Electronic Companion for "Monitoring Value-at-Risk and Expected Shortfall Forecasts"

Appendix A: Proofs

Let $\|\cdot\|$ denote any norm in $\mathbb{R}^d$. Due to the equivalence of norms in $\mathbb{R}^d$, the particular choice does not matter.

Proof of Theorem 1. Recall the definition of the $\mathbb{R}^d$-valued $\boldsymbol{y}_t$ in Section 3.1. In this proof we exploit that under $\mathcal{H}_0^{\mathrm{VaR}}$ the sequence $\{\boldsymbol{y}_t\}$ is a bounded, d-dependent random vector. It is easy to verify that there is no autocorrelation in $\{\boldsymbol{y}_t\}$. Thus, the strong invariance principle in Lemma 4.1 of Horváth et al. (1999) holds for $\boldsymbol{y}_t$. Noting that $\boldsymbol{D}_{\boldsymbol{y}} := \mathrm{Cov}(\boldsymbol{y}_1) = \boldsymbol{I}_{d\times d}$ is the $(d\times d)$-identity matrix, the aforementioned lemma implies

$$\left\|\sum_{i=1}^{k}\boldsymbol{y}_i - \boldsymbol{W}(k)\right\| = o(k^{1/2-\kappa}) \qquad \text{almost surely (a.s.)}$$

for $0<\kappa<1/[24(2+d)]$ and $\boldsymbol{W}(\cdot)$ a standard Brownian motion in $\mathbb{R}^d$. This in turn implies

$$\sup_{k\in\mathbb{N}}\frac{1}{k^{1/\nu}}\left\|\sum_{i=1}^{k}\boldsymbol{y}_i - \boldsymbol{W}(k)\right\| = \mathcal{O}_{\mathrm{P}}(1)$$

for $\nu = 1/(1/2-\kappa) > 2$. Recalling the notation $\overline{\boldsymbol{y}}_{l,n} = 1/(n-l+1)\sum_{t=l}^{n}\boldsymbol{y}_t$, we rewrite this as

$$\overline{\boldsymbol{y}}_{1,k} = \boldsymbol{W}(k)/k + \mathcal{O}_{\mathrm{P}}(k^{1/\nu-1}) \tag{EC.1}$$

uniformly in $k\in\mathbb{N}$. We use (EC.1) to prove Theorem 1. Write

$$\sup_{k=m,\dots,\lfloor m\tau\rfloor}\mathrm{VaR}^M(k)\cdot w\big(k/\lfloor m\tau\rfloor\big) = \sup_{k=0,\dots,\lfloor m(\tau-1)\rfloor} m\overline{\boldsymbol{y}}'_{k+1,k+m}\overline{\boldsymbol{y}}_{k+1,k+m}\cdot w\big((k+m)/\lfloor m\tau\rfloor\big).$$

From (EC.1),

$$\begin{aligned}\overline{\boldsymbol{y}}_{k+1,k+m} &= \big(1+k/m\big)\,\overline{\boldsymbol{y}}_{1,k+m} - (k/m)\overline{\boldsymbol{y}}_{1,k}\\ &= \frac{1}{m}\big[\boldsymbol{W}(k+m)-\boldsymbol{W}(k)\big] + \frac{\mathcal{O}_{\mathrm{P}}((k+m)^{1/\nu}) + \mathcal{O}_{\mathrm{P}}(k^{1/\nu})}{m}.\end{aligned} \tag{EC.2}$$

Using the scaling property of Brownian motion, this implies uniformly in $k\in\{0,\dots,\lfloor m(\tau-1)\rfloor\}$

$$\begin{aligned}\sqrt{m}\overline{\boldsymbol{y}}_{k+1,k+m} &= \frac{\boldsymbol{W}(k+m)-\boldsymbol{W}(k)}{\sqrt{m}} + \mathcal{O}_{\mathrm{P}}\left(\frac{(k+m)^{1/\nu}}{\sqrt{m}}\right)\\ &\overset{\mathcal{D}}{=} \big[\boldsymbol{W}(1+k/m)-\boldsymbol{W}(k/m)\big] + o_{\mathrm{P}}(1).\end{aligned}$$

We then obtain

$$\begin{aligned}\mathrm{VaR}^M(k)\cdot w(k/\lfloor m\tau\rfloor) &= m\overline{\boldsymbol{y}}'_{k+1,k+m}\overline{\boldsymbol{y}}_{k+1,k+m}\cdot w\big((k+m)/\lfloor m\tau\rfloor\big)\\ &= \big[\boldsymbol{W}(1+k/m)-\boldsymbol{W}(k/m)\big]'\big[\boldsymbol{W}(1+k/m)-\boldsymbol{W}(k/m)\big]\cdot w\big((k+m)/\lfloor m\tau\rfloor\big) + o_{\mathrm{P}}(1)\end{aligned}$$

uniformly in $k\in\{0,\dots,\lfloor m(\tau-1)\rfloor\}$. By the modulus of continuity of Brownian motion (cf. Horváth et al. 2006),

$$\sup_{k=0,\dots,\lfloor m(\tau-1)\rfloor}\big[\boldsymbol{W}(1+k/m)-\boldsymbol{W}(k/m)\big]'\big[\boldsymbol{W}(1+k/m)-\boldsymbol{W}(k/m)\big]\cdot w((k+m)/\lfloor m\tau\rfloor)$$

$$\underset{(m\to\infty)}{\overset{\text{a.s.}}{\longrightarrow}} \sup_{t\in[0,\tau-1]} \left[\boldsymbol{W}(1+t)-\boldsymbol{W}(t)\right]' \left[\boldsymbol{W}(1+t)-\boldsymbol{W}(t)\right]\cdot w((t+1)/\tau)$$
$$\overset{\mathcal{D}}{=} \sup_{t\in[0,\tau-1]} \sum_{i=1}^{d} \left[W_i(1+t)-W_i(t)\right]^2 \cdot w((t+1)/\tau),$$

where $W_i(\cdot)$, $i=1,\ldots,d$, are mutually independent standard Brownian motions in $\mathbb{R}$. The conclusion follows. $\square$

