Journal of Data Science 20 (2), 135–155
April 2022
DOI: 10.6339/22-JDS1039
Data Science in Action

# A Hybrid Monitoring Procedure for Detecting Abnormality with Application to Energy Consumption Data

Daeyoung Lim[1], Ming-Hui Chen[1,*], Nalini Ravishanker[1], Mark Bolduc[2], Brian McKeon[2], and Stanley Nolan[2]

[1]*Department of Statistics, University of Connecticut, 215 Glenbrook Rd. U-4120, Storrs, CT 06269-4120, United States*
[2]*Facilities Operations, University of Connecticut, 25 LeDoyt Road U-3252, Storrs, CT 06269-3252, United States*

**Abstract**

The complexity of energy infrastructure at large institutions increasingly calls for data-driven monitoring of energy usage. This article presents a hybrid monitoring algorithm for detecting consumption surges using statistical hypothesis testing, leveraging the posterior distribution and its information about uncertainty to introduce randomness in the parameter estimates, while retaining the frequentist testing framework. This hybrid approach is designed to be asymptotically equivalent to the Neyman-Pearson test. We show via extensive simulation studies that the hybrid approach enjoys control over type-1 error rate even with finite sample sizes whereas the naive plug-in method tends to exceed the specified level, resulting in overpowered tests. The proposed method is applied to the natural gas usage data at the University of Connecticut.

**Keywords** *Bayesian; computationally-intensive method; frequentist; hypothesis testing*

## 1 Introduction

Large academic institutions face challenges in energy management due to complex energy infrastructure (University of Michigan, 2011; Worcester Polytechnic Institute, 2007). The number of buildings is just one factor that makes energy management difficult. In most institutions, energy management envelops energy auditing, energy bills, life cycle costing, electrical distribution systems, boilers and fired systems, steam distribution, cogeneration, energy management control systems, insulation, compressed air, renewable energy sources and water management, distributed generation, and codes standards and legislation (Doty and Turner, 2004; Capehart et al., 2020). The sheer complexity of energy management calls for a screening process to identify anomalous energy usage with the goal of reducing the number of accounts that have to be inspected manually. Manual inspection is time-consuming and labor-intensive, and facilities professionals' time is wasted if an inspection turns out to have been unnecessary. A screening process to detect energy usage anomalies in buildings ultimately depends on what defines an anomalous behavior. In general, anomalies are observations in energy usage that substantially deviate from what is expected. Statistical monitoring learns normal behaviors from data to locate atypical behaviors. Zhang and Paschalidis (2018) derive a Hoeffding test statistic for network systems, leveraging large deviation theory and assuming that observations follow a finite-state Markov

*Corresponding author. Email: ming-hui.chen@uconn.edu.


Received December 11, 2021; Accepted February 13, 2022

chain. They use the *relative entropy* between the empirical and theoretical probability laws to define what is anomalous. Likewise, in Fu and Jeske (2014), anomaly detection in network traffic is formulated as a hypothesis test of $H_0 : \boldsymbol{\mu} = \boldsymbol{\beta}_0$ versus $H_1 : \boldsymbol{\mu} = c\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ indicates a coefficient vector of fixed effects (the daily or weekly mean network traffic counts) and $c$ is a percentage increment tuning parameter that represents the maximum acceptable multiplicative increase. The true parameter $\boldsymbol{\beta}_0$ is often unknown in anomaly detection problems, whereas it should usually be prespecified under a conventional hypothesis testing setting. The main contribution of Fu and Jeske (2014) is that their proposed method resolves the problem of unknown $\boldsymbol{\beta}_0$ and potential nuisance parameters in a model. However, they essentially take an estimate for $\boldsymbol{\beta}_0$ as truth, which inherently harbors unaddressed randomness by virtue of being estimated. Their confidence in using plug-in estimates stems from the large amount of network traffic data readily available due to its high-frequency nature.

Existing literature on "anomaly detection" methods for energy consumption proposes various ways to quantify "what is expected" and how severely the observed data depart from it. Seem (2007) proposes a *modified z-score* after identifying outliers using the *generalized studentized deviates*. Zhao (2014) assumes that the difference between adjacent observations should approximately be the same, which works as a measure of large discontinuity. Rashid and Singh (2018) propose distance-based abnormality scores. To the best of our knowledge, these papers on anomaly detection are mainly focused on high-frequency demand or usage data, which do not use properties available for coarser-grained monthly data. In the time series data framework, anomaly detection is formulated as a change point detection problem. For instance, Ross et al. (2011) and Ross et al. (2013) develop change point detection models based on hypothesis tests where $p$-values govern statistical decision-making, while Raftery and Akman (1986) model the data as a stream of Poisson random variables whose parameter changes from $\lambda_1$ to $\lambda_2$ after an unknown break point.

Our paper was primarily motivated by Fu and Jeske (2014) where the data set is partitioned into historical data and monitored data (see Section 2 for a detailed description of the data) and the historical data set is used to learn normal behavior, which is then used to test the monitored data. This framework inevitably involves estimating parameters to be used in hypothesis testing, which requires large data. However, unlike network traffic systems, the amount of data available to train the model is usually smaller for energy usage data, yielding estimates with higher uncertainty. To overcome this shortcoming, multiple buildings *a priori* believed to behave similarly are grouped together to borrow strength from each other. This alone cannot remedy the uncertainty incurred by replacing truth with an estimate. In that direction, the posterior distribution is a powerful object that encodes uncertainty in its own way and can be used to reintroduce randomness that was ignored in the plug-in scheme. Unfortunately, the posterior predictive distribution of the test statistic is not available in closed form, and a computationally intensive method is required to construct it via Monte Carlo sampling.

## 2 A Motivating Case: UConn Energy Data

Facilities Operations at the University of Connecticut (UConn) collects and monitors data on utility consumption. UConn's main campus in Storrs, Connecticut, has 120 buildings with over 400 separate external utility accounts for natural gas and electricity provided by local utility service companies, CNG and Eversource. The main campus has an additional 150 buildings with over 240 internal meter accounts for utilities produced by UConn's Central Utility Plant

providing the campus with electricity, steam, and chilled water. We mainly focus on UConn's natural gas consumption measured monthly by UConn Facilities Operations. The raw data consist of monthly measurements of natural gas consumption collected by UConn Facilities Operations. A total of 259 separate utility accounts are available across 120 buildings, each building having a varying number of meters. On a monthly basis, CNG, the local utility service provider for natural gas, provides UConn with a spreadsheet that contains all the account information including natural gas usage. In what follows, the building names and meter codes have been masked due to privacy considerations. UConn CNG data span 14 years from February 2007 to December 2020. Out of 245 CNG meters, we select 71 meters installed in residential buildings on UConn Storrs campus. Residential buildings have been grouped together *a priori* due to their similar energy consumption profiles. Energy meters have been installed and activated in university buildings at different times, making data availability different for each account. All 70 accounts for UConn's residential buildings were available from January 2008 and December 2016. One account was removed due to at least one unavailable measurement between Jan. 2008 and Dec. 2016. Each observation corresponds to a single utility bill for the associated meter in one billing period, measured as the difference between two readings in hundred cubic feet (abbreviated to CCF). The billing start and end dates do not align uniformly across accounts, and therefore, the meter readings are adjusted for the differing number of days within each billing cycle to amount to 30-day use. Furthermore, the building size must be accounted for, since larger buildings tend to use more gas. Each measurement is adjusted to represent gas usage for 100 sqft. The final normalization becomes as follows:

$$\frac{30 \times 100 \times y}{(\texttt{\#days})(\texttt{sqft})(\texttt{degree-days})}, \tag{1}$$

where $y$ is the observation, `sqft` indicates the square footage and `degree-days` is the number of degree days of the corresponding month (see Section 6). A set of 12 months forms a "cycle" regardless of the month the cycle starts with. For example, Feb. 2012 to Jan. 2013 can form a cycle. For practical purposes, most interesting cycles are based on calendar years, fiscal years, or academic years. We represent the measurement for the $i$th year, $j$th month, and $k$th account as $y^h_{ijk}$, where $i = 1, \ldots, T$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$. A batch of observations within a cycle need not be 12, and therefore, we let $j$ take on numbers up to an arbitrary $J$ smaller than (or equal to) 12—i.e., $j = 1, \ldots, J$ (see Section 4.2). The batch that we are interested in regarding the existence of anomalous activities is in the last cycle, which we call *monitored* or *test* data, denoted by $(\boldsymbol{y}_{T+1,1}, \ldots, \boldsymbol{y}_{T+1,K})$, where $\boldsymbol{y}_{T+1,k} = (y_{T+1,1,k}, \ldots, y_{T+1,J,k})^\top$. This naturally gives rise to $T$ cycles of *historical data*, which we write $(\boldsymbol{y}^h_{i1}, \ldots, \boldsymbol{y}^h_{iK})$ for $i = 1, \ldots, T$, where $\boldsymbol{y}^h_{ik} = (y^h_{ijk}, \ldots, y^h_{iJk})^\top$ and the superscript $h$ indicates historical data. Figure 1 contains two heat maps, showing the standard deviations of the normalized natural gas use data for 70 residential buildings on UConn's main campus, with higher standard deviations colored darker. The left panel illustrates the standard deviations across years given a month and an account. Except for the first account (leftmost column), cells in each row are colored similarly. This suggests that the normalization removed various sources of variation and allowed years to be assumed as exhibiting similar variability. The same observation is possible for the right panel, containing standard deviations across accounts given a month and a year. Account variability is noticeably consistent within a month across years.

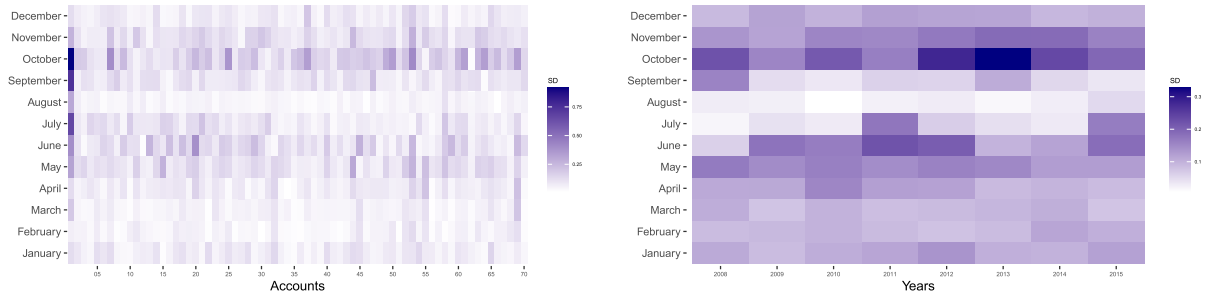



Figure 1: Heat maps of standard deviations. Left panel shows standard deviations across years for a given month and an account. Right panel shows standard deviations across accounts while month and account are held fixed.

