Bayesian hierarchical model for financial time series

Leonardo Bottolo
Petros Dellaportas

University of Cambridge, UK
University College London, UK

lb664@cam.ac.uk

CIRM 4th March 2016

Outline

1. Background
2. Motivating example: prediction in time series
3. Modeling multivariate time series
4. Computational issues
5. Evaluating prediction
6. Concluding remarks

Background

- In finance and econometrics, there has been an explosion of papers analyzing returns and volatility by adopting AR-GARCH-type models.

- Of crucial interest is the ability of these models to forecast/ predict returns and volatility at time $t+1$ conditional on the available information at time t.

- The choice of these models is based on forecasting power rather than the ability to explain the data generation mechanism.

Background (cont'd)

- An important practical question is how the length T of time series used for the analysis of data (training data set) affects the model forecasting power.

- Larger values of T provide more robust parameter estimates, smaller T may be more informative in volatility forecasting.

- In econometrics literature, non-parametric forecasting models are adopted where older points receive smaller weights in the inferential procedure.

Motivating example

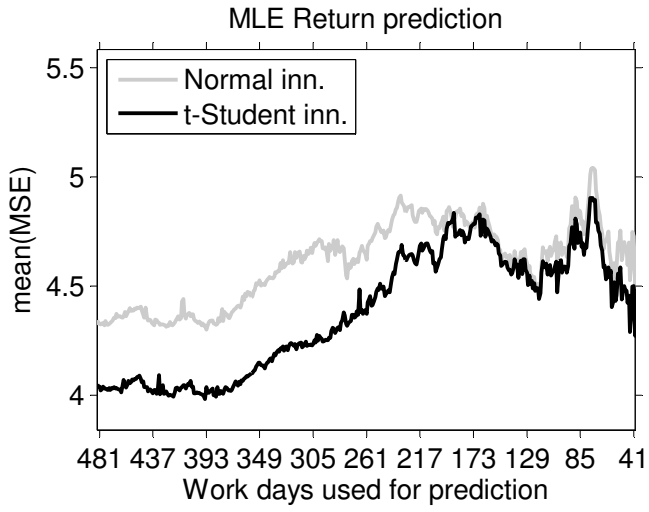
- Assume that at a particular day t, the prediction of returns and volatilities of the one hundred stocks of the financial index SP100 in day $t+1$ is required.

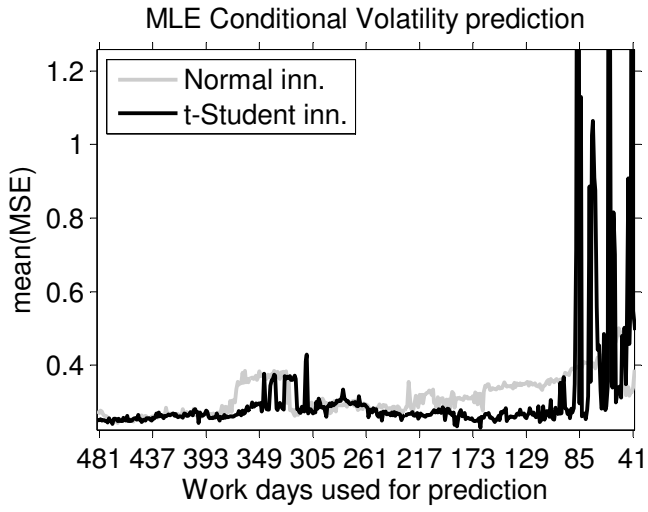
- Assume that AR(1)-GARCH(1,1) model which has five parameters is adopted for each stock.

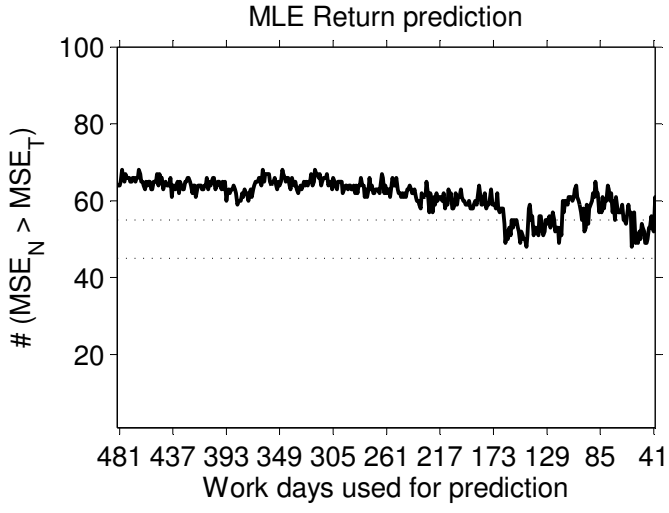
Motivating example (cont'd)

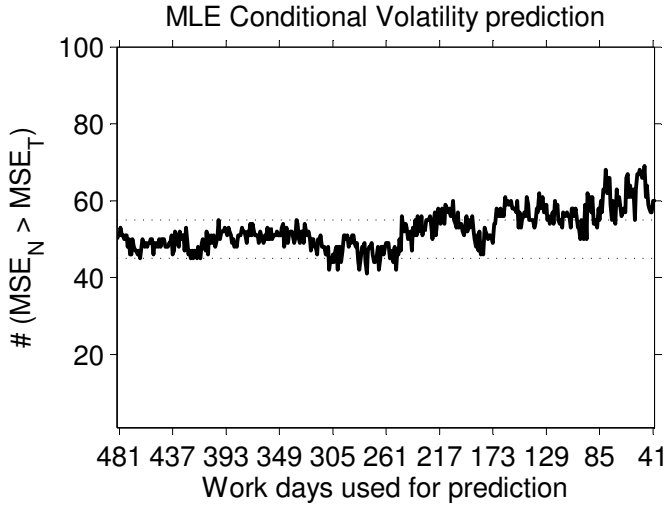
- We use stocks closing price at time $t+1$ to test the predictive power of the model for the returns.
- We take as a proxy for the true volatility at day $t+1$ the quadratic variation calculated using 5-minutes intra-day returns of day $t+1$.
- We choose as a testing criterion of the model forecasting power to be the mean squared error (MSE) of the model prediction at time $t+1$ against the return and quadratic variation at time $t+1$.

Motivating example (cont'd)

