Statistical Deep Learning for Dependent Establishment Data

Paul A. Parker[1]

Joint work with Qi Wang[1] and Robert Lund[1]

Aug. 4, 2024

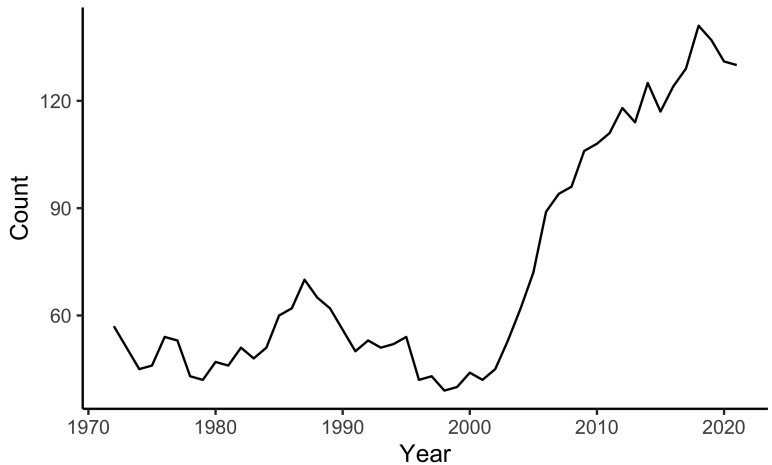
[1] Department of Statistics, University of California Santa Cruz

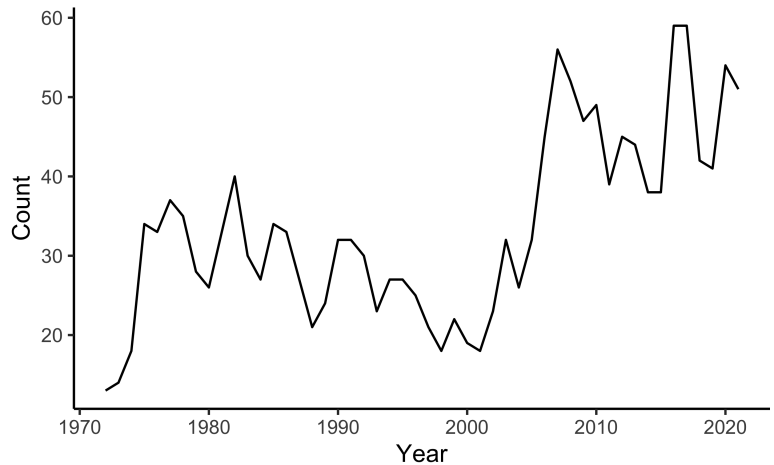
Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS)

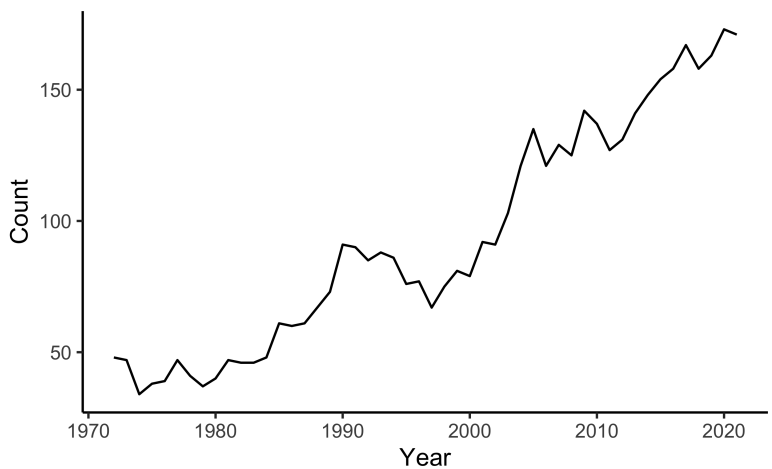
- Annual census of U.S. academic institutions.
- Sponsored by the National Center for Science and Engineering Statistics.
- Data available from 1972 to 2021.
- Provides insight into the demographics and distribution of graduate students across different fields and institutions.
  - We are looking are graduate student counts.
- As illustration, we focus specifically on schools in California.

