Screening-Based Bregman Divergence Estimation and the Application to Spike Train Data Analysis

by

Yi Chai

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: 2/13/14

Committee members:

Chunming Zhang, Professor, Statistics
Zhengjun Zhang, Associate Professor, Statistics
Karl Rohe, Assistant Professor, Statistics
Yingqi Zhao, Assistant Professor, Biostatistics & Medical Informatics
Wei-Yin Loh, Professor, Statistics

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor Professor Chunming Zhang, for her introduction to the amazing world of statistics, and her guidance and encouragement through the entire course of my research. Prof. Zhang has selflessly shared original research ideas with me and guided me patiently in every stage of my graduate study. Without her enlightening instruction, kindness and patience, I could not have completed my dissertation. Her dedication in statistical and scientific research motivated me to conduct original research in pursuit of knowledge and excellence not only in this dissertation but also in my future career. It has been my great honor and privilege to have had the opportunity to work closely and learn from her.

I would also like to extend my sincere thanks to Professor Zhengjun Zhang, Professor Karl Rohe, Professor Yingqi Zhao and Professor Wei-Yin Loh, for their service in my thesis committee. Their careful reading of this dissertation and valuable suggestions and comments help me to improve the thesis. Special thanks go to Professor Zhengjun Zhang, for his

insightful comments in the group meetings and unceasing encouragement during my time as a Ph.D. student.

Besides, I am very grateful for the time I spent in the Department of Statistics at University of Wisconsin-Madison. My solid background is built from receiving rigid training here in both theories and applications of statistics, which are not only beneficial to my research but also priceless for my career. Thanks go to all the faculty members, staff and graduate students in the department.

Finally I would like to thank my dear parents Xiaoming Chai and Taifen Zhang, for their years of support and understanding. Their selfless love supports me every day during my study. I dedicate this dissertation to them.

Contents

List of Tables

List of Figures

SCREENING-BASED BREGMAN DIVERGENCE ESTIMATION AND THE APPLICATION TO SPIKE TRAIN DATA ANALYSIS

Yi Chai

Under the supervision of Professor Chunming Zhang

At the University of Wisconsin-Madison

The thesis consists of two parts. In the first part, we extend the marginal screening procedure, by means of Bregman divergence as the loss function, to a broader class of models, including not only the GLM but also the quasi-likelihood model. A sure screening property is established for the proposed screening procedure and the simulation results and real data studies illustrate that a two-step procedure, which combines the feature screening in the first step and a penalized-BD estimation in the second step, is practically applicable to identify the set of important variables and achieve good estimation of model parameters, with the computational cost much less than those without using the screening step.

The second part of the theis presents one application to the spike train data obtained from the prelimbic region of the frontal cortex of adult male rats when performing a T-Maze based delayed-alternation task of working memory. We propose a structured regularization method based on Sparse Group Lasso under the GLM framework to estimate the functional connectivity among neurons. We also provide the theoretical properties of the proposed method and develop an efficient algorithm to handle the

large spike train dataset with the complex penalty. The simulation results show that our method performs better than other existing methods. Our results from the real data give some insight into the neuronal network in that region of rats' brain.

Chunming Zhang

Chapter 1

Introduction

The thesis consists of two parts. In the first part, we extend the marginal screening procedure, by means of Bregman divergence as the loss function, to a broader class of models, including not only the GLM but also the quasi-likelihood model. The second part presents one application to the spike train data obtained from the prelimbic region of the frontal cortex of adult male rats when performing a T-Maze based delayed-alternation task of working memory.

In the recent literature, there has been a tremendous amount of work on the high (and ultra-high) dimensional regression estimation and classification. These types of studies arise frequently from many different areas of scientific research, such as fMRI brain images, microarrays, genomics, financial data, and internet traffic data. With the development of new technologies, we are now able to collect data sets which are much larger and

more complex than we could have imagined a few years ago. In certain applications, we can even see that the dimensionality $p = p_n$ can grow with the sample size n.

For problems when p_n grows faster than n, the classical regression model with p_n parameters is not identifiable. On the other hand, in many applications only a small number of variables among all p_n predictors would really have actual impact on the response variable. Thus, a sparse structure is usually assumed in such cases. As a result, those techniques that can generate sparse solutions are preferred and extensively studied. Regularization is one of the most commonly used techniques aiming at obtaining well behaved solutions to overparameterized estimation problems. Numerous variable selection methods, based on regularization/penalization, have been developed, including the bridge regression (Frank and Friedman (1993)), the Lasso (Tibshirani (1996)), the SCAD (Fan (1997)), the MCP (Zhang (2010)), the Dantzig selector (Candes and Tao (2007)), among many others.

Fan and Lv (2008) proposed another approach, which is a screening procedure to select important variables based on their marginal correlations. The "sure independence screening" property was established under certain conditions in their work. Fan and Song (2010) extended the sure independence screening procedure to the generalized linear model (GLM) (McCullagh and Nelder (1989)). Their result works well for the GLM, but is somewhat restrictive, since their arguments largely depend on the

nice properties associated with the exponential family and the canonical link function. Zhu et al. (2011) proposed a model-free feature screening approach (SIRS) using a special marginal utility measure based on the conditional distribution of the response given predictors. They showed that their approach can rank the important variables above unimportant ones asymptotically in multi-index models.

Those works inspire us to explore the suitability of applying the screening procedures to more general models, for example without either the explicit form of distributions or any parametric forms between response and predictors. In this work, we extend the marginal screening procedure, by means of Bregman divergence (BD), to a wider class of screening procedure, including not only ranking by marginal maximum likelihood estimate in the GLM, which has been studied by Fan and Song (2010), but also ranking by the quasi-likelihood (McCullagh (1983)), which has been less developed, and a lot more. An interesting example is given in Section 2.4.1 for overdispersed Poisson responses, to which the conventional GLM is not applicable. Furthermore, there are a few loss functions widely used in machine learning systems, for example hinge loss for the support vector machine (Vapnik (2000)) and exponential loss for boosting (Hastie et al. (2001)), but not generally fulfill GLM model assumptions. Bregman divergence can be used to unify many commonly used loss functions and simultaneously study their asymptotic behavior.

The proposed method utilizes the marginal regression minimum-BD

estimator of each predictor and ranks their importance according to the absolute values of the marginal estimates. The sure screening property can be established based on a non-asymptotic probability bound for the occurrences of selection inconsistency. This means that all the truly important variables will be selected with overwhelming probabilities and the results are applicable under NP-dimensionality which allows the dimension p_n to grow as fast as $\log(p_n) = O(n^a)$ for some $a \in (0, 1)$. Note that such results are obtained under a very general framework in which the distribution of the response, conditional on the predictors, is allowed to be incompletely or not fully specified, and only certain moment conditions and tail properties are assumed. Our results also do not require that the predictor either follows an elliptically contoured distribution as in Fan and Song (2010) or satisfies linearity condition as in Zhu et al. (2011). Although we use certain functional form of the marginal estimates, the main result of sure screening property does not require a particular parametric model, and it reveals that different choices of BD will only affect some constants in the probability bound.

We carry out numerical assessment and comparison of the proposed screening method and the resulting estimation performance of several popular methods for the final model, via both simulation studies and real data analysis. Our screening method based marginal regression minimum-BD estimator performs well comparing with SIRS, especially in the most stringent simulation setting in which both response and predictors are

binary and very sparse. The results show that the two-step procedure is practically applicable. Another considerable advantage enjoyed by this two-step procedure is that by filtering out most of the unimportant variables in the first step, we can greatly reduce the computational expense for parameter estimation in the second step which is usually more costly. Thus, the computation times of the two-step procedure in the simulation are just a fraction of those without using the screening step.

Besides the above theoretical development of BD estimation, we also study one particular spike train data obtained from adult male rats when performing a T-Maze experiment. We adopt the generalized linear model (GLM) framework, which is a special case of BD and propose a structured regularization method for the spike train data to utilize certain structural information of the parameter space in the GLMs and better investigate the functional connectivity within a neuronal network. We also provide the theoretical properties of proposed method and develop an efficient algorithm to handle the complexity of the optimization problem for large spike train data. The simulation results show that our proposed method performs better than existing methods. Our results from the real data give some insights into the neuronal network in that region.

The rest of the thesis is organized as follows. Chapter 2 develops a screening procedure based on componentwise regression minimum-BD estimation. We establish its sure screening property, and provide simulation results and real data applications. Chapter 3 presents one

application to the spike train data analysis. We propose a structured regularization method based on Sparse Group Lasso under the GLM framework to estimate the functional connectivity among neurons, as well as an efficient algorithm. The theoretical properties and simulation results are provided to support the proposed method. Chapter 4 concludes with a brief discussion and future work. Details of technical assumptions and proofs are relegated to the Appendix.

Chapter 2

Screening-Based Bregman Divergence Estimation with NP-Dimensionality

2.1 Regression model and Bregman divergence

2.1.1 A general framework

Assume that the observed data $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$ are random samples from the population distribution of a p-dimensional predictor vector $\boldsymbol{X}$ and a scalar response Y where $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$. The number of variables p is allowed to grow with the sample size n, thus we denote it as p_n when needed. In this paper, we

are interested in predicting the response Y by its conditional expectation given $\boldsymbol{X}$,

$$m(\boldsymbol{X}) = E(Y \mid \boldsymbol{X}). \tag{2.1}$$

While it is possible that p_n can grow much faster than n, we assume that the true underlying model is sparse, which means that $m(\boldsymbol{X})$ functionally only depends on a small fraction of the predictors, denoted by

$$\mathcal{M}_n = \{1 \leq j \leq p_n : m(\boldsymbol{X}) \text{ functionally depends on } X_j\}$$

with cardinality $s_n = |\mathcal{M}_n|$. Without loss of generality, we can re-arrange the predictors such that $\mathcal{M}_n = \{1, \ldots, s_n\}$. Write $\boldsymbol{X} = (\boldsymbol{X}^{(\mathrm{I})T}, \boldsymbol{X}^{(\mathrm{II})T})^T$ where $\boldsymbol{X}^{(\mathrm{I})}$ collects all truly important predictors and $\boldsymbol{X}^{(\mathrm{II})}$ is just noise. Under this framework, our goal is to investigate the performance of a wide class of feature screening methods, by means of Bregman divergence that will be introduced in the next section.

2.1.2 Bregman divergence

Brègman (1967) introduced a device for constructing a bivariate function which can be used as a general loss function. For a given concave function q, define the Bregman divergence as

$$Q(\nu, \mu) = -q(\nu) + q(\mu) + (\nu - \mu)q'(\mu). \tag{2.2}$$

Conversely, for a given Q-loss, Zhang et al. (2009) provided necessary and sufficient conditions for Q being a BD, and further derived an explicit formula for solving the generating q-function. They also showed that the quadratic function, the Kullback-Leibler divergence (or the deviance loss) for the exponential family of probability functions, the (negative) quasi-likelihood function, and many margin-based loss functions, such as the misclassification loss, the hinge loss for the support vector machine (Vapnik (2000)), and the exponential loss used in AdaBoost (Hastie et al. (2001)), are all special cases of BD.

As an illustration, when we relax the distributional assumption on the response Y by only assuming $\mathrm{var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \sigma^2 V(m(\boldsymbol{x}))$ for a known continuous function $V(\cdot) > 0$, the quasi-likelihood function Q, given by the partial differential equation

$$\partial Q(Y, \mu)/\partial \mu = (Y - \mu)/V(\mu),$$

for a nuisance parameter $\sigma^2 > 0$, is usually used as an alternative of complete log-likelihood function. Zhang et al. (2009) verified that the (negative) quasi-likelihood function belongs to the BD and derived the generating q-function, given by

$$q(\mu) = \int_a^\mu \frac{s - \mu}{V(s)} ds, \tag{2.3}$$

where a is a finite constant such that the integral is well-defined.

2.2 Screening via componentwise regression minimum-BD estimation

Our proposed screening procedure is based on the componentwise regression minimum-BD estimators, defined as

$$\widehat{m}_j(\cdot) = \arg\min_{m} \frac{1}{n}\sum_{i=1}^{n} Q(Y_i, m(X_{ij})), \text{ for } j = 1, \ldots, p_n, \tag{2.4}$$

where the loss function Q is a BD as defined in (2.2) with a generating q-function. Furthermore, we restrict the $m_j(\cdot)$ to be of the form,

$$m_j(x) = F^{-1}(\alpha_j + x\beta_j), \tag{2.5}$$

where α_j and β_j are two parameters to be estimated, and F is a known link function for appropriate data type. Usually, an identity link $F(\mu) = \mu$ corresponds to the linear regression model for continuous responses; a logit link $F(\mu) = \log(\frac{\mu}{1-\mu})$ is utilized in the logistic regression for binary responses; a log link $F(\mu) = \log(\mu)$ is used in the Poisson regression of count responses.

The functional form in (2.5) is a linear approximation to the problem in (2.4) which appears somewhat restrictive, however our later theoretical results show that such class of functions is actually rich enough to detect the marginal importance of predictors for the screening purpose.

Thus, minimization problem in (2.4) is equivalent to estimate $(\widehat{\alpha}_j^{\mathrm{CR}}, \widehat{\beta}_j^{\mathrm{CR}})$,

for $j = 1, \ldots, p_n$, which are defined as

$$(\widehat{\alpha}_j^{\mathrm{CR}}, \widehat{\beta}_j^{\mathrm{CR}}) = \arg \min_{(\alpha_j, \beta_j) \in \mathbb{R}^2} \frac{1}{n} \sum_{i=1}^{n} Q(Y_i, F^{-1}(\alpha_j + X_{ij}\beta_j)). \tag{2.6}$$

We select the variables by choosing those whose componentwise coefficient estimators $|\widehat{\beta}_j^{\mathrm{CR}}|$ exceed a predefined threshold value $\gamma_n > 0$, i.e. variables X_j with indices j belonging to the set

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leq j \leq p_n : |\widehat{\beta}_j^{\mathrm{CR}}| \geq \gamma_n\}$$

will be selected; the remaining variables will be screened out.

The minimization problem (2.6) only involves a univariate predictor and an intercept. Thus fast and robust computation would be feasible even in NP-dimensional problems. When an appropriate γ_n is chosen, we can significantly reduce the dimension of the original parameter space to a much smaller one and thus make it more manageable. After the screening step, other variable selection methods, like those directly based on penalization, would be more feasible on survived variables.

In our screening procedure, the magnitude of the componentwise regression coefficient estimator, $\widehat{\beta}_j^{\mathrm{CR}}$, serves as a proxy for the importance of the corresponding feature X_j. Two questions arise naturally:

(I) how well the set $\widehat{\mathcal{M}}_{\gamma_n}$ preserves all important predictors, given that the estimators $\widehat{\beta}_j^{\mathrm{CR}}$ from the componentwise minimization problem

(2.6) only approximate the importance of predictors in the original model (2.1);

(II) how small the size of $\widehat{\mathcal{M}}_{\gamma_n}$ can be, given that $\widehat{\mathcal{M}}_{\gamma_n}$ should still include all truly important variables.

We will answer these two questions, by showing that the sure screening property holds under certain conditions, in the following sections.

2.2.1 Population version of componentwise regression minimum-BD estimator

Recall that the estimators in (2.6) are based on empirical minimization. To gain further insights, we define a population analogue of (2.6), denoted by

$$(\alpha_j^{\mathrm{CR}}, \beta_j^{\mathrm{CR}}) = \arg \min_{(\alpha_j, \beta_j) \in \mathbb{R}^2} E\{Q(Y, F^{-1}(\alpha_j + X_j \beta_j))\}, \tag{2.7}$$

where the expectation is taken with respect to the underlying joint distribution of $(\boldsymbol{X}, Y)$.

Note that the componentwise minimum-BD estimator $\widehat{\beta}_j^{\mathrm{CR}}$ will converge in probability to the population version β_j^{CR}. To guarantee the validity of the screening procedure, it is necessary that β_j^{CR} should at least preserve the significance of truly important predictors, i.e. those whose coefficient $\beta_{j:0} \neq 0$. Theorem 2.1 confirms that the significance of β_j^{CR} (i.e. $\beta_j^{\mathrm{CR}} \neq 0$) only depends on the correlation between Y and X_j.

Theorem 2.1. *Assume the conditions* A1–A4 *in the Appendix. For any* $j = 1, \ldots, p_n$*, it follows that* $\beta_j^{\mathrm{CR}} = 0$ *if and only if* $\mathrm{cov}(Y, X_j) = \mathrm{cov}(m(\boldsymbol{X}), X_j) = 0$*.*

Theorem 2.1 implies that if the response variable Y is correlated with an important variable X_j, whose coefficient $\beta_{j;0}$ is non-zero, then the componentwise regression coefficient β_j^{CR} will be non-zero. In contrast, for those unimportant variables X_j which are uncorrelated with Y, β_j^{CR} will be zero. Theorem 2.2 further indicates that the magnitude of β_j^{CR} is also closely related to the magnitude of correlation between X_j and Y.

Theorem 2.2. *Assume the conditions* A1–A5 *in the Appendix. For any positive sequences* $\mathcal{A}_n$ *and* $\mathcal{B}_n$*,*

(i) *if* $\min_{j=1,\ldots,s_n} |\mathrm{cov}(Y, X_j)| \geq \mathcal{A}_n$*, then there exists a positive constant* c_1 *such that*

$$\min_{j=1,\ldots,s_n} |\beta_j^{\mathrm{CR}}| \geq c_1 \mathcal{A}_n;$$

(ii) *if* $\max_{j=s_n+1,\ldots,p_n} |\mathrm{cov}(Y, X_j)| = O(\mathcal{B}_n)$*, then*

$$\max_{j=s_n+1,\ldots,p_n} |\beta_j^{\mathrm{CR}}| = O(\mathcal{B}_n).$$

The conditions used in Theorem 2.2 are typically regarded as mild and are often assumed in the literature (Zhang et al. (2009); Fan and Song (2010)). The assumptions A1 and A2 are some restriction on the

tail behavior of the population distribution. Assumptions A3 and A4 are about the convexity and smoothness of BD. The requirement of covariance between Y and X_j's for $j \in \mathcal{M}$ is to ensure that the minimal signal strength of important variables should not be too week and still identifiable.

If those conditions hold and $\mathcal{A}_n \succeq \mathcal{B}_n$, naturally we could utilize the gap between two groups of $\{|\beta_j^{\mathrm{CR}}|\}_{j=1}^{p_n}$ to identify the important variables, where $a_n \succeq b_n$ denotes that there exists a constant $c > 0$ such that $a_n \geq cb_n$ for all $n \geq 1$.

