第 8 章　预测评价与预测组合*

Francis X. Diebold 和 Jose A. Lopez

显然，在经济学和金融学中，预测非常重要，并且广为使用。很简单，好的预测导致好的决策。预测评价与预测组合技术的重要性紧随其后——预测使用者很自然地会热切关注预测的追踪和改进预测的表现。更一般地，预测评价在实证经济学和金融学的许多问题中都位置显赫，比如：
——预期是理性的吗？（例如Keane和Runkle，1990；Bonham和Cohen，1995）
——金融市场有效吗？（例如Fama，1970，1991）
——宏观经济冲击导致代理人在所有的时间跨度上修正他们的预期，或者只是在短期和中期的时间跨度上修正他们的预期？（例如Campbell和Mankiw，1987；Cochrane，1988）
——观察到的资产收益率是否“过于波动”？（例如Shiller，1979；LeRoy和Porter，1981）
——从长期来看，资产的收益率可预测吗？（例如Fama和French，1988；Mark，1995）
——远期汇率是否是不同时间跨度上的未来即期价格的无偏、并且/或者精确的预测？（例如Hansen和Hodrick，1980）
——政府预算计划可能是由于策略上的原因，系统性地过于乐观吗？（例如Auerbach，1994；Campbell和Ghysels，1995）
——名义利率是未来通货膨胀的良好预测吗？（例如Fama，1975；Nelson和Schwert，1977）

本文中，我们有选择性地分五节阐述预测评价和预测组合的方法。第一节，讨论单个预测的评价，尤其是评价能否以及如何改进预测。第2节，讨论相比较（competing）的预测精度的评价和比较。第3节，讨论是否以及如何把一系列预测组合在一起，得到一个更优良的复合预测（composite forecast）。第4节，阐述几个与经济学和金融学特别相关的预测评价问题，包括评价变动方向预测、概率预测和波动率（volatility）预测的方法。第5节，给出结论。

在处理预测评价的问题上，一般性和繁琐性之间会有一个权衡。因此，大部分篇幅集中于单变量协方差平稳过程的线性最小二乘预测，或者假定正态性，以使线性预测和条件期望一致。我们留待读者去充实其余方面的内容。然而，在某些特别值得关注的情形中，我们确实直接关注导致线性预测与条件均值不一致的非线性，以及需要特别注意的非平稳。

1　单个预测的评价

最优预测的性质众所周知，预测的评价基本上就是检验这些特性。首先，建立一些表示法并且回顾一些熟悉的结果。用 y_t 表示我们关注的协方差平稳时间序列。假设唯一的确定性分量是一个可能非零的均值 μ，Wold表达式为 $y_t = \mu + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \ldots$，其中 $\varepsilon_t \sim WN(0, \sigma^2)$，WN表示没有序列相关的（但不一定是高斯的，因而也不一定是独立的）白噪声。本文均假设可逆性，这等价于具有单边自回归表达式。

向前K步线性最小二乘预测为 $\hat{y}_{t+k,t} = \mu + b_k\varepsilon_t + b_{k+1}\varepsilon_{t-1} + \ldots$，相应的向前K步预测误差为

$$e_{t+k,t} = y_{t+k} - \hat{y}_{t+k,t} = \varepsilon_{t+k} + b_1\varepsilon_{t+k-1} + \cdots + b_{k-1}\varepsilon_{t+1}。 \qquad (1)$$

* 我们感谢 Clive Granger 有价值的评论，我们也感谢国家科学基金、斯隆基金及宾西法尼亚大学研究基金的资助。

最后，向前K步预测误差的方差为。

$$\sigma_k^2 = \mathrm{var}(e_{t+k,t}) = \sigma^2(\sum_{i=1}^{k-1} b_i^2) \quad 。 \qquad (2)$$

从最优预测立即可得到其误差的四个主要特性（下面有更详细的讨论）：

（1）最优预测误差的均值为零（由（1）得）；

（2）向前1步最优预测误差是白噪声（（1）相应于k=1时的特例）；

（3）向前K步最优预测误差最多为MA(k-1)（（1）的一般情形）；

（4）向前K步最优预测误差的方差随k非减（由（2）得）。

在继续探讨之前，我们现在阐述一些精确的，无分布（distribution-free）的检验，以检验独立（但不一定相同）的分布序列是否具有为零的中位数。这些检验对评价上述最优预测误差的特性以及检验其他以后将会涉及的假设是有用的。这样的检验有许多，其中最流行也反复使用的有两个，即符号检验和Wilcoxon符号-秩检验。

用 x_t 表示要检验的序列，假设有 T 个观测值。实施符号检验的零假设是观测值序列独立、中位数为零。[1] 检验统计量是直观的，构造简单明了——在零假设下，容量为 T 的样本中，正观测值数目的分布是参数为 T 和 1/2 的二项式分布。因此，检验统计量是简单的

$$S = \sum_{t=1}^{T} I_+(x_t),$$

其中

$$I_+(x_t) = \begin{cases} 1 & \text{如果} x_t > 0, \\ 0 & \text{其他。} \end{cases}$$

在大样本中，统计量的学生化形式是标准正态分布，

$$\frac{S - T/2}{\sqrt{T/4}} \overset{a}{\sim} N(0,1) 。$$

因此，可用标准的二项式分布表或正态分布表来评估统计量的显著性。

应注意的是，符号检验不要求分布对称。Wilcoxon 符号-秩检验，一个有关无分布的方法，*确实*要求分布对称，当分布对称时，它具有比符号检验更高的功效。除了需要增加对称性假定，零假设是一样的，检验统计量是正观测值的绝对值的秩之和，

$$W = \sum_{t=1}^{T} I_+(x_t) Rank(|x_t|),$$

其中排序是递增的（例如，指定最大绝对值的观测值的秩为 T，等等）。检验表面看来很简单——如果潜在分布（underlying distribution）关于零对称，正观测值绝对值的秩之和“非常大”（或者“非常小”）是“非常不可能的”。符号-秩统计量的精确有限样本的零假设分布没有多余参数（nuisance parameter），也不随真实潜在分布的改变而改变，它已被制成统计表。此外，在大样本时，统计量的学生化形式是标准正态分布，

$$\frac{W - [T(T+1)]/4}{\sqrt{[T(T+1)(2T+1)]/24}} \overset{a}{\sim} N(0,1) 。$$

[1] 如果序列为对称分布，那么中位数为零当然等同于均值为零。

最优预测特性的检验

给定预测的追踪纪录（track record）$\hat{y}_{t+k,t}$ 以及相应的实际值 y_{t+k}，预测使用者自然想评价预测的表现。上述分类的最优预测特性，很容易检验。

a. 最优预测误差的均值为零

取决于愿意保留什么样的假定，对这个假设可以做种种不同的标准检验。例如，如果 $e_{t+k,t}$ 为高斯白噪声（向前1步预测误差可能就是这种情况），那么显然可以选择标准 t 检验，因为它是精确而且始终最有效力。如果误差不是高斯分布但仍然是独立同分布（i.i.d），那么t检验在大样本时仍然是有用的。然而，如果有（或可能有）更复杂的依赖关系或异质结构，那么就需要用别的检验，比如那些基于广义矩方法的检验。

由于有时只可以得到短期的追踪纪录，如果非正态性或者更多的依赖/异质结构要求使用渐近检验方法，那将是不幸的。然而，事实并非如此，因为正如 Campbell和Ghysels（1995）所指出的那样，通常可以用无分布的非参数检验。尽管无分布的检验确实要求独立性（符号检验）或者独立性和对称性（符号-秩检验），但是这些检验不要求分布在不同时间是正态分布或同分布。这样，这些检验对各种预测误差分布，以及独立但不同分布类型的异方差是自动稳健的（automatically robust）。

然而，当k>1时，即使最优预测的误差也可能呈现序列相关，因此必须对非参数检验进行修正。在预测误差是（k-1）阶相关的假设下，下面k个预测误差序列中的每一个都无序列相关：$\{e_{1+k,1}, e_{1+2k,1+k}, e_{1+3k,1+2k}, \cdots\}$ ， $\{e_{2+k,2}, e_{2+2k,2+k}, e_{2+3k,2+2k}, \cdots\}$ ， $\{e_{3+k,3}, e_{3+2k,3+k}, e_{3+3k,3+2k}, \cdots\}$ ，$\cdots$，$\{e_{2k,k}, e_{3k,2k}, e_{4k,3k}, \cdots\}$。因此，对K个误差序列的每序列进行K次检验，每次检验的基准（size）为 α/k，如果*任何*一个序列拒绝零假设，则拒绝零假设，这就是Bonferroni边界检验（基准上界为α）。这个方法是保守的，甚至在渐近时也是保守的。作为另一种选择，可以仅使用k个误差序列中的一个，并在α水平上进行一次精确检验，其代价是由于舍弃了许多观测值，功效会降低。

在总结这一节时，我们强调，无分布的非参数检验与更常见的检验相比，既不是肯定“更好”也不是肯定“更差”，它们有各自适用的场合，因而是互补的。非参数检验的优点是，它们通常是精确的有限样本检验，具有很好的小样本功效。此外，要在小样本时使用更标准的检验方法，必须满足正态性和同方差假定，非参数检验对是否偏离这些标准假定并不敏感。然而，非参数检验的缺点是它们要求预测误差是独立的，这一假定甚至比条件均值独立还强，更不用说线性投影（linear-projection）独立的假定了。此外，尽管非参数检验经过修正，可以允许预测误差有k阶相关，但是必须付出很大的代价，这个代价可能是基准不精确或者功效下降。

b. 向前 1 步最优预测误差是白噪声

更准确地说，线性最小二乘预测的误差是线性投影独立的，而最小二乘预测的误差是条件均值独立的。从来不需要误差完全序列独立，例如，就像GARCH过程的条件方差相关，总是可以通过更高阶矩引入相关性。

