minimization problem.

where $w(y_i)$ refers to the sampling weights, with $w(y_i) \geq 0$. The RNE is given by the ratio of the variance of $h(y)$ to the asymptotic variance of $N^{1/2}\left(\bar{h}_N - \mu_h\right)$, and can be interpreted as the number of draws necessary to achieve any given numerical standard error using the target ("tilted") density relative to the number required using the importance density.

Under standard regularity conditions, Geweke (1989, Theorem 1) establishes that $\bar{h}_N$ is consistent for μ_h. In addition, assuming that the mean of $w(y)$ and $w(y)h(y)^2$ are finite, then Geweke (1989, Theorem 2) shows that $N^{1/2}\left(\bar{h}_N - \mu_h\right)$ is asymptotically Normal with mean zero and variance σ^2, where

$$\sigma^2 = \int [h(y) - \mu_h]^2 \, w(y) \, f^*(y) dy = \int [h(y) - \mu_h]^2 \, w(y)^2 \, f(y) dy .$$

Further, $N\hat{\sigma}_N^2$ is consistent for σ^2, where

$$\hat{\sigma}_N^2 = \sum_{i=1}^{N} [h(y_i) - \bar{h}_N]^2 w(y_i)^2 / [\sum_{i=1}^{N} w(y_i)]^2 .$$

Geweke refers to the quantity $\hat{\sigma}_N$ as the numerical standard error of $\bar{h}_N$. If it had been possible initially to sample directly from the target density f^* itself, then the weights would all be unity, and σ^2 would in fact be equal to the variance of $h(y)$ under the density f^*. But since the sample is actually drawn from f and then re-weighted, the re-weighting influences the accuracy of the estimator, and this is reflected in $\hat{\sigma}_N^2$. Other things equal, large values of the weight function drive up the numerical standard error, so the natural monitoring device is the ratio the variance of $h(y)$ to the scaled numerical variance $N\hat{\sigma}_N^2$, which is estimated by

$$\text{RNE} = \frac{\sum_{i=1}^{N}[h(y_i) - \bar{h}_N]^2 w(y_i) / [\sum_{i=1}^{N} w(y_i)]}{N\sum_{i=1}^{N}[h(y_i) - \bar{h}_N]^2 w(y_i)^2 / [\sum_{i=1}^{N} w(y_i)]^2}.$$

Clearly, if the weight function is constant, RNE is unity; values of *RNE* substantially less than unity reflect very unequal weights, and signal possible numerical inaccuracies in estimating the mean of $h(y)$. The unequal weights reflect what is effectively a reduction in sample size. Indeed, as Geweke notes, the numerical standard error of $\bar{h}_N$ is $(N \cdot \text{RNE})^{-1/2}$ times the tilted standard deviation, making $(N \cdot \text{RNE})$ a measure of the effective size of the sample from the target density.[3]

Other measures of inequality in the weights can also be helpful in practice. For example, "Lorenz curves" display the fraction of total weight attributable to a given fraction of the observations. The associated Gini coefficient (twice the area between the Lorenz curve and the "perfect equality" 45-degree line) reflects the degree of inequality in the weights, and is an alternative measure to Geweke's ω_1 and ω_{10}.

In principle, there is an RNE computable for every function $h(y)$ of interest, whereas the Lorenz curve, Gini coefficient and ω_m depend only upon the single set of weights. In what follows, we judge the adequacy of our predictive densities as importance samplers for the tilted densities by reporting the weight-only measures together with the RNEs for the functions $g(y)$ associated with the tilt itself. Of course, a low RNE for the tilting function $g(y)$ need not imply a low RNE for other functions of interest, though evidence of highly unequal weighting should always suggest caution in interpreting results.

[3] In our application, the weights $w(y_i)$ are the tilted weights π_i^*, and the so the sums in the denominator of the expressions for $\hat{\sigma}_N^2$ and the variance of $h(y)$ are equal to unity.

I.4 *Interpretation of the Weight Function as a Prior Distribution.* In our applications, the distribution of interest is a predictive distribution. Such a distribution arises as follows. First, a parametric model (likelihood) for the data y given parameters θ is specified: $p(y|\theta)$. Similarly, a prior distribution for θ is specified as $p(\theta)$. By Bayes' rule, the posterior distribution for θ is proportional to the product of prior and likelihood,

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta).$$

Given the data y and the parameters θ, the distribution of a future value of y, y' is given by $p(y' \mid y, \theta)$. Then the predictive distribution is

$$f(y' \mid y) \propto \int p(y' \mid y, \theta)\, p(\theta \mid y) d\theta.$$

To sample from the predictive distribution, one typically samples θ_i from the posterior $p(\theta|y)$ and then y'_i from $p(y' \mid y, \theta_i)$. That is, we can think of the drawing y'_i as being a function of the data and the underlying parameter draw: $y'(y, \theta_i)$. Similarly, the re-weighted draw from a tilted density for y' can be thought of as a drawing from the predictive density associated with the original likelihood but with a tilted *prior* proportional to $\exp\{\gamma' g(y'(y,\theta))\}\, p(\theta)$.[4] Thus the moment condition used to modify the original predictive density can be thought of as part of the prior itself, a very natural way to incorporate non-sample information into the analysis. With the weights π_i^* in hand it would therefore be straightforward to compute updated posterior distributions for functions of interest.

[4] In general, the dependence of y' on y is nontrivial, and the "tilted prior" is data dependent. Like Zellner's (1977) "maximal data information prior", it introduces as little extra information as possible, though in our case, some of that information is data-based. Alternatively, the moment condition associated with the tilt can be thought of as post-sample information, and the tilted predictive the update of the original predictive in light of the new information.

II. EXAMPLES

In this section we present three examples that implement the relative entropy forecasting technique. In each case the basic forecast model is a vector autoregression (VAR) of the form

(7) $$y_t = b + B_1 y_{t-1} + \cdots B_p y_{t-p} + u_t,\ t = 1, \cdots, T$$

where y_t denotes an $k\times1$ vector of current dated observations for period t on the m variables in the VAR; the B_i are $k\times k$ coefficient matrices; and b is an $k\times1$ vector of constant terms. The error term is assumed to be a Normal and independently distributed $k\times1$ vector such that $E[u_t \mid y_{t-s}, s > 0] = 0$, and $E[u_t u_t' \mid y_{t-s}, s > 0] = \Sigma > 0$ for all t. The time subscript t represents months in the first example and quarters in the other examples.

