Time Series that are Robust to Regime Changes

Dimitris Bertsimas
Sloan School of Management and Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Ivan Paskov
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Abstract

Time series exhibit regime changes over time, a phenomenon commonly encountered in fields such as economics, ecology, medicine, finance, and many others. In contrast to previous approaches which require strong assumptions on the regime switching process, or are designed to perform well after the regime shift has already occurred, we propose a methodology for constructing a single model that is explicitly designed to be robust against regime changes. The approach is non-parametric, requiring no description of a regime switching process, and most importantly, is inherently forward looking, in that the model is designed to smoothly pass through the break point between two regimes, without knowing of when the aforementioned break point will occur. Through experiments, we demonstrate that these stable models perform significantly better than their non-stable counterparts during periods of regime shift, with the magnitude of the improvement being inversely proportional to the frequency with which the models are refit. Just as important, we show that this increased robustness does not come at the price of performance during stationary periods. Finally, these stable time series models are highly efficient, and scale to data sets of essentially any desired size.

Keywords: stability, randomization, optimization, time series, robustness, regime change, interpretability

1. Introduction

Time series exhibit regime changes over time, a phenomenon commonly encountered in fields such as economics, ecology, medicine, finance, and many others. Traditional approaches to this problem involve the construction of multiple models and some mechanism of switching between the models, either via having an explicit model for the dynamics of the regime
switching process, or via statistical hypothesis tests which attempt to detect when a regime shift has occurred. The practical importance of either approach, however, has been limited. In the case of the former, it requires a detailed understanding of the dynamics of the regime switching process, which unfortunately is almost never available in practice. And in the case of the latter, it is inherently an “after the fact” form of analysis, conducted after a regime switch has already occurred.

In this paper, we propose a methodology for constructing a single model that is explicitly designed to be robust against regime changes. The approach is non-parametric, requiring no description of a regime switching process, and most importantly, is inherently forward looking, in that the model is designed to smoothly pass through the break point between two regimes, without knowing of when the aforementioned break point will occur. Through experiments, we demonstrate that these stable models perform significantly better than their non-stable counterparts during periods of regime shift, with the magnitude of the improvement being inversely proportional to the frequency with which the models are refit. Just as important, we show that this increased robustness does not come at the price of performance during stationary periods. Finally, these stable time series models are highly efficient, and scale to data sets of essentially any desired size.

1.1 Literature

One major school of thought for how to deal with regime changes is to first detect that a regime change has in fact occurred, and to then refit the model towards the new regime. The first step in this process, i.e., the detection of a regime change, has a few different variants. The first type is parametric, where a probability distribution is posited to model the data, and then classical methods such as the t-test are used to detect the break point, see Ducrè-Robitaille et al. (2003) for more information. There are also Bayesian variants of this analysis, where in addition to assuming a probability model for the data, one also specifies prior distributions for the parameters in that model (i.e., such as the mean of the observations before and after the regime shift), see Perreault et al. (2000) and Chu and Zhao (2004) for more information. A second variant is non-parametric, i.e., it assumes no probability model, and instead uses non-parametric hypothesis tests such as the Mann–Whitney U-test or the Wilcoxon rank sum test to detect the break point, see Mauget (2003) and Karl and Williams (1987) for more information. A third variant uses regression analysis coupled with binary indicator variables and interaction terms with the time variable to account for changes in the mean and slope before and after the regime change, see Lund and Reeves (2002) and Solow (1987) for more information. There are many other variants as well, see Lund and Reeves (2002) for more information, but fundamentally all of them first wait for the break point to occur, and then retrospectively analyze the data to attempt to detect the presence or absence of a regime shift before taking corrective action. The limitation of this style of analysis is that it is “after the fact” and hence of limited utility for anyone looking forward and attempting to avoid the potentially significant effects of a regime shift. To address this limitation, there is a second school of thought for dealing with regime changes that posits an explicit model for the dynamics of the regime switching process, which is
then integrated amongst multiple times series models, one per epoch of coherent behavior. These models are all generally of the form:

\[ X_t = a_0^{(J_t)} + \sum_{i=1}^{p} a_0^{(J_t)} X_{t-i} + b^{(J_t)} \epsilon_t, \]

where \( \epsilon_t \) are i.i.d. \( N(0, \sigma^2) \) and \( \{J_t\} \) is an indicator time series taking values in \( [J] = \{1, \ldots, J\} \) and which acts as the switching mechanism. The various models then differ in the assumptions they make on the \( \{J_t\} \) stochastic process. The simplest assumption to make is that \( \{J_t\} \) are independent and identically distributed according to some distribution. This gives rise to the so called Exponential Autoregressive-Moving Average EARMA \((p, q)\) Process, see Lawrance and Lewis (1980) for more information. An alternative assumption is that \( \{J_t\} \) is a finite-state Markov chain, giving the so called Markov Switching Model, see Tong and Lim (1980) and Hamilton (1989) for more information. Another axis along which the assumptions about \( \{J_t\} \) differ is whether \( \{J_t\} \) is observable or hidden. There are even various partially hidden extensions, the so called fuzzy extension models or partially-hidden switching models, as described in Wu and Chen (2007), as well as many other variants, see Pinto and Spezia (2016), Dong-Mei et al. (2017), and Elliott et al. (2018) for more information. In all these variants, the parameters are estimated via Maximum Likelihood estimation. While these class of models are an improvement over the previous class, they still possess various limitations such as ease of use, interpretability, and perhaps most critically of all, a strong reliance upon accurately pinning down the true underlying dynamics.

1.2 Contributions and Structure

In this paper, we propose a new methodology for constructing time series models that are robust to regime changes and use the term stable time series to describe them. Through experiments, we demonstrate that these stable models perform significantly better than their non-stable counterparts during periods of regime shift. Just as important, we show that this increased robustness does not come at the price of performance during stationary periods. Finally, these stable time series models are highly efficient, and scale to data sets of essentially any desired size.

The structure of the paper is as follows. In Section 2, we introduce the stable time series methodology. In Section 3, we discuss how to efficiently compute stable time series solutions. In Section 4, we present computational results comparing stable time series to regular time series for multiple data sets, at multiple frequencies, across a variety of refitting schedules, all over both a turbulent period as well as a calm period. In Section 5, we discuss our results and in Section 6, we report our conclusions.
2. The Stable Time Series Methodology

In this section, we briefly review popular time series models before introducing the stable time series methodology.

2.1 Traditional Time Series Models

One of the most widely used time series models is the Autoregressive Model of Order $p$, denoted $AR(p)$. It posits that the response variable depends linearly on its own previous values and on a stochastic term, i.e.,

$$X_t = \gamma + \sum_{i=1}^{p} \xi_i X_{t-i} + \epsilon_t,$$  \hspace{1cm} (1)

where $\xi_1, \ldots, \xi_p$ are the parameters of the model, $\gamma$ is a constant, and $\epsilon_t$ is white noise. A popular way to estimate the coefficients $\xi_1, \ldots, \xi_p$ is via OLS regression, see Hamilton (1994) for more information. Moreover, the order of the model, $p$, is typically either selected via the Akaike information criterion (AIC), Bayesian Information Criterion (BIC), or more directly via a validation set approach, see Hamilton (1994) for more information. It is also useful to point out that there are straightforward generalizations to this model, the most popular being the Autoregressive Integrated Moving Average model of order $(p, d, q)$, denoted $ARIMA(p, d, q)$. The moving average component of this model refers to positing that the response depends in a linear fashion on the lagged error terms. As these are not observable in practice, this paper will proceed as in Tong and Lim (1980) with the commonly applied assumption that $q = 0$, see also Hamilton (1994) for more information. The integrated component of this model corresponds to situations where the data shows evidence of non-stationarity, and so an initial differencing of the data is performed $d$ times to eliminate this non-stationarity. In what follows, we will always assume that the data has been appropriately differenced so we are working with the resulting stationary “residual process”, and hence can without loss of generality simply talk about Autoregressive Model of Order $p$, denoted $AR(p)$.

