Quality-Adjusted Price Indices Powered by AI

P. Bajari, Z. Cen, V. Chernozhukov (MIT), R. Huerta (UCSD),
J. Li, M. Manukonda, G. Monokrousos
Amazon Core AI

Motivation

- Inflation indices are important inputs into measuring aggregate productivity and cost of living, and conducting monetary and economic policy.
- We want to contribute to the science of inflation measurement based on **quality-adjusted (hedonic)** prices.
- Main challenges today:
    1. millions of products (global trade environment);
    2. prices change quite often;
    3. extremely high turnover for some products (e.g., apparel, electronics).
- Our teams addressed these challenges to produce a method that utilizes scalable **AI and Econometrics** tools to predict quality-adjusted prices using text and image embeddings

We want to share our findings:

- 1. Deep learning embeddings, produced by Deep Neural Networks (DNN) provide valuable input features for hedonic models.
- 2. Deep learning (utilizing DNN) leads superior price prediction compared to ML and other methods.

Outline

1) Basic Price Indices
2) Quality-Adjusted (**Hedonic**) Price Indices
3) Hedonic Prices Indices Using ML and AI
    1) Feature Engineering from Text
    2) Feature Engineering from Images
    3) Nonlinear Price Prediction using Neural Networks
4) Conclusion

Transaction-Price Quantity Index (TPQI)

- Price P_{jt} and quantity Q_{jt} for product *j* in period *t*
- Transaction-Price Quantity Indices are based on **matching**:

Paasche Index: $$R_t^{P,M} = \frac{\sum P_{jt} Q_{jt}}{\sum P_{j(t-1)} Q_{jt}}$$

Laspeyres Index: $$R_t^{L,M} = \frac{\sum P_{jt} Q_{j(t-1)}}{\sum P_{j(t-1)} Q_{j(t-1)}}$$

Fisher Index: $$R_t^{F,M} = \sqrt{R_t^{P,M} R_t^{L,M}}$$

where the summation in the denominator/numerator is over the matching set (largest common set).

- Missing products **create biases in the matching set**.

Need for Hedonics (Quality-Adjusted Pricing)

- To avoid biases in the **matching set**, we can predict prices of missing products in period-to-period comparisons. Relevant for product categories with high turn-over.
- In product groups like apparel, **about 50% of products get replaced** with new products **every month**.
- **Matching sets** can be non-representative of good baskets, creating systematic biases.
- **Use predicted prices**, **using product attributes or qualities**, instead of the missing and observed prices

Hedonic Price Quantity Index

- **Replace prices by quality-adjusted prices**

Paasche Index: $$\hat{R}_t^{P,M} = \frac{\sum \hat{P}_{jt} Q_{jt}}{\sum \hat{P}_{j(t-1)} Q_{jt}}$$

Laspeyres Index: $$\hat{R}_t^{P,M} = \frac{\sum \hat{P}_{jt} Q_{j(t-1)}}{\sum \hat{P}_{j(t-1)} Q_{j(t-1)}}$$

Fisher Index: $$\hat{R}_t^{F,M} = \sqrt{\hat{R}_t^{P,M} \hat{R}_t^{L,M}}$$

- **Summation is done over the union of products that transacted in two periods**

The Hedonic Price Model: Predictive Model for Price Using Product Features

The hedonic model is the predictive model for price given the product features:

$$\underbrace{P_{jt}}_{Y_{jt}} = P_t(\underbrace{W_{jt}, I_{jt}}_{X_{jt}}) + \epsilon_{jt},$$

where P_{jt} is the price of product j at time t, X_{jt} are the product embedding features, and the pricing function $x \mapsto P_t(x)$ can change from period to period, reflecting the fact that product attributes/features may be valued differently in different periods.

What are the Features?

Query: red dress

Customer behavior data

3,989 customer reviews | 259 answered questions

Image

$X \in \mathbb{R}^d$

Description

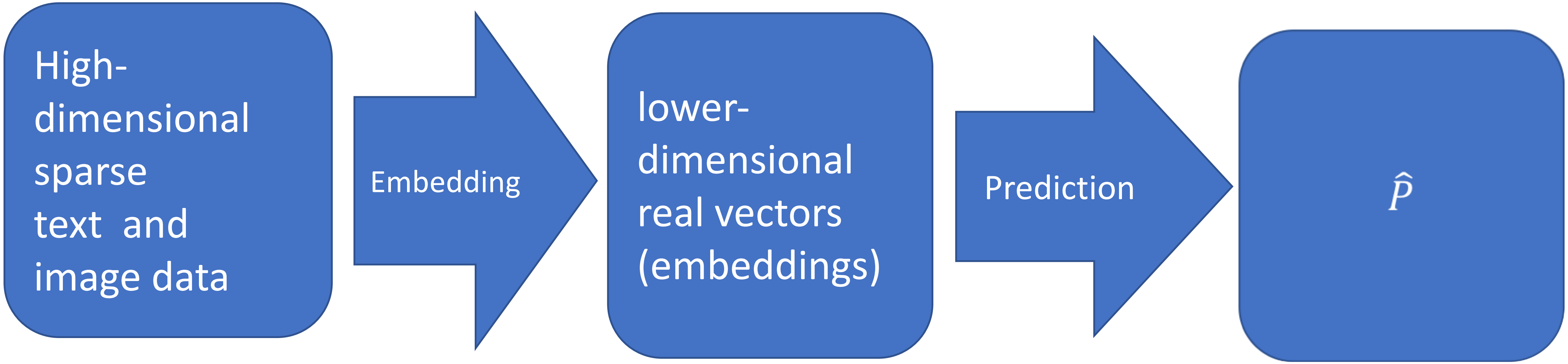
- Material - Cotton & Spandex.
- Imported
- Classic and Iconic Audrey Hepburn 50s Vintage Solid Color Swing Dress, Put on and Show Your Elegance and Charm.
- Features: Boat Neckline; Sleeveless; Full Circle Swing; Quick Access Zipper for Easy On and Off
- It's Great Choice for Daily Casual, Wedding , Ball, Party, Banquet and Other Occasion.
- [Size Chart] PLEASE Make Sure Your Measurements and Compare to the Size Chart From the picture on the left side or in the Following Description.
- Hand Wash Carefully,Low Temperature for Washing,Can not High Temperature Ironing, Line Dry