Given a prior distribution $p(\theta)$ for the model parameters θ = vec(B,Σ), where $B = \left[b, B_1, \ldots, B_p\right]'$, and given the data density $p(Y_T \mid \theta)$, where $Y_T = [y_1, \ldots, y_T]'$, we generate a sample from the predictive density $f(y_{T+h} \mid Y_T)$, $h > 0$, by combining draws from the posterior $p(\theta \mid Y_T)$ (obtained via Gibbs sampling techniques), with draws from $p(y_{T+h} \mid Y_T, \theta)$. In all the applications that follow, 10,000 draws are used to build up the empirical predictive distribution.

There are two practical questions that arise when applying this technique. The first is: are the moment restrictions valid, or do they severely distort the original forecast distributions? The greater the distortion, the more unequal the weights and the lower the RNE, so we use weight-inequality and RNE measures to assess the "lack of fit" of the moment restrictions. In the present context, "lack of fit" refers to divergence between the forecast distribution generated from the underlying model and the distribution that incorporates the moment ("tilting") restrictions.

The second question is: do the moment restrictions improve the forecast performance of the model over the period being examined? For that we rely on the relative RMSE of the mean forecasts as a guide. Imposing restrictions consistent with the actual data generating process will tend to improve the forecast performance irrespective of the distortion introduced into the empirical predictive distributions; however, in a practical setting we find that large distortions to the predictive distributions as indicated by low RNEs may or may not be associated with improved forecast accuracy. Conversely, a high—RNE value implies that the restrictions will have very little impact on the forecast performance of the model.

In the first example we use a Bayesian-style VAR model to produce an alternative forecast imposing information about the future course of the federal funds rate obtained from financial markets. Since it is possible (though cumbersome) to produce approximate conditional forecasts in such a model (see Waggoner and Zha, 1999), our procedure simply provides a computationally convenient substitute for existing methods. The second and third examples show how to impose moment conditions from economic theory on the predictions of a VAR model. Specifically, the second example imposes a Taylor-rule restriction onto the forecasts from a VAR model of output, inflation and interest rates. The third example uses the covariance restriction between the intertemporal marginal rate of substitution and returns implied by a consumption-based asset pricing model to restrict the forecasts from a VAR model of consumption growth and real returns.

II.1 Forecasting the Federal Funds Rate Using Information from the Futures Market. In this example the VAR model uses a random walk Normal-Wishart prior of the type described in Sims and Zha (1998). The data are monthly observations on the federal funds rate, the log of real GDP (distributed monthly using the Chow-Lin technique), the log of the CPI price index, the log

of the price of West Texas Intermediate oil, the unemployment rate, and the log of the M2 monetary aggregate. This particular VAR model has been shown to have reasonable forecast properties over the 1990's (see Robertson and Tallman, 1999), and is routinely used in forecasting exercises at the Federal Reserve Bank of Atlanta.

Letting y_t^q denote the quarterly average of the monthly data, we calculated a sequence of empirical predictive densities $f\left(y_{T+h}^q \middle| Y_T\right)$ for h = 1,…,8 quarters beginning in January 1992, and using data for the period 1960:02 to 1991:12 to fit the model. The forecasts were updated each month as new information became available, including re-fitting the model each quarter. The process was followed for 96 months until 1999:01, resulting in an ensemble of 96 overlapping sets of 1-8-quarter-ahead forecast distributions.[5]

We impose moment restrictions on the forecasts so that the mean funds rate for the next six months coincides with the forecasts implied by data on contracts in the federal funds rate futures market. Robertson and Tallman (2001) provide details on how the implicit forecasts are extracted from the futures market data. We take the implicit futures market forecasts of the funds rate and force the mean of predictive distribution of the VAR model to equal the futures market forecasts by optimally (in the KLIC sense) choosing a new set of weights for the predictive distribution. Stopping at this point leaves the conditions "soft" in the terminology of Waggoner and Zha (1999). One could also restrict the variation around the mean forecast to be very small, meaning the conditions are essentially "hard"—the traditional conditional forecast. Another possibility would be to restrict the variability of the funds rate forecast to match the historical

[5] The three-month lag in the availability of quarterly GDP data means that, the forecasts formed at the end of February, say, are for the 24 months including January, because there is no new real GDP observation yet. For March, the forecast follows the same procedure, but there is clearly more "data" that can be used for conditioning the forecasts of January, February and March real GDP. The tilting procedure could be readily adapted to take the advance and preliminary GDP estimates as mean estimates of "final" GDP, and use the historical variability of the revision errors as variance conditions.

sample variance of the futures market forecast errors, thereby imposing the same precision as the futures market.

Table 1a presents comparisons of the standard forecast accuracy measures from the VAR models forecast and the moment restricted forecast (with the mean restricted to match that of the futures market data).[6] First, the mean federal funds rate forecasts were more accurate when they are restricted to coincide with the futures market forecasts, consistent with those in Robertson and Tallman (2001), Evans and Kuttner (1998), and Rudebusch (1998). For instance, the one-quarter-ahead relative root mean squared error (RMSE) of the restricted federal funds forecast is 60 percent lower than the RMSE associated with the VAR model's mean forecast. As we move beyond the horizons directly affected by the futures market data (which is at most two quarters), the improvement in RMSE dissipates.

Despite notably improved forecast accuracy for the federal funds rate, there is no systematic evidence that the restricted forecasts contribute to a consistent improvement in the forecast accuracy of any of the other variables. Among the more notable results, the RMSE of the 4-quarter-ahead unemployment rate forecasts is around 10 percent smaller than that of the mean VAR forecast. However, at that same four-quarter horizon, the conditional forecast errors of inflation are 10 percent larger. Also, the RMSE for the restricted unemployment rate forecast at the eight-quarter horizon is noticeably worse than that from the VAR model.