2.2 The Stable Time Series Model

While Eq. (1) will clearly result in a model that performs reasonably well in environments similar to the one it was trained for, there is no reason to believe it will perform well under regime changes. In fact, we might even expect it to do quite poorly as the model is literally optimized precisely for the environment it was trained for, and no other. This property makes it fragile.

To make the time series model described in Eq. (1) more robust, we apply a variant of the central idea in Ensemble Learning: namely, rather than constructing multiple separate models and then averaging them, we construct multiple training sets, and require our final
time series model to be capable of performing well across all of them. Where will these multiple training sets come from? They will be generated synthetically from our original data, and then trained against simultaneously using robust optimization. More precisely, this idea can be formulated mathematically as:

\[
\min_{\xi, \gamma} \max_{z \in \mathcal{Z}} \sum_{t=p+1}^{n} z_t (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2.
\] (2)

The interpretation of Eq. (2) is that every vector \( z \in \mathcal{Z} \) corresponds to one synthetic data set. Then, each proposed model, which is parameterized by the pair \((\xi, \gamma)\), is exposed to all training sets \( z \) within \( \mathcal{Z} \), and that model which does the best simultaneously across all the synthetic training sets is finally chosen.

As an example, suppose \( \{x_t\} \) represents a time series of the S&P 500 over the past 50 years. Contained in that stretch of time would be a subsection, \( \{x_{l_1}, \ldots, x_{l_1+r_1}\} \) corresponding to the 2008 financial crisis. Also contained in that stretch of time would be a subsection, \( \{x_{l_2}, \ldots, x_{l_2+r_2}\} \) corresponding to the 1998 Russian default crisis, as well as many other such subsections corresponding to both “turbulent” times as well as “calm” times. Semantically then, what Eq. (2) is encoding is the following request: find a model that would have done well during the 2008 financial crisis, as well as the 1998 Russian default crisis, as well as many other such remarkable (and also unremarkable) scenarios. By design, the final model returned by performing the optimization described in Eq. (2) would be one that is robust to many different kinds of regimes, as well as the transitions between them.

What remains to be specified is the structure of the uncertainty set \( \mathcal{Z} \). One natural choice is all subsets of size \( k \):

\[
\mathcal{Z}_1 = \left\{ z : \sum_{i=1}^{n} z_i = k, \ z_i \in \{0, 1\}, \ i \in [n] \right\}.
\]

At an optimal solution of (2), each \( z_i \) will be equal to either 0 or 1, with the interpretation that if \( z_i = 1 \), then point \((x_i, y_i)\) is assigned to the training set, otherwise it is assigned to the validation set. The number \( k \) indicates the desired proportion between the size of the training and validation sets. Namely, by setting \( k = 0.7n \) we recover the typical 70/30 training/validation split and by setting \( k = 0.5n \) we recover the 50/50 training/validation split, etc.

While seemingly natural, the above formulation does not explicitly capture the contiguous structure present within time series data. And as the S&P 500 example above demonstrates, this is likely important. To amend this, we can instead use the following generalized uncer-
tainty set:

\[ Z_2 = \left\{ (z, s) \left| \begin{array}{c}
\sum_{i=1}^{n} s_i = k, s_t \leq z_{i+r}, \forall i, r \in [h-1], \\
\sum_{i=1}^{n} s_i = kh, \\
\sum_{v=0}^{h-1} s_{i+v} \leq 1, i \in [n-h+1], \\
z_i \in \{0,1\}, i \in [n], s_i \in \{0,1\}, i \in [n-h]\end{array} \right. \} \]

the interpretation being that the above uncertainty set chooses \( k \) “anchor points” in time, denoted by \( s_i \), and then takes the \( h - 1 \) consecutive points after each anchor point. Thus each synthetic data set contained within \( Z_2 \) is made up of \( k \) consecutive time periods, each of length \( h \). Note that when \( h = 1 \), that we recover the original proposed uncertainty set, \( Z_1 \), and hence \( Z_2 \) can be viewed as a generalization of \( Z_1 \).

2.3 Visualizing the Model

To better visualize how exactly Eq. (2) first generates many synthetic data sets out of the original data set, and then builds a model that performs well across all of them, consider Figure 1.

In Figure (1a) we have a plot of the original time series. In Figure (1b) we have a plot of 10 randomly generated time series from \( Z_1 \) where \( k = 0.7n \). The optimization problem (2) then amounts to finding a time series model that would have done well across all the “synthetic time series” contained in \( Z_1 \), ten of which are visualized in Figure (1b). Note that if all time series from \( Z_1 \) were visualized, we could identify as a subset those that would also be present as time series from \( Z_2 \) with appropriately chosen parameters. Indeed, they would exactly correspond to those series consisting of contiguous segments from the original time series.

3. Computing Stable Time Series Solutions

In this section, we describe how to compute stable solutions. We first do so using the formulation involving the uncertainty set \( Z_2 \), and then proceed to do the same with \( Z_1 \). While it is true that any solution using \( Z_1 \) can be derived from a solution to \( Z_2 \) with appropriately chosen parameters, as we will see, there are several benefits, both from a computational perspective as well as an interpretability perspective to having a specialized procedure for the \( Z_1 \) formulation.
Figure 1: In Figure (1a), we have a plot of the original time series. In Figure (1b) we have a plot of 10 randomly generated time series from $\mathcal{Z}$ where $k = 0.7n$. 
3.1 Tractable Robust Counterpart for $Z_2$ Formulation

Our approach will consist of first reformulating Eq. (2) into a form whose integer relaxation is tight, and then leveraging techniques from the field of robust optimization to solve the resulting problem.

3.1.1 Tight Integer Relaxation

The first step in the integer reformulation entails generalizing Eq. (2) to allow for consecutive sequences of potentially unequal lengths and by moving this constraint from the constraint set into the objective. Specifically, we accomplish this using a fusion regularizer, see Tibshirani et al. (2005) for more information:

$$\min_{\xi, \gamma} \max_{z \in \mathcal{Z}} \sum_{t=p+1}^{n} z_t(x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2 - \lambda \sum_{t=p+1}^{n-1} |z_{t+1} - z_t|$$

where $\mathcal{Z} = \left\{ z : \sum_{i=p+1}^{n} z_i = k, \, z_i \in \{0, 1\}, \, i \in [p+1, n-1] \right\}$. The interpretation of Eq. (3) is that $\lambda$ modulates the degree of “pressure” exerted upon $z_{t+1}$ and $z_t$ to take on the same value. In the limit as $\lambda \to \infty$, the solution returned is one of a consecutive sequence of $z_i$'s all taking the value 1 of length $n - p$. When $\lambda = 0$, the solution has the freedom to return $n - p$ singleton $z_i$'s equal to one having no relationship to one another (which if you recall, is exactly what $Z_1$ accomplishes). Finally $\lambda$ between 0 and $\infty$ traces out multiple sequences of varying lengths of $z_i$'s equal to one. In the sense described above, Eq. (3) is a generalization of Eq. (2) paired with $Z_2$, as the consecutive sequences returned are no longer constrained to all have the same length. Recalling our S&P 500 example, this is a very desirable property as, for example, the 2008 financial crisis occurred over a longer period of time than did the COVID-19 financial crisis, and it is important our model has the flexibility to capture this variability. We next linearize the absolute value introduced in Eq. (3) as well as move the constraint, $\sum_{i=p+1}^{n} z_i = k$, from the constraint set into the objective, yielding:

$$\min_{\xi, \gamma, z, v_1, v_2} \max_{z, v_1, v_2 \in \mathcal{Z}} \sum_{t=p+1}^{n} z_t(x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2 - \lambda \sum_{t=p+1}^{n-1} (v_{1t} + v_{2t}) - \Gamma \sum_{t=p+1}^{n} z_t$$

where:

$$\mathcal{Z} = \left\{ (z, v_1, v_2) \bigg| \begin{array}{l} v_{1t} \geq z_{t+1} - z_t, \, v_{2t} \geq z_t - z_{t+1}, \, i \in [p+1, n-1], \\
0 \leq v_{1t}, v_{2t} \leq 1, \, z_i \in \{0, 1\}, \, i \in [p+1, n] \end{array} \right\}.$$

Noting that for a fixed $(\xi, \gamma)$, the inner maximization problem is a binary linear optimization problem. Thus, instead of optimizing over $\mathcal{Z}$ we can optimize over the convex hull of $\mathcal{Z}$,
denoted by conv(\(Z\)). Hence, Problem (4) is equivalent to:

\[
\begin{align*}
\min_{\xi, \gamma} & \max_{z, v_1, v_2 \in \text{conv}(Z)} \sum_{t=p+1}^{n} z_t (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2 - \lambda \sum_{t=p+1}^{n-1} (v_{1t} + v_{2t}) - \Gamma \sum_{t=p+1}^{n} z_t,
\end{align*}
\]

(5)

We next characterize the set \(\text{conv}(Z)\).

**Theorem 1** We have

\[
\text{conv}(Z) = \left\{ (z, v_1, v_2) \mid v_{1t} \geq z_{t+1} - z_t, \ v_{2t} \geq z_t - z_{t+1}, \ t \in [p + 1, n - 1], \ 0 \leq v_{1t}, v_{2t} \leq 1, \ 0 \leq z_t \leq 1, \ t \in [p + 1, n] \right\}.
\]

(6)

**Proof** From Hoffman and Kruskal (1956) we have that if a matrix \(A\) is totally unimodular, and \(b\) is integral, then linear optimization problems of the form \(Ax \leq b\) have integral optima, for any \(c\). Equation (4) is indeed of this form, the only item requiring further justification is that our constraint matrix is totally unimodular. To demonstrate this, recall the following theorem from Ghouila-Houri (1962): a matrix \(A\) is totally unimodular if and only if the matrices \(A^T, -A, (A, I)\), and \((A, 0)\) are totally unimodular. By recursively applying this theorem to our constraint set, our task reduces to showing that the matrix encoding the constraints:

\[
v_{1t} \geq z_{t+1} - z_t, \ v_{2t} \geq z_t - z_{t+1}, \ i \in [p + 1, n - 1]
\]

is totally unimodular. By the above theorem, it suffices to show that the transpose of the matrix encoding the above constraints is totally unimodular. We can accomplish this by invoking the theorem by Hoffman and Kruskal (1956) which states that a \((0, +1, -1)\) matrix \(A\) is totally unimodular if it contains no more than one +1 and no more than one −1 in each column. Trivially noting that this is the case, we can now conclude that the matrix encoding the constraints in Equation (4) is totally unimodular, allowing us to conclude via the earlier Ghouila-Houri (1962) theorem that \(\text{conv}(Z)\) is given by (6).

\[\blacksquare\]

### 3.1.2 Robust Counterpart

Having succeeded in eliminating the binary constraints from Equation (2), we are now in a position to leverage the fact that our formulation belongs to the class of robust optimization (RO) problems, see Bertsimas et al. (2011) for a review. The two most frequently described methods in the literature for solving such problems are reformulation to a deterministic optimization problem (often called the robust counterpart) or an iterative cutting-plane method, see Bertsimas et al. (2015) for more information. In the case of Eq. (2) an exact reformulation is possible, see Bertsimas and Paskov (2020) and Bertsimas et al. (2020) for more information, and thus we will take that route.
To alleviate the multiplication of variables (i.e., the product of \( z_t \) with \( (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2 - \Gamma \)), we take the linear optimization dual of the inner maximization problem:

\[
\max_{z, v_1, v_2 \in \mathbb{Z}} \sum_{t=p+1}^n z_t \left[ (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2 - \Gamma \right] - \lambda \sum_{t=p+1}^{n-1} (v_{1t} + v_{2t})
\]

by introducing the dual variables:

- \( \{q_t\}_{t=p+1}^{n-1} \) for the set of constraints: \( z_{t+1} - z_t - v_{1t} \leq 0, \quad t \in [p+1, n-1] \),
- \( \{p_t\}_{t=p+1}^{n-1} \) for the set of constraints: \( z_t - z_{t+1} - v_{2t} \leq 0, \quad t \in [p+1, n-1] \),
- \( \{r_t\}_{t=p+1}^n \) for the set of constraints: \( z_t \leq 1, \quad t \in [p+1, n] \),
- \( \{S_{1t}\}_{t=p+1}^{n-1} \) for the set of constraints: \( v_{1t} \leq 1, \quad t \in [p+1, n-1] \),
- \( \{S_{2t}\}_{t=p+1}^{n-1} \) for the set of constraints: \( v_{2t} \leq 1, \quad t \in [p+1, n-1] \),

to arrive at:

\[
\min_{p, q, s_1, s_2} \sum_{t=p+1}^n r_t - \lambda \sum_{t=p+1}^{n-1} (s_{1t} + s_{2t})
\]

s.t.

\[
\begin{align*}
q_{t+1} + p_{t+1} + r_{t+1} & \geq \left[ (x_{t+1} - \gamma - \xi_1 x_{t} - \ldots - \xi_p x_1)^2 - \Gamma \right], \\
q_{t-1} - p_{t-1} + q_t + p_t + r_t & \geq \left[ (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2 - \Gamma \right], \\
q_{n-1} + p_{n-1} + r_n & \geq \left[ (x_n - \gamma - \xi_1 x_{n-1} - \ldots - \xi_p x_{n-p})^2 - \Gamma \right], \\
q_t + s_{1t} & \geq -\lambda, \quad t \in [p+1, n-1], \\
p_t + s_{2t} & \geq \lambda, \quad t \in [p+1, n-1], \\
p, q, s_1, s_2 & \geq 0,
\end{align*}
\]

where \( \mathcal{G} = [p+2, n-1] \). Substituting this minimization problem back into the outer minimization we arrive at the following problem:

\[
\min_{\xi, \gamma, p, q, s_1, s_2} \sum_{t=p+1}^n r_t - \lambda \sum_{t=p+1}^{n-1} (s_{1t} + s_{2t})
\]

s.t.

\[
\begin{align*}
q_{t+1} + p_{t+1} + r_{t+1} & \geq \left[ (x_{t+1} - \gamma - \xi_1 x_{t} - \ldots - \xi_p x_1)^2 - \Gamma \right], \\
q_{t-1} - p_{t-1} + q_t + p_t + r_t & \geq \left[ (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2 - \Gamma \right], \\
q_{n-1} + p_{n-1} + r_n & \geq \left[ (x_n - \gamma - \xi_1 x_{n-1} - \ldots - \xi_p x_{n-p})^2 - \Gamma \right], \\
q_t + s_{1t} & \geq -\lambda, \quad t \in [p+1, n-1], \\
p_t + s_{2t} & \geq \lambda, \quad t \in [p+1, n-1], \\
p, q, s_1, s_2 & \geq 0.
\end{align*}
\]

This is a convex quadratic optimization problem, and hence can be solved by commercial optimization software in very high dimensions.
3.2 Tractable Robust Counterpart for $Z_1$ Formulation

We next describe how to compute stable solutions for the formulation involving the uncertainty set $Z_1$. While it is true that any solution using $Z_1$ can be derived from a solution to $Z_2$ with appropriately chosen parameters (and hence the previous derivation is immediately usable as a tool for solving the formulation involving the uncertainty set $Z_1$), as we will see, there are several benefits, both from a computational perspective as well as an interpretability perspective to having a specialized procedure for the $Z_1$ formulation.