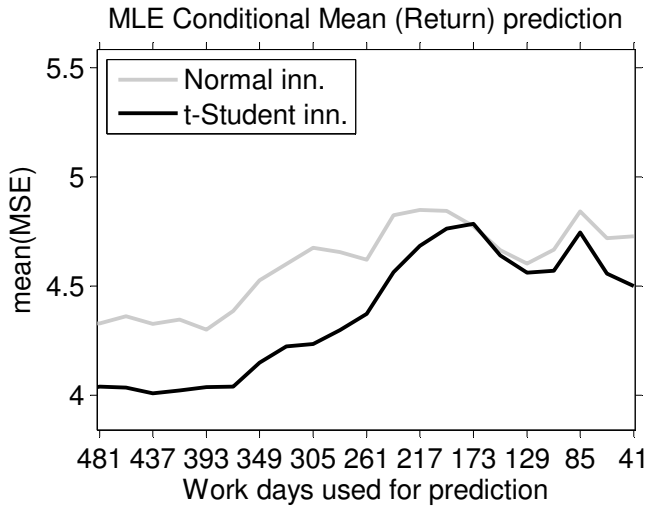


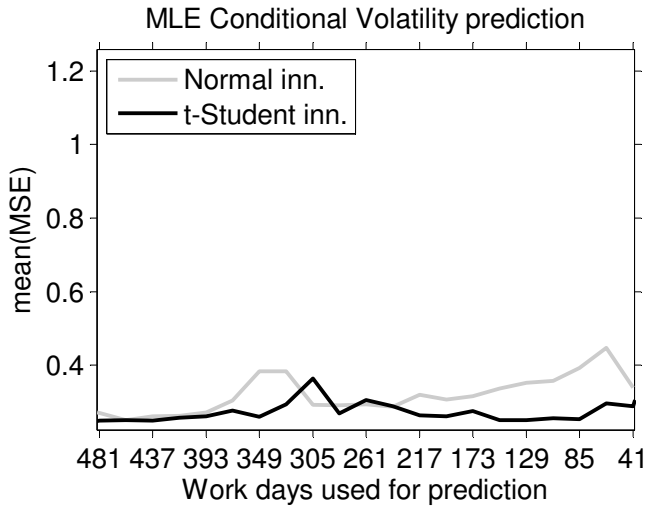


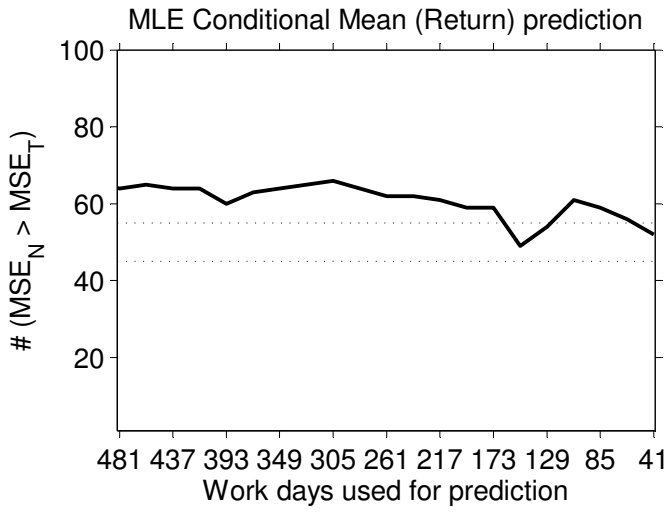


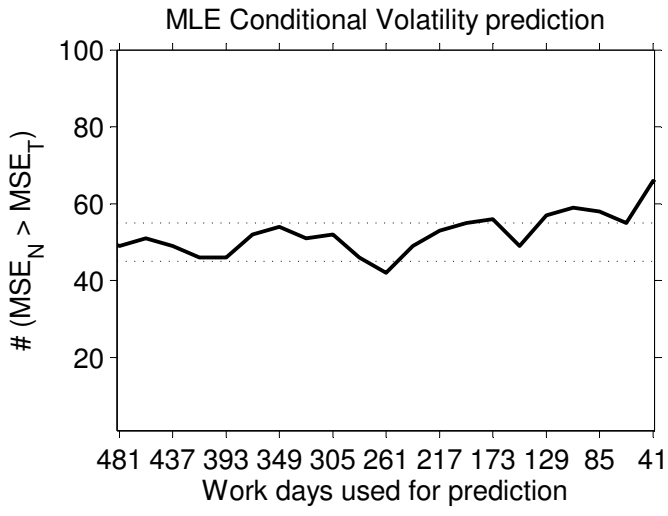


Motivating example (cont'd)









Modeling multivariate time series

- Suppose that for each stock i, $i = 1, \ldots, I$, a times series or returns $(x_{t,i})_{t=1}^{T}$ is observed.

- For each of them a model appropriate for forecasting is an AR(1)-GARCH(1,1) process of the form

$$x_{t,i} = \mu_i + \rho_i x_{t-1,i} + \epsilon_{t,i}$$

with the innovation $\epsilon_{t,i} \sim \sigma_{t,i} t(\nu_i)$, where $t(d)$ is a standard Student-t distribution with d degrees of freedom (DoF) and

$$\sigma_{t,i}^2 = \eta_i + \alpha_i \epsilon_{t-1}^2 + \beta_i \sigma_{t-1,i}^2.$$

Modeling multivariate time series (cont'd)

- We denote with $\boldsymbol{\psi}_i = (\mu_i, \rho_i, \eta_i, \alpha_i, \beta_i)^T$ the parameter vector of the AR(1)-GARCH(1,1) model of stock i. We use $\boldsymbol{\lambda}_i = (\lambda_i^{\mu}, \lambda_i^{\rho}, \lambda_i^{\eta}, \lambda_i^{\alpha}, \lambda_i^{\beta})^T$ for the transformed parameter vector.

- Standard Bayesian hierarchical setup assumes exchangeability across units in a population (*full exchangeable*), allowing borrowing strength and stable inference procedures.

- For example, a standard hierarchical modeling specification in the AR(1)-GARCH(1,1) model would assume that λ_i (without subscript for easy of notation) are exchangeable and follow some common distributions, i.e. $\lambda_i \sim N(\theta, \xi)$.

Modeling multivariate time series (cont'd)

- We assume that the parameter λ_i comes from the mixture model

$$\lambda_i \sim \sum_{h=1}^{k} \omega_h N(\theta_h, \xi_h),$$

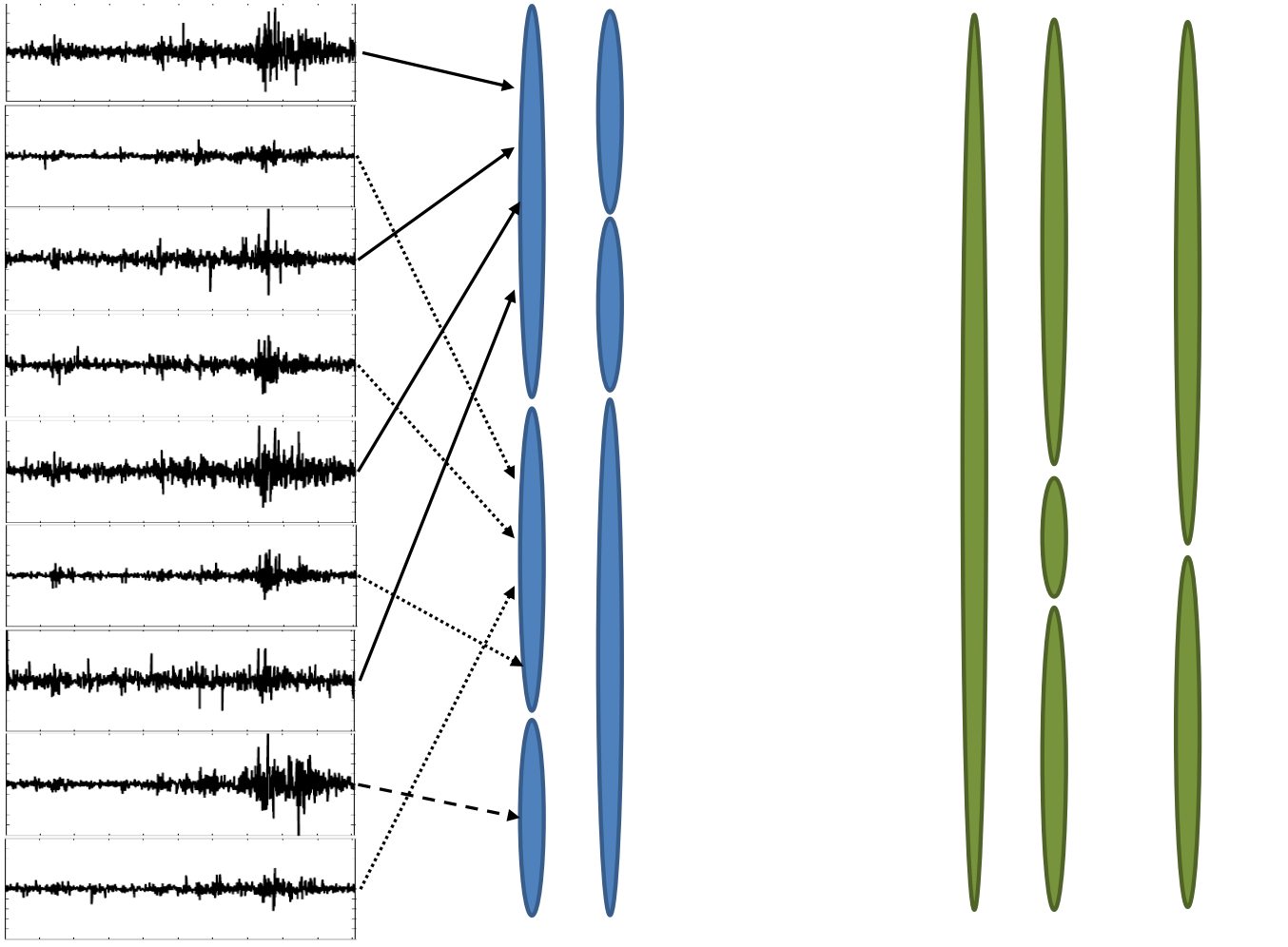
where $\sum_h \omega_h = 1$ and the number of components k is unknown.