Proof of (7). Under $\mathcal{H}_0^{\text{VaR}}$, $\sum_{t=1}^{k} I_{t,\alpha} \overset{\mathcal{D}}{=} X$ with $X\sim \text{Bin}(k,\alpha)$, where $\text{Bin}(k,\alpha)$ denotes the binomial distribution with parameters $k\in\mathbb{N}$ and $\alpha\in(0,1)$. Thus,

$$\begin{aligned}\mu_{uc}^{\text{VaR}}(k) &= \mathrm{E}\left|X-k\alpha\right| \\ &= \sum_{i=0}^{k} \left|i-k\alpha\right| \mathrm{P}\{X=i\} \\ &= \sum_{i=0}^{k} \left|i-k\alpha\right| \binom{k}{i} \alpha^i (1-\alpha)^{k-i},\end{aligned}$$

as desired. $\square$

Proof of (8). Define $S(k,m):=\sum_{t=k}^{m} I_{t,\alpha}$ and $S:=S(1,k)$. Put $I_t:=I_{t,\alpha}$ for short. Use the law of iterated expectation to write

$$\begin{aligned}\mu_{iid}(k) &= \mathrm{E}\left[1/\mathrm{P}\{S\}\sum_{i=1}^{S+1}(t_i-t_{i-1})^2\right] \\ &= \mathrm{E}\left\{\mathrm{E}\left[1/\mathrm{P}\{S\}\sum_{i=1}^{S+1}(t_i-t_{i-1})^2 \;\middle|\; S\right]\right\} \\ &= \sum_{s=0}^{k}\frac{\mathrm{P}\{S=s\}}{\mathrm{P}\{S=s\}}\mathrm{E}\left[\sum_{i=1}^{s+1}(t_i-t_{i-1})^2 \;\middle|\; S=s\right] \\ &= \sum_{s=0}^{k}\sum_{i=1}^{s+1}\mathrm{E}\left[(t_i-t_{i-1})^2 \;\middle|\; S=s\right]. \end{aligned} \tag{EC.3}$$

Note that for fixed k and s, t_i-t_{i-1} can only assume the values $1,\ldots,k-s+1$. We first consider the differences t_i-t_{i-1}, which only involve the $t_1,\ldots,t_s$; and treat the differences involving the boundary values t_0 and t_{s+1} later. For $d=1,\ldots,k-s+1$ we have for any $i=2,\ldots,s$,

$$\begin{aligned}&\mathrm{P}\{t_i-t_{i-1}=d \,|\, S=s\} \\ &= \mathrm{P}\{t_i-t_{i-1}=d,\ t_{i-1}\in\{i-1,\ldots,k-s-d+i\} \,|\, S=s\} \\ &= \sum_{m=i-1}^{k-s-d+i}\mathrm{P}\{t_i-t_{i-1}=d,\ t_{i-1}=m \,|\, S=s\} \\ &= \sum_{m=i-1}^{k-s-d+i}\mathrm{P}\left\{t_i=d+m,\ t_{i-1}=m,\ S=s\right\}/\mathrm{P}\{S=s\} \\ &= \sum_{m=i-1}^{k-s-d+i}\mathrm{P}\big\{S(1,m-1)=i-2,\ I_m=1, I_{m+1}=0,\ldots,I_{d+m-1}=0, I_{d+m}=1, \\ &\qquad\qquad S(d+m+1,k)=s-i\big\}/\mathrm{P}\{S=s\}.\end{aligned}$$

Use that under $\mathcal{H}_0^{\mathrm{VaR}}$, $S(k,m)\sim\mathrm{Bin}(m-k+1,\alpha)$, to derive that

$$\begin{aligned}
&\mathrm{P}\{t_i-t_{i-1}=d\mid S=s\}\\
&=\sum_{m=i-1}^{k-s-d+i}\frac{\left[\binom{m-1}{i-2}\alpha^{i-2}(1-\alpha)^{m-1-(i-2)}\right]\alpha(1-\alpha)^{d-1}\alpha\left[\binom{k-(d+m)}{s-i}\alpha^{s-i}(1-\alpha)^{k-(d+m)-(s-i)}\right]}{\binom{k}{s}\alpha^s(1-\alpha)^{k-s}}\\
&=\sum_{m=i-1}^{k-s-d+i}\frac{\binom{m-1}{i-2}\binom{k-(d+m)}{s-i}}{\binom{k}{s}}.
\end{aligned}\tag{EC.4}$$

Use the fact that $\binom{n}{k}=\binom{n}{n-k}$ and the Rothe–Hagen identity

$$\sum_{k=0}^{n}\frac{a}{a+bk}\binom{a+bk}{k}\binom{c-bk}{n-k}=\binom{a+c}{n}\qquad(a,b,c\in\mathbb{C})$$

to derive that

$$\begin{aligned}
\sum_{m=i-1}^{k-s-d+i}\binom{m-1}{i-2}\binom{k-(d+m)}{s-i}&=\sum_{\widetilde{m}=0}^{k-s-d+1}\binom{\widetilde{m}+i-2}{i-2}\binom{k-(d+\widetilde{m})-i+1}{s-i}\\
&=\sum_{\widetilde{m}=0}^{k-s-d+1}\binom{i-2+\widetilde{m}}{\widetilde{m}}\binom{k-d+1-i-\widetilde{m}}{k-d+1-s-\widetilde{m}}\\
&=\sum_{\widetilde{m}=0}^{k-s-d+1}\frac{i-1}{i-1+\widetilde{m}}\binom{i-1+\widetilde{m}}{\widetilde{m}}\binom{k-d+1-i-\widetilde{m}}{k-d+1-s-\widetilde{m}}\\
&=\binom{k-d}{k-d+1-s}\\
&=\binom{k-d}{s-1}.
\end{aligned}$$

Plugging this into (EC.4), we obtain

$$\mathrm{P}\{t_i-t_{i-1}=d\mid S=s\}=\binom{k-d}{s-1}\Big/\binom{k}{s}.$$

For $i\in\{1,S+1\}$ we can also show that

$$\mathrm{P}\{t_i-t_{i-1}=d\mid S=s\}=\binom{k-d}{s-1}\Big/\binom{k}{s}.$$

We do so for $i=S+1$, the case $i=1$ can be dealt with similarly. For $d=1,\dots,k-s+1$,

$$\begin{aligned}
&\mathrm{P}\{t_{S+1}-t_S=d\mid S=s\}=\mathrm{P}\{k+1-t_s=d\mid S=s\}\\
&=\mathrm{P}\{t_s=k+1-d\mid S=s\}\\
&=\mathrm{P}\{t_s=k+1-d,S=s\}/\mathrm{P}\{S=s\}\\
&=\mathrm{P}\{I_k=0,I_{k-1}=0,\dots,I_{k-d+2}=0,I_{k-d+1}=1,S=s\}/\mathrm{P}\{S=s\}\\
&=\mathrm{P}\{I_k=0,I_{k-1}=0,\dots,I_{k-d+2}=0,I_{k-d+1}=1,S(1,k-d)=s-1\}/\mathrm{P}\{S=s\}\\
&=\mathrm{P}\{I_k=0\}\,\mathrm{P}\{I_{k-1}=0\}\cdot\ldots\cdot\mathrm{P}\{I_{k-d+2}=0\}\,\mathrm{P}\{I_{k-d+1}=1\}\,\mathrm{P}\{S(1,k-d)=s-1\}/\mathrm{P}\{S=s\}\\
&=\frac{(1-\alpha)^{d-1}\alpha^1\binom{k-d}{s-1}\alpha^{s-1}(1-\alpha)^{k-d-(s-1)}}{\binom{k}{s}\alpha^s(1-\alpha)^{k-s}}\\
&=\binom{k-d}{s-1}\Big/\binom{k}{s}.
\end{aligned}$$