在保留不同假定的集合下，可以用标准的渐近检验来检验白噪声假设。例如，样本自相关和

偏自相关函数，再加上Bartlett的渐近标准误，可以对预测误差是否是白噪声进行有效的图形诊断。基于序列相关系数的标准检验以及Box-Pierce和有关的统计量可能也是有用的。

Dufour（1981）对符号检验和Wilcoxon符号-秩检验进行改进，得到了向前1步预测误差序列独立性的精确检验，不要求正态性或预测误差同分布。例如，考虑预测误差服从中位数为零的独立对称分布这个零假设。那么，中位数（$e_{t+1,t}e_{t+2,t+1}$）=0；即，两个中位数为零的独立对称随机变量的乘积本身也服从中位数为零的对称分布。在预测误差有正的序列相关的备择假设下，中位数（$e_{t+1,t}e_{t+2,t+1}$）>0；在预测误差有负的序列相关的备择假设下，中位数（$e_{t+1,t}e_{t+2,t+1}$）<0。这其实是检验交叉乘积序列 $z_t = e_{t+1,t}e_{t+2,t+1}$ 是否关于零对称。检验 z_t 是否关于零对称，显然应该用符号-秩检验，$W_D = \sum_{t=1}^{T} I_+(x_t)Rank(|z_t|)$。应注意的是，即使 $e_{t+1,t}$ 不是序列相关的，z_t 也可能是序列相关的，从表面上看来，这明显违反了有效的符号-秩检验应该符合（可以用于检验 z_t）的条件。因此，Dufour的贡献是重要的——Dufour证明，序列相关没有影响，W_D 的分布与 W 的分布是一样的。

c. 向前K步最优预测误差最多为MA(k-1)

Cumby和Huizinga（1992）对大于k-1阶的序列相关提出了一个有用的渐近检验。零假设是 $e_{t+k,t}$ 序列为MA(q)（$0 \le q \le k-1$），备择假设是大于k-1阶的滞后中至少有一个自相关系数不等于零。在零假设下，$e_{t+k,t}$ 的样本自相关系数 $\hat{\rho} = [\hat{\rho}_{q+1}, \cdots, \hat{\rho}_{q+s}]$ 具有渐近分布 $\sqrt{T}\hat{\rho} \sim N(0,V)$。[2]因此，

$$C = T\hat{\rho}/\hat{V}^{-1}\hat{\rho}$$

在零假设下的渐近分布是 χ_s^2 分布，其中 $\hat{V}$ 为 V 的一致估计量。

也可以对Dufour（1981）的无分布非参数检验进行调整，为大于k-1阶的序列相关提供有限样本边界检验。如前所述，将预测误差分成k个序列，在（k-1）阶相关的零假设下，每个序列都是序列独立的。那么，对每个序列，取 $z_{k,t} = e_{t+k,t}e_{t+2k,t+k}$，如果有一个或更多个子集的检验统计量在 α/k 水平上被拒绝，那么零假设在以 α 为上界的显著性水平上被拒绝。

d. 向前 k 步最优预测误差的方差随 k 非减

向前 k 步预测误差的方差 $\sigma_k^2 = \text{var}(e_{t+k,t}) = \sigma^2(\sum_{i=1}^{k-1} b_i^2)$ 随 k 非减。因此，有必要简单检查一下样本的向前 k 步预测误差的方差和 k 的函数关系。两者均要满足的条件以及观察预测误差的方差随 k 增长的模式，经常可以带来有用的信息。[3]只要注意允许不同时期的样本方差相关，也可

[2] s 是使用者选择的滞后截止阶数。
[3] Diebold 和 Lindner（1995）在非平稳长期-记忆的环境下，拓展了这一思路。

行进行正式的统计推断。

评价对于信息集的最优性

最优预测误差的重要性质，是在做预测时，预测误差在可得信息集的基础上是不可预测的（包括上述分类的特性在内）。所有其它性质都来源于这一性质。无论关注的是线性投影最优还是条件均值最优，无论有关的损失函数是否是二次型，无论预测序列是否平稳，这个说法都是正确的。

依据Brown和Maital（1981），区别局部最优（partial optimality）和完全最优（full optimality）是有益的。局部最优指在可得信息集Ω_t的一些子集而不是所有子集上，预测误差是不可预测的。例如，局部最优的特征是，一个预测就用来形成预测的信息而言是最优的，但是使用的这些信息不是本来*可以*用来形成预测的全部信息。因此，在相比较的预测中，如果就它自己的信息集而言每一个都是最优的，那么每一个预测都可能有局部最优特性。

可以通过$e_{t+k,t} = \alpha' x_t + u_t$形式的回归检验局部最优，其中$x_t \subset \Omega_t$。像在 Mincer 和 Zarnowitz（1969）一文中一样，关于$\hat{y}_{t+k,t}$局部最优检验的特例很受关注。有关的回归为$e_{t+k,t} = \alpha_0 + \alpha_1 \hat{y}_{t+k,t} + u_t$或$y_{t+k} = \beta_0 + \beta_1 \hat{y}_{t+k,t} + u_t$，其中局部最优对应于$(\alpha_0, \alpha_1) = (0,0)$或$(\beta_0, \beta_1) = (0,1)$。[4]也可对回归进行扩展，允许有不同类型的非线性。例如，依据 Ramsey（1969），可以检验在回归$e_{t+k,t} = \sum_{j=0}^{J} \alpha_j \hat{y}_{t+k,t}^{j} + u_t$中，是否所有系数均为零。

与局部最优不同，*完全*最优要求预测误差在进行预测时所有可得信息（即Ω_t的全部）的基础上不可预测。从概念上说，可以通过$e_{t+k,t} = \alpha' x_t + u_t$形式的回归检验完全最优。[*]如果对所有的$x_t \subseteq \Omega_t$，$\alpha = 0$，那么预测是完全最优的。在实践中，永远无法检验完全最优，只能检验信息集不断增大时的局部最优。

也可用无分布的非参数方法检验对于不同信息集的最优性。例如，象 Campbell 和 Dufour（1991，1995）所提出的，很容易对符号检验和符号-秩检验进行修改，用来检验预测误差与可得信息之间的正交性。例如，如果$e_{t+1,t}$是独立于$x_t \in \Omega_t$的线性投影，那么$\text{cov}(e_{t+1,t}, x_t) = 0$。因此，在对称的情况下，可以使用符号-秩检验来验证是否$E[z_t] = E[e_{t+1,t}, x_t] = 0$，或者更一般地，可以用符号检验来验证中位数（$z_t$）=中位数$(e_{t+1,t}, x_t)$=0。[5]相关的符号统计量和符号-秩统计量是$S_{\perp} = \sum_{t=1}^{T} I_+(z_t)$和$W_{\perp} = \sum_{t=1}^{T} I_+(z_t) Rank(|z_t|)$。此外，可以通过取$z_t = e_{t+1,t} g(x_t)$，其中

[4] 在这些回归中，向前 1 步预测的扰动项应该是白噪声，但向前多步预测的扰动项可能是序列相关的。

[*] 原文为完全理性 full rationality，疑有误。——译者注。

[5] 应用符号检验或者符号-秩检验来检验z_t的条件是否满足，并不是显而易见的，但是这些条件是满足的，详见 Campbell 和 Dufour（1995）。

g(.)为我们感兴趣的非线性函数，对信息集中的元素进行非线性变换，这对于评价预测是条件均值独立而不是简单的线性投影独立是有价值的。最后，与前面一样，可以将检验推广到向前 k 步预测误差。简单地取 $z_t = e_{t+k,t}g(x_t)$，将 z_t 序列分成通常的 k 个子集，如果子集中的任何一个检验统计量在 α /k 的水平上是显著的，那么在以 α 为上界的显著性水平上拒绝预测误差与可得信息正交这个零假设。[6]

2. 比较多个预测的精度

预测精度指标

事实上，任何人都不可能得到完全最优预测；取而代之，常常出现的情形是对几个预测（它们都是局部最优的）进行比较，或者把它们组合在一起。在衡量预测精度时，关键的目标是损失函数 $L(y_{t+k}, \hat{y}_{t+k,t})$，通常限指 $L(e_{t+k,t})$，这个函数描绘由各对预测值和实际值产生的“损失”、“代价”、或“负效用”的图形。除了损失函数的形状，预测的时间跨度（forecast horizon）（k）也相当重要。损失函数不同且/或时间跨度不同，预测精度的排序可能很不相同。这个问题导致有些人举出理由证明各种“普遍适用的”精度指标有种种优点。例如，Clements 和 Hendry（1993）主张用一种精度指标，用该指标衡量精度时，预测的排序对某些变换是不变的。

然而，合适的损失函数最终取决于待研究问题的状况。如 Diebold（1993）和其他许多人所强调的，预测通常用于特定的决策环境，例如，政府官员的政策决策或市场参与者的交易决策。因此，合适的精度指标是从预测使用者面对的损失函数中产生的。例如，经济学家可能对来自不同预测的利润流（例如 Leitch 和 Tanner，1991，1995；Engle 等，1993）或效用流（例如 McCulloch 和 Rossi，1990； West、 Edison 和 Cho，1993）感兴趣。

尽管如此，让我们讨论一些标准的统计损失函数，因为它们使用广泛，是常用的基准(benchmarks)。通常是对预测误差 $e_{t+k,t} = y_{t+k} - \hat{y}_{t+k,t}$ 或百分误差 $p_{t+k,t} = (y_{t+k} - \hat{y}_{t+k,t}) / y_{t+k}$ 定义精度指标。例如，平均误差（mean error）$ME = \frac{1}{T}\sum_{t=1}^{T} e_{t+k,t}$，以及平均百分误差 $MPE = \frac{1}{T}\sum_{t=1}^{T} p_{t+k,t}$ 是偏倚指标（measures of bias），偏倚指标是精度的一个组成部分。