- We assign a discrete uniform prior on the number of components, $k \sim Unif(0, I)$ with I is the number of stocks analyzed.

- The hierarchical structure is further specified by assigning a symmetric prior density on the mixture weights $\boldsymbol{\omega} \sim Dir(d_1, \ldots, d_k)$.

Modeling multivariate time series (cont'd)

$$x_{t,i} = \mu_i + \rho_i x_{t-1,i} + \epsilon_{t,i} \qquad \sigma^2_{t,i} = \eta_i + \alpha_i \epsilon^2_{t-1} + \beta_i \sigma^2_{t-1,i}$$

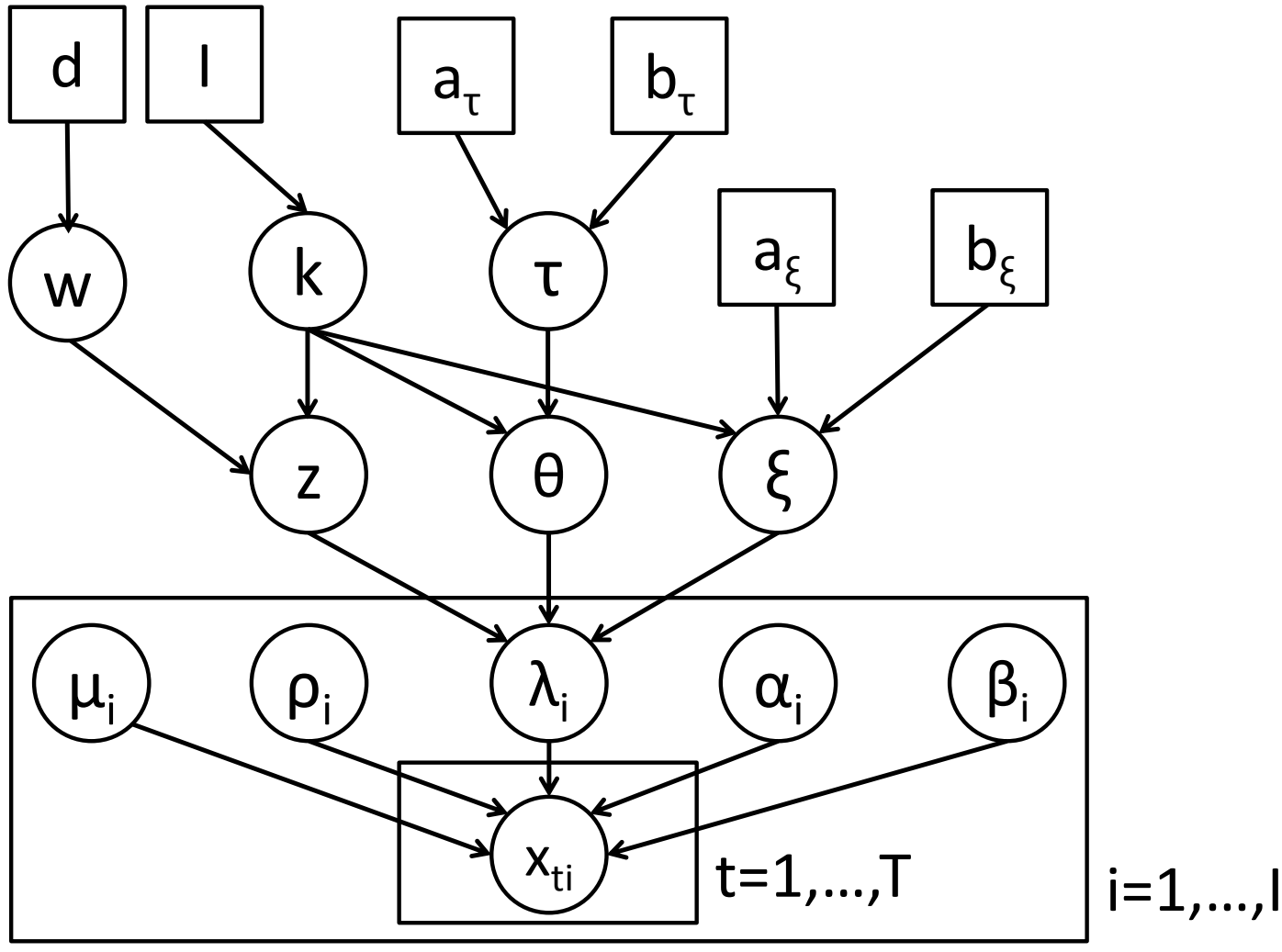


Priors setup

$$\lambda_i \sim \sum_{h=1}^{k} \omega_h N(\theta_h, \xi_h).$$

- The model is completed by specifying prior densities (Nobile and Green, 2000).
  - We assign a normal distribution on the centre of the components, $\theta_h \sim N(m_h, 1/\tau)$.
  - The variance of each mixture component (*within-component variance*) is assumed to have an inverse gamma distribution, $\xi_h^{-1} \sim Ga(a_\xi, b_\xi)$.
  - The *between-components variance* follows an inverse gamma distribution $\tau \sim Ga(a_\tau, b_\tau)$.

Priors setup (cont'd)



Components random proximity

- When the parameters λ_i do not have a physical interpretation or they characterize an unobservable process, prior judgement of their cluster membership appears difficult.

- Since the finite mixture models are defined in the AR(1)-GARCH(1,1) parameter space, we lack of the substantial knowledge needed to be incorporated in the prior specification, i.e. on b_ξ and b_τ.

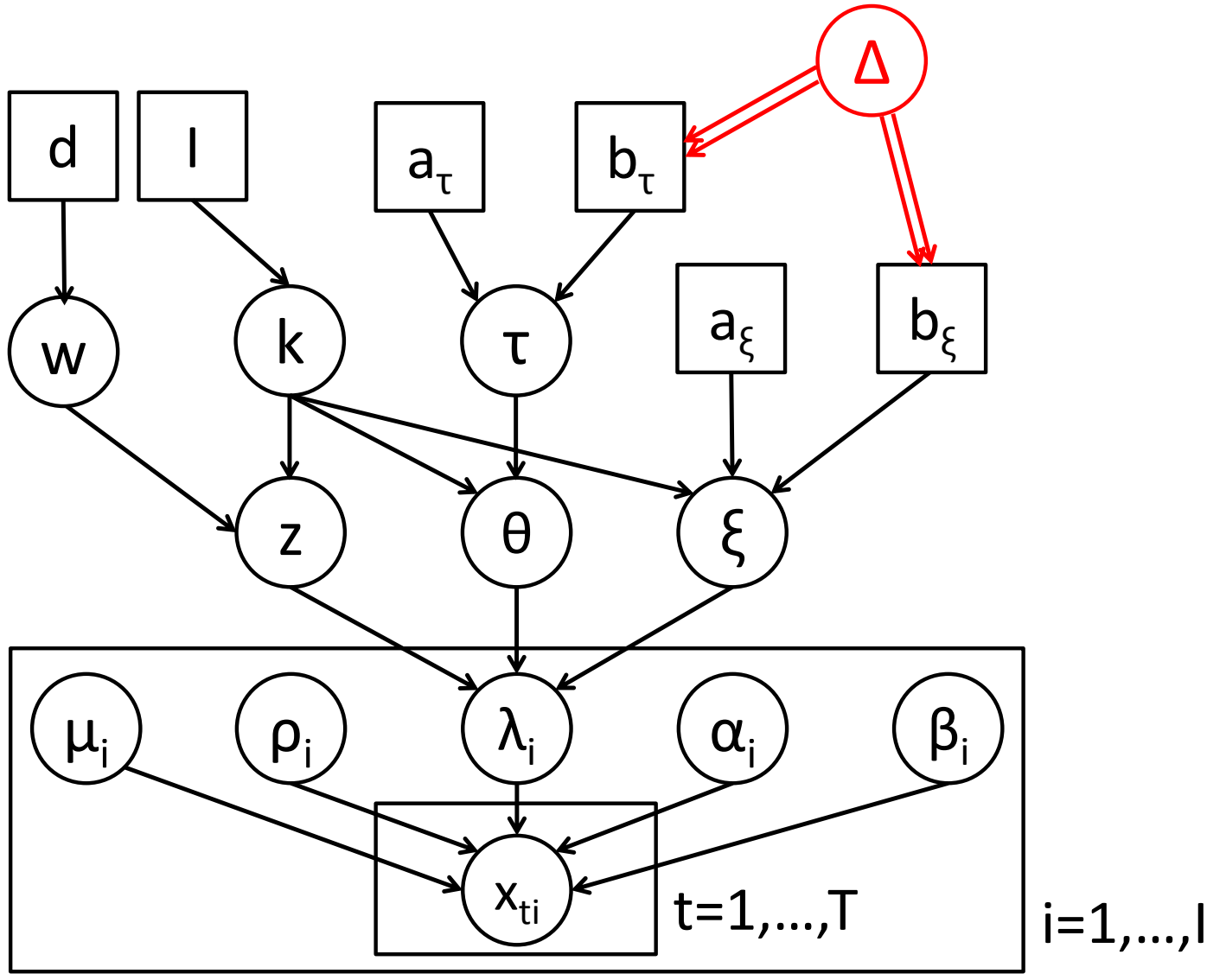
Components random proximity (cont'd)