2.2.2 Sure screening property of componentwise BD regression

We start by giving the uniform convergence of componentwise regression minimum-BD estimator (2.6). To facilitate the deviation, we assume $E(X_j) = 0$ and $E(X_j^2) = 1$, for $j = 1, \ldots, p_n$ in the following results.

Theorem 2.3. *Assume the conditions* A1–A5 *in the Appendix. Then for any positive sequence $\mathcal{A}_n$ satisfying $\mathcal{A}_n\sqrt{n}/\log(n) \to \infty$, there exists some positive constant c_2 such that*

$$P\Big(\max_{1\leq j\leq p_n} |\widehat{\beta}_j^{\mathrm{CR}} - \beta_j^{\mathrm{CR}}| \geq \mathcal{A}_n\Big) \leq p_n\{\exp(-c_2\mathcal{A}_n^2 n) + nm_0\exp(-m_1\mathcal{A}_n^2 n)\}$$

where m_0 and m_1 are the constants given in Condition A2 *of the Appendix.*

Theorem 2.3 is an application of the exponential bound for the Quasi-

MLE in Fan and Song (2010) (Theorem 1). It guarantees that the empirical estimator $\widehat{\beta}_j^{\mathrm{CR}}$ will be close enough to the population version β_j^{CR} with large probability. With Theorem 2.3, we obtain the following Corollary 2.4 which demonstrates the sure screening property of componentwise BD regression.

Corollary 2.4. *Assume the conditions in Theorem* 2.3. *Set* $\gamma_n = c_1\mathcal{A}_n/2$, *where* c_1 *is the constant given in Theorem* 2.2.

(i) (*Sure screening property*) *If* $\min_{j=1,\ldots,s_n} |\mathrm{cov}(Y, X_j)| \geq \mathcal{A}_n$, *then*

$$P(\mathcal{M}_n \subseteq \widehat{\mathcal{M}}_{\gamma_n}) \geq 1 - s_n\{\exp(-c_2c_1^2\mathcal{A}_n^2 n/4) + nm_0\exp(-m_1c_1^2\mathcal{A}_n^2 n/4\}.$$

(ii) *If* $\max_{j=s_n+1,\ldots,p_n} |\mathrm{cov}(Y, X_j)| \leq \mathcal{B}_n = o(\mathcal{A}_n)$, *then*

$$P(\widehat{\mathcal{M}}_{\gamma_n} \subseteq \mathcal{M}_n) \geq 1 - (p_n - s_n)\{\exp(-c_2c_1^2\mathcal{A}_n^2 n/4) + nm_0\exp(-m_1c_1^2\mathcal{A}_n^2 n/4)\}.$$

Corollary 2.4 addressed the first question raised at the beginning of Section 2.2. It is easy to see that if we assume $\mathcal{A}_n = c_0 n^{-\alpha}$ where $0 < \alpha < 1/2$, then probability bounds in Corollary 2.4 is approaching one with the order $1 - O\{p_n \exp(-c_3 n^{1-2\alpha})\}$ for a positive constant c_3, which is the same rate obtained in Fan and Lv (2008) and Fan and Song (2010). This implies that the correct model will be selected with probability tending to one even under NP-dimensionality where p_n is permitted to be as large as $\log(p_n) = o(n^{1-2\alpha})$.

Remark 2.5. *In Corollary* 2.4*, the conclusion from part* (i) *and the conclusion from part* (ii) *hold separately. In some cases, the assumption about the covariance between the response and predictors in part* (ii) *of Corollary* 2.4 *needs to be relaxed, but the sure screening property given in part* (i) *of Corollary* 2.4 *will still hold. It means that even if we can not eliminate all unimportant predictors due to certain correlation between predictors, it is still guaranteed that we will not miss any truly important variables.*

2.2.3 Comparison with sure independence screening in GLM

While our motivation is from Fan and Song (2010)'s work on the generalized linear model (GLM), our results largely enhance the capability of marginal screening methods by extending it to a broader class of models with any BD as a loss function. In fact, proving the sure screening property under such general framework is by no means straightforward. The main challenge is that certain relationships in GLM are not applicable under the arbitrary choice of BD and link function in our setting. For example the following equality holds under GLM and its canonical link,

$$E\left(F^{-1}(\alpha_j^{\mathrm{CR}} + \beta_j^{\mathrm{CR}} X_j)X_j\right) = E\left(F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0)X_j\right), \qquad (2.8)$$

which is the equation (14) used in Fan and Song (2010) and a significant part of the proof for their Theorems 2, 3 and 5. However, (2.8) no longer

holds in the BD estimation with an arbitrary link. To overcome such technical challenge, we need to introduce a different way to express the componentwise regression minimum BD estimate and also impose some assumptions on the uniform bound of predictors. For details, see the Appendix.

Although the proposed screening procedure is based on certain linear form, the sure screening property of the proposed screening method actually does not require any particular parametric form of relationship between the response Y and the predictors $\boldsymbol{X}$. Instead, the sure screening property is mainly built on the assumption about minimal signal strength of important variables measured by marginal covariance. Our results also reveal that the different choices of BD will only affect constants c_1 and c_2 in the probability bounds in Corollary 2.4.

2.3 Two-step procedure with penalized-BD estimation

The results in Section 2.2 show that the screening procedure based on componentwise regression minimum-BD estimation works well in selecting the truly important variables. However, it may not be a good way to build a predictive model and provide estimates. In the absence of screening, Zhang et al. (2010) investigated the penalized-BD estimation and its oracle

property in a large-dimensional model with the following form,

$$m(\boldsymbol{X}) = E(Y \mid \boldsymbol{X}) = F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0) \tag{2.9}$$

where $\beta_{0;0}$ and $\boldsymbol{\beta}_0 = (\beta_{1;0}, \ldots, \beta_{p;0})^T$ are unknown true parameters, and F is a known link function. Particularly, their penalized-BD estimator using weighted L_1 penalties minimizes the criterion function,

$$\ell(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Q(Y_i, F^{-1}(\beta_0 + \boldsymbol{X}_i^T\boldsymbol{\beta})) + \lambda_n \sum_{j=1}^{p_n} w_j|\beta_j|, \tag{2.10}$$

where λ_n is the tuning parameter and $\{w_j\}_{j=1}^{p_n}$ are given weights for parameters $\{\beta_j\}_{j=1}^{p_n}$. In this section, we adopt the same setting.

Now we could answer the second question given at the beginning of Section 2.2 by Theorem 2.6 which is to control the number of the selected variables in the set $\widehat{\mathcal{M}}_{\gamma_n}$.

Theorem 2.6. *Assume conditions in Theorem* 2.3 *and condition* C *in the Appendix. Let* $\gamma_n = c_1\mathcal{A}_n/2$, *where* c_1 *is the constant given in Theorem* 2.2. *It holds that*

$$P(|\widehat{\mathcal{M}}_{\gamma_n}| \leq O(\mathcal{A}_n^{-2}\lambda_{\max}(\Sigma))) \geq 1 - p_n\{\exp(-c_2c_1^2\mathcal{A}_n^2 n/4) + nm_0\exp(-m_1c_1^2\mathcal{A}_n^2 n/4)\}$$

where $\Sigma = \mathrm{var}(\boldsymbol{X})$ *and* $\lambda_{\max}(\Sigma)$ *denotes the maximum eigenvalue of* Σ.

When $\mathcal{A}_n = O(n^{-\alpha}), \lambda_{\max} = O(n^{-\tau})$ and $2\alpha + \tau < 1$, Theorem 2.6

indicates that the number of selected covariates will not exceed the order $n^{2\alpha+\tau} = o\{n/\log(n)\}$. Therefore, we propose a two-step procedure from screening features to estimating coefficients of selected variables as follows. Since the cutoff value γ_n involves some unknown constant, in practice we propose another easier and more straightforward scheme that choose γ_n to be the p'_nth largest values of $|\widehat{\beta}_j^{\mathrm{CR}}|$, where $p'_n = \lfloor n/\log(n) \rfloor$ and $\lfloor \cdot \rfloor$ denotes the floor function. The choice of p'_n is large enough so that all truly important covariates will be selected, and also suitable for further estimation method in the second stage.

Step 1: Obtain the componentwise regression minimum-BD estimators $\widehat{\beta}_j^{\mathrm{CR}}$ in (2.6) and select sufficiently many covariates, corresponding to p'_n largest absolute values of $|\widehat{\beta}_j^{\mathrm{CR}}|$. Denote by $\widehat{\mathcal{M}}$ the set of indices of selected variables, where $|\widehat{\mathcal{M}}| = p'_n$.

Step 2: Set the coefficients of $(p_n - p'_n)$ variables not in $\widehat{\mathcal{M}}$ equal to zero. Use (2.10) to estimate the other parameters of those p'_n features selected in Step 1.

The idea of two-step procedures is widely used in the literature, for example the multi-stage method in Wasserman and Roeder (2009). After the first step, we can greatly reduce the dimensionality and at the same time, by the sure screening property, still preserve all truly important variables with high probability.

Proposition 2.7 also indicates that our two-step procedure enjoys the oracle property under certain conditions.

Proposition 2.7. *Assume the conditions in Theorem* 2.6 *and* $\mathcal{A}_n = O(1)$, $\mathcal{B}_n = o(\mathcal{A}_n)$. *Suppose* $s_n = o(n^{1/5})$, $s_n(p'_n - s_n) = o(n)$ *and* $\lambda_{\max}(\Sigma)/(p'_n\mathcal{A}_n^2) = o(1)$. *Let* $\lambda_n = \mathcal{A}_n/\sqrt{n}$ *and select weights in* (2.10) *by*

$$\widehat{w}_j = |\widehat{\beta}_j^{\mathrm{PCR}}|^{-1} \quad \textit{for } j \in \widehat{\mathcal{M}},$$

where $\widehat{\beta}_j^{\mathrm{PCR}}$ *is based on the penalized componentwise regression* BD *estimation,*

$$(\widehat{\alpha}_j^{\mathrm{PCR}}, \widehat{\beta}_j^{\mathrm{PCR}}) = \arg\min_{(\alpha_j,\beta_j)\in\mathbb{R}^2} \frac{1}{n}\sum_{i=1}^{n} Q(Y_i, F^{-1}(\alpha_j + X_{ij}\beta_j)) + \kappa_n|\beta_j| \quad (2.11)$$

with $\kappa_n = \mathcal{A}_n$. *Let* $\widetilde{\boldsymbol{X}}^{(\mathrm{I})} = (1, \boldsymbol{X}^{(\mathrm{I})T})^T$, $\widetilde{\boldsymbol{\beta}}_0 = (\beta_{0;0}, \boldsymbol{\beta}_0^T)^T$, $\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})} = (\beta_{0;0}, \boldsymbol{\beta}_0^{(\mathrm{I})T})^T$. *Then we have the following results for the two-step estimator.*

(i) *There exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\| = O_P((s_n/n)^{1/2})$.

(ii) *Any* $(n/s_n)^{1/2}$*-consistent local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_0 = (\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})}, \widehat{\boldsymbol{\beta}}^{(\mathrm{II})T})^T$ *satisfies* $P(\widehat{\boldsymbol{\beta}}^{(\mathrm{II})} = \mathbf{0}) \to 1$.

(iii) *Assume the Condition* D *in the Appendix. If* $\min_{j=1,\ldots,s_n} |\beta_j|/(s_n/n)^{1/2} \to \infty$, *then for any fixed integer* k *and any* $k \times (s_n + 1)$ *matrix* A_n *such that* $A_n A_n^T \to G$ *with* G *being a* $k \times k$ *nonnegative-definite symmetric matrix,*

we have that

$$\sqrt{n}A_n\mathbf{\Omega}_n^{-1/2}\Big\{\mathbf{H}_n(\widehat{\widetilde{\boldsymbol{\beta}}}^{(\mathrm{I})} - \widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) + \lambda_n\mathbf{W}_n\mathrm{sign}(\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})})\Big\} \xrightarrow{\mathcal{L}} N(\mathbf{0}, G),$$

where

$$\begin{aligned}
\mathbf{\Omega}_n &= E[\mathrm{var}(Y \mid \boldsymbol{X})\{q''(m(\boldsymbol{X}))/F'(m(\boldsymbol{X}))\}^2\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}],\\
\mathbf{H}_n &= -E[q''(m(\boldsymbol{X}))/\{F'(m(\boldsymbol{X}))\}^2\widetilde{\boldsymbol{X}}^{(\mathrm{I})}\widetilde{\boldsymbol{X}}^{(\mathrm{I})T}],\\
\mathbf{W}_n &= \mathrm{diag}(0, \widehat{w}_1, \ldots, \widehat{w}_{s_n}),
\end{aligned}$$

and $\mathrm{sign}(\widetilde{\boldsymbol{\beta}}_0^{(\mathrm{I})}) = (\mathrm{sign}(\beta_{0;0}), \mathrm{sign}(\beta_{1;0}), \ldots, \mathrm{sign}(\beta_{s_n;0}))^T$.

Note that the proposed componentwise BD regression weight selection method in Zhang et al. (2010) excludes an intercept term. In contrast, our current weight selection method (2.11) includes the intercept term. Nevertheless, the assumptions for the oracle property are still satisfied and thus our procedure would also enjoy the oracle property.

2.4 Simulation study

In this section, we assess the performance of both the screening step and the estimation step in the two-step procedure. Two different settings of (n, p_n) are used in our simulation,

$$(250, 250), \qquad (350, 2500),$$

which correspond to high and ultra-high dimensionality, respectively. The computations are performed using Matlab 2011b on Linux CONDOR computation environment.

2.4.1 Performance of feature screening

To evaluate the performance of our proposed screening method, referred to as "SIS-BD", we will measure the accuracy of the importance ranking of the predictors by the minimum model size (MMS) needed to include all truly important predictors. We also provide a coverage measure as the percentage of times that all non-zero coefficients are picked up when setting $p_n' = \lfloor n/\log(n) \rfloor$. We compare our method with a model-free screening procedure "SIRS" of Zhu et al. (2011). All the results are averaged over 1000 simulation runs.

Overdispersed Poisson responses

Here we consider the overdispersed Poisson model with the response Y generated according to $\mathrm{var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = 2\, m(\boldsymbol{x})$, where $m(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x})$. The link function used for count data is the log link. Thus,

$$\log\{m(\boldsymbol{x})\} = \beta_{0;0} + \boldsymbol{x}^T \boldsymbol{\beta}_0,$$

where $\beta_{0;0} = 1$ and $\boldsymbol{\beta}_0 = (2.5, 2, 2, 1.5, 0, \ldots, 0)^T$. The predictors are generated by $X_{ij} = \Phi(Z_{ij}) - 0.5$, $i = 1, \ldots, n$ and $j = 1, \ldots, p_n$, where Φ is the

standard normal distribution function, and

$$(Z_{i1}, \ldots, Z_{ip_n})^T \sim N(\mathbf{0}, \rho \mathbf{J}_{p_n} + (1-\rho)\mathbf{I}_{p_n}), \tag{2.12}$$

with $\mathbf{J}_d$ a $d \times d$ matrix in which all entries are ones and $\mathbf{I}_d$ a $d \times d$ identity matrix. Thus $(X_{i1}, \ldots, X_{ip_n})$ are marginally $\mathrm{Uniform}(-0.5, 0.5)$ random variables and correlated if $\rho \neq 0$. The type of BD we used here is

$$Q(Y, \mu) = \mu - Y \log(\mu) - Y + Y \log(Y)$$

which is generated by the q-function in (2.3) when $V(\mu) = \mu$, explicitly, $q(\mu) = \mu - a - \mu\{\log(\mu) - \log(a)\}$.

We give the mean, standard deviation along with a five number summary of MMS for two screening methods in different settings. The results are recorded in Table 2.1. Both $\mathrm{SIS\text{-}BD}$ and SIRS procedures work well in this nonlinear model and rank the truly important predictor at the very top of the list, thus the resulting MMS's are very close to the true model size. The ultra-high dimensionality makes the feature selection problem harder, however the value of MMS does not increase much, which supports our theoretical results in Section 2.2. When the dependence parameter increases from $\rho = 0.2$ to $\rho = 0.5$, the correlation between predictors becomes larger and the unimportant predictors can be easily confounded with important predictors. In such cases, MMS's become a little bigger, but most of them still remain at a very low level. In all settings, the coverage

percentage of all non-zero coefficients are pretty good. Two methods generally have comparable performance and the SIS-BD method is slightly better than the SIRS method in terms of the mean and maximum of MMS as well as coverage percentage.

Table 2.1: ***(Simulation results: overdispersed Poisson count responses)*** *Mean, standard deviation, and five number summary of the minimum model size, out of* 1000 *replications. Here* std*: standard deviation;* Q_1*: first quartile;* Q_3*: third quartile; Coverage: percentage of times that all non-zero coefficients are selected in* $\widehat{\mathcal{M}}$*.*

ρ	Method	Mean (std)	(Min	Q_1	Q_2	Q_3	Max)	Coverage
$(n, p_n) = (250, 250)$								
0.2	SIS-BD	4.68 (3.4)	(4	4	4	4	71)	99.8%
	SIRS	4.80 (4.3)	(4	4	4	4	73)	99.6%
0.5	SIS-BD	7.41 (8.6)	(4	4	4	6	89)	98.7%
	SIRS	8.19 (11.0)	(4	4	4	7	121)	98.0%
$(n, p_n) = (350, 2500)$								
0.2	SIS-BD	5.42 (10.0)	(4	4	4	4	217)	99.6%
	SIRS	6.04 (15.6)	(4	4	4	4	355)	99.3%
0.5	SIS-BD	14.98 (33.8)	(4	4	5	9	390)	94.6%
	SIRS	18.11 (47.2)	(4	4	5	11	622)	93.2%

Bernoulli binary responses

We further investigate the logistic regression model with a binary response Y, which is generated as a Bernoulli random variable with

$$P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \exp\{-(\beta_{0;0} + \boldsymbol{x}^T \boldsymbol{\beta}_0)\}},$$

where $\beta_{0;0} = 3$ and $\boldsymbol{\beta}_0 = (3, 2, -3, -2, 0, \ldots, 0)^T$. The predictors are also generated by independent Bernoulli random variables with

$$P(X_{ij} = 1) = r, \qquad i = 1, \ldots, n,\ j = 1, \ldots, p_n. \tag{2.13}$$

The logit link function is used.