Thus, for any $i=1,\dots,s+1$,

$$\begin{aligned}\mathrm{E}\left[(t_i-t_{i-1})^2 \;\middle|\; S=s\right] &= \sum_{d=1}^{k-s+1} d^2\,\mathrm{P}\{t_i-t_{i-1}=d \mid S=s\}\\ &= \sum_{d=1}^{k-s+1} d^2 \frac{\binom{k-d}{s-1}}{\binom{k}{s}}\\ &= \frac{(k+1)(2k-s+2)}{(s+1)(s+2)}.\end{aligned}$$

Plugging this into (EC.3) gives,

$$\begin{aligned}\mu_{iid}(k) &= \sum_{s=0}^{k}(s+1)\frac{(k+1)(2k-s+2)}{(s+1)(s+2)}\\ &= (k+1)\left\{2(k+2)[\psi(k+3)-\psi(1)-1]-(k+1)\right\}.\end{aligned}$$

This concludes the proof. $\square$

Appendix B: Limit Theory under Estimation Risk

As argued in Section 5 of the main paper, we view a test of (11) as more relevant from a practical perspective, because it directly focuses on the quality of the VaR forecasts themselves. It may however still be of interest to know whether the model that produced the VaR forecasts could plausibly be the 'true model', such that tests of (12) are required. Likewise, for ES forecasts, instead of $\mathcal{H}_0^{\mathrm{ES}}$, one may want to be able to test

$$H_{t,\alpha}(\boldsymbol{\theta}) \text{ is i.i.d. with } H_{t,\alpha}(\boldsymbol{\theta}) \overset{\mathcal{D}}{=} \frac{\alpha-U}{\alpha} I_{\{U\leq\alpha\}}, \quad U\sim\mathcal{U}[0,1], \tag{EC.5}$$

where $H_{t,\alpha}(\boldsymbol{\theta})=1/\alpha(\alpha-U_t(\boldsymbol{\theta}))I_{\{U_t(\boldsymbol{\theta})\leq\alpha\}}$ and $\boldsymbol{\theta}$ denotes the true parameter vector from the parameter space $\boldsymbol{\Theta}$. Here, $U_t=U_t(\boldsymbol{\theta})=F_t(Y_t)=F_t(Y_t;\Omega_{t-1},\boldsymbol{\theta})$ with the dependence of the conditional distribution function on the model parameters and the information set now made explicit. In the sequel, $\widetilde{\boldsymbol{\theta}}$ denotes a generic parameter vector from the parameter space $\boldsymbol{\Theta}$. In this section, we derive monitoring procedures for (EC.5) assuming a fixed-window estimation scheme. For brevity, we do so only for the detector $\mathrm{ES}^C(k)$, but the arguments here could also be used for $\mathrm{ES}^M(k)$ and $\mathrm{ES}^S(k)$ as well as other estimation schemes.

To do so, we impose the following conditions, where we set $\varphi(U_t)=1/\alpha(\alpha-U_t)I_{\{U_t\leq\alpha\}}$, $c_\varphi=\mathrm{E}[\varphi(U_t)]$ and $v_\varphi=\mathrm{Var}(\varphi(U_t))$ as shorthand notation.

Assumption EC.1. *The conditional distribution of Y_t given Ω_{t-1} is given by $F_t(\cdot)=F_t(\,\cdot\,;\Omega_{t-1},\boldsymbol{\theta})=\mathrm{P}\{Y_t\leq\cdot\mid\Omega_{t-1}\}$, where $\Omega_{t-1}=\sigma(Y_{t-1},Y_{t-2},\dots;X_{t-1},X_{t-2},\dots)$ with X_t denoting additional exogenous variables.*

Assumption EC.2. *$\left\{(Y_t,X_t)'\right\}_{t\in\mathbb{Z}}$ is strictly stationary and ergodic.*

Assumption EC.3. *The estimator $\widehat{\boldsymbol{\theta}}_E$ is $\sqrt{E}$-consistent for $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is in the interior of the parameter space $\boldsymbol{\Theta}$. Moreover, $\widehat{\boldsymbol{\theta}}_E$ satisfies the following asymptotic (Bahadur) expansion,*

$$\sqrt{E}(\widehat{\boldsymbol{\theta}}_E-\boldsymbol{\theta})=\frac{1}{\sqrt{E}}\sum_{t=-E+1}^{0} l_t+o_{\mathrm{P}}(1),$$

where l_t is such that $\mathrm{E}[l_t\mid\Omega_{t-1}]=0$ and $\mathrm{E}[l_tl_t']$ exists and is positive definite.

Assumption EC.4. *The effect of information truncation satisfies*

$$\sup_{\widetilde{\boldsymbol{\theta}}\in\boldsymbol{\Theta},s\in[0,1]}\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Ts\rfloor}\Big|\varphi\big(F(Y_t;\widehat{\Omega}_{t-1},\widetilde{\boldsymbol{\theta}})\big)-\varphi\big(F(Y_t;\Omega_{t-1},\widetilde{\boldsymbol{\theta}})\big)\Big|=o_{\mathrm{P}}(1),$$

where $\widehat{\Omega}_{t-1}$ *denotes the truncated time-*$(t-1)$ *information set available to the forecaster (e.g.,* $\widehat{\Omega}_{t-1}=\sigma(Y_{t-1},\ldots,Y_{-E+1};X_{t-1},\ldots,X_{-E+1})$*).*

