

# Estimating Latent Base Demand Using the Kalman Filter

Shahr Bano Bokhari (MSDS25027)

## 1 Introduction

This report documents an end-to-end workflow to estimate **latent base demand** from observed daily *Units Sold* using a Kalman Filter (KF). The pipeline includes:

- Exploratory data analysis (EDA) on the full dataset.
- Subsetting to Store 5 and examining demand structure.
- Attempting linear regression to learn a transition model.
- Implementing a baseline KF with  $F = 1$  (random walk).
- Recovering latent increments to identify possible control variables.
- Using L1/L2 regularization to select variables that drive latent transitions.
- Re-estimating a KF with selected control inputs.
- Hyperparameter tuning of  $Q$  and  $R$  using the NIS statistic.

## 2 Dataset Overview and EDA

The dataset consists of 73,100 transactions across 20 products and 5 stores over 2 years. After cleaning, categorical fields were encoded, dates were parsed, and numerical distributions inspected.

## 2.1 Key Findings

- **Units Sold** is strongly right-skewed; most values lie below 100.
- **Demand Forecast** is highly correlated with actual sales ( $r = 0.9968$ ), indicating a very accurate forecasting rule.
- **Price** and **Competitor Pricing** are almost perfectly correlated, yet neither predicts sales well.
- **Inventory Level** correlates moderately with sales ( $r \approx 0.59$ ).
- Most categorical variables (*Seasonality, Weather, Discount*) produce only mild variation.

Based on these, we expect latent demand to be smooth with occasional shocks.

## 3 Store 5 Exploratory Analysis

Store 5 (S005) shows stable and high demand across categories, making it a suitable subset for latent demand estimation. Key findings from cross-tab analyses are summarized below.

### 3.1 Category-Level Patterns

Category averages are tightly clustered (135–140 units), indicating uniform performance and no outliers at the category level.

### 3.2 Product Effects

Product-level deviations are large compared to category-level variation:

- Some products (e.g., Electronics P0020, Furniture P0015) are far above their category means.
- Others (e.g., Groceries P0012) fall significantly below.

Thus, product granularity is essential for modeling demand.

### 3.3 Regional Influences

Geography meaningfully shapes demand: Clothing performs better in East/South, Electronics peak in North, and Furniture is lowest in West.

### 3.4 Discount Effects

Discount responsiveness varies by category: Electronics and Furniture respond strongly to 20% discounts, while Groceries and Clothing show weaker or inconsistent effects. Toys may decrease under higher discounts.

### 3.5 Weather Effects

Weather-driven patterns include: Clothing stronger on Sunny days, Groceries stronger on Cloudy days, and Electronics peaking under Sunny/Snowy conditions.

### 3.6 Holiday and Seasonal Patterns

Holidays shift demand differently by category (e.g., Clothing and Groceries increase, Electronics decrease). Seasonality shows strong structure: Clothing peaks in Summer, Electronics and Toys peak in Winter, and Groceries peak in Autumn.

### 3.7 Summary

- Store 5 is stable and representative for modeling.
- Product-level variation dominates category-level trends.
- Region, seasonality, weather, and promotions show structured, interpretable effects.

## Kalman Filter Performance Metric: NIS

The **Normalized Innovation Squared (NIS)** is used to assess the calibration of the Kalman Filter. It quantifies how well the filter's predicted observations match the actual measurements.

$$\text{NIS}_t = \nu_t^\top S_t^{-1} \nu_t, \quad \nu_t = z_t - \hat{z}_{t|t-1}, \quad S_t = H P_{t|t-1} H^\top + R$$

where  $\nu_t$  is the innovation (prediction error),  $S_t$  its covariance, and  $H$ ,  $P_{t|t-1}$ ,  $R$  are standard KF matrices.

For a 1D observation ( $H = 1$ ), the NIS follows a chi-squared distribution with 1 degree of freedom:

$$\text{NIS}_t \sim \chi_1^2, \quad E[\text{NIS}_t] = 1, \quad \text{Std}[\text{NIS}_t] = \sqrt{2} \approx 1.414$$

#### Interpretation in our context:

- Mean NIS  $\approx 1$  indicates well-calibrated filter predictions.
- NIS consistently  $> 4$  indicates an overconfident filter (uncertainties underestimated).
- NIS consistently  $< 1$  indicates an underconfident filter (uncertainties overestimated).

