Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2025 1–38

Variational Bayes Portfolio Construction

Nicolas Nguyen NICOLAS.NGUYEN@UNI-TUEBINGEN.DE
University of Tübingen

James Ridgway JAMES.RIDGWAY@CFM.COM
Capital Fund Management (CFM), Paris

Claire Vernade CLAIRE.VERNADE@UNI-TUEBINGEN.DE
University of Tübingen

Abstract

Portfolio construction is the science of balancing reward and risk; it is at the core of modern finance. In this paper, we tackle the question of optimal decision-making within a Bayesian paradigm, starting from a decision-theoretic formulation. Despite the inherent intractability of the optimal decision in any interesting scenarios, we manage to rewrite it as a saddle-point problem. Leveraging the literature on variational Bayes (VB), we propose a relaxation of the original problem. This novel methodology results in an efficient algorithm that not only performs well but is also provably convergent. Furthermore, we provide theoretical results on the statistical consistency of the resulting decision with the optimal Bayesian decision. Using real data, our proposal significantly enhances the speed and scalability of portfolio selection problems. We benchmark our results against state-of-the-art algorithms, as well as a Monte Carlo algorithm targeting the optimal decision.

1. Introduction

Portfolio construction (or selection) is a fundamental problem in modern finance (Markowitz, 1952; Merton, 1972), involving the strategic allocation of capital across multiple assets to achieve an optimal tradeoff between risk and return. As financial markets grow in complexity, designing robust portfolios that effectively account for market uncertainty has become increasingly critical. Traditional approaches, such as Markowitz's mean-variance optimization (Markowitz, 1952), have provided a foundational framework for portfolio construction but are now facing challenges in modern finance problems. Markets are becoming increasingly dynamic, with non-Gaussian asset returns, and, in some cases, small dataset sizes. This has led to suboptimal performance in real-world scenarios (see discussions in Benichou et al. (2016)). Formally, the mean-variance framework of a portfolio of d assets can be stated as choosing weights δ from a decision set[1] $\mathcal{D} \subset \mathbb{R}^d$ by solving

$$\operatorname*{argmax}_{\delta \in \mathcal{D}} \delta^T \mu \qquad \text{s.t.} \quad \delta^T \Sigma \delta \leq \lambda \,, \tag{1}$$

where $\mu \in \mathbb{R}^d$ is the mean of observations, $\Sigma \in \mathbb{R}^{d \times d}$ its covariance matrix, and λ is a risk tolerance parameter. Extensive research has focused on improving this framework by addressing key challenges, such as incorporating higher-order moments of return distributions

1. For example, we might consider the d-dimensional simplex as a decision set, $\mathcal{D} = \Delta_d := \{\delta = (\delta_i)_{i \in [d]} \in \mathbb{R}^d : \delta_i \geq 0 \quad \forall i \in [d] \,, \sum_{i=1}^d \delta_i = 1\}$.

© N. Nguyen, J. Ridgway & C. Vernade.

(Harvey et al., 2010), introducing robust optimization techniques (Ismail and Pham, 2019), and refining covariance matrix estimation from noisy data (Benichou et al., 2016; Bun et al., 2017; Agrawal et al., 2022; Benaych-Georges et al., 2023). The Markowitz portfolio provides a systematic approach to balancing return and risk, and despite its limitations, continues to serve as a hard-to-beat benchmark.

Beyond variance-focused methods, utility-based portfolio construction considers an investor's subjective perception of risk and reward, allowing for a nuanced approach to decision-making that can address concerns such as tail risks or other features not captured by variance alone. A particularly useful utility function in this context is the exponential utility function with risk parameter $\lambda > 0$, defined as

$$u_\lambda : (y, \delta) \mapsto \frac{1}{\lambda}(1 - e^{-\lambda y^\top \delta}), \tag{2}$$

which has the following remarkable property: when returns are Normally distributed, maximizing the expected exponential utility is equivalent to the mean-variance optimization in (1) (Merton, 1969),

$$\operatorname*{argmax}_{\delta \in \mathcal{D}} \mathbb{E}_{y \sim \mathcal{N}(\mu, \Sigma)} \left[u_\lambda(y, \delta) \right] \iff \operatorname*{argmax}_{\delta \in \mathcal{D}} \left\{ \mu^\top \delta - \lambda \delta^\top \Sigma \delta \right\}.$$

This equivalence establishes a strong link between the mean-variance theory and utility-based approaches, making the latter a compelling alternative for capturing investor preferences (Gerber and Pafum, 1998). However, this equivalence does not hold beyond Normally distributed returns, since the expected utility function cannot be computed in closed-form. Despite recent advances on this problem (Luxenberg and Boyd, 2024), the general problem of *decision in the face of uncertainty* remains. Specifically, beyond point estimates, we need a reliable estimator of the mean and covariance (μ, Σ) of the recorded historical time series.

Notations. For an arbitrary probability space $(\Theta, \mathcal{T}_\Theta, \pi)$ where Θ is a Polish space, we denote as $\mathcal{M}(\Theta)$ the set of probability distributions on Θ. Throughout this paper, π denotes a generic probability distribution, where its probability space will be clear from context. $\mathcal{S}^d$ is the set of squared positive definite matrices of size d. For a probability distribution π and a random variable θ, $\mathbb{E}_\rho[\theta]$ denotes the expectation of θ when $\theta \sim \rho$. The Kullback-Leibler (KL) divergence between two probability distributions π_1 and π_2 is denoted $\mathcal{K}(\pi_1, \pi_2)$. For a random variable θ and a measure π, we use the infinitesimal notation $\pi(\mathrm{d}\theta)$, generalizing the notation $\pi(\theta = \cdot)$ when Θ is countable.

2. Problem Setting

We consider a supervised learning setting where we have access to n observations $H_n = (Y_t)_{t \in [n]}$, where each $Y_t \in \mathbb{R}^d$. We assume that $(Y_t)_t$ is the realization of a stochastic process parameterized by an unknown parameter θ^*, $(Y_t)_t \sim P_{\theta^*}$. We do not make any additional assumption on P_{θ^*} for now. We define a probability space $(\Theta, \mathcal{T}_\Theta)$ associated with θ and express our initial uncertainty about this parameter through a prior distribution π_0.

Building on the discussions in Section 1, we formalize our portfolio construction problem in the lens of Bayesian decision theory (Robert, 2007, Chapter 2). The *Bayesian decision*

δ^* with respect to a utility function u is the decision $\delta \in \mathcal{D}$ that maximises the *posterior expected utility* (or *Bayesian risk*),

$$\delta^* = \operatorname*{argmax}_{\delta \in \mathcal{D}} \int_{\mathbb{R}^d} u(Y_{n+1}, \delta) \pi(\mathrm{d}Y_{n+1} \,|\, H_n) \,, \tag{3}$$

where $\pi(\mathrm{d}Y_{n+1} \,|\, H_n)$ is the *posterior predictive distribution* of new (unseen) observation Y_{n+1}. Note that δ^* is a function of H_n, $\delta^* = \delta^*(H_n)$, but this dependency is omitted to simplify notation. Under the particular choice of exponential utility (2), we can rewrite (3) as

$$\delta^* = \operatorname*{argmin}_{\delta \in \mathcal{D}} \int_{\mathbb{R}^d} e^{-\lambda \delta^\top Y_{n+1}} \pi(\mathrm{d}Y_{n+1} \,|\, H_n) \,, \tag{4}$$

for a given $\lambda > 0$ fixed by the user. One major challenge in Bayesian modelling is the lack of closed-form solutions for posterior predictive distributions, except for simple statistical models (*e.g.* Gaussian model with deterministic covariance). Hence, directly computing (4) in closed-form is generally infeasible. While various methods exist to numerically compute this integral (*e.g.* Monte-Carlo estimates), they tend to be computationally expensive, particularly in high-dimensional spaces. We address this in the following section by rewriting the objective function as a *saddle point*. We then make use of the same relaxation as in Variational Bayes (VB; Jordan et al., 1999) to approximate the inner optimisation.

3. Exponential Utility Maximization as a Saddle-Point Optimization

3.1. Main Observation

Our main contribution is to show that maximizing an exponential utility function is *equivalent* to solving a saddle-point optimization problem. We believe that the following result may be of independent interest to anyone seeking to maximize an exponential utility for various applications.

Theorem 1 *The optimal Bayesian decision* (4) *can be written as a saddle-point,*

$$\delta^* = \operatorname*{argmin}_{\delta \in \mathcal{D}} \max_{\rho \in \mathcal{M}(\mathbb{R}^d \times \Theta)} \left\{ -\mathcal{K}(\rho, \tilde{\pi}_n) + Z_\delta \right\} \,, \tag{5}$$

where π_n is the posterior distribution over the joint parameter $(y, \theta) \in \mathbb{R}^d \times \Theta$, $\tilde{\pi}_n$ is defined as $\mathrm{d}\tilde{\pi}_n = \frac{e^{-\lambda \delta^\top Y_{n+1}}}{\mathbb{E}_{\pi_n}\left[e^{-\lambda \delta^\top Y_{n+1}}\right]} \mathrm{d}\pi_n$ and $Z_\delta = -\mathbb{E}_{\pi_n}\left[e^{-\lambda \delta^\top Y_{n+1}}\right]$ is a term that does not depend on ρ.

The proof of Theorem 1 relies on a well-known change-of-measure (Donsker and Varadhan, 1983) on the log of the exponential utility. We also introduce the *risk* function $\mathcal{R}_{\mathcal{M}}$ over all probability measures;

$$\forall \delta \in \mathcal{D}, \quad \mathcal{R}_{\mathcal{M}}(\delta) = \sup_{\rho \in \mathcal{M}(\mathbb{R}^d \times \Theta)} \left\{ -\mathcal{K}(\rho, \tilde{\pi}_n) + Z_\delta \right\} \,,$$

and δ^* is the decision that minimizes the risk $\mathcal{R}_{\mathcal{M}}$.

3.2. Variational Bayes Approximation of δ^*

Computing $\tilde{\pi}_n$ is challenging because the normalization constant $\mathbb{E}_{\pi_n}\left[e^{-\lambda\delta^\top Y_{n+1}}\right]$ is intractable for non-conjugate models (for which it is equivalent to computing the integral (4)). We now demonstrate how the min-max formulation in (5) can be leveraged to enable the use of VB approximation.

Maximization over a subspace of measures. Fix an arbitrary decision $\delta \in \mathcal{D}$. Since Z_δ does not depend on ρ, the distribution that solves the maximum writes[2]

$$\rho^*(\delta) := \operatorname*{argmin}_{\rho \in \mathcal{M}(\mathbb{R}^d \times \Theta)} \mathcal{K}(\rho, \tilde{\pi}_n)\,, \tag{6}$$

where we emphasize that ρ^* depends on δ since $\tilde{\pi}_n$ does. VB approximations instead solve a restriction of (6): we define a family of measures $\mathcal{F} \subseteq \mathcal{M}(\mathbb{R}^d \times \Theta)$ for which the restricted problem (6) over this family is considered tractable. For example, the *mean-field family* (Parisi and Shankar, 1988; Bishop, 2006) assumes independence between parameters: assuming Θ factorizes as a product of $K \geq 1$ subspaces, $\Theta = \prod_{i=1}^K \Theta_i$, the mean-field family of $(\mathbb{R}^d, \Theta)$ is defined as

$$\mathcal{F}\left(\mathbb{R}^d \times \Theta\right) = \Big\{\rho \in \mathcal{M}(\mathbb{R}^d \times \Theta) : \rho(\mathrm{d}(Y_{n+1}, \theta)) = \rho_y(\mathrm{d}Y_{n+1}) \prod_{i=1}^K \rho_i(\mathrm{d}\theta_i) : \\ \rho_y \in \mathcal{M}(\mathbb{R}^d),\ \rho_i \in \mathcal{M}(\Theta_i)\, \forall i \in [K]\Big\}\,.$$

Notice that $\mathcal{F}$ does not make *any assumption* on the form of the distributions ρ_y or $(\rho_i)_i$'s, but only relies on the factorisation assumption and the underlying statistical model. We denote by $\hat{\rho}_{\mathrm{VB}}$ the Mean-field variational approximation of ρ^*, that is,

$$\hat{\rho}_{\mathrm{VB}}(\delta) = \operatorname*{argmin}_{\rho \in \mathcal{F}(\mathbb{R}^d \times \Theta)} \mathcal{K}(\rho, \tilde{\pi}_n)\,. \tag{7}$$

Since we deal with a parametric underlying statistical model, $\hat{\rho}_{\mathrm{VB}}$ is also parametric. The main advantage of using $\mathcal{F}$ is that $\hat{\rho}_{\mathrm{VB}}$ can be computed numerically since it is the solution of a fixed-point equation.

Proposition 2 *The variational distribution is written as $\hat{\rho}_{\mathrm{VB}} = \rho_y \otimes_{j=1}^K \rho_j$, where for any $\delta \in \mathcal{D}$ we have*

$$\log \rho_y(\mathrm{d}Y_{n+1}) \propto \mathbb{E}_{\rho_1, \dots \rho_K}\left[\log e^{-\lambda\delta^\top Y_{n+1}} \pi(Y_{n+1}, \theta, H_n)\right]$$
$$\log \rho_j(\mathrm{d}\theta_j) \propto \mathbb{E}_{\rho_y, \boldsymbol{\rho}_{-j}}\left[\log e^{-\lambda\delta^\top Y_{n+1}} \pi(Y_{n+1}, \theta, H_n)\right]\,,$$

where for any $j \in [K]$, $\mathbb{E}_{\boldsymbol{\rho}_{-j}}[\cdot]$ denotes the expectation with respect to the measures $(\rho_i)_{i \in [K] \setminus \{j\}}$.

The proof of the previous equation is provided in Appendix D, and is a direct consequence of a well-known result for Mean-field variational inference (Chapter 10; Bishop, 2006).

2. The negative sign is omitted for now but we will plug it in the final objective function.

Minimization over decisions. Once we found the variational approximation $\hat{\rho}_{\mathrm{VB}}$ for a given δ, we define the corresponding *variational decision* $\hat{\delta}_{\mathrm{VB}}$ by just plugging $\hat{\rho}_{\mathrm{VB}}$ into (5),

$$\hat{\delta}_{\mathrm{VB}} = \operatorname*{argmin}_{\delta \in \mathcal{D}} \mathcal{R}_{\mathcal{F}}(\delta)\,, \tag{8}$$

where we introduced the objective function $\delta \mapsto \mathcal{R}_{\mathcal{F}}(\delta)$,

$$\mathcal{R}_{\mathcal{F}}(\delta) = \sup_{\rho \in \mathcal{F}(\mathbb{R}^d \times \Theta)} \{-\mathcal{K}(\rho, \tilde{\pi}_n) + Z_\delta\} = -\mathcal{K}(\hat{\rho}_{\mathrm{VB}}(\delta), \tilde{\pi}_n) + Z_\delta\,. \tag{9}$$

Note that $\mathcal{R}_{\mathcal{F}}$ can be seen as an approximation of the risk function $\mathcal{R}_{\mathcal{M}}$, where for all $\delta \in \mathcal{D}$, $\mathcal{R}_{\mathcal{F}}(\delta) \leq \mathcal{R}_{\mathcal{M}}(\delta)$. Then, one key observation is that once we computed $\hat{\rho}_{\mathrm{VB}}$, we don't have to compute Z_δ because

$$-\mathcal{K}(\hat{\rho}_{\mathrm{VB}}(\delta), \tilde{\pi}_n) + Z_\delta = -\mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}[\log(\hat{\rho}_{\mathrm{VB}})] - \lambda \delta^\top \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}[Y_{n+1}] + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}[\log \pi_n] + C\,,$$

where C does not depend on δ, and hence (9) can be computed in closed-form. Since $\hat{\rho}_{\mathrm{VB}}$ depends on δ, optimizing with respect to δ requires to alternate Gradient-descent steps on (9) with adjustment steps (7) in the following way:

i) **Gradient-Descent step.** We perform one step of gradient descent with a constant step-size η,

$$\hat{\delta}^{(\mathrm{k}+1)} \leftarrow \hat{\delta}^{(\mathrm{k})} - \eta \nabla_\delta \mathcal{R}_{\mathcal{F}}(\hat{\delta}^{(\mathrm{k})}_{\mathrm{VB}})\,.$$

ii) **Adjustment step.**We recompute the variational distribution $\hat{\rho}_{\mathrm{VB}}$ solution to (5) for the decision $\hat{\delta}^{(\mathrm{k}+1)}$. The pseudo-code of our general method (denoted as `VB-Portfolio`) is shown in Algorithm 1. In the following section, we will introduce specific statistical models to which this algorithm can be applied.

Algorithm 1 `VB-Portfolio`: Portfolio Construction with Variational Bayes.

Input: Dataset H_n, Prior π_0 on θ, initial decision guess $\hat{\delta}^{(0)}$, decision space $\mathcal{D}$.
for $k = 1, \ldots,$ **do**
  while *Not converging* **do**
    $\hat{\rho}_{\mathrm{VB}} \leftarrow T(\hat{\rho}_{\mathrm{VB}})$ where T is the fixed-point operator defined in Theorem 14 for GW model, Theorem 16 for AR model, Theorem 18 for GP model.
  end
  $\hat{\delta}^{(k+1)} \leftarrow \mathrm{Proj}_{\mathcal{D}}\left(\hat{\delta}^{(k)} - \alpha_k \nabla_\delta \mathcal{R}_{\mathcal{F}}(\hat{\delta}^{(k)})\right)$, where $\mathcal{R}_{\mathcal{F}}$ is defined in Theorem 15 for GW model, Theorem 17 for AR model and Theorem 19 for GP model.
end
Return $\hat{\delta}^{(\infty)} = \hat{\delta}_{\mathrm{VB}}$.

Convexity properties of the objective function. An important property of the objective function (9) is that it enjoys remarkable properties such as *convexity* and *smoothness*. Therefore, applying Projected Gradient Descent on $\delta \mapsto \mathcal{R}_{\mathcal{F}}(\delta)$, where the projection set $\mathcal{D}$ is compact and convex ensures that the iterates $(\hat{\delta}^{(k)})_k$ will converge to an optimal point with value $\mathcal{R}_{\mathcal{F}}(\hat{\delta}_{\mathrm{VB}})$, that is, $\mathcal{R}_{\mathcal{F}}(\hat{\delta}^{(k)}) \to_k \mathcal{R}_{\mathcal{F}}(\hat{\delta}_{\mathrm{VB}})$ at rate $\mathcal{O}(1/k)$. We state formally these results in Theorem 3. These properties will also play a crucial role in establishing the statistical convergence of $\hat{\delta}_{\mathrm{VB}}$.

