Analyzing High-Speed Network Data

Kejia Hu*, Jaesik Choi+, Alex Sim+, Jiming Jiang*

University of California, Davis*, Lawrence Berkeley National Lab+

May 30, 2013

Outline

1. Introduction
   - NetFlow Data
   - GLMM: Generliazed Linear Mixed Model
   - Literature Review
2. Methodology
   - Gaussian Case
   - Poisson Case
   - Dimensionwise Selection of Tuning Parameters
3. NetFlow Data Study
   - GLMM on Duration
   - GLMM on Frequency of Congestion
4. Simulation
   - Prediction Accuracy
   - Modeling Time
5. Reference

Introduction

- Efficient data access is esstial for sharing massive amounts of data among many geographically distributed collaborators.
- The analysis of network traffic is important to efficiently utilize the limited resources offered by the network infrastrctures and plan wisely large data transfers.

Objectives

Data transfer performance for large dataset can be improved by learning the current condition and accurately predicting the future network performance.

- Short-term **prediction** of network traffic performance guides the immediate scientific data placements for network users.
- Long-term **forecast** of network traffic enables the capacity planning of the network infrastructure up to the future needs for network designers.

Chanllenges

Such predictions become non-trivial when

- amounts of network measurement data grow in unprecedented speed and volumes
- models are misspecified with high-dimension data

NetFlow Data

Start	End	SrcIPadd(masked)	Sif
SrcP	DstP	DstIPadd(masked)	Dif
P	Fl	Pkts	Octets
0930.23:59:37.920	0930.23:59:37.925	xxx.xxx.xxx.xxx	179
62362	22364	xxx.xxx.xxx.xxx	175
6	0	1	52
0930.23:59:38.345	0930.23:59:39.051	xxx.xxx.xxx.xxx	179
62362	28335	xxx.xxx.xxx.xxx	175
6	0	4	208
1001.00:00:00.372	1001.00:00:00.372	xxx.xxx.xxx.xxx	179
62362	20492	xxx.xxx.xxx.xxx	175
6	0	2	104

NetFlow Data: Practical Questions

- An accurate prediction of duration of a transfer can help choosing the start time, the path and the delivery condition.
- For network designers, accurate prediction will help prepare the future needs and match the network requirements and the bandwidth in long run.

GLMM: Generliazed Linear Mixed Model

GLMM with a vector of random effects v and the responses $y_1, ..., y_m$ of m groups that are conditionally independent. $f_i(y_i|v)$ follows the exponential family with

$$E(y_i|v) = \mu_i, g(\mu_i) = x_i'\beta + z_i'v \tag{1}$$

where $g(.)$ is the link function and $g^{-1} = h$, $v \sim N(0, \Phi)$

Motivation of Modeling the NetFlow using GLMM wigh Lasso

- Uneven colletting time stamps of NetFlow.
  GLMM can have a group of variables with no requirement for even-space time.
- NetFlow record showing mixed effects.
  GLMM is general in the sense of 1. relating the regression model to the response variable via link function 2. categorizes the variance source by measuring the random effects.
- NetFlow measurements are of super large volume and high dimension.
  Lasso brings in an efficent fast algorithm to select and fit the model for the dataset.

Review on Penalized methods-Lasso in Models

- Jiming Jiang, Thuan Nguyen and J. Sunil Rao (2011): Best Predictive Small Area Estimation, Journal of the American Statistical Association, 106:494, 732-745
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso.,Journal of the Royal Statistical Society Series B, 58, 267288.
- Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010) Joint variable selection for fixed and random effects in linear mixed-effects models., Biometrics, 66, 10691077.
- Ibrahim JG, Zhu H, Garcia RI, Guo R. (2011) Fixed and random effects selection in mixed effects models., Biometrics, 67, 495-503.

Predictive Generalized Linear Mixed Modeling via the Lasso

Prediction is our main goal. Thus we consider how to use Lasso to select random effects and fixed effects to gain the OBP with minimum MSPE.

Two advantages for our new methods:

- not influenced by the error caused by misspecification model
- much less computationally costly then Penalized Model Selection because Penalized Model Selection require computational intensive EM algorithm

Moreover, we develop a tuning parameters selection methods based on bootstrap.

Prediction with Lasso-Gaussian

The GLMM can model both continuous variable and discrete variable as response variable.
First focuses on the interest to predict continuous variable.
The BPE is $\check{\beta} = (X'\Gamma'\Gamma X)^{-1}X'\Gamma'\Gamma y$
The MSPE to be minimized is

$$MSPE(\check{\theta}) = E(|\check{\theta} - \theta|^2)$$
$$= E(|H'v + F'e|^2) - 2E((v'H + e'F)\Gamma(y - X\beta))$$
$$+E((y - X\beta)'\Gamma'\Gamma(y - X\beta))$$
$$= I_1 - 2I_2 + I_3$$

Prediction with Lasso-Gaussian

- Fixed Effects Selection:
  $\min_\beta Q(\beta) = (y - X\beta)'M'M(y - X\beta) + \lambda \sum |\beta_i|$
- Random Effects Selection:
  $\min_d Q(d) = (y - X\beta)'M'M(y - X\beta)$
  $+tr((2HBZ - HF'Z - Z'FR')G) - tr(FB\Sigma) + \lambda \sum |d_i|$

Prediction with Lasso-Poisson

Now focus on the prediction interest that the observations are counts.
Given the small area means $\mu_1, ..., \mu_m$, the observations $y_1, ..., y_m$ (with y_i being from the ith small area) are independent such that

$$y_i \sim Poisson(\mu_i); log(\mu_i) = x_i'\beta + d_i v_i$$

where x_i is a vector of covariates; v_i is an area-specific random effect, $v_i \sim N(0, 1)$.

