Forecasting US Inflation
Using Bayesian Nonparametric Models

Todd E. Clark¹  Florian Huber²  Gary Koop³  Massimiliano Marcellino⁴

¹Federal Reserve Bank of Cleveland ²University of Salzburg ³University of Strathclyde ⁴Bocconi University

SUERF Workshop, 28 November 2022

The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland or the Federal Reserve System.
Introduction: Forecasting US inflation

- Inflation forecasting is fraught with challenges (see Faust and Wright, 2013)
  - Structural economic models and simple economic reasoning imply that inflation should be forecastable with a range of indicators
  - Other work has explored forecasting inflation with forward-looking financial indicators, such as bond yields
  - However: Simple inflation forecasts from the unobserved components model of Stock and Watson (2007) are hard to beat
- Most research uses parametric and linear models but
  - some work has suggested that the nonlinear effects of economic activity on inflation kick in as the economy becomes very strong
  - evidence for a non-linear Phillips curve (see Babb and Detmeister, 2017)
- More recently, researchers adopt flexible nonlinear and nonparametric models
Nonlinear and nonparametric modeling of inflation

- A fast-growing literature evaluates the use of machine learning techniques for macroeconomic forecasting
  - Random forests work particularly well (see Goulet Coulombe, 2020, arXiv; Medeiros et al., 2021, JBES; Clark et al., 2021, arXiv)
  - Recent papers use neural networks for modeling inflation (Goulet Coulombe, 2021, arXiv; Hauzenberger et al., 2022, IJoF)
  - Other papers use approximation techniques to estimate complicated models of inflation (see, e.g., Korobilis, 2019, JBES; Hauzenberger et al. 2021, JBES)
- These models are capable of capturing outliers in macroeconomic data flexibly and thus produce superior density forecasts (see Huber et al. 2020, JoE)
- However: A common assumption is that the shocks to the inflation process are Gaussian and homoskedastic
- What about the Sims (2002) critique?
What we do in this paper

- We build a fully nonparametric model of inflation
  - Flexible and analytically tractable Gaussian process regression for the conditional mean (see Hauzenberger et al., 2021, arXiv)
  - Overfitting is avoided through a subspace shrinkage prior (see Shin et al., 2019, JASA)
  - We assume the shock distribution to be unknown and estimate it using a Dirichlet Process mixture (DPM) model

- This model is applied to forecast US inflation
  - We focus not only on point and density forecasts but pay particular attention to tail predictions
  - Shed light on whether nonparametrics/nonlinearities are relevant in different segments of the predictive distribution
Do we need to depart from Gaussianity?

The figure shows the density of a linear combination of identified structural shocks (supply, demand, etc.) for the period from 1975 to 2019Q4 (in dashed red) and through 2021Q3 (in solid black)
Do we need to depart from linearity?

The plot shows the PC coefficient estimated recursively using a simple OLS regression with lagged unemployment rate, inflation expectations and the term spread.
A nonparametric model for US inflation

- Our model assumes that inflation \( \{y_t\}_{t=1}^T \) depends on a vector of \( K \) appropriately lagged predictors \( x_t \)

- The relationship between \( y_t \) and \( x_t \) is given by:

\[
y_t = f(x_t) + \varepsilon_t.
\]

with \( f : \mathbb{R}^K \to \mathbb{R} \) denoting an unknown function and \( \varepsilon_t \) is a white noise shock with unknown distribution

- If \( f(x_t) = x_t' \beta \) we end up with the linear regression model

- Next we focus on how we treat \( f \) and then we discuss our treatment of \( \varepsilon_t \)
Nonparametric modeling of the conditional mean

- Learning $f$ can be achieved through, e.g., Bayesian additive regression trees (Chipman et al., 2010), B-splines (Shin et al., 2019), or (deep) neural networks (Goulet Coloumbe, 2021)

- We propose approximating the unknown function $f$ using a GP regression (see Rasmussen and Williams, 2006)
  - This approach places a GP prior on the function $f$

$$f(x_t) \sim \mathcal{GP}(0, k(x_t, x_t))$$

- This implies a Gaussian prior on $f = (f(x_1), \ldots, f(x_T))'$ of the form:

$$f \sim \mathcal{N}(0, K),$$

with $K$ being a $T \times T$-dimensional kernel matrix with typical element $k(x_t, x_\tau)$ for times $t$ and $\tau$

- GP priors are nonparametric → no assumption on the form of $f$ is made
- They can be viewed as a prior over an infinite dimensional function space
Some details on GPs

- Crucially depend on the kernel function
  - Suitable kernels allow for capturing many different functional shapes and dynamics for the function $f$
  - GPs can also approximate the behavior of more complex models such as splines and neural nets by using suitable kernels (see Neal, 1994)
  - Most popular choice for the $K$: Gaussian kernel
- A typical element of $K$ under a Gaussian kernel is given by:

  $$k(x_t, x_\tau) = \xi \times \exp\left(-\frac{\phi}{2} ||x_t - x_\tau||^2\right),$$

  with $\xi, \phi \in \mathbb{R}^+$ denoting the hyperparameters of the kernel
- We estimate $\xi$ and $\phi$ using Metropolis Hastings updates
- Very easy to implement and they scale well in $K$ (not in $T$!)
Some intuition on what GPs do

- One possible way of modeling nonlinearities would be to map the inputs into some, higher dimensional space.
- An example would be the space of powers: $x_t \rightarrow \psi(x_t)$ with $\psi(x_t) = (x_t, x_t^2, \ldots, x_t^m)'$
- However, the choice of the mapping is ad-hoc and infinitely many mappings exist
- GPs handle this problem elegantly by placing a Gaussian prior over all possible (smooth) functions
- One can combine different Kernels to match features of the data
Some intuition on what GPs do

- **Toy model:** $\text{Inflation}_t = f(\text{UNRATE}_{t-1}) + \epsilon_t$
- $\xi = 4$ and $\phi = 0.01$

(a) Prior

(b) Posterior
Some intuition on what GPs do

- **Toy model:** Inflation$_t = f(UNRATE_{t-1}) + \varepsilon_t$
- $\xi = 4$ and $\phi = 0.1$

(a) Prior

(b) Posterior
Some intuition on what GPs do

- **Toy model:** \( \text{Inflation}_t = f(\text{UNRATE}_{t-1}) + \varepsilon_t \)
- \( \xi = 4 \) and \( \phi = 4 \)

(a) Prior
(b) Posterior
Subspace shrinkage

What if the function $f$ takes a simpler form?

- In this case, the GP regression could overfit and regularization on the functional form might be necessary
- We shrink towards a linear regression model by introducing a subspace shrinkage prior (see Shin et al. 2019)

Let $\Phi_0 = X(X'X)^{-1}X'$ denote the linear projection matrix of $X = (x'_1, \ldots, x'_T)'$

Subspace shrinkage involves modifying the kernel as follows:

$$K_1 = \left(K^{-1} + \frac{(I - \Phi_0)}{\tau^2}\right)^{-1}.$$

with $\tau$ denoting a weighting parameter on the linear subspace

- The GP prior is then replaced by

$$f \sim \mathcal{N}(0, K_1).$$
Subspace shrinkage ctd’

- The scaling parameter $\tau$ is crucial
  - Large values imply little weight on the linear piece
  - Small values force the function towards a linear specification
  - We assume that the prior on $\tau$ has a density proportional to:
    \[
p(\tau) \propto \frac{(\tau^2)^{d_1-1/2}}{(1 + \tau^2)^{d_0+d_i}} \quad \text{for } \tau \in (0, \infty).
    \]
  - Reduces to the half-Cauchy distribution if $d_0 = d_1 = 1/2 \rightarrow$ this is why the prior is called functional Horseshoe (fHS)

- Main implication of the prior (see Lemma 3.1 of Shin et al.):
  \[
  \mathbb{E}(f|\omega, \bullet) = (1 - \omega)\bar{f} + \omega \Phi_0 Y,
  \]
  with $\omega = 1/(1 + \tau^2) \in [0, 1]$ and $\bar{f}$ being the posterior mean of $f$ with Gaussian kernel $K$
Nonparametric modeling of the error distribution

- Modeling the conditional mean through GPs provides substantial flexibility
- However:
  - Difficult to capture heavy tails and asymmetries in the shocks
  - Moreover, not clear where nonlinearities come from (changes in transmission channels vs. huge structural shocks)
  - Ignoring these questions and simply assuming that $\varepsilon_t$ is Gaussian with constant variances might imply that $f$ is turning nonlinear to control for unobserved heterogeneity with respect to the shocks
- We model $\varepsilon_t$ flexibly by estimating the distribution using a Dirichlet Process Mixture (DPM) model (see Escobar and West, 1995; Frühwirth-Schnatter and Malsiner-Walli, 2018)
Dirichlet Process Mixtures

- We assume that $\varepsilon_t$ arises from:

$$
\varepsilon_t \sim \sum_{j=1}^{\infty} w_j N(\mu_j, \sigma_j^2), \quad \sum_{j=1}^{\infty} w_j = 1 \text{ and } w_j \geq 0 \ \forall \ j
$$

- On the weights we use a stick-breaking process (SBP) such that:

$$
w_1 = \xi_1, \quad w_j = \xi_j \prod_{i=1}^{j-1} (1 - \xi_i), \quad \text{for } j > 1,
$$

with $\xi_j \sim B(1, \alpha)$ being Beta distributed auxiliary variables with $\alpha$ denoting a hyperparameter that we estimate using a Metropolis Hastings update.

- The component means $\mu_j$ are Gaussian with zero mean and the component-variances $\sigma_j^2$ arise from an inverse Gamma distribution a priori.
Intuition

- Intuitively speaking, this mixture specification soaks up any variation in $y_t$ not explained through the GP component
- The GP captures smoothly evolving trends in inflation (similar to what an UC would do)
- The DPM captures transitory and possibly large deviations from this trend
- Example:
  - Structural breaks in the PC coefficients during the Volcker disinflation (which would be captured through the GP) versus unexpected large shocks during the pandemic
  - The DPM handles this by clustering pandemic shocks together
  - The corresponding predictive distribution will imply very few such shocks
- One shortcoming of DPMs: no persistence in the regimes $\rightarrow$ in our empirical work we capture this by replacing the iid variances with an SV process
This is how it looks like with synthetic data

We assume the following DGP:

\[
y_t = 0.6 y_{t-1} + \sum_{i=1}^{K} e^{x_{it}} \sin(x_{it}) + \varepsilon_t, \quad x_{it} \sim \mathcal{N}(0, 1), \quad \varepsilon_t \sim t_3(0, 1)
\]
Data overview

- We use quarterly data that range from 1959:Q1 to 2021:Q3 from the FRED-QD database
  - Three model sizes:
    1. Small: Includes only lagged inflation
    2. Moderate: Includes 29 variables
    3. Large: Includes 169 variables
  - 5-quarters-ahead inflation expectation from the Survey of Professional Forecasters are included as well
- We focus on forecasting CPI inflation, measured as \((400/h) \ln(P_{t+h}/P_t)\) at horizon \(h = 1, 4\)
- Results on ex food and energy inflation provided in a robustness check.
Design of the forecasting exercise and competing models

- We use a combination of models for the cond. mean and variance (and for each of the datasets)
  - Cond. mean models differ whether we use GPs, GPs with subspace shrinkage (GP-sub) or linear models (linear)
  - Cond. variance/shocks differ: Gaussian and homoskedastic (Homosk.), Gaussian with SV (SV), DPM, and DPM-SV
- Natural competitor: UC-SV model of Stock and Watson (2007)
- The design of our forecasting exercise is recursive
  - We use the period from 1980:Q1 to 2021:Q3 as our forecast evaluation period and produce forecasts for horizons $h = 1$ and $h = 4$
  - These forecasts are produced by lagging the elements in $x_t$ appropriately (i.e., we compute direct forecasts)
- As measures of predictive accuracy we use mean squared forecast errors (MSE), log predictive likelihoods (LPLs) and quantile scores (QSs)
Summary of the results

- Our results confirm the benefits of our flexible, nonparametric models
  - Over the 1980 to 2021 period, our nonparametric models often improve upon the univariate model of Stock and Watson (2007)
  - Gains of our methods during the pandemic are much more pronounced
- The primary gains to flexible nonparametric modeling come from nonlinear modeling of the conditional mean, through Gaussian processes
  - Adding more information does not uniformly improve accuracy
  - The best performing models are the one that use the moderately-sized dataset
- When we focus on the tails we observe strong gains in predicting deflation (left-tail) risks
- Somewhat more challenging to capture right-tail risks
- Models produce predictive densities which are asymmetric and heavy tailed
- In this talk, I focus on the moderately-sized models
# Overall forecasting performance: moderately sized models

<table>
<thead>
<tr>
<th></th>
<th>Full hold-out period</th>
<th></th>
<th></th>
<th></th>
<th>Only pandemic observations</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$h = 1$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>0.975</td>
<td>0.947</td>
<td>1.067</td>
<td>1.056</td>
<td><strong>0.844</strong></td>
<td>0.861</td>
<td>0.849</td>
<td>0.854</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.142)</td>
<td>(0.095)</td>
<td>(0.118)</td>
<td>(0.754)</td>
<td>(0.628)</td>
<td>(0.682)</td>
<td>(0.613)</td>
</tr>
<tr>
<td>GP</td>
<td>0.941</td>
<td>0.953</td>
<td>0.938</td>
<td>0.937</td>
<td>0.868</td>
<td>0.934</td>
<td>0.899</td>
<td>0.924</td>
</tr>
<tr>
<td></td>
<td>(0.195)</td>
<td>(0.221)</td>
<td>(0.333)</td>
<td>(0.33)</td>
<td>(0.883)</td>
<td>(0.846)</td>
<td>(0.739)</td>
<td>(0.877)</td>
</tr>
<tr>
<td>GP-sub</td>
<td><strong>0.925</strong></td>
<td>0.943</td>
<td>0.996</td>
<td>0.95</td>
<td>0.879</td>
<td>0.925</td>
<td>0.89</td>
<td>0.908</td>
</tr>
<tr>
<td></td>
<td>(0.189)</td>
<td>(0.246)</td>
<td>(0.065)</td>
<td>(0.321)</td>
<td>(0.796)</td>
<td>(0.839)</td>
<td>(0.642)</td>
<td>(0.728)</td>
</tr>
<tr>
<td><strong>$h = 4$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>1.196</td>
<td>1.285</td>
<td>0.935</td>
<td>0.996</td>
<td>1.308</td>
<td>1.333</td>
<td>1.105</td>
<td>1.079</td>
</tr>
<tr>
<td></td>
<td>(-0.24)</td>
<td>(-0.317)</td>
<td>(0.029)</td>
<td>(-0.012)</td>
<td>(0.371)</td>
<td>(0.43)</td>
<td>(-0.233)</td>
<td>(-0.052)</td>
</tr>
<tr>
<td>GP</td>
<td>0.856</td>
<td>0.871</td>
<td>0.873</td>
<td><strong>0.848</strong></td>
<td>0.529</td>
<td>0.536</td>
<td>0.669</td>
<td>0.573</td>
</tr>
<tr>
<td></td>
<td>(0.171)</td>
<td>(0.126)</td>
<td>(0.255)</td>
<td>(0.249)</td>
<td>(1.193)</td>
<td>(1.174)</td>
<td>(1.186)</td>
<td>(1.292)</td>
</tr>
<tr>
<td>GP-sub</td>
<td>0.952</td>
<td>0.999</td>
<td>0.927</td>
<td>0.959</td>
<td>0.737</td>
<td>0.784</td>
<td>0.732</td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td>(0.111)</td>
<td>(0.059)</td>
<td>(0.177)</td>
<td>(0.162)</td>
<td>(1.135)</td>
<td>(1.061)</td>
<td>(1.031)</td>
<td>(1.057)</td>
</tr>
</tbody>
</table>

MSE and Average LPL Results. LPL results given in parentheses. Results are relative to the UC-SV.
Forecast performance over time: 4-steps-ahead

Cumulative log predictive likelihoods against the UC-SV model: Moderately sized models
Quantile scores: GP

1-quarter-ahead

Homosk.
DPM
SV
DPM-SV

4-quarters-ahead

Homosk.
DPM
SV
DPM-SV
Quantile scores: GP-sub

1–quarter–ahead

4–quarters–ahead

Homosk.

DPM

SV

DPM–SV

Chart showing quantile scores for different models and forecast horizons.
A deeper look at the predictive densities

Notes: Black solid lines refer to the GP regression and the DPM-SV specification on the shocks, while dashed blue lines denote the findings of the UC-SV model. The left panel refers to the one-quarter-ahead predictions, while the right panel shows the results for four-quarters-ahead.

