E-backtesting expected shortfall: What defines a “good” forecasting method for Chinese regulators?

Weihua Zhao

Jan 2024

Supervisor: Prof. Ruodu Wang

A project
In partial fulfillment of the requirement for
Master of Quantitative Finance

Table of Contents

1. Introduction

Following the implementation of Basel IV, the China Banking and Insurance Regulatory Commission promulgated the Administrative Measures for the Capital of Commercial Banks in Oct 2023, in which the standard market risk measure is moved from Value-at-Risk (VaR) to Expected Shortfall (ES). The Administrative Measures comes into force since Jan 2024, and as in Basel IV, it authorizes the discretion to decide the forecasting method of ES, urging banks in China to pivot their VaR forecasts to ES forecasts that is "good" in regulatory sense.

By definition, ES is a by-product of VaR. In the literature, methods for their joint forecast are often classified into three groups: fully parametric method with specified model form and distributional assumption; non-parametric method with minimal distributional assumption; semiparametric method with distributional assumption or specified model form (but not both). Examining which method explains the regulatory expectation under Basel IV is key for the on-going banking practices in China, and a calibration tool, called backtesting, that levels with the regulation standards is a necessity to fulfill this purpose. However, there is no Basel standard for backtesting ES yet mainly due to its nonelicitability (Gneiting, 2011), with a compromise in the current proposal to backtesting VaR instead. Many empirical studies evaluate the three groups using either backtests of VaR or backtests of ES that not characterizing regulatory features due to the challenging task. Recently, Wang et al. (2023) leverage e-values and e-tests to tailor a backtesting approach of ES for regulators, enabling the evaluation of forecasting methods to sail with the Basel IV trend. Their approach is called e-backtesting.

In this essay, I evaluate forecasting methods of ES via e-backtests to provide a perspective on what defines a "good" forecasting method of ES in the regulatory sense for the Chinese market. For comprehensiveness, my candidates cover three GARCH models, three extensions based on extreme value theory (McNeil and Frey, 2000), three extensions based on filtered historical simulation (Nolde and Ziegel, 2017), the empirical estimation, and the LSTM-AL model (Li et al., 2021) that incorporates long short-term memory (LSTM) network into the ES-CAViaR framework. For robustness, my evaluation is carried out both across four representative markets: equity, fixed income, foreign exchange, commodity and across portfolios optimized by four strategies: naïve 1/N, minimize ES, maximize return, Markovitz mean variance. For consistency, my ES forecasts proceed with Basel IV standard: at a confidence level of 97.5%, with an estimation window of the most recent 250 days, and on a daily basis.

My e-backtesting evidence strongly coincides with the regime shifts for the equity and foreign exchange markets, pointing to a key metric in defining a "good" ES model for Chinese regulators: the model must demonstrate strong, stable, and initiative adaptability to structure breaks due to frequent regime shifts. Unfortunately, none of my candidates is immune to the forces, though the LSTM-AL model exhibits much stronger resilience.

This essay is organized as follows. Chapter 2 addresses the theoretical framework and provides a cheat sheet for the GARCH-based methods; Chapter 3 addresses the theoretical framework and provides a cookbook for the LATM-AL model; Chapter 4 records a lesson on the

tail behavior under the price-limit mechanism learned from a failed attempt; Chapter 5 addresses the empirical data and the details of forecasting ES by the methods introduced in Chapter 2 and 3; Chapter 6 presents evidence form e-backtesting; Chapter 7 concludes.

2. The GARCH-based Methods

Given the conditional heteroscedasticity in volatility, financial returns are often modelled as a first-order autoregressive process with GARCH(1,1) errors. The AR(1)-GARCH(1,1) model specifies the dynamic of returns with an assumed conditional distribution and yields a fully parametric estimation for conditional VaR and ES. As in Nolde and Ziegel (2017), the AR(1)-GARCH(1,1) model can be extended to some semi-parametric methods, namely the GARCH-EVT which incorporates the extreme value theory (EVT) and the GARCH-FHS which incorporates filtered historical simulation (FHS). Such hybrid methods alleviate the distributional assumption and estimate VaR and ES using the standardized realized residuals. The following notations will be used throughout this chapter. Let $\{X_t\}_{t\in\mathbb{N}}$ be a sequence of random losses (e.g., negative log-returns) defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[|X_t|] < \infty$ for all t. Let $\mathcal{F}_t = \sigma(X_s | s \leq t)$ denoting the σ-algebra generated by the random variables $X_1, \dots, X_t$, representing the information on returns available up to time t. Denote the confidence level for ES and VaR by α (e.g., $\alpha = 97.5\%$). Denote the general inverse of a distribution F by $F^{\leftarrow}$ while that of a continuous distribution F by F^{-1}. The VaR (at level α) of a random loss X with CDF F_X is defined as $VaR(X) = \inf\{x \in \mathbb{R} | F_X(x) \geq \alpha\}$, and the corresponding ES (at level α) is defined as $ES(X) = 1/(1-\alpha) \int_\alpha^1 VaR_\alpha(X) d\alpha$.

2.1. The AR(1)-GARCH(1,1) Process

The sequence of returns $\{X_t\}_{t\in\mathbb{N}}$ can be modelled as an AR(1)-GARCH(1,1) process:

$$X_t = \mu_t + \sigma_t Z_t, \tag{2.1}$$

where $\{Z_t\}_{t\in\mathbb{N}}$ are independent and identically distributed with mean 0, variance 1, and CDF F_Z, and μ_t and σ_t are measurable with respect to $\mathcal{F}_{t-1}$. Specifically, the conditional mean μ_t follows an AR(1) process:

$$\mu_t = \mu + \varphi(X_{t-1} - \mu),$$

with mean μ and $|\varphi| < 1$; the conditional volatility σ_t evolves as a GARCH(1,1) process:

$$\sigma_t^2 = \alpha_0 + \alpha_1 (X_{t-1} - \mu_{t-1})^2 + \beta_1 \sigma_{t-1}^2, \tag{2.2}$$

where $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, and $\alpha_1 + \beta_1 < 1$. It follows that the conditional VaR and ES for day $t+1$ at confidence level α are defined as

$$VaR_{\alpha,t+1}(X_{t+1} | \mathcal{F}_t) = \mu_{t+1} + \sigma_{t+1} F_Z^{\leftarrow}(\alpha), \tag{2.2}$$

and

$$ES_{\alpha,t+1}(X_{t+1}|\mathcal{F}_t) = \mu_{t+1} + \sigma_{t+1}ES_\alpha(Z). \tag{2.3}$$

Three common choices for F_Z in the literature are the (standard) normal distribution $N(0,1)$, (standardized) student's t distribution $t_\nu(0,1)$ and the (standard) skewed-t distribution $st(0,1,\gamma,\nu)$. For $Z\sim N(0,1)$, with CDF and PDF denoted Φ and ϕ, we have

$$F_Z^{\leftarrow}(\alpha) = \Phi^{-1}(\alpha), \tag{2.4}$$

and

$$ES_\alpha(Z) = \frac{\phi\left(\Phi^{-1}(\alpha)\right)}{1-\alpha}. \tag{2.5}$$

For $Z\sim t_\nu(0,1)$, we have

$$F_Z^{\leftarrow}(\alpha) = F_{t,\nu}^{-1}(\alpha)\sqrt{\frac{\nu-2}{\nu}}, \nu > 2 \tag{2.6}$$

and

$$ES_\alpha(Z) = \frac{f_{t,\nu}\left(F_{t,\nu}^{-1}(\alpha)\right)}{1-\alpha}\left(\frac{\nu+\left(F_{t,\nu}^{-1}(\alpha)\right)^2}{\nu-1}\right)\sqrt{\frac{\nu-2}{\nu}}, \nu > 2 \tag{2.7}$$

where ν can be estimated via MLE or QMLE under the AR(1)-GARCH(1,1) specification, $f_{t,\nu}$ is the PDF, and $F_{t,\nu}^{-1}$ is the inverse CDF of a standard t distribution with degree of freedom ν. Likewise, for $Z\sim st(0,1,\gamma,\nu)$, γ and ν are the skewness and shape parameters, respectively, which can be estimated via MLE or QMLE under the AR(1)-GARCH(1,1) specification. According to measures of skewness and kurtosis in Fernandez & Steel (1998), we have

$$F_Z^{\leftarrow}(\alpha) = F_{st,\lambda,q}^{-1}(\alpha), \tag{2.8}$$

where $F_{st,\lambda,q}^{-1}$ is the inverse CDF of a standardized skewed-t distribution such that the skewness parameter $\lambda = (\gamma^2-1)/(\gamma^2+1)$ and the shape parameter $q = \nu/2$. ES can then be calculated explicitly using the expression presented in Patton et al. (2019). For illustrative purposes, define

$$c = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}},\ a = -4\lambda c\left(\frac{\nu-2}{\nu-1}\right), \text{ and } b = \sqrt{1+3\lambda^2-a^2}.$$

If $F_Z^{\leftarrow}(\alpha) \geq \frac{a}{b}$, we have

$$ES_\alpha(Z) = -\frac{\tilde{\alpha}}{1-\alpha}(1+\lambda)\left(-\frac{a}{b} - \frac{1+\lambda}{b}ES_Z^*\right), \tag{2.9}$$

where

$$\tilde{\alpha} = F_{st,0,q}\left(\frac{b}{(1+\lambda)}\left(-F_Z^{\leftarrow}(\alpha) + \frac{a}{b}\right)\right),$$

and

$$ES_Z^* = \sqrt{\frac{(\nu-2)}{\nu}}\frac{\nu^{\nu/2}}{2\tilde{\alpha}\sqrt{\pi}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\left(F_{t,\nu}^{-1}(1-\tilde{\alpha})^2+\nu\right)^{(1-\nu)/2}.$$

Otherwise, we have

$$ES_\alpha(Z) = -\frac{\tilde{\alpha}}{1-\alpha}(1-\lambda)\left(-\frac{a}{b}-\frac{1-\lambda}{b}ES_Z^*\right), \tag{2.10}$$

where

$$a = 4\lambda c\left(\frac{\nu-2}{\nu-1}\right),$$
$$\tilde{\alpha} = F_{st,0,q}\left(\frac{b}{(1-\lambda)}\left(F_{st,\lambda,q}^{-1}(\alpha)+\frac{a}{b}\right)\right),$$

and

$$ES_Z^* = \sqrt{\frac{(\nu-2)}{\nu}}\frac{\nu^{\nu/2}}{2\tilde{\alpha}\sqrt{\pi}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\left(F_{t,\nu}^{-1}(1-\tilde{\alpha})^2+\nu\right)^{(1-\nu)/2}.$$

2.2. GARCH-EVT

It is commonly admitted that the marginal distribution of asset returns is leptokurtic, and the distribution of negative log-returns has heavier tails than that of positive ones. A GARCH process with normal innovations is theoretically endowed with leptokurtosis, provided that the fourth-order moment exists (see Theorem 2 in Bollerslve, 1986, for conditions). With finite degree of freedom, student's t innovations can capture fatter tails than do the normal; with proper skewness, skewed-t innovations can further replicate the thickness of tails in an asymmetric fashion. However, introducing skewness and kurtosis may lead to inconsistent estimates of GARCH parameters (Fan et al., 2014). Moreover, as mentioned in Nolde and Ziegel (2017), there are often not enough data points in the tail regions to give a proper justification for a parametric model. A remedy for ARMA-GARCH process is proposed in McNeil and Frey (2000), which softens the impact of inappropriate innovations by incorporating EVT to model the tail.

Consider innovations $\{Z_t\}$ in (2.1) where $Z_t \sim$ SWN $(0,1)$ with CDF F_Z. The excess distribution of Z_t above a threshold u is given by

$$F_u(x) := \mathbb{P}(Z_t - u \le x | Z_t > u\} = \frac{F_Z(x+u)-F_Z(u)}{1-F_Z(u)},$$

where $0 \le x < x_0 - u$, and x_0 is the right endpoint of F_Z. Balkema and de Haan (1974) and Pickands (1975) proved for a large class of distributions F_Z that there exists a positive measurable function $\beta(u)$ such that

$$\lim_{u\to x_0}\sup_{0\le x<x_0-u}\left|F_u(x)-G_{\xi,\beta(u)}(x)\right| = 0,$$

where $G_{\xi,\beta}(x)$ is the CDF of a generalized Pareto distribution (GPD) given by

$$G_{\xi,\beta}(x) = \begin{cases} 1-(1+\xi x/\beta)^{-1/\xi} & \text{if } \xi \ne 0, \\ 1-\exp(-x/\beta) & \text{if } \xi = 0, \end{cases}$$

where $\beta > 0$, and the support is $y \geq 0$ if $\xi \geq 0$ and $0 \leq y \leq -\beta/\xi$ otherwise. This motivates the Peaks-over-Threshold (EVT-POT) method proposed in McNeil and Frey (2000). Specifically, assume the returns are specified as in (2.1). The standardized implied residuals from fitting an AR(1)-GARCH(1,1) process is

$$\hat{z}_t = \frac{x_t - \hat{\mu}_t}{\hat{\sigma}_t}, \tag{2.11}$$

where x_t is the realized return, and $\hat{\mu}_t$ and $\hat{\sigma}_t$ are the estimated conditional mean and volatility, respectively. Following the two-stage procedures in McNeil and Frey (2000), one checks whether the realized sequence in (2.11) is SWN(0,1); if satisfied, then select a threshold u for $\hat{z}_t$ and fit a GPD to the excesses. Denote by $(\hat{\xi}, \hat{\beta})$ the shape and scale estimates. The conditional VaR in (2.2) is estimated as

$$\widehat{VaR}_{\alpha,t+1}(X_{t+1}|\mathcal{F}_t) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\left(\widehat{F_Z^{\leftarrow}}(\alpha)\right) \tag{2.12}$$

where $\widehat{F_Z^{\leftarrow}}(\alpha)$ is the estimated VaR for F_Z, calculated as

$$\widehat{F_Z^{\leftarrow}}(\alpha) = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{1-\alpha}{N/n}\right)^{-\hat{\xi}} - 1\right), \hat{\xi} \neq 0, \tag{2.13}$$

with N/n being the proportion of $\hat{z}_t$ over the threshold u. The conditional ES in (2.3) is estimated as

$$\widehat{ES}_{\alpha,t+1}(X_{t+1}|\mathcal{F}_t) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\left(\frac{\widehat{F_Z^{\leftarrow}}(\alpha)}{1-\hat{\xi}} + \frac{\hat{\beta} - \hat{\xi}u}{1-\hat{\xi}}\right), \hat{\xi} < 1. \tag{2.14}$$

A reasonable u is frequently selected based on graphical representation. The optimal threshold should be large enough to reduce the bias but small enough to reduce the variance. As remarked in Echaust and Just (2020), there is extensive literature proposing approaches to threshold selection, but none of these conceptions outperforms others in all situations.

2.3. GARCH-FHS

A typical non-parametric method of prediction is empirical estimation, which screenshots the past returns and infers the near future VaR and ES from the sample quantile. The neglection of time-varying volatility in empirical VaR and ES often leads to lenient estimation during a calm period but parsimonious estimation during a stressed one. A refinement is filtered historical simulation (FHS), which dates back to the pioneering work in Hull and White (1998) and Barone-Adesi et al. (1999). Such methods rescale historical returns by the underlying conditional volatilities and perform empirical estimation on a simulated sample. In my empirical study discussed later, I consider the GARCH-FHS method in Nolde and Ziegel (2017), which randomly draws (with replacement) a sample of 10,000 observations from the standardized residuals $\{\hat{z}_t\}$ in (2.11) and then take the empirical VaR and ES for $F_Z^{\leftarrow}(\alpha)$ in (2.2) and $ES_\alpha(Z)$ in (2.3), respectively.

3. The LSTM-AL Model

As in Engle and Manganelli (2004), volatility of asset returns clusters over time; accordingly, the VaR, which is tied up to the standard deviation of the distribution of returns, is expected to exhibit similar serial dependency. In this regard, they propose a conditional autoregressive quantile model for VaR (CAViaR), which resembles the autoregressive specification in GARCH while exempts distributional assumptions. In light of the joint elicitability of VaR and ES raised in Fissler and Ziegel (2016), Taylor (2019) extends the CAViaR to the ES-CAViaR framework that can model VaR and ES jointly. Nonetheless, the weights on lagged terms in either the GARCH or the ES-CAViaR specification fail to articulate the well-documented empirical observation of long-range dependency in volatility. In order to resolve such drawback, Li et al. (2021) embed a long short-term memory (LSTM) network into the ES-CAViaR framework and leverage the asymmetric Laplace (AL) density to calibrate model estimation. They name their model LSTM-AL. Throughout, denote the confidence level by $\bar{\alpha}$ (i.e., $\bar{\alpha} = 1 - \alpha$, where α follows the notation in Chapter 1); MLE shortcuts maximum likelihood estimation.

3.1. The ES-CAViaR framework

Following the terminology in Taylor (2020), scoring function is the term for a loss function when being used to evaluate the prediction of some measure of a distribution; a measure is elicitable if the true forecast is the unique minimizer of the expectation for at least one scoring function; in Fissler and Ziegel (2016), such scoring functions are defined as being strictly consistent for the associated measure. In that sense, with some proposed model for a risk measure, elicitability would enable a way to estimate parameters by minimizing the expectation of corresponding strictly consistent scoring function given observed returns.

ES is not elicitable (Gneiting, 2011). Nevertheless, Fissler and Ziegel (2016) have proved that ES is jointly elicitable with VaR and formulate a set of rules for strictly consistent scoring functions of (VaR, ES). On grounds of these rules, Tayor (2019) shows that the function defined as

$$S(VaR_t, ES_t, Y_t) = -\log\left(\frac{\bar{\alpha}-1}{ES_t}\right) - \left(\frac{(Y_t - VaR_t)(\bar{\alpha} - \mathbb{1}\{Y_t \le VaR_t\})}{\bar{\alpha} ES_t}\right)$$

is a strictly consistent scoring function for (VaR, ES). Hence, given y_t as an observation for Y_t, the optimal forecast $\left(\widehat{VaR}_t,\ \widehat{ES}_t\right)$ is given by

$$\left(\widehat{VaR}_t,\ \widehat{ES}_t\right) = \underset{VaR_t,\, ES_t}{\operatorname{argmin}}\, \mathbb{E}[S(VaR_t, ES_t, Y_t) | Y_t = y_t]. \tag{3.1}$$

Expressing $S(VaR_t, ES_t, Y_t)$ in (3.1) as $exp\left(-S(VaR_t, ES_t, y_t)\right)$ implies

$$\left(\widehat{VaR}_t,\ \widehat{ES}_t\right) = \underset{VaR_t,\, ES_t}{\operatorname{argmin}}\, \mathbb{E}\left[-\log\left(exp\left(-S(VaR_t, ES_t, Y_t)\right) | Y_t\right) = y_t\right], \tag{3.2}$$

where $\exp\big(-S(VaR_t, ES_t, y_t)\big)$ coincides the likelihood of an AL distribution with density:

$$f_{AL}(y_t|VaR_t, ES_t) = \frac{\bar{\alpha}-1}{ES_t} exp\left(\frac{(y_t - VaR_t)(\bar{\alpha} - \mathbb{1}\{y_t \le VaR_t\})}{\bar{\alpha} ES_t}\right), \tag{3.3}$$

where $\bar{\alpha}ES_t$ equals to the scale parameter, and $\bar{\alpha}$ equals to the skewness parameter. That is, we get

$$\underset{VaR_t,\, ES_t}{\operatorname{argmin}} \mathbb{E}[S(VaR_t, ES_t, Y_t)|Y_t = y_t] = \underset{VaR_t,\, ES_t}{\operatorname{argmax}} \mathbb{E}[\log(f_{AL}(y_t|VaR_t, ES_t)|Y_t) = y_t]. \tag{3.4}$$

The nice equivalence in (3.4) rationalizes the convenience to estimate the pair (VaR, ES) via the MLE of f_{AL}.