Assumption EC.5. *The conditional distribution function* $F_t^U(x;\widetilde{\boldsymbol{\theta}})=\mathrm{P}\{U_t(\widetilde{\boldsymbol{\theta}})\leq x\mid\Omega_{t-1}\}$ *is continuously differentiable in* $\widetilde{\boldsymbol{\theta}}$ *and* $x\in[0,1]$ *almost surely. Moreover,* $\operatorname{Var}\Big(\varphi\big(U_t(\boldsymbol{\theta})\big)\Big)<\infty$*,*

$$\mathrm{E}\left[\sup_{\widetilde{\boldsymbol{\theta}}\in\boldsymbol{\Theta},x\in[0,1]}\left\|\frac{\partial F_t^U(x;\widetilde{\boldsymbol{\theta}})}{\partial x}\right\|\right]<C,$$
$$\mathrm{E}\left[\int_0^1\sup_{\widetilde{\boldsymbol{\theta}}\in\boldsymbol{\Theta}}\left\|\frac{\partial F_t^U(x;\widetilde{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\right\|\mathrm{d}\varphi(x)\right]<C.$$

While (EC.5) is based on $U_t(\boldsymbol{\theta})$, only the $\widehat{U}_t=F_t(Y_t;\widehat{\Omega}_{t-1},\widehat{\boldsymbol{\theta}}_E)$ are available to test it. In this section, we deal with both the estimation risk induced by $\widehat{\boldsymbol{\theta}}_E$ and the information truncation by $\widehat{\Omega}_{t-1}$. To do so, define

$$R_{T,j}(x,y)=\frac{1}{T}\sum_{t=1}^{T}\Big[I_{\{U_t\leq x\}}-x\Big]\Big[I_{\{U_{t-j}\leq y\}}-y\Big],$$
$$\widehat{R}_{T,j}(x,y)=\frac{1}{T}\sum_{t=1}^{T}\Big[I_{\{\widehat{U}_t\leq x\}}-x\Big]\Big[I_{\{\widehat{U}_{t-j}\leq y\}}-y\Big].$$

Assumption EC.6. *For* $T/E\to\lambda\in(0,\infty)$ *as* $E\to\infty$ *and* $T\to\infty$*, it holds that*

$$\sup_{x\in[0,1],y\in[0,1],s\in[0,1]}\Big|\sqrt{T}\big[\widehat{R}_{\lfloor Ts\rfloor,j}(x,y)-R_{\lfloor Ts\rfloor,j}(x,y)\big]-\sqrt{\lambda}\sqrt{E}(\widehat{\boldsymbol{\theta}}_E-\boldsymbol{\theta})'E_j(x,y)\Big|=o_{\mathrm{P}}(1),$$

where $E_j(x,y)=\mathrm{E}\Big\{\frac{\partial F_t^U(x,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\big[I_{\{U_{t-j}\leq y\}}-y\big]\Big\}$*.*

Assumptions EC.1–EC.5 are almost identical to Assumptions A0–A4 in Du and Escanciano (2017). We also refer to that paper for some discussion of these conditions. Assumption EC.6 is a high-level assumption that we impose here for convenience. Du and Escanciano (2017, Lemma A1) show that, for fixed $s=1$, Assumption EC.6 already follows from Assumptions EC.1–EC.5. To do so, they build on Du (2016), who in turn heavily draws on Delgado and Escanciano (2007). Thus, it is likely that Assumption EC.6 is actually implied by Assumptions EC.1–EC.5. However, proving this is beyond the scope of the present paper, as it requires extending the weak convergence results of Delgado and Escanciano (2007) and Du (2016) to a functional setting with the additional parameter $s\in[0,1]$.

Let $\gamma_j=\operatorname{Cov}\big(\varphi(U_t),\varphi(U_{t-j})\big)$ and $\rho_j=\gamma_j/\gamma_0$ denote the lag-j autocovariance and autocorrelation, respectively. The infeasible estimators based on the $\{U_t\}$ are

$$\gamma_{T,j}=\frac{1}{T}\sum_{t=1}^{T}\big(\varphi(U_t)-c_\varphi\big)\big(\varphi(U_{t-j})-c_\varphi\big)\quad\text{and}\quad\rho_{T,j}=\gamma_{T,j}/v_\varphi,$$

with feasible counterparts

$$\widehat{\gamma}_{T,j}=\frac{1}{T}\sum_{t=1}^{T}\big(\varphi(\widehat{U}_t)-c_\varphi\big)\big(\varphi(\widehat{U}_{t-j})-c_\varphi\big)\quad\text{and}\quad\widehat{\rho}_{T,j}=\widehat{\gamma}_{T,j}/v_\varphi.$$

Theorem EC.1. *Suppose Assumptions EC.1–EC.6 hold. Then, we have for $T=\lfloor m\tau\rfloor$ with $\tau>1$,*

$$\sup_{k=m,\ldots,\lfloor m\tau\rfloor}\mathrm{ES}^{C}(k)\xrightarrow[(m\to\infty)]{\mathcal{D}}\sup_{s\in[1/\tau,1]}\Big[\boldsymbol{W}(s)/\sqrt{s}+\boldsymbol{Z}\Big]'\Big[\boldsymbol{W}(s)/\sqrt{s}+\boldsymbol{Z}\Big],$$

where $\{\boldsymbol{W}(s)\}$ is a standard $\mathbb{R}^d$-dimensional Brownian motion, independent of the multivariate Gaussian $\boldsymbol{Z}$ with $\mathrm{E}[\boldsymbol{Z}]=0$ and $\mathrm{Var}(\boldsymbol{Z})=\boldsymbol{\Sigma}=(\sigma_{ij}=\lambda R_i'\,\mathrm{E}[l_t l_t']R_j)_{i,j=1,\ldots,d}$, where

$$R_j=-\frac{1}{v_\varphi}\,\mathrm{E}\left[\Big(\varphi(U_{t-j})-c_\varphi\Big)\int_0^1\frac{\partial F_t^U(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathrm{d}\varphi(x)\right],\qquad j=1,\ldots,d.$$

Estimation effects appear in the limiting distribution of $\sup_{k=m,\ldots,\lfloor m\tau\rfloor}\mathrm{ES}^{C}(k)$ through the Gaussian random variable $\boldsymbol{Z}$. However, for a sufficiently large estimation period such that $\lambda\downarrow 0$, $\boldsymbol{Z}$ vanishes and, hence, estimation effects vanish. This implies that even tests of $\mathcal{H}_0^{\mathrm{ES}}$, which do not factor out estimation effects, will not reject ES forecasts from the true model when the estimation period is sufficiently long compared to the evaluation period.