目前，最常见的总体精度指标是均方误差 $MSE = \frac{1}{T}\sum_{t=1}^{T} e_{t+k,t}^2$，或均方百分误差 $MSPE = \frac{1}{T}\sum_{t=1}^{T} p_{t+k,t}^2$。通常用这些指标的平方根以保持单位，这就产生了均方根误差 $RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_{t+k,t}^2}$，以及均方根百分误差 $RMSPE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} p_{t+k,t}^2}$。相对不是那么流行但还是常用的精度指标是平均绝对误差 $MAE = \frac{1}{T}\sum_{t=1}^{T} \left| e_{t+k,t} \right|$，和平均绝对百分误差

[6] 我们的讨论隐含的假定是 $e_{t+1,t}$ 和 $g(x_t)$ 都是以零为中心的。如果预测是无偏的，那么对 $e_{t+1,t}$ 来说，这个假定是成立的，但是，没有理由认为，对 $g(x_t)$ 来说，这个假定也应该成立。因此，一般来说，是对 $g(x_t)$ - μ_t 进行检验，其中 μ_t 为 $g(x_t)$ 的中心参数，比如均值、中值或趋势倾向。详见 Campbell 和 Dufour（1995）。

$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|p_{t+k,t}\right|$。

将均方误差（MSE）分解成预测误差的方差及预测偏倚的平方之和，可获得一些信息。即，

$$MSE = E\left[(y_{t+k} - \hat{y}_{t+k,t})^2\right] = \mathrm{var}(y_{t+k} - \hat{y}_{t+k,t}) + (E[y_{t+k}] - E[\hat{y}_{t+k,t}])^2,$$

或等价地

$$MSE = \mathrm{var}(y_{t+k}) + \mathrm{var}(\hat{y}_{t+k,t}) - 2\,\mathrm{cov}(y_{t+k}, \hat{y}_{t+k,t}) + (E[y_{t+k}] - E[\hat{y}_{t+k,t}])^2。$$

这一结果清楚地表明均方误差仅取决于实际序列和预测序列联合分布二阶矩的结构。因此，如Murphy和Winkler（1978，1992）所指出的，虽然对 y_{t+k} 和 $\hat{y}_{t+k,t}$ 的联合分布来说，均方误差是一个有用的概括统计量（summary statistic），但是它包含的信息通常比实际联合分布本身所包含的信息要少得多。因此，突出联合分布不同侧面的其他统计量可能也是有用的。当然，最终人们可能想直接致力于估计联合分布。如果样本容量足够大，可以估计相对精度，估计联合分布是可行的。

衡量可预测性

评价预测精度是自然的、并可提供信息。然而，我们得赶紧补充指出，即使是很好的预测，实际值与预测值也可能不相同。举一个极端的例子，注意到零均值白噪声过程的线性最小二乘预测就是零——预测值和实际值的路径看起来很不一样，然而在二次型损失函数下，不存在更好的线性预测。这个例子凸显了可预测性的内在局限性，可预测性取决于被预测的过程；一些过程天生就容易预测，而另一些过程则难以预测。换言之，预测器（forcaster）以之为条件达到最优的信息，有时是很有价值的，有时是没有价值的。

如何量化可预测性的问题立刻出现了。Granger 和 Newbold（1976）按大家熟悉的线性回归的 R^2 的模式，对在二次型损失函数下协方差平稳序列的可预测性，提出了一个很自然的定义

$$G = \frac{\mathrm{var}(\hat{y}_{t+1,t})}{\mathrm{var}(y_{t+1})} = 1 - \frac{\mathrm{var}(e_{t+1,t})}{\mathrm{var}(y_{t+1})}$$

其中，预测值和预测误差都是指最优（即线性最小二乘最优或条件均值最优）预测。

在结束本小节时，我们指出，虽然可预测性指标是有用的概念，但是可预测性的大小是由过程的总体性质及其最优预测决定的，所以可预测性指标不能帮助人们评价实际报告的（可能远非最优的）预测的“优度”。例如，如果 $\hat{y}_{t+1,t}$ 的方差不比协方差平稳序列 y_{t+1} 的方差低很多，那么，可能是预测很糟糕，也可能是序列天生就几乎不可预测，也可能两者兼而有之。

预测精度的统计比较[7]

损失函数一旦确定，了解相比较的预测中哪一个预测的期望损失最小，经常是令人感兴趣的。当然可以根据样本期平均损失的大小对预测排序，但是人们可能希望能够衡量这样的平均损失的抽样变异性（sampling variability）。或是希望能够检验预测 i 和预测 j 之间的期望损失之差为零（即

[7] 本小节很大一部分取自于 Diebold 和 Mariano（1995）。

$E[L(y_{t+k},\hat{y}^i_{t+k,t})]=E[L(y_{t+k},\hat{y}^j_{t+k,t})]$）的假设，备择假设是其中一个预测比另一个更好。

对一个预测集合中每一个预测的期望损失相等的假设，Stekler（1987）提出了一个以秩为基础的检验。[8] 给定 N 个相比较的预测，根据它们的精度在每一次预测时给每一个预测一个秩（最好的预测取秩为 N，次好的取秩 $N-1$，以此类推）。那么将每个预测的秩逐期加总，

$$H^i=\sum_{t=1}^{T} Rank（L(y_{t+k},\hat{y}^i_{t+k})），\ I=1,\ldots,N，$$

形成了拟合优度检验的卡方统计量，

$$H=\sum_{i=1}^{N}\frac{(H^i-NT/2)^2}{NT/2}。$$

在零假设下，$H\sim x^2_{N-1}$。如此处所述，这个检验要求排序是独立于空间和时间的，但是如果排序是时间上（k-1）阶相关，沿用 Bonferroni 边界检验法，可以对检验进行简单的修正。不仅如此，通过利用 Fisher 的随机化原理，可能可以求得该检验的精确形式。[9]

Stekler 以秩为基础的方法的一个局限，是舍弃了不同预测之间期望损失差异*程度*的信息。在许多应用中，人们不仅想知道期望损失之差*是否*不等于零（或比率不等于 1），而且想知道相差*多大*。实际上，人们想知道损失之差的样本均值的抽样分布（或单个损失的样本均值的抽样分布），知道抽样分布除了直接提供信息，也使得人们可以使用 Wald 检验来验证期望损失之差为零这个假设。Diebold 和 Mariano（1995）在 Granger 和 Newbold（1986）以及 Meese 和 Rogoff（1988）这些更早期的工作的基础上，发展了一种检验期望损失之差是否为零的方法，它可以允许预测误差非零均值、非高斯分布、序列相关以及同期相关。

一般而言，损失函数为 $L(y_{t+k},\hat{y}^i_{t+k,t})$。由于在许多应用中，损失函数是预测误差的直接函数，$L(y_{t+k},\hat{y}^i_{t+k,t})=L(e^i_{t+k,t})$，为简化符号起见，我们此后用 $L(e^i_{t+k,t})$ 表示损失函数，然而，我们要注意，某些损失函数（如变动方向预测的损失函数）不能简化成 $L(e^i_{t+k,t})$ 形式的损失函数。[10] 两个等预测精度的零假设为 $E[L(e^i_{t+k,t})]=E[L(e^j_{t+k,t})]$，或 $E[d_t]=0$，其中 $d_t\equiv L(e^i_{t+k,t})-L(e^j_{t+k,t})$ 为损失之差。

如果 d_t 为协方差平稳、短期记忆序列，那么可以用标准的结果推导损失之差的样本均值的渐近分布，

$$\sqrt{T}(\bar{d}-\mu)\overset{a}{\sim}N(0,2\pi f_d(0))，$$

其中 $\bar{d}=1/T\sum_{t=1}^{T}[L(e^i_{t+k,t})-L(e^j_{t+k,t})]$ 是损失之差的样本均值，$f_d(0)=1/2\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau)$ 是损

[8] Stekler 使用的损失函数是均方根误差，但是也可以使用其他损失函数。
[9] 例如，参见 Bradley（1968），第 4 章。
[10] 在这种情况下，应该使用 $L(y_{t+k},\hat{y}^i_{t+k,t})$ 形式的损失函数。

失之差在频率为零时的谱密度，$\gamma_d(\tau)=E[(d_t-\mu)(d_{t-\tau}-\mu)]$ 是损失之差在位移量为 τ 时的自协方差，而 μ 是损失之差的总体均值。$f_d(0)$ 的表达式表明，即使损失之差仅仅是弱序列相关，由于自协方差项的累积，序列相关的修正也可能相当大。在大样本中，用于检验等预测精度的零假设的统计量显然是经过标准化的损失之差的样本均值，

$$B=\frac{\bar{d}}{\sqrt{2\pi\hat{f}_d(0)/T}},$$

其中 $\hat{f}_d(0)$ 为 $f_d(0)$ 的一致估计。

获得预测精度的精确有限样本检验来补充渐近检验是有价值的。和往常一样，可以使用各种符号检验和符号-秩检验。在使用符号检验时，零假设是损失之差的中位数为零，即中位数 $(L(e^i_{t+k,t})-L(e^j_{t+k,t}))=0$。注意，零假设损失之差的中位数为零，与零假设损失的中位数之差为零是不一样的，即中位数$(L(e^i_{t+k,t})-L(e^j_{t+k,t}))\neq$中位数$(L(e^i_{t+k,t}))-$中位数$(L(e^j_{t+k,t}))$。由于这个原因，符号检验的零假设与渐近 Diebold-Mariano 检验的零假设在精神上略有差异，但是尽管如此，符号检验的零假设有直观且有意义的解释：$P(L(e^i_{t+k,t})>L(e^j_{t+k,t}))=P(L(e^i_{t+k,t})<L(e^j_{t+k,t}))$。

在使用 Wilcoxon 符号-秩检验时，零假设是损失之差序列关于零这个中位数（从而也是均值）对称，这个零假设准确地对应于渐近 Diebold-Mariano 检验的零假设。例如，如果将 $L(e^i_{t+k,t})$ 和 $L(e^j_{t+k,t})$ 的位置互换，分布是一样的，那么损失之差是关于零对称的。损失之差是否关于零对称说到底是一个经验性的问题，可用标准的方法进行评估。

无分布的非参数检验统计量是直观的，构造简单明了。符号检验统计量是 $S_B=\sum_{t=1}^{T}I_+(d_t)$，而符号-秩检验统计量是 $W_B=\sum_{t=1}^{T}I_+(d_t)Rank(|d_t|)$。与前面一样，可以用 Bonferroni 边界法来处理序列相关。有趣的是，注意在比较多步预测时，用 Engle 和 Kozicki（1993）的话说，预测误差序列相关可能是“共同特性”，因为预测误差序列相关在很大程度上是由于预测的时间跨度比数据抽样的时间间隔更长这一事实产生的，因此，即使在预测误差中有序列相关，在损失之差中也可能没有序列相关。当然可以对这种可能性进行实证检验。

West（1994）采用了一个与 Diebold 和 Mariano 的方法关系非常密切、然而又不尽相同的方法。主要的区别是 West 假设预测值是用估计出来的回归模型计算的，并且在这样的框架中直接考虑了参数不确定性的影响。当估计的样本较小时，两种检验方法会产生不同的结果。然而，随着估计期相对于预测期的长度不断增长，参数不确定性的影响消失，Diebold-Mariano 统计量和 West 统计量是相同的。

West 的方法比起 Diebold-Mariano 的方法既更广义又更狭义。更广义在于它修正了由于不断更新参数估计而导致的非平稳，更狭义在于这些修正是在比 Diebold-Mariano 的框架更严格的框架限制中进行的，在 Diebold-Mariano 的框架中，不需要对我们通常不知道或不完全知道的、构成预测基础的模型作任何假定。

在结束本小节时，我们指出，将预测的精度与“朴素（naive）”预测的精度进行比较，有时可以提供一些信息。Theil（1961）的 U 统计量实现了这样一种简单而常见的比较。U 统计量是某一给定预测的向前 1 步预测的均方误差与随机游走预测 $\hat{y}_{t+1,t} = y_t$ 的均方误差之比，即，

$$U = \frac{\sum_{t=1}^{T}(y_{t+1} - \hat{y}_{t+1,t})^2}{\sum_{t=1}^{T}(y_{t+1} - y_t)^2}。$$

可以直接推广到其它损失函数或其他时间跨度。可以用刚刚描述过的方法确定构成 U 统计量基础的、均方误差之比的统计显著性。当然，必须记住，特别是对许多经济变量和金融变量而言，随机游走预测未必“朴素”，所以 U 统计量的值接近 1，未必是“坏事”。包括 Armstrong 和 Fildes（1995）在内的一些作者，提倡用 U 统计量以及与 U 统计量密切相关的其他统计量来比较不同预测方法在预测不同序列时的精度。

3. 将预测组合在一起

在比较预测精度时，人们会问，就特定的损失函数而言，哪一个预测最好。然而，不管一个预测是否“最好”，关于能否将相比较的预测以类似于构建资产组合的模式有效地组合起来，形成一个优于所有原始预测的复合预测，仍然是一个问题。因此，预测组合，尽管明显与预测精度比较有关，在逻辑上是不同的，有其独立的意义。

预测包容检验（forecast encompassing test）

预测包容检验使我们可以确定某一预测是否包含（或涵盖）了相比较的预测所包含的所有相关信息。这个想法至少可以追溯到 Nelson（1972）以及 Cooper 和 Nelson（1975），而 Chong 和 Hendry（1986）将这个想法正式化并加以扩展。为简化起见，我们集中关注有两个预测 $\hat{y}^1_{t+k,t}$ 和 $\hat{y}^2_{t+k,t}$ 时的情况。考虑回归

$$y_{t+k} = \beta_0 + \beta_1 \hat{y}^1_{t+k,t} + \beta_2 \hat{y}^2_{t+k,t} + \varepsilon_{t+k,t}。$$

如果 $(\beta_0, \beta_1, \beta_2) = (0,1,0)$，我们说模型 1“预测包容”（forecast-encompass）模型 2，而如果 $(\beta_0, \beta_1, \beta_2) = (0,0,1)$，我们说模型 2 预测包容模型 1。对 $(\beta_0, \beta_1, \beta_2)$ 的任何其他值，两个模型不互相包容，两个预测都包含关于 y_{t+k} 的有用信息。在特定条件下，可用标准的方法检验包容假设。[11] 此外，尽管这种检验看来还没有出现在预测文献中，但是将前面讨论过的无分布的检验进行简单的推广，求解包容假设的精确有限样本检验（或是当 k>1 时的边界检验）是简单明了的。

Fair 和 Shiller（1989，1990）提出了一种与上述方法不同但与之相关的方法，它基于回归

$$y_{t+k} - y_t = \beta_0 + \beta_1(\hat{y}^1_{t+k,t} - y_t) + \beta_2(\hat{y}^2_{t+k,t} - y_t) + \varepsilon_{t+k,t}。$$

[11] 注意如果 k>1，$\varepsilon_{t+k,t}$ 通常会有 MA(k-1)序列相关。

与前面一样，预测包容对应于系数值（0，1，0）或（0，0，1）。在预测包容的零假设下，Chong-Hendry 和 Fair-Shiller 的回归是相同的。然而，当预测的包容是单整的，可以证明 Fair-Shiller 的框架更方便，因为，Fair-Shiller 的框架设定用的是变化量，便于使用高斯渐近分布理论。

预测组合

一个模型的预测不能包容其他模型的预测，表明所有被考察的模型都是错误设定的。在实践中，这种情况很常见，并不奇怪，因为所有的预测模型当然是错误设定的——它们总是有目的地对复杂得多的现实进行抽象。那么，预测组合技术的作用是什么？在一个信息集可以即刻且无成本地合并在一起的世界里，预测组合技术没什么用；合并信息集总是比预测好。在长期，有时通过改进模型的设定，可以将信息集合并在一起。但是在短期——特别是当必须在给定的时限之前做出及时的预测——合并信息集常常要么是不可能的，要么代价实在高。这个简单的见解推动了关于预测组合的一个务实的想法，这个想法认为，由于假定不能合并信息集，分析的基本对象是预测而不是模型。因此，可将预测组合看作连接短期的、实时（real time）的预测的产生过程和更长期的、正在进行的模型发展过程的重要纽带。

人们已经提出了许多组合方法，可以将这些组合方法粗略地分成两类，“方差-协方差”法和“基于回归”法。我们先考虑 Bates 和 Granger（1969）提出的方差-协方差法。假设我们有两个无偏预测，用这两个无偏预测构建一个组合预测 [12]

$$\hat{y}^{c}_{t+k,t} = \omega\hat{y}^{1}_{t+k,t} + (1-\omega)\hat{y}^{2}_{t+k,t}。$$

因为权重和为 1，复合预测必然是无偏的。此外，被组合在一起的预测的误差将满足和被组合在一起的预测一样的关系，即，

$$e^{c}_{t+k,t} = \omega e^{1}_{t+k,t} + (1-\omega)e^{2}_{t+k,t},$$

其方差为 $\sigma_c^2 = \omega^2\sigma_{11}^2 + (1-\omega)^2\sigma_{22}^2 + 2\omega(1-\omega)\sigma_{12}$，其中 σ_{11}^2 和 σ_{22}^2 为无条件预测误差的方差，而 σ_{12} 是它们的协方差。使组合预测误差的方差最小化（而且由无偏性，组合预测误差的均方误差也最小化了）的权重为

$$\omega^{*} = \frac{\sigma_{22}^2 - \sigma_{12}}{\sigma_{22}^2 + \sigma_{11}^2 - 2\sigma_{12}}。$$

注意最优权重由基础方差和基础协方差（underlying variance and covariance）共同决定。而且，很容易证明，除一个预测包容另一个预测的情形以外，最优复合预测误差的方差小于 $(\sigma_{11}^2, \sigma_{22}^2)$ 的最小值。因此，从总体上看，将预测组合在一起没损失什么，而且有可能很有好处。

在实践中，我们用一致估计代替最优组合权重中的未知方差和协方差；即，我们用 $\hat{\sigma}_{ij} = 1/T\sum_{t=1}^{T} e^{i}_{t+k,t}e^{j}_{t+k,t}$ 代替 σ_{ij} 估计 ω^{*}，得到

$$\omega^{*} = \frac{\hat{\sigma}_{22}^2 - \hat{\sigma}_{12}}{\hat{\sigma}_{22}^2 + \hat{\sigma}_{11}^2 - 2\hat{\sigma}_{12}}。$$

在通常可以取得的样本容量的有限样本中，抽样误差会损害组合权重的估计，原始预测之间经常

[12] 如 Newbold 及 Granger（1974）所证明的那样，推广到 M>2 个相比较的无偏预测是简单明了的。

有共线性，使抽样误差损害组合权重估计的问题更加严重。因此，虽然我们希望通过将预测组合在一起以减少样本外预测的均方误差，但是我们不能保证一定能做到。然而，就像 Clemen（1989）在回顾大量预测组合的文献时所发现的那样，在实践中，预测组合技术常常表现非常好。

现在考虑预测组合的“回归法”。Chong-Hendry 和 Fair-Shiller 的包容回归的形式直接表明，可以通过简单地将实际值对预测值回归来将预测组合在一起。Granger 和 Ramanathan（1984）指出，最优方差-协方差组合权重向量可以解释成受两个约束（权重和为 1 以及不包括截距项）时 y_{t+k} 的线性预测对预测值回归的系数向量。当然，在实践中，我们只是对可取得的数据进行回归。

一般而言，回归法是简单灵活的。因为任何“回归工具”都有使用的可能性，回归法有许多变形和扩展。关键是合理目标下的推广。我们给出四个例子：随时间变化的组合权重、动态组合回归、组合权重接近等权的贝叶斯退化（Bayesian shrinkage）、非线性组合回归。

a. 随时间变化的组合权重

Granger 和 Newbold（1973）在方差-协方差的背景下、Diebold 和 Pauly（1987）在回归的背景下提出了随时间变化的组合权重。例如，在回归框架中，人们可以对组合回归进行加权估计或滚动估计，或者估计参数会随时间变化的组合回归。

应该用随时间变化权重的原因有很多。首先，不同的学习速度导致某个预测相对于其它预测随时间不断改进，在这种情况下，人们自然想逐渐增加这个准确性正在不断提高的预测的权重。其次，不同预测模型的设计可能使这些预测方法在某些环境中相对于在其它环境而言是更好的预测工具。例如，在高通胀时期，工资-价格部分高度发达的结构模型可能大大优于较简单的模型。在这样的时期，更复杂成熟的模型将得到更高的权重。再次，代理人决策规则中的参数可能随时间漂移，某些预测技术可能相对而言更容易受到这种漂移的影响。

b. 动态组合回归

组合回归产生误差的序列相关是很自然的。Diebold（1988）考虑了协方差平稳时的情形，认为当 $\beta_1 + \beta_2 \neq 1$ 时，在无约束的预测组合回归中，可能会有序列相关。更一般地，在组合回归中允许序列相关以捕捉其它预测没有捕捉到的被预测变量的动态特性，可能是个好主意。在这方面，遵从 Hendry 和 Mizon（1978），Coulson 和 Robins（1993）指出，扰动项序列相关的组合回归是他们所提倡的、包括滞后因变量和滞后预测值的组合回归的特例。

c. 组合权重接近等权的贝叶斯退化

经常可以发现，各预测的简单算术平均即使相对于各预测的“最优”复合预测的表现也非常好。[13] 显然，施加等权重约束消除了估计权重间的变异，代价是有可能产生偏倚。然而，有证据表明，在二次型损失函数下，施加等权重约束的好处经常超过这种代价。了解到这一点，Clemen 和 Winker（1986）以及 Diebold 和 Pauly（1990）提出了贝叶斯退化技术，允许在估计组合权重时不同程度地结合先验信息；这样一来，最小二乘权重和先验权重就成为后验均值组合权重的极端特例。实际的后验均值组合权重是这两个极端权重矩阵的加权平均。例如，当使用自然共轭正态-伽马分布（natural conjugate normal-gamma prior）时，后验均值组合权重向量为

$$\beta^{\text{后验}} = (Q + F'F)^{-1}(Q\beta^{\text{先验}} + F'F\beta),$$

其中 $\beta^{\text{先验}}$ 为先验均值向量，Q 为先验精度矩阵，F 为组合回归的设计矩阵（design matrix），而 $\hat{\beta}$

[13] 参见 Winkler 和 Makridakis（1983），Cleman（1989），以及这两篇文章的许多参考文献。

为最小二乘组合权重向量。退化明显是朝向一个中心趋势指标（例如算术平均）。在这个过程中，组合权重逐渐向算术均值靠拢，但是当（或如果）数据有某些问题时，数据仍然会有所表现。

d. 非线性组合回归

当然，没有理由一定要把组合回归处理成线性回归，各种常见的非线性回归也是可以的。Deutsch、Granger 和 Terä svirta（1994）提出了一个特别有趣的可能性，他们建议

$$\hat{y}^{c}_{t+k,t} = I(s_t = 1)(\beta_{11}\hat{y}^{1}_{t+k,t} + \beta_{12}\hat{y}^{2}_{t+k,t}) + I(s_t = 2)(\beta_{21}\hat{y}^{1}_{t+k,t} + \beta_{22}\hat{y}^{2}_{t+k,t})。$$

支配组合权重的状态，可以取决于一个或两个模型过去的预测误差或是取决于各种经济变量。此外，指标权重不一定仅仅是一个二元变量；通过允许权重是预测误差或经济变量的函数，状态间的转换会变得更平缓。

4. 评价经济预测和金融预测的几个专题

评价变动方向预测

在经济和金融决策中，经常要用到变动方向预测（例如 Leitch 和 Tanner，1991，1995； Satchell 和 Timmermann，1992）。如何评价这种预测的问题也立即出现了。经过适当的修改，前面讲到的预测精度比较检验的结果仍然是成立的，所以就不在此重复了。取而代之，我们注意到人们经常看到对变动方向预测是否“有价值”的评价，我们应该讨论一下这个问题。

关于变动方向预测是否“有价值”的问题必然涉及到与朴素基准的比较——变动方向预测与“朴素的”抛硬币（成功的概率等于对应的边际概率）相比较。考虑一张 2×2 的或然事件表（contingency table）。为了简化记号，将预测值和实际值落入的两种状态称为 “i”和“j”。例如，通常 i=“上”而 j=“下”。表 1 和 2 清晰地显示出有关观测到的单元计数和未观测到的单元概率的记号。变动方向预测没有价值的零假设是预测值和实际值相互独立，在这种情况下，$P_{ij} = P_{i.}P_{.j}, \forall i, j$。如往常一样，在零假设下进行假设检验。实际的单元概率当然是未知的，所以我们使用一致估计 $\hat{P}_{i.} = O_{i.}/O$ 和 $\hat{P}_{.j} = O_{.j}/O$。那么在零假设下，可以用 $\hat{E}_{ij} = \hat{P}_{i.}\hat{P}_{.j}O = O_{i.}O_{.j}/O$ 求得单元计数的期望 $E_{ij} = P_{i.}P_{.j}O$ 的一致估计。最后，构造统计量 $C = \sum_{i,j=1}^{2}(O_{ij} - \hat{E}_{ij})^2/\hat{E}_{ij}$。在零假设下，$C \xrightarrow{d} \chi^2_1$。

Merton（1981）以及 Henriksson 和 Merton（1981）提出了一个与预测的价值密切相关的检验。他们断言，如果 $P_{ii}/P_{i.} + P_{jj}/P_{j.} > 1$，那么预测是有价值的。因此，他们对零假设 $P_{ii}/P_{i.} + P_{jj}/P_{j.} = 1$，备择假设为 $P_{ii}/P_{i.} + P_{jj}/P_{j.} \neq 1$ 给出了一个精确检验。Schnader 和 Stekler（1990）以及 Stekler（1994）不同程度地提到，由 Pesaran 和 Timmermann（1992）将其正式化的一个关键的见解，是如果边际概率固定在观测到的相对频率 $O_{i.}/O$ 和 $O_{.j}/O$，Henriksson-Merton 的零假设和或然事件表的零假设是等价的。在推导 Henriksson-Merton 检验统计量的精确有限样本分布时，需要同样令人不惬意的假定。

然而，在渐近条件下，所有的方法都好；Henriksson-Merton 统计量的平方，经过适当的正规

化，渐近地等价于 C——或然事件表的卡方统计量。此外，2×2 或然表检验很容易推广至 N×N 的情形，有

$$C_N = \sum_{i,j=1}^{N} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}。$$

表 1
观测到的单元计数

	实际值 i	实际值 j	边际
预测值 i	O_{ii}	O_{ij}	$O_{i.}$
预测值 j	O_{ji}	O_{jj}	$O_{j.}$
边际	$O_{.i}$	$O_{.j}$	总计：O

表 2
未观测到的单元概率

	实际值 i	实际值 j	边际
预测值 i	P_{ii}	P_{ij}	$P_{i.}$
预测值 j	P_{ji}	P_{jj}	$P_{j.}$
边际	$P_{.i}$	$P_{.j}$	总计：P

在零假设下，$C_N \overset{a}{\sim} \chi^2_{(N-1)(N-1)}$。如 Pesaran 和 Timmermann（1992）所指出的，这里出现了一个微妙问题。在 2×2 情形中，我们必须依据整个表检验，因为非对角线的元素由对角线上的元素决定，每一行的两个元素之和必须为 1。在 N×N 的情形中，与 2×2 时不同，对检验哪个单元有更大范围的选择，就预测评价的目的而言，可能最好是只关注对角线上的单元。

在结束本小节时，我们指出，尽管在变动方向预测中我们经常关注或然表检验（由于同样的原因，在更标准的预测中，我们经常关注基于 Theil 的 U 统计量的检验），在那个意义上的预测“价值”，既不是能够形成一个可获得超额收益的交易策略意义上的预测价值的充分条件，也不是必要条件。例如，一个预测可能优于边际预测，但是在扣除了交易成本之后，仍然不能获得超额收益。另一方面，一个预测可能比边际预测更差，但是如果预测准确，仍然可能获得巨额利润，Cumby 和 Modest（1987）强调了这一点。

评价概率预测

许多时候，经济预测和金融预测的问题是各种概率，比如明年出现商业周期转折点的概率，今年公司会在某个债券发行上违约的概率，或是今年 S&P500 股票指数的收益率会超过 10%的概率。在评价概率预测时，有许多特殊的规则，现在开始讨论这些特殊规则。令 $P_{t+k,t}$ 是在 t 时对 t+k

时的事件作的概率预测，并令当事件发生时 $R_{t+k}=1$，否则为 0。如果只有两个可能事件，$P_{t+k,t}$ 为标量。更一般地，如果有 N 个可能事件，那么 $P_{t+k,t}$ 为（N-1）×1 向量。[14] 为简化符号，我们将集中关注标量概率预测。

概率预测的精度指标一般称为“得分（score）”，最常见的得分是 Brier（1950）的二次概率得分，也叫 Brier 得分，

$$QPS=\frac{1}{T}\sum_{t=1}^{T}2(P_{t+k,t}-R_{t+k})^2 \text{ 。}$$

显然，$QPS\in[0,2]$而且有负指向（negative orientation）（更小的值意味着预测更精确）。[15] 为了理解 QPS，注意任何预测的精度指的是使用该预测时的期望损失，而损失通常取决于预测值和实际值间的差异。那么，在二次型损失函数下的概率预测，追踪 $P_{t+k,t}$ 和 R_{t+k} 之差的平方的平均数，看起来是合理的，这正是 QPS 所表达的。因此，QPS 大致相当于概率预测中的均方误差。

然而，QPS 只是*大致*相当于均方误差，因为 $P_{t+k,t}$ 实际上不是对*结果*（0—1）的*预测*，而是分派给结果的概率。评价概率预测的一个更自然和直接的方法就是直接将预测概率与观测到的相对频率作比较，即评估误差量度值（calibration）。误差量度值的一个全面衡量是整体平方偏误，

$$GSB=2(\overline{P}-\overline{R})^2$$

其中 $\overline{P}=1/T\sum_{t=1}^{T}P_{t+k,t}$，$\overline{R}=1/T\sum_{t=1}^{T}R_{t+k}$。$GSB\in[0,2]$有负指向。

也可以在单位区间的任意子集内计算局部误差量度值。例如，可以检查与预测的概率在 0.6 和 0.7 之间相对应的实际观测到的相对频率是否也在 0.6 和 0.7 之间。进一步地，可以根据使用者的兴趣和具体情况，将单位区间分割成 J 个子集，计算所有 J 个子集的局部误差量度值的加权平均数。[16] 这样算出来的指标就是局部平方偏误，

$$LSB=\frac{1}{T}\sum_{j=1}^{J}2T_j(\overline{P}_j-\overline{R}_j)^2 \text{ ，}$$

其中 T_j 是在 j 集合中概率预测的数目，$\overline{P}_j$ 是在 j 集合中的概率预测值的均值，而 $\overline{R}_j$ 是在 j 集合中的实际值的均值，j=1，…，J。注意 $LSB\in[0,2]$，而且 LSB=0 意味着 GSB=0，但反之不成立。

至少在实际值相互独立的条件下，检验量度值是否足够好是简单明了的。对某一给定的事件和对应的概率预测序列 $\{P_{t+k,t}\}_{t=1}^{T}$，构造预测值的 J 个互斥且全体无遗漏的（collectively exhaustive）子集，用 π_j，j=1，…，J 代表每个范围的中点。令 R_j 分别代表当预测在集合 j 中时观测到的事

[14] 概率之和为 1 的约束隐含确定了分派给第 N 个事件的概率预测。
[15] 在全向量（full vector）时，QPS 公式中的“2”是人为加上去的。我们当然可以去掉 2，不会影响相比较的预测的 QPS 排序，但是为了能够与其他文献相比较，我们没有去掉 2。
[16] 例如，Diebold 和 Rudebusch（1989）将单位区间分成十等份。

件数目，则“范围 j”的误差量度值统计量的定义是

$$Z_j = \frac{(R_j - T_j\pi_j)}{(T_j\pi_j(1-\pi_j))^{1/2}} \equiv \frac{(R_j - e_j)}{w_j^{1/2}}, j = 1,\ldots,J,$$

而整体误差量度值统计量

$$Z_0 = \frac{(R_+ - e_+)}{w_+^{1/2}},$$

其中 $R_+ = \sum_{j=1}^{J} R_j$，$e_+ = \sum_{j=1}^{J} T_j\pi_j$，$w_+ = \sum_{j=1}^{J} T_j\pi_j(1-\pi_j)$。$Z_0$ 是所有单元的局部量度值都足够好的联合检验，而 Z_j 统计量是逐个单元检验其局部误差量度值。[17] 在相互独立的条件下，二项式结构显然意味着 $Z_0 \overset{a}{\sim} N(0,1)$ 和 $Z_j \overset{a}{\sim} N(0,1), \forall j = 1,\ldots,J$。在一个迷人的扩展中，Seillier-Moiseiwitsch 和 Dawid（1993）证明 Z_0 和 Z_j 渐近正态性的成立更具一般性，包括在实际应用时放宽相互独立的条件，有序列相关的情形。

概率预测（或更精确地说，是概率预测对应的实际值）另一个令人感兴趣的特征，叫决断（resolution），

$$RES = \frac{1}{T}\sum_{j=1}^{J} 2T_j(\bar{R}_j - \bar{R})^2$$

RES 就是 $\bar{R}$ 和 $\bar{R}_j$ 之差平方的加权平均，是反映观测到的相对频率在不同单元间变动的指标。$RES \geq 0$ 且有正指向。如 Murphy（1973）所证明的，QPS 存在一个能提供一些信息的分解，

$$QPS = QPS_{\bar{R}} + LSB - RES$$

其中 $QPS_{\bar{R}}$ 是 QPS 在 $P_{t+k,t} = \bar{R}$ 时的值。这种分解凸显了概率预测的不同特性间的权衡。

就像在“标准”预测时 Theil 的 U 统计量一样，将某个概率预测的表现与基准的表现进行比较，有时能提供一些信息。例如，Murphy（1974）提出用统计量

$$M = QPS - QPS_{\bar{R}} = LSB - RES$$

衡量手头的预测和基准预测 $\bar{R}$ 在预测精度方面的差异。使用前面讨论过的 Diebold-Mariano 方法，我们也可以评价 QPS 和 $QPS_{\bar{R}}$ 差异的显著性、不同预测间的 QPS 或其他概率预测精度指标的差异的显著性、或是不同预测的局部或整体量度值差异的显著性。

评价波动率预测

在金融学中，许多有趣的问题，比如期权定价、风险对冲和组合管理，清楚明确地取决于资产价格的方差。因此，人们提出了各种方法来进行波动率预测。与点预测或概率预测相比，由于

[17] 我们当然可以通过将单位区间分割成单位区间本身来检验整体量度值是否足够好。

实际条件方差是*不可观测*的，波动率预测的评价变得复杂了。

解决这个不可观测问题的标准“方案”是用实际值的平方 ε_{t+k}^2 作为真实条件方差 h_{t+k} 的代理变量，因为 $E[\varepsilon_{t+k}^2 \mid \Omega_{t+k-1}] = E[h_{t+k} v_{t+k}^2 \mid \Omega_{t+k-1}] = h_{t+k}$，其中 $v_{t+k} \sim WN(0,1)$。[18] 因此，例如，有 $MSE = 1/T \sum_{t=1}^{T} (\varepsilon_{t+k}^2 - \hat{h}_{t+k,t})^2$。尽管人们经常用均方误差衡量波动率预测的精度，但 Bollerslev、Engle 和 Nelson（1994）指出，由于均方误差对正的波动率预测和负的波动率预测（负的波动率是无意义的）的惩罚是对称的，用均方误差来衡量波动率预测的精度是不合适的。对波动率的惩罚不对称的两个可供选择的损失函数，是 Pagan 和 Schwert（1990）采用的对数损失函数，

$$LL = \frac{1}{T} \sum_{t=1}^{T} [\ln(\varepsilon_{t+k}^2) - \ln(\hat{h}_{t+k,t})]^2 ,$$

以及 Bollerslev 和 Ghysels（1994）采用的对异方差进行过调整的均方误差，

$$HMSE = \frac{1}{T} \sum_{t=1}^{T} [\frac{\varepsilon_{t+k}^2}{\hat{h}_{t+k,t}} - 1]^2 。$$

Bollerslev、Engle 和 Nelson（1994）表明，高斯拟最大似然函数中的损失函数经常用于拟合波动率模型；即，

$$GMLE = \frac{1}{T} \sum_{t=1}^{T} [\ln(\hat{h}_{t+k,t}) + \frac{\varepsilon_{t+k}^2}{\hat{h}_{t+k,t}}] 。$$

正如所有的预测评价一样，预测使用者最感兴趣的波动率预测评价是在有关的损失函数下进行的。West、Edison 和 Cho（1993）以及 Engle 等（1993）在这方面做出了重要贡献，分别提出了基于效用最大化和利润最大化的经济损失函数。Lopez（1995）提出了一个可以使用各种经济损失函数的波动率预测评估框架。这个框架的基础是通过对 ε_t 假定的或估计的分布积分，将波动率预测转化为概率预测。通过选择与所关注的事件相对应的积分区间，预测使用者可以将他的损失函数的各项元素整合到概率预测中。

例如，给定 $\varepsilon_{t+k} \mid \Omega_t \sim D(0, h_{t+k,t})$ 以及波动率预测 $\hat{h}_{t+k,t}$，对事件 $\varepsilon_{t+k} \in [L_{\varepsilon,t+k}, U_{\varepsilon,t+k}]$ 感兴趣的期权交易者会形成概率预测

$$\begin{aligned} P_{t+k,t} &= \Pr(L_{\varepsilon,t+k} < \varepsilon_{t+k} < U_{\varepsilon,t+k}) \\ &= \Pr(\frac{L_{\varepsilon,t+k}}{\sqrt{\hat{h}_{t+k,t}}} < z_{t+k} < \frac{U_{\varepsilon,t+k}}{\sqrt{\hat{h}_{t+k,t}}}) \\ &= \int_{l_{\varepsilon,t+k}}^{u_{\varepsilon,t+k}} f(z_{t+k}) dz_{t+k} , \end{aligned}$$

[18] 虽然 ε_{t+k}^2 是 h_{t+k} 的无偏估计量，但它是 h_{t+k} 的不精确或“有噪声”的估计量。例如，如果 $v_{t+k} \sim N(0,1)$，那么因为 $v_{t+k}^2 \sim x_1^2$，$\varepsilon_{t+k}^2 = h_{t+k} v_{t+k}^2$ 的条件均值等于 h_{t+k}。然而，因为 x_1^2 分布的中位数是 0.455，在超过 50% 的时间里 $\varepsilon_{t+k}^2 < 1/2 h_{t+k}$。

