Bayesian Exponential Smoothing with the Local-Seasonal-Global Trend Model

Daniel F. Schmidt
Work with Slavek Smyl, Alexander Documentov, Erwin Wibowo, Xueying Long and Christoph Bergmeir

Department of Data Science and AI
Monash University

Meta 2024
June 14, 2024

Outline

1. Motivation
2. The Local-Seasonal-Global Trend Model
3. Bayesian Inference of the LSGT Model
4. Benchmarking Results

Outline

Exponential Smoothing

- Exponential smoothing is a widely used classical forecasting technique
  - It is simple to implement and understand
  - Reasonably flexible and robust to assumptions
  - Often exhibits good forecasting performance
- The most basic form forecasts the next point using

$$
\begin{aligned}
\hat{y}_{t+1} &\sim N(l_{t+1}, \sigma^2) \\
l_{t+1} &= \alpha l_t + (1-\alpha) y_t
\end{aligned}
$$

- Various extensions have been proposed
  - Seasonality (additive, multiplicative)
  - Local trends (additive, multiplicative, damped)
  - Different stochastic error models (additive, multiplicative)
- Usually fitted (point estimation) via maximum likelihood
- Intervals computed using asympototic arguments

Exponential Smoothing

- Exponential smoothing is a widely used classical forecasting technique
  - It is simple to implement and understand
  - Reasonably flexible and robust to assumptions
  - Often exhibits good forecasting performance
- The most basic form forecasts the next point using

$$\begin{aligned} \hat{y}_{t+1} &\sim N(l_{t+1}, \sigma^2) \\ l_{t+1} &= \alpha l_t + (1-\alpha) y_t \end{aligned}$$

- Various extensions have been proposed
  - Seasonality (additive, multiplicative)
  - Local trends (additive, multiplicative, damped)
  - Different stochastic error models (additive, multiplicative)
- Usually fitted (point estimation) via maximum likelihood
- Intervals computed using asympototic arguments

Weaknesses of Classical Exponential Smoothing

- The classical exponential smoothers have several shortcomings
  1. Limited functional form of trend
     - Either additive (linear) or multiplicative (exponential)
     - Unable to model super-linear but sub-exponential trends
  2. Normality of errors
     - Stochastic error terms are usually assumed to be normally distributed
     - Extensions to non-normal errors (particularly intervals) is difficult in an maximum likelihood framework
  3. Homoscedasticity of errors
     - The usual assumption is that errors are constant variance
     - Extensions usually assume simple proportionality; unable to handle super-linear/sub-exponential growth

Weaknesses of Classical Exponential Smoothing

- The classical exponential smoothers have several shortcomings
  1. Limited functional form of trend
     - Either additive (linear) or multiplicative (exponential)
     - Unable to model super-linear but sub-exponential trends
  2. Normality of errors
     - Stochastic error terms are usually assumed to be normally distributed
     - Extensions to non-normal errors (particularly intervals) is difficult in an maximum likelihood framework
  3. Homoscedasticity of errors
     - The usual assumption is that errors are constant variance
     - Extensions usually assume simple proportionality; unable to handle super-linear/sub-exponential growth

Weaknesses of Classical Exponential Smoothing

- The classical exponential smoothers have several shortcomings
  1. Limited functional form of trend
     - Either additive (linear) or multiplicative (exponential)
     - Unable to model super-linear but sub-exponential trends
  2. Normality of errors
     - Stochastic error terms are usually assumed to be normally distributed
     - Extensions to non-normal errors (particularly intervals) is difficult in an maximum likelihood framework
  3. Homoscedasticity of errors
     - The usual assumption is that errors are constant variance
     - Extensions usually assume simple proportionality; unable to handle super-linear/sub-exponential growth

Weaknesses of Classical Exponential Smoothing

- The classical exponential smoothers have several shortcomings
  1. Limited functional form of trend
     - Either additive (linear) or multiplicative (exponential)
     - Unable to model super-linear but sub-exponential trends
  2. Normality of errors
     - Stochastic error terms are usually assumed to be normally distributed
     - Extensions to non-normal errors (particularly intervals) is difficult in an maximum likelihood framework
  3. Homoscedasticity of errors
     - The usual assumption is that errors are constant variance
     - Extensions usually assume simple proportionality; unable to handle super-linear/sub-exponential growth

