Bayesian Techniques Applied in Counterfactual Analysis with High Dimensional Settings and Time Varying Properties *

Yaohan Chen

School of Economics, Singapore Management University

Last version: June 12, 2020

This version: December 15, 2020

Abstract

Bayesian methodologies are rapidly becoming tools in machine learning analysis. Among all these conventional and recently developed Bayesian analysis tools, Expectation Maximization (EM) algorithms plays the pivotal role in modern Bayesian analysis and so is its extension Variational Expectation and Maximization (VEM) algorithm. Both of these methodologies have been widely used in the application of modern Bayesian analysis in machine learning such as textual analysis established on Latent Dirichlet Allocation (LDA), which is usually referred to Topic Models. This note mainly discusses the modern technical tools and analytical framework for application of modern Bayesian analysis along with the associated properties with specific focus on Topic Models and how EM algorithm is set up from Variational Bayesian perspective.

Keywords: Bayesian; Machine Learning; Asset pricing; Dynamic Covariance Structure

* I am grateful for the discussion kindly shared literatures with Associate Professor Tao Zeng from Zhejiang University. All errors are my own. You may contact me at yaohan.chen.2017@phdecons.smu.edu.sg.

1 Introduction

Variational Bayes methods and Empirical Bayes methods (Armstrong et al., 2020) have recently gained much attention in academia especially in machine learning literature since under many circumstances shrinkage has to be imposed to improve the both the in-sample fitting and out-of-sample prediction accuracy. Another major reason for the increasing popularity of this analysis framework within academia is that currently there is no general way to account for uncertainties without imposing many parametric assumptions, this is actually an issue that has been pointed in some recent literature like Hansen (2016, p.116). Bayesian analysis instead offers a flexible and philosophically coherent framework based on posterior analysis for learning with data rich in hierarchical structure. However in most practical application settings, the posterior is in general intractable and consequently some modern sampling algorithms like MCMC have to be adopted for approximating the target posterior based on the sampling. One key feature of sampling methods like MCMC is the intrinsically required heavy block-loop structure and hence very often computational efficiency is deteriorated in large-scale applications. By comparison, Variational Bayes (VB) method as an alternative for sampling algorithm like MCMC has been empirically proved to be able to circumvent the computational inefficiency often encountered in sampling algorithm while keeping desired approximation as much as possible (Braun and McAuliffe, 2010; Blei, Kucukelbir, and McAuliffe, 2017). VB actually is a specific methodology originated from the more extensive variational approximation analytical framework which can be at least traced back to (Jordan et al., 1999; Winn and Bishop, 2005). One of the major reasons for the documented computation efficiency gained from VB, which is one of the main attractive properties justifying its increasing popularity in large-scale application as well, is the corresponding feature in formulating the posterior approximation as a specific optimization problem. Briefly speaking, the rough idea underlying the core of VB is to find the closest distribution to the exact posterior over some family of distributions (Guo, Wang, Fan, Broderick, and Dunson, 2016). Recently there is also some work justifying the ground theory for VB (Wang and Blei, 2019; Alquier and Ridgway, 2020), where they have formerly defined the VB methodology and related theoretical properties of VB approximation (specifically all these work discuss the frequentist concentration of VB approximation for posterior under different divergence concepts). In fact VB is an extensive framework such that many classical algorithms like EM algorithm are possible to be included in the discussion within this framework. As for the development of modern empirical methods in Economics and Finance, Armstrong, Kolesár, and Plagborg-Møller (2020) extends empirical Bayes confidence interval of Morris (1983) in a robust way free of the prior distribution imposed on means.

Another reason why Variational Bayes method has attained much attention in economics literature is the development of economic models with intrinsic requirement to incorporate stochastic variation of parameters since Cooley and Prescott (1976). In recent decades, accompanied with the fast development of data science and machine learning techniques, econometric models with time-varying parameters and high-dimensional data-structure are broadly acknowledged in academia

as an important topic, there has been a large amount of recent literature established on this framework.

Another area where such a kind of analytical framework is possibly to be applied is counterfactual analysis with synthetic controls in high-dimensional settings. Seminal work can be traced back to Abadie and Gardeazabal (2003) and Abadie et al. (2010). Recently there is a nice and comprehensive review work on this framework by Abadie (2020). Ever since the synthetic control method was proposed in Abadie and Gardeazabal (2003) for the first time, it has received widespread attention both in economic and machine learning literature. The reason why counterfactual analysis is an important tool for social science is that under many circumstances causality is of the most importance to be discussed in social science researches (for instance there is growing literature in finance discussing the publication effect of anomaly-based trading strategies (Mclean and Pontiff, 2016; Chen and Zimmermann, 2020b; Pelger and Xiong, 2020)) and usually the casual statements with respect to treatment (intervention) depends on the construction of counterfactuals, which are typically the unobserved outcomes that would have been if a unit had not been treated. Routinely, estimation of counterfactual effects of this kind is based on the aforementioned synthetic control method that makes the desired estimation upon the estimation of outcomes of similar group of individuals not affected by the intervention. This intuitive idea driving synthetic method is similar to the comparative studies widely adopted in other social science where the key idea is laid upon that the effect of intervention can be inferred by comparing the evolution of the target (outcome) variables of interest between the unit exposed to treatment and a group of units that are similar to the unit but were not affected by the treatment. A inherent key factor for ensuring the validity comparative analysis is the affinity between unit exposed to treatment and the units not exposed to treatment, which is however not that easy to be justified in practice. There is an emerging literature on applying machine learning methods in recovering the unobserved counterfactual effects from synthetic controls (Doudchenko and Imbens, 2017; Chernozhukov et al., 2020a,b; Athey et al., 2020).

One alternative for counterfactual construction discussed in Carvalho et al. (2018), the so-called `ArCo` method, is essentially a two-step procedure: In the first step, the data before the occurrence of the intervention is used to estimate a multivariate time-series regression model relating the variables in the treated (dependent variables) with the variables belonging to the untreated peers (explanatory variables); In the second step, the counterfactual is constructed by extrapolating the estimated model with data after intervention. Finally the estimated effect is the time-series average of the difference between the data actual data and counterfactual.

A natural follow-up question would be whether it is possible to account for the potential time-varying properties.

2 Basic Setup and Assumptions

2.1 Notation

All random variables are defined in a Probability Space $(\Omega, \mathcal{F}, \mathbb{P})$. Random variables are defined by upper case letter and the associated realization is defined by lower case letter such that $X(\omega) = x$. Matrices and vectors are written in bold letters $\mathbf{X}$. Hence each entry of the realized $\mathbf{X}$, i.e., x_{ij}, denotes the i-th realization (row) of the j-th random variable (column). Sets are denoted by calligraphic upper case $\mathcal{X}$ and the associated cardinality is denoted by $|\mathcal{X}|$. For the symbol of norm, we use $\|\cdot\|$ to denote the generic (semi)norm; $\|\cdot\|_q$ and $\|\cdot\|_{\mathcal{L}^q}$ are adopted to denote ℓ^q and $\mathcal{L}^q$ norms respectively for $q \in [1, \infty)$. Specifically these two norms are defined separately for the cases when $d > 1$ and $d = 1$ when the vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ degenerates to a scalar (For this case we just use X to denote the degenerated scalar random variable). That is

$$\begin{aligned} \|\mathbf{X}\|_q &:= \left(\sum_{i=1}^{d} |X_i|^q \right)^{1/q} \\ \|X\|_{\mathcal{L}^q} &:= \left(\mathbb{E}|X|^q \right)^{1/q} \end{aligned}$$

We further introduce other notation for norm such that

$$\begin{aligned} \|\mathbf{X}\|_\infty &:= \max_{i \leqslant d} |X_i| \text{ if } \mathbf{X} \text{ is } d \times 1 \text{ vector} \\ \|\mathbf{X}\|_{\max} &:= \max_{i \leqslant m, j \leqslant n} |X_{i,j}| \text{ if } \mathbf{X} \text{ is } m \times n \text{ matrix} \end{aligned}$$

Furthermore, we introduce the following notation (ℓ^0 norm)

$$\|\mathbf{X}\|_0 := |\{i : X_i \neq 0\}|$$

and quadratic form associated with d-dimensional matrix $\mathbf{M}$ takes the following form

$$\|\mathbf{X}\|_{\mathbf{M}}^2 := \mathbf{X}^\top \mathbf{M} \mathbf{X}$$

And we also adopt $\operatorname{diag}(\mathbf{X})$ to denote the diagonal matrix whose diagonal elements are extracted from the diagonal of $\mathbf{X}$; $\mathbf{1}(A)$ represents an indicator function on the event A, i.e., $\mathbf{1}(A) = 1$ if A is true or $\mathbf{1}(A) = 0$, otherwise. Yes you can build a VAR model for the target variable and the covariates using the approach by Koop and Korobilis. Then the estimated model together with updated values for the covariates can be used to estimate counterfactual values of the target variable. In the following discussion, I may use term "intervened" or "treated" interchangeably with the same meaning that intervention has taken place at a specific period.

Definition 1 *The probability density function for the random matrix* $\mathbf{X}$ $(n \times p)$ *that follows the*

matrix normal distribution $\mathcal{MN}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ *has the form:*

$$p(\mathbf{X} \mid \mathbf{M}, \mathbf{U}, \mathbf{V}) = \frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left[\mathbf{V}^{-1}(\mathbf{X}-\mathbf{M})^{\top}\mathbf{U}^{-1}(\mathbf{X}-\mathbf{M})\right]\right)}{(2\pi)^{np/2}|\mathbf{V}|^{n/2}|\mathbf{U}|^{p/2}} \tag{1}$$

The matrix norm is related to the multivariate normal distribution in the following way

$$\mathbf{X} \sim \mathcal{MN}(\mathbf{M}, \mathbf{U}, \mathbf{V}) \tag{2}$$

if and only if

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

where $\mathbf{V}$ *is* $p \times p$ *matrix and* $\mathbf{U}$ *is* $n \times n$ *matrix respectively.*

2.2 Counterfactual analysis and synthetic control

The procedure implemented in counterfactual analysis with synthetic controls is briefly discussed in this subsection. We assume the target variable is a scalar denoted by Y_t, which is exposed to a treatment (intervention) that occurred at $t = T_0 + 1$. Counterfactual effect is then estimated from the covariates $\mathbf{X}$ ($\mathbf{X}$ is generally a $P \times J$ matrix with the j-th column indicating the j-th unit, either treated or not; and for each column as a $P \times 1$ column vector refers to the vector collecting the covariates associated with each unit), i.e., characteristics of peers that are assumed to be unaffected by the intervention. Realization of real valued random vector is observed, which is denoted by

$$\mathbf{Z}_t = (Z_{1t}, \ldots, Z_{J+1,t})^{\top} = \left(\mathbf{Z}_{0t}^{\top} \vdots \mathbf{Z}_{1t}^{\top}\right)^{\top}$$

where we just introduce notation $\mathbf{Z}_{0t}^{\top}$ and $\mathbf{Z}_{1t}^{\top}$ as two sub column vectors of $\mathbf{Z}_t$ to emphasize that the observed random vector generally can be separated into treated group ($\mathbf{Z}_{1t}^{\top}$ accordingly) and untreated group ($\mathbf{Z}_{1t}^{\top}$ accordingly). $\mathcal{D}_t \in \{0, 1\}$ is introduced as a sequence of binary variable indicating the periods when treatment (intervention) was in place, that is

$$\mathcal{D}_t = \begin{cases} 1 & \text{if } t > T_0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

For each period t, the observed Z_{it} could be either from $Z_{it}^{(1)}$ which is the potential outcome when the unit i is exposed to the intervention or from $Z_{it}^{(0)}$ which is the potential outcome when the unit i is not exposed to the intervention. Hence we are able to jointly express $Z_{it} \in \mathbf{Z}_{1t}$ as

$$Z_{it} = \mathcal{D}_t Z_{it}^{(1)} + (1 - \mathcal{D}_t) Z_{it}^{(0)} \tag{4}$$

Without loss of generality, we may focus on the case when there is a single variable contained treated group and denote this scalar $Y_t = Z_{1t}$ as target variable; All the remained J units from donor

pool (peers assumed to be not affected by intervention) are collected in $\mathbf{Z}_{0t}$, that is

$$\mathbf{Z}_{0t} = (Z_{2t}, \ldots, Z_{J+1,t})^{\top}$$

Given this restricted case when Z_{1t} is the target variable potentially exposed to intervention, we are interested in evaluating what would Z_{1t} have been like had there been no intervention, i.e. we are interested in obtaining the evaluation of $Z_{1t}^{(0)}$ by approximating it using the following functional form [1]

$$Z_{1t}^{0} = \mathcal{M}\left(\mathbf{Z}_{0t}^{(0)}; \boldsymbol{\theta}_0\right) + V_t, \quad t = 1, \ldots, T \tag{5}$$

where

$$\mathcal{M} : \mathcal{Z} \times \Theta \to \mathbf{R}, \ \mathcal{Z} \subseteq \mathbf{R}^{n-1}$$

and then the estimated evaluation of $Z_{1t}^{(0)}$ is

$$\widehat{Z}_{1t}^{(0)} = \mathcal{M}\left(\mathbf{Z}_{0t}^{(0)}; \widehat{\boldsymbol{\theta}}_{T_0}\right)$$

The remained question of potential research interest is what exactly the functional form of $\mathcal{M}(\cdot)$ we are able to apply to approximate the target unobserved counterfactual effect. Machine learning method as one alternative of the statistician toolkit is widely known for its ability at this kind of objective, for instance the recently developed machine learning algorithms taking the advantage of Bayesian techniques featuring the tree structure (Chipman et al., 1998; Denison et al., 1998; Chipman et al., 2010; Ročková and van der Pas, 2019; Ročková, 2019; Ročková and Saha, 2019). Alternatively it is also worthwhile making an attempt to apply the framework proposed by Koop and Korobilis (2020) to model $\mathcal{M}(\cdot)$, since this would potentially bring the advantage of accounting for the time-varying properties of $\mathcal{M}(\cdot)$.

Alternatively, we may make our analysis restricted to linear framework and accordingly some theoretic properties are relatively easy to establish, hence this is by far the framework within which most of the discussions are made in literature. In particular, within linear framework, the general objective (Abadie, 2020) is to uncover the optimized $\mathbf{W}^{*}$ such that for a given $\mathbf{V}$ the following objective function is minimized

$$\|\mathbf{X}_1 - \mathbf{X}_0\mathbf{W}\| = \left(\sum_{p=1}^{P} v_p \left(X_{p1} - w_2 X_{p2} - \cdots - w_{J+1} X_{p,J+1}\right)^2\right)^{1/2} \tag{6}$$

[1] As argued in Abadie (2020), one of the key feature of synthetic control method is the implicit assumption that the combination of units in the donor pool may approximate the characteristics of the the affected unit substantially better than any unaffected unit alone, and this serves as one motivating reason why we want to introduce the general functional form of units in donor pool.

where

$$\begin{aligned}\mathbf{W} &= (w_2, \ldots, w_J)^\top \\ \mathbf{V} &= (v_1, \ldots, v_k)^\top\end{aligned}$$

and we abbreviate the timing subscript for both $\mathbf{X}_{1t}$ and $\mathbf{X}_{0t}$ with focus on the cross-sectional dimension. Then the functional form of counterfactual effect within this framework is

$$\widehat{Z}_{1t}^{(0)} = \mathcal{M}\left(\mathbf{Z}_{0t}; \widehat{\boldsymbol{\theta}}_{T_0}\right) = \sum_{j=2}^{J+1} w_j^* Z_{jt} \tag{7}$$

One of the key assumption justifying our discussion is

$$Z_{1t} \perp\!\!\!\perp \mathcal{D}_s \text{ for all } t, s.$$

Once we are able to back out $Z_{1t}^{(0)}$ as precise as possible, then conceptually the following difference is used as the proxy for evaluating counterfactual effects,

$$\mathcal{H}_0 : \delta_t \coloneqq Z_{1t}^{(1)} - Z_{1t}^{(0)} \quad \text{for } t > T_0 \tag{8}$$

and corresponding estimation for δ_t along with it is given as

$$\widehat{\delta}_t = Z_{1t} - \widehat{Z}_{1t}^{(0)}$$

and finally the evaluated difference due to the intervention is

$$\widehat{\Delta}_T = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \widehat{\delta}_t$$

and the associated $\frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \delta_t$. In this note I also attach the replicated figures separately in Figure 4(a) and Figure 4(b) as following. As emphasized in Abadie (2020), synthetic control method is essentially a weighted average algorithm using the weighted average of characteristics of units from donor pool (i.e. sometimes it is referred to fitting) to approximate the unobserved counterfactual effect of target variable by restricting the weights on probability simplex. As we have always emphasized, the goal of counterfactual analysis or synthetic control is to approximate the trajectory that would have been observed for $Y_t = Z_{1t}$ at $t > T_0$ in the absence of intervention.

To justify the intrinsic advantage of synthetic control method in comparison to other methodologies, it would be useful to compare synthetic control method with linear regression framework. The way we adhere to using linear regression is augmenting the characteristic space of units from treated group and untreated group by adding a row of ones to $\mathbf{X}_1$ and $\mathbf{X}_0$ respectively, denoted by $\overline{\mathbf{X}}_1$ and $\overline{\mathbf{X}}_0$. Further more, let $\mathbf{Y}_0$ as the $(T - T_0) \times J$ matrix denote the observed effects for J units

from untreated group. The regression of $\mathbf{Y}_0$ on $\overline{\mathbf{X}}_0$ yields the estimated counterfactual effect based on regression as

$$\widehat{\mathbf{B}}^\top \overline{\mathbf{X}}_1 \tag{9}$$

where $\widehat{\mathbf{B}}$ collects the regression coefficients in the following form

$$\widehat{\mathbf{B}} = \left(\overline{\mathbf{X}}_0 \overline{\mathbf{X}}_0^\top\right)^{-1} \overline{\mathbf{X}}_0 \mathbf{Y}_0^\top$$

Hence this implies that the estimated counterfactual effect is

$$\underbrace{\mathbf{Y}_0 \overline{\mathbf{X}}_1^\top \left(\overline{\mathbf{X}}_0 \overline{\mathbf{X}}_0^\top\right)^{-1} \overline{\mathbf{X}}_1}_{\mathbf{W}^{\text{reg}}} \tag{10}$$

Remark 2.1 *For the restricted case where there is a single unit in the treated group, both* $\mathbf{X}_1$ *and* $\overline{\mathbf{X}}_1$ *are* $P \times 1$ *column vector, which implies the final estimated counterfactual effect collected in* $\mathbf{Y}_0 \overline{\mathbf{X}}_1^\top \left(\overline{\mathbf{X}}_0 \overline{\mathbf{X}}_0^\top\right)^{-1} \overline{\mathbf{X}}_1$ *is a* $(T - T_0) \times 1$ *column vector with each entry indicating the estimated counterfactual effect at each period after intervention.*

Remark 2.2 *One prominent feature of the prevalent synthetic analysis framework is the underlying implicit assumption that researchers are informed about how the treated and untreated group is divided, but it seems not always the case in practice. Actually once the intervention is imposed, which units are exposed to the intervention generally is not known for researchers. Or in other words, for the target variable that is supposed to be contained in treated group, the corresponding treated effect on the target variable is time-varying (after the period when the intervention is imposed). Hence the counterfactual effect like a "phantom" should be treated as latent variable in general, and recently there is an inspiring work by Bojinov and Shepard (2019). In this work they have emphasized the timing scheme of counterfactual effect, which is demonstrated in the following adapted figure*

[Place Figure 1 about here]

The basic idea is as following: it is obvious from the figure that for a realized sequence of target variable(s), the status whether it is (they are) affected by the intervention (treatment) is totally determined by the unobserved path

$$\mathbf{W}_{1:T} = (W_1, \ldots, W_T)$$

where each element W_t $(1 \leqslant t \leqslant T)$ *as the binary variable indicating whether the associated target variable is exposed to intervention or not.* W_t $(1 \leqslant t \leqslant T)$ *is also referred to treatment path in literature.*

An straightforward extension in Bayesian framework is by considering the Dirichlet allocation or the variational bayes method widely adopted in Bayesian literature. Since all these techniques

automatically satisfy the inherent requirement of synthetic control method that the average weights have to be added up to 1. Moreover, as has been widely discussed in statistical literature, Dirichlet allocation intrinsically features the sparsity desired in the application of synthetic method. A specific discussion within Bayesian framework is discussed as following. Moreover, it is a natural idea to regard the unobserved counterfactual of units from treated group as latent variables within Bayesian framework.

Generally, there are two kinds of available information (data) for backing out the counterfactual effect: (i) the inherent trend information of target variable(s) contained in treated group, which is potentially able to be extracted by using (V)AR model; (ii) information contained in covariates collected in donor pool (untreated group), information of this kind in general features the dynamic of covariates. Finally the evaluated counterfactual effect can be written as the linear combination of the counterfactual effects estimated from these two kinds of information. Formally in mathematics,

$$\begin{cases} \mathbf{z}_{1t} = A\mathbf{z}_{1t-1} + \boldsymbol{\epsilon}_t & (11) \\ \mathbf{z}_{0t} = A\mathbf{x}_{0t} + \boldsymbol{\epsilon}_t & (12) \end{cases}$$

3 Latent Dirichlet Allocation

3.1 Basic concepts

As proposed in Blei et al. (2003), some basic concepts about text documents are introduced as following

- A **word** refers to a item drawn from vocabulary indexed by $\{1, \ldots, V\}$, each word is represented by a V-vector such that if this word belongs to the v-th term then w is specifically a vector $\mathrm{w}^v = 1$ and $\mathrm{w}^u = 0$ if $u \neq v$.
- A **document** refers to a sequence of N words collected in $\mathbf{w} = (\mathrm{w}_1, \ldots, \mathrm{w}_N)$, where w_n is the n-th word in sequence.
- A **corpus** is a collection of M documents denoted by $D = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$.

3.2 Estimation

Basically, Latent Dirichlet Allocation (LDA) assumes the following generating process,

Step 1 The term distribution $\boldsymbol{\beta}$ is determined for each topic by

$$\boldsymbol{\beta} \sim \mathrm{Dir}(\delta)$$

Step 2 Choose $\theta \sim \mathrm{Dir}(\alpha)$.

Step 3 For each of N words w_n:

3a) Choose a topic $\mathrm{z}_n \sim \text{Multinomial}(\theta)$.

3b) Choose a word w_n from $p(\mathrm{w}_n \mid \mathrm{z}_n, \boldsymbol{\beta})$, a multinomial probability conditional on the topic z_n.

where $\boldsymbol{\beta}$ refers to a word probabilities matrix of dimension $k \times V$ with,

$$\beta_{ij} = \mathrm{P}\left(\mathrm{w}^j = 1 \mid \mathrm{z}^i = 1\right).$$

Specifically more in details, a k-dimensional Dirichlet random variable can take values in the $(k-1)$ simplex (a k vector lies in the $(k-1)$-simplex if $\theta_i \geqslant 0$ and $\sum_{i=1}^{k} \theta_i = 1$), and has the following probability density function over its simplex

$$\begin{aligned} p(\theta \mid \alpha) &= \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma\left(\alpha_i\right)} \theta_i^{\alpha_i - 1} \cdots \theta_k^{\alpha_k - 1} \\ &= \frac{\Gamma(\alpha k)}{\Gamma(\alpha)^k} \prod_{i=1}^{k} \theta_i^{\alpha - 1} \text{ (A special case in which } \alpha_1 = \cdots = \alpha_k) \end{aligned}$$

Remark 3.1 *How Dirichlet distribution is updated based on data observations. Suppose that random variables are drawn from multinomial distribution (thus sampling from $1, \cdots, k$ with repetition) with the sampling probability drawn from the special case of Dirichlet distribution mentioned in the context. There are totally N-times samplings such that*

$$x_1 +, \ldots, +x_i +, \ldots, +x_k = N \text{ and } 1 \leqslant i \leqslant k$$

where x_i represents how many times k is selected. The joint density is

$$\begin{aligned} p(\mathbf{x}, \theta \mid \alpha) &= \frac{N!}{x_1!, \cdots, x_k!} \prod_{i=1}^{k} \theta_i^{x_i} \cdot \frac{\Gamma(\alpha k)}{\Gamma(\alpha)^k} \prod_{i=1}^{k} \theta_i^{\alpha - 1} \\ &= \frac{N!}{x_1!, \cdots, x_k!} \frac{\Gamma(\alpha k)}{\Gamma(\alpha)^k} \cdot \prod_{i=1}^{k} \theta_i^{x_i + \alpha - 1} \end{aligned} \tag{13}$$

where (13) is of the functional form of the density function of Dirichlet distribution, hence it is easy to obtain the marginalization of the distribution of data $\mathbf{x}$ by integrating out θ, which yields

$$p(\mathbf{x} \mid \alpha) = \frac{N!}{x_1!, \cdots, x_k!}$$

Consequently according to Byes rule, we have the posterior update about θ based on data observations as

$$p(\theta \mid \mathbf{x}, \alpha) = \frac{p(\mathbf{x}, \theta \mid \alpha)}{p(\mathbf{x} \mid \alpha)} = \frac{\Gamma(\alpha k)}{\Gamma(\alpha)^k} \cdot \prod_{i=1}^{k} \theta_i^{x_i + \alpha - 1} = \text{Dir}(x_1 + \alpha, \cdots, x_k + \alpha) \tag{14}$$

Discussion following conventional style in machine learning and natural language processing

literature generally adopts conventional methodologies for processing unstructured information extracted from text, the corresponding analysis of which relies much on counting the occurrences of words. These methodologies however are justified by the implicit assumption that the order of words collected in documents (filings) is not related to the information to be recovered. By contrast, topic models featuring the application of LDA (Latent Dirichlet Allocation) provides a relatively successful extension of the conventional methodologies in natural language processing by assuming that the words collected in documents are implicitly in connection with several different topics (to some extent, topics or themes are regarded as the latent variables which are not observed and required to be estimated). To the end, topic models are essentially the modelling framework through which the aggregate information contained in documents (each document is regarded as a sequence of words in order) is extracted. Generally there are two alternatives for estimating this this model, VEM (Variational Expectation Maximization) and Gibbs Sampling, which are to be briefly discussed in the following.