The first panel of Figure 1 displays the time series of RNE for the 1-step ahead predictive density (normalized by subtracting the corresponding futures market forecast). Numbers close to unity imply little difference between the VAR model's mean forecasts and those of the futures market. This relationship is further demonstrated in the second panel that shows the absolute

[6] The forecast accuracy results are essentially the same as those obtained using the "hard conditioning" methods of Waggoner and Zha (1999).

difference between the implied futures market 1-month ahead forecast and the VAR model forecast. The lowest RNE values are associated with periods of time when the gap between the mean of the VAR forecast and the futures market forecast are largest. The mean RNE of the 96 *h* step forecasts ranged between 0.75 and 0.79 (see Table 1b), suggesting that the moment restrictions are not severe in general. The third panel of Figure 1 displays the actual 1-month-ahead forecast errors of the futures market data and the mean VAR model forecast. The VAR model generated substantially larger forecast errors than the futures market for the period 1994 - 1995 — a period of rising interest rates and one that followed several years of low and stable interest rates. The VAR model adjusts too slowly to the local upward trend in interest rates likely due to near random walk nature of the VAR model combined with the downward trajectory of inflation and low money growth over that period. In contrast, the futures market adjusted quickly to the new policy environment.

To get a sense of the magnitude of the effect on the predictive distributions, Figures 2 and 3 depict the (smoothed) histograms of the *j*-month-ahead funds rate predictions (normalized by subtracting off the corresponding futures market forecast) formed at two distinct forecast dates; the end of September 1993 and September 1994, respectively. In each plot the dashed line is the histogram of the equally weighted draws, while the solid line is the histogram using the tilted weights. The vertical dotted line in each plot represents the unconditional sample mean, and would be zero if the restriction held unconditionally. For the September 1993 forecasts the RNE values are uniformly high and the two histograms almost lie on top of each other, suggesting a close correspondence between the model's predictions and the futures market forecasts in each period. In contrast, for a forecast formed at the end of September of 1994 the RNE values of uniformly low. In addition, as can been seen in Figure 3, when the sample mean is considerably below zero the corresponding shift in the histograms is substantial. In particular, positive draws

are up-weighted considerably relative to negative draws and the bounded support of the draws means that the titled histogram is heavily truncated.

Figure 4 displays the implied Lorenz curve for the weights at the two forecast dates. The 45-degree line corresponds to the case of equally weighting each draw. The dotted and dashed lines are the accumulated tilted weights for the September 1993 and the September 1994 forecasts. This graph highlights the difference between the weighting schemes. For example, under equal weighting 50 percent of the sample receives 50 percent of the weight. For the tilted September 1993 forecasts half the sample receives close to 40 percent of the cumulative weight, whereas by September 1994 the funds market and raw VAR forecasts are very different, requiring a substantial tilt and very unequal weighting: in this case, half the sample receives less than 10 percent of the weight, while the other half receives over 90%.

II.2 Forecasting Using Information from a Taylor Rule. In the previous example, the moment restrictions applied to a single variable. In this example, we incorporate forecast information that restricts the behavior of a linear combination of variables. The model is a quarterly VAR for the funds rate (r), CPI inflation (π) and the output gap (x).[7] The moment restriction is that the implied residual from a standard Taylor rule for given set of parameter values has mean zero over the forecast horizon. Specifically, we assume that for $h = 1,\ldots,8$,

$$r_{T+h} - 2.5 - \pi_{T+h} - 0.5(\pi_{T+h} - \pi^*) - 0.5x_{T+h}$$

has mean zero. We use an inflation target $\pi^* = 1.5$ percent, making the equilibrium real funds rate 2.5 percent; these values are typical of the literature on inflation targeting.

The VAR model uses a diffuse prior (rather than a random walk prior) because the data do not exhibit any global trends. We generated a sequence of quarterly predictive densities for h

= 1,…,8 quarters beginning in the first quarter of 1994, using data for the period 1960:1 to 1993:4 to fit the model. Sequentially, for each quarter until 1997:4, a new observation was added to the "fitting" data set, and new 1-8 step predictive distributions were simulated, resulting in an ensemble of 16 sets of 8-quarter-ahead forecast distributions.

Table 2a presents comparisons of the standard forecast accuracy measures from the VAR model forecast and the Taylor rule restricted forecast. Over the forecast period (and for this particular specification of the Taylor rule), it is clear that the moment restrictions improve forecast performance, especially for the funds rate in the short-term, and for inflation and the output gap at longer horizons. That is, the Taylor-rule appears to describe the behavior of the variables more accurately than does the unrestricted VAR model. More specifically, an examination of the individual forecasts reveals that the mean forecast from the VAR model tended to under-forecast the funds rate and over-forecast inflation during much of the forecast period. The Taylor rule, in contrast, better captured the increases in the funds rate in 1994 and the relatively tame inflation profile.

The first panel of Figure 5 displays the time series of the RNE computed for the Taylor-rule restriction applied to one-quarter-ahead forecasts. The mean RNE is 0.45, and the RNE values vary considerably, ranging from 0.54 to 0.07.[8] Thus, the distortion introduced by the Taylor-rule restriction is substantial in some periods, particularly early in the forecast period. The variability of the RNE reflects the fact that on occasion there is considerable difference between the VAR model's mean forecasts for the Taylor-rule residual and the restricted value of zero. The absolute size of VAR mean forecasts of the Taylor rule residual is presented in the second panel of Figure 5. Consistent with the previous example, the lowest RNE values are

[7] The data are taken from Leeper and Zha (2001).
[8] The mean *RNE* across the 8 moment restrictions (one per forecast step) varies between .45 and .48. (See Table 2b.)

associated with periods of time when the Taylor rule residual is the largest. Taken together these results suggest that the forecast of the chosen VAR model fitted over the whole sample is not markedly inconsistent with a particular specification of a Taylor-rule and that, in this case, the Taylor-rule introduced information that improved forecasting accuracy.

II.3 Forecasting Consumption and Returns by Incorporating Asset Pricing Model Information. In this example, we use the Euler equation from a standard specification of the inter-temporal consumption capital asset pricing model (CCAPM) as a moment restriction on forecasts of real consumption growth and interest rates. Specifically, we restrict the mean of the forecast of the product of the gross real return and the stochastic discount factor,

$$\beta\left[\left(\frac{c_{T+h}}{c_{T+h-1}}\right)^{-\alpha} r_{T+h}\right]$$

to equal unity; where r is the gross real return; c is the level of real consumption; α is the constant relative risk aversion parameter; and β is the discount factor. Unlike the previous two examples, the CCAPM moment restriction involves a non-linear function of forecasts.