Statistical guarantees of Variational Bayes. The restriction of the variational formulation to a smaller set of measures introduces a bias in the resulting decisions. There is a growing literature studying the statistical properties of variational Bayes approximation (Alquier et al., 2016; Wang and Blei, 2019; Alquier and Ridgway, 2020; Yang et al., 2020; Ray and Szabó, 2022; Huix et al., 2024). Those results are not directly transferable to our problem because we do not only require the convergence of the approximate measure but the convergence with respect to the argmin of the objective function; we show that our VB algorithm converges asymptotically with respect to the sample size n (see Section 4).

4. Theoretical Guarantees

In this section, we assume that $\mathcal{D}$ is compact and convex, as is the case when $\mathcal{D} = \Delta_d$, the standard simplex. We introduce the notation $\hat{\delta}^{(k)}$ for the result of our algorithm after k iterations of Algorithm 1. Under the formal conditions stated below, we establish the following two key theoretical results. Complete proofs are given in Appendix C.

4.1. Numerical Convergence

We begin by addressing the convergence of our algorithm, which relies on the convexity and smoothness properties of the objective function $\mathcal{R}_{\mathcal{F}}$.

Proposition 3 (Properties of $\mathcal{R}_{\mathcal{F}}$) *For a fixed dataset size n, $\delta \mapsto \mathcal{R}_{\mathcal{F}}(\delta)$ is convex and smooth. Let L denote the smoothness constant of $\mathcal{R}_{\mathcal{F}}$. Using Gradient descent with a step size of $\eta = 1/L$ and initial point $\hat{\delta}^{(0)}$ ensures*

$$\mathcal{R}_{\mathcal{F}}(\hat{\delta}^{(k)}) - \mathcal{R}_{\mathcal{F}}(\hat{\delta}_{\mathrm{VB}}) \leq \frac{2L}{k-1} \|\hat{\delta}^{(0)} - \hat{\delta}_{\mathrm{VB}}\|_2 \,.$$

4.2. Statistical Consistency

Next, we establish asymptotic consistency results as the sample size $n \to +\infty$, focusing on the behaviour of the *decision* $\hat{\delta}_{\mathrm{VB}}$. For this, we introduce an assumption regarding the asymptotic behaviour of the fixed-point equation.

ASSUMPTION 1:
As the sample size $n \to +\infty$, the variational distribution converges pointwise,

$$\forall \delta \in \mathcal{D}\,, \quad \hat{\rho}_{\mathrm{VB}}(\delta) \xrightarrow{n \to +\infty} \hat{\rho}_{\infty}(\delta),$$

where $\hat{\rho}_{\infty}$ is solution to the asymptotic fixed-point operator: denoting T_n the fixed point operator defined in Theorem 2, for any $\delta \in \mathcal{D}$, if $T_n(\hat{\rho}_{\mathrm{VB}}(\delta)) = \hat{\rho}_{\mathrm{VB}}(\delta)$ then the operator $T_{\infty} = \lim_{n \to +\infty} T_n$ satisfies $T_{\infty}(\hat{\rho}_{\infty}(\delta)) = \hat{\rho}_{\infty}(\delta)$.

Next we derive asymptotic consistency of the variational decision $\hat{\delta}_{\mathrm{VB}}$ under this assumption.

Theorem 4 (Consistency of the variational decision) *Under both* GW (10) *and* AR (11) *models, the variational decision converges almost surely to the Markovitz decision in $\mathcal{D}$,*

$$\hat{\delta}_{\mathrm{VB}} \xrightarrow[n \to +\infty]{a.s.} \operatorname*{argmin}_{\delta \in \mathcal{D}} \left\{ \frac{1}{2} \delta^{\top} \hat{\Sigma}_{\infty}^{-1} \delta - \lambda \delta^{\top} \hat{\mu}_{\infty} \right\} .$$

where $\hat{\Sigma}_\infty$ and $\hat{\mu}_\infty$ are the empirical estimates of the covariance matrix and mean vector, respectively, in the limit as $n \to +\infty$.

5. Conclusion

We showed that Bayesian optimal decision for exponential utility can be interpreted as a saddle-point problem. We developed a computationally efficient algorithm based on variational Bayes with provable convergence guarantees, demonstrating its effectiveness in real-world portfolio optimization problems.

Maximizing exponential utility functions. Our min-max formulation (Theorem 1) provides a versatile framework for scenarios where the expected utility lacks a closed-form solution. This methodology not only bridges theoretical and practical domains but also holds promise for broader applications, particularly in areas like reinforcement learning (Marthe et al., 2024), where exponential utility functions are pivotal for navigating decision-making under uncertainty.

Beyond Gradient-Descent. Although our objective function is convex and smooth, leveraging advanced optimization techniques could unlock further potential. Techniques such as Nesterov's acceleration (Nesterov et al., 2018) and mirror-descent-based methods (Nemirovski, 2004) for saddle-point optimization present opportunities to enhance convergence rates and scalability. These methods could prove especially beneficial for portfolio construction in high-dimensional settings.

6. Acknowledgements

The authors would like to warmly thank Emmanuel Sérié for insightful discussions and remarks on this work. We also thank Vincent Fortuin and Pierre Laforgue for valuable feedback.

Nicolas Nguyen and Claire Vernade are funded by the Deutsche Forschungsgemeinschaft (DFG) under both the project 468806714 of the Emmy Noether Programme and under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

Nicolas Nguyen and Claire Vernade thank the international Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

Raj Agrawal, Uma Roy, and Caroline Uhler. Covariance matrix estimation under total positivity for portfolio selection. *Journal of Financial Econometrics*, 20(2):367–389, 2022.

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 2020.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1): 8374–8414, 2016.

Francis Bach. *Learning theory from first principles*. MIT press, 2024.

Florent Benaych-Georges, Jean-Philippe Bouchaud, and Marc Potters. Optimal cleaning for singular values of cross-covariance matrices. *The Annals of Applied Probability*, 33(2): 1295–1326, 2023.

Raphael Benichou, Yves Lempérière, Julien Kockelkoren, Philip Seager, Jean-Philippe Bouchaud, Marc Potters, et al. Agnostic risk parity: Taming known and unknown-unknowns. *arXiv preprint arXiv:1610.08818*, 2016.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

C Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:531–537, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer, 2002.

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

Zexun Chen, Bo Wang, and Alexander N Gorban. Multivariate gaussian and student-t process regression for multi-output prediction. *Neural Computing and Applications*, 32: 3005–3028, 2020.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, pages 721–741, 1984.

Hans U Gerber and Gérard Pafum. Utility functions: from risk theory to finance. *North American Actuarial Journal*, 2(3):74–91, 1998.

Campbell R Harvey, John C Liechty, Merrill W Liechty, and Peter Müller. Portfolio selection with higher moments. *Quantitative Finance*, 10(5):469–485, 2010.

Tom Huix, Anna Korba, Alain Durmus, and Eric Moulines. Theoretical guarantees for variational inference with fixed-variance mixture of gaussians. *arXiv preprint arXiv:2406.04012*, 2024.

Amine Ismail and Huyên Pham. Robust markowitz mean-variance portfolio selection under ambiguous covariance matrix. *Mathematical Finance*, 29(1):174–207, 2019.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5): 603–621, 2003.

Eric Luxenberg and Stephen Boyd. Portfolio construction with gaussian mixture returns and exponential utility via convex optimization. *Optimization and Engineering*, 25(1): 555–574, 2024.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Alexandre Marthe, Aurélien Garivier, and Claire Vernade. Beyond average return in markov decision processes. *Advances in Neural Information Processing Systems*, 36, 2024.

Robert C Merton. Lifetime portfolio selection under uncertainty: The continuous-time case. *The review of Economics and Statistics*, pages 247–257, 1969.

Robert C Merton. An analytic derivation of the efficient portfolio frontier. *Journal of financial and quantitative analysis*, 7(4):1851–1872, 1972.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Giorgio Parisi and Ramamurti Shankar. Statistical field theory. *American Journal of Physics*, 1988.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6: 1939–1959, 2005.

Jose Mario Quintana. *Multivariate Bayesian forecasting models*. PhD thesis, University of Warwick, 1987.

Kolyan Ray and Botond Szabó. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281, 2022.

Christian P Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

William F Sharpe. Mutual fund performance. *The Journal of business*, 39(1):119–138, 1966.

William F Sharpe. The sharpe ratio. *Journal of portfolio management*, 21(1):49–58, 1994.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

Appendix A. Application to Specific Statistical Models

So far, Algorithm 1 remained theoretical since we do not introduce assumptions on the statistical model P_{θ^*} yet, *i.e.* we derived a general algorithm that holds for any parametric statistical model P_{θ^*}. We now introduce relevant statistical models in the context of finance, where the core problem is to estimate the mean of investment returns, and the correlation between these returns. For all these models, we derive the corresponding fixed-point operator and the objective function $\mathcal{R}_{\mathcal{F}}$ in closed form in Appendix D.

A.1. Gaussian-Wishart (GW)

For this model, observations (returns) are assumed independent and Normally distributed with unknown mean μ and precision Λ; putting a Gaussian prior on the mean and a Wishart prior on the precision matrix,

$$\begin{aligned} &Y_t \,|\, \mu, \Lambda \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Lambda^{-1}) && \forall t \in \mathbb{N} \\ &\mu \sim \mathcal{N}(\mu_0, \Lambda_0^{-1})\,, \quad \Lambda \sim \mathcal{W}(\nu_0, \psi_0)\,. \end{aligned} \tag{10}$$

Since we put prior on both mean and covariance $\theta = (\mu, \Lambda)$, this model does not have closed-form moments for its *joint* posterior distribution $\pi(\mathrm{d}(\mu, \Lambda) \,|\, H_n)$, so we cannot compute the integral (4) directly. The full expression of the fixed-point operator in Theorem 2 for this model is derived in Theorem 14, along with the corresponding objective function $\mathcal{R}_{\mathcal{F}}$ detailed in Theorem 15.

A.2. Autoregressive Model (AR)

The non-dynamic model defined in Equation (10) is rather conservative, as it assumes no autocorrelation in returns, treating them as independent across time. While this simplifies learning and estimation, it overlooks the temporal dependencies often present in financial data such as market trends. Ignoring these patterns may limit the model's capacity to capture the true structure of returns. To circumvent this limitation, we introduce a model that incorporates a dynamic in the observations. We first outline a few definitions.

Definition 5 (Matrix normal distribution (Quintana, 1987)) *We say that $X \in \mathbb{R}^{d\times d}$ follows a matrix normal distribution with mean parameter $M \in \mathbb{R}^{d\times d}$, row-variance $U \in \mathbb{R}^{d\times d}$ and column variance $V \in \mathbb{R}^{d\times d}$ and denote $X \sim \mathcal{MN}(M, U, V)$ if and only if*

$$\mathrm{vec}(X) \sim \mathcal{N}(\mathrm{vec}(M), V \otimes U)\,,$$

where we define $\mathrm{vec}(A)$ as the concatenated vector in $\mathbb{R}^{mn}$ of a matrix $A \in \mathbb{R}^{m\times n}$, $\mathrm{vec}(A) = (A_1, \cdots A_n)$.

In this model, we arbitrarily[3] set the initial value $Y_0 \sim \delta_y$ with $y \in \mathbb{R}^d$. Then, the model writes

$$\begin{aligned} &Y_t \,|\, Y_{t-1}, \Gamma, \Lambda \sim \mathcal{N}(\Gamma Y_{t-1}, \Lambda) && \forall t \in \mathbb{N}^* \\ &\Gamma \sim \mathcal{MN}(M_0, U_0, V_0)\,, \quad \Lambda \sim \mathcal{W}(\nu_0, \psi_0)\,. \end{aligned} \tag{11}$$

The full expression of the fixed-point operator in Theorem 2 for this model is derived in Theorem 16, along with the corresponding objective function $\mathcal{R}_{\mathcal{F}}$ detailed in Theorem 17.

3. This first observation can be set thanks to previously collected data, which may be available in practice.

A.3. Gaussian Process Model (GP)

Gaussian processes (GPs; Williams and Rasmussen, 2006) can model the correlations between returns without assuming a specific functional form, making them particularly well-suited for environments with non-linear dependencies. We first define formally multivariate GPs.

Definition 6 (Multivariate Gaussian process (MGP) (Chen et al., 2020)) *f follows a multivariate Gaussian process with mean function $\mu : \mathbb{R} \to \mathbb{R}$, row variance function $k : \mathbb{R}^2 \to \mathbb{R}$ and column variance Ω, and we denote $f(\cdot) \sim \mathcal{MGP}(\mu(\cdot), k(\cdot,\cdot), \Omega)$, if, for every set of points $\{1, \dots, m\}$ with m any integer, we have $\left(f(t_1)^\top, \dots, f(t_m)^\top\right)^\top \sim \mathcal{MN}(M_0^m, \Sigma_0^m, \Omega)$, where $[M_0^m]_{ij} = \mu(t_i)_j$ and $[\Sigma_0^m]_{ij} = k(t_i, t_j)$.*

Putting a multivariate GP prior on the mean returns μ, the GP model is defined as

$$\begin{aligned} &Y_t \,|\, \mu(\cdot), \Lambda \sim \mathcal{N}(\mu(t), \Lambda^{-1}) \qquad \forall t \in \mathbb{N} \\ &\mu(\cdot) \sim \mathcal{MGP}(\mu_0(\cdot), K_0(\cdot), \Omega_0), \quad \Lambda \sim \mathcal{W}(\nu_0, \psi_0)\,, \end{aligned} \tag{12}$$

where we emphasise that at time step t, $\mu(t) \in \mathbb{R}^d$ (*i.e.* $(Y_t)_{t\geq 1}$ is a multivariate stochastic process). We will use specific kernel functions K_0 in numerical experiments. The full expression of the fixed-point operator in Theorem 2 for this model is derived in Theorem 18, and the corresponding objective function $\mathcal{R}_\mathcal{F}$ in Theorem 19.

Remark 7 (Computing the gradient $\nabla_\delta \mathcal{R}_\mathcal{F}$) *For the objective functions presented in Theorems 15, 17 and 19, we use automatic differentiation techniques to compute their gradients (Bradbury et al., 2018).*

Appendix B. Numerical Experiments

B.1. Experiments on Real-world Dataset

Dataset. We use financial indices associated with the G20 member countries, spanning the period from 2012 to 2024; these data are publicly available[4]. These indices are chosen over individual stock prices to minimize selection and survivorship biases. We apply Exponential Moving Averages (EMA) (Brockwell and Davis, 2002) with 8 different scales to each index and compute the corresponding EMA for all indices. The EMA-transformed signals are then aggregated by averaging across scales, producing dataset with $d = 8$ experts, where each column represents the averaged EMA signal at a specific scale, capturing smoothed trends across the indices. Monthly observations are extracted from this transformed dataset, yielding three different settings of increasing sample sizes: $(n, d) = (12, 8)$ **(Setting 1)**, $(n, d) = (48, 8)$ **(Setting 2)**, and $(n, d) = (84, 8)$ **(Setting 3)**.

Baselines. We compare our algorithm, `VB-Portfolio`, instantiated with models (10), (11) and (12) against several baseline portfolios. The first is the Equal Weights portfolio (**EW**, also called the $1/d$ portfolio), which assigns uniform weights to all assets: $\hat{\delta}_{\text{EW}} = \frac{1}{d} 1_d$. We also consider the Markowitz portfolio (**Mwz**), as described in Equation (1), $\hat{\delta}_{\text{Mwz}} = \frac{1}{\lambda} \hat{\Sigma}_n^{-1} \hat{\mu}_n$. Additionally, a more refined approach is to regularize the covariance matrix estimate, which

4. https://finance.yahoo.com/markets/world-indices/

is particularly advantageous in data-poor regimes. The Ledoit-Wolf method (Ledoit and Wolf, 2003) employs a *shrinkage* technique to stabilize the sample covariance matrix by combining it with a structured target matrix, typically a scaled identity matrix. The resulting shrunk covariance matrix is defined as $\hat{\Sigma}_n^{\mathrm{LW}} = (1-\alpha)\hat{\Sigma}_n + \alpha I_d$, where α is the shrinkage intensity. Notably, the optimal value of α can be explicitly computed, as derived in Ledoit and Wolf (2003). This adjustment balances bias and variance, resulting in a better-conditioned estimator for high-dimensional settings. We define the corresponding Markowitz-based portfolio as $\hat{\delta}_{\mathrm{Mtz}}^{\mathrm{LW}} = \frac{1}{\lambda}\left(\hat{\Sigma}_n^{\mathrm{LW}}\right)^{-1}\hat{\mu}_n$. Finally, we compare our approach against a state-of-the-art method, the Exponential Utility for Gaussian Mixtures (**EGM**), which maximizes the exponential utility under a Gaussian mixture model assumption for returns (Luxenberg and Boyd, 2024).

Prior parameters. For all models, the Wishart prior is set as $\nu_0 = d$ and $\psi_0 = \frac{1}{\nu_0}\hat{\Sigma}_n^{-1}$, where $\hat{\Sigma}_n$ is the empirical estimate of covariance matrix. For the GW model, we set the prior mean as the empirical mean, $\mu_0 = \hat{\mu}_n$ and $\Sigma_0 = I_d$. For the AR model, we set M_0 as the MLE estimate of the transition matrix Γ obtained via linear regression on the observations H_n, $M_0 = \left(\sum_{t=1}^n Y_t Y_{t-1}^\top\right)\left(\sum_{t=1}^n Y_t Y_t^\top\right)^{-1}$. We set the row-covariance U_0 and column-covariance V_0 as identities. For the GP model (12), we choose a Radial basis function kernel parameterized by γ *i.e.* $\forall t_1, t_2, \quad k_\gamma(t_1, t_2) = \exp\left(\frac{(t_1-t_2)^2}{2\gamma^2}\right)$, and tune γ trough Gradient-based optimization (with respect to marginal likelihood). We set the mean function to 0 and the prior column variance as $\Omega_0 = I_d$. The hyperparameter choices are discussed in Appendix F.

Results (Cumulative wealth and regret). For each allocation strategy, we compute the out-of-sample cumulative wealth, $\delta^\top \sum_{y\in\mathcal{E}_{\mathrm{test}}} y$ ($\mathcal{E}_{\mathrm{test}}$ are observations of the testing set, *i.e.* observations from 2013 for Setting 1, from 2016 for Setting 2 and from 2018 for Setting 3). To enable a fair comparison across strategies with varying levels of risk, we rescale each cumulative wealth by its standard deviation. This risk-adjusted rescaling is a standard convention in portfolio construction literature. Additionally, we plot the strategy corresponding to allocating all mass on the *best index in hindsight*, defined as the index with the highest cumulative wealth at the end of the testing horizon. To further assess performance, we plot the *cumulative regret* for each strategy against this best index in hindsight, that is, the difference between the cumulative wealth of the best index in hindsight and the one of the given strategy. We rescale this cumulative difference by its standard deviation. Results are displayed in Figure 1, and show that while `VB-Portfolio`(GW) exhibits performance comparable to the Markowitz-based portfolio, `VB-Portfolio`, when instantiated with both the AR and GP models, outperforms the other strategies overall. This superior performance suggests that these models are particularly effective at adapting to evolving market conditions.