Prediction with Lasso-Poisson

BP of μ_i under the assumed model is

$$E_{M,\psi}(\mu_i|y) = g_{M,i}(\psi, y_i)$$

where $E_{M,\psi}$ denotes conditional expectation under model M and paramter vector $\psi = \{\beta, d\}$
The overall MSPE can be expressed as

$$\begin{aligned} MSPE &= \sum_{i=1}^{m} E\{g_{M,i}(\psi, y_i) - \mu_i\}^2 \\ &= E\{\sum_{i=1}^{m} g_{M,i}^2(\psi, y_i)\} - 2\sum_{i=1}^{m} E\{g_{M,i}(\psi, y_i)\mu_i\} + \sum_{i=1}^{m} E(\mu_i^2) \\ &= I_1 + 2I_2 + I_3 \end{aligned}$$

Prediction with Lasso-Poisson

$$
\begin{aligned}
E\{g_{M,i}(\psi, y_i)\mu_i\} &= E\{\mu_i E\{g_{M,i}(\psi, y_i)|\mu\}\} \\
&= \sum_{k=0}^{\infty} g_{M,i}(\psi, k) E(e^{-\mu_i}\mu_i^{k+1}/k!) \\
&= \sum_{k=0}^{\infty} (k+1) E(e^{-\mu_i}\mu_i^{k+1}/(k+1)!)
\end{aligned}
$$

Furthermore $E(e^{-\mu_i}\mu_i^{k+1}/(k+1)!) = E\{1_{(y_i=k+1)}\}$

Prediction with Lasso-Poisson

Thus, if we define $g_{M,i}(\psi, -1) = 0$, then we have

$$E\{g_{M,i}(\psi, y_i)\mu_i\} = E\{\sum_{k=0}^{\infty} g_{M,i}(\psi, k)(k+1)1_{(y_i=k+1)}\}$$
$$= E\{g_{M,i}(\psi, y_i - 1)y_i\}$$

Prediction with Lasso-Poisson

Till now, we have the expression

$$MSPE = E\{\sum_{i=1}^{m} g_{M,i}^2(\psi, y_i) - 2\sum_{i=1}^{m} g_{M,i}(\psi, y_i - 1)y_i + I_3\}$$

To minimize MSPE is equivalent to minimizing

$$Q(\psi) = \sum_{i=1}^{m} g_{M,i}^2(\psi, y_i) - 2\sum_{i=1}^{m} g_{M,i}(\psi, y_i - 1)y_i$$

To minimize the fixed effect, $Q(\beta) = Q(\psi) + \lambda_\beta \sum_{j=1}^{p} |\beta_j|$

To minimize the random efffect, $Q(d) = Q(\psi) + \lambda_d \sum_{i=1}^{m} |d_i|$

Prediction with Lasso-Poisson

To expand $g_{M,i}(\psi, y_i)$, we first write down the distribution.
$v_i \sim N(0,1)$ and $f(v_i) = \frac{1}{\sqrt{(2\pi)}} exp(-\frac{1}{2} v_i^2)$

$y_i \sim Poisson(\mu_i)$ and $f(y_i|\mu_i) = \frac{exp(-\mu_i)\mu_i^{y_i}}{y_i!}$
Since $log(\mu_i) = x_i'\beta + d_i v_i$, then

$$f(y_i|v_i) = \frac{exp(-[exp(x_i'\beta + d_i v_i)])[exp(x_i'\beta + d_i v_i)]^{y_i}}{y_i!}$$

Prediction with Lasso-Poisson

The BP can be written as the following

$$
\begin{aligned}
g_{M,i}(\psi, y_i) &= E_{M,\psi}(\mu_i | y_i) \\
&= \int \mu_i f(\mu_i | y_i)\, d\mu_i \\
&= \int \mu_i \frac{f(\mu_i, y_i)}{f(y_i)}\, d\mu_i \\
&= \int \mu_i^2 d_i \frac{f(v_i, y_i)}{\int f(y_i, vi)\, dv_i}\, dv_i \\
&= \int \mu_i^2 d_i \frac{f(y_i | v_i) f(v_i)}{\int f(y_i | vi) f(v_i)\, dv_i}\, dv_i
\end{aligned}
$$

Select Tuning Parameters

M_{opt} represents the optimal model
$M_0 = M_0(\lambda)$ represent the model selected by PGLMM given Lasso with λ

We want to set our tuning parameter λ as the one
$argmax P(M_0(\lambda) = M_{opt})$

The idea is nice, but two problems rise.

- What is P?
- What is M_{opt}?

Select Tuning Parameters

Reference: Fence methods for mixed model selection by Jiang et.al from Annals of Statistics 2008, Vol. 36, No. 4, 1669-1692

For P

- The assumptions that full model M_f is the model contains all information
- Boostrap samples under M_f and it almost duplicate the information from original data
- The sample allows us to approximate the probability distribution P

Select Tuning Parameters

For M_{opt}, use the idea of maximum likelihood.