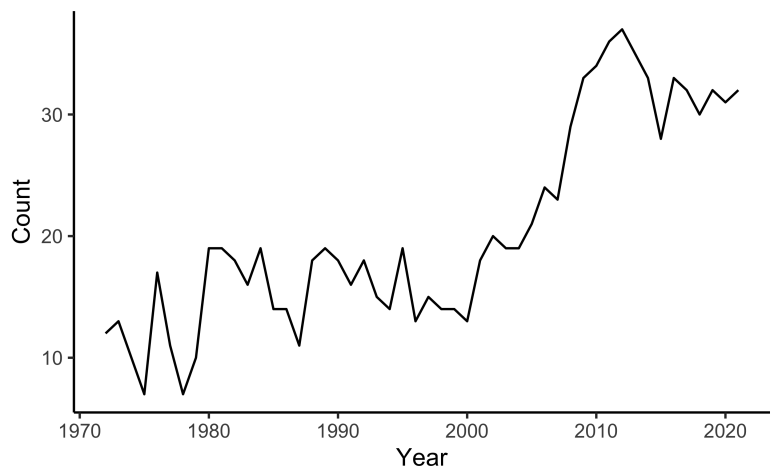
Exploratory Analysis

Time Series Plots



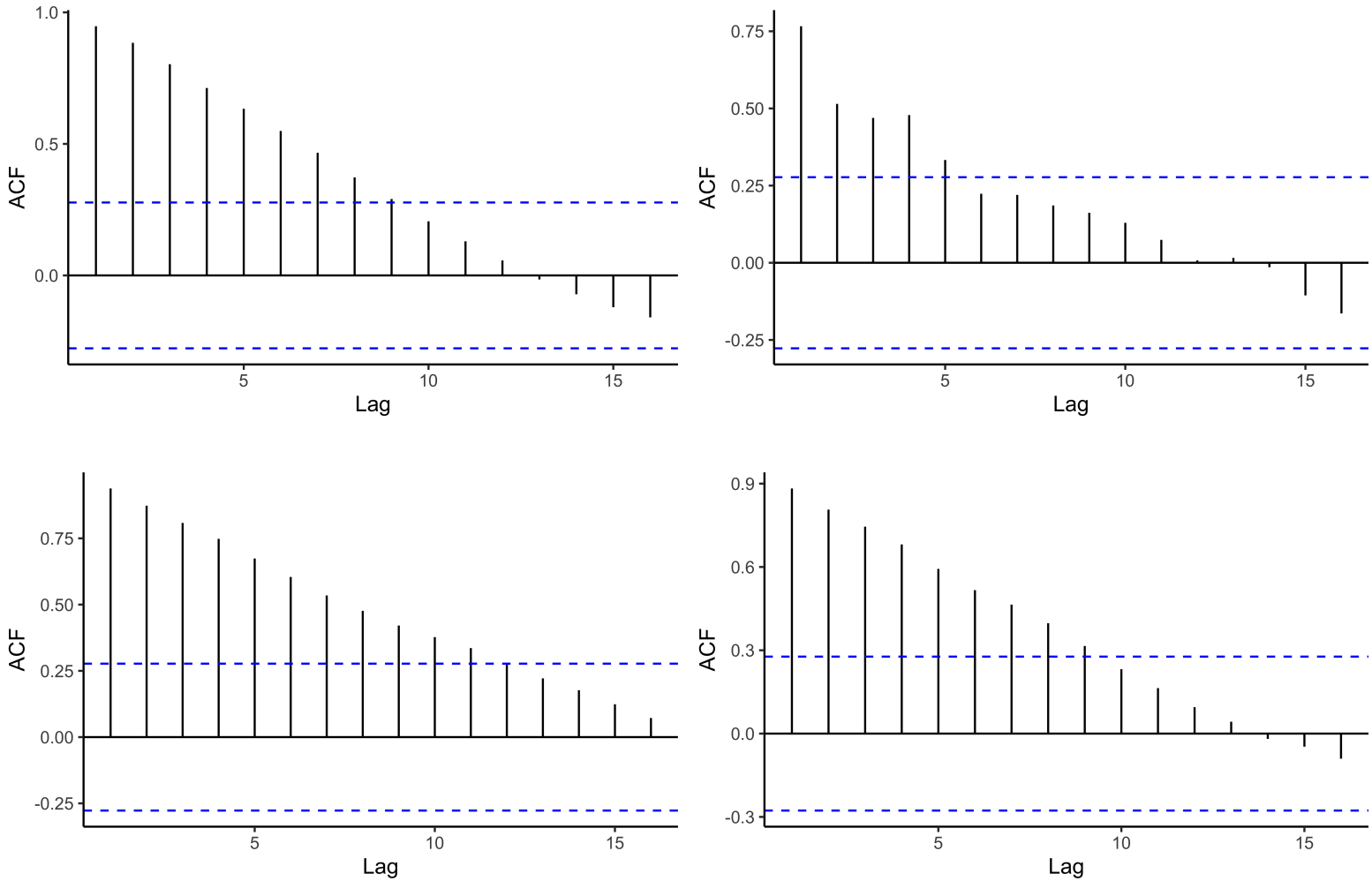






Exploratory Analysis

ACF Plots



Goals

- Construct a one-size-fits-all approach for GSS graduate student data.
- Address the count nature of the data.
- Incorporate temporal dependence in a flexible manner.
- Quantify uncertainty around forecasts/estimates.

Deep Learning

Feed-forward Neural Networks

- Feed-forwards neural networks (FNNs) are widely use in the deep learning literature for their ability to model general nonlinear functions.

$$\hat{y}_i = g_o\left(\boldsymbol{h}_i'\boldsymbol{\eta}\right)$$
$$\boldsymbol{h}_i = g\left(\boldsymbol{W}\boldsymbol{x}_i\right)$$

- The function $g(\cdot)$ is known as an activation function and is applied element-wise. Typical choices are the sigmoid and hyperbolic tangent functions.
- The function $g_o(\cdot)$ is an output layer activation function, and is chosen based on the support of the data. For continuous data this is typically the identity function.
- The $n_h \times p$ parameters $\boldsymbol{W}$ and $n_h \times 1$ parameters $\boldsymbol{\eta}$ are usually estimated with gradient descent techniques.

Deep Learning

Recurrent Neural Networks

- Recurrent neural networks are a deep learning technique that more efficiently models dependence present in sequential data such as time series.

$$\hat{y}_t = g_o\left(\boldsymbol{h}_t'\boldsymbol{\eta}\right)$$
$$\boldsymbol{h}_t = g(\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t)$$

- Here, the hidden layers are connected across time.