Sharpe Ratios Comparison. The Sharpe ratio (Sharpe, 1966, 1994) is a widely used metric for assessing the risk-adjusted performance of investment strategies. It is defined as the ratio of the mean return to the standard deviation, quantifying the return per unit of risk. Let $\hat{\mu}_{\mathrm{test}}(\delta)$ denotes the mean return of strategy δ over the testing set, and $\hat{\sigma}_{\mathrm{test}}(\delta)$ its standard deviation. The annualized Sharpe ratio is then computed as $\mathrm{SR}(\delta) = \sqrt{12}\hat{\mu}_{\mathrm{test}}(\delta)/\hat{\sigma}_{\mathrm{test}}(\delta)$.

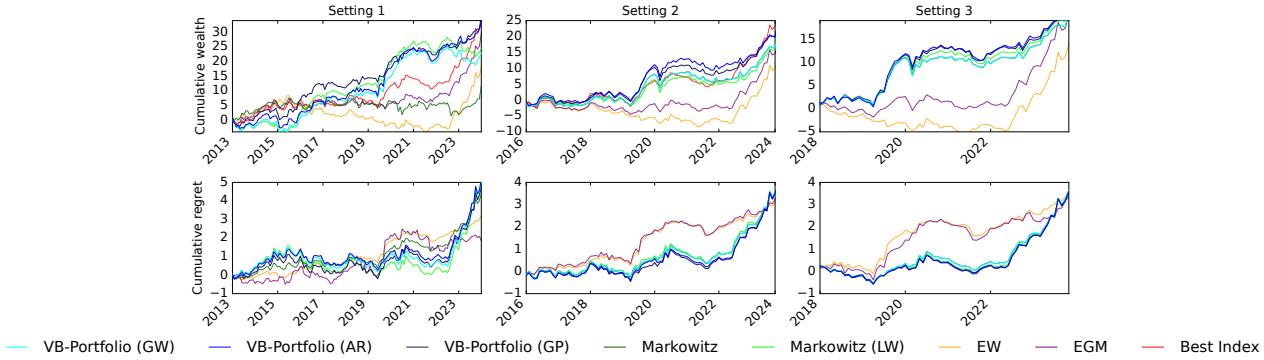

Figure 1: Cumulative wealth (row 1) and cumulative regret with respect to best index in hindsight (row 2) in 3 settings. For each strategy, we rescale the cumulative plot by the standard deviation of the returns.

Table 1: Out-of-sample annualized Sharpe ratios of each portfolio for different settings.

Allocation Strategy	**Annualized Sharpe Ratio**		
	Setting 1	Setting 2	Setting 3
`algVB (GW)`	0.59	0.77	1.03
`algVB (AR)`	0.88	**0.90**	**1.16**
`algVB (GP)`	**0.90**	**0.90**	1.12
`Markowitz`	0.31	0.77	1.03
`Markowitz (LW)`	0.63	0.77	1.11
`Equal Weights`	0.52	0.52	0.74
`EGM`	0.80	0.80	1.05

This metric facilitates meaningful comparisons across strategies by highlighting those that deliver higher returns relative to risk. As illustrated in Table 1, `VB-Portfolio`(AR) achieves the highest Sharpe ratio overall, closely followed by `VB-Portfolio`(GP). Both methods consistently outperform traditional approaches, such as the Markowitz portfolio, particularly in Settings 1 and 2, where the smaller sample sizes lead to less stable estimates for the other strategies. These findings underscore the robustness of the proposed methods in data-poor environments, demonstrating that Bayesian approaches are well-suited to regularizing estimates and mitigating the impact of limited training data.

B.2. Numerical Consistency

To assess the consistency of our approximation $\hat{\delta}_{\mathrm{VB}}$ to δ^*, we use a gradient descent-based algorithm that leverages Markov Chain Monte Carlo (MCMC) sampling to estimate the gradient of the objective function.

Approximating the objective function. First, we want to approximate the integral (4) for any $\delta \in \mathcal{D}$;

$$\mathcal{R}_{\mathcal{M}}(\delta) = \int_{\mathcal{Y}} e^{-\lambda \delta^\top Y_{n+1}} \pi(\mathrm{d}Y_{n+1} \,|\, H_n) \approx \frac{1}{M} \sum_{m=1}^{M} e^{-\lambda \delta^\top y^{(k)}} \,,$$

where $(y^{(k)})_{k\in[M]}$ are M samples from the predictive posterior distribution $\pi(\cdot \,|\, H_n)$. This can be done with the Gibbs sampling algorithm (Geman and Geman, 1984): in fact, by remarking that

$$\pi(\mathrm{d}Y_{n+1} \,|\, H_n) = \int_\Theta \pi(\mathrm{d}Y_{n+1} \,|\, \theta)\pi(\mathrm{d}\theta \,|\, H_n)\,,$$

and since we now how to sample from the *conditional posterior* $\pi(\mathrm{d}\theta_i \,|\, \boldsymbol{\theta}_{-i}, H_n)$, we can generate M samples $(\theta^{(m)})_{m\leq M}$ from the joint posterior distribution $\pi(\mathrm{d}\theta \,|\, H_n)$. From this sequence, we can now sample from the distribution $\breve{\pi}(\mathrm{d}Y_{n+1}) \propto e^{-\lambda\delta^\top Y_{n+1}}\pi(\mathrm{d}Y_{n+1} \mid H_n)$ *conditionally* on one sample $\theta^{(k)}$, by drawing a sample $y^{(k)}$ from the distribution

$$\breve{\pi}_k(\mathrm{d}Y_{n+1}) \propto e^{-\lambda^\top Y_{n+1}}\pi(\mathrm{d}Y_{n+1}; \theta^{(k)})\,,$$

resulting in a sequence of M samples $(y^{(k)})_{k\in[M]}$.

Approximating the gradient $\nabla_\delta \mathcal{R}$. By Leibniz rule, we have

$$\nabla_\delta \mathcal{R}_{\mathcal{M}}(\delta) = -\lambda \int_{\mathcal{Y}} Y_{n+1} \frac{e^{-\lambda\delta^\top Y_{n+1}}\pi(\mathrm{d}Y_{n+1} \mid H_n)}{\int_{\mathcal{Y}} e^{-\lambda\delta^\top Y_{n+1}}\pi(\mathrm{d}Y_{n+1} \mid H_n)} = -\lambda \mathbb{E}_{\breve{\pi}}\left[Y_{n+1}\right]\,,$$

for which we can approximate by

$$\nabla_\delta \mathcal{R}_{\mathcal{M}}(\delta) \approx -\lambda \frac{1}{M}\sum_{k=1}^{M} z^{(k)}\,, \qquad \text{where } z^{(k)} \sim \breve{\pi}_k(\cdot)\,.$$

The pseudo-code of this MCMC algorithm is shown in Algorithm 2. We instantiate this algorithm for GW and AR model in Appendix E, where we provide in particular the expressions of the conditional posteriors and $\breve{\pi}$.

Algorithm 2 `MCMC-Portfolio`: Portfolio Construction with Markov Chain Monte-Carlo.

Input: Dataset H_n, initial decision $\hat{\delta}^{(0)}$, number of Monte-Carlo samples M, risk parameter λ, step-size η.

while *Not converging* **do**

  Get M samples $(\theta^{(k)})_{k\in[M]}$ from Gibbs sampler.

  For all $k \in [M]$, sample $z^{(k)} \sim \breve{\pi}_k$.

  $\hat{\delta}^{(k+1)} \leftarrow \mathrm{Proj}_{\mathcal{D}}\left(\hat{\delta}^{(k)} + \eta\lambda\frac{1}{M}\sum_{k\in[M]} z^{(k)}\right)$

end

Return $\hat{\delta}^{(\infty)} = \hat{\delta}^{\mathrm{MCMC}}$.

Consistency with synthetic data. To evaluate the numerical consistency of our method, we show that $\mathbb{E}\left[\|\hat{\delta}^{\mathrm{MCMC}} - \hat{\delta}_{\mathrm{VB}}\|_2\right]$ converges to 0 as the sample size n grows. We generate synthetic datasets for both GW model and AR model with $d = 30$. For the GW model, we randomly set each component of the true mean return, μ_i^*, according to a uniform distribution, $\mu_i^* \sim \mathcal{U}([0,1])$, and use an identity covariance matrix, $\Sigma^* = I_d$. For the AR

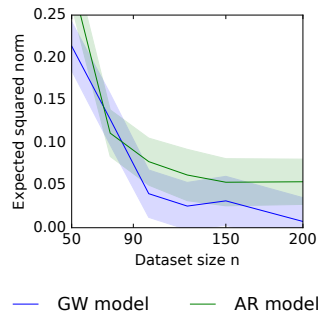

Figure 2: Expected 2-norm $\mathbb{E}\left[\|\hat{\delta}^{\mathrm{MCMC}} - \hat{\delta}_{\mathrm{VB}}\|_2\right]$ as a function of dataset size n for both the GW and AR(1) models. Each point represents the average over 50 iterations, with error bars indicating one standard deviation to represent the confidence interval. The dimensionality of the data is fixed at $d = 30$.

model, the true transition matrix Γ^* is set to a diagonal matrix with values evenly spaced from 0.6 to 0.99, while the covariance matrix is $\Sigma^* = 0.1 I_d$. We generate datasets of varying sizes, with n ranging from 50 to 200. For the GW model, we simulate data according to (10), and for the AR model, we follow (11). For each dataset, we compute the decision vector using both `VB-Portfolio` and `MCMC-Portfolio`, where the Gibbs sampler is run for $M = 20{,}000$ iterations. We compute $\mathbb{E}\left[\|\hat{\delta}^{\mathrm{MCMC}} - \hat{\delta}_{\mathrm{VB}}\|_2\right]$, where the expectation is taken over 50 repetitions with newly generated datasets.

Figure 2 illustrates the relationship between this expected norm difference and the dataset size n. As n increases, we observe that the difference between the two decision methods diminishes, confirming the numerical consistency of `VB-Portfolio` with respect to `MCMC-Portfolio` as the sample size grows.

Appendix C. Complete Proofs

C.1. Proof of Theorem 1

Proof (Theorem 1) By rewriting (4),

$$\delta^* = \operatorname*{argmin}_{\delta \in \mathcal{D}} \int_{\mathbb{R}^d} e^{-\lambda \delta^\top Y_{n+1}} \pi(\mathrm{d}Y_{n+1} \,|\, H_n) \overset{\text{(i)}}{=} \operatorname*{argmin}_{\delta \in \mathcal{D}} \log \int_{\mathbb{R}^d} e^{-\lambda \delta^\top Y_{n+1}} \pi(\mathrm{d}Y_{n+1} \,|\, H_n)$$
$$\overset{\text{(ii)}}{=} \operatorname*{argmin}_{\delta \in \mathcal{D}} \log \int_{\mathbb{R}^d \times \Theta} e^{-\lambda \delta^\top Y_{n+1}} \underbrace{\pi(\mathrm{d}(Y_{n+1}, \theta) \,|\, H_n)}_{:= \pi_n((\mathrm{d}(Y_{n+1}, \theta))},$$

where in (i) we took the log in front of the objective function, and in (ii) we marginalized out θ conditionally on H_n (since $\int_\Theta \pi(\mathrm{d}(Y_{n+1}, , \theta) \,|\, H_n) = \pi(\mathrm{d}Y_{n+1} \,|\, H_n)$). Applying the Donsker-and-Varadhan change-of-measure (Donsker and Varadhan, 1983) on $h : x \mapsto -\lambda \delta^\top x$ gives

$$\delta^* = \operatorname*{argmin}_{\delta \in \mathcal{D}} \sup_{\rho \in \mathcal{M}(\mathbb{R}^d \times \Theta)} \left\{ -\lambda \delta^\top \mathbb{E}_\rho \left[Y_{n+1} \right] - \mathcal{K}(\rho, \pi_n) \right\}. \tag{13}$$

We next have to show that the expression inside the supremum can be expressed as a KL divergence (up to an additive constant that does not depend on ρ); we observe that for any $\delta \in \mathcal{D}$,

$$-\lambda\delta^\top \mathbb{E}_\rho\left[Y_{n+1}\right] = -\int_{\mathbb{R}^d\times\Theta} \log \frac{1}{e^{-\lambda\delta^\top Y_{n+1}}} \rho(\mathrm{d}(Y_{n+1},\theta))$$

and therefore, by introducing the probability measure $\tilde{\pi}_n$ defined as $\mathrm{d}\tilde{\pi}_n = \frac{e^{-\lambda\delta^\top Y_{n+1}}}{\mathbb{E}_{\pi_n}\left[e^{-\lambda\delta^\top Y_{n+1}}\right]}\mathrm{d}\pi_n$, we have

$$-\lambda\delta^\top \mathbb{E}_\rho\left[Y_{n+1}\right] - \mathcal{K}(\rho,\pi_n) = -\int_{\mathbb{R}^d\times\Theta} \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\tilde{\pi}_n}(Y_{n+1},\theta)\right)\rho(\mathrm{d}(Y_{n+1},\theta)) + Z_\delta = -\mathcal{K}(\rho,\tilde{\pi}_n) + Z_\delta\,. \tag{14}$$

Combining (13) with (14), we can rewrite the Bayes optimal decision as

$$\delta^* = \operatorname*{argmin}_{\delta\in\mathcal{D}} \sup_{\rho\in\mathcal{M}(\mathbb{R}^d\times\Theta)} \left\{-\mathcal{K}(\rho,\tilde{\pi}_n) + Z_\delta\right\}\,,$$

where the supremum is indeed achieved for $\rho = \tilde{\pi}_n$. ■

C.2. Auxiliary Lemmas

Lemma 8 *The mean-field space $\mathcal{F}(\mathbb{R}^d\times\Theta)$ is closed in the space of probability measures $\mathcal{M}$.*

Proof Consider any sequence $(\rho^i)_{i\in\mathbb{N}}$ in $\mathcal{F}(\mathbb{R}^d\times\Theta)$ that has a limit in $\mathcal{M}(\mathbb{R}^d\times\Theta)$. Then for any $i\in\mathbb{N}$, ρ^i can be factorized as

$$\rho^i(\mathrm{d}(y,\theta)) = \rho^i_y(\mathrm{d}y)\prod_{k=1}^K \rho^i_k(\mathrm{d}\theta_k)\,,$$

where we recall that we assume that Θ factorizes as a product of K subspaces, $\Theta = \prod_{k=1}^K \Theta_k$. Then limit of $(\rho^i)_i$ exists by construction, and

$$\lim_{i\to+\infty}\rho^i(\mathrm{d}(y,\theta)) = \lim_{i\to+\infty}\left(\rho^i_y(\mathrm{d}y)\prod_{k=1}^K\rho^i_k(\mathrm{d}\theta_k)\right) = \lim_{i\to+\infty}\rho^i_y(\mathrm{d}y)\prod_{k=1}^K\lim_{i\to+\infty}\rho^i_k(\mathrm{d}\theta_k) \in \mathcal{F}(\mathbb{R}\times\Theta)\,,$$

which proves that the limit of any convergent sequence in $\mathcal{F}(\mathbb{R}^d\times\Theta)$ has its limit in $\mathcal{F}(\mathbb{R}^d\times\Theta)$, and hence $\mathcal{F}(\mathbb{R}\times\Theta)$ is closed in $\mathcal{M}$. ■

C.3. Proof of Theorem 3

For any $\delta \in \mathcal{D}$ and any measurable h in $\mathbb{R}^d \times \Theta$, $h \mapsto g(h) = \sup_{\rho \in \mathcal{M}(\mathcal{Y}\times\Theta)}(\langle h, \rho\rangle - \mathcal{K}(\rho, \pi))$ is a convex map since it is defined as the Fenchel-Legendre transformation of $\rho \mapsto \mathcal{K}(\rho, \pi_n)$, with $\rho \in \mathcal{M}(\mathcal{Y} \times \Theta)$ (and the space of probability measures $\mathcal{M}(\mathcal{Y} \times \Theta)$ is a convex set). Thus, we also have that the map $h \mapsto \tilde{g}(h) = \sup_{\rho \in \mathcal{F}(\mathcal{Y}\times\Theta)}(\langle h, \rho\rangle - \mathcal{K}(\rho, \pi))$ is convex since for any probability measure $\rho \in \mathcal{M}(\mathcal{Y} \times \Theta)$, $h \mapsto \langle h, \rho\rangle - \mathcal{K}(\rho, \pi)$ is convex (as a linear map) and $\tilde{g}$ is the *pointwise supremum* of a family of convex function (the supremum conserves convexity (Boyd and Vandenberghe, 2004)). Taking h as $h_\delta(y) = -\lambda\delta^\top y$ and remarking that it is a convex function with respect to δ for a given $\lambda > 0$ shows that $\mathcal{R}_\mathcal{F}$ is convex: indeed, by composition of convex functions, $\tilde{g}(h_\delta(y))$ is convex with respect to δ, and so $\mathcal{R}_\mathcal{F}(\delta) = \tilde{g}(h_\delta(y))$. Moreover, $\rho \mapsto \mathcal{K}(\rho, \pi_n)$ is *strongly convex* on the space of probability measures $\mathcal{M}(\mathbb{R}^d \times \Theta)$, and hence g is *smooth* (as a convex conjugate of a strongly convex function). Hence, $\mathcal{R}_\mathcal{F}$ is also smooth by composition (with the same arguments as above for the convexity).

The convergence rate mentioned follows directly from the classical results of gradient descent applied to convex, smooth functions (see, for instance, Bach (2024, Chapter 5)).

C.4. Proof of Theorem 4

The first step involves interchanging the limit and the argmin over $\delta \in \mathcal{D}$, allowing us to express it as:

$$\lim_{n\to+\infty} \operatorname*{argmin}_{\delta\in\mathcal{D}} \mathcal{R}_\mathcal{F}(\delta) \triangleq \operatorname*{argmin}_{\delta\in\mathcal{D}} \lim_{n\to+\infty} \mathcal{R}_\mathcal{F}(\delta)\,.$$

This step $\triangle$ is non-trivial because the argmin function is a *set*, necessitating the use of general regularity conditions. Additionally, $\mathcal{R}_\mathcal{F}$ is defined implicitly as a supremum over a space of measures. The second step involves analyzing $\lim_{n\to+\infty} \mathcal{R}_\mathcal{F}(\delta)$, for which we know how to proceed based on the statistical model introduced in Appendix A.