Results from two types of BD will be presented in this section, the Bernoulli deviance (DEV) loss,

$$Q(Y, \mu) = -2\{Y \log(\mu) + (1 - Y) \log(1 - \mu)\}$$

which corresponds to $q(\mu) = -2\{\mu \log(\mu) + (1 - \mu) \log(1 - \mu)\}$, and the Exponential (EXP) loss,

$$Q(Y, \mu) = \exp[-(Y - .5) \log\{\mu/(1 - \mu)\}]$$

which corresponds to $q(\mu) = 2\{\mu(1 - \mu)\}^{1/2}$.

In this particular setting, all the response and predictors are binary which provide very limited information from either part. Table 2.2 reports the mean and the five number summary of MMS from our simulation. Two choices of q-functions give similar results in all cases. The MMS increases as the dimensionality grows, but at a much smaller rate so that the identification of important predictors and further parameter estimation would be still possible. When $r = 0.2$, the signal from the data is more

scarce and it is more difficult to identify the truly important variables. The SIS-BD outperforms the SIRS in this setting. Even under the most stringent situation when p_n far more exceeds n and $r = 0.2$, the median MMS of SIS-BD is still quite small, compared with the total number of predictors.

Table 2.2: ***(Simulation results: parameter estimation for Bernoulli binary responses)*** *Mean, standard deviation, and five number summary of the minimum model size, out of* 1000 *replications.*

r	Method (Loss)	Mean (std)	(Min	Q_1	Q_2	Q_3	Max)	Coverage
$(n, p_n) = (250, 250)$								
0.2	SIS-BD (DEV)	26.78 (35.5)	(4	7	13	28	250)	83.5%
	SIS-BD (EXP)	28.35 (38.7)	(4	8	13	28	250)	82.3%
	SIRS	29.97 (36.8)	(4	8	15	34	250)	80.8%
0.5	SIS-BD (DEV)	9.15 (11.9)	(4	4	5	8	117)	97.4%
	SIS-BD (EXP)	9.15 (11.8)	(4	4	5	8	114)	97.4%
	SIRS	9.17 (12.0)	(4	4	5	8	116)	97.6%
$(n, p_n) = (350, 2500)$								
0.2	SIS-BD (DEV)	98.80 (208.4)	(4	14	33	87	2267)	66.4%
	SIS-BD (EXP)	110.01 (241.2)	(4	14	33	94	2267)	65.7%
	SIRS	114.60 (219.2)	(4	16	42	115	2264)	60.3%
0.5	SIS-BD (DEV)	20.97 (85.8)	(4	4	5	11	1899)	93.7%
	SIS-BD (EXP)	20.95 (85.6)	(4	4	5	11	1886)	93.8%
	SIRS	20.99 (86.3)	(4	4	5	11	1906)	93.7%

2.4.2 Performance of parameter estimation

Here we will compare the performance of the two-step procedure with those variable selection methods using penalization which are directly

applied to all variables. Namely, they minimize a criterion function similar to (2.10), except that the choice of the penalty can be the following:

(I) (SCAD) the SCAD penalty, with an accompanying parameter $a = 3.7$ (Fan (1997));

(II) (MCP) the MCP penalty, with an accompanying parameter $a = 3.7$ (Zhang (2010));

(III) (L_1) the L_1 penalty (Tibshirani (1996));

(IV) (WL1PCR) the weighted-L_1 penalty with weights selected by (2.11).

Let $p_n' = \lfloor n/\log(n) \rfloor$ in the first step. For brevity, the two-step procedures are referred to as S-SCAD, S-MCP, S-L_1, and S-WL1PCR, respectively. We also include the oracle estimator as an optimal but unpractical solution, which is obtained by the unpenalized BD estimation with only those covariates whose coefficients are indeed non-zero.

The tuning constants λ_n and κ_n are selected via a grid search separately to minimize the Akaike's information criterion (AIC). All the results are averaged over 100 simulation runs.

Overdispersed Poisson responses

The setting is similar to that in Section 2.4.1 except that we fix the dependence parameter between predictors at $\rho = 0.2$. To compare the accuracy of the estimated parameters by different methods, the average of bias

$\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\|$ across those 100 training sets is calculated. The test error (TE) is obtained from an independently generated test set $\{(\boldsymbol{x}_\ell, y_\ell)\}_{\ell=1}^{L=10000}$ by $\sum_{\ell=1}^{L} Q(y_\ell, \widehat{m}(\boldsymbol{x}_\ell))/L$. We also provide the model selection performance via C-Z which is the total number of coefficients which are correctly estimated to be zero when the true coefficients are zero, and C-NZ which is the total number of coefficients which are correctly estimated to be non-zero when the true coefficients are non-zero. Finally, we record the average running time of each method under different settings. The ultra-high dimensional problem usually imposes a big challenge in computations as well as model selection and estimation, so considerably faster speed can be viewed as an advantage. Table 2.3 summarizes the simulation results.

We can see that all methods perform reasonably well, while all two-step procedures are slightly better than the counterparts which directly apply the estimation step. The last method using a screening step followed by the weighted-L_1 penalized estimation with weights selected by the PCR method is the best with a small margin, which supports our theoretical results in Section 2.3. While the gain of the accuracy from the screening step does not seem to be very dramatic, we notice that the speed of the two-step procedures is much faster, where the screening step can reduce the computation time by a factor of 5 to 20. This indicates that the screening step can indeed filter out most irrelevant predictors without sacrificing the accuracy, so that we can make better use of the computational resources.

Table 2.3: ***(Simulation results: overdispersed Poisson count responses)*** *Predictors are marginally* $\mathrm{Uniform}(-0.5, 0.5)$ *with dependence parameter* $\rho = 0.2$ *in* (2.12). *Results are averaged over* 100 *replications. Here* TE *is the test error obtained from an independent test set; time is the average running time in seconds.*

(n, p_n)	Method	Bias	TE	#C-Z (std)	#C-NZ (std)	Time (sec)
(250, 250)	SCAD	0.72	2.29	231.3(5.4)	4.0 (0.0)	5.53
	S-SCAD	0.55	2.24	237.2(3.1)	4.0 (0.0)	0.39
	MCP	0.60	2.26	235.8(5.8)	4.0 (0.0)	5.18
	S-MCP	0.51	2.24	239.2(2.8)	4.0 (0.0)	0.38
	L_1	0.81	2.31	232.4(5.2)	4.0 (0.0)	1.96
	S-L_1	0.70	2.28	235.3(3.6)	4.0 (0.0)	0.38
	WL1PCR	0.53	2.24	241.2(3.4)	4.0 (0.0)	13.43
	S-WL1PCR	0.53	2.24	241.4(3.0)	4.0 (0.0)	2.24
	Oracle	0.28	2.19	246	4	0.02
(350, 2500)	SCAD	0.85	2.34	2470.7(8.9)	4.0 (0.0)	33.20
	S-SCAD	0.53	2.26	2484.4(3.5)	4.0 (0.1)	2.36
	MCP	0.67	2.29	2479.4(7.4)	4.0 (0.0)	32.28
	S-MCP	0.51	2.25	2486.7(3.7)	4.0 (0.1)	2.35
	L_1	0.92	2.36	2473.3(8.0)	4.0 (0.0)	13.30
	S-L_1	0.74	2.31	2481.1(3.6)	4.0 (0.1)	2.35
	WL1PCR	0.56	2.26	2489.8(4.1)	4.0 (0.0)	101.33
	S-WL1PCR	0.57	2.27	2489.4(4.2)	4.0 (0.1)	7.15
	Oracle	0.24	2.20	2496	4	0.13

Bernoulli binary responses

The setting is similar to that in Section 2.4.1 except that we only present the result with $r = 0.5$. For this type of classification problem, we calculate the misclassification rate (MR) for an independent test set instead of the test error. Other metrics are similar to those in Section 2.4.2. The results are

summarized in Tables 2.4. We can see that two-step procedures perform slightly better than the counterparts without applying the screening step, and the misclassification rates are almost the same as the optimal rate obtained by the oracle estimator which in not practical in reality. Again, in our comparison, the screening step can make the computation much faster than those methods that directly work with all possible predictors. In summary, the two-step procedures can outperform raw estimation method and have much faster speed.

2.5 Real data application

In this section, we apply the methods considered in Section 2.4.2 to two real data sets to illustrate the practical usefulness of the screening procedures. The tuning constants λ_n and κ_n in the second step is selected by the Akaike's information criterion (AIC).

2.5.1 Leukemia data

The Leukemia data set, available at `http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`, is widely used as a benchmark in the literature of high-dimensional classification problems. It was previously analyzed in Golub et al. (1999) to select features that can better classify the AML (acute mylogenous leukemia) and ALL (acute lymphocytic leukemia) cases. The data set consists of $p = 7129$ genes and $n = 72$ samples from two classes:

Table 2.4: ***(Simulation results: Bernoulli binary responses)*** *Predictors are independent Bernoulli random variables with* $r = 0.5$ *in* (2.13)*. Here* MR *is the misclassification rate on an independent test set. Results are averaged over* 100 *replications.*

(n, p_n)	Loss	Method	Bias	MR	#C-Z (std)	#C-NZ (std)	Time (sec)
(250, 250)	DEV	SCAD	1.94	0.13	245.5 (0.8)	3.6 (0.6)	0.25
		S-SCAD	1.94	0.13	245.5 (0.8)	3.7 (0.6)	0.18
		MCP	1.94	0.13	245.6 (0.7)	3.6 (0.6)	0.24
		S-MCP	1.93	0.13	245.6 (0.7)	3.6 (0.6)	0.18
		L_1	3.76	0.16	244.6 (1.4)	3.7 (0.6)	0.45
		S-L_1	3.75	0.16	244.6 (1.5)	3.7 (0.6)	0.25
		WL1PCR	2.56	0.13	244.7 (1.6)	3.7 (0.6)	2.46
		S-WL1PCR	2.56	0.13	244.7 (1.6)	3.7 (0.6)	0.74
		Oracle	1.31	0.13	246	4	0.02
	EXP	SCAD	2.35	0.14	245.7 (0.8)	3.4 (0.7)	0.20
		S-SCAD	2.30	0.13	245.6 (0.9)	3.5 (0.7)	0.16
		MCP	2.34	0.15	245.8 (0.6)	3.2 (0.9)	0.19
		S-MCP	2.38	0.15	245.7 (0.6)	3.3 (0.9)	0.16
		L_1	3.38	0.16	244.7 (1.4)	3.7 (0.6)	0.32
		S-L_1	3.37	0.15	244.6 (1.4)	3.7 (0.6)	0.20
		WL1PCR	2.23	0.13	244.6 (1.7)	3.6 (0.6)	2.43
		S-WL1PCR	2.23	0.13	244.6 (1.7)	3.6 (0.6)	0.75
		Oracle	1.67	0.13	246	4	0.02
(350, 2500)	DEV	SCAD	1.70	0.13	2495.5 (1.0)	3.7 (0.5)	2.87
		S-SCAD	1.69	0.13	2495.4 (1.1)	3.8 (0.5)	1.45
		MCP	1.75	0.13	2495.6 (0.7)	3.6 (0.7)	2.69
		S-MCP	1.76	0.13	2495.5 (0.7)	3.7 (0.7)	1.44
		L_1	3.88	0.17	2494.1 (1.8)	3.7 (0.6)	4.85
		S-L_1	3.87	0.17	2494.0 (1.9)	3.7 (0.6)	1.56
		WL1PCR	2.67	0.14	2493.9 (2.3)	3.8 (0.5)	27.56
		S-WL1PCR	2.67	0.14	2493.9 (2.3)	3.8 (0.5)	2.92
		Oracle	1.03	0.13	2496	4	0.21
	EXP	SCAD	2.02	0.13	2495.5 (0.9)	3.5 (0.6)	2.32
		S-SCAD	2.01	0.13	2495.5 (0.8)	3.6 (0.6)	1.33
		MCP	2.06	0.15	2495.7 (0.6)	3.4 (0.9)	2.16
		S-MCP	2.07	0.15	2495.7 (0.6)	3.5 (0.9)	1.32
		L_1	3.54	0.17	2493.9 (1.9)	3.7 (0.5)	3.44
		S-L_1	3.51	0.17	2493.7 (2.4)	3.7 (0.5)	1.39
		WL1PCR	2.32	0.13	2494.3 (2.0)	3.7 (0.6)	25.81
		S-WL1PCR	2.31	0.13	2494.3 (2.1)	3.7 (0.6)	2.77
		Oracle	1.30	0.13	2496	4	0.21

47 in class ALL and 25 in class AML. Among those 72 samples, 38 (with 27 in class ALL and 11 in class AML) of them were set as the training sample and 34 (with 20 in class ALL and 14 in class AML) of them were set to be the test sample.

We provide the number of misclassified cases in both training and test samples, as well as the number of selected variables and computation time. The results are given in Table 2.5. It shows that the screening step can help to select fewer predictors and improve the classification performance. When using the deviance loss, only 2-3 errors were made in the test set by selecting less than 5 genes for most of the two-step procedures. The number of errors and selected variables are significantly greater for those directly using the penalized estimation step. The computation time required for the two-step procedures is just a small fraction of the time required by their original version. Similar observations can be made from using the exponential loss.

2.5.2 Colon data

The classification of colon cancer is discussed in Alon et al. (1999) and the data set can be downloaded from `http://genomics-pubs.princeton.edu/oncology/`. It consists of $p = 2000$ genes and $n = 62$ samples, in which 22 samples are from normal colon tissues and 40 samples are from tumor tissues.

In our analysis, the data set is randomly split into two parts, with 45

Table 2.5: ***(Real data: Leukemia)*** *Number of misclassified cases among* 38 *training samples and* 34 *test samples, number of selected variables among all* 7129 *predictors and the computation time in seconds.*

Loss	Method	# Error (training)	# Error (test)	# Selected	Time (sec)
DEV	SCAD	0	7	3	26.4
	S-SCAD	0	3	2	5.0
	MCP	0	7	3	22.2
	S-MCP	0	3	2	4.9
	L_1	0	4	15	54.9
	S-L_1	0	2	7	5.8
	WL1PCR	0	3	6	441.5
	S-WL1PCR	0	3	5	11.8
EXP	SCAD	0	7	3	28.0
	S-SCAD	0	3	2	6.3
	MCP	0	7	3	29.3
	S-MCP	0	3	2	6.3
	L_1	0	7	11	56.8
	S-L_1	0	3	5	6.6
	WL1PCR	0	4	3	452.6
	S-WL1PCR	0	3	3	13.8

samples as training samples and the rest 17 as test samples. We repeat the random split 100 times and calculate the average number of misclassified cases in both sets. The results are summarized in Table 2.6. It again provides evidence that the two-step procedures are capable of identifying those important variables and obtaining a good estimation and prediction while considerably less computational resources are needed.

Table 2.6: ***(Real data: Colon)*** *Average number of misclassified cases among* 45 *training samples and* 17 *test samples, average number of selected variables among all* 2000 *predictors and the average computation time in seconds, over* 100 *replications.*

Loss	Method	# Error (training)	# Error (test)	# Selected	Time (sec)
DEV	SCAD	0.1	3.9	5.3	9.3
	S-SCAD	0.2	3.6	6.2	2.4
	MCP	0.1	4.0	5.4	8.6
	S-MCP	0.2	3.8	6.1	2.5
	L_1	1.0	3.8	16.1	19.2
	S-L_1	0.6	3.2	11.1	3.2
	WL1PCR	0.5	3.3	10.3	111.5
	S-WL1PCR	1.0	3.3	8.5	13.0
EXP	SCAD	1.0	4.0	5.0	9.7
	S-SCAD	0.6	3.8	5.7	3.2
	MCP	1.1	4.1	5.1	12.8
	S-MCP	0.5	3.8	5.8	3.2
	L_1	0.7	3.7	15.6	19.4
	S-L_1	0.7	3.3	10.5	3.5
	WL1PCR	0.4	3.2	9.6	117.3
	S-WL1PCR	1.0	3.4	8.4	13.6

Chapter 3

Analysis of neuronal functional connectivity using structured regularization in generalized linear models

3.1 Background

The capacity to simultaneously record spike trains of many neurons from awake, behaving subjects, has surpassed our ability to describe putative neural codes distributed across populations of neurons. Identifying correlation structure of a neuron ensemble beyond pairwise measures is critical for understanding how information is transferred within such a neural

population. However, the spike train data pose significant challenges to statistical researchers due to not only their complexity but also high dimensionality. First, neural spike trains of this type are generally non-stationary and exhibit a great amount of variability among repeated trials. Second, there are evidences showing that the neural spike activity patterns among neurons may be changed under different experimental conditions. Last but not the least, the high dimensional nature of the spike train data from these experiments makes the analysis computationally extremely intensive.

Several statistical methods have been proposed to identify functional connections within an ensemble of neurons. Early nonparametric methods, for example cross-correlogram (Perkel et al. (1967)) and joint peri-stimulus time histogram (JPSTH) (Gerstein and Perkel (1969)) provide a lot of insights and are still commonly used for analyzing the interactions between neurons. However, these methods also have serious drawbacks: they focus on the pairwise relationship, but ignore possible connections with other neurons in the ensemble or the influences from external stimulus. Furthermore, correlation-based analysis is limited to linear aspects of connectivity, which might be inadequate for neuronal spike activities. Recently, the model-based approaches draw a great deal of attention from researchers, in which a specific point process model is assumed, including the inhomogeneous Poisson process or more generally Cox point process (Cox and Isham (1980)).

In this thesis, we will study the spike train data by adopting the generalized linear model (GLM) framework from Brillinger (1992), which has shown the superiority in modeling the spiking activities of neurons (Truccolo et al. (2005)). However, we find that certain structural information of the parameter space in the GLMs has not been fully utilized in the previous works. Therefore, we propose a structured regularization method for the spike train data to better investigate the functional connectivity within a neuronal network.

In the literature, the generalized linear model (McCullagh and Nelder (1989)) has been widely used to analyze the spike point processes where the firing rate of a single neuron is modeled as a function of spiking history of concurrently recorded ensemble neurons (Brillinger (1992)). This GLM approach has been successfully applied to spike train data from many different types of experiments and becomes a very powerful and efficient tool of neural encoding and decoding (Truccolo et al. (2005); Pillow et al. (2008)). One significant advantage of the GLM approach is that the connections between every pair of neurons can be modeled as parameters of the GLM and then analyzed simultaneously under the general framework (Okatan et al. (2005)). In this setting, the influence from one neuron to another one is described by a temporal coupling kernel function with a group of parameters whose sign and magnitude indicate how neurons interact with each other. A more comprehensive review can be found in Truccolo (2010). We will mainly focus on the GLM

framework in the rest of our paper, nevertheless it is worth noting that other statistical models such as the Cox model (Berry et al. (2012)) and the dynamical Bayesian network (Eldawlatly et al. (2010)), are also useful tools for inferring the functional connections among ensemble neurons.