To compute critical values based on Theorem EC.1, we require consistent estimates of σ_{ij}. Since we may only use data from the estimation period, we slightly adapt the estimators proposed by Du and Escanciano (2017) to our setting. Specifically, we suggest to use

$$\widehat{\sigma}_{ij}=\frac{T}{E}\widehat{R}_i'W_E\widehat{R}_j.$$

Here,

$$W_E=\frac{1}{E}\sum_{t=-E+1}^{0}\widehat{l}_t\widehat{l}_t',$$
$$\widehat{R}_j=\frac{1}{\alpha(1/3-\alpha/4)}\frac{1}{E}\sum_{t=-E+1}^{0}\left[\Big(H_{t,\alpha}-\frac{\alpha}{2}\Big)\int_0^\alpha\frac{\partial\widehat{F}_t^U(x;\widehat{\boldsymbol{\theta}}_E)}{\partial\boldsymbol{\theta}}\mathrm{d}x\right],$$

where $\widehat{l}_t$ is a consistent estimator of l_t from Assumption EC.3 such that $W_E=\mathrm{E}[l_t l_t']+o_{\mathrm{P}}(1)$. We refer to Section 4 and Appendix B of Du and Escanciano (2017) for specific estimators $\widehat{l}_t$ and $\widehat{F}_t^U(x;\widehat{\boldsymbol{\theta}}_E)$ in location–scale models.

With the above estimators, one can approximate $\boldsymbol{Z}$ in Theorem EC.1 by generating a zero-mean Gaussian random variable with variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}=(\widehat{\sigma}_{ij})_{i,j=1,\ldots,d}$. The $\mathbb{R}^d$-dimensional Brownian motion $\boldsymbol{W}(\cdot)$ is likewise easy to simulate. Thus, critical values for a test of (EC.5) can easily be computed from Theorem EC.1 via simulations.

Finally, we note that, while Theorem EC.1 is reasonably general, Assumption EC.1 that the Y_t are modeled parametrically via $F_t(\cdot;\Omega_{t-1},\boldsymbol{\theta})$ may be binding. For instance, the CAViaR model of Engle and Manganelli (2004) or the model of Patton et al. (2019) directly focus on VaR (respectively, (VaR, ES)), without any attempt to model the Y_t themselves.

Proof of Theorem EC.1. The outline of the proof is similar to that of Theorem A2 in Du and Escanciano (2017). First, we consider the case of no information truncation, where $\widehat{\Omega}_{t-1}=\Omega_{t-1}$. Use integration by parts (e.g., Shiryaev 1996, Theorem 11) to obtain that

$$\int_0^1\int_0^1 R_{T,j}(x,y)\mathrm{d}\varphi(x)\mathrm{d}\varphi(y)$$

$$
\begin{aligned}
&= \frac{1}{T}\sum_{t=1}^{T}\int_0^1 \big(I_{\{U_t\leq x\}}-x\big)\mathrm{d}\varphi(x)\cdot\int_0^1 \big(I_{\{U_{t-j}\leq y\}}-y\big)\mathrm{d}\varphi(y)\\
&= \frac{1}{T}\sum_{t=1}^{T}\Big[\varphi(1)-\varphi(U_t)-\int_0^1 x\mathrm{d}\varphi(x)\Big]\cdot\Big[\varphi(1)-\varphi(U_{t-j})-\int_0^1 y\mathrm{d}\varphi(y)\Big]\\
&= \frac{1}{T}\sum_{t=1}^{T}\big[-\varphi(U_t)+c_\varphi\big]\cdot\big[-\varphi(U_{t-j})+c_\varphi\big]\\
&= \gamma_{T,j}.
\end{aligned}
$$

Thus, from Assumption EC.6 and the continuous mapping theorem (Davidson 1994, Theorem 26.13), it follows that

$$\sqrt{T}\widehat{\gamma}_{\lfloor Ts\rfloor,j}=\sqrt{T}\gamma_{\lfloor Ts\rfloor,j}+\sqrt{\lambda}\sqrt{E}(\widehat{\boldsymbol{\theta}}_E-\boldsymbol{\theta})'\int_0^1\int_0^1 E_j(x,y)\mathrm{d}\varphi(x)\mathrm{d}\varphi(y)+o_{\mathrm{P}}(1)$$

uniformly in $s\in[0,1]$. As in the proof of Theorem A2 in Du and Escanciano (2017), we have

$$\int_0^1\int_0^1 E_j(x,y)\mathrm{d}\varphi(x)\mathrm{d}\varphi(y)=v_\varphi R_j.$$

Hence, uniformly in $s\in[0,1]$,

$$\sqrt{T}\big(\widehat{\gamma}_{\lfloor Ts\rfloor,j}-\gamma_{\lfloor Ts\rfloor,j}\big)=\sqrt{\lambda}\sqrt{E}(\widehat{\boldsymbol{\theta}}_E-\boldsymbol{\theta})'v_\varphi R_j+o_{\mathrm{P}}(1).$$

Using also Assumption EC.3, this implies

$$
\begin{aligned}
\sqrt{T}\widehat{\rho}_{\lfloor Ts\rfloor,j}&=\sqrt{T}\rho_{\lfloor Ts\rfloor,j}+\sqrt{\lambda}\sqrt{E}(\widehat{\boldsymbol{\theta}}_E-\boldsymbol{\theta})'R_j+o_{\mathrm{P}}(1)\\
&=\sqrt{T}\rho_{\lfloor Ts\rfloor,j}+R_j'\sqrt{\lambda}\frac{1}{\sqrt{E}}\sum_{t=-E+1}^{0}l_t+o_{\mathrm{P}}(1)
\end{aligned}
$$

uniformly in $s\in[0,1]$. Then,

$$\sqrt{k}(\widehat{\rho}_{k,1},\dots,\widehat{\rho}_{k,d})'=\sqrt{k}(\rho_{k,1},\dots,\rho_{k,d})'+(R_1,\dots,R_d)'\sqrt{\lambda}\frac{1}{\sqrt{E}}\sum_{t=-E+1}^{0}l_t+o_{\mathrm{P}}(1). \tag{EC.6}$$

By independence of the out-of-sample U_t in $\sqrt{k}(\rho_{k,1},\dots,\rho_{k,d})'$ from the in-sample $\frac{1}{\sqrt{E}}\sum_{t=-E+1}^{0}l_t$, the distributional limits of the two right hand-side terms in (EC.6) are independent.

We consider the first two summands on the right-hand side of (EC.6) separately. By a central limit theorem for martingale differences, it follows that

$$(R_1,\dots,R_d)'\sqrt{\lambda}\frac{1}{\sqrt{E}}\sum_{t=-E+1}^{0}l_t\xrightarrow{\mathcal{D}}(Z_1,\dots,Z_d)'=:\boldsymbol{Z}, \tag{EC.7}$$

where $\boldsymbol{Z}$ is multivariate Gaussian with $\mathrm{E}[\boldsymbol{Z}]=0$ and $\mathrm{Var}(\boldsymbol{Z})=\boldsymbol{\Sigma}=(\sigma_{ij}=\lambda R_i'\,\mathrm{E}[l_tl_t']R_j)_{i,j=1,\dots,d}$.

For the first summand in (EC.6), we replace $\boldsymbol{y}_t$ with $\boldsymbol{h}_t$ in the proof of Theorem 1 to obtain in analogy to (EC.1) that

$$(\rho_{k,1},\dots,\rho_{k,d})'=\overline{\boldsymbol{h}}_{1,k}=\boldsymbol{W}(k)/k+\mathcal{O}_{\mathrm{P}}(k^{1/\nu-1})$$

for $\nu>2$ and a standard $\mathbb{R}^d$-dimensional Brownian motion $\boldsymbol{W}(\cdot)$. Hence,

$$\sqrt{k}(\rho_{k,1},\dots,\rho_{k,d})'=\frac{\sqrt{T}}{\sqrt{k}}\frac{\boldsymbol{W}(k)}{\sqrt{T}}+\mathcal{O}_{\mathrm{P}}(k^{1/\nu-1/2})$$

$$= \frac{\sqrt{T}}{\sqrt{k}} \boldsymbol{W}(k/T) + o_{\mathrm{P}}(1). \tag{EC.8}$$

Use (EC.6)–(EC.8) and the modulus of continuity of Brownian motion to get that

$$\begin{aligned}\sup_{k=m,\ldots,\lfloor m\tau\rfloor} \mathrm{ES}^{C}(k) &= \sup_{k=m,\ldots,\lfloor m\tau\rfloor} k(\widehat{\rho}_{k,1},\ldots,\widehat{\rho}_{k,d})(\widehat{\rho}_{k,1},\ldots,\widehat{\rho}_{k,d})' \\ &\underset{(m\to\infty)}{\xrightarrow{\mathcal{D}}} \sup_{s\in[1/\tau,1]} \Big[\boldsymbol{W}(s)/\sqrt{s}+\boldsymbol{Z}\Big]'\Big[\boldsymbol{W}(s)/\sqrt{s}+\boldsymbol{Z}\Big],\end{aligned}$$

where $\{\boldsymbol{W}(s)\}$ and $\boldsymbol{Z}$ are independent.

Now, consider the case of information truncation. To that end, define $\widetilde{U}_t = F_t(Y_t;\Omega_{t-1},\widehat{\boldsymbol{\theta}}_E)$ and $\widetilde{\gamma}_{T,j} = \frac{1}{T}\sum_{t=1}^{T}\big(\varphi(\widetilde{U}_t)-c_\varphi\big)\big(\varphi(\widetilde{U}_{t-j})-c_\varphi\big)$. Consider the decomposition

$$\sqrt{k}(\widehat{\gamma}_{k,j}-\gamma_{k,j}) = \sqrt{k}(\widehat{\gamma}_{k,j}-\widetilde{\gamma}_{k,j}) + \sqrt{k}(\widetilde{\gamma}_{k,j}-\gamma_{k,j}).$$

We show for the first term on the right hand-side that $\sqrt{k}(\widehat{\gamma}_{k,j}-\widetilde{\gamma}_{k,j}) = o_{\mathrm{P}}(1)$ uniformly in $k\in\{m,\ldots,\lfloor m\tau\rfloor\}$. It holds uniformly in $s\in[1/\tau,1]$ that

$$\begin{aligned}&\sqrt{\lfloor Ts\rfloor}\,|\widehat{\gamma}_{\lfloor Ts\rfloor,j}-\widetilde{\gamma}_{\lfloor Ts\rfloor,j}| \\ &\quad\le \frac{\sqrt{\lfloor Ts\rfloor}}{\lfloor Ts\rfloor}\sqrt{T}\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Ts\rfloor}\Big|\big[\varphi(\widehat{U}_t)-c_\varphi\big]\big[\varphi(\widehat{U}_{t-j})-c_\varphi\big] \\ &\qquad\qquad - \big[\varphi(\widetilde{U}_t)-c_\varphi\big]\big[\varphi(\widetilde{U}_{t-j})-c_\varphi\big]\Big| \\ &\quad= \sqrt{\frac{T}{\lfloor Ts\rfloor}}\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Ts\rfloor}\Big|\big[\varphi(\widehat{U}_t)-\varphi(\widetilde{U}_t)\big]\big[\varphi(\widehat{U}_{t-j})-c_\varphi\big]\Big| \\ &\qquad\qquad + \sqrt{\frac{T}{\lfloor Ts\rfloor}}\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Ts\rfloor}\Big|\big[\varphi(\widetilde{U}_t)-c_\varphi\big]\big[\varphi(\widehat{U}_{t-j})-\varphi(\widetilde{U}_{t-j})\big]\Big| \\ &\quad\le C\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Ts\rfloor}\Big\{\big|\varphi(\widehat{U}_t)-\varphi(\widetilde{U}_t)\big| + \big|\varphi(\widehat{U}_{t-j})-\varphi(\widetilde{U}_{t-j})\big|\Big\} \\ &\quad= o_{\mathrm{P}}(1),\end{aligned}$$

where we have used the boundedness of $|\varphi(\widehat{U}_t)-c_\varphi|$ and $|\varphi(\widetilde{U}_t)-c_\varphi|$ in the second to last step, and Assumption EC.4 in the final step. Hence, we have established that $\sup_{k\in\{m,\ldots,\lfloor m\tau\rfloor\}}|\sqrt{k}(\widehat{\gamma}_{k,j}-\widetilde{\gamma}_{k,j})| = o_{\mathrm{P}}(1)$.

