Xu *Appl Inform* (2018) 5:11
https://doi.org/10.1186/s40535-018-0058-5

Applied Informatics

REVIEW Open Access

Machine learning and causal analyses for modeling financial and economic data

Lei Xu[1,2*]

*Correspondence:
lxu@cse.cuhk.edu.hk
[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China
Full list of author information is available at the end of the article

Abstract

Instead of aiming at a systematic survey, we consider further developments on several typical linear models and their mixture extensions for prediction modeling, portfolio management and market analyses. The focus is put on outlining the studies by the author's research group, featured by (a) extensions of AR, ARCH and GARCH models into finite mixture or mixture-of-experts; (b) improvements of Sharpe ratio by maximizing the expected return and the upside volatility while minimizing the downside risk, with the help of a priori aided diversification; (c) developments of arbitrage pricing theory (APT) into temporal factor analysis (TFA)-based temporal APT, macroeconomics-modulated temporal APT and a general formulation for market modeling, together with applications to temporal prediction and dynamic portfolio management; (d) Bayesian Ying–Yang (BYY) harmony learning is adopted to implement these developments, featured with automatic model selection. After a brief introduction on BYY harmony learning, gradient-based algorithms and EM-like algorithms are provided for learning alternative mixture-of-experts-based AR, ARCH and GARCH models; and (e) path analysis for linear causal analyses is briefly reviewed, a recent development on ρ-diagram is refined for cofounder discovery, and a causal potential theory is proposed. Also, further discussions are made on structural equation modeling and its relations to modulated TFA-APT and nGCH-driven M-TFA-O.

Keywords: Prediction modeling, Portfolio management, Mixture-of-experts, Conditional heteroskedasticity, Arbitrage pricing theory, Temporal factor analysis, Macroeconomics modulated, Path analyses, Structural equation modeling, Cofounder discovery, Causal potential theory

Introduction

Financial and economic data are naturally recorded as temporal sequences or time series, and thus one of major tasks on those data is making time series analysis. Typically, a mathematical model is obtained to describe the regression relation of the current observation from its past observations, such that the future observation is predicted. Such a prediction task has been extensively studied in both the literature of time series analysis and the literature of machine learning and neural networks.

One most classic tool for time series analyses is the autoregressive (AR) model or generally autoregressive–moving-average (ARMA) model, which describes a linear dependence of the current observation on past values and noise disturbances. Extended from

Springer Open

describing stationary processes to data with some identifiable trend of a polynomial growth (Box and Jenkins 1970), an initial differencing step can be applied to remove such a non-stationarity. See Box 1 in Fig. 1; the autoregressive integrated moving average (ARIMA) model is used to refer a "cascade" of this initialization and ARMA. For simplicity, we still prefer to use AMRA to refer ARIMA by regarding such an initialization as a pre-processing stage.

In the literatures of statistics and econometrics, as outlined in Fig. 1 by Box 2, generalizations of ARMA have also been made toward Autoregressive Conditional Heteroskedasticity (ARCH) and generalized ARCH (GARCH) for considering conditional heteroskedasticity of variables (Engle 1982; Bollerslev 1986), to nonlinear ARMA for modeling nonlinear dependence (Leontaritis and Billings 1985), and Vector AR (VAR) for capturing the linear interdependencies among multiple time series (Sims 1980; Engle and Granger 1987).

The field of NN-ML in economics and finance involves each of the three streams of studies. In the early stage, most efforts were put on using multilayer neural networks or recurrent networks for a sophisticated nonlinear dependence of the current observation on past values and noise disturbances, as outlined in Fig. 1 by Box 3. There have been already several books on these studies (e.g., Azoff 1994; Gately 1995; Zhang 2003), and thus this chapter does not cover this type of studies.

Since 1994, the author's group has made many efforts on extending AR, ARMA, ARCH and GARCH models into finite mixture or mixture-of-experts (Xu 1994, 1995a, b; Cheung et al. 1996, 1997; Leung 1997; Kwok et al. 1998; Wong et al. 1998; Chiu and

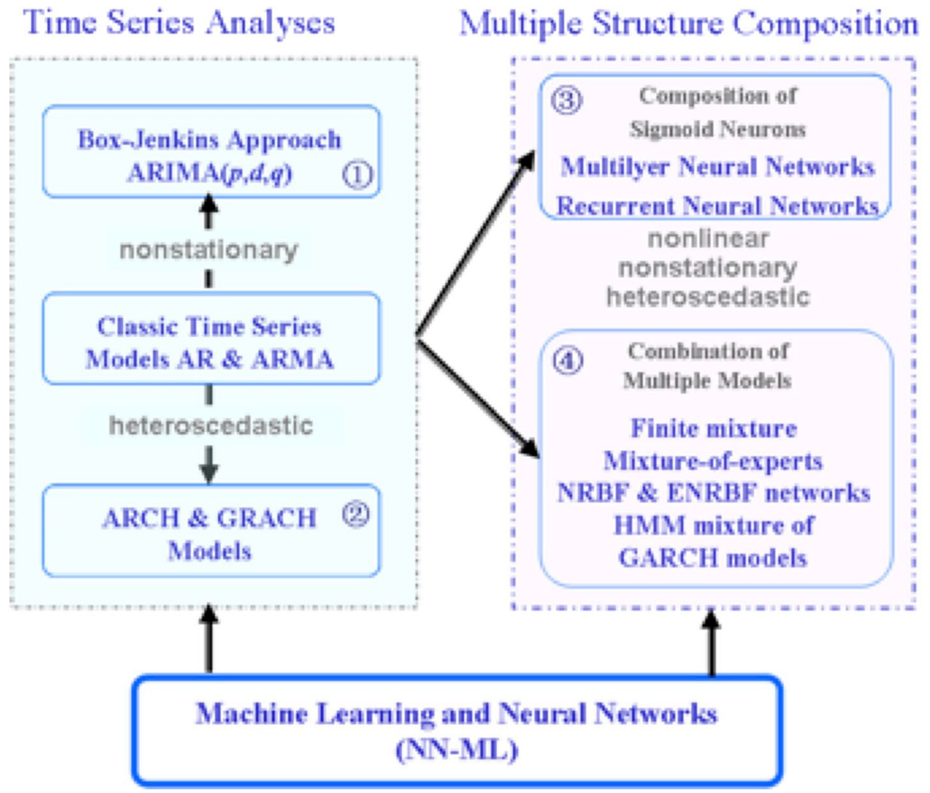


Fig. 1 A road map on studies of time series prediction

Xu 2002a, 2003; Tang et al. 2003). Outlined in Fig. 1 by Box 4, studies actually proceed along an alternative road for modeling temporal dependence featured with nonlinearity, heteroskedasticity and non-stationarity. "Financial prediction: time series models and three finite mixture extensions" section is dedicated to the studies summarized in Fig. 1, together with introductions on learning implementations by the maximum likelihood (ML) learning, the rival penalized competitive learning (RPCL) (Xu et al. 1992, 1993), and approaches of learning with model selection.

"Dynamic trading and portfolio management" section is dedicated to the studies summarized in Fig. 2, toward portfolio management directly, instead of making nonlinear modeling for analyses and predictions. Around the second half of the 1990s, efforts in the literature of neural networks and machine learning in economics and finance started to shift to adaptive trading; see Box 1. Subsequently, these efforts converge to the road pioneered by the Markowitz portfolio theory (Markowitz 1952) that maximizes the portfolio expected return for a given amount of portfolio risk by carefully choosing the proportions of assets; see Box 2. Based on Markowitz's mean–variance paradigm, Sharpe (1966, 1994) further suggests evaluating the goodness of an asset by a ratio of the excess asset return; see Box 3. Later, it is further realized that the return variance is not an appropriate measure of portfolio risk because it counts the positive fluctuation above the expected returns (called *upside volatility*) also as the part of risk. The downside risk thus becomes a topic to study, as illustrated in Fig. 2 by Box 4; e.g., Markowitz (1959) counts the volatility below the expected returns only.

After a brief introduction on the above-mentioned boxes in Fig. 2, "Dynamic trading and portfolio management" section further reexamines the Markowitz paradigm and Sharpe ratio with extensions that maximizes the expected returns and the upside volatility while minimizing the downside risk, with the help of a priori aided diversification (Hung et al. 2000, 2003), see Box 5 in Fig. 2. Moreover, several extensions have been

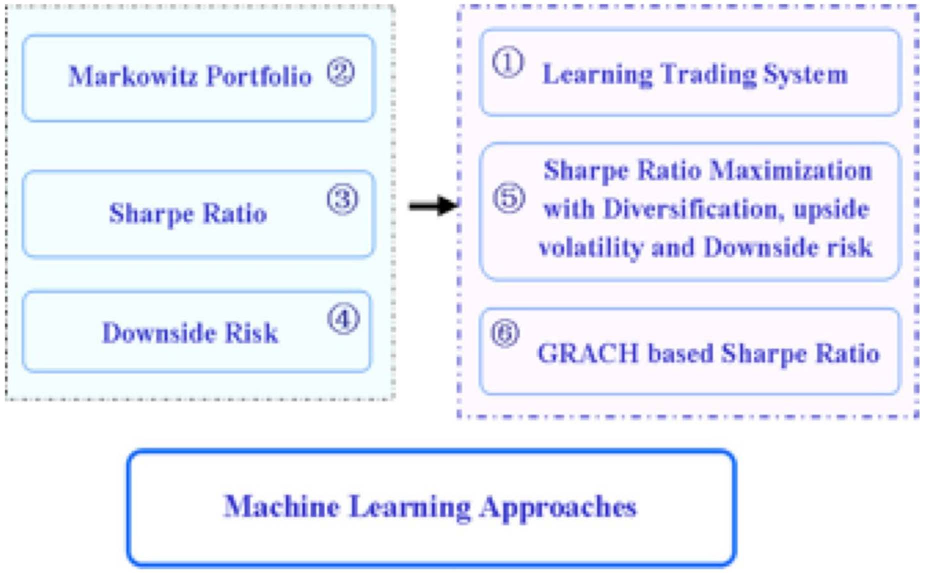

Markowitz Portfolio ②
Sharpe Ratio ③
Downside Risk ④
① Learning Trading System
⑤ Sharpe Ratio Maximization with Diversification, upside volatility and Downside risk
⑥ GRACH based Sharpe Ratio
Machine Learning Approaches

Fig. 2 A road map on studies of portfolio management

proposed along this direction in Sect III(C) of Xu (2001), including that nonparametric estimates of the expected return and volatilities are improved by ARCH or GARCH models; see Box 6 in Fig. 2.

Next, "Market modeling: APT theory and temporal factor analysis" section is dedicated to the efforts summarized in Fig. 3. The Markowitz scheme also leads to the Capital Asset Pricing Model (CAPM) (Sharpe 1964). However, the CAPM is criticized to be not enough to describe a market behavior merely via one endogenous factor. Then, a general linear model of multiple factors has been proposed under the name of Arbitrage Pricing Theory (APT) (Ross 1976). Unfortunately, the APT has not been widely accepted in popularity similar to the CAPM. The reason lies largely with its significant drawback: namely, its implementation is difficult due to the lack of specificity regarding the number and nature of the factors that systematically affect asset return (Dhrymes et al. 1984; Abeysekera and Mahajan 1987).

In "Market modeling: APT theory and temporal factor analysis" section, we start from introducing three approaches that are usually applied for the implementation of APT and address their drawbacks as outlined in "Introduction" section of (Xu 2001), which leads to an observation that the lack of specificity regarding the endogenous factors is not just regarding the number and nature of the factors, but even more seriously arising from the so-called rotation indeterminacy implemented by factor analysis. Thus, further efforts should explore how to add certain structure to remove or remedy this indeterminacy. As outlined in Fig. 3 by Box 1 and Box 2, temporal factor analysis (TFA) (Xu 1997, 2000) is suggested as a generalization of the original APT theory (Xu 2001) to tackle such an incompleteness, featured with a first-order autoregressive dependence added to each factor such that the incompleteness caused by a notorious rotation indeterminacy is removed. Such a generalization is thus called temporal APT in a sense that temporal relation is taken into consideration.

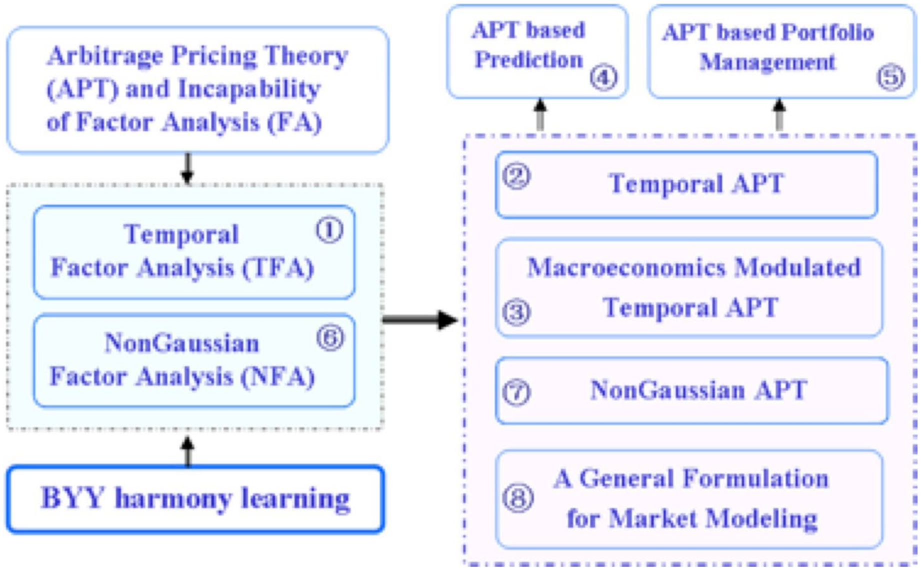

Arbitrage Pricing Theory (APT) and Incapability of Factor Analysis (FA)
APT based Prediction ④
APT based Portfolio Management ⑤
Temporal Factor Analysis (TFA) ①
NonGaussian Factor Analysis (NFA) ⑥
② Temporal APT
Macroeconomics Modulated ③ Temporal APT
⑦ NonGaussian APT
⑧ A General Formulation for Market Modeling
BYY harmony learning

Fig. 3 A road map on studies of generalized apt theories and applications

This section further considers the influences of macroeconomic indexes such as GDP, inflation, investor confidence and yield curve, via their roles in controlling or modulating the temporal factors, which leads to a macroeconomics-modulated temporal APT shown in Fig. 3 by Box 3. Alternatively, TFA may also be replaced by non-Gaussian factor analyses (NFA) such that the incompleteness caused by rotation indeterminacy can also be removed; see Box 6 and Box 7 in Fig. 3. Actually, both the temporal factors and non-Gaussian factors are two aspects of one market model: one observes a dynamic market process, while the other describes the market with all the time points projected to one reference spot. Even generally, conditional heteroskedasticity may also be added to the factors, which finally leads to Box 8 in Fig. 3, namely, a general formulation for financial market modeling that systematically integrates all the ingredients. As illustrated in Fig. 3 by Box 4, various prediction tasks and investment managements can also be conducted with the help of the temporal APT and the macroeconomics-modulated temporal APT.

Further developments of these linear models introduced are suggested to be implemented by the Bayesian Ying–Yang (BYY) harmony learning. In "Bayesian Ying–Yang harmony learning and two exemplar learning algorithms" section, the fundamentals of BYY harmony learning are briefly introduced. For learning alternative mixture-of-experts-based AR, ARCH and GARCH models, both gradient-based algorithms and EM-like algorithms are provided for implementations, featured with automatic model selection and in reference of the well-known EM algorithm.

Except for the first column in Fig. 1, where only one time series is considered, mostly we consider dependences across more than one channel of time series. Prediction and decision making in portfolio management are based on such dependences that may not necessarily reflect causal structure underlying data, while it will be better to make prediction and decision based on casual structure. In "Linear causal analyses" section, path analyses (Wright 1934) for linear causal analyses is briefly reviewed, a recent development on ρ-diagram (Xu 2018) is refined for cofounder discovery and a causal potential theory is proposed. Further discussions are made on structural equation modeling (SEM) (Ullman 2006; Pearl 2010a; Westland 2015; Kline 2015) and its relations to modulated TFA-APT and nGCH-driven M-TFA-O.

Financial prediction: time series models and three finite mixture extensions

Time series models and neural networks

One most classic tool for time series analyses is the autoregressive (AR) model or generally autoregressive–moving-average (ARMA) model as follows:

$$x_t = a_0 + \varepsilon_t + \sum_{j=1}^{q} a_j x_{t-j} + \sum_{i=1}^{p} b_i \varepsilon_{t-i}, \varepsilon_t \sim^{\text{i.i.d.}} G(\varepsilon|0, \sigma^2), \tag{1}$$

where $\varepsilon_t \sim^{\text{i.i.d.}} G(\varepsilon|0,\sigma^2)$ denotes that $\varepsilon_1, \ldots, \varepsilon_t, \ldots$ are i.i.d. samples from $G(\varepsilon|0,\sigma^2)$, while $G(u|\mu,\sigma^2)$ denotes a Gaussian distribution of u with the mean μ and the variance σ^2. Particularly, the ARMA model degenerates to the AR model when $q=0$.

The ARMA model is appropriate to describe a wide sense stationary sequence. Extension has been made to describe data ξ_t that have some clearly identifiable trend of a polynomial growth (Box and Jenkins 1970); see Box 1 in Fig. 1. It is made simply by an initial differencing to remove the non-stationarity. That is, we get

$$x_t = \Delta^d \xi_t, \quad \text{where } d > 0,\ \Delta u_t = u_t - u_{t-1} \text{and } u_t = \Delta^d \xi_t. \tag{2}$$

A cascade of this initialization and ARMA is called the autoregressive integrated moving average (ARIMA) model. For simplicity, we prefer to still use AMRA to indicate ARIMA by regarding such an initialization as a pre-processing stage.

In the literature of statistics, econometrics, control and signal processing, generalizations of ARMA have been made toward Autoregressive Conditional Heteroskedasticity (ARCH) and generalized ARCH (GARCH) for considering variables conditional to heteroskedasticity (Engle 1982; Bollerslev 1986); see Box 8 in Fig. 1. Namely, we consider

$$x_t = a_0 + \sum_{j=1}^{q} a_j x_{t-j} + \varepsilon_t, \varepsilon_t = \sigma_t z_t, z_t \sim^{\text{i.i.d.}} G(z|0,1),$$

where σ_t is not a constant, but given by the following regression:

$$\begin{aligned} \sigma_t^2(\vartheta) &= \sigma_0^2 + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \omega_j \sigma_{t-j}^2, \\ \vartheta &= \{\sigma_0^2 > 0, \beta_i \geq 0,\ for\ \ i > 0, \omega_j \geq 0,\ for\ \ j \geq 0\}, \end{aligned} \tag{3}$$

which is usually denoted by $GARCH(p,q)$ and degenerates to the ARCH model when $p=0$.

Extensions of the ARMA model have also been made under the name of nonlinear ARMA (NARMA) for modeling nonlinear dependence (Leontaritis and Billings 1985) and to Vector AR (VAR) for capturing the linear interdependencies among multiple time series (Sims 1980; Engle and Granger 1987). In the literature, many efforts have been made on using multilayer neural networks or recurrent networks for a sophisticated nonlinear dependence of the current observation on past values and noise disturbances, as illustrated by Box 3 in Fig. There are already several books on these studies (e.g., Azoff 1994; Gately 1995; Zhang 2003), and thus this chapter does not cover this type of studies. Instead, the subsequent two subsections will focus on Box 4 in Fig. 1, namely, learning mixture of multiple models.

Learning mixture of AR, ARMA, ARCH and GRACH models

Studies on finite mixture extensions of AR, ARMA, ARCH and GARCH models can be summarized into the following general expression:

$$\begin{aligned} &P(\varepsilon_t|\boldsymbol{x}_{t-1}^{q}, \theta) = \sum_{\ell=1}^{k} \alpha_\ell G(x_t - \mu_{\ell,t}|0, \sigma_{\ell,t}^2), \\ &\mu_{i,t} = \widehat{x}_t\left(\boldsymbol{x}_{t-1}^{q_i}, \boldsymbol{a}_i\right), \mathbf{x}_{t-1}^{m} = \left[x_{\mathrm{t\,-\,1}}, \ldots, x_{\mathrm{t\,-\,m}}\right]^{\mathrm{T}}, \boldsymbol{a}_i = \left[a_{0,i}, a_{1,i}, \ldots, a_{q_i,i}\right]^{\mathrm{T}}, \\ &\varepsilon_t = x_t - \widehat{x}_t\left(\boldsymbol{x}_{t-1}^{q}, \theta\right), q = \max\{q_{1,}, \ldots, q_m\}, \end{aligned} \tag{4}$$

where we consider k regression models $x_t = \mu_{i,t} + \varepsilon_{i,t}, i = 1, , \ldots, k$ with each $\mu_{i,t} = \widehat{x}_t\left(\boldsymbol{x}_{t-1}^{q_i}, \boldsymbol{a}_i\right)$ being either of AR, ARMA, ARCH and GARCH models, and with the corresponding residual $\varepsilon_{i,t}$ from $G(\varepsilon_{i,t}|0, \sigma_{i,t}^2)$. Typically, the studies of the AR, ARCH and GARCH models share the following detailed expression (Xu 1995a, b; Cheung et al. 1997; Kwok et al. 1998; Wong et al. 1998; Chiu and Xu 2003, 2004a; Tang et al. 2003):

$$\mu_{i,t} = \widehat{x}_t\left(\boldsymbol{x}_{t-1}^{q_i}, \boldsymbol{a}_i\right) = \boldsymbol{a}_i^{\mathrm{T}} \begin{bmatrix} 1 \\ \boldsymbol{x}_{t-1}^{q_i} \end{bmatrix}, \sigma_{i,t}^2 = \begin{cases} \sigma_{i,0}^2 > 0, & \text{(a)AR} \\ \sigma_{i,0}^2 + \boldsymbol{b}_i^{\mathrm{T}} E_{i,t-1}^{q_i}, & \text{(b)ARCH} \\ \sigma_{i,0}^2 + \boldsymbol{b}_i^{\mathrm{T}} E_{i,t-1}^{q_i} + \boldsymbol{w}_i^{\mathrm{T}} \sum_{i,t-1}^{p_i}, & \text{(c)GARCH} \end{cases}$$
$$E_{i,t-1}^n = \left[\varepsilon_{i,t-1}^2, \ldots, \varepsilon_{i,t-n}^2\right]^{\mathrm{T}}, \mathbf{w}_i = \left[w_{1,i}, \ldots, w_{p_i,i}\right]^{\mathrm{T}},\ \omega_{j,i} \geq 0, j = 1, \ldots, p_i$$
$$\Sigma_{i,t-1}^n = \left[\sigma_{i,t-1}^2, \ldots, \sigma_{i,t-n}^2\right]^{\mathrm{T}}, \mathbf{b}_i = \left[\beta_{1,i}, \ldots, \beta_{q_i,i}\right]^{\mathrm{T}},\ \beta_{j,i} \geq 0, j = 1, \ldots, q_i. \tag{5}$$

For ARMA (Kwok et al. 1998; Tang et al. 2003), the detailed expression of $\mu_{i,t} = \widehat{x}_t\left(\boldsymbol{x}_{t-1}^{q_i}, \boldsymbol{a}_i\right)$ is given by Eq. (1). Moreover, $\widehat{x}_t\left(\boldsymbol{x}_{t-1}^{q_i}, \boldsymbol{a}_i\right)$ can be also a specific nonlinear function, e.g., given by three-layer neural networks (Cheung et al. 1996, 1997) or the normalized radial basis function (NRBF) and extended NRBF (ENRBF) (Xu 1998, Xu 2009).