## 3 Testing Parameters Under Uncertainty

In energy monitoring, there is no one-size-fits-all number representing the energy use that is considered normal for all buildings. This very lack of reference point complicates the formulation of hypothesis testing. An intuitive solution would be to check whether observations have significantly changed relative to the historical data, assuming observations were stable and consistent for several years (Fu and Jeske, 2014). Unfortunately, this involves replacing truth with an estimate, which leaves the uncertainty unresolved. Assume $\boldsymbol{y}_{T+1,k}$ independently follows a $J$-dimensional multivariate normal distribution, $N(\boldsymbol{\mu}_k, \Sigma)$, where $\boldsymbol{\mu}_k$ is a mean vector and $\Sigma$ is an unknown covariance matrix for $k = 1, \ldots, K$. Consider the following hypotheses:

$$H_0 : \boldsymbol{\mu}_k = \boldsymbol{\beta}_{0k} \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_k = c\boldsymbol{\beta}_{0k}, \tag{2}$$

where $c$ is a prespecified scalar greater than one, indicating multiplicative excess energy use beyond a tolerable level, and $\boldsymbol{\beta}_{0k}$ is a $J$-dimensional vector that reflects acceptable energy usage, prespecified using historical data. Note that $\Sigma$ also needs to be specified using historical data. In essence, this hypothesis test compares the mean energy consumption represented by $\boldsymbol{\mu}_k$ to an acceptable level, $\boldsymbol{\beta}_{0k}$. The only *truly* known, or user-specified, component in our setting is the percentage increment tuning parameter $c$. Provided that $\boldsymbol{\beta}_{0k}$'s and $\Sigma$ are given, a traditional test would proceed to find the most powerful rejection region given by the Neyman-Pearson lemma (Neyman and Pearson, 1933; Casella and Berger, 2002). Estimated $\boldsymbol{\beta}_{0k}$'s in a classical testing procedure may yield biased statistical inference due to the variability in the estimate. In the most general case, repeatedly splitting the data at random provides a way to reintroduce some randomness in the estimate by employing one data set for parameter estimation and the rest for testing—known as the *train-test split*. This bootstrapping scheme extracts information about uncertainty from subsetting the data to inject missing randomness into the test statistic (Efron and Tibshirani, 1994). When data are more structured than the general case, random splits no longer work. For example, the motivating UConn CNG data (see Section 2) are observed one batch per cycle, with a set of historical data $\{\boldsymbol{y}^h_{i1}, \ldots, \boldsymbol{y}^h_{iK}\}$ for $i = 1, \ldots, T$ and a set of test data $\{\boldsymbol{y}_{T+1,1}, \ldots, \boldsymbol{y}_{T+1,K}\}$ corresponding to the last cycle. The subscript $T+1$ in test data is henceforth omitted. We are only interested in testing the last cycle, where the historical data are assumed to have been generated from the null model, $H_0 : \boldsymbol{\mu}_k = \boldsymbol{\beta}_{0k}$. The specific interest in testing only the last cycle restricts our ability to generate random splits. Although it is technically possible to bootstrap estimates from the historical data, it is challenged by prohibitively small bootstrap

sample sizes, and will produce an overpowered procedure similar to the naive plug-in procedures. Instead, we turn to an alternative object that contains the uncertainty information: the posterior distribution.

**A Motivating Special Case** We illustrate the single-account setting where only one utility account is of interest (i.e., $k$ is given) and lay out the proposed algorithm as a motivating example. Let $\boldsymbol{y}$ be a random vector from a $J$-dimensional multivariate normal distribution $N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_J)$. This is a special case of $N(\boldsymbol{\mu}, \Sigma)$ where the covariance matrix $\Sigma = \sigma^2 \boldsymbol{I}_J$. This assumption implies that the energy consumption is independent across different months and shares a common variance $\sigma^2$, allowing the estimation of $\sigma^2$ without borrowing information from other accounts. Under this simple setting, the historical data are expressed as $\{\boldsymbol{y}_1^h, \ldots, \boldsymbol{y}_T^h\}$ (or $\{\boldsymbol{y}_{ik}^h\}$ for a given $k$ and $i = 1, \ldots, T$ for notational consistency), and the test data are $\boldsymbol{y}$. We drop the subscript $k$ henceforth. The procedure is twofold: (1) the historical data are used to estimate the parameters; and (2) the test data are used to derive the rejection region and compute the test statistic. The hypotheses remain identical to those in Equation (2). The historical data $\boldsymbol{y}_i^h$ for $i = 1, \ldots, T$ are used to estimate the parameters $\boldsymbol{\beta}_0$ and $\sigma^2$. This model can be recast into the following matrix-vector form, $\boldsymbol{y}^h = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where $\boldsymbol{y}^h = \mathrm{vec}([\boldsymbol{y}_1^h \ \boldsymbol{y}_2^h \ \cdots \ \boldsymbol{y}_T^h]^\top)$ for which $\mathrm{vec}(\cdot)$ denotes vectorization stacking the column vectors, $\boldsymbol{X} = \oplus_{j=1}^J \boldsymbol{1}_T$ for which $\oplus$ indicates direct sum and $\boldsymbol{1}_T$ is the $T$-dimensional vector of ones, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{TJ})$. Then the least-squares estimator of $\boldsymbol{\beta}_0$ is $\widehat{\boldsymbol{\beta}}_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}^h$ where $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{I}_J / T$. Therefore, $\widehat{\boldsymbol{\beta}}_0 = (\overline{y}_{\cdot 1}^h, \overline{y}_{\cdot 2}^h, \ldots, \overline{y}_{\cdot J}^h)^\top$, where

$$\overline{y}_{\cdot j}^h = \frac{1}{T} \sum_{i=1}^{T} y_{ij}^h.$$

It is well-known that the least-squares estimator follows a normal distribution, i.e., $\widehat{\boldsymbol{\beta}}_0 \sim N(\boldsymbol{\beta}_0, \frac{\sigma^2}{T} \boldsymbol{I}_J)$. Also, the unbiased estimator of $\sigma^2$ and its sampling distribution are given as follows:

$$\widehat{\sigma}^2 = \frac{\| \boldsymbol{y}^h - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0 \|^2}{J(T-1)}, \quad \frac{J(T-1)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{J(T-1)},$$

where $\chi^2_\nu$ indicates a chi-squared distribution with $\nu$ degrees of freedom. This ultimately gives the posterior distributions under a uniform prior, $\pi(\boldsymbol{\beta}_0, \sigma^2) \propto 1$, as follows:

$$\boldsymbol{\beta}_0 \mid \sigma^2, \boldsymbol{y}^h \sim N\left(\widehat{\boldsymbol{\beta}}_0, \frac{\sigma^2}{T} \boldsymbol{I}_J\right), \quad \sigma^2 \mid \boldsymbol{y}^h \sim \mathrm{IG}\left(\frac{J(T-1)}{2}, \frac{J(T-1)\widehat{\sigma}^2}{2}\right),$$

where $X \sim \mathrm{IG}(a, b)$ denotes the inverse-gamma distribution whose density function is proportional to $x^{-(a+1)} \mathrm{e}^{-b/x}$. Subsequently, the test data is used to derive the test statistic and rejection region corresponding to the most powerful test through the Neyman-Pearson lemma. Writing the likelihood ratio of the alternative hypothesis to the null hypothesis as $\Lambda$ (i.e., $\Lambda = \frac{L(c\boldsymbol{\beta}_0, \sigma^2)}{L(\boldsymbol{\beta}_0, \sigma^2)}$), $\log \Lambda$ follows a normal distribution under $H_0$ (see Appendix B), which upon standardization yields

$$W := \frac{\log \Lambda - \mathsf{E}_{H_0}(\log \Lambda)}{\sqrt{\mathrm{Var}_{H_0}(\log \Lambda)}} = \frac{\sum_{j=1}^J (y_j - \beta_{0j})\beta_{0j}}{\sigma \sqrt{\sum_{j=1}^J \beta_{0j}^2}} \sim N(0, 1).$$

The posterior distributions of the parameters capture the variability introduced in the estimation step. The main idea of our procedure lies in using the posterior distribution to *calibrate* the

**Algorithm 1** Single-account testing procedure.

**procedure** SINGLETESTING($\boldsymbol{y}^h$, $\boldsymbol{y}$, $\alpha$, $B_1$, $B_2$, $B_3$)
  $\widehat{\boldsymbol{\beta}}_0 \leftarrow (\overline{y}^h_{\cdot 1}, \overline{y}^h_{\cdot 2}, \ldots, \overline{y}^h_{\cdot J})^\top$ and $\widehat{\sigma}^2 \leftarrow [J(T-1)]^{-1} \sum_{i=1}^{T} \sum_{j=1}^{J} (y^h_{ij} - \overline{y}^h_{\cdot j})^2$
  **for** $b_2 = 1 : B_2$ **do**
    Generate $\sigma_f^2 \sim \text{IG}(J(T-1)/2, J(T-1)\widehat{\sigma}^2/2)$ and $\boldsymbol{\beta}^f \sim N(\widehat{\boldsymbol{\beta}}_0, \frac{\sigma^2}{T}\mathbf{I}_J)$
    Generate $\boldsymbol{y}^f \sim N(\boldsymbol{\beta}^f, \sigma_f^2 \mathbf{I}_{12})$
    **for** $b_1 = 1 : B_1$ **do**
      Generate $\sigma^2 \sim \text{IG}(J(T-1)/2, J(T-1)\widehat{\sigma}^2/2)$ and $\boldsymbol{\beta} \sim N(\widehat{\boldsymbol{\beta}}_0, \frac{\sigma^2}{T}\mathbf{I}_J)$
      $\delta_{b_1} \leftarrow I\left(\frac{\boldsymbol{\beta}'(\boldsymbol{y}^f - \boldsymbol{\beta})}{\sigma\sqrt{\boldsymbol{\beta}'\boldsymbol{\beta}}} > z_{1-\alpha}\right)$
    $\widehat{p}_{b_2} \leftarrow \frac{1}{B_1} \sum_{b_1=1}^{B_1} \delta_{b_1}$
  $\gamma_\alpha \leftarrow (1-\alpha)^{\text{th}}$-quantile$(\widehat{\mathbf{p}})$
  **for** $b = 1 : B_3$ **do**
    Generate $\sigma^2 \sim \text{IG}(J(T-1)/2, J(T-1)\widehat{\sigma}^2/2)$ and $\boldsymbol{\beta} \sim N(\widehat{\boldsymbol{\beta}}_0, \frac{\sigma^2}{T}\mathbf{I}_J)$
    $\delta_b \leftarrow I\left(\frac{\boldsymbol{\beta}'(\boldsymbol{y} - \boldsymbol{\beta})}{\sigma\sqrt{\boldsymbol{\beta}'\boldsymbol{\beta}}} > z_{1-\alpha}\right)$
  $\widehat{p}_{\text{obs}} \leftarrow \frac{1}{B_3} \sum_{b=1}^{B_3} \delta_b$
  **return** $H_1$ if $\widehat{p}_{\text{obs}} \geqslant \gamma_\alpha$ and $H_0$ otherwise

frequentist testing procedure. The step-by-step description of our hybrid procedure is given in Algorithm 1. In our procedure, the test statistic is computed repeatedly using the parameter values generated from the posterior distribution, and the corresponding simulated data. This simulated data set is written as $\boldsymbol{y}^f$ where the superscript $f$ indicates *future values*. Then, the same calibration is performed on the test data to compute the calibrated $p$-value $\widehat{p}_{\text{obs}}$, which is ultimately compared to $\mathbf{p}$, the sample of posterior predictive $p$-values. If the observed $p$-value $\widehat{p}_{\text{obs}}$ associated with $\boldsymbol{y}$ exceeds the upper $\alpha$-th quantile of $\mathbf{p}$, the null hypothesis is rejected.