The optimal model is the model from which the data is generated, then this model should be the most likely given the data.
Thus, given λ, we look for the model that is most supproted by the data or alternatively the bootstrap sample since it almost duplicates the information in the original data.

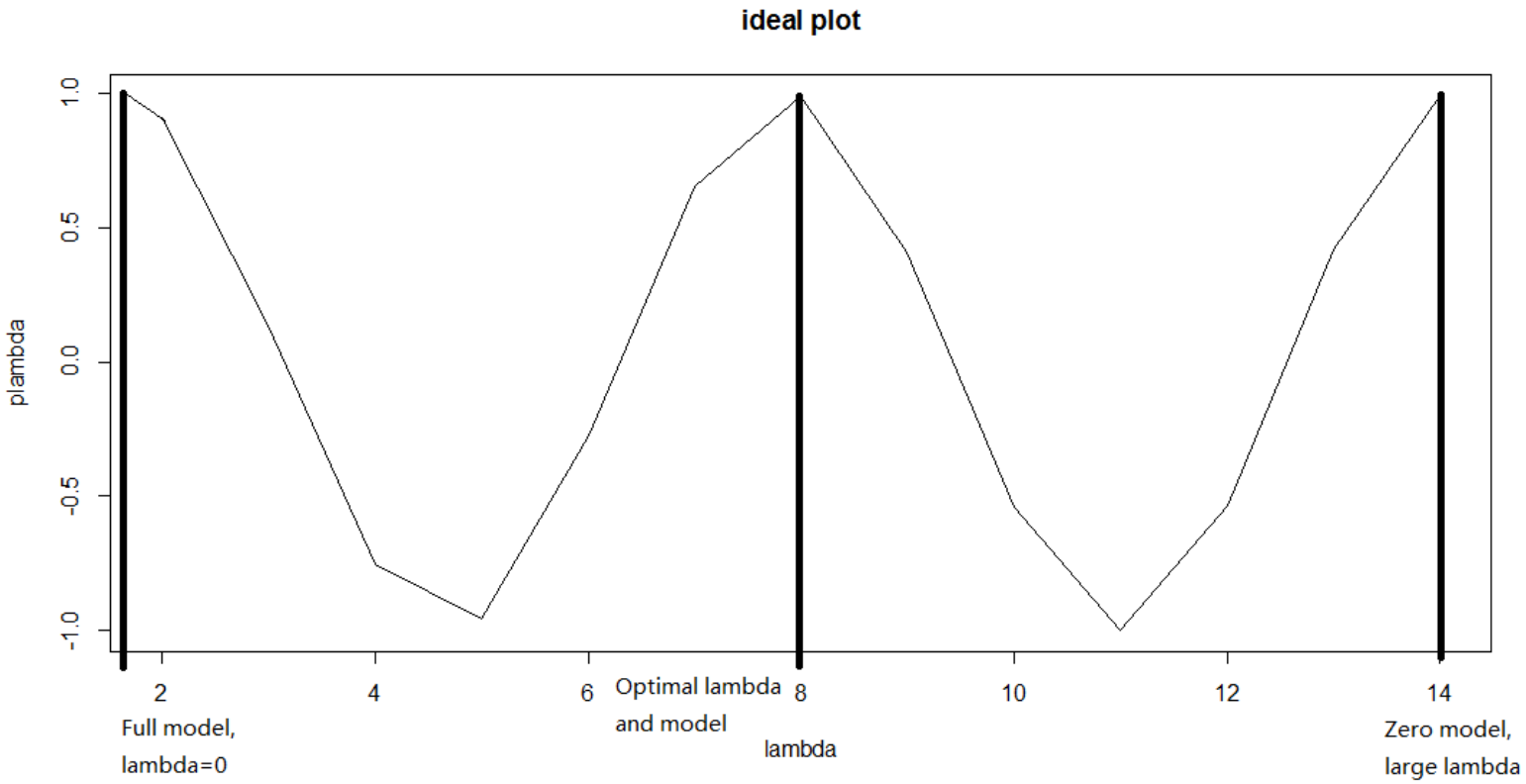
To define most supported,$p^*(M) = P^*(M_0 = M)$ where P^* denotes the empirical probability obtained by bootstrapping.
$\max_{M \in M_{set}} p^*(M)$

Select Tuning Parameters

The procedure is:

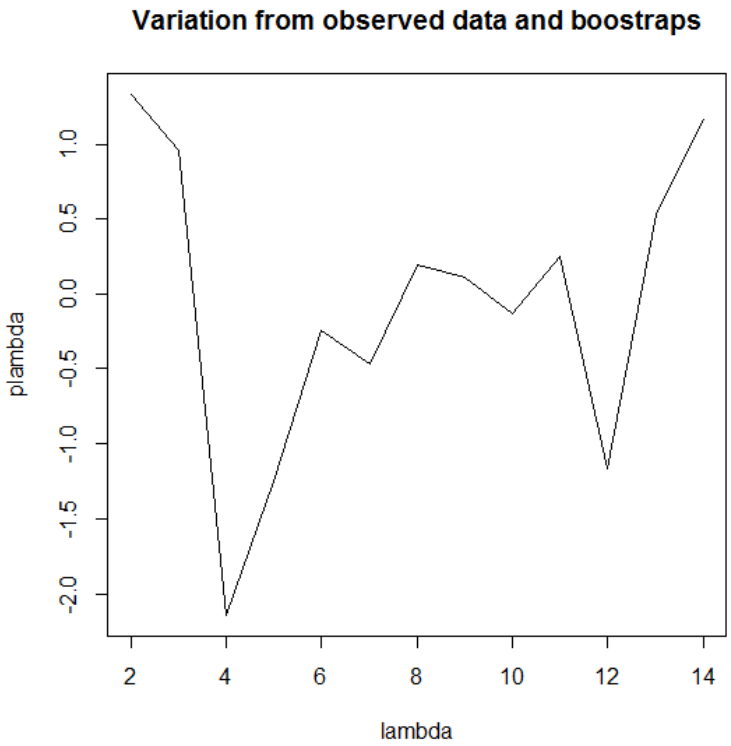
- Step 1: Select a grid of λ
- Step 2: For each λ, calculate the model selected most across all Bootstrap samples B
  The model $M*$ satisfies $p^*(M*, \lambda) = \max_{M \in M_{set}} p^*(M, \lambda)$
  where $p^*(M, \lambda) = P^*(M_0 = M|\lambda)$
- Step 3: Plot $p^*(\lambda)$ against λ. Find the peak in the middle and its corresponding λ_{opt}
  $\lambda_{opt} = argmax_\lambda P(M*(\lambda) = M_{opt})$

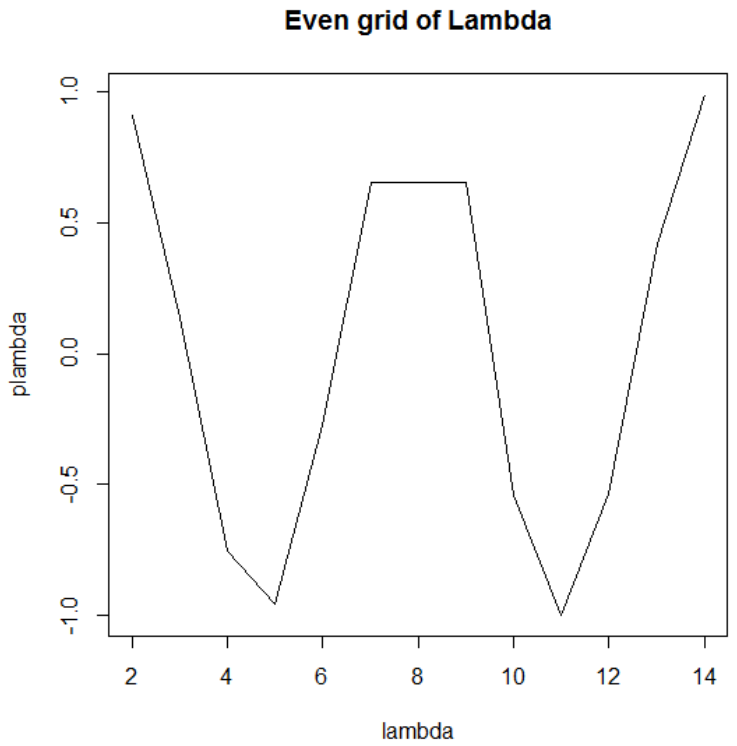
Select Tuning Parameters



Select Tuning Parameters

But the ideal situation doesn't alwasy come, many times we will end up in the following cases.





Dimensionwise Selection of Tuning Parameters

We no longer use even grid of lambda, rather use a dimension-related lambda.

For λ_j, it gives us the selected model with jth predictors.

Solving the platform, because each λ_j now selects model with different number of predictors and the corresponding probability $p^*(M*, \lambda)$ are not likely to be the same in the neighboring range.

Solving the variation,a more robust and sophiscated choice of λ_j eliminate the unwanted wigling in the plot

Dimensionwise Selection of Tuning Parameters

- Over a grid of λ, start from smallest $\lambda_p = 0$ return a P model. Keep increasing λ, until a P-1 model returns and record the current value as λ_{p-1}
  From $[\lambda_p, \lambda_{p-1})$ is the range that model with p predictors are chosen.
- Continue the above step, until we get all the range for model selection.
  For $i = 0, ..., p$, $[\lambda_i, \lambda_{i-1})$ is the range that model with i predictors are chosen.

Dimensionwise Selection of Tuning Parameters

- For each range, evenly seperate it into a grid by k candidate λ. For each λ, compute the model across all Bootstrap sample and chosen the optimal λ within this range as λ_i^* and the corresponding $p_i^*(M_i^*, \lambda_i^*)$
- Plot $p_i^*(M_i^*, \lambda_i^*)$ against λ_i^* and select the middle peak to be the overall optimal λ^* and its corresponding model as the optimal model M^*.

Dimensionwise Selection of Tuning Parameters

Based on Bootstrapped sample, use a dimension-related lambda to eliminate the unwanted wigling

dim	P	P-1	...	i	...	0
λ range	$[\lambda_p, \lambda_{p-1})$	$[\lambda_{p-1}, \lambda_{p-2})$	...	$[\lambda_i, \lambda_{i-1})$	...	$[\lambda_0, \lambda_{-1})$
λ_i^*	λ_p^*	λ_{p-1}^*	...	λ_i^*	...	λ_0^*
p_i^*	p_p^*	p_{p-1}^*	...	p_i^*	...	p_0^*

Dimensionwise Selection of Tuning Parameters

Based on Bootstrapped sample, the dimensionwise selection of tuning parameters solve the two original problems

- The platform case is solved because each λ now selects the model with different number of predictors, and the corresponding probability is not likely to be the same in the neighboring range.
- The variation case is solved because more robust and sophisticated choice of λ_j eliminates the unwanted wiggling in the plot.