Extending two CAViaR examples in Engle and Manganelli (2004) with two proposals for ES, Tayor (2019) synthesizes the ES-CAViaR framework which consists of four joint models for (VaR, ES). Here I list one of the combinations that will be referred to later:

$$VaR_t = \tau_0 + \tau_1 VaR_{t-1} + \tau_2 |y_{t-1}|, \tag{3.5}$$

which is called symmetric absolute value (SAV) in Engle and Manganelli (2004), and

$$ES_t = (1 + e^{\tau_3}) VaR_t, \tag{3.6}$$

which avoids the crossing between VaR and ES via multiplying VaR by a factor greater than 1. As underscored beforehand, parameters $(\tau_0, \tau_1, \tau_2, \tau_3)$ can be estimated via the MLE of (3.3).

3.2. The LSTM-AL Model

Statistically, volatility clustering is interpreted as profound positive autocorrelations of the absolute or squared returns at non-zero lags. Much of the empirical research on econometrics (see, e.g., Lo, 1991; Cont, 2001) have remarked that the autocorrelation function $\rho(h)$ of squared or absolute returns decays in lag h slower than the exponential rate. However, for an AR(1)-GARCH(1,1) process decomposed in (2.1) to (2.3), the autocorrelation function of $\{X_t^2\}_{t\in\mathbb{N}}$ decays exponentially to zero with rate $\alpha_1 + \beta_1$ (where $\alpha_1 + \beta_1 < 1$). Due to the exemption of distributional assumptions, it is not possible to derive a closed-form $\rho(h)$ of the VaR_t in (3.5); nevertheless, expanding the lagged terms in (3.5) gives some insight:

$$\begin{aligned} VaR_t &= \tau_0 + \tau_1 VaR_{t-1} + \tau_2 |y_{t-1}| \\ &= \tau_0 + \tau_1 \tau_0 + \tau_1^2 VaR_{t-2} + \tau_1 \tau_2 |y_{t-2}| + \tau_2 |y_{t-1}| \\ &= \cdots = \textstyle\sum_{k=0}^{n-1} \tau_0 {\tau_1}^k + \tau_1^n VaR_{t-n} + \sum_{k=1}^{n} |y_{t-k}| {\tau_1}^{k-1} \tau_2, \end{aligned}$$

where much of empirical studies (see e.g., Engle and Manganelli, 2004; Gerlach et al., 2011; Taylor, 2019) show that the autoregressive parameter τ_1 is lower than but relatively close to 1. Hence, the autoregressive term $\tau_1^n VaR_{t-n} \to 0$ as $n \to \infty$, which suggests that the persistence between VaR_t and VaR_{t-n} is profound but diminishes to zero exponentially with rate τ_1. As such, both the GARCH and the ES-CAViaR fail to mimic long-term memory in volatility.

LSTMs are a special kind of recurrent neural network (RNN), well-known for its capability in learning long-term dependencies without the problem of vanishing and exploding gradients (see e.g., Hochreiter and Schmidhuber, 1997; Gers et al., 2003, for in-depth discussion). Inspired by its advantage, Li et al. (2021) revised (3.5) and (3.6) with an LSTM architecture in the following way:

$$VaR_t = \eta_t + \beta_0|y_{t-1}| + \beta_1 VaR_{t-1} \tag{3.7}$$

$$ES_t = (1 + exp\,(\gamma_0 + \gamma_1 h_t))VaR_t \tag{3.8}$$

$$\eta_t = \alpha_0 + \alpha_1 h_t \tag{3.9}$$

$$h_t = LSTM(\eta_{t-1},\ h_{t-1}, C_{t-1}), \tag{3.10}$$

where $LSTM$ abstracts the architecture of an LSTM unit on step t ($t \geq 1$). Specifically,

$$LSTM(\eta_{t-1},\ h_{t-1}, C_{t-1}) = g_t^o\ tanh\ (C_t) \tag{3.11}$$

$$C_t = g_t^f C_{t-1} + g_t^i\, \eta_t^d \tag{3.12}$$

$$g_t^o = \sigma(b_o + \mu_o \eta_{t-1} + \omega_o h_{t-1}) \tag{3.13}$$

$$g_t^f = \sigma\big(b_f + \mu_f \eta_{t-1} + \omega_f h_{t-1}\big) \tag{3.14}$$

$$g_t^i = \sigma(b_i + \mu_i \eta_{t-1} + \omega_i h_{t-1}) \tag{3.15}$$

$$\eta_t^d = \sigma(b_d + \mu_d \eta_{t-1} + \omega_d h_{t-1}), \tag{3.16}$$

where σ is the sigmoid function valued between 0 and 1; $g_t^i,\ g_t^f, g_t^o$ refers to the input gate, forget gate, and output gate, respectively, which regulate what new story to memorize, what historical event to discard, and what information to output in synergy; C_t is the cell state, empowering the LSTM to learn long-term dependency in an elegant way; η_t^d is the new candidate to be stored in cell state at time t. Repeating steps (3.7) to (3.10) in a sequential manner generates a series of forecasts for (VaR, ES).

Let $\boldsymbol{y} = (y_1, \dots, y_N)$ represent a sequence of observed returns. Following the nice equivalence in (3.4), the parameter vector of the LSTM-AL model:

$$\boldsymbol{\theta} = \big(\beta_0, \beta_1, \gamma_0, \gamma_1, \alpha_0, \alpha_1, b_o, \mu_o, \omega_o, b_f, \mu_f, \omega_f, b_i, \mu_i, \omega_i, b_d, \mu_d, \omega_d\big) \tag{3.17}$$

can be estimated via MLE of the AL density, f_{AL}, defined previously in (3.3), as

$$\widehat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L_{AL}(\boldsymbol{\theta}; \boldsymbol{y}) \tag{3.18}$$

where

$$L_{AL}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{t=1}^{N} f_{AL}\big(y_t | VaR_t^{\boldsymbol{\theta}}, ES_t^{\boldsymbol{\theta}}\big) \tag{3.19}$$

refers to the conditional likelihood of f_{AL} given $\boldsymbol{y}$, and $\big(VaR_t^{\boldsymbol{\theta}}, ES_t^{\boldsymbol{\theta}}\big)$ denotes the VaR and ES forecasted with $\boldsymbol{\theta}$. This MLE approach is a frequentist estimation in the sense that $\boldsymbol{\theta}$ is unknown but fixed. Taking the Maximum a Posteriori (MAP) estimation instead can help upgrade $\boldsymbol{\theta}$ to be a

random variable $\boldsymbol{\Theta}$, allowing prior knowledge on $\boldsymbol{\Theta}$ to be sharpened each time a latest return is quoted. The transformation from MLE in (3.18) to MAP estimation proceeds with Bayes' Theorem:

$$f_{\boldsymbol{\Theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}) \propto L_{AL}(\boldsymbol{\theta};\boldsymbol{y}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}),$$

where $f_{\boldsymbol{\Theta}|\boldsymbol{Y}}$ is the (joint) posterior density of $\boldsymbol{\Theta}$ associated with the AL likelihood, interpreted as the "true" distribution of $\boldsymbol{\Theta}$; $f_{\boldsymbol{\Theta}}$ is the (joint) prior density, interpreted as a "guess" on $\boldsymbol{\Theta}$. Now the optimal estimate becomes:

$$\widehat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, f_{\boldsymbol{\Theta}|\boldsymbol{Y}}(\boldsymbol{\theta}|\boldsymbol{y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L_{AL}(\boldsymbol{\theta};\boldsymbol{y}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}). \tag{3.20}$$

In practice, the posterior $f_{\boldsymbol{\Theta}|\boldsymbol{Y}}$ is intractable, which means it is neither in a form permitting direct inference, nor in a form permitting direct sampling. Appreciated the explicit expression for L_{AL} in (3.18), the product $L_{AL}(\boldsymbol{\theta};\boldsymbol{y}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ gives rise to calibrate $\boldsymbol{\theta}$ in computational scheme.

3.3. An adaptive Metropolis MCMC Algorithm

The Metropolis-Hastings (MH) MCMC algorithm (Hastings, 1970; Metropolis et al.,1953) is widely used in optimization problems analogous to (3.20), which generates an irreducible and aperiodic Markov chain whose stationary probability density is regarded as the target intractable posterior. Let d denote the dimension of the parameter vector throughout. Following the notation in the previous section, starting from an initial state $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, at each t $(t = 1,2,\dots)$, a candidate $\boldsymbol{\theta}_t^*$ is randomly drawn from a proposal symmetric distribution centered as the previous state $\boldsymbol{\theta}_{t-1}$; the sampled $\boldsymbol{\theta}_t^*$ is accepted as $\boldsymbol{\theta}_t$ with probability

$$\alpha(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t^*) = 1 \wedge \frac{L_{AL}(\boldsymbol{\theta}_t^*;\boldsymbol{y}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t^*)}{L_{AL}(\boldsymbol{\theta}_{t-1};\boldsymbol{y}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_{t-1})};$$

if rejected, the chain stands still. The asymptotic average of $\alpha(\cdot,\cdot)$ as $d \to \infty$ is referred to as the asymptotic accept rate of the algorithm. In other words, the chain $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots\}$ travels over the support of the target distribution with step size confined to the proposal distribution and converges to the maximum posterior density that represents the model best fitting the observed returns. In particular, a sophisticated covariance structure is key to the success of convergence: if the proposed covariance yields a too narrow ambient space, the steps can be too small to visit all important regions (e.g., the mode of the target distribution), whereas if the space is too wide, the speed to converge can be very low due to too many rejections (e.g., see Figure 7 in Andrieu et al., 2003 for an illustration of the univariate case). Denote the covariance matrix of the target distribution by Σ. Gelman et al. (1996) and Roberts et al. (1997) have proved the weak convergence for a sequence of the d-dimensional MH algorithms to a Langevin diffusion as $d \to \infty$, whose speed measure is maximized by the proposal covariance matrix $2.38^2\Sigma/d$. (see e.g., Theorem 1.1 in Roberts et al., 1997; Theorem 3.1 in Gelman et al., 1996, for details). This proposal leads to an asymptotic acceptance rate 0.234, which is the so-called (theoretical) optimal acceptance rate for MH

algorithms and is admitted as a benchmark for tuning the scale of proposal such that, in practice, the chain explores the effective region efficiently.

In order to rescale the proposal covariance structure automatically, an adaptive metropolis algorithm was proposed in Roberts and Rosenthal (2009), which splits the journey of a chain into two stages: if $t \leq 2d$, then states are sampled from $N(\boldsymbol{\theta}_{t-1}, 0.1^2 I_d/d)$, with I_d denoting a $d \times d$ identity matrix; otherwise, the sampler is $0.95N(\boldsymbol{\theta}_{t-1}, 2.38^2\Sigma_t/d) + 0.05N(\boldsymbol{\theta}_{t-1}, 0.1^2 I_d/d)$, where $N(\boldsymbol{\theta}_{t-1}, 2.38^2\Sigma_t/d)$ approximates the optimal proposal stated beforehand by the current empirical covariance estimate Σ_t. As remarked in Roberts and Rosenthal (2009), here the sampler $N(\boldsymbol{\theta}_{t-1}, 0.1^2 I_d/d)$ serves as a "safety measure" to avoid the chain stuck at singular values of Σ_t. A pseudo-code summarizing how to estimate the LSTM model using the adaptive Metropolis algorithm can be found in Appendix D.1.

4. The Gaussian Mixture Model

Monte Carlo methods offer a flexible way to forecast risks, which generate random scenarios from some assumed distribution for returns and estimate risk measures based on the quantile of simulated trajectories. As seen in the previous chapter, a mixture of distributions can incorporate global explorers to sketch peaks and valleys and local explorers to fill in finer details. Empirical studies in Seyfi et al. (2021) show that a Monte Carlo method based on Gaussian mixture model (GMM) can excel in portfolio risk forecasts. Unexpectedly, my attempt of their method with four portfolios of China A-share stocks failed. Given that the failure illustrates a lesson on the importance of tail behavior under the price-limit mechanism, I record my attempt and analysis in this chapter. Description of my attempt will involve portfolio optimization and backtest, which are discussed in later chapters and are disregarded here as they do not add interpretational value.

4.1. Portfolio return simulation

GMM is a special type of normal mean-variance mixture, which groups data into several clusters and model each cluster with a multivariate normal distribution (MVN). Formally, a random vector $\boldsymbol{X} = (X_1, \dots, X_k)$ follows a GMM of N_c clusters, denoted $\boldsymbol{X} \sim Gmm_{N_c}(\boldsymbol{\omega}, \boldsymbol{\mu}, \Sigma)$, if its density can be written as

$$f(\boldsymbol{x}) = \sum_{j=1}^{N_c} \omega_j \, \phi(\boldsymbol{x} \,|\boldsymbol{\mu}_j, \Sigma_j),$$

where $\boldsymbol{x} = (x_1, \dots, x_k) \in \mathbb{R}^k$, $\boldsymbol{\omega} = (\omega_1, \dots, \omega_{N_c}) \in \mathbb{R}^{N_c}$ with $\sum_{j=1}^{N_c} \omega_j = 1$, $\boldsymbol{\mu}_j \in \mathbb{R}^k$, and Σ_j is a $k \times k$ covariance matrix for $j = 1, \dots, N_c$. Here, $\phi(\boldsymbol{x} \,|\boldsymbol{\mu}_j, \Sigma_j)$ is the density of MVN, given by

$$\phi(\boldsymbol{x} \,|\boldsymbol{\mu}_j, \Sigma_j) = (2\pi)^{-k/2} \det(\Sigma_j)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)\right\}.$$

Let $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ be a sequence of observed samples for $\boldsymbol{X}$ with $\boldsymbol{x}^{(i)} \in \mathbb{R}^k$ for $i = 1, \dots, n$. Let $z_1, \dots, z_{N_c}$ be a sequence of latent variables that follows a binary distribution such that $z_j \in \{0,1\}$

with $\mathbb{P}(z_j = 1) = \omega_j$ for $j = 1, \dots, N_c$. That is, for each i, $z_j = 1$ iff $\boldsymbol{x}^{(i)}$ is in cluster j. Following procedures in Seyfi et al. (2021), $\{\boldsymbol{\omega}, \boldsymbol{\mu}, \Sigma\}$ can be estimated by the Expectation–Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977) as follows: (1) initialize $\{\boldsymbol{\omega}, \boldsymbol{\mu}, \Sigma\}$ via the K-means algorithm (2) calculate $r_{ij} = \omega_j \phi(\boldsymbol{x}^{(i)} | \boldsymbol{\mu}_j, \Sigma_j) / (\sum_{j=1}^{N_c} \omega_j \phi(\boldsymbol{x}^{(i)} | \boldsymbol{\mu}_j, \Sigma_j))$, for each $i = 1, \dots, n$ and $j = 1, \dots, N_c$; (3) update $\omega_j = \sum_{i=1}^{n} r_{ij} / n$, $\boldsymbol{\mu}_j = (\sum_{i=1}^{n} r_{ij} \boldsymbol{x}^{(i)}) / (\sum_{i=1}^{n} r_{ij})$, and $\Sigma_j = \sum_{i=1}^{n} r_{ij} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)^T / (\sum_{i=1}^{n} r_{ij})$ for each $j = 1, \dots, N_c$; (4) repeat (1) to (3) until some convergence condition(s) met (e.g., number of iterations, error tolerance, etc.).

Now suppose we have a portfolio of k assets with a weight vector $\boldsymbol{w} = (w_1, \dots, w_k) \in \mathbb{R}^k$. Let $\{\boldsymbol{r}^{(1)}, \dots, \boldsymbol{r}^{(n)}\}$ be a collection of realized log returns with $\boldsymbol{r}^{(i)} \in \mathbb{R}^k$ for each $i = 1, \dots, n$. Seyfi et al. (2021) suggested a way to simulate portfolio returns and estimate T-day ahead ($T = 1, 2, \dots$) risk measures accordingly by sampling from a GMM. Their approach is summarized through the following steps.

1. Determine the number of clusters, N_c, of the model and estimate parameters $\{\boldsymbol{\omega}, \boldsymbol{\mu}, \Sigma\}$ by the EM algorithm using realized log returns $\boldsymbol{r}^{(1)}, \dots, \boldsymbol{r}^{(n)}$.
2. For $j = 1, \dots, N_c$, let $\lceil \omega_j T \rceil$ be the number of samples to be drawn from the jth cluster, MVN$(\boldsymbol{\mu}_j, \Sigma_j)$; hence the total number of samples drawn is $N = \sum_{j=1}^{N_c} \lceil \omega_j T \rceil$. Each sample consists of m independent returns. Denote the generated samples by $\{\tilde{R}_t\}_{t=1,\dots,N}$, where

$$\tilde{R}_t = \begin{bmatrix} \tilde{r}_t^{(1)} & \dots & \tilde{r}_t^{(m)} \end{bmatrix}^T = \begin{bmatrix} \tilde{r}_{t,1}^{(1)} & \dots & \tilde{r}_{t,k}^{(1)} \\ \vdots & \ddots & \vdots \\ \tilde{r}_{t,1}^{(m)} & \dots & \tilde{r}_{t,k}^{(m)} \end{bmatrix}.$$

3. Sum the N matrices $\{\tilde{R}_t\}_{t=1,\dots,N}$ element-wise to get T-holding period returns

$$\tilde{R} = \begin{bmatrix} \sum_{t=1}^{N} \tilde{r}_{t,1}^{(1)} & \dots & \sum_{t=1}^{N} \tilde{r}_{t,k}^{(1)} \\ \vdots & \ddots & \vdots \\ \sum_{t=1}^{N} \tilde{r}_{t,1}^{(m)} & \dots & \sum_{t=1}^{N} \tilde{r}_{t,k}^{(m)} \end{bmatrix}.$$

4. Adjust the simulated returns $\tilde{R}$ for volatility to produce more conservative risk forecasts. For each stock $s = 1, \dots, k$, the adjustment factor is $v_s = \sigma_{short,s} / \sigma_{long,s}$, where σ_{short} refers to standard deviation of stock returns over a short period (e.g., 70 days) while σ_{long} refers to that over a longer period (e.g., 250 days). The rescaled returns are

$$\tilde{R}_{adj} = \begin{bmatrix} v_1 \sum_{t=1}^{N} \tilde{r}_{t,1}^{(1)} & \dots & v_k \sum_{t=1}^{N} \tilde{r}_{t,k}^{(1)} \\ \vdots & \ddots & \vdots \\ v_1 \sum_{t=1}^{N} \tilde{r}_{t,1}^{(m)} & \dots & v_k \sum_{t=1}^{N} \tilde{r}_{t,k}^{(m)} \end{bmatrix}.$$

5. Get the simulated portfolio returns

$$\tilde{r}_p = \begin{bmatrix} \tilde{r}_p^{(1)} \\ \vdots \\ \tilde{r}_p^{(m)} \end{bmatrix} = \tilde{R}_{adj}\boldsymbol{w}.$$

6. Forecast T-day ahead VaR and ES at confidence level α (e.g., 0.025 refers to 97.5%) as

$$\widehat{VaR}_\alpha = \text{quantile}\big(\tilde{r}_p, \alpha\big)$$

$$\widehat{ES}_\alpha = \frac{1}{m'}\sum_{l=1}^{m} \tilde{r}_p^{(l)} \mathbb{1}\big\{\tilde{r}_p^{(l)} \le \widehat{VaR}_\alpha\big\},$$

where m' denotes the number of $\tilde{r}_p^{(l)}$ such that $\tilde{r}_p^{(l)} \le \widehat{VaR}_\alpha$ for $l = 1, \dots, m$.