Title

Anni Coco

Anni Coco Women's Classy Audrey Hepburn 1950s Vintage Rockabilly Swing Dress

The Process: Embeddings and Predictions



- **Embedding** is done by Deep Learning (Neural Network) methods:

a) Text (*W*) : ***ELMo (*****Bert, W2Vec)**

b) Images (*I*) : ***ResNet*** **(AlexNet, GoogLeNet)**

- **Prediction** is done again by Neural Networks (other regression methods).

The Benefits of Text and Image Features in Predicting Prices/Hedonic Regression

- Using only conventional features in linear regression gives ***R^2* for predicting log-price around 30%.**

- Using ***W*** and ***I*** features in **Linear Regression** gives ***R^2* about 50%.**

- Using ***W* and *I* features** plus **Random Forest** brings ***R^2* = 60-65%.**

- Using ***W* and *I* features** plus **Neural Networks (NN)** brings ***R^2* = 75-90%.**

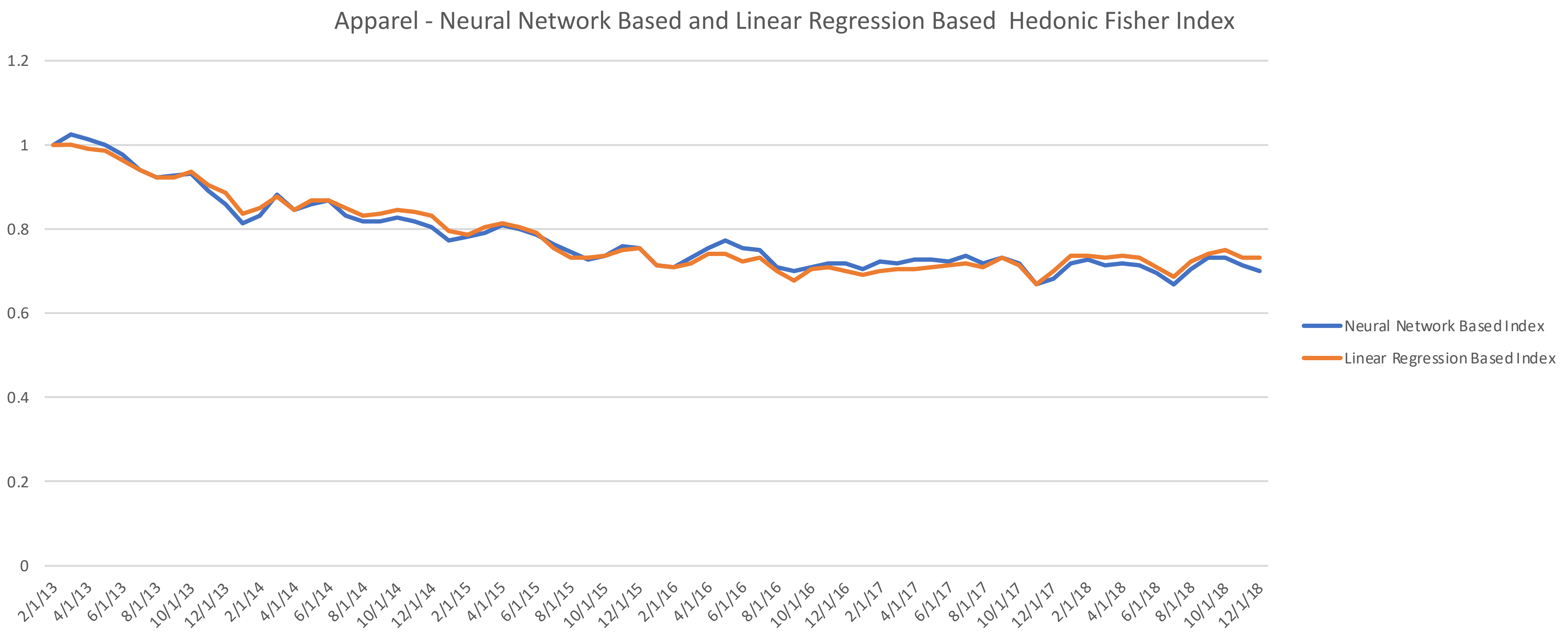
- All *R^2* results are out-of-sample.