## 4 Hybrid Monitoring for Multiple Accounts

The common variance assumption and independence of energy usage across months in Section 2 are too restrictive. Estimating an unstructured $\boldsymbol{\Sigma}$ typically requires a large amount of historical data (i.e., a sufficiently large $T$) if we are using data from a single account. There are two key motivations: (1) estimating $\boldsymbol{\Sigma}$ by pooling information; and (2) monitoring energy usage of all accounts. However, this approach requires that all accounts share the same covariance matrix, which is well-supported by Figure 1. Assume $\boldsymbol{y}_k$ follows a $J$-dimensional multivariate normal distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with a mean vector $\boldsymbol{\mu}_k$ and an unknown covariance matrix $\boldsymbol{\Sigma}$ for $k = 1, \ldots, K$. The data are now modeled as being correlated within a cycle but independent between cycles. Now consider the following hypotheses:

$$H_0 : \boldsymbol{\mu}_k = \boldsymbol{\beta}_{0k} \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_k = \boldsymbol{C}\boldsymbol{\beta}_{0k} \quad \text{for all } k, \tag{3}$$

where $\boldsymbol{C}$ is the percentage increment matrix. An example of $\boldsymbol{C}$ is $\text{diag}(c_1, \ldots, c_J)$ with $c_j > 1$ for all $j$. The historical data can be expressed as $\boldsymbol{y}^h_{ik} = \boldsymbol{\beta}_{0k} + \boldsymbol{\epsilon}_{ik}$ for $i = 1, \ldots, T$ and $k = 1, \ldots, K$, where $\boldsymbol{\beta}_{0k} = (\beta_{1,0k}, \ldots, \beta_{J,0k})^\top$ is a mean vector for the $k$th account, $\boldsymbol{\epsilon}_{ik} \overset{\text{IID}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma}$ is an unknown $J \times J$ symmetric positive-definite matrix. Unbiased estimators for $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}$ are

well-known, computed using the historical data:

$$\widehat{\boldsymbol{\beta}}_{0k} = \overline{\boldsymbol{y}}^h_{\cdot k} = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{y}^h_{ik}, \quad \widehat{\Sigma} = \frac{1}{K(T-1)}\sum_{i=1}^{T}\sum_{k=1}^{K}(\boldsymbol{y}^h_{ik} - \widehat{\boldsymbol{\beta}}_{0k})(\boldsymbol{y}^h_{ik} - \widehat{\boldsymbol{\beta}}_{0k})^\top.$$

The hypothesis testing expressed in Equation (3) is inherently a multiple testing problem. Denoting the test data by $\boldsymbol{y}_k = (y_{1k}, \ldots, y_{Jk})^\top$, the logarithm of the likelihood ratio follows a normal distribution under $H_0$ by the same logic as in the motivating special case, where assuming $c = c_1 = \cdots = c_J$ yields under $H_0$

$$W_k := \frac{\log \Lambda_k - \mathsf{E}_{H_0}(\log \Lambda_k)}{\sqrt{\mathrm{Var}_{H_0}(\log \Lambda_k)}} = \frac{\boldsymbol{\beta}_{0k}^\top \Sigma^{-1}(\boldsymbol{y}_k - \boldsymbol{\beta}_{0k})}{\sqrt{\boldsymbol{\beta}_{0k}^\top \Sigma^{-1}\boldsymbol{\beta}_{0k}}} \sim N(0,1). \tag{4}$$

The sum-of-squares matrix $\boldsymbol{S} := K(T-1)\widehat{\Sigma} = \sum_{i=1}^{T}\sum_{k=1}^{K}(\boldsymbol{y}^h_{ik} - \widehat{\boldsymbol{\beta}}_{0k})(\boldsymbol{y}^h_{ik} - \widehat{\boldsymbol{\beta}}_{0k})^\top$ follows a Wishart distribution, $\mathrm{W}(K(T-1), \Sigma)$, where $\mathrm{W}_J(\nu, \Sigma)$ is the Wishart distribution with $\nu$ degrees of freedom and a $J \times J$ scale matrix $\Sigma$ whose density function is

$$f_{\mathrm{W}}(\boldsymbol{X} \mid \nu, \Sigma) = \frac{|\boldsymbol{X}|^{(\nu-J-1)/2}\mathrm{etr}(-\Sigma^{-1}\boldsymbol{X}/2)}{2^{\nu J/2}|\Sigma|^{\nu/2}\Gamma_J(\nu/2)} I(\boldsymbol{X} \in \mathcal{S}^J_{++}),$$

$\Gamma_J$ is the multivariate gamma function defined by

$$\Gamma_J(z) = \pi^{J(J-1)/4}\prod_{j=1}^{J}\Gamma[z + (1-j)/2],$$

and $\mathcal{S}^J_{++}$ is the space of $J \times J$ symmetric positive-definite matrices. This in turn yields a posterior distribution $\Sigma \mid \boldsymbol{y}^h \sim \mathrm{IW}(K(T-1) - J - 1, \boldsymbol{S})$ under a noninformative prior, $\pi(\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0K}, \Sigma) \propto 1$, where $\mathrm{IW}_J(\nu, \Sigma)$ is the Inverse-Wishart distribution with $\nu$ degrees of freedom and a $J \times J$ scale matrix $\Sigma$ whose density function is

$$f_{\mathrm{IW}}(\boldsymbol{X} \mid \nu, \Sigma) = \frac{|\Sigma|^{\nu/2}\mathrm{etr}\left(-\Sigma\boldsymbol{X}^{-1}/2\right)}{2^{J\nu}\Gamma_J(\nu/2)|\boldsymbol{X}|^{(\nu+J+1)/2}} I(\boldsymbol{X} \in \mathcal{S}^J_{++}).$$

Likewise, the (conditional) posterior distribution of $\boldsymbol{\beta}_{0k}$ is given by $\boldsymbol{\beta}_{0k} \mid \Sigma, \boldsymbol{y}^h \overset{\text{IND}}{\sim} N(\widehat{\boldsymbol{\beta}}_{0k}, \Sigma/T)$. For every $k$, the observed value of $W_k = w_k$ will produce a $p$-value $u_k = 1 - \Phi(w_k)$. There is extensive research on how to adjust $K$ $p$-values to control the familywise error rate or the false discovery rate. In multiple-testing literature, the null hypothesis, $H_0 : \boldsymbol{\mu}_k = \boldsymbol{\beta}_{0k}$ for all $k$, is referred to as the *grand null*, and $H_1 : \boldsymbol{\mu}_k \neq \boldsymbol{\beta}_{0k}$ for at least one $k$ is equivalently the grand alternative hypothesis. For our procedure (see Algorithm 2), we consider five multiplicity corrections including the Bonferroni correction (Dunn, 1961; Holm, 1979; Hochberg, 1988; Hommel, 1988; Simes, 1986; Šidák, 1967; Benjamini and Hochberg, 1995). The adjusted $p$-value (Wright, 1992) for each method is computed as follows:

- Bonferroni: $\tilde{p}_k = Ku_k$
- Holm: $\tilde{p}_{(k)} = \max_{j \leqslant k}\{(K+1-j)u_{(j)}\}_1$, where $\{x\}_1 = \min(x, 1)$
- Hochberg: $\tilde{p}_{(k)} = \min_{j \geqslant k}\{(K+1-j)u_{(j)}\}_1$
- Hommel: no one-line expression for adjusting $p$-values
- Benjamini-Hochberg: $\tilde{p}_{(k)} = \min_{j \geqslant k}\{K/(K+1-j)u_{(j)}\}_1$

**Algorithm 2** Multiple-account testing procedure.

1: **procedure** MULTIPLETESTING($\boldsymbol{y}^h$, $\boldsymbol{y}$, $\alpha$, $B_1$, $B_2$, $B_3$, $h$)
2: $\widehat{\boldsymbol{\beta}}_{0k} \leftarrow \overline{\boldsymbol{y}}^h_{\cdot k}$ for $k = 1, \ldots, K$ and $\boldsymbol{S} \leftarrow \sum_{i=1}^{T}\sum_{k=1}^{K}(\boldsymbol{y}^h_{ik} - \overline{\boldsymbol{y}}^h_{\cdot k})(\boldsymbol{y}^h_{ik} - \overline{\boldsymbol{y}}^h_{\cdot k})^\top$
3: **for** $b_2 = 1 : B_2$ **do**
4: Generate $\Sigma_f \sim \text{IW}(K(T-1) - J - 1, \boldsymbol{S})$ and $\boldsymbol{\beta}^f_k \overset{\text{IID}}{\sim} N(\overline{\boldsymbol{y}}_{\cdot k}, \Sigma_f / T)$ for $k = 1, \ldots, K$
5: Generate $\boldsymbol{y}^f_k \sim N(\boldsymbol{\beta}^f_k, \Sigma_f)$
6: **for** $b_1 = 1 : B_1$ **do**
7: Generate $\Sigma \sim \text{IW}(K(T-1) - J - 1, \boldsymbol{S})$
8: Generate $\boldsymbol{\beta}_k \overset{\text{IID}}{\sim} N(\overline{\boldsymbol{y}}_{\cdot k}, \Sigma / T)$ for $k = 1, \ldots, K$
9: $u_k \leftarrow 1 - \Phi\left(\frac{\boldsymbol{\beta}_k^\top \Sigma^{-1}(\boldsymbol{y}^f_k - \boldsymbol{\beta}_k)}{\sqrt{\boldsymbol{\beta}_k^\top \Sigma^{-1} \boldsymbol{\beta}_k}}\right)$ for $k = 1, \ldots, K$
10: Compute $(\tilde{p}_1, \ldots, \tilde{p}_K) \leftarrow h(u_1, \ldots, u_K)$
11: $\delta_{b_1} \leftarrow \mathsf{any}(\tilde{p}_1 < \alpha, \ldots, \tilde{p}_K < \alpha)$
12: $\widehat{p}_{b_2} \leftarrow \frac{1}{B_1}\sum_{b_1=1}^{B_1} \delta_{b_1}$
13: $\gamma_\alpha \leftarrow (1-\alpha)^{\text{th}}$-quantile$(\widehat{\mathbf{p}})$
14: **for** $b = 1 : B_3$ **do**
15: Generate $\Sigma \sim \text{IW}(K(T-1) - J - 1, \boldsymbol{S})$ and $\boldsymbol{\beta}_k \overset{\text{IID}}{\sim} N(\overline{\boldsymbol{y}}_{\cdot k}, \Sigma / T)$ for $k = 1, \ldots, K$
16: $u_k \leftarrow 1 - \Phi\left(\frac{\boldsymbol{\beta}_k^\top \Sigma^{-1}(\boldsymbol{y}_k - \boldsymbol{\beta}_k)}{\sqrt{\boldsymbol{\beta}_k^\top \Sigma^{-1} \boldsymbol{\beta}_k}}\right)$ for $k = 1, \ldots, K$
17: Compute $(\tilde{p}_1, \ldots, \tilde{p}_K) \leftarrow h(u_1, \ldots, u_K)$
18: $\delta_b \leftarrow \mathsf{any}(\tilde{p}_1 < \alpha, \ldots, \tilde{p}_K < \alpha)$
19: $\widehat{p}_{\text{obs}} \leftarrow \frac{1}{B_3}\sum_{b=1}^{B_3} \delta_b$
20: **return** $H_1$ if $\widehat{p}_{\text{obs}} \geqslant \gamma_\alpha$ and $H_0$ otherwise

Holm's procedure is replaced with the Holm-Šidák method in our simulation studies to increase power. The corresponding adjusted $p$-value becomes $\tilde{p}_{(k)} = \max_{j \leqslant k}\{1 - (1 - u_{(j)})^{K+1-j}\}_1$.

