Case Study: Projecting Canadian Bankruptcy Rate in 2011-2012 with 1987 - 2010 Data
Authors: Xiaowen Zhang, Sooraj Subrahmannian, Si Chen, Ernest Kim

Introduction

With a goal of precisely and accurately forecasting monthly bankruptcy rates for Canada from 2011-2012, we took a weighted average of various univariate and multivariate time series models. Using monthly data from January 1987 to December 2010 of the bankruptcy rate, housing price index, unemployment rate, and population, we fit Holt-Winters, VAR(p), SARIMAX and VARX models.

Using Root Mean Squared Error (RMSE)[1] as our metric, we constructed models utilizing data from 1987 to 2008, tested their performance on 2009-2010, and then refit the models with the complete data set to predict 2011-2012. With an aim to reduce RMSE in our validation set (2009-2010), we tested parameters on the various models and applied the best-performing parameters to generate our prediction.

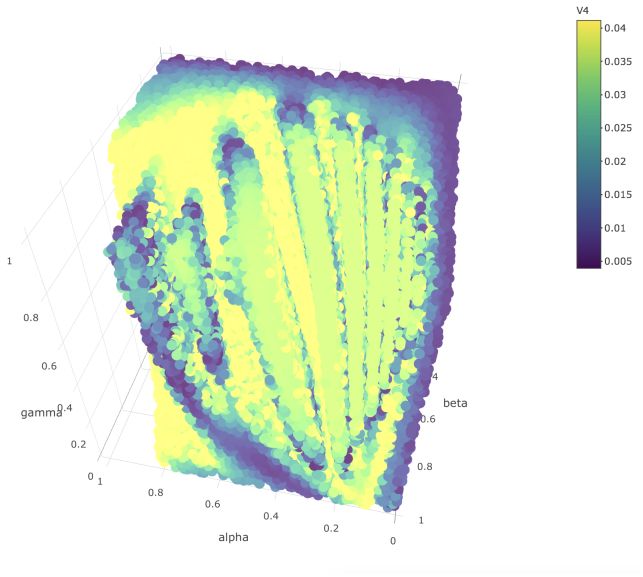
Our prediction, with intervals, are included in the conclusion of this report. Overall, we found that the Holt-Winters models were able to generate the most accurate predictions, however we chose to ensemble multiple models in order to avoid issues with over-fitting. We also built and tested our models on the following time period: 2009-2010 (two years).

Univariate Methods: Holt-Winters

Holt-Winters, also known as “Triple Exponential Smoothing”, applies an exponential smoothing parameter to level, trend and seasonal components of a time series. The function contains three hyper-parameters which allows the user to adjust weights assigned to previous values (with values further back in time being weighted down exponentially). Unlike ARIMA models, Holt-Winters require no parametric assumptions and thus, requires none of the normal validations (zero-mean residuals, homoscedasticity, no autocorrelation, nor normality).

However, Holt-Winters models requires a grid search to find the optimal levels for the three parameters. The following image was created from a representative sample of a grid search of two million parameter combinations:

[1] RMSE is a measure of how much our predictions varied from the true value. By building a training set and leaving out the last two years to test our predictions, we are able to build optimal models.



This graphic demonstrates that ideal values for a Holt-Winters model exists along a volumetric gradient. The dark regions indicate where RMSE is minimized. Certain regions of parameters (the volume around alpha = 1, beta = 1, gamma = 0 for instance) were excluded from the sample due to particularly poor performance. For the purposes of this study, we used the best set of parameters found in the grid search.

Multivariate Methods: SARIMAX

SARIMAX models incorporate multivariate time series data in order to fit a model and generate predictions. Built on similar Box-Jenkins approaches, SARIMAX models stationary time series only, using differencing operations to remove trend and seasonality. Lastly, under SARIMAX, we assume that while a variable may influence the response, the reverse is not true (these external variables are known as exogenous).

We think there may be dependencies of our response variable (Bankruptcy Rate) on exogenous variables. In other words, unemployment and housing price may influence the bankruptcy rate, but the bankruptcy rate does not influence these variables. Thus, we try to build a SARIMAX model to fit the data.

First, we explore autocorrelation plots to identify trends and seasonality in the Bankruptcy Rate data. We discovered that after a difference with lag twelve, the autocorrelation plots looked stationary and ready to build a model. Next, based on differencing for trend and seasonality, we used a loop function to iterate over all possible reasonable values for p, q, P, Q, along with all possible combinations of explanatory variable (House Price, Population, Unemployment Rate).

From the loop, we generate a table listing model comparison diagnostics which allow us to choose the optimal model. We collected the top two models from the ranking table and used a likelihood ratio test to compare them. Likelihood ratio tests allow us to generate a test statistics

to compare two models (one being more complex) and indicate if the more complex model is distinct and better. Finally we get a conclusion that a SARIMAX model with order of (2,1,2)(2,1,2) plus X = Unemployment Rate is optimal, with a RMSE 0.004489097.