Our overall approach will be similar to one the employed previously, i.e., first reformulating Eq. (2) into a form whose integer relaxation is tight, and then leveraging techniques from the field of robust optimization to solve the resulting problem.

3.2.1 Tight Integer Relaxation

In the case of the uncertainty set $Z_1$, the first step (i.e., reformulation into a form whose integer relaxation is tight) is considerably simpler than before. This is because if we refer to Eq. (2), we observe that the inner maximization problem is linear in $z$, and hence the problem is equivalent to optimizing over the convex hull of $Z$

$$\text{conv}(Z) = \left\{ z : \sum_{i=p+1}^{n} z_i = k, \ 0 \leq z_i \leq 1, \ i \in [p+1,n] \right\}. $$

Thus, Problem (2) is equivalent to

$$\min_{\xi, \gamma} \max_{z \in \text{conv}(Z)} \sum_{t=p+1}^{n} z_t (x_t - \gamma - \xi_1 x_{t-1} - \cdots - \xi_p x_{t-p})^2 \tag{7}$$

$$\text{conv}(Z) = \left\{ z : \sum_{i=p+1}^{n} z_i = k, \ 0 \leq z_i \leq 1, \ i \in [p+1,n] \right\}. $$

3.2.2 Robust Counterpart

Having succeeded in eliminating the binary constraints from Equation (2), we are now in a position to leverage the fact that our formulation given in Eq. (7) belongs to the class of robust optimization (RO) problems, see Bertsimas et al. (2011) for a review. As before, the two most frequently described methods in the literature for solving such problems are reformulation to a deterministic optimization problem (often called the robust counterpart) or an iterative cutting-plane method, see Bertsimas et al. (2015) for more information. In the case of Eq. (7) an exact reformulation is possible, see Bertsimas and Paskov (2020) and Bertsimas et al. (2020) for more information, and thus we will take that route.
To alleviate the multiplication of variables (i.e., the product of $z_t$ with $(x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2$) we take the linear optimization dual of the inner maximization problem

$$\max_{z_t} \sum_{t=p+1}^{n} z_t (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2 \quad \text{s.t.} \quad \sum_{i=p+1}^{n} z_i = k, \quad 0 \leq z_i \leq 1, \quad i \in [p+1, n]$$

by introducing the dual variable $\theta$ for the first constraint and the dual variables $u_i, \ i \in [p+1, n]$ for the second set of constraints to arrive at:

$$\min_{\theta, u_i} k\theta + \sum_{i=p+1}^{n} u_i \quad \text{s.t.} \quad \theta + u_i \geq (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2, \quad u_i \geq 0, \quad i \in [p+1, n].$$

Substituting this minimization problem back into the outer minimization we arrive at the following problem:

$$\min_{\xi, \gamma, \theta, u_i} k\theta + \sum_{i=p+1}^{n} u_i \quad \text{s.t.} \quad \theta + u_i \geq (x_t - \gamma - \xi_1 x_{t-1} - \ldots - \xi_p x_{t-p})^2, \quad u_i \geq 0, \quad i \in [p+1, n].$$

This is a convex quadratic optimization problem, and hence can be solved by commercial optimization software in very high dimensions. Indeed further note that as compared with the solution to the formulation employing $Z_2$, the above formulation involves far fewer decision variables and constraints, and hence optimization theory would predict run times should be faster for this formulation, a prediction which is later validated in our timing section. A further advantage is that the above formulation is perhaps more interpretable than the previous one in that it has one fewer hyperparameters to tune, and even the nature of that hyperparameter has a simple interpretation: the size of the subsets to be employed in training. These two points together suggest that while in theory it is true that any solution using Eq. (2) paired with $Z_1$ can be derived from a solution to Eq. (2) paired with $Z_2$ with appropriately chosen parameters, in practice is preferable to use the specialized procedure derived above if one’s intent is to run Eq. (2) paired with $Z_1$. Indeed in all experiments in the rest of the paper employing this model we do exactly this.

4. Computational Results

In this section, we present computational results comparing stable time series ($Z_1$) and stable time series ($Z_2$) to regular time series across several metrics. Note that in the remainder of the paper, when we say stable time series ($Z_2$), we are referring to the methodology described by Eq. (2) and employing uncertainty set $Z_2$. When we say stable time series ($Z_1$), we are referring to the methodology described by Eq. (2) and employing uncertainty set $Z_1$.

First, we compare the three methods during a turbulent time period to investigate the additional stability conferred by stable time series ($Z_2$) and stable time series ($Z_1$). This
Stable Time Series

is done at two different prediction frequencies: where predictions are made once every 24 hours, as well as at a higher frequency, once every 8 hours. The above analysis is also replicated at three different “re-fitting” schedules, the first where the models are refit daily, the second where the models are refit weekly, and the third where the models are refit monthly. The rationale behind additionally investigating the models at different refitting schedules is that if stable time series are indeed robust to regime changes, then they should perform well even during less frequent re-fittings. Indeed, we may even hypothesize that the magnitude of the improvement would be inversely proportional to the frequency with which the models are refit.

Finally, the same analysis is also performed during a non-turbulent, calm period to investigate whether stability during turbulent periods comes at the price of lower performance during more calm time periods.

The above analysis is performed on eight separate time series provided by the Beth Israel Deaconess Medical Center: the first four represent time series of the number of arrivals to the emergency department, sampled at 8 and 24 hours respectively, all culled across a turbulent time period as well as a calm time period. Time series 5 through 8 represent time series of the number of bed requests within the hospital, again sampled at 8 and 24 hours respectively, and all culled across a turbulent time period as well as a calm time period.

4.1 Testing Methodology

To compare stable time series (Z2) and stable time series (Z1) to regular time series, we employ the following methodology:

1. We first collected several data sets from the Beth Israel Deaconess Medical Center (Bertsimas et al. (2016)). The first class of data sets represent time series of the number of arrivals to the emergency department over multiple time periods sampled every 24 hours and 8 hours, respectively, all culled across a turbulent time period as well as a calm time period. Time series 5 through 8 represent time series of the number of bed requests within the hospital, again sampled at 8 and 24 hours respectively, and all culled across a turbulent time period as well as a calm time period.

2. We next separated the two classes of data sets (i.e., number of arrivals to the emergency department at all available frequencies and number of bed requests within the hospital at all available frequencies) into data over a turbulent period (December 2019 through October 2020, i.e., a period of time capturing the COVID-19 dislocation) as well as data over a non-turbulent, more “typical” period (January 2019 through November 2019). Both segments of time were confirmed by Beth Israel Deaconess Medical Center to be atypical and typical, respectively.

3. Next, the data was split into a testing set (15% of the total data), validation set (30% of the non-test data), and training data (70% of the non-test data). Note that as this is time series data, this split cannot be done randomly but rather must be done
chronologically. At this time, the data was also log transformed and first differences were taken to ensure any trend and seasonal components were removed.

4. Then, for a given data set (i.e., either emergency department arrivals or hospital bed requests) at a specific frequency (i.e., either 24 hours or 8 hours) during a coherent period of time (i.e., either turbulent or non turbulent), and for a fixed refitting window (i.e., either daily, weekly or monthly), denote this period $k$, the following procedure was applied: stable time series ($Z_2$), stable time series ($Z_1$) and regular time series were trained on the data set from the beginning up to through (initially) one month. Predictions were then made for the next period of length $k$ (i.e., if the refitting frequency is weekly, then predictions would be made for the next week).

5. For the sake of thoroughness, this is done in two different ways, the first where predictions are made before the period begins for the whole duration of the period (i.e., in the case of weekly refitting, before the week begins we would issue predictions for Monday through Sunday), and the second in a rolling window fashion (i.e., before making predictions for Tuesday, we wait to receive the actual numbers from Monday). Note that in the case of the rolling window predictions, even though predictions are made in a rolling window fashion, the model fitting is not (i.e., it is kept fixed for the duration of the period). Only at the end of the period is the model refit (i.e., in the case of weekly refitting, if the first model was fit using the data from the beginning up through one month, then the second model would be fit using data from the beginning up through one month and one week).

6. Accuracy (mean squared error) is reported for both prediction schemes as is stability of the model (i.e., the variability of the coefficients over the entire period of refitting). Standard errors are included alongside each metric to ensure statistical significance of the results.

4.2 Daily Refitting

In this section, we present results on both classes of data sets, first on the number of arrivals to the emergency department, and second on the number of bed requests within the hospital, at all available frequencies (i.e., 24 hours and 8 hours) under daily refitting. Note that while both standard and rolling window predictions are reported for each schema above, in the case of daily refitting for the 24 hour frequency data sets, standard and rolling window predictions necessarily coincide. Finally, the above analysis is performed during both a turbulent period as well as a calm period, and results from both are reported.

4.2.1 Turbulent Period

Results comparing stable time series to regular time series for both data sets at all available frequencies during a turbulent period are reported in Table 1.
Table 1: Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a turbulent period under daily refitting. The accuracy (MSE) of both standard, as well as rolling window predictions, are reported alongside model stability metrics (i.e., variance of the coefficients of the model). Standard errors are included with each metric. TS, STS, STS1, RW refer to Time Series, Stable Time Series (Z2), Stable Time Series (Z1) and rolling window predictions, respectively.

The structure of Table 1 is as follows: Each row corresponds to one method. For example, TS corresponds to regular time series, STS corresponds to stable time series (Z2), and STS1 corresponds to stable time series (Z1). Furthermore, when the name appears without the modifier _RW, those are results that correspond to regular predictions, whereas those that include the _RW modifier correspond to rolling window predictions. Each column corresponds to a particular type of metric. The two main ones present are accuracy, which is measured by mean squared error and shows up in the table as MSE, as well as model stability, which represents the standard deviation of the model coefficients over time and shows up in the table as Model Stability. Furthermore, each metric in the table is accompanied by a frequency modifier: (24 hours) indicating that the corresponding data set was sampled at a daily frequency, whereas (8 hours) indicates a data set that was sampled at a tri-daily frequency. Also note that every metric is accompanied by a standard error, which shows up in the table as the column S.E. immediately to the right of the column corresponding to each metric. Finally, the above analysis was conducted for both the emergency department arrivals data set as well as the number of bed requests within the hospital data set, and results for each appear under the corresponding titles in the table: “Emergency Department Arrivals” and “Hospital Bed Requests” respectively.

Table 1 indicates that for the regime of daily refitting during a turbulent period, stable time series (Z2) models improve upon both the accuracy of traditional times series models as well their stability; for stable time series (Z1) we see the improvement in stability pushed farther still, though this now comes at the price of a smaller improvement in accuracy over regular time series as compared with stable time series (Z2). In particular, for the emergency department arrivals data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 1.95 percent with an associated p-value of < 0.001, and in the case of rolling window predictions,
we have a statistically significant improvement of 1.95 percent with an associated p-value of < 0.001. In terms of model stability, the gains are larger, now with a statistically significant improvement of 17.47 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 24 hours, we have a statistically significant improvement of 0.91 percent with an associated p-value of 0.006 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 0.91 percent with an associated p-value of 0.006. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 23.99 percent over regular time series with an associated p-value of < 0.001.

For the emergency department arrivals data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 2.85 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 1.56 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 9.24 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 8 hours, we have a statistically significant improvement of 1.42 percent with an associated p-value of 0.008 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 0.68 percent with an associated p-value of 0.004. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 19.68 percent over regular time series with an associated p-value of < 0.001.

Now for the hospital bed requests data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 2.92 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 2.92 percent with an associated p-value of < 0.001. In terms of model stability, the gains are larger, now with a statistically significant improvement of 16.84 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the hospital bed requests data set at a frequency of 24 hours, we have a statistically significant improvement of 2.37 percent with an associated p-value of 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 2.37 percent with an associated p-value of 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 19.90 percent over regular time series with an associated p-value of < 0.001.

For the hospital bed requests data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have an improvement of 1.33 percent with an associated p-value of 0.212, and in the case of rolling window predictions, we have a statistically significant improvement of 1.22 percent with an associated p-value of 0.010. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 27.31 percent with an associated p-value of < 0.001. In the case of regular
### Stable Time Series

#### 4.2.2 Calm Period

Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a calm period are reported in Table 2. The structure of Table 2 is identical to that of Table 1, the only difference being that the former corresponds to results for a turbulent period whereas the latter corresponds to results for a calm period.

Table 2 indicates that for the regime of daily refitting during a calm period, stable time series (Z2) models improve upon both the accuracy of traditional times series models as well their stability; for stable time series (Z1) we see the improvement in stability pushed farther still, though this now comes at the price of a smaller improvement in accuracy over regular time series as compared with stable time series (Z2). In particular, for the emergency department arrivals data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 3.85 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 3.85 percent with an associated p-value of < 0.001. In terms of model stability, the gains are larger, now with a statistically significant improvement of 10.03 percent with an associated p-value of < 0.001.

### Table 2: Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a calm period under daily refitting.

<table>
<thead>
<tr>
<th></th>
<th>Emergency Department Arrivals</th>
<th>Hospital Bed Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE (24 Hours)</td>
<td>S.E.</td>
</tr>
<tr>
<td><strong>TS</strong></td>
<td>1.785E-02</td>
<td>6.433E-04</td>
</tr>
<tr>
<td><strong>TS_RW</strong></td>
<td>1.785E-02</td>
<td>6.433E-04</td>
</tr>
<tr>
<td><strong>STS2</strong></td>
<td>1.716E-02</td>
<td>5.858E-04</td>
</tr>
<tr>
<td><strong>STS2_RW</strong></td>
<td>1.716E-02</td>
<td>5.858E-04</td>
</tr>
<tr>
<td><strong>STS1</strong></td>
<td>1.749E-02</td>
<td>6.093E-04</td>
</tr>
<tr>
<td><strong>STS1_RW</strong></td>
<td>1.749E-02</td>
<td>6.093E-04</td>
</tr>
</tbody>
</table>

The accuracy (MSE) of both standard, as well as rolling window predictions, are reported alongside model stability metrics (i.e., variance of the coefficients of the model). Standard errors are included with each metric. TS, STS, STS1, RW refer to Time Series, Stable Time Series (Z2), Stable Time Series (Z1) and rolling window predictions, respectively.
predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 24 hours, we have a statistically significant improvement of 2.02 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 2.02 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 11.22 percent over regular time series with an associated p-value of < 0.001.

For the emergency department arrivals data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 5.19 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 2.68 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 5.94 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 8 hours, we have a statistically significant improvement of 4.00 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 1.97 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 7.07 percent over regular time series with an associated p-value of < 0.001.