- Two parameters are considered coming from the same mixture component if $p_0 = \Pr(|\lambda_i - \lambda_l| \leq \Delta)$.

- We build the empirical distribution of the random variable Δ based on all the $I(I-1)/2$ pairwise absolute differences of the MLE, $|\hat{\lambda}_i - \hat{\lambda}_l|$, $i, l = 1, \ldots, I$ with $i > l$.

- We apply an extra layer of hierarchy by assuming that Δ has a gamma distribution center on the median of Δ and variance based on the interquartile range of Δ.

Components random proximity (cont'd)



Components random proximity (cont'd)

- For a specific value of Δ different b_ξ and b_τ are derived.

- Integrating out ξ_h, $\lambda_i - \lambda_l$ follows a t-distribution with $2a_\xi$ DoF, center in 0 and with precision $a_\xi/(2b_\xi)$

$$p_0 = 2F_{2a_\xi}\left\{\Delta\left(\frac{a_\xi}{2b_\xi}\right)^{1/2}\right\} - 1$$

and

$$b_\xi = \frac{a_\xi}{2}\Delta^2\left\{F^{-1}_{2a_\xi}\left(\frac{1+p_0}{2}\right)\right\}^{-2}.$$

Components random proximity (cont'd)

- However the derivation of the two hyper-parameters is computational expensive since a time consuming root-finding algorithm must be applied for b_τ.

- For implementation purposes, a crucial element in our setup is the discretiation of the support of gamma prior, storing the value of b_ξ and b_T for each Δ_d, $d = 1, \ldots, D$. For instance,

$$b_{\xi,d} = \frac{a_\xi}{2}\Delta_d^2 \left\{ F_{2a_\xi}^{-1}\left(\frac{1+p_0}{2}\right)\right\}^{-2}.$$

Computational strategy

- Posterior inference for z, θ_h, ξ_h, τ and Δ is performed using Gibbs sampling.

- We use a reversible jump algorithm Green (1995) to sample from the posterior distribution of k. Both τ and Δ do not depend on the k so they are not included in the acceptance ratio.

- Sampling parameters $\boldsymbol{\psi}_i = (\mu_i, \rho_i, \eta_i, \alpha_i, \beta_i)^T$ is achieved with random walk Metropolis-within-Gibbs algorithm and scaling of all these chains is necessary.

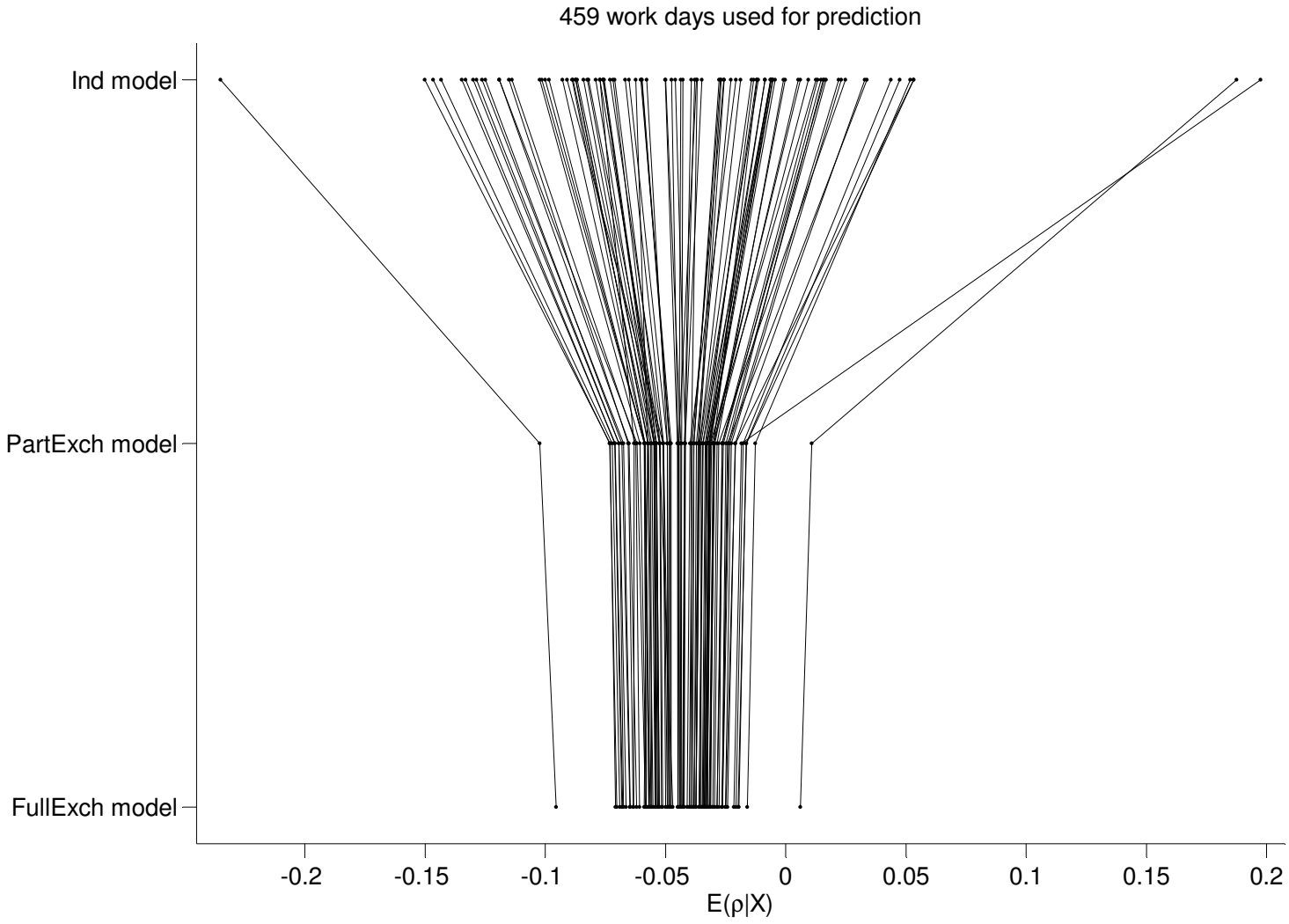
Computational strategy (cont'd)

- For each stock i and for each proposed values $(\mu_i^*, \rho_i^*, \eta_i^*, \alpha_i^*, \beta_i^*)$, we need to evaluated the likelihood $\ell\left((x_{it}|\sigma_{it}^2)_{t=1}^T\right)$ with

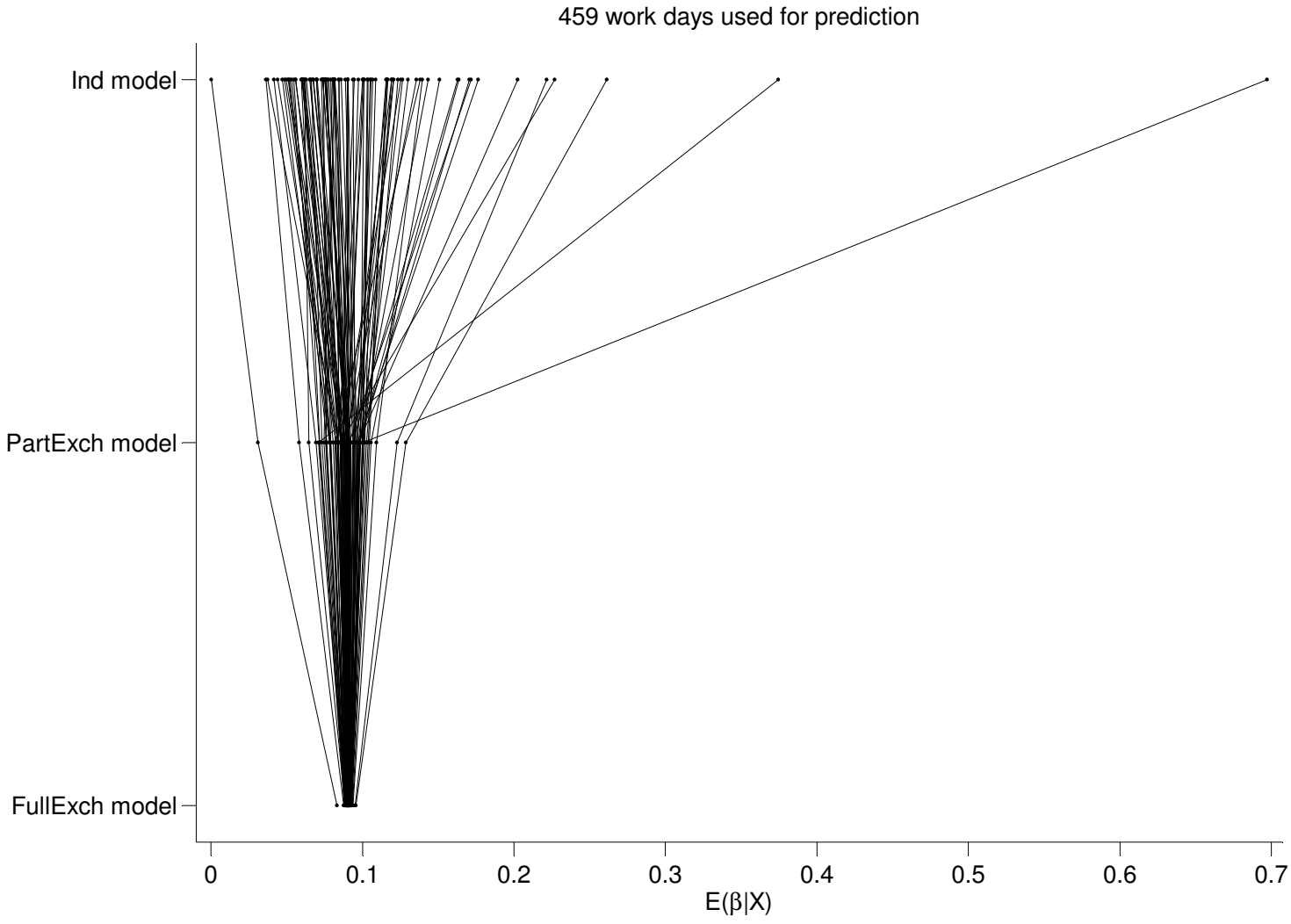
$$\sigma_{t,i}^2 = \eta_i^* + \alpha_i^* \epsilon_{t-1}^2 + \beta_i^* \sigma_{t-1,i}^2.$$

- This is extremely time expensive since we have to simulate the entire path using the recursive volatility equation.

Posterior inference: shrinkage (cont'd)



Posterior inference: shrinkage (cont'd)



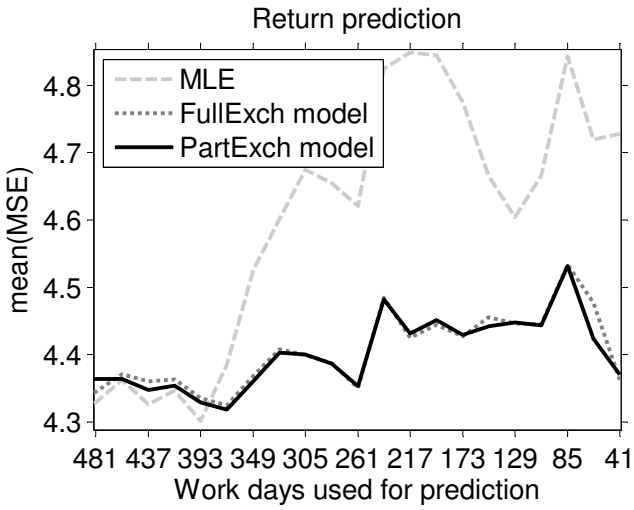
Bayesian predictive densities

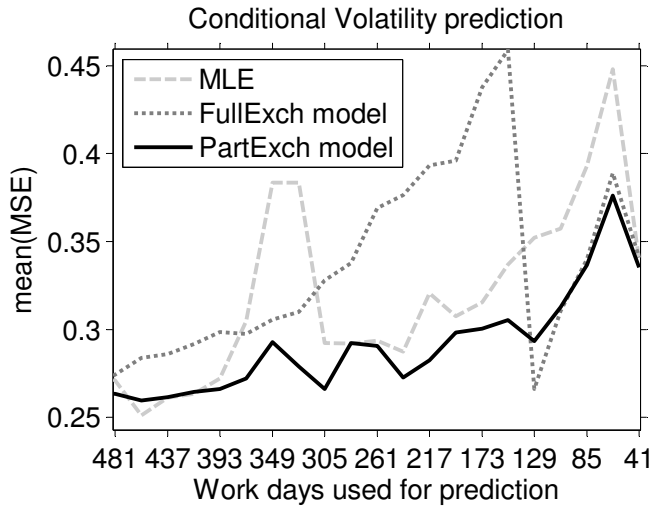
- Bayesian inference allows to derive of the *predictive density* for the next-day return $\pi(x_{t+1,i}|X)$ and the next-day volatility $\pi(\sigma_{t+1,i}|X)$ under any innovation density.

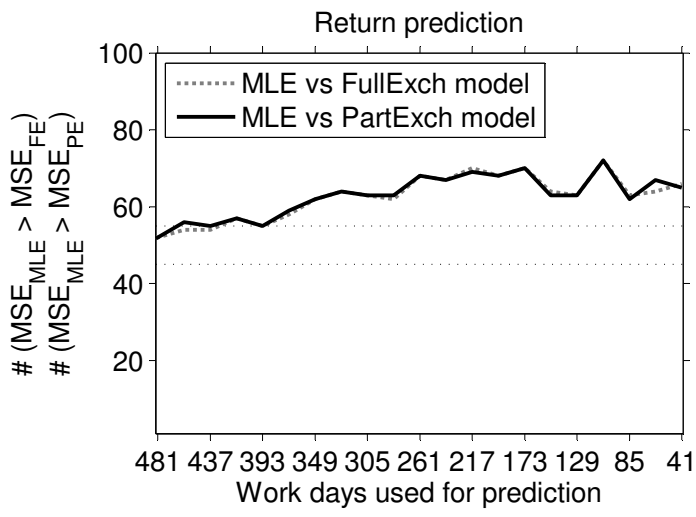
- From these densities moments can be calculated. For instance,

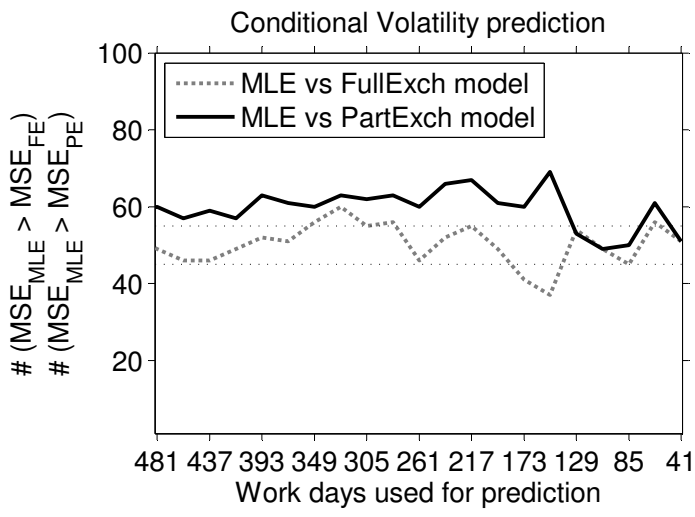
$$E(x_{t+1,i}|X) = \frac{1}{S}\sum_{s=1}^{S} \mu_i^{(s)} + \rho_i^{(s)} x_{t,i}.$$

Comparison with MLE and FullExch model, $t+1$

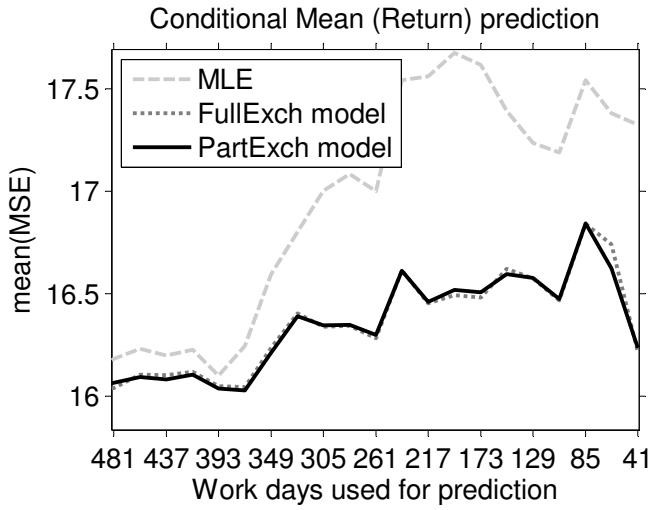


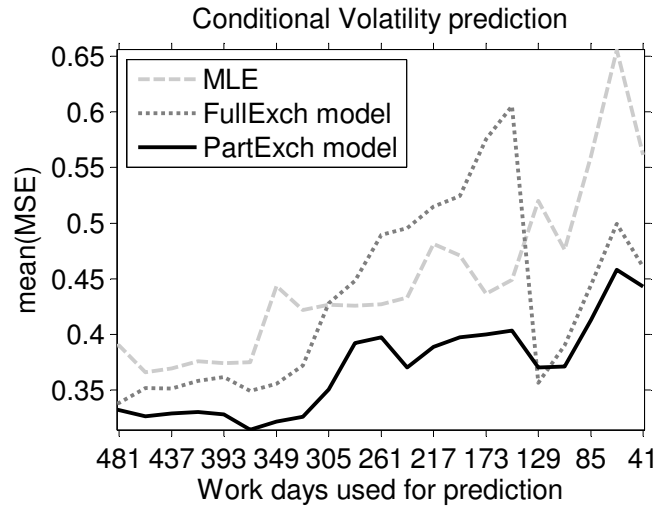


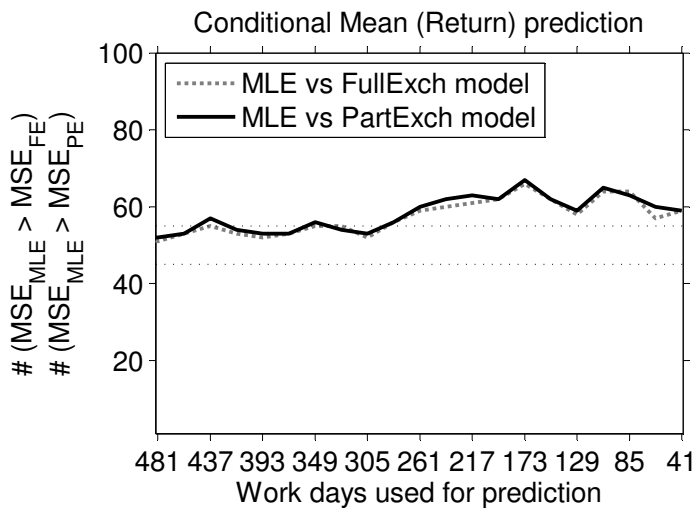


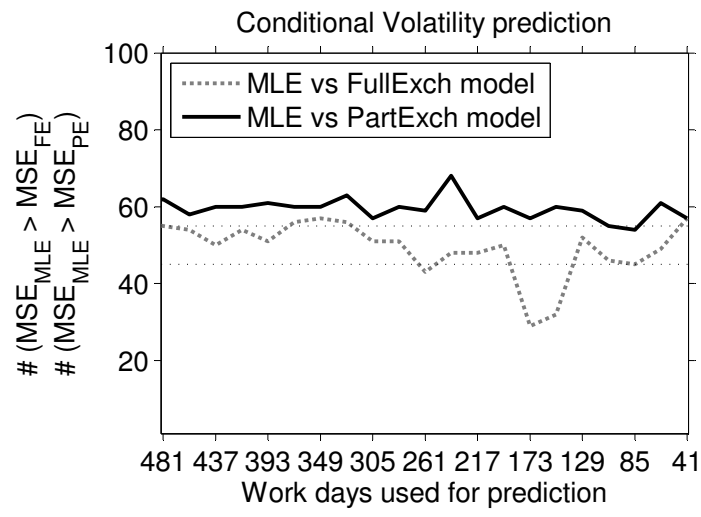


Comparison with MLE and FullExch model, $t+2$









Model assessment using predictive densities

- Return and volatility MSE is defined as

$$MSE(x_{t+1,i}|X) = (x_{t+1,i} - E(x_{t+1,i}|X))^2$$

and

$$MSE(\sigma_{t+1,i}|X) = (\sigma_{t+1,i} - E(\sigma_{t+1,i}|X))^2.$$

- We called them "point vs point" model assessment.

- They do not capture *model uncertainty* that is fully summarised in the predictive densities.

Model assessment using predictive densities (cont'd)

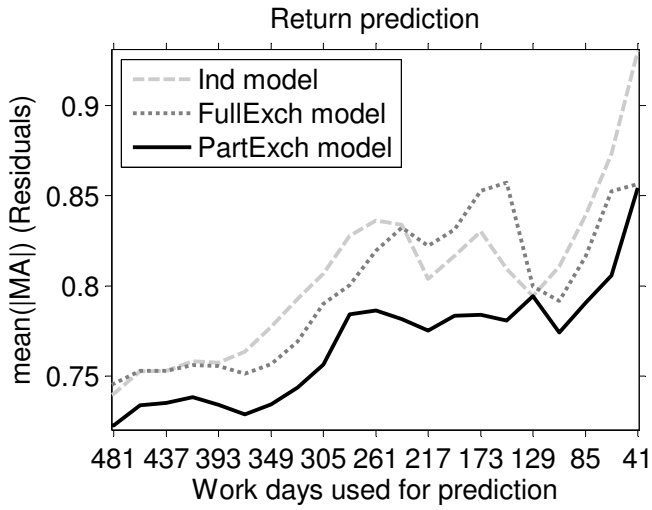
- Other summary statistics have to be used (Gelfand, Dey and Chang, 1992):

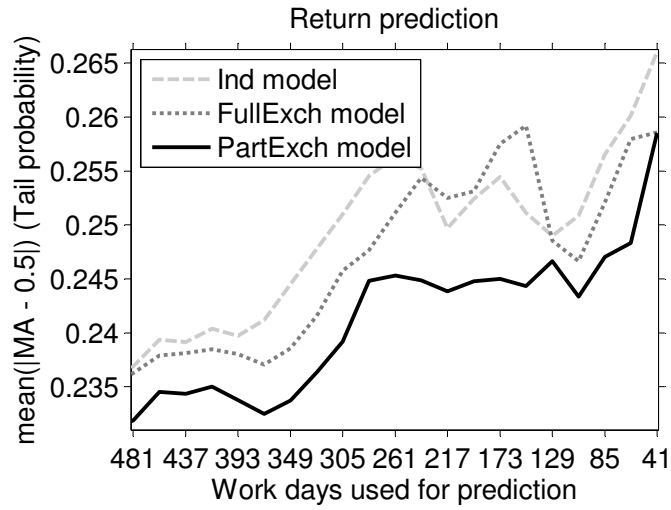
$$MA_1(x_{t+1,i}|X) = \frac{x_{t+1,i} - E(x_{t+1,i}|X)}{\sqrt{V(x_{t+1,i}|X)}} \quad (\textit{Residuals})$$