Deep Learning

Recurrent Neural Networks

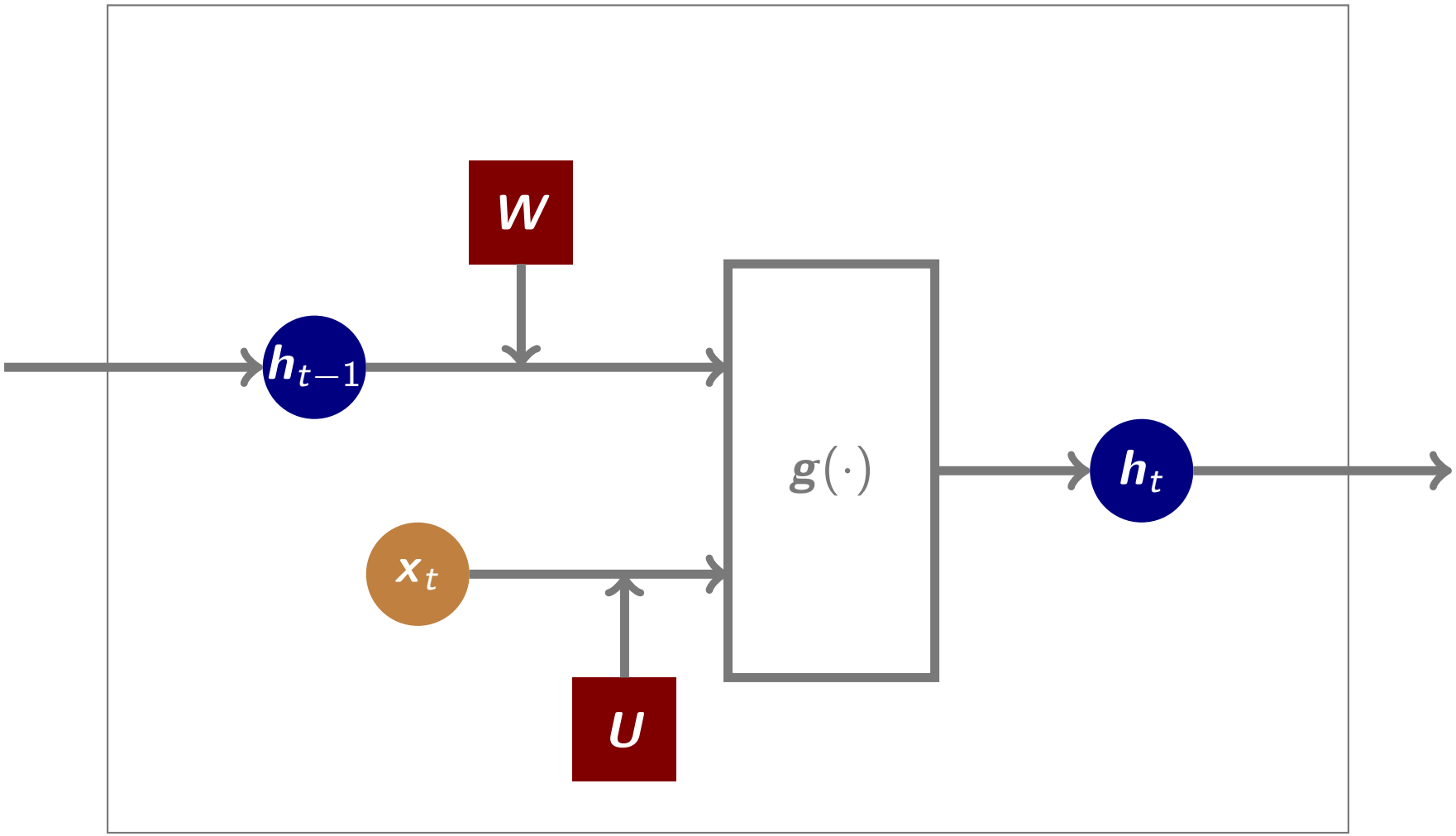


Figure: Graphical depiction of a recurrent layer.

Deep Learning

Echo State Networks

- One alternative to traditional RNNs is the Echo State Network (ESN).

- Instead of estimating all parameters in the model, an ESN randomly samples and fixes the hidden layer parameters before model fitting.

$$\hat{y}_t = g_o(\boldsymbol{h}_t'\boldsymbol{\eta})$$
$$\boldsymbol{h}_t = g\left(\frac{\nu}{|\lambda_W|}\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t\right)$$

- The weight matrices $\boldsymbol{W}$ and $\boldsymbol{U}$ are randomly chosen and fixed. Thus, the only parameters that must be learned are $\boldsymbol{\eta}$.
- Here, $\frac{\nu}{|\lambda_W|}$ is a scaling term recommended by McDermott and Wikle (2017) for stability.

Note:

The deep learning approaches discussed so far are algorithmic and not considered statistical models.

Deep Learning

Echo State Networks

- The ESN can be linked to a likelihood, allowing for maximum likelihood estimation of η (McDermott and Wikle, 2017).
- An ensemble ESN can be fit by resampling the hidden layer parameters many times (McDermott and Wikle, 2017).
  - This allows for uncertainty quantification.
- A Bayesian ESN can also be fit by placing a prior distribution on η (McDermott and Wikle, 2019).

Count Echo State Network