Performance Across Product Categories Varies

- **Best Cases:** Apparel, Electronics, Baby; Neural Network R-square > 75%

- **Tough Cases:** Toys, Books, Shoes; Neural Network R-square ranges from 40 to 60 %

(Preliminary) Empirical Result: Amazon Hedonic Index for Apparel Product Group



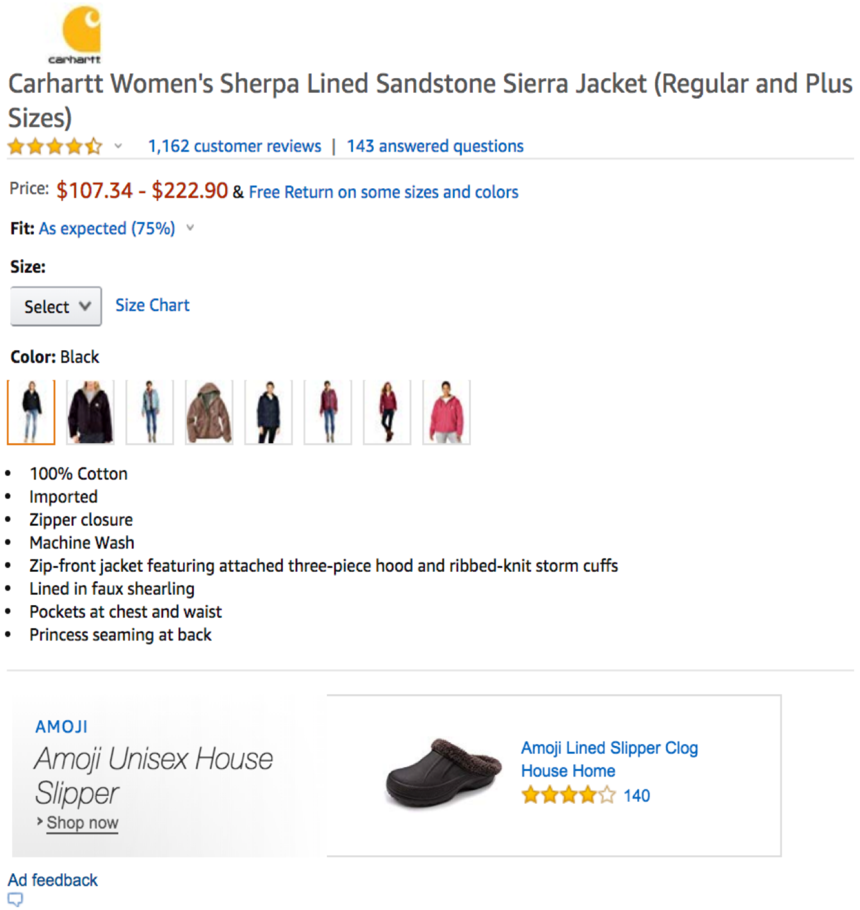
The Index is Chained Month-over-Month

- For comparison, the non-hedonic Fisher index *decreases much more* over the period 2013-2018.

- The simple hedonic index, that is based on simple catalogue information, also decreases much more than the hedonic index (but less drastically than the non-hedonic Fisher index).