The maximum likelihood estimation (MLE) is generally applicable to obtain the estimate of model parameters in GLM, but it is also known to be vulnerable to over-fitting problems which frequently arise from high dimensional data. For a typical real neural spike train data set, the spiking rate is often very low and the design matrix could be very sparse, i.e. more than 90% of entries are zeros. In such cases, the MLE may not be reliable. Therefore, some regularization techniques can be applied to obtain well behaved solutions to over-parameterized estimation problems (Bickel and Li (2006)). On the other hand, a sparse solution is desirable since the neurons may not be connected to all other neurons in a large population and the sparsity could improve the interpretability. Since standard MLE can not automatically produce a sparse solution and the estimated parameters by MLE are almost surely nonzero, one extra step is usually needed to select the significant parameters which correspond to those truly functional connections among neurons, for example multiple hypothesis testing with control of family wise error rate or false discovery rate (Gerhard et al. (2011); Kim et al. (2011); Berry et al. (2012)).

For above reasons, a class of non-smooth penalties are introduced to achieve the goal of both obtaining a stable estimate and encouraging a

sparse solution, including the L_1 penalty or Lasso (Tibshirani (1996)), the SCAD (Fan (1997)), among many others. The L_1 regularized GLM has been studied by many researchers in neuroscience due to its simplicity, easy implementation and good performance (Kelly et al. (2010); Chen et al. (2011); Mishchenko et al. (2011); Zhao et al. (2012)). Furthermore, regularization or penalization can also be interpreted as imposing a prior on the parameters from a Bayesian perspective, and the regularized log-likelihood can be viewed as the log posterior density of the parameters (Stevenson et al. (2009)). Thus, the regularized maximum likelihood estimation is equivalent to the maximum a posteriori estimate from the Bayesian point of view.

However, none of these works has considered the structural information of the parameter space, in which all parameters related to the interaction between one specific pair of neurons naturally form a group and the whole group could be better estimated together. Instead of treating all parameters individually, we wish to be able to make decisions jointly with the grouping information and promote the structured sparsity, i.e. select entire group of parameters, or make the inclusion of some parameters depend on the inclusion of other parameters. Some recent works in statistics literature have explored this direction by the use of group or hierarchical norm penalties and illustrated some promising properties, particularly the Group Lasso (Yuan and Lin (2006); Meier et al. (2008)) and the Sparse Group Lasso (Simon et al. (2013); Chatterjee et al. (2012)).

among many others (Wang et al. (2009); Liu and Ye (2010)). In this paper, we propose a new structured regularization method for the spike train data to incorporate the prior structural information into the modeling and guide the selection of parameters according to the underlying structure of the parameter space. The main contributions of our work are as follows.

- We introduce the group-structured regularized GLM, in the context of multiple spike train data and show that our method performs better on simulated spike train data.
- The proposed method can be shown to be estimation consistent and asymptotically select the correct model with the smallest number of covariates.
- A fast and efficient algorithm called Accelerated Full Gradient Update (AFGU) is developed to handle the complex structured penalty in the generalized linear models for large sparse data sets.
- The application of our method to a large dataset recorded from the prelimbic region of the frontal cortex (plPFC) of adult male rats when performing a T-Maze based delayed-alternation task of working memory shows some insightful results.

3.2 Methodology

3.2.1 Background on CI-GLM

Consider an ensemble of C neurons. The spike train data are usually collected as multivariate (C-dimensional) point processes, denoted by

$$0 < u_{c,1} < u_{c,2} < \cdots < u_{c,J_c} \leq T, \quad \text{for } c = 1, \ldots, C,$$

where J_c is the total number of observed spikes by neuron c during the experiment, $\{u_{c,i}\}_{i=1}^{J_c}$ are the timestamps when neuron c spikes, and T is the time length of the experiment trial. Here we will take a discrete-time approach and split the whole experiment time period $(0, T]$ into n equally-spaced time bins $\tau_k \equiv (t_{k-1}, t_k]$, $k = 1, \ldots, n$, of length $\delta = T/n$, i.e. $t_k = k\delta$. The number of spikes fired by neuron c within the kth time bin is denoted by $N_c(\tau_k)$ and the spiking history of all neurons up to the time point t_k is denoted by $N_{1:C}(\tau_{1:k})$.

We model the distribution of $N_c(\tau_k)$ by a conditional intensity-generalized linear model (CI-GLM) in which the conditional intensity function of neuron c, $\lambda_c(\tau_k \mid \cdot) = E\{N_c(\tau_k) \mid \cdot\}$, takes into account several inputs at that

time and has the following form,

$$\log\Big\{\lambda_c(\tau_k \mid N_{1:C}(\tau_{1:(k-1)}), \boldsymbol{X}(\tau_k))\Big\} = \gamma_{c,0} + \sum_{p=1}^{P} \gamma_{c,p} N_c(\tau_{k-p}) + \sum_{\substack{i=1 \\ i\neq c}}^{C} \Big\{ \sum_{q=1}^{Q} \gamma_{c,i,q} N_i(\tau_{k-q}) \Big\} + \boldsymbol{X}(\tau_k)^T \boldsymbol{\beta}_c, \tag{3.1}$$

where the intercept $\gamma_{c,0}$ represents the baseline firing rate of neuron c; $\sum_{p=1}^{P} \gamma_{c,p} N_c(\tau_{k-p})$ is the effect of the intrinsic spiking history of neuron c, up to P bins into the past with $\gamma_{c,p}$ representing the autoregressive coefficients at lag p; $\sum_{q=1}^{Q} \gamma_{c,i,q} N_i(\tau_{k-q})$ is the coupling kernel function describing the influence of neuron i on neuron c. Such influence is computed base on a spiking history window, up to Q bins into the past where $\gamma_{c,i,q}$ represents the coupling coefficient at lag q. For example, if $\gamma_{c,i,q}$ is positive, neuron i would functionally excite neuron c after lag of q bins. The larger the value of $\gamma_{c,i,q}$, the stronger the excitatory drive; $\boldsymbol{X}(\tau_k) = (X_{\tau_k,1}, \ldots, X_{\tau_k,M})^T$ is a M-dimensional vector representing some extrinsic covariates measured during the kth time bin, τ_k, while $\boldsymbol{\beta}_c = (\beta_{c,1}, \ldots, \beta_{c,M})^T$ represents the vector of corresponding coefficients. For example, $\boldsymbol{X}(\tau_k)$ includes several status indicators related to the performance of the animals in a behavioral task for electrophysiological recordings (described in Section 3.5). Now we are able to simultaneously handle several factors including the spike history, neuronal network structure, and other covariates. Note that we

distinguish the influence of its own spiking history from the influences coming from other neurons, and allow the length of the history windows (P and Q) to be different for greater flexibility.

Under this CI-GLM framework, the likelihood of the observed spike trains is then given by

$$\prod_{c=1}^{C} \mathrm{P}(N_c(\tau_1), \ldots, N_c(\tau_n)) = \prod_{c=1}^{C} \prod_{k=1}^{n} \frac{\{\lambda_c(\tau_k \mid \cdot)\}^{N_c(\tau_k)} \exp\{-\lambda_c(\tau_k \mid \cdot)\}}{N_c(\tau_k)!}.$$

The conventional MLE can be obtained by minimizing the negative log-likelihood,

$$\sum_{c=1}^{C} \ell_c(\widetilde{\boldsymbol{\theta}}_c), \tag{3.2}$$

where

$$\ell_c(\widetilde{\boldsymbol{\theta}}_c) = -\frac{1}{n} \sum_{k=1}^{n} [N_c(\tau_k) \log\{\lambda_c(\tau_k \mid \cdot)\} - \lambda_c(\tau_k \mid \cdot) - \log\{N_c(\tau_k)!\}] \tag{3.3}$$

is the negative log-likelihood of a single neuron c and

$$\widetilde{\boldsymbol{\theta}}_c = (\gamma_{c,0}, \{\gamma_{c,p}\}_{p=1,\ldots,P}, \{\gamma_{c,i,q}\}_{i=1,\ldots,C,\, i\neq c;\, q=1,\ldots,Q}, \{\beta_{c,m}\}_{m=1,\ldots,M})^T$$

is the vector collecting all parameters for neuron c, including the baseline firing rate $\gamma_{c,0}$ as the intercept. We shall see that the minimization problem in (3.2) is separable for each neuron, so we can equivalently solve it by minimizing the negative log-likelihood neuron by neuron. Therefore, we

will analyze each neuron individually in the rest of the paper.

3.2.2 Proposed structured regularization in GLM

Regularization or penalization is a very useful technique aiming at obtaining well behaved solutions for sparse signals in some large dimensional problems. Given the large dimensionality and large portion of zero entries in spike train data, the un-regularized estimation usually suffers from the over-fitting problem and results in unstable solutions. Furthermore, the coefficients $\gamma_{c,i,q}$ of coupling kernel function representing the functional connection between neuron cells are assumed to be sparse. However, un-regularized estimator, i.e. MLE, fails to provide a sparse estimation of those parameters.

Here we wish to propose an appropriate regularization for the model in (3.1). Among numerous variable selection methods that are recently developed based on regularization, Lasso using the L_1 penalty (Tibshirani (1996)) is one of the most popular methods due to its simplicity and good performance. The general form of Lasso is defined as follows,

$$\min_{\widetilde{\boldsymbol{\theta}}} \left\{ \ell(\widetilde{\boldsymbol{\theta}}) + \eta \|\boldsymbol{\theta}\|_1 \right\},$$

where $\ell(\widetilde{\boldsymbol{\theta}})$ is usually the negative log-likelihood function or other appropriate loss function, $\boldsymbol{\theta}$ is the vector of all parameters excluding the intercept, η is a tuning parameter, and $\|\cdot\|_1$ denotes the L_1 norm of a vector, i.e. the

sum of absolute values of all coordinates. Such L_1 regularized GLM has been studied by several researchers recently, including Kelly et al. (2010) and Zhao et al. (2012), among many others and it shows a lot of advantages and provides significant improvement.

However, Lasso by design can only treat all parameters individually, and it is not able to incorporate the structural information inherent in a particular data set. As mentioned earlier, for any two fixed neurons, neuron c and neuron i, the coefficients $\{\gamma_{c,i,q} : q = 1, \ldots, Q\}$ together describe one coupling kernel function, which induces a natural "grouping" of the coefficients in our problem. It arises the need to design other regularization method that can promote structured sparse solution. Recently, Simon et al. (2013) proposed the Sparse Group Lasso (SGL) which can perform the selection at both the group level and the individual level. The general form of SGL is

$$\min_{\widetilde{\boldsymbol{\theta}}} \left\{ \ell(\widetilde{\boldsymbol{\theta}}) + (1-\alpha)\eta \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\theta}^{(g)}\|_2 + \alpha\eta \|\boldsymbol{\theta}\|_1 \right\},$$

where G is the number of groups, p_g is the size of the gth group, $\boldsymbol{\theta}^{(g)}$ is the vector of coefficients in the gth group so that $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)T}, \ldots, \boldsymbol{\theta}^{(G)T})^T$, and $\|\cdot\|_2$ denotes the L_2 norm of a vector. Here, both $\alpha \in [0, 1]$ and $\eta > 0$ are tuning parameters. The SGL is a convex combination of the Lasso and Group Lasso (Yuan and Lin (2006)) controlled by α. Smaller values of α promote the sparsity at group level, while larger values of α encourage

individual sparsity.

Motivated by the structured penalties mentioned above, we now formulate the regularized GLM problem in our setting as follows. There are two types of groups, the coefficients of the autoregressive kernel functions

$$\boldsymbol{\gamma}^{(c)} = (\gamma_{c,1}, \ldots, \gamma_{c,P})^T$$

and those of the coupling kernel functions

$$\boldsymbol{\gamma}^{(c,i)} = (\gamma_{c,i,1}, \ldots, \gamma_{c,i,Q})^T.$$

In the meantime, the coefficients for the covariates $\boldsymbol{X}(\cdot)$,

$$\boldsymbol{\beta}_c = (\beta_{c,1}, \ldots, \beta_{c,M})^T,$$

are not grouped together and we treat them as stand-alone variables. Thus our solution minimizes

$$L_c(\widetilde{\boldsymbol{\theta}}_c) = \ell_c(\widetilde{\boldsymbol{\theta}}_c) + \mathcal{P}_{(\eta_c,\alpha_c)}(\widetilde{\boldsymbol{\theta}}_c), \tag{3.4}$$

where

$$\mathcal{P}_{(\eta_c,\alpha_c)}(\widetilde{\boldsymbol{\theta}}_c) \;=\; (1-\alpha_c)\eta_c \left\{ \sqrt{P}\|\boldsymbol{\gamma}^{(c)}\|_2 + \sqrt{Q} \sum_{i=1,\ldots,C; i\neq c} \|\boldsymbol{\gamma}^{(c,i)}\|_2 \right\}$$

$$+\alpha_c\eta_c\left\{\|\boldsymbol{\gamma}^{(c)}\|_1+\sum_{i=1,\ldots,C;i\neq c}\|\boldsymbol{\gamma}^{(c,i)}\|_1\right\}+\eta_c\|\boldsymbol{\beta}_c\|_1$$

denotes the structured penalty term.

Note that the penalties can also be easily extended to more complex hierarchy structure; more details in Liu and Ye (2010).

3.2.3 Computation of the proposed regularized GLM

The main reason that Lasso and SGL can produce sparse solutions is that the penalty terms are not differentiable at 0. However this feature, on the other hand, also brings computational difficulty in solving the minimization problem. The standard convex optimization used to obtain MLE, like Newton-Raphson method, is not directly applicable to non-smooth objective functions. There has been a tremendous amount of work on regularized optimization from both statistics and computer science perspective. Recently, the coordinate descent (CD) algorithm, rediscovered by Friedman et al. (2007), has gained lots of attention in regularized linear and logistic regression and was shown to have computational superiority. Friedman et al. (2010) and Simon et al. (2013) applied similar idea and developed a block-wise coordinate descent (BCD) algorithm for SGL.

However, we also find that while the BCD algorithm is very fast and scales well in the linear model, it can be quite costly in GLM when the sample size is very large, i.e. number of bins in the spike train data. Based on

our initial simulation study, the BCD algorithm could not handle the large data set well at the size similar to the real data. Therefore, we develop an Accelerated Full Gradient Update (AFGU) algorithm motivated from the previous work by Kim et al. (2008), Beck and Teboulle (2009) and Wright (2012). Our proposed AFGU algorithm is based on the full gradient of log-likelihood function and a specific shrinkage-thresholding operator depending on the penalty function, combined with a Newton-based acceleration techniques over active parameters. Compared with BCD, the AFGU algorithm could improve the performance in two ways. First, we use the full gradient instead of coordinate-wise/block-wise gradient to avoid evaluating gradient at each coordinate one at a time. Support $\widetilde{\boldsymbol{\theta}}_c^{(0)}$ is the current estimate in the optimization. The full gradient update is equivalent to minimize

$$\frac{\rho}{2}\|\widetilde{\boldsymbol{\theta}}_c - \widetilde{\boldsymbol{\theta}}_c^{(0)}\|_2^2 + \nabla\ell_c(\widetilde{\boldsymbol{\theta}}_c^{(0)})^T(\widetilde{\boldsymbol{\theta}}_c - \widetilde{\boldsymbol{\theta}}_c^{(0)}) + \mathcal{P}_{(\eta_c,\alpha_c)}(\widetilde{\boldsymbol{\theta}}_c), \tag{3.5}$$

where ρ is some small positive constant to control the step size of the gradient update. The minimization in (3.5) is separable for different groups and the sub-problem for each group can be analytically solved as in Simon et al. (2013).

Second, a Newton-type acceleration step after the gradient update step is performed on the reduced parameter space which only consists of the coordinates with nonzero coefficient at current estimate. Let $\mathcal{A}$ denote

the index set of reduced parameter space and $g_{\mathcal{A}}$, $H_{\mathcal{A}}$ denote the reduce gradient vector and Hessian matrix of $L_c(\widetilde{\boldsymbol{\theta}}_c)$ for those parameters only in $\mathcal{A}$. The reduced-Newton update on the current estimate is then simply

$$\widetilde{\boldsymbol{\theta}}_{c,\mathcal{A}} = \widetilde{\boldsymbol{\theta}}^{(0)}_{c,\mathcal{A}} - \{H_{\mathcal{A}}(\widetilde{\boldsymbol{\theta}}^{(0)}_c)\}^{-1} g_{\mathcal{A}}(\widetilde{\boldsymbol{\theta}}^{(0)}_c)$$

where $\widetilde{\boldsymbol{\theta}}_{c,\mathcal{A}}$ denote the sub-vector of $\widetilde{\boldsymbol{\theta}}_c$ with the coordinates that are in $\mathcal{A}$. This acceleration step generally yields a vast performance improvement over the gradient methods that use only first-order information and the improvement is worth the cost of evaluating the reduced Hessian. When the true parameters are sparse, the reduced Hessian is small and its inverse can be easily computed. Therefore, the AFGU algorithm is expected to run faster and theoretically converges to the optimum point Q-quadratically which is better than simple gradient or coordinate-wise gradient method with Q-linear convergence rate (Wright (2012)). Some fine adjustments are also made to efficiently handle the sparse neural data which have a large portion of zero entries and stabilize some matrix operations.

For the tuning parameters η_c, we compute the regularization path, i.e. a sequence of solutions for a decreasing sequence $(\eta_c^{(1)}, \ldots, \eta_c^{(I)})$. Together with a grid search over the interval $[0, 1]$ for the tuning parameters α_c, we find the pair (α_c, η_c) that minimizes the Bayesian Information Criterion

(BIC) defined as

$$\mathrm{BIC}_c(\widehat{\widetilde{\boldsymbol{\theta}}}_c) = 2\ell_c(\widehat{\widetilde{\boldsymbol{\theta}}}_c) + \mathrm{df}(\widehat{\boldsymbol{\theta}}_c)\log(n)/n, \tag{3.6}$$

where $\ell_c(\cdot)$ is as defined in (3.3) and $\mathrm{df}(\widehat{\boldsymbol{\theta}}_c)$ is the number of nonzero elements of $\widehat{\boldsymbol{\theta}}_c$.