To verify that this model's assumptions hold for future prediction's accuracy, we use following residuals test:

1. The first is a t-test for zero-mean assumption. It passed with p value of 0.6899.
2. The second is constant variance. Our assumption and null hypothesis for this model is that our data exhibits constant variance. We pass this with p value of 0.273.
3. The third assumption tests that the residuals are uncorrelated. Here we use tsdiag() function to combine ACF and Ljung-Box test all in one and passed only for the first several lags.
4. The last is test for normality. We use a QQ-plot and Shapiro-Wilke and pass with a p-value of 0.09444.

In this way, we get a SARIMAX model with order of (2,1,2)(2,1,2) plus X = Unemployment Rate is the best, with a RMSE 0.004489097.

Multivariate Methods: VAR(p)

Unlike SARIMAX that treats variables as exogenous, vector autoregression (VAR) model treats all variables as endogenous, meaning that they could mutually influence each other. Each endogenous variables is modeled similarly to a regression model where the predictors are fit with other endogenous variables with maximum-*p* lags. The VAR model can capture quite complex patterns, however, VAR models can quickly overfit the training data since the model has a large number of parameters. Therefore when tuning parameters, one should consider using a smaller *p*-lag even if the resulting accuracy is slightly worse than a model with a larger *p*.

The model selection depends on three components: the different variables, the max-*p* lag and the amount of data included in the training process. The endogenous variables available in the data apart from bankruptcy rate are the unemployment rate, population, and house price index. We used the "VARselect" function in R to find the optimal order of the model and then modeled every possible combination of the response-endogenous explanatory serieses. Then, we calculated the RMSE for each model using the predicted bankruptcy rate and the bankruptcy rate from the validation set.

Based on the principle of getting the minimum RMSE, VAR(p) model includes only Bankruptcy Rate and Population series with p = 10 is the "optimal" model for Vector Autoregressive process. Our "optimal" model in this case produced a RMSE of 0.003846416.

For this model, we further performed residual diagnostics to see if the "optimal" model satisfies the following assumptions: I. zero- mean, II. zero-correlation, III. constant variance, IV. normality. It turns out that variable the "Bankruptcy Rate" passed all assumptions tests except

that the zero-correlation assumption is not satisfied. Unfortunately, the variable Population did not pass any of the four assumption tests. Without passing all four assumptions, our model can no longer rely on a statistical distribution, which we need to generate valid intervals. In this case, our "optimal" model has limitations: the fitted model and the point predictions are valid, but the prediction intervals are no longer valid.

Multivariate Methods: VARX

This model is an extension of the VAR model. The VARX model is used when the data has a mix of both endogenous and exogenous variables. An ideal way of modelling such a data would be as follows:

- Construct a VAR(p) model
- Check for normality and if satisfied, look at the summary of the model to identify and separate homogenous and endogenous variables.

Unfortunately, the VAR(p) model does not satisfy normality for all of the variables. As in the VAR(p) model,we cannot rely on a statistical distribution.

To find an optimal VARX, we tried an iterative method. Since VARX model work only on multivariate analysis, the number of endogenous variables must be at least two and the number of exogenous variables must be at least one.

Based on the definition of population, we treated it as an exogenous variable. Our primary response variable must be kept as endogenous. That leaves us two free variables: "Housing Price Index" and "Unemployment rate"

Variables	Is the variable Included	Exogenous/Endogenous
Bankruptcy rate	True	Exo
Unemployment rate	True/False	Exo/Endo
Housing Price Index	True/False	Exo/Endo
Population	True/False	Endo

Only one exclusion is allowed at a time because of the reasons mentioned above. Therefore there are 12 different models across a grid of *p*-lags (whose min and max were defined using the VAR select function which checks the optimal value of p with using metrics such as AIC,BIC). A total of 72 models were tested. As in previous model, we found a seasonality of lag 12.

Observations:

For all the models, training RMSE and validation RMSE was tracked across a grid of *p*-lags. The training RMSE decreased from 0.002 to 0.0018 whereas validation RMSE decreased from 0.005 to 0.0039 when p was varied from 1 to 8. This suggests that there is no overfitting. Also dropping Population or Unemployment did not give any consistent improvement, suggesting no improvements could be made by dropping these variables. However, dropping HPI showed a clear sign of loss of model accuracy.

Assumptions and Limitations:

1. Zero- Mean test : Passed for all variables
2. Levene's test for constant variance : Passed for all variables
3. Ljung-Box test for uncorrelatedness: Failed for all variables
4. Shapiro-Wilk Test for Normality: Passed for all variables except Housing Price Index.