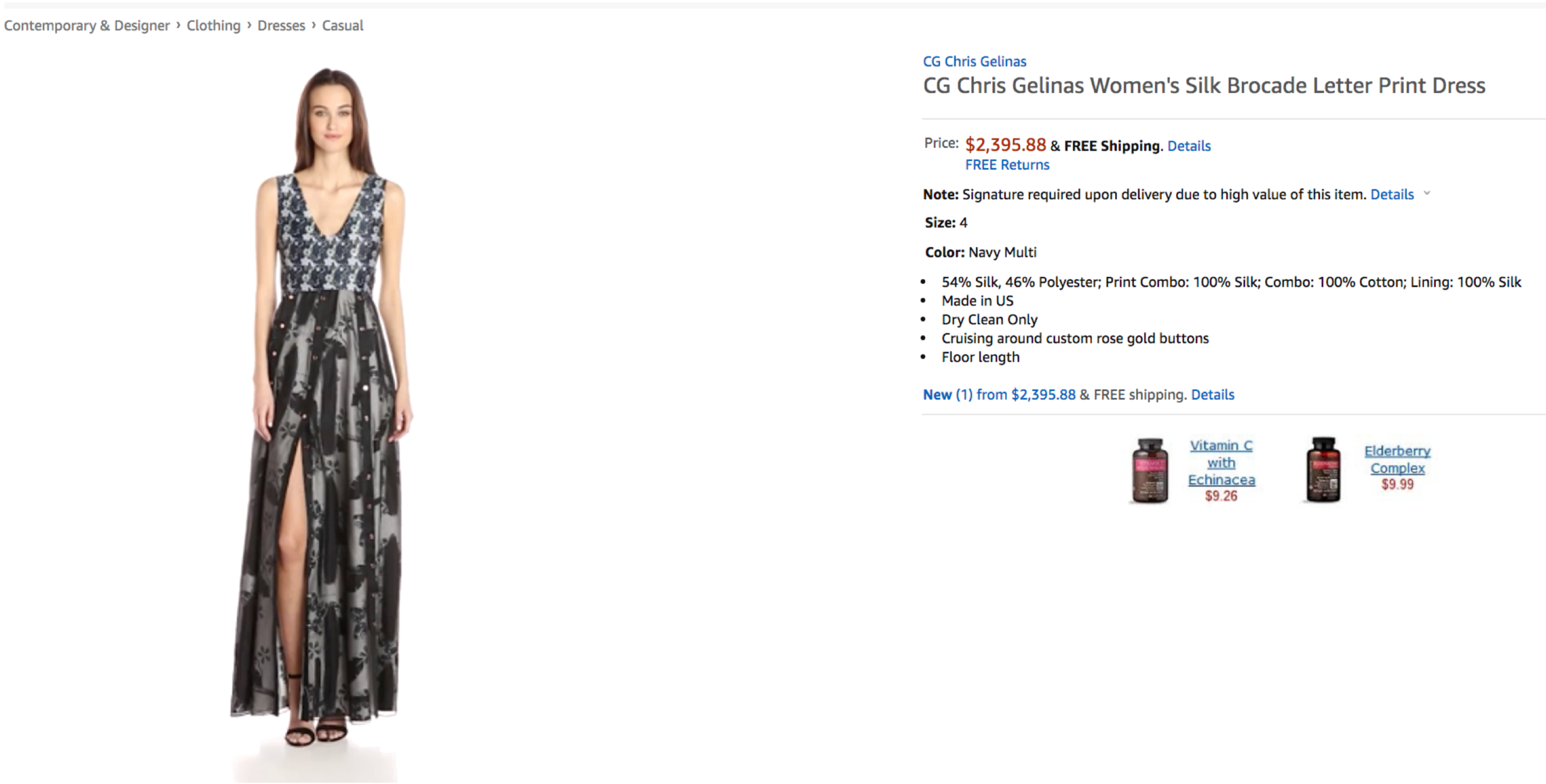
- N.B. *The results are not representative of the general inflation/deflation patterns in the U.S. At best they provide a data-point in the e-commerce subsector of the overall retail sector.*

Hits and Misses: Examples



NN and Linear Regression gave accurate predictions for this product about 100 (with Linear Regression underpredicting the price).

Hits and Misses: Examples



NN and Linear Regression underpredicted the price. NN predicts the price of 230; Linear Regression predicts the price of 90.

This CG designer dress is a hard item to predict: many other items by the same brand are priced at around 200. Do the models miss some attributes? (e.g., custom-made rose gold buttons).

Dynamic Approach: Using Leads and Lags To Predict Missing Prices

- We can also use leads and lags of prices to predict missing prices . A simple motivation can be given by considering a model where (approximately):

$$P_{jt} = P_{jt}(X_j, U_j)$$

- Where U_j is the latent unobserved quality characteristics. Assume the price is monotonically increasing in the latent quality. Then conditioning on the previous and future values of P_{jt} is equivalent to conditioning on U_j.

- In constructing the index, either a leads or lag is always available, since the product *j* has transacted at least in one period.

The Benefits of Dynamic Model

- Using only leads and lags brings R^2 **for predicting log-price around 93%.** Using ***W*** and ***I*** features in addition only adds about **.5-1%** in fit

- While the dynamic approach can be good with the **Month-over-Month** chaining (still biases may occur, since predictions are based on the matching set of products).

- It may perform poorly with the **Year-over-Year** chaining (since the set of surviving products to train the dynamic predictive model can be small and biased). By contrast, hedonics performs well with the YoY chaining.

- Dynamic Approach gives results very close to non-hedonic Fisher index (showing more drastic deflation over 6 years).