Now the resulted plot is more close the shape in the ideal case.

NetFlow Data Study

The NetFlow data is provided by ESnet for the duration from May 1 2013 to June 30 2013. The data size is of 10^9 observations. Two models are built:

- GLMM on short-term transfer duration time.
- GLMM on long-term congestestion

The estimates of the model and its prediction accuracy compared with two traditional GLMM algorithms: Backward-Forward selection and Estimation-based Lasso.

GLMM on Duration

The full model predicts the transfer duration, assuming influences from the fixed effects including transfer start time, transfer size (Octets and Packets) and the random effects including network transfer condition such as Flag and Protocol, source and destination Port numbers and transfer path such as source and destination IP addresses and source and destination Interfaces. After selecting and fitting the model via our Predictive Lasso procedure, the final model is

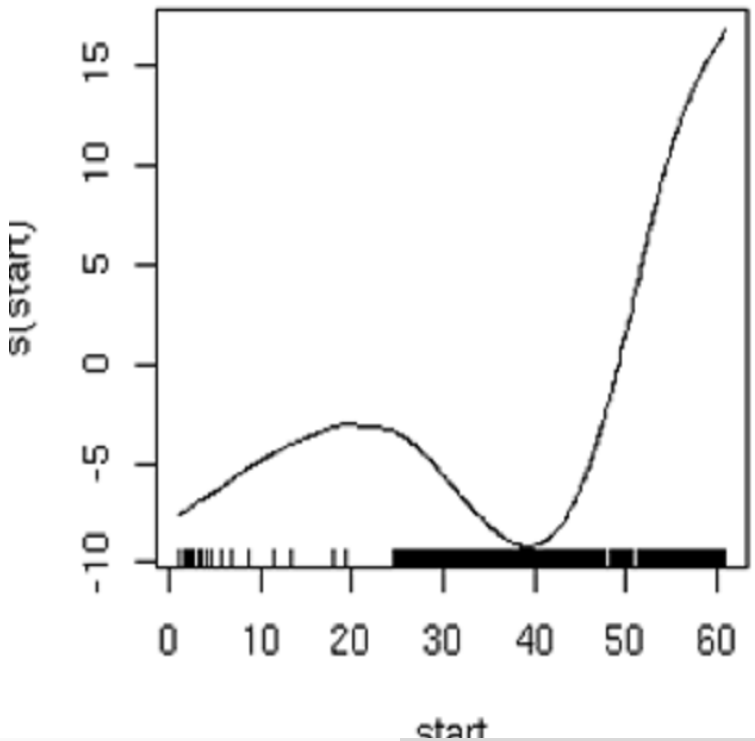
$$y = \beta_{start} s(x_{start}) + \beta_{pkt} x_{pkt} + Z_{ip-path} v_{ip-path} + e \qquad (2)$$

GLMM on Duration

- Since time variable is usually not linear related with the response variable, smoothing spline transformation $s(.)$ is implemented and shwon in Figure 1.
- P-value in the final model is all less than 2e-16. The significant fixed effects in the model are start time and number of packets, as shown in Table 1.
- The random effects' standard deviance estimates are plotted in Figure 2, and the plot shows the traffic duration varies with the different IP paths.
- Besides the uncertainty resulted from each IP path, the background noise is estimated with a standard deviation of 11.2392.

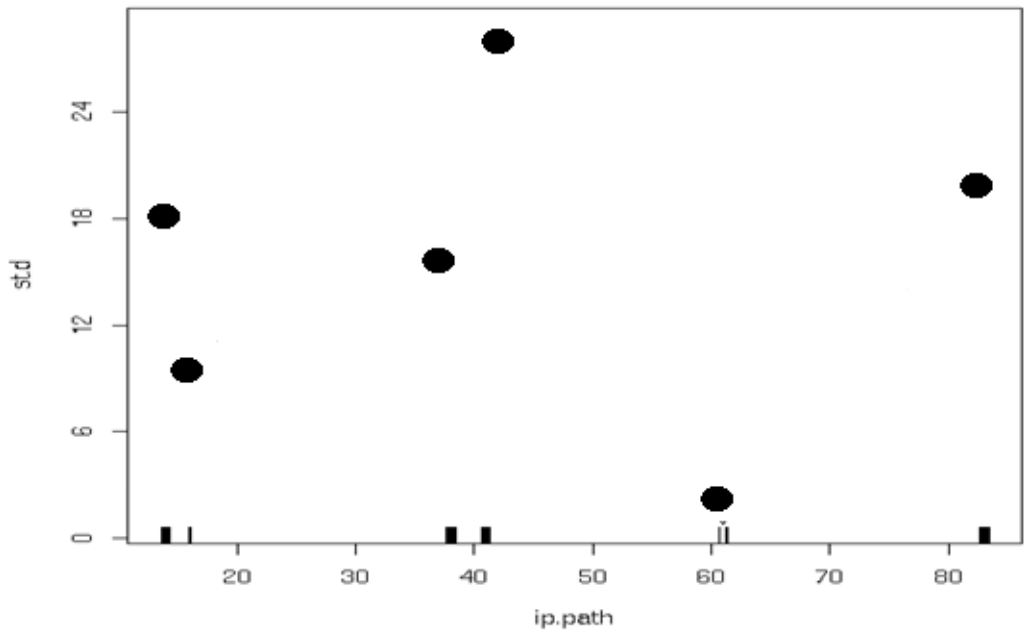
GLMM on Duration

Figure : Smoothing Spline Transformed Start Time Variable, showing the nonlinear relationship between start time and transfer duration