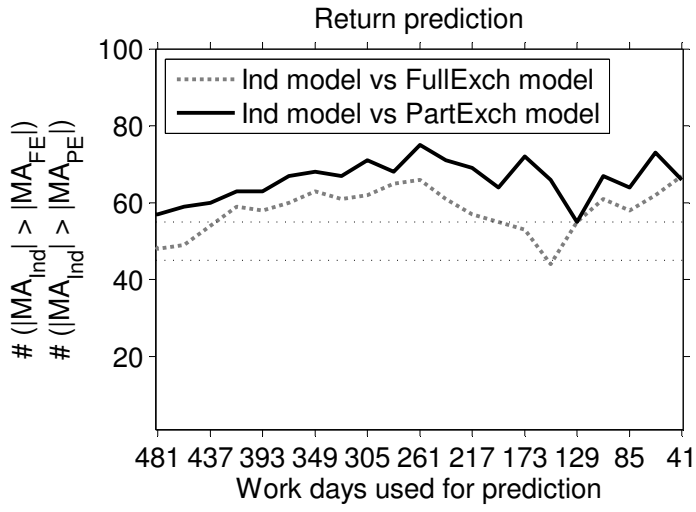
$$MA_2(x_{t+1,i}) = P(X_{t+1,i} \leq x_{t+1,i}|X) \quad (\textit{Tail probability})$$

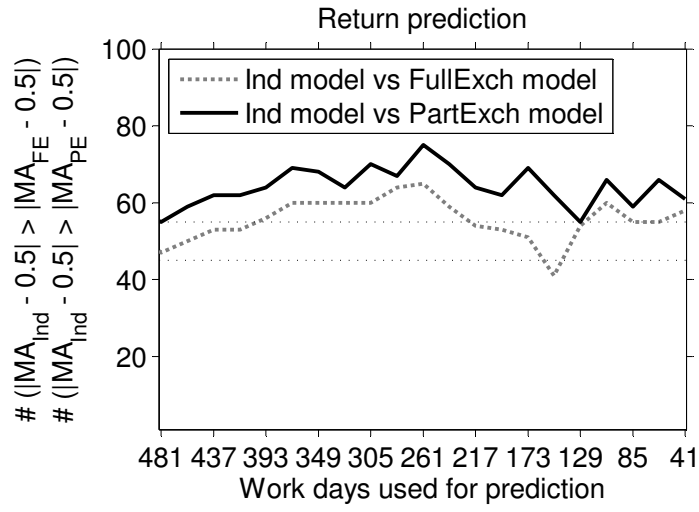
- This is "point vs density" model assessment.