VEM (Variational Expectation Maximization) is introduced for estimation with the likelihood function specified as following

$$\begin{aligned}\ell(\alpha, \boldsymbol{\beta}) &= \log(p(\mathbf{w} \mid \alpha, \boldsymbol{\beta})) \\ &= \log \int \left\{ \sum_{\mathbf{z}} \left[\prod_{i=1}^{N} p\left(\mathrm{w}_i \mid \mathrm{z}_i, \boldsymbol{\beta}\right) p\left(\mathrm{z}_i \mid \theta\right) \right] \right\} p(\theta \mid \alpha) d\theta.\end{aligned}$$

where $\mathbf{z} = (\mathrm{z}_i)_{i=1,\ldots,N}$ includes all the combinations of assigning N words into k topics. VEM is specifically a combination of Variational inference and E-M algorithms. More details about the Variational inference refer to (Mclachlan and Krishnan, 2008; Dempster et al., 1977; Wainwright and Jordan, 2008). Since here entries of $\boldsymbol{\beta}$ are directly regarded as parameters to be estimated if VEM is applied, hyperparameter δ is no more the parameter of interest.

Gibbs sampling can be done as well for approximating $p(\mathbf{z} \mid \mathbf{w})$, which essentially corresponds to the posterior probability of topic conditional on the observed words. α and δ are fixed at suggested values with $\alpha = 50/k$ and $\delta = 0.1$. Theoretical work done by Griffiths and Steyvers (2004); Phan et al. (2008) has justified that the whole loop of Gibbs sampling is compositionally based on sampling from the following distribution Fama and French (1993, 1996, 2015)

$$p\left(\mathrm{z}_i = K \mid \mathbf{w}, \mathrm{z}_{-i}\right) \propto \frac{n_{-i,K}^{(j)} + \delta}{n_{-i,K}^{(\cdot)} + V\delta} \frac{n_{-i,K}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + k\alpha} \tag{15}$$

- where $n_{-i,K}^{(j)}$ denotes how often the j-th term collected in vocabulary is assigned to topic K without the i-th word.
- d_i indexes the document to which word w_i belongs.
- Dot $\cdot$ implies that the summation over this index is implemented.

The term distribution and topic distribution is updated as following

$$\hat{\beta}_K^{(j)} = \frac{n_K^{(j)} + \delta}{n_K^{(\cdot)} + V\delta} \qquad \hat{\theta}_K^{(d)} = \frac{n_K^{(d)} + \alpha}{n^{(d)} + k\alpha} \tag{16}$$

for $j = 1, \ldots, V$ and $d = 1, \ldots, D$.

4 How to define Disaster Measure using LDA

The output from LDA generally gives us the key words extracted from text. The remained question is about how to match the extracted keywords (topics) with the existing dictionary indicating whether the associated contents are about disasters.

5 Notes for Variational Methods in Bayesian Analysis

5.1 EM Algorithm

Rather than sampling as in MCMC, EM algorithm initially proposed by (Dempster et al., 1977) chooses the target density function from a family of approximate density functions $\mathcal{Q}$ such that,

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z})\in\mathcal{Q}} \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}\mid\mathbf{x})) \tag{17}$$

Here we just denote $q(\mathbf{z}) \in \mathcal{Q}$ without specifying any parameters for it, however sometimes for the implementability concern in practice we usually consider the parametrized family $\mathcal{Q}$ such that each element contained in $\mathcal{Q}$ takes the form of $q_{\boldsymbol{\phi}}(\mathbf{z})$. Actually specification of the functional form of $q(\mathbf{z})$ depends on how much the flexibility to be taken into account for analysis and for the scenario when we make the assumption that $q(\mathbf{z})$ is factorizable, it can be demonstrated how optimization is set up in the discussion about Variational Expectation Maximization algorithm contained in the next subsection.

The most popular estimation technique adopted in statistical literature is MLE (Maximum Likelihood Estimation) where parameters are estimated based on the observed data by solving the following optimization

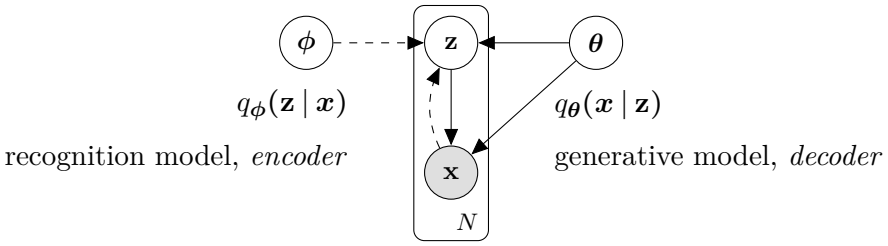
$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}). \tag{18}$$

Instead of being directly connected with observed data $\mathbf{x}$ through data generating process, it is practical and sometimes for convenience concern to introduce latent variables $\mathbf{z}$ such that the

likelihood function is actually the marginalization of joint likelihood as following [2]

$$p(\mathbf{x};\boldsymbol{\theta}) = \int p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})d\mathbf{z} = \int p(\mathbf{x}\mid\mathbf{z};\boldsymbol{\theta})p(\mathbf{z};\boldsymbol{\theta})d\mathbf{z} \tag{19}$$

It would also be helpful to summarize the general ideas for the framework in the following figure where latent variables are introduced, which is usually the case where Variational Bayes method plays the role.



Proposition 5.1 *Under some regular conditions,*

$$p(\mathbf{x};\boldsymbol{\theta}) = F(q,\boldsymbol{\theta}) + \mathrm{KL}(q\|p) \tag{20}$$

with

$$F(q,\boldsymbol{\theta}) = \int q(\mathbf{z})\ln\left(\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}\right)d\mathbf{z} \tag{21}$$

Proof.

$$\begin{aligned}
\ln p(\mathbf{x};\boldsymbol{\theta}) &= \int q(\mathbf{z})\ln\frac{p(\mathbf{x};\boldsymbol{\theta})p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\theta})}d\mathbf{z} \\
&= \int q(\mathbf{z})\ln\frac{p(\mathbf{z},\mathbf{x};\boldsymbol{\theta})}{p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\theta})}d\mathbf{z} \\
&= \int q(\mathbf{z})\ln\frac{p(\mathbf{z},\mathbf{x};\boldsymbol{\theta})q(\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\theta})q(\mathbf{z})}d\mathbf{z} \\
&= \int q(\mathbf{z})\ln\left(\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}\right)d\mathbf{z} + \int q(\mathbf{z})\ln\frac{q(\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x};\boldsymbol{\theta})}d\mathbf{z}
\end{aligned}$$

It is known that Kullback-Leibler divergence is not negative and hence $F(q,\boldsymbol{\theta})$ refers to the Evidence Lower Bound (ELBO). □

[2] Actually an inspiring perspective on this would be regarding $\mathbf{z}$ as local latent variables, one per observation, and $\boldsymbol{\theta}$ as (a d-dimensional vector and usually d is less than the sample size denoted by n) as global latent variables, which intrinsically does not change with sample size.

Remark 5.1 Alternatively, we may note that similar trick applies to the posterior with contingency on data. Usually this is for the case where Bayesian techniques apply when our target is to obtain the posterior mode, that is to maximize $\ln p(\boldsymbol{\theta}; \mathbf{x})$. Note that analogously we have

$$\begin{aligned}
\ln p(\boldsymbol{\theta}; \mathbf{x}) &= \int q(\mathbf{z}) \ln \frac{p(\boldsymbol{\theta}; \mathbf{x}) p(\mathbf{z} \mid \boldsymbol{\theta}; \mathbf{x})}{p(\mathbf{z} \mid \boldsymbol{\theta}; \mathbf{x})} d\mathbf{z} \\
&= \int q(\mathbf{z}) \ln \frac{p(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x})}{p(\mathbf{z} \mid \boldsymbol{\theta}; \mathbf{x})} d\mathbf{z} \\
&= \int q(\mathbf{z}) \ln \frac{p(\mathbf{z}, \boldsymbol{\theta}; \mathbf{x}) q(\mathbf{z})}{p(\mathbf{z} \mid \boldsymbol{\theta}; \mathbf{x}) q(\mathbf{z})} d\mathbf{z} \\
&= \int q(\mathbf{z}) \ln \left(\frac{p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right) d\mathbf{z} + \int q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z} \mid \boldsymbol{\theta}; \mathbf{x})} d\mathbf{z}
\end{aligned}$$

Consequently this implies that EM algorithm still applies for the maximization of posterior with contingency on data and the corresponding $q(\mathbf{z})$ is replaced with $q\left(\mathbf{z} \mid \boldsymbol{\theta}^{\mathrm{OLD}}; \mathbf{x}\right)$, which essentially shares the same functional form with $q\left(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\mathrm{OLD}}\right)$ at each iteration.

Remark 5.2 A direct implication from Proposition 5.1 is that the objective $F(q, \boldsymbol{\theta})$ as the lower bound of log marginal likelihood $\ln p(\mathbf{x}; \boldsymbol{\theta})$, the gap between $\ln p(\mathbf{x}; \boldsymbol{\theta})$ and $F(q, \boldsymbol{\theta})$ is KL distance and hence maximizing ELBO minimizes KL, which is essentially the objective of Variational Bayes.

EM algorithm is essentially an iterative algorithm with parameters $\boldsymbol{\theta}$ constantly updated in a scheme such that $\boldsymbol{\theta}^{\mathrm{OLD}} \to \boldsymbol{\theta}^{\mathrm{NEW}} \to \boldsymbol{\theta}^{\mathrm{OLD}} \cdots$ At E-step, replacing $q(\mathbf{z})$ with $p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\mathrm{OLD}})$ and substituting back to (21) yields

$$\begin{aligned}
F(\boldsymbol{\theta}, q) &= \int p\left(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\mathrm{OLD}}\right) \ln \left(\frac{p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})}{p\left(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\mathrm{OLD}}\right)} \right) d\mathbf{z} \\
&= \int p\left(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\mathrm{OLD}}\right) \ln \left(p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})\right) d\mathbf{z} - \int p\left(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\mathrm{OLD}}\right) \ln \left(p\left(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\mathrm{OLD}}\right)\right) d\mathbf{z} \\
&= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{OLD}}) + \text{const}
\end{aligned}$$

Hence equivalently in the M-step, $\boldsymbol{\theta}$ is updated such that

$$\boldsymbol{\theta}^{\mathrm{NEW}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{OLD}}) \tag{22}$$

To sum up, EM algorithm essentially relies on the following two iterative steps repeatedly in a closed loop until difference between $\boldsymbol{\theta}^{\mathrm{NEW}}$ and $\boldsymbol{\theta}^{\mathrm{OLD}}$ is less than the tolerance rate specified.

E-step Compute $p\left(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\mathrm{OLD}}\right)$ and $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{OLD}})$

M-step Evaluate

$$\boldsymbol{\theta}^{\mathrm{NEW}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{OLD}})$$

Example 5.1 *Consider a Gaussian mixture with G components indexed by $j = 1, \ldots, G$ and for each component along with a specific observation x_i we know exactly the density form $f_j(\mathrm{x}_i)$ and denote the parameters as the probabilities assigned to each component, π_j and $\sum_{j=1}^{G} \pi_j = 1$, collected in $\boldsymbol{\theta} = \{\pi_j\}_{j=1}^{G}$. All these lead to the log marginal likelihood of observed data $\mathbf{x}$ as*

$$\ln p(\mathbf{x} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \left[\sum_{j=1}^{G} \pi_j f_j(\mathrm{x}_i) \right] \tag{23}$$

An obvious difficulty for applying MLE with respect to this log marginal likelihood is the summation over g as the input of $\ln(\cdot)$. However this problem could be circumvented by introducing latent variable $\mathbf{z}$ such that unobserved element contained in $\mathbf{z}$ indexed by z_{ij} is 0-1 variable indicating whether the i-th unobserved component z_i associated with x_i is from the j-th component. Consequently the joint log likelihood of $\mathbf{x}$ and $\mathbf{z}$ (sometime this is also referred to complete log likelihood) is

$$\ln p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{G} \mathrm{z}_{ij} \left[\ln \pi_j + \ln f_j(\mathrm{x}_i) \right] \tag{24}$$

*Expectation of $\ln p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})$ over $\mathbf{z}$ taken with respect to $p\left(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{OLD}\right)$, which is required for the **E-step**, is given as*

$$\sum_{i=1}^{N} \sum_{j=1}^{G} \frac{\pi_j^{OLD} f_j(\mathrm{x}_i)}{\sum_{g=1}^{G} \pi_g^{OLD} f_g(\mathrm{x}_i)} \left[\ln \pi_j + \ln f_j(\mathrm{x}_i) \right] \tag{25}$$

*Then for the **M-step**, we consider the following constrained optimization problem*

$$\begin{aligned} \max \quad & \sum_{i=1}^{N} \sum_{j=1}^{G} \frac{\pi_j^{OLD} f_j(\mathrm{x}_i)}{\sum_{g=1}^{G} \pi_g^{OLD} f_g(\mathrm{x}_i)} \left[\ln \pi_j + \ln f_j(\mathrm{x}_i) \right] \\ \text{s.t.} \quad & \pi_1 + \cdots + \pi_G = 1 \end{aligned}$$

the standard Lagrangian technique works here with corresponding multiplier denoted as λ and the related first order conditions for π_j implies that

$$\frac{1}{\pi_j} \sum_{i=1}^{N} \frac{\pi_j^{OLD} f_j(\mathrm{x}_i)}{\sum_{g=1}^{G} \pi_g^{OLD} f_g(\mathrm{x}_i)} - \lambda = 0, \quad 1 \leqslant j \leqslant G$$

and

$$\pi_j = \sum_{i=1}^{N} \frac{\pi_j^{OLD} f_j(\mathrm{x}_i)}{\sum_{g=1}^{G} \pi_g^{OLD} f_g(\mathrm{x}_i)} \Big/ \lambda$$

Taking summation over j on both sides of the above equation and using the required constraint solves λ as

$$\lambda = \sum_{i=1}^{N} \sum_{j=1}^{G} \frac{\pi_j^{OLD} f_j(\mathrm{x}_i)}{\sum_{g=1}^{G} \pi_g^{OLD} f_g(\mathrm{x}_i)} = N$$

*Finally all the previous discussion yields that $\boldsymbol{\theta}$ updated from $\boldsymbol{\theta}^{OLD}$ to $\boldsymbol{\theta}^{NEW}$ in the **M-step** is as following*

$$\boldsymbol{\theta}^{NEW} = \left(\sum_{i=1}^{N} \frac{\pi_1^{OLD} f_j(\mathrm{x}_i)}{\sum_{g=1}^{G} \pi_g^{OLD} f_g(\mathrm{x}_i)} \Big/ N, \cdots, \sum_{i=1}^{N} \frac{\pi_G^{OLD} f_j(\mathrm{x}_i)}{\sum_{g=1}^{G} \pi_g^{OLD} f_g(\mathrm{x}_i)} \Big/ N \right) \tag{26}$$

A simple `R` code are provided for demonstrating the main logic of this procedure.

Example 5.2 *Another direct application of EM algorithm is variable selection. Normally in Bayesian literature this target is achieved by imposing "spike-and-slab" prior (which is essentially a special case of Gaussian mixture) on coefficients and the corresponding importance weights of different variables are relied upon the posterior updates (George and McCulloch, 1993, 1997; Ročková and George, 2014). Specifically, we focus on the setting where there exists $n \times 1$ response vector (collection of observations of dependent variables), $\mathbf{y}$; and potential variables (independent variables), $\mathbf{X} = [\mathrm{x}_1, \ldots, \mathrm{x}_p]$, which is a $n \times p$ matrix. Temporarily a Gaussian linear model is assumed to relate $\mathbf{X}$ to $\mathbf{y}$. That is, we generally have the following specified normal density,*

$$f(\mathbf{y} \mid \mathbf{X}) = \mathcal{N}_n\left(\iota_n \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right) \tag{27}$$

As with many Bayesian variable selection approaches, a binary latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ is introduced where $\gamma_i \in \{0, 1\}$ and $\gamma_i = 1$ indicates that x_i is included in model. The intrinsic logic for the variable selection objective is that in combination with some suitable prior for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ and $\boldsymbol{\gamma}$, the induced posterior $\pi(\boldsymbol{\gamma} \mid \text{Data}) = \pi(\boldsymbol{\gamma} \mid \boldsymbol{y}, \mathbf{X})$ then summarizes all the postdata variable selection uncertainty. In particular, (George and McCulloch, 1997, henceforth GM97) proposed the following specified prior,

$$\pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma}, v_0, v_1) = \mathcal{N}_p(\mathbf{0}, D_{\sigma,\gamma}) \tag{28}$$

where

$$D_{\sigma,\gamma} = \sigma^2 \operatorname{diag}(a_1, \ldots, a_p).$$

with $a_i = (1 - \gamma_i)v_0 + \gamma_i v_1$ for $0 \leqslant v_0 < v_1$. Such a kind of specification implies that it is essentially a Gaussian mixture with two components. GM97 recommended setting v_0 and v_1 to be small and large values, respectively, to distinguish those β_i values which warrant exclusion of x_i from that inclusion of x_i. The key discussion corresponds closely to the specified distribution for $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$. A natural

default choice for this would be the i.i.d. Bernoulli prior of the following form

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta}) = \theta^{|\boldsymbol{\gamma}|}(1-\theta)^{p-|\boldsymbol{\gamma}|} \tag{29}$$

where

$$\theta \in [0,1] \text{ and } |\boldsymbol{\gamma}| = \sum_i \gamma_i.$$

For the prior of θ, normally we adopt beta distribution such that

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

For the prior of σ^2, GM97 proposes the use of inverse gamma prior

$$\pi\left(\sigma^2 \mid \boldsymbol{\gamma}\right) = \mathrm{IG}\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right)$$

where this inverse gamma prior takes the following functional PDF form [3]*,*

$$\frac{(\nu\lambda/2)^{\nu/2}}{\Gamma(\nu/2)}\left(\sigma^2\right)^{-\nu/2-1}\exp\left(-\frac{\nu\lambda}{2\sigma^2}\right)$$

Note that for any practical application of EM algorithm, our target is never to exactly pin down the posterior distribution of latent variables but the posterior updates of parameters instead. Consequently given our setting here, parameters are collected in $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma)$ [4] *and our ultimate target is geared toward finding **posterior modes** of **parameter** posterior $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \mathbf{y}, \mathbf{X})$ rather than simulating from the entire **model** posterior $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$.*

Following the conventional procedure implemented in EM algorithm, maximization of $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \mathbf{y}, \mathbf{X})$ is implemented indirectly by proceeding iteratively in terms of the "complete log likelihood" as we did in the previous example, that is, $\ln \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$, where the inclusion indicators collected in $\boldsymbol{\gamma}$ are treated as "missing data". More precisely, the whole procedure comprises the iterative maximization of the following objective function

$$\begin{aligned} &\mathbf{Q}\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) \\ = \; &\mathbb{E}_{\boldsymbol{\gamma} \mid .}\left[\ln \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma, \gamma \mid \mathbf{y}, \mathbf{X}) \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}, \mathbf{y}\right] \end{aligned} \tag{30}$$

where the operator $\mathbb{E}_{\boldsymbol{\gamma} \mid .}(\cdot)$ denotes the conditional expectation, that is,

$$\mathbb{E}_{\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}, \mathbf{y}, \mathbf{X}}(\cdot)$$

[3] In general for inverse-gamma distribution with parameter α and β, the corresponding PDF (probability density function) takes the form of $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1}\exp\left(-\frac{\beta}{x}\right)$, where x refers to the random variable.

[4] Here we ignore the emphasis of α since generally it can be subsumed in $\boldsymbol{\beta}$ by treating it as the coefficient of fixed constant.

Specifically, (30) can be decomposed into the addition of a constant term and the following additional two terms [5]

$$\begin{aligned}
&\mathbf{Q}_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) \\
&= -\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} - \frac{n+p+v+2}{2}\ln\left(\sigma^2\right) - \frac{\nu\lambda}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2 \mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right]
\end{aligned}$$

$$\begin{aligned}
&\mathbf{Q}_2\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) \\
&= \sum_{i=1}^{p}\ln\left(\frac{\theta}{1-\theta}\right)\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\gamma_i + (a-1)\ln(\theta) + (p+b-1)\ln(\theta)
\end{aligned}$$

Thus in general we have

$$\mathbf{Q}\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = \mathbf{C} + \mathbf{Q}_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) + \mathbf{Q}_2\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) \tag{31}$$

Such a kind of separability into two distinct functions yields a M-step that is obtained by optimizing each function. To see why $\mathbf{Q}_1(\cdot)$ *and* $\mathbf{Q}_2(\cdot)$ *take the demonstrated functional form respectively, firstly we observe the generic fact that any well-defined posterior is proportional to the corresponding jointly likelihood up to a constant which is the result of marginalizing out the parameters. The everything remained is just obtained by taking log of the product of prior and and the likelihood of data given specified parameter. Analogous discussion applies for* $\mathbf{Q}_2(\cdot)$*. More specifically, for* $\mathbf{Q}_1(\cdot)$*, we have*

$$\begin{array}{ccccc}
f(\mathbf{y} \mid \mathbf{X}) & \cdot & \pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma}, v_0, v_1) & \cdot & \pi\left(\sigma^2 \mid \boldsymbol{\gamma}\right) \\
\uparrow & & \uparrow & & \uparrow \\
\mathcal{N}_n\left(\iota_n\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right) & & \mathcal{N}_p\left(\mathbf{0}, D_{\sigma,\boldsymbol{\gamma}}\right) & & \mathrm{IG}\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right)
\end{array}$$

Taking log of the above product yields the desired functional form of $\mathbf{Q}_1(\cdot)$*; For* $\mathbf{Q}_2(\cdot)$*, we have*

$$\begin{aligned}
&\pi(\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta}) \\
&\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{|\boldsymbol{\gamma}|}(1-\theta)^{p-|\boldsymbol{\gamma}|} = \theta^{a-1}(1-\theta)^{p+b-1}\prod_{i=1}^{p}\left(\frac{\theta}{1-\theta}\right)^{\gamma_i}
\end{aligned}$$

Taking log of the above formula yields the desired functional form of $\mathbf{Q}_2(\cdot)$

$$(a-1)\ln(\theta) + (p+b-1)\ln(1-\theta) + \sum_{i=1}^{p}\ln\left(\frac{\theta}{1-\theta}\right)\gamma_i$$

hence in the E-step where γ_i *is replaced with its corresponding conditional expectation,* $\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\gamma_i$*, we*

[5] It is a little bit different from the formula derived in GM97, where the multiplier before $\ln\left(\sigma^2\right)$ is given as $\frac{n-1+p+\nu}{2}$. I personally hold the opinion that the formula given in GM97 should be a typo.

obtain the desired functional form of $\mathbf{Q}_2(\cdot)$. *From the previous discussion, two conditional expectation plays the vital role in proceeding the E-step, thus* $\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right]$ *and* $\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\gamma_i$ *respectively. For this reason we briefly discuss how this two conditional expectation is calculated,*

- $\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\gamma_i$:

$$\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\gamma_i = \mathbf{P}\left(\gamma_i = 1 \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = p_i^\star$$

where

$$p_i^\star = \frac{a_i}{a_i + b_i}$$

and

$$\begin{aligned} a_i &= \pi\left(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1\right)\mathbf{P}\left(\gamma_i = 1 \mid \boldsymbol{\theta}^{(k)}\right) \\ b_i &= \pi\left(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 0\right)\mathbf{P}\left(\gamma_i = 0 \mid \boldsymbol{\theta}^{(k)}\right) \end{aligned}$$

- $\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right]$:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\left[\frac{1}{v_0\left(1-\gamma_i\right)+v_1\gamma_i}\right] \\ =\ &\frac{\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\left(1-\gamma_i\right)}{v_0} + \frac{\mathbb{E}_{\boldsymbol{\gamma} \mid \cdot}\gamma_i}{v_1} = \frac{1-p_i^\star}{v_0} + \frac{p_i^\star}{v_1} \equiv d_i^\star \end{aligned}$$

One attractive feature of the previously discussed setting is that all the required maximization in M-step yields analytical solution so that $\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\theta}^{(k)}$ *and* $\sigma^{(k)}$ *could be updated quickly to* $\boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\theta}^{(k+1)}$ *and* $\sigma^{(k+1)}$ *with* $\boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\theta}^{(k+1)}$ *and* $\sigma^{(k+1)}$ *replaced by the corresponding optimal solution respectively. Under some regular conditions, this sequence would converge at* $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\theta}}$ *and* $\widehat{\sigma}$. *Then based on these estimated parameters, the conditional component inclusion probabilities are given by*

$$\mathbf{P}\left(\gamma_i \mid \widehat{\beta}_i, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}\right) = \frac{c_i}{c_i + d_i} \tag{32}$$

where

$$\begin{aligned} c_i &= \pi\left(\widehat{\beta}_i, \widehat{\sigma}, \gamma_i = 1\right)\pi\left(\gamma_i = 1 \mid \widehat{\boldsymbol{\theta}}\right) \\ d_i &= \pi\left(\widehat{\beta}_i, \widehat{\sigma}, \gamma_i = 0\right)\pi\left(\gamma_i = 0 \mid \widehat{\boldsymbol{\theta}}\right) \end{aligned}$$

Definition 2 α*-Rényi divergence is defined for measuring the relative divergence between two probability measures. Suppose* $\alpha \in (0,1)$ *and* P *and* Q *are two probability measures. Let* μ *be any measure such that* $P \ll \mu$ *and* $R \ll \mu$, *for example* $\mu = P + Q$, *then* α*-Rényi divergence is defined as following*

$$D_\alpha\left(P, Q\right) = \frac{1}{\alpha - 1}\log\int\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^\alpha\left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1-\alpha}\mathrm{d}\mu \tag{33}$$

Comprehensive discussion about α-Rényi divergence is available in Van Erven and Harremoës (2014). An important convergence would be as following,

$$\lim_{\alpha\to 0}\frac{1}{\alpha-1}\ln\int p^{\alpha}q^{1-\alpha}\mathrm{d}\mu \;=\; -\ln\int \mathbf{1}_{\{p>0\}}q\mathrm{d}\mu \;=\; -\ln Q\left(p>0\right)$$
$$\lim_{\alpha\to 1}\frac{1}{\alpha-1}\ln\int p^{\alpha}q^{1-\alpha}\mathrm{d}\mu \;=\; \int p\ln\left(\frac{p}{q}\right)\mathrm{d}\mu$$

where $p=\frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q=\frac{\mathrm{d}R}{\mathrm{d}\mu}$.