Note that AIC as another popular model selection criterion is not suitable for our purpose, because AIC usually performs better for prediction but not very well in the variable selection. While the main objective of the study is to detect the functional connectivity between neurons, i.e. nonzero coupling coefficients, BIC is more capable of selecting truly nonzero coefficients. Also as we illustrate in the next section, BIC enjoys some model selection consistency result, whereas AIC does not.

3.3 Theoretical properties

In this section we provide some theoretical properties of our proposed method. Here we allow the total number of neurons to diverge slowly with the length of the experiment, i.e. C can grow as n increases. Let $\widetilde{\boldsymbol{\theta}}_c^* = (\theta_0^*, \theta_1^*, \ldots, \theta_d^*)^T$ be the vector of unknown true parameters, where $d = P + (C-1)Q + M$. Theorem 3.1 guarantees the existence of a consistent local minimizer of (3.4) and such minimizer is $\sqrt{n/C}$-consistent.

Theorem 3.1. *Assume Conditions* B1–B3 *in the Appendix. If* $\eta_c\sqrt{n} = O(1)$

and $C^4/n = o(1)$ *as* $n \to \infty$, *then there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\theta}}}_c$ *of* (3.4) *such that* $\|\widehat{\widetilde{\boldsymbol{\theta}}}_c - \widetilde{\boldsymbol{\theta}}_c^*\| = O_P(\sqrt{C/n})$.

We would like to point out that the proof of Theorem 3.1 uses the penalized Bregman Divergence (BD) framework in Zhang (2010), but differs in the two parts. First, we need to treat the BD loss more carefully since the data set is not an i.i.d. sample. The martingale version of central limit theorem (Theorem 7.4 in Durrett (2010)) is used to establish the desired result in the proof. Second, the sparse group Lasso penalty is more complex than L_1 and those penalties that are imposed on individual variables. Details can be found in the Appendix.

Next we will study the asymptotic property of the BIC selector. Before we state our result, some notations are needed. Let $\mathcal{M}_F = \{1, \ldots, d\}$ denote the full model and $\mathcal{M}^* = \{1 \le j \le d : \theta_j^* \neq 0\}$ denote the index set of all truly important variables. Also define the class of correct models $\mathcal{C} = \{\mathcal{M} : \mathcal{M}^* \subseteq \mathcal{M},\ \mathcal{M} \subseteq \mathcal{M}_F\}$ as the collection of models that correctly select all truly important variables, and define the class of wrong models as $\mathcal{W} = \{\mathcal{M} : \mathcal{M}^* \not\subseteq \mathcal{M},\ \mathcal{M} \subseteq \mathcal{M}_F\}$, where at least one important variable is excluded. Then we have the following theorem.

Theorem 3.2. *Assume Conditions* B1–B4 *in the Appendix. If* $C = o\{\log(n)\}$, *then with probability approaching* 1, *the* BIC *in* (3.6) *can select the correct model in* $\mathcal{C}$ *with the smallest number of covariates, among all* $\sqrt{n/C}$*-consistent minimizer of* (3.4).

Remark 3.3. *In the proof of Theorem 3.2, we can see that for two choices of tuning parameters that both result in a* $\sqrt{n/C}$*-consistent estimator, the difference of* $\ell_c(\widehat{\widetilde{\boldsymbol{\theta}}}_c)$ *is of the order* $Op(C/n)$*. Therefore, the penalty term of* BIC *in* (3.6)*, which is of the order* $Op\{\log(n)/n\}$ *and higher than* $Op(C/n)$*, is able to select the correct model with the smallest number of covariates. In contrast, the penalty term used in* AIC*,* $2\mathrm{df}(\widehat{\boldsymbol{\theta}}_c)/n$*, just has the same order as* $Op(C/n)$*, thus* AIC *may not perform the similar variable selection as* BIC *does.*

3.4 Simulation studies

3.4.1 Simple network

To illustrate the application and performance of the proposed regularized GLM, we first simulate an ensemble of 10 neurons with the network structure given by Figure 3.1. There are 10 connections in the constructed network, of which 7 connections have type A kernel and 3 connections have type B kernel (red and blue in Figure 3.1 respectively). For each neuron, the baseline rate was set at $\gamma_{c,0} = -3$, which gives an average firing rate at $e^{-3}/0.1 = 0.5$ Hz when the bin size is 0.1 second. The length of the history window of both autoregressive and coupling kernel is set to be 10 bins in the generating process. Refractory periods of these neurons were modeled by letting first few parameters $\gamma_{c,p}$ to be very negative then rise to be slightly positive before going back to zero similar to Truccolo et al. (2005) (Figure 3.2). Two possible forms of coupling kernel functions

between neurons are also illustrated in Figure 3.2. For simplicity, the covariate term $\boldsymbol{X}(\tau_k)^T\boldsymbol{\beta}_c$ is not included in the simulation.

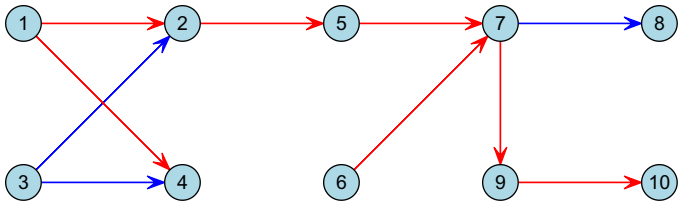

Figure 3.1: *A simulated network with* 10 *neurons. The colors of the edges indicate two difference types of coupling kernel functions* (*Figure* 3.2)*, where red represents type A kernel* (*more excitatory*) *and blue represents type B kernel* (*more inhibitory*).

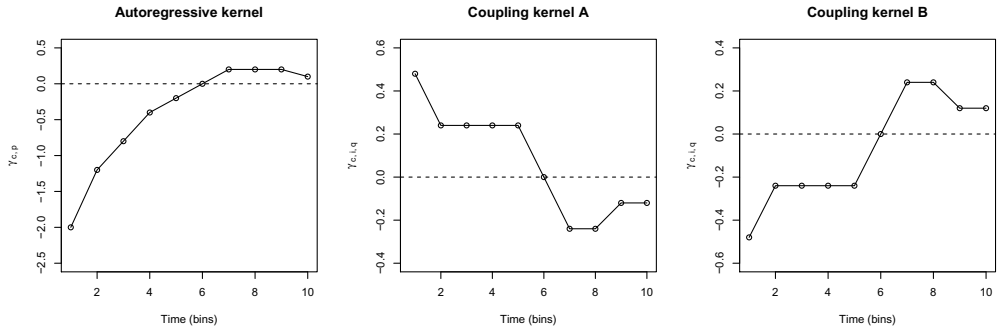

Figure 3.2: *True values of parameters in the kernel functions for the simulation.*

We compare the performance of the following four methods, two existing L_1 regularized methods in different settings and the proposed SGL regularized method with varying length of the history window. We also tried the SCAD penalty (Fan (1997)), which performs similar to the L_1 regularized methods (thus not shown here).

(I) (L_1-short) L_1 penalty with combined short-term history window (Zhao et al. (2012)), $P = 10$, $Q = 3$;

(II) (L_1-10-10) L_1 penalty with history window $P = 10$, $Q = 10$;

(III) (SGL-10-10) SGL penalty with history window $P = 10$, $Q = 10$;

(IV) (SGL-15-15) SGL penalty with history window $P = 15$, $Q = 15$;

Note that the method (I) uses a different parameterization as follows,

$$F^{-1}\left\{\lambda_c(\tau_k \mid N_{1:C}(\tau_{1:(k-1)}))\right\} = \gamma_{c,0} + \sum_{p=1}^{P} \gamma_{c,p} N_c(\tau_{k-p}) + \sum_{\substack{i=1 \\ i \neq c}}^{C} \gamma_{c,i,1} \left(\sum_{q=1}^{Q} N_i(\tau_{k-q}) \right),$$

where Q is chosen to be small and the coupling kernel coefficients $\{\gamma_{c,i,q} : q = 1, \ldots, Q\}$ are simplified to a single coefficient $\gamma_{c,i,1}$. The corresponding parameter vector is

$$\widetilde{\boldsymbol{\theta}}_c = \left(\gamma_{c,0}, \{\gamma_{c,p}\}_{p=1,\ldots,P}, \{\gamma_{c,i,1}\}_{i=1,\ldots,C,\, i \neq c}\right)^T,$$

and the imposed L_1 penalty is

$$\mathcal{P}_{\eta_c}(\widetilde{\boldsymbol{\theta}}_c) = \eta_c \left(\sum_{p=1}^{P} |\gamma_{c,p}| + \sum_{\substack{i=1 \\ i \neq c}}^{C} |\{\gamma_{c,i,1}| \right).$$

The method (IV) is to investigate the performance of the SGL regularized method with an inexact length of the history window since we do not know how long the interaction between neurons would last in reality. For a better comparison, we also carry out the simulation under the setting that the coupling kernel function is 25% stronger, i.e. the magnitude of coefficient is 25% larger than those given in Figure 3.2. We simulate spike trains with different length to see how the result changes.

The performance of all methods are summarized in Table 3.1. "CorrectAll" represents the number of all connections correctly detected by methods where there is indeed a connection between two neurons in the true network. "CorrectA" and "CorrectB" are number of occurrences that type A kernel and type B kernel are correctly detected respectively. "CorrectNC" is the number of pairs correctly identified as no connections where there is actually no connection. For all 4 metrics, "CorrectAll", "CorrectA", "CorrectB", and "CorrectNC", larger value means better performance. All values represent the average across 100 simulation runs for each spike train length and kernel function strength.

From Table 3.1, all methods can successfully detect the sparse structure of the network, i.e. most parameters are estimated as zero reflecting very good levels of specificity. However, the sensitivity of detecting significant coupling kernel functions is not as good as specificity. When the relative strength of coupling kernel functions increases, we can see that the sensitivity becomes better. Such an effect is reasonable since stronger signals

Table 3.1: ***(Simulation of simple network)*** *Performance comparison among regularized methods with varying kernel function strength and spike train length. Results are averaged over* 100 *replications. Here* std *is standard deviation.*

Relative Strength	Method	CorrectAll (std)	CorrectA (std)	CorrectB (std)	CorrectNC (std)
Length $= 10000$ bins					
100%	L_1-short	1.98 (1.5)	1.68 (1.3)	0.30 (0.5)	79.01 (1.3)
	L_1-10-10	0.75 (1.0)	0.72 (0.9)	0.03 (0.2)	79.72 (0.5)
	SGL-10-10	3.71 (1.7)	2.97 (1.4)	0.74 (0.7)	77.94 (2.1)
	SGL-15-15	1.75 (1.1)	1.53 (1.0)	0.22 (0.4)	79.30 (1.4)
	True network	10	7	3	80
125%	L_1-short	4.46 (1.7)	3.68 (1.4)	0.78 (0.7)	78.04 (1.7)
	L_1-10-10	1.98 (1.4)	1.91 (1.3)	0.07 (0.3)	79.66 (0.6)
	SGL-10-10	6.82 (1.5)	5.38 (1.1)	1.44 (0.8)	76.21 (2.9)
	SGL-15-15	4.72 (1.7)	3.96 (1.4)	0.76 (0.8)	77.96 (2.2)
	True network	10	7	3	80
Length $= 15000$ bins					
100%	L_1-short	6.57 (1.6)	5.05 (1.3)	1.52 (0.8)	74.45 (3.0)
	L_1-10-10	2.72 (1.7)	2.39 (1.4)	0.33 (0.6)	79.00 (1.1)
	SGL-10-10	6.87 (1.7)	5.22 (1.3)	1.65 (0.8)	76.47 (2.6)
	SGL-15-15	5.19 (1.8)	4.09 (1.4)	1.10 (0.9)	77.76 (2.2)
	True network	10	7	3	80
125%	L_1-short	8.36 (1.3)	6.29 (0.8)	2.07 (0.8)	73.55 (2.9)
	L_1-10-10	5.01 (1.6)	4.51 (1.3)	0.50 (0.6)	78.58 (1.3)
	SGL-10-10	9.18 (0.8)	6.72 (0.5)	2.46 (0.6)	75.04 (3.2)
	SGL-15-15	8.32 (1.2)	6.32 (0.7)	2.00 (0.8)	76.75 (2.2)
	True network	10	7	3	80

will result in bigger changes to the conditional intensity, which makes the interaction easier to be detected. When we increase the length of the spike train to 15000 bins, we see better performance for all methods, which would be expected since we simply have larger data set and more infor-

mation to study the network structure of the neuron population. Among all methods, SGL regularized method with the exact length of kernels' history window has the best performance of finding the interactions between neuron. Interestingly, even when we misspecified the length of the history window of coupling kernel functions in method (IV), it can still get the comparable results. Both of SGL regularized methods outperform the other two L_1 regularized methods.

3.4.2 Complex network

Next, we simulate a more complex ensemble of 60 neurons with the network structure given by Figure 3.3 and Figure 3.4. There are 60 connections in the constructed network, of which 30 connections have type A kernel and 30 connections have type B kernel. The other settings are similar to the previous simulation except that we simulate with longer spike train in order to handle the larger network.

The results of this simulation are summarized in Table 3.2. Besides those patterns that are already discussed in the previous simulation, we can clearly see from Figure 3.5 that SGL regularized methods perform considerably better than two L_1 based methods, for e.g. method (III) is able to identify 10 more true connections on average and much fewer false discoveries than method (I) when $n = 25000$ and the relative strength is 125%. The estimated network structures by the SGL regularized method are already very close to the true simulation network in most cases.

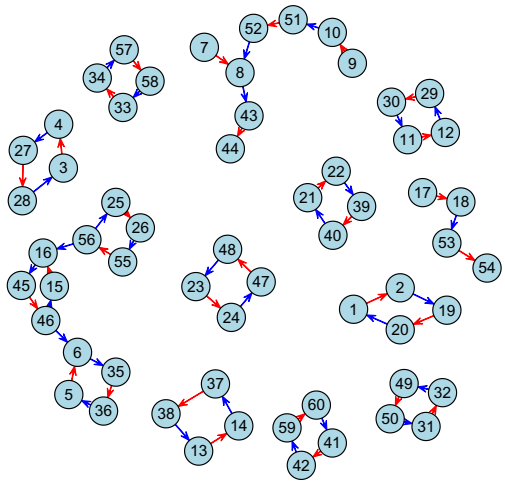

Figure 3.3: *A simulated network with* 60 *neurons. The colors of the edges indicate two difference types of coupling kernel functions* (*Figure* 3.2)*, where red represents type A kernel and blue represents type B kernel.*

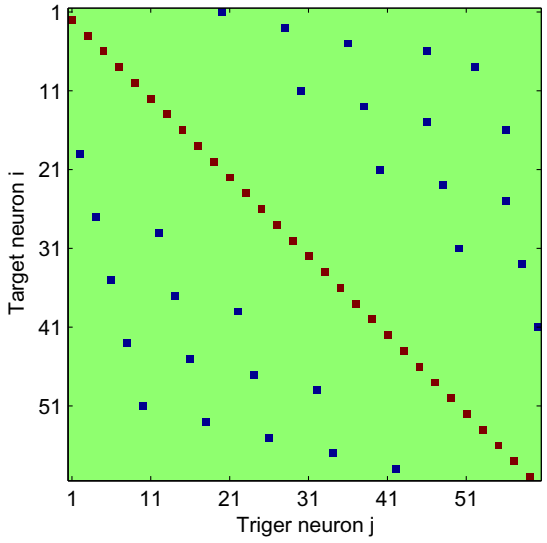

Figure 3.4: *The connectivity matrix of the simulated network. The color of the cell at* i*th row and* j *column indicates how neuron* j*'s spiking history would affect the firing rate of neuron* i*. Green means no connection; Red and Blue represents two difference types of coupling kernel functions respectively.*

Table 3.2: ***(Simulation of complex network)*** *Performance comparison among regularized methods with varying kernel function strength and spike train length. Results are averaged over* 100 *replications. Here* std *is standard deviation.*

Relative Strength	Method	CorrectAll (std)	CorrectA (std)	CorrectB (std)	CorrectNC (std)
Length $= 15000$ bins					
100%	L_1-short	8.87 (2.4)	7.44 (2.3)	1.43 (1.2)	3475.32 (2.9)
	L_1-10-10	6.10 (2.5)	5.47 (2.3)	0.63 (0.8)	3475.40 (2.4)
	SGL-10-10	18.21 (3.3)	12.46 (2.7)	5.75 (2.2)	3473.82 (2.9)
	SGL-15-15	10.00 (2.6)	7.55 (2.1)	2.45 (1.5)	3477.51 (1.6)
	True network	60	30	30	3480
125%	L_1-short	19.11 (3.6)	15.30 (2.7)	3.81 (1.8)	3472.37 (3.3)
	L_1-10-10	13.76 (3.4)	12.43 (2.9)	1.33 (1.2)	3474.64 (2.6)
	SGL-10-10	36.83 (4.1)	22.97 (2.5)	13.86 (3.0)	3471.46 (3.3)
	SGL-15-15	24.85 (3.9)	17.86 (3.0)	6.99 (2.4)	3476.32 (2.1)
	True network	60	30	30	3480
Length $= 20000$ bins					
100%	L_1-short	26.69 (3.7)	19.00 (2.7)	7.69 (2.2)	3455.69 (7.2)
	L_1-10-10	16.37 (3.2)	13.10 (2.5)	3.27 (1.8)	3466.82 (3.9)
	SGL-10-10	30.45 (3.8)	18.84 (2.6)	11.61 (2.7)	3472.86 (3.0)
	SGL-15-15	21.53 (3.7)	14.34 (2.7)	7.19 (2.5)	3476.18 (2.1)
	True network	60	30	30	3480
125%	L_1-short	38.99 (3.4)	26.19 (1.8)	12.80 (2.6)	3448.41 (8.7)
	L_1-10-10	27.80 (3.3)	22.21 (2.4)	5.59 (2.1)	3466.01 (4.1)
	SGL-10-10	50.53 (2.9)	28.15 (1.3)	22.38 (2.5)	3468.84 (3.8)
	SGL-15-15	42.61 (3.2)	25.59 (1.7)	17.02 (2.5)	3473.92 (2.8)
	True network	60	30	30	3480
Length $= 25000$ bins					
100%	L_1-short	42.55 (3.5)	26.18 (1.8)	16.37 (2.8)	3426.78 (9.1)
	L_1-10-10	26.72 (3.8)	19.08 (2.7)	7.64 (2.4)	3458.81 (5.4)
	SGL-10-10	45.52 (2.8)	25.04 (2.1)	20.48 (2.3)	3469.71 (3.5)
	SGL-15-15	35.72 (3.5)	20.91 (2.5)	14.81 (2.7)	3474.99 (2.2)
	True network	60	30	30	3480
125%	L_1-short	51.57 (2.5)	29.39 (0.8)	22.18 (2.2)	3419.09 (9.9)
	L_1-10-10	40.02 (2.9)	26.84 (1.6)	13.18 (2.6)	3459.92 (5.6)
	SGL-10-10	57.77 (1.4)	29.60 (0.7)	28.17 (1.3)	3467.06 (4.1)
	SGL-15-15	54.21 (2.4)	28.88 (1.2)	25.33 (2.1)	3472.29 (3.3)
	True network	60	30	30	3480

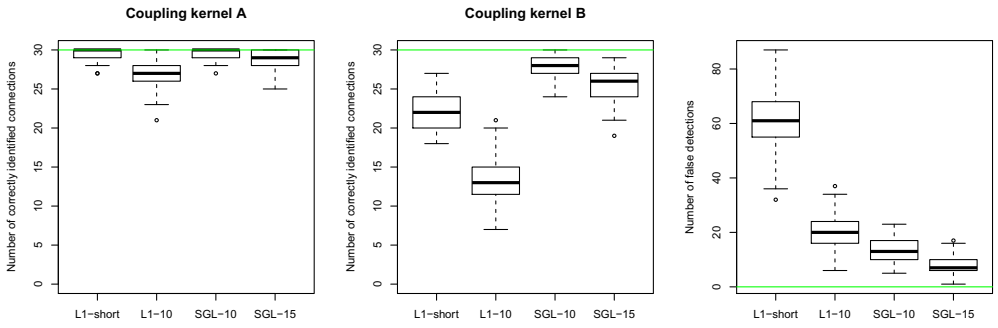

Figure 3.5: *Comparison of the numbers of correctly identified connections and false detections among 4 methods in the simulation of the complex network when* $n = 25000$ *and relative strength is 125%.*

We also find that across all methods the connections of type B coupling kernel is much harder to detect than connections of type A kernel. Since type A kernel is more excitatory (more positive coupling coefficients $\gamma_{c,i,q}$) and type B kernel is more inhibitory (more negative $\gamma_{c,i,q}$), it is likely that the initially low baseline firing rate could be difficult for the detection of inhibitory influences given the relatively short length of simulated spike train. The estimated kernel functional form by L_1-10-10 and SGL-10-10 are provided in Figure 3.6, in which red lines represent the true functional forms and blue lines are estimated kernels from different trials. (Since L_1-short treats the whole coupling kernel function as one single coefficient, it can not provide the detailed functional form of kernels.) We find that the L_1-10-10 method usually can only detect the high peak at the beginning while the SGL-10-10 method can somewhat provide more informative

estimation of the kernel functions.