Algorithm 2 describes the proposed multiple-account procedure, where $h(u_1, \ldots, u_K)$ is the chosen algorithm for computing the adjusted $p$-values. The multiple-account procedure follows the same logic as that of Algorithm 1. However, unlike Algorithm 1, this multiple-account procedure defies an analytical expression of its statistical power for a given multiplicity correction. We therefore provide a Monte Carlo scheme for computing the statistical power once we have full knowledge of the data generation process. See Section 5 for detail.

### 4.1 Complexity Analysis

The computationally intensive nature of Algorithm 2 merits a brief discussion about its computational cost. Since it is well-known that nested `for` loops are costly, we focus our discussion on the innermost `for` loop that repeats a sequence of test statistic calculation $B_1$ times. The most computational overhead comes from either line 7 or line 9 in Algorithm 2. Generating a $J \times J$ inverse-Wishart random matrix $\Sigma$ produces its inverse $\Sigma^{-1}$ as a byproduct. The inversion here is of $O(J^3)$, where $O(f(n))$ indicates a class of algorithms whose order is $f(n)$—i.e., an algorithm $g(n)$ is an element of $O(f(n))$ if there exist a positive integer $N$ and a positive real number $c$ such that $0 \leqslant g(n) \leqslant cf(n)$ for all $n \geqslant N$ (Cormen et al., 2022). In line 9, $\Sigma^{-1}(\boldsymbol{y}^f_k - \boldsymbol{\beta}_k)$ can be done in one step for all $k = 1, \ldots, K$ by stacking $\{\boldsymbol{y}^f_k - \boldsymbol{\beta}_k\}_{k=1}^K$ into a $J \times K$ matrix $\boldsymbol{D}$

whose $k$-th column is $\boldsymbol{y}_k^f - \boldsymbol{\beta}_k$. Then, using the fact that the computational complexity of matrix multiplication between an $m \times n$ matrix and an $n \times p$ matrix is $O(mnp)$, calculating $\Sigma^{-1}\boldsymbol{D}$ is of $O(J^2K)$. In theory, the computational complexity of one iteration of the innermost `for` loop is therefore $O(J^2 \max(K, J))$. In practice, however, $J$ is the number of months in a cycle, bounded above by 12. This makes it highly probable that $K$ is greater than $J$. All in all, the complexity of Algorithm 2 is therefore $O(B_1 B_2 J^2 \max(K, J))$. As daunting as it may seem, there is a source of significant computational gains in the proposed method: parellelization. The posterior distribution $\pi(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \Sigma)$ does not require Markov chain Monte Carlo sampling, which lends itself to parallelization. There are various ways to parallelize a program (see Section 5 for our settings). Although parallelizing Algorithm 2 considerably speeds up computation, a high value of $B_1 B_2$ can easily cancel out the acceleration. We recommend that both $B_1$ and $B_2$ be smaller than 2,000 for a good balance between performance and computational feasibility.

### 4.2 Testing Before a Full Cycle

Testing the last full cycle of data was initially proposed in Fu and Jeske (2014) for high-frequency data collection. However, testing full-cycle may be too prohibitive for practical use. For instance, we may want to test energy usage before having observed a full 12 months. The aggregate nature of the UConn CNG data lends itself well to the central limit theorem, allowing the use of the multivariate normal distribution as our sampling distribution. Fortunately, the affine property of the multivariate normal distribution (Rao, 1973; Ravishanker et al., 2022) allows the restriction to be relaxed to any subset of the observations. The index $j = 1, \ldots, J$ has been kept arbitrary on purpose to accommodate this relaxation. That is,

$$\tilde{\boldsymbol{y}}_k := E^\top \boldsymbol{y}_k \sim N(E^\top \boldsymbol{\mu}_k, E^\top \Sigma E),$$

where $E$ is a $J \times a$ selection matrix whose columns are standard unit vectors $\boldsymbol{e}_j = (0, 0, \ldots, 1, \ldots, 0)^\top$ whose $j$th element is one and zeros elsewhere, and $\tilde{\boldsymbol{y}}_k$ indicates an $a$-dimensional subvector of $\boldsymbol{y}_k$ of selected elements to be tested, for which $a$ is the number of selected elements. Redefining $J := a$ and $\boldsymbol{y}_k := \tilde{\boldsymbol{y}}_k$ brings us back to the original formulation—Equation (3)—with a mild abuse of notation.

### 4.3 Bayesian Interpretation

Our proposed procedures permit a straightforward Bayesian interpretation under noninformative priors. The posterior distribution of $\Sigma \mid \boldsymbol{y}$ and the conditional posterior distribution $\boldsymbol{\beta} \mid \Sigma, \boldsymbol{y}$ are given in Sun and Berger (2007). In our case, the distribution $\boldsymbol{\beta}_{0k} \mid \Sigma, \boldsymbol{y}$ is easily obtained as $N(\overline{\boldsymbol{y}}_{\cdot k}, \Sigma/T)$. As for the marginal posterior of $\Sigma$, assume the Jeffreys-type prior: $\pi(\boldsymbol{\beta}, \Sigma) \propto |\Sigma|^{-d/2}$ (Jeffreys, 1998; Geisser and Cornfield, 1963; Sun and Berger, 2007). Then the marginal posterior density becomes

$$\pi(\Sigma \mid \boldsymbol{y}) \propto |\Sigma|^{-\frac{K(T-1)+d}{2}} \operatorname{etr}(-\Sigma^{-1}\boldsymbol{S}/2) \sim \mathrm{IW}(K(T-1) + d - J - 1, \boldsymbol{S}). \tag{5}$$

If we choose $d$ to be the dimension of $\Sigma$, that is let $\pi(\boldsymbol{\beta}, \Sigma)$ to be Jeffreys' prior, then $\Sigma \mid \boldsymbol{y} \sim \mathrm{IW}(K(T-1) - 1, \boldsymbol{S})$. The first nested `for` loops in Algorithm 2 correspond to a Monte Carlo sampling scheme to construct the distribution of the random variables defined as an expectation as follows. Writing $U_k = 1 - \Phi(W_k(\boldsymbol{y}^f, \boldsymbol{\beta}_k, \Sigma))$ with $\Phi$ being the distribution function of the

standard normal distribution,

$$\begin{aligned}&\mathsf{E}_{\pi(\boldsymbol{\beta},\Sigma|\boldsymbol{y})}\left\{h(U_1(\boldsymbol{y}^f,\boldsymbol{\beta}_1,\Sigma),\ldots,U_K(\boldsymbol{y}^f,\boldsymbol{\beta}_K,\Sigma))\right\}\\&\quad=\iint h(U_1(\boldsymbol{y}^f,\boldsymbol{\beta}_1,\Sigma),\ldots,U_K(\boldsymbol{y}^f,\boldsymbol{\beta}_K,\Sigma))\,\pi(\boldsymbol{\beta}\mid\Sigma,\boldsymbol{y})\pi(\Sigma\mid\boldsymbol{y})\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\Sigma,\end{aligned}\tag{6}$$

where $\boldsymbol{y}^f$ follows the posterior predictive distribution. This expectation is a random variable whose randomness stems from $\boldsymbol{y}^f$. By sequentially sampling from $\Sigma^{(b)}\sim[\Sigma\mid\boldsymbol{y}]$ and $\boldsymbol{\beta}^{(b)}\sim[\boldsymbol{\beta}\mid\Sigma^{(b)},\boldsymbol{y}]$ for $b=1,\ldots,B$, the expectation in Equation 6 is approximated by

$$\begin{aligned}&\frac{1}{B}\sum_{b=1}^{B}h(U_1(\boldsymbol{y}^f,\boldsymbol{\beta}_1^{(b)},\Sigma^{(b)}),\ldots,U_K(\boldsymbol{y}^f,\boldsymbol{\beta}_K^{(b)},\Sigma^{(b)}))\\&\qquad\to\mathsf{E}_{\pi(\boldsymbol{\beta},\Sigma|\boldsymbol{y})}\left\{h(U_1(\boldsymbol{y}^f,\boldsymbol{\beta}_1,\Sigma),\ldots,U_K(\boldsymbol{y}^f,\boldsymbol{\beta}_K,\Sigma))\right\},\end{aligned}\tag{7}$$

as $B\to\infty$. The entire procedure is asymptotically equivalent to the frequentist Neyman-Pearson test due to posterior consistency (see Proposition 1 in Appendix A).