Remark 5.3 *Essentially the justification for these two limits is based on the justification for commuting the integration and the limit taken with respect to α. This is carefully discussed in Van Erven and Harremoës (2014), where **Lemma 1** and the corresponding derivation for **Theorem 5** jointly justify such a kind of desired communication. The fundamental tool used for proving these claims is monotone convergence theorem. Rather than discussing the proof rigorously here, an important and useful observation would be that $\frac{\mathrm{d}}{\mathrm{d}\alpha}p^{\alpha}q^{1-\alpha}=p^{\alpha}q^{1-\alpha}\ln\left(\frac{p}{q}\right)$, accordingly once communication between integration and limit taken with respect to α is rigorously justified, a heuristic demonstration of why the claimed limits hold is by applying L'Hôpital's rule such that*

$$\lim_{\alpha\to 1}\frac{1}{\alpha-1}p^{\alpha}q^{1-\alpha}=\lim_{\alpha\to 1}p^{\alpha}q^{1-\alpha}\ln\left(\frac{p}{q}\right)=p\ln\left(\frac{p}{q}\right).$$

Another useful facts are as long as $P\ll Q$, then we have

- *$p<q$ almost surely holds with respect to Q.*
- *If we define $x_{\alpha}=\int p^{\alpha}q^{1-\alpha}\mathrm{d}\mu$, then it is appropriate to have $\lim_{\alpha\to 1}\int p^{\alpha}q^{1-\alpha}\mathrm{d}\mu=\int p\mathrm{d}\mu=1$. More specifically, this convergence is depicted from different directions as following*

$$\int p^{\alpha}q^{1-\alpha}\mathrm{d}\mu=\int\left(\frac{p}{q}\right)^{\alpha}q\,\mathrm{d}\mu$$

$$\alpha\uparrow 1\quad x_{\alpha}\downarrow 1\quad x_{\alpha}-1\geqslant\ln x_{\alpha}\to 0$$
$$\alpha\downarrow 1\quad x_{\alpha}\uparrow 1\quad \ln x_{\alpha}\geqslant x_{\alpha}-1\to 0$$

where $\uparrow 1$ refers to the convergence below 1 while $\downarrow 1$ refers to the convergence above 1.

Consequently,

$$\lim_{\alpha\to 1}\frac{1}{\alpha-1}\ln\int p^{\alpha}q^{1-\alpha}\mathrm{d}\mu=\lim_{\alpha\to 1}\frac{\int p^{\alpha}q^{1-\alpha}\mathrm{d}\mu-1}{\alpha-1}=\lim_{\alpha\to 1}\int_{p,q>0}\frac{p-p^{\alpha}q^{1-\alpha}}{1-\alpha}\mathrm{d}\mu \tag{34}$$

and this implies as long as we are able to change the order of limits and integration and then taking limit with respect to α by applying L'Hôpital's rule yields

$$\lim_{\alpha\to 1}D_{\alpha}\left(P,Q\right)=\mathrm{KL}\left(P,Q\right) \tag{35}$$

With this introduced α-Rényi divergence, we introduce the following concepts for later discussion

- Let Q denote the dominating measure for the family of distributions for data and hence the associated Radon-Nikodym derivative (probability density function in some sense) is as following
$$p_{\boldsymbol{\theta}} = \frac{\mathrm{d}P_{\boldsymbol{\theta}}}{\mathrm{d}Q}$$
- $\boldsymbol{\theta} \in \Theta$ and Θ is equipped with proper σ-algebra $\mathcal{T}$, $\mathcal{M}_1^+(\Theta, \mathcal{T})$ refers to the set of all probability distributions on measurable space $(\Theta, \mathcal{T})$. Accordingly, $\pi \in \mathcal{M}_1^+(\Theta, \mathcal{T})$ denotes prior.
- Likelihood as the function of $\boldsymbol{\theta}$ is [6]
$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathrm{x}_i)$$
- $\forall (\boldsymbol{\theta}, \boldsymbol{\theta}')$, negative log-likelihood ratio is defined as following
$$r_n(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{i=1}^{n} \log \frac{p_{\boldsymbol{\theta}'}(\mathrm{x}_i)}{p_{\boldsymbol{\theta}}(\mathrm{x}_i)}$$
- Fractional posterior is then defined as following
$$\pi_{n,\alpha}(\mathrm{d}\boldsymbol{\theta} \mid \mathbf{x}_1^n) = \frac{\mathrm{e}^{-\alpha r_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)} \pi(\mathrm{d}\boldsymbol{\theta})}{\int \mathrm{e}^{-\alpha r_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)} \pi(\mathrm{d}\boldsymbol{\theta}) \pi(\mathrm{d}\boldsymbol{\theta})} \tag{36}$$

With these introduced concepts and notations, we summarize some important results discussed in Alquier and Ridgway (2020):

Theorem 5.1 (PAC-Bayesian Inequality) *For any $\alpha \in (0,1)$ and any $\epsilon \in (0,1)$,*

$$\mathbb{P}\Bigg(\forall \rho \in \mathcal{M}_1^+(\Theta), \int D_\alpha(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}_0})\, \rho(\mathrm{d}\boldsymbol{\theta}) \leqslant \frac{\alpha}{1-\alpha} \int \frac{r_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)}{n} \rho(\mathrm{d}\boldsymbol{\theta}) + \frac{\mathrm{KL}(\rho, \pi) + \log(1/\epsilon)}{n(1-\alpha)} \Bigg) \geqslant 1 - \epsilon \tag{37}$$

This result originates from Catoni (2004, 2007) and is extended to variational approximation by alternatively defining the following approximate posterior as

$$\begin{aligned} \tilde{\pi}_{n,\alpha}(\cdot \mid \mathbf{x}_1^n) &= \underset{\rho \in \mathcal{F}}{\operatorname{argmin}} \left\{ \alpha \int r_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \rho(\mathrm{d}\boldsymbol{\theta}) + \mathrm{KL}(\rho, \pi) \right\} \\ &= \underset{\rho \in \mathcal{F}}{\operatorname{argmin}} \left\{ -\alpha \int \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathrm{x}_i)\, \rho(\mathrm{d}\boldsymbol{\theta}) + \mathrm{KL}(\rho, \pi) \right\} \end{aligned}$$

[6] x_i refers to the i-th realization of random variables (either scalars or vectors) $\mathbf{x}$.

Corollary 5.1 (Concentration of VB approximation) *For any $\alpha \in (0,1)$ and $\epsilon \in (0,1)$, with probability at least $1-\epsilon$*

$$\int D_\alpha\left(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}_0}\right) \tilde{\pi}_{n,\alpha}\left(\mathrm{d}\boldsymbol{\theta} \mid \mathbf{x}_1^n\right) \leqslant \inf_{\rho \in \mathcal{F}} \left\{ \frac{\alpha}{1-\alpha} \int \frac{r_n\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0\right)}{n} \rho\left(\mathrm{d}\boldsymbol{\theta}\right) + \frac{\mathrm{KL}(\rho, \pi) + \log(1/\epsilon)}{n(1-\alpha)} \right\} \tag{38}$$

Theorem 5.1 and Corollary 5.1 jointly leads to the following claimed main result in Alquier and Ridgway (2020)

Theorem 5.2 *For any fixed $\mathcal{F} \subset \mathcal{M}_1^+(\Theta)$, if we assume that there exists $\varepsilon_n > 0$ such that*

$$\int \mathrm{KL}\left(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}_0}\right) \rho_n\left(\mathrm{d}\boldsymbol{\theta}\right) \leqslant \varepsilon_n, \quad \int \mathbb{E}\left[\log^2\left(\frac{p_{\boldsymbol{\theta}}(\mathrm{x}_i)}{p_{\boldsymbol{\theta}_0}(\mathrm{x}_i)}\right)\right] \rho_n\left(\mathrm{d}\boldsymbol{\theta}\right) \leqslant \varepsilon_n$$

and

$$\mathrm{KL}\left(\rho_n, \pi\right) \leqslant n\varepsilon_n$$

Then for any $\alpha \in (0,1)$, for any $(\varepsilon, \eta) \in (0,1)^2$, the following inequality holds

$$\mathbb{P}\left[\int D_\alpha\left(P_{\boldsymbol{\theta}_0}, P_{\boldsymbol{\theta}}\right) \tilde{\pi}_{n,\alpha}\left(\mathrm{d}\boldsymbol{\theta} \mid \mathbf{x}_1^n\right) \leqslant \frac{(\alpha+1)\varepsilon + \alpha\sqrt{\frac{\varepsilon_n}{n\eta}} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{n}}{1-\alpha}\right] \geqslant 1-\varepsilon-\eta \tag{39}$$

Remark 5.4 *There are some remarks to be emphasized here for discussing this main result.*

- *How ρ_n is selected. Define the following restricted area*

  $$B(r) = \left\{ \boldsymbol{\theta} \in \Theta : \mathrm{KL}\left(P_{\boldsymbol{\theta}_0}, P_{\boldsymbol{\theta}}\right) \leqslant r,\ \mathbb{E}\left[\log^2\left(\frac{p_{\boldsymbol{\theta}}(\mathrm{x}_i)}{p_{\boldsymbol{\theta}_0}(\mathrm{x}_i)}\right)\right] \leqslant r \right\}$$

  and then ρ_n is selected such that

  $$\rho_n = \pi_{|B(\varepsilon_n)}$$

  i.e. ρ_n is selected as the π restricted to $B(\varepsilon_n)$ and hence for this case the required condition rewrites as following

  $$\mathrm{KL}\left(\rho_n, \pi\right) = -\log \pi\left(B(\varepsilon_n)\right) \leqslant n\varepsilon_n$$

- *An alternative explanation for the results implied from this theorem is that Expected (with respect to VB posterior) α-Rényi divergence of data distribution associated with different $\boldsymbol{\theta}$ cannot be that divergent with probability $1-\varepsilon-\eta$ and this probability is with respect to data.*

Alternatively as we have mentioned earlier that we are allowed to regard $\boldsymbol{\theta}$ as latent variables but temporarily treat it as the fixed parameter. By doing so, it is appropriate to redefine the evidence lower bound (ELBO) by simultaneously accounting for both local latent variables $\mathbf{z}$ and

global latent variables $\boldsymbol{\theta}$. This trick treating the global variable $\boldsymbol{\theta}$ as temporarily fixed also justifies our discussion temporarily within frequentist domain and later the following discussion about how Variational Frequentist and Variational Bayes are connected. To formally set up the connection, it would be helpful to summarize the associated definitions as following

[Place Table 1 about here]

5.2 Variational Expectation Maximization

The key difference between Variational Expectation Maximization (VEM) and Expectation Maximization (EM) is mainly at E-step where there is an implicitly crucial assumption that $p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\text{OLD}})$ is available so that $q(\mathbf{z})$ can be replaced with $p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\text{OLD}})$. However this is not generally the case when $p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{\text{OLD}})$ is tractable, hence we turn to find the optimal functionality of $q(\cdot)$ instead by relying on the assumption that $q(\mathbf{z}) = \prod_i q(\mathrm{z}_i)$, then ELBO can be factorized as following

$$
\begin{aligned}
F(q, \boldsymbol{\theta}) &= \int q(\mathbf{z}) \ln \left(\frac{p(\mathbf{z}, \mathbf{x}; \boldsymbol{\theta})}{q(\mathbf{z})} \right) d\mathbf{z} \\
&= \int q(\mathrm{z}_j) \left(\int \prod_{i \neq j} q(\mathrm{z}_i) \ln p(\mathbf{z}, \mathbf{x}; \boldsymbol{\theta}) \right) \prod_{i \neq j} d\mathrm{z}_i d\mathrm{z}_j \; - \int q(\mathrm{z}_j) \ln q(\mathrm{z}_j) d\mathrm{z}_j - \sum_{i \neq j} \int q(\mathrm{z}_i) \ln q(\mathrm{z}_i) d\mathrm{z}_i \\
&= \int q(\mathrm{z}_j) \ln \left(\frac{\exp \left(\langle \ln p(\mathbf{z}, \mathbf{x} \mid \boldsymbol{\theta}) \rangle_{i \neq j} \right)}{q(\mathrm{z}_j)} \right) d\mathrm{z}_j - \sum_{i \neq j} \int q(\mathrm{z}_i) \ln q(\mathrm{z}_i) d\mathrm{z}_i \\
&= \int q(\mathrm{z}_j) \ln \left(\frac{\tilde{p}_{i \neq j}}{q(\mathrm{z}_j)} \right) d\mathrm{z}_j + H(\mathrm{z}_{i \neq j}) + \text{const} \\
&= -\mathrm{KL} \left(q_j \| \tilde{p}_{i \neq j} \right) + H(\mathrm{z}_{i \neq j}) + \text{const}
\end{aligned}
$$

where $\tilde{p}_{i \neq j}$ is normalized pdf. Since K-L divergence is non-negative, with $\mathrm{z}_{i \neq j}$ to be integrated out under proper expectation measure, the sufficient and necessary condition for ELBO to be maximized is $\mathrm{KL} \left(q_j \| \tilde{p}_{i \neq j} \right) = 0$, which sufficiently holds when

$$
q(\mathrm{z}_j) = \tilde{p}_{i \neq j} \propto \exp \left(\langle \ln p(\mathbf{z}, \mathbf{x} \mid \boldsymbol{\theta}) \rangle_{i \neq j} \right)
$$

or in other words the log of optimal density $q_j^*(\mathrm{z}_j)$ is

$$
\ln q_j^* \left(\mathrm{z}_j \right) = \langle \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \rangle_{i \neq j} + \text{const} \tag{40}
$$

Consequently, the EM algorithm is given by the following two steps

Variational E-step Evaluate $q^{\text{NEW}}(\mathbf{z})$ to maximize $F(q, \boldsymbol{\theta}^{\text{OLD}})$ solving (40)

Variational M-step Find

$$\boldsymbol{\theta}^{\text{NEW}} = \arg\max_{\boldsymbol{\theta}} F\left(\boldsymbol{q}^{\text{NEW}}, \boldsymbol{\theta}\right)$$

In comparison to conventional data used for economics and finance research, information encoded in textual data is relatively more rich and serves as good complement to the conventional structured data. Recently there seems to be explosion of social science researches using text data directly or text data extracted from videos. One reason that there were relatively less empirical researches in this norm is the intrinsic features that data of these kinds are high dimensional while most of the analysis tool is not as powerful as the available analysis tools nowadays. Conventionally, classical MCMC methods serve as the major tools for the application of Bayesian methods in practice but one fatal problems encountered in the MCMC is its deteriorated computational efficiency especially for the case when both the covariates dimension and sample size are large. Variational Bayes (VB) by comparison replaces the conventional MCMC sampling procedure with optimization problem described previously so that to some extent alleviates the computational inefficiency, and VB is nowadays is increasingly becoming the major tools for many computationally demanding applications such as image and video processing, and natural language processing (NLP).

Unstructured data recorded in texts, videos can reveal some useful information for financial market and one important mechanism through which the unstructured data play the pivotal role is that the people's emotion revealed from data is contagious and hence generates impact on people's decisions corresponding to economic activities, this is also the mechanism referred by some existing literature as social transmission. To some extent, such a hypothesized mechanism is quite natural since psychologically people's subjective perceptions are easily influenced by others so are associated economic behaviours.

Development in modern information technologies and computer science provides analysis tools both convenient and powerful to extract information collected in aforementioned unstructured data like texts and videos. Platforms like Google Cloud and Microsoft Azure both provide extensive support for modern data science language such as `R` and `Python` for natural language processing. Besides, community within the ecosystems of both `R` and `Python` have contributed a lot of integrated convenient packages such as `tm`, `topicmodels`, `googlelanguageR`, `tuneR`, `av` for `R` and `youtube_dl` for `Python`. Gentzkow, Kelly, and Tady (2019) recently provides a relatively comprehensive review about how the recently developed techniques have been applied for analysing textual data in economics and finance.

As what is mentioned previously, data extracted from text or video is highly unstructured and hence to make it manageable, some preliminary procedures have to be implemented. One important measure commonly used in textual analysis for this purpose is called "term frequency-inverse document frequency" (abbreviated as `tf-idf`). Generally, it is defined as following

$$\texttt{tf-idf: } \texttt{tf}_{ij} \times \texttt{idf}_{j} \tag{41}$$

where $\texttt{tf}_{ij}$ refers to the term frequency of word j in document i such that

$$\texttt{tf}_{ij} = \frac{c_{ij}}{\sum_k c_{ik}}$$

with c_{ij} defined as the count of occurrences of word j in document i; And $\texttt{idf}_j$ refers to: $\log(n/d_j)$ with

$$n : \text{total number of documents}$$

$$d_j : \sum_i \mathbf{1}(c_{ij} > 0),\ \mathbf{1}(\cdot) \text{ is indicator function.}$$

Since in general we want to make data extracted from text reveal distinguishable information, those words rarely appearing in each document and commonly appearing in most or all of the documents should be eliminated. This can be achieved by setting a threshold value for the `tf-idf` measure since by construction small $\texttt{tf}_{ij}$ implies that the j-th word is rare in document i and similarly small $\texttt{idf}_j$ implies that the j-th word is relatively much common in all the documents. Essentially `tf-idf` as the transformed measure defines the most representative terms in a given document to be those that appear infrequently overall, but frequently in that specific document (see Engle et al., 2020, henceforth EGLKS2020). Based on the `tf-idf` measure, EGLKS2020 constructs a climate change index by calculating the "cosine similarity" between `tf-idf` measures for "climate change vocabulary (CCV)" and each daily *The Wall Street Journal* (WSJ) edition. Another index referred to Crimson Hexagon's negative sentiment climate change index is constructed along with climate change index as well. One reason motivating the construction of Crimson Hexagon's (CH) negative sentiment climate change index in companion with climate change index for comparison, as articulated in EGLKS2020, is that the construction of WSJ Climate Change News Index implicitly embeds the assumption that there is no discrepancy between good news and bad news about climate changes given the coverage of all WSJ corpus, which is potentially at the risk of inaccurately capturing the positive news about climate change. Consequently, Crimson Hexagon's (CH) negative sentiment climate change index as one alternative for addressing this concern is forwarded in EGLKS2020 by confining the focus specifically on negative climate change news. Particularly, the CH Negative Climate Change Index is constructed as the "share of all news articles that are both about 'climate change' and that have been assigned to the 'negative sentiment' category."

5.2.1 Discussion about parametrized optimization

It is obvious from the previous discussion that optimization plays the pivotal role in the implementation of VB method and consequently we may briefly discuss the general optimization procedure we often implement in practice. Stochastic gradient ascent method proposed in Robbins and Monro (1951) is widely used in practice for solving the optimization in **M-step**. For notation simplicity and consistency while emphasizing the associated dependency of ELBO on parameters collected in

$\boldsymbol{\phi}$, we denote the parametrized ELBO as

$$\mathcal{F}(\boldsymbol{\phi}) = F(q, \boldsymbol{\phi}, \boldsymbol{\theta}) \quad = \int q_{\boldsymbol{\phi}}(\mathbf{z}) \ln\left(\frac{p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})}{q_{\boldsymbol{\phi}}(\mathbf{z})}\right) d\mathbf{z}$$

where we emphasize that $\mathcal{F}(\cdot)$ only depends on the $\boldsymbol{\phi}$ with $\boldsymbol{\theta}$ fixed at constant and then we are able to apply stochastic approximation method by calculating the unbiased gradient estimates $\widehat{\nabla_{\boldsymbol{\phi}}\mathcal{F}(\boldsymbol{\phi})}$. To obtain the explicit formula for $\nabla_{\boldsymbol{\phi}}$, it would be useful to note a key fact that

$$\begin{aligned}
\mathbb{E}_q\left[\nabla_{\boldsymbol{\phi}} \ln q_{\boldsymbol{\phi}}(\mathbf{z})\right] &= \mathbb{E}_q\left[\frac{1}{q_{\boldsymbol{\phi}}(\mathbf{z})} \nabla_{\boldsymbol{\phi}} q_{\boldsymbol{\phi}}(\mathbf{z})\right] \\
&= \int q_{\boldsymbol{\phi}}(\mathbf{z}) \frac{1}{q_{\boldsymbol{\phi}}(\mathbf{z})} \nabla_{\boldsymbol{\phi}} q_{\boldsymbol{\phi}}(\mathbf{z}) d\mathbf{z} \\
&= \int \nabla_{\boldsymbol{\phi}} q_{\boldsymbol{\phi}}(\mathbf{z}) d\mathbf{z} = 0 \quad \left(\text{since} \int q_{\boldsymbol{\phi}}(\mathbf{z}) d\mathbf{z} = 1\right)
\end{aligned}$$

This fact is also referred to the so-called log-derivative trick. With this observed fact, we are able to derive as following

$$\begin{aligned}
\nabla_{\boldsymbol{\phi}}\mathcal{F}(\boldsymbol{\phi}) &= \int \nabla_{\boldsymbol{\phi}} q_{\boldsymbol{\phi}}(\mathbf{z}) \left[\ln p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z})\right] d\mathbf{z} - \int q_{\boldsymbol{\phi}}(\mathbf{z}) \frac{1}{q_{\boldsymbol{\phi}}(\mathbf{z})} \nabla_{\boldsymbol{\phi}} q_{\boldsymbol{\phi}}(\mathbf{z}) d\mathbf{z} \\
&= \mathbb{E}_q\left[\nabla_{\boldsymbol{\phi}} \ln q_{\boldsymbol{\phi}}(\mathbf{z}) \left(\ln p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z})\right)\right] \qquad (42)
\end{aligned}$$

since gradient calculated as in (42) is expectation taken with respect to $q_{\boldsymbol{\phi}}(\cdot)$, it is easy to calculate the estimated gradient $\widehat{\nabla_{\boldsymbol{\phi}}\mathcal{F}(\boldsymbol{\phi})}$ unbiasedly using samples from $q_{\boldsymbol{\phi}}(\mathbf{z})$ provided the analytical form of $q_{\boldsymbol{\phi}}(\mathbf{z})$ and the corresponding sampling from $q_{\boldsymbol{\phi}}(\mathbf{z})$ is readily available. Then the basic idea of stochastic gradient descent method is as following. With $\mathcal{F}(\boldsymbol{\phi})$ as objective function to optimized and assume that under some regular conditions $\mathcal{F}(\boldsymbol{\phi})$ is differentiable, thus $\nabla_{\boldsymbol{\phi}}\mathcal{F}(\boldsymbol{\phi})$ is its gradient and $\widehat{\nabla_{\boldsymbol{\phi}}\mathcal{F}(\boldsymbol{\phi})}$, we start from an initial value $\boldsymbol{\phi}^{(0)}$ and implement the following recursion for $t = 0, 1, \ldots$

$$\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)} + \rho_t \circ \widehat{\nabla_{\boldsymbol{\phi}}\mathcal{F}\left(\boldsymbol{\phi}^{(t)}\right)}$$

until a certain stopping criteria is satisfied. It can be justified that as long as $\rho_t, t \geqslant 0$ (learning rate) satisfies Robbins-Monro conditions, that is $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 = \infty$, then convergence of the sequence $\boldsymbol{\phi}^{(t)}$ will be to local optimum. There are many discussions on choosing learning rate ρ_t in literature and ADADELTA method (Zeiler, 2012) is one relatively easy to implement and hence widely used in practice. Specifically for this setting at iteration $t+1$, the i-th element ϕ_i of $\boldsymbol{\phi}$ is updated as

$$\phi_i^{(t+1)} = \phi_i^{(t)} + \Delta\phi_i^{(t)}$$

where the incremental step size $\Delta\phi_i^{(t)}$ is $\rho_i^{(t)} g_{\phi_i}^{(t)}$ and $g_{\phi_i}^{(t)}$ denotes the i-th component $\widehat{\nabla_{\boldsymbol{\phi}} \mathcal{F}\left(\boldsymbol{\phi}^{(t)}\right)}$ with $\rho_i^{(t)} g_{\phi_i}^{(t)}$ adaptively selected as

$$\rho_i^{(t)} = \frac{\sqrt{\mathbb{E}\left(\Delta_{\phi_i}^2\right)^{(t-1)} + \epsilon}}{\sqrt{\mathbb{E}\left(g_{\phi_i}^2\right)^{(t)} + \epsilon}}$$

where ϵ is a small positive constant and $\mathbb{E}\left(\Delta_{\phi_i}^2\right)^{(t)}$, $\mathbb{E}\left(g_{\phi_i}^2\right)^{(t)}$ are decayed running average of $\Delta\phi_i^{(t)^2}$ and $g_{\phi_i}^{(t)^2}$ respectively, defined by

$$\begin{aligned}
\mathbb{E}\left(\Delta_{\phi_i}^2\right)^{(t)} &= \zeta \mathbb{E}\left(g_{\phi_i}^2\right)^{(t-1)} + (1-\zeta)\,\Delta\phi_i^{(t)^2} \\
\mathbb{E}\left(g_{\phi_i}^2\right)^{(t)} &= \zeta \mathbb{E}\left(g_{\phi_i}^2\right)^{(t-1)} + (1-\zeta)\, g_{\phi_i}^{(t)^2}
\end{aligned}$$

Typically tuning parameters are specified as $\epsilon = 10^{-6}$ and $\zeta = 0.95$ the recursion is initialized at $\mathbb{E}\left(\Delta_{\phi_i}^2\right)^{(t)} = \mathbb{E}\left(g_{\phi_i}^2\right)^{(t)} = 0$.