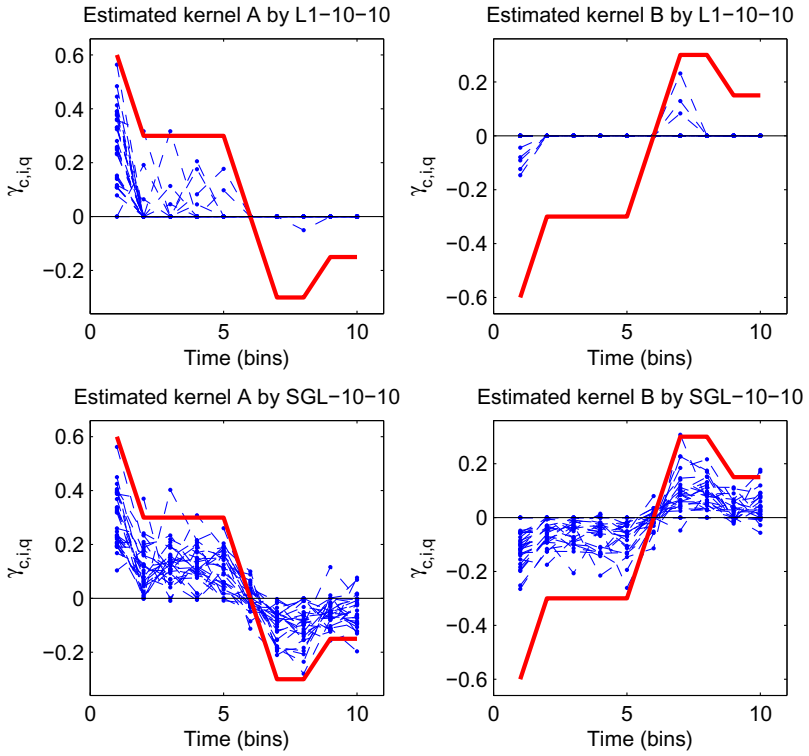

Figure 3.6: *The estimated kernel functional forms in the simulation of the complex network when* $n = 25000$ *and relative strength is 125%.. The left two plots are for type A kernel and the right plots are for type B kernel. The estimation obtained by* L_1*-10-10 is given by two plots on top, and estimation by* SGL*-10-10 is given by the other two plots at bottom. Red sick lines represent the true functional forms and blue dashed lines are estimated forms.*

3.5 Application to neurophysiological data

3.5.1 T-Maze task of working memory

We aim to estimate the functional connectivity structure of neurons in the rat prelimbic region of the frontal cortex (plPFC) using the SGL regularized GLM. Neural data were obtained from adult male Sprague-Dawley rats performing a T-Maze based delayed-alternation task of working memory (Devilbiss and Waterhouse (2004); Devilbiss et al. (2006), Devilbiss et al. (2012)). In brief, animals were initially trained to navigate down the runway of the T-Maze and choose one of two arms opposite to the one previously visited for food rewards (chocolate chips 1.6 gm) delivered by the experimenter's hand. For each trial, the animal was placed in a start-box for a pre-determined interval (a delay) and released with the removal of a starting gate. The rat would move to the junction of the "T" and choose the left or right arm (Figure 3.7). On a correct choice, the animal was rewarded and replaced in the start-box, yielding a correct sequence of "left, right, left, right . . .". On incorrect trials the animals was removed from the incorrectly chosen arm and returned to the start-box. Training continued in the T-maze task to 90% accuracy on 10 trials (0 seconds delay, 1 testing session per day). A restricted feeding schedule (16g-20g of standard chow) maintained motivation with quantities of food titrated for each animal to maintain motivation for each 40 trial session. Animals were then surgically implanted with recording electrodes and returned to ad-lib feeding

for the duration of recovery (7-10 days). Following recovery, restricted feeding was reinstated and training continued until animal performance was stable across 40 trials at 90-100% correct with at least a 10 second delay. Training/testing occurred at the same time each day.

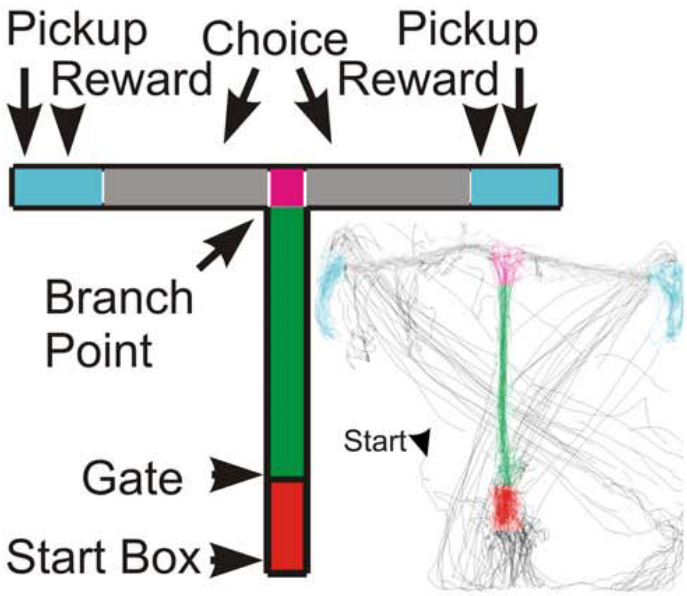

Figure 3.7: ***(Neurophysiological data)*** *T-Maze task: for each trial, the animal was placed in a start-box for a pre-determined interval (a delay) and released with the removal of a starting gate. The rat would move to the junction of the "T" and choose the left or right arm. On a correct choice, the animal was rewarded and replaced in the start-box.*

3.5.2 Neural data discrimination and recording

Online Discrimination and Recording: On the day of a recording session, the animal was tethered to customized multi-channel electrophysiological

hardware (Plexon Inc, Dallas, TX). Neural activity was amplified, discriminated, time stamped, and recorded from putative single units of the plPFC. The animal remained in its home cage, placed above the T-maze, during this discrimination phase to allow the animal to habituate to the tether and the recording arena. Template matching algorithms were applied to neural activity to preliminarily discriminate action potentials exhibiting a $3:1$ signal to noise ratio. Following discrimination of plPFC units, the animals remained tethered to the recording hardware for the remainder of the day. During each testing session, videotape recordings were made of the entire experimental session with a video counter timer providing time stamps (resolution = 0.0125 sec) synchronized to the multi-unit recording systems.

Offline Unit Analysis: After each experimental session, pre-established offline criteria were used to verify that waveforms assigned to each online discriminated united originated from a single neuron. These previously described criteria were based on unit waveform properties and spike train discharge patterns including: 1) variability of peak waveform voltage, 2) variability of waveform slope(s) from peak to peak, 3) separability of clustering of scattergram points from the waveform's first two principal components, and 4) spike train autocorrelogram. Neurons that did not meet these criteria were excluded from the study. Neuronal waveform shape, discharge pattern (inter-spike interval), and response properties were further examined to verify that neurons were not recorded across mul-

tiple recording sessions. Further analyses of data from neurons recorded from multiple session days were limited to the first recording session of that neuron. Lastly, plPFC units were further classified as "broad spike" (BS-type) or "narrow spike" (NS) to putatively identify large projection pyramidal neurons (BS-type). Essentially, the peak-peak (P-P) duration of waveforms from verified neurons were calculated. Neurons with P-P intervals greater than 200 μs were classified as BS-type neurons. Other cells classified as NS neurons (P-P intervals between 100-200 μs) or neurons not meeting either category were eliminated from further analyses. Following each experimental session, the behavioral intervals were visually identified by the experimenter from time-stamped video recordings and manually entered into each neural recording data file. These event-states were defined by events occurring as the animal performs the T-Maze task that included

- placement of the rat into the start box,
- removal of the start gate,
- rat reaching the branch point of the "T",
- the rat entering one of two goal arms (choice),
- receipt of food reward, and
- removal of the rat to begin another trial.

Each trial was further classified as a correct or incorrect trial as well as by chosen spatial goal (i.e. left vs. right arm). Dividing the task into these behavioral intervals should not be interpreted as representing individual and discrete cognitive functions. Instead, these are convenient delineations of different behaviors and task events where different PFC-dependent processes may be involved. Certainly, one can expect that some PFC-dependent processes may be involved in adjacent behavioral intervals, and like all delayed-response tasks, a clear temporal delineation of onset and termination of different cognitive functions is difficult.

3.5.3 Neurophysiological data analysis

Data set comprises 8 experiments, each having 40 trails (40 turns). Basic statistics are provided in Table 3.3. Given the length of the spike trains and the relatively low firing rate on average, we bin the spikes at 100 ms (0.1 sec) which gives about 10000 to 20000 bins for each experiment. We set $P = 30$, i.e. the spike history may affect its own spontaneous firing rate up to 3 seconds; and $Q = 10$, i.e. the influence of spikes from other neurons lasts up to 1 second. Other choices of P and Q are also considered, which give similar results. Finally we code the current state of the experimental animals into 8 indicator columns which are listed in Table 3.4. Note that we only model the states parameters as linear terms which only change the overall firing rate and do not change the connectivity network structure. Then we applied the SGL regularized method to recorded

neurophysiological data for each of the eight experiments.

Table 3.3: ***(Neurophysiological data)*** *Summary statistics of the T-Maze task dataset.*

		Total length	Total number of neurons			Average
No.	Subject	(sec)	BS-type	NS-type	Total	firing rate (Hz)
1	TM19	1449.2	53	4	73	0.567
2	TM19	1150.8	60	5	83	0.408
3	TM44	2243.7	44	6	59	0.772
4	TM44	2150.3	62	3	77	0.364
5	TM44	2191.1	57	3	71	0.192
6	TM44	2011.4	53	3	63	0.241
7	TM57	930.3	30	8	39	0.375
8	TM58	765.2	23	0	26	0.342

Table 3.4: ***(Neurophysiological data)*** *Description of the covariates used in the analysis.*

Var. #	Name	Description
1	Delay	Stay in the resting area until the gate is released.
2	Run	Run through from the gate to the branch point.
3	Decision	At the branch area of the T maze and make a turn.
4	Choice	Run to the left or right end.
5	Reward	Eat reward in the correct trail.
6	Pick-up	Be picked up and put back to the resting area.
7	Left	Whether made the left turn in the trial.
8	Correct	Whether made the correct choice in the trial. (different from the previous one.)

For the parameters of the autoregressive kernel functions, $\gamma_{c,p}$, a lot of them are estimated to be zero. Most of the identified nonzero parameters

are positive, which indicate that there is some self-exciting processes during a certain period of time. Figure 3.8 gives the estimated autoregressive kernel parameters for selected neurons (No.14, No.16 and No.73) in experiment 4, which illustrate some typical autoregressive kernel functional curves. The refractory effect was rarely seen in the study. It is possible that the 0.1 sec resolution we used may be already longer than the refractory period of neurons. Note that some hilly curves are observed from Neurons 14 and 16. This cyclic phenomenon is possibly caused by mutually excitatory connections between these two neurons, rather than solely the individual self-exciting process. Such excitatory pair or clique can also be found in the network structure of other experiments.

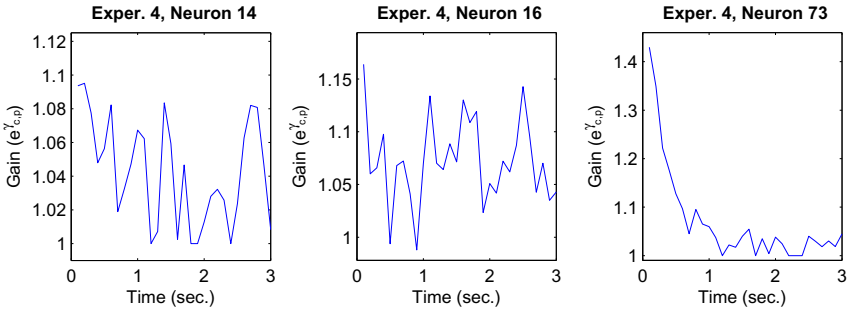

Figure 3.8: ***(Neurophysiological data)*** *The estimated parameters of autoregressive kernel functions for selected neurons in experiment* 4.

Table 3.5 summarized the number of detected functional connections into two categories: excitatory and inhibitory connections. The type of detected connection from neuron c to neuron i is determined by the sign of the sum of the coupling kernel coefficients, $\text{sign}(\sum_{q=1}^{Q} \gamma_{c,i,q})$. We also

compare the results with two L_1 based methods discussed in Section 3.4. We can see that among over thousands of possible connection pairs, only a few of them are selected by regularized methods. The short-term L_1 method detect more connections than the other two methods, but based on the simulation results we may doubt that it could report several false discoveries. The long-term L_1 method reports smallest number of connections, and the SGL should provide a good balance and better results as those in the simulation studies. For example, the estimated networks from experiment 1 and 4 are given in Figure 3.9.

Table 3.5: ***(Neurophysiological data)*** *Summary of estimated networks by regularized methods.*

	Possible	L_1-short		L_1		SGL	
Experiment	pairs	Excit.	Inhib.	Excit.	Inhib.	Excit.	Inhib.
1	5256	51	41	20	13	34	18
2	6806	18	23	13	11	11	13
3	3422	114	59	52	18	57	10
4	5852	100	52	25	10	57	17
5	4970	17	5	3	0	6	1
6	3906	30	7	16	4	12	3
7	1482	52	9	16	1	19	2
8	650	0	0	0	0	0	0

Also we provide the colored maps of the interaction matrix in Figure 3.10, whose color at ith row and jth column represents influence from neuron j onto neuron i. Red is for excitatory effect, blue for inhibitory effect and green for no effect, where the exact color is determined by the

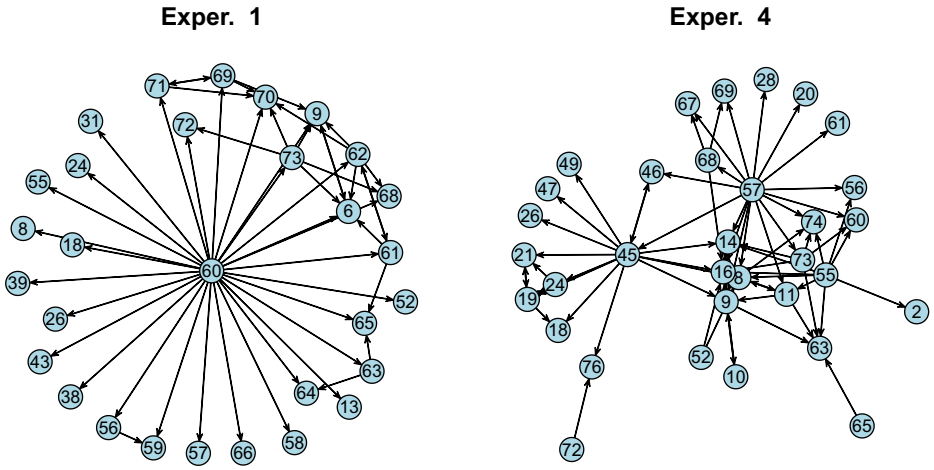

Figure 3.9: ***(Neurophysiological data)*** *Estimated network structures by* SGL *regularized method.*

strength of the interaction, i.e. the Euclidean (L_2) norm of the coupling kernel coefficients, $\sqrt{\sum_{q=1}^{Q} \gamma_{c,i,q}^2}$. (Experiment 5 and 8 are omitted in Figure 3.8 since too few connections are identified, which may be due to the low baseline firing rate and short experimental time.)

The detected connections among neurons share some common characteristic across experiments. For example, there are several excitatory pairs and cliques along the diagonal lines, which means the neurons in the self-exciting groups tend to be close. On the other hand, the inhibitory connections are usually found between long distance pairs. The number of detected inhibitory connections is fewer than that of the excitatory ones. This is probably because that there is much more BS-type neurons than NS-type neurons and BS-type neurons tend to have excitatory effect

on other neurons while NS-type neurons tend to have inhibitory effect. We need to note that the term "functional connections" here refers to the statistical dependencies between spike trains (neurons) and does not necessarily imply the existence of an anatomical connection between the corresponding neurons (Okatan et al. (2005)), nevertheless those results would be still useful to guide further research.