## 5 Simulation Studies

In this section, we conduct extensive simulation studies to investigate the performance of the proposed algorithms relative to the naive plug-in method. The size, or equivalently the type-1 error, and the power are the quantities of interest. Both the size analysis and the power analysis consist of a simulation study for the motivating special case with $\Sigma=\sigma^2\boldsymbol{I}_J$ for a given $k$ and another for the multiple-account case. We detail the data generation setting here; note that the historical data are assumed to follow the null hypothesis whereas the test data follow the null hypothesis and alternative hypothesis, respectively, in size analysis and power analysis. The related data sets are generated accordingly using R (R Core Team, 2021). All algorithms are implemented in C++ using Rcpp (Eddelbuettel and Balamuta, 2018) and RcppArmadillo (Eddelbuettel and Sanderson, 2014) for linear algebra, and OpenMP (OpenMP Architecture Review Board, 2018) for parallel programming. For the single-account case, we generate 10,000 data sets each for $T=3$, $T=5$, $T=10$, and $T=15$ from the model specification, where true $\boldsymbol{\beta}$ is set to $(6,7,8,9,10,11,12,13,14,15,16,17)^\top$ and $\sigma^2=2.5$. Recall that $T$ indicates the number of cycles in the historical data used to estimate the parameters. We run the algorithm for $\alpha=0.01$, $\alpha=0.025$, and $\alpha=0.05$. There are three tuning parameters: $B_1$, $B_2$, and $B_3$ indicating the numbers of Monte Carlo iterations within Algorithm 1 and Algorithm 2. The single-account case involves relatively few computational intensive matrix operations like matrix inversion or the Cholesky decomposition—the tuning parameters were chosen to be large numbers ($B_1=4000$, $B_2=5000$, $B_3=4000$). The seed number was set to 2797542 for data generation and 18007 for running the algorithms, respectively. For simulations related to the multiple-account case illustrated in Section 4, we generate 1,000 data sets each for $T=3$, $T=5$, $T=10$, $T=15$ from the appropriate model. That is, all historical data were generated from the model $\boldsymbol{y}_k\overset{\text{IND}}{\sim}N(\boldsymbol{\beta}_k,\Sigma)$ for $k=1,\ldots,70$, while the test data for power analysis were generated from $\boldsymbol{y}_k\overset{\text{IND}}{\sim}N(c\boldsymbol{\beta}_k,\Sigma)$ for $k=10,11,12$ where $c=1.2$. Note that $\boldsymbol{y}_k$ for $k\notin\{10,11,12\}$ still follow $N(\boldsymbol{\beta}_k,\Sigma)$ even in power analysis. The specific values of the true parameters, $\boldsymbol{\beta}_k$ and $\Sigma$ were selected to mirror the UConn CNG data, in which case $J=12$, and $c=1.2$ to indicate 20% increase in natural gas usage, a threshold beyond which utility engineers at the Facilities Operations deem excessive.

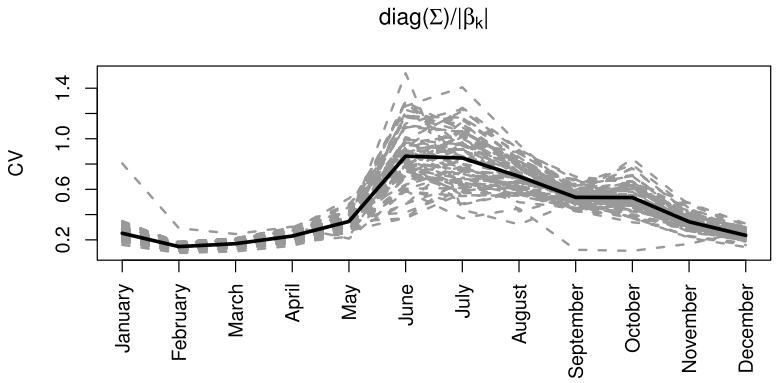



Figure 2: The coefficients of variation between $\boldsymbol{\beta}_k$ and the diagonal entries of $\boldsymbol{\Sigma}$ across 70 CNG meters installed in residential buildings at UConn.

### 5.1 Size Analysis

The three panels on the left column of Figure 3 visualize the comparison between Algorithm 1 and the naive plug-in method in the single-account case. It is clearly observed that the type-1 error is not controlled when plug-in estimates are used. Meanwhile, the proposed method effectively contains the type-1 error below the desired level. The discrepancy between the methods is the most pronounced when $T = 3$ since the number of samples used to estimate the (plug-in) unbiased estimator falls well short of that required for the law of large numbers to be at play, leaving much of the randomness in the estimator unaccounted for.

Similarly, the comparison between Algorithm 2 and the naive plug-in methods in the multiple-account case is summarized in Figure 3. The proposed methods, as well as the naive plug-in methods, depend on the multiple testing corrections denoted by $h(\cdot)$ in Algorithm 2. Four multiplicity corrections are considered—Holm, Hochberg, Hommel, and Benjamini-Hochberg. The familywise error rate, the multivariate equivalent of the type-1 error rate, is slightly elevated to avoid too low a probability of anomaly detection ($\alpha \in \{0.025, 0.05, 0.1\}$).

The bar plots show that the inflated type-1 error issue that plagued the plug-in methods in the single-account case carries over to the multiple-account case regarding the FWER. All three panels in the right column of Figure 3 have bars labeled with the letter "N", indicating that the corresponding values were computed using "naive" (plug-in) methods, exceeding the desired FWERs marked as red lines. Note that the inflated FWERs of the plug-in estimates subside as more data become available, i.e., $T$ increases. On the other hand, the proposed methods exhibit a rather conservative behavior. We observe that the proposed algorithms "play it safe" with scarce data, or equivalently with large randomness, and grow more confident to reject the grand null hypothesis as more data come in. Little difference in the performances exists across multiplicity corrections. It has been observed that the proposed method exhibits type-1 error rate that approaches the desired FWER from below, whereas the plug-in method does so from above. In theory, both methods should converge in the presence of sufficient data, due to frequentist consistency. To confirm this, an additional set of simulations with 4,629 data sets has been conducted for $T = 100$. Note that 100 years of historical data are unrealistic in practice. As evidenced in Figure 4, the two methods return indistinguishable size estimates with a wealth of data. This finding supports our method's control over size in circumstances with small-to-moderate data.

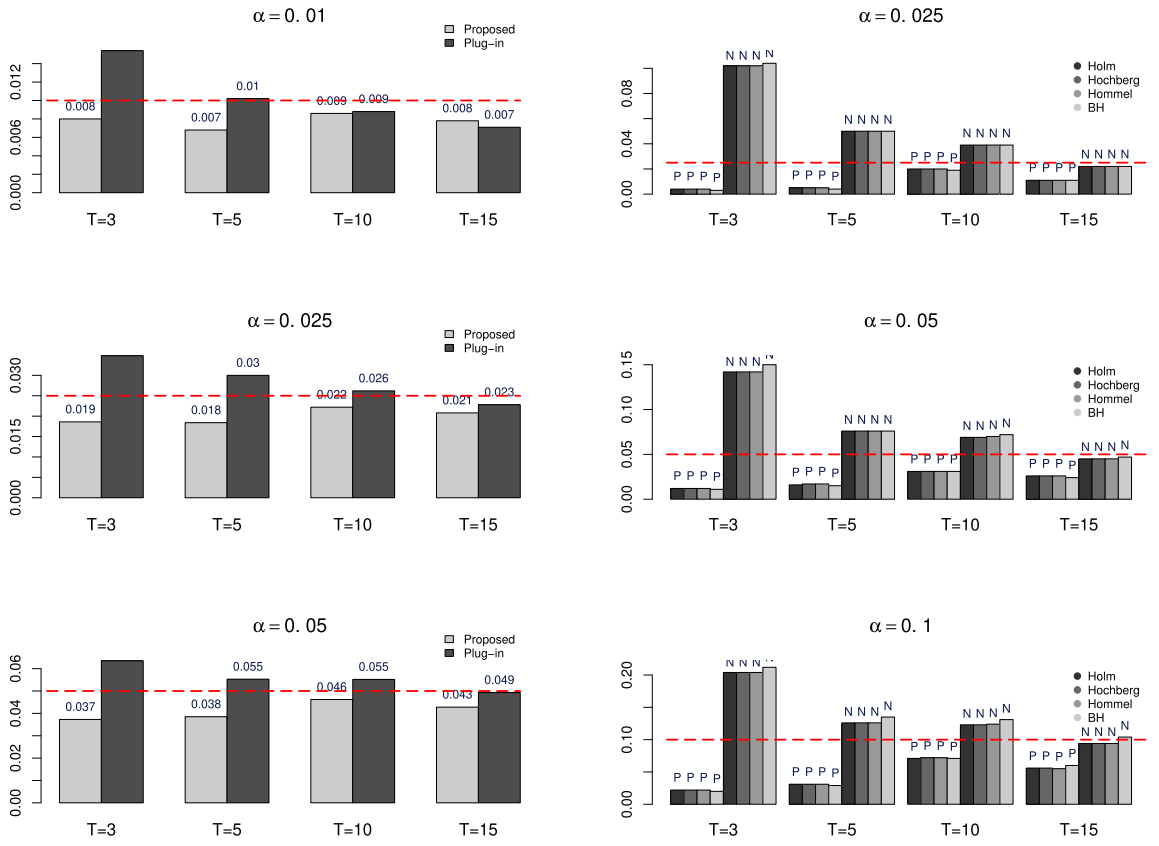
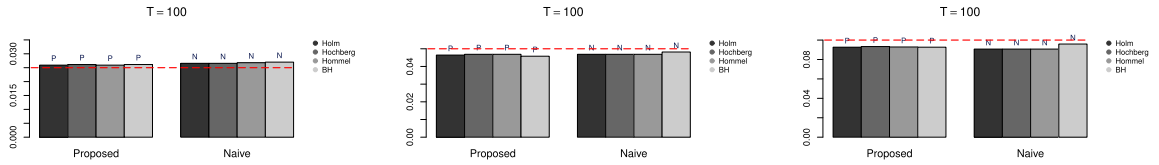



Figure 3: (Left column) Simulation results for the size analysis in the single-account case. The red dashed lines are the desired $\alpha$-level. (Right column) Simulation results from 1,000 data sets in the multiple-account case. The letters "P" and "N" are short for "Proposed" and "Naive" respectively, indicating the proposed methods and naive plug-in methods. Multiple testing corrections are distinguished by colors, as shown in the legend. The red dashed lines are the desired FWER.




(a) $\alpha = 0.025$ (b) $\alpha = 0.05$ (c) $\alpha = 0.1$

Figure 4: Simulation results for size convergence from 4,629 data sets in the multiple-account case for $T = 100$. The letters "P" and "N" are short for "Proposed" and "Naive", respectively, indicating the proposed methods and naive plug-in methods. Multiplicity corrections are distinguished by colors, as shown in the legend. The red dashed lines are the desired FWER.