4.2. A failed attempt: price limit mechanism and the magnet effect

Since Dec 26, 1996, the Chinese regulator started to impose a limit of $\pm 10\%$ for daily price change on both Shanghai stock exchange and Shenzhen stock exchange, with exemption on the following date: (1) the IPO date; (2) the first trading date after the stock split-structure reform; (3) the first trading date after seasonal offerings; (4) The first trading date after material assets restructuring; (5) The first re-listing date of de-listed stocks. Consequently, simulation for Chinese stocks must be consistent with the price-limit mechanism.

Motivated by the nice backtest performance of VaR for a portfolio of US stocks in Seyfi et al. (2021), I attempt to apply their method on 4 portfolios of $k = 22$ representative China A-share stocks from Aug 28, 2001 to Dec 31, 2021 to make daily ES forecasts with a moving window of 250 days. Following steps in 4.1, stock prices are transformed into log-returns; on each moving window, I fit a GMM of $N_c = 4$ (see Appendix A for sufficiency) and simulate $m = 3{,}000$ return series for each portfolio to forecast 97.5% VaR and ES. As shown in Appendix B.2, my stock returns are almost bounded above by log(1.1) but are not bounded below by log(0.9) mainly due to large price drops during the 2005–2007 Chinese Split-Share Structure Reform (see Li et al., 2014 for review). Therefore, I only impose an upper truncation of log(1.1) upon sampling for the sake of conservativeness. In addition, multiple settings for the rescale factor v_s proposed in Seyfi et al. (2021) are compared; as even the best leads to poor forecast, the details are omitted. My ES forecasts for all portfolios are rejected with strong evidence in 300 days on average since backtest.

Figure 4.1 shows the simulated return series for each portfolio and hints a reason of the failure: the simulated losses (i.e., negative returns) are insufficiently extreme while the simulated returns (i.e., positive returns) are unnecessarily thick, leading to the level at which VaR and ES fluctuate close to zero. When making daily forecast using the method in Seyfi et al. (2021), each time I actually sample from a MVN that is associated with the largest weight (i.e., set $T = 1$ in step 2).

In this regard, I try to discretize $T = 1$ into multiple bins to see if sampling from a mixture instead can make the situation better. Figure 4.2 shows 264 return series simulated with different size of bins for the naïve portfolio. One point should be noted: as samples are drawn independently, simulating more paths will only make the colored region wider without changing much the pattern. Taking a closer look at the plots, we can see that sampling from mixture models amplifies the asymmetry in gains and losses; however, all the simulated paths tend to expand towards the tail regions uniformly, and such behavior seems to be more pronounced for gains for which a limit was imposed upon sampling. This observation violates the "magnet effect" documented in much of empirical research: the asset price will accelerate towards the bounds as it approaches the limits set for the prices (Wang et al., 2021). Admittedly, any observation involves subjectiveness; nevertheless, it is worth exploring the question: is there a sampling that accommodates the price-limit mechanism while conforms to the magnet effect simultaneously?

The truncated distribution seems to provide an answer. For clarity, I focus on the one-dimensional case. A truncated normal distribution has CDF

$$\Psi_{\mu,\sigma,a,b}(x) = \begin{cases} 0 & x \le a \\ \frac{\Phi_{\mu,\sigma}(x) - \Phi_{\mu,\sigma}(a)}{\Phi_{\mu,\sigma}(b) - \Phi_{\mu,\sigma}(a)} & a < x < b \\ 1 & b \le x \end{cases}$$

where $\Phi_{\mu,\sigma}$ is the CDF of a normal distribution with mean μ and standard deviation σ, a is the lower truncation, and b is the upper truncation. For some $u \in [0,1]$, the inverse CDF of $\Psi_{\mu,\sigma,a,b}$ is

$$\Psi^{-1}_{\mu,\sigma,a,b}(u) = \Phi^{-1}_{\mu,\sigma}\Big(\Phi_{\mu,\sigma}(a) + u\left(\Phi_{\mu,\sigma}(a) - \Phi_{\mu,\sigma}(a)\right)\Big). \tag{4.1}$$

Suppose $\mu = 0, \sigma = 1, a = -\infty$, and $b = \log(1.1)$, which refers to a standard normal distribution truncated at $\log(1.1)$. The middle panel in Figure 4.3 portraits 1,024 draws from this truncated distribution using (4.1) and 1,024 draws from the standard normal with values greater than $\log(1.1)$ rejected; the left panel zooms in the upper tail region while the right panel zooms in the lower tail region. For both methods, I use a same randomized Sobol sequence to improve sampling quality.

Although the two samplers are the same in the sense of descriptive statistics (see notes in Figure 4.3 for information), they show different behaviors in expansion to the tails. More precisely, the truncated normal sampler extends by layer, with the thickness of each layer tapering off as it approaches the tail regions, whereas the normal sampler extends in a more uniform manner. Compared to lower tails, the phenomenon is more significant for the upper tail region, where a limit was imposed. Such decrease in thickness as the sample approaches the limit seems to be in line with the price acceleration assumed in magnet effect; with that being said, whether truncated distribution indeed explains the magnet effect cannot be ascertained without further investigation. What is clear, at this moment, is the difference between sampling from a truncated normal distribution and sampling from a normal distribution subject to the same bound(s). I notice some papers emphasize the impracticality of drawing truncated normal samples from a normal rejection sampler in the sense of efficiency (see e.g., Geweke, 1991; Li and Ghosh, 2015). I would like to

underline their features in terms of tail behavior, and my failed attempt has demonstrated how an inaccurate portrait of the tail thickness of gains and losses can squeeze risk estimates.

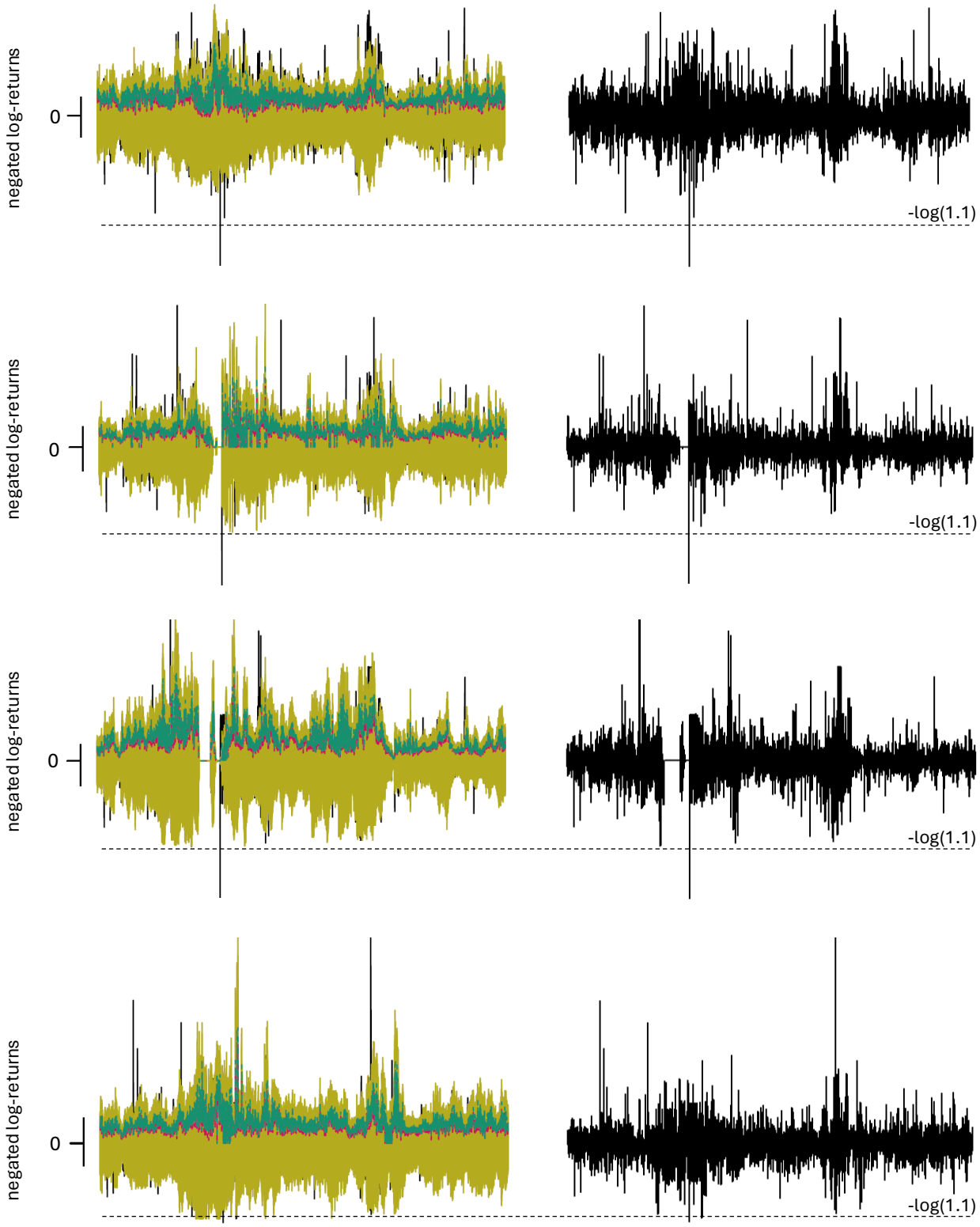

Figure 4.1: The failed attempt; gold color refers to 3,000 simulated portfolio returns, red refers to 97.5% VaR forecast, green refers to 97.5% ES forecast, and black refers to the realized portfolio return; x-axis refers to 2002 to 2021; from top to bottom: naïve 1/N portfolio, min-ES portfolio, max-return portfolio, and mean-variance portfolio; dotted line refers to the 10% price limit on gains

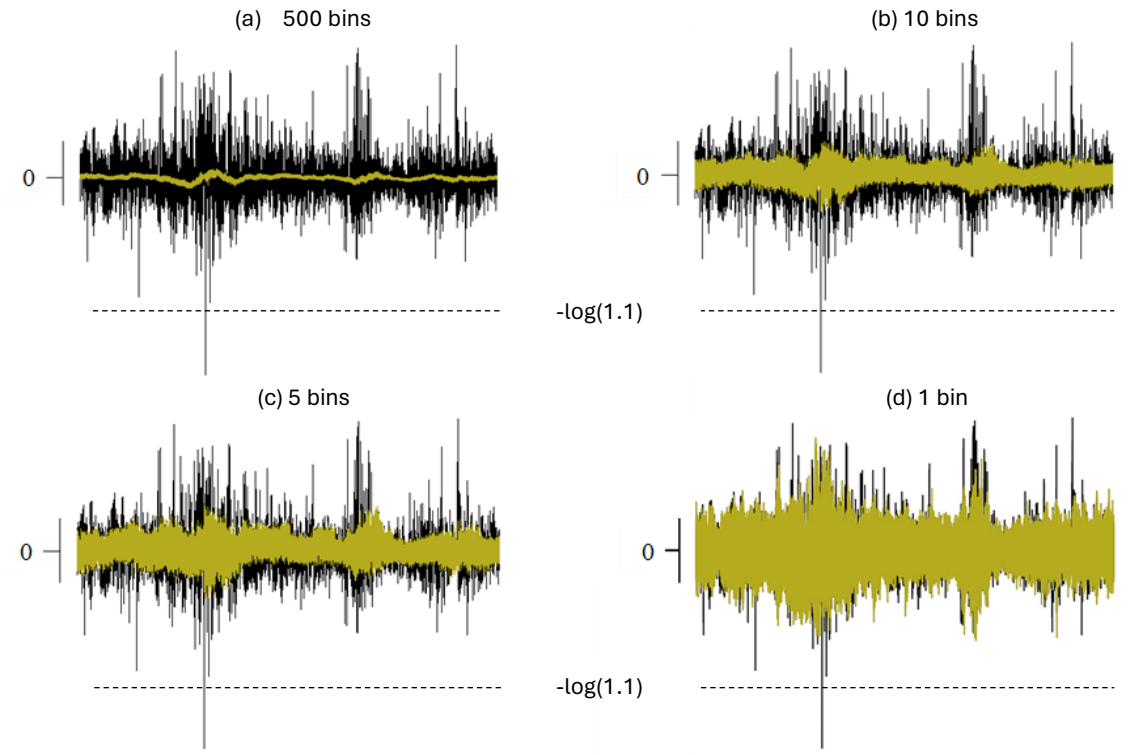

Figure 4.2: 264 return series simulated with different size of bins for the naïve 1/N portfolio; y-axis is negated log-return; x-axis is 2002 to 2021; black is the realized returns while gold is the simulated returns; T=(1/500)*500 for (a), T=(1/10)*10 for (b), T=(1/5)*5 for (c), and T=1 for (d) (i.e., no discretization); dotted line refers to the 10% limit imposed on gains

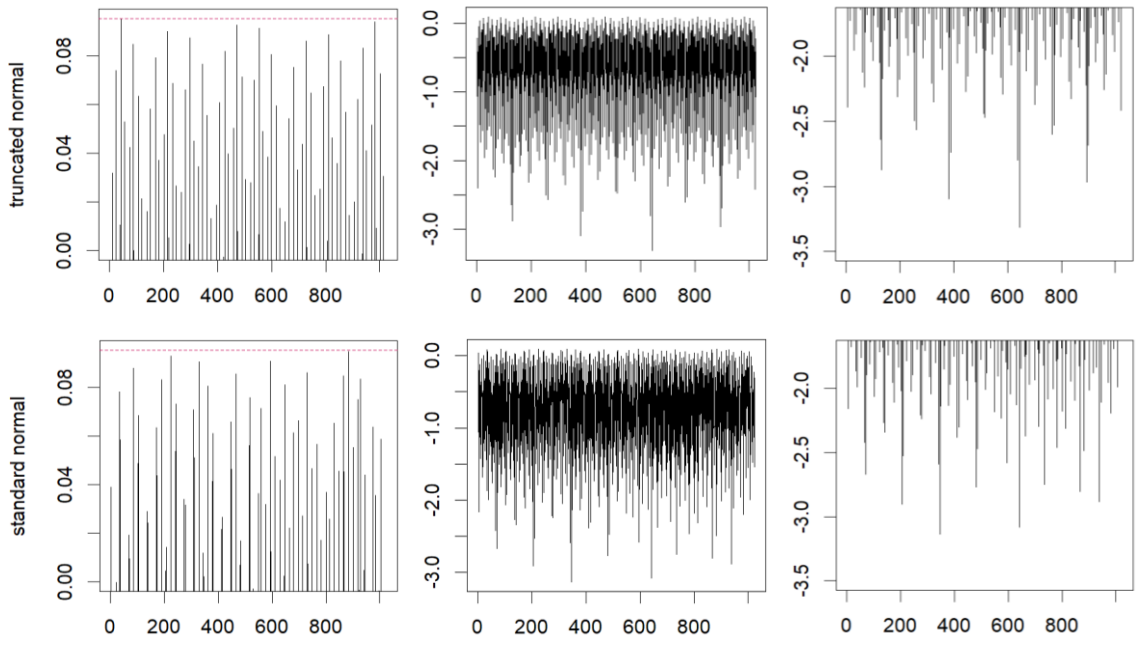

Figure 4.3: middle panel: 1,024 draws normal truncated at log(1.1) and 1,024 draws from normal with samples > log(1.1) rejected; left: upper tail region; right: lower tail region; dotted line in red is log(1.1); x-axis is the index of draws; truncated: min -3.13 mean -0.74 max 0.09 kurtosis 3.60 skewness -0.92; normal: min -3.31 mean -0.74 max 0.09 kurtosis 3.54 skewness -0.93

5. Empirical Setting: Portfolio construction and risk forecast

The Basel Committee initiated a Fundamental Review of the Trading book (FRTB) in 2012 (as a part of Basel III) to adapt existing market risk framework to the lessons learned during the global financial crisis. A key update is the change of standard market risk measure from the 99% VaR in Basel 2.5 to 97.5% ES. The revision is finalized in 2016, included as a part of the so-called Basel IV. Following the implementation of Basel IV, the Chinese regulatory commission CBIRC recently updated the VaR to ES, making ES the most important market risk metric for the ongoing banking practices in China. In this chapter, I discuss the details regarding forecasting 97.5% ES for the Chinese market using methodologies presented in Chapter 2 and Chapter 3, contemplating four major classes of market risk and four strategies of portfolio optimization.

5.1. Data Description

5.1.1. Financial return series

Under Basel III, market risk exposure is mapped to four asset classes: equity, fixed income, foreign exchange, and commodity. For comprehensiveness, for each class, I select a proxy covering the turbulent periods: 2008–2009 global financial crisis and 2015–2016 Chinese market crash to perform risk estimation. Specifically, for equity, I use daily closing prices of the Shanghai Composite Index (000001); for fixed income, I use daily closing prices of the ChinaBond New Composite Index (CBA00101); for foreign exchange, I use daily ratios of RMB to the US dollar (RMB/USD); for commodity, I use daily closing physical price of Au99.99 traded on the Shanghai Gold Exchange. Observation horizon and descriptive statistics for each series can be found in Appendix B.1.

5.1.2. Portfolios

Rather than market indices, a small-sized portfolio can exhibit more specific uncertainties due to idiosyncratic risk and the risk preference of investors. Such uncertainty can improve the soundness in evaluating risk estimation methods. Similar to the portfolio setup in Wang et al. (2023), for each of the 11 industries in the Global Industry Classification Standard (GICS), I pick two stocks from China A-shares with the largest market capitalization as of year-end 2021, and construct portfolios using four strategies: naïve 1/N, minimizing ES, maximizing return, and Markowitz mean-variance model. Listed firms in China are subjected to frequent trading suspensions, in either a mandatory or voluntary manner, for the sake of transparency and equality in market information (see He et al., 2019 for example and effect). The time horizon of my stocks is from Aug 27, 2001 to Dec 31, 2021, during when all of them has undergone trading suspensions for 26 to 820 days in total. Missing values due to such halt and resumption are addressed as follows:

if a stock is halted for one day, I fill in the average of the adjacent values; otherwise, I fill in the most recent available quote. Details of stock profiles can be found in Appendix B.1.