In-sample applications of this specification of the CCAPM typically fit the data poorly for economically reasonable values of α and β. For our out-of-sample application, we use data on the nominal three-month Treasury bill rate as the nominal interest rate measure and the (annualized) percentage change in the CPI (average of monthly CPI levels over the quarter) as the inflation measure. To proxy the real rate of interest, we use the nominal three-month Treasury bill rate less the quarterly inflation rate measured by the CPI. For the real consumption growth rate, we add nominal consumption expenditures for services and nominal consumption expenditures for non-durable goods and then deflate that number by a geometric weighted-average of the relevant implicit deflators.

In this example, the data are stationary, so again a diffuse prior was used. Here, we generated a sequence of quarterly predictive densities for $h = 1,\ldots,8$ quarters beginning in the first quarter of 1995, using data for the period 1960:1 to 1994:4 to fit the model. Analogous to the previous examples, sequentially, for each quarter until 1999:4, a new observation was added to the "fitting" data set, and new h step predictive distributions were simulated, resulting in an ensemble of 20 sets of h-quarter-ahead forecast distributions.

For the CCAPM parameters, we set β equal to 0.96 and α equal to 2, implying a moderate degree of risk aversion. Because it is more likely that the CCAPM restriction holds as a longer run restriction rather than describing period-to-period movements we enforce the CCAPM restriction on the last forecast period (quarter 8) only.

The first panel of Figure 6 shows the time series of the relative numerical efficiency from applying the CCAPM restriction on the predictive distribution generated by the VAR model. The second panel of Figure 3 displays the absolute value of the difference between unity and the CCAPM transformation of the VAR forecasts for the real interest rate and real consumption growth. These charts show how restricting the furthest forecast period to satisfy the CCAPM restriction results in a substantial adjustment to the VAR model's predictive distribution. [9] The time-series mean *RNE* for the 8-quarter ahead prediction is 0.04, suggesting that the predictive distribution must be altered radically in order to satisfy the moment condition. [10]

Table 3a presents comparisons of the standard forecast accuracy measures from the VAR models forecast, and the CCAPM restricted forecast. Even though we impose the restriction

[9] Enforcing the restriction on earlier forecast periods in addition to the final forecast period exacerbates the distortion to the predictive distribution. Searching across values for α we find that smallest KLIC value is generated by setting the relative risk aversion equal to -0.375, consistent with non-concave utility, and comparable to the empirical results of Hansen and Singleton (1996). See Neely, Roy, and Whiteman (2001) for a demonstration that such estimates can be traced to near non-identification of the model due to poor predictability of consumption growth and returns.

[10] See Table 3b.

only in the final forecast period, there are noticeable impacts on the accuracy of the restricted forecasts in earlier periods as well. For 8-quarters-ahead, the RMSEs for the restricted forecasts are around twice that those of the VAR model. At a 4-quarter horizon, the RMSEs for the restricted forecasts are about 1.5 times those of the VAR model, while 1-quarter ahead the difference is negligible. Hence, in this case, the large distortion introduced by imposing this particular specification of the CCAPM coincided with poor forecast performance as well. Despite the distortion, economic interpretations of the mean forecasts for the real interest rate and the growth rate of real consumption are consistent with the CCAPM restriction: forecasts of the real interest rate are increased and the forecasts of the real consumption growth rate are lowered relative to the respective VAR forecasts.

III. CONCLUSION

This paper has described a relative entropy procedure for imposing moment restrictions on simulated distributions from a variety of models. The technique produces a set of weights that imply a distribution that is as close as possible to the original in the sense of minimizing the associated Kullback-Leibler Information Criterion, or relative entropy. The technique is illustrated by three examples that progress from *atheoretic* conditional forecasting, to imposing restrictions from a theoretical model on a forecast. The preliminary results from the application of the technique are encouraging, and the potential breadth of application seems to be large.

References

Csiszár, I., (1975). "*I*-Divergence Geometry of Probability Distributions and Minimization Problems," *The Annals of Probability* 3:146-158.

Doan, T., Litterman, R., and C. Sims (1984), "Forecasting and Conditional Projection Using Realistic Prior Distributions," Econometric Reviews 3:1-100.

Evans, C. L. and Kuttner, K. N., (1998), "Can VARs Describe Monetary Policy?" in *Topics in Monetary Policy Modeling*. Basle: Bank of International Settlements, 93-109.

Foster, F.D., and C.H. Whiteman, (2002), "Bayesian Prediction, Entropy, and Option Pricing in the U.S. Soybean Market, 1993-1997," University of Iowa manuscript.

Geweke, J., (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration," Econometrica, Vol. 57, No. 6 (November), 1317-1339.

Hansen, L.P. and K. Singleton, (1983) "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns." Journal of Political Economy, Vol 91, no 2, pp 249-265.

Hanesen, L.P., and K. Singleton (1996), "Efficient Estimation of Linear Asset Pricing Models With Moving Average Errors," Journal of Business and Economic Statistics 14:53-68.

Kitamura, Y., and M. Stutzer (1997), "An Information–Theoretic Alternative to Generalized Method of Moments Estimation," Econometrica 65:861-874.

Neely, C.J., Roy, A., and C.H. Whiteman, (2001), "Risk Aversion Versus Intertemporal Substitution: A Case Study of Identification Failure in the Intertemporal Consumption Capital Asset Pricing Model," Journal of Business and Economic Statistics 19:395-403.

Qin, J., and J. Lawless, (1994). "Empirical Likelihood and General Estimating Equations," Annals of Statistics, Vol. 22, No. 1. (March), pp. 300-325.

Robertson, John C. and Ellis W. Tallman. 1999. "Vector Autoregressions: Forecasting and Reality." Federal Reserve Bank of Atlanta *Economic Review*, First Quarter, 4-18.

Robertson, John C. and Ellis W. Tallman. 2001. "Improving Federal-Funds Rate Forecasts in VAR Models Used for Policy Analysis," *Journal of Business and Economic Statistics* 19 (July): 324-30.