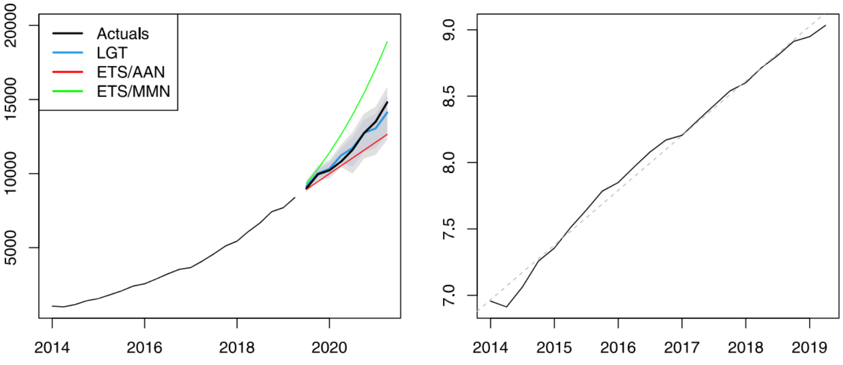
Our Contribution: A New Exponential Smoother

- To address these we introduced a new family of exponential smoothers
  - The local-global-trend (LGT) model
  - The seasonal-global-trend (SGT) model
- Flexible trend terms, heavy tailed errors, heteroscedasticity
- Implemented using the Stan HMC sampler
- This was effective but very slow
- Our latest work has consolidated the model into an single local-seasonal-global-trend (LSGT) model
  - We tidied up some statistical aspects of the model
  - Implemented bespoke Gibbs sampler that is much faster than Stan HMC

Our Contribution: A New Exponential Smoother

- To address these we introduced a new family of exponential smoothers
  - The local-global-trend (LGT) model
  - The seasonal-global-trend (SGT) model
- Flexible trend terms, heavy tailed errors, heteroscedasticity
- Implemented using the Stan HMC sampler
- This was effective but very slow
- Our latest work has consolidated the model into an single local-seasonal-global-trend (LSGT) model
  - We tidied up some statistical aspects of the model
  - Implemented bespoke Gibbs sampler that is much faster than Stan HMC

A Motivating Example (1)

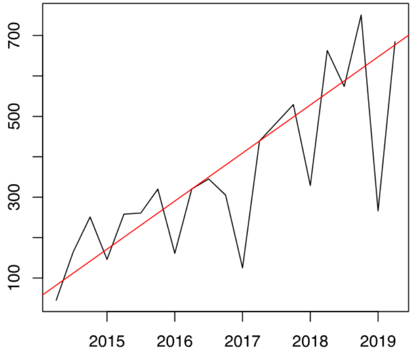


(a) Original series

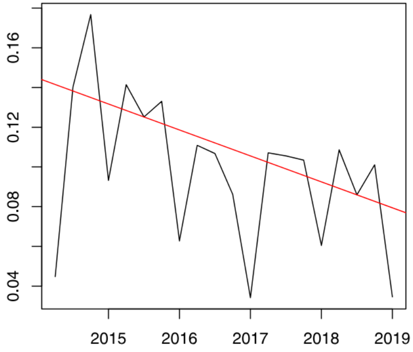
(b) Log of series

Quarterly revenue of Amazon Web Services (in million USD, source: statista.com)
Logarithm suggests non-exponential growth.

A Motivating Example (2)



(c) Absolute differences of the series



(d) Absolute percentage differences of the series

Assuming errors are bigger than changes in trend, errors grow over time.
We see that errors are not proportional to the trend.