Naïve 1/N refers to the rule in which the fraction 1/N of wealth is allocated to each of the N assets available for investment at each rebalancing date (DeMiguel et al., 2009). I perform this strategy with the initial wealth assumed to be one and rebalance the weight at the end of each year. Minimizing ES (min-ES) refers to the portfolio that minimizes the estimated ES exposure subject to a minimum required return using the dual representation of ES based on VaR. Let $\boldsymbol{r}_t = (r_{t,1}, \dots, r_{t,22})^T$ be the random vector of log-returns for the 22 stocks at time t, $f(\boldsymbol{r}_t)$ be the density of $\boldsymbol{r}_t$, and $\boldsymbol{x}_t$ be the weight vector. Following Rockafellar and Uryasev (2002), the optimization problem is given by

$$\begin{aligned} & \min_{\xi, \boldsymbol{x}_t} \left\{ \xi + \frac{1}{1-\alpha} \int_{\boldsymbol{r}_t \in \mathbb{R}^n} (-\boldsymbol{x}_t^T \boldsymbol{r}_t - \xi)^+ f(\boldsymbol{r}_t) d\boldsymbol{r}_t \right\} \\ \approx \; & \min_{\xi, \boldsymbol{x}_t} \left\{ \xi + \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} \left(-\boldsymbol{x}_t^T \boldsymbol{r}_{t-i} - \xi \right)^+ \right\} \\ = \; & \min_{\xi, \boldsymbol{x}_t} \left\{ \xi + \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} y_{t,i} \right\} \end{aligned} \tag{5.1}$$

$$\begin{aligned} s.t \quad & \boldsymbol{x}_t \geq \mathbf{0}, \;\; \boldsymbol{x}_t^T \mathbf{1} = 1 \\ & y_{t,i} \geq 0, \;\; y_{t,i} \geq -\boldsymbol{x}_t^T \boldsymbol{r}_{t-i} - \xi, \;\; \mathrm{i} = 1, 2, \dots, \mathrm{n} \\ & \boldsymbol{x}_t^T \mathbb{E}[\boldsymbol{r}_t] \geq \mu, \end{aligned}$$

where I set μ as the average of past n-day realized returns of the naïve portfolio, $n = 250$ as the size of estimation window, and $\alpha = 97.5\%$ as the confidence level. The expected return $\mathbb{E}[\boldsymbol{r}_t]$ is estimated by the CH-4-factor model proposed in Liu et al. (2019), in which

$$\mathbb{E}[r_{t,j}] = RF_t + \beta_j MKT_t + s_j SMB_t + v_j VMG_t + t_j PMO_t, \tag{5.2}$$

with RF, MKT, SMB, VMG, PMO denoting the risk free rate, market factor, size factor, value factor, and turnover factor, respectively, for $j = 1, \dots, 22$ (see Prof. Robert F. Stambaugh's website for daily factor values). Likewise, maximizing return (max-R) refers to the portfolio that maximizes the estimated expected return subject to a maximum ES tolerance. The algorithm is

$$\begin{aligned} & \max_{\boldsymbol{x}_t} \boldsymbol{x}_t^T \mathbb{E}[\boldsymbol{r}_t] \\ s.t \quad & \boldsymbol{x}_t \geq \mathbf{0}, \;\; \boldsymbol{x}_t^T \mathbf{1} = 1 \\ & y_{t,i} \geq 0, \;\; y_{t,i} \geq -\boldsymbol{x}_t^T \boldsymbol{r}_{t-i} - \xi, \;\; i = 1, 2, \dots, n \\ & \xi + \frac{1}{(1-\alpha)n} \sum_{i=1}^{n} y_{t,i} \leq \gamma, \end{aligned}$$

where I set $n = 250$ as the size of estimation window and γ as the average of the past n-day empirical ES estimates for the naïve portfolio. Again, $\mathbb{E}[\boldsymbol{r}_t]$ is estimated using (5.2). For the Markowitz mean-variance portfolio (MV), the optimization problem becomes

$$\min_{\boldsymbol{x}_t} \; -\boldsymbol{x}_t^T \mathbb{E}[\boldsymbol{r}_t] + \frac{1}{2}\boldsymbol{x}_t^T \Sigma_t \boldsymbol{x}_t$$
$$s.t \qquad \boldsymbol{x}_t \geq \mathbf{0}, \;\; \boldsymbol{x}_t^T \mathbf{1} = 1,$$

where I estimate Σ_t by the empirical covariance matrix of past 250-day returns and estimate $\mathbb{E}[\boldsymbol{r}_t]$ using (5.2). In summary, the weight vector $\boldsymbol{x}_t$ is updated daily for each portfolio, with that of naïve 1/N being rebalanced to 1/22 on the last trading day of each year and that of others being estimated using the most recent 250 observations. A visualization of how $\boldsymbol{x}_t$ evolves can be found in Appendix B.3. Portraits for portfolios can be found in Appendix B.2, and descriptive statistics can be found in Appendix B.1.

5.2. **Forecasting ES**

5.2.1. The GARCH-based family

Under Basel IV, banks are required to compute 97.5% ES on a daily basis (for the desk-level and bank-wide internal model of the RWA) with a 12-month estimation window (consult Section C.3, BCBS, 2019, for computation). To align with the regulatory standards, an AR(1)-GARCH(1,1) model is fitted to all eight series with a moving estimation window of 250 observations to forecast daily VaR and ES at a 97.5th percentile. Each series is transformed into negated percentage log-returns and has passed the ADF test for stationarity with p-value less than 0.01. Outliers are not cleaned for the sake of conservativeness in risk forecasting. On each moving window, the model with normal, t, or skewed-t innovations is estimated using the solnp solver in the rugarch package, and VaR and ES are forecasted following (2.2) to (2.10). The convergence of the solver is noted to vary with regime shifts: for portfolios, the minimum relative tolerance upsurges during the pre- and post-periods of the global financial crisis; for RMB/USD, the tolerance peaks during the 2005–2007 RMB exchange rate reform. Numerical details on convergence are provided in Table 5.1.

Standardized residuals from each moving window are then used as input for the GARCH-EVT and the GARCH-FHS methods. For GARCH-EVT, the threshold u is chosen as the first u after which the linear pattern in the mean residual life plot, which represents the mean of the excesses of u, seems to disappear. Plots for standardized residuals and the mean residual life for each series can be found in Appendix C, and the selection of u is as follows: $u = 9$ for RMB/USD; $u = 1.5$ for MV; $u = 2$ for min-ES; $u = 1$ for all others. In order to control the variance from inadequate exceedances, I set a decision rule: on each moving window, if the u selected above yields at least 10% of observations on that window, then VaR and ES are forecasted by GARCH-EVT; otherwise, the forecasts are still produced from AR(1)-GARCH(1,1). The choice of 10% takes in to account the rule-of-thumb that the threshold is chosen around the 90th quantile (see, e.g., McNeil and Frey, 2000; Karmakar and Shukla, 2015; Bee et al., 2016). Hence, there is no difference in my risk forecasts using the GARCH-EVT and the AR(1)-GARCH(1,1) unless the estimation window witnesses adequate extreme events.

For GARCH-FHS, as mentioned in Chapter 2, I follow the implementation in Nolde and Ziegel (2017). Their method differs from the AR(1)-GARCH(1,1) model in the sense that it exploits the empirical quantile of the realized standardized residuals rather than the assumed innovations when estimating $F_Z^{\leftarrow}(\alpha)$ in (2.2) and $ES_\alpha(Z)$ in (2.3). As such, the gap between the forecast made by GARCH-FHS and that made by AR(1)-GARCH(1,1) diagnoses how much heteroskedasticity and leptokurtosis the AR(1)-GARCH(1,1) model has actually captured. Unsurprisingly, it is observed that the GARCH-FHS exhibits dramatic reaction to shocks both in individual and in cluster. A portrait for ES forecasts produced by the GARCH-based family is provided in Figures 5.2 and 5.3.

5.2.2. The LSTM-AL model

Turning the focus to the LSTM-AL model, all the eight series are transformed into percentage log-returns, with 4688 observations for portfolios, 5094 for SSE, 4625 for AU99.99, 5254 for CBA, and 5282 for RMB/USD, respectively. For computational efficiency, I use an expanding window to estimate $\boldsymbol{\theta}$ in (3.17) and predict out-of-sample risk forecasts for the same length. Specifically, for each series, let the first observation refer to day one; the first 250 observations are used to predict (VaR, ES) at a 97.5^{th} percentile for days 251 through 500; the first 500 observations are used to predict for days 501 through 1000; the first 1000 observations are used to predict for days 1001 through 2000; the first 2000 observations are used to predict for days 2001 through 4000; observations from day 2001 to day 4000 are used to predict for days 4001 through the end.

It is commonly agreed in the literature that a mature MCMC algorithm should not be affected much by initialization. That is, the chain travels to the same destination no matter where it departs. In reality, however, a chain can never be ascertained to actually converge within finite iterations, and there exists a tradeoff between numerical stability and a reasonable amount of running time. Therefore, in practice, the chain can be sensitive to initial state. To control such sensitiveness and reinforce the robustness of risk forecasts, on each estimation window, six chains are generated following the pseudocode in Appendix D.1, where each chain starts at a different state and consists of 15,000 runs with the first 4,500 disregarded (i.e., the burn-in period). Initial states are selected to be diversified, and the choice of 15,000 and 4,500 is supported by visual inspection of trace plots over dozens of trails; details on consideration can be found in Appendix D.2.

For each chain, the average of accepted samples (i.e., the average of distinct states after the burn-in period) refers to an estimate for $\boldsymbol{\theta}$ and is employed to generate one forecast series for (VaR, ES) using (3.7) to (3.16). For simplicity, initial values for VaR in (3.7) and $LSTM$ in (3.10) are all set to zero. Based on my observation, the ES forecasted on each estimation window can be characterized into three scenarios: (1) the six series coincide with each other in either a conservative or a parsimonious manner (see e.g., the top panel in Figure 5.1); (2) the six series spread out, with all being conservative (see e.g., the middle panel in Figure 5.1); (3) the six series spread out, with some being conservative while some being parsimonious (see e.g., the bottom

panel in Figure 6.1). For the first two scenarios, the final forecast for (VaR, ES) on that window is taken by the average; for the last scenario, it is taken by the average where the (VaR, ES) with the least conservative ES is excluded. The final forecast over each window constitutes the forecast series that is used for backtesting.

Li et al. (2021) try to target the optimal acceptance rate (23.4%) when training their model without mentioning much the purpose. Here, I would like to clarify that the acceptance rate of a chain not necessarily suggests the performance of risk forecasts. I have observed multiple chains with almost optimal acceptance rate but poor VaR and ES forecasts and vice versa. As discussed previously in Chapter 3, a low acceptance rate only suggests the inefficiency and ineffectiveness of the proposal density. Therefore, I do not aim at the optimal acceptance rate but rather take it as a measure of how feasible the algorithm is for an estimation window. The average acceptance rate on each estimation window is summarized in Table 5.2. Similar to what is noted for the GARCH model, the feasibility of the algorithm also varies with regime shifts: for RMB/USD, the acceptance rate sinks to its lowest on the window during when the RMB exchange rate reform was initiated; for all other series, the rate hits a 20-year low during the global financial crisis. A portrait for ES forecasts produced by the LSTM-AL model is provided in Figure 5.2 and 5.3.

A key point: how to tune the LSTM-AL model to achieve accuracy and conservativeness in risk forecasting is neglected in Li et al. (2021). Comparing forecasts produced by each chain, I notice that what differentiates a "good" forecast from its "poor" peer mainly lies in γ_0, γ_1 in (3.8) and the 12 parameters of $LSTM$ in (3.10). An example is provided in Table 5.3. In this sense, how to tune the proposal density of these 14 parameters to adapt to regime shifts should be a topic for further investigation.

	Normal		Student's t		Skewed-t	
	window	tol	window	tol	window	tol
SSE	all	1e-8	all	1e-8	all	1e-8
AU99.99	all	1e-8	all	1e-8	all	1e-8
CBA	all	1e-8	all	1e-8	all	1e-8
RMB/USD	2052nd	1e-5	656th	1e-5	all	1e-8
Naïve 1/N	all	1e-8	all	1e-8	all	1e-8
Min-ES	1000th	1e-6	1085th	1e-4	1113th	1e-5
MV	994th	1e-5	956th	1e-4	1194th	1e-5
Max-R	all	1e-8	all	1e-8	all	1e-8

Table 5.1: The maximum of the minimum relative tolerance for convergence for each series; window refers to the index of the estimation window where the maximum is achieved; tol refers to the numerical level of tolerance; the first estimation window starts from 2002-09-13 for min-ES and MV; from 2002-01-04 for RMB/USD; each window contains 250 days

	Estimation Window				
	1–250	1–500	1–1000	1–2000	2001–4000
SSE	14.97	6.87	5.76	17.11	26.06
AU99.99	8.10	3.10	0.90	11.73	16.96
CBA	25.13	10.68	8.59	31.62	13.08
RMB/USD	1.73	0.46	2.83	3.80	20.60
Naïve 1/N	8.14	11.76	17.67	6.96	24.73
Min-ES	13.32	16.01	7.25	29.93	13.61
MV	17.58	12.42	3.64	9.79	6.97
Max-R	23.06	8.77	6.91	3.88	16.63

Table 5.2: The average acceptance rate (%) on each estimation window for each series; six chains, each of 15,000 runs, are trained per window, and the average acceptance rate is calculated as the total number of accepted samples over all chains divided by 90,000

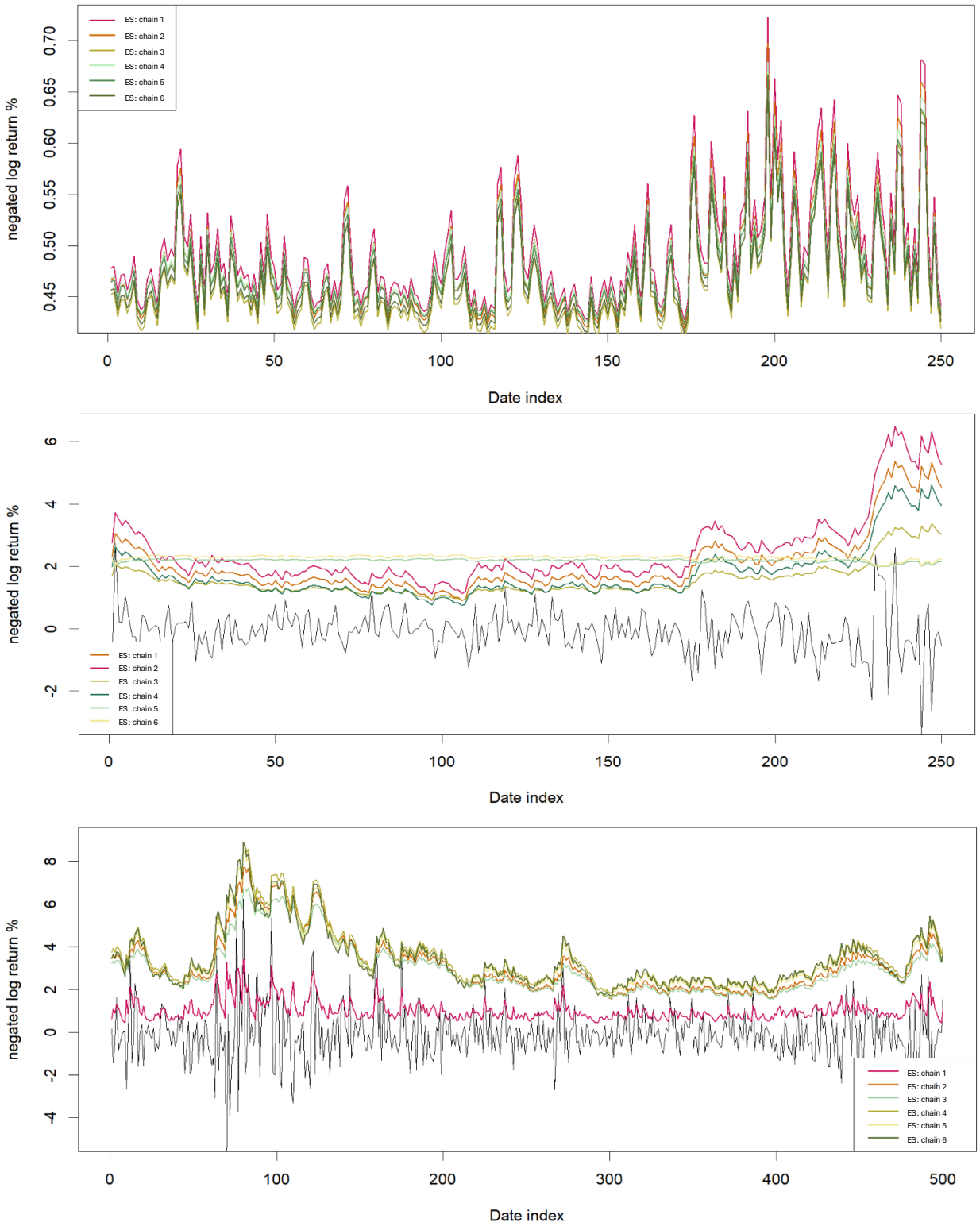

Figure 5.1: Three scenarios of ES forecasts produced by six chains on an estimation window; top panel: 97.5% ES for CBA on the first moving window; middle panel: 97.5% ES for AU99.99 on the first moving window; bottom panel: 97.5% ES for AU99.99 on the second moving window

	Index of Chain						
	1	2	3	4	5	6	variance %
β_0	-0.311007	-0.165284	-0.133114	-0.212591	-0.223983	-0.247478	0.390673
β_1	0.428585	0.928215	0.939077	0.902918	0.899536	0.877425	3.901533
$\boldsymbol{\gamma_0}$	-0.972224	-0.098071	0.035963	-0.306136	-0.393220	-0.444244	**12.225206**
$\boldsymbol{\gamma_1}$	-0.333574	-0.888750	-1.172942	-0.348624	-0.690556	-0.046477	**17.160718**
α_0	-0.325103	-0.452600	-0.621833	-0.205915	-0.209205	-0.167938	3.129453
α_1	0.669185	1.016076	1.188637	0.463720	0.557916	0.532991	8.696474
μ_f	-0.057387	-0.123725	0.536567	0.274915	-0.204641	0.174153	7.886287
$\boldsymbol{\omega_f}$	0.093502	0.574316	0.589908	0.482447	0.490329	-0.224424	**10.787836**
b_f	0.280358	0.272671	0.208127	0.315524	0.044410	-0.004232	1.798343
μ_i	0.203016	-0.177204	-0.330279	-0.547399	-0.040773	-0.242907	6.542054
ω_i	-0.364225	0.403780	0.247932	-0.011639	0.230732	-0.015626	7.412260
b_i	0.294775	0.396108	0.448369	0.170702	0.477489	-0.150198	5.555223
$\boldsymbol{\mu_d}$	-0.205460	0.634319	0.637368	0.205629	-0.088263	0.132303	**12.595863**
ω_d	-0.135147	0.425125	0.484809	0.139769	0.560613	0.217190	6.753667
b_d	-0.212224	-0.215484	-0.257302	-0.090316	-0.365100	0.160827	3.301228
$\boldsymbol{\mu_o}$	-0.373114	-0.263379	0.703383	0.471332	-0.223538	-0.457801	**23.632420**
$\boldsymbol{\omega_o}$	0.332246	0.643929	0.888021	0.440560	0.143811	-0.234409	**15.325066**
$\boldsymbol{b_o}$	-0.292018	0.315947	0.654489	0.422382	0.282876	0.128238	**10.146841**

Table 5.3: The estimated parameter set of the LSTM-AL model by using different chains; here chain 1–6 corresponds to the six ES forecasts plotted in the bottom panel of Figure 6.1 (e.g., chain 1 refers to the red line); variance is the sample variance of the six parameter estimates; parameters whose variance is over 10% is highlighted in boldface

