MSc Thesis Proposals

MSc: Statistics and Actuarial-Financial Mathematics

Konstantinos Bourazas
kbourazas@aegean.gr

Department of Statistics and Actuarial-Financial Mathematics

Research directions

May 14, 2026

1A. Detecting heterogeneity in multi-site replication studies[1]

- The **replication crisis** in empirical science has motivated multi-site replication projects, in which the same experiment is conducted simultaneously at several independent sites
- A primary question in such designs is whether the effect size is consistent across sites or whether substantial **heterogeneity** exists between them, and how many sites and subjects per site are needed to answer that question with confidence
- Frequentist power analysis conditions on a single fixed value of the parameter of interest. We instead develop a Bayesian framework that places a full distribution on the heterogeneity parameter and uses the **Bayes factor** as a measure of evidence
- A key advantage of the Bayes factor is that it can provide evidence *for* or *against* the presence of heterogeneity, whereas a non-significant frequentist test gives no support for the null

[1] Bourazas, K., Consonni, G., & Deldossi, L. (2024). Bayesian sample size determination for detecting heterogeneity in multi-site replication studies. *TEST*, 33(3), 697–716.

1B. The hierarchical model and the Bayes factor

- With m sites and n subjects per site, let t_j be the estimated effect size at site j. The hierarchical model is

$$t_j \mid \mu_j \sim N\Big(\mu_j, \sigma^2/n\Big), \qquad \mu_j \mid \mu, \tau^2 \sim N(\mu, \tau^2)$$

  where μ is the overall effect size and τ^2 the between-site variance

- Integrating out the site means gives the two competing models

$$\mathcal{M}_1 : \; t_j \mid \mu, \tau^2 \sim N\Big(\mu, \tau^2 + \sigma^2/n\Big), \qquad \mathcal{M}_0 : \; t_j \mid \mu \sim N\Big(\mu, \sigma^2/n\Big),$$

  with $\mathcal{M}_0$ recovered from $\mathcal{M}_1$ at $\tau^2 = 0$

- The two models are compared through the **Bayes factor** BF_{01}. Inference uses a weakly informative **analysis prior** on the relative heterogeneity τ/σ, while design computations rely on a separate, informative **design prior** expressing the level of heterogeneity worth detecting

- Sample size determination then selects (n, m) so that the prior predictive distribution of BF_{01} yields a high probability of correctly identifying heterogeneity, with a controlled probability of misleading evidence under $\mathcal{M}_0$

1C. SSD in the constrained setting via power priors

- The framework above is **unconstrained**, in the sense that the original publication that motivated the replication enters neither the design nor the analysis. The proposal is to extend it to the **constrained** setting, where the original study is brought into the model through a **power prior** on its likelihood
- The power prior involves a borrowing parameter $\alpha_0 \in [0, 1]$, with full borrowing at one and no borrowing at zero, and the choice of α_0 is itself an open question. Part of the work is to develop a principled way to set it, appropriate for the replication context
- Given that, the second strand is the corresponding **SSD**, that is to determine the number of sites m and subjects per site n for which the Bayes factor delivers compelling evidence on heterogeneity inside the power-prior-augmented model
- Compared to the unconstrained design, the expected payoff is a more efficient use of available information when the original study is informative, while preserving a smooth fall-back to the unconstrained SSD when it is not

2A. Self-starting change point detection: the Shiryaev framework[2]

- In short-run and startup applications there is no Phase I sample to calibrate the chart against, so the model must learn the IC state and monitor for shifts **simultaneously** from the very first observation
- Under the **At Most One Change (AMOC)** scenario we observe data $\mathsf{x}_n = (x_1, \ldots, x_n)$ sequentially, with a known IC density f_0, a known OOC density f_1, and an unknown change point $\tau \sim \text{Geom}(p)$
- The Shiryaev likelihood combines the two regimes around τ

$$f(\mathsf{x}_n \mid \tau) = \begin{cases} \prod_{i=1}^{\tau-1} f_0(x_i) \prod_{i=\tau}^{n} f_1(x_i), & \tau \le n \\ \prod_{i=1}^{n} f_0(x_i), & \tau > n \end{cases}$$