Another way for implementing optimization is by directly minimizing the divergence between the approximate distribution and target posterior distribution. For discussion simplicity, introducing the following shorthand notation

$$\mathcal{D}(q) := \mathcal{D}(q, p_{\mathbf{x}}) \tag{43}$$

$$\widetilde{\mathcal{D}}(q) := \mathcal{D}(q, f) \tag{44}$$

where $p_{\mathbf{x}}$ as the shorthand notation is in consistence with the aforementioned target posterior $p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta})$ and correspondingly f denotes the joint density of observed data $\mathbf{x}$ and latent variables $\mathbf{z}$ (In other words, $\widetilde{\mathcal{D}}(q, f)$ refers to the negative of ELBO). Given the relation demonstrated in (20), our objective is to minimize $\mathcal{D}(q)$ by choosing q. Gradient Boosting is one frequently used method for this objective and specifically for this case the rough idea for implementing gradient boosting is by considering perturbation from q to $(1-\epsilon)q + \epsilon h$. It would be useful to derive the *functional derivative* of $\mathcal{D}(q)$, denoted by $\nabla\mathcal{D}(q)$, a direct derivation with shorthand notation is as following [7]

$$\begin{aligned}
\nabla\mathcal{D}(q) &= \nabla_q \left(q \log \frac{q}{f}\right) \\
&= \log \frac{q}{f} + q \frac{f}{q} \frac{1}{f} = \log \frac{q}{f} + 1
\end{aligned}$$

Recall the gradient boosting dynamics we introduced previously,

$$q_t = (1-\alpha_t)\, q_{t-1} + \alpha_t h_t$$

[7] Here we emphasize that q as the function value should be regarded as a univariate variable

and with $\epsilon \to 0$, Taylor expansion with respect to ϵ yields

$$\begin{aligned}\mathcal{D}\left((1-\epsilon)q+\epsilon h\right) &= \mathcal{D}\left(q+(h-q)\epsilon\right)\\ &= \mathcal{D}\left(q\right)+\epsilon\langle h-q, g\rangle + o\left(\epsilon^2\right)\end{aligned}$$

Consequently for a fixed q we just need to choose h to minimize $\langle h, g\rangle = \langle h, \nabla\mathcal{D}\left(q\right)\rangle$ and this also suggests that intuitively we need to approximately choose h to match the "negative" direction $\nabla\mathcal{D}\left(q\right)$, i.e. $-\nabla\mathcal{D}\left(q\right)$ if possible. And with the derived $\nabla\mathcal{D}\left(q\right)$ we focus on the following optimization problem instead

$$\mathcal{D}\left(q_t\right) = \mathcal{D}\left(q_{t-1}\right) + \alpha_t\langle h_t, \log\left(q_{t-1}/f\right)\rangle - \alpha_t\langle q_{t-1}, \log\left(q_{t-1}/f\right)\rangle + o\left(\alpha_t^2\right) \tag{45}$$

5.2.2 Application of variational bayes method

Recently there are some emerging literature discussing the application of variational expectation maximization method in estimating time-varying parameter models (TVP) (Kowal et al., 2019; Koop and Korobilis, 2020, henceforth KK20) based on the recent progress (Wang and Blei, 2019; Alquier and Ridgway, 2020) in theoretically justifying the asymptotic properties of variational bayes method. Following discussion is mainly about the basic data generating process (DGP) under TVP framework and how variational bayes method is applied for estimating the parameters specifying TVP modelling. Data is assumed to be generated as following

$$y_t = \beta_{1t}x_{1t} + \beta_{2t}x_{2t} + \ldots + \beta_{pt}x_{pt} + \sigma_t\varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1) \tag{46}$$

$$x_{j,t} \sim \mathcal{N}(0,1),\ j = 1,\ldots,p \tag{47}$$

$$\beta_{j,t} = s_{j,t}\times\theta_{j,t},\ s_{j,t} \text{ is 0-1 indicator variable} \tag{48}$$

$$\theta_{j,t} = \underline{\theta}_j + \underline{\rho}\left(\theta_{j,t-1} - \underline{\theta}_j\right) + \underline{\delta}\eta_{j,t},\ \eta_{j,t} \sim \mathcal{N}(0,1) \tag{49}$$

$$\log\left(\sigma_t^2\right) = \underline{\sigma}^2 + \underline{\phi}\left(\log\left(\sigma_{t-1}^2\right) - \underline{\sigma}^2\right) + \underline{\xi}\zeta_t,\ \zeta_t \sim \mathcal{N}(0,1) \tag{50}$$

$$\theta_{j,0} = \underline{\theta}_j,\quad \log\left(\sigma_0^2\right) = \underline{\sigma}^2 \tag{51}$$

Generally the model specified under the time-varying parameter setting is as following,

$$y_t = \mathbf{X}_t\boldsymbol{\beta}_t + \epsilon_t \tag{52}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t \tag{53}$$

where $\boldsymbol{\beta}_t = (\beta_{1t},\ldots,\beta_{pt})^\top$; $\epsilon_t \sim \mathcal{N}(0,\sigma_t^2)$ with σ_t^2 as time-varying parameter; $\boldsymbol{\eta}_t \sim \mathcal{N}(0,\boldsymbol{W}_t)$ with $\boldsymbol{W}_t$ as diagonal matrix $\boldsymbol{W}_t = \text{diag}\left(w_{1t},\ldots,w_{pt}\right)$. For later description simplicity, we introduce the notation $\boldsymbol{w}_t = \left[w_{1t},\ldots,w_{pt}\right]^\top$ as the $p\times 1$ vector collecting diagonal elements of $\boldsymbol{W}_t$. To see how variational Bayes method is applied in this time-varying framework. It is the generally the interests

of researchers to recover $p(\boldsymbol{\beta}_t, \boldsymbol{w}_t \mid \boldsymbol{y}_{1:t})$ from Bayesian perspective. We further assume the factorized structure for the variational density

$$q(\boldsymbol{\beta}_t, \boldsymbol{w}_t \mid \boldsymbol{y}_{1:t}) = q(\boldsymbol{\beta}_t \mid \boldsymbol{y}_{1:t}) \prod_{j=1}^{p} q(w_{j,t} \mid \boldsymbol{y}_{1:t}) \tag{54}$$

Similar logic applies to get the optimized functional form of ELBO akin to the the discussion of generic variation bayes method, which yields that ELBO is maximized by iterating through the following recursions

$$\begin{aligned}
q(\boldsymbol{\beta}_t \mid \boldsymbol{y}_{1:t}) &\propto \exp\left(\int \log p(\boldsymbol{\beta}_t, \boldsymbol{w}_t \mid \boldsymbol{y}_{1:t}) \prod_{j}^{p} q(w_{j,t} \mid \boldsymbol{y}_{1:t})\, d\boldsymbol{w}_t \right) \\
&\propto \exp\left(\int \log p(y_t, \boldsymbol{\beta}_t, \boldsymbol{w}_t \mid \boldsymbol{y}_{1:t-1}) \prod_{j}^{p} q(w_{j,t} \mid \boldsymbol{y}_{1:t})\, d\boldsymbol{w}_t \right) \qquad (55) \\
q(w_{j,t} \mid \boldsymbol{y}_{1:t}) &\propto \exp\left(\int \log p(\boldsymbol{\beta}_t, \boldsymbol{w}_t \mid \boldsymbol{y}_{1:t})\, q(\boldsymbol{\beta}_t \mid \boldsymbol{y}_{1:t})\, d\boldsymbol{\beta}_t \right) \\
&\propto \exp\left(\int \log p(y_t, \boldsymbol{\beta}_t, \boldsymbol{w}_t \mid \boldsymbol{y}_{1:t-1})\, q(\boldsymbol{\beta}_t \mid \boldsymbol{y}_{1:t})\, d\boldsymbol{\beta}_t \right), \; j = 1, \ldots, p \qquad (56)
\end{aligned}$$

Then the following recursive relationship can be set up

$$\begin{aligned}
q(\boldsymbol{\beta}_t \mid \boldsymbol{y}_{1:t}) &\propto \exp\left[\mathbb{E}_{q(w_t \mid \boldsymbol{y}_{1:t})} \left(\log p(y_t, \boldsymbol{\beta}_t, \boldsymbol{w}_t \mid \boldsymbol{y}_{1:t-1}) \right) \right] \\
&= \exp\left\{ \mathbb{E}_{q(w_t \mid \boldsymbol{y}_{1:t})} \log \left[p(y_t \mid \boldsymbol{\beta}_t, \boldsymbol{w}_t) p(\boldsymbol{\beta}_t \mid \boldsymbol{y}_{1:t-1}) p(\boldsymbol{w}_t \mid \boldsymbol{y}_{1:t-1}) \right] \right\} \\
&= p(y_t \mid \boldsymbol{\beta}_t, \boldsymbol{w}_t) \exp\left\{ \mathbb{E}_{q(\boldsymbol{w}_t \mid \boldsymbol{y}_{1:t})} \left[\log p(\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1}) + \log q(\boldsymbol{\beta}_t \mid \boldsymbol{y}_{1:t-1}) \right] \right\} \qquad (57) \\
&\quad \times \underbrace{\exp\left\{ \mathbb{E}_{q(\boldsymbol{w}_t \mid \boldsymbol{y}_{1:t})} \left[\log p(\boldsymbol{w}_t \mid \boldsymbol{y}_{1:t-1}) \right] \right\} \times \left[\log q(\boldsymbol{\beta}_t \mid \boldsymbol{y}_{1:t-1}) \right]^{-1}}_{\propto \text{ const}}
\end{aligned}$$

As suggested by George and McCulloch (1993), variable selection under this high-dimensional and dynamic setting is implemented based on the following spike-and-slab prior specification on the time-varying coefficients,

$$\begin{aligned}
\beta_{j,t} \mid \gamma_{j,t}, \tau_{j,t}^2 &\sim (1 - \gamma_{j,t}) \mathcal{N}\left(0, \underline{c}\tau_{j,t}^2\right) + \gamma_{j,t} \mathcal{N}\left(0, \tau_{j,t}^2\right) && (58) \\
\gamma_{j,t} \mid \pi_{0,t} &\sim \text{Bernoulli}(\pi_{0,t}) && (59) \\
\frac{1}{\tau_{j,t}^2} &\sim \text{Gamma}(g_0, h_0) && (60) \\
\pi_{0,t} &\sim \text{Beta}(1, 1). && (61)
\end{aligned}$$

That is , If $\gamma_{j,t} = 1$, the prior for $\beta_{j,t}$ has a normal prior with zero mean and variance $\tau_{j,t}^2$, while if $\gamma_{j,t} = 0$ the prior variance becomes $\underline{c}\tau_{j,t}^2$ which also implies that whenever variable selection is

required $\underline{c}$ has to be set as $\underline{c} \to 0$. But along with (58), we also have dynamic transition such that

$$\beta_{j,t} \mid \beta_{j,t-1}, w_{j,t} \sim \mathcal{N}\left(\beta_{j,t-1}, w_{j,t}\right) \tag{62}$$

Then it is possible for us to combine the prior information imposed on $\beta_{j,t}$ jointly with (58) and (62) such that

$$\boldsymbol{\beta}_t = \widetilde{\boldsymbol{F}}_t \boldsymbol{\beta}_{t-1} + \widetilde{\boldsymbol{\eta}}_t \tag{63}$$

where

$$\begin{aligned}
\widetilde{\boldsymbol{\eta}}_t &\sim \mathcal{N}\left(\mathbf{0}, \widetilde{\boldsymbol{W}}_t\right) \\
\widetilde{\boldsymbol{W}}_t &= \left[\mathbb{E}\left(\boldsymbol{W}_t\right)^{-1} + \mathbb{E}\left(\boldsymbol{V}_t\right)^{-1}\right]^{-1} \\
\widetilde{\boldsymbol{F}}_t &= \widetilde{\boldsymbol{W}} \times \mathbb{E}\left(\boldsymbol{W}_t\right)^{-1} \\
\boldsymbol{W}_t &= \operatorname{diag}\left(w_{1,t}, \ldots, w_{p,t}\right) \\
\boldsymbol{V}_t &= \operatorname{diag}\left(v_{1,t}, \ldots, v_{p,t}\right)
\end{aligned}$$

A quick demonstration for the replicated Monte Carlo experiment of KK20 is as following

[Place Figure 2 about here]

[Place Figure 3 about here]

My implementation instead is based on the optimized hybrid `R` and `C++` codes relying on Mircrosoft R Open distribution. This implementation automatically implements parallel matrix computation based on the `C++` Armadillo linear algebra template and consequently would improve computational efficiency significantly on multi-cores system. Armadillo as the `C++` library for linear algebra and & scientific computing is initially developed and actively maintained by Conrad Sanderson from Griffith University. More details about hybrid coding in `R` and `C++` are covered comprehensively in the classic textbook Eddelbuettel (2013). The associated `R` package `vbdvsarmadillo` and more details about implementation are both available at

https://www.yaohanchen.com/post/vbdvsrcpp/

5.3 Variational Bayes with Intractable Likelihood

As in Tran et al. (2017), let us consider the a generic scenario where there are observed random variables $\mathbf{y}$, latent variables $\mathbf{z}$ and the associated parameters $\boldsymbol{\theta} \in \Theta$. Within the Bayesian analytical framework, in general we are interested in

$$\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{64}$$

which is usually referred to the posterior distribution on which we make posterior inference. However, $p(\mathbf{y} \mid \boldsymbol{\theta})$ is required for constructing sampler like MCMC, which under some circumstance is not exactly available. Hence variational bayes method instead provides an alternative for recovering $\pi(\boldsymbol{\theta})$ when the the likelihood function $p(\mathbf{y} \mid \boldsymbol{\theta})$ is not available. The key idea is discussed as following. In fact, we can firstly apply filtering techniques like particle filter to estimate the likelihood function $p(\mathbf{y} \mid \boldsymbol{\theta})$, denoted by $\widehat{p}_N(\mathbf{y} \mid \boldsymbol{\theta})$ with N specifying the number of particles used for estimating, and we use $\mathbf{z}$ to denote the difference between the estimated likelihood function and the true likelihood function in log scale such that

$$\mathbf{z} = \log \widehat{p}_N(\mathbf{y} \mid \boldsymbol{\theta}) - \log p(\boldsymbol{y} \mid \boldsymbol{\theta})$$

which is the form that we can alternatively write as $\widehat{p}_N(\boldsymbol{y} \mid \boldsymbol{\theta}) = e^{\mathbf{z}} p(\boldsymbol{y} \mid \boldsymbol{\theta})$. Given the unbiasedness of this estimated likelihood function, we thus have the following fact

$$\int e^{\mathbf{z}} g_N(\mathbf{z} \mid \boldsymbol{\theta}) d\mathbf{z} = 1$$

where $g_N(\mathbf{z} \mid \boldsymbol{\theta})$ refers to the density function of random variable z conditional on parameter $\boldsymbol{\theta}$. With this newly introduced random variable z, we can define the following augmented likelihood function $\pi_N(\boldsymbol{\theta}, \mathbf{z})$ on $\Theta \times \mathbf{R}$, [8]

$$\pi_N(\boldsymbol{\theta}, \mathbf{z}) = \frac{p(\boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}) e^{\mathbf{z}} g_N(\mathbf{z} \mid \boldsymbol{\theta})}{p(\mathbf{y})} = \pi(\boldsymbol{\theta}) e^{\mathbf{z}} g_N(\mathbf{z} \mid \boldsymbol{\theta}). \tag{65}$$

From which we can derive that

$$\begin{aligned} \log p(\mathbf{y}) &= \log \left[\frac{p(\boldsymbol{\theta}) P(\boldsymbol{y} \mid \boldsymbol{\theta}) e^{\mathbf{z}} g_N(\mathbf{z} \mid \boldsymbol{\theta})}{\pi_N(\boldsymbol{\theta}, \mathbf{z})} \right] \\ &= \log \left[\frac{p(\boldsymbol{\theta}) \widehat{p}_N(\mathbf{z} \mid \boldsymbol{\theta})}{q_{\lambda,N}(\boldsymbol{\theta}, \mathbf{z})} \right] + \log \left[\frac{q_{\lambda,N}(\boldsymbol{\theta}, \mathbf{z}) g_N(\boldsymbol{\theta}, \mathbf{z})}{\pi_N(\boldsymbol{\theta}, \mathbf{z})} \right] \end{aligned} \tag{66}$$

where $q_{\lambda,N}(\boldsymbol{\theta}, \mathbf{z})$ is the correspondingly introduced density function over the space of $\boldsymbol{\theta}$ and $\mathbf{z}$. Hence taking integration on both sides over $\boldsymbol{\theta}$, $\mathbf{z}$ with respect to $q_{\lambda,N}(\boldsymbol{\theta}, \mathbf{z})$ on both sides of the above

[8] Augmented likelihood function defined here is guaranteed to be well-defined density function given that $\int e^{z} g_N(z \mid \boldsymbol{\theta}) dz = 1$ and $\pi(\boldsymbol{\theta})$ is well-defined posterior. In other words, integration of $\pi_N(\boldsymbol{\theta}, z)$ over $\boldsymbol{\theta}$ and $\boldsymbol{z}$ is exactly equal 1.

equation yields

$$\log p(\mathbf{y}) = \underbrace{\int q_{\lambda,N}(\boldsymbol{\theta},\mathbf{z}) \log\left[\frac{p(\boldsymbol{\theta})\widehat{p}_N(\mathbf{z}\mid\boldsymbol{\theta})}{q_{\lambda,N}(\boldsymbol{\theta},\mathbf{z})}\right] d\boldsymbol{\theta} d\mathbf{z}}_{\mathrm{LB}(\lambda)} + \underbrace{\int q_{\lambda,N}(\boldsymbol{\theta},\mathbf{z}) \log\left[\frac{q_{\lambda,N}(\boldsymbol{\theta},\mathbf{z}) g_N(\boldsymbol{\theta},\mathbf{z})}{\pi_N(\boldsymbol{\theta},\mathbf{z})}\right] d\boldsymbol{\theta} d\mathbf{z}}_{\mathrm{KL}(\lambda)} \quad (67)$$

Remark 5.5

- $q_{\lambda,N}(\boldsymbol{\theta},\mathbf{z})$ as the constructed density function can especially take the following form

  $$q_{\lambda,N}(\boldsymbol{\theta},\mathbf{z}) = q_\lambda(\boldsymbol{\theta}) g_N(\boldsymbol{\theta},\mathbf{z})$$

  and this structure separates the dependence of λ on N and hence could bring some convenience, this will be discussed later.

- $\mathrm{KL}(\lambda)$ is the associated Kullback-Leibler divergence by comparing $q_{\lambda,N}(\boldsymbol{\theta},\mathbf{z})$ with the true augmented likelihood function $\pi_N(\boldsymbol{\theta},\mathbf{z})$. This is a function of λ and always no less than 0.

- $\mathrm{LB}(\lambda)$ as the function of λ is justified as the lower bound of the marginal log likelihood $\log p(\boldsymbol{y})$ given the fact that $\mathrm{KL}(\lambda) \geqslant 0$. $\mathrm{LB}(\lambda)$ is later applied for checking the convergence of gradient.

5.4 Dynamic Shrinking Process

Before we move on to the application of variational bayes method, it necessary to briefly discuss a related branch of literature recently appearing in statistics discussing shrinkage dynamically. One of the representative work is done by Kowal et al. (2019), and the following discussion is mainly based on the framework of this paper. All the stuff to be discussed is closely related with the concepts about global-local prior defined as following

$$\omega_t \mid \tau, \lambda_t \overset{indep}{\sim} \mathcal{N}\left(0, \tau^2\lambda_t^2\right), \quad (68)$$

And $h_t = \log\left(\tau^2\lambda_t^2\right)$ is a process following the general dependent data model

$$h_t = \mu + \psi_t + \eta_t, \quad \eta_t \overset{iid}{\sim} Z(\alpha,\beta,0,1) \quad (69)$$

where

$$\begin{aligned} \mu &= \log\left(\tau^2\right) \\ \psi_t + \eta_t &= \log\left(\lambda_t^2\right) \end{aligned}$$

and $Z(\alpha, \beta, \mu_z, \sigma_z)$ denotes the Z-distribution with density function specified as following

$$[z] = [\sigma_z B(\alpha, \beta)]^{-1} \{\exp[(z - \mu_z)/\sigma_z]\}^{\alpha} \{1 + \exp[(z - \mu_z)/\sigma_z]\}^{-(\alpha+\beta)}, z \in \mathbb{R} \tag{70}$$

and $B(\cdot, \cdot)$ refers to the Beta function. Then the dynamic shrinkage process is modelled as following

$$h_{t+1} = \mu + \phi(h_t - \mu) + \eta_t, \quad \eta_t \overset{iid}{\sim} Z(\alpha, \beta, 0, 1) \tag{71}$$

Remark 5.6 ψ_t plays the role governing the adaptive shrinkage process with the following two emphasized special case

- Replacing ψ_t in (70) with $\phi(h_t - \mu)$ in (71) yields equivalence.
- Replacing ψ_t in (70) with $\boldsymbol{z}_t^{\top}\boldsymbol{\alpha}$ for a vector of predictors yields the equivalence for linear regression framework.