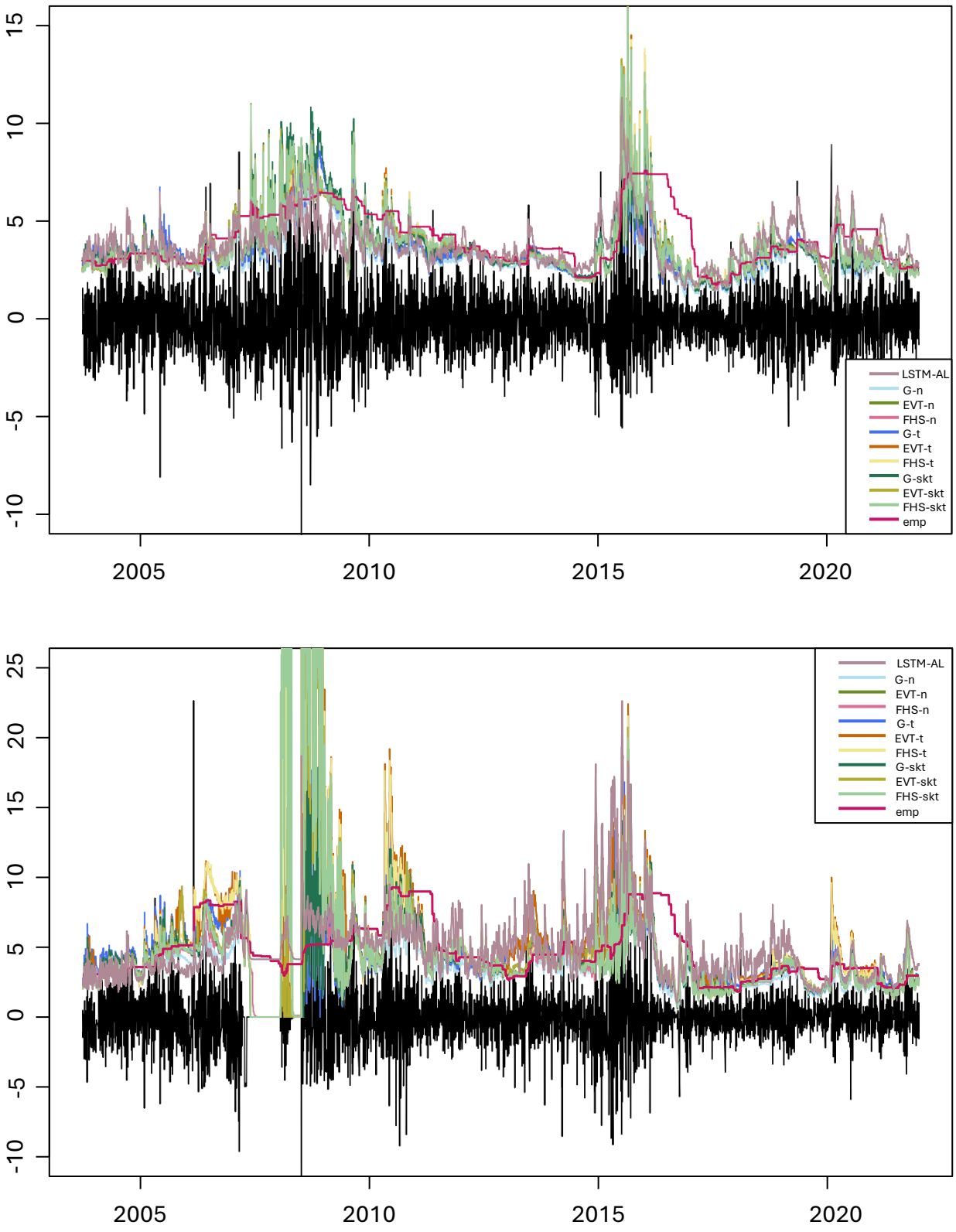

Figure 5.2: 97.5% ES forecasts (negative log-return %) for portfolios; top panel: Naïve 1/N; bottom: MV

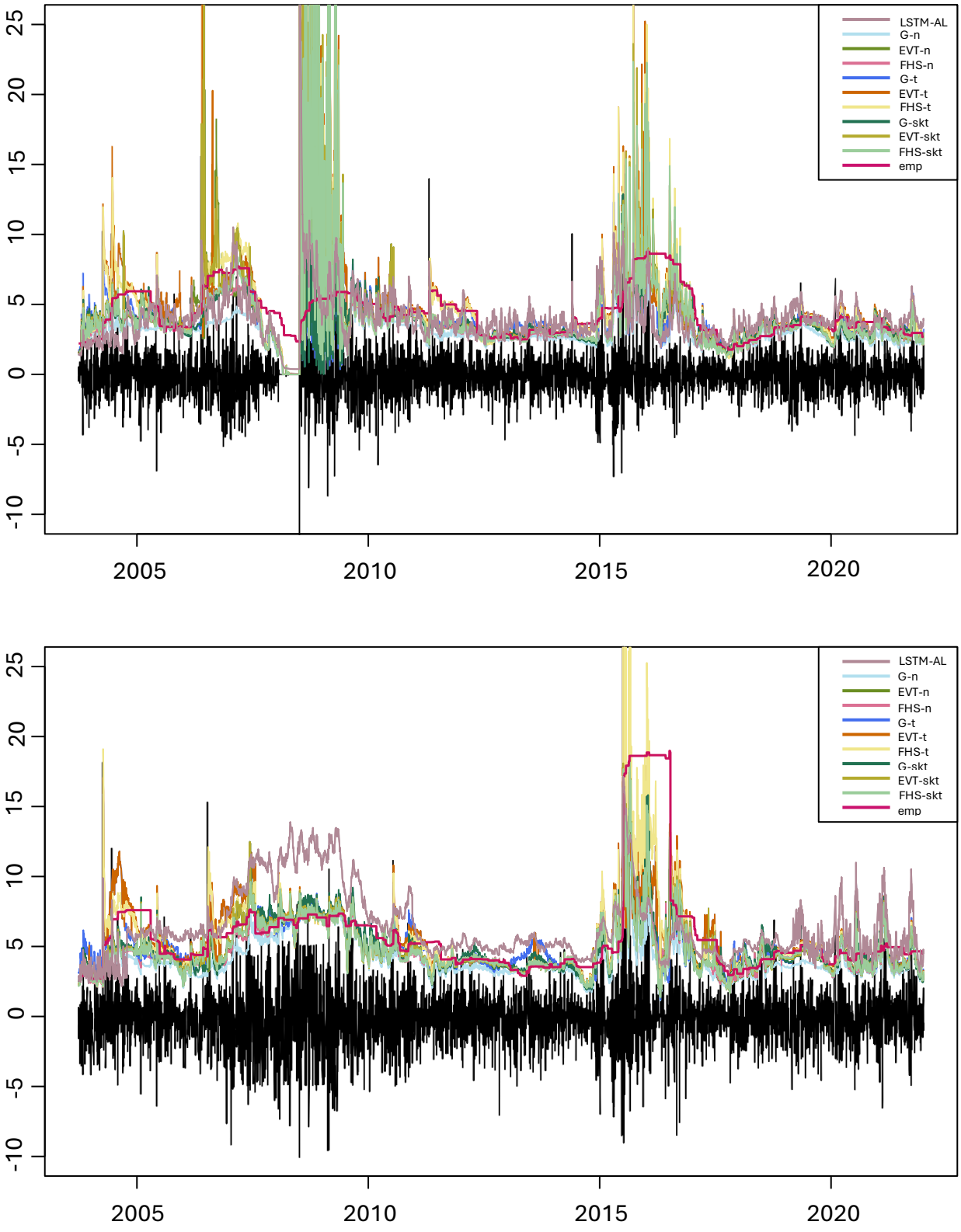

Figure 5.2: 97.5% ES forecasts (negative log-return %) for portfolios; top panel: min-ES; bottom: max-R

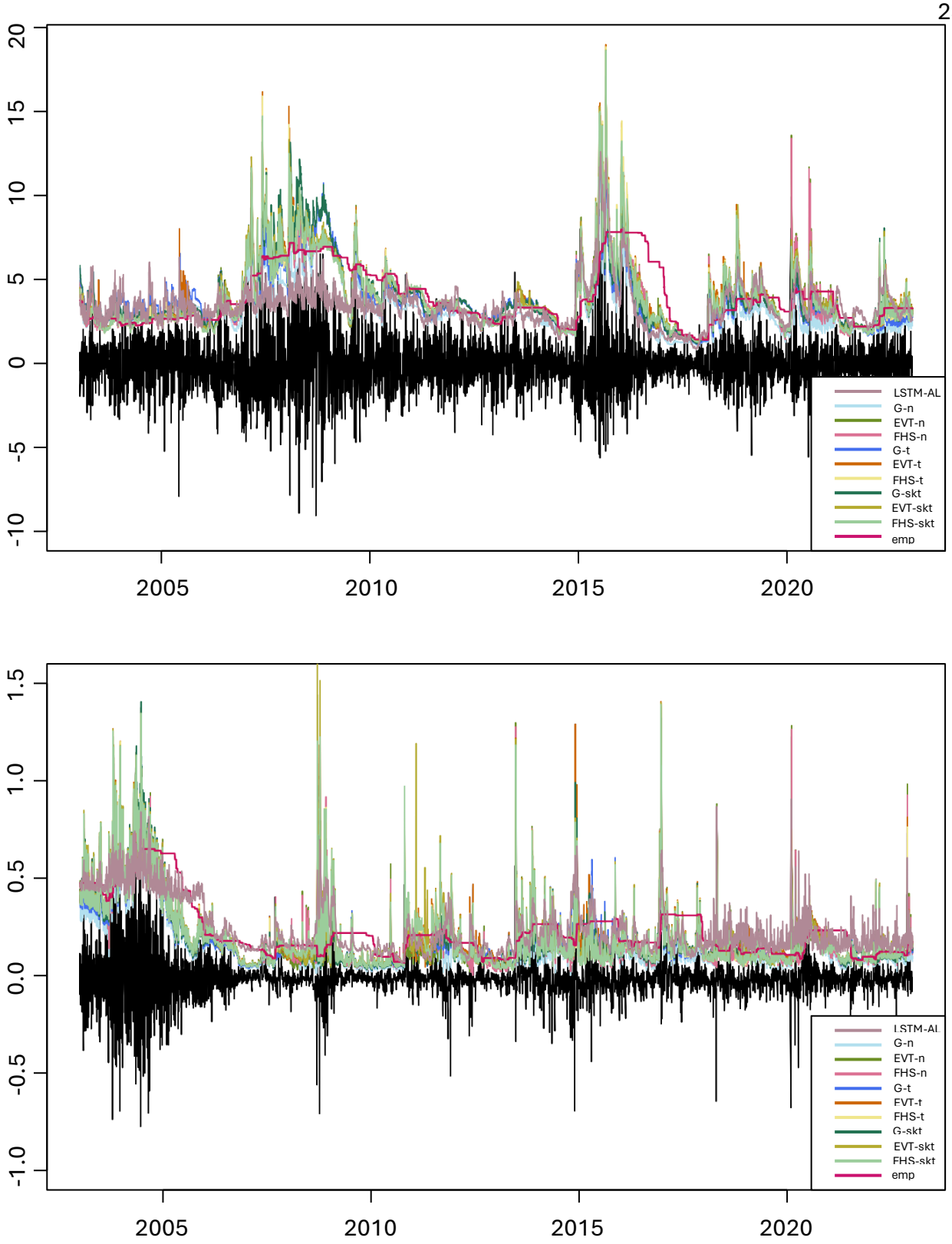

Figure 5.3: 97.5% ES forecasts (negative log-return %) for return series; top panel: SSE; bottom: CBA

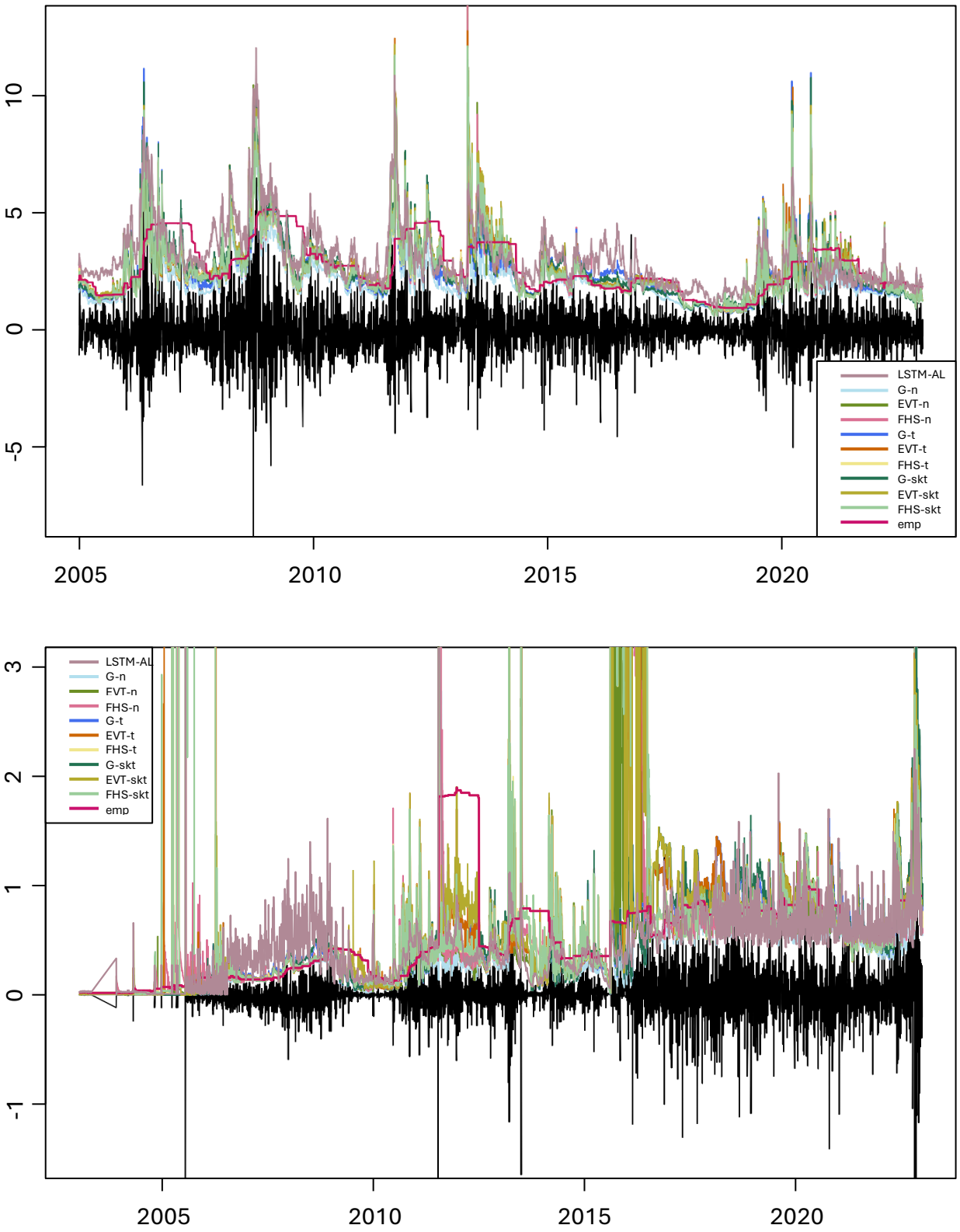

Figure 5.3: 97.5% ES (negative log-return %) return series; top panel: AU99.99; bottom: RMB/USD

6. E-backtesting ES

6.1. E-backtesting

Backtesting risk measures refers to hypothesis tests that gauge accuracy and effectiveness of risk forecast using realized losses. The transformation from 99% VaR to 97% ES is accompanied by a number of follow-up amendments on the model approval process and backtesting requirements. The current proposal in Basel IV is to backtest VaR forecasts at the 97.5th and 99th quantile on a 12-month basis instead of backtesting ES, mainly due to the nonelicitability of ES. Undoubtedly, a well-established backtesting framework for ES will be a major topic in further revisions. Wang et al. (2023) leverage the superiority of e-values over conventional p-values to tailor a backtest approach for regulators, and they call their method e-backtesting.

Borrowing their idea, how e-backtesting could feature regulatory expectation is three-fold. First, according to the Basel standard introduced in Chapter 5, ES must be calculated on a daily basis with an estimation window of 250 observations, which raises the need for sequential monitoring and a hypothesis test that makes logical sense for a small sample size. E-backtesting employes a stopped (non-negative) supermartingale with a countably infinite index set as the test statistic, enabling the regulator to sequentially monitor ES without asymptotic approximation. Second, recent revisions to Basel III yield a more stringent approval process by supervisory authority for the internal model proposed by banks (see e.g., MAR30 of BCBS, 2022). In this regard, a portable backtesting approach that embraces inclusiveness to different models will be more favored. E-backtesting is model-free in the sense that it is defined for all probability measures on an arbitrary measurable space, enabling regulators to scrutinize models without knowing nuts and bolts. Third, regulators target prudence rather than accuracy. By construction, the test statistic for e-backtesting signals early alerts on risk estimation with an accumulated flow of information and evidence (Wang and Ramdas, 2022), endowing the prudence with objectiveness and comprehensiveness.

To keep terminology simplified, I leave out some concepts in Wang et al. (2023) and focus on the computation. Let $(\Omega, \mathcal{F})$ be a measurable space. A hypothesis H is a subset of probability measures on $(\Omega, \mathcal{F})$. A non-negative stochastic process $(E_t)_{t\in K}$, $K \subseteq \mathbb{N} \cup \{\infty\}$, adapted to a given filtration, is an e-process for H if $\mathbb{E}^{\mathbb{P}}[E_\tau] \leq 1$ for all stopping times τ taking values in K and each $\mathbb{P} \in H$. An e-test rejects the Null H if a realized E_τ, called an e-value (Vovk and Wang, 2021), is larger than a threshold. Let L_s be the realized loss at time s, $\widehat{ES}_s$ be the ES forecast at time s, and $\widehat{VaR}_s$ be the VaR forecast at time s. The e-process used for e-backtesting ES at confidence level p is defined as

$$\begin{cases} M_0 = 1 \\ M_t(\boldsymbol{\lambda}) = \prod_{s=1}^{t} \left(1 - \lambda_s + \lambda_s \, e_p^{ES}\left(L_s, \widehat{ES}_s, \widehat{VaR}_s\right)\right), \ t \in \mathbb{N} \cup \{\infty\} \end{cases} \tag{6.1}$$

where $e_p^{ES}(\,)$ is called a model-free e-statistic for (ES_p, VaR_p) testing ES_p, defined as

$$e_p^{ES}\left(L_s, \widehat{ES}_s, \widehat{VaR}_s\right) = \frac{(L_s - \widehat{VaR}_s)^+}{(1-p)(\widehat{ES}_s - \widehat{VaR}_s)},$$

and $\boldsymbol{\lambda}$ is a so-called betting process optimized by maximizing the expected log-capital with three methods: GREE, GREL, GREM (see Section 5 in Wang et al., 2023 for differences). According to Taylor expansion, the optimal betting process using GREE and GREL can be approximated as

$$\lambda_t^{GREE} \approx 0 \vee \frac{\sum_{s=1}^{t-1} e_p^{ES}(L_s, \widehat{ES}_s, \widehat{VaR}_s) - (t-1)}{\sum_{s=1}^{t-1}\left(e_p^{ES}(L_s, \widehat{ES}_s, \widehat{VaR}_s) - 1\right)^2} \wedge 0.5, \tag{6.2}$$

$$\lambda_t^{GREL} \approx 0 \vee \frac{\sum_{s=1}^{t-1} e_p^{ES}(L_s, \widehat{ES}_t, \widehat{VaR}_t) - (t-1)}{\sum_{s=1}^{t-1}\left(e_p^{ES}(L_s, \widehat{ES}_t, \widehat{VaR}_t) - 1\right)^2} \wedge 0.5. \tag{6.3}$$

The GREM method is a mixture of the GREE and GREL methods, for which the e-process in (6.1) is calculated as

$$M_t(\boldsymbol{\lambda}^{GREM}) = \frac{M_t(\boldsymbol{\lambda}^{GREE})}{2} + \frac{M_t(\boldsymbol{\lambda}^{GREL})}{2}. \tag{6.4}$$

Evidence against the forecast (i.e., the forecast is not conservative), called a detection, is reported when the e-process exceeds thresholds 2, 5, or 10. Detection with size 2 signals early warning that risk predictions might not be prudent enough. According to the rule of thumb in Jeffreys (1961), 5 refers to "substantial", and 10 refers to "decisive".