C.4.1. Inverting Limit and Argmin

We rely on Rockafellar and Wets (2009, Theorem 7.33) for this purpose; this strong result requires to show the two following conditions:

C.1. $(\mathcal{R}_\mathcal{F})_n$ epi-converges to $\mathcal{R}^*_\mathcal{F}$, where $\mathcal{R}^*_\mathcal{F}$ is lower-semi-continuous and proper.

C.2. $(\mathcal{R}_\mathcal{F})_n$ is a lower-semi-continuous and proper sequence.

To show that $(\mathcal{R}_\mathcal{F})_n$ epi-converges, we first establish some general regularity properties of this sequence, from which the epi-convergence will naturally follow. For any $\lambda > 0$ and $n \in \mathbb{N}$, we introduce the functional

$$f_n(\delta, \rho) = -\lambda\delta^\top \mathbb{E}_\rho\left[Y_{n+1}\right] - \mathcal{K}(\rho, \pi_n)\,. \tag{15}$$

$(\mathcal{R}_\mathcal{F})_n$ is a uniformly continuous sequence. Since Θ is a Polish space, $\mathbb{R}^d \times \Theta$ is also a Polish space. According to Billingsley (2013, Th. 1.3), $\mathcal{M}(\mathbb{R}^d \times \Theta)$ is a space of *tight* measures. By Prokhorov's theorem, $\mathcal{M}(\mathbb{R}^d \times \Theta)$ is *relatively compact* in the weak-*

topology. Given that $\mathcal{F}(\mathbb{R}^d \times \Theta)$ is closed in $\mathcal{F}$ (Theorem 8), it follows that $\mathcal{F}(\mathbb{R}^d \times \Theta)$ is also compact as a closed subset of a relatively compact space. By the *Maximum theorem*, $\mathcal{R}_{\mathcal{F}}$ is continuous on $\mathcal{D}$. Furthermore, for any $\rho \in \mathcal{F}$, the function $f_n(\delta, \rho)$, as defined in (15), is linear with respect to δ (since the KL term is independent of δ), making f_n *uniformly continuous* in δ. Consequently, since $\mathcal{R}_{\mathcal{F}}$ is continuous on a compact set and f_n is uniformly continuous with respect to δ, we have that for all $\rho \in \mathcal{F}$ and for any $\varepsilon > 0$, there exists a $\gamma > 0$ such that

$$\|\delta_1 - \delta_2\| < \gamma \implies |f_n(\delta_1, \rho) - f_n(\delta_2, \rho)| \leq \epsilon\,,$$

and therefore,

$$\forall \varepsilon > 0,\, \exists \gamma > 0,\, \|\delta_1 - \delta_2\| < \gamma \implies |\mathcal{R}_{\mathcal{F}}(\delta_1) - \mathcal{R}_{\mathcal{F}}(\delta_2)| \leq \sup_{\rho \in \mathcal{F}} |f_n(\delta_1, \rho) - f_n(\delta_2, \rho)| \leq \epsilon$$

independently in n, which is the definition of uniform continuity of $\mathcal{R}_{\mathcal{F}}$.

$(\mathcal{R}_{\mathcal{F}})_n$ is epi-convergent sequence. Since $\mathcal{R}_{\mathcal{F}}$ is uniformly continuous, smooth, and convex on a compact domain for any $n \in \mathbb{N}$ (Theorem 3), it follows that $\mathcal{R}_{\mathcal{F}}$ is uniformly bounded on this compact space, implying that $(\mathcal{R}_{\mathcal{F}})_n$ is *equicontinuous*. Additionally, $(\mathcal{R}_{\mathcal{F}})_n$ converges pointwise to a limit $\mathcal{R}^*_{\mathcal{F}}$, as established by Assumption 1. By the Arzelà-Ascoli theorem, the equicontinuous sequence $(\mathcal{R}_{\mathcal{F}})_n$ converges uniformly to $\mathcal{R}^*_{\mathcal{F}}$. According to Rockafellar and Wets (2009, Theorem 7.11), $(\mathcal{R}_{\mathcal{F}})_n$ epi-converges if and only if it converges continuously. Since uniform convergence implies continuous convergence, we conclude that $(\mathcal{R}_{\mathcal{F}})_n$ is *epi-convergent* to $\mathcal{R}^*_{\mathcal{F}}$, thereby verifying **C.1**.

Lower semi-continuity and proper conditions. Since $(\mathcal{R}_{\mathcal{F}})_n$ is continuous, it is also lower semi-continuous. Furthermore, because $(\mathcal{R}_{\mathcal{F}})_n$ converges continuously, $\mathcal{R}^*_{\mathcal{F}}$ is continuous and thus lower semi-continuous as well. To show that any preimage of a set $I \subset \mathbb{R}$ (e.g., a closed interval) is compact, note that since $\mathcal{R}_{\mathcal{F}}$ is continuous, $\mathcal{R}^{-1}_{\mathcal{F}}(I)$ is a closed subset of $\mathcal{D}$. Given that $\mathcal{D}$ is compact, $\mathcal{R}^{-1}_{\mathcal{F}}(I)$ is a closed subset of a compact space, and hence compact. The same reasoning applies to $\mathcal{R}^*_{\mathcal{F}}$ due to continuous convergence. Therefore, we have verified **C.2**.

C.4.2. Asymptotic Objective Function

We now turn our attention to computing $\lim_{n\to+\infty} \mathcal{R}_{\mathcal{F}} = \mathcal{R}^*_{\mathcal{F}}$. Since $\hat{\rho}_{\text{VB}}$ converges to a distribution $\hat{\rho}_\infty$, where $\hat{\rho}\infty$ is the fixed point of T_∞ (Assumption 1), we can explicitly compute $\mathcal{R}^*_{\mathcal{F}}$ using the known form of $\mathcal{R}_{\mathcal{F}}$ as a function of n for the statistical models under consideration. The following lemma formalizes this result.

Lemma 9 *For both* GW (10) *and* AR (11) *models, we have*

$$\forall \delta \in \mathcal{D}\,, \quad \mathcal{R}^*_{\mathcal{F}}(\delta) = \frac{1}{2}(\lambda\delta)^\top \hat{\Sigma}^{-1}_\infty (\lambda\delta) - \lambda\delta^\top \hat{\mu}_\infty\,,$$

where for the GW model we have

$$\hat{\mu}_\infty = \lim_{n\to+\infty} \left(\frac{1}{n} \sum_{t=1}^{n} Y_t \right)\,, \qquad \hat{\Sigma}_\infty = \lim_{n\to+\infty} \left(\frac{1}{n} \sum_{t=1}^{n} (Y_t - \hat{\mu}_\infty)(Y_t - \hat{\mu}_\infty)^\top \right)\,,$$

and for the AR model

$$\hat{\mu}_\infty = \lim_{n\to+\infty} \left(\sum_{t=1}^n Y_t Y_{t-1}\right)\left(\sum_{t=1}^n Y_t Y_t\right)^{-1} Y_\infty\,, \quad \hat{\Sigma}_\infty = \lim_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^n (Y_t-\hat{\mu}_\infty)\,(Y_t-\hat{\mu}_\infty)^\top\right)\,,$$

which is the corresponding sample mean estimate for the GW and the AR model respectively, and the corresponding sample covariance estimate. Hence, the asymptotic variational decision writes

$$\lim_{n\to+\infty} \hat{\delta}_{\mathrm{VB}} = \operatorname*{argmin}_{\delta\in\mathcal{D}} \left\{\frac{1}{2}(\lambda\delta)^\top \hat{\Sigma}_\infty^{-1}(\lambda\delta) - \lambda\delta^\top \hat{\mu}_\infty\right\}\,,$$

i.e. the Markowitz decision in $\mathcal{D}$.

We now prove Theorem 9 for both GW and AR model.

Proof [Proof of Theorem 9 for GW model] Starting from Theorem 14, the asymptotic operator T_∞ is the limit of the operator T_n defined in Theorem 14 when $n \to +\infty$ (by Assumption 1), giving

$$T_\infty : (\xi_y, \Lambda_y, \xi_\mu, \Lambda_\mu, \psi_\Lambda) \mapsto \begin{pmatrix} \xi_\mu - \lambda(\nu\psi_\Lambda)^{-1}\delta \\ \nu\psi_\Lambda \\ \lim_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^n Y_t\right) \\ n\nu\psi_\Lambda \\ \lim_{n\to+\infty}\left(n\left(\Lambda_\mu^{-1}+\xi_\mu\xi_\mu^\top\right) + \sum_{t=1}^n Y_t Y_t^\top - 2\sum_{t=1}^n Y_t\xi_\mu^\top\right)^{-1} \end{pmatrix}.$$

Thanks to Assumption 1, the fixed point of T_∞ denoted by $(\xi_y^\infty, \Lambda_y^\infty, \xi_\mu, \Lambda_\mu^\infty, \psi_\Lambda^\infty)$ satisfies

$$\begin{cases} \xi_y^\infty = \hat{\mu}_\infty - \lambda\hat{\Sigma}_\infty\delta \\ \Lambda_y^\infty = \hat{\Lambda}_\infty \\ \xi_\mu^\infty = \hat{\mu}_\infty \\ \Lambda_\mu^\infty = \lim_{n\to+\infty} n\hat{\Lambda}_\infty \\ \psi_\Lambda^\infty = \left(\sum_{s=1}^n (Y_t-\hat{\mu}_\infty)(Y_t-\hat{\mu}_\infty)^\top\right)^{-1}\,, \end{cases}$$

where

$$\hat{\mu}_\infty = \lim_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^n Y_t\right) \quad \hat{\Sigma}_\infty = \lim_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^n (Y_t-\hat{\mu}_\infty)\,(Y_t-\hat{\mu}_\infty)^\top\right)\,, \quad \hat{\Lambda}_\infty = \hat{\Sigma}_\infty^{-1}\,.$$

Plugging this solution the objective function in Theorem 15 and keeping only terms depending on δ, the asymptotic objective function writes

$$\mathcal{R}^*_{\mathcal{F}}(\delta) = -\lambda\delta^\top\left(\hat{\mu}_\infty - \lambda\hat{\Sigma}_\infty\delta\right)\,.$$

■

Proof (Proof of Theorem 9 for AR model) Starting from Theorem 16, the asymptotic operator T_∞ writes

$$T_\infty : \left(\xi_y, \Lambda_y, M_\Gamma, V_\Gamma \otimes U_\Gamma, \psi_\Lambda^{-1}\right) \mapsto \begin{pmatrix} M_\Gamma Y_\infty - \lambda(\nu\psi_\Lambda)^{-1}\delta \\ \nu_\Lambda \psi_\Lambda \\ \lim_{n\to+\infty} \left(\sum_{t=1}^n Y_t Y_{t-1}\right)\left(\sum_{t=1}^n Y_t Y_t\right)^{-1} \\ \lim_{n\to+\infty} \left(\sum_{t=1}^n Y_t Y_t^\top \otimes \nu_\Lambda \psi_\Lambda\right)^{-1} \\ \lim_{n\to+\infty} \left(\sum_{t=0}^n Y_t Y_t^\top - 2\left(\sum_{t=1}^n Y_{t-1} Y_t^\top\right)^\top \left(\sum_{t=1}^n Y_t Y_t^\top\right)^{-1} \left(\sum_{t=1}^n Y_{t-1} Y_t^\top\right)\right) \end{pmatrix},$$

where Y_∞ denotes the last observation of the dataset.

Thanks to Assumption 1, the fixed point of T_∞ denoted by $(\xi_\infty, \Lambda_\infty, M_\Gamma^\infty, (V_\Gamma \otimes U_\Gamma)^\infty, \psi_\Lambda^\infty)$ satisfy

$$\begin{cases} \xi_y^\infty = \hat{\mu} - \lambda \hat{\Sigma}_\infty \delta \\ \Lambda_y^\infty = \hat{\Lambda}_\infty \\ M_\Gamma^\infty = \lim_{n\to+\infty} \left(\sum_{t=1}^n Y_t Y_{t-1}\right)\left(\sum_{t=1}^n Y_t Y_t\right)^{-1} \\ (V_\Gamma \otimes U_\Gamma)^\infty = \lim_{n\to+\infty} \left(\sum_{t=1}^n Y_t Y_t^\top \otimes \left(\sum_{t=0}^n Y_t Y_t^\top - 2\frac{1}{n}\left(\sum_{t=1}^n Y_{t-1} Y_t^\top\right)^\top \left(\sum_{t=1}^n Y_t Y_t^\top\right)\left(\sum_{t=1}^n Y_{t-1} Y_t^\top\right)\right)^{-1}\right)^{-1} \\ \psi_\Lambda^\infty = \lim_{n\to+\infty} \left(\frac{1}{n}\left(\sum_{t=0}^n Y_t Y_t^\top - 2\frac{1}{n}\left(\sum_{t=1}^n Y_{t-1} Y_t^\top\right)^\top \left(\sum_{t=1}^n Y_t Y_t^\top\right)\left(\sum_{t=1}^n Y_{t-1} Y_t^\top\right)\right)^{-1}\right), \end{cases}$$

where

$$\hat{\mu}_\infty = \lim_{n\to+\infty} \left(\sum_{t=1}^n Y_t Y_{t-1}\right)\left(\sum_{t=1}^n Y_t Y_t\right)^{-1} Y_\infty, \quad \hat{\Sigma}_\infty = \sum_{t=0}^n Y_t Y_t^\top - 2\frac{1}{n}\left(\sum_{t=1}^n Y_{t-1} Y_t^\top\right)^\top \left(\sum_{t=1}^n Y_t Y_t^\top\right)\left(\sum_{t=1}^n Y_{t-1} Y_t^\top\right)$$
$$\hat{\Lambda}_\infty = (\hat{\Sigma}_\infty)^{-1}.$$

Plugging this solution to the objective function (Theorem 17) and keeping only terms depending on δ, the asymptotic objective function writes

$$\lim_{n\to+\infty} \mathcal{R}_{\mathcal{F}}^*(\delta) = -\lambda\delta^\top \left(\hat{\mu}_\infty - \lambda\hat{\Sigma}_\infty\delta\right).$$

■

Appendix D. Derivation of `VB-Portfolio` for Specific Models

Let us begin by introducing some convenient compact notations.

Definition 10 (Kronecker product) *Let $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{m\times q}$. Then the* kronecker product *between A and B, denoted by $A \otimes B \in \mathbb{R}^{mn\times pq}$ is defined as follows,*

$$A \otimes B = \begin{pmatrix} a_{1,1}B & \dots & a_{1,p}B \\ a_{2,1}B & \dots & a_{2,p}B \\ \vdots & \vdots & \vdots \\ a_{n,1}B & \dots & a_{n,p}B \end{pmatrix}.$$

Definition 11 (Vectorization operator) *Let $A \in \mathbb{R}^{n\times p}$, such that*

$$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,p} \\ \vdots & \vdots & \vdots \\ a_{n,1} & \dots & a_{n,p} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a_1} & \dots & \boldsymbol{a_p} \end{pmatrix} .$$

Then, we denote $\mathrm{vec}(A)$ *the vector of size np such that* $\mathrm{vec}(A) = \begin{pmatrix} \boldsymbol{a_1} \\ \vdots \\ \boldsymbol{a_p} \end{pmatrix}$.

Remark 12 *Using Theorem 5, the* AR *model* (11) *is* equivalent *to the following formulation,*

$$\begin{aligned} &Y_t \,|\, Y_{t-1}, \Gamma, \Lambda \sim \mathcal{N}(\Gamma Y_{t-1}, \Lambda) \qquad \forall t \in \mathbb{N}^* \\ &\mathrm{vec}(\Gamma) \sim \mathcal{N}(\mathrm{vec}(M_0), V_0 \otimes U_0)\,, \quad \Lambda \sim \mathcal{W}(\nu_0, \psi_0)\,. \end{aligned}$$

Proposition 13 (Properties of vec **operator)**

- $\mathrm{vec}(AXB) = (B^\top \otimes A)\mathrm{vec}(X)$.
- $\mathrm{Tr}(ABC) = \mathrm{vec}(A^\top)^\top (I \otimes B)\mathrm{vec}(C)$.
- $\mathrm{Tr}(A^\top BCD^\top) = \mathrm{vec}(A)^\top (D \otimes B)\mathrm{vec}(C)$.

We refer to Quintana (1987) for the proofs of these results.

D.1. General Fixed-Point Equation (Proof of Theorem 2)

This proof is a direct application of Bishop (Chapter 10; 2006) to our problem; we have an additional risk term $e^{-\lambda\delta^\top y}$, which modifies the computation of the complete joint distribution. We recall that $\tilde{\pi}_n$ is the probability distribution defined as $\mathrm{d}\tilde{\pi}_n = \frac{e^{-\lambda\delta^\top Y_{n+1}}}{\mathbb{E}_{\pi_n}\left[e^{-\lambda\delta^\top Y_{n+1}}\right]}\mathrm{d}\pi_n$. Then for any $\rho \in \mathcal{M}$, we have

$$\mathcal{K}(\rho, \tilde{\pi}_n) = \log \pi(H_n) - \mathrm{E}(\rho) + \log Z_\delta\,,$$

where

$$\mathrm{E}(\rho) = \int_{\mathcal{Y}\times\Theta} \log\left(\frac{e^{-\lambda\delta^\top Y_{n+1}}\pi\left(Y_{n+1}, \theta, H_n\right)}{\rho(Y_{n+1}, \theta)}\right)\rho(\mathrm{d}(Y_{n+1}, \theta))\,.$$

$E(\rho)$ is called the *evidence lower bound* (ELBO), and is the only term that depends on ρ. Hence, minimizing $\rho \mapsto \mathcal{K}(\rho, \tilde{\pi}_n)$ is equivalent at maximizing $\rho \mapsto \mathrm{E}(\rho)$. Then, for any $\rho \in \mathcal{F}$, we have

$$\mathrm{E}(\rho) = \int_{\mathcal{Y}\times\Theta} \left(\log\left(\mathrm{e}^{-\lambda\delta^\top \mathrm{Y_{n+1}}}\pi\left(\mathrm{Y_{n+1}}, \theta, \mathrm{H_n}\right)\right) - \log \rho_{\mathrm{y}}(\mathrm{Y_{n+1}}) - \sum_{\mathrm{i}=1}^{\mathrm{K}} \rho_{\mathrm{i}}(\theta_{\mathrm{i}})\right)\rho_{\mathrm{y}}(\mathrm{dY_{n+1}})\prod_{\mathrm{i}=1}^{\mathrm{K}}\rho_{\mathrm{i}}(\mathrm{d}\theta_{\mathrm{i}})\,. \tag{16}$$

Keeping terms that depend on θ_j only and applying Fubini theorem, we have

$$\mathrm{E}(\rho) \propto_{\theta_j} \int_{\Theta_j} \left(\int_{\mathcal{Y}\times\Theta\setminus\Theta_j} e^{-\lambda\delta^\top Y_{n+1}} \pi\left(Y_{n+1}, \theta, H_n\right) \rho_y(\mathrm{d}Y_{n+1}) \prod_{i\neq j}^{K} \rho_i(\mathrm{d}\theta_i) \right) \rho_j(\mathrm{d}\theta_j) - \int_{\Theta_j} \log \rho_j(\mathrm{d}\theta_j)\rho_j(\mathrm{d}\theta_j)\,.$$

The maximizer of $\rho \mapsto \mathrm{E}(\rho)$ with respect to each of the θ_i's can be derived (by computing the Lagragian of (16) (Jordan et al., 1999)), and we can show that the maximum is reached when

$$\forall j \in [K]\,, \quad \log \rho_j(\mathrm{d}\theta_j) \propto \exp\left(\int_{\mathcal{Y}\times\Theta\setminus\Theta_j} e^{-\lambda\delta^\top Y_{n+1}} \pi\left(Y_{n+1}, \theta, H_n\right) \rho_y(\mathrm{d}Y_{n+1}) \prod_{i\neq j}^{K} \rho_i(\mathrm{d}\theta_i) \right)\,, \tag{17}$$

which can be seen as the expectation of $e^{-\lambda\delta^\top Y_{n+1}} \pi\left(Y_{n+1}, \theta, H_n\right)$ taken with respect to all parameters θ_i with measure ρ_i except θ_j. The main advantage of using mean-field assumption is that (17) yields to a natural algorithm where we update successively each ρ_i's until stabilization.