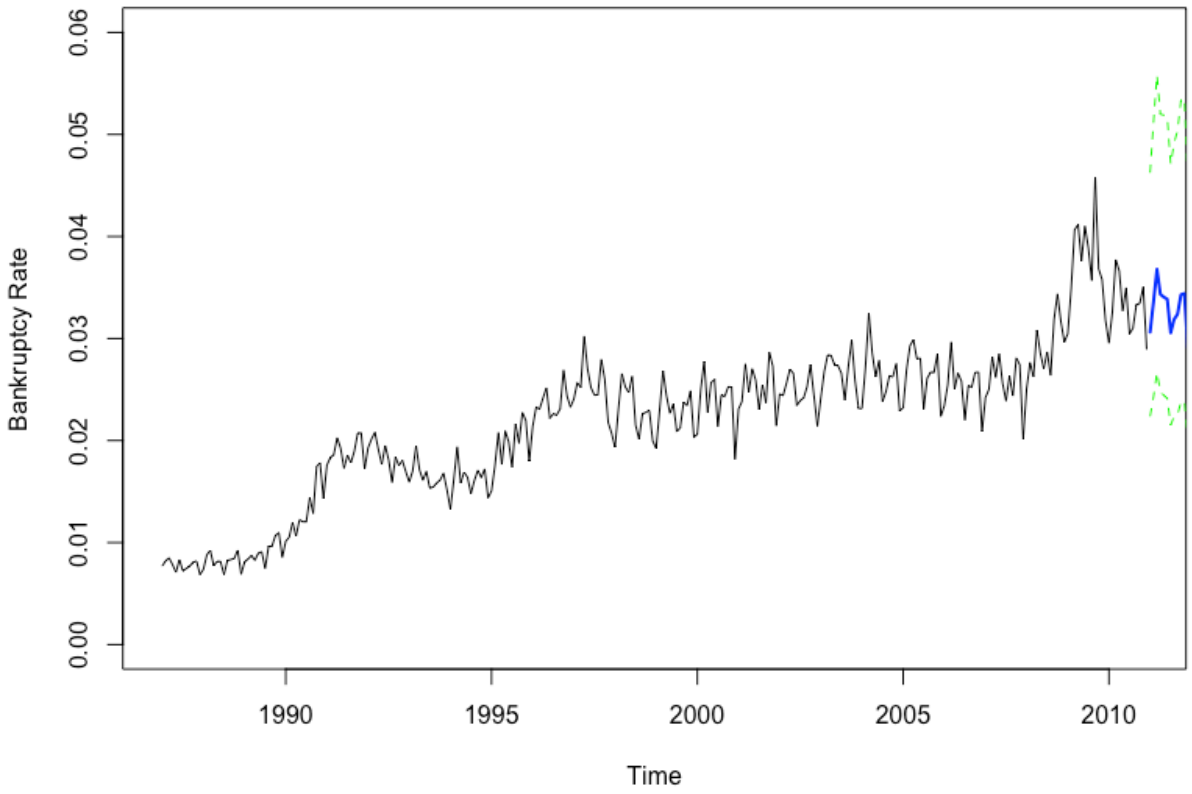
Ensembling

We decided to take a weighted average to ensemble all four models that we generated. We used an average in order to capture some information from each model. In order to generate the weights, we created a heuristic that gave additional weight to performance on our validation set but penalized model complexity. As many time series models can easily overfit the data with increased complexity, we wanted to ensure that our ensemble would generalize well.

Complexity	**Model**	**Validation RMSE**	**Model Weights**
14 (coefficients)	HW	0.00343	0.357
42 (r + pr^2)	VAR(p)	0.00385	0.183
93 (r + pr^2)	VAR(x)	0.00395	0.120
9 (coefficients)	SARIMAX	0.00449	0.340

* 1/(sqrt(complexity) * rmse) / sum(1/sqrt(complexity) * rmse)
* 0.07835778 0.26385784 0.59184439 0.06593998

Conclusion



Date	Point Estimate	Lower Estimate	Upper Estimate
Jan-2011	0.03062832	0.02236706	0.04624546
Feb-2011	0.03359887	0.02429433	0.05122409
Mar-2011	0.03678506	0.02668417	0.05583537
Apr-2011	0.03433334	0.02480042	0.05199276
May-2011	0.03405655	0.02442092	0.05189436
Jun-2011	0.03386036	0.02414477	0.05170226
Jul-2011	0.03055665	0.02150961	0.04706880
Aug-2011	0.03187083	0.02236923	0.04914597
Sep-2011	0.03237974	0.02250262	0.05039081
Oct-2011	0.03429561	0.02382081	0.05333377
Nov-2011	0.03441084	0.02385663	0.05356490
Dec-2011	0.02910082	0.02006772	0.04536672
Jan-2012	0.03097600	0.02106948	0.04893087
Feb-2012	0.03372487	0.02269222	0.05403709
Mar-2012	0.03623668	0.02431072	0.05837594
Apr-2012	0.03431398	0.02306960	0.05507993
May-2012	0.03460860	0.02313866	0.05595289
Jun-2012	0.03356806	0.02222598	0.05488587

Jul-2012	0.03123906	0.02058972	0.05131633
Aug-2012	0.03239858	0.02124784	0.05364089
Sep-2012	0.03322942	0.02162153	0.05568279
Oct-2012	0.03493450	0.02264427	0.05897476
Nov-2012	0.03438039	0.02213178	0.05867853
Dec-2012	0.02944778	0.01896783	0.05030879