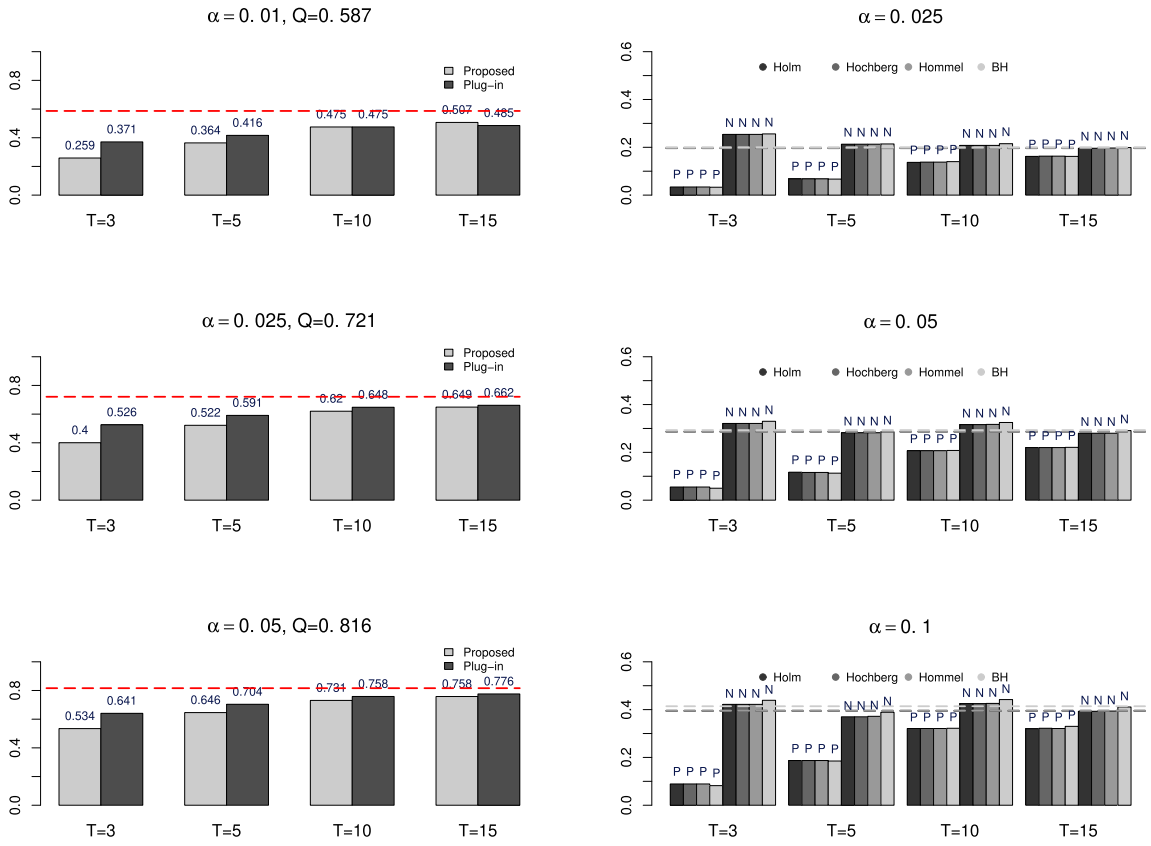



Figure 5: (Left column) Simulation results in the single-account case for power analysis. The red dashed lines indicate the corresponding theoretical power when the true $\boldsymbol{\beta}$ and $\sigma^2$ were used, computed by Equation (11). (Right column) Simulation results in the multiple-account case for power analysis. The letters "P" and "N" are short for "Proposed" and "Naive", respectively, indicating the proposed methods and naive plug-in methods. The multiple testing corrections are distinguished by colors, as shown in the legend. True power is marked as a dashed line for each method, computed through Monte Carlo simulations of 10,000 iterations.

## 5.2 Power Analysis

The left column of Figure 5 contains three bar plots comparing the estimated statistical power using the proposed method to that of plug-in estimates. The red lines indicate the theoretical power for the corresponding configuration, computed by $1 - \Phi(z_{1-\alpha} - (c-1)\|\boldsymbol{\beta}_0\|/\sigma)$, where $z_\alpha$ is the $\alpha$-th quantile of the standard normal distribution. The statistical power is by and large comparable except when $T$ is low where the plug-in estimate yields moderately higher statistical power than the proposed method. However, this comes at the expense of allowing exceedingly high type-1 error as demonstrated in Figure 3. The gains in statistical power are not worth relinquishing control over type-1 error, considering the differences in type-1 error and statistical power.

In the multiple-account case, the right column of Figure 5 shows the statistical power of the proposed methods and plug-in methods labeled as "P" and "N", respectively. Each multiple testing correction is distinguished by a different color. Despite the lack of a closed-form expression

of the true power, it can be computed with Monte Carlo simulations, i.e.,

$$\frac{1}{S}\sum_{s=1}^{S} \text{any}\left(h(U_1, \ldots, U_{K-3}, V_{10}, V_{11}, V_{12}) < \alpha\right) \to 1-\beta, \tag{8}$$

as $S \to \infty$, where $U_k \overset{\text{IID}}{\sim} U(0,1)$, $V_k = 1 - \Phi\left(Z_k + (c-1)\sqrt{\boldsymbol{\beta}_k^\top \Sigma^{-1} \boldsymbol{\beta}_k}\right)$ and $Z_k \overset{\text{IID}}{\sim} N(0,1)$ for $k = 10, 11, 12$, $\beta$ is the type-2 error with respect to the grand hypotheses, and $h$ is the multiple testing correction chosen from {Holm, Hochberg, Hommel, BH} which yields the adjusted $p$-values. Considering the true power computed through Monte Carlo simulations is marked as dashed lines in all three panels, it is easily observed that the naive plug-in methods outperform the Neyman-Pearson test which is supposed to be the most powerful. Despite its ostensible—yet misleading—outperformance, this is in fact due to the uncontrolled FWER, as shown in Figure 3. In the context of energy use monitoring, this indicates that the plug-in methods set off false alarms twice as many times as the acceptable rate specified by the type-1 error, incurring wasteful human inspection. Even with 15 cycles of historical data, the plug-in methods are still overpowered. On the other hand, the proposed procedure produces statistical power that falls short of that of the Neyman-Pearson test; however, it catches up reasonably quickly with 10 cycles of historical data, with which the naive plug-in procedures are still overpowered.

## 6 Real Data Analysis

In this section, we apply the proposed hybrid energy monitoring algorithm detailed in Section 4 to UConn CNG data described in Section 2. As briefly mentioned in Section 2, it is not optimal to flag an already expected surge in natural gas use as abnormal in cases where the building itself is large or the temperature for a particular month drops unusually. To that end, prior to analysis, we normalize each observation into consumption for a 100 square-foot building for a 30-day month per degree day (as in Equation (1)). For buildings with more than one meter, the square footage is divided by the number of meters.

Facilities Operations collects daily weather data from the National Oceanic and Atmospheric Administration (NOAA) through the web application programming interface (API) provided by the Applied Climate Information System. Of all available weather information, the heating degree days and cooling degree days defined as $\text{HDD} = \max(0, 65 - \bar{u})$ and $\text{CDD} = \max(0, \bar{u} - 65)$ are used to encapsulate the weather on a given day, where $\bar{u}$ is the average of the maximum and minimum temperatures of a day, and 65 degrees Fahrenheit is the neutral balance point. Thus, the observations are further divided by the sum of degree days to carve out the uneven impact of the weather on the natural gas usage. We set the tuning parameters to $B_1 = 1000$, $B_2 = 2000$, and $B_3 = 2000$, and run our procedure for the first four months of a calendar year (January, February, March, and April) of 2008 through 2016. Figure 7 contains the histograms of $\mathbf{p}$ of the proposed procedure for four multiplicity corrections: Holm, Hochberg, Hommel, and BH. The solid vertical line in each panel indicates the upper $\alpha$-th quantile of the empirical distribution $\mathbf{p}$, the exact value of which is written by the line with the Greek letter $\gamma_\alpha$. And the dashed line is the corresponding observed $p$-value of the test data, next to which the corresponding value is written as $\widehat{p}_{\text{obs}}$. All four multiplicity-correction methods agree that there exists an anomalous account in the first four months of 2017.

Upon the grand null hypothesis being rejected, individual adjusted $p$-values have been examined to locate the possible anomalous surge in natural gas usage. It is beneficial to check each

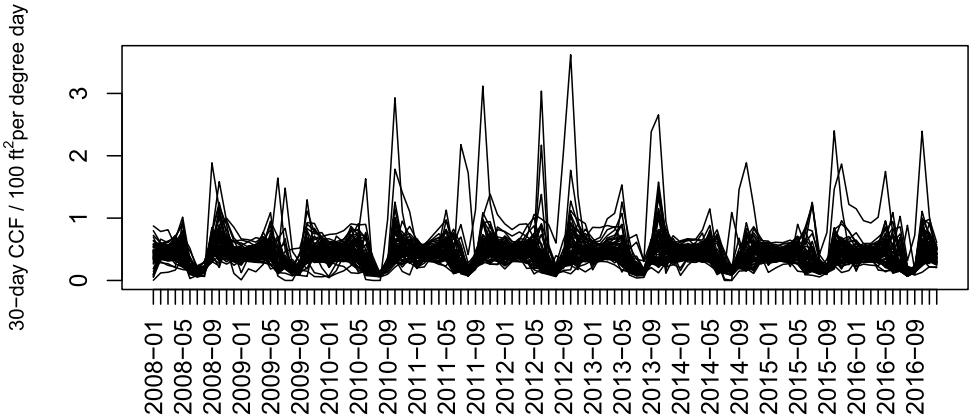



Figure 6: A line plot of the degree-day-adjusted data from UConn residential buildings.

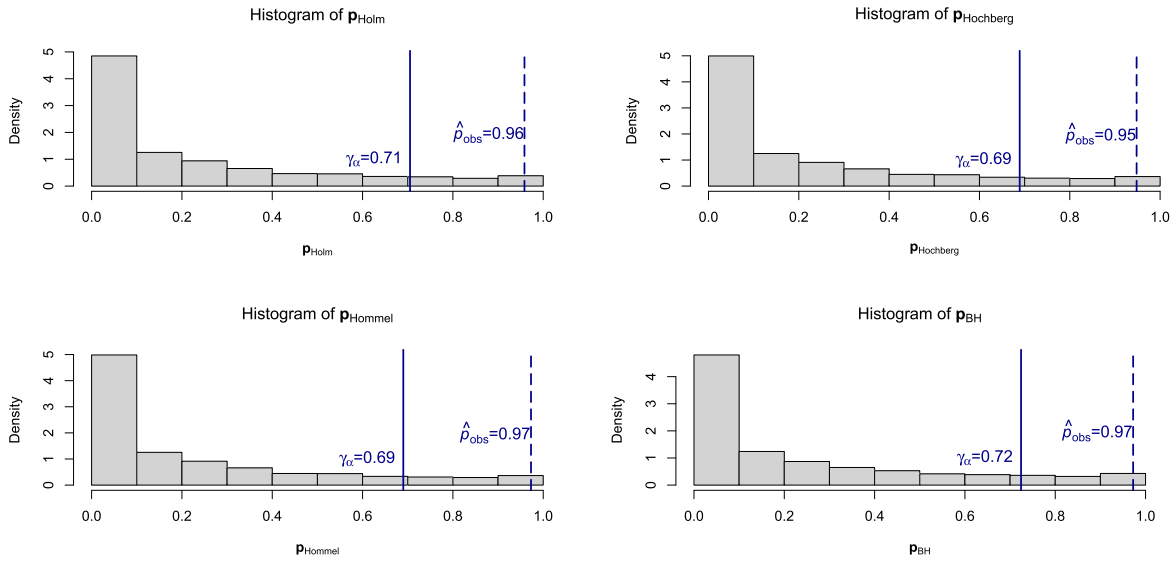



Figure 7: The histograms of constructed distributions of $p$-values regarding the grand null. Only the first four months were selected to be tested—January, February, March, and April.