Table 3.6 provides a summary of estimated coefficients for 8 state indicator covariates described in Table 3.4. Among all coefficients, many are estimated to be insignificant, which means the spike activities are generally similar across different states of the T-maze trials. Among all covariates, the first, second and third covariates show somewhat stronger sign of difference in the firing rate at the "Delay", "Run" and "Decision" states. There are more negative coefficients selected than positive coefficients for the "Delay" variable, which suggests that the overall firing rate is lower when the subject was waiting in the resting area. On the other hand, more positive coefficients for "Run" and "Decision" variables suggest that the firing rate could be higher when the subject was moving and deciding the turning direction. However, such evidence is not strong enough to make further conclusion.

Table 3.6: ***(Neurophysiological data)*** *Summary of estimated coefficients for* 8 *state indicator covariates described in Table* 3.4.

Exper.	Count number of positive and negative coefficients															
	Delay		Run		Decision		Choice		Reward		Pick-up		Left		Correct	
	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−
1	5	3	1	0	12	5	0	0	0	0	3	2	2	3	1	5
2	9	7	2	0	0	1	1	0	1	1	2	1	0	2	0	1
3	8	29	5	1	5	3	2	0	0	0	7	1	1	7	0	2
4	9	23	7	0	6	0	8	0	0	0	5	4	8	4	6	3
5	6	13	2	2	1	0	2	0	0	1	5	1	1	4	1	3
6	11	15	2	0	3	0	3	0	5	0	3	3	1	4	0	1
7	10	7	15	1	6	0	1	1	1	6	1	10	4	6	0	5
8	0	0	0	0	0	0	0	0	1	0	4	0	1	0	1	0

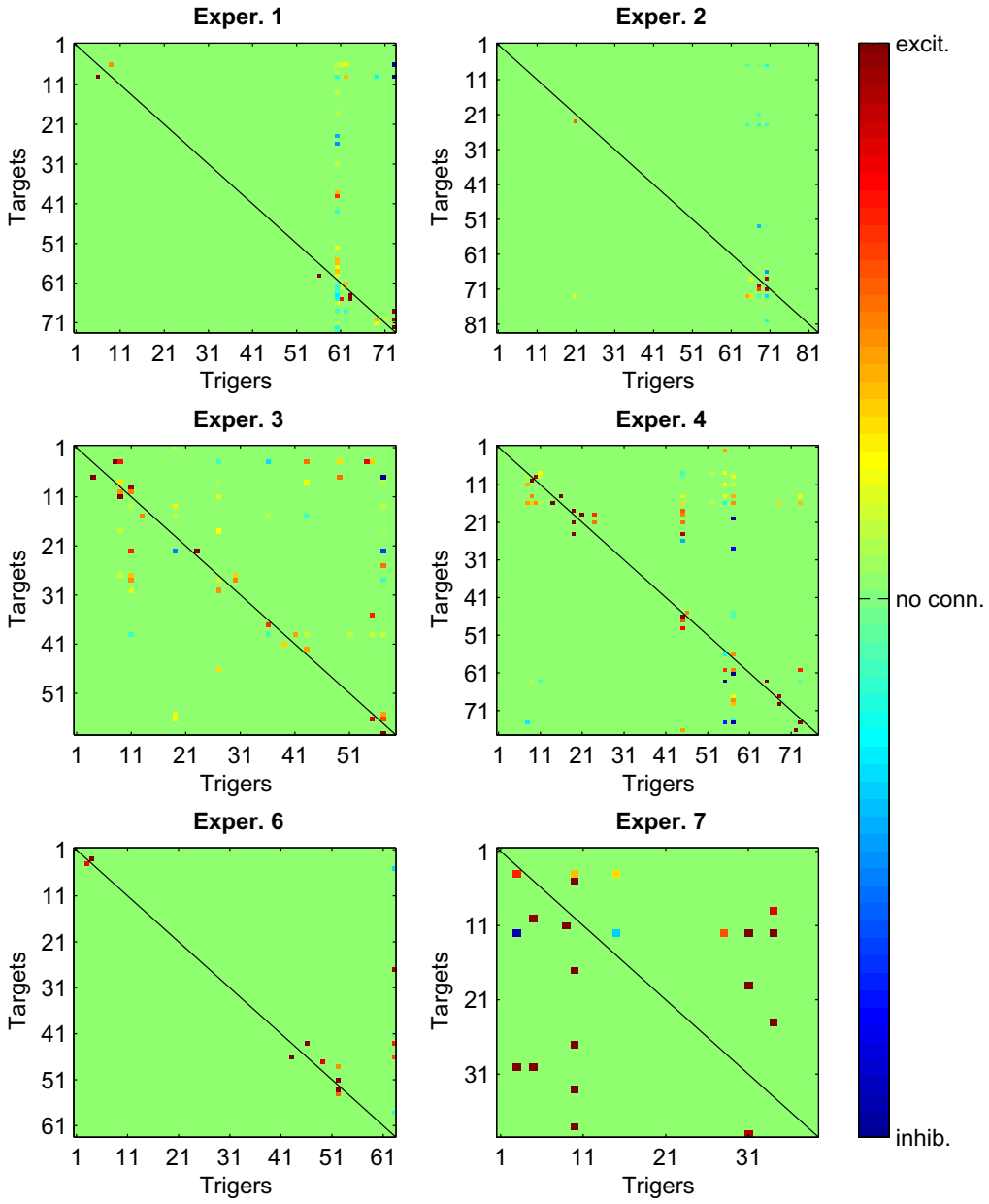

Figure 3.10: ***(Neurophysiological data)*** *Estimated network matrices by* SGL *regularized method. The red and blue colors indicate the excitation and inhibition between the pair, while green means no functional connections between the pair.*

Chapter 4

Discussion

When analyzing simultaneously recorded spike trains, it is desired to have one unified model to include several factors like the autoregressive effect for each neuron, the functional connections between pairs of neurons as well as the experimental state indicators. In this work, we use the regularized GLM with Poisson response to achieve such a goal. CI-GLM framework is able to deal with different types of the factors that can affect the neuron activities, meanwhile the appropriate penalty can force the coefficients of those insignificant factors to be exactly zero so that the functional connections between neurons can be detected and separated from noises.

Here our approach is based on the belief that among the huge amount of all possible connections, only a very small portion are really significant, so that a sparse network can be constructed using the estimated

connections from regularized method. Comparing with previous works (for example, Zhao et al. (2012)), We further propose the use of structured regularization which provides more flexibility to incorporate the prior structure information of parameter space and it improves the performance. We give more rigorous theoretical results when considering the dependence within the spike trains. From the simulation results, the SGL regularized method can indeed improve the sensitivity for detecting the true connections without sacrificing the specificity. The computation is very efficient using the proposed AFGU algorithm under the sparse network assumption.

We then applied the proposed method to real spike train recordings from neurons in the prelimbic region of the frontal cortex of adult male rats. Some interesting findings about the ensemble of neurons includes:

- more excitatory connections are detected than inhibitory connections;
- excitatory connections are more likely to appear among neighbor neurons, while inhibitory connections tend to be long-distance;
- the firing rate tends to be lower when the subject was waiting in the resting area and higher when the subject was making decisions.

Although the statistically significant functional connection does not infer synaptic connections, it provides useful information to guide further

research on the details of the interactions within neuronal network. In summary, our proposed methods is applicable to a board type of simultaneously spike train recordings and expected to have better performance than existing methods.

On the other hand, we could continue our studies by exploring the following in the future research:

- We manually choose bin size and the history window length in current analysis, however an automatic way to adaptively choose those parameters might be desired.

- Several types of structured regularization are proposed in recent statistics literature (Wang et al. (2009), Geng et al. (2013), and Liu and Ye (2010)). We could compare their performance in the context.

Appendix A

Proofs

A.1 Proofs of theorems in Chapter 2

Notation. For notational brevity, let $\mathsf{X}_j = (1, X_j)^T$ and $\mathbf{b}_j = (\alpha_j, \beta_j)^T$ denote the two-dimensional covariate and parameter of the componentwise regression minimum-BD estimation in (2.6), respectively. Denote $\widehat{\mathbf{b}}_j^{\mathrm{CR}} = (\widehat{\alpha}_j^{\mathrm{CR}}, \widehat{\beta}_j^{\mathrm{CR}})^T$ and $\mathbf{b}_j^{\mathrm{CR}} = (\alpha_j^{\mathrm{CR}}, \beta_j^{\mathrm{CR}})^T$. Throughout this section, $\|\cdot\|_1$ is the L_1-norm, $\|\cdot\|$ is the Euclidean L_2-norm, and $\|\cdot\|_\infty$ is used to denote the L_∞-norm.

Condition. We have the following assumptions in which M, B, B' are sufficiently large constants. Those are not the weakest possible, but serve to facilitate the technical derivations.

A1. For all j, X_j are uniformly bounded, i.e. $\|\boldsymbol{X}\|_\infty \leq M$. $\Sigma = \mathrm{var}(\boldsymbol{X})$

exists finitelyand is nonsingular and $\liminf_{n\to\infty} \min_{j=1,\ldots,p_n} \mathrm{var}(X_j) > \delta$ for some positive number δ.

A2. $\mathrm{var}(Y \mid \boldsymbol{X}) > 0$, $E(Y^2) < \infty$ and the tail probability of Y satisfies that there exist some positive constants m_0 and m_1 such that for sufficiently large t, $P(|Y| > t) \le m_0 \exp(-m_1 t)$.

A3. Assume that the quantities $\mathrm{q}_k(y;\theta) = (\partial^k/\partial\theta^k)Q(y, F^{-1}(\theta))$, $k = 0, 1, \ldots$, exist finitely up to any order required. Suppose $\mathrm{q}_2(y;\theta) > 0$ for all $\theta \in \mathbb{R}$ and all y in the range of Y.

A4. $F(\cdot)$ is a bijection and F''' is continuous. Without loss of generality, assume $F'(\cdot) > 0$.

A5. For all j, $\mathbf{b}_j^{\mathrm{CR}}$ is an interior point of $\mathbb{R}_B^2 = \{(a,b) \in \mathbb{R}^2 : |a| \le B, |b| \le B\}$.

C. $\|\boldsymbol{\beta}_0^{(\mathrm{I})}\|_1 \le B'$.

D. Assume that the eigenvalues of $\boldsymbol{\Omega}_n$ and $\mathbf{H}_n$ are uniformly bounded away from 0; $\|\mathbf{H}_n^{-1}\boldsymbol{\Omega}_n\|$ is bounded away from ∞.

Proof of Theorem 2.1. Since $m(\boldsymbol{X}) = E(Y \mid \boldsymbol{X})$,

$$\begin{aligned}
\mathrm{cov}(E(Y \mid \boldsymbol{X}), X_j) &= E\{E(Y \mid \boldsymbol{X})X_j\} - E\{E(Y \mid \boldsymbol{X})\}E(X_j) \\
&= E\{E(YX_j \mid \boldsymbol{X})\} - E(Y)E(X_j) \\
&= E(YX_j) - E(Y)E(X_j) = \mathrm{cov}(Y, X_j).
\end{aligned}$$

It follows from Condition A3 that $Q(y, F^{-1}(\theta))$ is strictly convex in θ. Therefore, the minimizer of (2.7) is the solution of the score equations of (2.7) which are given by

$$E\{\mathrm{q}_1(Y; \alpha_j^{\mathrm{CR}} + X_j\beta_j^{\mathrm{CR}})\} = E\Big\{\frac{(Y - \mu_j^{\mathrm{CR}})q''(\mu_j^{\mathrm{CR}})}{F'(\mu_j^{\mathrm{CR}})}\Big\} = 0,$$
$$E\{\mathrm{q}_1(Y; \alpha_j^{\mathrm{CR}} + X_j\beta_j^{\mathrm{CR}})X_j\} = E\Big\{\frac{X_j(Y - \mu_j^{\mathrm{CR}})q''(\mu_j^{\mathrm{CR}})}{F'(\mu_j^{\mathrm{CR}})}\Big\} = 0,$$

where $\mu_j^{\mathrm{CR}} = F^{-1}(\alpha_j^{\mathrm{CR}} + X_j\beta_j^{\mathrm{CR}})$.

We first show that if $\beta_j^{\mathrm{CR}} = 0$, then $\mathrm{cov}(Y, X_j) = 0$. When $\beta_j^{\mathrm{CR}} = 0$, μ_j^{CR} is a constant. Two score equations become

$$\mu_j^{\mathrm{CR}} = F^{-1}(\alpha_j^{\mathrm{CR}}) = E(Y), \quad E(X_jY) - \mu_j^{\mathrm{CR}}E(X_j) = 0,$$

which implies $\mathrm{cov}(Y, X_j) = 0$.

On the other side, if $\mathrm{cov}(Y, X_j) = 0$, it is easy to verify that $(F(E(Y)), 0)$ satisfies the score equations,

$$\begin{aligned} E\Big[\frac{\{Y - E(Y)\}q''(E(Y))}{F'(E(Y))}\Big] &= E\{Y - E(Y)\}\frac{q''(E(Y))}{F'(E(Y))} = 0, \\ E\Big[\frac{X_j\{Y - E(Y)\}q''(E(Y))}{F'(E(Y))}\Big] &= \mathrm{cov}(Y, X_j)\frac{q''(E(Y))}{F'(E(Y))} = 0. \end{aligned}$$

Therefore $\beta_j^{\mathrm{CR}} = 0$. ■

Proof of Theorem 2.2. We first show part (i). The first two partial derivatives of $\ell_j^{\mathrm{CR}}(\alpha_j, \beta_j) = E\{Q(Y, F^{-1}(\alpha_j + X_j\beta_j))\}$ with respect to α_j are given

by

$$\begin{aligned}\frac{\partial \ell_j^{\mathrm{CR}}(\alpha_j,\beta_j)}{\partial \alpha_j} &= E\{\mathrm{q}_1(Y;\alpha_j+X_j\beta_j)\},\\ \frac{\partial^2 \ell_j^{\mathrm{CR}}(\alpha_j,\beta_j)}{\partial \alpha_j^2} &= E\{\mathrm{q}_2(Y;\alpha_j+X_j\beta_j)\}.\end{aligned}$$

By Condition A3, it follows that $\ell_j^{\mathrm{CR}}(\alpha_j,\beta_j)$ is convex in α_j. Then, for any given $\beta_j = b$, the minimizer of $\ell_j^{\mathrm{CR}}(\alpha_j, b)$ will be the solution to the following equation,

$$h_j(\alpha; b) = E\{\mathrm{q}_1(Y;\alpha + X_j b)\} = 0.$$

Denote by $\alpha_j(b)$ the solution of $h_j(\alpha; b) = 0$. Thus, $\alpha_j(b)$ is unique and a well-defined function of b.

Now we have an equivalent definition of β_j^{CR} given by

$$\beta_j^{\mathrm{CR}} = \arg\min_b \ell_j(b),$$

where $\ell_j(b) = E\{Q(Y, F^{-1}(\alpha_j(b) + X_j b))\}$ and the first and second derivatives of $\ell_j(b)$ are given by

$$\begin{aligned}\ell_j'(b) = \frac{d\ell_j(b)}{db} &= E[\mathrm{q}_1(Y;\alpha_j(b)+X_jb)\{\alpha_j'(b)+X_j\}],\\ \ell_j''(b) = \frac{d^2\ell_j(b)}{db^2} &= E[\mathrm{q}_2(Y;\alpha_j(b)+X_jb)\{\alpha_j'(b)+X_j\}^2] > 0,\end{aligned}$$

which implies that $\ell_j(b)$ is convex in b and β_j^{CR} is unique and satisfies

$\ell_j'(\beta_j^{\mathrm{CR}}) = 0$.

By the mean-value theorem,

$$\ell_j'(b) = \ell_j'(0) + b\ell_j''(b^*), \tag{A.1.1}$$

where b^* is between 0 and b. It can be shown that

$$\alpha_j(0) = F(E(Y)) \quad \text{for any } j = 1, \ldots, p_n.$$

Since $E\{\mathrm{q}_1(Y; \alpha_j(0))\} = 0$, we observe that

$$\ell_j'(0) = E\{\mathrm{q}_1(Y; \alpha_j(0))X_j\} = C_0 \operatorname{cov}(Y, X_j),$$

where $C_0 = q''(E(Y))/F'(E(Y))$. By conditions A1 and A3, $|X_j| \leq M$ and $\mathrm{q}_2(y; \theta) > 0$, we observe that for any b and $j = 1, \ldots, p_n$,

$$|\alpha_j'(b)| = \Big| - \frac{E\{\mathrm{q}_2(Y; \alpha_j(b) + X_j b)X_j\}}{E\{\mathrm{q}_2(Y; \alpha_j(b) + X_j b)\}} \Big| \leq M.$$

Let $K_1 = \sup_{|\theta| \leq (M+1)B} E\{\mathrm{q}_2(Y; \theta)\}$. Then for any $j = 1, \ldots, p_n$ and any $-B < b < B$,

$$\ell_j''(b) \leq (2M)^2 K_1.$$

Let $c_1 = C_0/(4K_1M^2)$. By (A.1.1), for any $j = 1, \ldots, s_n$,

$$|\beta_j^{\mathrm{CR}}| \geq \frac{|C_0 \operatorname{cov}(Y, X_j)|}{4K_1M^2} = c_1 |\operatorname{cov}(Y, X_j)| \geq c_1 \mathcal{A}_n.$$

We now show part (ii). Let $K_2 = \inf_{|\theta| \leq (M+1)B} E\{\mathrm{q}_2(Y;\theta)\}$. Similar to part (i), for any $j = 1, \ldots, p_n$ and any $-B < b < B$,

$$\ell_j''(b) \geq K_2 \mathrm{var}(X_j) \geq K_2 \delta^2$$

for some positive constant δ by condition A1. Again by (A.1.1), for any $j = s_n + 1, \ldots, p_n$,

$$|\beta_j^{\mathrm{CR}}| \leq |C_0 \operatorname{cov}(Y, X_j)|/(K_2\delta^2) = O(\mathcal{B}_n). \tag{A.1.2}$$

■

Proof of Theorem 2.3. To prove Theorem 2.3, the following lemma (which is Theorem 1 of Fan and Song (2010)) will be needed.