Technical Details of Feature Engineering

Query: red dress

Customer behavior data

3,989 customer reviews | 259 answered questions

Image

$$X \in \mathbb{R}^d$$

Description

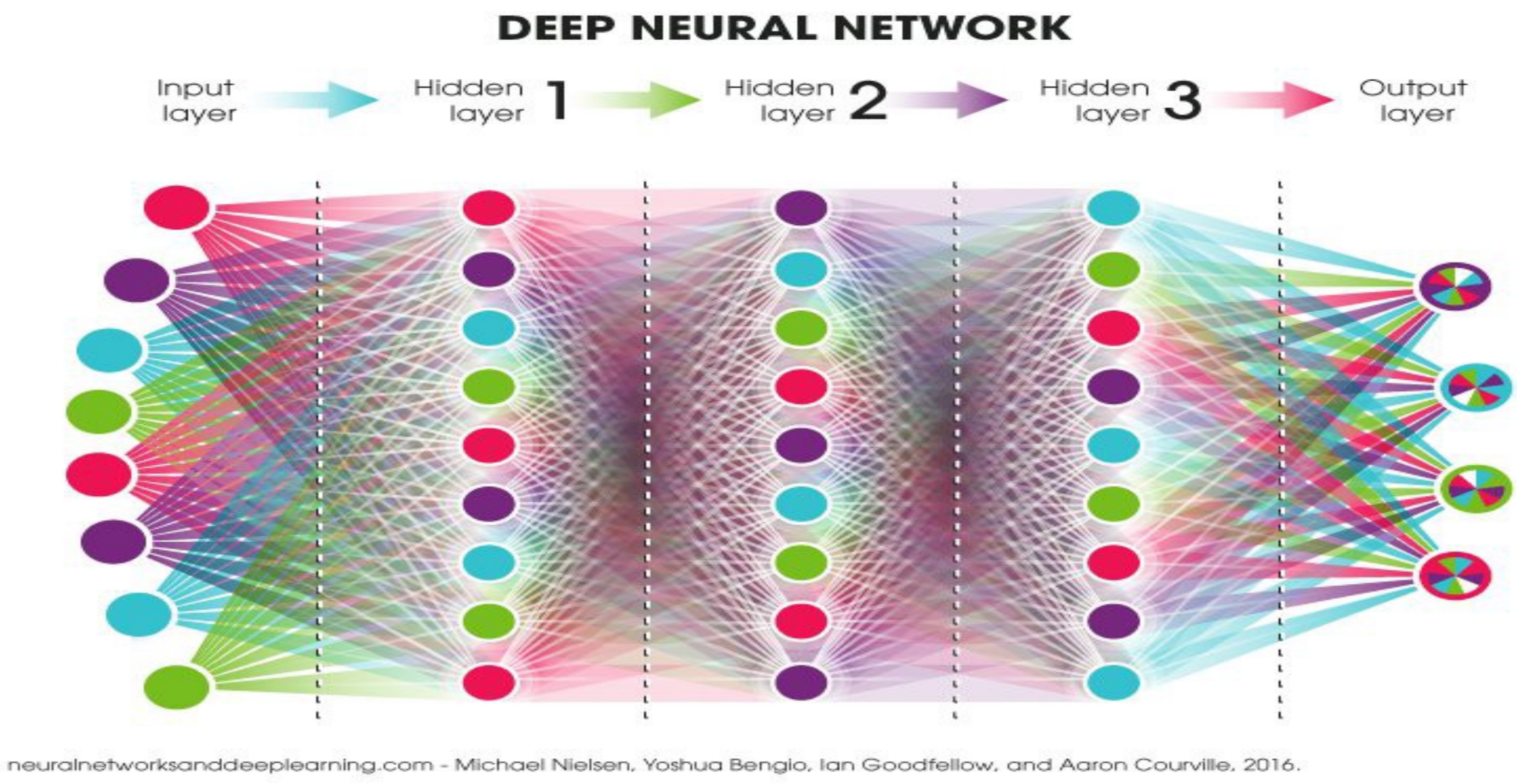
- Material - Cotton & Spandex.
- Imported
- Classic and Iconic Audrey Hepburn 50s Vintage Solid Color Swing Dress, Put on and Show Your Elegance and Charm.
- Features: Boat Neckline; Sleeveless; Full Circle Swing; Quick Access Zipper for Easy On and Off
- It's Great Choice for Daily Casual, Wedding , Ball, Party, Banquet and Other Occasion.
- [Size Chart] PLEASE Make Sure Your Measurements and Compare to the Size Chart From the picture on the left side or in the Following Description.
- Hand Wash Carefully,Low Temperature for Washing,Can not High Temperature Ironing, Line Dry

Title

Anni Coco

Anni Coco Women's Classy Audrey Hepburn 1950s Vintage Rockabilly Swing Dress

Features are created by (Deep) Neural Nets



Word2vec

In summary, dictionary V consists of binary words, and embeddings W are generated by applying a rank $m \ll d$ matrix to produce embeddings:

$$\underbrace{W}_{m\times d} = \underbrace{w}_{m\times d}\underbrace{V}_{d\times d}.$$

Given a sequence of K words, $\{V_{j,m}\}_{j=1}^{K}$, indexed by m, we have a middle word $V_{i,m}$ whose identity we have to predict from the surrounding words $\{V_{j,m}\}_{j=1}^{K-1}$.