According to Eq. (4), a sequence x_1, ..., x_t, ... may come from the ith one of the k models with the probability α_i, and jointly the k models describe the sequence x_1, ..., x_t, ... with a residual ε_t that comes from a Gaussian mixture $P(\varepsilon_t|\boldsymbol{x}_{t-1}^q, \theta)$. In such a way, a nonlinear dependence of the current observation on past values and noise disturbances is modeled by probabilistically combining a mixture of linear models, which keeps the model structure simple and easy to learn. Moreover, non-stationarity beyond ones handled by ARIMA and GARCH models is able to be modeled via switching among individual linear models.

Also, a sequence x_1, ..., x_t, ... may be segmented into pieces with different statistical properties, simply by Bayesian posterior as follows (Xu 1994, 1995a, b):

$$P(j_t|x_t, \boldsymbol{x}_{t-1}^q, \theta) = \frac{\alpha_{j_t} G(x_t - \mu_{j_t,t}|0, \sigma_{j_t,t}^2)}{\sum_{j_t=1}^k \alpha_{j_t} G(x_t - \mu_{j_t,t}|0, \sigma_{j_t,t}^2)}, \tag{6}$$

that is, x_t is identified as coming from the j^*th model by

$$j^* = \mathrm{argmax}_j P(j|x_t, \boldsymbol{x}_{t-1}^q, \theta) \quad \text{or } j^* = \mathrm{argmax}_j [\alpha_j G(x_t - \mu_{j,t}|0, \sigma_{j,t}^2)].$$

To reduce the number of small fragments, some post-processing or smoothing regularization may be added. Moreover, we may extend a finite mixture into a hidden Markov model (HMM) (Rabiner 1989), in which each hidden state is associated with one $G(x_t - \mu_{j,t}|0, \sigma_{j,t}^2)$ and the transition between state is described by

$$\begin{aligned} &\boldsymbol{\alpha_t} = \boldsymbol{Q\alpha_{t-1}}, \boldsymbol{\alpha_t} = [\boldsymbol{\alpha_{1,t}}, \ldots, \boldsymbol{\alpha_{k,t}}]^{\mathrm{T}}, 0 \leq \boldsymbol{\alpha_{j,t}} \leq 1, \sum_{\boldsymbol{j}} \boldsymbol{\alpha_{j,t}} = 1, \\ &\boldsymbol{Q} = \left[\boldsymbol{q_{j|i}}\right], 0 \leq \boldsymbol{q_{j|i}} \leq 1, \sum_{\boldsymbol{i}=1}^{\boldsymbol{k}} \boldsymbol{q_{j|i}} = 1, \end{aligned} \tag{7}$$

with $\alpha_{j,t}$ estimated as time proceeds and then used in Eq. (5) and Eq. (6). Moreover, we can also further modify Eq. (5) and Eq. (6) into

$$P(j_t|x_t, j_{t-1}, \boldsymbol{x}^q_{t-1}, \theta) = \frac{q_{j_t|j_{t-1}} G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t})}{\sum_{j_t=1}^{k} q_{j_t|j_{t-1}} G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t})},$$
$$j^*_t = \text{argmax}_j P(j|x_t, j_{t-1}, \boldsymbol{x}^q_{t-1}, \theta) \text{ or } j^*_t = \text{argmax}_j [q_{j|j_{t-1}} G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t})]. \tag{8}$$

Next, we proceed to estimate x_t from the finite mixture by Eq. (4). It follows that

$$\widehat{x}_t\left(\boldsymbol{x}^q_{t-1}, \theta\right) = \int x_t p(\varepsilon_t|\boldsymbol{x}^q_{t-1}, \theta)\, \mathrm{d}x_t = \sum_{i=1}^{k} \alpha_i \mu_{i,t}, \tag{9}$$

that is, we improve the prediction of x_t via each individual model by a line combination weighted by each α_i. However, this improvement is limited because α_i is a constant that does not change as the samples vary with time.

Each α_i in Eq. (4) cannot directly be replaced by its corresponding Bayes posterior by Eq. (5). First, $P(j_t|x_t, \boldsymbol{x}^q_{t-1}, \theta)$ cannot be moved out of the integral $\int x_t P(j_t|x_t, \boldsymbol{x}^q_{t-1}, \theta)$ $G(x_t|\mu_{j,t}, \sigma^2_{j,t})dx_t$, though the integral can be made approximately. Second, the calculation needs to know x_t. Getting $\widehat{x}_t$ from knowing x_t is applicable to a filtering problem that gets a smoothed or filtered version from x_t, but it is not applicable to a prediction problem that targets at getting $\widehat{x}_t$ from its past observations.

Instead, we use a predictive $P(j_t|\boldsymbol{x}^q_{t-1}, \varphi)$ based on the immediate past observations $\boldsymbol{x}^q_{t-1}$ to combine the prediction of individual prediction model adaptively; that is, we have

$$p(\varepsilon_t|\boldsymbol{x}^q_{t-1}, \theta) = \sum_{j_t=1}^{k} P(j_t|\boldsymbol{x}^q_{t-1}, \varphi) G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t}),$$
$$\widehat{x}_t\left(\boldsymbol{x}^q_{t-1}, \theta\right) = \int x_t p(\varepsilon_t|\boldsymbol{x}^q_{t-1}, \theta)\, \mathrm{d}x_t = \sum_{j_t=1}^{k} P(j_t|\boldsymbol{x}^q_{t-1}, \varphi)\mu_{j_t,t}, \tag{10}$$

which summarizes extensions of the AR, ARMA, ARCH and GARCH models with the help of the mixture-of-experts (ME). In the implementation of the original ME (Jacobs et al. 1991; Jordan and Xu 1995), $P(j|\boldsymbol{x}^q_{t-1}, \varphi)$ is called the gating net and given as follows:

$$P(j|\boldsymbol{x}^q_{t-1}, \varphi) = e^{g_j\left(\boldsymbol{x}^q_{t-1}, \varphi\right)} / \sum_{j=1}^{k} e^{g_j\left(\boldsymbol{x}^q_{t-1}, \varphi\right)},$$

with $g_1\left(\boldsymbol{x}^q_{t-1}, \varphi\right), \ldots, g_k\left(\boldsymbol{x}^q_{t-1}, \varphi\right)$ being the output of multilayer networks.

In an implementation of an alternative ME model (Xu et al. 1994, 1995), we consider a predictive Bayesian posteriori

$$P(j|\boldsymbol{x}^q_{t-1}, \varphi) = \frac{\alpha_j q(\boldsymbol{x}^q_{t-1}|\psi_j)}{q(\boldsymbol{x}^q_{t-1}|\psi)}, q(\mathbf{x}^q_{t-1}|\psi) = \sum_{j=1}^{k} \alpha_j q(\boldsymbol{x}^q_{t-1}|\psi_j). \tag{11}$$

For the AR, ARCH and GARCH models, we further have

$$q(\boldsymbol{x}^q_{t-1}|\psi_j) = q(x_{t-1}|x_{t-2},\dots,x_{t-q})q(x_{t-2}|x_{t-3},\dots,x_{t-q})\cdots q\big(x_{t-q}\big).$$

To simplify the computation, we may consider the following approximation:

$$q(\boldsymbol{x}^q_{t-1}|\psi_j) \approx q(x_{t-1}|x_{t-2},\dots,x_{t-q-1}) = G(x_{t-1}-\mu_{j,t-1}|0,\sigma^2_{j,t-1}). \tag{12}$$

A further insight into Eq. (11) can be obtained at a setting that $\sigma^2_{j,t-1}=\sigma^2_j$ and $x_{t-1}=\mu_{j,t-1}$.; in this special case, we have a further simplification:

$$P(j|\boldsymbol{x}^q_{t-1},\varphi) = \frac{\alpha_j/\sigma_j}{\sum_{j=1}^k \alpha_j/\sigma_j}, \tag{13}$$

which shares a similar concept to the mixture-using variance (MUV) and actually degenerates to this MUV (Perrone and Cooper 1993, Perrone 1994) when $\alpha_j \propto \sigma^{-1}_{j,t}$. Another special case is that $\alpha_i/\sigma_{i,t}$ is constant, and it follows from Eqs. (11) to (12) that we have

$$P(j|\boldsymbol{x}^q_{t-1},\varphi) = e^{-\frac{1}{2\sigma^2_{j,t-1}}(x_{t-1}-\mu_{j,t-1})^2} \Big/ \sum_{j=1}^k e^{-\frac{1}{2\sigma^2_{j,t-1}}(x_{t-1}-\mu_{j,t-1})^2}, \tag{14}$$

by which we get the counterparts of NRBF and ENRBF (Xu 1998, Xu 2009).

The other choices of $P(j|\boldsymbol{x}^q_{t-1},\varphi)$ may also be obtained or modified from Table 3 in Xu and Amari (2008). Moreover, similar to Eq. (8), it still follows from $q(\boldsymbol{x}^q_{t-1}|\psi_j)$ given by Eqs. (11) and (12) that we may further incorporate the HMM model from Eq. (7) into Eq. (11) and get

$$P(j_t|x_t,j_{t-1},\boldsymbol{x}^q_{t-1},\varphi) = q_{j_t|j_{t-1}}q(\boldsymbol{x}^q_{t-1}|\psi_j) \Big/ \sum_{j=1}^k q_{j_t|j_{t-1}}q(\boldsymbol{x}^q_{t-1}|\psi_j). \tag{15}$$

Maximum likelihood, RPCL learning and learning with model selection

Typically, unknown parameters in the models in Eqs. (4), (8), (10) and (11) are estimated by the maximum likelihood (ML) learning, that is, the following maximization:

$$\Theta^* = \arg\max_{\Theta} L(\{x_t\}_{t=1}^N|\Theta),$$

$$L(\{x_t\}_{t=1}^N|\Theta) = \begin{cases} \sum_t \ln \sum_{j_t=1}^k \alpha_{j_t} G(x_{\mathrm{t}}-\mu_{j_t,t}\,|0,\sigma^2_{j_t,t}), & \text{(a) for finite mixture by Eq. (4),} \\ \sum_t \ln \left\{ \sum_{j_t=1}^k P(j_t|\mathbf{x}^q_{t-1},\phi)G(x_{\mathrm{t}}-\mu_{j_t,t}\,|0,\sigma^2_{j_t,t}) \right\}, & \text{(b) for ME by Eq. (10),} \\ \sum_t \ln \left\{ \sum_{j_t=1}^k \alpha_{j_t} q(\mathbf{x}^q_{t-1}|\psi_{j_t})G(x_{\mathrm{t}}-\mu_{j_t,t}\,|0,\sigma^2_{j_t,t}) \right\}, & \text{(c) for AME by Eq. (11),} \\ \ln \left\{ \sum_{j_1,\dots,j_N} \prod_t q_{j_t|j_{t-1}} G(x_{\mathrm{t}}-\mu_{j_t,t}\,|0,\sigma^2_{j_t,t}) \right\}, & \text{(d) for HMM mixture .} \end{cases} \tag{16}$$

This maximization is implemented by the EM algorithm (Redner and Walker 1984), e.g., see the EM algorithms for finite mixture of AR models in Xu (1994, 1995a, b), finite mixture of GARCH models in Wong et al. (1998), finite mixture of ARMA–GARCH models in Tang et al. (2003) and the original ME in Jordan and Xu (1995), as well as the alternative ME model, NRBF and ENRBF in Xu et al. (1994, 1995) and Xu (1998, 2009).

For an HMM mixture, we may also have the following approximate likelihood:

$$L(x_{t(t=1)}^{N}|\Theta) = \begin{cases} \sum_t \ln \sum_{j_t=1}^{k} q_{j_t|j_{t-1}} G(x_t - \mu_{j_t,t}|0, \sigma_{j_t,t}^2), & \text{(i)}, \\ \sum_t \ln \left\{ \sum_{j_t=1}^{k} q_{j_t j_{t-1}} q(\boldsymbol{x}_{t-1}^{q}|\psi_{j_t}) G(x_t - \mu_{j_t,t}|0, \sigma_{j_t,t}^2) \right\}, & \text{(ii)}. \end{cases} \quad (17)$$

One critical problem for the ML learning is that a good performance on a training set is not necessarily good on a testing set, especially when the training set consists of a small size of samples. The reason is that there may be too many free parameters. As introduced in the third section of Xu (2009), efforts on this problem are mainly featured by learning with *model selection*. Model selection refers to select a model with an appropriate complexity $\boldsymbol{k}$. For the models considered in the previous subsection, $\boldsymbol{k}$ consists of the number of individual models, the autoregression order and the moving average order for each individual model. Typically, the ML learning is not good for model selection. However, whether the EM algorithm works well depends on whether an appropriate $\boldsymbol{k}$ is selected.

Classically, model selection is made in a two-stage implementation. First, enumerate a candidate set $\mathbf{K}$ of $\boldsymbol{k}$ and estimate a solution $\Theta_{\boldsymbol{k}}^{*}$ for the unknown set Θ_k of parameters by the ML learning at each $\boldsymbol{k} \in \mathbf{K}$. Second, use a model selection criterion $J\left(\Theta_{\boldsymbol{k}}^{*}\right)$ to select a best $\boldsymbol{k}^{*}$. Several classical criteria are available for the purpose, such as AIC, CAIC and BIC/MDL, and readers are referred to Xu (2009, 2010) for a recent outline. Unfortunately, any one of these criteria usually provides a rough estimate that may not yield a satisfactory performance. Even with a criterion $J(\Theta_{\boldsymbol{k}})$ available, this two-stage approach usually incurs a huge computing cost. Still, the parameter learning performance deteriorates rapidly as $\boldsymbol{k}$ increases, which makes the value of $J(\Theta_{\boldsymbol{k}})$ to be evaluated unreliably.

One direction that tackles this challenge is called automatic model selection, which is associated with a learning algorithm or a learning principle with the following two features:

- When there is an indicator $\rho(\theta_{\boldsymbol{r}})$ on a subset $\theta_{\boldsymbol{r}} \in \Theta_{\boldsymbol{k}}$, we have $\rho(\theta_{\boldsymbol{r}}) = 0$ if $\theta_{\boldsymbol{r}}$ consists of parameters of a redundant structural part.
- In implementation of this algorithm or principle, there is a mechanism that automatically drives $\rho(\theta_{\boldsymbol{r}}) \to 0$ as $\theta_{\boldsymbol{r}}$ toward a specific value. Thus, the corresponding redundant structural part is effectively discarded.

An early effort along this direction is rival penalized competitive learning (RPCL) (Xu et al. 1992, 1993) for adaptively learning a model that consists of k substructures as follows:

$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + p_{j,t}\eta \frac{\partial \pi_{j,t}\left(\theta_j^{\text{old}}\right)}{\partial \theta_j}, \quad p_{j,t} = \begin{cases} 1, & j = c = \text{argmax}_j \pi_{j,t}\left(\theta_j^{\text{old}}\right), \\ \gamma, & j = \text{argmax}_{j \neq c} \pi_{j,t}\left(\theta_j^{\text{old}}\right), \\ 0, & \text{otherwise}. \end{cases} \quad (18)$$

where $\eta > 0$ is a learning step size and γ is a small positive number, e.g., $\gamma = 0.005$–0.01. With k initially at a value large enough, a current input sample x_t is allocated to one of the k substructures via competition. The winner adapts to this sample by a little bit, while the rival is de-learned a little bit to reduce a duplicated allocation. This rival penalized mechanism will discard extra substructures, making model selection automatically during learning. Readers are referred to Xu (2007) for a recent overview and extensions.

Corresponding to Eq. (16), $\pi_{j,t}(\theta_j^{\mathrm{old}})$ in Eq. (18) is given as follows:

$$\pi_{j_t,t}\left(\theta_{j_t}\right) = \begin{cases} \ln[\alpha_{j_t} G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t})], & \text{(a) for finite mixture by Eq.(4)}, \\ \ln[P(j_t|\boldsymbol{x}^q_{t-1}, \varphi) G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t})], & \text{(b) for ME by Eq.(10)}, \\ \ln[\alpha_{j_t} q(\boldsymbol{x}^q_{t-1}|\psi_{j_t}) G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t})], & \text{(c) for AME by Eq.(11)} \end{cases} . \tag{19}$$

For an HMM mixture, we may also approximately have

$$\pi_{j_t,t}\left(\theta_{j_t}\right) = \begin{cases} \ln[q_{j_t|j_{t-1}} G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t})], & \text{(i) for HMM mixture by Eq.(8)}, \\ \ln[q_{j_t|j_{t-1}} q(\boldsymbol{x}^q_{t-1}|\psi_{j_t}) G(x_t - \mu_{j_t,t}|0, \sigma^2_{j_t,t})], & \text{(ii) for HMM AME by Eq.(15)}. \end{cases} \tag{20}$$

Another stream of automatic model selection is featured by those appropriate prior-based efforts. By a Laplace prior in a regression task, sparse learning or Lasso shrinkage prunes away extra weights (Williams 1995; Tibshirani 1996). For pruning away Gaussian components on Gaussian mixture, a Jeffreys priori is used in the implementation of the minimum message length (MML) that minimizes a two-part message for a statement of model and a statement of data encoded by that model (Figueiredo and Jain 2002), and also Dirichlet–Normal–Wishart priories is added on Gaussian components in the implementation of the variational Bayes (VB) that computes a lower bound of the marginal likelihood (McGrory and Titterington 2007).

However, these efforts highly depend on choosing an appropriate prior, which is usually a difficult task, while an inappropriate prior may deteriorate the performance of model selection seriously. Without any priors on the parameters, VB and MML all degenerate to the maximum likelihood learning, while the RPCL learning is still capable of automatic model selection. Firstly proposed in Xu (1995a, b) and systematically developed over a decade and half (Xu 2001, 2007, 2010, 2012), the third stream of efforts has been made under the name of Bayesian Ying–Yang (BYY) harmony learning. The BYY harmony learning shares a mechanism similar to the RPCL learning. Also, the performances of BYY harmony learning can be further improved by incorporating appropriate priors. Further details about the BYY harmony learning are referred to "Bayesian Ying–Yang harmony learning and two exemplar learning algorithms" section, where a tutorial is also provided on one BYY harmony learning algorithm for alternative mixture-of-experts-based GARCH models.

Dynamic trading and portfolio management

Dynamic trading by supervised learning and reinforcement learning

Instead of building a mathematical model for understanding and forecasting time series, studies of neural networks and machine learning in economics and finance started to shift from nonlinear forecasting modeling to adaptive trading and dynamic portfolio management (Neuneier 1996; Choey and Weigend 1997; Xu and Cheung 1997; Moody et al. 1998; Hung et al. 2000; Moody and Saffell 2001; Hung et al. 2003; Chiu and Xu 2004b; Jangmin 2006). Efforts on portfolio management will be addressed in the next subsection. In the sequel, we introduce efforts on learning dynamic trading based on one single time series, with the help of supervised learning, reinforcement learning and Sharpe ratio maximization.

Given a sequence x_1, ..., x_t, e.g., the sequence of one asset, Gold, FOREX index,..., etc., at any time point $t \leq \tau$ we may infer a sequence $I_1^p, \ldots I_t^p$ each I_τ^p being the following desired trading signal:

$$I_\tau^p = \begin{cases} +1, & \text{to buy,} \\ -1, & \text{to sell,} \\ 0, & \text{no action,} \end{cases} \tag{21}$$

based on a trading strategy (e.g., maximum return) or an external expertise.

The task of learning decision, as illustrated by Box 1 in Fig. 2, can be formulated as a nonlinear regression model:

$$\tilde{I}_t^p = \frac{1 - e^{-f\left(XF_t^q, \left\{I_{t-\tau}^p\right\}_{t=1}^q, \Theta\right)}}{1 + e^{-f\left(XF_t^q, \left\{I_{t-\tau}^p\right\}_{t=1}^q, \Theta\right)}} \tag{22}$$

where $f\left(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \mathit{\Theta}\right)$ is implemented by an ENRBF network in Xu & Cheung (1997). Also, it can be implemented by three-layer neural networks. Supervised learning is used to determine the unknown parametric $\mathit{\Theta}$ by minimizing

$$E_2(\mathit{\Theta}) = \sum_t [I_t^p - f(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \mathit{\Theta})]^2, \tag{23}$$

where XF_t^q may be directly a number of past observations $\{x_{t-\tau}\}_{t=1}^q$ or certain features $\{F_t^{(i)}\}$ extracted from$\{x_{t-\tau}\}_{t=1}^q$, e.g., $F_t^{(i)}$ may be MACD, RSI, %K, %D, as well as features from candlestick charts and configurations from waves, etc. Also, we may put both together to consider $XF_t^q = \left\{\{x_{t-\tau}\}_{t=1}^q, \left\{F_t^{(i)}\right\}\right\}$.

One key problem is how to keep a good generalization ability by training with a small length of sequence x_1, ..., x_t. One way is adding some regularization term $E_2(\mathit{\Theta}) + \lambda \mathit{\Gamma}(\mathit{\Theta})$. Without a priori knowledge, however, it is not an easy task to get an appropriate term $\mathit{\Gamma}(\mathit{\Theta})$ and its strength λ. The other way is to describe the model as follows:

$$\begin{gathered} q\left(\tilde{I}_t^p \mid XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta\right) = \frac{exp\left[z_t^{(1)} f^{(1)}\left(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta\right) + z_t^{(2)} f^{(2)}\left(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta\right)\right]}{1 + exp\left[f^{(1)}\left(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta\right)\right] + \exp\left[f^{(2)}\left(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta\right)\right]} \\ f(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta) = [f^{(1)}(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta), f^{(2)}(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta)]^{\mathrm{T}}, \end{gathered} \tag{24}$$

with $I_t^p = [z_t^{(1)}, z_t^{(2)}, z_t^{(3)}]^{\mathrm{T}}, z_t^{(2)} = 0 \text{ or } 1 \text{ and } z_t^{(1)} + z_t^{(2)} + z_t^{(3)} = 1$. Correspondingly, min $_{\mathit{\Theta}}E_2(\mathit{\Theta})$ is replaced by maximizing the likelihood $L(\mathit{\Theta}) = \sum_t \ln q(I_t^p | f(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \mathit{\Theta}))$. In the formulation, learning regularization may be implemented via Bayesian learning with help of a priori distribution $q(\mathit{\Theta})$, i.e., max $_{\mathit{\Theta}}[L(\mathit{\Theta}) + \ln q(\mathit{\Theta})]$. For a better generalization ability, we may also put $q(I_t^p | f(XF_t^q, \{I_{t-\tau}^p\}_{t=1}^q, \mathit{\Theta}))$ into a Bayesian Ying–Yang system and making BYY harmony learning with automatic model selection; see Sect. 4.4 in Xu (2010).