Comparison with Ind and FullExch model, $t+1$

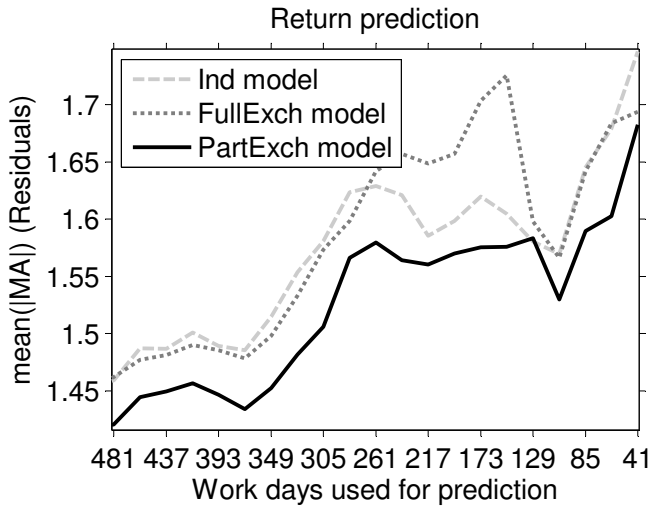


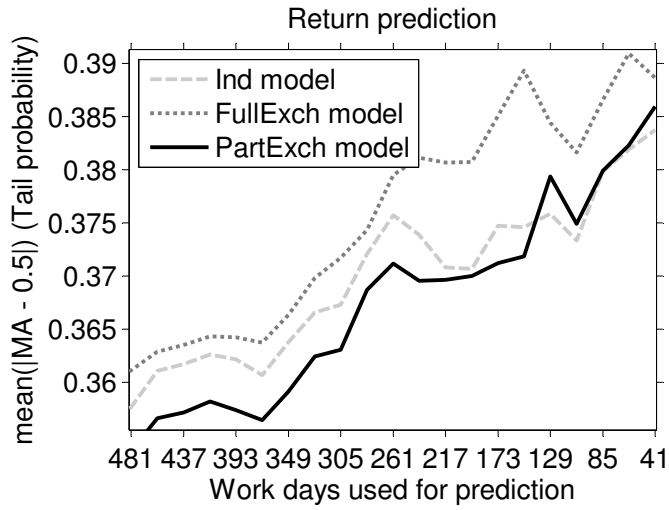


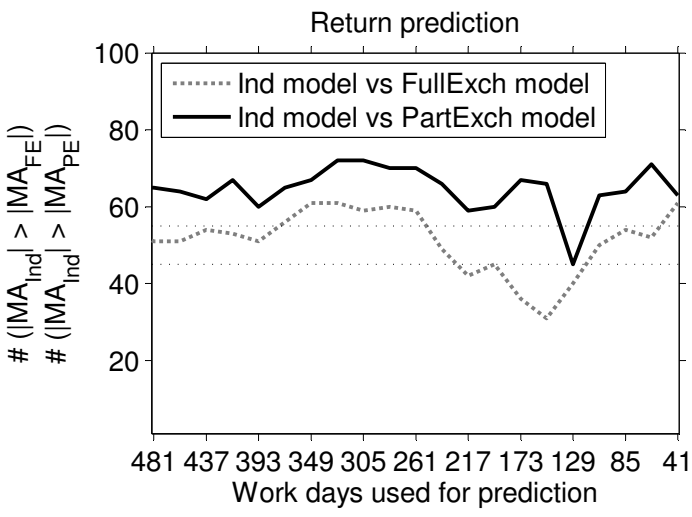


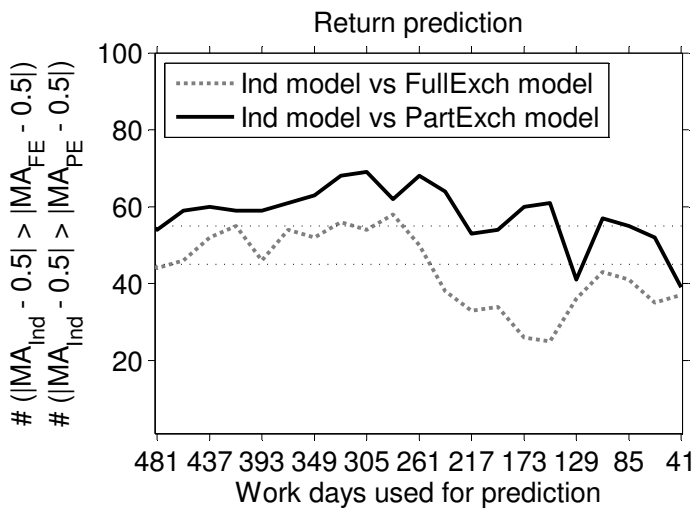


Comparison with Ind and FullExch model, $t + 2$







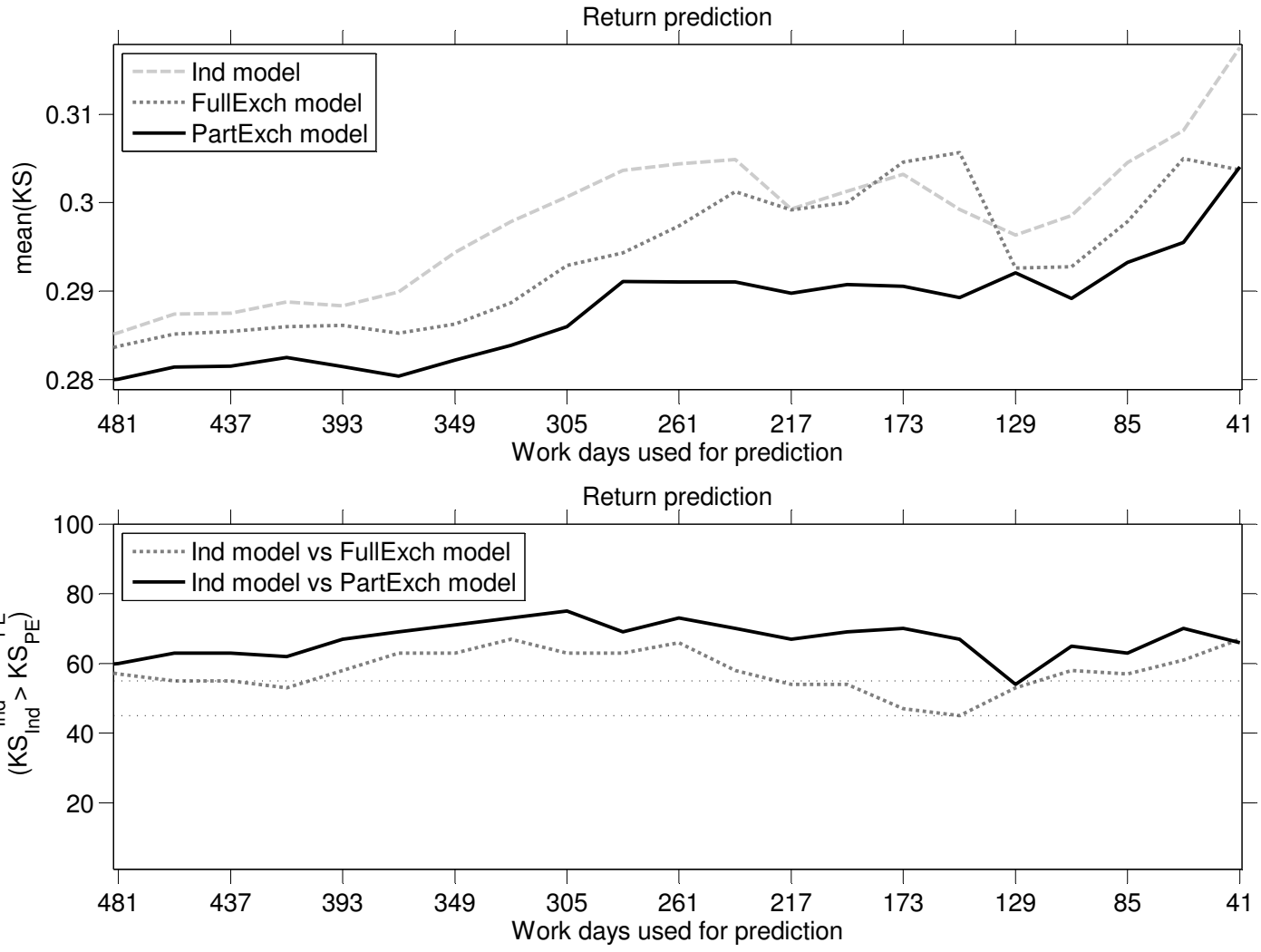


Fully Bayesian model comparison

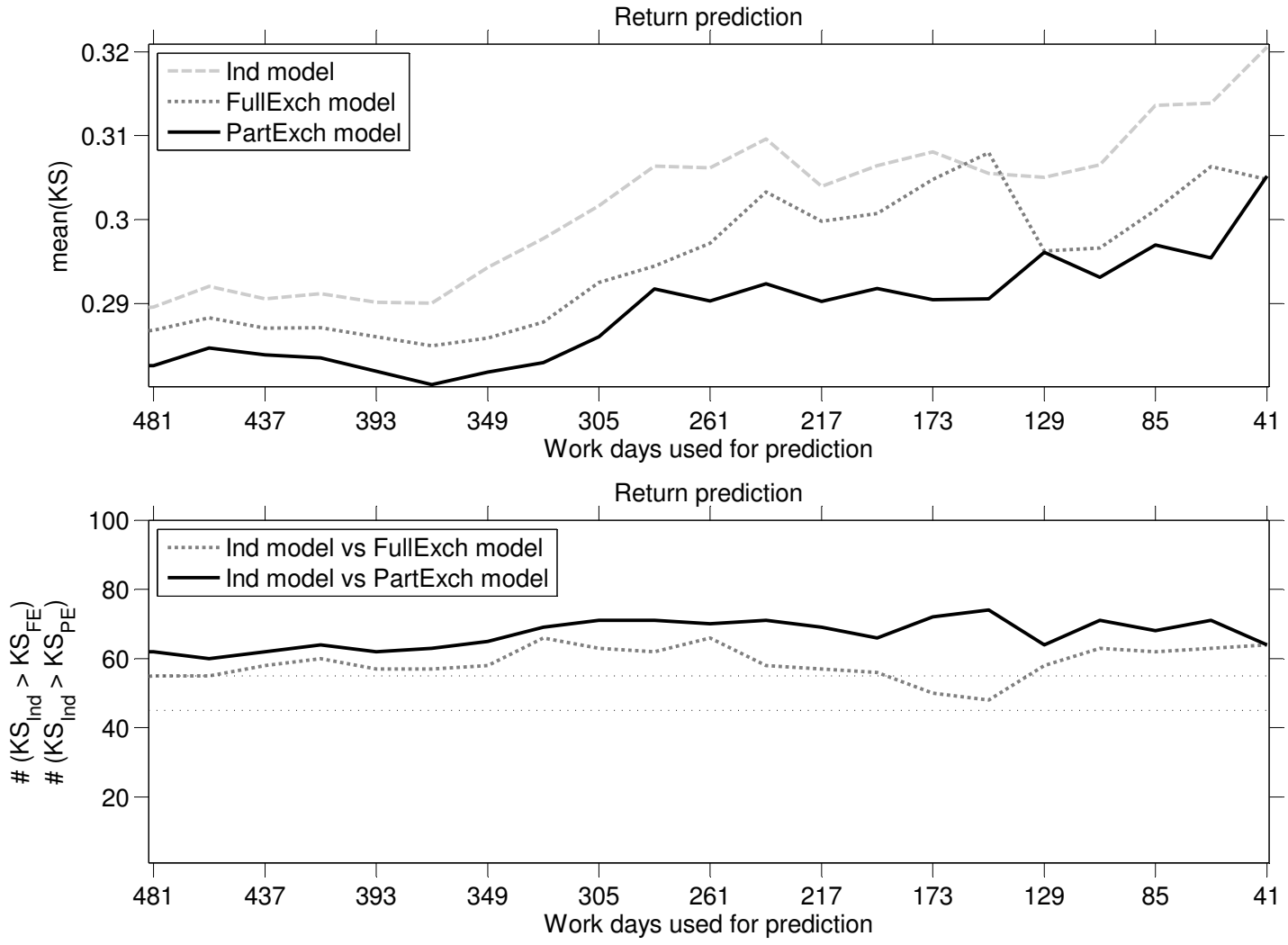
- Since the exact predictive density for normal innovations is known with mean $x_{t+1,i}$ and variance $\sigma^2_{t+1,i}$, we can calculate the distance with the predictive density.

- We called it "density vs density" model assessment.

Comparison with Ind and FullExch model, $t+1$



Comparison with Ind and FullExch model, $t + 2$



Conclusions

- Our *partial exchangeable* model offers more robust return and volatility forecasts.

- It outperforms standard MLE, the bayesian equivalent – *independent* model - and *full exchangeable* model forecasts when the length T of time series is short (less than 2 years).

- Posterior evidence of clustering for each parameter of the AR(1)-GARCH(1,1) model can be used in a variety of financial applications.