Collapse the embeddings for context words by a sum,

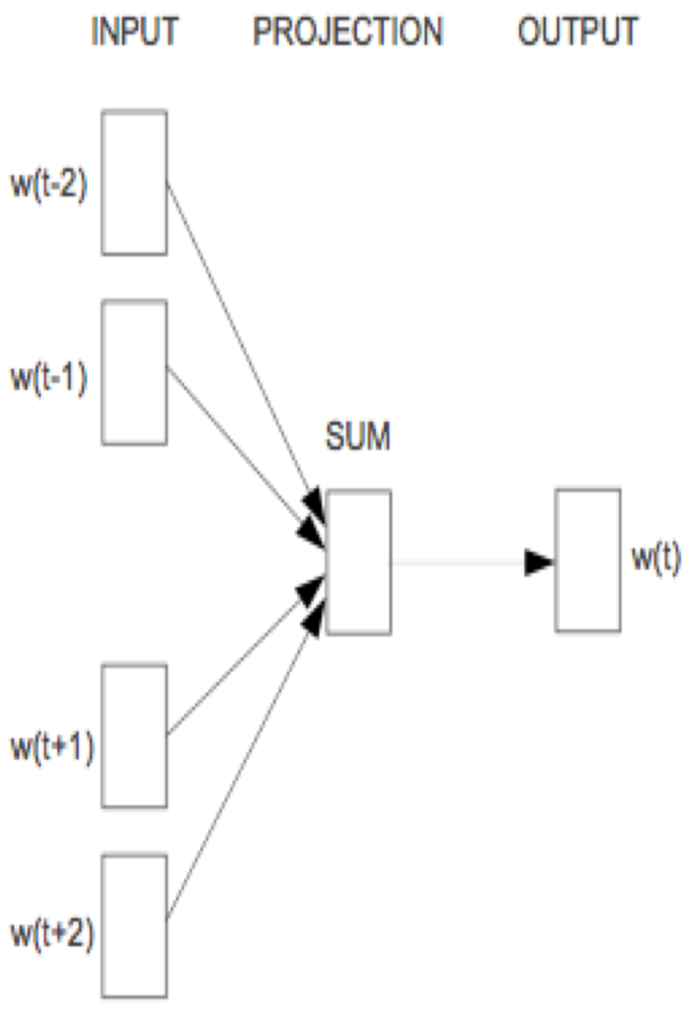
$$\bar{W}_m = \sum_j W_{j,m}.$$

This reflect the assumption that the context words are exchangeable, an assumption that simplifies the model, but is clearly very strong.

The probability of middle word $V_{i,m}$ being equal to the k-th word in the dictionary is modeled as

$$P(V_{i,m} = v_u \mid \{V_{j,m}\}_{j\in\{1,\ldots,K-1\}\setminus i}; \pi, w) = Z_u = \frac{e^{\pi_u' \bar{W}_m}}{\sum_{l=1}^{d} e^{\pi_l' \bar{W}_m}},$$

where π_k are conformable parameter vectors, describing the choice probability.



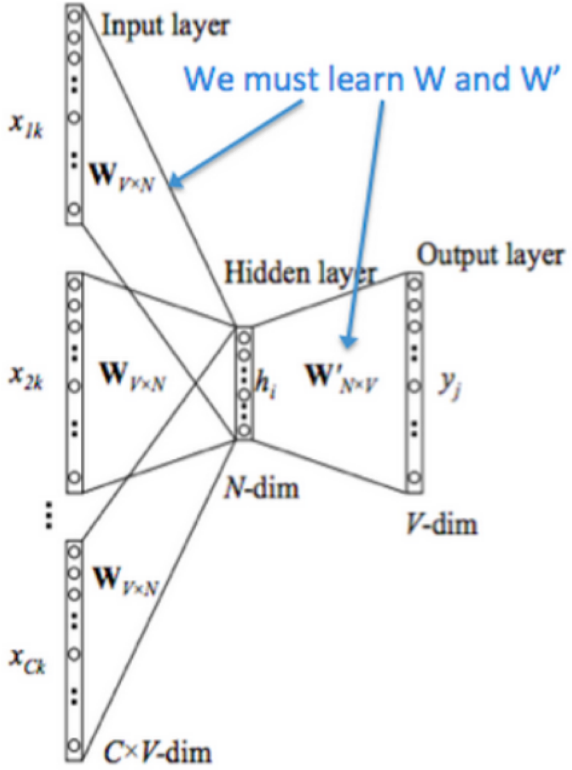
CBOW

Word2vec

The key parameters are the parameters of multinomal logistic distribution: $\pi = (\pi'_1, ...\pi'_d)$ and the parameters w of the embedding hidden layer.

The are jointly estimated using the maximum quasi-likelihood method:

$$(\hat{\pi}, \hat{w}) = \arg\max_{w,\pi} \sum_m \begin{pmatrix} 1(V_{i,m} = V_k) \log P(V_{i,m} = V_k \mid \{W_{j,m}\}) \\ +1(V_{i,m} \neq V_k) \log P(V_{i,m} \neq V_k \mid \{W_{j,m}\}) \end{pmatrix}.$$



Word Embeddings: Examples

womens	0.387542	0.03051	-0.19703	0.179724	-0.222901	-0.606905	0.306091	-0.597467
mens	0.758868	0.372418	0.370116	0.706623	-0.124954	0.5088	0.106177	0.208935
clothing	0.149283	0.5161	-0.027684	0.218484	-0.851416	-0.409885	0.386088	0.170605
shoes	1.323812	-0.358704	-0.007683	-0.552144	0.011261	0.365239	0.228273	-0.565655
women	0.601477	-0.045845	-0.099481	0.010576	-0.096852	-0.605281	0.25606	-0.550759
girls	0.417473	-0.005265	-0.40939	-0.531189	-1.31938	-0.034746	-0.940507	-0.361215
men	0.778298	0.406613	0.426292	0.534272	-0.056103	0.51756	0.107846	0.245275
boys	0.896637	-0.016821	-0.001602	-0.181901	-1.313441	0.449006	-0.828408	0.52121
accessories	0.86825	-0.378385	-1.247708	1.541265	0.323952	0.282909	-0.491176	0.081314
socks	0.27636	0.354296	0.185734	0.301311	-0.643142	-0.021945	0.320751	0.240676
luggage	0.796763	1.749548	-2.30671	-0.559585	0.03054	0.921458	0.417333	0.313436
dress	0.282053	0.233192	0.043318	0.174759	-0.50114	-0.381047	0.297995	-0.026033
baby	0.346065	-0.550016	-1.136202	-0.043899	-2.004979	0.689747	-1.091575	0.009901
jewelry	-0.315784	0.347808	-0.308736	0.878713	-0.766016	1.124318	-0.079883	-2.039485
black	0.427496	0.030204	-0.019082	0.224096	-0.162242	-0.325359	0.170407	-0.172714
boots	1.009074	-0.30359	0.03197	-0.334004	-0.095679	0.111328	0.11769	-0.51878
shirts	0.444152	0.452918	0.393656	0.517929	-0.531462	0.099621	0.146202	0.204338
shirt	0.328998	0.421561	0.226565	0.455649	-0.700352	0.067224	0.106364	0.233862
underwear	0.230821	0.490978	0.226338	0.202376	-0.774363	0.004693	0.228712	0.310215

Word2Vec Embeddings have interesting properties

Word2Vec("handbag")+ Word2Vec("men")- Word2Vec("woman")

≈Word2Vec("briefcase")

Word2Vec("tie")+ Word2Vec("woman")- Word2Vec("men")

≈ Word2Vec("pashmina") , Word2Vec("scarf")

- Distance is the Cosine Distance = Euclidian distance after normalizing vector norms to unity
- Embeddings for Title and Product Description are obtained by summing of the individual words

ELMO exploits the context more

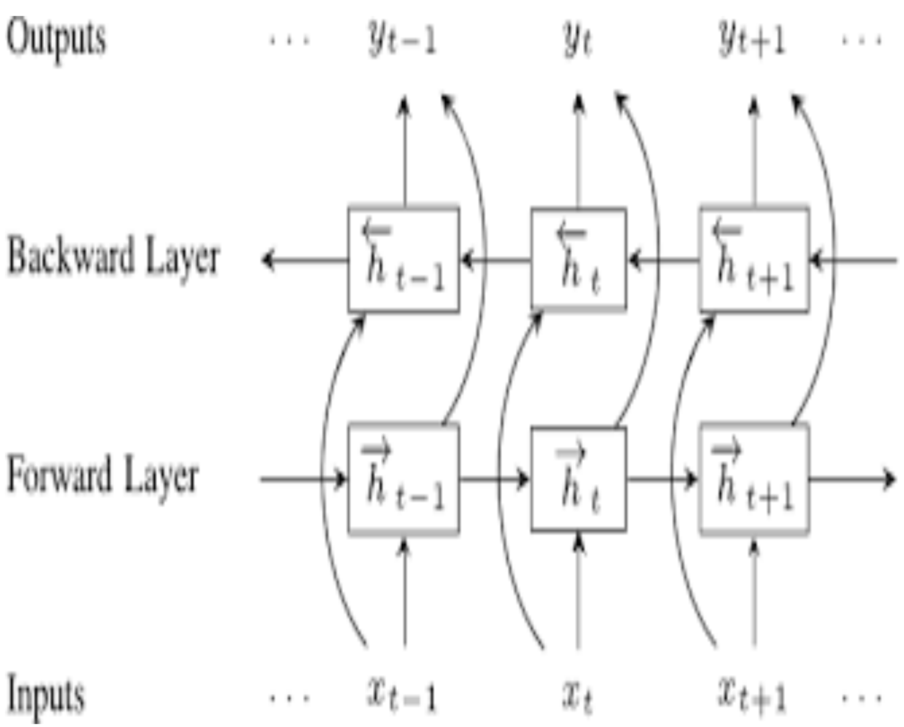
ELMO Embeddings. The ELMO algorithm uses the ideas of the Shannon game, where we guess the next word in the sentence m, i.e.

$$P[V_{k,m}|V_{1,m}, ..., V_{k-1,m}; \theta]$$

and also uses the reverse guessing as well:

$$P[V_{k,m}|V_{k+1,m}, ...V_{K_m,m}; \theta]$$

Recursive neural networks with a single or multiple hidden layers are used to model these probabilities. Parameters are estimated using quasi maximum log-likelihood methods, where the forward and backward quasi-likelihoods are added together.