Finally, the arguments that were used for $\sqrt{k}(\widehat{\gamma}_{k,j}-\gamma_{k,j})$ in the case of no information truncation, apply verbatim for $\sqrt{k}(\widetilde{\gamma}_{k,j}-\gamma_{k,j})$. This completes the proof. □

As in the context of VaR forecasts (see Remark 6), we stress that asymptotics as in Theorem EC.1 are a poor guide in finite samples because the sample size T is often too short relative to α. It may therefore be quite appealing to resort to bootstrap implementations of the monitoring procedures. Any bootstrap scheme would however have to replicate the limiting distribution from Theorem EC.1 under both the null and the alternative. Moreover, it would have to do so in what is essentially a *forward-looking* setup, which complicates the implementation. Concretely, the critical values are already needed from the time $t=1$ of the monitoring period onward. Therefore, the sampling behavior of the ES forecasts based on estimated parameters needs to be replicated correctly for *the whole* monitoring period, *already at its beginning*. Under Assumption EC.2 (strict stationarity of the data generating process), this is feasible, since the distribution properties of (Y_t, X_t) may be extrapolated to the monitoring period. But, without strict stationarity or exact knowledge of the data generating process for the entire monitoring period, there is no resampling method that may resample data not yet observed.

For fixed-window estimation, one may for instance use the following bootstrap scheme:

1. Using a suitable bootstrap scheme,[11] generate bootstrap samples $\{(Y_t^*, X_t^*)'\}_{t=-E+1,\ldots,T}$.

2. Based on $\{(Y_t^*, X_t^*)'\}_{t=-E+1,\ldots,0}$, compute the corresponding bootstrap version of the relevant parameter estimators, $\widehat{\boldsymbol{\theta}}^*$.

3. Compute the resulting sequence of resampled estimated conditional d.f.s, $F_t^* = F_t(\cdot;\Omega_{t-1}^*, \widehat{\boldsymbol{\theta}}^*)$, and the corresponding probability bootstrapped integral transforms $\widehat{U}_t^* = F_t^*(Y_t^*)$, where $\Omega_{t-1}^* = \sigma\big((Y_{t-1}^*, X_{t-1}^*)', (Y_{t-2}^*, X_{t-2}^*)', \ldots\big)$.

4. Compute the bootstrap cumulative violations $H_{t,\alpha}^* = \varphi\big(\widehat{U}_t^*\big)$, $t = 1, \ldots, T$.

5. Compute the desired bootstrap monitoring statistic using $H_{t,\alpha}^*$, $t = 1, \ldots, T$.

6. Use the quantiles of the resulting bootstrap distribution[12] for inference.

A rigorous proof of validity is beyond the scope of this appendix. We emphasize again that choosing an appropriate bootstrap scheme is a key step in the procedure, and that it may be difficult to resample data for the entire monitoring period without additional assumptions such as strict stationarity.

Appendix C: Additional Simulations With Different Moving Window Sizes

Here, we illustrate the performance of the MOSUM detectors $\mathrm{VaR}^M(m,k)$, $\mathrm{VaR}^M_{\mathrm{MCS}}(m,k)$, $\mathrm{ES}^M(m,k)$ and $\mathrm{ES}^M_{\mathrm{MCS}}(m,k)$ for different choices of m. Specifically, we choose $m \in \{25, 50, 100\}$, where $m = 50$ was used in Section 6. We use the same simulation design as in Section 6. Figures EC.1 and EC.2 are then the analogues of Figures 1 and 2, respectively.

Figure EC.1 shows the results for the VaR detectors. The differences between $m = 25$ and $m = 50$ are relatively small, yet choosing $m = 100$ leads to worse detection properties. For ES forecasts, the differences between the three choices of m are smaller, particularly for $\mathrm{ES}^M(m,k)$. Overall, a slight preference for $m = 50$ emerges due to its power (top panel) and its ability to improve risk forecasts (bottom panel). This evidence justifies our use of $m = 50$ in the simulations.

References

Christoffersen P (1998) Evaluating interval forecasts. *International Economic Review* 39(4):841–862.

Davidson J (1994) *Stochastic Limit Theory* (Oxford: Oxford University Press).

Delgado MA, Escanciano JC (2007) Nonparametric tests for conditional symmetry in dynamic models. *Journal of Econometrics* 141(2):652–682.

[11] Without precise assumptions on the data generating processes it is actually not clear at all which bootstrap should be used. Given strict stationarity, the moving-block bootstrap would plausibly work in many cases; furthermore, the parametric specification of the conditional d.f. F_t may justify a form of residual-based bootstrap. At a minimum, the analogue of Assumption EC.3 should hold true for the resampled estimators.

[12] Since these are seldom available analytically, proceed as usual, and repeat Steps 1–5 a large number B of times to approximate the cdf of the desired bootstrap distribution via the empirical cdf of the B resampled statistics.

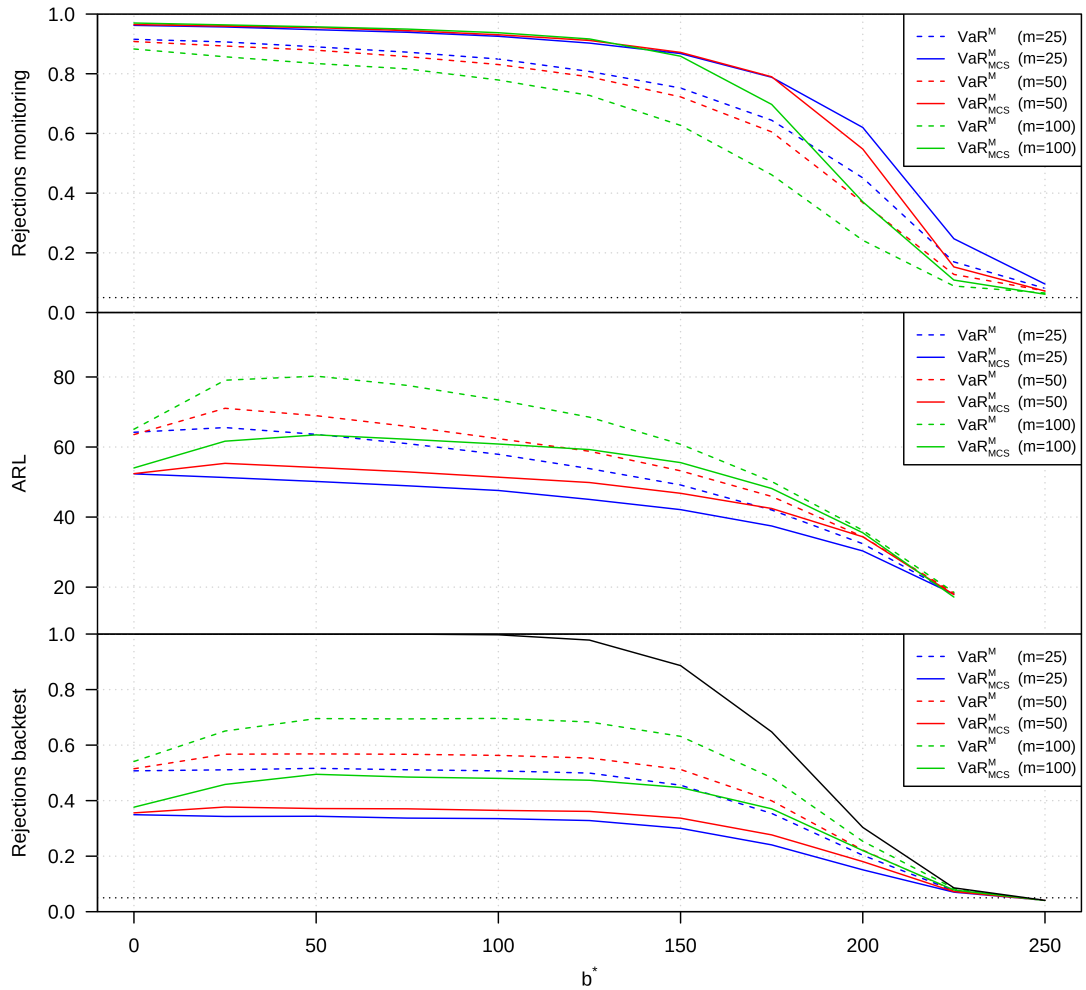

Figure EC.1 Top panel: Empirical rejection frequencies of MOSUM VaR monitoring procedures for $m=25$, $m=50$ and $m=100$. Dotted horizontal line indicates nominal level of 5%. Middle panel: ARLs of monitoring procedures. Bottom panel: Empirical rejection frequencies of backtest of Christoffersen (1998) for non-monitored (black solid line) and monitored VaR forecasts. Dotted horizontal line indicates nominal level of 5% of the backtest.

Du Z (2016) Nonparametric bootstrap tests for independence of generalized errors. *The Econometrics Journal* 19(1):55–83.

Du Z, Escanciano JC (2017) Backtesting expected shortfall: Accounting for tail risk. *Management Science* 63(4):940–958.

Engle RF, Manganelli S (2004) CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* 22(4):367–381.

Horváth L, Kokoszka P, Steinebach J (1999) Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis* 68(1):96–119.

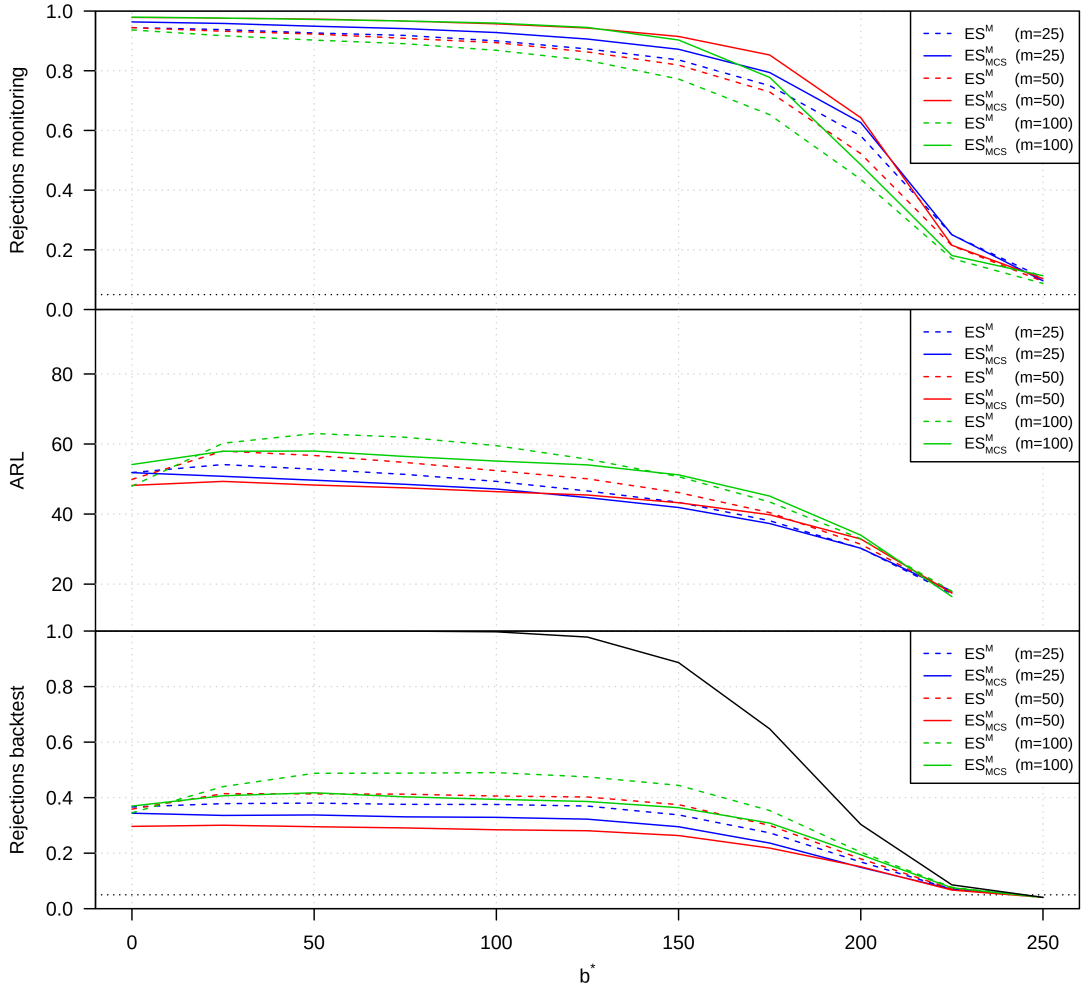

Figure EC.2 Top panel: Empirical rejection frequencies of MOSUM ES monitoring procedures for $m = 25$, $m = 50$ and $m = 100$. Dotted horizontal line indicates nominal level of 5%. Middle panel: ARLs of monitoring procedures. Bottom panel: Empirical rejection frequencies of backtest of Christoffersen (1998) for non-monitored (black solid line) and monitored ES forecasts. Dotted horizontal line indicates nominal level of 5% of the backtest.

Horváth L, Kokoszka P, Zhang A (2006) Monitoring constancy of variance in conditionally heteroskedastic time series. *Econometric Theory* 22(3):373–402.

Patton AJ, Ziegel JF, Chen R (2019) Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics* 211(2):388–413.

Shiryaev AN (1996) *Probability* (Berlin: Springer), 2nd edition.