adjusted $p$-value as the algorithm tends to drive up the critical value for the grand hypothesis test, which makes it more difficult to reject than in each individual case. This is because it requires just one $\tilde{p}_k$ smaller than $\alpha$ (line 11 of Algorithm 2) for $\delta_{b_1}$ to be counted as one, whereas the account-specific equivalent of $\delta_{b_1}$ may remain zero. All four correction methods agreed on $k = 45$, Apartment Building 45. All methods agreed upon $k = 41$, Apartment Building 41, being anomalous as well. Figure 8 contains the time plots for Apartment Building 41 and Apartment Building 45. The top two panels display the unadjusted native use recorded in CCF whereas the bottom two panels show the adjusted usage for 30 days and 100 square feet after each observation is divided by the number of degree days. Observations for each year's first four months (Jan–Apr) are colored red. Focusing on the last four red points that were monitored for anomalous behavior, the unadjusted native use in the top two panels do not particularly deviate from historical patterns. However, once adjusted, it becomes visible that the last cycle stands

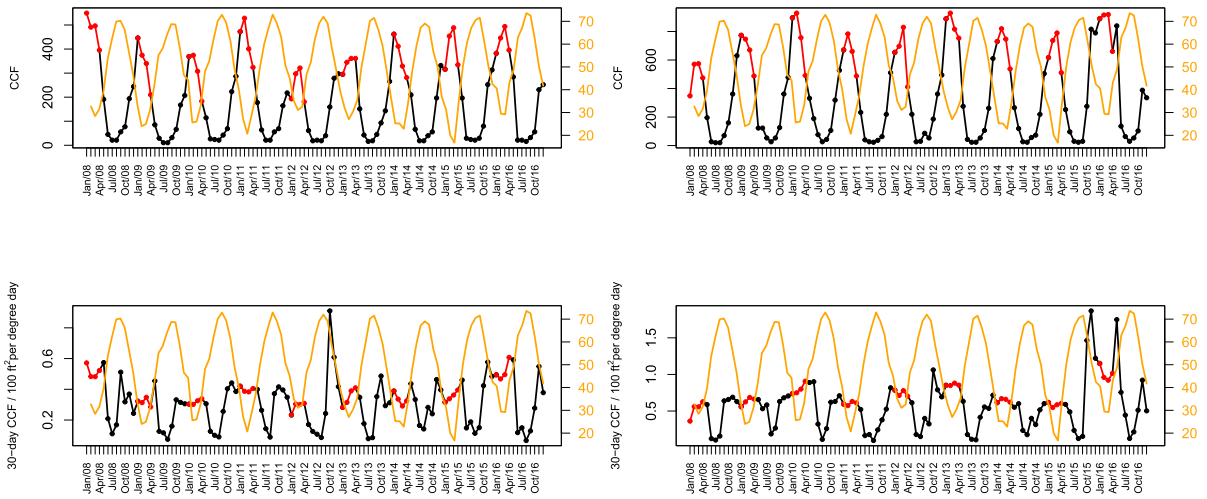



Figure 8: Two buildings detected as anomalous according to the proposed procedure. The top two panels show the (unadjusted) native use in CCF whereas the bottom two panels are adjusted for the number of days in a billing cycle, the square footage of each building, and the degree days. Points marked red are the first four months (Jan–Apr) in each year. The last four red dots were monitored. Orange lines indicate the average monthly temperature in Fahrenheit, whose axis is on the right side of each plot.

out from the other cycles. For Apartment Building 45, in particular, the average native use of the four months in 2016 was 27.5 percent higher than that of the previous year. The monthly average temperature, colored orange in the plots, indicates that the first four months of 2016 were warmer than previous years' months, which defies the expectation that gas usage would go down with warm weather. There are no cut-and-dried explanations for the anomalies discovered. However, there are a few common scenarios that could have led to such increases in gas usage. First, utilities managers from UConn Facilities Operations have reported that everyone has a different "temperature comfort zone." The residents in 2016 may have had higher-than-normal temperature preferences. Second, utilities managers have also recounted that residents in these buildings frequently forget to shut their windows, especially when they leave their rooms. If these behaviors happened unusually frequently between January and April of 2016 in the two apartment buildings, it could have caused the gas consumption to balloon beyond what is considered normal as defined by the historical data. However, that only two buildings—out of 70—were flagged as anomalous increases in gas usage suggests that the gas usage within UConn's residential buildings is consistent and well-synchronized with the average temperature each month.

To further compare the performance of our proposed algorithm to that of the plug-in scheme, we derive prediction intervals using both methods. For the naive approach, the $100(1-\alpha)\%$ prediction interval is given by $\widehat{\beta}+z_{1-\alpha}\widehat{\sigma}$, whereas for our proposed method, we obtain $\gamma_\alpha$ following Algorithm 1 and compute the upper $\gamma_\alpha$-th quantile of the realizations of $\{\beta^{(s)}+z_{1-\alpha}\sigma^{(s)}\}_{s=1}^{S}$, where $\beta^{(s)}$ and $\sigma^{(s)}$ are the $s$-th realized values from their posterior distribution. We have obtained these prediction intervals for Apartment Building 41 and Apartment Building 45. The results for $1-\alpha = 0.95$ are shown in Figure 9, where blue lines indicate the upper bounds of the 95% posterior predictive interval under the proposed model, and the long-dashed lines mark the

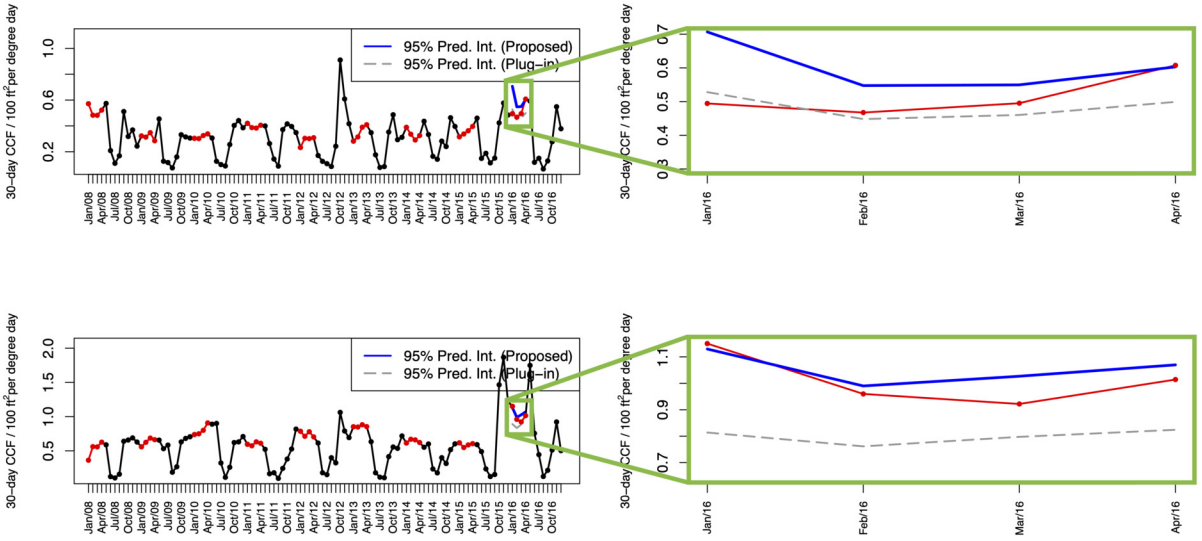



Figure 9: Prediction bands of natural gas consumption in Apartments 41 and 45 from Jan. to Apr., 2016. Blue solid lines indicate the upper bounds of the 95% Bayesian posterior predictive band under the proposed model, and the long-dashed lines mark the upper 95% prediction bound using the naive plug-in approach.

upper 95% prediction bounds using the naive plug-in approach. We see from this figure that our proposed method produces a markedly wider interval than the plug-in scheme as anticipated, since the estimation randomness accounted for in our method guards against overconfidence. These results are consistent with the simulation results in Section 5 where the proposed method was noticeably more conservative in rejecting the null hypothesis.

## 7 Discussions and Conclusion

We proposed a hybrid monitoring algorithm that takes advantage of the posterior distributions of the parameters to inject randomness in order to account for unaddressed uncertainty in statistical anomaly detection. To the best of our knowledge, the proposed algorithm is the first of its kind. From uncertainty calibration's perspective, our general idea in this application bears a resemblance to the posterior or prior predictive $p$-values (Meng, 1994; Hjort et al., 2006; Gelman et al., 1996). The proposed algorithm estimates the unknown parameters using historical data and tests the last cycle by computationally constructing the predictive distribution of the $p$-values, to which the observed $p$-value is compared. We have shown that our method adjusts for the estimation uncertainty and properly controls type-1 error rate.

There are several possible extensions for future research. As more data become available, both incorporating new information and phasing out old information affect the performance of the algorithm. The former can be handled efficiently through an online updating scheme, which significantly alleviates computational burden. This online updating scheme will be useful for high-frequency data where repeated parameter estimation with an increasing amount of data can quickly strain computational resources. The latter can be addressed by retiring old data, which goes back to the question of what the minimum amount of data to achieve a desired level of statistical power is. Furthermore, it is crucial that all past anomalous data points be handled

in a way that they do not undermine subsequent monitoring. Moreover, these fields are by nature highly collaborative, and domain experts play a substantial role in modeling the system correctly. We emphasize the importance of domain expert opinions in handling the detected anomalies. Methodologically, the hypothesis tests can be extended to composite hypotheses such as $H_0 : \boldsymbol{\mu} \geqslant c\boldsymbol{\beta}_{0k}$ versus $H_1 : \boldsymbol{\mu}_k < c\boldsymbol{\beta}_{0k}$ or $H_0 : \boldsymbol{\mu}_k \leqslant c\boldsymbol{\beta}_{0k}$ versus $H_1 : \boldsymbol{\mu}_k > c\boldsymbol{\beta}_{0k}$ with inequalities applied elementwise. By picking arbitrary $c_0$ and $c_1$ for null hypothesis and alternative hypothesis respectively such that either $c_0 > c_1$ or $c_1 > c_0$, the likelihood ratio becomes a monotone function of the data due to the fixed sign of $c_1 - c_0$, which gives a uniformly most powerful test. In this setting, the test statistic is easily generalized as

$$W_k := \frac{\log \Lambda_k - \mathsf{E}_{H_0}(\log \Lambda_k)}{\sqrt{\mathrm{Var}_{H_0}(\log \Lambda_k)}} = \frac{\mathrm{sign}(c_1 - c_0)\boldsymbol{\beta}_{0k}^\top \Sigma^{-1}(\boldsymbol{y}_k - c_0\boldsymbol{\beta}_{0k})}{\sqrt{\boldsymbol{\beta}_{0k}^\top \Sigma^{-1}\boldsymbol{\beta}_{0k}}}. \tag{9}$$

For future research, combining anomalous past data and online updating is currently under investigation. One potential useful method of handling flagged data is assigning weights. The online updating scheme will also need to be adjusted accordingly since deleting past data will disrupt updates. Gradually downweighting past data will automatically address retiring old data.