Outline

LSGT Model

$$y_{t+1} \mid \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \nu \sim t(\nu, \hat{y}_{t+1}, \hat{\sigma}_{t+1}),$$

where

$$\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}_{t+1}^2 &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}$$

subject to

$$\sum_{i}^{m} \log s_i = 0.$$

LSGT Model

$$y_{t+1} \mid \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \nu \sim t(\nu, \hat{y}_{t+1}, \hat{\sigma}_{t+1}),$$

where

$$\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}_{t+1}^2 &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}$$

subject to

$$\sum_i^m \log s_i = 0.$$

$\Longrightarrow$ Heavy tailed error distribution

LSGT Model

$$y_{t+1} \mid \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \nu \sim t(\nu, \hat{y}_{t+1}, \hat{\sigma}_{t+1}),$$

where

$$\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}_{t+1}^2 &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}$$

subject to

$$\sum_{i}^{m} \log s_i = 0.$$

$\Longrightarrow$ Global (powered) trend

LSGT Model

$$y_{t+1} \mid \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \nu \sim t(\nu, \hat{y}_{t+1}, \hat{\sigma}_{t+1}),$$

where

$$\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}_{t+1}^2 &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}$$

subject to

$$\sum_{i}^{m} \log s_i = 0.$$

$\Longrightarrow$ Local trend

LSGT Model

$$y_{t+1} \mid \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \nu \sim t(\nu, \hat{y}_{t+1}, \hat{\sigma}_{t+1}),$$

where

$$\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}_{t+1}^2 &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}$$

subject to

$$\sum_i^m \log s_i = 0.$$

$\Longrightarrow$ Seasonal components

LSGT Model

$$y_{t+1} \,|\, \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \nu \sim t(\nu, \hat{y}_{t+1}, \hat{\sigma}_{t+1}),$$

where

$$\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}_{t+1}^2 &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}$$

subject to

$$\sum_i^m \log s_i = 0.$$

$\Longrightarrow$ Heteroskedastic errors (powered global trend)

LSGT Model

$$y_{t+1} \mid \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \nu \sim t(\nu, \hat{y}_{t+1}, \hat{\sigma}_{t+1}),$$

where

$$\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}_{t+1}^2 &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}$$

subject to

$$\sum_i^m \log s_i = 0.$$

$\Longrightarrow$ Homoskedastic/heteroskedastic mixture control

LSGT Model

$$y_{t+1} \mid \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \nu \sim t(\nu, \hat{y}_{t+1}, \hat{\sigma}_{t+1}),$$

where

$$\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}_{t+1}^2 &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}$$

subject to

$$\sum_{i}^{m} \log s_i = 0.$$

$\implies$ Scale invariant seasonality adjustment

Outline

Bayesian Inference (1)

- We fit the LSGT Model using Bayesian inference
  - Why?
- The model has a large number of free parameters
- In contrast to fitting via maximum likelihood
  1. use of prior distributions can help stabilize estimates
  2. confidence intervals based on asymptotic arguments can be inaccurate
- This is particularly the case for shorter time series and heavy tailed error distributions
- Specifically, we use an Monte-Carlo Markov-chain (MCMC) approach

Bayesian Inference (1)

- We fit the LSGT Model using Bayesian inference
  - Why?
- The model has a large number of free parameters
- In contrast to fitting via maximum likelihood
  1. use of prior distributions can help stabilize estimates
  2. confidence intervals based on asymptotic arguments can be inaccurate
- This is particularly the case for shorter time series and heavy tailed error distributions
- Specifically, we use an Monte-Carlo Markov-chain (MCMC) approach

Bayesian Inference (2)

- In the Bayesian approach, we have
  1. a probability model, $p(\mathbf{y} \,|\, \boldsymbol{\theta})$
  2. a prior distribution, $\pi(\boldsymbol{\theta})$

  where $\mathbf{y} = (y_1, \ldots, y_n)$ is our data (time series) and $\boldsymbol{\theta} \in \Theta$ is a vector of model parameters
- The prior $\pi(\boldsymbol{\theta})$ expresses our *a priori* (before data) beliefs about how likely different values from Θ are to be the population values
- Upon observing data, we form the posterior distribution

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{y} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- This expresses our beliefs about how likely different values from Θ are to be the population values, having observed the data $\mathbf{y}$

Bayesian Inference (2)

- In the Bayesian approach, we have
  1. a probability model, $p(\mathbf{y} \,|\, \boldsymbol{\theta})$
  2. a prior distribution, $\pi(\boldsymbol{\theta})$

  where $\mathbf{y} = (y_1, \ldots, y_n)$ is our data (time series) and $\boldsymbol{\theta} \in \Theta$ is a vector of model parameters
- The prior $\pi(\boldsymbol{\theta})$ expresses our *a priori* (before data) beliefs about how likely different values from Θ are to be the population values
- Upon observing data, we form the posterior distribution

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{y} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

- This expresses our beliefs about how likely different values from Θ are to be the population values, having observed the data $\mathbf{y}$

Bayesian Inference (3)

- It is usual to use a plug-in distribution for forecasting
  - We find a point estimate $\hat{\boldsymbol{\theta}}$ and plug it into our probability model
- This ignores the uncertainty in our estimates, which can be considerable for "small" n
- In contrast, we use the posterior predictive distribution

$$p(y_{n+1}, \ldots, y_{n+h} \,|\, \mathbf{y}) = \int p(y_{n+1}, \ldots, y_{n+h} \,|\, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta} \,|\, \mathbf{y}) d\boldsymbol{\theta}$$

- A mixture of all models, weighted by their posterior probability $\Longrightarrow$ takes into account the uncertainty in our estimates

Bayesian Inference (4)

- The mixture integral is, in general, intractable
  - This is where MCMC helps
- We randomly sample m models, $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}$ from the posterior
- Then, for each model we generate one (or more) future realisations of the time series $\mathbf{y}$ from

$$(y_{n+1}, \ldots, y_{n+h}) \sim p(y_{n+1}, \ldots, y_{n+h} \,|\, \boldsymbol{\theta}^{(i)})$$

- We have a collection of future realisations; we may then use the empirical statistics of these for forecasting
  - the empirical median is our default point-forecast
  - the empirical quantiles give us forecast intervals
- We used the Stan probabilistic programming language to implement a first version of the BLSGT model
- We then created a Gibbs sampler to dramatically improve speed

Prior Distributions (1)

- Careful choice of priors can result in increased performance, particularly for small n
- We chose weakly informative priors for most parameters, e.g.,
  - a scale invariant prior $\pi(\sigma^2) \propto 1/\sigma^2$ for σ^2
  - Cauchy distributions for γ and λ (trend weights)
- For the exponential smoothing parameters

$$\alpha, \beta, \zeta \sim \mathrm{Beta}(a, b)$$

  with default choices $a = 1$, $b = 1/2$.
- This concentrates more probability mass near $\alpha = 1$, as the models are more densely concentrated near $\alpha = 1$ than $\alpha = 0$

Prior Distributions (2)

- Of interest is our prior for the initial seasonal terms $\log s_1, \ldots, \log s_m$
- Note that $\log s_j = 0$ implies $s_j = 1$
  $\Longrightarrow$ if all $s_j = 1$ and $\zeta = 0$ there is no seasonality
- We gave the $\log s_j$ horseshoe priors

$$\begin{aligned} \log s_j \,|\, \xi_j, \nu &\sim N(0, \xi_j^2 \nu^2) \\ \xi_j &\sim C^+(0,1) \\ \nu &\sim C^+(0,1) \end{aligned}$$

  where $C^+(0,1)$ denotes a half-Cauchy distribution
- The horseshoe prior has a spike of probability at $\log s_j = 0$ (sparsity inducing) and heavy tails
- If there is no real evidence for seasonality, the horseshoe will shrink the $\log s_j$ terms to zero
  $\Longrightarrow$ provides some robustness against mis-specification of seasonality

Gibbs Sampling

- The initial implementation was through the HMC sampler in Stan
  - It worked, but was very slow
- We developed a bespoke Gibbs sampler to sample from the posterior
- If $p(X, Y)$ is a density, then a Gibb sampler works by iterating
  - Sample X from the conditional $p(X \mid Y)$
  - Sample Y from the conditional $p(Y \mid X)$
- The power of Gibbs sampling is that it breaks the sampling problem into small sub-problems
- We are free to use any type of sampling algorithm for these

Scale Mixtures

- The error distribution in LSGT is a Student-t
  - The heavy tails lead to flat, difficult to sample likelihoods
- A key technique we use is a scale-mixture-of-normals (SMN) representation
- The scale-mixture-of-normals representation of a t-distribution; if

$$\begin{aligned} x \,|\, \omega^2 &\sim N(\mu, \sigma^2 \omega^2) \\ \omega^2 &\sim \mathrm{IG}(\delta/2, \delta/2) \end{aligned}$$

  then $x \sim t(\delta, \mu, \sigma)$
- The advantage of this is that conditional on the latent variable ω^2, x is now normally distributed
- We use this to simplify sampling

LSGT Model with SMN Representation

$$
\begin{aligned}
y_{t+1} \mid \hat{y}_{t+1}, \hat{\sigma}_{t+1}, \omega_{t+1} &\sim N(\hat{y}_{t+1}, \hat{\sigma}^2_{t+1}\omega^2_{t+1}), \\
\omega^2_{t+1} \mid \nu &\sim \mathrm{IG}(\nu/2, \nu/2)
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{y}_{t+1} &= (l_t + \gamma l_t^{\rho} + \lambda b_t) s_{t+1-m}, \\
l_t &= \alpha \left(\frac{y_t}{s_{t-m}} \right) + (1-\alpha) l_{t-1}, \\
b_t &= \beta (l_t - l_{t-1}) + (1-\beta) b_{t-1}, \\
\log s_t &= \zeta \log \frac{y_t}{l_t} + (1-\zeta) \log s_{t-m}, \\
\hat{\sigma}^2_{t+1} &= \chi^2 (\phi^2 + (1-\phi)^2 l_t^{2\tau}),
\end{aligned}
$$

$\Longrightarrow$ Conditional on ω^2_{t+1} and s_{t+1-m}, this is a Gaussian linear regression

Sampling the Other Parameters

- We use a mix of techniques for the other parameters
  - Gradient-assisted Metropolis-Hastings for α, β, ζ and seasonality terms
  - Grid samplers for power parameters ρ, τ and DOF ν
- Most other parameters have well known conditionals (inverse gamma, normal, etc.)
- Implemented in R and C++
  $\Longrightarrow$ resulted in very substantial speedups over Stan HMC
- The Gibbs sampler is based on the LSGT model presented here, which also tidied up the model specification over the original Stan version

Outline

Benchmark Datasets and Metrics

- We tested BLSGT on the M3 benchmark competition dataset
  - 645 yearly, 1,428 monthly, 756 quarterly and 174 "other" series
- We compared against the existing state-of-the-art univariate forecasting algorithms
- We used sMAPE

$$\text{sMAPE} = \frac{200}{h} \sum_{t=1}^{h} \frac{|y_{n+t} - \hat{y}_{n+t}|}{|y_{n+t}| + |\hat{y}_{n+t}|},$$

and MASE

$$\text{MASE} = \frac{h^{-1} \sum_{t=1}^{h} |y_{n+t} - \hat{y}_{n+t}|}{(n-s)^{-1} \sum_{t=s+1}^{n} |y_t - y_{t-s}|}.$$

Results (1)

	sMAPE	MASE
LSGT Gibbs (Homosc.)	**14.91**	2.55
LSGT Gibbs (Heterosc.)	14.99	2.50
LGT Stan	15.18	**2.48**
Best algorithm in M3	(RBF) 16.42	(ROBUST-Trend) 2.63
Hybrid	16.73	2.85
ETS (ZZZ)	17.37	2.86
ARIMA	17.12	2.96
Best of Bagged ETS	(BLD.Sieve) 17.80	(BLD.MBB) 3.15
THETA	16.97	2.81
MAPA	16.91	2.86

Performance on 645 yearly series

Results (2)

	sMAPE	MASE
	Monthly series	
LSGT Gibbs (Homosc.)	13.94	0.83
LSGT Gibbs (Heterosc.)	13.76	**0.82**
LGT Stan	13.77	0.83
Best algorithm in M3	(THETA) 13.89	(ForecastPro) 0.85
Hybrid	13.91	0.84
ETS (ZZZ)	14.15	0.86
ARIMA	14.98	0.88
Best of Bagged ETS	(BLD.MBB) 13.64	(BLD.MBB) 0.85
THETA	13.89	0.86
MAPA	**13.62**	0.84