6.2. Who wins?

The GREM method is recommended as a default choice in Wang et al. (2023) due to its stability in empirical studies. Hence, I use the e-process in (6.4) to backtest ES forecasts produced by all the methods discussed in Chapter 5. The empirical mean in (6.2) and (6.3) is calculated using a moving window of data in the past 250 days, and I kick off the backtest at the 251st observation of all return series (negated log-returns). Specifically, the start of backtesting date for each series is as follows: Oct 15, 2004 for portfolios (sample size for backtesting is 4188); Feb 16, 2004 for SSE (sample size for backtesting is 4594); Jan 13, 2006 for AU99.99 (sample size for backtesting is 4125); Jan 5, 2004 for CBA (sample size for backtesting is 4754); Aug 18, 2004 for RMB/USD (sample size for backtesting is 4782); In addition, for the sake of prudence, I focus on early detection (i.e., the case when an e-process goes beyond 2; hereinafter referred to as a "detection").

Backtesting results are summarized in Figure 6.1, Figure 6.2, Table 6.1, and Table 6.2. Overall, the LSTM-AL gives the most prudent forecasts with no detection over a nine-year horizon for most of the series, though such superiority is impaired by MV and SSE. The AR(1)-GARCH(1,1) model gives the most lenient forecasts for naïve 1/N and SSE, which, as indicated in Appendix B.1, have the least kurtosis than their counterparts. In general, GARCH-EVT with t or skewed-t innovations helps pivot the forecast to better conservativeness. GARCH-FHS does not show much advantage over its empirical peer, and consistent with the discussion in Chapter 5, it exhibits a dramatic action for the foreign exchange data, where the whole GARCH family suffers. Moreover,

comparing the magnitude of ES forecasts across all methods, max-R seems to involve the highest underlying risk, which signals the potential pitfall of imposing risk constraint using empirical estimate in portfolio optimization.

I start the backtest only 2 years before the global financial crisis. Detection over a stressed scenario is penalized in risk modeling yet may not be representative in decision making. Therefore, I re-backtest three methods at every 1,000 days to give a more general view: LSTM-AL model (for SSE and MV); GARCH-EVT with t innovation (for SSE); GARCH with skewed-t innovation (for RMB/USD). Figure 6.3 charts new evidence. As in the top panels, in the aftermath of the 2007–2008 crisis, the LSTM-AL model resumes its superiority with no detection even during the 2015 market crash. Interestingly, detections for the GARCH-EVT (t) method and the GARCH (skt) model, as shown in the bottom panels, exhibit striking consistency with regime shifts. For GARCH-EVT (t), when the backtest starts from Feb 16, 2004, detection occurs between the split-share structure reform and the global financial crisis; if the backtest is deferred for 1,000 days, detection occurs shortly after the launch of the most stringent control policy in China's real estate history: New National Ten Measures in 2010; if the test is deferred for 2,000 or more days, detections occur during the phased lockdown in 2022.

The phenomenon is even stronger for foreign exchange market, where detections for the GARCH (skt) model coincide with the four major regimes of China's exchange rate market: first, the move-away announced in 2005 from a fixed exchange rate to a floating one; second, the inclusion of RMB to IMF's Supplementary International Reserve (SDR) basket in 2015, followed by an immediate announcement on the RMB/USD central parity quoting mechanism; third, the volatility spillover of RMB during the COVID-19 crisis; fourth, the depreciation of RMB due to the rise of the US dollar index driven by the Fed interest rate hike and the increase in demand for the US dollar due to uncertainties over the Russian-Ukraine war in 2022. Such evidence points to the conclusion: whether an ES model is prudent in the regulatory sense hinges on how well it survives during regime shifts.

With no doubt, the LSTM-AL model wins the game. However, the win pays for efficiency: the model undergoes 90,000 iterations (i.e., six chains, each of 15,000 runs, as illustrated in Chapter 5) on every estimation window. More importantly, the win is not robust to all scenarios (e.g., observed detections for SSE and MV), suggesting its instable resilience during structural breaks.

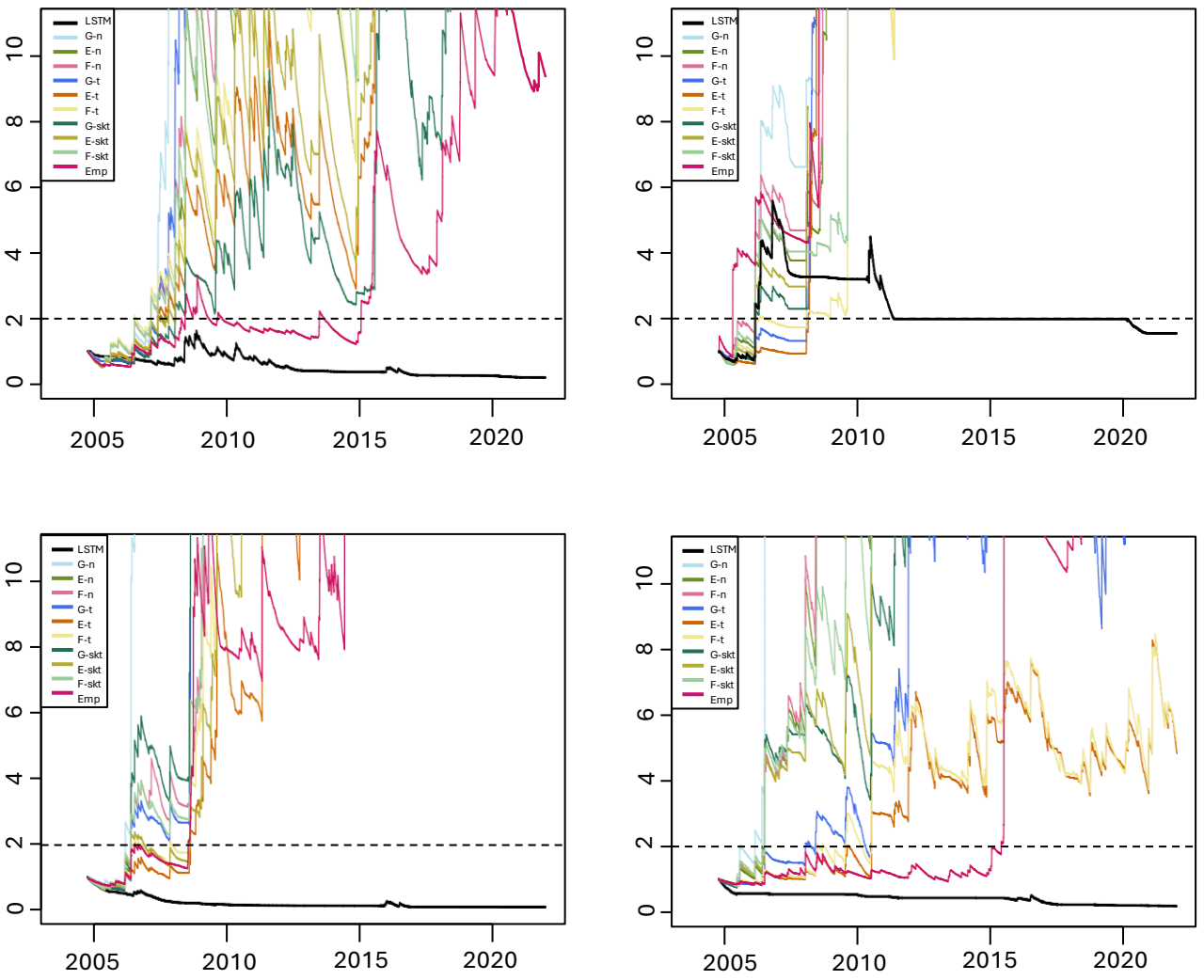

Figure 6.1: The realized e-processes for portfolios; top left: naïve 1/N; top right: MV; bottom left: min-ES; bottom right: max-R; y-axis refers to the realized e-values; dotted line refers to the detection at size 2; the backtesting horizon ends on Dec 31, 2021

Naïve 1/N			MV			Min-ES			Max-R		
	days	avg ES%		days	avg ES%		days	avg ES%		days	avg ES%
G-n	573	3.35	G-n	333	3.63	G-n	337	3.14	G-n	207	4.15
E-n	637	3.74	E-n	333	3.96	E-n	397	3.84	E-n	423	4.69
F-n	573	3.72	F-n	333	3.93	F-n	388	3.69	F-n	423	4.64
G-t	634	3.69	G-t	804	4.59	G-t	388	4.19	G-t	796	5.43
E-t	634	3.80	E-t	815	8.76	E-t	936	5.24	E-t	1179	5.64
F-t	423	3.77	F-t	333	6.25	F-t	388	5.05	F-t	900	5.60
G-skt	800	3.86	G-skt	333	4.18	G-skt	388	3.89	G-skt	397	4.91
E-skt	637	3.80	E-skt	333	5.15	E-skt	388	4.44	E-skt	423	4.81
F-skt	573	3.77	F-skt	333	5.84	F-skt	388	4.21	F-skt	423	4.72
LSTM	**NA**	3.81	LSTM	333	4.88	LSTM	**NA**	4.03	LSTM	**NA**	6.39
Emp	849	4.13	Emp	131	4.75	Emp	423	4.40	Emp	2501	5.81

Table 6.1: The number of days taken to observe a detection at size 2 for portfolios; avg ES% refers to the average ES forecasts (in negated log-returns); NA refers to no detection during the backtesting horizon; G shortcuts GARCH, E for EVT, F for FHS, Emp for empirical estimation; day 1 refers to M_0

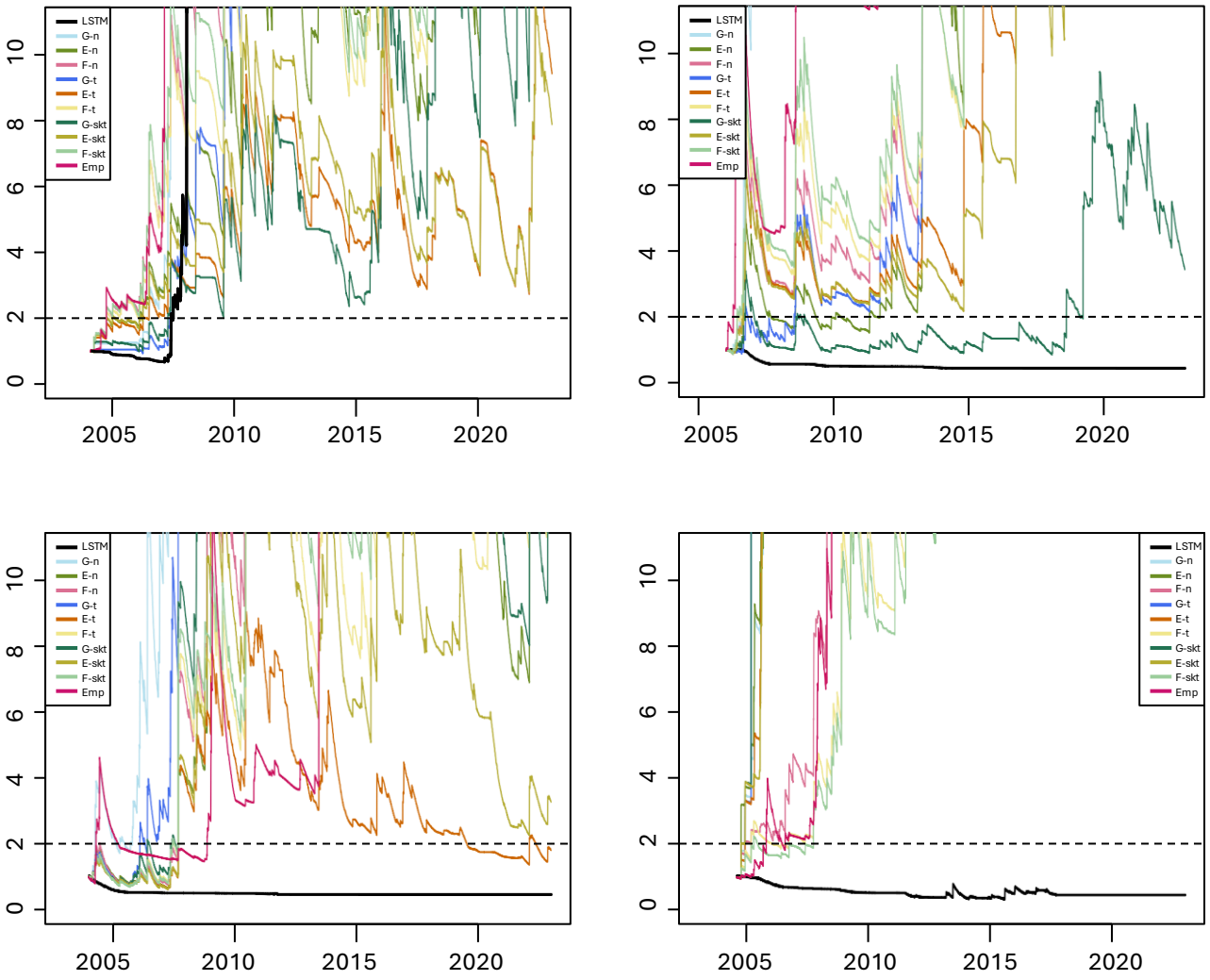

Figure 6.2: : The realized e-processes for market return series; top left: SSE; top right: AU99.99; bottom left: CBA; bottom right: RMB/USD; y-axis refers to the realized e-values; dotted line refers to the detection at size 2; the test horizon ends on Dec 31, 2022

SSE			AU99.99			CBA			RMB/USD		
	days	avg ES%		days	avg ES%		days	avg ES%		days	avg ES%
G-n	587	3.25	G-n	144	2.18	G-n	72	0.16	G-n	92	0.37
E-n	164	3.90	E-n	159	2.60	E-n	929	0.19	E-n	51	0.64
F-n	164	3.82	F-n	159	2.53	F-n	115	0.19	F-n	92	0.89
G-t	798	3.87	G-t	175	2.58	G-t	529	0.20	G-t	92	0.50
E-t	527	4.00	E-t	159	2.66	E-t	929	0.20	E-t	92	0.64
F-t	164	3.88	F-t	97	2.55	F-t	868	0.19	F-t	157	3328.34
G-skt	774	3.97	G-skt	160	2.66	G-skt	607	0.19	G-skt	51	0.52
E-skt	164	3.97	E-skt	159	2.64	E-skt	929	0.20	E-skt	51	0.72
F-skt	164	3.86	F-skt	97	2.55	F-skt	868	0.19	F-skt	186	1597.27
LSTM	801	3.03	LSTM	**NA**	3.12	LSTM	**NA**	0.24	LSTM	**NA**	0.45
Emp	164	3.94	Emp	64	2.70	Emp	81	0.23	Emp	317	0.53

Table 6.2: The number of days taken to observe a detection at size 2 for market return series; avg ES% refers to the average ES forecasts (in negated log-returns); NA refers to no detection during the backtesting horizon; G shortcuts GARCH, E for EVT, F for FHS, Emp for empirical estimation; day 1 refers to M_0

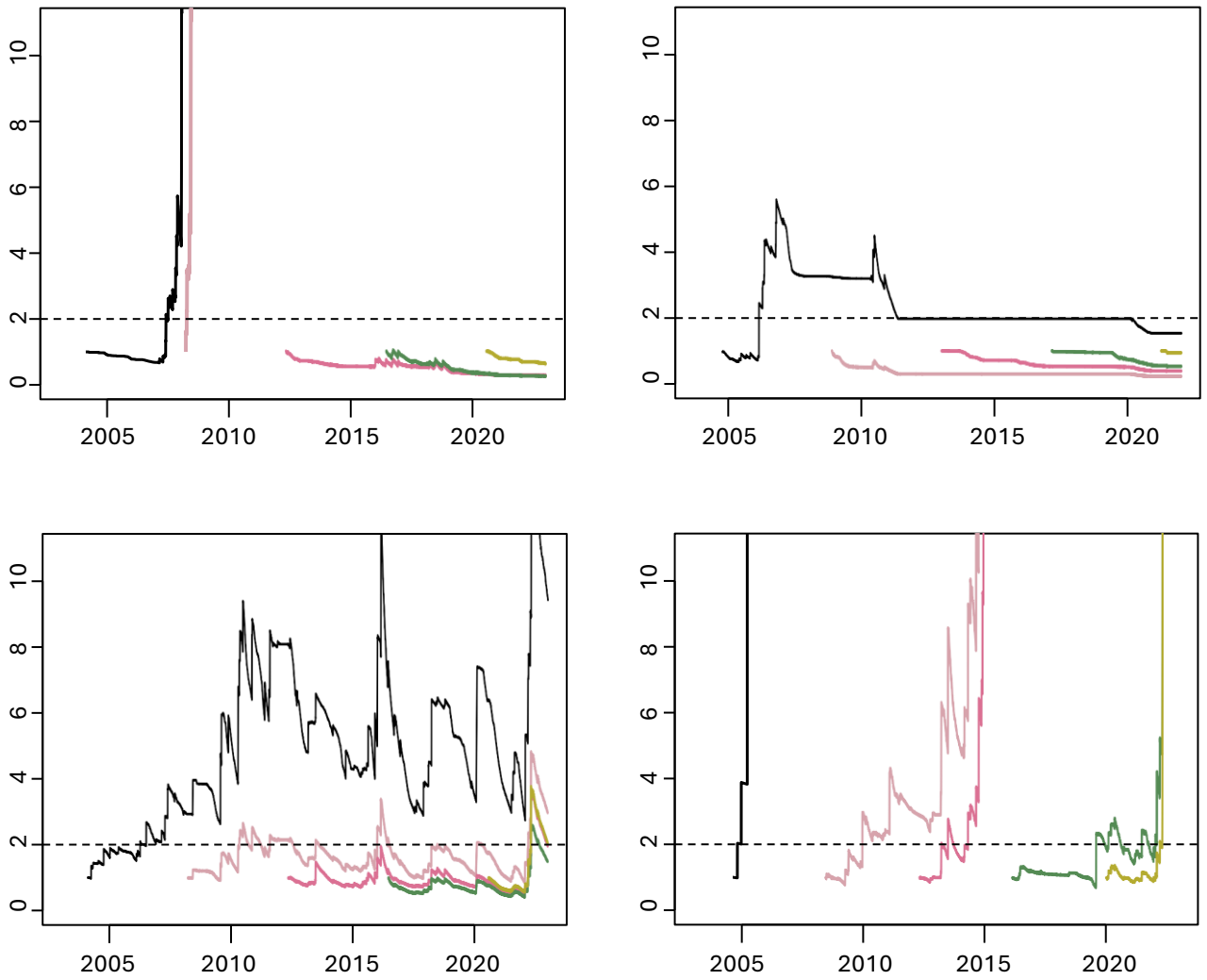

Figure 6.3: Backtesting three methods every 1,000 days; top left: backtesting LSTM-AL on SSE; top right: backtesting LSTM-AL on MV; bottom left: backtesting GARCH-EVT (t) on SSE; bottom right: backtesting GARCH-skt on RMB/USD; Black line refers the earlies test, light-pink line refers to the test starting 1,000 days later than the black, dark-pink line refers to that starting 2,000 days later, and so on;

Notes: for GARCH-EVT (t) on SSE (i.e., the bottom left), the detection date for each test (i.e., the first date the process goes beyond 2) is: 2006-04-13 for the first test, 2010-05-06 for the second test, and 2022-04-25 for all other tests; for GARCH-skt on RMB/USD, the detection date is: 2004-10-27 for the first, 2009-12-24 for the second; 2013-03-18 for the third, 2019-08-05 for the fourth, and 2022-03-15 for the last test

7. Concluding Remark: what defines a "good" model for Chinese regulators?

In summary, numerical, and graphical evidence from my backtesting in Chapter 6 light a key metric in defining a "good" forecasting model of ES for Chinese regulators: the model must demonstrate strong, stable, and initiative adaptability to regime shifts. As pointed out in much of the literature, active government interventions through all the four major markets: equity, fixed income, foreign exchange, and commodity has characterized the Chinese market, leading to more frequent and subtle structural breaks in financial time series.

What does that suggest? The so-called regime shift in financial time series is driven by events (e.g., reform, crisis, etc.). As a common sense, "forward in time" is the direction in which events occur, and in which we collect information. This makes it non-trivial to adapt a model to regime shift in an automated fashion as we already convince ourselves that future is unpredictable. What if we leverage some idea from theoretical physics? Informally, let us suppose that an event is nothing but a collection of some unknow but fixed "states", and the occurrence of the event is just a manifestation of the interaction between those states with reference to time. With this idea, we can try to find an exogenous stochastic process that describes the interaction of those "states" and incorporate it into the LSTM-AL model in some way such that the model configuration, which includes, but are not limited to, model parameters, initialization of algorithm, and the LSTM architecture, can adapt to regime shifts consistently. Admittedly, it might be ambitious to find a reasonable process that concretes over such imagination; given the extensive success of Markov-switching models in the literature of modeling structural change, a Markov process may be a good start to try.

References

C. Andrieu, N. D. Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

A. A. Balkema and L. D. Haan. Residual life time at great age, *The Annals of Probability*, 2:792–804, 1974.

G. Barone-Adesi, K. Giannopoulos, and L. Vosper. VaR without corrections for portfolios of derivative securities, *Journal of Futures Markets*, 19(5):583–602, 1999.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

M. Bee, D. J. Dupuis, and L. Trapin. Realizing the extremes: Estimation of tail-risk measures from a high-frequency perspective. *Journal of Empirical Finance*, 36:86–99, 2016.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

BCBS. MAR30-Internal models approach: general provisions. Bank for International Settlements, 2022.

BCBS. Minimum capital requirements for market risk. Bank for International Settlements, 2019.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Series B*, 39(1):1–38, 1977.

V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.

K. Echaust and M. Just. Value-at-Risk Estimation Using the GARCH-EVT approach with optimal tail selection. *Mathematics*, 8(1):1–24, 2020.

C. Fernandez and M. F. J. Steel. On Bayesian modelling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

J. Fan, L. Qi, and D. Xiu. Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *Journal of Business and Economic Statistics*, 32(2):178–191, 2014.

R. F. Engle and S. Manganelli. CAViaR: Conditional autoregressive Value-at-Risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381, 2004.

T. Fissler and J. F. Ziegel. Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4):1680–1707, 2016.

A. Gelman, G.O. Roberts, and W.R. Gilks (1996), Efficient Metropolis jumping. *Bayesian Statistics 5*, ed. J. Bernardo et al., 599–607. Oxford University Press.

F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. Learning precise timing with LSTM recurrent networks, *Journal of Machine Learning Research*. 3:115–143, 2003.

J. Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constrains. *Computer Sciences and Statistics Proceedings*, 23:571–578, 1991.

R. H. Gerlach, C. W. Chen, and N. Y. Chan. Bayesian time-varying quantile forecasting for value-at-risk in financial markets. *Journal of Business & Economic Statistics*, 29(4):481–492, 2011.

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011.

Q. He, J. Gan, S. Wang, and T. T. Chong. The effects of trading suspensions in China. *The North American Journal of Economics and Finance*, 50(C): 100985, 2019.

S. Hochreiter, F. F. Informatik, and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

T. Hull and A. White. Incorporating volatility updating into the historical simulation method for value-at-risk. *Journal of Risk*, 1(1):5–19, 1998.

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, third edition, 1961.

M. Karmakar and G. K. Shukla. Managing extreme risk in some major stock markets: An extreme value approach. *International Review of Economics & Finance*, 35:1–25, 2015.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

A. Lo. Long-term memory in stock market prices. *Econometrica*, 59(5):1279–313, 1991.

J. Liu, R. F. Stambaugh, and Y. Yuan. Size and value in China. *Journal of Financial Economics*, 134(1):48–69, 2019.

L. Liao, B. Liu, and H. Wang. China's secondary privatization: Perspectives from the Split-Share Structure Reform. *Journal of Financial Economics*, 113(3):500–518, 2014.

Y. Li and S. K. Ghosh. Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints. Journal of Statistics Theory and Practice, 9:712–732, 2015.

Z. Li., M. N. Tran, C. Wang, R. H. Gerlach, J. Gao. A Bayesian long-short term memory model for Value-at-risk and expected shortfall joint forecasting. *Preprint*, *arXiv:2001.08374v2*, 2021.

A. J. McNeil and R. Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, 7(3–4):271–300, 2000.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, Equations of state calculations by fast computing machines. *Journal of Chemical Physics*. 21:1087–1091, 1953.

N. Nolde and J. F. Ziegel. Elicitability and backtesting: Perspectives for banking regulation (with discussion). *Annals of Applied Statistics*, 11(4):1833–1874, 2017.

A. Patton, J. F. Ziegel, and R. Chen. Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *Journal of Econometrics*, 211(2):388–413, 2019.

J. Pickands. Statistical inference using extreme order statistics, *The Annals of Statistics*, 3:119–131, 1975.

G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annuals of Applied Probability*, 7(1):110–120, 1997.

R. T. Rockafellar and S. Uryasev. Conditional Value-at-Risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

J. W. Taylor. Forecast Combinations for Value-at-Risk and expected shortfall. *International Journal of Forecasting*, 36(2):428–441, 2020.

J. W. Taylor. Forecasting Value-at-Risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics*, 37(1):121–133, 2019.

V. Vovk and R. Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

D. Wang, J. Ding, G. Chu, D. Xu, and T. S. Wirjanto. Modelling asset returns in the presence of price limits with Markov-switching mixture of truncated normal GARCH distribution: Evidence from China. *Applied Economics*, 53(7):781–804, 2021.S. M. S. Seyfi, A. Sharifi, and H. Arian. Portfolio Value-at-Risk and expected shortfall using an efficient simulation approach based on Gaussian Mixture Model. *Mathematics and Computers in Simulation*, 190:1056–1079, 2021.

Q. Wang, R. Wang, and J. F. Ziegel. E-backtesting. *Preprint, arXiv: 2209.00991v3*, 2023.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society*, *Series B*, 84(3):822–852, 2022.

Appendix A

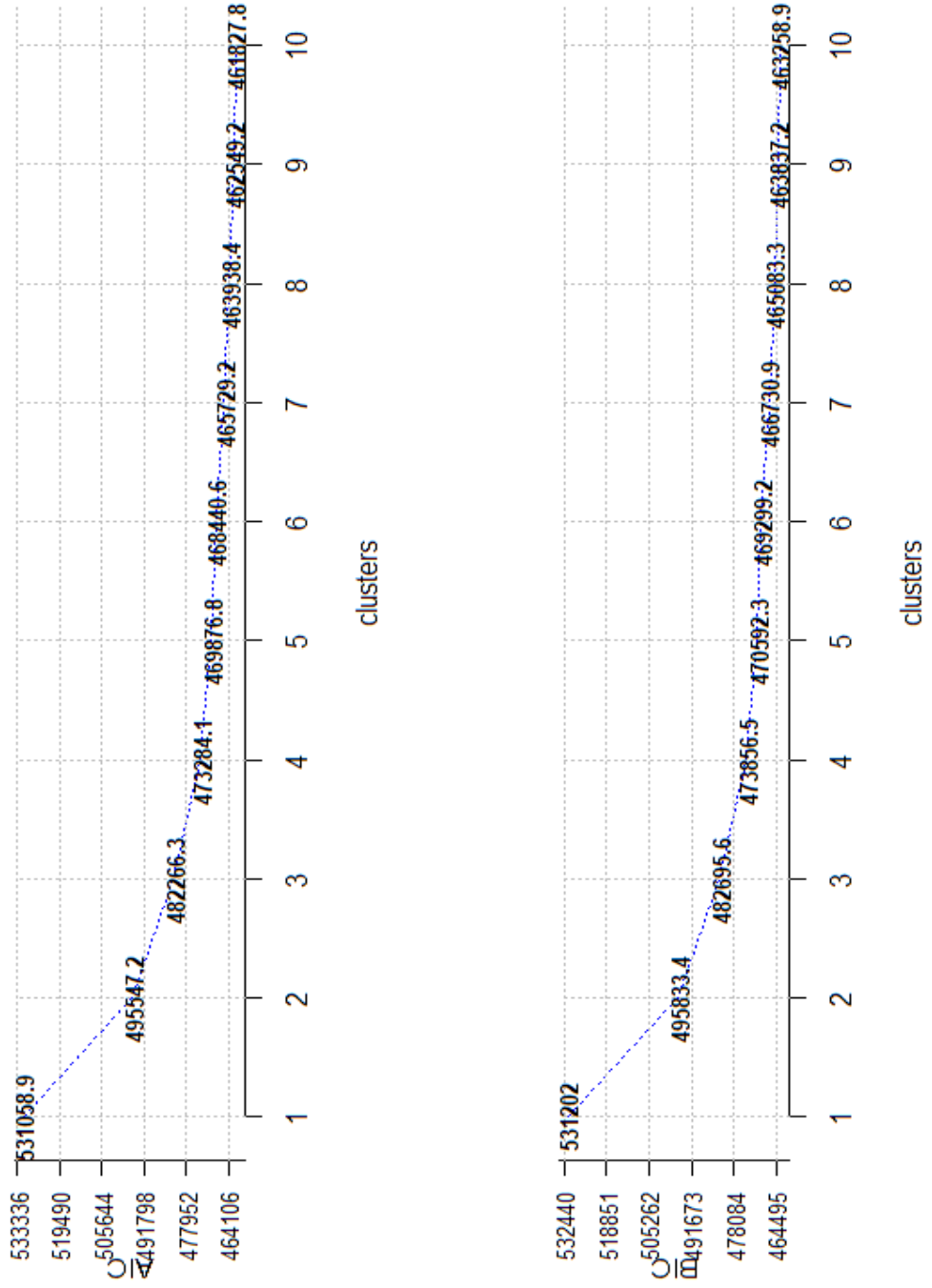

Figure A.1: The AIC and BIC score for the Gmm model with respect to different number of clusters; apparently, the derivative of the curve becomes gentle at 4, which suggests the sufficiency of setting four clusters

Appendix B.1

Ticker	GICS Level I Industry	Max	Min	Mean	History of suspensions and resumptions
600519	Consumer Staples	0.0965	-0.8368	8.21e-04	2 times; 52 days in total
000858	Consumer Staples	0.0955	-0.7035	3.62e-04	5 times; 105 days in total
000725	Information Technology	0.2462	-0.4091	-2.55e-04	5 times; 56 days in total
600703	Information Technology	0.9411	-0.7783	2.27e-04	6 times; 320 days in total
600886	Utilities	0.0957	-0.6986	5.88e-05	8 times; 219 days in total
600795	Utilities	0.0965	-0.8053	-3.60e-04	6 times; 123 days in total
600276	Health Care	0.0955	-0.4336	2.12e-04	2 times; 26 days in total
000538	Health Care	0.0955	-0.5154	3.12e-04	6 times; 202 days in total
600309	Materials	0.0956	-0.5442	2.83e-04	5 times; 169 days in total
600019	Materials	0.0962	-0.1154	1.10e-04	7 times; 126 days in total
000651	Customer Discretionary	0.0958	-0.8508	2.61e-04	5 times; 172 days in total
600104	Customer Discretionary	0.0959	-0.4026	1.93e-04	5 times; 77 days in total
600760	Industrials	0.0965	-0.3549	4.27e-04	9 times; 198 days in total
600150	Industrials	0.0957	-0.4659	1.74e-04	6 times; 160 days in total
000002	Real Estate	0.0958	-0.7099	6.18e-05	5 times; 176 days in total
600606	Real Estate	0.0962	-0.3623	-2.52e-04	10 times; 373 days in total
600028	Energy	0.0967	-0.3052	1.34e-05	3 times; 29 days in total
600346	Energy	0.0961	-0.9393	5.71e-05	12 times; 334 days in total
600745	Communication Services	0.2597	-0.1070	4.46e-04	16 times;820 days in total
000063	Communication Services	0.0956	-0.4275	4.23e-05	7 times; 107 days in total
000001	Financials	0.0959	-0.5429	5.15e-05	10 times; 114 days in total
600000	Financials	0.0956	-0.4019	-1.16e-04	8 times; 90 days in total

Table B.1: Information on the 22 stocks (in log-returns); History of suspensions and resumptions indicates how many times the stock has experienced suspensions and the total length

	Start date	End date	# of observations	Kurtosis	Skewness
SSE	2002-01-07	2022-12-31	5094	4.9710	-0.4456
AU99.99	2004-01-05	2022-12-30	4625	8.1417	-0.4010
CBA	2002-01-07	2022-12-31	5254	10.7104	0.0189
RMB/USD	2002-01-04	2022-12-30	5282	604.5195	-1.3550
Naïve 1/N	2002-09-13	2021-12-31	4688	4.4148	-0.3210
MV	2002-09-13	2021-12-31	4688	168.0006	4.9923
Min-ES	2002-09-13	2021-12-31	4688	970.0177	20.5460
Max-R	2002-09-13	2021-12-31	4688	323.1376	-9.4391

Table B.2: Descriptive statistics of the eight series; all the data are in log-returns

Appendix B.2

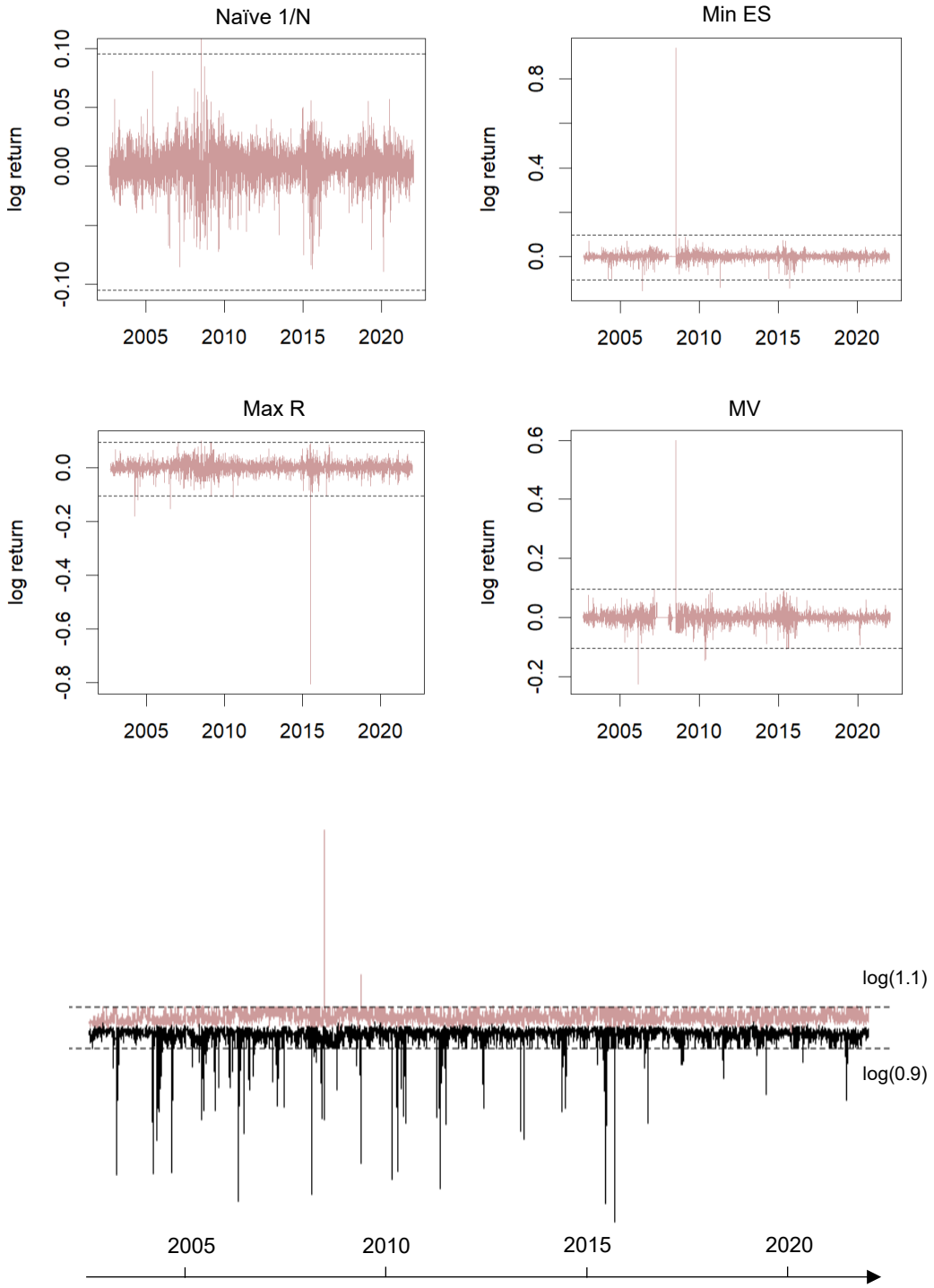

Figure B.1: The profile of portfolios; the top four panels chart the realized returns (log-returns); dotted line refers to log (1.1) (top) and log(0.9) (bottom); the bottom panel plots the largest individual loss (in black) and the largest individual gain (in pink) of my 22 stocks on each day. Obviously, the losses are not bounded below by log(0.9), mainly due to the price drop after stock splits

Appendix B.3

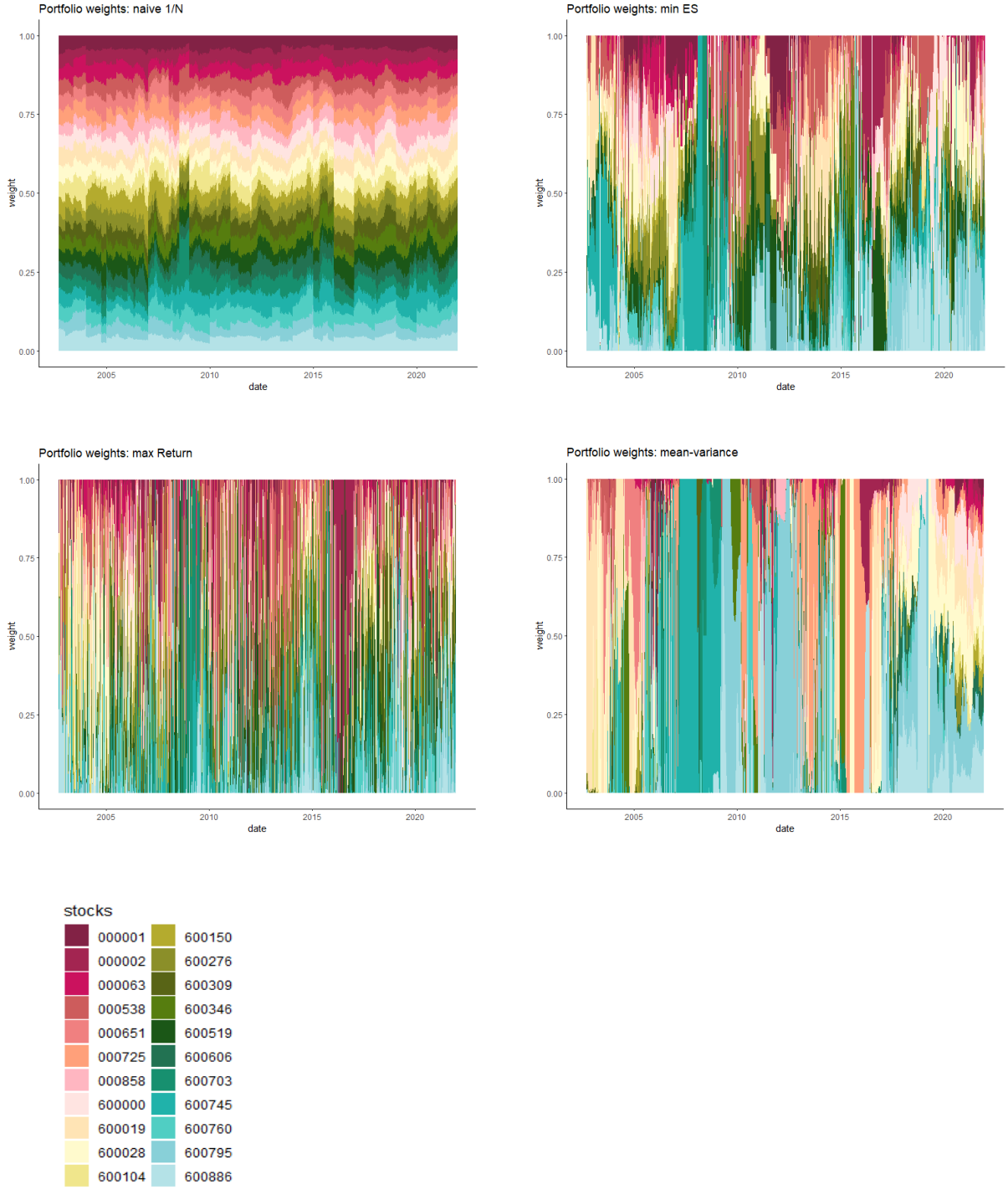

Figure B.2: visualization on how the weights in each portfolio update every day; each stock is assigned a color, with the length (vertical) of the colored region representing the size of the weight.