## Supplementary Material

An R package for our method can be found at https://github.com/daeyounglim/energystuff. This repository contains R functions running our proposed method, an R program for generating simulation data sets, and another R wrapper function simplifying user interface for when running simulations over a large number of data sets.

## Appendix A Posterior Consistency

**Proposition 1.** *Recall that*

$$\widehat{\boldsymbol{\beta}}_k = \overline{\boldsymbol{y}}_{\cdot k}^h = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{y}_{ik}^h, \quad \widehat{\Sigma} = \frac{1}{K(T-1)}\sum_{i=1}^{T}\sum_{k=1}^{K}(\boldsymbol{y}_{ik}^h - \widehat{\boldsymbol{\beta}}_k)(\boldsymbol{y}_{ik}^h - \widehat{\boldsymbol{\beta}}_k)^\top.$$

*Let the estimators of $\boldsymbol{\beta}_{0k}$ and $\Sigma_0$ be indexed by $T$, the number of data points. We omit the subscript $k$ since $\widehat{\boldsymbol{\beta}}_k$'s are conditionally independent given $\Sigma_0$. Denote the posterior probability measure associated with $N(\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}_T, \Sigma/T) \times \mathrm{IW}(\Sigma \mid K(T-1) + d^*, K(T-1)\widehat{\Sigma}_T)$ by $\Pi(\cdot,\cdot)$. Then for any compactly supported function $f$, as $T \to \infty$,*

$$\int f(\boldsymbol{\beta}, \Sigma)\Pi(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\Sigma) \to \int f(\boldsymbol{\beta}, \Sigma)\delta_{\boldsymbol{\beta}_0, \Sigma_0}(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\Sigma) = f(\boldsymbol{\beta}_0, \Sigma_0). \tag{10}$$

*This implies that, as $T \to \infty$, the posterior probability measure weakly converges to the degenerate point mass around the true parameter values, or equivalently posterior consistency.*

## Appendix B Derivation of Test Statistics

By the Neyman-Pearson lemma in the single-account case,

$$\log \Lambda = \log \frac{\phi(\boldsymbol{y} \mid c\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{I}_J)}{\phi(\boldsymbol{y} \mid \boldsymbol{\beta}_0, \sigma^2 \boldsymbol{I}_J)} = \sum_{j=1}^{J} \frac{y_j(c-1)\beta_{0j} - (c^2-1)\boldsymbol{\beta}_{0j}^2/2}{\sigma^2},$$

where $\phi(\cdot \mid \mu, \Sigma)$ is the density of a multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$. Under $H_0$, $\log \Lambda$ follows the following normal distribution:

$$\log \Lambda \sim N\left(-\sum_{j=1}^{J} \frac{(c-1)^2\beta_{0j}^2}{2\sigma^2}, \sum_{j=1}^{J} \frac{(c-1)^2\beta_{0j}^2}{\sigma^2}\right),$$

yielding

$$W := \frac{\log \Lambda - \mathsf{E}_{H_0}(\log \Lambda)}{\sqrt{\mathrm{Var}_{H_0}(\log \Lambda)}} = \frac{\sum_{j=1}^{J}(y_j - \beta_{0j})\beta_{0j}}{\sigma\sqrt{\sum_{j=1}^{J}\beta_{0j}^2}}.$$

Although the rejection region does not depend on $c$, the increment parameter $c$ plays a role in the power function, given by

$$Q(c, \boldsymbol{\beta}_0) = \mathsf{P}_{H_1}\left(\frac{\boldsymbol{\beta}_0^\top \boldsymbol{y} - c\boldsymbol{\beta}_0^\top \boldsymbol{\beta}_0}{\sigma\sqrt{\boldsymbol{\beta}_0^\top \boldsymbol{\beta}_0}} \geqslant z_{1-\alpha} + \frac{1-c}{\sigma}\sqrt{\boldsymbol{\beta}_0^\top \boldsymbol{\beta}_0}\right) = 1 - \Phi\left(z_{1-\alpha} - \frac{c-1}{\sigma}\|\boldsymbol{\beta}_0\|\right), \quad (11)$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution, and $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of the standard normal distribution, i.e., $\Phi(z_{1-\alpha}) = 1-\alpha$. In the multivariate case, the logarithm of the likelihood ratio follows a normal distribution under $H_0$, i.e.,

$$\begin{aligned}\log \Lambda_k &= \log \frac{\phi(\boldsymbol{y}_k \mid \boldsymbol{C}\boldsymbol{\beta}_{0k}, \Sigma)}{\phi(\boldsymbol{y}_k \mid \boldsymbol{\beta}_{0k}, \Sigma)} = -\frac{1}{2}\left\{\boldsymbol{\beta}_{0k}^\top(\boldsymbol{C}\Sigma^{-1}\boldsymbol{C} - \Sigma^{-1})\boldsymbol{\beta}_{0k} - 2\boldsymbol{\beta}_{0k}^\top(\boldsymbol{C} - \boldsymbol{I}_J)\Sigma^{-1}\boldsymbol{y}_k\right\} \\ &\sim N\left(-\frac{1}{2}\boldsymbol{\beta}_{0k}^\top\left\{\boldsymbol{C}\Sigma^{-1}\boldsymbol{C} - \Sigma^{-1} - 2(\boldsymbol{C} - \boldsymbol{I}_J)\Sigma^{-1}\right\}\boldsymbol{\beta}_{0k}, \boldsymbol{\beta}_{0k}^\top(\boldsymbol{C} - \boldsymbol{I}_J)\Sigma^{-1}(\boldsymbol{C} - \boldsymbol{I}_J)\boldsymbol{\beta}_{0k}\right),\end{aligned}$$

which gives

$$\frac{\log \Lambda_k - \mathsf{E}_{H_0}(\log \Lambda_k)}{\sqrt{\mathrm{Var}_{H_0}(\log \Lambda_k)}} = \frac{\boldsymbol{\beta}_{0k}^\top(\boldsymbol{C} - \boldsymbol{I}_J)\Sigma^{-1}(\boldsymbol{y}_k - \boldsymbol{\beta}_{0k})}{\sqrt{\boldsymbol{\beta}_{0k}^\top(\boldsymbol{C} - \boldsymbol{I}_J)\Sigma^{-1}(\boldsymbol{C} - \boldsymbol{I}_J)\boldsymbol{\beta}_{0k}}} \sim N(0, 1). \quad (12)$$

Assuming $c = c_1 = \cdots = c_J$, $\boldsymbol{C} - \boldsymbol{I}_J$ is reduced to a scalar, $c - 1$, and disappears from Equation (12), yielding

$$W_k := \frac{\log \Lambda_k - \mathsf{E}_{H_0}(\log \Lambda_k)}{\sqrt{\mathrm{Var}_{H_0}(\log \Lambda_k)}} = \frac{\boldsymbol{\beta}_{0k}^\top \Sigma^{-1}(\boldsymbol{y}_k - \boldsymbol{\beta}_{0k})}{\sqrt{\boldsymbol{\beta}_{0k}^\top \Sigma^{-1}\boldsymbol{\beta}_{0k}}} \sim N(0, 1) \quad \text{under } H_0.$$

## Acknowledgement

We would like to thank the editor, the associate editor, and the two anonymous reviewers for their invaluable comments and useful suggestions, which improved the quality of our paper.

## References

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300.

Capehart BL, Turner WC, Kennedy WJ (2020). *Guide to Energy Management*. River Publishers.

Casella G, Berger RL (2002). *Statistical Inference*. Cengage Learning, 2nd edition.

Cormen TH, Leiserson CE, Rivest RL, Stein C (2022). *Introduction to Algorithms*. The MIT Press, 4th edition.

Doty S, Turner WC (2004). *Energy Management Handbook*. CRC Press.

Dunn OJ (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293): 52–64.

Eddelbuettel D, Balamuta JJ (2018). Extending R with C++: A brief introduction to Rcpp. *The American Statistician*, 72(1): 28–36.

Eddelbuettel D, Sanderson C (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71: 1054–1063.

Efron B, Tibshirani RJ (1994). *An Introduction to the Bootstrap*. CRC Press.

Fu Y, Jeske DR (2014). SPC methods for nonstationary correlated count data with application to network surveillance. *Applied Stochastic Models in Business and Industry*, 30(6): 708–722.

Geisser S, Cornfield J (1963). Posterior distributions for multivariate normal parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(2): 368–376.

Gelman A, Meng XL, Stern H (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4): 733–760.

Hjort NL, Dahl FA, Steinbakk GH (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, 101(475): 1157–1174.

Hochberg Y (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4): 800–802.

Holm S (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70.

Hommel G (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2): 383–386.

Jeffreys H (1998). *The Theory of Probability*. OUP Oxford.

Meng XL (1994). Posterior predictive *p*-values. *The Annals of Statistics*, 22(3): 1142–1160.

Neyman J, Pearson ES (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694–706): 289–337.

OpenMP Architecture Review Board (2018). OpenMP application programming interface version 5.0.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery AE, Akman V (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika*, 73(1): 85–89.

Rao CR (1973). *Linear Statistical Inference and Its Applications*. John Wiley & Sons.

Rashid H, Singh P (2018). Monitor: An abnormality detection approach in buildings energy consumption. In: *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, 16–25. IEEE.

Ravishanker N, Chi Z, Dey DK (2022). *A First Course in Linear Model Theory*. Chapman and

Hall/CRC, 2nd edition.

Ross GJ, Tasoulis DK, Adams NM (2011). Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, 53(4): 379–389.

Ross GJ, Tasoulis DK, Adams NM (2013). Sequential monitoring of a Bernoulli sequence when the pre-change parameter is unknown. *Computational Statistics*, 28(2): 463–479.

Seem JE (2007). Using intelligent data analysis to detect abnormal energy consumption in buildings. *Energy and Buildings*, 39(1): 52–58.

Šidák Z (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318): 626–633.

Simes RJ (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3): 751–754.

Sun D, Berger JO (2007). Objective Bayesian analysis for the multivariate normal model. *Bayesian Statistics*, 8: 525–562.

University of Michigan (2011). Final Report: Assessing a Campus Energy Monitoring System. http://graham.umich.edu/media/files/campus-course-reports/CEMS%20Final%20Report.pdf. Accessed: 2021-10-25.

Worcester Polytechnic Institute (2007). Monitoring Electricity Consumption on the WPI Campus. The Reduction of Carbon Emissions Through the Implementation of Energy Information Tracking Technology. https://web.wpi.edu/Pubs/E-project/Available/E-project-060107-130245/unrestricted/iqpfinaldraft.pdf. Accessed: 2021-10-25.

Wright SP (1992). Adjusted p-values for simultaneous inference. *Biometrics*, 48(4): 1005–1013.

Zhang J, Paschalidis IC (2018). Statistical anomaly detection via composite hypothesis testing for Markov models. *IEEE Transactions on Signal Processing*, 66(3): 589–602.

Zhao L (2014). A novel method for detecting abnormal energy data in building energy monitoring system. *Journal of Energy*, 2014: 231571.