GLMM on Duration

Figure : Standard Deviation Estimates for Random Effects in GLMM (2) to Predict The Transfer Duration, showing the busier paths bring higher variation to the transfer duration



GLMM on Duration

Table : Coefficient Estimates for Fixed Effects in GLMM (2) to Predict the Transfer Duration

Fixed Effects	Estimates	Standard Deviation	P-value
Intercept	-13.809	0.914	<2e-16
Start Time	0.574	0.0169	<2e-16
Packets	1.115	0.035	<2e-16

GLMM on Duration

Table : Comparison of MSPE and Modeling Time to Predict Duration

	Est.Lasso s	BF Selection	Pred. Lasso
MSPE	2306	42230	127.3
Time (in seconds)	6.26e+7	5.43e+10	142

GLMM on Frequency of Congestion

This model predicts the frequency of congestion occurred in each link of the network. The response variable y is the number of congestion measured by the speed, BytesPerSecs. A congestion event is defined when BytesPerSecs is less than 50, which is the slowest 10% of network transfer speed.
The full model to predict number of congestion assuming influence from two sources: fixed effects and random effects. The fixed effects includes transfer size (Octets and Packets), number of transfers with their Protocol is 6, 17, 47 and 50 respectively and number of transfers with their Flag is 0,1,2 and 4 respectively. And random effect is transfer path, the source and destination IP address.

GLMM on Frequency of Congestion

After selecting and fitting the model via our Predictive Lasso procedure, the final model is

$$logE(y|v) = \beta_{pkts}x_{pkts} + \sum \beta_{p=i}x_{p_i} + Z_{ip-path}v_{ip-path} \quad (3)$$

- The significant fixed effects in the selected model are the transfer size, number of packets and the protocol used, as shown in Table 3.
- The random effects' standard deviance estimates are plotted in Figure 3 showing that the traffic duration in Y axis varies in different transfer IP paths in X axis.
- The background noise is estimated with a standard deviation of 25.316.

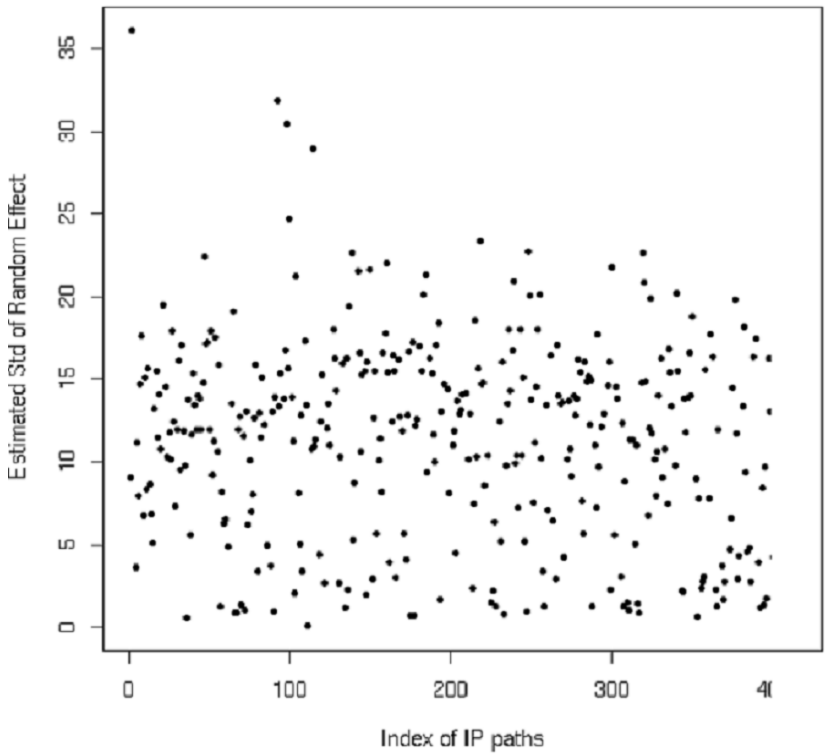
GLMM on Frequency of Congestion

Table : Coefficient Estimation for Fixed Effects in GLMM (3) to Predict The Frequency of Congestion

Fixed Effects	Estimates	Std	P-value
Intercept	816.95627	0.78305	<2e-16
Packets	28.53924	0.02284	<2e-16
Protocol=6	45.43606	0.01608	<2e-16
Protocol=17	-1.58644	0.14088	<2e-16
Protocol=47	-8.39576	0.36338	<2e-16
Protocol=50	-4.96028	0.05175	<2e-16

GLMM on Frequency of Congestion

Figure : St.D Estimates for Random Effects in GLMM (3) to Predict The Frequency of Congestion, showing the busier paths bring higher variation to the frequency of congestion



GLMM on Frequency of Congestion

Table : Comparison of MSPE and Modeling Time to Predict Counts of Congestion

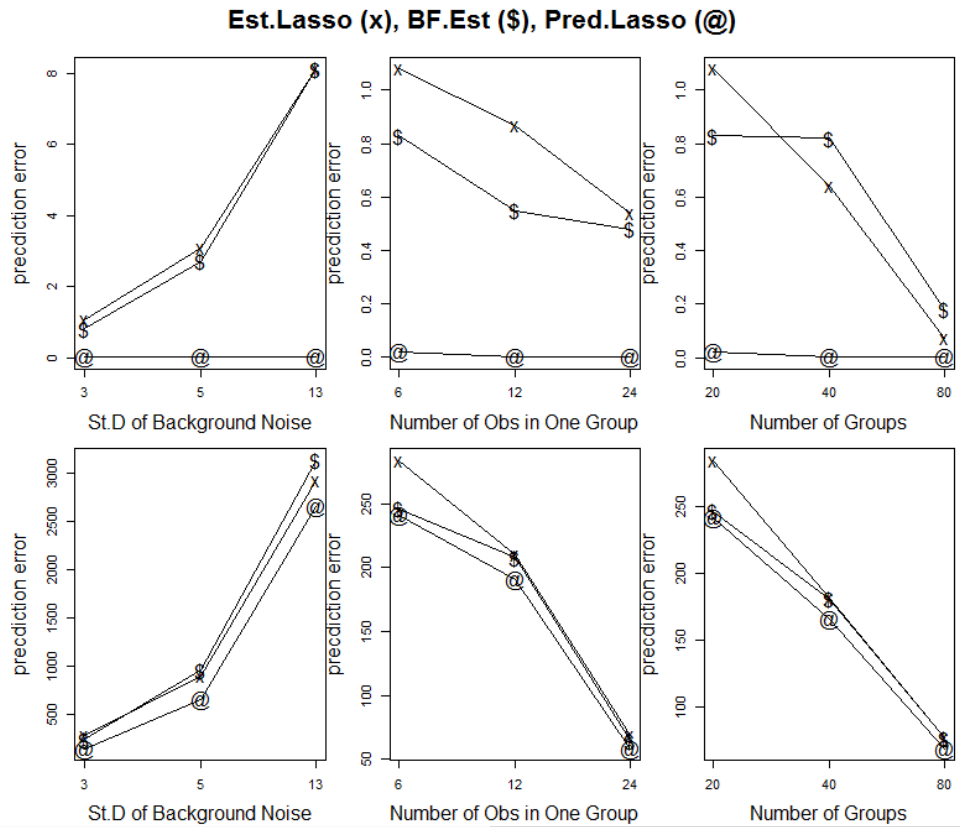
	Est.Lasso s	BF Selection	Pred. Lasso
MSPE	27.7	42.5	12.73
Time (in seconds)	7.31e+8	1.24e+10	10.06e+2

Simulation Procedure

- Step 1: Generate simulated data
- Step 2: Raise up an assumed model with redundant fixed effects and mixed effects
- Step 3: Three model selection/estimation techniques
- Step 4: Evaluate the **Prediction Accuracy and Modeling Time.**

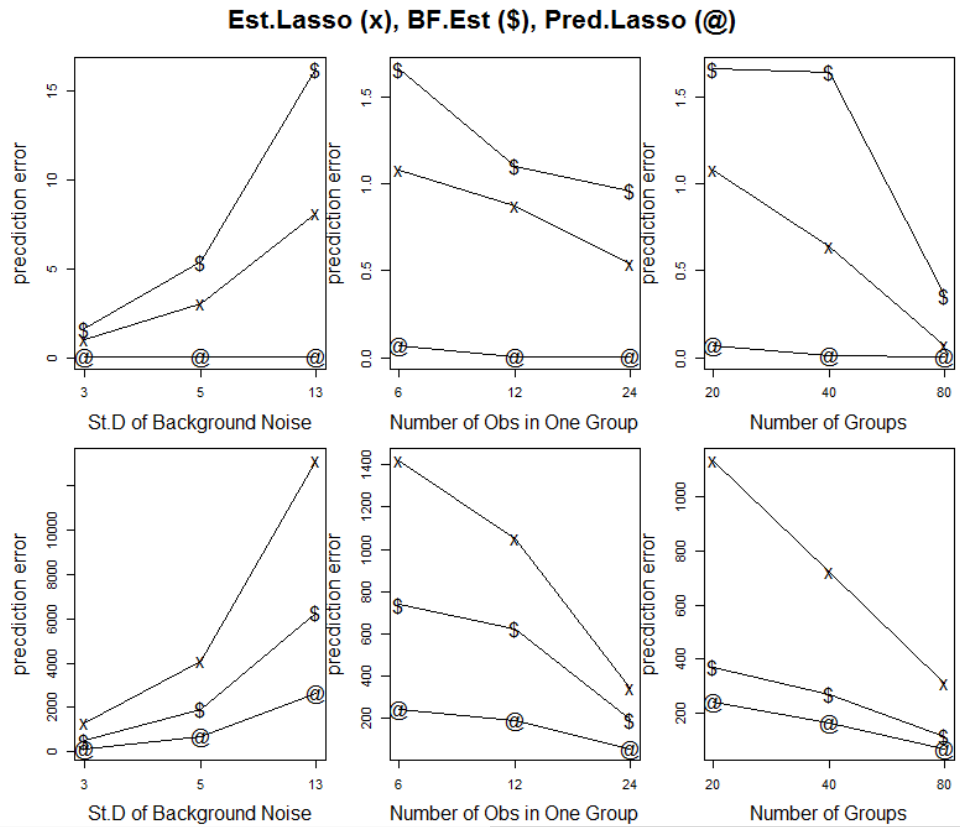
Simulation Result

Figure : Prediction Accuracy under Case 1: Gaussian Model, showing the Predictive Lasso has the smallest prediction error