Rudebusch, G. D. (1998), "Do Measures of Monetary Policy in a VAR Make Sense?" *International Economic Review*, 39, 907-31.

Sims, Christopher A. and Tao A. Zha. 1998. “Bayesian Methods for Dynamic Multivariate Models.” *International Economic Review.”* 39, 4: 949–968.

Stutzer, M., (1996). "A Simple Nonparametric Approach to Derivative Security Valuation," Journal of Finance, Vol. 51, December.

Theil, Henri. (1971). *Principles of Econometrics*, John Wiley and Sons, New York.

Waggoner, D.F., and T. Zha, (1999), “Conditional Forecasts in Dynamic Multivariate Models,” The Review of Economics and Statistics 81(4):639-651.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics, J. Wiley and Sons, Inc., New York.

Zellner, A., (1977). "Maximal Data Information Prior Distributions," in A. Aykac and C. Brumat (editors), New Developments in the Applications of Bayesian Methods, Amsterdam: North-Holland, 211-232.

Table 1a: Federal Funds Futures Market Example Forecasting Accuracy Results

Relative Root Mean Squared Forecast Error: Model Restricted to Match Federal Funds Rate Futures Market Forecast relative to Bayesian Vector Autoregression Model

Forecast period 1992Q1 to 2001Q4

Quarters Ahead:	1	2	3	4	5	6	7	8
Federal Funds Rate	0.39	0.61	0.74	0.77	0.84	0.90	0.93	0.95
CPI Inflation Rate	1.02	1.04	1.10	1.11	1.06	1.02	1.03	1.01
Unemployment Rate	0.95	0.95	0.93	0.90	0.93	0.98	1.02	1.06
Real GDP Growth Rate	0.97	0.97	0.98	1.05	1.02	1.03	1.05	1.05

Table 1b: Importance Sampling Diagnostics

Relative Numerical Efficiency

Steps:	1	2	3	4	5	6
Mean	0.79	0.79	0.78	0.77	0.76	0.75
Median	0.87	0.87	0.87	0.86	0.86	0.85
Standard Deviation	0.22	0.22	0.23	0.24	0.24	0.25
Range	0.92	0.94	0.95	0.96	0.96	0.96

Diagnostic	Mean	Median
KLIC	0.13	0.06
Largest Weight	7.29	3.36
ω_1	48.21	10.05
ω_{10}	20.86	6.74
GINI	.23	.19

Table 2a: Taylor Rule Example Forecasting Accuracy Results

Root Mean Squared Forecast Error of Taylor Rule Restricted Model relative to Unrestricted Vector Autoregression Model:

Forecast period 1992Q1 to 1999Q4

Steps:	1	2	3	4	5	6	7	8
Federal Funds rate	0.83	0.73	0.68	0.70	0.71	0.82	0.98	1.08
Inflation	1.08	1.10	1.00	0.92	0.92	0.84	0.82	0.85
Output Gap	1.00	0.99	0.94	0.91	0.91	0.90	0.89	0.90

Table 2b: Importance Sampling Diagnostics – Taylor Rule Example

Relative Numerical Efficiency – Taylor Rule

Steps:	1	2	3	4	5	6	7	8
Mean	0.45	0.48	0.48	0.48	0.46	0.46	0.46	0.45
Median	0.45	0.47	0.51	0.56	0.53	0.53	0.51	0.51
Standard Deviation	0.23	0.23	0.23	0.22	0.23	0.22	0.21	0.21
Range	0.72	0.68	0.70	0.67	0.66	0.62	0.62	0.61

Diagnostic	Mean	Median
KLIC	0.35	0.28
Largest Weight	23.84	13.39
ω_1	270.53	113.57
ω_{10}	105.66	54.216
GINI	.42	.40

KLIC is the Kullbach-Leibler information criterion.

Table 3a: Forecast Comparison Results for Consumption CAPM Restriction

Root Mean Squared Forecast Error of Consumption CAPM Restricted Model relative to Unrestricted Vector Autoregression Model:

Forecast period 1995Q1 to 2001Q4

Steps:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Consumption	1.09	1.27	1.41	1.61	1.75	1.86	2.07	2.14
Real interest rate	0.98	1.06	1.08	1.38	0.96	1.05	1.18	1.88

Table 3b: Importance Sampling Diagnostics - Consumption CAPM

Relative Numerical Efficiency - CAPM

Step:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mean	0.19	0.16	0.13	0.12	0.10	0.08	0.07	0.05
Median	0.20	0.17	0.14	0.14	0.11	0.08	0.08	0.06
Standard Deviation	0.07	0.05	0.05	0.06	0.05	0.04	0.04	0.03
Range	0.24	0.17	0.21	0.18	0.16	0.13	0.12	0.09

Diagnostic	**Mean**	**Median**
KLIC	0.66	0.63
Largest Weight	119.53	88.74
ω_1	2420.1	1456.3
ω_{10}	443.29	384.16
GINI	.59	.58