Remark 5.7 Z-distribution arises from Beta distribution by the following discussion

- $B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du = \int_0^1 (1-u)^{\alpha-1} u^{\beta-1} du = B(\beta, \alpha)$.
- Define $u = \frac{1}{1+\exp[(z-\mu_z)/\sigma_z]}$, we can readily check the $[z]$ of the previous functional form is indeed a probability density function,

$$\int [\sigma_z B(\alpha, \beta)]^{-1} \{\exp[(z - \mu_z)/\sigma_z]\}^{\alpha} \{1 + \exp[(z - \mu_z)/\sigma_z]\}^{-(\alpha+\beta)} dz$$

$$= \frac{1}{\sigma_z B(\alpha, \beta)} \int_0^1 (1-u)^{\alpha} u^{\beta} \left| \frac{-\sigma_z}{(1-u)u} \right| du = \frac{\sigma_z}{B(\alpha, \beta)} \int_0^1 (1-u)^{\alpha-1} u^{\beta-1} du = 1$$

- The first moment of Z-distribution is μ_z,

$$\int [\sigma_z B(\alpha, \beta)]^{-1} \{\exp[(z - \mu_z)/\sigma_z]\}^{\alpha} \{1 + \exp[(z - \mu_z)/\sigma_z]\}^{-(\alpha+\beta)} z \, dz$$

$$= \frac{1}{\sigma_z B(\alpha, \beta)} \int_0^1 (1-u)^{\alpha} u^{\beta} \left[\log\left(\frac{1}{u} - 1\right) \sigma_z + \mu_z \right] \left| \frac{-\sigma_z}{(1-u)u} \right| du = \mu_z$$

where we use the fact that

$$\begin{aligned}\frac{1}{B(\alpha,\beta)}\int_0^1 (1-u)^{\alpha-1}u^{\beta-1}\log(1-u)\ du &= \frac{1}{B(\alpha,\beta)}\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\log u\ du \\ &= \psi(\alpha)-\psi(\alpha+\beta) \\ \frac{1}{B(\alpha,\beta)}\int_0^1 (1-u)^{\alpha-1}u^{\beta-1}\log u\ du &= \psi(\beta)-\psi(\beta+\alpha)\end{aligned}$$

which implies that the first moment of $Z(\alpha,\beta,\mu_z,\sigma_z)$ should be $\psi(\alpha)-\psi(\beta)+\mu_z$ and hence it is symmetric with the origin concentrating on μ_z if and only inf $\alpha=\beta$. $\psi(\cdot)$ refers to digamma function, which is essentially the logarithm derivative of gamma function.

Remark 5.8 Some useful discussion about special distributions

- If specifically a random variable is following Z distribution such that $\eta \sim Z(\alpha,\beta,\mu_z,1)$, $\mu_z \in \mathbb{R}$. Then $\kappa = 1/(1+\exp(\eta)) \sim TPB(\beta,\alpha,\exp(\mu_z))$, where $\kappa \sim TPB(\beta,\alpha,\gamma)$ denotes the three parameter Beta distribution (Armagan et al., 2011) with density specified as following

$$[\kappa] = [B(\beta,\alpha)]^{-1}\gamma^{\beta}\kappa^{\beta-1}(1-\kappa)^{\alpha-1}[1+(\gamma-1)\kappa]^{-(\alpha+\beta)}, \kappa \in (0,1), \gamma > 0 \tag{72}$$

Proof. Recall that $\eta \sim Z(\alpha,\beta,\mu_z,1)$ has the following distribution with $\sigma_z = 1$

$$[z] = [\sigma_z B(\alpha,\beta)]^{-1}\left\{\exp\left[(z-\mu_z)/\sigma_z\right]\right\}^{\alpha}\left\{1+\exp\left[(z-\mu_z)/\sigma_z\right]\right\}^{-(\alpha+\beta)}$$

which implies that if we define $\lambda^2 = \exp(\eta)$, then

$$\begin{aligned}\left[\lambda^2\right] &\propto \left(\lambda^2\right)^{-1}\left\{\exp\left[\log\left(\lambda^2\right)-\mu_z\right]\right\}^{\alpha}\left\{1+\exp\left[\log\left(\lambda^2\right)-\mu_z\right]\right\}^{-(\alpha+\beta)} \\ &\propto \left(\lambda^2\right)^{\alpha-1}\left[1+\lambda^2/\exp(\mu_z)\right]^{-(\alpha+\beta)}\end{aligned}$$

and for $\kappa = 1/\left(1+\lambda^2\right)$,

$$\begin{aligned}[\kappa] &\propto \kappa^{-2}\left[\kappa^{-1}-1\right]^{\alpha-1}\left[1+\left(\kappa^{-1}-1\right)/\exp(\mu_z)\right]^{-(\alpha+\beta)} \\ &\propto \kappa^{-2-(\alpha-1)}(1-\kappa)^{\alpha-1}\left\{\kappa^{-1}\left[\kappa\exp(\mu_z)+(1-\kappa)\right]\right\}^{-(\alpha+\beta)} \\ &\propto (1-\kappa)^{\alpha-1}\kappa^{\beta-1}\left[\kappa\exp(\mu_z)+(1-\kappa)\right]^{-(\alpha+\beta)}\end{aligned}$$

□

- In a general framework, we consider λ^2 distributed with density

$$\left[\lambda^2\right] \propto \left(\lambda^2\right)^{\alpha-1}\left(1+\lambda^2\right)^{-(\alpha+\beta)}, \lambda > 0 \tag{73}$$

which is referred to inverted-Beta distribution, $IB(\beta,\alpha)$. Similar to the previous discussion, we can show that $\kappa = 1/\left(1+\lambda^2\right) \sim Beta\left(\beta,\alpha\right)$.

Proof.

$$\begin{aligned}\left[\kappa\right] &\propto \kappa^{-2}\left(\frac{1}{\kappa}-1\right)^{\alpha-1}\left(\frac{1}{\kappa}\right)^{-(\alpha+\beta)} \\ &\propto \left(1-\kappa\right)^{\alpha-1}\kappa^{-2+1-\alpha+\alpha+\beta} = \left(1-\kappa\right)^{\alpha-1}\kappa^{\beta-1},\end{aligned}$$

□

- Previous discussion hence can be summarized as following claim

$$\lambda^2 \sim IB(\beta,\alpha) \Leftrightarrow \kappa = 1/\left(1+\lambda^2\right) \sim Beta(\beta,\alpha) \Leftrightarrow \eta = \log\left(\lambda^2\right) = \log\left(\kappa^{-1}-1\right) \sim Z(\alpha,\beta,0,1).$$

Let us temporarily focus on the following process and discuss some associated properties

$$\kappa_{t+1} = \frac{1}{1+\tau^2\lambda_{t+1}^2} \tag{74}$$

The following theorem plays the pivotal rule

Theorem 5.3 *For the dynamic process specified as in (71), the conditional prior distribution of* κ_{t+1} *is*

$$\left[\kappa_{t+1} \mid \{\kappa_s\}_{s\leq t},\phi,\tau\right] \sim TPB\left(\beta,\alpha,\tau^{2(1-\phi)}\left[\frac{1-\kappa_t}{\kappa_t}\right]^{\phi}\right) \tag{75}$$

Proof. Recall the dynamic process specified as in (71), we have

$$\left[h_{t+1} \mid h_t,\phi,\mu\right] \sim Z\left(\alpha,\beta,\mu+\phi\left(h_t-\mu\right),1\right)$$

By using the results discussed in the aforementioned remarks, we have the conditional distribution of κ_{t+1} is

$$\left[\kappa_{t+1} \mid h_t,\phi,\mu\right] \sim TPB\left(\beta,\alpha,\exp\left(\mu+\phi\left(h_t-\mu\right)\right)\right) \tag{76}$$

□

Theorem 5.4 *For the dynamic process specified as in (71) and* $\alpha = \beta = \frac{1}{2}$*, the conditional prior distribution satisfies*

$$\mathbb{P}\left(\kappa_{t+1} < \varepsilon \mid \{\kappa_s\}_{s\leq t},\phi\right) \to 1 \tag{77}$$

as $\kappa_t \to 0$ *or any fixed* $\varepsilon \in (0,1)$ *and* $\phi \neq 0$

It is worthwhile discussing Pólya-Gamma distribution for introducing techniques mentioned later that have widely discussed in literature such as Barndorff-Nielsen et al. (1982) and Polson et al. (2013). The Pólya-Gamma distributions, denoted bt $PG(b,c)$ is a subset of the class of infinite

convolutions of gamma distributions. As a special case, $PG(1,0)$ is referred to the carefully chosen element of the the class of infinite convolutions of exponentials. In general, we have the following factorization, all the listed results are the consequence of Weierstrass factorization theorem

$$\begin{aligned}
\sin(\pi z) &= \pi z \prod_{k=1}^{\infty}\left(1-\left(\frac{1}{k}\right)^2\right) \\
\cos(\pi z) &= \prod_{k=0}^{\infty}\left(1-\left(\frac{z}{k+\frac{1}{2}}\right)^2\right) \\
\sinh(z) &= z\prod_{k=1}^{\infty}\left(1+\frac{z^2}{(\pi k)^2}\right) \\
\cosh(z) &= \prod_{k=1}^{\infty}\left(1+\frac{z^2}{(\pi k-\pi/2)^2}\right)
\end{aligned}$$

Based on these facts implied from Weierstrass factorization and that

- If a random variable ω is distributed following $PG(1,0)$, we are able to obtain the Laplace transform (with characterization function evaluated at it, where i denotes imaginary unit) as following

  $$\mathbb{E}\left[\exp\left(-\omega t\right)\right] = \cosh^{-1}\left(\sqrt{t/2}\right)$$

  Moreover, if ω is distributed following $PG(b,0)$, the corresponding Laplace transform is as following

  $$\mathbb{E}\left[\exp\left(-\omega t\right)\right] = \prod_{k=1}^{\infty}\left(1+\frac{t}{2\pi^2\left(k-1/2\right)^2}\right)^{-b} = \frac{1}{\cosh^b\left(\sqrt{t/2}\right)}. \tag{78}$$

- Gamma distribution denoted as $\mathrm{Ga}(\alpha,\beta)$, the characterization function generally takes the following form

  $$\left(1-\frac{i\tau}{\beta}\right)^{-\alpha}$$

  and accordingly with τ evaluated at it (complex number with only imaginary part), we have Laplace transform of Gamma distribution as following

  $$\left(1+\frac{t}{\beta}\right)^{-\alpha}$$

  Replacing $\alpha = b$ and $\beta = 1$ directly implies that Laplace transform of $\mathrm{Ga}(b,1)$ is

  $$\left(1+t\right)^{-b}$$

- Suppose that random variable ω is constructed as following

$$\omega \overset{D}{=\!=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2}$$

and $g_k \sim \text{Ga}(b,1)$ (g_k are mutually independent), then given that the Laplace transform of $\text{Ga}(b,1)$ is $(1+t)^{-1}$, we have the Laplace transform of this constructed random variable ω takes exactly the form of (78). This is why ω refers to Pólya-Gamma distribution.

As for the extension from $PG(b,0)$ to $PG(b,c)$, just introduce the following density function for $PG(b,c)$.

$$p(\omega \mid b,c) = \frac{\exp\left(-\frac{c^2}{2}\omega\right) p(\omega \mid b,0)}{\mathbb{E}_\omega\left\{\exp\left(-\frac{c^2}{2}\omega\right)\right\}} \tag{79}$$

where $\mathbb{E}_\omega$ is taken with respect to $p(\omega \mid b,0)$. Based on the detailed discussion in Polson et al. (2013) (equation (6) specifically), it is possible to demonstrate that $\omega \sim PG(b,c)$ is equivalent as following

$$\omega \overset{D}{=\!=} \sum_{k=1}^{\infty} \frac{\text{Ga}(b,1)}{d_k} = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{\text{Ga}(b,1)}{\left(k-\frac{1}{2}\right)^2 + c^2/\left(4\pi^2\right)} \tag{80}$$

5.5 Quick note EB coverage interval

Useful facts summarized from (Morris, 1983; Carlin and Louis, 2000), If we specify likelihood as normal such that $f(y \mid \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right)$, and $\theta \sim \mathcal{N}(\mu, \tau^2)$, $\theta \in \mathbf{R}$, $\mu \in \mathbf{R}$. Furthermore, we assume that σ is a constant positive known parameter and μ, τ are known hyperparameters as well. The posterior of θ for this specification is

$$p(\theta \mid y) = \mathcal{N}\left(\theta \;\middle|\; \frac{\sigma^2\mu + \tau^2 y}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$$

For a specific setting ($\mu = 0$), EB estimator for θ_i is given as $w_{EB}Y_i$ with $w_{EB} = \frac{\tau^2}{\sigma^2+\tau^2}$. we want to check whether θ_i is contained in a posterior stochastic interval for a specific critical value χ (in some sense, χ could be interpreted as radius of neighbourhood with its center on $w_{EB}Y_i$). Or in other words, non-coverage posterior confidence interval refers to those stochastic intervals, CI, $w_{EB}Y_i \pm \chi \cdot w_{EB}\sigma$ in which a fixed θ_i is not contained. All these intervals correspond to the case

$$|\theta_i - w_{EB}Y_i| \gtrless \chi \cdot w_{EB}\sigma \Leftrightarrow \left|\frac{\theta_i - w_{EB}Y_i}{w_{EB}\sigma}\right| \gtrless \chi$$

But note that

$$\mathbb{E}\left[w_{EB}Y_i - \theta_i\right] = \left(w_{EB}-1\right)\theta_i \quad \text{Var}\left[w_{EB}Y_i - \theta_i\right] = w_{EB}^2\sigma^2$$

hence the corresponding defined statistics $\frac{w_{EB}Y_i - \theta_i}{w_{EB}\sigma}$ is normally distributed with mean equal to b_i

and variance equal to 1, where b_i is of the following form

$$b_i = \frac{(w_{EB} - 1)\theta_i}{w_{EB}\sigma}$$

which justifies the definition such that

$$\frac{\theta_i - w_{EB}Y_i}{w_{EB}\sigma} - (-b_i) = Z \sim \mathcal{N}(0,1) \text{ and } \left|\frac{\theta_i - w_{EB}Y_i}{w_{EB}\sigma}\right| = |Z - b_i|$$

and the claim that probability of non-coverage is

$$r(b_i, \chi) = \mathbb{P}\left(\left|\frac{\theta_i - w_{EB}Y_i}{w_{EB}\sigma}\right| \geqslant \chi\right) = \mathbb{P}\left(|Z - b_i| \geqslant \chi\right) = \Phi\left(-\chi - b\right) + \Phi\left(-\chi + b\right) \tag{81}$$

where $\Phi(\cdot)$ denotes the cdf of standard normal distribution. $\Phi(\cdot)$ is differentiable and hence $r(b_i, \chi)$ is differentiable w.r.t. b_i, which is important for constructing the optimal functional form of $r\left(b, \chi\right)$. Note for the all the previous discussion, we keep our discussion based on fixing θ_i. Under Bayesian framework or the scenario where random effect is allowed, b_i by definition as the function of θ_i inherits the randomness from θ_i. In general θ_i can follow any form of feasible distribution but we obtain our EB estimator by pretending imposing normality. As the result, once we consider integrating out the uncertainty from b_i, which is actually inherited from θ_i, we just impose the structure that $b_i \sim F$, where F is used for generally describing the distribution of b_i.

Remark 5.9

- Given that Z as the standard normal distribution is symmetric on $\mathbf{R}$, $r(b, \chi)$ is symmetric in b.
- For all $t \geqslant 0$, we can define $r_0(t, \chi) = r(\sqrt{t}, \chi)$ and hence the following two optimization problems are equivalent

$$\begin{aligned} &\sup_F \mathbb{E}_F\left[r(t, \chi)\right] \\ \text{s.t.} \quad &\mathbb{E}_F[t^2] = m_2 \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} &\sup_F \mathbb{E}_F\left[r_0(t, \chi)\right] \\ \text{s.t.} \quad &\mathbb{E}_F[t] = m_2 \end{aligned} \tag{82}$$

  and we denote the corresponding optimal value as $\rho(m_2, \chi)$.

- The following statement (or lemma) discussed in Carolan (2002) gives a relatively formal definition about *least concave majorant* for the univariate case.
  Claim: *Suppose g is a function defined on a set containing $[a, b]$, where a can equal $-\infty$ or b can equal ∞. Then g^* defined as the least concave majorant of g for $t \in [a, b]$ is*

$$g^*(t) = \sup_{a \leq x_1 \leq t} \sup_{t \leq x_2 \leq b} \left\{\frac{(x_2 - t)g(x_1) + (t - x_1)g(x_2)}{x_2 - x_1}\right\} \tag{†}$$

where 0/0 is defined as $g(t)$ when $x_1 = x_2 = t$. Alternatively, we can concisely define g^* as

$$g^*(t) = \inf\{\tilde{g}(t) : \tilde{g}_t \geqslant g(t), \tilde{g} \text{ concave}\}.$$

- Let $\bar{r}(t,\chi)$ be the *least concave majorant* of $r_0(t,\chi)$, it can be shown that $\rho(m_2,\chi) = \bar{r}(m_2,\chi)$, where $\rho(m_2,\chi)$ refers to the optimal functional form that we want to obtain when t is evaluated at $t = m_2$.

- Previous bullet point implies that optimal functional form of $\rho(t,\chi)$ with t evaluated at the moment restriction $t = m_2$ is essentially given by the least concave majorant of $r_0(t,\chi)$, $\bar{r}(t,\chi)$. A natural question then will be what is the exactly the functional form of $\bar{r}(t,\chi)$. By the definition of least concave majorant, a straightforward motivating construction will be using linear combination as the approximation. It can be proved that $\bar{r}(t,\chi)$ takes the following functional form akin to (†)

$$\bar{r}(t,\chi) = \sup_{u \geq t} \left\{ (1 - t/u) r_0(0,\chi) + \frac{t}{u} r_0(u,\chi) \right\}, \tag{83}$$

6 Latent Factors

Dimension reduction is not only the major concern in machine learning literature but also other fields like Finance and Statistics. Factor structure as the major dimension reduction tools have been widely discussed in literature, both from theoretical and empirical perspective. Recently there are some meaningful discussions (see Lettu and Pelger, 2020a,b) about extending the conventional methodologies of principal component analysis (PCA) to a relatively more general framework by constructing the objective function accounting for the factor approximation error simultaneously from the time-series and cross-section dimension.

Basically, we observe excess return of N assets over T periods and factor structure is imposed as following

$$X_{t,n} = F_t^\top \Lambda_n + e_{t,n} \quad n = 1,\ldots,N \quad t = 1,\ldots,T \tag{84}$$

where if there are K factors, then both F_t and Λ_n are a $K \times 1$ vectors. In matrix notation this reads as

$$\underbrace{\mathbf{X}}_{T\times N} = \underbrace{\mathbf{F}}_{T\times K} \underbrace{\mathbf{\Lambda}^\top}_{K\times N} + \underbrace{\mathbf{e}}_{T\times N} \tag{85}$$

In practice, data is usually demeaned firstly and then applied with PCA. Moreover since Factor structure is imposed for summarizing the time-series variation shared commonly by cross-sectional data, the demeaned process is taken over time-series dimension. For description simplicity, introducing

the following notation first,

$$\text{projection matrix demeaning time-series} : \mathbf{M}_1 = \mathbf{I}_T - \tfrac{1}{T}\iota\iota^\top$$

$$\text{projection matrix} : \mathbf{M}_{\mathbf{\Lambda}} = \mathbf{I}_N - \mathbf{\Lambda}\left(\mathbf{\Lambda}^\top\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}^\top$$

where ι is a $T \times 1$ vector of 1's. With this notation introduced, the demeaned data (across time-series dimension) is readily represented in matrix notation as

$$\tilde{\mathbf{X}} = \mathbf{M}_1\mathbf{X}$$

$$\tilde{\mathbf{F}} = \mathbf{M}_1\mathbf{F}$$

The key step in implementing PCA with respect to $\tilde{\mathbf{X}}$ is calculating the sample second moment of $\tilde{\mathbf{X}}$:

$$\frac{1}{T}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \frac{1}{T}\mathbf{X}^\top\mathbf{M}_1\mathbf{X} = \frac{1}{T}\mathbf{X}^\top\mathbf{X} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top$$

Denote $\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{F}}$ as the estimated factor loadings and factors respectively, the following equivalence can be proved

Proposition 6.1 *The k-th column of $\hat{\mathbf{\Lambda}}$ is proportional to the k-th eigenvector extracted from*

$$\frac{1}{T}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \frac{1}{T}\mathbf{X}^\top\mathbf{X} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^\top$$

where $\overline{\mathbf{X}}$ as the $N \times 1$ vector denotes the time-series sample mean of cross-sectional excess returns, that is

$$\overline{\mathbf{X}} = \frac{1}{T}\mathbf{X}^\top\iota$$

and $\hat{\mathbf{\Lambda}}$, $\hat{\tilde{\mathbf{F}}}$ jointly solves the following minimization problem

$$\min_{\mathbf{\Lambda},\tilde{\mathbf{F}}} \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\left(\tilde{\mathbf{X}}_{t,n} - \tilde{F}_t^\top\Lambda_n\right)^2 = \min_{\mathbf{\Lambda}} \frac{1}{NT}\operatorname{trace}\left[\left(\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right)^\top\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right] \tag{86}$$

and

$$\tilde{\mathbf{F}} = \tilde{\mathbf{X}}\mathbf{\Lambda}\left(\mathbf{\Lambda}^\top\mathbf{\Lambda}\right)^{-1}$$

Proof. Justification for $\tilde{\mathbf{F}} = \tilde{\mathbf{X}}\mathbf{\Lambda}\left(\mathbf{\Lambda}^\top\mathbf{\Lambda}\right)^{-1}$ is simply based on the observation that we are able to commute the summation order of the objective function in a way such that the first summation is taken over n with t fixed and then the second summation is taken over t. First order condition for the summation over n with t fixed is the simply the OLS-style first order condition. Then substituting a given $\tilde{\mathbf{F}}$ of this form back yields the error term formulated as

$$\tilde{\mathbf{X}} - \tilde{\mathbf{F}}\mathbf{\Lambda}^\top = \tilde{\mathbf{X}} - \tilde{\mathbf{X}}\mathbf{\Lambda}\left(\mathbf{\Lambda}^\top\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}^\top = \tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}. \tag{87}$$

It is obvious that the objective function of this minimization problem could be represented as

$$\begin{aligned} & \frac{1}{NT}\operatorname{trace}\left[\left(\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right)^{\top}\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right] \\ = & \frac{1}{NT}\operatorname{trace}\left[\mathbf{M}_{\mathbf{\Lambda}}\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right] \\ = & \frac{1}{NT}\operatorname{trace}\left[\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right] \\ = & \frac{1}{NT}\operatorname{trace}\left[\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}}\right] - \frac{1}{NT}\operatorname{trace}\left[\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}}\mathbf{\Lambda}\left(\mathbf{\Lambda}^{\top}\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}^{\top}\right] \end{aligned}$$

Following discussion is useful for the final proof.