Quantile scores ($p = 0.95$) and 10, 50 and 90th percentiles of the one-step-ahead predictive distribution for the moderately sized data set
Which variables determine the predictive distribution?

- Difficult to answer due to the nonlinear/nonparametric conditional mean model
- However: we could search for a linear approximation to the predictive density
- Approximate the quantiles of the predictive distribution $Q_{p,t+h}$ using a linear and possibly sparse regression model
- For each $p$, we solve the following optimization problem:

$$\beta^*_p = \arg\min_{\beta_p} \sum_{t=t_0}^{T} (Q_{p,t+h} - \beta'_p x_t)^2 + \lambda \sum_{j=1}^{K} |\beta_{p,j}|,$$

where $t_0$ marks the beginning of the hold-out period, $\beta_p = (\beta_{p,1}, \ldots, \beta_{p,K})'$ are coefficients, and $\lambda \geq 0$ is a penalty term
- This resembles a LASSO problem
### Predictive variable relevance

#### 1-quarter-ahead

<table>
<thead>
<tr>
<th>Variable</th>
<th>1-quarter-ahead</th>
<th>4-quarters-ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCEC96</td>
<td>-0.01</td>
<td>+0.05</td>
</tr>
<tr>
<td>FPIx</td>
<td>-0.02</td>
<td>+0.03</td>
</tr>
<tr>
<td>GCEC1</td>
<td>-0.02</td>
<td>+0.02</td>
</tr>
<tr>
<td>INDPRO</td>
<td>-0.01</td>
<td>+0.01</td>
</tr>
<tr>
<td>CUMFNS</td>
<td>+0.02</td>
<td>+0.02</td>
</tr>
<tr>
<td>PAYEMS</td>
<td>-0.00</td>
<td>+0.10</td>
</tr>
<tr>
<td>CE16OV</td>
<td>-0.00</td>
<td>+0.01</td>
</tr>
<tr>
<td>AWHMAN</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td>CES6060000007</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td>GDPCTPI</td>
<td>-0.00</td>
<td>-0.04</td>
</tr>
<tr>
<td>CPIAUCSL</td>
<td>+0.30</td>
<td>+0.03</td>
</tr>
<tr>
<td>FEDFUNDS</td>
<td>-0.00</td>
<td>-0.00</td>
</tr>
<tr>
<td>GS10TB3Mx</td>
<td>+0.07</td>
<td>+0.06</td>
</tr>
<tr>
<td>BUSLOANsx</td>
<td>+0.09</td>
<td>+0.07</td>
</tr>
<tr>
<td>CONSUMERx</td>
<td>-0.05</td>
<td>-0.03</td>
</tr>
<tr>
<td>S.P.500</td>
<td>-0.08</td>
<td>-0.09</td>
</tr>
<tr>
<td>PPIACO</td>
<td>+0.05</td>
<td>+0.05</td>
</tr>
<tr>
<td>COMPRNFB</td>
<td>+0.01</td>
<td>+0.10</td>
</tr>
<tr>
<td>ULCNFNFB</td>
<td>-0.03</td>
<td>+0.07</td>
</tr>
<tr>
<td>BAA10YM</td>
<td>+0.07</td>
<td>+0.05</td>
</tr>
<tr>
<td>INFEXP</td>
<td>+0.27</td>
<td>+0.28</td>
</tr>
</tbody>
</table>
Conclusions

- Forecasting inflation is hard due to changing relations between it and its predictors and partly due to occasionally large and asymmetric shocks.
- The model we develop is designed to address these.
- In our empirical work, we have shown that the model is capable of producing accurate point and density forecasts.
  - These forecasts are often more precise than the ones obtained from simpler alternatives such as the UC-SV model.
  - The forecasting performance is driven by a superior overall performance in the left tail and the center of the distribution.
  - However, the performance in the right tail is somewhat weaker.
    - This is mainly driven by slightly inflated predictive intervals during the Great Moderation.
    - In the high-inflation period of the early 1980s and during the second year of the pandemic, our model also improves upon the UC-SV model in the right tail.
- Single equation techniques → direct forecasts might mask nonlinearities that you would capture through iterative forecasts (see Marcellino, Stock and Watson, 2005).