其中 z_{t+k} 是经过标准化的新生（innovation），$f(z_{t+k})$ 是 $D(0,1)$ 的函数形式，$[l_{\varepsilon,t+k}, u_{\varepsilon,t+k}]$ 是经过标准化的积分区间。与之相对照，对标的资产（underlying asset）的行为 $y_{t+k} = \mu_{t+k,t} + \varepsilon_{t+k}$，其中 $\mu_{t+k,t} = E[y_{t+k} \mid \Omega_t]$，感兴趣的预测者会形成概率预测

$$
\begin{aligned}
P_{t+k,t} &= \Pr(L_{y,t+k} < y_{t+k} < U_{y,t+k}) \\
&= \Pr\left(\frac{L_{y,t+k} - \hat{\mu}_{t+k,t}}{\sqrt{\hat{h}_{t+k,t}}} < z_{t+k} < \frac{U_{y,t+k} - \hat{\mu}_{t+k,t}}{\sqrt{\hat{h}_{t+k,t}}}\right) \\
&= \int_{l_{y,t+k}}^{u_{y,t+k}} f(z_{t+k}) dz_{t+k}
\end{aligned}
$$

其中 $\hat{\mu}_{t+k,t}$ 是所预测的条件均值，$[l_{y,t+k}, u_{y,t+k}]$ 是经过标准化的积分区间。

这些概率预测一旦产生，可以用上述的得分规则评价，并可以用 Diebold-Mariano 方法检验不同模型差异的显著性。这个框架的主要优点是它使评价建立在可观测事件的基础上，因此无需用 ε_{t+k}^2 替代不可观测的真实方差。

评价波动率预测的 Lopez 方法的基础是把随时间变化的概率分派给一个固定区间。另一种选择是，与传统置信区间的构造一样，我们可以固定概率，让区间的宽度变化。在这方面，Christoffersen（1995）建议研究这个事实：如果（1-α）%置信区间（用 $[L_{y,t+k}, U_{y,t+k}]$ 表示）对误差的量度是正确的，那么

$$E[I_{t+k,t} \mid I_{t,t-k}, I_{t-1,t-k-1}, \cdots, I_{k+1,1}] = (1-\alpha),$$

其中

$$I_{t+k,t} = \begin{cases} 1, & \text{如果 } y_{t+k} \in [L_{y,t+k}, U_{y,t+k}], \\ 0, & \text{其他。} \end{cases}$$

即，Christoffersen 建议检验条件覆盖（conditional coverage）。[19]

对区间预测的标准评价方法常常将注意力限制在无条件覆盖 $E[I_{t+k|t}] = (1-\alpha)$。但是，由于有着正确无条件覆盖的区间预测在任何特定时间仍然可能有不正确的*条件*覆盖，仅仅检验无条件覆盖一般来说是不够的。

对向前 1 步的区间预测（k=1），条件覆盖的判别准则变为

$$E[I_{t+1,t} \mid I_{t,t-1}, I_{t-1,t-2}, \cdots, I_{2,1}] = (1-\alpha),$$

或等价地

$$I_{t+1|t} \overset{i.i.d}{\sim} Bern(1-\alpha)。$$

[19] 一般而言，人们想检验是否 $E[I_{t+k|t} \mid \Omega_t] = (1-\alpha)$，其中 Ω_t 是在 t 时可得的全部信息。就现在的目的来说，为了构建一般且容易使用的检验，Ω_t 仅限于指标序列的过去值。

给定 T 个区间预测的 T 个指示器变量值，我们可以通过检验指示器变量是一个独立同分布的贝努利（Bernoulli）（1-α）随机变量这个假设来确定预测区间是否有正确的条件覆盖。通过比较指示器序列 $\{I_{t+1,t}\}$ 有约束和无约束的马尔可夫过程的对数似然函数，很容易构造独立同分布贝努利假设的似然比检验。无约束的转换概率矩阵为

$$\Pi = \begin{bmatrix} \pi_{11} & 1-\pi_{11} \\ 1-\pi_{00} & \pi_{00} \end{bmatrix},$$

其中 $\pi_{11} = P(I_{t+1|t} = 1 \mid I_{t|t-1} = 1)$，依此类推。在零假设下，转换概率矩阵为 $\begin{bmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \end{bmatrix}$。相应的近似似然函数为

$$L(\Pi \mid I) = (\pi_{11})^{n_{11}} (1-\pi_{11})^{n_{10}} (1-\pi_{00})^{n_{01}} \pi_{00}^{n_{00}}$$

和

$$L(\alpha \mid I) = (1-\alpha)^{(n_{11}+n_{01})} (\alpha)^{(n_{10}+n_{00})}。$$

其中 n_{ij} 为观测到的、从 i 转换到 j 的数目，而 i 为指示器序列。[20] 条件覆盖假设的似然比统计量为

$$LR_{cc} = 2[\ln L(\hat{\Pi} \mid I) - \ln L(\alpha \mid I)],$$

其中 $\hat{\Pi}$ 为最大似然估计。在零假设下，$LR_{cc} \overset{a}{\sim} \chi_2^2$。

条件覆盖的似然比检验可以分解成两个不同的、令人感兴趣的假设：正确的无条件覆盖假设 $E[I_{t+k|t}] = (1-\alpha)$，和独立性假设 $\pi_{11} = 1-\pi_{00}$。（在给定独立性的条件下）对正确的无条件覆盖假设的似然比检验是

$$LR_{uc} = 2[\ln L(\hat{\pi} \mid I) - \ln L(\alpha \mid I)],$$

其中 $L(\pi \mid I) = (1-\pi)^{(n_{11}+n_{01})} (\pi)^{(n_{10}+n_{00})}$。在零假设下，$LR_{uc} \overset{a}{\sim} \chi_1^2$。独立性假设可以单独用以下统计量检验

$$LR_{ind} = 2[\ln L(\hat{\Pi} \mid I) - \ln L(\hat{\pi} \mid I)]。$$

在零假设下，$LR_{ind} \overset{a}{\sim} \chi_1^2$。在小样本和大样本中，很明显都有 $LR_{cc} = LR_{uc} + LR_{ind}$。

当 k=1 时，也可使用 David（1947）的组检验（group test）来验证独立性，组检验是针对一阶相关的精确，并且同样也是最有效的检验。将一串连续的 0 或 1 定义为一个组，并令 r 为序列 $\{I_{t+1,t}\}$ 中的组数。在序列是独立同分布的零假设下，给定 1 的总数 n_1 及 0 的总数 n_0，r 的分布为

[20] 因为去掉了初始项，似然函数是近似的。当然，由于所有似然比检验都是渐近的，所以初始项的处理不会影响结果。

$$P(r \mid n_0, n_1) = \frac{f_r}{\binom{n}{n_0}}，\ 当\ r \geq 2\ 时，$$

其中 $n = n_0 + n_1$，而

$$f_r = \begin{cases} f_{2s} = 2\binom{n_0 - 1}{s-1}\binom{n_1 - 1}{s-1}，\ 当r为偶数时， \\ f_{2s+1} = \dfrac{f_{2s}(n-2s)}{(2s)}，\ 当r为奇数时。 \end{cases}$$

最后，尽管向前 k 步预测误差一般是序列相关的，但在似然比的框架中，将组检验推广到 k>1 是简单的。除了要求 k 阶马尔可夫链，基本的框架保持不变。然而，k 阶链总是可以写成状态空间扩展了的一阶链，因此，可以将一阶时的结果直接类推到 k 阶。

5. 结束语

三个现代主题弥漫在这一综述中，因此值得明确地突出它们。第一个主题是不同类型的预测，比如概率预测和波动率预测，正在进一步整合到经济预测和金融预测中，衍生出对新的预测评价方法的需求。

第二个主题是通常都以无分布的非参数检验为基础的精确有限样本假设检验的使用。在预测误差无偏、k 阶相关、与可得信息的正交以及在有不止一个预测时检验期望损失是否相等以及检验变动方向预测是否有价值等背景下，我们对这样的检验进行了明确扼要的描述。

第三个主题是相关损失函数的使用。在许多场合（比如可预测性的衡量和预测精度的比较检验），都有这个问题，并且可能很容易在其他场合（比如正交性检验、包容检验和组合回归）中引入这个问题。事实上，很快就会有在相关的损失函数下估计、预测以及预测评价（从而选择模型以及进行非嵌套的（nonnested）假设检验）的一体化工具包；参见 Weiss 和 Andersen（1984），Weiss（1995），Diebold 和 Mariano（1995），Christoffersen 和 Diebold（1994，1995）以及 Diebold，Ohanian 和 Berkowitz（1995）。

参考文献

Armstrong, J. S. and R. Fildes (1995). On the selection of error measures for comparisons among forecasting methods. J. Forecasting 14, 67-71.

Auerbach, A. (1994). The U.S. fiscal problem: Where we are, how we got here and where we're going. NBER Macroeconomics Annual, MIT Press, Cambridge, MA.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. Oper. Res. Quart. 20, 451-468.

Bollerslev, T., R. F. Engle and D. B. Nelson (1994). ARCH models. In: R. F. Engle and D. McFadden, eds., Handbook of Econometrics, Vol. 4, North-Holland, Amsterdam.

Bollerslev, T. and E. Ghysels (1994). Periodic autoregressive conditional heteroskedasticity. Working Paper No. 178, Department of Finance, Kellogg School of Management, Northwestern University.

Bonham, C. and R. Cohen (1995). Testing the rationality of price forecasts: Comment. Amer. Eeonora. Rev. 85, 284-289.

Bradley, J. V. (1968). Distributidn-free statistical tests. Prentice Hall, Englewood Cliffs, NJ.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review 75, 1-3.

Brown, B. W. and S. Maital (1981). What do economists know? An empirical study of experts' expectations. Econometrica 49, 491-504.

Campbell, B. and J.-M. Dufour (1991 Over-rejections in rational expectations models: A nonparametric approach to the Mankiw-Shapiro problem. Econom. Lett. 35, 285-290.

Campbell, B. and J.-M. Dufour (1995). Exact nonparametric orthogonality and random walk tests. Rev. Econom. Statist. 77, 1-16.

Campbell, B. and E. Ghysels (1995). Federal budget projections: A nonparametric assessment of bias and efficiency. Rev. Econom. Statist. 77, 17-31.

Campbell, J. Y. and N. G. Mankiw (1987). Are output fluctuations transitory? Quart. J. Econom. 102, 857-880.