Lemma A.1. *Consider data $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$ which are n i.i.d. samples of $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ for some space $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} \in \mathbb{R}$. A regression model for $\boldsymbol{X}$ and Y is assumed with loss function $\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$. Let*

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} E\{\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)\}$$

be the population parameter. Assume that $\boldsymbol{\beta}_0$ is an interior point of a sufficiently large, compact and convex set $\mathbb{B} \in \mathbb{R}^d$. Assume the following conditions on the model,

(F1) *The matrix,*

$$I(\boldsymbol{\beta}) = E\Big[\Big\{\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)\Big\}\Big\{\frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)\Big\}^T\Big],$$

exists finitely and is positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$*. Moreover,*

$$\|I(\boldsymbol{\beta})\|_{\mathbb{B}} = \sup_{\boldsymbol{\beta}\in\mathbb{B}, \|\boldsymbol{x}\|=1} \|I(\boldsymbol{\beta})^{1/2}\boldsymbol{x}\|$$

exists.

(F2) *The function* $\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$ *satisfies the Lipschitz property with a positive constant* k_n*,*

$$|\ell(\boldsymbol{x}^T\boldsymbol{\beta}, y) - \ell(\boldsymbol{x}^T\boldsymbol{\beta}', y)|I_n(\boldsymbol{x}, y) \le k_n|\boldsymbol{x}^T\boldsymbol{\beta} - \boldsymbol{x}^T\boldsymbol{\beta}'|I_n(\boldsymbol{x}, y)$$

for any $\boldsymbol{\beta} \in \mathbb{B}$ *and* $\boldsymbol{\beta}' \in \mathbb{B}$*, where* $I_n(\boldsymbol{x}, y) = \mathrm{I}\{(\boldsymbol{x}, y) \in \Omega_n\}$ *with*

$$\Omega_n = \{(\boldsymbol{x}, y) : \|\boldsymbol{x}\|_\infty \le K_n,\ |y| \le K_n^*\}$$

for some sufficiently large positive constants K_n *and* K_n^**. In addition, there exists a sufficiently large constant* C *such that with* $b_n = Ck_nV_n^{-1}(d/n)^{1/2}$ *and* V_n *given in Condition* (F3)*,*

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le b_n} |E[\{\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y) - \ell(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\}\{1 - I_n(\boldsymbol{X}, Y)\}]| \le o(d/n).$$

(F3) *The function $\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y)$ is convex in $\boldsymbol{\beta}$, satisfying*

$$E\{\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y) - \ell(\boldsymbol{X}^T\boldsymbol{\beta}_0, Y)\} \geq V_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2$$

for all $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq b_n$ and some positive constants V_n.

Then for any $t > 0$,

$$P(\sqrt{n}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \geq 16k_n(1+t)/V_n) \leq \exp(-2t^2/K_n^2) + nP(\Omega_n^c).$$

We now prove Theorem 2.3. The main idea is to apply Lemma A.1 by letting $d = 2$, $\boldsymbol{X} = \mathsf{X}_j$, $\boldsymbol{\beta} = \mathbf{b}_j$ and

$$\ell(\boldsymbol{X}^T\boldsymbol{\beta}, Y) = Q(Y, F^{-1}(\mathsf{X}_j^T\mathbf{b}_j)).$$

So we need to show that the condition (F1)-(F3) hold under our assumptions.

For condition (F1),

$$\begin{aligned} I_j(\mathbf{b}_j) &= E\{\mathrm{q}_1^2(Y; \mathsf{X}_j^T\mathbf{b}_j)\mathsf{X}_j\mathsf{X}_j^T\} \\ &= E\Big[\Big\{(Y-\mu_j)\frac{q''(\mu_j)}{F'(\mu_j)}\Big\}^2\mathsf{X}_j\mathsf{X}_j^T\Big], \end{aligned}$$

where $\mu_j = F^{-1}(\mathsf{X}_j^T\mathbf{b}_j)$. By assumption A1 and A2, $I_j(\mathbf{b}_j)$ is bounded and positive definite at $\mathbf{b}_j = \mathbf{b}_j^{\mathrm{CR}}$.

For condition (F3), it suffices to show that $E\{\mathrm{q}_2(Y; \mathsf{x}_j^T\mathbf{b})\mathsf{X}_j\mathsf{X}_j^T\} \geq VI_2$

for some $V > 0$, where I_2 is the 2×2 identity matrix. Since $E(\mathsf{X}_j \mathsf{X}_j^T) = I_2$ and $\mathrm{q}_2(Y; \mathsf{x}_j^T \mathbf{b}) > 0$, we only need to show that $E\{\mathrm{q}_2(Y; \mathsf{x}_j^T \mathbf{b})\} \geq V$ which is followed by

$$E\{\mathrm{q}_2(Y; \mathsf{x}_j^T \mathbf{b})\} \geq P(|Y| \leq K)\xi$$

where K is some sufficiently large positive constant such that $P(|Y| \leq K) > 0$ and $\xi = \inf_{|y| \leq K, |\theta| \leq (M+1)B} \mathrm{q}_2(y; \theta)$.

Lastly for condition (F2), let $K_n = M$ and $K_n^* = \mathcal{A}_n^2 n$ and $V_n = V$. For $(\mathsf{x}, y) \in \Omega_n$, we have that, for any $\mathbf{b} \in \mathbb{B}$ and $\mathbf{b}' \in \mathbb{B}$,

$$\begin{aligned}
& Q(y, F^{-1}(\mathsf{x}^T \mathbf{b})) - Q(y, F^{-1}(\mathsf{x}^T \mathbf{b}')) \\
= \; & \{q(F^{-1}(\mathsf{x}^T \mathbf{b})) - q(F^{-1}(\mathsf{x}^T \mathbf{b}'))\} + y\{q'(F^{-1}(\mathsf{x}^T \mathbf{b})) - q'(F^{-1}(\mathsf{x}^T \mathbf{b}'))\} \\
& -\{F^{-1}(\mathsf{x}^T \mathbf{b}) q'(F^{-1}(\mathsf{x}^T \mathbf{b})) - F^{-1}(\mathsf{x}^T \mathbf{b}') q'(F^{-1}(\mathsf{x}^T \mathbf{b}'))\} \\
= \; & \{f_1(\mathsf{x}^T \mathbf{b}) - f_1(\mathsf{x}^T \mathbf{b}')\} + y\{f_2(\mathsf{x}^T \mathbf{b}) - f_2(\mathsf{x}^T \mathbf{b}')\} - \{f_3(\mathsf{x}^T \mathbf{b}) - f_3(\mathsf{x}^T \mathbf{b}')\},
\end{aligned}$$

where $f_1(t) = q(F^{-1}(t))$, $f_2(t) = q'(F^{-1}(t))$ and $f_3(t) = F^{-1}(t) q'(F^{-1}(t))$. Let $C_1 = \sup_{|t| < (B+1)M} |f_1'(t)|$, $C_2 = \sup_{|t| < (B+1)M} |f_2'(t)|$ and $C_3 = \sup_{|t| < (B+1)M} |f_3'(t)|$. Then

$$|Q(y, F^{-1}(\mathsf{x}^T \mathbf{b})) - Q(y, F^{-1}(\mathsf{x}^T \mathbf{b}'))| I_n(\mathsf{x}, y) \leq (C_1 + K_n^* C_2 + C_3) |\mathsf{x}^T \mathbf{b} - \mathsf{x}^T \mathbf{b}'| I_n(\mathsf{x}, y)$$

which verifies the first part of condition (F2) with $k_n = C_1 + K_n^* C_2 + C_3$.

For the second part of condition (F2), for all $j = 1, \ldots, p_n$,

$$\begin{aligned} & |E[Q(Y, F^{-1}(\mathsf{X}_j^T \mathbf{b})) - Q(Y, F^{-1}(\mathsf{X}_j^T \mathbf{b}_j^{\mathrm{CR}}))\{1 - I_n(\mathsf{X}_j, Y)\}]| \\ \leq & \; E\{(C_1' + |Y|C_2' + C_3')\mathrm{I}(|Y| > K_n^*)\} \\ \leq & \; \sqrt{E\{(C_1' + |Y|C_2' + C_3')^2\}}\sqrt{P(|Y| > K_n^*)} \\ \leq & \; O(\exp(-m_1 \mathcal{A}_n \sqrt{n}/2)) = o(1/n), \end{aligned}$$

where

$$\begin{aligned} C_1' &= \sup_{|t|<(M+1)B} f_1(t) - \inf_{|t|<(M+1)B} f_1(t), \\ C_2' &= \sup_{|t|<(M+1)B} f_2(t) - \inf_{|t|<(M+1)B} f_2(t), \\ C_3' &= \sup_{|t|<(M+1)B} f_3(t) - \inf_{|t|<(M+1)B} f_3(t). \end{aligned}$$

Then by Lemma A.1, for any t and any $j = 1, \ldots, p_n$,

$$P(\sqrt{n}\|\widehat{\mathbf{b}}_j^{\mathrm{CR}} - \mathbf{b}_j^{\mathrm{CR}}\| \geq 16k_n(1+t)/V_n) \leq \exp(-2t^2/K_n^2) + nm_0 \exp(-m_1 K_n^*).$$

Taking $(1 + t) = \mathcal{A}_n \sqrt{n} V_n/(16k_n)$ and noting $K_n^* = \mathcal{A}_n^2 n$ yield

$$P(\|\widehat{\mathbf{b}}_j^{\mathrm{CR}} - \mathbf{b}_j^{\mathrm{CR}}\| \geq \mathcal{A}_n) \leq \exp(-c_2 \mathcal{A}_n^2 n) + nm_0 \exp(-m_1 \mathcal{A}_n^2 n),$$

where c_2 is a suitable positive constant. The desired result follows from using $P(|\widehat{\beta}_j^{\mathrm{CR}} - \beta_j^{\mathrm{CR}}| \geq \mathcal{A}_n) \leq P(\|\widehat{\mathbf{b}}_j^{\mathrm{CR}} - \mathbf{b}_j^{\mathrm{CR}}\| \geq \mathcal{A}_n)$ and Bonferroni inequality. ■

Proof of Theorem 2.6. Define $\boldsymbol{\beta}^{\mathrm{CR}} = (\beta_1^{\mathrm{CR}}, \ldots, \beta_{p_n}^{\mathrm{CR}})^T$. We first prove that

$$\|\boldsymbol{\beta}^{\mathrm{CR}}\|^2 = \sum_{j=1}^{p_n} |\beta_j^{\mathrm{CR}}|^2 = O(\lambda_{\max}(\Sigma)).$$

Let $C_4 = C_0/(K_2 M^2)$. By (A.1.2), for all $j = 1, \ldots, p_n$,

$$\begin{aligned} |\beta_j^{\mathrm{CR}}| &\leq C_4 |\mathrm{cov}(X_j, Y)| \\ &= C_4 |E[\{X_j - E(X_j)\}E(Y \mid \boldsymbol{X})]| \\ &= C_4 |E[\{X_j - E(X_j)\}F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0)]|. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} & |\{X_j - E(X_j)\}\{F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0) - F^{-1}(\beta_{0;0} + E(\boldsymbol{X}^T\boldsymbol{\beta}_0))\}| \\ \leq \quad & C_5 |\{X_j - E(X_j)\}\{\boldsymbol{X} - E(\boldsymbol{X})\}^T\boldsymbol{\beta}_0|, \end{aligned}$$

where $C_5 = \sup_{|t|<B'M+B}(F^{-1})'(t)$. Again, by taking the expectation on both sides and then putting it into the vector form, we have

$$\begin{aligned} & \|E[\{\boldsymbol{X} - E(\boldsymbol{X})\}\{F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0)\}]\|^2 \\ = \quad & \|E[\{\boldsymbol{X} - E(\boldsymbol{X})\}\{F^{-1}(\beta_{0;0} + \boldsymbol{X}^T\boldsymbol{\beta}_0) - F^{-1}(\beta_{0;0} + E(\boldsymbol{X}^T\boldsymbol{\beta}_0))\}]\|^2 \\ \leq \quad & C_5^2 \|E[\{\boldsymbol{X} - E(\boldsymbol{X})\}\{\boldsymbol{X} - E(\boldsymbol{X})\}^T]\boldsymbol{\beta}_0\|^2 = C_5^2 \|\Sigma\boldsymbol{\beta}_0\|^2 \\ \leq \quad & C_5^2 \lambda_{\max}(\Sigma) \|\Sigma^{1/2}\boldsymbol{\beta}_0\|^2. \end{aligned}$$

Since $\|\Sigma^{1/2}\boldsymbol{\beta}_0\|^2 = \mathrm{var}(\boldsymbol{X}^T\boldsymbol{\beta}_0) \leq M'$, it follows that

$$\|\boldsymbol{\beta}^{\mathrm{CR}}\|^2 \leq C_4^2 C_5^2 M' \lambda_{\max}(\Sigma).$$

Finally by the above result, the cardinality of the set $\{j : |\beta_j^{\mathrm{CR}}| > \epsilon \mathcal{A}_n\}$ can not be bigger than $O(\mathcal{A}_n^{-2}\lambda_{\max})$ for any $\epsilon > 0$. The desired result can be easily seen from Theorem 2.3. ■

Proof of Proposition 2.7. The oracle property were obtained by Zhang et al. (2010) for BD estimation when $s_n^5/n \to 0$ and $s_n(p_n - s_n) = o(n)$. If we can prove that when the number $p_n' = |\widehat{\mathcal{M}}|$ of variables selected in the screening step can be set appropriately, the event $\{\mathcal{M}_n \subseteq \widehat{\mathcal{M}}\}$ happens with probability approaching 1, then the conclusion should follow.

By Theorem 2.6, we have

$$P(|\widehat{\mathcal{M}}_{\gamma_n}| \leq O(\mathcal{A}_n^{-2}\lambda_{\max}(\Sigma))) = 1 - o(1).$$

Since we choose p_n' such that $\lambda_{\max}(\Sigma)/(p_n'\mathcal{A}_n^2) = o(1)$, it is equivalent to choose another appropriate $\gamma_n' \leq \gamma_n$. Thus, $\widehat{\mathcal{M}}_{\gamma_n} \subseteq \widehat{\mathcal{M}}$ and by Corollary 2.4,

$$P(\mathcal{M}_n \subseteq \widehat{\mathcal{M}}) \geq P(\mathcal{M}_n \subseteq \widehat{\mathcal{M}}_{\gamma_n}) = 1 - o(1).$$

The oracle property can be expressed as $P(\text{event O}) \to 1$ as $n \to \infty$.

Since $s_n = o(n^{1/5})$ and $p'_n = o(n/s_n)$, we have

$$P(\text{event O} \mid \mathcal{M}_n \subseteq \widehat{\mathcal{M}}) = 1 - o(1).$$

The desired result follows from

$$P(\text{event O}) \geq P(\text{event O} \mid \mathcal{M}_n \subseteq \widehat{\mathcal{M}}) P(\mathcal{M}_n \subseteq \widehat{\mathcal{M}}) = 1 - o(1).$$

■

A.2 Proofs of theorems in Chapter 3

List of notations.

- C: total number of neurons in the ensemble.
- T: total length of the experiment trial.
- n: total number of time bins.
- $\delta = T/n$: the length of each time bin.
- $\tau_k = (t_{k-1}, t_k]$: the start and end time points of the kth time bin.
- $N_c(\tau_k)$: count the number of spikes fired by neuron c during τ_k.
- $N_{1:C}(\tau_{1:k}) = \{N_c(\tau_l) : c = 1, \ldots, C; l = 1, \ldots, k\}$: the spiking history of all neurons up to the time point t_k.
- $\lambda_c(\tau_k \mid \cdot) = E\{N_c(\tau_k) \mid \cdot\}$: the conditional intensity function of neuron c during τ_k.
- $\gamma_{c,0}$: the baseline firing rate of neuron c.
- $\boldsymbol{\gamma}^{(c)} = (\gamma_{c,1}, \ldots, \gamma_{c,P})$: the coefficients of the autoregressive kernel function of neuron c at different lags.
- $\boldsymbol{\gamma}^{(c,i)} = (\gamma_{c,i,1}, \ldots, \gamma_{c,i,Q})$: the coefficients of the coupling kernel function representing the influence of neuron i onto neuron c at different lags.

- $\boldsymbol{X}(\tau_k)$: a M dimensional covariate vector measured during τ_k.
- $\boldsymbol{\beta}_c = (\beta_{c,1}, \ldots, \beta_{c,M})$: the coefficients that correspond to the covariates for neuron c.
- P: the history windows of autoregressive kernel functions.
- Q: the history windows of coupling kernel functions.
- M: the length of covariate vector $\boldsymbol{X}(\tau_k)$.
- $\widetilde{\boldsymbol{\theta}}_c$: the vector form of all parameters for neuron c.
- $\boldsymbol{\theta}_c$: the vector form of all parameters for neuron c excluding the intercept $\gamma_{c,0}$.
- $\widetilde{\boldsymbol{\theta}}_c^*$: the true unknown parameter vector.
- $d = P + (C-1)Q + M$: the total number of covariates
- $\mathcal{M}_F$: the full model, i.e. $\{1, \ldots, d\}$.
- $\mathcal{M}^*$: the true model, i.e. index set of all truly important covariates.
- $\mathcal{C} = \{\mathcal{M} : \mathcal{M}^* \subseteq \mathcal{M},\ \mathcal{M} \subseteq \mathcal{M}_F\}$: the class of correct models.
- $\mathcal{W} = \{\mathcal{M} : \mathcal{M}^* \not\subseteq \mathcal{M},\ \mathcal{M} \subseteq \mathcal{M}_F\}$: the class of wrong models.
- $\ell_c(\widetilde{\boldsymbol{\theta}}_c)$: negative log-likelihood of neuron c.
- $\mathcal{P}_{(\eta_c,\alpha_c)}(\widetilde{\boldsymbol{\theta}}_c)$: structured penalty of $\widetilde{\boldsymbol{\theta}}_c$.

- $L_c(\widetilde{\boldsymbol{\theta}}_c)$: negative log-likelihood of neuron c with the penalty.
- (α_c, η_c): tuning parameters of the penalty.

Further define $\mathsf{Y}_{c,k}$ and $\widetilde{\mathsf{X}}_{c,k}$ as the response and covariate vector of (3.4) at the kth time bin, $k = 1, \ldots, n$. Let $\widetilde{\mathcal{X}_c} = (\widetilde{\mathsf{X}}_{c,1}, \ldots, \widetilde{\mathsf{X}}_{c,n})^T$ denote the whole design matrix. Adopting the BD framework in Zhang *et al.* (2010), the problem of (3.4) is equivalent to the penalized BD estimation with

$$Q(Y, \mu) = \mu - