A Randomized Missing Data Approach to Robust Filtering and Forecasting

Dobrislav Dobrev[†]   Derek Hansen[*]   Paweł J. Szerszeń[†]

December 31, 2021

Abstract

We put forward a simple new randomized missing data (RMD) approach to robust filtering of state-space models, motivated by the idea that the inclusion of only a small fraction of available highly precise measurements can still extract most of the attainable efficiency gains for filtering latent states, estimating model parameters, and producing out-of-sample forecasts. In our general RMD framework we develop two alternative implementations: endogenous (RMD-N) and exogenous (RMD-X) randomization of missing data. A degree of robustness to outliers and model misspecification is achieved by purposely randomizing over the utilized subset of seemingly highly precise but possibly misspecified or outlier contaminated data measurements in their original time series order, while treating the rest as if missing. Time-series dependence is thus fully preserved and all available measurements can get utilized subject to a degree of downweighting depending on the loss function of interest. The arising robustness-efficiency trade-off is controlled by varying the fraction of randomly utilized measurements or the incurred relative efficiency loss. As an empirical illustration, we show consistently attractive performance of our RMD framework in popular unobserved components models for extracting inflation trends. We further consider model extensions that more directly reflect inflation targeting by central banks and reveal its effectiveness through improved inflation forecasting.

Keywords: state space models, robustness, outliers, misspecification, inflation modelling.

JEL classification: C11; C15; C22; C53; E37.

1. Introduction

State-space models play an important role in fields as diverse as engineering, medicine, economics, and finance. However, their empirical performance, particularly in forecasting applications, suffers when there is even a small degree of model misspecification. This misspecification can be induced, for example, by the presence of outliers in collected measurement data. These problems have motivated the development of methods for estimation of state-space models with specific focus on robustness to complex data imperfections. One popular approach imposes heavy-tailed distributional assumptions as in Durbin and Koopman (2000) or Harvey and Luati (2014). Another alternative adds thresholding or an outlier detection step as in Calvet et al. (2015), Crevits and Croux (2017), Maìz et al. (2012), among others.

In this paper we propose a randomized missing data (RMD) framework for robust estimation of state-space models. Our framework achieves a degree of robustness to outliers and model misspecification by purposely randomizing the subset of included but possibly misspecifed or outlier contaminated data while treating the rest as if missing. This ensures that all available measurements would still get utilized but subject to a degree of downweighting. What makes our randomization approach work is that in many cases, most of the information for filtering latent states, estimating model parameters, and producing out-of-sample forecasts can be extracted from a small subset of available measurements. This is particularly true when the latent process is highly persistent through time. Thus, randomizing the inclusion of data points can achieve robustness to outliers and model misspecification without much efficiency loss. Moreover, this trade-off between robustness and efficiency loss can be controlled by tuning it to optimize out-of-sample forecasting metrics or, in general, by minimizing a loss function of interest.

A similar idea rationalizing the exclusion of data measurements was previously

considered by Sims (2003) and later Sims (2011) who develop the notion of rational inattention. They show that, in the presence of explicit costs to collecting and processing measurements, it is optimal from a decision-theoretic point of view to not use all available data. We argue that the presence of measurement outliers and misspecification leads to a similar trade-off between bias and efficiency. Since the inclusion of all available observations can yield biased state and parameter estimates, the risk of including contaminated observations is an implicit cost to using more data. Hence, by inducing a degree of "randomized inattention" in our RMD framework, we can control for the trade-off between bias and efficiency subject to minimizing the loss function of interest. As previously noted by Hamilton (1986), among others, the total uncertainty of filtered states in state-space models can be decomposed into separate filter uncertainty and parameter uncertainty components. This decomposition further entails an inherent bias-efficiency trade-off with respect to the set of measurements taken into consideration in the presence of model misspecification, e.g. due to intermittent measurement distortions.

In Section 2, we lay out the general properties of the RMD framework in a Bayesian setting and then show two alternative implementations. The first one, denoted RMD-N, allows for endogenous randomization of missing data based on an indicator that is subject to a learning assumption. The second implementation, denoted RMD-X, is subject to a no-learning assumption with exogenous randomization of missing data. We establish that both RMD-N and RMD-X arise as special cases in our general RMD framework under specific choices of the generative model for missing observations. The imposed fraction of randomly utilized measurements plays the role of a regularization parameter controlling the arising robustness-efficiency trade-off in the presence of misspecified observations with favorable learning properties. As such, the RMD framework reflects more generally also the vision by Fan et al. (2021) for cross-fertilization between the two cultures of statistics (data/generative modeling) and machine learning (algorithmic modeling).

To illustrate the empirical performance of our approach, in Section 3 we consider state

space models for extracting inflation trends and document favorable performance of our RMD framework and the resulting out-of-sample forecasts. We first apply RMD to the classical unobserved components (UC) model. It has been well documented by Stock and Watson (2007) and Stock and Watson (2016), among others, that the forecasting performance of the UC model is hindered by the presence of time-variation in the precision of inflation rate measurements and additional measurement distortions due to outliers. We therefore study an RMD-N and RMD-X augmented versions of the classical UC model vis-à-vis the unobserved components/stochastic volatility outlier-adjustment (UCSVO) model that was put forward by Stock and Watson (2016) as another way to minimize the detrimental impact of outliers by way of subjecting them to particular distributional assumptions.

We find that our RMD-N and RMD-X augmented UC model that avoids overfitting via "randomized inattention" to the available measurements can offer modest improvements over the UCSVO model in terms of out-of-sample forecasting performance, especially at longer 8- and 12- month horizons. We conclude that the favorable forecasting results obtained when applying RMD filtering to the UC model underscore the findings in Stock and Watson (2007) and Stock and Watson (2016) regarding the need for time-series inflation forecasts to address the presence of potentially complex inflation measurement imperfections. We also show improvement in mean squared forecast error (MSFE) over a UC model with t-distributed innovations without stochastic volatility (UC-T), demonstrating that the filtering and forecasting benefits offered by RMD cannot be attained by simply imposing a heavier-tailed measurement distribution.

We further extend the empirical study of RMD extraction of inflation trends by considering also a more general AR model for the unobserved inflation component which allows for reversion to a long-run mean either estimated from the data or fixed at the target rate of 2% as a way to more directly reflect central bank inflation targeting that has been in place for many years. The much superior forecasting performance we document when the mean is fixed to the *a priori* 2% inflation target in place suggests that inflation targeting has

been effective and the RMD framework works best when paired with additional economic theory sources of information about the underlying model. Both when applied to the classical UC and better-performing AR models with fixed inflation target mean our RMD approach to robust filtering and forecasting offers attractive empirical performance gains and ease of implementation in comparison to existing alternatives.

2.   The RMD Modelling Framework

For $t \geq 1$, the following equations characterize the Randomized Missing Data (RMD) modelling framework:

$$x_0 \sim \mu_\theta(\cdot) \tag{1}$$

$$x_t | x_{t-1}, \theta \sim g_\theta(\cdot | x_{t-1}, \theta) \tag{2}$$

$$y_t | x_t, \theta \sim f_\theta(\cdot | x_t) \tag{3}$$

$$y_t^{(j)} \sim f_j(\cdot | \hat{y}^{t-1}) \tag{4}$$

$$C_t \sim P(.|\beta) \tag{5}$$

$$\hat{y}_t = \begin{cases} y_t \text{ if } C_t = 1 \\ y_t^{(j)} \text{ if } C_t = 0 \end{cases} \tag{6}$$

As in a standard state-space model, $x_t \in \mathbb{R}^{d_x}$ is a latent Markov process parameterized by $\theta \in \Theta$ with an initial distribution μ_θ (Equation 1) and a transition kernel g_θ (Equation 2). The observation $y_t \in \mathbb{R}^{d_y}$ contains information about the latent states x_t through the kernel f_θ (Equation 3). However, unlike a standard state-space model, we further suppose that y_t is only observed a fraction of the time as specified by the three additional equations (4)-(6). Specifically, we observe $\hat{y}_t$, which equals the correctly-specified y_t only when the auxiliary indicator variable $C_t \in \{0, 1\}$ equals 1 (Equation 6). In the case that $C_t = 0$,
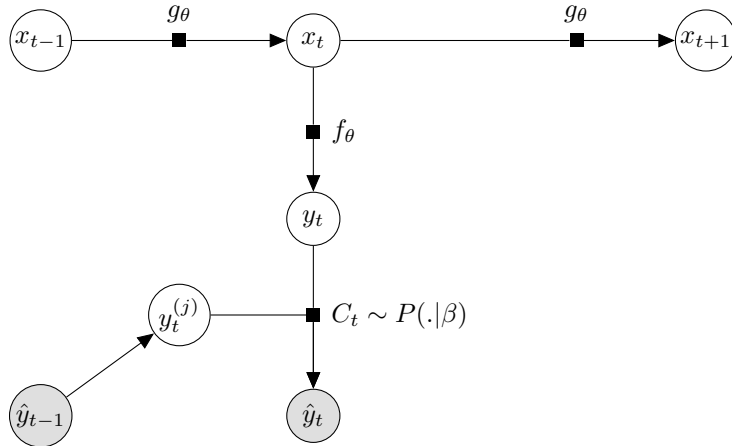
Figure 1: State space model with observed measurement $\hat{y}_t$ that, conditional on the value of the auxiliary indicator C_t, can be possibly corrupted/uninformative and given by $y_t^{(j)}$ or a correctly-specified one and given by y_t. Shaded nodes are the observed data.

the observed $\hat{y}_t$ instead equals a corrupted value $y_j \in \mathbb{R}^{d_y}$ from the *alternative kernel* f_j which, conditional on previous observations $\hat{y}^{t-1}$, is independent of the model parameters θ, the latent process x_t, and current and future realizations of y_t. Unlike a typical missing data problem, the observed $\hat{y}_t$ has a recorded value at every index from 1 to T. However, based on the structure of the model, if the true C^T were known, then the framework would simply reduce to standard estimation with missing data, since the $y_t^{(j)}$ have no additional information about θ or x_t. Figure 1 provides a block-diagram illustration of the model with observed measurement $\hat{y}_t$ that, conditional on the value of the auxiliary indicator C_t, can be possibly corrupted/uninformative and given by $y_t^{(j)}$ or a correctly-specified one and given by y_t.

The goal of adopting the RMD modelling framework is to optimize the trade-off between the gain of including correctly-specified observations and the loss of including misspecified observations. This is performed through controlling the distribution of C_t through the parameter β. Although the distribution of C_t can be general, the simplest case is to assume the C_t are i.i.d. with distribution $Bernoulli(\beta)$, $\beta \in [0, 1]$. In this formulation, $\beta = 1$ corresponds to including all measurements as if correctly specified, while the limiting case of

$\beta = 0$ has no inclusion of data at all. For $\beta \in (0, 1)$, the level of β reflects the degree to which the observed $\hat{y}^T$ are included via randomization inducing downweighting of the observed $\hat{y}^T$ to robustify model estimates and forecasts. As such, the three additional equations (4)-(6) characterizing our RMD framework can be viewed as an add-on to any standard state space model specification (1)-(3) for the purposes of robust filtering and forecasting in the presence of measurement outliers and model mis-specification even without knowledge of the exact form of such imperfections.

In addition to the prior distribution for C^T, the nature of the induced missing data randomization depends on the distribution of the alternative measurement f_j. In general, the alternative measurement at time t is assumed to be independent of model parameters θ and states x^T conditional on observations up to time $t - 1$. Using the general case, we are able to derive the posterior distribution $P(x^T, \theta, C^T | \hat{y}^T, \beta)$ sequentially, showing that at each t the distribution is a mixture between including each $\hat{y}_t$ as if it were equal to its informative counterpart y_t and treating the measurement as if it were missing. The model posterior distriubution is given by

$$P(x^T, \theta | \hat{y}^T, \beta) = \prod_{t=1}^{T} \left(\hat{\beta}(\hat{y}^t) \frac{f_\theta(\hat{y}_t | x_t)}{F(\hat{y}_t | \hat{y}^{t-1}, \beta)} + (1 - \hat{\beta}(\hat{y}^t)) \right) g_\theta(x_t | x_{t-1}) \mu_\theta(x_0) P(\theta) \qquad (7)$$

Here, $\hat{\beta}(\hat{y}^t)$ is shorthand for the filtered posterior probability $P(C_t = 1 | \hat{y}^t, \beta)$ derived in Appendix A, and $F(\hat{y}_t | \hat{y}^{t-1}, \beta) = \int_{x_t, \theta} f_\theta(\hat{y}_t | x_t) dP(x_t, \theta | \hat{y}^{t-1}, \beta)$ is the one-step-ahead predictive density of y_t evaluated at $\hat{y}_t$. A detailed derivation of the posterior distribution can be found in Appendix A.

Given this posterior form in general, a natural question that arises is how to select a particular choice of f_j. We show two different assumptions which lead to different explicit forms for f_j and respective alternative RMD implementations based on endogenous randomization (RMD-N) or exogenous randomization (RMD-X). In the RMD-N specification, we assume that the missing indicator C_t cannot be learned from past or present observations

$\hat{y}_t$, but can be learned from future observations. From this assumption, we can derive the exact form that the alternative distribution f_j must take. This approach has favorable learning properties, because one can infer where in the past $E(C_t|\hat{y}^T)$ is low, and also imposes neglible computational cost compared to the original state-space model defined by equations (1)-(3).

In contrast, in the RMD-X specification, we assume that the missing indicator C_t cannot be inferred even after observing all data points. We show that this leads to another exact form for f_j which induces a posterior distribution which randomizes over $P(C^T|\beta)$.

2.1   RMD-N: Endogenous Randomization

The RMD-N approach models the indicator variable C^T endogenously within a Bayesian hierarchical model using the relationships outlined in Equations (1) - (6). This method requires an explicit specification of the alternative distribution f_j. As indicated above, there is much flexibility in how f_j is chosen. In our RMD-N framework we assume that $y_t^{(j)}$ has distribution given by the one-step-ahead predictive density $F(\cdot|\hat{y}^{t-1}, \beta)$ and is conditionally independent of all other past, present, and future realizations. This is equivalent to assuming the contaminated measurements should be indistinguishable from the true measurements at the time of observing them, but allowing for the possibility that future values may be informative about C_t. This particular choice of f_j pairs well with existing state-of-the-art Sequential Monte Carlo (SMC) techniques for jointly estimating parameters and states such as SMC^2 from Chopin et al. (2013). This is because at each point in time the latent C_t can be marginalized out, and $f_j(\hat{y}_t)$ can be calculated as a density forecast over the available filtered particles sampled from $P(x_{t-1}, \theta|\hat{y}^{t-1}, \beta)$.

We start by supposing that C_t is not inferable with information up to time t. Specifically, $P(C_t = 1|\hat{y}^t, \beta) = P(C_t = 1|\beta) = \beta$. Following Equation (A.1'), this assumption implies that $f_j(\hat{y}_t|\hat{y}^{t-1}, \beta) = F(\hat{y}_t|\hat{y}^{t-1}, \beta)$, which is the one-step-ahead predictive density

evaluated at the observed data point. The posterior distribution in Equation (A.3) can also be further simplified

$$P(x^t, \theta | \hat{y}^t) = (\beta \frac{f(\hat{y}_t | x_t, \theta)}{F(\hat{y}_t | \hat{y}^{t-1}, \beta)} + (1 - \beta)) g_\theta(x_t | x_{t-1}) P(x^{t-1}, \theta | \hat{y}^{t-1}) \tag{8}$$

Because the contaminated measurement $y_t^{(j)}$ is indistinguishable in distribution from the true measurement y_t with present information, the inclusion of the observed $\hat{y}_t$ at the time of observation is only weighted by the prespecified prior probability β. However, while the filtered probability $P(C_t = 1 | \hat{y}^t, \beta)$ equals β, the smoothed probability $P(C_t = 1 | \hat{y}^T, \beta)$ will not necessarily equal β. This is because future realizations of $\hat{y}_t$ will be informative about C_t. Thus, at the end of estimation, the inclusion and weighting of a measurement at a particular point in time will depend on the extent to which it is corroborated by future observations. These smoothed probabilities can also be of interest to the researcher in diagnosing where the original model breaks down.

2.2 RMD-X: Exogenous Randomization

The data randomization prescribed by RMD-X fixes the weight on each subset of the data beforehand. More specifically, the RMD-X approach exogenously draws each path C^T from some prespecified distribution satisfying the condition $P(C_t = 1) = \beta$. Parameter estimates and forecasts are calculated for each C^T as they would be in the usual missing-data context. The estimates and forecasts are then averaged across paths, making them more robust to data imperfections at particular points. Because the missing data randomization is performed outside of the estimation process, there is no learning about the indicator C_t.

We derive a specific choice of f_j in our general RMD framework which elicits a posterior distribution equivalent to the purely exogenous randomization of RMD-X. Instead of assuming that only C_t cannot be inferred with information up to time t, as in the case of

RMD-N, we now assume that the entire path C^t up to time t cannot be learned from the data. This leads to another explicit formulation of f_j inducing a mixture distribution over all posterior distributions conditional on C^T which are weighted only by the prior weight assigned to C^T. Thus, we are able to justify averaging estimates over repeat data copies with randomly induced missing values as a key simplifying feature of the RMD-X approach.

To obtain an RMD-X implementation that fixes the weight on each subset of the data beforehand we set f_j as follows:

$$f_j(\hat{y}_{\{t:C_t=0\}}|C^T, \hat{y}_{\{t:C_t=1\}}) = P_y(\hat{y}_{\{C_t=0\}}|\hat{y}_{\{C_t=1\}})$$

where P_y is the uncontaminated observation distribution. With this selection of f_j, $\hat{y}^T$ is independent of C^T:

$$P(\hat{y}^T|C^T) = P_y(\hat{y}_{\{C_t=1\}})P_y(\hat{y}_{\{C_t=0\}}|\hat{y}_{\{C_t=1\}}) = P_y(\hat{y}^T)$$

From this, $P(C^T|\hat{y}^T, \beta) = P(C^T|\beta)$, and we obtain the posterior as follows:

$$
\begin{aligned}
P(x^T, \theta|\hat{y}^T, \beta) &= \sum_{C^T} P(C^T|\hat{y}^T, \beta)P(x^T, \theta|C^T, \hat{y}^T) \\
&= \sum_{C^T} P(C^T|\beta)\frac{P(x^T, \theta)P(\hat{y}^T|C^T, x^T, \theta)}{P(\hat{y}^T|C^T)} \\
&= \sum_{C^T} P(C^T|\beta)\frac{P(x^T, \theta)\prod_{t:C_t=1} f(\hat{y}_t|x_t, \theta)f_j(\hat{y}_{\{C_t=0\}}|\hat{y}_{\{C_t=1\}})}{P(\hat{y}_{\{C_t=1\}})f_j(\hat{y}_{\{C_t=0\}}|\hat{y}_{C_t=1})} \\
&= \sum_{C^T} P(C^T|\beta)\frac{P(x^T, \theta)\prod_{t:C_t=1} f(\hat{y}_t|x_t, \theta)}{P(\hat{y}_{\{C_t=1\}})} \\
&= \sum_{C^T} P(C^T|\beta)P(x^T, \theta|\hat{y}_{\{C_t=1\}})
\end{aligned}
$$

Thus, if the estimator of interest is the posterior mode, f_j integrates out and this leads

to RMD-X implementation equivalent to a weighted mixture over all possible C^T:

$$P(x^T, \theta | \hat{y}^T, \beta) = \sum_{C^T} P(C^T | \beta) P(x^T, \theta | \hat{y}_{\{C_t = 1\}}) \tag{9}$$

This result justifies casting RMD-X in simpler terms allowing for MLE estimation with the use of bootstrap-like aggregation over data replicas that preserve time series dependence by randomly drawing missing values and retaining the original time series order of the rest. More formally, for a designated weighting scheme $P(C^T | \beta)$, the two-step algorithm for computing the RMD-X filtered states and parameter estimates is as follows:

<u>Step 1</u>: Path-by-path filtering and forecasting for each path $C_i^T \in \{0, 1\}^T$:

- Assume $\hat{y}_{\{t:C_{t,i}=1\}} \overset{D}{=} y_{\{t:C_{t,i}=1\}}$ conditional on θ, x^T, with $\hat{y}_{\{t:C_{t,i}=0\}}$ missing at random.

- Obtain parameter estimates $\tilde{\theta}_i := \tilde{\theta}(\hat{y}_{\{t:C_{t,i}=1\}})$.

- Obtain filtered and forecasted states $\tilde{x}_i^{T+h} = \tilde{x}^{T+h}(\hat{y}_{\{t:\ C_{t,i}=1\}}, \tilde{\theta}_i)$.

<u>Step 2</u>: Aggregation across all paths $C_i^T \in \{0, 1\}^T$:

- Obtain RMD parameter estimates $\bar{\theta} = \bar{\theta}(\hat{y}^T, \beta) = \sum_i \tilde{\theta}_i P(C_i^T | \beta)$.

- Obtain RMD filtered and forecasted states $\bar{x}^{T+h} = \bar{x}^{T+h}(\hat{y}^T, \beta) = \sum_i \tilde{x}_i^{T+h} P(C_i^T | \beta)$.

The estimates calculated in RMD-X can be in principle any function of the data, but generally $\tilde{\theta}_i$ will either be the maximum-likelihood estimator of θ or a statistic of the posterior distribution if performing a Bayesian analysis. Likewise, if the goal is to minimize squared-error loss, then the estimated $\tilde{x}_{t+h,i}$ and forecasted $y_{t+h,i}$ will generally be calculated as expectations over the given subset of data and the estimated $\tilde{\theta}_i$.

The weighting scheme chosen for each path C^T is flexible. The most basic scheme one could use is sampling each $C_t \sim Bernoulli(\beta)$ independently. However, especially for low

β, this design assigns significant probability to using a low number of observations, which can lead to identifiability issues due to an insufficient number of observations. Therefore, a more practical approach is to fix the number of observations at $[\beta T]$, where $[x]$ is the nearest integer to x. This leads to the following distribution for C^T:

$$P(C^T) = \begin{cases} \binom{T}{[\beta T]}^{-1} \text{ if } |C^T| = [\beta T] \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

In practice, since the sample space of C^T is growing exponentially, it is infeasible to calculate the expectation across all possible C^T. Therefore, in applications, the expectation can be replaced with an average over Monte Carlo samples from $P(.|\beta)$. While this introduces some stochastic noise to the estimation, in practice we find that this noise is small provided a large enough number of samples drawn.

2.3 RMD-X versus RMD-N: Advantages and Disadvantages

A key advantage to the RMD-X method is its simplicity by purposefully inducing and randomizing over missing data in the original model. Focusing specifically on linear Gaussian state-space models, many software packages already have the built-in capability to handle missing data. All that is left to the user is to implement the sampling scheme for C^T, which can be as simple as sampling from $\{1, \ldots, T\}$ without replacement. Finally, RMD-X also allows for model estimation, filtering and forecasting robust to data contamination without the need to specify a specific model for that contamination.

However, because the missing data randomization is handled exogenously outside of the statistical model, one cannot learn about C^T using RMD-X. For forming robust predictions about $\hat{y}_t$, this is less important, but there may be interest in inferring which values of $E(C_t|\hat{y}^T)$ are particularly low. Another realted drawback to both RMD-X and f_j taken

in the "no-learning" approach is that calculation of the likelihood requires conditioning on an entire path of C^T.

By contrast, in the case of RMD-N and f_j with "smoothed" learning, C_t can be marginalized out. This means that $f_j(\cdot)$ can be estimated over the existing filtered particles of x_{t-1} and θ already available in an SMC^2 estimation. Hence, this choice of f_j within RMD-N adds negligible cost to the run-time of the algorithm.

As such, RMD-X introduces significant computational overhead because it estimates the model independently for each draw of C_i^T. Hence, it becomes impractical with a large number of observations T. In cases where the original model has a tractable likelihood function, such as via the Kalman filter, this overhead is manageable. However, many processes of interest follow a non-linear or non-Gaussian specification. These more general models require stochastic estimation methods such as Sequential Monte Carlo (SMC) which are computationally expensive, making this additional overhead impractical. Further, RMD-X becomes unstable with small values of β and a small number of observations T, where the sample size effectively used for estimation becomes too small for a model to be estimated.

Summing up, RMD-X offers a simple way to implement our RMD framework but cannot be applied in all scenarios, while RMD-N offers a tractable way to estimate models with very large T or small values of β as well as to learn about C^T and improve conditionally the amount of utilized information from each available measurement.

2.4 Distinction from Related Work

Relationship to Sim's Rational Inattention

Similar to the idea of rational inattention by Sims (2003) and Sims (2011), the RMD framework allows for the possibility that the optimal strategy for estimation and forecasting is to not fully fit to all observed data. However, rather than supposing an exogenous cost

to including more data, the RMD framework considers the risk of information loss from overfitting to misspecified values as an implicit cost of fitting to all observed data. This gives rise to "randomized innatention" to measurements and an optimal randomized usage rate of the data to a degree where the marginal loss of including a possibly misspecified observation, e.g., an outlier, equals the marginal gain of including it. While in practice the true expectation of information gain and loss cannot be calculated, this tradeoff can be achieved by setting the hyperparameter $\beta = P(C_t)$, via cross-validation or other related methods, to optimize a chosen loss function such as the log-forecasting density or Mean Squared Forecasting Error (MSFE) at a set forecasting horizon.

Relationship to Heavy Tailed Modelling

Assuming that the choice of f_j is more diffuse than the observation kernel f, the implementation of RMD naturally induces a heavier-tailed measurement equation, which has precedence in the literature to make state space models more robust to outliers. Models with a Gaussian measurement are particularly susceptible to outliers, and often need to be adapted to achieve robustness in applications. Such adaptations include using a mixture of Gaussians as in Kitagawa (1989) or a T distribution as in Harvey and Luati (2014) to better approximate the true heavy-tailed measurement. However, an advantage of the RMD framework is that it can be applied to any state-space model to safeguard from misspecification, even in cases where there are no obvious heavy-tailed modifications to the observation kernel. Moreover, existing robustification methods work by directly modifying f to allow for more extreme realizations, and these methods are not as flexible as the RMD framework since they usually assume that the degree of kurtosis exhibited by f is constant conditional on some hidden states and parameters.

Relationship to Bootstrap Aggregation

The RMD-X implementation of the RMD framework effectively can be viewed as a time-series extension of bootstrap aggregation (bagging), originally developed by Breiman (1996). As a key distinction from bagging, RMD-X samples without replacement and preserves the original index of each observation while randomly sampling missing values. This latter feature is a key difference from the extant literature on bagging, which focuses mostly on cross-sectional settings rather than time-series models where temporal dependence puts natural restrictions on resampling methods. A handful of studies that have considered the benefits of bootstrapping and bagging in a time-series context have done so either by resampling the full-sample model-based i.i.d. residuals to assess estimation uncertainty such as Stoffer and Wall (1991) or by resampling entire blocks of dependent observations to robustify forecasts as in Inoue and Kilian (2008) and Bergmeir et al. (2016).

Relationship to Machine Learning

More generally, apart from offering a time-series extension of bagging, the RMD framework can also be viewed as embodying the vision by Fan et al. (2021) for cross-fertilization between the two cultures of statistics (data/generative modeling) and machine learning (algorithmic modeling). From a machine learning standpoint, the fraction of randomly utilized measurements in the RMD framework constitutes a regularization parameter that controls the arising robustness-efficiency trade-off in the presence of misspecified observations with favorable learning properties. As such, it is possible to cast RMD alternatively as a machine learning method featuring a regularization parameter in the spirit of many other regularization techniques such as those collectively covered by Fan et al. (2020), among others.

2.5 Bias-Variance Decomposition of Expected Prediction Error

Both the RMD-N and RMD-X implementations of the RMD framework robustify the forecasts $\bar{x}_{T+h}(\beta)$ of future states x_{T+h} by inducing randomization rate $\beta \in (0, 1]$ between the model-implied kernel $f(.)$ and the alternative kernel $f_j(.)$ for the available measurements y_t up to time T. Thanks to avoiding explicit parametrization of the induced alternative kernel $f_j(.)$ for misspecified measurements this leads to the following bias-variance decomposition of the expected square prediction error as a function of β:

$$
\begin{aligned}
& \mathbb{E}\left(\bar{x}_{T+h}(\beta) - x_{T+h}\right)^2 \\
&= \mathbb{E}\left(\left(\bar{x}_{T+h}(\beta) - \mathbb{E}\bar{x}_{T+h}(\beta)\right) + \left(\mathbb{E}\bar{x}_{T+h}(\beta) - \mathbb{E}x_{T+h}\right) + \left(\mathbb{E}x_{T+h} - x_{T+h}\right)\right)^2 \\
&= \underbrace{\mathbb{E}\left(\bar{x}_{T+h}(\beta) - \mathbb{E}\bar{x}_{T+h}(\beta)\right)^2}_{\text{Reducible variance term } \downarrow \text{ as } \beta \uparrow} + \underbrace{\left(\mathbb{E}\bar{x}_{T+h}(\beta) - \mathbb{E}x_{T+h}\right)^2}_{\text{Reducible bias term } \downarrow \text{ as } \beta \downarrow} + \underbrace{\mathbb{E}\left(\mathbb{E}x_{T+h} - x_{T+h}\right)^2}_{\text{Irreducible variance term}} \quad (11)
\end{aligned}
$$

In the machine learning literature such bias-variance decomposition is known to arise in many prediction problems under square loss function, with all expectations reflecting the distribution of measurements and latent states given information up to time T. The first term represents reducible variance of the predictor. In the case of RMD the variance of $\bar{x}_{T+h}(\beta)$ decreases as the randomization rate β goes up allowing the model to more closely fit the full data sample at the expense of potentially larger bias. The second term captures reducible bias of the predictor. In the case of RMD the bias of $\bar{x}_{T+h}(\beta)$ goes down as the randomization rate β drops, thereby alleviating possible over-fitting of the full data sample at the expense of potentially larger variance. The third term represents irreducible variance reflecting the inherent uncertainty about future states x_{T+h} also when given full information up to time T.

Thus, the bias-variance decomposition (11) implies that for minimizing the expected prediction error under square loss function it should be beneficial to use the RMD framework

whenever full utilization of all available data (i.e. $\beta = 1$) entails a non-negligible bias term due to data over-fitting and model mis-specification. In such cases a data-driven estimate of the optimal value of $\beta \ll 1$ can be obtained using standard cross-validation techniques. By contrast, under the null of an unbiased model the zero bias term in the decomposition (11) would imply an optimal value of $\beta = 1$. From this perspective, the RMD framework offers two complementary uses: (i) an easy to attain model-free improvement for robust filtering and forecasting (optimal $\beta \ll 1$) based on any biased model of choice e.g. preferred due to simplicity; (ii) a simple test for unbiasedness (optimal $\beta \approx 1$) of any considered alternative more complex model-based improvement over a simpler biased model of choice. If one further takes the view that "all models are wrong but some are useful" then the RMD framework can thus offer an advantageous tool to help achieve model parsimony in applications of state-space models to forecasting (Occam's Razor).

To broaden the scope of these insights beyond the popular special case of a square loss function, it is useful to note that the bias-variance decomposition (11) readily extends also to a wide range of other loss functions. First, James and Hastie (1997) and James (2003) have shown how to obtain a straightforward generalization of the bias-variance decomposition (11) for any symmetric loss function. Second, Heskes (1998), Wu and Vos (2012) and Vos and Wu (2015) have further obtained an analogous bias-variance decomposition valid for any kind of error measure that can be derived from a Kullback-Leibler divergence or loglikelihood stemming from the underlying probability model. Therefore, the above insights regarding the RMD-induced bias-variance tradeoff as a function of the randomization rate β readily extend for any symmetric loss function as well as for the case when the expected prediction error captures the difference between the entire predictive distribution of $\bar{x}_{T+h}(\beta)$ and the respective target distribution of x_{T+h} and associated likelihoods.

3. Empirical Illustration: Inflation forecasting

3.1 Unobserved Components Model

Market participants and central bankers alike traditionally pay considerable attention to data-driven assessments of long-run inflation expectations. A large part of the literature on inflation forecasting has considered alternative econometric approaches to estimating inflation trends based on time series modelling of officially released price index data.[1] Stock and Watson (2007) and Stock and Watson (2016) provide compelling evidence for time-variation in the precision of inflation rate measurements as well as the presence of additional measurement distortions due to outliers. Inspired by these findings, we consider the ability of our robust filtering approach to successfully guard against the impact of inflation measurement imperfections without the need to explicitly model them for the purposes of improved forecasting of long-run inflation trends. In particular, we apply our RMD framework to the classical unobserved components (UC) representation of the IMA(1,1) benchmark model for inflation forecasting scrutinized also by Stock and Watson (2007). This RMD augmented specification is given by the following set of equations:

---

[1]   Another large strand of the literature aims to exploit the predictive content of collected survey-based measures of inflation. For a recent literature survey on inflation forecasting see for example Faust and Wright (2013).

$$x_t | x_{t-1}, \theta \sim g(.|x_{t-1}) = N(x_{t-1}, \sigma_\varepsilon^2) \tag{2'}$$

$$y_t | x_t, \theta \sim f(.|x_t) = N(x_t, \sigma_\eta^2) \tag{3'}$$

$$y_t^{(j)} \sim f_j(.) \tag{4'}$$

$$C_t \sim Bernoulli(\beta) \tag{5'}$$

$$\hat{y}_t = \begin{cases} y_t \text{ if } C_t = 1 \\ y_t^{(j)} \text{ if } C_t = 0 \end{cases} \tag{6'}$$

where $\hat{y}_t$ denotes the observed change in the log price-level (PCE) at quarter t subject to potential distortions tackled by the RMD equations (5')-(6'). C_t is the Bernoulli indicator variable which controls whether $\hat{y}_t$ is informative about the underlying rate $\beta \in [0,1]$. In this RMD augmented setting, y_t is the undistorted and hence informative measurement which depends on x_t, the latent trend component of inflation, and model parameters $\theta := (\sigma_\eta, \sigma_\epsilon)$. In contrast, $y_t^{(j)}$ is the alternative distorted measurement which contains no information about x or the model parameters. We consider both the RMD-N and RMD-X implementations of the RMD framework. In the case of RMD-N, as described in Section 2.1, we set the alternative distribution f_j as $F(\hat{y}_t | Y^{t-1}, \beta) := \int f(\hat{y}_t | x_t, \theta) dP(x_t, \theta | Y^{t-1})$, the one-step-ahead predictive density. In the case of RMD-X, we consider the simplified implementation given in Section 2.2.

We estimate this model with the same aggregate PCE price index quarterly data series (PCE-all) used in Stock and Watson (2016) from 1960Q1 to 2015Q2 with a forecast evaluation period set to 1990Q1-2015Q2, in order to directly compare with the more recent UCSVO model presented in that paper.[2] First, to illustrate how RMD impacts the estimation of latent states, we compare the filtered mean from the original UC model ($\beta = 1$) and the

---

[2]   We thank Stock and Watson (2016) for providing as an online supplement the data and program codes necessary for replicating their results.

Figure 2: Filtered mean comparison between $\beta = 1$ (dashed red) and $\beta = 0.15$ (solid purple) for RMD-X applied to the Unobserved Components (UC) model of Stock & Watson (2007), where black dots represent the observed log-quarterly inflation.
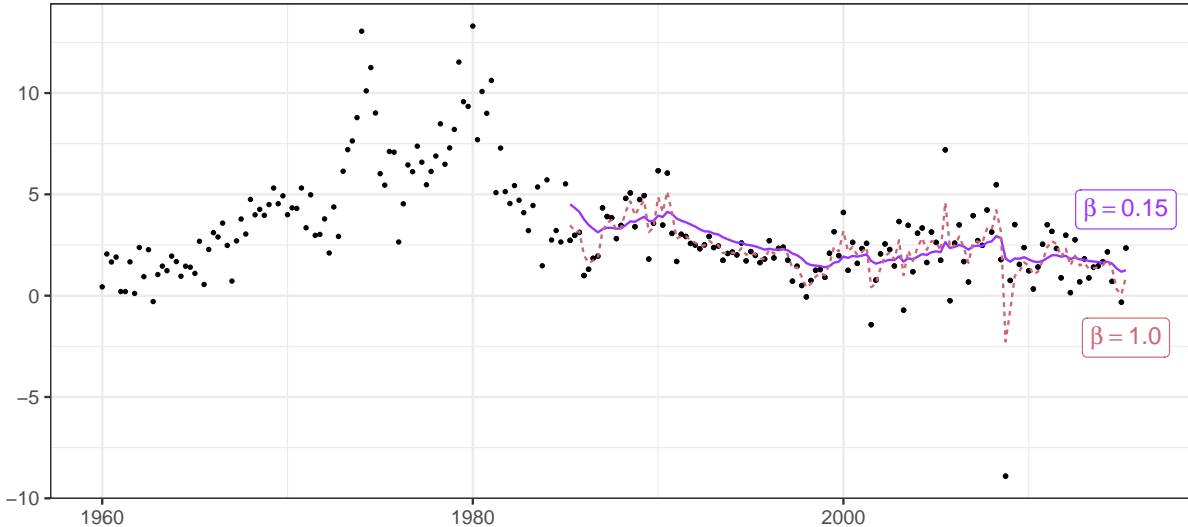
UC model augmented with either RMD-X (Figure 2) or RMD-N (Figure 3) at ($\beta = 0.15$). Without applying missing data randomization, the filtered mean for the original UC model visually overfits the observed inflation in each quarter. On the other hand, the filtered mean for the UC models augmented with RMD-X and RMD-N appear to follow a smoother path which better tracks the long-run trend of the process. This supports the notion in Section 3 that RMD favors smoother dynamics which follow long-term trend at the cost of short-term fit. Focusing on RMD-N specifically, the bottom panel of Figure 3 shows the inferred posterior smoothed probability that $C_t = 1$ at each point in time using all available data. The model identifies several points (indicated as large black dots) which are very unlikely to have information about the unobserved component ($P(C_t < 0.005|\hat{y}^T)$), and these points coincide with what one might visually determine as outliers. However, rather than relying on strict *apriori* assumptions to filter outliers, these points are identified simply from the dynamics of the process. Specfically, if future values do not follow the trend implied by an observation, the smoothed probability $P(C_t = 1|\hat{y}^T)$ decreases, allowing for an interpretable mechanism to filter our or downweight contaminated measurements.

Figure 3: Filtered mean comparison between $\beta = 1$ (dashed red) and $\beta = 0.15$ (solid blue) for RMD-N applied to the Unobserved Components (UC) model of Stock & Watson (2007), where black dots are the observed log-quarterly inflation. The smoothed posterior probability $P(C_t = 1|\hat{y}^T)$ for $\beta = 0.15$ is shown below. Large black dots indicate observations for which $P(C_t = 1|\hat{y}^T) < 0.005$.
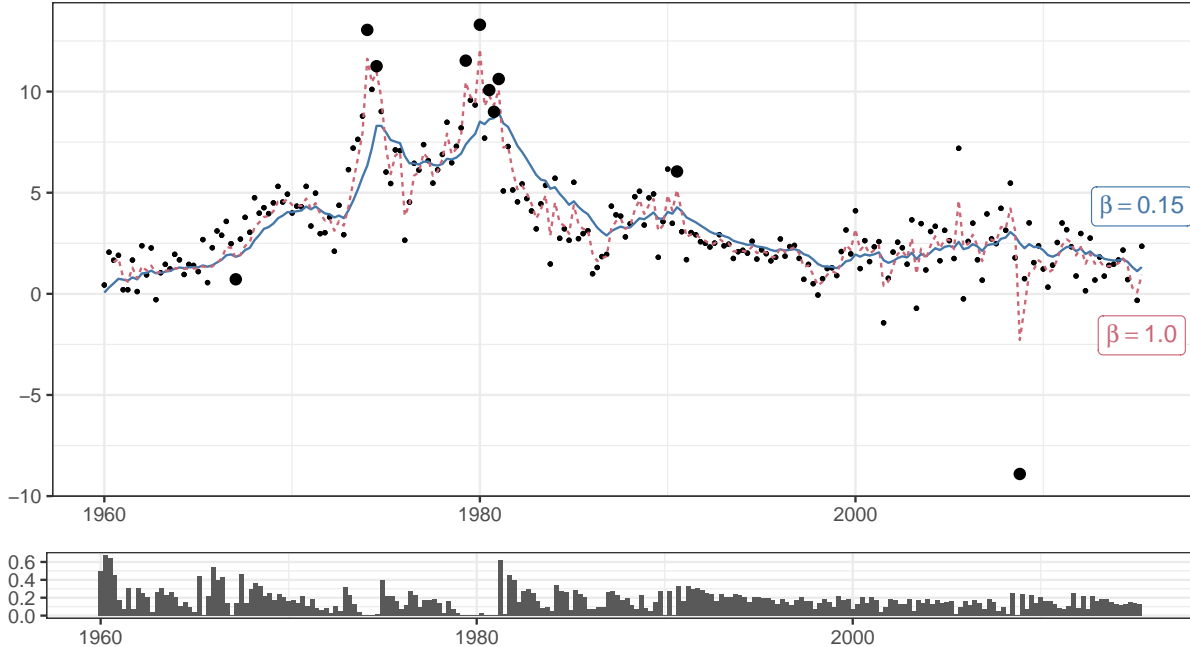
Next, to demonstrate the change in estimated parameters from applying RMD, Table 1 presents parameter estimates for the UC model with RMD-N using varying values of β estimated over the full sample from 1960Q1 to 2015Q2. It is evident that both the observation and state equation variance increases with the increase of β. This is explained by the requirement of the UC model to fit increasingly more and larger outliers present in the data. The increase is more pronounced for the parameter σ_η that controls the precision of the observation equation. To better illustrate the changes in the estimates of these parameters as we change the randomization parameter β we also present the plots with their estimates in Figure 4. As it is evident from Figure 4, the low and medium values of β do not have a significant impact on the estimates of σ_ε, while the estimates of σ_η grow steadily with the values of β. Since higher values of β naturally imply that more outliers are to be fitted by the model we should expect the variances of both the observation and state equations to

21

Table 1: UC model parameters estimated with RMD-N for different values of β based on data from 1960Q1 to 2015Q2.

β	Posterior Percentile	σ_ϵ	σ_η
0.15	50%	0.350	0.144
	2.5%	0.191	0.000
	97.5%	0.641	0.781
0.25	50%	0.360	0.435
	2.5%	0.222	0.002
	97.5%	0.645	0.870
0.90	50%	0.627	1.068
	2.5%	0.445	0.868
	97.5%	0.841	1.290
1.00	50%	0.653	1.375
	2.5%	0.467	1.205
	97.5%	0.891	1.564

increase with this parameter. However, this effect is stronger for the observation equation.

To evaluate the forecasting performance of the UC model with RMD, we first consider whether the improvement over the original UC model is significant. We use the Amisano and Giacomini (2007) test statistics, $\widehat{WLR}$, to compare the forecasting performance of the model augmented with RMD-N versus the original model. The $\widehat{WLR}$ statistics are computed relative to the model with $\beta = 1$ utilizing all inflation measurements with no randomization, which corresponds to the original UC specification in Stock and Watson (2007). Table 2 presents results for the UC model estimated recursively using the sample from 1960Q1 to 2005Q2 to generate forecast densities evaluated at 1-, 4-, 8-, and 12-quarter-ahead horizons starting from 1990Q1. For brevity, we document the results for three different choices of the parameter β of 0.15, 0.25 and 0.9. It is evident that all specifications with data randomization perform better than the benchmark UC model producing positive t-statistics with statistically significant improvements for all horizons for the specifications with lower values of β of 0.15 and 0.25.

After showing significant improvement in density forecasts, we compare the Mean Squared Forecasting Error (MSFE) of the UC model augmented with either RMD-X or RMD-N to that of the original UC model ($\beta = 1$). Following Stock and Watson (2007) and

Figure 4: Posterior distributions of UC model parameters estimated with RMD-N for different values of β based on quarterly inflation data from 1960Q1 to 2015Q2. From lightest to darkest, the shaded regions correspond to the middle 50%, 90% and 98% of the distribution.
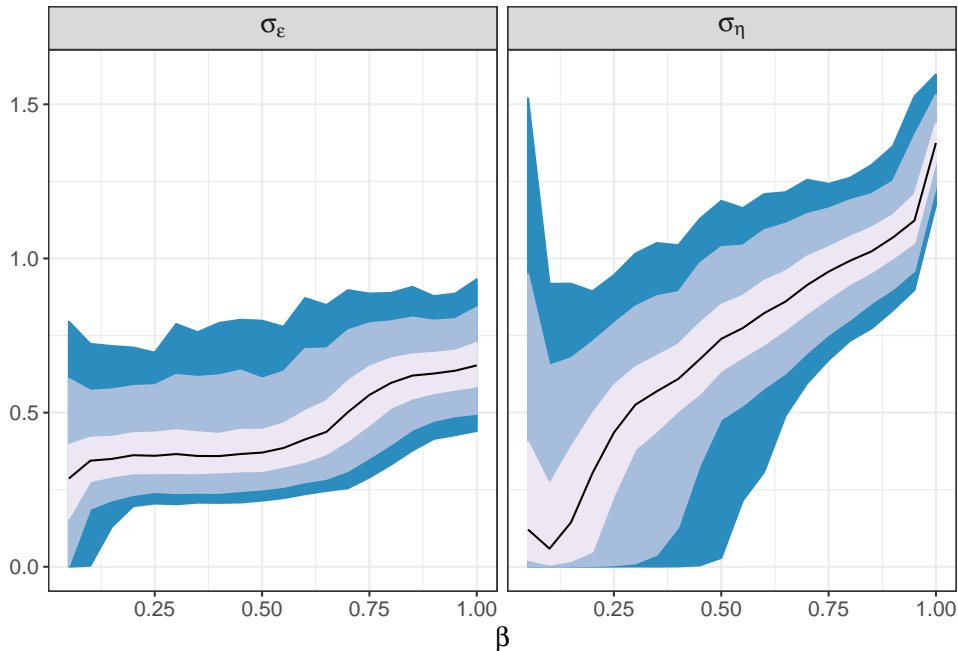
Stock and Watson (2016), we target the average inflation over 1-, 4-, 8-, and 12-quarters ahead. For completeness, we apply these criteria also to two other existing approaches for attaining robustness to outliers as alternative benchmarks against which we evaluate the relative performance of RMD. The first approach, which we denote as UC-T, follows Harvey and Luati (2014) in replacing the Gaussian measurement in Equation 3' with a scaled T distribution with $\nu > 0$ degrees of freedom. The second approach is the unobserved components/stochastic volatility outlier-adjustment (UCSVO) model representing a more recent extenstion of the UC model by Stock and Watson (2016). The UCSVO model also minimizes the detrimental impact of outliers by subjecting them to particular distributional assumptions while also allowing for stochastic volatility.

For a fair comparison, the latent states and parameters are re-estimated online at each point in time from 1960Q1 to 2015Q1 using only past information (no look ahead), with Mean

Table 2: **Comparison of forecasting performance of the UC model with RMD-N for different values of** β. The UC model is estimated based on data from 1960Q1 to 2005Q2, evaluating forecast density at 1-,4-,8-, and 12-quarter-ahead horizons starting in 1990Q1. The Weighted Likelihood Ratio (WLR) test from Amisano and Giacomini (2007) reported in the table compares forecasts based on the UC model with RMD-N for different values of β versus the ones for $\beta = 1$ standing for the original UC model as a benchmark.

β	h	$\widehat{\mathrm{WLR}}$	$\hat{\sigma}_{\mathrm{WLR}}$	t	P
0.15	1	0.2981	1.75	1.71	0.95
	4	0.2661	0.86	3.08	0.99
	8	0.2593	0.45	5.56	0.99
	12	0.2688	0.33	7.85	0.99
0.25	1	0.2748	1.52	1.82	0.96
	4	0.2507	0.73	3.39	0.99
	8	0.2514	0.47	5.15	0.99
	12	0.2688	0.37	6.93	0.99
0.90	1	0.0403	0.26	1.58	0.94
	4	0.0222	0.16	1.36	0.91
	8	0.0130	0.12	1.03	0.84
	12	0.0170	0.19	0.84	0.80

Squared Forecasting Error (MSFE) evaluated across 1, 4, 8, and 12-quarter horizons from 1990Q1 onward. Moreover, the hyperparameter β is chosen at each point in time as the value minimizing MSFE in targeting average inflation over 1-, 4-, 8-, and 12-quarter horizons also based on past data up to that point in time with no look-ahead (see Figure 5, top row). Thus, at each point in time and for each of the different optimizing criteria, the chosen optimal β is used to formulate forecasts. The resulting MSFE from each of these four strategies can then be used to objectively compare RMD-X and RMD-N to the UC-T and UCSVO benchmarks. As shown in Table 3 and in the top row of Figure 6, there is a clear improvement in adopting RMD-X or RMD-N over the base UC model, particularly when aiming to optimize over a 12-quarter horizon. Specifically, RMD-N achieves a 51% reduction in MSFE over the base UC model at 12-quarter horizons, as well as a 41% reduction compared to the UC-T model. Moreover, compared to UCSVO across the same 12-quarter horizon, the UC model augmented with RMD-N leads to a 14% reduction in MSFE. These gains demonstrate the ability of our RMD approach to robustly extract the persistent inflation trend component

Table 3: MSFE comparison for different considered state space models for extracting inflation trends augmented with RMD-X and RMD-N for a range of values of β optimized recursively to different forecasting horizons, compared to the UC and UCSVO benchmark models from Stock & Watson (2007) and Stock & Watson (2016) respectively. The considered forecast horizons are: one quarter (Q1), four quarter (Q2), eight quarter (Q8) and twelve quarters (Q12).

Model	RMD Approach	Strategy for β	Q1	Q4	Q8	Q12
UC	RMD-X	Q1	2.96	1.14	0.86	0.82
		Q4	2.95	1.10	0.84	0.80
		Q8	2.94	1.02	0.74	0.66
		Q12	3.03	1.04	0.74	0.68
	N/A	$\beta = 1$	3.33	1.58	1.26	1.22
	RMD-N	Q1	2.92	1.13	0.85	0.78
		Q4	2.93	1.11	0.85	0.78
		Q8	2.88	0.97	0.67	0.59
		Q12	2.96	0.98	0.66	0.59
	N/A	$\beta = 1$	3.33	1.58	1.26	1.22
AR	RMD-X	Q1	3.16	1.46	1.44	1.75
		Q4	3.34	1.54	1.49	1.79
		Q8	3.44	1.61	1.53	1.84
		Q12	3.55	1.66	1.59	1.92
	N/A	$\beta = 1$	3.31	1.61	1.48	1.74
	RMD-N	Q1	3.38	1.80	2.13	3.09
		Q4	3.50	1.79	2.10	2.98
		Q8	3.97	2.19	2.17	2.44
		Q12	4.36	2.40	2.19	2.24
	N/A	$\beta = 1$	5.42	3.51	3.11	3.02
ARMF	RMD-X	Q1	2.89	1.01	0.64	0.52
		Q4	3.01	1.11	0.77	0.60
		Q8	2.96	1.00	0.62	0.44
		Q12	2.98	0.92	0.51	0.33
	N/A	$\beta = 1$	3.27	1.39	0.92	0.75
	RMD-N	Q1	2.86	1.00	0.67	0.55
		Q4	2.96	1.06	0.73	0.57
		Q8	2.92	0.90	0.54	0.41
		Q12	2.91	0.86	0.48	0.33
	N/A	$\beta = 1$	3.23	1.31	0.84	0.66

Benchmarks			Q1	Q4	Q8	Q12
		UC-T ($\beta = 1$)	3.24	1.41	1.13	1.01
		UCSVO ($\beta = 1$)	–	1.09	0.81	0.69
		Naive Forecast of 2%	8.97	7.40	6.96	6.67

Figure 5: Optimal value of β selected over time using information up to a given date. Each box compares β selected by RMD-X and RMD-N optimizing a given forecasting MSFE criterion (1, 4, 8, or 12-quarters ahead) for different underlying models (UC, AR, or ARMF).
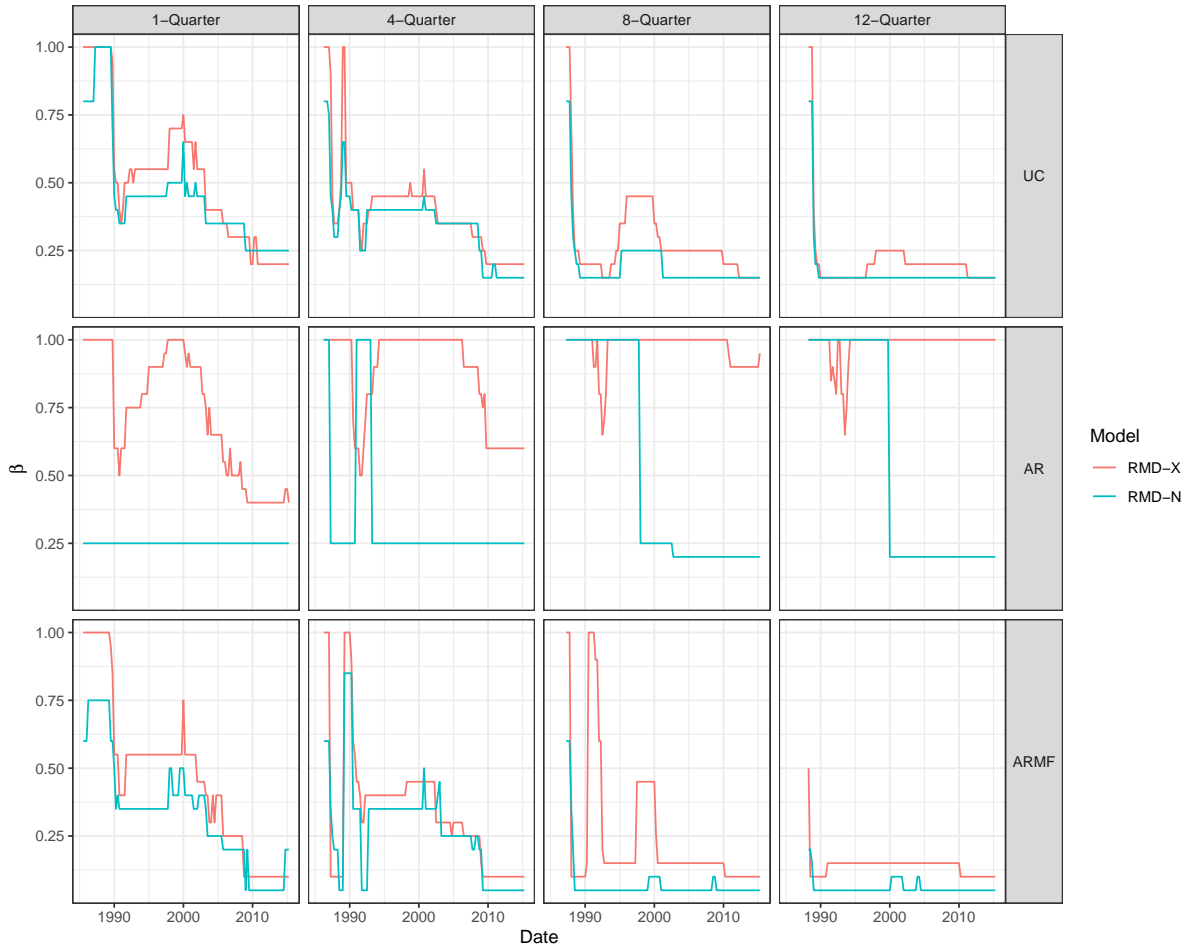
in the UC model from noisy and potentially misspecified or outlier-contaminated inflation measurements without the need to parametrize such more complex features as done in the UCSVO model. Overall, these results indicate that employing our RMD filtering framework significantly improves the forecasting performance of the classical UC model by naturally safeguarding against overfitting data outliers, thereby offering an attractive alternative to the UCSVO model.

3.2 Generalized AR Model Reflecting Inflation Targeting

The transition equation of the state process in the UC model, as shown in equation (2'), can be viewed as a corner case of the following more general autoregressive (AR) model:

$$x_t|x_{t-1}, \theta \sim g(x_t|x_{t-1}) = N(\mu + \kappa(x_{t-1} - \mu), \sigma_\varepsilon^2) \tag{12}$$

Here, κ reflects the degree of mean-reversion of the process, and the original UC model is equivalent to fixing $\kappa = 0$, which simultaneously removes μ as an identifiable parameter. Inflation targeting long adopted by central banks further provides justification from economic theory for modelling inflation with a long-run or stationary mean. For example, in the case of the United States an inflation target of 2% has been maintained by the Federal Reserve for many years (see Federal Reserve Board (2015)). However, even if the model were correctly specified, the long-run mean inflation rate cannot be estimated precisely without a large sample if $\kappa \approx 0$ (i.e., the process is close to a random walk). We therefore consider embedding knowledge about the inflation target policy in place directly into the model by fixing the long run rate at 2%. Thus, in addition to examining the performance of an unrestricted AR model, we also assess the forecasting performance of an alternative version of the same model (denoted ARMF) where μ is fixed a priori at the 2% inflation target.

Figure 6: Mean Squared Forecasting Error (MSFE) over different horizons (lower values are better). Each box compares the performance of RMD-X (red solid lines) and RMD-N (green solid lines) using a given criterion for selecting β (1, 4, 8, or 12-quarters ahead) and a given model (UC, AR, or ARMF). Latent states, parameters, and β are estimated from 1960Q1 to 2015Q2, and the MSFE is calculated with forecasts formulated in 1990Q1 onward. MSFEs from the original model with $\beta = 1$ (dashed lines) and the UCSVO model (blue dots) from Stock and Watson (2016) are plotted for reference.
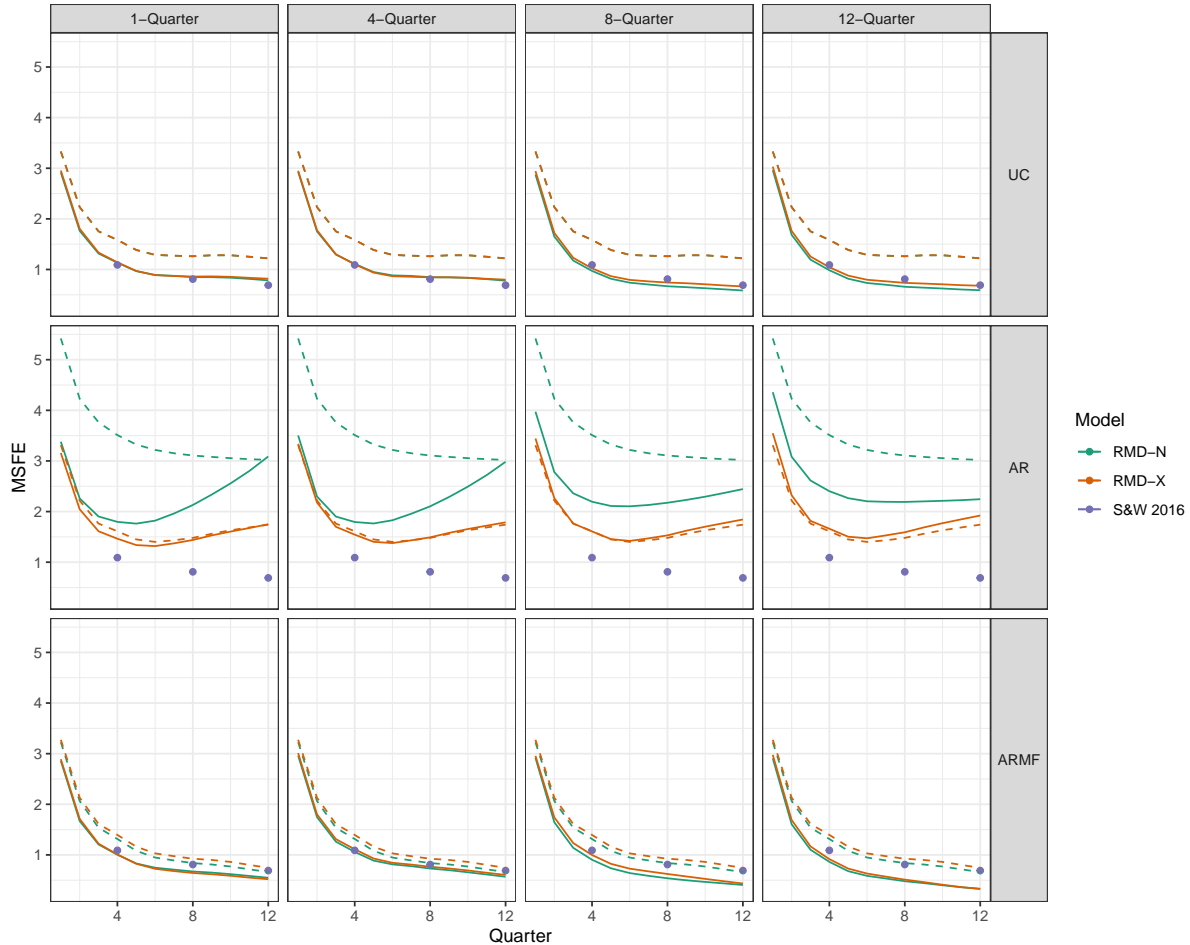
When applying our RMD framework to the AR model as we did for UC, Figure 5 shows that RMD-X selects $\beta = 1$ for most t at longer horizons for the unrestricted AR model with hard to estimate long-run mean. The RMD-N model favors a lower β and shows improvement over the base model, but the forecasting performance of RMD-N is worse than that of RMD-X across the board. This discrepancy is likely due to the fact that, as a fully Bayesian estimation, RMD-N includes the large uncertainty about μ in its forecasts,

which is detrimental to forecasting performance. Neither implementation of RMD reaches the performance of the benchmark UCSVO model. These findings are consistent with results from the prior literature also favoring UC over AR models for extracting inflation trends.

On the other hand, explicitly reflecting the inflation target in place by way of imposing a long-run mean of 2% in the ARMF model vastly improves forecasting performance compared to the UC model, even without using the RMD framework. Furthermore, applying the RMD-N framework to ARMF leads to a substantial improvement in the MSFE performance in the 12-quarter horizon over both the UCSVO benchmark (52% reduction) and that of the UC model augmented with RMD-N (43% reduction). These improvements in inflation forecasting can also be taken as new evidence that inflation targeting has been effective. [3]

Thus, in all three considered state space models for inflation (UC, AR and ARMF), adopting our RMD filtering framework consistently leads to often notable degrees of improvement in forecasting performance. Moreover, based on the vast RMD improvements attained for the UC and ARMF models which impose specific parameter constraints, we note that RMD performs best when paired with additional sources of information from economic theory. As RMD has the effect to regularize the estimated latent states toward path dynamics dictated by θ, it should not be surprising that having extra information about θ *a priori*, e.g. stemming from economic theory, can be crucial for filtering and forecasting performance. This further underscores that expert modelling still remains important for successful application of machine learning techniques by leveraging also particular theory-implied features of the data generating process.

---

[3] As a sanity check, we also compared forecasting results to a trival plug-in method where we set the 2% inflation target for every forecast. The performance of this naive method was significantly worse than that of any of the considered state space models. See last row in Table 3.

4. Conclusions

Similar to the justification for rational inattention (Sims (2003), Sims (2011)), in this paper we argue that the presence of measurement outliers and misspecification in state-space models may limit the scope for utilizing or putting equal weight on every available data measurement. More specifically, we note that even absent any explicit costs to data collection and processing, the optimal degree of utilization of imperfect data would implicitly depend on loss function sensitivities to bias versus efficiency. This motivates us to put forward a randomized missing data (RMD) approach to robust filtering of state-space models aiming to induce a degree of "randomized innatention" to available measurements. We accomplish this by purposely randomizing over the utilized subset of seemingly highly precise but possibly misspecified or outlier contaminated data measurements in their original time series order, while treating the rest as if missing. Such randomization ensures that time-series dependence is fully preserved and all available measurements can get utilized subject to a degree of downweighting depending on the loss function of interest. It also exploits the idea that the inclusion of only a small fraction of available highly precise measurements can still extract most of the attainable efficiency gains for filtering latent states, estimating model parameters, and producing out-of-sample forecasts. The arising robustness-efficiency trade-off is controlled by varying the fraction of randomly utilized measurements or the incurred relative efficiency loss from such randomized utilization of the available measurements.

On the theory side, we first lay down the general properties of our RMD framework in a Bayesian setting and then provide two alternative implementations. The first one, denoted RMD-N, allows for endogenous randomization of missing data based on an indicator that is subject to a learning assumption. The second implementation, denoted RMD-X, is subject to a no-learning assumption with exogenous randomization of missing data. We establish that both RMD-N and RMD-X arise as special cases of our general RMD framework under specific convenient choices of the underlying conditional distribution of the missing or misspecified

observations.

On the empirical side, we show consistently attractive performance of our RMD framework and the resulting out-of-sample forecasts in popular state space models for extracting inflation trends along with model extensions that more directly reflect inflation targeting by central banks. First, we demonstrate that RMD-N and RMD-X augmented versions of the classical UC model offer a viable alternative to the well-established unobserved components/stochastic volatility outlier-adjustment (UCSVO) model that was put forward by Stock and Watson (2016) and can improve forecasting average inflation, especially over longer horizons such as twelve quarters ahead. Second, we document even greater forecast improvements over a UC model with t-distributed innovations without stochastic volatility (UC-T), as another indication that the filtering and forecasting benefits offered by RMD cannot be attained by simply imposing a heavier-tailed measurement distribution.

Finally, we also consider a more general AR model for the unobserved inflation component which allows for reversion to a long-run mean either estimated from the data or fixed at the inflation target rate of 2% as a way to more directly reflect central bank inflation targeting that has been in place for many years. The much superior forecasting performance we document when the mean is fixed to the *a priori* 2% inflation target in place suggests that inflation targeting has been effective and the RMD framework works best when paired with additional economic theory sources of information about the underlying model. This further underscores that expert modelling can still be important for successful application of machine learning techniques by leveraging also more specific theory-implied features of the data generating process. That said, both when applied to the classical UC and better-performing AR models with fixed inflation target mean our RMD approach to robust filtering and forecasting offers attractive empirical performance gains and ease of implementation in comparison to existing alternatives.

Overall, the proposed RMD framework for robust filtering and forecasting shows

promising avenues for further exploration on both the theory and empirical side in a host of other areas of applications of state-space modelling and forecasting in the presence of data imperfections. Ongoing work in progress suggests clear value added of the RMD framework also in mainstream finance applications such as return density forecasting based on popular parametric and semi-parametric models.

Acknowledgements

References

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2):177–190.

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2014). A robust neighborhood truncation approach to estimation of integrated quarticity. *Econometric Theory*, 30(1):3–59.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280.

Barndorff-Nielsen, O. E. and Shephard, N. (2005a). How accurate is the asymptotic approximation to the distribution of realized variance. *Identification and inference for econometric models. A Festschrift in honour of TJ Rothenberg*, pages 306–311.

Barndorff-Nielsen, O. E. and Shephard, N. (2005b). Variation, jumps, market frictions and high frequency data in financial econometrics. In *Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress*. Cambridge University Press.

Barndorff-Nielsen, O. E., Shephard, N., and Winkel, M. (2006). Limit theorems for multipower variation in the presence of jumps. *Stochastic processes and their applications*, 116(5):796–806.

Bergmeir, C., Hyndman, R., and Benítez, J. (2016). Bagging exponential smoothing methods using stl decomposition and box-cox transformation. *International Journal of Forecasting*, 32(2):303 – 312.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Calvet, L. E., Czellar, V., and Ronchetti, E. (2015). Robust filtering. *Journal of the American Statistical Association*, 110(512):1591–1606.

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). SMC2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society*, 75(3):397–426.

Crevits, R. and Croux, C. (2017). Robust estimation of linear state space models.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-gaussian observations based on state space models from both classical and bayesian perspectives. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):3–56.

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical Foundations of Data Science*. Chapman & Hall.

Fan, J., Ma, C., Wang, K., and Zhu, Z. (2021). Modern data modeling: Cross-fertilization of the two cultures. *Observational Studies*, 7(1):65–76.

Faust, J. and Wright, J. H. (2013). Forecasting inflation. volume 2, chapter Chapter 1, pages 2–56. Elsevier.

Federal Reserve Board (2015). Why does the federal reserve aim for 2 percent inflation over time? https://www.federalreserve.gov/faqs/economy_14400.htm. Last Update: 2015-01-26.

Hamilton, J. D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 33(3):387–397.

Harvey, A. and Luati, A. (2014). Filtering with heavy tails. *Journal of the American Statistical Association*, 109(507):1112–1122.

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2):247–264.

Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10:1425–1433.

Inoue, A. and Kilian, L. (2008). How useful is bagging in forecasting economic time series? a case study of u.s. consumer price inflation. *Journal of the American Statistical Association*, 103(482):511–522.

James, G. (2003). Variance and bias for general loss functions. *Machine Learning*, 51:115—-135.

James, G. and Hastie, T. (1997). Generalizations of the bias/variance decomposition for prediction error. Technical report.

Kitagawa, G. (1989). Non-gaussian seasonal adjustment. *Computers & Mathematics with Applications*, 18(6-7):503–514.

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo methods in practice*, pages 197–223.

Maìz, C., Molanes-Lòpez, E., Mìguez, J., and Djurìc, P. (2012). A particle filtering scheme for processing time series corrupted by outliers. *IEEE Transactions on Signal Processing*, 60(9):4611–4627.

Shephard, N. (1996). Statistical aspects of arch and stochastic volatility. *Monographs on Statistics and Applied Probability*, 65:1–68.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50:665–690.

Sims, C. A. (2011). Rational inattention and monetary economics. In *Handbook of Monetary Economics*, volume 3A, chapter 4, pages 155–181. Elsevier.

Stock, J. H. and Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39(s1):3–33.

Stock, J. H. and Watson, M. W. (2016). Core inflation and trend inflation. *Review of Economics and Statistics*, 98(4):770–784.

Stoffer, D. S. and Wall, K. D. (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and the kalman filter. *Journal of the American Statistical Association*, 86(416):1024–1033.

Vos, P. and Wu, Q. (2015). Maximum likelihood estimators uniformly minimize distribution variance among distribution unbiased estimators in exponential families. *Bernoulli*, 21(4):2120–2138.

Wu, Q. and Vos, P. (2012). Decomposition of Kullback–Leibler risk and unbiasedness for parameter-free estimators. *Journal of Statistical Planning and Inference*, 142:1525–1536.

Appendix A RMD: Posterior distribution

We start with the most general form for $y_t^{(j)} \sim f_j^t$, where $y_t^{(j)}$ is independent of θ and x^T conditional on all information up to time $t-1$ (i.e. $\hat{y}^{t-1}$). Within a Bayesian paradigm, independence from θ is well-defined through the choice of prior. The goal is to sequentially infer the joint posterior distribution $P(\theta, x^T, C^T | \hat{y}^T, \beta)$ as observations are added. Specifically, we show how one can update the posterior $P(x^{t-1}, \theta | \hat{y}^{t-1}, \beta)$ after observing $\hat{y}_t$ to obtain $P(x^t, \theta | \hat{y}^t, \beta)$. From here on, we assume that the unconditional probability $P(C_t = 1)$ is β and that C_t is independent of all past random variables and x_t, θ. Thus, $P(C_t = 1 | \hat{y}^{t-1}, x^t, \theta, \beta) = \beta$. We then derive the posterior distribution of $P(C_t = 1 | \hat{y}^t, \beta)$ using Bayes' rule:

$$P(C_t = 1 | \hat{y}^t, \beta) = \frac{\beta P(\hat{y}_t | C_t = 1, \hat{y}^{t-1})}{\beta P(\hat{y}_t | C_t = 1, \hat{y}^{t-1}\beta) + (1 - \beta)P(\hat{y}_t | \hat{y}^{t-1}, \beta)} \tag{A.1}$$

From the setup, the distribution of $\hat{y}_t$ conditional on $(C_t, x^t, \theta, \hat{y}^{t-1}, \beta)$ is either f_θ if $C_t = 1$ or f_j^t if $C_t = 0$. Define $F(y_t | \hat{y}^{t-1}, \beta) := \int_{x_t, \theta} f_\theta(y_t | x_t) dP(x_t, \theta | \hat{y}^{t-1}, \beta)$, which is the one-step-ahead predictive distribution of y_t given $\hat{y}^{t-1}$ and β. Equation A.1 can be rewritten directly using F and f_j.

$$P(C_t = 1 | \hat{y}^t, \beta) = \frac{\beta F(\hat{y}_t | \hat{y}^{t-1}, \beta)}{\beta F(\hat{y}_t | \hat{y}^{t-1}, \beta) + (1 - \beta)f_j(\hat{y}_t | \hat{y}^{t-1}, \beta)} =: \hat{\beta}(\hat{y}^t) \tag{A.1'}$$

We then focus on the conditional posterior distribution $P(x^t, \theta | C_t = i, \hat{y}^t, x^t, \theta, \beta)$ for $i \in \{0, 1\}$. To do so, we consider the joint distribution with the true measurement y_t included as well.

$$P(x^t, \theta | C_t, \hat{y}^t, \beta) = \int_{y_t} P(x^t, \theta | y_t, C_t, \hat{y}^t, \beta) dP(y_t | C_t, \hat{y}^t, \beta)$$

$$= \int_{y_t} P(x^t, \theta | y_t, C_t, \hat{y}^{t-1}, \beta) dP(y_t | C_t, \hat{y}^t, \beta)$$

$$= \left(\int_{y_t} \frac{f_\theta(y_t | x_t, \theta)}{F(y_t | \hat{y}^{t-1}, \beta)} dP(y_t | C_t, \hat{y}^t, \beta) \right) P(x^t, \theta | C_t, \hat{y}^{t-1}, \beta) \qquad \text{(A.2)}$$

$$= \left(\int_{y_t} \frac{f_\theta(y_t | x_t, \theta)}{F(y_t | \hat{y}^{t-1}, \beta)} dP(y_t | C_t, \hat{y}^t, \beta) \right) P(x^t, \theta | \hat{y}^{t-1}, \beta)$$

The value of the integral depends on the value of C_t. If $C_t = 1$, then $y_t = \hat{y}_t$ almost surely, so $P(y_t | C_t = 1, \hat{y}^t)$ equals the Dirac probability measure at $\hat{y}_t$. Consequently, the integral will evaluate to $\frac{f_\theta(\hat{y}_t | x_t, \theta)}{F(\hat{y}_t | \hat{y}^{t-1}, \beta)}$. On the other hand, if $C_t = 0$, then, because $\hat{y}_t = y_t^{(j)}$ almost surely and $y_t^{(j)}$ is conditionally independent of y_t, $P(y_t | C_t = 0, \hat{y}_t, \beta) = P(y_t | \hat{y}^{t-1}, \beta) = F(y_t | \hat{y}^{t-1}, \beta)$. Since this cancels out with the denominator, and the numerator is a probability density, the integral will evaluate to 1. We then substitute these values into Equation A.2:

$$P(x^t, \theta | C_t, \hat{y}^t, \beta) = \begin{cases} \frac{f_\theta(\hat{y}_t | x_t, \theta)}{F(\hat{y}_t | \hat{y}^{t-1}, \beta)} P(x^t, \theta | \hat{y}^{t-1}, \beta) & \text{if } C_t = 1 \\[2ex] P(x^t, \theta | \hat{y}^{t-1}, \beta) & \text{if } C_t = 0 \end{cases} \qquad \text{(A.2')}$$

These findings confirm the intuition described earlier. If $C_t = 1$, then the observed $\hat{y}_t$ is treated just like it were the true measurement y_t. If $C_t = 0$, then $\hat{y}_t$ is discarded and the posterior belief about x^t and θ remains unchanged. Combining Equations A.1' and A.2', the full update of the posterior distribution of x^t, θ with the information $\hat{y}_t$ is characterized below:

$$P(x^t, \theta | \hat{y}^t, \beta) = \left(\hat{\beta}(\hat{y}^t) \frac{f_\theta(\hat{y}_t | x_t)}{F(\hat{y}_t | \hat{y}^{t-1}, \beta)} + (1 - \hat{\beta}(\hat{y}^t)) \right) g_\theta(x_t | x_{t-1}) P(x^{t-1}, \theta | \hat{y}^{t-1}, \beta) \qquad \text{(A.3)}$$

Recursing back to $t = 0$, the posterior distribution can be written as:

$$P(x^T, \theta | \hat{y}^T, \beta) = \prod_{t=1}^{T} \left(\hat{\beta}(\hat{y}^t) \frac{f_\theta(\hat{y}_t | x_t)}{F(\hat{y}_t | \hat{y}^{t-1}, \beta)} + (1 - \hat{\beta}(\hat{y}^t)) \right) g_\theta(x_t | x_{t-1}) \mu_\theta(x_0) P(\theta) \qquad \text{(A.4)}$$

We notice that the posterior distribution as characterized in Equations (A.3) and (A.4) resembles how the posterior distribution of the originally specified model would update with new information, and are in fact equal if $\beta = 1$. For $\beta \in (0, 1)$, we see that the resulting posterior distribution at each stage is exactly a mixture between including or excluding the observed $\hat{y}_t$, with weights governed by the posterior probability $\hat{\beta}$. This shows that the RMD framework introduces a natural and interpretable way to incorporate misspecfied measurements into a model by way of downweighting via randomization, with the degree to which a data point is added to the model given exactly by the belief that it corresponds to the correctly specified measurement equation.