Figure 1: Federal Funds Futures Market Restriction Example

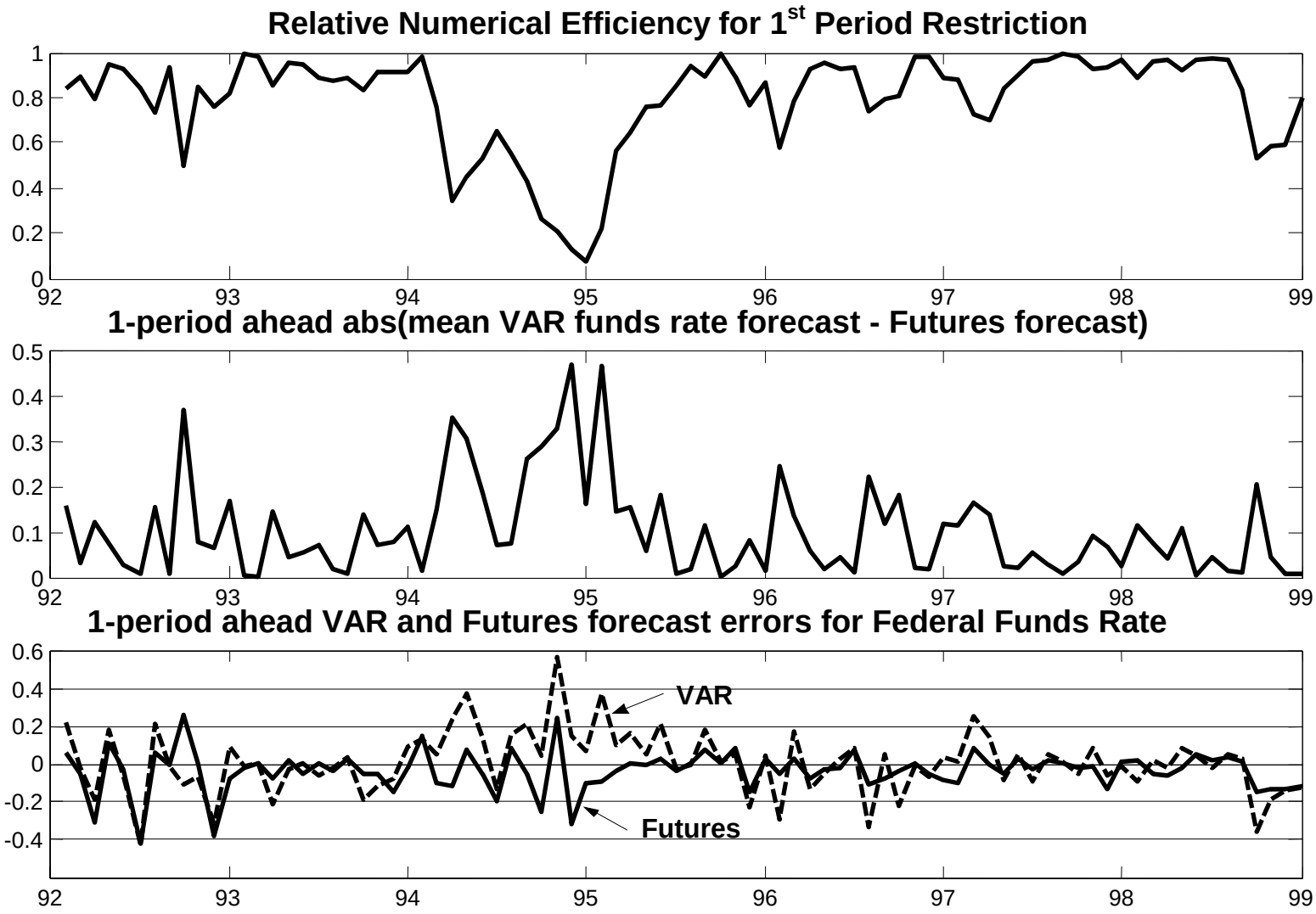

Figure 2: Tilted vs 1/n Empirical Error Distributions (93:09 Forecast)

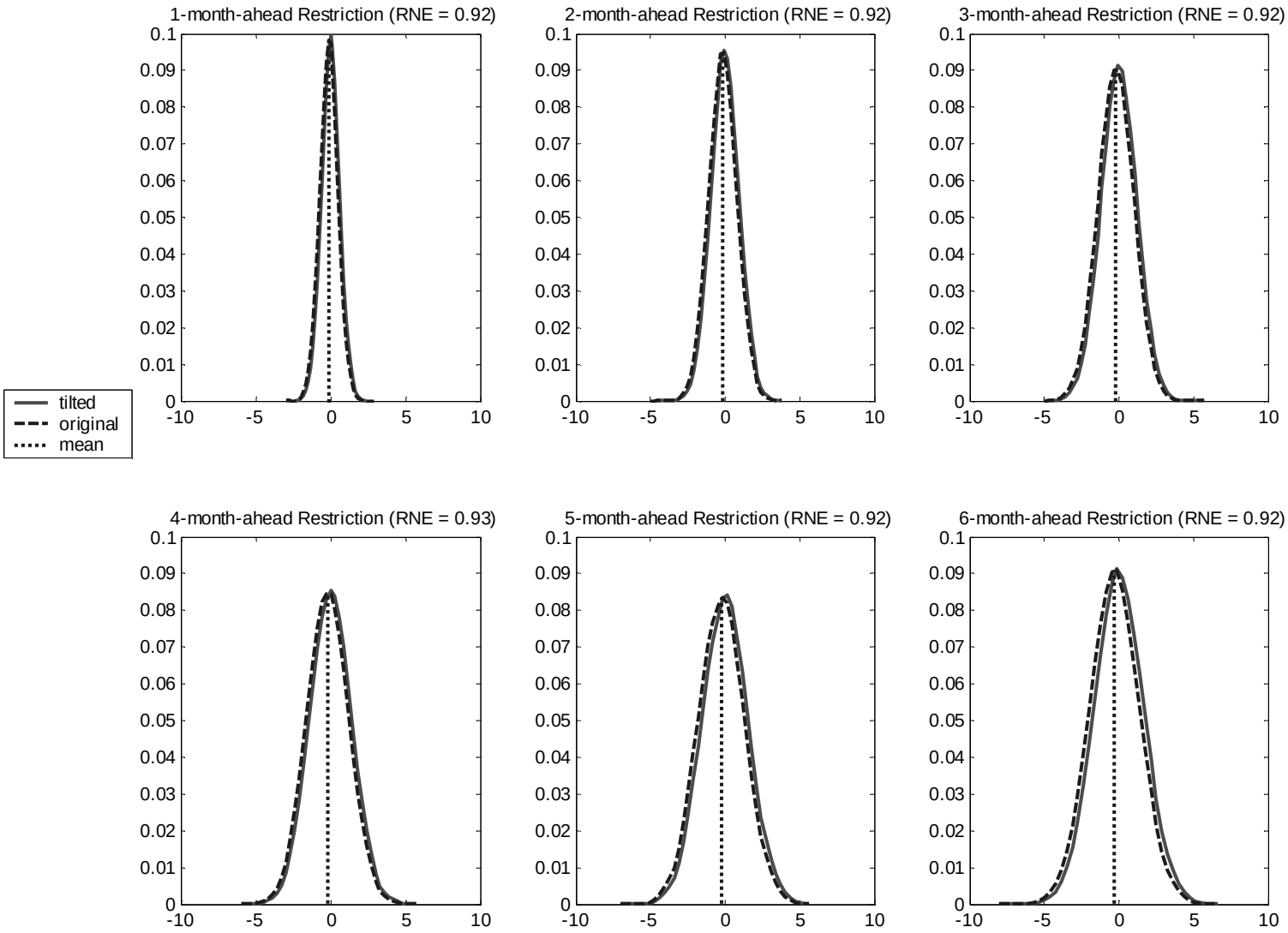

Figure 3: Tilted vs 1/n Empirical Error Distributions (94:09 Forecast)