Appendix C

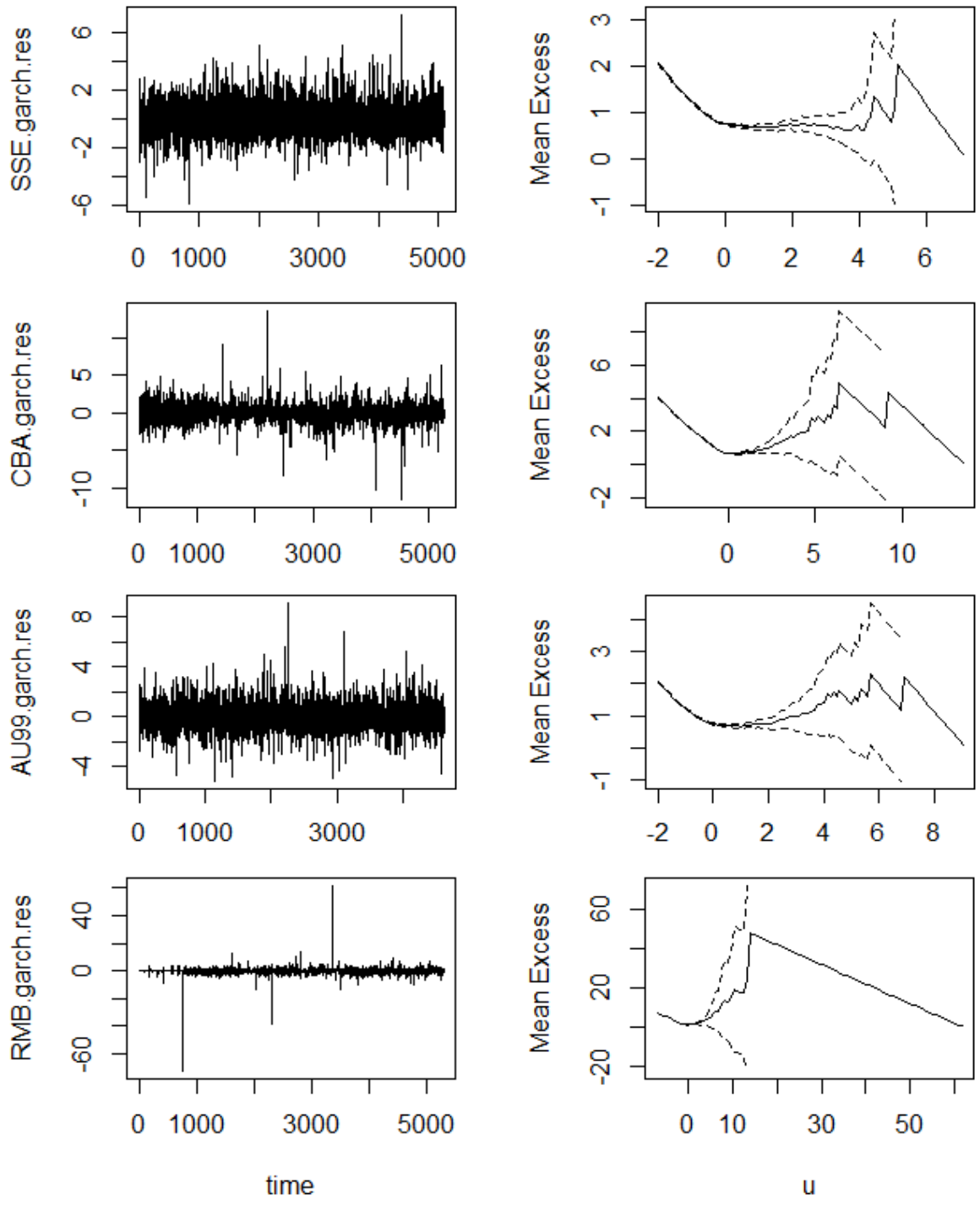

Figure C.1: The graphical tool to help determine the threshold in GARCH-EVT for market return series; left panel refers to the standardized residuals of fitting a GARCH-n model to all the series without rolling window; right panel refers to the mean excess plot; dotted lines refer to 95% confidence intervals for each mean excess value; from top to bottom: SSE, CBA, AU99.99, and RMB/USD; the threshold is selected as follows: $u = 9$ for RMB/USD and $u = 1$ for all others

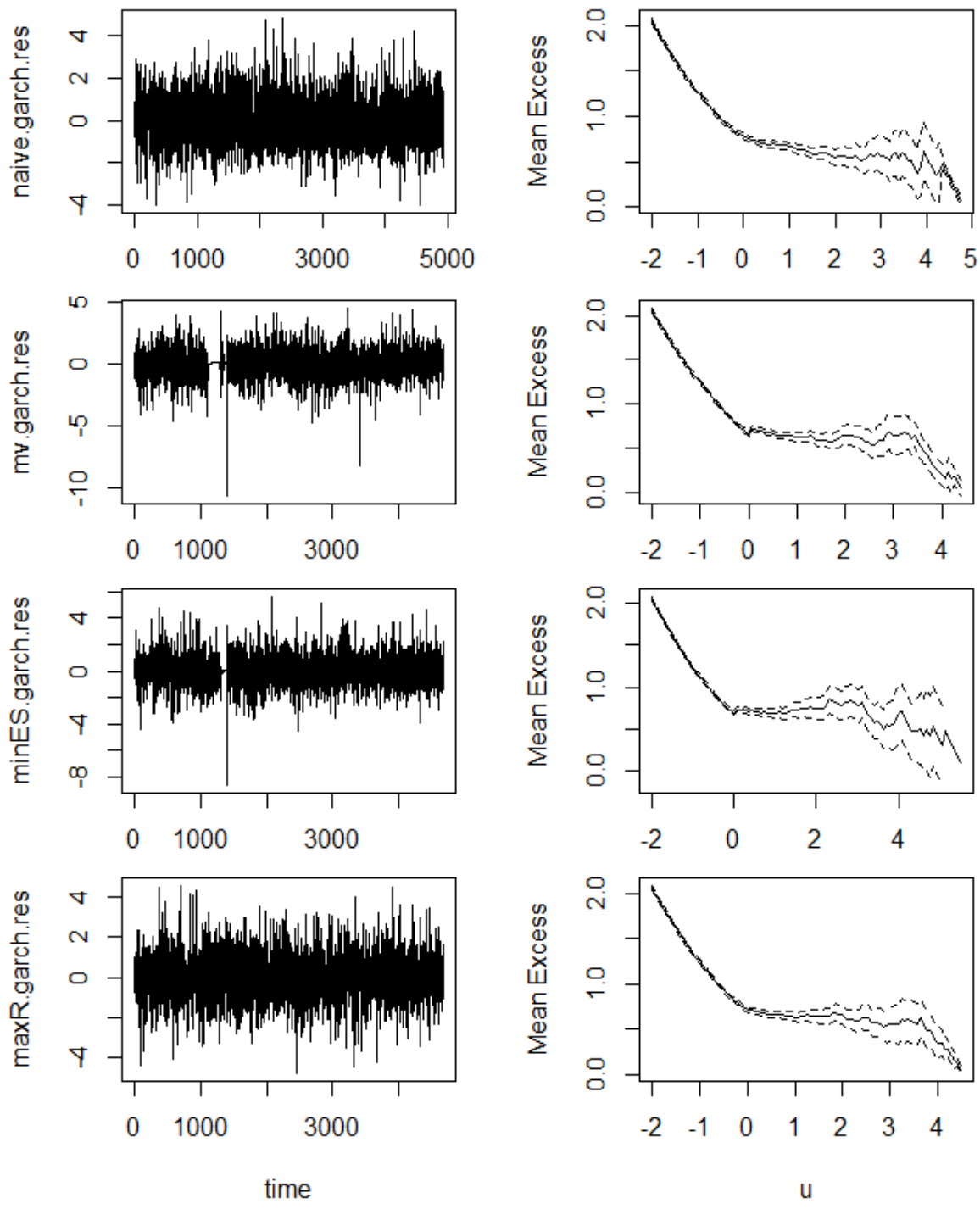

Figure C.2: The graphical tool to help determine the threshold in GARCH-EVT for portfolio return series; left panel refers to the standardized residuals of fitting a GARCH-n model to all the series without rolling window; right panel refers to the mean excess plot; dotted lines refer to 95% confidence intervals for each mean excess value; from top to bottom: naïve 1/N, MV, min-ES, and max-R; the threshold is selected as follows: $u = 1.5$ for MV; $u = 2$ for min-ES; $u = 1$ for all others

Appendix D.1

The adaptive Metropolis MCMC algorithm for LSTM-AL estimation

Input: initial parameter set $\boldsymbol{\theta_0}$;
realized return $\boldsymbol{r_1, \dots, r_T}$ as in-sample;
number of iterations $\boldsymbol{n}$ and probability level $\boldsymbol{\alpha}$ for VaR and ES;
Output: Posterior samples of the LSTM-AL parameters

For $\boldsymbol{n \leq 36}$ do
Sample $\boldsymbol{\theta^*}$ from the proposal density $\boldsymbol{N(\theta_n, (0.1)^2 I_{18}/18)}$
Estimate $(\boldsymbol{ES_t^*, VaR_t^*})$ for $\boldsymbol{t = 1, 2, \dots, T}$ with $\boldsymbol{\theta^*}$using (10) to (19)

Compute the ratio $\boldsymbol{p_r^* = \frac{\prod_{t=1}^{T} f_{AL}(r_t|ES_t^*, VaR_t^*)}{\prod_{t=1}^{T} f_{AL}(r_t|ES_t, VaR_t)} * \frac{\prod_{i=1}^{18} prior(\theta^*[i])}{\prod_{i=1}^{18} prior(\theta_n[i])}}$

Compute the acceptance probability $\boldsymbol{\alpha(\theta^*, \theta_n) = min\{1, p_r^*\}}$
Sample $\boldsymbol{u \sim U(0,1)}$
If $\boldsymbol{u < \alpha(\theta^*, \theta_n)}$ then Accept $\boldsymbol{\theta^*}$ and update $\boldsymbol{\theta_{n+1} = \theta^*}$
Else $\boldsymbol{\theta_{n+1} = \theta_n}$
End for

For $\boldsymbol{n > 36}$ do
Sample $\boldsymbol{\theta^*}$ from $\boldsymbol{0.95N(\theta_n, (2.38)^2 \Sigma_n/18) + 0.05N(\theta_n, (0.1)^2 I_{18}/18)}$
Estimate $(\boldsymbol{ES_t^*, VaR_t^*})$ for $\boldsymbol{t = 1, 2, \dots, T}$ with $\boldsymbol{\theta^*}$ following (9) to (18)

Compute the posterior ratio $\boldsymbol{p_r^* = \frac{\prod_{t=1}^{T} f_{AL}(r_t|ES_t^*, VaR_t^*)}{\prod_{t=1}^{T} f_{AL}(r_t|ES_t, VaR_t)} * \frac{\prod_{i=1}^{18} prior(\theta^*[i])}{\prod_{i=1}^{18} prior(\theta[i])}}$

Compute the acceptance probability $\boldsymbol{\alpha(\theta^*, \theta_n) = min\{1, p_r^*\}}$
Sample $\boldsymbol{u \sim U(0,1)}$
If $\boldsymbol{u < \alpha(\theta^*, \theta_n)}$ then
Accept $\boldsymbol{\theta^*}$ and update $\boldsymbol{\theta_{n+1} = \theta^*}$

$$\boldsymbol{m_{n+1} = \frac{n}{n+1} m_n + \frac{1}{n+1} \theta_n}, \text{ where } \boldsymbol{m_n} \text{ is the empirical mean}$$

$$\boldsymbol{\Sigma_{n+1} = \frac{n-1}{n} \Sigma_n + m_n{}^T m_n + \frac{1}{n} \theta_n{}^T \theta_n - \frac{n-1}{n} m_{n+1}{}^T m_{n+1}}$$

Else $\boldsymbol{\theta_{n+1} = \theta_n}$
End for

Assumed prior distributions for parameter set $\boldsymbol{\theta}$ of the LSTM-AL model							
Parameter	β_0	β_1	γ_0	γ_1	α_0	α_1	LSTM
Prior	Flat	Flat	Flat	Flat	$N(0, 0.1)$	$IG(2.5, 0.25)$	$N(0, 0.1)$

Note: The pseudo code is a revision for Algorithm 1 in Li et al., 2021, and their typos are corrected here; "Flat" prior in the bottom table can be interpreted as a uniform (0,1) distribution

Appendix D.2

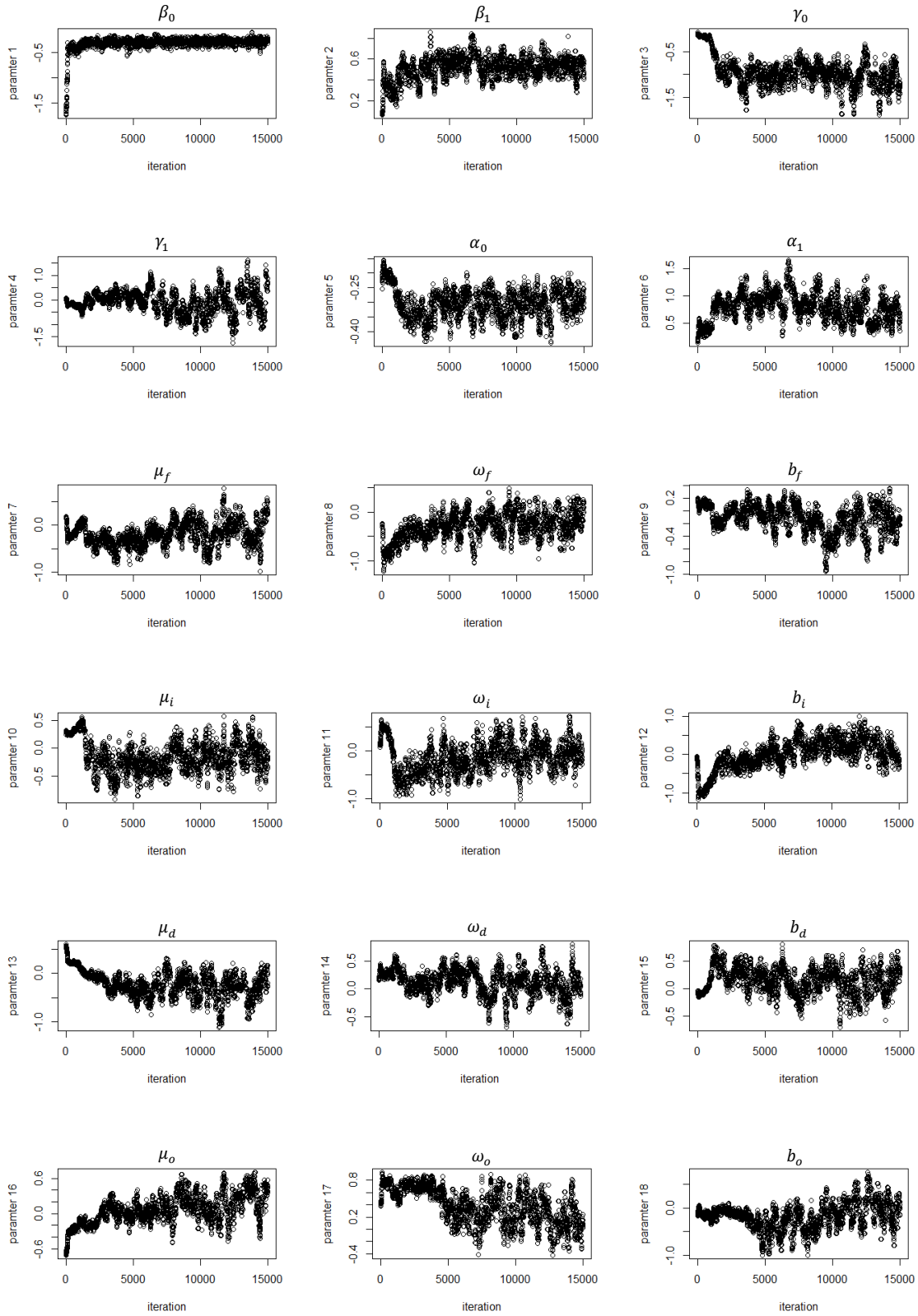

Figure D.1: The trace plot for CBA with a randomly chosen initial sample; all the 18 parameters become relatively stable within 15,000 runs, though such stability is subject to some fluctuations; for most parameters, the chain seems to step into a new phase after around 4,500 runs; Trace plots with different initial setting are compared and the plots for other series are examined; I get the similar observation; The initial samples are randomly drawn from the assumed prior distribution (as in the pseudo code on the previous page).