Now for the hospital bed requests data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have an improvement of 1.33 percent with an associated p-value of 0.098, and in the case of rolling window predictions, we have an improvement of 1.33 percent with an associated p-value of 0.098. In terms of model stability, the gains are larger, now with a statistically significant improvement of 12.72 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the hospital bed requests data set at a frequency of 24 hours, we have an improvement of 0.42 percent with an associated p-value of 0.640 over regular time series, and in the case of rolling window predictions, we have an improvement of 0.42 percent with an associated p-value of 0.640. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 19.85 percent over regular time series with an associated p-value of < 0.001.

For the hospital bed requests data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 4.62 percent with an associated p-value of 0.004, and in the case of rolling window predictions, we have a statistically significant improvement of 5.25 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 18.86 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the hospital bed requests data set at a frequency of 8 hours, we have a statistically significant improvement of 3.70 percent with an associated p-value of 0.021 over regular time series, and in the case of rolling window predictions, we have a statistically significant improve-
### Stable Time Series

#### Weekly Refitting Across All Available Frequencies During Turbulent Period

<table>
<thead>
<tr>
<th></th>
<th>Emergency Department Arrivals</th>
<th>Hospital Bed Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE (24 Hours)</td>
<td>S.E.</td>
</tr>
<tr>
<td>TS</td>
<td>3.313E-02</td>
<td>6.32E-04</td>
</tr>
<tr>
<td>TS_RW</td>
<td>2.12E-02</td>
<td>1.52E-04</td>
</tr>
<tr>
<td>STS2</td>
<td>3.01E-02</td>
<td>4.33E-04</td>
</tr>
<tr>
<td>STS2_RW</td>
<td>2.09E-02</td>
<td>1.57E-04</td>
</tr>
<tr>
<td>STS1</td>
<td>3.32E-02</td>
<td>4.37E-04</td>
</tr>
<tr>
<td>STS1_RW</td>
<td>2.14E-02</td>
<td>1.37E-04</td>
</tr>
</tbody>
</table>

**Table 3:** Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a turbulent period under weekly refitting. The accuracy (MSE) of both standard, as well as rolling window predictions, are reported alongside model stability metrics (i.e., variance of the coefficients of the model). Standard errors are included with each metric. TS, STS, STS1, RW refer to Time Series, Stable Time Series (Z2), Stable Time Series (Z1) and rolling window predictions, respectively.

#### 4.3 Weekly Refitting

In this section, we present results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets (i.e., the emergency department arrivals data set and number of bed requests within the hospital data set), at all available frequencies (i.e., 24 hours and 8 hours), for both calm and turbulent regimes, all under weekly refitting.

#### 4.3.1 Turbulent Period

Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a turbulent period are reported in Table 3. The structure of Table 3 is identical to that of Table 1, the only difference being that the former corresponds to weekly refitting during a turbulent period whereas the latter corresponds to daily refitting during a turbulent period.

Table 3 indicates that for the regime of weekly refitting during a turbulent period, stable time series (Z2) models improve upon both the accuracy of traditional times series models as well their stability; for stable time series (Z1) we see the improvement in stability pushed farther still, though this now comes at the price of a smaller improvement in accuracy over regular time series as compared with stable time series (Z2). In particular, for the emergency
department arrivals data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 0.91 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 4.02 percent with an associated p-value of < 0.001. In terms of model stability, the gains are larger, now with a statistically significant improvement of 12.03 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 24 hours, we have a statistically significant improvement of 0.38 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 0.81 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 44.18 percent over regular time series with an associated p-value of < 0.001.

For the emergency department arrivals data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 1.69 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 4.48 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 27.88 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 8 hours, we have a statistically significant improvement of 0.63 percent with an associated p-value of 0.012 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 2.13 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 42.56 percent over regular time series with an associated p-value of < 0.001.

Now for the hospital bed requests data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 1.45 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 5.82 percent with an associated p-value of < 0.001. In terms of model stability, the gains are larger, now with a statistically significant improvement of 38.15 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the hospital bed requests data set at a frequency of 24 hours, we have a statistically significant improvement of 0.79 percent with an associated p-value of 0.025 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 2.09 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 49.87 percent over regular time series with an associated p-value of < 0.001.

For the hospital bed requests data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 1.46 percent with an associated p-value of < 0.001, and in the case of rolling window
Table 4: Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a calm period under weekly refitting. The accuracy (MSE) of both standard, as well as rolling window predictions, are reported alongside model stability metrics (i.e., variance of the coefficients of the model). Standard errors are included with each metric. TS, STS, STS1, RW refer to Time Series, Stable Time Series (Z2), Stable Time Series (Z1) and rolling window predictions, respectively.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Frequency</th>
<th>TS</th>
<th>S.E.</th>
<th>Model Stability (24 Hours)</th>
<th>S.E.</th>
<th>MSE (8 Hours)</th>
<th>S.E.</th>
<th>Model Stability (8 Hours)</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emergency Department Arrivals</td>
<td></td>
<td>Model Stability (24 Hours)</td>
<td>S.E.</td>
<td>MSE (8 Hours)</td>
<td>S.E.</td>
<td>Model Stability (8 Hours)</td>
<td>S.E.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>4.17E-02</td>
<td>1.39E-04</td>
<td>3.49E-03</td>
<td>4.04E-06</td>
<td>1.20E-01</td>
<td>3.66E-03</td>
<td>9.15E-06</td>
<td>9.15E-06</td>
<td></td>
</tr>
<tr>
<td>TS_RW</td>
<td>1.82E-02</td>
<td>1.22E-04</td>
<td>3.49E-03</td>
<td>4.04E-06</td>
<td>1.20E-01</td>
<td>3.66E-03</td>
<td>9.15E-06</td>
<td>9.15E-06</td>
<td></td>
</tr>
<tr>
<td>STS2</td>
<td>1.21E-02</td>
<td>4.92E-04</td>
<td>1.23E-03</td>
<td>3.75E-06</td>
<td>1.30E-01</td>
<td>3.66E-03</td>
<td>9.15E-06</td>
<td>9.15E-06</td>
<td></td>
</tr>
<tr>
<td>STS2_RW</td>
<td>1.72E-02</td>
<td>1.04E-04</td>
<td>3.75E-06</td>
<td>6.08E-02</td>
<td>6.26E-04</td>
<td>3.66E-03</td>
<td>9.15E-06</td>
<td>9.15E-06</td>
<td></td>
</tr>
<tr>
<td>STS1</td>
<td>3.94E-02</td>
<td>7.38E-04</td>
<td>3.49E-03</td>
<td>4.04E-06</td>
<td>1.30E-01</td>
<td>3.66E-03</td>
<td>9.15E-06</td>
<td>9.15E-06</td>
<td></td>
</tr>
<tr>
<td>STS1_RW</td>
<td>1.90E-02</td>
<td>1.52E-04</td>
<td>3.49E-03</td>
<td>4.04E-06</td>
<td>1.30E-01</td>
<td>3.66E-03</td>
<td>9.15E-06</td>
<td>9.15E-06</td>
<td></td>
</tr>
<tr>
<td>Hospital Bed Requests</td>
<td></td>
<td>Model Stability (24 Hours)</td>
<td>S.E.</td>
<td>MSE (8 Hours)</td>
<td>S.E.</td>
<td>Model Stability (8 Hours)</td>
<td>S.E.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>4.70E-02</td>
<td>2.06E-03</td>
<td>9.56E-03</td>
<td>7.68E-05</td>
<td>2.63E-01</td>
<td>1.78E-02</td>
<td>9.13E-06</td>
<td>9.13E-06</td>
<td></td>
</tr>
<tr>
<td>STS2</td>
<td>4.94E-02</td>
<td>1.77E-03</td>
<td>3.06E-03</td>
<td>3.06E-05</td>
<td>2.51E-01</td>
<td>1.51E-02</td>
<td>9.63E-05</td>
<td>9.63E-05</td>
<td></td>
</tr>
<tr>
<td>STS2_RW</td>
<td>2.75E-02</td>
<td>1.34E-03</td>
<td>3.06E-03</td>
<td>3.06E-05</td>
<td>2.51E-01</td>
<td>1.51E-02</td>
<td>9.63E-05</td>
<td>9.63E-05</td>
<td></td>
</tr>
<tr>
<td>STS1</td>
<td>4.14E-02</td>
<td>1.82E-03</td>
<td>6.19E-03</td>
<td>3.61E-05</td>
<td>2.51E-01</td>
<td>1.51E-02</td>
<td>9.63E-05</td>
<td>9.63E-05</td>
<td></td>
</tr>
<tr>
<td>STS1_RW</td>
<td>2.67E-02</td>
<td>1.47E-03</td>
<td>6.19E-03</td>
<td>3.61E-05</td>
<td>2.51E-01</td>
<td>1.51E-02</td>
<td>9.63E-05</td>
<td>9.63E-05</td>
<td></td>
</tr>
</tbody>
</table>