The other key problem is how to make a pre-processing stage for getting a desired sequence $I_1^p, \ldots, I_t^p$, which can be obtained automatically by a trading strategy, e.g., getting a profit and cutting a loss beyond a pre-specified threshold as follows:

$$I_t^p = \begin{cases} +1, & \text{if } (x_t - x_{t-1}) \big/ \sigma_t \geq g_0^+ > 0, \\ -1, & \text{if } (x_t - x_{t-1}) \big/ \sigma_t \leq g_0^- \leq 0, \\ 0, & \text{no action,} \end{cases}$$

where σ_t is an estimation of the *volatility* about this asset. Also, $I_1^p, \ldots, I_t^p$ may come from an outcome of market technical analysis, which is difficult to get $I_1^p, \ldots, I_t^p$ adaptively in a dynamic trading.

From the studies (Moody et al. 1998; Moody and Saffell 2001; Jangmin 2006), $I_1^p, \ldots, I_t^p$ is a sequence of actions that are dynamically learned by reinforcement learning. Typically, a reinforcement learning model consists of a set S of environment states (e.g., differences in the current price of asset and the volumes in holding) and a set A (e.g., buy, sell, no action) of actions. There is also a policy π that chooses an action $a_t \in A$ at an environment state s_t. The action a_t makes the environment move to a new state s_{t+1}. Associated with the transition (s_t, a_t, s_{t+1}), there is a scalar immediate reward $r_{t+1}(s_t, a_t, s_{t+1})$ that is estimated according to a utility function, e.g., a maximum profit. The goal is to collect as much reward as possible by determining a sequence of actions $a_1, \ldots, a_t$.

In the literature of reinforcement learning, one popular approach is called Q-learning, by which a_t is chosen according to a table $Q(s_t, a_t)$ that is learned from $r_{t+1}(s_t, a_t, s_{t+1})$. For a dynamic trading, the S of environment states are featured by differences in the current price of asset and the volumes in holding. Quantizing the differences into the states is not an easy task. Also, there will be a large number states to be considered. As a result, we need to learn a large $Q(s_t, a_t)$ table, which not only increases computing cost rapidly, but also makes the problem of a small sample size become more serious because $Q(s_t, a_t)$ consists of too many free parameters to be determined. Instead of Q-learning, the action a_t in $r_{t+1}(s_t, a_t, s_{t+1})$ can be approximately replaced by the value of I_t^p given by Eq. (22) such that $r_{t+1}(s_t, a_t, s_{t+1})$ is replaced by an expression $r_{t+1}(s_t, s_{t+1}, \{x_{t-\tau}\}_{t=1}^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta)$. As a result, the maximization of $\sum_{t=1}^{\infty} \gamma^t r_{t+1}(s_t, a_t, s_{t+1})$ with respect to a sequence of discrete actions $a_1, \ldots, a_t$ is replaced by the maximization of $\sum_{t=1}^{\infty} \gamma^t r_{t+1}(s_t, s_{t+1}, \{x_{t-\tau}\}_{t=1}^q, \{I_{t-\tau}^p\}_{t=1}^q, \Theta)$ with respect to Θ. Similar to learning regularization, the problem of a small sample size may also be handled by adding a a priori term, e.g., $\sum_{t=1}^{\infty} \gamma^t r_{t+1}\left(s_t, s_{t+1}, \{x_{t-\tau}\}_{t=1}^q, \left\{I_{t-\tau}^p\right\}_{t=1}^q, \Theta\right) + \lambda \ln q(\Theta)$.

The last but not the least, the specific expression of $r_{t+1}(s_t, a_t, s_{t+1})$ is an important practical issue, related to the current price of asset, the volume in holding, the transaction cost and the tax, as well as personal preference. There could be a number of choices. See Fig. 2 by Box 3; a widely used one is the Sharpe ratio, which is originally suggested for evaluating the goodness of an asset in market by a ratio of the excess asset return (i.e., after minus the benchmark return) over the standard deviation of the excess asset return (Sharpe 1966, 1994). For dynamic trading, it is not the Sharpe ratio of the asset in market that has to be calculated, but the Sharpe ratio of the dynamic trading system, which depends on a sequence of actions $a_1, \ldots, a_t$.

Dynamic portfolio management by maximizing Sharpe ratio and extensions

Instead of only considering one single asset, a common and more reliable practice is considering a portfolio of assets, and thus portfolio management is one important topic in the finance literature. For the supervised learning by Eq. (22), its extension can be made simply by considering $I^p_{j,t}(XF^q_t, \{I^p_{j,t-\tau}\}^q_{t=1}, \Theta_j), j = 1, \ldots, k$ with each in the format of Eq. (22), and learning is made by minimizing the total sum $\sum_j E_2(\Theta_j)$. Simply, we get the training signals $I^p_{j,1}, \ldots, I^p_{j,t}$ per asset individually. Still, further studies are needed on how to get the training signals bases on the whole portfolio of assets. Conceptually, extension of reinforcement learning to multiple assets is rather straightforward too. However, both the set S of environment states and the set A of possible actions increase rapidly, which makes learning a large table $Q(s_t, a_t)$ seriously suffer the problem of a small sample size. Thus, it becomes more critical to get $a_1, \ldots, a_t$ to be approximately replaced by $\{I^p_{j,t}(XF^q_t, \{I^p_{j,t-\tau}\}^q_{t=1}, \Theta_j)\}^k_{j=1}$ in evaluating the reward r_{t+1} (Moody et al. 1998; Moody and Saffell 2001). Similar to supervised learning, one direction for tackling the problem of a small sample size is incorporating with learning regularization.

Alternatively, another direction to pursuit portfolio management is exploring the road pioneered by the Markowitz portfolio theory (Markowitz 1952), see Box 2 in Fig. 2. By this theory, the return of an investment portfolio is the proportion-weighted combination of the constituent assets' returns, while the portfolio volatility is a function of the correlations between the component assets. The portfolio expected return is maximized subject to a given amount of portfolio risk, or equivalently risk is minimized for a given level of expected return. Moreover, the Markowitz mean–variance scheme also leads to the suggestion of Sharpe ratio (Sharpe 1966, 1994), which is typically used to evaluate the performance of a portfolio.

In both the standard Markowitz mean–variance scheme and Sharpe ratio approach, a risk is defined as the return variance, which has been subsequently realized that the variance is not an appropriate measure because it counts the positive fluctuation above the expected returns (also called *upside volatility*) as a part of the risk. See Box 4 in Fig. 2; the *downside risk* thus becomes a topic to study. Markowitz (1959) counts the volatility below the expected returns only. Fishburn (1977) makes a mean-risk analysis with risk associated with below-target returns and proposes a more sophisticated measure of risk associated with below-target return, which has been further refined by Sortino and Meer (1991). Basically, this downside risk is the volatility of return below the minimal acceptable return (also called *target return G*).

$$\mathrm{down}V_\gamma(G) = \int_{-\infty}^{G} (G - r)^\gamma \, \mathrm{d}F(\mathrm{r}) \tag{25}$$

Moreover, the downside risk of a single asset has been extended into the following covariance (Hung et al. 2000, 2003):

$$\boldsymbol{D} = \left[d_{i,j}\right], d_{i,j} = \int\limits_{-\infty}^{G}\int\limits_{-\infty}^{G} (G - r_i)^{\frac{\gamma}{2}} (G - r_j)^{\frac{\gamma}{2}} p(r_i, r_j) \, \mathrm{d}r_i \mathrm{d}r_j, \tag{26}$$

for the returns r_j, $j = 1, \ldots, k$ of multiple assets. Also, we have the following matrix for the upside volatility:

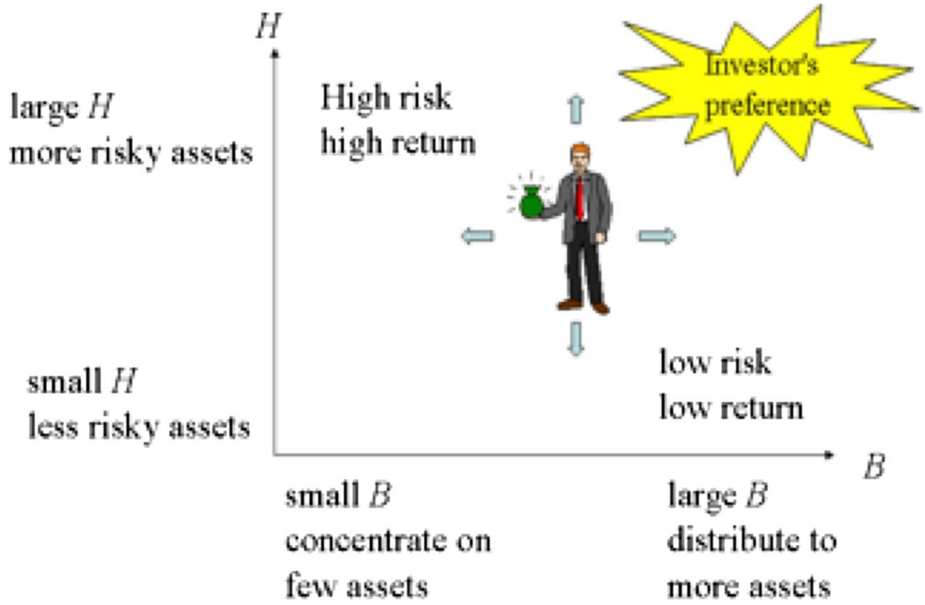


Fig. 4 Use of the parameters *H*, *B*

$$\boldsymbol{U} = \left[u_{i,j}\right],\ u_{i,j} = \int_{G}^{+\infty}\int_{G}^{+\infty} (r_i - G)^{\frac{\gamma}{2}} (r_j - G)^{\frac{\gamma}{2}} p(r_i, r_j)\, \mathrm{d}r_i \mathrm{d}r_j. \tag{27}$$

The sprit of the Markowitz theory and the Shape ratio, i.e., maximizing the expected returns while minimizing the risk, is reasonably modified into one extended Sharpe ratio featured by maximizing both the expected returns and the upside volatility while minimizing the downside risk; see Box 5 in Fig. 2. In Hung et al. (2000, 2003), this generalization is implemented by the following maximizaon:

$$\begin{aligned} &\operatorname*{Max}_{\boldsymbol{w}} \left[\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{r} + H\boldsymbol{w}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{w}} + B\boldsymbol{w}^{\mathrm{T}}(1 - \boldsymbol{w})\right], 1 = [1, \ldots, 1]^{\mathrm{T}}, \\ &\boldsymbol{r} = \left[r_1, \ldots, r_k\right]^{\mathrm{T}}, \mathbf{w} = \left[w_1, \ldots, w_k\right]^{\mathrm{T}}, \quad \sum_{i=1}^{k} w_i = 1, \quad w_i \geq 0. \end{aligned} \tag{28}$$

As shown in Fig. 4, we use the parameters *H*, *B* to adapt the investor's preference. The parameter *H* represents a strength of maximizing upside volatility and *B* represents a strength of diversification or regularization. The term $\boldsymbol{w}^{\mathrm{T}}(1 - \boldsymbol{w})$ is a diversification term that reaches its minimum when one w_i is 1 and others are 0, and its maximum when all the elements $\boldsymbol{w}$ are equal.

It has been experimentally shown that this generalization of Sharpe ratio can effectively reduce the risk while obtaining great returns, in comparison with the standard Markowitz mean–variance scheme and Sharpe ratio. Moreover, investors expect a constant return with a minimum downward risk, for which we can simply set $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{r} = r_{\mathrm{spec}}$, while the others expect a maximum return under a constant downward risk, for which we can simply set $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{w} = v_{\mathrm{spec}}$.

In Sect III(C) of Xu (2001), several developments have been proposed along this direction. First, a more practical scenario is considered, featured with a portfolio of risk securities with returns $r_{j,t}, j = 1, \ldots, k$, a risk-free bond with return r^f and transaction cost with a rate r_c. That is, $r_t = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{r}$ is replaced by

$$\begin{aligned} r_t &= (1-\alpha_0)r^f + \alpha_0 \sum_{j=1}^{k} \left[w_{j,t}r_{j,t} - r_c \sum_{j=1}^{k} \left| w_{j,t} - w_{j,t-1} \right| \left(1 + r_{j,t}\right) \right] \\ &= (1-\alpha_0)r^f + \alpha_0 \left[\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{r}_t - r_c \delta \boldsymbol{w}_t^{\mathrm{T}}(1 + \boldsymbol{r}_t) \right], \alpha_0 > 0, \mathbf{w}_t^{\mathrm{T}}1 = 1, \\ \delta \boldsymbol{w}_t &= \left[\left| w_{1,t} - w_{1,t-1} \right|, \ldots, \left| w_{k,t} - w_{k,t-1} \right| \right]^{\mathrm{T}}, \end{aligned} \tag{29}$$

where each $w_{j,t}$ may be nonnegative as in Eq. (28). In this case, short of a risk security is not permitted but borrowing from the risk-free bond is allowed, i.e., we can have $1-\alpha_0 < 0$. Also, we may allow a negative $w_{j,t}$, i.e., short of a risk security is permitted.

Second, instead of considering $E\boldsymbol{w}^{\mathrm{T}}\boldsymbol{r} = \boldsymbol{w}^{\mathrm{T}}E\boldsymbol{r}$ and $E\left[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{r} - E\boldsymbol{w}^{\mathrm{T}}\boldsymbol{r}\right]\left[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{r} - E\boldsymbol{w}^{\mathrm{T}}\boldsymbol{r}\right]^{\mathrm{T}}$ for the expected return and its volatility, we compute their estimations directly from samples $R_T = \{r_t,\ t = 1,\ \ldots,\ T\}$ within a time window. Accordingly, it follows from Eq. (25) that we get the counterpart of Eq. (28) as follows:

$$\begin{aligned} Sp &= \frac{M(R_T)}{\sqrt[\gamma]{V_G^D(R_T)}} + \beta_V \frac{\sqrt[\gamma]{V_G^U(R_T)}}{\sqrt[\gamma]{V_G^D(R_T)}} + \beta_{\boldsymbol{w}} D(\boldsymbol{w}),\ M(R_T) = \frac{1}{T}\sum_{t=1}^{\mathrm{T}} r_t, \\ V_G^D(R_T) &= \frac{1}{\#(r_t \le G)} \sum_{r_t \le G} (G - r_t)^{\gamma},\ V_G^U(R_T) = \frac{1}{\#(r_t > G)} \sum_{r_t > G} (r_t - G)^{\gamma}, \end{aligned} \tag{30}$$

where $\#S$ denotes the cardinality of the set S, and the parameter $\beta_V, \beta_{\boldsymbol{w}}$ are the counterparts of H, B in Eq. (28). Moreover, $D(\boldsymbol{w})$ is a diversification term that reaches its minimum when one w_i is 1 and the others are 0, and reaches its maximum when all the elements $\boldsymbol{w}$ are equal. There could be several choices for $D(\boldsymbol{w})$. One example is $\boldsymbol{w}^{\mathrm{T}}(1-\boldsymbol{w})$ in Eq. (28) or equivalently $-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$. One other example is

$$D(\boldsymbol{w}) = -\sum_{j=1}^{k} w_{j,t} \ \ln\, w_{j,t}, \sum_{j=1}^{k} w_{j,t} = 1, \quad w_{j,t} \ge 0. \tag{31}$$

Moreover, $M(R_T) \Big/ \sqrt[\gamma]{V_G^D(R_T)}$ is a ratio which is also an improvement over $\boldsymbol{w}^{\mathrm{T}}E\boldsymbol{r}/\boldsymbol{w}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{w}$ in Eq. (28), and actually $\boldsymbol{w}^{\mathrm{T}}E\boldsymbol{r}/\boldsymbol{w}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{w}$ is not really a ratio. Third, instead of directly searching the parameters $\alpha_0, \boldsymbol{w}_t$, we may let

$$\begin{aligned} \alpha_0 &= e^{-g(\boldsymbol{r}_t,\psi)}, w_{j,t} = \frac{e^{f^{(j)}(\boldsymbol{r}_t,\varphi)}}{\sum_{i=1}^{k} e^{f^{(i)}(\boldsymbol{r}_t,\varphi)}}, \\ f(\boldsymbol{r}_t,\varphi) &= \left[f^{(j)}(\boldsymbol{r}_t,\varphi), \ldots, f^{(j)}(\boldsymbol{r}_t,\varphi) \right]^{\mathrm{T}}, \end{aligned} \tag{32}$$

with $g(\boldsymbol{r}_t, \psi), f(\boldsymbol{r}_t, \varphi)$ implemented by neural networks, e.g., an ENRBF network. In the next section, we will show that a portfolio of security returns $\boldsymbol{r}_t$ may also be modeled by a temporal extension of arbitrage pricing theory such that $\boldsymbol{r}_t$ is mapped into inner factor $\boldsymbol{y}_t$ with a much lowered dimension. Instead of depending on the security returns $\boldsymbol{r}_t$, we use $\boldsymbol{y}_t$ to replace $\boldsymbol{r}_t$ in Eq. (28) for a further improvement.

Following the extension proposed in Xu (2001), most of the above addressed extensions have been investigated together with detailed algorithm, experiments on real market data and comparative studies (Chiu and Xu 2002b, 2003, 2004b). Still, at the end of Sect III(C) in Xu (2001), there was one briefly introduced idea that has not been further investigated yet. Here, some further details are addressed.

In Eq. (30) and also in Eq. (28), as well as in the existing studies on the Markowitz portfolio optimization and the Sharpe ratio, the expected return and volatilities are nonparametric estimates directly from samples $R_T = \{\boldsymbol{r}_t, t = 1, \ldots, T\}$. To capture a temporal dependence better, one idea is using an ARCH or GARCH model to describe a sequence $\{r_t, t = 1, \ldots, T\}$ of the portfolio return $r_t = \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t$; see Box in Fig. 2. It follows from Eq. (3) that we have

$$r_{t+1} = a_0 + \sum_{j=1}^{q} a_j r_{t+1-j} + \sigma_t \varepsilon_t, \varepsilon_t \sim^{\text{i.i.d.}} G(\varepsilon|0,1), \text{and } \sigma_t = \sigma_t^2(\vartheta) \text{ by Eq. (3)}. \tag{33}$$

Taking the expectation and separating the first term from the rest, as well as approximately considering $E\boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t \approx a_1 \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t$, we further get

$$\begin{aligned}
&Er_{t+1} = a_0 + \sum_{j=1}^{q} a_j E r_{t+1-j} = a_1 E\boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t + E\hat{r}_{t-1} \approx a_1 \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t + E\hat{r}_{t-1}, \\
&\sigma_{t+1}^2 = \sigma_0^2 + \sum_{i=1}^{q} \beta_i \varepsilon_{t+1-i}^2 + \sum_{j=1}^{p} \omega_j \sigma_{t+1-j}^2 = \beta(\boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t - r_t^{AR})^2 + \hat{\sigma}_t^2, r_t^{AR} = a_0 + \sum_{j=1}^{q} a_j r_{t-j}, \\
&E\hat{r}_{t-1} = a_0 + \sum_{j=2}^{q} a_j E r_{t+1-j}, \hat{\sigma}_t^2 = \sigma_0^2 + \sum_{i=2}^{q} \beta_i \varepsilon_{t+1-i}^2 + \sum_{j=1}^{p} \omega_j \sigma_{t+1-j}^2,
\end{aligned} \tag{34}$$

from which we get the following GARCH-based Shape ratio

$$J(\boldsymbol{w}_t) = \frac{Er_{t+1}}{\sigma_{t+1}} = \frac{a_1 \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t + E\hat{r}_{t-1}}{\beta_1 (\boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t - r_t^{AR})^2 + \hat{\sigma}_t^2}. \tag{35}$$

Given the GARCH model and the past $Er_{t-j}, r_{t-j}, \quad j = 1, \ldots, k$, we have $E\hat{r}_{t-1}$, $\hat{\sigma}_t^2$, r_t^{AR}, a_1, β_1 available. As $\boldsymbol{r}_t$ is obtained, we compute the gradient of $J(\boldsymbol{w}_t)$ and update

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} + \eta \nabla_{\boldsymbol{w}_t} J(\boldsymbol{w}_t), \text{for a learning step size } \eta > 0. \tag{36}$$

Then, we get $\varepsilon_t^2 = (\boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{r}_t - r_t^{AR})^2$ and update $a_i^{\mathrm{new}} = e^{c_1^{\mathrm{new}}}, c_i^{\mathrm{new}} = c_i^{\mathrm{old}} - \eta \frac{\mathrm{d}\varepsilon_t^2}{\mathrm{d}c_i^{\mathrm{old}}}$, for $i = 0, 1, a_j^{new} = a_j^{old} - \eta \frac{d\varepsilon_t^2}{da_j^{old}}, \quad$ for $j = 2, \ldots, q$.

Also, we update the parameters ϑ in the same way as one standard GARCH solving approach. Next, we use Eq. (36) for updating $\boldsymbol{w}_{t+1}$ again.

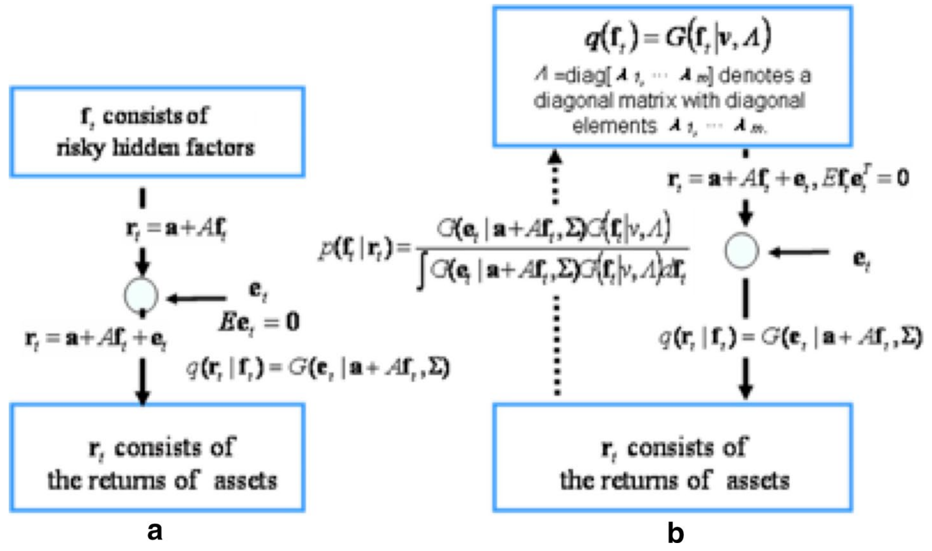

$q(\mathbf{f}_t)=G(\mathbf{f}_t|\mathbf{v},\Lambda)$
Λ =diag[$\Lambda_1, \cdots \Lambda_m$] denotes a diagonal matrix with diagonal elements $\Lambda_1, \cdots \Lambda_m$
$\mathbf{f}_t$ consists of risky hidden factors
$\mathbf{r}_t=\mathbf{a}+A\mathbf{f}_t$
$\mathbf{e}_t$
$E\mathbf{e}_t=\mathbf{0}$
$\mathbf{r}_t=\mathbf{a}+A\mathbf{f}_t+\mathbf{e}_t$
$q(\mathbf{r}_t|\mathbf{f}_t)=G(\mathbf{e}_t|\mathbf{a}+A\mathbf{f}_t,\Sigma)$
$p(\mathbf{f}_t|\mathbf{r}_t)=\frac{G(\mathbf{e}_t|\mathbf{a}+A\mathbf{f}_t,\Sigma)G(\mathbf{f}_t|\mathbf{v},\Lambda)}{\int G(\mathbf{e}_t|\mathbf{a}+A\mathbf{f}_t,\Sigma)G(\mathbf{f}_t|\mathbf{v},\Lambda)d\mathbf{f}_t}$
$\mathbf{r}_t=\mathbf{a}+A\mathbf{f}_t+\mathbf{e}_t, E\mathbf{f}_t\mathbf{e}_t^T=\mathbf{0}$
$\mathbf{e}_t$
$q(\mathbf{r}_t|\mathbf{f}_t)=G(\mathbf{e}_t|\mathbf{a}+A\mathbf{f}_t,\Sigma)$
$\mathbf{r}_t$ consists of the returns of assets
$\mathbf{r}_t$ consists of the returns of assets
a
b

Fig. 5 Arbitrage pricing theory and factor analysis

Market modeling: APT theory and temporal factor analysis

Arbitrage pricing theory and factor analysis's incapability

Beyond only optimizing the outcome by investing a portfolio of multiple assets, the Markowitz mean–variance scheme also leads to the linear modeling of the market. The most famous one is the well-known capital asset pricing model (CAPM) (Sharpe 1964). However, the CAPM is criticized as being not sufficient to describe market behavior merely via one endogenous factor.

Under the name of arbitrage pricing theory (APT), Ross (1976) proposed the following linear model of multiple hidden or endogenous factors:

$$\boldsymbol{r}_t=\boldsymbol{a}+A\boldsymbol{f}_t+\boldsymbol{e}_t, A=[a_{ij}], \mathbf{r}_t=[r_{1,t},\ldots,r_{k,t}]^{\mathrm{T}}, \boldsymbol{f}_t=[f_{1,t},\ldots,f_{n,t}]^{\mathrm{T}}, \boldsymbol{e}_t=[e_{1,t},\ldots,e_{k,t}]^{\mathrm{T}}. \tag{37}$$

As illustrated in Fig. 5a, $\boldsymbol{r}_t$ consists of the returns of k assets in this market, $\boldsymbol{f}_t$ consists of m risky hidden factors that will affect the rate of returns on all assets by different degrees of sensitivity and a_{ij} is the sensitivity of the ith asset to factor j, also called factor loading, Moreover, each element of $\boldsymbol{e}_t$ is the risky asset's idiosyncratic random shock with mean zero, and each element of $\boldsymbol{a}$ is a constant part of the corresponding risky asset.

Since its inception, the APT has attracted a considerable interest as a tool for interpreting investment results and controlling portfolio risk. However, the APT has been accepted by the investment community, but is not as popular as the CAPM. The reason largely relates to APT's serious drawback, namely, its implementation is difficult due to the lack of specificity regarding the nature of the factors that systematically affect asset returns. As outlined in Sect. I of (Xu 2001), typically three types of approaches have been applied for the APT implementation.

Most of the studies are featured with $\boldsymbol{f}_t$ given by the so-called *fundamental factor*s, i.e., historic time series of a set of macroeconomic or fundamental indexes. With the hidden factors chosen, the problem becomes a typical multivariate linear regression problem: $\boldsymbol{r}_t = \boldsymbol{a} + A\boldsymbol{f}_t + \boldsymbol{e}_t$. However, choosing these *fundamental factors* is not an easy task. Chen et al. (1986) chose five macroeconomic factors, including surprises in GDP, inflation, investor confidence, and yield curve. Also, others consider index or spot or future market price, e.g., short-term interest rate, a diversified stock index, oil price, gold or precious metal prices, and currency exchange rate in place of macroeconomic factors. With efforts over decades, little progress has been achieved on identifying the number and nature of these *fundamental factors*. Many researchers believe that this issue is essentially empirical in nature, because the factors change over time and between economies.

There have been also efforts under the name of the *cross-sectional approaches* that observes the correlations of all the assets of $\boldsymbol{r}_t$ to each of the hidden factor in $\boldsymbol{f}_t$ by a certain period, resulting in estimates of elements of A that reflect the assets' sensitivities to these hidden factors. Then, the task is to estimate $\boldsymbol{f}_t$ upon $\boldsymbol{r}_t$ and A, which is typically handled as a linear *cross-sectional* regression and solved by the least square error method in the literature of economics and finance. In Sect. I of Xu (2001), it is formulated as an inverse mapping problem, a topic that has been widely studied in the neural network and machine learning literature.

Observation of an implementation of the least square error method actually shows that the residuals $\boldsymbol{e}_t$ are uncorrelated among the elements and also with the factors $\boldsymbol{f}_t$ and that each element of $\boldsymbol{e}_t$ reflects a collective effect of many random noise, that is, we have $E\boldsymbol{f}_t\boldsymbol{e}_t^{\mathrm{T}} = 0$ and also $q(\boldsymbol{r}_t|\boldsymbol{f}_t)$ as shown by the top-down pathway on the right part of Fig. 5b. An inverse of the top-down path is a bottom-up path on the left part of Fig. 5b, for which the optimal solution is the following Bayesian inverse:

$$p(\boldsymbol{f}_t|\boldsymbol{r}_t) = \frac{G(\boldsymbol{e}_t|\boldsymbol{a} + A\boldsymbol{f}_t, \Sigma)q(\boldsymbol{f}_t)}{\int G(\boldsymbol{e}_t|\boldsymbol{a} + A\boldsymbol{f}_t, \Sigma)q(\boldsymbol{f}_t)d\boldsymbol{f}_t}. \tag{38}$$

Here, we encounter a probabilistic structure $q(\boldsymbol{f}_t)$ of hidden factors. Approximately, if only considering its statistics up to the second order, $q(\boldsymbol{f}_t)$ is approximated by a Gaussian $G(\boldsymbol{f}_t|\nu, \Lambda)$ as shown in Fig. 5b. In such a case, we have the following analytical solution:

$$\hat{\boldsymbol{f}}_t = \int \boldsymbol{f}_t p(\boldsymbol{f}_t|\boldsymbol{r}_t)\,d\boldsymbol{f}_t = \left(A^{\mathrm{T}}\Sigma^{-1}A + \Lambda^{-1}\right)\left[A^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{r}_t - \boldsymbol{a}) + \Lambda^{-1}\nu\right], \tag{39}$$

which returns to a least square error solution when there is no information about $q(\boldsymbol{f}_t)$ for which we may simply set $\Lambda = 0$, $\nu = 0$.

Similar to the first approach, the second approach is also essentially empirical in nature, which needs not only a manual help to identify the number and nature of hidden factors, but also at least an enough long period of historic data about factors for estimating of elements of A. Moreover, getting elements of A by the correlations between $\boldsymbol{f}_t$ upon $\boldsymbol{r}_t$ actually imposes additional constraints on the values that A may take. The second approach is supplementary to the first approach, but it still cannot get rid of the nature that the factors are chosen heuristically and even rather arbitrarily. We may regard that the second approach actually consists of two steps. First, estimation of elements of A bases on a period historic data of macroeconomic or fundamental indexes takes the same role of the first approach or even just an implementation of the first approach. Second, we estimate $\boldsymbol{f}_t$ upon $\boldsymbol{r}_t$ and A, e.g., typically by Eq. (39).

The third type of efforts are called factor-analytic approach, attempting to use a statistical approach called factor analysis (FA) to get both the unknown and the unknown factors estimated from the observed return series $\{\boldsymbol{r}_t\}$. There is no need of external heuristics, and thus it seems more appealing. As shown in Fig. 5b, an FA model comes from modifying Fig. 5a with an additional structure that $\boldsymbol{f}_t$ comes from a Gaussian $G(\boldsymbol{f}_t|\nu, \Lambda)$ with a diagonal Λ or even $\Lambda = I$. Unfortunately, empirical tests showed that factor analysis does not explain economic variables well. As addressed in Sect. I of Xu (2001), some incapability of factor analysis mainly comes from two kinds of intrinsic indeterminacy. One is the rotation indeterminacy, i.e.,

$$\text{if} A, \boldsymbol{f}_t \text{ is a solution, } A\varphi^{\mathrm{T}}, \varphi\boldsymbol{f}_t \text{ is also a solution for any rotation matrix } \varphi, \tag{40}$$

while such a rotation may lead to a solution far from the correct one. The other comes from an intrinsic indeterminacy of an appropriate number of factors, while the selection of a correct number of factors is essential to the performance of using the APT model. Usually, it is set by a rule of thumb. Actually, factor analysis also suffers other types of indeterminacy. One is any rescaling $D\boldsymbol{f}_t$ of a solution $\boldsymbol{f}_t$ is still a solution for a diagonal matrix D, which is not critical because it reserves the waveform of each element in $\boldsymbol{f}_t$. The other is additive indeterminacy, i.e., A, Λ, Σ and A^*, Λ^*, $\Sigma^*$are both the solutions as long as $A\Lambda \mathrm{A}^{\mathrm{T}}+\Sigma=A^*\Lambda^*A^{*\mathrm{T}}+\Sigma^*$. However, the effect of this indeterminacy can be reduced significantly when $\Sigma=\sigma^2 I$. Therefore, our attention should be mainly on the first two key challenges, namely, removing the rotation indeterminacy by Eq. (40) and determining an appropriate number of factors.

The first challenge has been seldom considered by the APT studies in the fields of economics and finance, while there are some efforts on the second challenge, i.e., determining an appropriate number of factors with the help of statistical testing. The simplest one is making maximum likelihood factor analysis (MLFA) followed by the likelihood ratio (LR) test, shortly MLFA-LR. Empirical evidences show that the minimum number of factors accepted by the LR test tends to increase with the number of securities. Alternatively, Chamberlain and Rothschild (1983) suggest analyzing eigenvalues of the population covariance matrix, shortly eigenvalue approach. Still, Brown (1989) empirically found that this approach biases toward too few factors and the result consistent with one factor may be equally consistent with multiple equally weighted factors.

On one hand, being essentially empirical in nature, both the fundamental factor-based approaches and the cross-sectional approaches rely on pre-knowledge or external beliefs to choose the factors heuristically, in lack of consensus and consistency over what should be the real factors in APT. On the other hand, the implementation of factor analysis suffers the rotation indeterminacy by Eq. (40) and the difficulty of determining an appropriate number of factors. These problems incur for criticisms on the APT theory, e.g., see Dhrymes et al. (1984); Abeysekera and Mahajan (1987).

Instead of doubting the incorrectness of the APT theory, our understanding is that the APT theory is correct but incomplete. The APT suggests to model a market at no arbitrage equilibrium by a linear model, which is justifiable. However, this theory is incomplete because this linear model cannot be uniquely or even reasonably specified merely from the observed return series $\{\boldsymbol{r}_t\}$. To complete the theory, further specification should be imposed on the components of this model. The fundamental factor-based approaches fix the hidden factors by heuristically and empirically picking a set of macroeconomic or fundamental indexes, which removes the indeterminacy but leaves the difficult questions on how to choose these factors and whether the factors should come directly from macroeconomic or fundamental indexes. The cross-sectional approaches aim at estimating A, which leaves the difficult question on how A can be estimated correctly. To get A by the assets' sensitivities to these hidden factors, we still need to heuristically and empirically pick a set of macroeconomic or fundamental indexes, Finally, the FA model is also unable to remove the incompleteness of the APT, because imposing an additional Gaussian $G\left(\boldsymbol{f}_t|\nu, \Lambda\right)$ is still not enough to remove the critical indeterminacy by Eq. (40). In a summary, the original APT (Ross 1976) is reasonable but incomplete, and further efforts should explore how to add certain structure to remove or remedy the incompleteness.

Temporal factor analysis and temporal APT

The famous CAPM model is featured by one factor that is not a manually chosen exogenous macroeconomic or fundamental index but an invisible and intrinsic market indicator. The APT was motivated by following the basic sprit of CAPM to answer the criticism that merely one factor is not enough to describe the market behavior. However, implementing APT by manually picking macroeconomic or fundamental indices actually deviates from the original motivation. Encouragingly, the direction of FA implementation is still consistent with the original motivation that seeks intrinsic factors, and thus we further proceed along this direction. Keeping Eq. (37), we extend the Gaussian structure $G\left(\boldsymbol{f}_t|\nu, \Lambda\right)$ into a better structure such that the indeterminacy by Eq. (40) or the incompleteness of the FA model can be removed or at least remedied.

Temporal factor analysis (TFA) is such a further development of FA; see Box 1 in Fig. 3. The early study was started in 1997, firstly introduced briefly by Xu (1997) and further addressed in Xu (2000) (this manuscript actually reached the editorial office also in 1997). See Box 2 in Fig. 3: the key idea is modifying Eq. (37) as follows:

$$
\begin{aligned}
&\boldsymbol{r}_t = \boldsymbol{a} + A\boldsymbol{f}_t + \boldsymbol{e}_t, A = \left[a_{ij}\right], \mathbf{r}_t = [r_{1,t}, \ldots, r_{k,t}]^{\mathrm{T}},\\
&\boldsymbol{f}_t = \left[f_{1,t}, \ldots, f_{n,t}\right]^{\mathrm{T}}, \boldsymbol{e}_t = \left[e_{1,t}, \ldots, e_{k,t}\right]^{\mathrm{T}}, \mathrm{E}\,\mathbf{f}_t\boldsymbol{e}_t^{\mathrm{T}} = 0,\\
&\boldsymbol{f}_t = B\boldsymbol{f}_{t-1} + \varepsilon_t, B = \mathrm{diag}[b_1, \ldots, b_m] \neq bI \text{ with } b \neq 0,\\
&\quad \varepsilon_t \sim G(\varepsilon_t|0, \Lambda) \text{ with a diagonal } \Lambda, \mathrm{E}\,\mathbf{f}_{t-1}\varepsilon_t^{\mathrm{T}} = 0.
\end{aligned}
\tag{41}
$$

That is, the first-order autoregressive dependence is added to each factor in $\boldsymbol{f}_t$ via B, and Eq. (41) returns to FA by Eq. (37) when $B = 0$.

It is this temporal dependence that removes the rotation indeterminacy by Eq. (40); see Sect IV (A) in Xu (2000) and Sect. II in Xu (2002). Roughly, the following points may be understood:

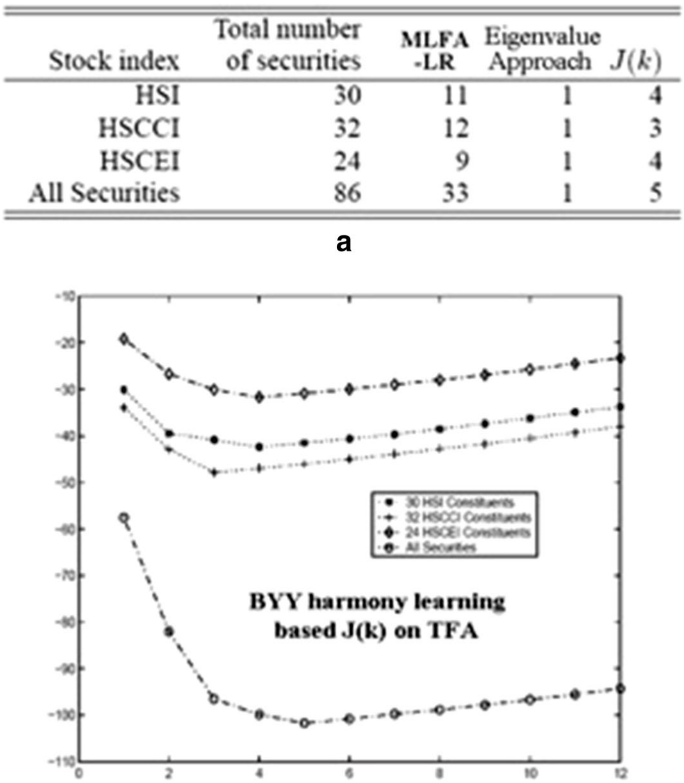
- For any diagonal matrix D, we have $A\boldsymbol{f} = \tilde{A}\tilde{\boldsymbol{f}}, \tilde{A} = AD, \tilde{\boldsymbol{f}} = D^{-1}\boldsymbol{f}$, which keeps the format $\boldsymbol{r}_t = \boldsymbol{a} + A\boldsymbol{f}_t + \boldsymbol{e}_t$ unchanged and also the elements of $\tilde{\boldsymbol{f}}$ remain mutually independent. i.e., Equation (37) has an indeterminacy of unknown scaling on factors of $\tilde{\boldsymbol{f}}$. Thus, we may simply consider $\boldsymbol{f}_t \sim G(\boldsymbol{f}_t|0, I)$. For any rotation matrix ϕ with $\varphi^{\mathrm{T}}\varphi = I$, we have $A\boldsymbol{f} = \tilde{A}\tilde{\boldsymbol{f}}$, and $\tilde{A} = A\varphi^{\mathrm{T}}, \tilde{\boldsymbol{f}} = \varphi\boldsymbol{f}$ with $\tilde{f}_t \sim G\left(\tilde{f}_t|0, I\right)$. That is, Eq. (37) has also an indeterminacy of unknown rotation on factors $\tilde{\boldsymbol{f}}$.
- For any diagonal matrix D, we also have $D^{-1}\boldsymbol{f}_t = D^{-1}BDD^{-1}\boldsymbol{f}_{t-1} + D^{-1}\varepsilon_t$ and $\tilde{\boldsymbol{f}}_t = B\tilde{\boldsymbol{f}}_{t-1} + \tilde{\varepsilon}_t,$, where $\tilde{\varepsilon}_t = D^{-1}\varepsilon_t$ comes from $G\left(\tilde{\varepsilon}_t\middle|0, D^{-1}\Lambda D^{-1}\right)$ and $D^{-1}\Lambda D^{-1}$ is still diagonal. That is, Eq. (41) still has an indeterminacy of unknown scaling on factors $\tilde{\boldsymbol{f}}$. Again, we may consider $\varepsilon_t \sim G(\varepsilon_t|0, I)$. For any rotation matrix ϕ with $\phi^{\mathrm{T}}\phi = I$, we have $\tilde{\boldsymbol{f}}_t = \tilde{B}\tilde{\boldsymbol{f}}_{t-1} + \tilde{\varepsilon}_t$ with $\tilde{\varepsilon}_t \sim G(\tilde{\varepsilon}_t|0, I)$, while $\tilde{B} = \varphi B\varphi^{\mathrm{T}}$ is no longer diagonal and even B is diagonal. If $\tilde{B} = \varphi B\varphi^{\mathrm{T}}$ is required to be diagonal, the only rotation matrix is $\phi = I$ and thus the rotation indeterminacy is removed.

Still there is an indeterminacy of unknown scaling on factors of $\tilde{\boldsymbol{f}}$, but it will not change the waveform of $f_{1,t}$, …, $f_{n,t}$. Also, we may normalize each factor to remove such indeterminacy.

In Xu (2001), the TFA by Eq. (41) is thus suggested as a refinement of the original APT theory, by which the original part of APT is kept without modification, while a temporal structure $\boldsymbol{f}_t = B\boldsymbol{f}_{t-1} + \varepsilon_t$ is added such that the incompleteness caused by the rotation indeterminacy has been removed. Such a refinement may be called temporal APT in a sense that temporal relation is taken into consideration of market modeling. That is, a static equation by Eq. (37) is not enough to describe a market equilibrium, but a temporal structure should be an important ingredient of a market equilibrium.

Why is an AR model of merely order one $\boldsymbol{f}_t = B\boldsymbol{f}_{t-1} + \varepsilon_t$ considered as this temporal structure? First, we consider that hidden factors $\boldsymbol{f}_t$ are driven by Gaussian noise $\varepsilon_t \sim G(\varepsilon_t|0, \Lambda)$, following a general consensus that the noisy component in most econometric and statistical models is Gaussian distributed. The rationale comes from the central limit theorem which implies that the compounding of a large number of unknown distributions will be approximately normal. Second, the first-order AR model can be attributed to the weak form of efficient market hypothesis (EMH), that is, stock price today is conditionally independent of all previous prices given the price of yesterday. Third, though observable economic indices are seldom independent, it cannot rule out that hidden factors that denominate a market equilibrium are mutually independent. Instead, independent factors may help to make market equilibrium simpler.

As addressed in the previous subsection, past efforts on determining an appropriate number of factors have not provided much support on the APT. For one example, the MLFA-LR test shows that the number of factors tends to increase with the number of securities. For another example, the identification via eigenvalue approach (Chamberlain and Rothschild 1983) biases toward a smaller factor number. In one IJCNN 02 paper (Chiu and Xu 2002a), empirical tests on Hong Kong stock market data show not only that these two unfavorable biases are again observed, but also that the TFA-based APT can provide a reasonable answer to the number of factors in the Hong Kong stock market. As shown in Fig. 6, the number of factors identified by MLFA-LR test varies as the



Stock index	Total number of securities	MLFA -LR	Eigenvalue Approach	$J(k)$
HSI	30	11	1	4
HSCCI	32	12	1	3
HSCEI	24	9	1	4
All Securities	86	33	1	5

a

b

Fig. 6 Comparison on finding the number of factors identified by MLFA-LR test, eigenvalue approach and BYY harmony learning-based TFA

numbers of securities, while the number of factors identified by the eigenvalue approach is always 1. In contrast, BYY harmony learning based TFA stably identifies four or five factors regardless of the numbers of securities, which is quite consistent with the number identified via heuristic empirical analysis, e.g., in Chen et al. (1986).

The above introduced nature of TFA and preliminary studies suggest that there may need a renewed interest in the literature of finance and economics to further investigate APT and its further developments. To consider which topics to pursue, it is helpful to observe the differences of TFA from related methods.

First, $\boldsymbol{f}_t = B\boldsymbol{f}_{t-1} + \varepsilon_t$ in Eq. (41) is actually a special type of the first-order vector AR (VAR). Being different from the conventional VAR that are used for capturing linear interdependencies among multiple time series (Sims 1980; Engle and Granger 1987), the TFA captures the interdependencies among multiple time series by $\boldsymbol{r}_t = \boldsymbol{a} + A\boldsymbol{f}_t + \boldsymbol{e}_t$ and temporal dependences by $\boldsymbol{f}_t = B\boldsymbol{f}_{t-1} + \varepsilon_t$. As addressed in Sect. 3.2.1 in Xu (2012), it is more efficient to separately treat these two types of dependences.

Second, if we do not constrain B,Λ to be diagonal, Eq. (41) becomes a general state–space model (SSM) or a linear dynamical system (LDS), which has been widely studied in the literature of control theory and signal processing. As outlined in Sect. 5.2.1 of Xu (2012), in a period that is more or less the same as the studies on TFA (Xu 1997; 2000), there was a renewed interest on a general LDS, featured by using the EM algorithm for parameter estimation under the ML learning (Ghahramani and Hinton 2000).

Accordingly, this EM algorithm was originally derived in the early 1980s and re-introduced in the early 1990s (Shumway and Stoffer 1991). Neither these studies suggest using the LDS as a further development of APT, nor the notorious rotation indeterminacy in Eq. (40) has been taken into consideration. On the contrary, more problems of indeterminacy than the FA are actually incurred in this general LDS model due to many extra free parameters, which makes identifiability even worse. For an example, applied to radar automatic target recognition based on high-resolution range profile, it has been shown in Wang et al. (2011) that the recognition performance of the general LDS is actually even inferior to that of the FA, while TFA obtains better performances than the FA.

Third, many efforts have been made on determining the factor number of FA in the literature of statistics and machine learning, typically in a two-stage implementation. The first stage uses the EM algorithm to make the ML learning for unknown parameters in the FA while the second stage selects an appropriate number of factors with help of a model selection criterion. In Tu and Xu (2011), a systematic comparative investigation has been made on a number of typical model selection criteria, including not only Akaike's AIC, Schwarz's BIC, Bozdogan's CAIC, Hannan–Quinn criterion, but also recent Minka's PCA criterion, Kritchman and Nadler's tests, and Perry and Wolfe's rank, as well as the criterion obtained from the BYY harmony learning theory (Xu 2001).

As discussed above, there is not really a need to further consider the relations to VAR and LDS. Instead, further explorations may start from continuing the study in the IJCNN02 paper (Chiu and Xu 2002b) and proceed to clarify the following issues:

- Does using one of the above model selection criteria in a two-stage implementation improve the number of FA factors identified by the MLFA-LR test and the eigenvalue approach? If yes, does this improvement help the FA-based implementation of APT, even still suffering the rotation indeterminacy by Eq. (40).
- Still using one of the above model selection criteria in a two-stage implementation, how much improvement TFA can be obtained after removing the rotation indeterminacy by $\boldsymbol{f}_t = B\boldsymbol{f}_{t-1} + \varepsilon_t$?

Additionally, studies may be made on data from other major international markets, with those past empirical analyses (e.g., Chen et al. 1986; Azeez and Yonezawa 2006) as references. In addition to a two-stage implementation, one promising feature of implementing the TFA by the BYY harmony learning (Xu 2001) is that the number of temporal factors is determined automatically during learning, which saves computational costs greatly and also improves the learning performance of TFA, for which details are referred to Sect. 5 of Xu (2010) and Sect. 5.2 of Xu (2012).

Macroeconomics-modulated TFA-APT and nGCH-driven M-TFA-O

In those empirical APT studies, the practice that uses macroeconomic indexes as $\boldsymbol{f}_t$ leads to an understanding that $\boldsymbol{f}_t$ typically consists of a set of macroeconomic or fundamental indexes. In an FA implementation or a TFA implementation by Eq. (41), such an understanding may not be correct. Actually, $\boldsymbol{f}_t$ may vary much slower than the return $\boldsymbol{r}_t$ and thus be regarded as a macroeconomic type of indices. However, $\boldsymbol{f}_t$ may also vary in a timescale similar to the changes of $\boldsymbol{r}_t$. Moreover, $\boldsymbol{f}_t$ in Eq. (41) is intrinsically determined from real data $\boldsymbol{r}_t$ and usually will not coincide with exogenous macroeconomic indexes, such as GDP, inflation, investor confidence, and yield curve. Therefore, we need to further investigate how the market is influenced by these exogenous variables or macroeconomic indexes.

Being quite different from many existing studies that explicitly model the relation between market return $\boldsymbol{r}_t$ and macroeconomic indices, the influences of these indices to $\boldsymbol{r}_t$ are considered via their roles in modulating the temporal factors in $\boldsymbol{f}_t$, as shown in Fig. 3 by Box 3. This idea is realized via extending Eq. (41) into the following macroeconomics-modulated TFA–APT:

$$\begin{aligned} \boldsymbol{r}_t &= \boldsymbol{a} + A\boldsymbol{f}_t + \boldsymbol{e}_t, \mathrm{E}\, \mathbf{f}_t\boldsymbol{e}_t^{\mathrm{T}} = 0, \\ \boldsymbol{f}_t &= B\boldsymbol{f}_{t-1} + H\boldsymbol{m}_t + \varepsilon_t, \mathrm{E}\, \mathbf{f}_{t-1}\varepsilon_t^{\mathrm{T}} = 0, \mathrm{E}\, \mathbf{m}_{t-1}\varepsilon_t^{\mathrm{T}} = 0, \\ \boldsymbol{m}_t &= C\boldsymbol{v}_t + \eta_t, \mathrm{E}\, \boldsymbol{v}_t\, \eta_t^{\mathrm{T}} = 0, \end{aligned} \tag{42}$$

where $\boldsymbol{e}_t$, ε_t, and η_t are Gaussian white noises and independent of each other. Typically, $\boldsymbol{m}_t$ consists of several macroeconomic indices, and $\boldsymbol{v}_t$ consists of several known non-market factors that affect the macroeconomy. Specifically, $H\boldsymbol{m}_t$ describes the effect of the macroeconomic indices to the security market via the hidden factors $\boldsymbol{f}_t$. Actually, Eq. (42) comes from a simplification of one proposed in Sect. III(C) of (Xu 2001) and its Eq. (101), in particular, under the name of macroeconomics-modulated independent state–space model.

In one CIFEr2003 conference paper (Chiu and Xu 2003), empirical investigation is made on the model by Eq. (42). First, white noise tests are made on $\boldsymbol{e}_t$, ε_t, and η_t to ensure model specification adequacy. Second, the performances in return prediction and index forecasting are compared with that of the TFA model. Empirical results reveal that the model is not only well specified, but also superior to the TFA model in stock price and index forecasting.

See Box 4 in Fig. 3, there are two ways to perform prediction based on Eq. (41) and Eq. (42). The first way is intrinsically to get $\boldsymbol{r}_{t-1} \to \boldsymbol{f}_{t-1}$ and predict $\hat{\boldsymbol{r}}_t = \boldsymbol{a} + AB\boldsymbol{f}_{t-1}$ for Eq. (41) and $\hat{\boldsymbol{r}}_t = \boldsymbol{a} + A\left(B\boldsymbol{f}_{t-1} + H\boldsymbol{m}_t\right)$ for Eq. (42), while the second way is considering a given prediction $\boldsymbol{r}_{t-1} \to \boldsymbol{y}_t$ via $\boldsymbol{r}_{t-1} \to \boldsymbol{f}_{t-1}$, $B\boldsymbol{f}_{t-1} \to \boldsymbol{f}_t$ and then $\boldsymbol{f}_t \to \boldsymbol{y}_t$ by learning either linear or nonlinear regression, where y_t could be either $\boldsymbol{r}_t$ or any type of market indices. In one paper (Chiu and Xu 2002), $\boldsymbol{f}_t \to \boldsymbol{y}_t$ is implemented by the normalized radial basis function (NRBF) and extended NRBF (ENRBF) (Xu 1998, 2009) and predicts the stock price or return $\boldsymbol{r}_t$. Empirical studies on Hong Kong market data have shown the superiority of this prediction over not only a conventional prediction $\boldsymbol{f}_t \to \boldsymbol{y}_t$, but also the prediction $\hat{\boldsymbol{r}}_t = \boldsymbol{a} + AB\boldsymbol{f}_{t-1}$.

Based on Eqs. (41) and (42), in addition to making a prediction featured with learning a regression $\boldsymbol{f}_t \to \boldsymbol{y}_t$, we may also use $\boldsymbol{f}_t$ to replace $\boldsymbol{r}_t$ in the previous Eq. (29) for adaptive portfolio management; see Box 5 in Fig. 3. This APT based portfolio management was firstly suggested in Sect. III(c) and especially by Eqs. (96) and (97) in Xu (2001). Extensive simulation results reveal that this $\boldsymbol{f}_t$-based portfolio management generally excels the return $\boldsymbol{r}_t$ based portfolio management by Eq. (29) (Chiu and Xu 2004b).

In general, a parametric $\boldsymbol{y}_t = g\left(\boldsymbol{f}_t, \theta\right)$ can be added to Eq. (41) to provide the outputs of this model for application purposes for such prediction and portfolio management.

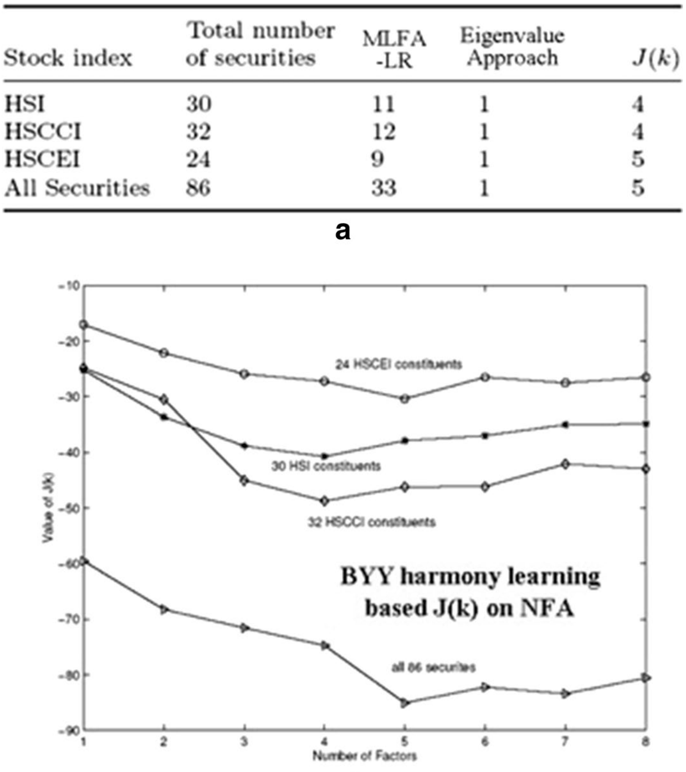
Moreover, beyond the consideration of Gaussian white noises as the driven noise ε_t, we may consider a non-Gaussian driven noise ε_t or a driven noise ε_t with a conditional heteroskedasticity. In summary, we further generalize Eq. (42) into the following model

$$
\begin{aligned}
&\text{(a)}\ \ \mathbf{r}_t = \boldsymbol{a} + A\boldsymbol{f}_t + \boldsymbol{e}_t, \mathrm{E}\,\mathbf{f}_t\boldsymbol{e}_t^{\mathrm{T}} = 0,\\
&\quad\ \ \mathbf{e}_t \sim^{\text{i.i.d.}} G(\boldsymbol{e}_t|0, \Sigma_e) \text{ with a diagonal covariance } \Sigma_e\\
&\text{(b)}\ \ \mathbf{y}_t = g\left(\boldsymbol{f}_t, \theta\right);\\
&\text{(c)}\\
&\qquad\quad \mathbf{f}_t = B\boldsymbol{f}_{t-1} + H\boldsymbol{m}_t + \mathrm{diag}\left[\sigma_t^{(1)}, \ldots, \sigma_t^{(m)}\right]\varepsilon_t,\ q(\varepsilon_t) = \prod_j q\left(\varepsilon_t^{(j)}\right),\\
&\qquad\quad \varepsilon_t = [\varepsilon_t^{(1)}, \ldots, \varepsilon_t^{(m)}]^{\mathrm{T}}_{,}\ \ \mathrm{E}\,\mathbf{f}_{t-1}\varepsilon_t^{\mathrm{T}} = 0,\ \mathrm{E}\mathbf{m}_t\varepsilon_t^{\mathrm{T}} = 0,\ \mathrm{E}\varepsilon_t^{(j)} = 0, \mathrm{E}\varepsilon_t^{(j)2} = 1,\\
&\quad q\left(\varepsilon_t^{(j)}\right) = \begin{cases} G(\varepsilon_t^{(j)}|0,1), & \text{(i) one Gaussian,}\\ \sum_i \alpha_i^{(j)} G(\varepsilon_t^{(j)}|\mu_i^{(j)}, \lambda_i^{(j)}), & \text{(ii) Gaussian mixture;} \end{cases}\\
&\qquad \sigma_t^{(j)} = \begin{cases} \text{a constant } \sigma^{(j)}, & \text{(a) nonheteroskedasticity,}\\ \sigma_t^{(j)}\left(\vartheta^{(j)}\right) \text{given by Eq. (3)}, & \text{(b) heteroskedasticity;} \end{cases}\\
&\text{(d)}\ \ \mathbf{m}_t = C\boldsymbol{v}_t + \eta_t, \mathrm{E}\,\mathbf{v}_t\eta_t^{\mathrm{T}} = 0,\\
&\quad\ \ \eta_t \sim^{\text{i.i.d.}} G(\eta_t|0, \Sigma_\eta) \text{with a digognal covariance } \Sigma_\eta.
\end{aligned}
\tag{43}
$$

Its basic part consists of ingredients (a)(b)(c). In the special case $\boldsymbol{H=0}$, its function is TFA with two extensions. One is outputting $\mathbf{y_t}$, thus shortly denoted by TFA-O. The other is that ingredient (c) drives $\boldsymbol{f}^t$ by its last term that is either or both of non-Gaussian (nG) and conditional heteroscedasticity (CH), for which we use nGCH-driven TFA-O to refer this formulation. When $\boldsymbol{H \neq 0}, \boldsymbol{f}^t$ is also modulated by the macroeconomic market force $\boldsymbol{m}_t$, it leads to the general formulation shortly named nGCH-driven M-TFA-O.

The central role is taken by the statistical nature of ingredient (c), with several scenarios as follows:

- For the case that $B = 0, H = 0$ and $q(\boldsymbol{\varepsilon}_t^{(j)})$ in Choice (i) as well as $\sigma_t^{(j)}$ in Choice (a), ingredient (a) and ingredient (c) jointly degenerate back to the FA-based implementation of the original APT by Eq. (37).
- For the case that $B=0$, $\varepsilon_t = 0$, it follows from $\tilde{A} = AH$ that ingredient (a) and ingredient (c) jointly degenerate back to the fundamental factors based implementation of the original APT by Eq. (37).
- For the case that $B=0$, $q(\varepsilon_t^{(j)})$ in Choice (i), and $\sigma_t^{(j)}$ in Choice (a), ingredient (a) and ingredient (c) jointly act as a combination of the above two implementations.
- For the case that $H=0$, $q(\varepsilon_t^{(j)})$ in Choice (i), and $\sigma_t^{(j)}$ in Choice (a), as well as $B=\mathrm{diag}[b_1, \ldots, b_m]^{\mathrm{T}}$, ingredient (a) and ingredient (c) jointly become the TFA-based implementation by Eq. (41). It further becomes Eq. (42) when $H\neq 0$. Moreover, conditional heteroskedasticity is further considered in ε_t via Choice (i) of $\sigma_t^{(j)}$ to be replaced by Choice (b). As shown by empirical investigation in the CIEF'2003 conference paper (Chiu and Xu 2003), we consider that the conditional heteroskedasticity in the TFA-based implementation is considerably better than the TFA-based implementation without such a consideration.



Stock index	Total number of securities	MLFA -LR	Eigenvalue Approach	$J(k)$
HSI	30	11	1	4
HSCCI	32	12	1	4
HSCEI	24	9	1	5
All Securities	86	33	1	5

a

b

Fig. 7 Comparison on finding the number of factors identified by MLFA-LR test, eigenvalue approach and the BYY harmony learning-based NFA

Another alternative is that Choice (i) of a Gaussian $q(\varepsilon_t^{(j)})$ is replaced by Choice (ii) of a non-Gaussian $q(\varepsilon_t^{(j)})$. In the simplest case, $B=0$, $H=0$, and $\sigma_t^{(j)}$ in Choice (a), ingredient (a) and ingredient (c) jointly degenerate back to the non-Gaussian FA (NFA) as outlined in Fig. 3 by Box 6, for which details are referred to Sect. III(A) in Xu (2001), Sect. IV in Xu (2004), and Sect. 3.2 in Xu (2010). Accordingly, we get a Non-Gaussian APT as shown in Fig. 3 by Box 7. Interestingly, NFA can also remove the FA's rotation indeterminacy by Eq. (40), though there is no temporal structure $\boldsymbol{f}_t$ in consideration because $B=0$, $H=0$. Similar to Fig. 6, shown in Fig. 7 are the results of empirical investigation made on determining the appropriate factor number of APT by NFA (Chiu and Xu 2004a), still in comparison with the results of the MLFA-LR test and the eigenvalue approach as listed in Fig. 7a. Again, the BYY harmony learning-based NFA stably identified four or five factors regardless of the numbers of securities.

This alternative provides a different perspective on how to remove the indeterminacy by Eq. (40) or the incompleteness of APT. Without the additional equation about $\boldsymbol{f}_t$, the formulation of NFA implementation seems closer than the TFA implementation to the original APT formulation by Eq. (37). Naturally, there rises a question on which one is right, TFA or NFA? Actually, they are two aspects of one market model. TFA observes a dynamic market process while NFA describes the market with all the time points projected to one observation spot such that a Gaussian process is projected to be observed as a mixture of Gaussian distributions. Generally, we may have two natures to be considered in the same market, that is, considering both $B=\mathrm{diag}[b_1, \ldots, b_m]^{\mathrm{T}}$ and the choice (ii) of a non-Gaussian $q(\varepsilon_t^{(j)})$. Even generally, the conditional heteroskedasticity may also be added in via letting $\sigma_t^{(j)}$ in the choice (b). Systematically integrating all the parts and all the ingredients together, Eq. (43) may serve as a general formulation for financial market modeling.

Bayesian Ying–Yang harmony learning and two exemplar learning algorithms

Bayesian Ying–Yang (BYY) harmony learning

The Bayesian Ying–Yang (BYY) harmony learning was proposed in Xu (1995a, b) and subsequently developed systematically (Xu 2001, 2007, 2010, 2012), which provides not only a framework that accommodates typical learning approaches from a unified perspective, but also a new road that leads to improved model selection criteria, Ying–Yang alternative learning with automatic model selection, as well as coordinated implementation of Ying-based model selection and Yang-based learning regularization.

From a modern science perspective that regards the famous ancient Yin–Yang philosophy as a meta theory of system sciences and intelligent systems, a system that survives and interacts with its world can be regarded as a Ying–Yang system that functionally composes of two complement parts. One is called Ying, from its inside into its external world, by which a set $\boldsymbol{X}_N = \{x_t\}_{t=1}^N$ of samples are regarded as generated from its representation $\boldsymbol{R}$, while the other is called Yang, from an external world into its inside. A two directional view is considered via the joint distribution of $\boldsymbol{X}, \boldsymbol{R}$ in two types of Bayesian decomposition. The decomposition of $p(\boldsymbol{X}, \boldsymbol{R})$ coincides the Yang concept with a visible domain $p(\boldsymbol{X})$ for a Yang space and a $\boldsymbol{X} \to \boldsymbol{R}$ pathway by $p(\boldsymbol{R}|\boldsymbol{X})$ as a Yang pathway. Thus, $p(\boldsymbol{X}, \boldsymbol{R})$ is called Yang machine. Also, $q(\boldsymbol{X}, \boldsymbol{R})$ is called Ying machine with an invisible domain $q(\boldsymbol{R})$ for a Ying space and a $\boldsymbol{R} \to \boldsymbol{X}$ pathway by $q(\boldsymbol{X}|\boldsymbol{R})$ as a Ying pathway. Such a Ying–Yang pair is called Bayesian Ying–Yang (BYY) system. Ying–Yang pair interact with each other under the principle of best harmony, which is mathematically implemented by maximizing

$$H(p||q) = \int p(\boldsymbol{R}|\boldsymbol{X})p(\boldsymbol{X})\ln[q(\boldsymbol{X}|\boldsymbol{R})q(\boldsymbol{R})]\, \mathrm{d}\boldsymbol{X}\mathrm{d}\boldsymbol{R}. \tag{44}$$

For a machine learning or modeling purpose, we first need to consider a mathematical representation for $\boldsymbol{R}$. The first column of Table lists several typical examples. Usually, $\boldsymbol{R}$ consists of two parts. One is a long-term memory θ that consists of all unknown parameters in the system for collectively representing the underlying structure of $\boldsymbol{X}_N$, while the other is a short-term memory YL with each element being either or both of a categorical label $\ell \in L$ and a vector $y \in Y$ as the corresponding inner representation of one element $x \in X$. For examples, we have a vector y for describing $\boldsymbol{f}_t$ in the APT model by Eq. (37), while we simply have a label ℓ in the time series model by Eq. (4).

The probabilistic structure $q(Y, L)$ is considered jointly with $q(\boldsymbol{X}|\boldsymbol{R}) = q(\boldsymbol{X}|Y, L, \theta)$, depending on both the tasks in consideration and a trade-off between the complexity of

$q(Y, L)$ and the complexity of $q(\boldsymbol{X}|Y, L, \theta)$. For the task of TFA modeling by Eq. (41), we have $q(\boldsymbol{X}|Y, L, \theta)$ by $q(\boldsymbol{r}_t|\boldsymbol{f}_t)$ and $q(Y, L)$ by $q\left(\boldsymbol{f}_t|\boldsymbol{f}_{t-1}\right)$ as follows:

$$\begin{aligned} q(\boldsymbol{r}_t|\boldsymbol{f}_t) &= G(\boldsymbol{r}_t|\boldsymbol{a} + A\boldsymbol{f}_t, \Sigma) \quad \text{with a diagonal } \Sigma, \\ q\left(\boldsymbol{f}_t|\boldsymbol{f}_{t-1}\right) &= G\left(\boldsymbol{f}_t|B\boldsymbol{f}_{t-1}, \Lambda\right) \quad \text{with a diagonal } \Lambda. \end{aligned} \tag{45}$$

Moreover, the remaining part in $q(R) = q(Y, L|\theta)q(\theta)$ is usually called a priori $q(\theta)$ that is chosen depending on the types of parameters and their positions in the Ying machine. In general, a Ying machine $q(X, R) = q(X|R)q(R)$ is designed according to a least complexity principle, featured with designing $q(R) = q(Y, L|\theta)q(\theta)$ in a least redundancy principle and designing $q(\boldsymbol{X}|\boldsymbol{R}) = q(\boldsymbol{X}|Y, L, \theta)$ in a divide–conquer principle.

For the Yang machine $p(X, R) = p(R|X)p(X)$, $p(X)$ directly comes from samples $\boldsymbol{X}_N$, while $p(R|X)$ is designed based on the Ying machine $q(X, R) = q(X|R)q(R)$ according to the variety preservation principle, that is

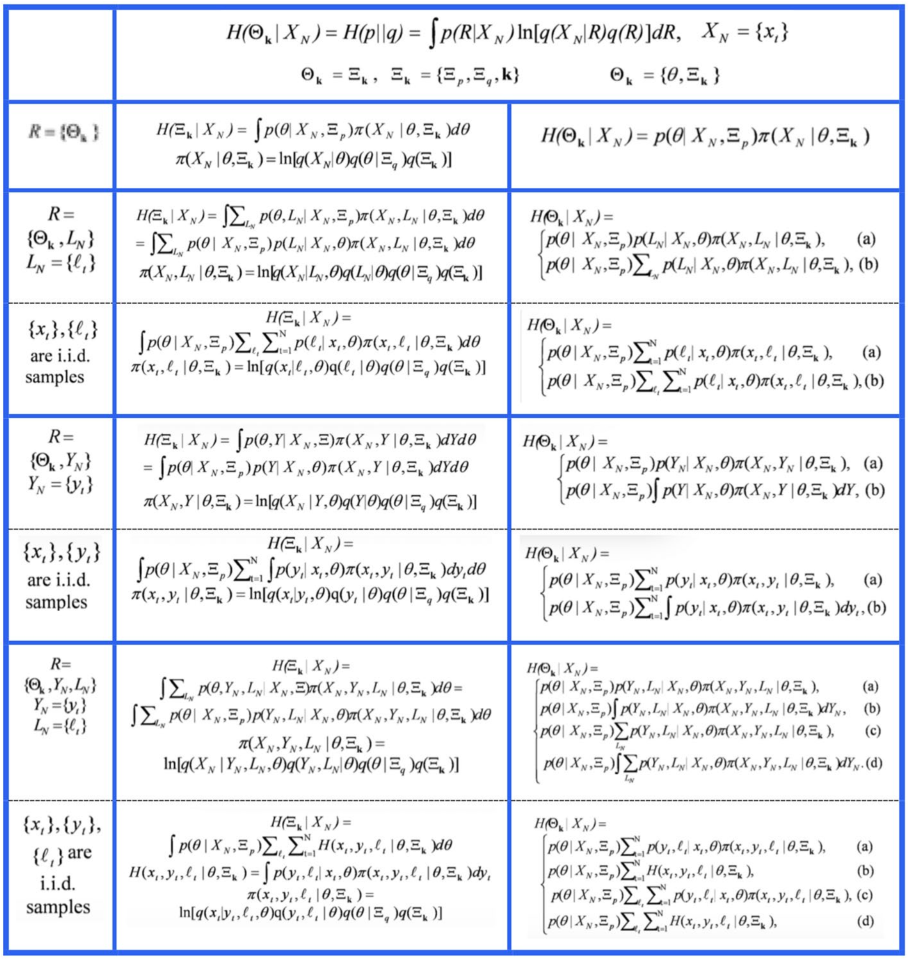
$$\begin{gathered} p(R|X) = q(R|X) \quad \text{in a strong sense} \\ \text{or} \\ \mathrm{Cov}_{R|X} \; of \; p(R|X) = \mathrm{Cov}_{R|X} \; of \; q(R|X) \quad \text{in a week sense.} \\ q(R|X) = q(X|R)q(R) \Big/ \int q(X|R)q(R)\, \mathrm{d}X \mathrm{d}R, \end{gathered} \tag{46}$$

where $\mathrm{Cov}_{R|X}$ indicates a covariance matrix of R conditioning on X. Readers are referred to Xu (2010, 2012) for recent systematic outlines on major issues for designing Ying–Yang machines. To be specific, reading is suggested to start with Sect. 3.2 in Xu (2012) and refer to Sect. 4.2 in Xu (2010) for supplementary materials. Also, readers are referred to Xu (2011) for another perspective that a co-dimensional matrix pair forms a building unit and a hierarchy of such building units sets up the BYY system.

With a BYY system designed, all the remaining unknowns in the system are determined via maximizing the harmony functional by Eq. (44). Typically, there are two types of unknowns. Given the structure of a BYY system or a parametric model in general, it actually means a family of infinite many candidate structures with everyone in a same configuration but in different scales. That is, each candidate is featured by a scale parameter $\boldsymbol{k}$ in terms of one integer or a set of integers. For examples, $\boldsymbol{k}$ consists of the model number k and the orders $\{q_i\}$ for the model in Eq. (3), while merely of the dimension k in the APT model by Eq. (37).

The second type of unknown is featured by a set $\theta_{\boldsymbol{k}}$ of unknown parameters within the candidate structure featured by a specific $\boldsymbol{k}$. Accordingly, maximizing the harmony functional $H(p||q)$ by Eq. (44) makes both parameter learning on determining $\theta_{\boldsymbol{k}}$ and model selection on determining $\boldsymbol{k}$. This BYY best harmony learning provides a favorable mechanism for model selection. Readers are referred to Xu (2010, 2012) for recent systematic overviews on the fundamentals, the novelties and favorable natures of the BYY best harmony learning. To be specific, reading is suggested to start with Sect. 4.1 in Xu (2012) on two different aspects of measuring bi-entity proximity and Sect. 4.2 on the BYY harmony learning from the perspectives of Ying–Yang best matching versus Ying–Yang best harmony, and then proceed to Sect. 7 for a

Table 1 $\boldsymbol{H(p\|q)}$ **in four specific types of implementations**



	$H(\Theta_{\mathbf{k}}\mid X_N)=H(p\|q)=\int p(R\mid X_N)\ln[q(X_N\mid R)q(R)]dR,\quad X_N=\{x_t\}$ $\Theta_{\mathbf{k}}=\Xi_{\mathbf{k}},\ \ \Xi_{\mathbf{k}}=\{\Xi_p,\Xi_q,\mathbf{k}\}$	$\Theta_{\mathbf{k}}=\{\theta,\Xi_{\mathbf{k}}\}$
$R=\{\Theta_{\mathbf{k}}\}$	$H(\Xi_{\mathbf{k}}\mid X_N)=\int p(\theta\mid X_N,\Xi_p)\pi(X_N\mid\theta,\Xi_{\mathbf{k}})d\theta$ $\pi(X_N\mid\theta,\Xi_{\mathbf{k}})=\ln[q(X_N\mid\theta)q(\theta\mid\Xi_q)q(\Xi_{\mathbf{k}})]$	$H(\Theta_{\mathbf{k}}\mid X_N)=p(\theta\mid X_N,\Xi_p)\pi(X_N\mid\theta,\Xi_{\mathbf{k}})$
$R=$ $\{\Theta_{\mathbf{k}},L_N\}$ $L_N=\{\ell_t\}$	$H(\Xi_{\mathbf{k}}\mid X_N)=\int\sum_{L_N}p(\theta,L_N\mid X_N,\Xi_p)\pi(X_N,L_N\mid\theta,\Xi_{\mathbf{k}})d\theta$ $=\int\sum_{L_N}p(\theta\mid X_N,\Xi_p)p(L_N\mid X_N,\theta)\pi(X_N,L_N\mid\theta,\Xi_{\mathbf{k}})d\theta$ $\pi(X_N,L_N\mid\theta,\Xi_{\mathbf{k}})=\ln[q(X_N\mid L_N,\theta)q(L_N\mid\theta)q(\theta\mid\Xi_q)q(\Xi_{\mathbf{k}})]$	$H(\Theta_{\mathbf{k}}\mid X_N)=$ $\begin{cases}p(\theta\mid X_N,\Xi_p)p(L_N\mid X_N,\theta)\pi(X_N,L_N\mid\theta,\Xi_{\mathbf{k}}), & \text{(a)}\\ p(\theta\mid X_N,\Xi_p)\sum_{L_N}p(L_N\mid X_N,\theta)\pi(X_N,L_N\mid\theta,\Xi_{\mathbf{k}}), & \text{(b)}\end{cases}$
$\{x_t\},\{\ell_t\}$ are i.i.d. samples	$H(\Xi_{\mathbf{k}}\mid X_N)=$ $\int p(\theta\mid X_N,\Xi_p)\sum_{\ell_t}\sum_{t=1}^{N}p(\ell_t\mid x_t,\theta)\pi(x_t,\ell_t\mid\theta,\Xi_{\mathbf{k}})d\theta$ $\pi(x_t,\ell_t\mid\theta,\Xi_{\mathbf{k}})=\ln[q(x_t\mid\ell_t,\theta)q(\ell_t\mid\theta)q(\theta\mid\Xi_q)q(\Xi_{\mathbf{k}})]$	$H(\Theta_{\mathbf{k}}\mid X_N)=$ $\begin{cases}p(\theta\mid X_N,\Xi_p)\sum_{t=1}^{N}p(\ell_t\mid x_t,\theta)\pi(x_t,\ell_t\mid\theta,\Xi_{\mathbf{k}}), & \text{(a)}\\ p(\theta\mid X_N,\Xi_p)\sum_{\ell_t}\sum_{t=1}^{N}p(\ell_t\mid x_t,\theta)\pi(x_t,\ell_t\mid\theta,\Xi_{\mathbf{k}}), & \text{(b)}\end{cases}$
$R=$ $\{\Theta_{\mathbf{k}},Y_N\}$ $Y_N=\{y_t\}$	$H(\Xi_{\mathbf{k}}\mid X_N)=\int p(\theta,Y\mid X_N,\Xi)\pi(X_N,Y\mid\theta,\Xi_{\mathbf{k}})dYd\theta$ $=\int p(\theta\mid X_N,\Xi_p)p(Y\mid X_N,\theta)\pi(X_N,Y\mid\theta,\Xi_{\mathbf{k}})dYd\theta$ $\pi(X_N,Y\mid\theta,\Xi_{\mathbf{k}})=\ln[q(X_N\mid Y,\theta)q(Y\mid\theta)q(\theta\mid\Xi_q)q(\Xi_{\mathbf{k}})]$	$H(\Theta_{\mathbf{k}}\mid X_N)=$ $\begin{cases}p(\theta\mid X_N,\Xi_p)p(Y_N\mid X_N,\theta)\pi(X_N,Y_N\mid\theta,\Xi_{\mathbf{k}}), & \text{(a)}\\ p(\theta\mid X_N,\Xi_p)\int p(Y\mid X_N,\theta)\pi(X_N,Y\mid\theta,\Xi_{\mathbf{k}})dY, & \text{(b)}\end{cases}$
$\{x_t\},\{y_t\}$ are i.i.d. samples	$H(\Xi_{\mathbf{k}}\mid X_N)=$ $\int p(\theta\mid X_N,\Xi_p)\sum_{t=1}^{N}\int p(y_t\mid x_t,\theta)\pi(x_t,y_t\mid\theta,\Xi_{\mathbf{k}})dy_t d\theta$ $\pi(x_t,y_t\mid\theta,\Xi_{\mathbf{k}})=\ln[q(x_t\mid y_t,\theta)q(y_t\mid\theta)q(\theta\mid\Xi_q)q(\Xi_{\mathbf{k}})]$	$H(\Theta_{\mathbf{k}}\mid X_N)=$ $\begin{cases}p(\theta\mid X_N,\Xi_p)\sum_{t=1}^{N}p(y_t\mid x_t,\theta)\pi(x_t,y_t\mid\theta,\Xi_{\mathbf{k}}), & \text{(a)}\\ p(\theta\mid X_N,\Xi_p)\sum_{t=1}^{N}\int p(y_t\mid x_t,\theta)\pi(x_t,y_t\mid\theta,\Xi_{\mathbf{k}})dy_t, & \text{(b)}\end{cases}$
$R=$ $\{\Theta_{\mathbf{k}},Y_N,L_N\}$ $Y_N=\{y_t\}$ $L_N=\{\ell_t\}$	$H(\Xi_{\mathbf{k}}\mid X_N)=$ $\int\sum_{L_N}p(\theta,Y_N,L_N\mid X_N,\Xi)\pi(X_N,Y_N,L_N\mid\theta,\Xi_{\mathbf{k}})d\theta=$ $\int\sum_{L_N}p(\theta\mid X_N,\Xi_p)p(Y_N,L_N\mid X_N,\theta)\pi(X_N,Y_N,L_N\mid\theta,\Xi_{\mathbf{k}})d\theta$ $\pi(X_N,Y_N,L_N\mid\theta,\Xi_{\mathbf{k}})=$ $\ln[q(X_N\mid Y_N,L_N,\theta)q(Y_N,L_N\mid\theta)q(\theta\mid\Xi_q)q(\Xi_{\mathbf{k}})]$	$H(\Theta_{\mathbf{k}}\mid X_N)=$ $\begin{cases}p(\theta\mid X_N,\Xi_p)p(Y_N,L_N\mid X_N,\theta)\pi(X_N,Y_N,L_N\mid\theta,\Xi_{\mathbf{k}}), & \text{(a)}\\ p(\theta\mid X_N,\Xi_p)\int p(Y_N,L_N\mid X_N,\theta)\pi(X_N,Y_N,L_N\mid\theta,\Xi_{\mathbf{k}})dY_N, & \text{(b)}\\ p(\theta\mid X_N,\Xi_p)\sum_{L_N}p(Y_N,L_N\mid X_N,\theta)\pi(X_N,Y_N,L_N\mid\theta,\Xi_{\mathbf{k}}), & \text{(c)}\\ p(\theta\mid X_N,\Xi_p)\int\sum_{L_N}p(Y_N,L_N\mid X_N,\theta)\pi(X_N,Y_N,L_N\mid\theta,\Xi_{\mathbf{k}})dY_N. & \text{(d)}\end{cases}$
$\{x_t\},\{y_t\},$ $\{\ell_t\}$ are i.i.d. samples	$H(\Xi_{\mathbf{k}}\mid X_N)=$ $\int p(\theta\mid X_N,\Xi_p)\sum_{\ell_t}\sum_{t=1}^{N}H(x_t,y_t,\ell_t\mid\theta,\Xi_{\mathbf{k}})d\theta$ $H(x_t,y_t,\ell_t\mid\theta,\Xi_{\mathbf{k}})=\int p(y_t,\ell_t\mid x_t,\theta)\pi(x_t,y_t,\ell_t\mid\theta,\Xi_{\mathbf{k}})dy_t$ $\pi(x_t,y_t,\ell_t\mid\theta,\Xi_{\mathbf{k}})=$ $\ln[q(x_t\mid y_t,\ell_t,\theta)q(y_t,\ell_t\mid\theta)q(\theta\mid\Xi_q)q(\Xi_{\mathbf{k}})]$	$H(\Theta_{\mathbf{k}}\mid X_N)=$ $\begin{cases}p(\theta\mid X_N,\Xi_p)\sum_{t=1}^{N}p(y_t,\ell_t\mid x_t,\theta)\pi(x_t,y_t,\ell_t\mid\theta,\Xi_{\mathbf{k}}), & \text{(a)}\\ p(\theta\mid X_N,\Xi_p)\sum_{t=1}^{N}H(x_t,y_t,\ell_t\mid\theta,\Xi_{\mathbf{k}}), & \text{(b)}\\ p(\theta\mid X_N,\Xi_p)\sum_{\ell_t}\sum_{t=1}^{N}p(y_t,\ell_t\mid x_t,\theta)\pi(x_t,y_t,\ell_t\mid\theta,\Xi_{\mathbf{k}}), & \text{(c)}\\ p(\theta\mid X_N,\Xi_p)\sum_{\ell_t}\sum_{t=1}^{N}H(x_t,y_t,\ell_t\mid\theta,\Xi_{\mathbf{k}}), & \text{(d)}\end{cases}$

systematic outline on the thirteen topics about the BYY best harmony learning. Also, readers are referred to Xu (2010) for supplementary materials in Sect. 4.1 and the roadmap shown in Fig. A2 for the relations to other typical learning approaches.

The implementation of maximizing $H(p\|q)$ consists of different specific cases for different learning problems and application tasks. Inputting the samples $\boldsymbol{X}_N$ by $p(\boldsymbol{X})=\delta(\boldsymbol{X}-\boldsymbol{X}_N)$, $H(p\|q)$ in Eq. (44) is simplified into the one on the top of Table 1. As $\boldsymbol{R}$ takes different specific forms given in the first column of Table 1, we have four types of $H(p\|q)$ as listed in the second column of the table, plus their corresponding special cases of i.i.d. samples $\{x_t\}_{t=1}^{N}$.

Moreover, the collective operations $\int[\bullet]\,dY_N$ and $\sum_L[\bullet]$ may be simplified by removing the integral or the summation to merely consider their optimal values, from which those of $H(p\|q)$ in the second column of Table 1 result in the corresponding counterparts of $H(\Theta_{\boldsymbol{k}}|X_N)$ in the third column of the table. Each type in the second column may have more than one counterparts by removing either or both of the two collective operations. Such a removal makes learning implementation of $H(\Xi_{\boldsymbol{k}}|X_N)$ easier but the learned system become more prone to an overfitting of a small size of samples.

As addressed at the end of "Learning mixture of AR, ARMA, ARCH and GRACH models" section, the BYY harmony learning has an automatic model selection mechanism similar to the RPCL learning. Additionally, $H(\Theta_{\boldsymbol{k}}|X_N)$ in the third column of Table 1 provides another angle to view such a mechanism. For example, observing the choice (a) in the last-bottom box of the table, maximizing $H(\Theta_{\boldsymbol{k}}|X_N)$ consists of maximizing not only $p(\theta|X_N, \Xi)$ that is same as the Bayesian learning, but also $\sum_{t=1}^{N} p(y_t, \ell_t|x_t, \theta)\pi(x_t, y_t, \ell_t|\theta_{\ell_t})$ that includes maximizing a term $\omega_{y_t,\ell_t} \ln \omega_{y_t,\ell_t}$ with $\omega_{\ell_t} = q(x_t|y_t, \ell_t, \theta_{\ell_t})q(y_t, \ell_t|\theta_{\ell_t})$. Noticing that $\omega_{y_t,\ell_t} \ln \omega_{y_t,\ell_t}$. monotonically increasing for $\omega_{\ell_t} > e^{-1}$ but decreasing for $\omega_{\ell_t} < e^{-1}$, a value $\omega_{\ell_t} = q(x_t|y_t, \ell_t, \theta_{\ell_t})q(y_t, \ell_t|\theta_{\ell_t}) > e^{-1}$ indicates the current fit to x_t is bigger than this threshold and increasing $\omega_{\ell_t} \ln \omega_{\ell_t}$ enhances learning by $q\left(x_t|y_t, \ell_t, \theta_{\ell_t}\right)q(y_t, \ell_t|\theta_{\ell_t})$ to fit x_t; while a value $\omega_{\ell_t} < e^{-1}$ indicates that this fit is below a threshold and increasing $\omega_{\ell_t} \ln \omega_{\ell_t}$ actually reduces this fit, i.e., a de-learning occurs. This is similar to the RPCL learning.

For the existing Bayes approaches, it is crucial to choosing an appropriate prior, which is usually a difficult task, while an inappropriate prior may deteriorate the performance of model selection seriously. Without any priors on the parameters, Bayes approaches degenerate to the maximum likelihood learning, while the BYY harny learning is still capable of automatic model selection. Also in Table 1, if a priori distribution $q(\theta|\Xi_q)$ is also considered, the performances of BYY harmony learning will be further improved. A simple choice of $q(\theta|\Xi_q)$ is a Jeffreys prior, for which there is no parameter Ξ_q. Alternatively, we may also consider a parametric distribution. Typically, a priori $q(\theta|\Xi_q)$ and a posteriori $p(\theta|X_N, \Xi_p)$ are either jointly a conjugate parametric pair or approximately two parametric distributions with each having a set of hyper-parameters, namely, Ξ_p,Ξ_q. Actually, a hyper-priori $q(\Xi)$ is further considered for $\Xi = \left\{\Xi_p, \Xi_q\right\}$, for which $q(\Xi)$ is a distribution usually with no more prior, e.g., by a Jeffreys prior.

The implementation of maximizing $H(p||q)$ is featured by jointly determining $\Theta_{\boldsymbol{k}}$ and $\boldsymbol{k}$, namely

$$\max_{\boldsymbol{k},\Theta_{\boldsymbol{k}}} H(\Theta_{\boldsymbol{k}}|X_N). \tag{47}$$

Moreover, determining $\Theta_{\boldsymbol{k}}$ further consists of determining $\theta_{\boldsymbol{k}}$ and $\Xi_{\boldsymbol{k}}$ (if any), as well as updating y_t, ℓ_t per sample x_t. Generally, the implementation of Eq. (47) is an alternative iterative process that consists of Step $y\ell$ for updating y_t, ℓ_t, Step θ for parameter learning, Step Ξ for learning hyper-parameters (if any), and Step $\boldsymbol{k}$ for model selection. This process is featured by apex approximation, manifold shrinking, and balanced operation. Readers are referred to Sect. 4.3 in Xu (2012) for a recent systematic overview on major issues about the BYY harmony learning implementation and to Sect. 4.3 in Xu (2010) for further supplementary materials. Considering two typical learning tasks, readers are referred to Sect. 2 in Xu (2012) and Sect. 3 in Xu (2010) for the BYY harmony learning algorithms on Gaussian mixture and factor analysis as well as their extensions.

Learning implementation: gradient algorithms versus EM-like algorithms

The maximization by Eq. (47) can be implemented by different types of learning algorithms. The simplest and widely applicable type is featured by the following gradient based updating:

$$\Theta_{\boldsymbol{k}}^{\text{new}} \leftarrow \Theta_{\boldsymbol{k}}^{\text{old}} + \Delta\Theta_{\boldsymbol{k}} \in D_{\Theta_{\boldsymbol{k}}}, \Delta\Theta_{\boldsymbol{k}} \propto \nabla_{\Theta_{\boldsymbol{k}} \in D_{\Theta_{\boldsymbol{k}}}} H(\Theta_{\boldsymbol{k}}|X_N), \tag{48}$$

where $\Delta u \propto g_u$ means $\Delta u = \gamma g_u$ with a small $\gamma > 0$, $\nabla_{u \in D_u} f(u)$ is the gradient of $f(u)$ with respect to u within the domain D_u of u, and $u + \Delta u \in D_u$ means updating within the domain D_u of u. In the sequel, the use of $\Delta u \propto g_u$ includes the updating $u^{\text{new}} = u^{\text{old}} + \Delta u \in D_u$ even without writing it explicitly. For those choices of $H(\Theta_{\boldsymbol{k}}|X_N)$ in Table 1, if integrals are involved, we need to first handle the integrals and then take gradient on a mathematical expression without integrals, for which we approximately use a Taylor expansion around a maximal point up to the second order. Readers are referred to Sect. 4.3 in Xu (2012) for further details.

To show how a BYY harmony learning algorithm is obtained via the gradient based updating by Eq. (48). Further details are provided on learning the following alternative mixture-of-experts:

$$\begin{aligned} p(x_t|\boldsymbol{x}_{t-1}^q, \theta) &= \sum_{j=1}^{k} P(j|x_{\text{t-1}} - \mu_{j,t-1}, \theta) G(x_t - \mu_{j,t}|0, \sigma_{j,t}^2), \\ P(j|x_{\text{t-1}} - \mu_{j,t-1}, \theta) &= \frac{\alpha_j G(x_{\text{t-1}} - \mu_{j,t-1}|0, \sigma_{j,t-1}^2)}{\sum_{j=1}^{k} \alpha_j G(x_{\text{t-1}} - \mu_{j,t-1}|0, \sigma_{j,t-1}^2)}, \end{aligned} \tag{49}$$

which comes from Eqs. (10), (11) and (12), while $\mu_{j,t}$ comes from the GARCH model given by Eq. (5). To develop algorithms for the ML learning by Eq. (16)(c) and the RPCL learning by Eq. (18), we consider the following likelihood:

$$\begin{aligned} &L(\{x_t\}_{t=1}^N|\Theta) = \sum\nolimits_t \ln\left\{\sum_{j=1}^{k} \alpha_j G(x_{t-1} - \mu_{j,t-1}|0, \sigma_{j,t-1}^2) G(x_t - \mu_{j,t}|0, \sigma_{j,t}^2)\right\}, \\ &\text{and } \pi_{j,t}\left(\theta_j\right) = \ln\{\alpha_j G(x_{t-1} - \mu_{j,t-1}|0, \sigma_{j,t-1}^2) G(x_t - \mu_{j,t}|0, \sigma_{j,t}^2)\}. \end{aligned} \tag{50}$$

Instead of maximizing the likelihood, learning algorithm is derived for maximizing

$$\begin{aligned} &H(p||q) = \smallint p(\theta|X_N, \Xi_p) H(\Theta_{\boldsymbol{k}}|X_N)\, d\theta \\ &H(\Theta_{\boldsymbol{k}}|X_N) = \ln[q(\theta|\Xi_q) q(\Xi)] + \sum_t \sum_{j=1}^{k} p_{t,t-1}(j|\theta) \pi_{j,t}\left(\theta_j\right), \\ &p_{t,t-1}(j|\theta) = \frac{\alpha_j G(x_{t-1} - \mu_{j,t-1}|0, \sigma_{j,t-1}^2) G(x_t - \mu_{j,t}|0, \sigma_{j,t}^2)}{\sum_{j=1}^{k} \alpha_j G(x_{t-1} - \mu_{j,t-1}|0, \sigma_{j,t-1}^2) G(x_t - \mu_{j,t}|0, \sigma_{j,t}^2)}, \end{aligned} \tag{51}$$

where $q(\theta|\Xi_q)$ is a priori distribution typically in a least redundant factorization as follows:

$$\begin{aligned}
&q\left(\theta|\Xi_q\right)=q\left(\left\{\alpha_j\right\}_{j=1}^{k}\right)\prod_{j,i}q\left(a_{j,i}\right)\prod_{j,i}q\left(\beta_{j,i}\right)\prod_{j,i}q\left(\omega_{j,i}\right),\\
&\text{Usually, we have}\\
&q(\{\alpha_j\}_{j=1}^{k})\ :\ \text{Dirichlet},\\
&q\left(\beta_{j,i}\right),q\left(\omega_{j,i}\right):\text{nonnegative densities, e .g., exponential or gamma,}\\
&q\left(a_{j,i}\right):\text{Gaussian or Laplacian, e.g., a Gaussian } G(a_{j,i}|0,\rho_{j,i}^2)\\
&\text{with } q\left(\rho_{j,i}^2\right)\text{being a Jeffreys prior or an inverse gamma.}
\end{aligned}\tag{52}$$

Alternatively, each factor may be simply a Jeffreys prior. The posterior $p(\theta|X_N, \Xi_p)$ also have choices. First, $p(\theta|X_N, \Xi_p)$ and $q(\theta|\Xi_q)$ are a conjugate pair such that the integral over θ can be handled analytically; see Sect. 4.3 of Xu (2012). Second, we may simply consider that $p(\theta|X_N, \Xi_p)$ is free of structure and maximizing $H(p||q)$ with respect to $p(\theta|X_N, \Xi_p)$ is simplified into the maximization of $H(\Theta_{\boldsymbol{k}}|X_N)$ with respect to $\Theta_{\boldsymbol{k}}$. It follows from Eq. (48) that we consider the following gradient updating

$$\begin{aligned}
\Delta\phi \propto \nabla_\phi H\left(\Theta_{\boldsymbol{k}}^{old}|X_N\right),\phi\subset\Theta_{\boldsymbol{k}}=\{\theta,\Xi_{\boldsymbol{k}}\},\\
\nabla_\phi H(\Theta_{\boldsymbol{k}}|X_N)=g_\phi(\Theta_{\boldsymbol{k}})+\sum_t\sum_{j=1}^{k}\rho_{j,t}(\theta)\nabla_\phi\pi_{j,t}(\theta),\\
g_\phi(\Theta_{\boldsymbol{k}})=\nabla_\phi\ln[q(\theta|\Xi_q)q(\Xi)],\\
\rho_{j,t}(\theta)=p_{t,t-1}(j|\theta)\left[1+\Delta\pi_{j,t}(\theta)\right],\\
\Delta\pi_{j,t}(\theta)=\pi_{j,t}\left(\theta_j\right)-\sum_{j=1}^{k}p_{t,t-1}(j|\theta)\pi_{j,t}\left(\theta_j\right),
\end{aligned}\tag{53}$$

where ϕ is a subset of $\Theta_{\boldsymbol{k}}=\{\theta,\Xi_{\boldsymbol{k}}\}$, e.g., either of $\{\boldsymbol{a}_j\},\{\mu_j\},\{\boldsymbol{b}_j\},\{\boldsymbol{w}_j\},\ldots$ etc. One particular example of ϕ is $\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_k]^{\mathrm{T}}$ subject to each $\alpha_j\geq 0$ and $\boldsymbol{\alpha}^{\mathrm{T}}1=1$ with $1=[1, ..., 1]^{\mathrm{T}}$, for which we get $\boldsymbol{\alpha}$ via updating $\boldsymbol{c}=[c_1,\ldots,c_k]^{\mathrm{T}}$ as follows:

$$\begin{aligned}
&\alpha_j=e^{c_j}/\sum\nolimits_\ell e^{c_\ell},\Delta\mathbf{c}\propto\nabla_{\boldsymbol{c}}H(\theta,\Xi_{\boldsymbol{k}}|X_N)=\left(I-\boldsymbol{\alpha}^{\mathrm{old}}1^{\mathrm{T}}\right)\mathrm{diag}\left[\sum_t p_{1,t},\ldots,\sum_t p_{k,t}\right],\\
&\textbf{If a } \alpha_j\rightarrow 0,\ \textbf{discard the corresponding structure and its } \theta_j.
\end{aligned}\tag{54}$$

As addressed in Eq. (5) in Xu (2010) and in Sect. 4.3.2 of Xu (2012), the maximization of Eq. (47) has a mechanism that pushes $\alpha_j\rightarrow 0$ if the corresponding expert is extra, i.e., automatic model selection occurs. Each of nonnegative parameters in $\{\boldsymbol{b}_j\},\{\boldsymbol{w}_j\}$ may also be updated in a similar way, e.g., considering $\xi=\nu^2$ or $\xi=\exp(\nu)$ such that ξ is updated via $\Delta\nu\propto\nabla_\nu H(\Theta_{\boldsymbol{k}}^{\mathrm{old}}|X_N)$. With the help of the priories $q\left(\beta_{j,i}\right)$ and $q(\omega_{j,i})$ in Eq. (52), the maximization of Eq. (47) also pushes $\beta_{j,i}\rightarrow 0$ and $\omega_{j,i}\rightarrow 0$ if some order of the GARCH part in Eq. (4) and Eq. (5) is extra. Moreover, with help of the priori $q(a_{j,i})$ in Eq. (52), the maximization of Eq. (47) also pushes $\rho_{j,i}^2\rightarrow 0$ if some order of the AR part in Eq. (4) and Eq. (5) is extra.

The learning implementation by Eq. (53) covers not only the gradient based ML learning by simply setting $\Delta\pi_{j,t}(\theta_j^{\rm old})=0$ in the Yang step, but also the RPCL learning algorithm simply with $p_{j,t}$ given by Eq. (18). Moreover, setting $\boldsymbol{w}_i = 0$ leads to learning a mixture of ARCH models, while setting $\boldsymbol{w}_i = 0$ and $\boldsymbol{b}_i = 0$ degenerates to learning a mixture of AR models.

For implementing the ML learning, it also been widely regarded that the EM algorithm is preferred over the gradient-based algorithm (Redner and Walker 1984; Xu and Jordan 1996). In addition to the gradient-based implementation by Eq. (53), the BYY harmony learning may also be implemented by the following EM-like procedure:

$$\begin{aligned}
&\text{Yang Step: } p_{j,t} = \rho_{j,t}\left(\theta^{\rm old}\right), \text{see Eq. (53)},\\
&\text{Ying Step: Let } \tilde{\Theta}_{\boldsymbol{k}} = \Theta_k - \phi \ \text{ and } \ \tilde{\theta} = \theta - \phi,\\
&\text{Solve the root } \phi^{*}\text{of } \chi\left(\phi\right) = 0 \text{ or approxaimtely (if difficult)},\\
&\chi\left(\phi\right) = g_{\phi}\left(\tilde{\Theta}_k^{\rm old} \cup \phi\right) + \sum_t \sum_{j=1}^{k} p_{j,t} \nabla_{\phi} \pi_{j,t}\left(\tilde{\theta}^{\rm old} \cup \phi\right),\\
&\text{Then, update } \phi^{\rm new} = \phi^{*},
\end{aligned} \tag{55}$$

where $\boldsymbol{A}$–$\boldsymbol{B}$ denotes the complement of **A** with respect to **B**, i.e., $\boldsymbol{A}-\boldsymbol{B} = \{x \in \boldsymbol{A} | x \notin \boldsymbol{B}\}$. When the root ϕ^{*} of $\chi(\phi)=0$ is solved analytically, setting $\Delta\pi_{j,t}(\theta)=0$ makes Eq. (53) degenerate to the EM algorithm for the ML learning if $g_{\phi}(\Theta_{\boldsymbol{k}}) = 0$ or the Bayes learning if $g_{\phi}(\Theta_{\boldsymbol{k}}) \neq 0$. Generally, the algorithm by Eq. (55) is different from the EM algorithm by the factor $1+\Delta\pi_{j,t}(\theta)$, which takes an important role in making model selection. However, the EM algorithm is guaranteed to converge (Redner and Walker 1984), while the factor $1+\Delta\pi_{j,t}(\theta)$ makes the Ying–Yang iteration lose such a guarantee.

Efforts are made on remedying this weakness. One simple way is replacing $\phi^{\rm new}=\phi^{*}$ in Eq. (55) by the following linear combination

$$\phi^{\rm new} = \phi^{\rm old} + \eta\left(\phi^{*} - \phi^{\rm old}\right), \quad 0 \leq \eta \leq 1. \tag{56}$$

E.g., see Box 3 and Remark (c) in Fig. 7 and Box 7 in Fig. 8 of Xu (2010). However, how to choose an appropriate $0 \leq \eta \leq 1$ remains a problem, which can be handled in one of the following two ways:

- Initialize $\eta \leq 1$, get $\phi^{\rm new}$ by Eq. (56) and check whether $H(\tilde{\Theta}_k^{\rm old} \cup \phi^{new} | X_N) > H(\tilde{\Theta}_k^{\rm old} \cup \phi^{old} | X_N)$

  If yes, we move to the next Ying step in Eq. (55), otherwise reduce η in some way to get $\phi^{\rm new}$ and make such a check again.
- Seek an optimal η^{*} that maximizes $H(\eta) = H(\tilde{\Theta}_k^{\rm old} \cup \left[\phi^{\rm old} + \eta\left(\phi^{*} - \phi^{\rm old}\right)\right] | X_N)$, which can be handled by one of many techniques for one variable optimization. One example is solving the root of $\mathrm{d}H(\eta)/\mathrm{d}\eta = 0$.

Alternatively, another way to get ϕ^{new} from ϕ^{*} and ϕ^{old} is a reconsideration of $\nabla_{\phi}H(\Theta_{\boldsymbol{k}}|X_N)$ in Eq. (53). Making a first order Taylor expansion of $\rho_{j,t}(\theta)$ around θ^{old} and of $\nabla_{\varphi}\pi_{j,t}(\theta)$ around ϕ^{*}, we consider

$$\begin{aligned}
&\rho_{j,t}(\theta)\nabla_{\phi}\pi_{j,t}(\theta) \approx \\
&\quad \left[\rho_{j,t}\left(\theta^{\mathrm{old}}\right) + \nabla_{\phi}\rho_{j,t}\left(\theta^{\mathrm{old}})^{\mathrm{T}}\left(\phi - \varphi^{old}\right)\right]\left[\nabla_{\phi}\pi_{j,t}\left(\tilde{\theta}^{\mathrm{old}}, \varphi^{*}\right) + \nabla_{\phi\phi^{\mathrm{T}}}\pi_{j,t}\left(\tilde{\theta}^{\mathrm{old}}, \varphi^{*}\right)(\phi - \varphi^{*})\right] \\
&\approx \rho_{j,t}\left(\theta^{\mathrm{old}}\right)\nabla_{\phi}\pi_{j,t}\left(\tilde{\theta}^{\mathrm{old}}, \varphi^{*}\right) + U_{j,t}\left(\phi - \varphi^{\mathrm{old}}\right) + V_{j,t}(\phi - \varphi^{*}) \\
&U_{j,t} = \nabla_{\phi}\pi_{j,t}\left(\tilde{\theta}^{\mathrm{old}}, \varphi^{*}\right)\nabla_{\phi}\rho_{j,t}(\theta^{\mathrm{old}})^{\mathrm{T}}, \;\; V_{j,t} = \rho_{j,t}\left(\theta^{\mathrm{old}}\right)\nabla_{\phi\phi^{\mathrm{T}}}\pi_{j,t}\left(\tilde{\theta}^{\mathrm{old}}, \varphi^{*}\right),
\end{aligned}$$

where the second $\approx$ comes from dropping the second order term $\left(\phi - \varphi^{\mathrm{old}}\right)^{\mathrm{T}}\nabla_{\phi}\rho_{j,t}\left(\theta^{\mathrm{old}}\right)\,\nabla_{\phi\phi^{\mathrm{T}}}\pi_{j,t}\left(\tilde{\theta}^{\mathrm{old}}, \varphi^{*}\right)(\phi - \varphi^{*})$. Taking the sum over *j*, *t*, the counterpart of the first term becomes $\chi(\phi^{*}) = 0$ and thus disappears, from which we are led to

$$\psi(\phi) = \nabla_{\phi}H(\Theta_{\boldsymbol{k}}|X_N) \approx g_{\phi}(\Theta_{\boldsymbol{k}}) + \sum_{t}\sum_{j=1}^{k}\left[U_{j,t}\left(\phi - \varphi^{\mathrm{old}}\right) + V_{j,t}(\phi - \varphi^{*})\right]. \quad (57)$$

Then, we solve $\psi(\phi^{\mathrm{new}}) = 0$ to get ϕ^{new} from ϕ^{*} and ϕ^{old}. Particularly, when $g_{\phi}(\Theta_{\boldsymbol{k}}) = 0$ we simply have

$$\phi^{\mathrm{new}} = \left[\sum_{t}\sum_{j=1}^{k}(U_{j,t} + V_{j,t})\right]^{-1}\sum_{t}\sum_{j=1}^{k}\left(U_{j,t}\varphi^{\mathrm{old}} + V_{j,t}\varphi^{*}\right). \quad (58)$$

It is still a linear function of ϕ^{*} and ϕ^{old}, but becomes much advanced than the one by Eq. (56).

Linear causal analyses

Path analyses and a recent development on ρ-diagram

Path analyses is one earliest causal analysis approach, proposed around 1918 by Sewall Wright who made its developments more extensively in the 1920s (Wright 1921, 1934). It has been not only further investigated in the formulation of structural equation modeling (SEM) (Ullman 2006; Hooper et al. 2008; Pearl 2010a; Kline 2015) with wide applications, but also found its uses in many complex modeling areas, including biology, psychology, sociology, and econometrics. Details are left to a vast volume of publications in literature. Here, we introduce a recent development on a modified formulation named ρ-diagram (Xu 2018).

The formulation considers a directed acyclic graph (DAG) or Bayesian networks, with visible nodes x_1, x_2,..., x_n and hidden nodes w_1,...,w_m. Each x_i is normalized to be zero mean and unit variance and each w_j is assumed to be zero mean and unit variance too; while each edge is associated with the correlation coefficient between its two nodes. In other words, such a diagram is completely defined by pairwise correlation coefficients, and thus called ρ-diagram in that each correlation coefficient is denoted by ρ shortly. Being different from the classical procedure for path analyses, namely getting topology by prior, estimating unknown parameters and causal effects, and making model-fit assessment on alternative models, a TPC procedure is suggested for ρ-diagram (Xu 2018), which begins at *T*opology discovery from data based on ρ-diagram, and then makes *P*arameter estimation and *C*ausality embedded model-fit assessment.

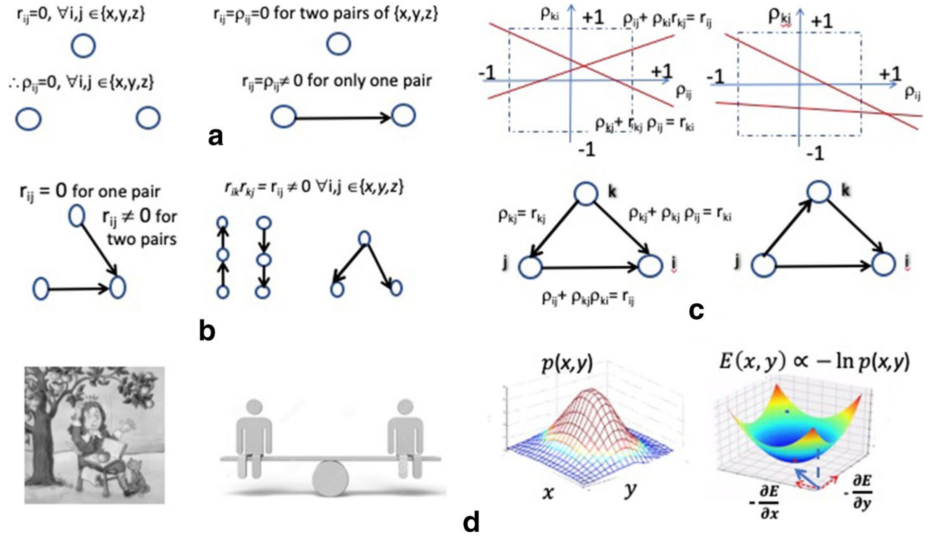
Topology discovery is based on equations that are obtained from path tracing in a way similar to Wright's system of tracing rules. The difference is that unknowns in equations involve only the within-diagram ρ-variables, while knowns are pairwise correlation r-coefficients obtained from visible nodes x_1, x_2,..., x_n, subject to the constraints that all the ρ-variables vary between $[-1,+1]$. We discover a topology underlying data by checking whether a set of constrained equations is deterministically solved, that is, having (1) no solution, (2) a unique solution (or few solutions), and (3) infinite many of solutions.

For details refer to Xu (2018). Here, an illustration is made on topologies of 3-node diagrams, as illustrated in Fig. 8. Given a diagram with nodes x, y, z, the simplest case is illustrated in Fig. 8a, featured by that every pairwise correlation is zero or there is only one pair that gets $r_{ij}\neq 0$, which can be directly identified by observing r_{ij}, $\forall i,j \in\{x,y,z\}$. Shown in Fig. 8b are topologies that have two edges. The first one gets two edges in a fork, which can be identified by observing $r_{ij}=0$ for only one pair while $r_{ij}\neq 0$ for other two pairs. The other topologies describes the causality from conditional independence analysis, which can be identified by observing $r_{ik}r_{kj} = r_{ij}\neq 0$ $\forall i,j \in\{x,y,z\}$ on all the permutations of x, y, z.

Shown in Fig. 8c are two typical topologies of widely encountered causal structure called cofounder. Via path tracing, the following equations are obtained:

$$\rho_{ki} + \rho_{kj}\rho_{ji} = r_{ki}, \quad \rho_{ji} + \rho_{kj}\rho_{ki} = r_{ji}, \quad \rho_{kj} = r_{kj}; \qquad -1 \leq \rho_{ji}, \rho_{kj}, \rho_{ki} \leq 1 \tag{59}$$

As shown in Fig. 8c, we may check whether two lines get cross within the dashed box. If yes, a cofounder is identified in either of two topologies on the bottom of



r_{ij}=0, ∀i,j ∈{x,y,z}
∴ ρ_{ij}=0, ∀i,j ∈{x,y,z}
r_{ij}=ρ_{ij}=0 for two pairs of {x,y,z}
r_{ij}=ρ_{ij}≠ 0 for only one pair
a
ρ_{ki}
+1
ρ_{ij}+ $\rho_{ki}r_{kj}$= r_{ij}
-1
+1
ρ_{ij}
ρ_{kj}+ r_{kj} ρ_{ij} = r_{ki}
-1
ρ_{kj}
+1
-1
+1
ρ_{ij}
-1
r_{ij} = 0 for one pair
r_{ij} ≠ 0 for two pairs
$r_{ik}r_{kj}$ = r_{ij} ≠ 0 ∀i,j ∈{x,y,z}
b
k
ρ_{kj}= r_{kj}
ρ_{kj}+ ρ_{kj} ρ_{ij} = r_{ki}
j
i
ρ_{ij}+ $\rho_{kj}\rho_{ki}$= r_{ij}
k
j
i
c
$p(x,y)$
x
y
$E(x,y) \propto -\ln p(x,y)$
$-\frac{\partial E}{\partial x}$
$-\frac{\partial E}{\partial y}$
d

Fig. 8 Causal analyses: path analysis on ρ-diagram and causal potential theory

Fig. 8c. However, the direction between j and k cannot be identified. Even so, the direct causal direction and effect

$$\rho_{ji} = \left(r_{ji} - r_{kj} r_{ki}\right) \Big/ \left(1 - r_{kj}^2\right) \tag{60}$$

is uniquely determined, i.e., the cofounder effect can be remedied.

If two lines do not intersect within the box, one may further check one other permutation of labels i, j, k. It is unlikely that two different permutations are both identified because it merely happens when not only $\rho = r$ holds on two edges but also four linear equations have consistent solution for unknowns. If no permutation can be identified, it means that there is not such a cofounder causality underlying data. However, there may be still other causality. On one hand, we may check whether there is some causality in types of Fig. 8a, b. On the other hand, we may continue to diagrams with four nodes or more.

Causal potential theory

As already mentioned above, the direction between j and k in Fig. 8c cannot be identified. Also, edge directions in Fig. 8b cannot be identified too. There have been extensive studies on detecting causal direction and evaluating causal strength (Peters et al. 2009; Zhang and Hyvärinen 2009; Hoyer et al. 2009; Rubin and John 2011), via analyzing certain types of asymmetry between two variables X and Y. One most authoritative definition of causality is $p(Y|do\ X=x)$ with '$do\ X=x$' indicating the action that imposes $X=x$ (Pearl 2010b). In these studies, causality is actually examined from a descriptive perspective.

As illustrated in Fig. 8d, possible movements that apple falls and balance loses are actually caused by physics mechanism, i.e., the law of universal gravitation and the lever principle, where causality is actually an issue of dynamics, about how movements are caused by forces that come from potential difference. It follows from the viewpoint of grand unification that we are thus motivated to believe that causality in terms of probability, information, and intelligence should be also governed by similar dynamics.

Consider the relationship described by density distribution $p(x,y)$, as illustrated in Fig. 8d, the quantity $\mathrm{E}(x,y) \propto -\ln p(x,y)$ actually describes a sort of potential energy density on an infinitesimal piece $\mathrm{d}x\mathrm{d}y$, and represents a difference of potential energy density in reference of a uniform distribution on the space x, y, while we can get

$$[I_x, I_y] = \left[-\frac{\partial \mathrm{E}(x,y)}{\partial x}, -\frac{\partial \mathrm{E}(x,y)}{\partial y}\right] \tag{61}$$

to represent a force field that drives information flow toward the area with the lowest energy, or equivalently driving that information flows from rare occurring locations toward high occurring locations.

Changes of x, y and the rates of changes are described by I_x, I_y, respectively, and both are actually driven by the difference of potential energy density of $\mathrm{E}(x, y)$. The problems about whether one of X, Y causes the other or whether two are mutually caused each other may be examined through I_x, I_y. Typically, we may encounter the following cases:

$$\begin{aligned}
&\text{Case } O : I_x = \frac{\partial \ln p(x,y)}{\partial x} = f(x),\ I_y = \frac{\partial \ln p(x,y)}{\partial y} = g(y);\\
&\text{Case } A : I_x = \frac{\partial \ln p(x,y)}{\partial x} = f(x),\ I_y = \frac{\partial \ln p(x,y)}{\partial y} = g(x,y);\\
&\text{Case } B : I_x = \frac{\partial \ln p(x,y)}{\partial x} = f(x,y),\ I_y = \frac{\partial \ln p(x,y)}{\partial y} = g(y);\\
&\text{Case } C : I_x = \frac{\partial \ln p(x,y)}{\partial x} = f(x,y),\ I_y = \frac{\partial \ln p(x,y)}{\partial y} = g(x,y).
\end{aligned} \tag{62}$$

For Case O, changes of x merely relates to itself, while changes of y merely relates to itself, that is, changing x is independent of change of y. For Case A, changes of x merely relates to itself, while changes of y relate to both of x, y, where we may regard that changing x causes change of y. For Case B, changes of y merely relates to itself, while changes of x relate to both of x, y, where we may regard that changing y causes change of x. For Case C, changes of x, y are mutually related.

From a set of samples of x, y, we may develop certain statistics to identify which case is actually encountered. Due to noise and a finite sample size, the first three cases are rarely found. What are often encountered is Case C. In such cases, we may further check whether one of x, y takes a dominant role, while the other maybe ignored, that is, whether we have either or both of

$$f(x,y) \approx f(x), g(x,y) \approx g(y). \tag{63}$$

Further insights on causality may be obtained from this perspective, not only a pair X, Y may be identified in one of the four cases on the entire domain that x, y vary, but also a pair may be identified in one case on some subdomain but in a different case on some different subdomain. That is, causal direction may reverse, disappear, and emerge as x, y vary on different subdomains.

To be more specific, we observe two typical examples. The first considers binary x, y from

$$p(x,y) = p(y|x)p(x), \quad \text{for } x, y = 0, 1 \tag{64}$$

$$p(y|x) = s^y(bx+c)[1-s(bx+c)]^{1-y},\ q(x) = q^x(1-q)^{1-x},$$

where $s(r)$ is a sigmoid function and $p(y|x)$ describes a logistic regression, for which we get

$$\begin{aligned}
&I_x = \ln\frac{q}{1-q} + bs'^{(bx+c)}\left[\frac{y}{s(bx+c)} - \frac{1-y}{1-s(bx+c)}\right] = \ln\frac{q}{1-q} + b\delta,\\
&\delta = s(bx+c) - y,\ I_y = \ln\frac{s(bx+c)}{1-s(bx+c)}.
\end{aligned} \tag{65}$$

We usually have $\delta \approx 0$ if the logistic regression fits well, thus it leads to Case A above, i.e., the causal direction is $x \rightarrow y$, which is consistent to our existing understanding on this model.

The second example considers $p(x,y)$ from a joint density of Gaussian variables x, y with zero mean and unit variance as well as their correlation coefficient ρ. It follows that

$$-I_x = x + \rho y,\ -I_y = y + \rho x, \tag{66}$$

which leads to Case 0 when $\rho=0$, Case A when $\rho y\approx 0$, Case B when $\rho x\approx 0$, and Case C in general. That is, we are unable to identify causal direction on the entire domain, which is also consistent to our existing understanding. Interestingly, we get new insight that it is possible to detect causal direction in some particular subdomains.It also may deserve to extend these studies to consider a density $p(\boldsymbol{x},\boldsymbol{y})$ with $\boldsymbol{x},\boldsymbol{y}$ being vectors such that we examine causality between two groups of variables.

SEM and its relations to modulated TFA-APT and nGCH-driven M-TFA-O

In its early stages of developments, modeling by equations in path analyses and structural equation modeling (SEM) were used without a particular clarification. In recent decades, SEM is gradually developed into the following formulation (Ullman 2006; Kline 2016):

$$\boldsymbol{x}=\boldsymbol{\Lambda}_x\boldsymbol{\xi}+\boldsymbol{\delta},\ \boldsymbol{y}=\boldsymbol{\Lambda}_y\boldsymbol{\eta}+\boldsymbol{\varepsilon},\ \boldsymbol{\eta}=\boldsymbol{B}\boldsymbol{\eta}+\boldsymbol{\Gamma}\boldsymbol{\xi}+\boldsymbol{\varsigma} \tag{67}$$

To compare modulated TFA-APT and nGCH-driven M-TFA-O, we observe the following equations from Eq. (42) and in Eq. (43):

$$\boldsymbol{r}_t=\boldsymbol{a}+\boldsymbol{A}\boldsymbol{f}_t+\boldsymbol{e}_t,\ \boldsymbol{f}_t=\boldsymbol{B}\boldsymbol{f}_{t-1}+\boldsymbol{H}\boldsymbol{m}_t+\boldsymbol{\varepsilon}_t,\ \boldsymbol{m}_t=\boldsymbol{C}\boldsymbol{v}_t+\boldsymbol{\eta}_t,$$

Putting the last one into the second one, we may rewrite

$$\begin{aligned}&\boldsymbol{f}_t=\boldsymbol{B}\boldsymbol{f}_{t-1}+\boldsymbol{H}\boldsymbol{C}\boldsymbol{v}_t+\boldsymbol{H}\boldsymbol{\eta}_t+\boldsymbol{\varepsilon}_t,\\ &\boldsymbol{r}_t=\boldsymbol{a}+\boldsymbol{A}\boldsymbol{f}_t+\boldsymbol{e}_t,\ \boldsymbol{m}_t=\boldsymbol{C}\boldsymbol{v}_t+\boldsymbol{\eta}_t.\end{aligned} \tag{68}$$

Table 2 compares the notations in Eqs. (62) and (63).

The two are actually the same at the special case $\boldsymbol{H}=\boldsymbol{0}$. Generally, we observe that modulated TFA-APT may be regarded as a variant or extension of SEM.

Coming from different perspectives, SEM and the modulated TFA–APT aim at causal analysis in a closely related way. Both consist of FA as basic ingredient that suffers the intrinsic rotation indeterminacy by Eq. (40). In path analysis and SEM study, the problem is avoided by making hidden factors $\boldsymbol{f}^t$ and/or the elements of $\boldsymbol{A}$ partly known with human-aide. While in the modulated TFA-APT, the problem is solved by considering both independence cross hidden factors and temporal dependence $\boldsymbol{B}\boldsymbol{f}^t\_^{1}$ among each factor. We may combine the ideas to improve each other. On one hand, SEM motivates us to prune away extra edges that correspond to elements of $\boldsymbol{A}$, which may be implemented by sparse learning. On the other hand, we may improve SEM by considering temporal dependence among endogenous factors.

Moreover, rotation indeterminacy may also be removed by changing the driving noise of hidden factors from Gaussian $q(\varepsilon_t^{(j)})$ into non-Gaussian $q(\varepsilon_t^{(j)})$ (Xu 2001, 2004). Furthermore, conditional heteroskedasticity (Chiu and Xu 2003) has also been included in the driving noise to encode non-stationarity. The two points are actually included in Item (c) in Eq. (43), which extends the modulated TFA-APT into nGCH-driven M-TFA-O, which may also be used to improve SEM. Furthermore, a non-diagonal matrix B may be considered to replace a diagnal matrix B in TFA, such that Granger causality

Table 2 In comparison with modulated TFA-APT and GMCH-driven M-TFA

In Eq. (61)	y	$\Lambda_y\eta$	ε	$B\eta$	$\Gamma\xi$	ς	x	$\Lambda_x\xi$	δ
In Eq. (62)	r_t -a	Af_t	e_t	Bf_{t-1}	$HC\ v_t$	$H\eta_t+\varepsilon_t$	m_t	Ev_t	η_t

like problem (Granger 1969) may be taken in consideration together with the previous cofounder problem further examined.

Abbreviations

AIC: Akaike information criterion; APT: arbitrage pricing theory; AR: autoregressive; ARCH: Autoregressive Conditional Heteroskedasticity; ARIMA: autoregressive integrated moving average; ARMA: autoregressive–moving average; BYY: Bayesian Ying Yang; BIC: Bayesian information criterion; CAIC: consistent AIC; CAPM: capital asset pricing model; EMH: efficient market hypothesis; HMM: hidden Markov model; GARCH: generalized ARCH; LDS: linear dynamical system; LR: likelihood ratio; MDL: minimum description length; ME: mixture-of-experts; ML: maximum likelihood; MLFA: maximum likelihood factor analysis; MML: minimum message length; MUV: mixture using variance; NFA: non-Gaussian factor analyses; NRBF: normalized radial basis function; ρ-diagram: a diagram defined by a set of pairwise correlation coefficients; RPCL: rival penalized competitive learning; SEM: structural equation modeling; SSM: state space model; TFA: temporal factor analysis; VAR: vector autoregressive; VB: variational Bayes.

Authors' contributions

All from the sole author LX. The author read and approved the final manuscript.

Author details

[1] CMaCH Centre, School of Electronic Information and Electrical Engineering, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Minhang District, Shanghai 200240, China. [2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China.

Acknowledgements

This work was supported by the Zhi-Yuan chair professorship start-up Grant (WF220103010) from Shanghai Jiao Tong University.

Competing interests

The author declares that there is no competing interests.

Availability of data and materials

Not applicable.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

WF220103010, Shanghai Jiao Tong University.

Publisher's Note

Received: 5 December 2018 Accepted: 20 December 2018
Published online: 29 December 2018

References

Abeysekera SP, Mahajan A (1987) A test of the APT in pricing UK stocks. J Account Finance 17(3):377–391

Azeez AA, Yonezawa Y (2006) Macroeconomic factors and the empirical content of the Arbitrage Pricing Theory in the Japanese stock market. Jpn World Econ 18(4):568–591

Azoff ME (1994) Neural network time series forecasting of financial markets. Wiley, New York

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econom 31:307–327

Box G, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco

Brown SJ (1989) The number of factors in security returns. J Finance 44(5):1247–1262

Chamberlain G, Rothschild M (1983) Arbitrage, factor structure, and mean–variance analysis on large asset markets. Econometrica 51(5):1281–1304

Chen NF, Roll R, Ross S (1986) Economic forces and the stock market. J Bus 59(3):383–403

Cheung YM, Leung WM, Xu L (1996) Combination of buffered back-propagation and RPCL-CLP by mixture-of-experts model for foreign exchange rate forecasting. In: Proceedings of 3rd international conference on neural networks in the capital markets, London, UK, Oct 11–13, 1996. World Scientific Pub, Singapore, pp 554–563

Cheung Y, Leung WM, Xu L (1997) Adaptive rival penalized competitive learning and combined linear predictor model for financial forecast and investment. Int J Neural Syst 8:517–534

Chiu KC, Xu L (2002) Stock price and index forecasting by arbitrage pricing theory-based Gaussian TFA learning. In: Yin HJ (ed) Lecture notes in computer sciences (LNCS), vol 2412. Springer, Berlin, pp 366–371

Chiu KC, Xu L (2002) A comparative study of Gaussian TFA learning and statistical tests on the factor number in APT. In: Proceedings of international joint conference on neural networks 2002 (IJCNN '02), Honolulu, Hawaii, USA, May 12–17, 2002. pp 2243–2248

Chiu KC, Xu L (2003) Stock forecasting by ARCH driven Gaussian TFA and alternative mixture experts models. In: Proceedings of 3rd international workshop on computational intelligence in economics and finance, North Carolina, USA, Sept 26–30. pp 1096–1099

Chiu KC, Xu L (2003) On generalized arbitrage pricing theory analysis: empirical investigation of the macroeconomics modulated independent state–space model. In: Proceedings of 2003 international conference on computational intelligence for financial engineering, Hong Kong, March 20–23. pp 139–144

Chiu KC, Xu L (2004a) Arbitrage pricing theory based Gaussian temporal factor analysis for adaptive portfolio management. J Decis Support Syst 37:485–500

Chiu KC, Xu L (2004b) NFA for factor number determination in APT. Int J Theor Appl Finance 7:253–267

Choey M, Weigend AS (1997) Nonlinear trading models through Sharpe ratio optimization. Int J Neural Syst 8(3):417–431

Dhrymes PJ, Friend I, Gultekin B (1984) A critical reexamination of the empirical evidence on the arbitrage pricing theory. J Finance 39(2):323–346

Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom Inflation. Econometrica 50:987–1008

Engle RF, Granger CWJ (1987) Co-integration and error–correction: representation, estimation and testing. Econometrica 55(2):251–276

Figueiredo MAT, Jain AK (2002) Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Mach Intell 24(3):381–396

Fishburn PC (1977) Mean-risk analysis with risk associated with below-target returns. Am Econ Rev 67(2):116–126

Gately E (1995) Neural networks for financial forecasting. John Wiley & Sons, New York

Ghahramani Z, Hinton GE (2000) Variational learning for switching state–space models. Neural Comput 12(4):831–864

Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3):424–438

Hooper D, Coughlan J, Mullen MR (2008) Structural equation modelling: guidelines for determining model fit. Electron J Bus Res Methods 6(1):53–60

Hoyer PO, Janzing D, Mooij JM, Peters J, Schölkopf B (2009) Nonlinear causal discovery with additive noise models. In: Advances in neural information processing systems, pp 689–696

Hung KK, Cheung CC, Xu L (2000) New Sharpe-ratio-related methods for portfolio selection. In: IEEE/IAFE/INFORMS 2000 conference on computational intelligence for financial engineering, New York City, USA, March 26–28, pp 34–37

Hung KK, Cheung Y, Xu L (2003) An extended ASLD trading system to enhance portfolio management. IEEE Trans Neural Networks 14:413–425

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991) Adaptive mixtures of local experts. Neural Comput 3:79–87

Jangmin O, Jongwoo L, Lee JW, Zhang BT (2006) Adaptive stock trading with dynamic asset allocation using reinforcement learning Inform Sci 176(15):2121–2147

Jordan MI, Xu L (1995) Convergence results for the EM approach to mixtures of experts architectures. Neural Netw 8:1409–1431

Kline RB (2015) Principles and practice of structural equation modeling, 4th edn. Guilford Publications, New York

Kwok HY, Chen CM, Xu L (1998) Comparison between mixture of ARMA and mixture of AR model with application to time series forecasting. In: Proceedings of international conference on neural information processing, Kitakyushu, Japan, October 21–23, vol 2. pp 1049–1052

Leontaritis IJ, Billings SA (1985) Input-output parametric models for non-linear systems Part I: deterministic non-linear systems and Part II: stochastic non-linear systems. Int J Control 41:303–344

Leung WM, Cheung Y, Xu L (1997) Application of mixture of experts models to nonlinear financial forecasting. In: Caldwell RB (ed) Nonlinear financial forecasting: proceedings of the first INFFC, (Finance & Technology Publishing, 1997), pp 153–168

Markowitz HM (1952) Portfolio selection. J Finance 7(1):77–91

Markowitz HM (1959) Portfolio selection: efficient diversification of investments. John Wiley & Sons, New York

McGrory CA, Titterington DM (2007) Variational approximations in Bayesian model selection for finite mixture distributions. Comput Stat Data Anal 51(11):5352–5367

Moody J, Saffell M (2001) Q learning to trade via direct reinforcement. IEEE Trans Neural Networks 12(4):875–889

Moody J, Lizhong W, Liao Y, Saffell M (1998) Performance functions and reinforcement learning for trading systems and portfolios. J Forecasting 17:441–470

Neuneier R (1996) Optimal asset allocation using adaptive dynamic programming. In: Touretzky DS (ed) Advances in neural information processing systems, 8th edn. MIT Press, Cambridge, pp 952–958

Pearl J (2010) An introduction to causal inference. Int J Biostat 6(2):1–62

Perrone MP (1994) Putting it all together: methods for combining neural networks. In: Cowan JD, Tesauro G, Alspector J (eds) Advances in neural information processing systems. Morgan Kaufmann, San Francisco, pp 1188–1189

Perrone MP, Cooper LN (1993) When networks disagree: ensemble methods for neural networks. In: Mammone RJ (ed) Neural networks for speech and image processing. Chapman & Hall, New York, pp 126–142

Peters J, Janzing D, Gretton A, Schölkopf B (2009) Detecting the direction of causal time series. In: Proceedings of the 26th annual international conference on machine learning. ACM, New York, pp 801–808

Rabiner LR (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc IEEE 77(2):257–286

Redner RA, Walker HF (1984) Mixture densities, maximum likelihood, and the EM algorithm. SIAM Rev 26:195–239

Ross S (1976) The arbitrage theory of capital asset pricing. J Econ Theory 13(3):341–360

Rubin DB, John L (2011) Rubin causal model. International encyclopedia of statistical science. Springer, Berlin, pp 1263–1265

Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Finance XIX(3):425–442

Sharpe FW (1966) Mutual fund performance. J Bus 39(S1):119–138

Sharpe WF (1994) The Sharpe ratio-properly used, it can improve investment. J Portfolio Manag Fall 21:49–58

Shumway RH, Stoffer DS (1991) Dynamic linear models with switching. J Am Stat Assoc 86(415):763–769

Sims C (1980) Macroeconomics and reality. Econometrica 48(1):1–48

Sortino FA, van der Meer R (1991) Downside risk: capturing what's at stake in investment situations. J Portfolio Manag 17(4):27–31

Tang H, Chiu K-C, Xu L (2003) Finite mixture of ARMA-GARCH model for stock price prediction. In: Proceedings of 3rd international workshop on computational intelligence in economics and finance, North Carolina, USA, Sep 26–30, pp 1112–1119

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B 58(1):267–288

Tu S, Xu L (2011) An investigation of several typical model selection criteria for detecting the number of signals. Front Electr Electron Eng China 6(2):245–255

Ullman JB (2006) Structural equation modeling reviewing the basics and moving forward. J Pers Assess 87(1):35–50

Wang P et al (2011) Radar HRRP statistical recognition with temporal factor analysis by automatic Bayesian Ying–Yang harmony learning. Front Electr Electron Eng China 6(2):300–317

Westland JC (2015) Structural equation modeling: from paths to networks. Springer, New York

Williams PM (1995) Bayesian regularization and pruning using a Laplace prior. Neural Comput 7(1):117–143

Wong WC, Yip F, Xu L (1998) Financial prediction by finite mixture GARCH model. In: Proceedings of international conference on neural information processing, Kitakyushu, Japan, Oct 21–23, 3(1998), pp 1351–1354

Wright S (1921) Correlation and causation. J Agric Res 20(7):557–585

Wright S (1934) The method of path coefficients. Ann Math Stat 5(3):161–215

Xu L (1994) Signal segmentation by finite mixture model and EM algorithm. In: Proceedings of international symposium on artificial neural networks, Tainan, Dec 15–17, pp 453–458

Xu L (1995) Channel equalization by finite mixtures and the EM algorithm. In: Proceedings of IEEE neural networks and signal processing workshop. Cambridge, MA, Aug 31–Sep 2, vol 5, pp 603–612

Xu L (1995) Ying–Yang machines: a Bayesian–Kullback scheme for unified learning and new results on vector quantization. In: Proceedings of the international conference on neural information processing, Beijing, China, Oct 30–Nov 3, pp 977–988 (A further version Advances in NIPS8, Touretzky DS et al (ed), MIT Press, Cambridge MA, 1996: 444–450)

Xu L (1997) Bayesian Ying Yang system and theory as a unified statistical learning approach: (II) from unsupervised learning to supervised learning, and temporal modeling. In: Wong KM et al (eds) Proceedings of theoretical aspects of neural computation: a multidisciplinary perspective. Springer, Berlin, pp 29–42

Xu L (1998) RBF nets, mixture experts, and Bayesian Ying–Yang learning. Neurocomputing 19:223–257

Xu L (2000) Temporal BYY learning for state space approach, hidden Markov model, and blind source separation. IEEE Trans Signal Process 48(7):2132–2144

Xu L (2001) BYY harmony learning, independent state space and generalized APT financial analyses. IEEE Trans Neural Netw 12:822–849

Xu L (2002) Temporal factor analysis: stable-identifiable family, orthogonal flow learning, and automated model selection. In: Proceedings of international joint conference on neural networks. Honolulu, HI, USA, 12–17 May, pp 472–476

Xu L (2004) Advances on BYY harmony learning: information theoretic perspective, generalized projection geometry, and independent factor autodetermination. IEEE Trans Neural Netw 15(4):885–902

Xu L (2007) A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving. Pattern Recogn 40:2129–2153

Xu L (2009) Learning algorithms for RBF functions and subspace based functions. In: Olivas ES et al (eds) Handbook of research on machine learning applications and trends: algorithms, methods and techniques. IGI Global, Hershey, pp 60–94

Xu L (2010) Bayesian Ying–Yang system, best harmony learning, and five action circling. J Front Electr Electron Eng China 5(3):281–328 **(A special issue on Emerging Themes on Information Theory and Bayesian Approach)**

Xu L (2012) On essential topics of BYY harmony learning: current status, challenging issues, and gene analysis applications. J Front Electr Electron Eng 7(1):147–196 **(A special issue on Machine learning and intelligence science: IScIDE (C))**

Xu L (2018) Deep bidirectional intelligence: AlphaZero, deep IA-search, deep IA-infer, and TPC causal learning. Appl Inform 5(5):38

Xu L, Amari S (2008) Combining classifiers and learning mixture of experts. In: Rabuñal Dopico JR (ed) Encyclopedia of artificial intelligence. IGI Global, Hershey, pp 318–326

Xu L, Cheung Y (1997) Adaptive supervised learning decision networks for traders and portfolios. J Comput Intell Finance 5(6):11–16 **(A short version also in Proceedings of IEEE-IAFE 1997 International Conference on Computational Intelligence for Financial Engineering (CIFEr), New York City, March 23-25, 1997, 206–212)**

Xu L, Jordan MI (1996) On convergence properties of the EM algorithm for Gaussian mixtures. Neural Comput 8(1):129–151

Xu L, Krzyzak A, Oja E (1992) Unsupervised and supervised classifications by rival Penalized competitive learning. In: Proceedings of 11th international conference on pattern recognition. Hague, Netherlands, Aug 30–Sep 3, pp 672–675

Xu L, Krzyzak A, Oja E (1993) Rival penalized competitive learning for clustering analysis, RBF net and curve detection. IEEE Trans Neural Netw 4:636–649

Xu L, Jordan MI, Hinton GE (1994) A modified gating network for the mixtures of experts architecture. Proceedings of 1994 world congress on neural networks, vol 2. San Diego, CA, June 4–9, pp 405–410

Xu L, Jordan MI, Hinton GE (1995) An alternative model for mixtures of experts. In: Tesauro G et al (eds) Advances in neural information processing systems 7. MIT Press, Cambridge, pp 633–640

Zhang PG (ed) (2003) Neural networks in business forecasting, forecasting and control. IRM Press, London

Zhang K, Hyvärinen A (2009) On the identifiability of the post-nonlinear causal model. Proceedings of the 25th conference on uncertainty in artificial intelligence (UAI 2009). Montreal, Canada, 2009, pp 647–655