- We propose the count echo state network (CESN) to model GSS counts.

$$
\begin{aligned}
Y_t|\lambda_t &\overset{ind}{\sim} \text{Poisson}(\lambda_t) \\
\log(\lambda_t) &= \boldsymbol{h}_t'\boldsymbol{\eta} \\
\boldsymbol{h}_t &= g\left(\frac{\nu}{|\lambda_W|}\boldsymbol{W}\boldsymbol{h}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t\right),\ t = 2, \ldots, T \\
\boldsymbol{h}_1 &= g\left(\boldsymbol{U}\boldsymbol{x}_1\right)
\end{aligned}
$$

Note:

The CESN can be fit a single time, providing only a point estimate, or an ensemble can be used for uncertainty quantification.

Count Echo State Network

Parameter Estimation

- The CESN model can be estimated via penalized maximum likelihood to prevent overfitting.
- For example, with an L1 penalty, to estimate $\boldsymbol{\eta}$, we would minimize the loss function

$$\mathcal{L} = -\sum_{t=1}^{T} \left(Y_t \boldsymbol{h}_t' \boldsymbol{\eta} - \exp(\boldsymbol{h}_t' \boldsymbol{\eta}) \right) + \tau \sum_{j=1}^{n_h} |\eta_j|.$$

- This can be done via standard software such as the `glmnet` package (Friedman et al., 2010).
- For the purposes of illustration, we fix $\tau = 2$ (this parameter could be tuned via cross validation to further improve predictive performance).

Bayesian Count Echo State Network

- We also explore a Bayesian CESN model by placing a prior distribution on $\boldsymbol{\eta}$.