D.2. `VB-Portfolio` for the Gaussian-Wishart Model

We start by deriving the fixed-point equation, and then we derive the corresponding objective function $\mathcal{R}_{\mathcal{F}}$.

Lemma 14 (Solution of (7) under Gaussian-Wishart model) *Under the stationary Gaussian-Wishart model (10), for any $\delta \in \mathcal{D}$, the corresponding variational distribution $\hat{\rho}_{\mathrm{VB}}$ can be factorised as follows,*

$$\hat{\rho}_{\mathrm{VB}}(\mathrm{d}(Y_{n+1}, \mu, \Lambda)) = \rho_y(\mathrm{d}Y_{n+1})\rho_\mu(\mathrm{d}\mu)\rho_\Lambda(\mathrm{d}\Lambda)\,,$$

where $\rho_y(\mathrm{d}Y_{n+1}) = \mathcal{N}(\mathrm{d}Y_{n+1}; \xi_y, \Lambda_y^{-1})$, $\rho_\mu(\mathrm{d}\mu) = \mathcal{N}(\mathrm{d}\mu; \xi_\mu, \Lambda_\mu^{-1})$ and $\rho_\Lambda(\mathrm{d}\Lambda) = \mathcal{W}(\mathrm{d}\Lambda; \nu_\Lambda, \psi_\Lambda)$. Moreover, the variational parameters $(\xi_y, \Lambda_y, \xi_\mu, \Lambda_\mu, \nu_\Lambda, \psi_\Lambda)$ satisfy a fixed-point equation $T_n(\xi_y, \Lambda_y, \xi_\mu, \Lambda_\mu, \nu_\Lambda, \psi_\Lambda) = (\xi_y, \Lambda_y, \xi_\mu, \Lambda_\mu, \nu_\Lambda, \psi_\Lambda)$, where T_n is given as follows:

$$T_n : (\xi_y, \Lambda_y, \xi_\mu, \Lambda_\mu, \nu_\Lambda, \psi_\Lambda) \mapsto \begin{pmatrix} \xi_\mu - \frac{\lambda}{\nu_\Lambda}\psi_\Lambda^{-1}\delta \\ \nu_\Lambda\psi_\Lambda \\ \frac{1}{n+1}\left(\nu_\Lambda\psi_\Lambda + \frac{1}{n+1}\Lambda_0\right)^{-1}\left(\nu_\Lambda\psi_\Lambda(\xi_y + \sum_{t\in[n]} Y_t) + \Lambda_0\mu_0\right) \\ (n+1)\nu_\Lambda\psi_\Lambda + \Lambda_0 \\ n + \nu_0 + 1 \\ \left(\Lambda_y^{-1} + \xi_y\xi_y^\top + (n+1)(\Lambda_\mu^{-1} + \xi_\mu\xi_\mu^\top) + \sum_{t\in[n]} Y_tY_t^\top - 2(\xi_y + \sum_{t\in[n]} Y_t)\xi_\mu^\top) + \psi_0^{-1}\right)^{-1} \end{pmatrix}.$$

Proof We first express $\tilde{\pi}_n$ independently of the underlying statistical model;

$$\begin{aligned} \tilde{\pi}_n(Y_{n+1}, \mu, \Lambda) &\propto e^{-\lambda\delta^\top Y_{n+1}} \pi(Y_{n+1}, \mu, \Lambda \,|\, H_n) \\ &\propto e^{-\lambda\delta^\top Y_{n+1}} \pi(H_n \,|\, Y_{n+1}, \mu, \Lambda)\pi(Y_{n+1}, \mu, \Lambda) \\ &\propto e^{-\lambda\delta^\top Y_{n+1}} \pi(H_n \,|\, Y_{n+1}, \mu, \Lambda)\pi(Y_{n+1} \,|\, \mu, \Lambda)\pi_0(\mu, \Lambda)\,, \end{aligned} \tag{18}$$

where the first equation follows from the definition of $\tilde{\pi}_n$ in Theorem 1 and the second equation follows from Bayes rule. Since the model (10) involves n *i.i.d.* observations conditionally on the parameters, (18) gives

$$\begin{aligned}\tilde{\pi}_n(Y_{n+1},\mu,\Lambda) &\propto e^{-\lambda\delta^\top Y_{n+1}}\pi(H_n\,|\,Y_{n+1},\mu,\Lambda)\pi(Y_{n+1}\,|\,\mu,\Lambda)\pi_0(\mu,\Lambda)\\ &\propto e^{-\lambda\delta^\top Y_{n+1}}\left(\prod_{t=1}^n\mathcal{N}(Y_t;\mu,\Lambda^{-1})\right)\mathcal{N}(Y_{n+1};\mu,\Lambda^{-1})\mathcal{N}(\mu;\mu_0,\Lambda_0^{-1})\mathcal{W}(\Lambda;\nu_0,\psi_0)\,.\end{aligned}$$

By Theorem 2, we can compute each variational distribution ρ_y, ρ_μ and ρ^*_Λ:

$$\begin{aligned}\log\rho_y(Y_{n+1}) &\propto_{Y_{n+1}} \mathbb{E}_{\rho_\mu,\rho_\Lambda}\left[\log\tilde{\pi}_n(Y_{n+1},\mu,\Lambda)\right]\\ &\propto \mathbb{E}_{\rho_\mu,\rho_\Lambda}\left[-\frac{1}{2}\left(2\lambda\delta^\top Y_{n+1}+(Y_{n+1}-\mu)^\top\Lambda(Y_{n+1}-\mu)\right)\right]\\ &\propto -\frac{1}{2}\left(Y_{n+1}^\top\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]Y_{n+1}-2Y_{n+1}^\top\left(\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\mathbb{E}_{\rho_\mu}\left[\mu\right]-\lambda\delta\right)\right)\,,\end{aligned}$$

which gives, by completing the Gaussian square with respect to the variable Y_{n+1},

$$\begin{aligned}\rho_{Y_{n+1}}(\mathrm{d}Y_{n+1}) &= \mathcal{N}(Y_{n+1};\xi_y,\Lambda_y^{-1})\\ \text{where } \xi_y &= \mathbb{E}_{\rho_\mu}\left[\mu\right]-\lambda\left(\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\right)^{-1}\delta\\ \Lambda_y &= \mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\,.\end{aligned}\tag{19}$$

For the mean variational distribution ρ_μ,

$$\begin{aligned}\log\rho_\mu(\mu) &\propto_\mu \mathbb{E}_{\rho_y,\rho_\Lambda}\left[\log\left(\prod_{t=1}^n\mathcal{N}(Y_t;\mu,\Lambda^{-1})\mathcal{N}(Y_{n+1};\mu,\Lambda^{-1})\mathcal{N}(\mu;\mu_0,\Lambda_0^{-1})\right)\right]\\ &\propto -\frac{1}{2}\left(\mu^\top\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\mu-2\mu^\top\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\mathbb{E}_{\rho_{Y_{n+1}}}\left[Y_{n+1}\right]+n\mu^\top\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\mu-2\mu^\top\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\sum_{t=1}^nY_t+\mu^\top\Lambda_0\mu-2\mu^\top\Lambda_0\mu_0\right)\\ &\propto -\frac{1}{2}\left(\mu^\top\left(\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]+n\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]+\Lambda_0\right)\mu-2\mu^\top\left(\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\mathbb{E}_{\rho_{Y_{n+1}}}\left[Y_{n+1}\right]+\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\sum_{t=1}^n+\Lambda_0\mu_0\right)\right)\,,\end{aligned}$$

which gives, by completing the Gaussian square with respect to the variable μ,

$$\begin{aligned}&\rho_\mu(\mathrm{d}\mu)=\mathcal{N}(\mathrm{d}\mu;\xi_\mu,\Lambda_\mu^{-1})\\ &\text{where } \xi_\mu=\left((n+1)\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]+\Lambda_0\right)^{-1}\left(\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]\left(\mathbb{E}_{\rho_{Y_{n+1}}}\left[Y_{n+1}\right]+\sum_{t=1}^n\right)+\Lambda_0\mu_0\right)\\ &\qquad\Lambda_\mu=(n+1)\mathbb{E}_{\rho_\Lambda}\left[\Lambda\right]+\Lambda_0\,.\end{aligned}$$

■

Finally, for the precision variational distribution ρ_Λ,

$$\begin{aligned}
\log \rho_\Lambda(\Lambda) &\propto_\Lambda \mathbb{E}_{\rho_{Y_{n+1}},\rho_\mu}\left[\log\left(\prod_{t=1}^{n}\mathcal{N}(Y_t;\mu,\Lambda^{-1})\mathcal{N}(Y_{n+1};\mu,\Lambda^{-1})\mathcal{W}(\Lambda;\nu_0,\psi_0)\right)\right] \\
&\propto \mathbb{E}_{\rho_{Y_{n+1}},\rho_\mu}\left[-\frac{1}{2}\sum_{t=1}^{n}(Y_t-\mu)\Lambda(Y_t-\mu)+\frac{n}{2}\log|\Lambda|-\frac{1}{2}(Y_{n+1}-\mu)^\top\Lambda(Y_t-\mu)\right] \\
&+\frac{1}{2}\log|\Lambda|+\frac{\nu_0-d+1}{2}\log|\Lambda|-\frac{1}{2}\mathrm{Tr}(\psi_0^{-1}\Lambda) \\
&\propto \mathbb{E}_{\rho_{Y_{n+1}},\rho_\mu}\Bigg[-\frac{1}{2}\left(\mathbb{E}_{\rho_{Y_{n+1}}}\left[\mathrm{Tr}(Y_{n+1}Y_{n+1}^\top\Lambda)\right]\right)-2\mathbb{E}_{\rho_\mu}[\mu]^\top\Lambda\mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]+\mathbb{E}_{\rho_{Y_{n+1}}}\left[\mathrm{Tr}(\mu\mu^\top\Lambda)\right] \\
&+\mathrm{Tr}(\sum_{t=1}^{n}Y_tY_t^\top\Lambda)-2\mathbb{E}_{\rho_\mu}\left[\mathrm{Tr}(\mu\sum_{t=1}^{n}Y_t^\top\Lambda)\right]+\mathbb{E}_{\rho_\mu}\left[n\mathrm{Tr}(\mu\mu^\top\Lambda)\right]+\mathrm{Tr}(\psi_0^{-1}\Lambda)+\frac{1}{2}\log|\Lambda| \\
&+\frac{n}{2}\log|\Lambda|+\frac{\nu_0-d-1}{2}\log|\Lambda|\Bigg] \\
&\propto -\frac{1}{2}\mathrm{Tr}\Bigg(\Big(\mathbb{E}_{Y_{n+1}}\left[Y_{n+1}Y_{n+1}^\top\right]+\mathbb{E}_{\rho_\mu}\left[\mu\mu^\top\right]-2\mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]\,\mathbb{E}_{\rho_\mu}[\mu]^\top+\sum_{t=1}^{n}Y_tY_t^\top-2\sum_{t=1}^{n}Y_t\mathbb{E}_{\rho_\mu}[\mu]^\top \\
&+n\mathbb{E}_{\rho_\mu}\left[\mu\mu^\top\right]+\psi_0^{-1}\Big)\Lambda\Bigg)+\frac{1}{2}(n+\nu_0-d-1)\log|\Lambda|\,.
\end{aligned}$$

Identifying the corresponding terms with a Wishart distribution yields to

$\rho_\mu(\mathrm{d}\Lambda)=\mathcal{W}(\mathrm{d}\Lambda;\nu_\Lambda,\psi_\Lambda)$
where $\nu_\Lambda=n+d+1$

$$\begin{aligned}
\psi_\Lambda^{-1} &= \mathbb{E}_{Y_{n+1}}\left[Y_{n+1}Y_{n+1}^\top\right]+\mathbb{E}_{\rho_\mu}\left[\mu\mu^\top\right]-2\mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]\,\mathbb{E}_{\rho_\mu}[\mu]^\top+\sum_{t=1}^{n}Y_tY_t^\top-2\sum_{t=1}^{n}Y_t\mathbb{E}_{\rho_\mu}[\mu]^\top \\
&+n\mathbb{E}_{\rho_\mu}\left[\mu\mu^\top\right]+\psi_0^{-1}\,.
\end{aligned}$$

Lemma 15 (Objective function under stationary Gaussian-Wishart model) *Fir any $\delta\in\mathcal{D}$, let $(\xi_y,\Lambda_y,\xi_\mu,\Lambda_\mu,\nu_\Lambda,\psi_\Lambda)$ be the parameters of the corresponding variational distribution $\hat{\rho}_{\mathrm{VB}}$ under the stationary Gaussian-Wishart model (10). Then, the objective function can be written as*

$$\begin{aligned}
\mathcal{R}_{\mathcal{F}}(\delta) &= -\frac{\nu_\Lambda}{2}\mathrm{Tr}\left(\left(\sum_{t\in[n]}Y_tY_t^\top-2\left(\sum_{t\in[n]}Y_t+\xi_y\right)\xi_\mu^\top+(n+1)(\Lambda_\mu^{-1}+\xi_\mu\xi_\mu^\top)+\Lambda_y^{-1}+\xi_y\xi_y^\top+\psi_0^{-1}\right)\psi_\Lambda\right) \\
&-\frac{1}{2}\mathrm{Tr}\big((\Lambda_\mu^{-1}+\xi_\mu\xi_\mu^\top)\Lambda_0\big)+\xi_\mu^\top\Lambda_0\mu_0+\frac{1}{2}(n+\nu_0+1)\log\det(\psi_\Lambda) \\
&-\frac{1}{2}\big(\log\det(\Lambda_y)+\log\det(\Lambda_\mu)\big)-\lambda\delta^\top\xi_y\,.
\end{aligned}$$

Proof From Theorem 14, we found that the variational distribution for the GW model can be written as

$$\hat{\rho}_{\mathrm{VB}}(\mathrm{d}(Y_{n+1},\mu,\Lambda))=\mathcal{N}(Y_{n+1};:\xi_y,\Lambda_y^{-1})\mathcal{N}(\mu;\xi_\mu,\Lambda_\mu^{-1})\mathcal{W}(\Lambda;\nu_\Lambda,\psi_\Lambda)\,.$$

Starting from the definition of $\mathcal{R}_{\mathcal{M}}$, we have, for any $\delta \in \mathcal{D}$,

$$\begin{aligned}\mathcal{R}_{\mathcal{M}}(\delta) = &-\mathbb{E}_{\rho_y}\left[\log \rho_y(Y_{n+1})\right] - \mathbb{E}_{\rho_\mu}\left[\log \rho_\mu(\mu)\right] - \mathbb{E}_{\rho_\Lambda}\left[\log \rho_\Lambda(\Lambda)\right] - \lambda\delta^\top \mathbb{E}_{\rho_y}\left[Y_{n+1}\right] \\ &+ \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \pi\left(Y_{n+1}, \mu, \Lambda \,|\, H_n\right)\right] + C\,,\end{aligned}$$

where C is a constant that does not depend on δ. We now aim at computing each of these terms. One can easily verify that

$$-\mathbb{E}_{\rho_y}\left[\log \rho_y(Y_{n+1})\right] \propto_\delta -\frac{1}{2}\log|\Lambda_y| \quad -\mathbb{E}_{\rho_\mu}\left[\log \rho_\mu(\mu)\right] \propto_\delta -\frac{1}{2}\log|\Lambda_\mu| \quad -\mathbb{E}_{\rho_\Lambda}\left[\log \rho_\Lambda(\Lambda)\right] \propto_\delta \frac{d+1}{2}\log|\psi_\Lambda|\,.$$

Moreover,

$$\begin{aligned}\mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \pi(Y_{n+1}, \mu, \Lambda \,|\, H_n)\right] \propto_\delta &\, \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \prod_{t=1}^n \mathcal{N}(Y_t; \mu, \Lambda^{-1})\right] + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{N}(Y_{n+1}; \mu, \Lambda^{-1})\right] \\ &+ \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{N}(\mu; \mu_0, \Lambda_0^{-1})\right] + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{W}(\Lambda; \nu_0, \psi_0\right]\,.\end{aligned}$$

We can compute each of these terms exactly the same way we did in the proof of Theorem 14, and combining these with the terms above give the desited expression. ■