Simulation Result

Figure : Prediction Accuracy under Case 2: Poisson Model, showing the Predictive Lasso has the smallest prediction error



Complexity and Modeling Time for Three Comparison Methods

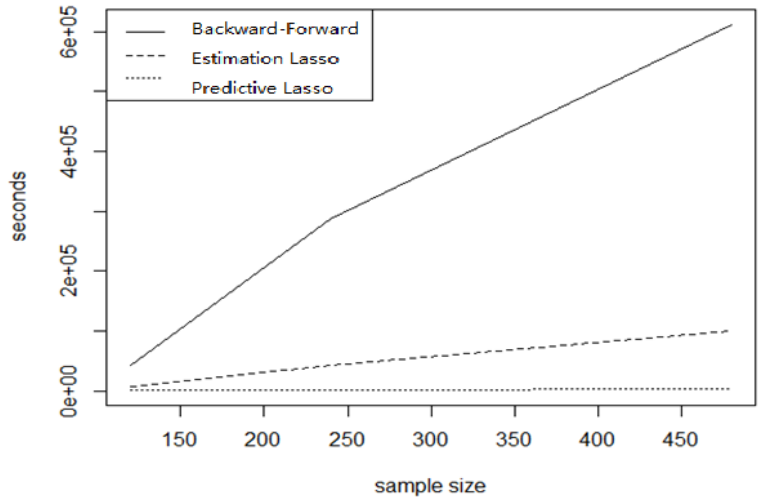
Table : Complexity and Modeling Time for Three Comparison Methods

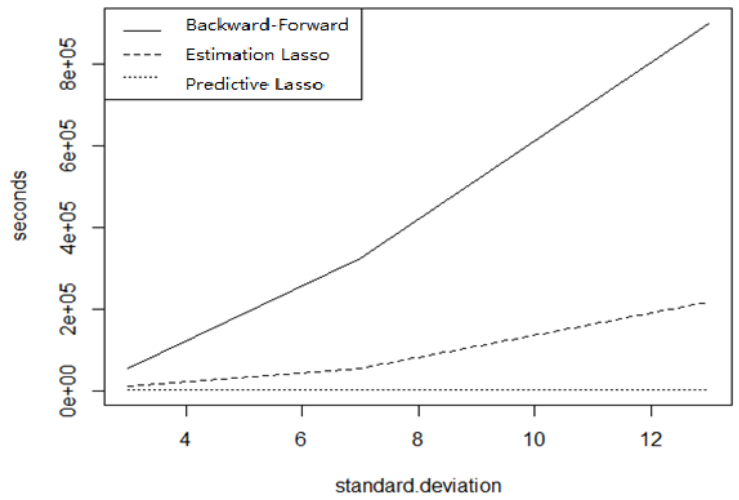
	Est.Lasso s	BF Selection	Pred. Lasso
Time	$MCEM \times k$	$MCEM \times \sum_{i=1}^{I} k_i \sum_{j=1}^{J_i} n_{ij}$	$Opt \times 1$
Complexity	$O(npk)$	$O(np \sum_{i=1}^{I} k_i \sum_{j=1}^{J_i} n_{ij})$	$O(np)$

Note: MCEM=MC+Opt(imization).
k, k_i, n_{ij} are measurement of iteration steps in each method and are strictly greater then 1.

Simulation Procedure

Figure : Computational Costs of Three Methods, showing the Predictive Lasso has the least computation time for both the size of the data increases (left) and the uncertainty in the data increases (right)





Reference

- Stallings, W.(1999). SNMP, SNMthp2, SNMPv3 and RMON 1 and 2, Addison-Wesley
- Cisco Systems Inc.(1966).NetFlow Services and Applications - White paper
- Tibshirani, R.(2011).Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B 58,267288.
- Jiang, J.,Nguyena T. and Rao,J. S.(2011).Best Predictive Small Area Estimation.Journal of the American Statistical Association 106:494, 732-745
- Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010).Joint variable selection for fixed and random effects in linear mixed effect models. Biometrics 66, 10691077

Reference Ctd

- Jiang, J., Rao, J.S., Gu, Z. and Nguyena T. (2008).Fence methods for mixed model selection. Annals of Statistics. 36-4,1669-1692
- Hu, K., Sim, A., Antoniades, D. and Dovrolis, C.(2013).Estimating and Forecasting Network Traffic Performance Based on Statistical Patterns Observed in SNMP Data. MLDM 2013 601-615
- Antoniades, D., Hu, K., Sim, A., Dovrolis, C.(2013) What SNMP Data Can Tell Us about Edge-to-Edge Network Performance. PAM 2013. 267-269
- Hu, K., Choi, J., Jiang, J. and Sim, A.(2013) Best Predictive GLMM using LASSO with Application on High-Speed Network. LBNL Tech Report 6327E, 2013

Reference Ctd

- Ibrahim, JG., Zhu, H., Garcia, RI. and Guo R.(2011) Fixed and Random Effects Selection in Mixed Effects Models. Biometrics, 67, 495-503
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 1-38

Q and A

Q and A
Thank you