## 4 Attempt to Estimate Transition Dynamics via Regression

To estimate the state transition matrix  $F$  in

$$x_t = Fx_{t-1} + Bu_t + w_t,$$

I regressed observed sales  $x_t$  on lagged values  $x_{t-1}$  and candidate covariates (e.g., Inventory Level, product indicators) using OLS, Ridge, and Lasso:

$$\hat{x}_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 \text{InventoryLevel}_t + \dots$$

Key findings:

- The linear models achieved modest fit ( $R^2 \approx 0.35$ ), indicating weak predictive power.
- Ridge and Lasso revealed only a few non-zero coefficients, with  $x_{t-1}$  itself showing limited influence.

- The weak linear relationship suggests that a deterministic  $F$  cannot adequately capture latent dynamics.

Consequently, regression-based estimation of  $F$  and input coefficients was abandoned in favor of a baseline random-walk Kalman Filter.

## 5 Baseline Kalman Filter

Given the weak regression relationship between  $x_t$  and  $x_{t-1}$ , a simple random-walk model was adopted as the baseline:

$$x_t = x_{t-1} + w_t, \quad F = 1,$$

with measurement model

$$z_t = Hx_t + v_t, \quad H = 1,$$

and process and measurement noise variances  $Q$  and  $R$ .

### 5.1 Baseline KF Results

The baseline Kalman Filter was run using initial estimates:

$$R = 74.49, \quad Q = 23840.07$$

yielding:

- Kalman Gain: [0.9979, 0.9969, ...]
- Posterior covariance  $P$ : [35830.75, 23914.40, ...]
- Mean innovation: -0.0065
- Std innovation: 154.16
- Mean NIS: 0.991
- Std NIS: 1.401
- MSE: 0.229
- RMSE: 0.479

## 5.2 Interpretation

- Even without hyperparameter tuning, the baseline KF provides a reasonable estimate of latent demand.
- The mean NIS ( $\approx 1$ ) indicates the filter is fairly well-calibrated, although the slightly high standard deviation of NIS suggests occasional deviations in prediction uncertainty.
- The Kalman Gain values near 1 reflect that the filter heavily weights incoming observations relative to prior estimates, consistent with the random-walk assumption.
- MSE and RMSE values indicate a moderate level of prediction error, leaving room for improvement through the inclusion of informative control inputs or state augmentation.

## 5.3 Latent Increments: Extracting Prior and Posterior Estimates

After running the baseline KF, prior and posterior state estimates were collected:

$$x_{t|t-1} = \text{prior}, \quad x_{t|t} = \text{posterior}.$$

Latent increments were computed:

$$\Delta x_t = x_{t|t} - x_{t|t-1}.$$

These increments represent shocks that the model could not explain using  $F = 1$  alone.

## 6 Learning Control Inputs from Latent Increments

Since

$$\Delta x_t \approx Bu_t + w_t,$$

we regressed  $\Delta x_t$  on all candidate explanatory variables:

$$\Delta x_t = \gamma_0 + \gamma^\top u_t + \epsilon_t.$$

Both L1 (Lasso) and L2 (Ridge) were used:

## 7 Kalman Filter with Control Inputs

To investigate whether external covariates could explain short-term variations in the latent demand process, we regressed the latent increments  $\Delta x_t = x_{t|t} - x_{t|t-1}$  on all available features using both Ridge and Lasso regularization. The resulting coefficients were used to identify potential control inputs for an exogenous Kalman Filter.

## 8 Ridge Regression for Control Input Selection

Ridge regression was first applied to the latent increments. Using a threshold of  $|\beta_j| > 2$ , the following variables were selected:

```
[ 'Inventory_Level', 'Product_ID_P0015', 'Product_ID_P0014',
  'Weather_Condition_Rainy', 'Product_ID_P0019', 'Category_Furniture',
  'Product_ID_P0003', 'Product_ID_P0012', 'Product_ID_P0020',
  'Category_Groceries', 'Product_ID_P0006', 'Product_ID_P0010',
  'Product_ID_P0002', 'Region_South', 'Product_ID_P0013', 'Product_ID_P0005',
  'Seasonality_Spring', 'Weather_Condition_Sunny', 'Region_West',
  'Category_Toys', 'Holiday/Promotion_True', 'Weather_Condition_Snowy',
  'Price' ]
```

### 8.1 Ridge Coefficient Path Interpretation

Although many coefficients remain technically nonzero, the coefficient paths reveal that only two predictors demonstrate substantial magnitude at small  $\lambda$  values. The remaining coefficients stay near zero and shift only under very large penalties ( $\lambda \approx 10^4\text{--}10^6$ ).

The  $\ell_2$  norm decreases smoothly from the OLS level ( $\|\beta\|_2 \approx 70$ ) without indicating a richer linear structure. This suggests that **linear effects of the**

**raw covariates on  $\Delta x_t$  are weak.** However, this does not imply that the features lack predictive value.

## 8.2 Do Ridge-Based Inputs Improve the KF?

Given the model structure and the EDA's nonlinear patterns, only modest improvements over the baseline KF are expected. Empirically:

Metric	Baseline KF	Ridge-Control KF
Mean NIS	0.99	0.81
RMSE	0.479	0.433
Innovation Std	154.16	139.56

Ridge yields slightly better calibration and slightly lower RMSE, but the gains are modest.

## 9 Lasso Regression for Control Input Selection

Lasso regression was also applied to the latent increments to identify sparse and interpretable control inputs. Using the same  $|\beta_j| > 2$  threshold, only a single predictor was selected:

[‘Inventory Level’]

### 9.1 Lasso Coefficient Path Interpretation

The Lasso paths show a clear sparsity pattern:

- most coefficients remain at zero across all  $\lambda$ ,
- one dominant coefficient starts near  $\beta \approx 20$  and shrinks around  $\lambda \approx 1$ ,
- another moderate coefficient begins above 60 and decreases around  $\lambda \approx 10^2$ .

The shrinkage-level curve  $\|\beta_{\text{lasso}}(\lambda)\|_1 / \|\beta_{\text{OLS}}\|_1$  and the declining  $\ell_2$  norm confirm progressive regularization and effective feature selection. Lasso therefore identifies the dominant linear driver of latent increments while pushing weaker features to zero.

## 9.2 Do Lasso-Based Inputs Improve the KF?

As with Ridge, Lasso provides modest improvements relative to the baseline:

Metric	Baseline KF	Lasso-Control KF
Mean NIS	0.99	0.81
RMSE	0.479	0.434
Innovation Std	154.16	139.68

Lasso and Ridge KFs perform very similarly, reflecting the limited linear structure in the latent increments.

## 9.3 Comparison of Baseline, Ridge, and Lasso Kalman Filters

To assess the impact of incorporating control inputs derived from Ridge and Lasso regularization, the baseline random-walk KF was compared to Ridge- and Lasso-controlled variants. The key metrics are summarized below:

KF Variant	Mean Innovation	Std Innovation	Mean NIS	Std NIS	MSE	RMSE
Baseline	-0.0065	154.16	0.991	1.401	0.229	0.479
Ridge-Control	-1.028	139.56	0.812	1.144	0.188	0.433
Lasso-Control	0.293	139.68	0.813	1.141	0.188	0.434

### Interpretation:

- Both Ridge- and Lasso-controlled KFs show improved performance over the baseline in terms of RMSE and MSE, indicating that including covariates helps explain some short-term variations in latent demand.
- The mean NIS decreases from  $\sim 0.99$  (baseline) to  $\sim 0.81$  for both Ridge and Lasso, suggesting slightly better calibration and more realistic estimation of prediction uncertainty.
- Despite Ridge including a larger set of predictors, its performance is almost identical to the Lasso variant, which selected only **Inventory Level** as a control input. This indicates that a single, strong predictor can be as effective as a larger set of weak predictors in a linear KF.

- The small differences in mean innovation and standard deviation of innovation show that Ridge slightly reduces volatility, but both linear-controlled KFs remain limited by the assumption of linear additive effects.
- Overall, Lasso's sparsity leads to a more interpretable model without loss of performance, highlighting that **parsimonious models can capture most of the linear structure in latent increments**.

## 9.4 Implication

Across both Ridge and Lasso analyses, the limited improvements should not be interpreted as evidence that the covariates lack predictive value. Instead, they highlight the limitations of purely linear, additive models applied to a dataset whose EDA revealed strong **nonlinear** and **interaction-driven** patterns.

*The Ridge and Lasso results reflect the constraints of linear modeling—not feature irrelevance. The Store 5 analysis strongly suggests that nonlinearities and cross-feature interactions (e.g., season × category, weather × product, region × product) likely contain meaningful predictive structure that linear models cannot capture.*

Thus, expanding the feature space through interaction terms, polynomial features, spline bases, or nonlinear feature selection methods (e.g., tree-based models) may provide substantially richer control inputs for the Kalman Filter. The modest gains observed in the linear models are consistent with only a small portion of the underlying structure being accessible through linear effects alone.

## 10 Kalman Filter with Augmented State: Units Sold & Inventory Level

After identifying **Units Sold** and **Inventory Level** as the main control inputs from the Lasso analysis, I implemented a Kalman Filter using these two variables as the state vector:

$$\mathbf{x}_t = \begin{bmatrix} \text{Units Sold}_t \\ \text{Inventory Level}_t \end{bmatrix}.$$

## 10.1 Initial KF Results

Using the initial estimates for  $Q$  and  $R$ :

$$R = [[74.4898]], \quad Q = \begin{bmatrix} 23840.069 & 0 \\ 0 & 2.0023 \end{bmatrix},$$

the KF produced:

- Kalman Gain: [0.9981, 0.9979, ...]
- Covariance  $P$ : [39895.41, 35708.29, ...]
- Mean innovation: 0.292
- Std innovation: 139.75
- Mean NIS: 0.458
- Std NIS: 0.642
- MSE: 0.0596
- RMSE: 0.244

## 10.2 Hyperparameter Tuning of $Q$ and $R$

Several grids of  $Q$  and  $R$  scaling factors were explored to improve filter calibration and prediction accuracy. Key findings:

- Top candidates after multiple tuning attempts:
  - $Q_{\text{scale}} = 0.5$ ,  $R_{\text{scale}} = 0.13$ , mean NIS = 0.921, RMSE = 0.064
  - $Q_{\text{scale}} = 1$ ,  $R_{\text{scale}} = 0.12$ , mean NIS = 0.461, RMSE = 0.029
  - $Q_{\text{scale}} = 3$ ,  $R_{\text{scale}} = 0.1$ , mean NIS = 0.154, RMSE = 0.008
- The final choice was:

$$Q_{\text{scale}} = 0.5, \quad R_{\text{scale}} = 0.13$$

### 10.3 Final Augmented KF Results

Using the tuned parameters, the augmented KF produced:

- $R = [[9.6837]]$ ,  $Q = \begin{bmatrix} 11920.03 & 0 \\ 0 & 1.0011 \end{bmatrix}$
- Kalman Gain:  $[0.99965, 0.99950, \dots]$
- Covariance  $P$ :  $[27975.38, 19470.53, \dots]$
- Mean innovation: 0.292
- Std innovation: 139.83
- Mean NIS: 0.921
- Std NIS: 1.291
- MSE: 0.00407
- RMSE: 0.064

### 10.4 Comparison of All KF Variants

KF Variant	Mean NIS	Std NIS	MSE	RMSE
Baseline	0.991	1.401	0.229	0.479
Ridge-Control	0.812	1.144	0.188	0.433
Lasso-Control	0.813	1.141	0.188	0.434
Augmented Lasso KF	0.921	1.291	0.0041	0.064

### 10.5 Interpretation

- Augmenting the state with Units Sold and Inventory Level dramatically reduced both RMSE and MSE.
- The tuned  $Q$  and  $R$  improved NIS calibration, bringing mean NIS closer to 1.
- Compared to Ridge and Lasso linear KF variants, the augmented KF captures latent dynamics much more accurately, confirming the predictive importance of Inventory Level and interaction with Units Sold.

## 11 Final Conclusion

This study demonstrates a systematic approach to estimating latent base demand using the Kalman Filter across multiple modeling strategies. Key takeaways include:

- **Baseline KF:** A simple random-walk filter provides a reasonable approximation of latent demand but exhibits relatively high RMSE and suboptimal NIS calibration.
- **Linear KF with Control Inputs:** Ridge and Lasso regularization identified Inventory Level and a few product-specific variables as potential linear drivers. Incorporating these as control inputs improved RMSE and NIS modestly, yet the linear assumptions limited the ability to fully capture latent dynamics.
- **Augmented KF with Units Sold and Inventory Level:** Expanding the state to include Units Sold alongside Inventory Level, combined with careful tuning of  $Q$  and  $R$ , yielded substantial performance gains. RMSE dropped dramatically (from  $\sim 0.43$  to  $\sim 0.064$ ), and mean NIS approached the ideal value of 1, indicating a well-calibrated filter.
- **Implications:** The results confirm that latent demand is strongly influenced by Inventory Level and its interaction with observed sales. Linear additive models capture only part of the underlying structure; augmenting the state vector allows the KF to account for richer dynamics without explicitly modeling nonlinearities.
- **Future Directions:** Further improvements could involve hierarchical or multi-store state-space models, explicit modeling of nonlinear interactions, or joint EM-based estimation of  $F$ ,  $Q$ ,  $R$ , and  $B$  matrices to fully exploit the data structure.

Overall, the augmented Kalman Filter approach provides an effective and interpretable framework for recovering latent demand, with clear evidence that incorporating relevant covariates directly into the state vector can dramatically enhance predictive accuracy and filter calibration.