Chong, Y. y. and D. F. Hendry (1986). Econometric evaluation of linear macroeconomic models. Rev. Econom. Stud. 53, 671-490.

Christoffersen, P. F. (1995). Predicting uncertainty in the foreign exchange markets. Manuscript, Department of Economics, University of Pennsylvania.

Christoffersen, P. F. and F. X. Diebold (1994). Optimal prediction under asymmetric loss. Technical Working Paper No. 167, National Bureau of Economic Research, Cambridge, MA.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. Internat. J. Forecasting 5, 559-581.

Clemen, R. T. and R. L. Winkler (1986). Combining economic forecasts. J. Econom. Business Statist. 4, 39-46.

Clements, M. P. and D. F. Hendry (1993). On the limitations of comparing mean squared forecast errors. J. Forecasting 12, 617-638.

Cochrane, J. H. (1988). How big is the random walk in GNP? J. Politic. Econom. 96, 893-920.

Cooper, D. M. and C. R. Nelson (1975). The ex-ante prediction performance of the St. Louis and F.R.B.-M.I.T.-Penn econometric models and some results on composite predictors. J. Money, Credit and Banking 7, 1-32.

Coulson, N. E. and R. P. Robins (1993). Forecast combination in a dynamic setting. J. Forecasting 12, 63-67.

Cumby, R. E. and J. Huizinga (1992). Testing the autocorrelation structure of disturbances in ordinary least squares and instrumental variables regressions. Econometrica 60, 185-195.

Cumby, R. E. and D. M. Modest (1987). Testing for market timing ability: A framework for forecast evaluation. J. Financ. Econom. 19, 169-189.

David, F. N. (1947). A power function for tests of randomness in a sequence of alternatives. Biometrika 34, 335-339.

Deutsch, M., C. W. J. Granger and T. Tersvirta (1994). The combination of forecasts using changing weights, lnternat. J. Forecasting 10, 47-57.

Diebold, F. X. (1988). Serial correlation and the combination of forecasts. J. Business Econom. Statist. 6, 105-111.

Diebold, F. X. (1993). On the limitations of comparing mean square forecast errors: Comment. J. Forecasting 12, 641-642.

Diebold, F. X. and P. Lindner (1995). Fractional integration and interval prediction. Econom. Lett., to appear.

Diebold, F. X. and R. Mariano (1995). Comparing predictive accuracy. J. Business Econom. Statist. 13, 253-264.

Diebold, F. X. L. Ohanian and J. Berkowitz (1995). Dynamic equilibrium economies: A framework for comparing models and data. Technical Working Paper No. 174, National Bureau of Economic Research, Cambridge, MA.

Diebold, F. X. and P. Pauly (1987). Structural change and the combination of forecasts. J. Forecasting 6, 21-40.

Diebold, F. X. and P. Pauly (1990). The use of prior information in forecast combination. Internat. J. Forecasting 6, 503-508.

Diebold, F. X. and G. D. Rudebusch (1989). Scoring the leading indicators. J. Business 62, 369-391.

Dufour, J.-M. (1981). Rank tests for serial dependence. J. Time Ser. Anal. 2, 117-128.

Engle, R. F., C.-H. Hong A. Kane and J. Noh (1993). Arbitrage valuation of variance forecasts with simulated options. In: D. Chance and R. Tripp, eds., Advances in Futures and Options Research, JlA Press, Greenwich, CT.

Engle, R. F. and S. Kozicki (1993). Testing for common features. J. Business Econom. Statist. 11, 369-395.

Fair, R. C. and R. J. Shiller (1989). The informational content of ex-ante forecasts. Rev. Econom. Statist. 71, 325-331.

Fair, R. C. and R. J. Shiller (1990). Comparing information in forecasts from econometric models. Amer. Econom. Rev. 80, 375-389.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. J. Finance 25, 383-417.

Fama, E. F. (1975). Short-term interest rates as predictors of inflation, Amer. Econom. Rev. 65, 269-282.

Fama, E. F. (1991). Efficient markets II. J. Finance 46, 1575-1617.

Fnma, E. F. and K. R. French (1988). Permanent and temporary components of stock prices. J. Politic. Econom. 96, 246-273.

Granger, C. W. J. and P. Newbold(1973). Some comments on the evaluation of economic forecasts. Appl. Econom. 5, 35-47.

Granger, C. W. J. and P. Newbold (1976). Forecasting transformed series. J. Roy. Statist. Soc. B 38, 189-203.

Granger, C. W. J. and P. Newbold (1986). Forecasting economic time series. 2nd ed., Academic Press, San Diego.

Granger, C. W. J. and R. Ramanathan (1984). Improved methods of forecasting. J, Forecasting 3, 197-204.

Hansen, L. P. and R. J. Hodrick (1980). Forward exchange rates as optimal predictors of future spot rates: An econometric investigation. J. Politic. Econom. 88, 829-853.

Hendry, D. F. and G. E. Mizon (1978). Serial correlation as a convenient simplification, not a nuisance: A comment on a study of the demand for money by the Bank of England. Econom. J. 88, 549--563.

Henriksson, R. D. and R. C. Merton (1981). On market timing and investment performance II: Statistical procedures for evaluating forecast skills. J. Business 54, 513-533.

Keane, M. P. and D. E. Runkle (1990). Testing the rationality of price forecasts: New evidence from panel data. Amer. Econom. Rev. 80, 714--735.

Leitch, G. and J. E. Tanner (1991). Economic forecast evaluation: Profits versus the conventional error measures. Amer. Econom. Rev. 81, 580-590.

Leitch, G. and J. E. Tanner (1995). Professional economic forecasts: Are they worth their costs? J. Forecasting 14, 143---157.

LeRoy, S. F. and R. D. Porter (1981). The present value relation: Tests based on implied variance bounds. Econometrica 49, 555-574.

Lopez, J. A. (1995). Evaluating the predictive accuracy of volatility models. Manuscript, Research and Market Analysis Group, Federal Reserve Bank of New York.

Mark, N. C. (1995). Exchange rates and fundamentals: Evidence on long-horizon predictability. Amer. Econ. Rev. 85, 201-218.

McCulloch, R. and P. E. Rossi (1990). Posterior, predictive and utility-based approaches to testing the arbitrage pricing theory. J. Financ. Econ. 28, 7-38.

Meese, R. A. and K. Rogoff (1988). Was it real? The exchange rate - interest differential relation over the modern floating-rate period. J. Finance 43, 933-948.

Merton, R. C. (1981). On market timing and investment performance I: An equilibrium theory of value for market forecasts. J. Business 54, 513-533.

Mincer, J. and V. Zarnowitz (1969). The evaluation of economic forecasts. In: J. Mincer, ed., Economic forecasts and expectations, National Bureau of Economic Research, New York.

Murphy, A. H. (1973). A new vector partition of the probability score. J. Appl. Meteor. 12, 595--600.

Murphy, A. H. (1974). A sample skill score for probability forecasts. Mronthly Weather Review 102, 48-55.

Murphy, A. H. and R. L. Winkler (1987). A general framework for forecast evaluation. Monthly Weather Review 115, 1330-1338.

Murphy, A. H. and R. L. Winkler (1992). Diagnostic verification of probability forecasts. Internat. J. Forecasting 7, 435-455.

Nelson, C. R. (1972). The prediction performance of the F.R.B.-M.I.T.-Penn model of the U.S. economy. Amer. Econom. Rev. 62, 902-917.

Nelson, C. R. and G. W. Schwert (1977). Short term interest rates as predictors of inflation: On testing the hypothesis that the real rate of interest is constant. Amer. Econom. Rev. 67, 478-486.

Newbold, P. and C. W. J. Granger (1974). Experience with forecasting univariate time series and the combination of forecasts. J. Roy. Statist. Soc. A 137, 131-146.

Pagan, A. R. and G. W. Schwert (1990). Alternative models for conditional stock volatility. J. Econometrics 45, 267-290.

Pesaran, M. H. (1974). On the general problem of model selection. Rev. Econom. Stud. 41, 153-171.

Pesaran, M. H. and A. Timmermann (1992). A simple nonparametric test of predictive performance. J. Business Econom. Statist. 10, 461-465.

Ramsey, J. B. (1969). Tests for specification errors in classical least-squares regression analysis. J. Roy. Statist. Soc. B 2, 350---371.

Satchell, S. and A. Timmermann (1992). An assessment of the economic value of nonlinear foreign exchange rate forecasts. Financial Economics Discussion Paper FE-6/92, Birkbeck College, Cambridge University.

Schnader, M. H. and H. O. Stekler (1990). Evaluating predictions of change. J. Business 63, 99-107.

Seillier-Moiseiwitsch, F. and A. P. Dawid (1993). On testing the validity of sequential probability forecasts. J, Amer. Statist. Assoc. 88, 355-359.

Shiller, R. J. (1979). The volatility of long term interest rates and expectations models of the term structure, ar. Politic. Econom. 87, 1190-1219.

Stekler, H. O. (1987). Who forecasts better? J. Business Econom. Statist. 5, 155-158.

Stekler, H. O. (1994). Are economic forecasts valuable? J. Forecasting 13, 495-505.

Theil, H. (1961). Economic Forecasts and Policy. North-Holland, Amsterdam.

Weiss, A, A. (1995). Estimating time series models using the relevant cost function. Manuscript, Department of Economics, University of Southern California.

Weiss, A. A. and A. P. Andersen (1984). Estimating forecasting models using the relevant forecast evaluation criterion, J. Roy. Statist. Soc. A 137, 484-487.

West, K. D. (1994). Asymptotic inference about predictive ability. Manuscript, Department of Economics, University of Wisconsin.

West, K. D., H. J. Edison and D. Cho (1993). A utility-based comparison of some models of exchange rate volatility. J. Internat. Econom. 35, 23-45.

Winkler, R. L. and S. Makridakis (1983). The combination of forecasts. J. Roy. Statist. Soc. A 146, 150-157.