D.3. `VB-Portfolio` for the AR Model

Lemma 16 (Solution of (7) under AR(1) model) *Under* AR *model* (11)*, for any* $\delta \in \mathcal{D}$*, the corresponding variational distribution* $\hat{\rho}_{\mathrm{VB}}$ *can be factorised as follows,*

$$\hat{\rho}_{\mathrm{VB}}(\mathrm{d}(Y_{n+1}, \Gamma, \Lambda)) = \rho_y(\mathrm{d}Y_{n+1})\rho_\Gamma(\mathrm{d}(\mathrm{vec}(\Gamma)))\rho_\Lambda(\mathrm{d}\Lambda)\,,$$

where $\rho_y(\mathrm{d}Y_{n+1}) = \mathcal{N}(\mathrm{d}Y_{n+1}; \xi_y, \Lambda_y^{-1})$*,* $\rho_\Gamma(\mathrm{d}(\mathrm{vec}(\Gamma))) = \mathcal{N}(\mathrm{d}(\mathrm{vec}(\Gamma)); \mathrm{vec}(M_\Gamma), V_\Gamma \otimes U_\Gamma)$*,* $\rho_\Lambda(\mathrm{d}\Lambda) = \mathcal{W}(\mathrm{d}\Lambda; \nu_\Lambda, \psi_\Lambda)$*. Moreover, the variational parameters* $\phi = (\xi_y, \Lambda_y, \mathrm{vec}(M_\Gamma), V_\Gamma \otimes U_\Gamma, \nu_\Lambda, \psi_\Lambda)$ *satisfy a fixed-point equation* $T_n(\phi) = \phi$*, where* T_n *is given as follows:*

$$T_n : \phi \mapsto \begin{pmatrix} M_\Gamma Y_n - \frac{\lambda}{\nu_\Lambda}\psi_\Lambda^{-1}\delta \\ \nu_\Lambda \psi_\Lambda \\ (V_\Gamma \otimes U_\Gamma)\left[(I_d \otimes \nu_\Lambda \psi_\Lambda)\,\mathrm{vec}(\sum_{t=1}^n Y_t Y_{t-1}^\top + \xi_y Y_n^\top) + (V_0^{-1} \otimes U_0^{-1})\mathrm{vec}(M_0)\right] \\ \left(\sum_{t=0}^n Y_t Y_t^\top \otimes \nu_\Lambda \psi_\Lambda + V_0^{-1} \otimes U_0^{-1}\right)^{-1} \\ n + \nu_0 + 1 \\ \psi_0^{-1} + \xi_y \xi_y^\top + \Lambda_y^{-1} + \sum_{t=1}^n Y_t Y_t^\top - 2M_\Gamma\left(\sum_{t\in[n]} Y_{t-1}Y_t^\top + Y_n \xi_y^\top\right) + M_\Gamma \sum_{t=0}^n Y_t Y_t^\top M_\Lambda^\top + \sum_{i=1}^{d^2} \sigma_i \mathrm{vec}^{-1}(u_i)\left(\sum_{t=0}^n Y_t Y_t^\top\right)\mathrm{vec}^{-1}(u_i)^\top \end{pmatrix},$$

where $(\sigma_i, u_i)_{i\in[d]}$ *is the spectral decomposition of* $V_\Lambda \otimes U_\Lambda$*.*

Proof First, we write $\tilde{\pi}_n$ by remarking

$$\begin{aligned}\tilde{\pi}_n(Y_{n+1}, \Gamma, \Lambda) &\propto e^{-\lambda\delta^\top Y_{n+1}}\pi(Y_{n+1}, \Gamma, \Lambda \,|\, H_n) \\ &\propto e^{-\lambda\delta^\top Y_{n+1}}\pi(H_n \,|\, Y_{n+1}, \Gamma, \Lambda)\pi(Y_{n+1}, \Gamma, \Lambda) \\ &\propto e^{-\lambda\delta^\top Y_{n+1}}\pi(H_n \,|\, Y_{n+1}, \Gamma, \Lambda)\pi(Y_{n+1} \,|\, \Gamma, \Lambda)\pi_0(\Gamma, \Lambda)\,, \qquad (20)\end{aligned}$$

On behalf of (11), (20) becomes

$$\tilde{\pi}_n(Y_{n+1}, \Gamma, \Lambda) \propto e^{-\lambda\delta^\top Y_{n+1}}\left(\prod_{t=1}^n \mathcal{N}(Y_t; \Gamma Y_{t-1}, \Lambda^{-1})\right)\pi(Y_{n+1}; \Gamma Y_t, \Lambda^{-1})\mathcal{MN}(\Gamma; M_0, U_0, V_0)\mathcal{W}(\Lambda; \nu_0, \psi_0)\,.$$

Keeping only terms that depend on the variable Y_{n+1},

$$\begin{aligned}\log \rho_y(Y_{n+1}) &\propto_{Y_{n+1}} \mathbb{E}_{\rho_\Gamma,\rho_\Lambda}\left[\log\left(e^{-\lambda\delta^\top Y_{n+1}}\mathcal{N}(Y_{n+1};\Gamma Y_n,\Lambda^{-1})\right)\right]\\ &\propto -\lambda\delta^\top Y_{n+1} - \frac{1}{2}\left(Y_{n+1}^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\, Y_{n+1} - 2Y_{n+1}^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\,\mathbb{E}_{\rho_\Gamma}[\Gamma]\, Y_n\right)\\ &\propto -\frac{1}{2}\left(Y_{n+1}^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\, Y_{n+1} - 2Y_{n+1}^\top\left(\mathbb{E}_{\rho_\Lambda}[\Lambda]\,\mathbb{E}_{\rho_\Gamma}[\Gamma]\, Y_n - \lambda\delta\right)\right),\end{aligned}$$

which gives, by completying the Gaussian square with respect tp Y_{n+1},

$$\begin{aligned}\rho_y(\mathrm{d}Y_{n+1}) &= \mathcal{N}(\mathrm{d}Y_{n+1};\xi_y,\Lambda_y^{-1})\\ \text{where } \xi_y &= \mathbb{E}_{\rho_\Gamma}[\Gamma]\left(Y_n - \lambda\mathbb{E}_{\rho_\Gamma}[\Gamma]^{-1}\,\mathbb{E}_{\rho_\Lambda}[\Lambda]^{-1}\,\delta\right)\\ \Lambda_y &= \mathbb{E}_{\rho_\Lambda}[\Lambda]\,.\end{aligned}$$

For the variational distribution with respect to Γ,

$$\begin{aligned}\log \rho_\Gamma(\Gamma) &\propto_\Gamma \mathbb{E}_{\rho_{Y_{n+1}},\rho_\Lambda}\left[\log\left(\prod_{t=1}^n \mathcal{N}(Y_t;\Gamma Y_{t-1},\Lambda^{-1})\mathcal{N}(Y_{n+1};\Gamma Y_n,\Lambda^{-1})\mathcal{MN}(\Gamma;M_0,U_0,V_0)\right)\right]\\ &\propto \underbrace{-\frac{1}{2}\sum_{t=1}^n (Y_t-\Gamma Y_{t-1})^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\,(Y_t-\Gamma Y_{t-1})}_{(1)} \underbrace{-\frac{1}{2}\mathbb{E}_{\rho_{Y_{n+1}}}\left[(Y_{n+1}-\Gamma Y_n)^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\,(Y_{n+1}-\Gamma Y_n)\right]}_{(2)}\\ &\quad \underbrace{-\frac{1}{2}\left(\mathrm{vec}(\Gamma)-\mathrm{vec}(M_0)\right)^\top (V_0\otimes U_0)^{-1}\left(\mathrm{vec}(\Gamma)-\mathrm{vec}(M_0)\right)}_{(3)}\,.\end{aligned}$$

$$\begin{aligned}(1) &\propto_\Gamma -\frac{1}{2}\sum_{t=1}^n\left((\Gamma Y_{t-1})^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\,\Gamma Y_{t-1} - 2Y_t^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\,\Gamma Y_{t-1}\right)\\ &\propto -\frac{1}{2}\left(\mathrm{Tr}\left(\Gamma^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\,\Gamma\sum_{t=1}^n Y_{t-1}Y_{t-1}^\top\right) - 2\mathrm{Tr}\left(\sum_{t=1}^n Y_{t-1}Y_t^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\,\Gamma\right)\right)\\ &\propto -\frac{1}{2}\left(\mathrm{vec}(\Gamma)^\top\left(\sum_{t=1}^n Y_{t-1}Y_{t-1}^\top\otimes\mathbb{E}_{\rho_\Lambda}[\Lambda]\right)\mathrm{vec}(\Gamma) - 2\mathrm{vec}\left(\sum_{t=1}^n Y_tY_{t-1}^\top\right)^\top (I_d\otimes\mathbb{E}_{\rho_\Lambda}[\Lambda])\mathrm{vec}(\Gamma)\right),\end{aligned}$$

where the last line follows from Theorem 13. We can show with the exact same arguments that

$$(2) \propto_\Gamma -\frac{1}{2}\left(\mathrm{vec}(\Gamma)^\top\left(Y_nY_n^\top\otimes\mathbb{E}_{\rho_\Lambda}[\Lambda]\right)\mathrm{vec}(\Gamma) - 2\mathrm{vec}\left(\mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]\,Y_{n-1}^\top\right)^\top (I_d\otimes\mathbb{E}_{\rho_\Lambda}[\Lambda])\mathrm{vec}(\Gamma)\right),$$

and for the third term,

$$(3) \propto -\frac{1}{2}\left(\mathrm{vec}(\Gamma)^\top (V_0\otimes U_0)^{-1}\,\mathrm{vec}(\Gamma) - 2\mathrm{vec}(\Gamma)^\top (V_0\otimes U_0)^{-1}\,\mathrm{vec}(M_0)\right).$$

Adding (1), (2) and (3) and completing the corresponding Gaussian squares gives

$$\begin{aligned}
&\log \rho_\Gamma(\mathrm{d}\Gamma) = \mathcal{N}(\mathrm{dvec}(\Gamma); \xi_\Gamma, \Lambda_\Gamma^{-1}) \\
&\quad \text{where } \Lambda_\Gamma = \sum_{t=0}^{n} Y_t Y_t^\top \otimes \mathbb{E}_{\rho_\Lambda}[\Lambda] + V_0^{-1} \otimes U_0^{-1} \\
&\qquad \xi_\Gamma = \Lambda_\Gamma^{-1} \left((I_d \otimes \mathbb{E}_{\rho_\Lambda}[\Lambda]) \left(\mathrm{vec}\left(\sum_{t=1}^{n} Y_t Y_{t-1}^\top \right) + \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}] Y_n^\top \right) + (V_0^{-1} \otimes U_0^{-1}) \mathrm{vec}(M_0) \right).
\end{aligned}$$

The variational distribution with respect to Λ requires more efforts:

$$\begin{aligned}
\log \rho_\Lambda(\Lambda) &\propto_\Lambda \mathbb{E}_{\rho_{Y_{n+1}}, \rho_\Gamma} \left[\log \prod_{t=1}^{n} \mathcal{N}(Y_t; \Gamma Y_{t-1}, \Lambda^{-1}) \mathcal{N}(Y_{n+1}; \Gamma Y_n, \Lambda^{-1}) \mathcal{W}(\Lambda; \nu_0, \psi_0) \right] \\
&\propto \mathbb{E}_{\rho_{Y_{n+1}}, \rho_\Gamma} \left[-\frac{1}{2} \sum_{t=1}^{n} (Y_t - \Gamma Y_{t-1})^\top \Lambda (Y_t - \Gamma Y_{t-1}) - \frac{1}{2} (Y_{n+1} - \Gamma Y_n)^\top \Lambda (Y_{n+1} - \Gamma Y_{n-1}) \right] \\
&\quad + \frac{n + \nu_0 - d - 1}{2} \log |\Lambda| - \frac{1}{2} \mathrm{Tr}(\Lambda \psi_0^{-1}) \\
&\propto -\frac{1}{2} \mathrm{Tr} \left(\Lambda \underbrace{\left(\mathbb{E}_{\rho_\Gamma} \left[\sum_{t=1}^{n} (Y_t - \Gamma Y_{t-1})(Y_t - \Gamma Y_{t-1})^\top \right] + \mathbb{E}_{\rho_{Y_{n+1}}, \rho_\Gamma} \left[(Y_{n+1} - \Gamma Y_n)(Y_{n+1} - \Gamma Y_n)^\top \right] + \psi_0^{-1} \right)}_{(\star)} \right) \\
&\quad + \frac{n + \nu_0 - d - 1}{2} \log |\Lambda| .
\end{aligned}$$

where we refer to the proof in Theorem 14 for the computation in the second line. Then term inside the trace writes

$$\begin{aligned}
(\star) &= \sum_{t=1}^{n} Y_t Y_t^\top + \sum_{t=1}^{n} \mathbb{E}_{\rho_\Gamma} \left[\Gamma Y_{t-1} Y_{t-1} \Gamma^\top \right] - \mathbb{E}_{\rho_\Gamma}[\Gamma] \sum_{t=1}^{n} Y_{t-1} Y_t^\top - \sum_{t=1}^{n} Y_t Y_{t-1}^\top \mathbb{E}_{\rho_\Gamma}[\Gamma]^\top + \psi_0^{-1} \\
&\quad + \mathbb{E}_{\rho_{Y_{n+1}}} \left[Y_{n+1} Y_{n+1}^\top \right] + \mathbb{E}_{\rho_\Gamma} \left[\Gamma Y_n Y_n \Gamma^\top \right] - \mathbb{E}_{\rho_\Gamma}[\Gamma] Y_n \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]^\top - \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}] Y_n \Gamma^\top \\
&= \sum_{t=1}^{n} Y_t Y_t^\top + \sum_{t=1}^{n} \mathbb{E}_{\rho_\Gamma} \left[\Gamma Y_{t-1} Y_{t-1} \Gamma^\top \right] - \mathbb{E}_{\rho_\Gamma}[\Gamma] \sum_{t=1}^{n} Y_{t-1} Y_t^\top - \sum_{t=1}^{n} Y_t Y_{t-1}^\top \mathbb{E}_{\rho_\Gamma}[\Gamma]^\top \\
&\quad + \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}] \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]^\top + \mathrm{cov}_{\rho_{Y_{n+1}}}(Y_{n+1}) + \mathbb{E}_{\rho_\Gamma} \left[\Gamma Y_n Y_n \Gamma^\top \right] - \mathbb{E}_{\rho_\Gamma}[\Gamma] Y_n \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]^\top \\
&\quad - \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}] Y_n \Gamma^\top .
\end{aligned}$$

The main difficulty here is to compute the term related to $\sum_{t=1}^{n} \mathbb{E}_{\rho_\Gamma}\left[\Gamma Y_{t-1} Y_{t-1} \Gamma^\top\right] + \mathbb{E}_{\rho_\Gamma}\left[\Gamma Y_n Y_n \Gamma^\top\right] = \sum_{t=0}^{n} \mathbb{E}_{\rho_\Gamma}\left[\Gamma Y_t Y_t \Gamma^\top\right]$; extracting this term gives

$$\begin{aligned}
\mathrm{Tr}\left(\Lambda \sum_{t=0}^{n} \mathbb{E}_{\rho_\Gamma}\left[\Gamma Y_t Y_t \Gamma^\top\right]\right) &= \mathbb{E}_{\rho_\Gamma}\left[\mathrm{Tr}\left(\Gamma^\top \Lambda \Gamma \sum_{t=0}^{n} Y_t Y_t^\top\right)\right] \\
&= \mathbb{E}_{\rho_\Gamma}\left[\mathrm{vec}(\Gamma)^\top \left(\sum_{t=0}^{n} Y_t Y_t^\top \otimes \Lambda\right) \mathrm{vec}(\Gamma)\right] \\
&= \mathrm{Tr}\left(\left(\sum_{t=0}^{n} Y_t Y_t^\top \otimes \Lambda\right) \mathbb{E}_{\rho_\Gamma}\left[\mathrm{vec}(\Gamma)\mathrm{vec}(\Gamma)^\top\right]\right) \\
&= \underbrace{\mathrm{Tr}\left(\left(\sum_{t=0}^{n} Y_t Y_t^\top \otimes \Lambda\right) \mathrm{vec}(\mathbb{E}_{\rho_\Gamma}\left[\mathrm{vec}(\Gamma)\right])\mathrm{vec}(\mathbb{E}_{\rho_\Gamma}\left[\mathrm{vec}(\Gamma)\right])^\top\right)}_{(1)} \\
&\quad + \underbrace{\mathrm{Tr}\left(\left(\sum_{t=0}^{n} Y_t Y_t^\top \otimes \Lambda\right) \mathrm{cov}_{\rho_\Gamma}\left(\mathrm{vec}(\Gamma)\right)\right)}_{(2)} .
\end{aligned}$$

The first trace term of the last line writes

$$(1) = \mathbb{E}_{\rho_\Gamma}\left[\mathrm{vec}(\Gamma)\right]^\top \left(\sum_{t=0}^{n} Y_t Y_t^\top \otimes \Lambda\right) \mathbb{E}_{\rho_\Gamma}\left[\mathrm{vec}(\Gamma)\right] = \mathrm{Tr}\left(\mathbb{E}_{\rho_\Gamma}\left[\Gamma\right]^\top \Lambda \mathbb{E}_{\rho_\Gamma}\left[\Gamma\right] \sum_{t=0}^{n} Y_t Y_t^\top\right) = \mathrm{Tr}\left(\Lambda \mathbb{E}_{\rho_\Gamma}\left[\Gamma\right] \sum_{t=0}^{n} Y_t Y_t^\top \mathbb{E}_{\rho_\Gamma}\left[\Gamma\right]^\top\right) ,$$

while the second term gives,

$$(2) = \mathrm{Tr}\left(\left(\sum_{t=0}^{n} Y_t Y_t^\top \otimes \Lambda\right) \sum_{i=1}^{d^2} \sigma_i u_i u_i^\top\right) ,$$

where we denote $(\sigma_i, u_i)_{i=1}^{d^2}$ the spectral decomposition of the covariance matrix $\mathrm{cov}_{\rho_\Gamma}\left(\mathrm{vec}(\Gamma)\right)$. Then we have

$$\begin{aligned}
(2) = \sum_{i=1}^{d^2} \sigma_i u_i^\top \left(\sum_{t=0}^{n} Y_t Y_t^\top \otimes \Lambda\right) u_i &= \sum_{i=1}^{d^2} \sigma_i \mathrm{Tr}\left(\mathrm{vec}^{-1}(u_i)^\top \Lambda \mathrm{vec}^{-1}(u_i) \sum_{t=0}^{n} Y_t Y_t^\top\right) \\
&= \sum_{i=1}^{d^2} \sigma_i \mathrm{Tr}\left(\Lambda \mathrm{vec}^{-1}(u_i) \sum_{t=0}^{n} Y_t Y_t^\top \mathrm{vec}^{-1}(u_i)^\top\right) .
\end{aligned}$$

Combining all these yields to

$$\rho_\Lambda(\mathrm{d}\Lambda) = \mathcal{W}(\mathrm{d}\Lambda; \nu_\Lambda, \psi_\Lambda)$$
$$\text{where } \nu_\Lambda = n + d + 1$$
$$\psi_\Lambda = \psi_0^{-1} + \sum_{t=1}^{n} Y_t Y_t^\top + \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]\,\mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]^\top + \mathrm{cov}_{\rho_{Y_{n+1}}}(Y_{n+1}) - \mathbb{E}_{\rho_\Gamma}[\Gamma]\,Y_n \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]^\top$$
$$- \mathbb{E}_{\rho_{Y_{n+1}}}[Y_{n+1}]\,Y_n \mathbb{E}_{\rho_\Gamma}[\Gamma]^\top - \mathbb{E}_{\rho_\Gamma}[\Gamma] \sum_{t=1}^{n} Y_{t-1} Y_t^\top - \sum_{t=1}^{n} Y_t Y_{t-1}^\top \mathbb{E}_{\rho_\Gamma}[\Gamma]^\top$$
$$+ \mathbb{E}_{\rho_\Gamma}[\Gamma] \sum_{t=0}^{n} Y_t Y_t^\top \mathbb{E}_{\rho_\Gamma}[\Gamma]^\top + \sum_{i=1}^{d^2} \sigma_i \mathrm{vec}^{-1}(u_i) \sum_{t=0}^{n} Y_t Y_t^\top \mathrm{vec}^{-1}(u_i)^\top .$$

■

We next derive the corresponding objective function.