4.3.2 Calm Period

Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a calm period are reported in Table 4. The structure of Table 4 is identical to that of Table 3, the only difference being that the former corresponds to weekly refitting during a calm period whereas the latter corresponds to weekly refitting during a turbulent period.

Table 4 indicates that for the regime of weekly refitting during a calm period, stable time series (Z2) models improve upon both the accuracy of traditional times series models as well their stability; for stable time series (Z1) we see the improvement in stability pushed farther still, though this now comes at the price of a smaller improvement in accuracy over regular time series as compared with stable time series (Z2). In particular, for the emergency department arrivals data set at a frequency of 24 hours, we see that in the case of regular
predictions for stable time series (Z2), we have a statistically significant improvement of 7.46 percent with an associated p-value of $< 0.001$, and in the case of rolling window predictions, we have a statistically significant improvement of 3.73 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains are larger, now with a statistically significant improvement of 32.06 percent with an associated p-value of $< 0.001$. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 24 hours, we have a statistically significant improvement of 3.73 percent with an associated p-value of $< 0.001$ over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 1.03 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 35.90 percent over regular time series with an associated p-value of $< 0.001$.

For the emergency department arrivals data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 1.30 percent with an associated p-value of $< 0.001$, and in the case of rolling window predictions, we have a statistically significant improvement of 3.72 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 8.70 percent with an associated p-value of $< 0.001$. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 8 hours, we have an improvement of 0.32 percent with an associated p-value of 0.055 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 1.52 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 17.42 percent over regular time series with an associated p-value of $< 0.001$.

Now for the hospital bed requests data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 4.45 percent with an associated p-value of $< 0.001$, and in the case of rolling window predictions, we have a statistically significant improvement of 6.72 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains are larger, now with a statistically significant improvement of 20.50 percent with an associated p-value of $< 0.001$. In the case of regular predictions for stable time series (Z1), again for the hospital bed requests data set at a frequency of 24 hours, we have a statistically significant improvement of 3.27 percent with an associated p-value of $< 0.001$ over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 1.98 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 35.27 percent over regular time series with an associated p-value of $< 0.001$.

For the hospital bed requests data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 4.59 percent with an associated p-value of $< 0.001$, and in the case of rolling window predictions, we have a statistically significant improvement of 4.50 percent with an asso-
ciated p-value of $< 0.001$. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 1.22 percent with an associated p-value of $< 0.001$. In the case of regular predictions for stable time series (Z1), again for the for the hospital bed requests data set at a frequency of 8 hours, we have a statistically significant improvement of 3.19 percent with an associated p-value of $< 0.001$ over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 2.21 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 1.65 percent over regular time series with an associated p-value of $< 0.001$.

4.4 Monthly Refitting

In this section, we present results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets (i.e., the emergency department arrivals data set and number of bed requests within the hospital data set), at all available frequencies (i.e., 24 hours and 8 hours), for both calm and turbulent regimes, all under monthly refitting.

4.4.1 Turbulent Period

Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a turbulent period are reported in Table 5. The structure of Table 5 is identical to that of Table 1, the only difference being that the former corresponds to monthly refitting during a turbulent period whereas the latter corresponds to daily refitting during a turbulent period.

Table 5 indicates that for the regime of monthly refitting during a turbulent period, stable time series (Z2) models improve upon both the accuracy of traditional times series models as well their stability; for stable time series (Z1) we see the improvement in stability pushed farther still, though this now comes at the price of a smaller improvement in accuracy over regular time series as compared with stable time series (Z2). In particular, for the emergency department arrivals data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 6.07 percent with an associated p-value of $< 0.001$, and in the case of rolling window predictions, we have a statistically significant improvement of 8.95 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains are larger, now with a statistically significant improvement of 38.52 percent with an associated p-value of $< 0.001$. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 24 hours, we have a statistically significant improvement of 3.05 percent with an associated p-value of $< 0.001$ over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 4.94 percent with an associated p-value of $< 0.001$. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 55.89 percent over regular time series with an associated p-value of $< 0.001$. 

23
Table 5: Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a turbulent period under monthly refitting. The accuracy (MSE) of both standard, as well as rolling window predictions, are reported alongside model stability metrics (i.e., variance of the coefficients of the model). Standard errors are included with each metric. TS, STS, STS1, RW refer to Time Series, Stable Time Series (Z2), Stable Time Series (Z1) and rolling window predictions, respectively.

For the emergency department arrivals data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 9.44 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 4.47 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 27.55 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 8 hours, we have a statistically significant improvement of 2.57 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 3.08 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 44.71 percent over regular time series with an associated p-value of < 0.001.

Now for the hospital bed requests data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 4.92 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 7.93 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the for the hospital bed requests data set at a frequency of 24 hours, we have a statistically significant improvement of 3.04 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 6.67 percent with an associated p-value of < 0.001. In terms of model
Table 6: Results comparing stable time series (Z2) and stable time series (Z1) to regular time series for both data sets at all available frequencies during a calm period under monthly refitting. The accuracy (MSE) of both standard, as well as rolling window predictions, are reported alongside model stability metrics (i.e., variance of the coefficients of the model). Standard errors are included with each metric. TS, STS, STS1, RW refer to Time Series, Stable Time Series (Z2), Stable Time Series (Z1) and rolling window predictions, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Emergency Department Arrivals</th>
<th>Hospital Bed Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE (24 Hours)</td>
<td>S.E.</td>
</tr>
<tr>
<td>TS</td>
<td>3.101E-02</td>
<td>5.090E-05</td>
</tr>
<tr>
<td>TS_RW</td>
<td>2.822E-02</td>
<td>4.637E-05</td>
</tr>
<tr>
<td>STS</td>
<td>3.101E-02</td>
<td>5.090E-05</td>
</tr>
<tr>
<td>STS_RW</td>
<td>2.822E-02</td>
<td>4.637E-05</td>
</tr>
</tbody>
</table>

For the hospital bed requests data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 4.57 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 9.92 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 57.14 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the the for the hospital bed requests data set at a frequency of 8 hours, we have a statistically significant improvement of 3.43 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 4.02 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 65.04 percent over regular time series with an associated p-value of < 0.001.