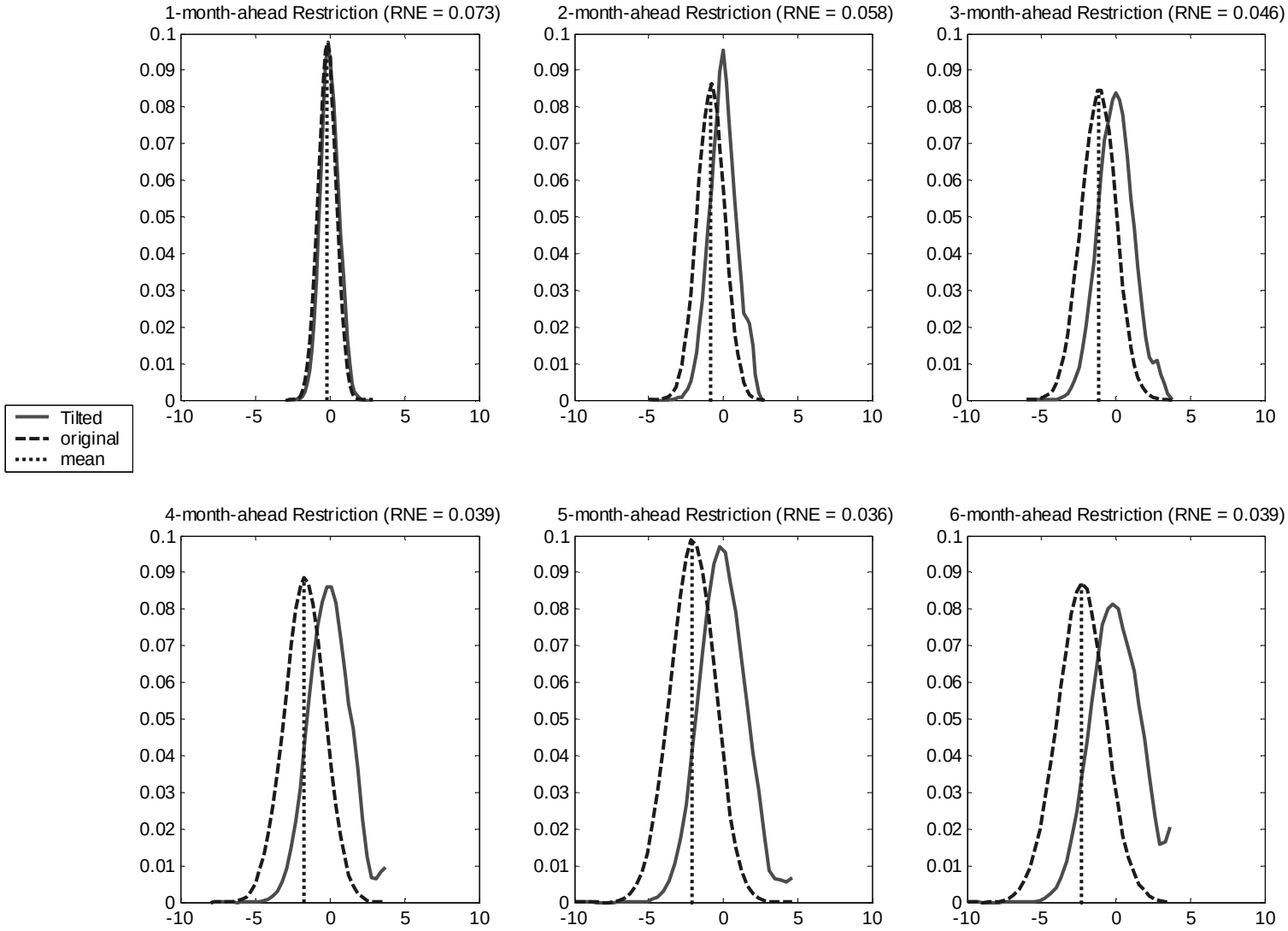

Figure 4: Lorenz Curve of Normalized Weights
(93:09 and 94:09 Funds Rate Forecasts)