1. $\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}}$ is symmetric real matrix and hence theoretically this is supposed to be represented as

$$\mathbf{QUQ}^{\top}, \quad \mathbf{U} = \operatorname{diag}\left(u_1, \ldots, u_N\right), \quad u_1 \geqslant, \ldots, \geqslant u_N \geqslant 0, \quad \mathbf{Q}^{\top}\mathbf{Q} = \mathbf{QQ}^{\top} = \mathbf{I}_N$$

2. $\mathbf{\Lambda}\left(\mathbf{\Lambda}^{\top}\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}^{\top}$ is an idempotent matrix, which implies it is supposed to be able to be represented as

$$\mathbf{PVP}^{\top}, \quad \mathbf{V} \text{ is a diagonal matrix with diagonal entries either 1 or 0}, \quad \mathbf{P}^{\top}\mathbf{P} = \mathbf{PP}^{\top} = \mathbf{I}_N$$

These two points jointly imply that

$$\begin{aligned} \min_{\mathbf{\Lambda}} \frac{1}{NT}\operatorname{trace}\left[\left(\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right)^{\top}\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right] \Leftrightarrow & \max_{\mathbf{\Lambda}} \frac{1}{NT}\operatorname{trace}\left[\mathbf{QUQ}^{\top}\mathbf{PVP}^{\top}\right] \\ = & \frac{1}{NT}\operatorname{trace}\left[\mathbf{UQ}^{\top}\mathbf{PVP}^{\top}\mathbf{Q}\right] \end{aligned}$$

Hence the upper bound for $\frac{1}{NT}\operatorname{trace}\left[\mathbf{UQ}^{\top}\mathbf{PVP}^{\top}\mathbf{Q}\right]$ is the summation of the largest K eigenvalues $u_1, \ldots, u_K$ of $\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}}$, [9] which is necessarily guaranteed by setting the first K columns of $\mathbf{Q}$ proportional to $\mathbf{\Lambda}$. □

However, arbitrage-theory predicts that factors should approximately price the cross-section of expected excess returns well, which also implies that the excess returns should directly enter into the objective function rather than the demeaned counterpart substituted in objective function as in the standard PCA procedure. Hence the following minimization problem should be considered

[9] A key fact for justifying this claim is $\mathbf{Q}^{\top}\mathbf{Q} = \mathbf{QQ}^{\top} = \mathbf{I}_N$ and $\mathbf{P}^{\top}\mathbf{P} = \mathbf{PP}^{\top} = \mathbf{I}_K$.

instead. [10]

$$\min_{\mathbf{\Lambda},\mathbf{F}} \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{T}X_n^\top\iota - \Lambda_n^\top\frac{1}{T}\mathbf{F}^\top\iota\right)^2 = \min_{\mathbf{\Lambda}}\frac{1}{N}\operatorname{trace}\left[\left(\frac{1}{T}\iota^\top\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right)\left(\frac{1}{T}\iota^\top\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right)^\top\right] \tag{88}$$

However, this objective function (88) does not identify a set of factors and loadings and the problem admits an infinite number of solutions. In fact, any $\mathbf{\Lambda}$ such that $\mathbf{X}^\top\iota$ belongs to the space spanned by the columns of $\mathbf{\Lambda}$ will be a solution. A natural step following for extension is by considering the combination of tow objective functions in the following way, which has been proposed in Lettu and Pelger (2020a)

$$\min_{\mathbf{\Lambda}}\left\{\frac{1}{NT}\operatorname{trace}\left[\left(\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right)^\top\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right] + (1+\gamma)\frac{1}{N}\operatorname{trace}\left[\left(\frac{1}{T}\iota^\top\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right)\left(\frac{1}{T}\iota^\top\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right)^\top\right]\right\} \tag{89}$$

$$\begin{aligned}
&= \min_{\mathbf{\Lambda}}\left\{\frac{1}{NT}\operatorname{trace}\left[\left(\mathbf{M}_{\mathbf{\Lambda}}\mathbf{X}^\top\mathbf{M}_1\right)\mathbf{M}_1\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right] + (1+\gamma)\frac{1}{N}\operatorname{trace}\left[\left(\frac{1}{T}\iota^\top\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right)\left(\frac{1}{T}\iota^\top\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right)^\top\right]\right\} \\
&= \min_{\mathbf{\Lambda}}\left\{\frac{1}{NT}\operatorname{trace}\left[\mathbf{M}_{\mathbf{\Lambda}}\mathbf{X}^\top\left(\mathbf{I}_T - \frac{1}{T}\iota\iota^\top\right)\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right] + (1+\gamma)\frac{1}{N}\operatorname{trace}\left[\left(\frac{1}{T}\mathbf{M}_{\mathbf{\Lambda}}\mathbf{X}^\top\iota\right)\left(\frac{1}{T}\iota^\top\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right)\right]\right\} \\
&= \min_{\mathbf{\Lambda}}\left\{\frac{1}{NT}\operatorname{trace}\left[\mathbf{M}_{\mathbf{\Lambda}}\mathbf{X}^\top\left(\mathbf{I}_T - \frac{1}{T}\iota\iota^\top\right)\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right] + (1+\gamma)\frac{1}{NT}\operatorname{trace}\left[\mathbf{M}_{\mathbf{\Lambda}}\mathbf{X}^\top\frac{1}{T}\iota\iota^\top\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right]\right\} \\
&= \min_{\mathbf{\Lambda}}\frac{1}{NT}\operatorname{trace}\left[\mathbf{M}_{\mathbf{\Lambda}}\mathbf{X}^\top\left(\mathbf{I}_T + \frac{\gamma}{T}\iota\iota^\top\right)\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right]
\end{aligned} \tag{90}$$

The implication from (90) is that the objective function is minimized by choosing $\mathbf{\Lambda}$ as the eigenvectors of the largest K eigenvalues of $\mathbf{X}^\top\left(\mathbf{I}_T + \frac{\gamma}{T}\iota\iota^\top\right)\mathbf{X}$.

Remark 6.1 To see how this modified PCA is connected with conventional PCA, we make the following discussion for comparison

	Conventional PCA	Modified PCA
Objective function	$\min_{\mathbf{\Lambda}}\frac{1}{NT}\operatorname{trace}\left[\mathbf{M}_{\mathbf{\Lambda}}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\mathbf{M}_{\mathbf{\Lambda}}\right]$	$\min_{\mathbf{\Lambda}}\frac{1}{NT}\operatorname{trace}\left[\mathbf{M}_{\mathbf{\Lambda}}\mathbf{X}^\top\left(\mathbf{I}_T + \frac{\gamma}{T}\iota\iota^\top\right)\mathbf{X}\mathbf{M}_{\mathbf{\Lambda}}\right]$

From this comparison we are able to see that both the modified PCA and conventional PCA share a similar functional form of objective function and once $\gamma = -1$, the objective function of modified PCA reduces to objective function of conventional PCA. Moreover, it is obvious that the application of PCA to $\mathbf{X}^\top\left(\mathbf{I}_T + \frac{\gamma}{T}\iota\iota^\top\right)\mathbf{X}$ is equivalent to the application of PCA to $\frac{1}{T}\mathbf{X}^\top\mathbf{X} + \gamma\overline{\mathbf{X}}\overline{\mathbf{X}}^\top$, where $\overline{\mathbf{X}} = \frac{1}{T}\mathbf{X}^\top\iota$.

[10] Actually (88) corresponds to a two-step optimization: for a fixed $\mathbf{\Lambda}$, $\mathbf{F}$ is chosen following the OLS-style first order condition, which implies that $\frac{1}{T}\mathbf{F}^\top\iota = \left(\mathbf{\Lambda}^\top\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}^\top\left(\frac{1}{T}\mathbf{X}^\top\iota\right)$, then substituting back $\mathbf{F}$ of this form yields the R.H.S. of (88).

7 Empirical Application

7.1 Data

Nowadays there are several benchmark datasets corresponding to the firm-level characteristics and returns collected at monthly-frequency, including the work done by Green et al. (2017), Gu et al. (2019), Kozak et al. (2020) and Chen and Zimmermann (2020a) (cited as CZ20). Among which CZ20 is in particular by far the most recent and comprehensive one that has successfully covered almost all the major documented anomalies in literature. [11]

The major reason why these firm-level characteristics are important is that essentially the discrepancy of stock returns at cross-sectional dimension is attributed to these documented characteristics and actually the expected return is a function of specific form of these characteristics. Conceptionally it is possible for us to regard each characteristic as one measure to distinguish different firms, and for dimension-reduction concern a specific factor structure summarizing the major source of cross-sectional variation is impose in literature. factor model or the factor-structure imposed in research is not simply confined to the discussion of stock market but applies broadly in other financial market like mutual funds (Warther, 1995; Barber, Huang, and Odean, 2016; Edelen and Warner, 2001; Berk and Green, 2004; Bergstresser and Poterba, 2004; Del Guercio and Tkac, 2002; Frazzini and Lamont, 2008; Ben-Rephael, Kandel, and Wohl, 2011; Ferson and Kim, 2012; Kogan and Papanikolau, 2013; Dou, Kogan, and Wu, 2020; Lou, 2012).

For decades after the factor framework proposed by Fama and French (1993) in finance literature, which is broadly acknowledged as the initial attempt to complement CAPM, there has a vast majority of researches with focus on searching new anomalies that lead to abnormal returns that cannot be explained benchmark model. However there is still ongoing debate about how the results claimed in academic research generate impact on the associated anomalies-based investment strategies and despite such a kind of research framework is overwhelming in empirical asset pricing literature, recently it has received widely known criticism either for the underlying methodologies or the credibility of the widely documented empirical results. The initially forwarded question about this issue is what emphasized in Cochrane (2017) that the presence of a vast collection of noisy and highly correlated predictors motivates the adoption of new methodologies instead of the conventional cross-sectional regressions and portfolio sorts. However, the prevalent methodology implemented in finance literature to uncover intervention (publication) effect is based on the direct comparison between the average returns of anomaly-based portfolios before and after publication, rarely is there literature discussing the counterfactual effect, i.e. the return if the anomaly had not been published. By contrast, counterfactual effect is of broad interest in Microeconometrics researches and it is intrinsically deserve more attention and recently there is an evolution accommodating this idea in finance literature like the work done in Pelger and Xiong (2020)

[11] We acknowledge the codes and data kindly shared by the authors and their efforts on constantly maintaining and updating the data. Both the codes and data are available at the authors' maintained website https://sites.google.com/site/chenandrewy/open-source-ap?authuser=0.

For the dataset constructed in CZ20, there are in total 210 portfolios based on different anomalies (characteristics). Specifically, each anomaly is based on a firm-specific variable (characteristic), e.g. the size and book-to-market ratio and then all the stocks traded on U.S. market are sorted into five quintile portfolios based on the corresponding firm-specific characteristic. Return associated with anomaly is Long-short portfolios that buy the highest quintile and sell the lowest quintile portfolio.

One of their claimed result for the simple case when there is only a single factor is as following

$$\sqrt{T}\left(\widetilde{\lambda}_i - \lambda_i\right) \xrightarrow{d} \begin{cases} \mathcal{N}\left(0, \frac{T}{T_0}\frac{\sigma_e^2}{\sigma_F^2} + 2\frac{T-T_0}{T_0}\right) & \text{for } i = 1, \ldots, N_0 \\ \mathcal{N}\left(0, \frac{\sigma_e^2}{\sigma_F^2}\right) & \text{for } i = N_0 + 1, \ldots, N \end{cases}$$

One of the main objective for Pelger and Xiong (2020) is to uncover the difference between common component of units after treatment adoption C_{it}^{treat} and the common component of the synthetic control C_{it}^{ctrl}, that is

$$\tau_{it} = C_{it}^{\text{treat}} - C_{it}^{\text{ctrl}} \tag{91}$$

Generally there are three kinds of patterns associated with data structure with missing values: (i) random missing pattern where whether the entry of data is observed or not does not depend on the entries or other observable covariates; (ii) All the data for treated panel are not observed simultaneously after a specific period; (iii) The periods after which the data for units collected in treated panel are not observed is staggered. This categorization is demonstrated visually as following

[Place Figure 5 about here]

For the inspiring insights on the practical empirical application, Let us first visualize the cumulative return associated with these long-short portfolios based sorting on anomalies (characteristics)

[Place Figure 6 about here]

This may not be that obvious, another direct measure would be calculating and comparing the average returns of different anomaly-based portfolios before and after the publication of specific anomaly.

[Place Figure 7 about here]

Figures and Tables

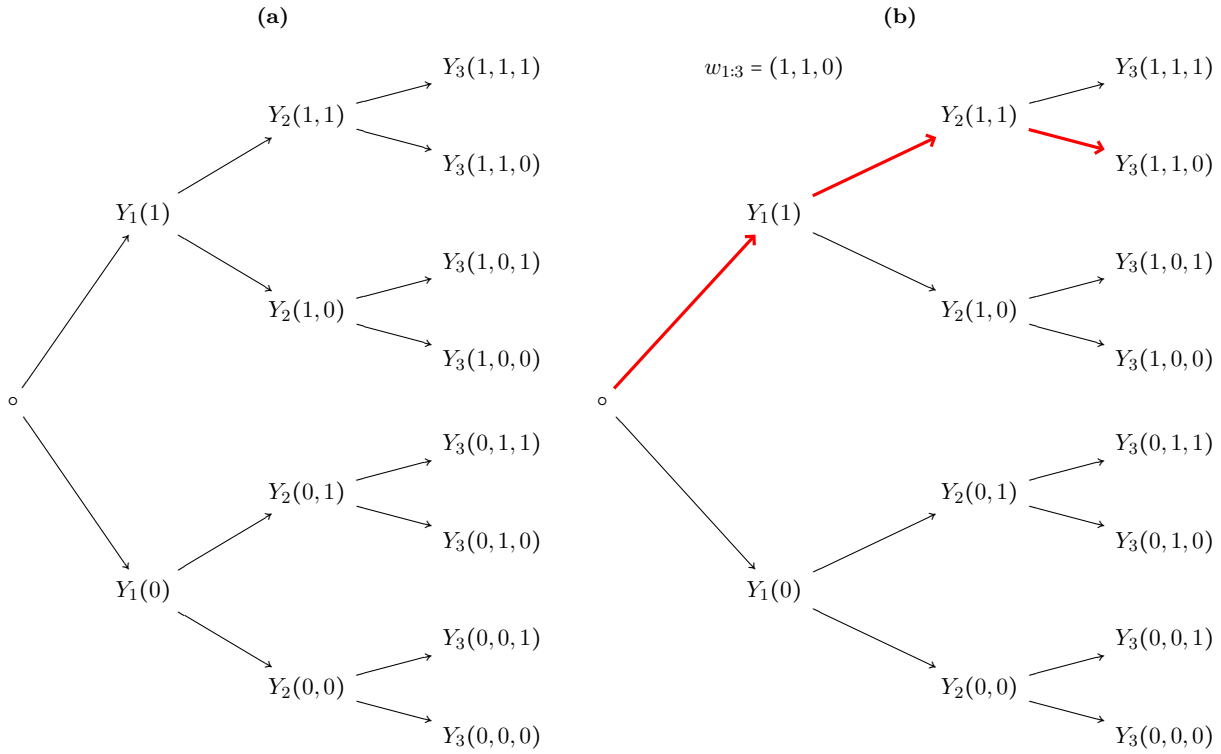


Figure 1. (a) demonstrates all the potential outcomes from intervention for $T = 3$. (b) demonstrates the observed outcome path (possibly not affected by intervention at some period) $Y_{1:3}(w_{1:3})$, indicated by the thick red line.

Table 1. Summary of basic concepts associated variational bayes method

Name	Definition
Variational log likelihood	$M_n(\boldsymbol{\theta};\mathbf{x}) := \sup_{q(\mathbf{z})\in\mathcal{Q}^n} F(q,\boldsymbol{\theta}) = \sup_{q(\mathbf{z})\in\mathcal{Q}^n} \int q(\mathbf{z}) \ln\left(\frac{p(\mathbf{x},\mathbf{z}\vert\boldsymbol{\theta})}{q(\mathbf{z})}\right) d\mathbf{z}$
Variational frequentist estimate (VFE)	$\operatorname{argmax}_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta},\mathbf{x})$
VB ideal	$\pi^*(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\boldsymbol{\theta}) M_n(\boldsymbol{\theta};\mathbf{x})}{\int p(\boldsymbol{\theta}) M_n(\boldsymbol{\theta};\mathbf{x}) \mathrm{d}\boldsymbol{\theta}}$
Evidence Lower Bound (ELBO)	$\mathrm{ELBO}(q(\boldsymbol{\theta},\mathbf{z})) := \iint q(\boldsymbol{\theta}) q(\mathbf{z}) \ln \frac{p(\mathbf{x},\mathbf{z},\boldsymbol{\theta})}{q(\boldsymbol{\theta}) q(\mathbf{z})} \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{z}$
VB posterior	$q^*(\boldsymbol{\theta}) := \arg\max_{q(\boldsymbol{\theta})\in\mathcal{Q}^d} \sup_{q(\mathbf{z})\in\mathcal{Q}^n} \mathrm{ELBO}(q(\boldsymbol{\theta},\mathbf{z}))$
VB estimate (VBE)	$\hat{\boldsymbol{\theta}}_n^* = \int \boldsymbol{\theta} \cdot q^*(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$

Note: (i) $M_n(\boldsymbol{\theta};\mathbf{x})$ is introduced with subscript n to emphasize the dependency on sample size. (ii) $\mathcal{Q}^n$ is introduced with superscript n to emphasize that the optimization is implemented over the family of distributions of local latent variables and all the alternatives contained in the family of these distributions are factorizable.

Figure 2. In this figure we demonstrate how the posterior probabilities assigned to the first 20 covariates evolve over time. The length of time-series observations is $T = 200$ and the total number of covariates is $p = 200$.

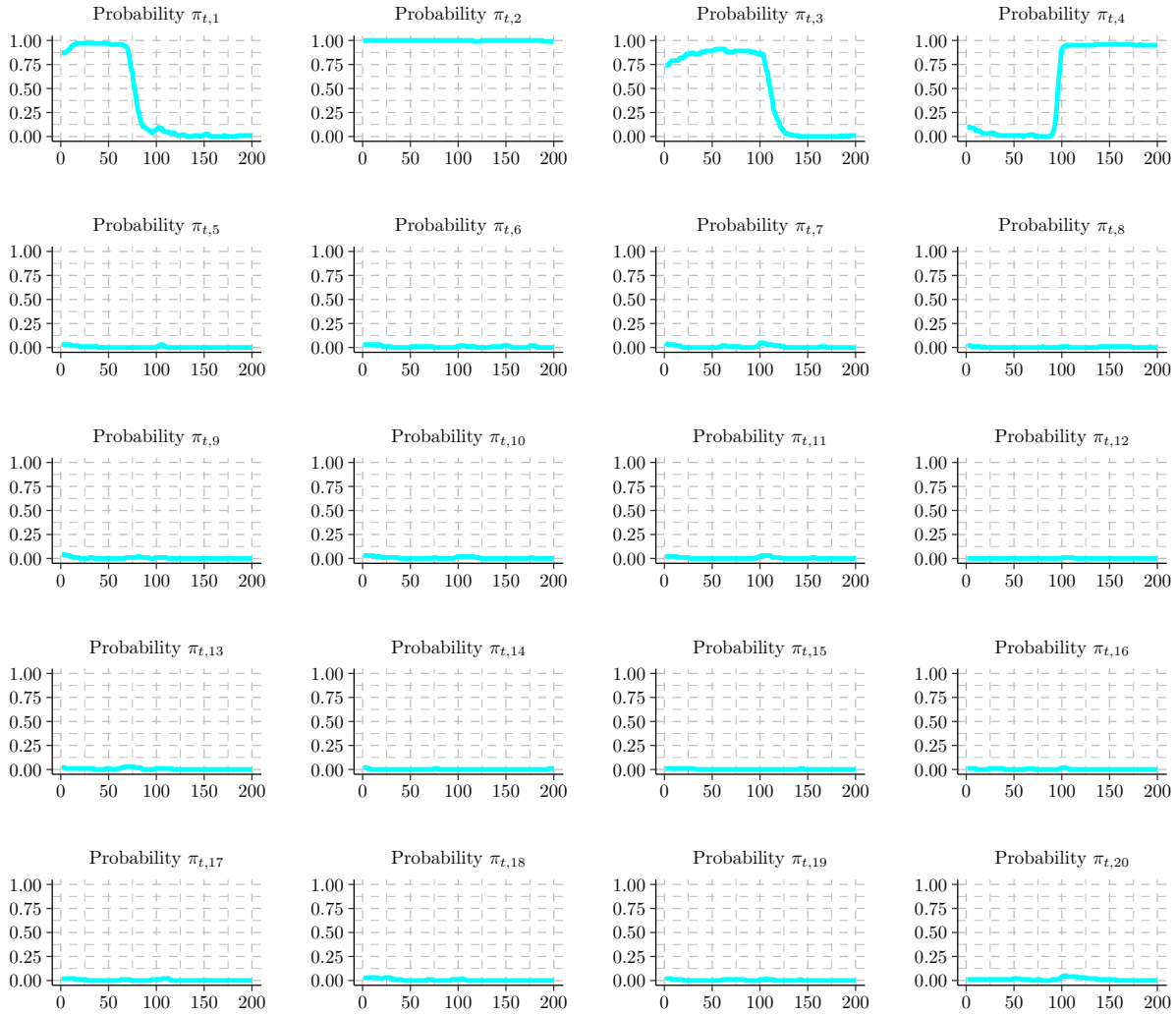


Figure 3. In this figure, we demonstrate how coefficients are estimated (medians over 100 Monte Carlo iterations) over time and the corresponding true values used for generating data as well. Cyan solid lines refer to coefficients fitted from the variational bayes algorithm and black long-dashed lines refer to true values used to generate data. The sub-figure contained in each panel represents how the related values evolve over time. And grey areas refer to the 84% and 16% quantiles (over 100 Monte Carlo iterations) respectively. The length of time-series observations is $T = 200$ and the total number of covariates is $p = 200$.

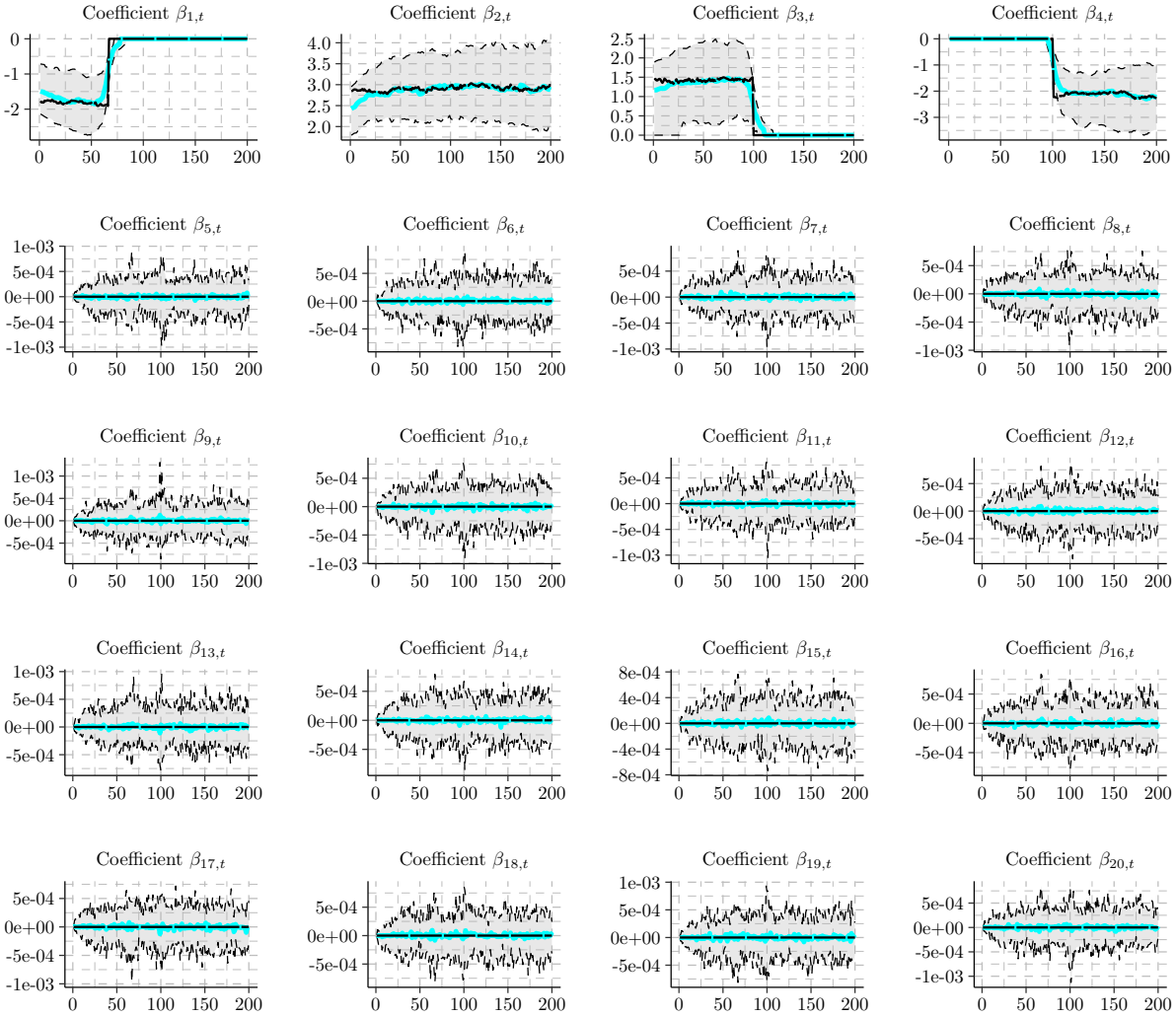


Figure 4. In this figure we demonstrate how the posterior probabilities assigned to the first 20 covariates evolve over time. The length of time-series observations is $T = 200$ and the total number of covariates is $p = 200$.