Lemma 17 (Objective function under AR model) *For any $\delta \in \mathcal{D}$, let $(\xi_y, \Lambda_y, M_\Gamma, V_\Gamma \otimes U_\Gamma, \nu_\Lambda, \psi_\Lambda)$ the parameters of the corresponding variational distribution $\hat{\rho}_{\mathrm{VB}}$ under the* AR *model* (11)*. Then, the objective function can is written as*

$$\mathcal{R}_{\mathcal{F}}(\delta) = -\frac{\nu_\Lambda}{2}\mathrm{Tr}\Bigg(\Bigg(\psi_0^{-1} + \xi_y \xi_y^\top + \Lambda_y^{-1} + \sum_{t=1}^{n} Y_t Y_t^\top - 2M_\Gamma \left(\sum_{t=1}^{n} Y_{t-1} Y_t^\top + Y_n \xi_y^\top\right) M_\Lambda^\top + M_\Lambda \sum_{t=0}^{n} Y_t Y_t^\top M_\Lambda^\top +$$
$$\sum_{i=1}^{d^2} \sigma_i \mathrm{vec}^{-1}(u_i) \sum_{t=0}^{n} Y_t Y_t^\top \mathrm{vec}^{-1}(u_i)^\top\Bigg)\psi_\Lambda\Bigg) - \frac{1}{2}\mathrm{Tr}\left((V_0^{-1} \otimes U_0^{-1})\left(\mathrm{vec}(M_\Lambda)\mathrm{vec}(M_\Lambda)^\top + V_\Lambda \otimes U_\Lambda\right)\right)$$
$$+ \mathrm{vec}(M_\Lambda)^\top (V_0 \otimes U_0)^{-1} \mathrm{vec}(M_0) + \frac{1}{2}(n + \nu_0 + 1)\log\det(\psi_\Lambda) - \frac{1}{2}\log\det(\Lambda_y)$$
$$- \frac{1}{2}\log\det(V_\Lambda \otimes U_\Lambda) - \lambda \delta^\top \xi_y .$$

Proof From Theorem 18, we found that the variational distribution for the AR model can be written as

$$\hat{\rho}_{\mathrm{VB}}(\mathrm{d}(Y_{n+1}, \Gamma, \Lambda)) = \mathcal{N}(\mathrm{d}Y_{n+1}; \xi_y, \Lambda_y^{-1})\mathcal{N}(\mathrm{d}(\mathrm{vec}(\Gamma)); \mathrm{vec}(M_\Gamma), V_\Gamma \otimes U_\Gamma)\mathcal{W}(\mathrm{d}\Lambda; \nu_\Lambda, \psi_\Lambda) .$$

Starting again from the definition of $\mathcal{R}_{\mathcal{M}}$, we have, for any $\delta \in \mathcal{D}$,

$$\mathcal{R}_{\mathcal{M}}(\delta) \propto_\delta -\mathbb{E}_{\rho_y}[\log \rho_y(Y_{n+1})] - \mathbb{E}_{\rho_\Gamma}[\log \rho_\Gamma(\Gamma)] - \mathbb{E}_{\rho_\Lambda}[\log \rho_\Lambda(\Lambda)] - \lambda \delta^\top \mathbb{E}_{\rho_y}[Y_{n+1}]$$
$$+ \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}[\log \pi(Y_{n+1}, \Gamma, \Lambda \mid H_n)] ,$$

where

$$- \mathbb{E}_{\rho_y}[\log \rho_y(Y_{n+1})] \propto_\delta -\frac{1}{2}\log|\Lambda_y| \qquad - \mathbb{E}_{\rho_\Gamma}[\log \rho_\Gamma(\Gamma)] \propto_\delta -\frac{1}{2}\log|V_\Lambda \otimes U_\Lambda|$$
$$- \mathbb{E}_{\rho_\Lambda}[\log \rho_\Lambda(\Lambda)] \propto_\delta \frac{d+1}{2}\log|\psi_\Lambda| .$$

Moreover,

$$\mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \pi(Y_{n+1}, \Gamma, \Lambda \mid H_n)\right] \propto_{\delta} \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \prod_{t=1}^{n} \mathcal{N}(Y_t; \Gamma Y_{t-1}, \Lambda^{-1})\right] + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{N}(Y_{n+1}; \Gamma Y_n, \Lambda^{-1})\right]$$
$$+ \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{MN}(\Gamma; M_0, U_0, V_0)\right] + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{W}(\Lambda; \nu_0, \psi_0\right] .$$

We can compute each of these terms exactly the same way we did in the proof of Theorem 16, and combining these with the terms above give the desited expression. ■

D.4. `VB-Portfolio` for the Gaussian Process Model

Lemma 18 (Solution of (7) under Gaussian-process Wishart model) *Under* GP *model (10), for any $\delta \in \mathcal{D}$, the corresponding variational distribution $\hat{\rho}_{\mathrm{VB}}$ can be factorised as follows,*

$$\hat{\rho}_{\mathrm{VB}}(\mathrm{d}(Y_{n+1}, \mu, \Lambda)) = \rho_y(\mathrm{d}Y_{n+1})\rho_\mu(\mathrm{d}\mu)\rho_\Lambda(\mathrm{d}\Lambda)\,,$$

where $\rho_y(\mathrm{d}Y_{n+1}) = \mathcal{N}(\mathrm{d}Y_{n+1}; \xi_y, \Lambda_y^{-1})$, $\rho_\mu(\mathrm{d}\mu) = \mathcal{MGP}(\mathrm{d}\mu; m_\mu(\cdot), k_\mu(\cdot), \Omega_\mu)$, $\rho_\Lambda(\mathrm{d}\Lambda) = \mathcal{W}(\mathrm{d}\Lambda; \nu_\Lambda, \psi_\Lambda)$. At time step $n+1$, the variational parameters $\phi_{n+1} = (\xi_y, \Lambda_y, m_\mu^{1:n+1}, (\Omega_\mu \otimes K_\mu)^{1:n+1}, \nu_\Lambda, \psi_\Lambda)$ satisfy a fixed-point equation $T_n(\phi_{n+1}) = \phi_{n+1}$, where T_n is given as follows:

$$T_n : \phi \mapsto \begin{pmatrix} m_\mu(n+1) - \frac{\lambda}{\nu_\Lambda}\psi_\Lambda^{-1}\delta \\ \nu_\Lambda \psi_\Lambda \\ (\Omega_\mu \otimes K_\mu)\left((I_{n+1} \otimes \nu_\Lambda \psi_\Lambda)\,\mathrm{vec}(\boldsymbol{Y}) + \left(\Omega_0^{-1} \otimes K_0^{-1}\right)\mathrm{vec}(M_0)\right) \\ \left(I_{n+1} \otimes \nu_\Lambda \psi_\Lambda + \Omega_0^{-1} \otimes K_0^{-1}\right)^{-1} \\ \nu_0 + n + 1 \\ \psi_0^{-1} + \xi_y \xi_y^\top + \Lambda_y^{-1} + \sum_{t\in[n]} Y_t Y_t^\top - 2\sum_{t\in[n]} m_\mu(t) Y_t^\top - 2 m_\mu(n+1)\xi_y^\top + \sum_{t=1}^{n+1} m_\mu(t) m_\mu(t)^\top + \sum_{t=1}^{n+1} Cov(\mu(t), \mu(t)) \end{pmatrix}.$$

Proof First, we write $\tilde{\pi}_n$ as

$$\begin{aligned}\tilde{\pi}_n(Y_{n+1}, \mu, \Lambda) &\propto e^{-\lambda\delta^\top Y_{n+1}} \pi(Y_{n+1}, \mu, \Lambda \mid H_n) \\ &\propto e^{-\lambda\delta^\top Y_{n+1}} \pi(H_n \mid \mu, \Lambda)\pi(Y_{n+1} \mid \mu, \Lambda)\pi_0(\mu, \Lambda) \\ &\propto e^{-\lambda\delta^\top Y_{n+1}} \prod_{t=1}^{n} \mathcal{N}(Y_t; \mu(t), \Lambda^{-1})\mathcal{N}(Y_{n+1}; \mu(t+1), \Lambda^{-1})\mathcal{MGP}(\mu; \mu_0(\cdot), K_0, \Omega_0)\mathcal{W}(\Lambda; \nu_0, \psi_0)\,.\end{aligned}$$

First, the variational distribution ρ_y can be derived as

$$\begin{aligned}\log \rho_y(Y_{n+1}) &\propto_{Y_{n+1}} \mathbb{E}_{\rho_\mu, \rho_\Lambda}\left[\log e^{-\lambda\delta^\top Y_{n+1}} \mathcal{N}(Y_{n+1}; \mu(t+1), \Lambda^{-1})\right] \\ &\propto \mathbb{E}_{\rho_\mu, \rho_\Lambda}\left[-\lambda\delta^\top Y_{n+1} - \frac{1}{2}\left(Y_{n+1} - \mu(t+1)\right)^\top \Lambda \left(Y_{n+1} - \mu(t+1)\right)\right] \\ &\propto -\frac{1}{2}\left(Y_{n+1}^\top \Lambda Y_{n+1} - 2Y_{n+1}^\top\left(\mathbb{E}_{\rho_\Lambda}[\Lambda]\,\mathbb{E}_{\rho_\mu}[\mu(t+1)] - \lambda\delta\right)\right) \\ &\propto \log \mathcal{N}\left(Y_{n+1}; \xi_y, \Lambda_y^{-1}\right)\,.\end{aligned}$$

The variational distribution ρ_y can be derived by deploying the GP prior on the indices $\{1, \dots n+1\}$;

$$\begin{aligned}
&\log \rho_\mu(\mu) \\
&\propto_\mu \mathbb{E}_{\rho_y, \rho_\Lambda}\left[\log \prod_{t=1}^{n} \mathcal{N}(Y_t; \mu, \Lambda^{-1}) \mathcal{N}(Y_{n+1}; \mu, \Lambda^{-1}) \mathcal{MGP}(\mu; \mu_0(\cdot), K_0, \Omega_0)\right] \\
&\propto \mathbb{E}_{\rho_y, \rho_\Lambda}\Bigg[-\frac{1}{2}\sum_{t=1}^{n} \left(Y_t - \mu(t)\right)^\top \Lambda \left(Y_t - \mu(t)\right) \\
&- \frac{1}{2}\left(Y_{n+1} - \mu(n+1)\right)^\top \Lambda \left(Y_{n+1} - \mu(n+1)\right) - \frac{1}{2}\boldsymbol{\mu}_{1:n+1}^\top \left(\Omega_0 \otimes \boldsymbol{K}_{\mathbf{0}}^{1:n+1}\right)^{-1} \boldsymbol{\mu}_{1:n+1}\Bigg],
\end{aligned}$$

where we define $\boldsymbol{\mu}_{1:n+1}$ as the concatenated vector $(\mu(1), \dots, \mu(n+1))$ of size $(n+1) \times d$, and $\boldsymbol{K}_{\mathbf{0}}^{1:n+1}$ as the matrix of size $(n+1, n+1)$ whose entries are $K_0(i,j)$ for $i, j \in [n+1]$. Then we have

$$\begin{aligned}
\log \rho_\mu(\mu) \propto_\mu &-\frac{1}{2}\Bigg(\sum_{t=1}^{n} \mu(t)^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\, \mu(t) - 2\sum_{t=1}^{n} \mu(t)^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\, Y_t + \mu(n+1)^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\, \mu(n+1) \\
&- 2\mu(n+1)^\top \mathbb{E}_{\rho_\Lambda}[\Lambda]\, \mathbb{E}_{\rho_y}[Y_{n+1}] + \boldsymbol{\mu}_{1:n+1}\left(\Omega_0 \otimes \boldsymbol{K}_{\mathbf{0}}^{1:n+1}\right)^{-1} - 2\boldsymbol{\mu}_{1:n+1}\left(\Omega_0 \otimes \boldsymbol{K}_{\mathbf{0}}^{1:n+1}\right)^{-1} \boldsymbol{\mu}_{1:n+1}\Bigg) \\
\propto &-\frac{1}{2}\Bigg(\boldsymbol{\mu}_{1:n+1}^\top \left(I_{n+1} \otimes \mathbb{E}_{\rho_\Lambda}[\Lambda] + \left(\Omega_0 \otimes \boldsymbol{K}_{\mathbf{0}}^{1:n+1}\right)^{-1}\right) \boldsymbol{\mu}_{1:n+1} \\
&- 2\boldsymbol{\mu}_{1:n+1}\left(\left(\Omega_0 \otimes \boldsymbol{K}_{\mathbf{0}}^{1:n+1}\right)^{-1} \boldsymbol{\mu}_{0,1:n+1} + \left(I_{n+1} \otimes \mathbb{E}_{\rho_\Lambda}[\Lambda]\right) \boldsymbol{Y}_{1:n+1}\right)\Bigg),
\end{aligned}$$

where we define $\boldsymbol{\mu}_{0,1:n+1}$ as the concatenated vector $(\mu_0(1), \dots, \mu_0(n+1))$, and $\boldsymbol{Y}_{1:n+1}$ the concatenated vector $(Y_1, \dots, Y_n, \xi_y)$. Therefore, we have

$$\log \rho_\mu(\mu) \propto \log \mathcal{MGP}(\mu; m_\mu(\cdot), \Omega_\mu, K_\mu),$$

where we define the mean function and covariance functions on indices $\{1, \dots n+1\}$,

$$\begin{aligned}
m_\mu^{1:n+1} &= \left(\Omega_\mu \otimes K_\mu^{1:n+1}\right)\left(\left(\Omega_0 \otimes \boldsymbol{K}_{\mathbf{0}}^{1:n+1}\right)^{-1} \boldsymbol{\mu}_{0,1:n+1} + \left(I_{n+1} \otimes \mathbb{E}_{\rho_\Lambda}[\Lambda]\right) \boldsymbol{Y}_{1:n+1}\right) \\
\left(\Omega_\mu \otimes K_\mu^{1:n+1}\right)^{-1} &= \left(I_{n+1} \otimes \mathbb{E}_{\rho_\Lambda}[\Lambda] + \left(\Omega_0 \otimes \boldsymbol{K}_{\mathbf{0}}^{1:n+1}\right)^{-1}\right).
\end{aligned}$$

Finally, the variational distribution ρ_Λ can be derived as

$$\begin{aligned}
\log \rho_\Lambda(\Lambda) &\propto_\Lambda \mathbb{E}_{\rho_y, \rho_\mu}\left[\prod_{t=1}^{n} \mathcal{N}(Y_y; \mu(t), \Lambda^{-1}) \mathcal{N}(Y_{n+1}; \mu(n+1), \Lambda^{-1}) \mathcal{W}(\Lambda; \nu_0, \psi_0)\right] \\
&\propto \underbrace{\mathbb{E}_{\rho_y, \rho_\mu}\left[-\frac{1}{2}\sum_{t=1}^{n} \left(Y_t - \mu(t)\right)^\top \Lambda \left(Y_t - \mu(t)\right) - \frac{1}{2}\left(Y_{n+1} - \mu(n+1)\right)^\top \Lambda \left(Y_{n+1} - \mu(n+1)\right)\right]}_{(*)} \\
&- \frac{1}{2}\mathrm{Tr}(\Lambda \psi_0^{-1}) + \frac{\nu_0 + n - d - 1}{2}\log|\Lambda|.
\end{aligned}$$

The term inside the expectation gives

$$
\begin{aligned}
(*) \propto_\Lambda & -\frac{1}{2}\mathrm{Tr}\Bigg(\Lambda\Bigg(\sum_{t=1}^{n} Y_t Y_t^\top - 2\sum_{t=1}^{n}\mathbb{E}_{\rho_\mu}\left[\mu(t)\right] Y_t^\top + \sum_{t=1}^{n}\mathbb{E}_{\rho_\mu}\left[\mu(t)\mu(t)^\top\right] + \mathbb{E}_{\rho_y}\left[Y_{n+1}Y_{n+1}^\top\right] \\
& - 2\mathbb{E}_{\rho_y}\left[Y_{n+1}\right]\mathbb{E}_{\rho_\mu}\left[\mu(n+1)\right] + \mathbb{E}_{\rho_\mu}\left[\mu(n+1)\mu(n+1)^\top\right]\Bigg)\Bigg) \\
& \propto -\frac{1}{2}\mathrm{Tr}\Bigg(\Lambda\Bigg(\sum_{t=1}^{n} Y_t Y_t^\top - 2\sum_{t=1}^{n} m_\mu(t) Y_t^\top + \sum_{t=1}^{n+1} m_\mu(t) m_\mu(t) + \xi_y\xi_y^\top + \Lambda_y^{-1} - 2\xi_y m_\mu(n+1) \\
& + \sum_{t=1}^{n+1}\mathrm{Cov}(\mu(t),\mu(t))\Bigg)\Bigg).
\end{aligned}
$$

Hence,

$$\log \rho_\Lambda(\Lambda) \propto \log \mathcal{W}(\Lambda;\nu_\Lambda,\psi_\Lambda)\,,$$

where

$$
\begin{aligned}
\nu_\Lambda &= \nu_0 + n + 1 \\
\psi_\Lambda &= \psi_0^{-1} + \xi_y\xi_y^\top + \Lambda_y^{-1} + \sum_{t\in[n]} Y_t Y_t^\top - 2\sum_{t\in[n]} m_\mu(t) Y_t^\top - 2m_\mu(n+1)\xi_y^\top + \sum_{t=1}^{n+1} m_\mu(t) m_\mu(t)^\top \\
&+ \sum_{t=1}^{n+1}\mathrm{Cov}(\mu(t),\mu(t))\,.
\end{aligned}
$$

■

Lemma 19 (Objective function under Gaussian Process model) *For any $\delta \in \mathcal{D}$, let $(\xi_y, \Lambda_y, m_\mu^{1:n+1}, (\Omega_\mu \otimes K_\mu)^{1:n+1}, \nu_\Lambda, \psi_\Lambda)$ the parameters of the corresponding variational distribution $\hat{\rho}_{\mathrm{VB}}$ under the* GP *model (12). Then, the objective function can is written as*

$$
\begin{aligned}
\mathcal{R}_{\mathcal{F}}(\delta) = & -\frac{\nu_\Lambda}{2}\mathrm{Tr}\Bigg(\Bigg(\psi_0^{-1} + \xi_y\xi_y^\top + \Lambda_y^{-1} + \sum_{t\in[n]} Y_t Y_t^\top - 2\sum_{t\in[n]} m_\mu(t) Y_t^\top - 2m_\mu(n+1)\xi_y^\top \\
& + \sum_{t=1}^{n+1} m_\mu(t) m_\mu(t)^\top + \sum_{t=1}^{n+1} Cov(\mu(t),\mu(t))\Bigg)\psi_\Lambda\Bigg) - \frac{1}{2}\mathrm{Tr}\left(\Omega\otimes K_0^{n+1}\right)^{-1}\left(m_\mu^{1:n+1}(m_\mu^{1:n+1})^\top + \Omega_\mu\otimes K_\mu^{n+1}\right) \\
& + (m_\mu^{1:n+1})^\top\left(\Omega_0\otimes K_0^{n+1}\right)^{-1}\mu_0^{1:n+1} + \frac{1}{2}(n+\nu_0+1)\log\det(\psi_\Lambda) - \frac{1}{2}\left(\log\det(\Lambda_y) + \log\det(\Omega_\mu\otimes K_\mu^{1:n+1})\right) \\
& - \lambda\delta^\top\xi_y
\end{aligned}
$$

Proof From Theorem 18, we found that the variational distribution for the AR model can be written as

$$\hat{\rho}_{\mathrm{VB}}(\mathrm{d}(Y_{n+1},\mu,\Lambda)) = \mathcal{N}(\mathrm{d}Y_{n+1};\xi_y,\Lambda_y^{-1})\mathcal{MGP}(\mathrm{d}\mu;m_\mu(\cdot),k_\mu(\cdot),\Omega_\mu)\mathcal{W}(\mathrm{d}\Lambda;\nu_\Lambda,\psi_\Lambda)\,,$$

Starting again from the definition of $\mathcal{R}_{\mathcal{M}}$, we have, for any $\delta \in \mathcal{D}$,

$$\begin{aligned}\mathcal{R}_{\mathcal{M}}(\delta) \propto_\delta & -\mathbb{E}_{\rho_y}\left[\log \rho_y(Y_{n+1})\right] - \mathbb{E}_{\rho_\mu}\left[\log \rho_\Gamma(\mu)\right] - \mathbb{E}_{\rho_\Lambda}\left[\log \rho_\Lambda(\Lambda)\right] - \lambda \delta^\top \mathbb{E}_{\rho_y}\left[Y_{n+1}\right] \\ & + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \pi\left(Y_{n+1}, \mu, \Lambda \,|\, H_n\right)\right] ,\end{aligned}$$

where

$$\begin{aligned}& - \mathbb{E}_{\rho_y}\left[\log \rho_y(Y_{n+1})\right] \propto_\delta -\frac{1}{2}\log|\Lambda_y| \qquad - \mathbb{E}_{\rho_\mu}\left[\log \rho_\mu(\mu)\right] \propto_\delta -\frac{1}{2}\log|\Omega_\mu \otimes K_\mu^{1:n+1}| \\ & - \mathbb{E}_{\rho_\Lambda}\left[\log \rho_\Lambda(\Lambda)\right] \propto_\delta \frac{d+1}{2}\log|\psi_\Lambda| .\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \pi(Y_{n+1}, \mu, \Lambda \,|\, H_n)\right] \propto_\delta & \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \prod_{t=1}^n \mathcal{N}(Y_t; \Gamma\mu(t), \Lambda^{-1})\right] + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{N}(Y_{n+1}; \mu(n+1), \Lambda^{-1})\right] \\ & + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{MGP}(\mu; \mu_0, K_0, \Omega_0)\right] + \mathbb{E}_{\hat{\rho}_{\mathrm{VB}}}\left[\log \mathcal{W}(\Lambda; \nu_0, \psi_0\right] .\end{aligned}$$

We can compute each of these terms exactly the same way we did in the proof of Theorem 18, and combining these with the terms above give the desited expression. ■

Appendix E. Derivation of `MCMC-Portfolio` for Specific Models

We derive specific instances of Algorithm 2 for both the GW and AR models, with a particular focus on detailing the form of the conditional posteriors.

E.1. GW Model

The joint parameter posterior $\pi(\mu, \Lambda \,|\, H_n)$ cannot be derived in closed-form, but we have the following conditional posteriors

$$\begin{aligned}\pi(\mathrm{d}\mu \,|\, \Lambda, H_n) &= \mathcal{N}\left(\mathrm{d}\mu \,;\, \left(\Lambda + \frac{1}{n}\Lambda_0\right)^{-1}\left(\Lambda \frac{1}{n}\sum_{t=1}^n Y_t + \frac{1}{n}\Lambda_0\mu_0\right), \frac{1}{n}\left(\Lambda + \frac{1}{n}\Lambda_0\right)^{-1}\right) \\ \pi(\mathrm{d}\Lambda \,|\, \mu, H_n) &= \mathcal{W}\left(\mathrm{d}\Lambda \,|\, \mu, H_n\right) = \mathcal{W}\left(\mathrm{d}\Lambda \,;\, n + \nu_0, \left(\sum_{t=1}^n (Y_t - \mu)(Y_t - \mu)^\top + \psi_0^{-1}\right)^{-1}\right) .\end{aligned}$$

Applying Gibbs sampling with these two conditional posteriors yield to a chain $(\mu^{(k)}, \Lambda^{(k)})_{k=1}^M$. For a given δ, the distribution $\check{\pi}_k$ is defined as

$$\check{\pi}_k(\mathrm{d}Y_{n+1}) = \mathcal{N}\left(\mathrm{d}Y_{n+1}; \mu^{(k)} - \lambda\Sigma^{(k)}\delta, \Sigma^{(k)}\right) .$$

Hence, the algorithm `MCMC-Portfolio`(GW) is defined as follows:

Algorithm 3 `MCMC-Portfolio` (GW): Portfolio Construction with MCMC for GW model.

Input: Dataset H_n, initial decision $\hat{\delta}^{(0)}$, number of Monte-Carlo samples M, risk parameter λ, step-size η, initial parameters $(\mu^{(0)}, \Lambda^{(0)})$.

while *Not converging* **do**

  for $k = 1, \ldots, M$ **do**

    $\mu^{(k)} \sim \pi(\mathrm{d}\mu \,|\, \Lambda^{(k-1)}, H_n)$ and $\Lambda^{(k)} \sim \pi(\mathrm{d}\Lambda \,|\, \mu^{(k)}, H_n)$.

  end

  For all $k \in [M]$, sample $z^{(k)} \sim \mathcal{N}\left(\mathrm{d}Y_{n+1}; \mu^{(k)} - \lambda\Sigma^{(k)}\delta, \Sigma^{(k)}\right)$.

  $\hat{\delta}^{(k+1)} \leftarrow \mathrm{Proj}_{\mathcal{D}}\left(\hat{\delta}^{(k)} + \eta\lambda\frac{1}{M}\sum_{k\in[M]} z^{(k)}\right)$

end

Return $\hat{\delta}^{(\infty)} = \hat{\delta}^{\mathrm{MCMC}}$.

E.2. AR Model

Here again, the joint posterior distribution $\pi(\Gamma, \Lambda \,|\, H_n)$ cannot be computed in closed-form; we can compute the conditional posterior as

$$\begin{aligned}
\pi(\mathrm{d}\Gamma \mid H_n, \Lambda) &\propto \pi(H_n \mid \Gamma, \Lambda)\pi(\mathrm{d}\Gamma) \\
&\propto \prod_{t=1}^{n} \exp\left(-\frac{1}{2}\left((\Gamma Y_{t-1})^\top \Lambda(\Gamma Y_{t-1}) - 2(\Gamma Y_{t-1})^\top \Lambda Y_t\right)\right)\pi(\mathrm{d}\Gamma) \\
&\propto \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\Gamma^\top\Lambda\Gamma G_n - 2\Gamma^\top\Lambda\sum_{t=1}^{n} Y_t Y_{t-1}\right)\right)\pi(\mathrm{d}\Gamma) \\
&\propto \exp\left(-\frac{1}{2}\left(\mathrm{vec}(\Gamma)^\top(G_n\otimes\Lambda)\mathrm{vec}(\Gamma) - 2\mathrm{vec}(\Gamma)^\top(I\otimes\Lambda)\mathrm{vec}(\sum_{t=1}^{n} Y_t Y_{t-1})\right)\right. \\
&\times \exp\left(-\frac{1}{2}\left(\mathrm{vec}(\Gamma)^\top(G_n\otimes\Lambda)^{-1}\mathrm{vec}(\Gamma)\right) - 2\mathrm{vec}(\Gamma)^\top(V_0\otimes U_0)^{-1}\mathrm{vec}(M_0)\right)\Bigg) \\
&\propto \exp\Bigg(-\frac{1}{2}\Bigg(\mathrm{vec}(\Gamma)^\top\left((G_n\otimes\Lambda)^{-1} + (V_0\otimes U_0)^{-1}\right) \\
&- 2\mathrm{Vec}(\Gamma)^\top\left((I\otimes\Lambda)\mathrm{vec}(\sum_{t=1}^{n} Y_t Y_{t-1}) + (V_0\otimes U_0)^{-1}\mathrm{vec}(M_0)\right)\Bigg)\Bigg)
\end{aligned}$$

Therefore, we have $\pi(\mathrm{d}\Gamma \,|\, H_n, \Lambda) = \mathcal{N}\left(\mathrm{dvec}(\Gamma)\,;\, \mu_\Gamma(\Lambda), \Sigma_\Gamma(\Lambda)\right)$, where

$$\begin{aligned}
\Sigma_\Gamma(\Lambda) &= \left((G_n\otimes\Lambda)^{-1} + (V_0\otimes U_0)^{-1}\right)^{-1} \\
\mu_\Gamma(\Lambda) &= \Sigma_\Gamma(\Lambda)\left((I\otimes\Lambda)\mathrm{vec}(\sum_{t=1}^{n} Y_t Y_{t-1}) + (V_0\otimes U_0)^{-1}\mathrm{vec}(M_0)\right).
\end{aligned}$$

We also have the conditional posterior for the precision matrix,

$$\pi\left(\mathrm{d}\Lambda \,|\, \Gamma, H_n) = \mathcal{W}(\mathrm{d}\Lambda\,;\, n+\nu_0, \left(\sum_{t=1}^{n}(Y_t - \Gamma Y_{t-1})(Y_t - \Gamma Y_{t-1})^\top + \psi_0^{-1}\right)^{-1}\right)$$

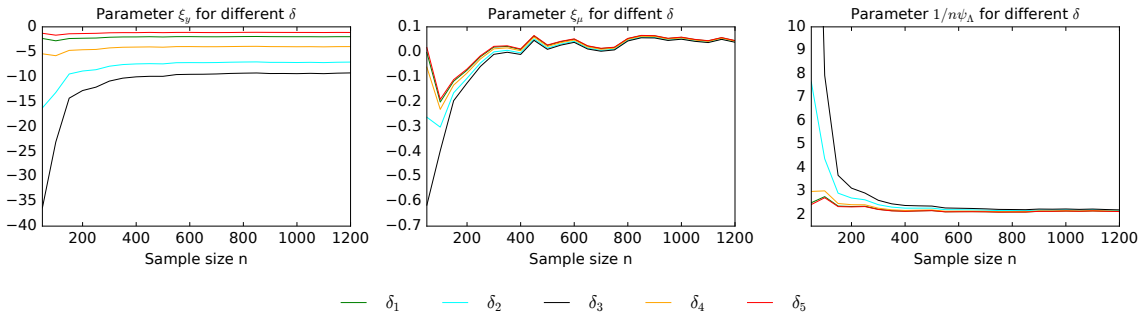

Figure 3: Convergence of variational parameters $(\xi_y, \xi_\mu, \psi_\Lambda)$ with respect to the sample size n in dimension $d = 1$ for the GW model. We take 5 different values of δ randomly (5 different colors).

Therefore, we can have samples from the joint posterior distribution $\pi(\delta(\Gamma, \Lambda) \mid H_n)$. For a given sample k, conditionally on the posterior samples $(\Gamma^{(k)}, \Lambda^{(k)})_{k=1}^M$, and the distribution $\check{\pi}_k$ is given by

$$\check{\pi}_k(\mathrm{d}Y_{n+1}) \propto e^{-\lambda^\top Y_{n+1}} \mathcal{N}(\mathrm{d}Y_{n+1}\ ;\ \Gamma^{(k)} Y_n, (\Lambda^{(k)})^{-1}) \propto \mathcal{N}\left(Y_{n+1}\ ;\ \Gamma^{(k)} Y_n - \lambda(\Lambda^{(k)})^{-1}\delta, (\Lambda^{(k)})^{-1}\right)$$

The algorithm `MCMC-Portfolio`(AR) can be instantiated as follows:

Algorithm 4 `MCMC-Portfolio` (AR): Portfolio Construction with MCMC for AR model.

Input: Dataset H_n, initial decision $\hat{\delta}^{(0)}$, number of Monte-Carlo samples M, risk parameter λ, step-size η, initial parameters $(\Gamma^{(0)}, \Lambda^{(0)})$.
while *Not converging* **do**
  for $k = 1, \ldots, M$ **do**
    $\Gamma^{(k)} \sim \pi(\mathrm{d}\Gamma \,|\, \Lambda^{(k-1)}, H_n)$ and $\Lambda^{(k)} \sim \pi(\mathrm{d}\Lambda \,|\, \Gamma^{(k)}, H_n)$.
  end
  For all $k \in [M]$, sample $z^{(k)} \sim \mathcal{N}\left(\mathrm{d}Y_{n+1}; \Gamma^{(k)} Y_n - \lambda \Sigma^{(k)} \delta, \Sigma^{(k)}\right)$.
  $\hat{\delta}^{(k+1)} \leftarrow \mathrm{Proj}_{\mathcal{D}}\left(\hat{\delta}^{(k)} + \eta\lambda \frac{1}{M} \sum_{k \in [M]} z^{(k)}\right)$
end
Return $\hat{\delta}^{(\infty)} = \hat{\delta}^{\mathrm{MCMC}}$.

Appendix F. Additional Numerical Experiments

The code is provided in the supplementary material.

F.1. Numerical Discussions on Assumption 1

We evaluate the convergence of the variational parameters as the sample size n increases. Specifically, we generate synthetic data with dimension $d = 1$ under the GW model, and for values of $n \in [50, 1200]$, we compute the variational parameters $(\xi_y, \xi_\mu, \psi_\Lambda)$ for different random values of δ, using the corresponding fixed-point computation. Figure 3 illustrates

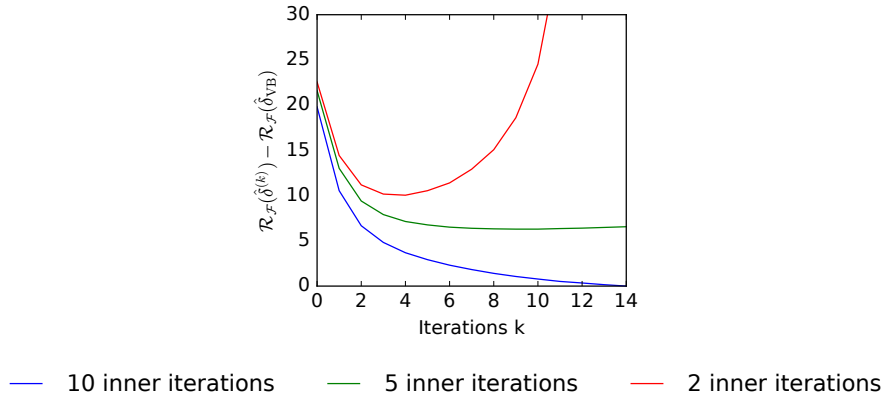

Figure 4: Convergence of the difference $\mathcal{R}_{\mathcal{F}}(\hat{\delta}_{\mathrm{VB}}) - \mathcal{R}_{\mathcal{F}}(\delta^{(k)})$ with respect to the kth iteration of the algorithm on the GW model. We repeat the experiment for a different amount of inner iterations. We set $d = 50$.

that the variational distribution derived from the fixed-point equation converges to the variational distribution obtained from the asymptotic fixed-point operator.

F.2. Number of Fixed-point Iterations

We evaluate the difference in the value function, $\mathcal{R}_{\mathcal{F}}(\hat{\delta}_{\mathrm{VB}}) - \mathcal{R}_{\mathcal{F}}(\delta^{(k)})$, of our algorithm when applied to the GW model on a synthetic dataset. We set $d = 50$ and fix $n = 200$, and examine the impact of the number of inner iterations performed on the fixed-point computation. Figure 4 demonstrates that insufficient inner iterations result in non-convergence of the value function $\mathcal{R}_{\mathcal{F}}(\hat{\delta}^{(k)})$, as the supremum in the objective function is not reached.

F.3. Computational Complexities

By employing the mean-field assumption, our algorithm `VB-Portfolio` simplifies the optimization problem from a measure-based setting (Equation (5)) to a finite-dimensional parametric optimization problem. The primary computational burden lies in the inversion of (d, d) matrices for both the stationary and autoregressive Gaussian-Wishart models, resulting in a computational complexity of $\mathcal{O}(d^3)$. In contrast, the Gaussian Process (GP) model faces scalability challenges, as it requires the inversion of (nd, nd) matrices, leading to cubic complexity with respect to both dimension d dataset size n. Existing works such as those proposed by Quinonero-Candela and Rasmussen (2005) aim at reducing this inversion complexity; we leave this challenging task as future work.

Ethics and Accessibility

Although our work is mainly theoretical, applications can vary depending on the user, and the same model might have different effects for individual investors, institutions, or automated systems. Making sure these methods are used in a fair and transparent way, and

that they are understandable and accessible to a wide range of users, is an important ethical concern.