### 4.4.2 Calm Period

Results comparing stable time series (Z2) to regular time series for both data sets at all available frequencies during a calm period are reported in Table 6. The structure of Table 6 is identical to that of Table 5, the only difference being that the former corresponds to monthly refitting during a calm period whereas the latter corresponds to monthly refitting during a turbulent period.
Table 6 indicates that for the regime of monthly refitting during a calm period, stable time series (Z2) models improve upon both the accuracy of traditional times series models as well their stability; for stable time series (Z1) we see the improvement in stability pushed farther still, though this now comes at the price of a smaller improvement in accuracy over regular time series as compared with stable time series (Z2). In particular, for the emergency department arrivals data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 8.97 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 6.62 percent with an associated p-value of < 0.001. In terms of model stability, the gains are larger, now with a statistically significant improvement of 67.02 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 24 hours, we have a statistically significant improvement of 3.19 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 4.27 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 61.88 percent over regular time series with an associated p-value of < 0.001.

For the emergency department arrivals data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 8.37 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 6.79 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 30.65 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the emergency department arrivals data set at a frequency of 8 hours, we have a statistically significant improvement of 3.87 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 4.77 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 42.52 percent over regular time series with an associated p-value of < 0.001.

Now for the hospital bed requests data set at a frequency of 24 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 6.01 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 10.87 percent with an associated p-value of < 0.001. In terms of model stability, the gains are larger, now with a statistically significant improvement of 40.43 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the hospital bed requests data set at a frequency of 24 hours, we have a statistically significant improvement of 2.67 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 6.29 percent with an associated p-value of < 0.001. In terms of model
Stable Time Series

stability, the gains are even larger, now with a statistically significant improvement of 65.91 percent over regular time series with an associated p-value of < 0.001.

For the hospital bed requests data set at a frequency of 8 hours, we see that in the case of regular predictions for stable time series (Z2), we have a statistically significant improvement of 5.92 percent with an associated p-value of < 0.001, and in the case of rolling window predictions, we have a statistically significant improvement of 7.30 percent with an associated p-value of < 0.001. In terms of model stability, the gains here are also larger, now with a statistically significant improvement of 22.29 percent with an associated p-value of < 0.001. In the case of regular predictions for stable time series (Z1), again for the for the hospital bed requests data set at a frequency of 8 hours, we have a statistically significant improvement of 3.62 percent with an associated p-value of < 0.001 over regular time series, and in the case of rolling window predictions, we have a statistically significant improvement of 6.07 percent with an associated p-value of < 0.001. In terms of model stability, the gains are even larger, now with a statistically significant improvement of 29.07 percent over regular time series with an associated p-value of < 0.001.

5. Discussion

Overall, stable time series models have a clear advantage over traditional time series models in terms of accuracy, and especially in terms of stability. This is true during both turbulent periods, and just as importantly, also true during calm periods.

One important observation is that essentially across the board, stable time series (Z2) outperformed stables time series (Z1) in terms of accuracy, and the reverse situation held for the case of model stability (i.e., stable time series (Z1) outperformed stable time series (Z2)). Importantly, despite this trade off, both stable time series (Z2) and stable time series (Z1) outperformed regular time series on both accuracy and model stability. The existence of this trade-off makes sense, as stable time series (Z1) employ a smaller uncertainty set than do stable time series (Z2), and hence can be viewed as more conservative. By actively attempting to inoculate themselves against a greater variety of situations, stable time series (Z1) achieve the strongest model stability of all three models, though pay a price for this conservatism relative to stable time series (Z2) in terms of accuracy. Ultimately, it will be up to the user to decide which objective, accuracy or stability is more important for his or her application, and these priorities will naturally guide the choice of stable time series (Z2) or stable time series (Z1), respectively. Regardless of the decision, it is encouraging to know that improvements can be had on both the accuracy front as well as the model stability front over regular time series with only a small overhead cost in terms of time.

Another important observation, is that generally the less frequently the models are refit, the greater the relative improvement in both accuracy and model stability for stable time series (Z2) and stable time series (Z1) over regular time series. This makes sense, as stable time series (Z2) and stable time series (Z1) are explicitly designed to be robust towards regime changes. If the models are being refit very frequently, it is less likely for a regime

27
shift to be unexpectedly encountered, and hence the relative benefit of using stable time series models will be smaller. Indeed we see this very clearly in the case of the daily refitting results, where in terms of accuracy, stable time series (Z1) and regular time series performed very similarly, and some cases where there appeared to be an edge proved in fact to be statistically insignificant. It is useful to note that stable time series (Z2) fared better in this regime than did stable time series (Z1), though directionally the observation still holds. In contrast, when we moved towards a less frequent weekly refitting regime, the benefits of stable time series (Z2) and stable time series (Z1) began to show themselves more clearly, and this improvement continued even more dramatically in the case of monthly refitting. This property of being stable over time is a particularly important property for a time series model to possess, as in practice, it is often impractical to refit a model with great frequency. This can be due to monetary constraints, institutional constraints, or countless other reasons. The fact that stable time series models perform well in the middle of an ever-changing global pandemic, even when refit only once a month provides a strong vote of confidence on their behalf.

Another interesting observation is that generally, the relative benefit of stable time series (Z2) and stable time series (Z1) appear to be greater for rolling window predictions than for regular predictions. At first glance, this appears reasonable, as one would expect having access to the most recent information would allow a method to perform better than it if it were only allowed access to slightly stale information. Indeed, the smaller error bars around these rolling window predictions further support this view. What is interesting, however, is that the relative improvement also appears to be greater in this setting for both stable time series (Z2) and stable time series (Z1). This suggests that perhaps regular time series models pounce too greedily onto the fresh information they receive from a rolling window scheme and overreact, whereas stable time series models are better able to place that information into context and react less excitedly. Indeed, this is further supported by the observation that the relative improvement for the stable time series models over regular time series in the case of rolling window predictions is the greatest during the calm regime. A possible reason being that during a calm regime, one knows that tomorrow is likely to be similar to today, and hence one shouldn’t take aberrant information too seriously.

The second point above provides strong evidence that stable time series models do in fact accomplish what they were designed to do, i.e., be stable across time and be robust towards regime changes. But it is the third point that makes them actually usable in practice. Because it shows that the increased robustness of stable time series models during turbulent periods does not come at the price of performance during calm periods. Indeed, they even enjoy a modest advantage over traditional time series models during such calm periods. Hence in practice, when one does not know apriori when a calm period will turn into a turbulent period, one will be well served by running a stable time series model at all times.
6. Conclusion

In this paper, we propose a robust optimization based framework for addressing the issue that time series exhibit regime switches over time. In contrast to previous approaches which require strong assumptions on the regime switching process, or are designed to perform well after the regime shift has already occurred, we propose a methodology for constructing a single model that is explicitly designed to be robust against regime changes. The approach is non-parametric, requiring no description of a regime switching process, and most importantly, is inherently forward looking, in that the model is designed to smoothly pass through the break point between two regimes, without knowing of when the aforementioned break point will occur. Through experiments on multiple data sets, at multiple sampling frequencies, across multiple fitting schedules, we demonstrate that these stable models perform significantly better than their non-stable counterparts during turbulent periods of regime shift, with the magnitude of the improvement being inversely proportional to the frequency with which the models are refit. Just as important, we show that this increased robustness does not come at the price of performance during calm periods. Indeed, they even enjoy a small advantage over traditional time series models during such calm periods. Finally, these stable time series models are highly efficient, and scale to data sets of essentially any desired size.

Acknowledgements

The research was partially supported by a grant to MIT (PI: Professor Bertsimas) by IBM, and by an NSF graduate fellowship to Ivan Paskov.
References