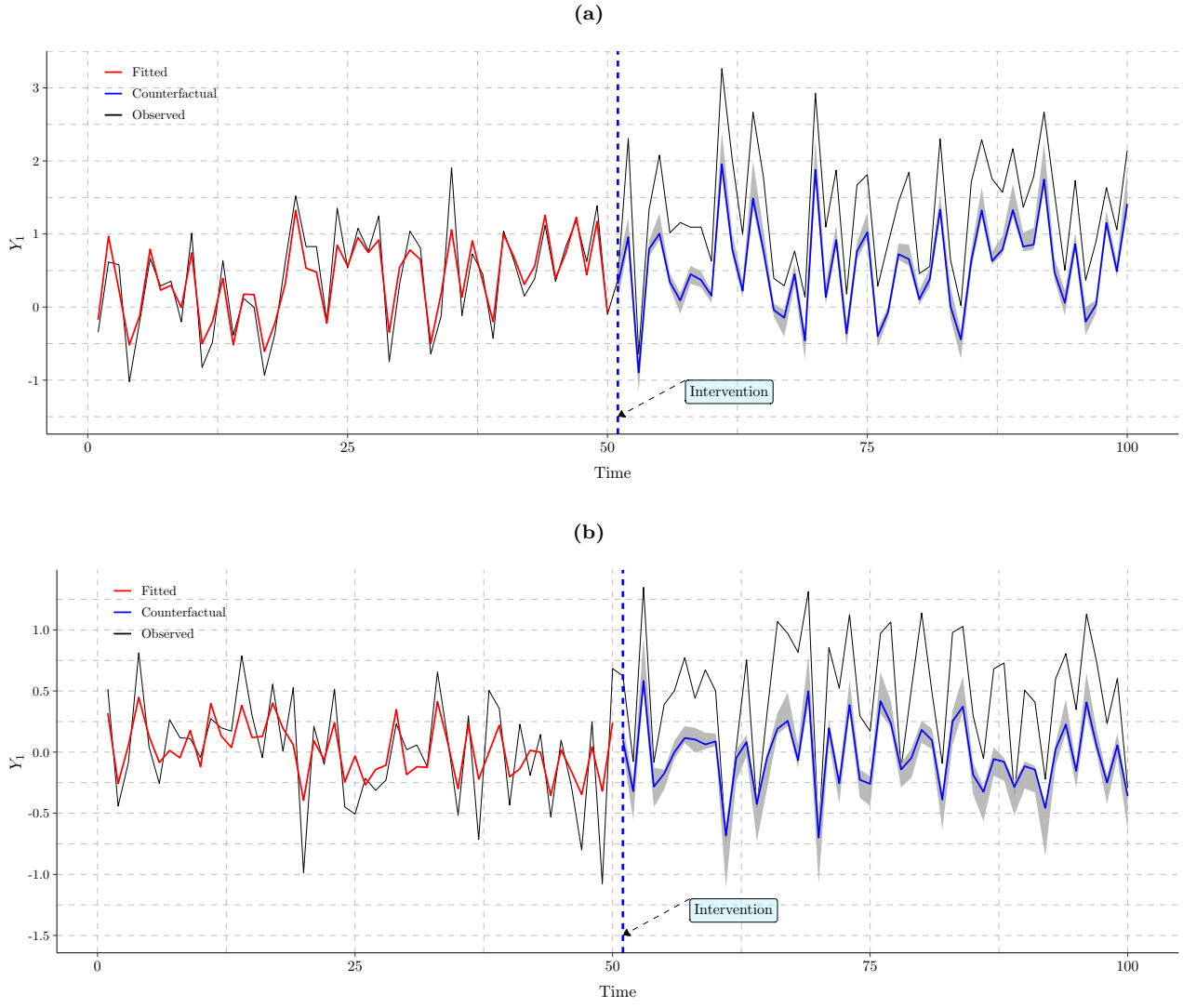


Figure 5. This is figure is for demonstrating visually the ideas of three data missing patterns mentioned in the context. Each panel refers to a specific missing pattern, i.e., (a), (b) and (c) refers to random missing pattern, simultaneously missing pattern and staggered missing pattern respectively. To make it as a comparable discussion in panel causal inference setting, horizontal axis refers to the time while the vertical axis refers to the units. Blue shaded entries indicate the missing observations.

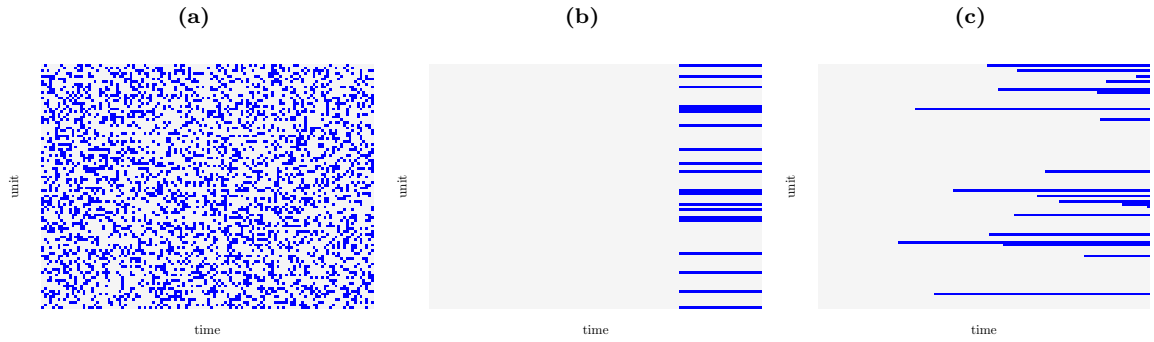
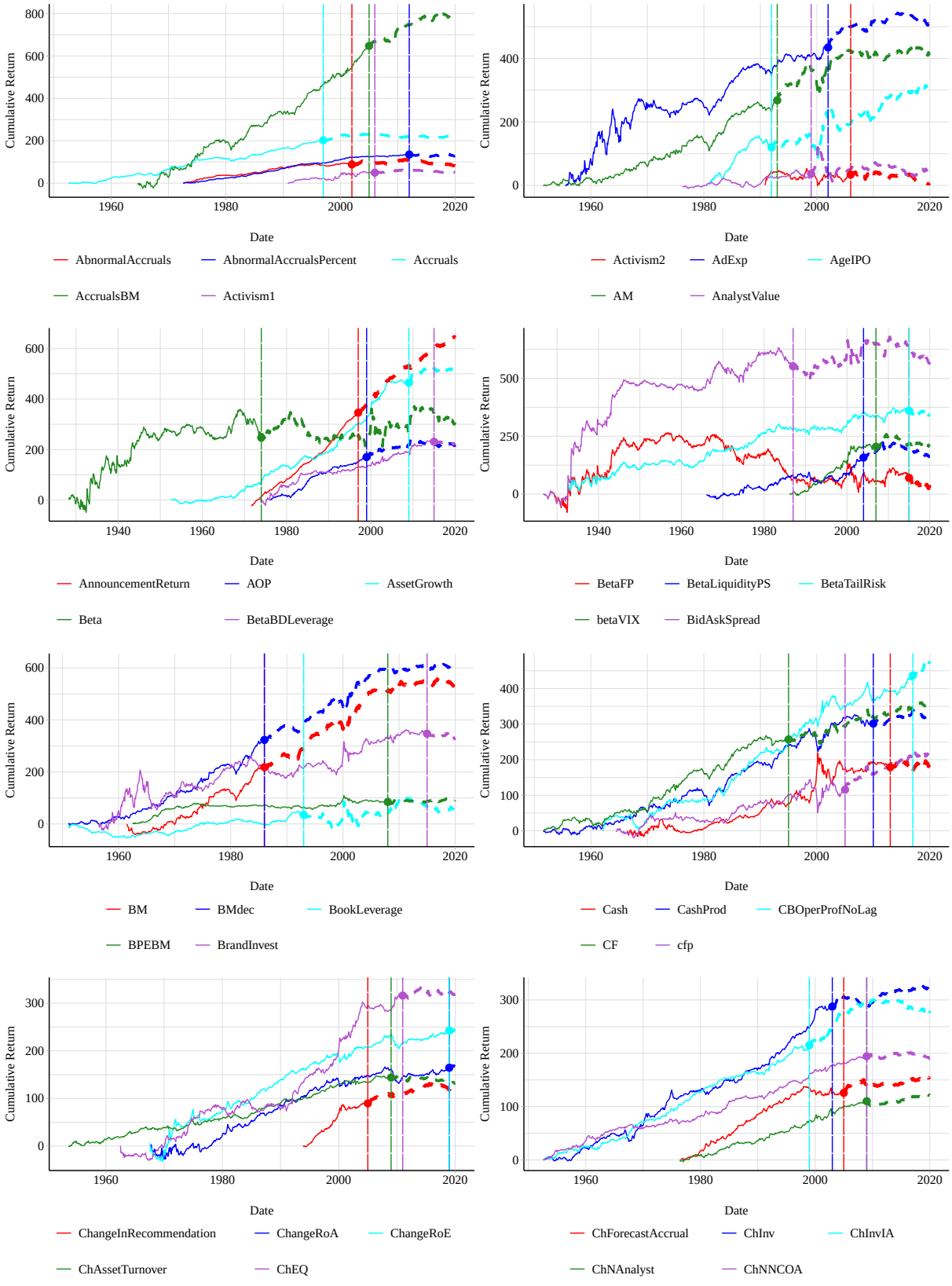
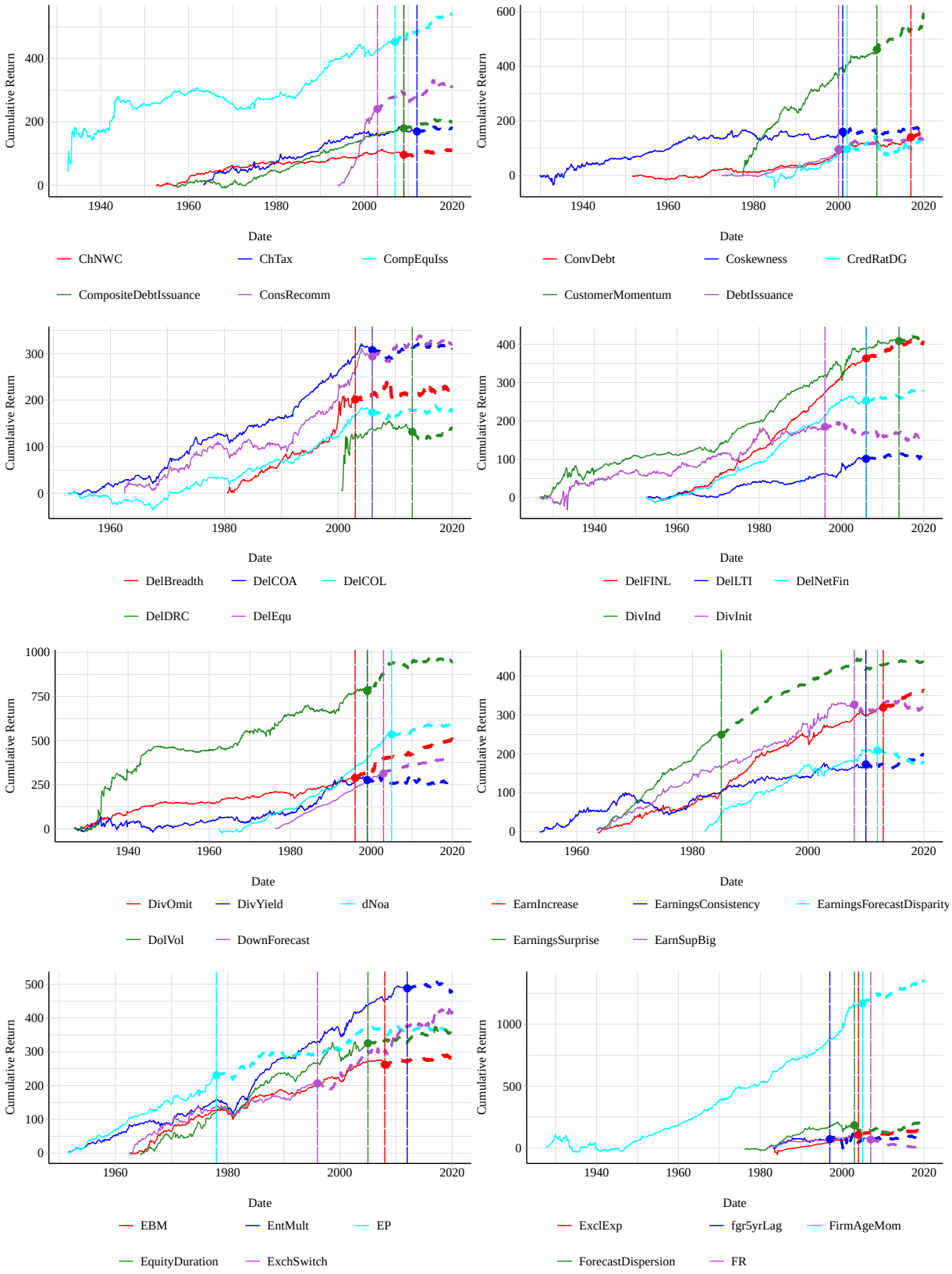
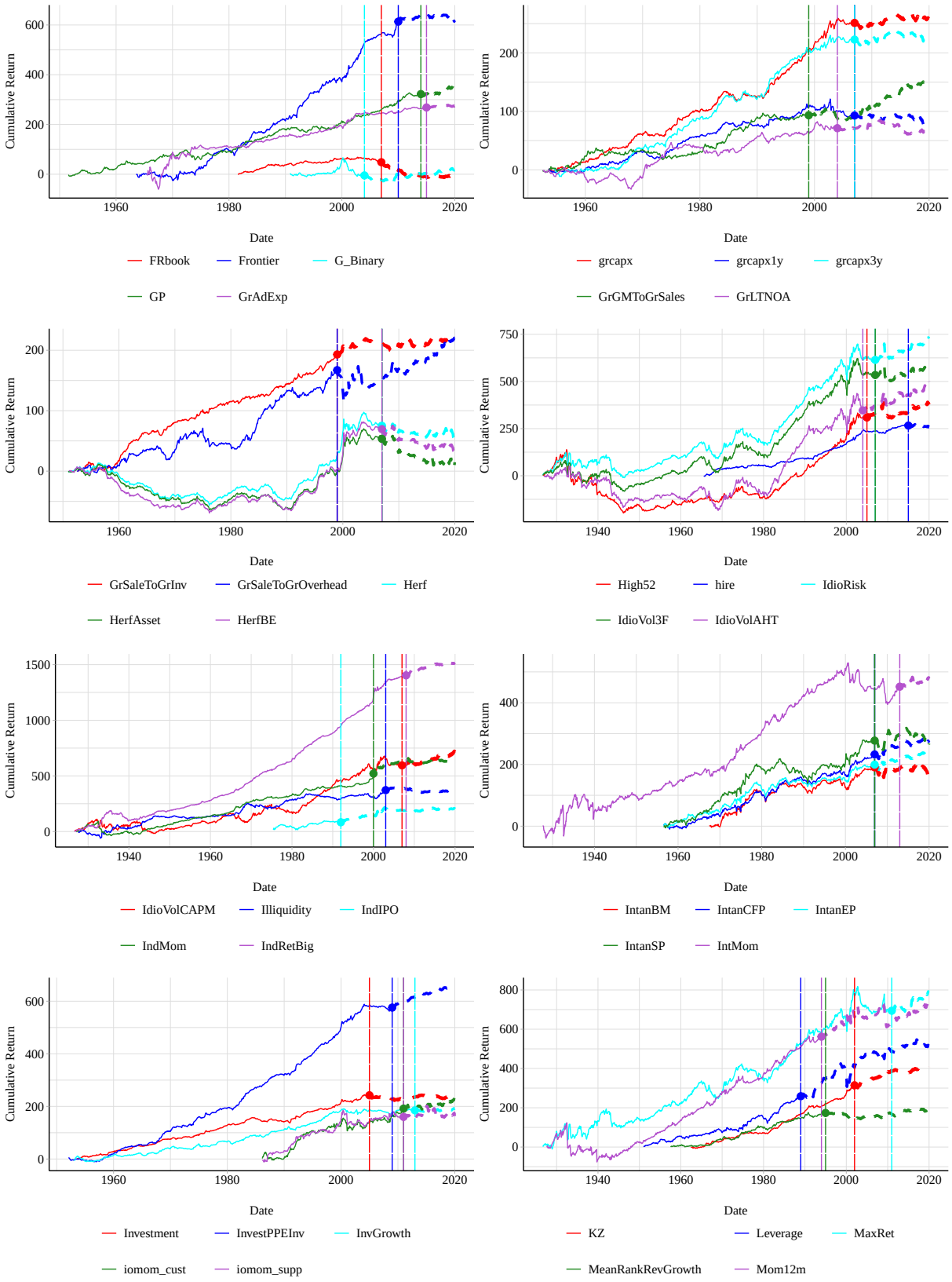
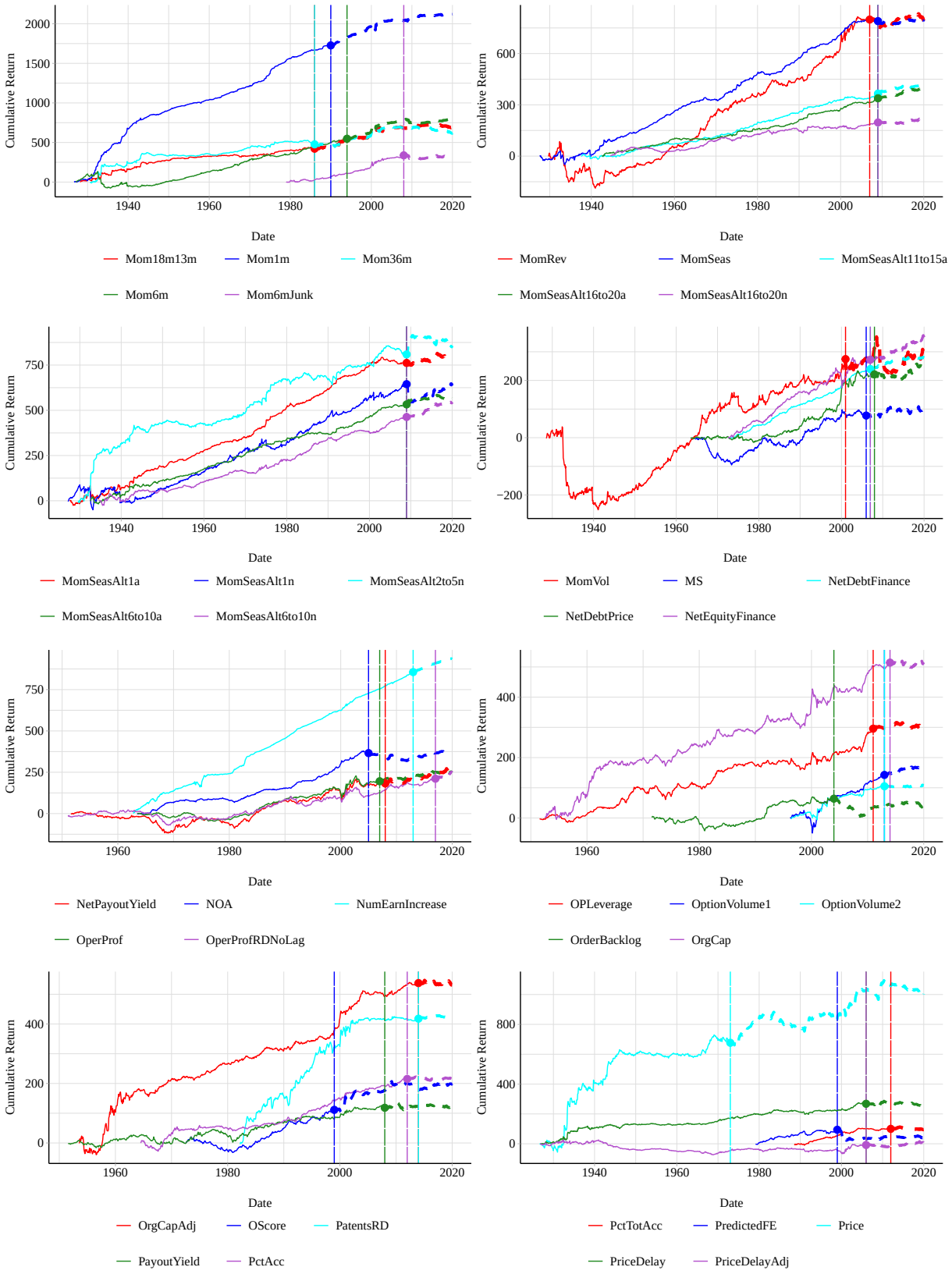


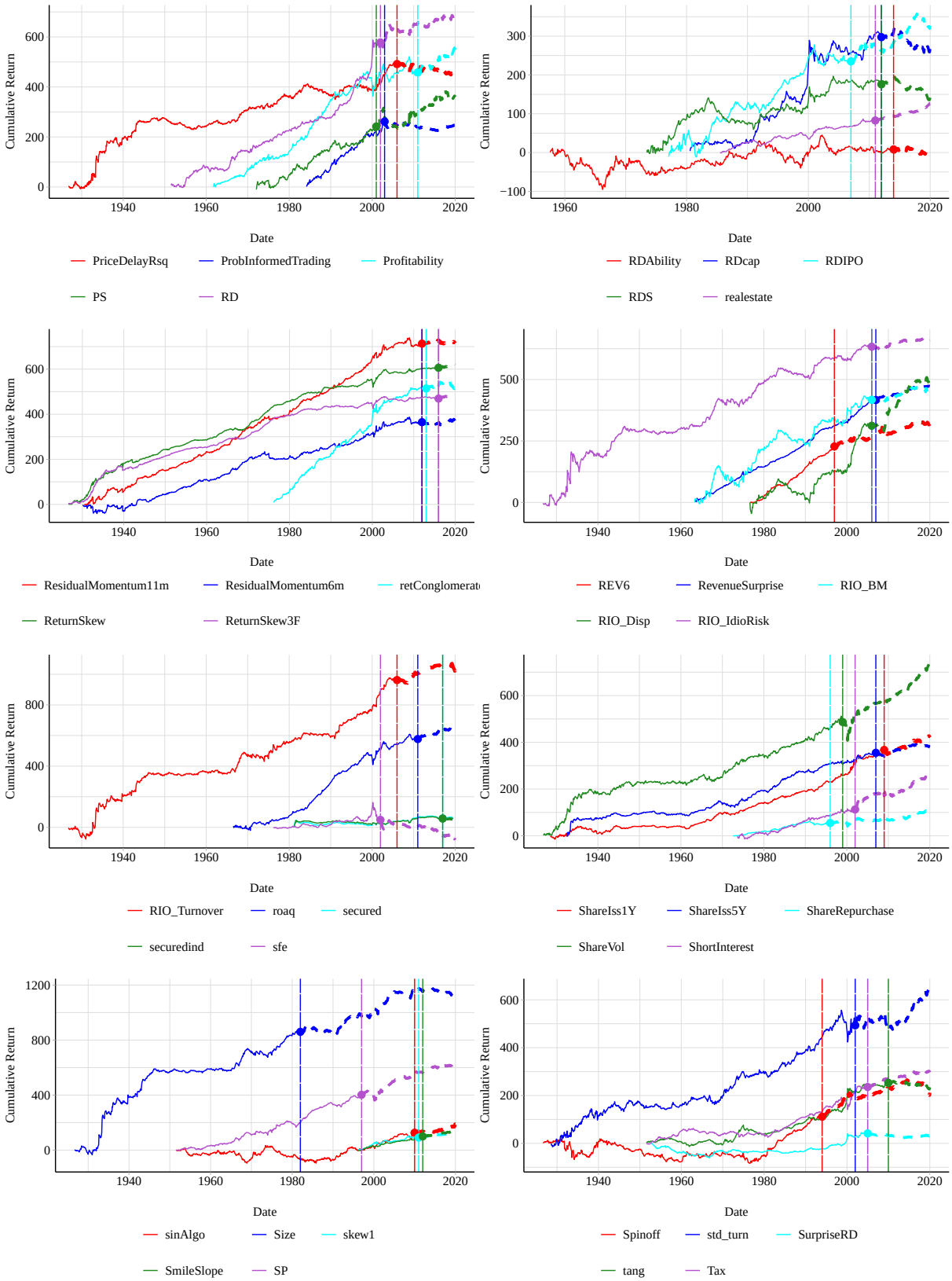
Figure 6

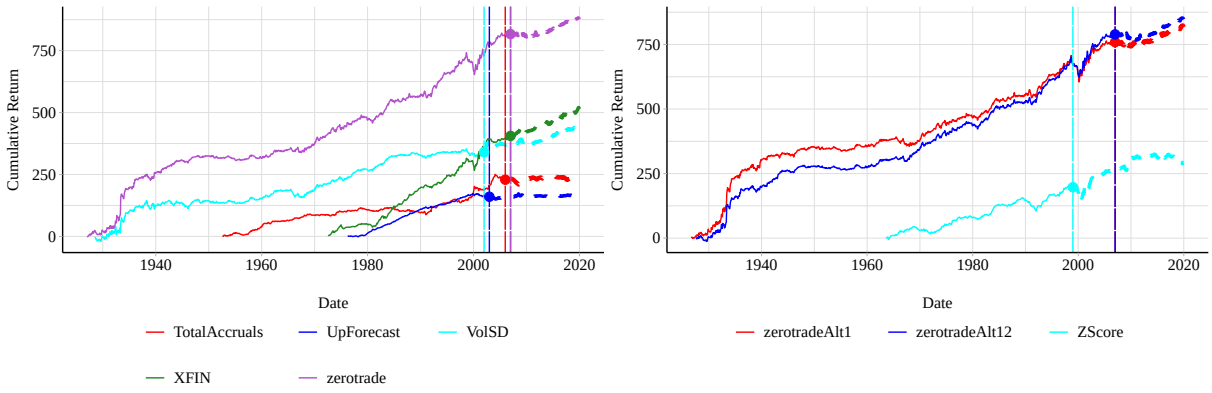

Cumulative Return
Date
AbnormalAccruals
AbnormalAccrualsPercent
Accruals
AccrualsBM
Activism1
Activism2
AdExp
AgeIPO
AM
AnalystValue
AnnouncementReturn
AOP
AssetGrowth
Beta
BetaBDLeverage
BetaFP
BetaLiquidityPS
BetaTailRisk
betaVIX
BidAskSpread
BM
BMdec
BookLeverage
BPEBM
BrandInvest
Cash
CashProd
CBOperProfNoLag
CF
cfp
ChangeInRecommendation
ChangeRoA
ChangeRoE
ChAssetTurnover
ChEQ
ChForecastAccrual
ChInv
ChInvIA
ChNAnalyst
ChNNCOA