- At each time n we compute the posterior marginal probability of a change, $\Pr(\tau \le n \mid \mathsf{x}_n)$, and raise an alarm the first time it crosses a decision limit $p^\star$ chosen to control the false alarm rate

[2] Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*,

2B. A Bayesian generalisation of Shiryaev's framework[3]

- Shiryaev assumes that both f_0 and f_1 are **fully known**, with a Geometric prior for τ. Both assumptions are restrictive in practice
- The Univariate Self-Starting Shiryaev (**U3S**) treats both the IC and the OOC parameters as unknown, with proper priors on each
- The Geometric prior is replaced by a **Discrete Weibull** on τ, which gives flexible control over the hazard (constant, increasing, or decreasing)
- The OOC parameter receives a **two-component mixture prior**, so a single chart can detect shifts in either direction (positive and negative shifts in the mean, or inflation and deflation of the variance)
- Closed-form models have been derived for the **Normal mean** and the **Normal variance**. After an alarm, the joint posterior also delivers inference on the size of the shift and on the change point itself

[3] Bourazas, K., & Tsiamyrtzis, P. (2024). Self-Starting Shiryaev (3S): A Bayesian online change point model and monitoring for short runs. *Working paper.*

2C. Resistance to absorption of a shift

- Self-starting charts learn the IC state from the same data they monitor. A shift that is missed early gets absorbed into the running parameter estimates and contaminates them
- This creates a **window of opportunity**, where missed early signals become harder and harder to detect, because the chart starts treating the contaminated data as if it were IC
- U3S is **resistant to this absorption**. Even when a shift is not detected at its onset, the chart does not lose the ability to detect it later, and the window of opportunity stays open
- The reason is structural. At every time n, the posterior probability of a change aggregates evidence across **every** candidate change point in the sample, so as long as a shift leaves a trace, U3S can still recover it, even after standard self-starting alternatives have absorbed it

2D. A Poisson U3S for count data

- Count data is ubiquitous in SPC, with applications ranging from defects per item to hospital admissions to network alarms, and the U3S framework has not yet been developed for this case
- A natural model places a multiplicative rate shift κ between the IC and OOC states, with conjugate Gamma priors on both rates and a Discrete Weibull on τ

$$\text{IC: } x_i \mid \lambda \overset{\text{iid}}{\sim} \text{Poisson}(\lambda), \quad \text{OOC: } x_i \mid (\lambda, \kappa) \overset{\text{iid}}{\sim} \text{Poisson}(\kappa\lambda),$$

$$\lambda \sim \text{Gamma}(a, b), \quad \kappa \sim \text{Gamma}(c, d)$$

- The Gamma-Poisson conjugacy delivers every marginal likelihood in closed form, so the U3S machinery transfers directly. In particular, the **resistance to absorption** carries over to the count setting, where it is especially valuable because small early rate shifts are notoriously easy to lose
- Natural further extensions include Negative Binomial data (overdispersion), zero-inflated counts, and bivariate count streams

3A. Bayesian classification with imperfect labels

- In many real problems the observed labels are not the truth. Audited firms may be flagged as fraudulent, but some genuine fraudsters never get caught, and a medical test may declare a healthy patient sick
- For each unit i we observe features X_i and a noisy label $Y_i \in \{0, 1\}$. The truth is a **latent class** $T_i \in \{0, 1\}$ that we never see directly
- A logistic link connects features to the true class

$$\Pr(T_i = 1 \mid X_i) = \sigma(f_i), \qquad \sigma(u) = \frac{1}{1 + e^{-u}}$$

  where f_i is a latent function of the features
- Our current model assumes **one-sided noise**. Audits are trustworthy when they flag a positive, but they miss some true positives

$$\Pr(Y_i = 1 \mid T_i = 0) = 0, \qquad \Pr(Y_i = 0 \mid T_i = 1) = \eta$$

- Marginalising over the latent class, the observed-label likelihood is $\Pr(Y_i = 1 \mid f_i, \eta) = (1 - \eta)\,\sigma(f_i)$