Performance on $1,428$ monthly series

Results (3)

	sMAPE	MASE
LSGT Gibbs (Homosc.)	8.78	1.06
LSGT Gibbs (Heterosc.)	**8.78**	**1.06**
LGT Stan	8.87	1.07
Best algorithm in M3	(THETA) 8.96	(THETA) 1.09
Hybrid	9.39	1.14
ETS (ZZZ)	9.61	1.18
ARIMA	10.00	1.19
Best of Bagged ETS	(BLD.Sieve) 10.03	(BLD.MBB) 1.22
MAPA	9.31	1.13

Performance on 756 quarterly series

Results (4)

	sMAPE	MASE
LSGT Gibbs (Homosc.)	4.21	1.70
LSGT Gibbs (Heterosc.)	**4.16**	**1.69**
LGT Stan	4.25	1.72
Best algorithm in M3	(ARARMA) 4.38	(AutoBox2) 1.86
Hybrid	4.33	1.79
ETS (ZZZ)	4.33	1.79
ARIMA	4.46	1.83
THETA	4.41	1.90
MAPA	4.48	1.87

Performance on 174 other series

Results (5)

	sMAPE	MASE
LSGT Gibbs (Homosc.)	12.27	1.30
LSGT Gibbs (Heterosc.)	**12.21**	**1.29**
LGT Stan	12.27	1.30
Best algorithm in M3	(THETA) 12.76	(THETA) 1.39
Hybrid	12.82	1.40
ETS (ZZZ)	13.13	1.43
ARIMA	13.58	1.46
MAPA	12.71	1.41

Overall performance

{S,L}GT (Stan) vs LSGT (Gibbs)

	MSIS 90p	MSIS 98p	Time (s)
Yearly series			
LGT Gibbs (heteroscedastic error)	**16.47**	**27.92**	4.63
LGT Stan	17.38	32.64	60.03
Monthly series			
SGT Gibbs (heteroscedastic error)	5.22	8.52	14.67
SGT Stan	**5.10**	**8.20**	163.84
Quarterly series			
SGT Gibbs (heteroscedastic error)	**7.19**	**13.06**	10.70
SGT Stan	7.64	15.95	374.12
Other series			
LGT Gibbs (heteroscedastic error)	**10.51**	**16.53**	7.71
LGT Stan	10.69	16.68	150.88

Mean Scaled Interval Scores (MSIS) and run times for Stan vs Gibbs

Robustness to Mis-specified Seasonality

	sMAPE	MASE
Monthly series		
SGT Gibbs (homoscedastic, horseshoe prior)	13.94	0.83
SGT Gibbs (heteroscedastic, horseshoe prior)	**13.76**	**0.82**
SGT Gibbs (homoscedastic, Cauchy prior)	13.92	0.83
SGT Gibbs (heteroscedastic, Cauchy prior)	13.78	0.83
Yearly series		
LGT Gibbs (homoscedastic error)	**14.91**	2.55
LGT Gibbs (heteroscedastic error)	14.99	**2.50**
SGT Gibbs (homoscedastic, horseshoe prior)	15.35	2.62
SGT Gibbs (heteroscedastic, horseshoe prior)	15.37	2.55
SGT Gibbs (homoscedastic, Cauchy prior)	15.81	2.72
SGT Gibbs (heteroscedastic, Cauchy prior)	15.56	2.61
SGT Stan (Cauchy prior $C(0,4)$)	16.56	2.74

Ablation study of the priors for the initial seasonal factors.

Thank you!

- We are finalizing the paper on new LSGT model and Gibbs sampler
- Current work-in-progress
  - Variational approach to speed things up further
  - Learning of priors over multiple series (global prior modelling)
- Code is available through CRAN
  - The `Rlgt`
- Details of the first version of LGT in
  - Slawek Smyl, Christoph Bergmeir, Alexander Dokumentov, Xueying Long, Erwin Wibowo, Daniel Schmidt, "*Local and global trend Bayesian exponential smoothing models*", International Journal of Forecasting
- Thank you – questions?