- Specifically we use a multivariate log-Gamma prior (Bradley et al. (2018) and Bradley et al. (2020)).

$$\boldsymbol{\eta} \sim \text{MLG}(\mathbf{0}, \alpha^{1/2}\kappa \boldsymbol{I}, \alpha\mathbf{1}, \alpha\mathbf{1})$$

  - Acts as a conjugate prior for Poisson regression.
  - Easy to sample from the posterior distribution.
  - Asymptotically equivalent to a multivariate Normal prior.

- Again, κ acts as a regularization parameter, where we fix $\kappa = 10$ for illustration.

Cross-Validation

Setup

- We construct one-step-ahead predictions for each of the years 2017-2021.
- We compare a variety of models:
  - An intercept only baseline model
  - An integer-valued generalized autoregressive conditional heteroscedasticity model (INGARCH(1,1))
  - A single CESN model
  - An ensemble CESN model
  - A Bayesian CESN model

Cross-Validation

Comparison Metrics

- Mean squared prediction error:
  $\text{MSPE}_t = \frac{1}{S}\sum_{s=1}^{S}(\hat{Y}_{st} - Y_{st})^2$
- Mean squared logarithmic prediction error:
  $\text{MSLPE}_t = \frac{1}{S}\sum_{s=1}^{S}\left(\log(\hat{Y}_{st}+1) - \log(Y_{st}+1)\right)^2$
- Interval score:
  $\frac{1}{S}\sum_{s=1}^{S}\left\{(u_{st}-\ell_{st}) + \frac{2}{\alpha}(\ell_{st}-Y_{st})I(Y_{st}<\ell_{st}) + \frac{2}{\alpha}(Y_{st}-u_{st})I(Y_{st}>u_{st})\right\}$

Cross-Validation Results

MSPE

Table: Mean square prediction error for one step ahead predictions of graduate student counts from 2017-2021.

Model	2017	2018	2019	2020	2021	5 Year Avg.
Intercept Only	3103	4908	3451	2454	3413	3466
INGARCH(1,1)	1116	1605	1033	1735	1245	1347
Single CESN	489	699	2483	1968	311	1190
Ensemble CESN	495	677	2404	1845	324	**1149**
Bayesian CESN	512	651	2645	1766	247	1164

Cross-Validation Results

MSLPE

Table: Mean square logarithmic prediction error ($\times 10^2$) for one step ahead predictions of graduate student counts from 2017-2021.

Model	2017	2018	2019	2020	2021	5 Year Avg.
Intercept Only	17.97	38.32	36.31	42.03	45.07	35.94
INGARCH(1,1)	8.33	19.77	15.11	8.02	15.58	13.30
Single CESN	8.23	20.41	7.17	9.90	8.02	10.75
Ensemble CESN	8.28	20.33	7.21	10.59	7.93	10.87
Bayesian CESN	8.13	17.80	5.02	9.53	2.84	**8.66**

Cross-Validation Results

Interval Score

Table: 95% prediction interval score for one step ahead predictions of graduate student counts from 2017-2021.

Model	2017	2018	2019	2020	2021	5 Year Avg.
INGARCH(1,1)	307	285	280	315	314	300
Ensemble CESN	188	174	292	282	109	209
Bayesian CESN	179	154	303	290	86	**202**

Summary

- We introduce the count echo state network in order to model graduate student counts from the GSS.
- Linking to a Poisson likelihood effectively captures the count nature of the dataset.
- We also introduce a Bayesian variant of the model, that can be fit in a computationally efficient manner.
- The proposed model(s) outperformed baseline models on out-of-sample predictions in terms of both point predictions and uncertainty quantification.

Future Work

- Explore modeling of other dependence structures in the data, such as spatial dependence.
- Explore other count distributions.
- Investigate recently developed methods for MCMC-free sampling with MLG models (Bradley and Clinch, 2024).

Thank you!

paulparker@ucsc.edu

Bradley, J. R. and Clinch, M. (2024). Generating independent replicates directly from the posterior distribution for a class of spatial hierarchical models. *Journal of Computational and Graphical Statistics*, (just-accepted):1–32.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018). Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion).

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *Journal of the American Statistical Association*, 115(532):2037–2052.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

McDermott, P. L. and Wikle, C. K. (2017). An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat*, 6(1):315–330.

McDermott, P. L. and Wikle, C. K. (2019). Bayesian recurrent neural

network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy*, 21(2):184.

Uses the Beamer simple theme from http://github.com/famuvie/beamerthemesimple