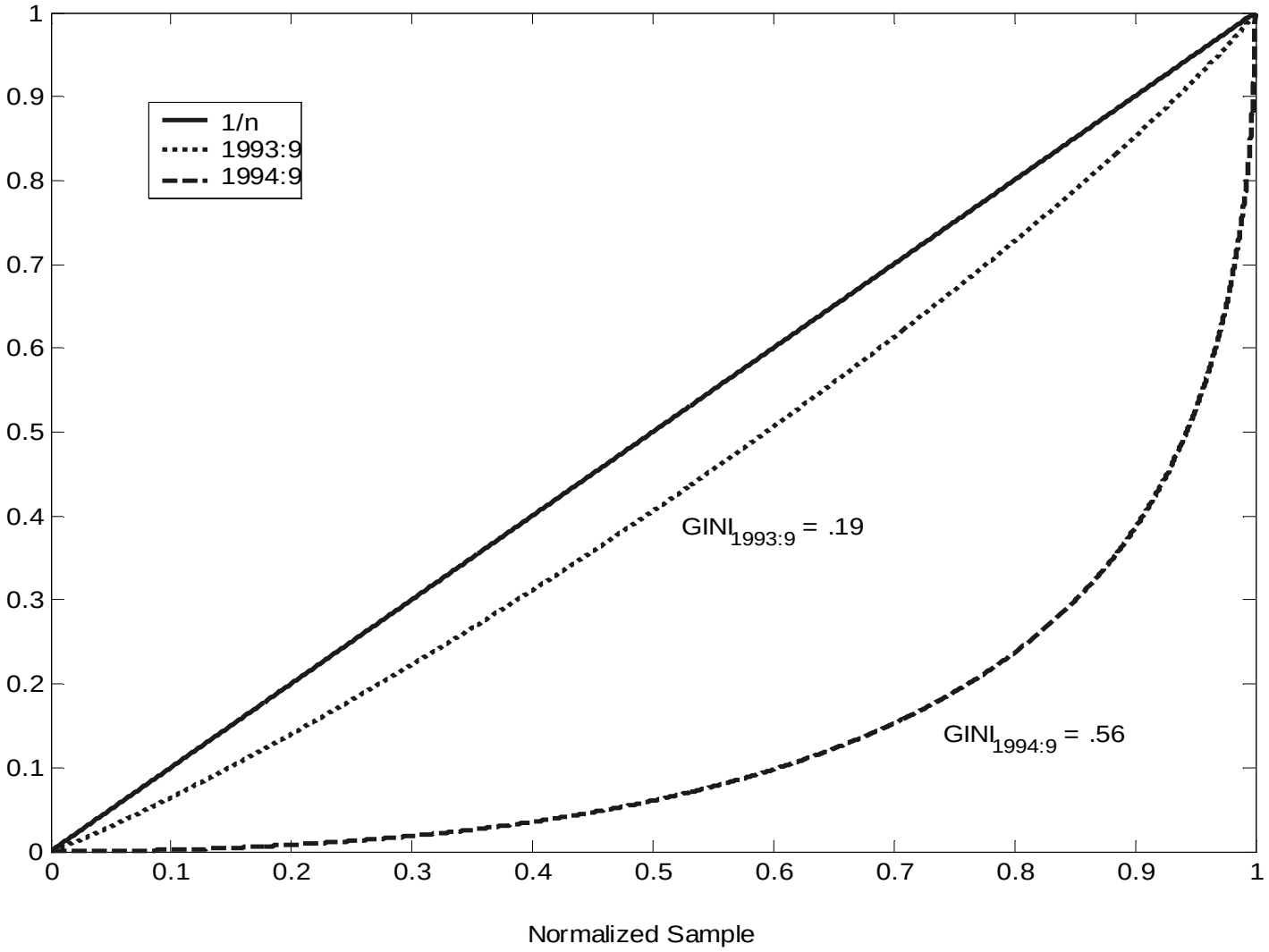

Figure 5: Taylor Rule Restriction Example

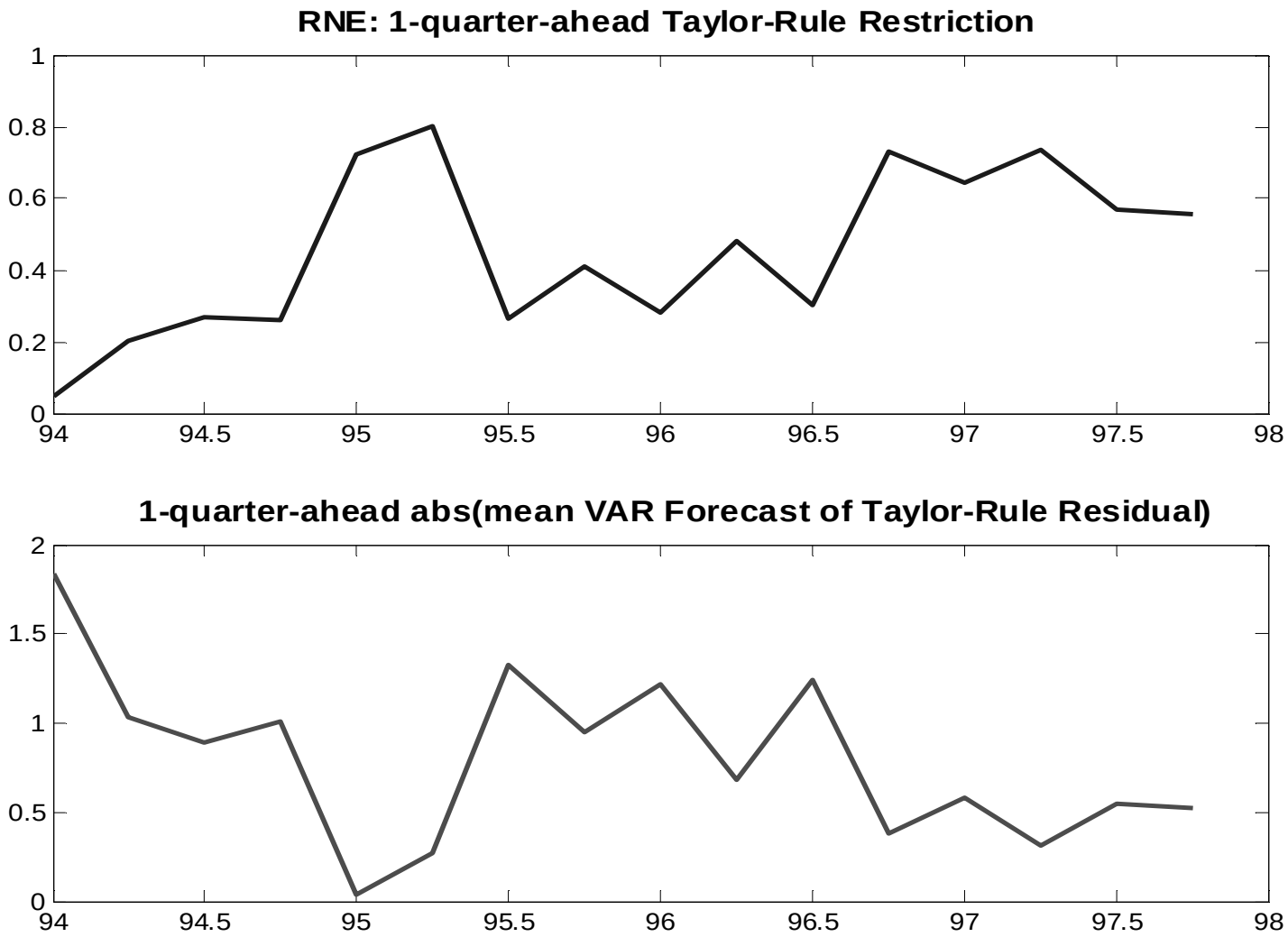

Figure 6: Consumption CAPM Restriction Example