ResNet50 Image Embedding

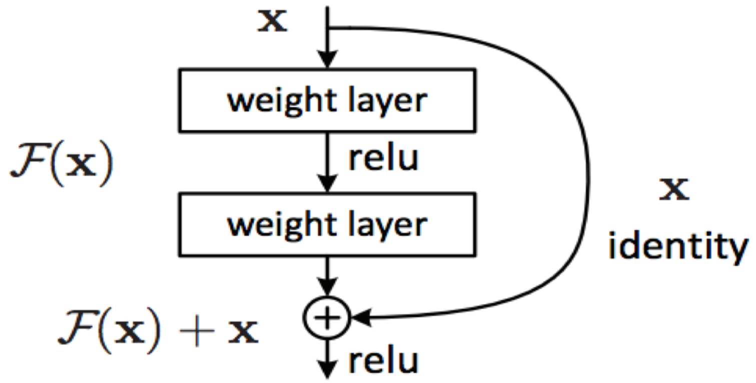
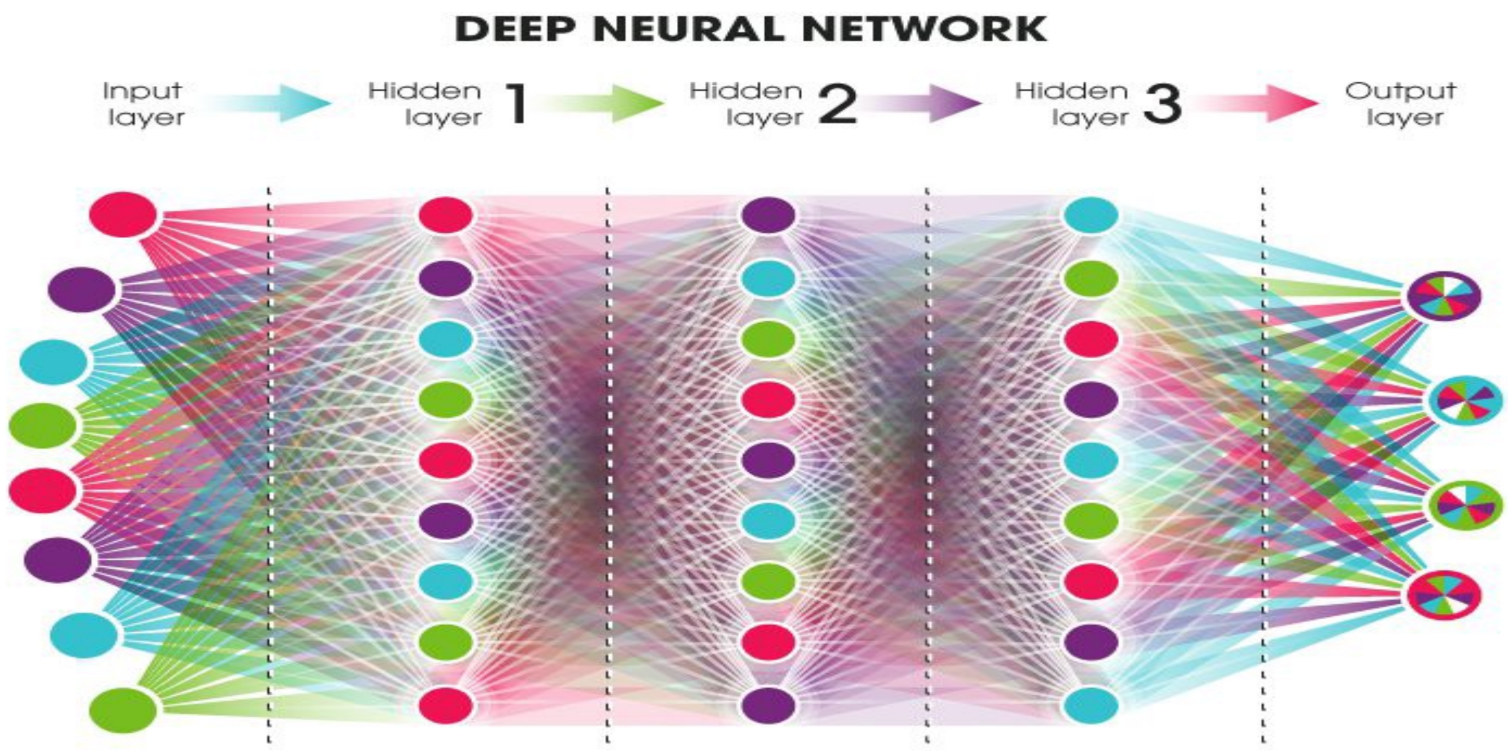


Figure 2. Residual learning: a building block.

a repeated composition of
the superposition of the linear (identity map)
with a nonlinearly generated layer

```
('Predicted:', [(u'n03450230', u'gown',
0.4549656), (u'n03534580', u'hoopskirt',
0.3363025), (u'n03866082', u'overskirt',
0.20369802)])
```

Predictions are created by Neural Nets with just several layers



neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

Conclusions

- Inflation indices are important inputs into measuring productivity and cost of living, and monetary and economic policy
- Our work addresses the *methodological challenges* in measuring inflation via hedonic approaches that arise due to
  - millions of products, with rapidly changing prices;
  - extremely high turnover for some product groups;
  - unstructured product attributes (title, product description, images).
- We do so by building hedonic price indices, which utilize
  - modern scalable computation that handles large amount of data
  - modern, open-source ML and AI tools to predict missing prices using product attributes.