3B. Building the latent score as a product of experts[4]

- Information about each unit usually comes from several sources, for instance firm-level covariates and one or more relational networks (shared installers, shared brand, common membership)
- Each source j contributes its own latent function $x_{j,i}$, modelled as a zero-mean Gaussian process with a source-specific covariance kernel

$$\mathsf{x}_j \sim \mathcal{N}(0,\ \mathsf{C}_j), \qquad j = 0, 1, \ldots, m$$

- Under the logistic link the product-of-experts construction collapses to a clean additive sum

$$f_i = \sum_{j=0}^{m} x_{j,i}, \qquad \Pr(T_i = 1 \mid X_i) = \sigma(f_i)$$

  so the joint model behaves like a single GP with additive covariance $\mathsf{C}_{\text{sum}} = \sum_j \mathsf{C}_j$
- Bayesian inference then delivers, for each unit, a calibrated class probability with credible interval, an estimate of the false-negative rate η, and the expected number of **missed positives** among the unflagged units

[4] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 1771–1800.

3C. Proposed extension: two-sided label uncertainty

- One-sided noise assumes that whenever the audit flags a positive, it is correct. In many domains this is too strong, and flagged units may also be misclassified
- The extension allows for **both** types of error

$$\Pr(Y_i = 1 \mid T_i = 0) = \alpha, \qquad \Pr(Y_i = 0 \mid T_i = 1) = \beta$$

  with weakly informative priors $\alpha \sim \text{Beta}(a_\alpha, b_\alpha)$ and $\beta \sim \text{Beta}(a_\beta, b_\beta)$
- Marginalising over T_i, the observed-label likelihood becomes

$$\Pr(Y_i = 1 \mid f_i, \alpha, \beta) = (1 - \beta)\,\sigma(f_i) + \alpha\,\big(1 - \sigma(f_i)\big)$$

  which recovers the one-sided model exactly when $\alpha = 0$
- Identifiability requires $\alpha + \beta < 1$, so the observed labels are at least informative. In practice this is reinforced by informative priors, a small validation set with verified labels, or anchor points
- The framework is **domain-agnostic**, with natural applications in medical diagnosis without a gold standard, financial distress under an ambiguous default definition, and compliance in administrative data

4. Bayesian label propagation

- In many real settings, only a small fraction of nodes in a network carry reliable labels and the rest are unknown. **Label propagation** is a classical semi-supervised technique that exploits the network structure to spread the limited label information from known cases to their unlabeled neighbors
- The intuition is that connected nodes tend to share their true status, so confirmed positives inform our beliefs about nodes in their vicinity, and risk diffuses naturally across the graph
- A **Bayesian version** places a graph-structured prior on the latent label of every node and adds a layer for label noise. Beliefs then propagate through the network in a principled way, yielding posterior probabilities with credible intervals for every node, rather than point estimates
- The result is a coherent map of posterior risk across the entire network, from which **connected suspicious subnetworks naturally emerge** as candidates for further investigation

5. Bayesian Anomaly detection in random finite sets[5]

- In many real applications the observation at each time step is not a single fixed-dimensional vector but a *set* of points whose number varies from step to step. Examples include objects detected by a sensor in each frame, transactions arriving in a time window, or events recorded across a region
- Such observations are formalised as **Random Finite Sets (RFS)**: finite sets whose cardinality is itself random and whose elements are jointly random. Standard anomaly-detection methods, which assume fixed-dimensional inputs, do not naturally apply
- **Sequential anomaly detection** in this setting monitors an incoming stream of such sets and decides whether each new set conforms to the IC behavior of the process or signals a deviation, taking both the number of points and their spatial pattern into account
- Our recent work formulates this in a Bayesian framework that learns the IC distribution online from past sets and flags anomalies via posterior predictive checks, with a discounting mechanism that lets the model adapt smoothly to systematic shifts

[5] Bourazas, K., Papaioannou, S., & Kolios, P. (2025). Adaptive OOC point pattern detection in sequential random finite set observations. In *Proceedings of the 23rd European Control Conference (ECC)*, pp. 1549–1556, Thessaloniki, Greece.