Cumulative Return
Date
ChNWC
ChTax
CompEquIss
CompositeDebtIssuance
ConsRecomm
ConvDebt
Coskewness
CredRatDG
CustomerMomentum
DebtIssuance
DelBreadth
DelCOA
DelCOL
DelDRC
DelEqu
DelFINL
DelLTI
DelNetFin
DivInd
DivInit
DivOmit
DivYield
dNoa
DolVol
DownForecast
EarnIncrease
EarningsConsistency
EarningsForecastDisparity
EarningsSurprise
EarnSupBig
EBM
EntMult
EP
EquityDuration
ExchSwitch
ExclExp
fgr5yrLag
FirmAgeMom
ForecastDispersion
FR


Cumulative Return
Date
FRbook
Frontier
G_Binary
GP
GrAdExp
grcapx
grcapx1y
grcapx3y
GrGMToGrSales
GrLTNOA
GrSaleToGrInv
GrSaleToGrOverhead
Herf
HerfAsset
HerfBE
High52
hire
IdioRisk
IdioVol3F
IdioVolAHT
IdioVolCAPM
Illiquidity
IndIPO
IndMom
IndRetBig
IntanBM
IntanCFP
IntanEP
IntanSP
IntMom
Investment
InvestPPEInv
InvGrowth
iomom_cust
iomom_supp
KZ
Leverage
MaxRet
MeanRankRevGrowth
Mom12m

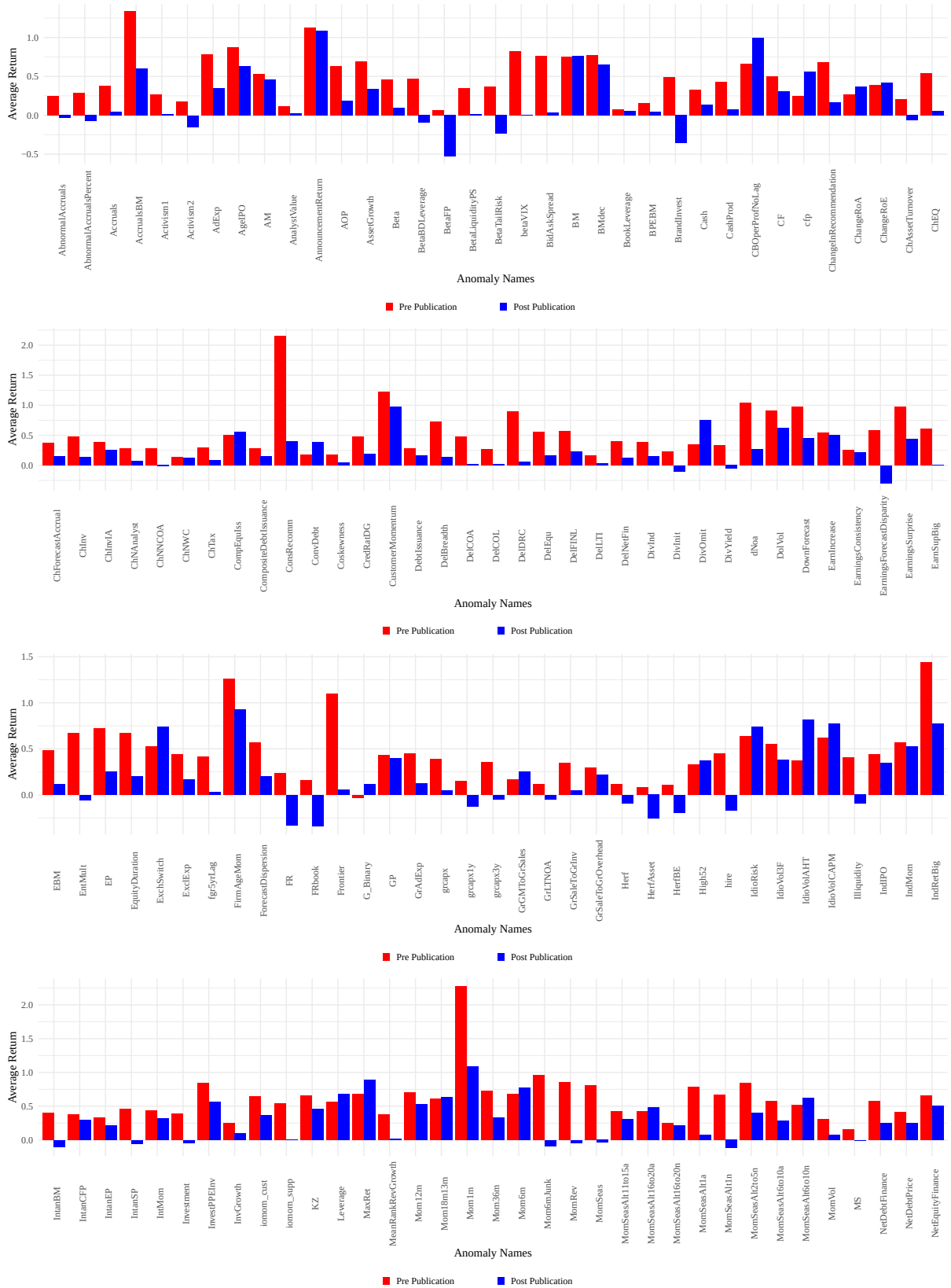

Cumulative Return
Date
Mom18m13m
Mom1m
Mom36m
Mom6m
Mom6mJunk
MomRev
MomSeas
MomSeasAlt11to15a
MomSeasAlt16to20a
MomSeasAlt16to20n
MomSeasAlt1a
MomSeasAlt1n
MomSeasAlt2to5n
MomSeasAlt6to10a
MomSeasAlt6to10n
MomVol
MS
NetDebtFinance
NetDebtPrice
NetEquityFinance
NetPayoutYield
NOA
NumEarnIncrease
OperProf
OperProfRDNoLag
OPLeverage
OptionVolume1
OptionVolume2
OrderBacklog
OrgCap
OrgCapAdj
OScore
PatentsRD
PayoutYield
PctAcc
PctTotAcc
PredictedFE
Price
PriceDelay
PriceDelayAdj

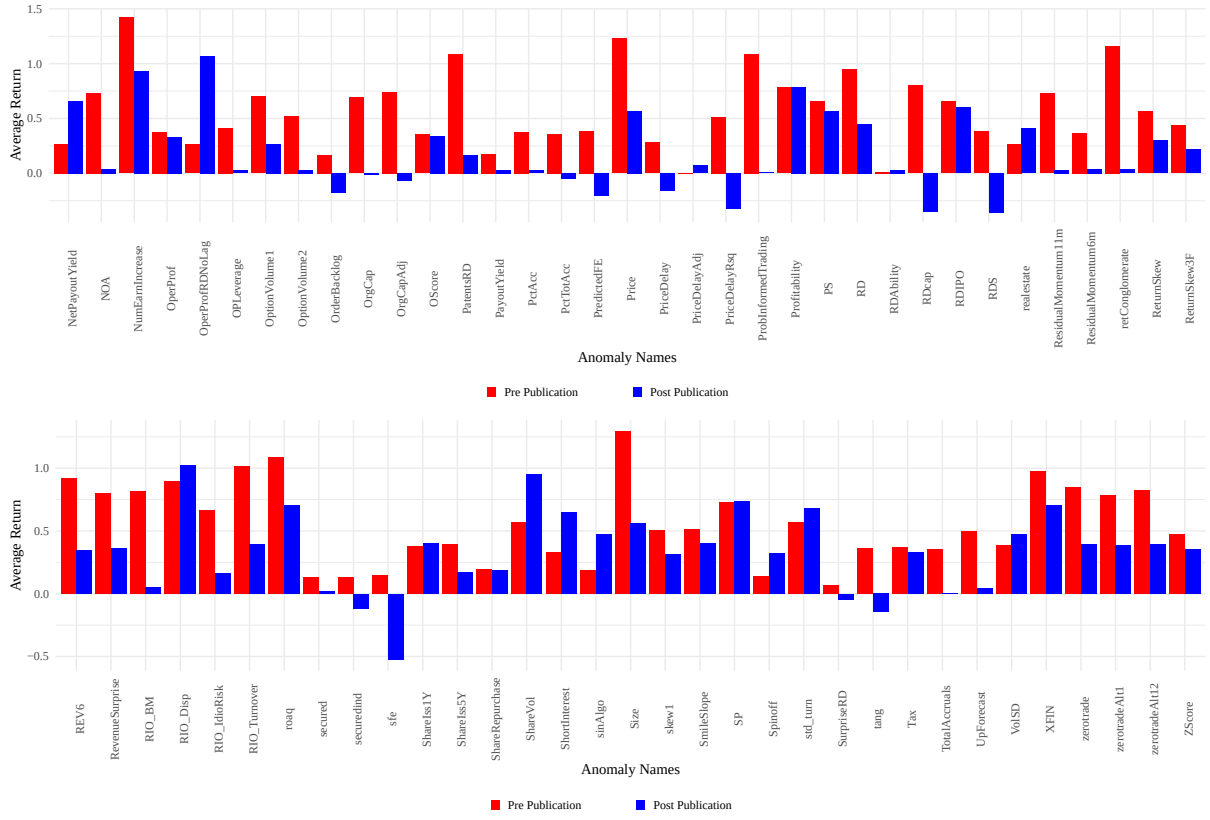

Cumulative Return
Date
PriceDelayRsq
ProbInformedTrading
Profitability
PS
RD
RDAbility
RDcap
RDIPO
RDS
realestate
ResidualMomentum11m
ResidualMomentum6m
retConglomerat
ReturnSkew
ReturnSkew3F
REV6
RevenueSurprise
RIO_BM
RIO_Disp
RIO_IdioRisk
RIO_Turnover
roaq
secured
securedind
sfe
ShareIss1Y
ShareIss5Y
ShareRepurchase
ShareVol
ShortInterest
sinAlgo
Size
skew1
SmileSlope
SP
Spinoff
std_turn
SurpriseRD
tang
Tax



Note: In each panel of the above figure, there are 5 plots of cumulative returns corresponding to different long-short portfolios based on different anomalies (characteristics) respectively. The intersection of vertical line and cumulative return plot indicates timing when the corresponding anomaly was published. Consequently, each cumulative return plot is separated into two parts: the solid part refers to the period before publication of specific anomaly while the dashed part refers to the period after publication of specific anomaly.

Figure 7


Average Return
1.0
0.5
0.0
−0.5
AbnormalAccruals
AbnormalAccrualsPercent
Accruals
AccrualsBM
Activism1
Activism2
AdExp
AgePO
AM
AnalystValue
AnnouncementReturn
AOP
AssetGrowth
Beta
BetaBDLeverage
BetaFP
BetaLiquidityPS
BetaTailRisk
betaVIX
BidAskSpread
BM
BMdec
BookLeverage
BPEBM
BrandInvest
Cash
CashProd
CBOperProfNoLag
CF
cfp
ChangeInRecommendation
ChangeRoA
ChangeRoE
ChAssetTurnover
ChEQ
Anomaly Names
Pre Publication
Post Publication
Average Return
2.0
1.5
1.0
0.5
0.0
ChForecastAccrual
ChInv
ChInvIA
ChNAnalyst
ChNNCOA
ChNWC
ChTax
CompEquIss
CompositeDebtIssuance
ConsRecomm
ConvDebt
Coskewness
CredRatDG
CustomerMomentum
DebtIssuance
DelBreadth
DelCOA
DelCOL
DelDRC
DelEqu
DelFINL
DelLTI
DelNetFin
DivInd
DivInit
DivOmit
DivYield
dNoa
DolVol
DownForecast
EarnIncrease
EarningsConsistency
EarningsForecastDisparity
EarningsSurprise
EarnSupBig
Anomaly Names
Pre Publication
Post Publication
Average Return
1.5
1.0
0.5
0.0
EBM
EntMult
EP
EquityDuration
ExchSwitch
ExclExp
fgr5yrLag
FirmAgeMom
ForecastDispersion
FR
FRbook
Frontier
G_Binary
GP
GrAdExp
grcapx
grcapx1y
grcapx3y
GrGMToGrSales
GrLTNOA
GrSaleToGrInv
GrSaleToGrOverhead
Herf
HerfAsset
HerfBE
High52
hire
IdioRisk
IdioVol3F
IdioVolAHT
IdioVolCAPM
Illiquidity
IndIPO
IndMom
IndRetBig
Anomaly Names
Pre Publication
Post Publication
Average Return
2.0
1.5
1.0
0.5
0.0
IntanBM
IntanCFP
IntanEP
IntanSP
IntMom
Investment
InvestPPEInv
InvGrowth
iomom_cust
iomom_supp
KZ
Leverage
MaxRet
MeanRankRevGrowth
Mom12m
Mom18m13m
Mom1m
Mom36m
Mom6m
Mom6mJunk
MomRev
MomSeas
MomSeasAlt11to15a
MomSeasAlt16to20a
MomSeasAlt16to20n
MomSeasAlt1a
MomSeasAlt1n
MomSeasAlt2to5n
MomSeasAlt6to10a
MomSeasAlt6to10n
MomVol
MS
NetDebtFinance
NetDebtPrice
NetEquityFinance
Anomaly Names
Pre Publication
Post Publication



Note: In each panel of the above figure, average returns of different anomaly-based long-short portfolios are demonstrated separately. Specifically, the red bar plot indicates the the average return of a specific anomaly-based long-short portfolio before that anomaly was published while the blue bar indicates the average return of long-short portfolio constructed from that anomaly after publication.

References

Abadie, A. (2020): "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," Working paper, Massachusetts Institute of Technology, forthcoming in *Journal of Economic Literature*. [Cited on pages 2, 5, and 6.]

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. [Cited on page 2.]

Abadie, A. and J. Gardeazabal (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132. [Cited on page 2.]

Alquier, P. and J. Ridgway (2020): "Concentration of Tempered Posteriors and of Their Variational Approximations," *The Annals of Statistics*, 48, 1475–1497. [Cited on pages 1, 20, 21, and 27.]

Armagan, A., D. B. Dunson, and M. Clyde (2011): "Generalized Beta Mixtures of Gaussians," *Advances in neural information processing systems*, 523–531. [Cited on page 33.]

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2020): "Robust Empirical Bayes Confidence Intervals," Working paper, Yale University and Princeton University. [Cited on page 1.]

Athey, S., M. Bayati, N. Doudchenkou, G. Imbens, and K. Khosravi (2020): "Matrix Completion Methods for Causal Panel Data Models," Working paper, arXiv:1710.10251. [Cited on page 2.]

Barber, B. M., X. Huang, and T. Odean (2016): 29, 2600–2642. [Cited on page 42.]

Barndorff-Nielsen, O., J. Kent, and M. Sørensen (1982): "Normal Variance-Mean Mixtures and z Distributions," *International Statistical Institute (ISI)*, 50, 145–159. [Cited on page 34.]

Ben-Rephael, A., S. Kandel, and A. Wohl (2011): "The Price Pressure of Aggregate Mutual Fund Flows," *The Journal of Financial and Quantiative Analysis*, 46, 585–603. [Cited on page 42.]

Bergstresser, D. and J. Poterba (2004): "Do after-tax returns affect mutual fund inflows," *Journal of Financial Economics*, 63, 381–414. [Cited on page 42.]

Berk, J. B. and R. C. Green (2004): "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, 112, 1269–1295. [Cited on page 42.]

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017): "Variational Inference: A Review for Statisticians," ArXiv:1601.00670. [Cited on page 1.]

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [Cited on page 8.]

Bojinov, I. and N. Shepard (2019): "Time Series Experiments and Causal Estimands: Exact Randomization Tests and Trading," *Journal of the American Statistical Association*, 114, 1665–1682. [Cited on page 7.]

Braun, M. and J. McAuliffe (2010): "Variational Inference for Large-scale Models of Discrete Choice," *Journal of American Statistical Association*, 105, 324–334. [Cited on page 1.]

Carlin, B. P. and T. A. Louis (2000): *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall/CRC, New York, NY, 2nd edition. [Cited on page 36.]

Carolan, C. A. (2002): "The Least Concave Majorant of the Empirical Distribution Function," *The Canadian Journal of Statistics*, 30, 317–328. [Cited on page 37.]

Carvalho, C. V., R. Masini, and M. C. Medeiros (2018): *Journal of Econometrics*, 207, 352–380. [Cited on page 2.]

Catoni, O. (2004): *Statistical Learning Theory and Stochastic Optimization.* Saint- Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer. [Cited on page 20.]

——— (2007): *PAC-Bayesian supervised classification: the thermodynamics of statistical learning.* Institute of Mathematical Statistics Lecture Notes— Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH. [Cited on page 20.]

Chen, A. Y. and T. Zimmermann (2020a): "Open Source Cross-Sectional Asset Pricing," Working paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3604626. [Cited on pages 42 and 43.]

——— (2020b): "Publication Bias and the Cross-Section of Stock Returns," 10, 249–289. [Cited on page 2.]

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2020a): "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," Working paper, arXiv:1712.09089. [Cited on page 2.]

——— (2020b): "Practical and Robust t-test based Inference for Synthetic Control and Related Methods," Working paper, arXiv:1812.10820. [Cited on page 2.]

Chipman, H. A., E. I. George, and R. E. McCulloch (1998): "Bayesian CART Model Search," *Journal of the American Statistical Association*, 93, 935–948. [Cited on page 5.]

——— (2010): "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [Cited on page 5.]

Cochrane, J. H. (2017): "Presidential Address: Discount rates," *Journal of Finance*, 66, 1047–1108. [Cited on page 42.]

Cooley, T. F. and E. C. Prescott (1976): "Estimation in the Presence of Stochastic Parameter Variation," *Econometrica*, 44, 167–184. [Cited on page 1.]

Del Guercio, D. and P. A. Tkac (2002): "Star Power: The Effect of Morningstar Ratings on Mutual Fund Flow," *The Journal of Financial and Quantiative Analysis*, 43, 907–936. [Cited on page 42.]

Dempster, A. P., N. M. Laird, and B. Rubin (1977): "Maximum Likelihood from Imcomplete Data via the EM Algorithm," *Journal of Royal Statistical Society. Series B (Methodological)*, 39, 1–38. [Cited on pages 10 and 11.]

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998): "A Bayesian CART Algorithm," *Biometrica*, 85, 363–377. [Cited on page 5.]

Dou, W. W., L. Kogan, and W. Wu (2020): "Common Fund Flows: Flow Hedging and Factor Pricing," Working paper, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3543675. [Cited on page 42.]

Doudchenko, N. and G. W. Imbens (2017): "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Working paper, Stanford University. [Cited on page 2.]

Eddelbuettel, D. (2013): *Seamless R and C++ Integration with Rcpp*. Use R! Series. Springer. [Cited on page 29.]

Edelen, R. M. and J. B. Warner (2001): "Aggregate Price Effects of Institutional Trading: A Study of Mutual Fund Flow and Market Returns," *Journal of Financial Economics*, 59, 195–220. [Cited on page 42.]

Engle, R., S. Giglio, H. Lee, B. Kelly, and J. Stroebel (2020): "Hedging Climate Change News," *Review of Financial Studies*, 33, 1184–1216. [Cited on page 24.]

Fama, E. F. and K. R. French (1993): "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33, 3 – 56. [Cited on pages 10 and 42.]

——— (1996): "Multifactor Explanations of Asset Pricing Anomalies," *The Journal of Finance*, 51, 55–84. [Cited on page 10.]

——— (2015): "A Five-factor Asset Pricing Model," *Journal of Financial Economics*, 116, 1 – 22. [Cited on page 10.]

Ferson, W. E. and M. S. Kim (2012): "The factor structure of mutual fund flows," *International Journal of Portfolio Analysis and Management*, 1, 112–143. [Cited on page 42.]

Frazzini, A. and O. A. Lamont (2008): "Dumb money: Mutual fund flows and the cross-section of stock returns," *Journal of Financial Economics*, 88, 299–322. [Cited on page 42.]

Gentzkow, M., B. Kelly, and M. Tady (2019): "Text as Data," *Journal of Economic Literature*, 57, 535–574. [Cited on page 23.]

George, E. I. and R. E. McCulloch (1993): "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [Cited on pages 15 and 28.]

——— (1997): "Approaches for Bayesian Variable Selection," *Journal of the American Statistical Association*, 7, 339–373. [Cited on pages 15, 16, and 17.]

Green, J., J. R. M. Hand, and X. F. Zhang (2017): "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *The Review of Financial Studies*, 30, 4389–4436. [Cited on page 42.]

Griffiths, T. L. and M. Steyvers (2004): "Finding Scientific Topics," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 5228–5235. [Cited on page 10.]

Gu, S., B. Kelly, and D. Xiu (2019): "Empirical Asset Pricing via Maching Learning," Working paper, forthcoming in the *The Review of Financial Studies*. [Cited on page 42.]

Guo, F., X. Wang, K. Fan, T. Broderick, and D. B. Dunson (2016): "Boosting Variational Inference," ArXiv:1611.05559. [Cited on page 1.]

Hansen, B. E. (2016): "Efficient Shrinkage in Parametric Models," *Journal of Econometrics*, 190, 115–132. [Cited on page 1.]

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999): "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [Cited on page 1.]

Kogan, L. and D. Papanikolau (2013): "Firm Characteristics and Stock Returns: The Role of Investment-Specific Shocks," *The Review of Financial Studies*, 26, 2718–2759. [Cited on page 42.]

Koop, G. and D. Korobilis (2020): "Bayesian Dynamic Variable Selection in High Dimensions," R&R in *Journal of Econometrics*. [Cited on pages 5, 27, and 29.]

Kowal, D. R., D. S. Matteson, and D. Ruppert (2019): "Dynamic Shrinkage Processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 781–804. [Cited on pages 27 and 31.]

Kozak, S., S. Nagel, and S. Santosh (2020): "Shrinking the Cross Section," *Journal of Financial Economics*, 135, 271–292. [Cited on page 42.]

Lettu, M. and M. Pelger (2020a): "Estimating Latent Asset-Pricing Factors," Forthcoming in *The Journal of Finance*. [Cited on pages 38 and 41.]

——— (2020b): "Factors that Fit the Time-Series and Cross-Section of Stock Returns," *Review of Financial Studies*, 33, 2274–2325. [Cited on page 38.]

Lou, D. (2012): "A flow-based explanation for return predictability," *The Review of Financial Studies*, 25, 3457–3489. [Cited on page 42.]

Mclachlan, G. J. and T. Krishnan (2008): *The EM Algorithms and Extensions.* New Jersey: John Wiley & Sons. [Cited on page 10.]

Mclean, R. D. and J. Pontiff (2016): "Does Academic Research Destroy Stock Return Predictability," *Journal of Finance*, 71. [Cited on page 2.]

Morris, C. N. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–55. [Cited on pages 1 and 36.]

Pelger, M. and R. Xiong (2020): "Large Dimensional Latent Factor Modeling with Missing Observations and Applications to Causal Inference," Working paper, https://mpelger.people.stanford.edu/research. [Cited on pages 2, 42, and 43.]

Phan, X., L. Nguyen, and S. Horiguchi (2008): "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections," In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)*, 91–100, Beijing, China. [Cited on page 10.]

Polson, N. G., J. G. Scott, and J. Windle (2013): "Bayesian Inference for Logistic Models using Pólya-Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349. [Cited on pages 34 and 36.]

Robbins, H. and S. Monro (1951): "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, 22, 400–407. [Cited on page 24.]

Ročková, V. (2019): "On Semi-parametric Bernstein-von Mises Theorems for BART," Working paper. [Cited on page 5.]

Ročková, V. and E. I. George (2014): "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846. [Cited on page 15.]

Ročková, V. and E. Saha (2019): "On Theory for BART," Working paper, 22^{nd} International Conference on Artificial Intelligence and Statistics. [Cited on page 5.]

Ročková, V. and S. van der Pas (2019): "Posterior Concentration for Bayesian Regression Trees and Forests," Manuscript, just accepted for *Annals of Statistics*. [Cited on page 5.]

Tran, M.-N., D. J. Nott, and R. Kohn (2017): "Variational Bayes With Intractable Likelihood," *Jounral of Computational and Graphical Statistics*, 26, 873–882. [Cited on page 29.]

Van Erven, T. and P. Harremoës (2014): "Rényi Divergence and Kullback-Leibler Divergence," *IEEE Transactions on Information Theory*, 60, 3797–3820. [Cited on page 19.]

Wainwright, M. J. and M. I. Jordan (2008): "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends in Machine Learning*, 1, 1–305. [Cited on page 10.]

Wang, Y. and D. M. Blei (2019): "Frequentist Consistency of Variational Bayes," *Journal of the American Statistical Association*, 114, 1147–1161. [Cited on pages 1 and 27.]

Warther, V. A. (1995): "Aggregate Mutual Fund Flows and Security Returns," *Journal of Financial Economics*, 39, 209–235. [Cited on page 42.]

Winn, J. and M. Bishop (2005): "Variational Message Passing," *Journal of Machine Learning Research*, 6, 661–694. [Cited on page 1.]

Zeiler, M. D. (2012): "ADADELTA: An Adaptive Learning Rate Method," ArXiv:1212.5701. [Cited on page 25.]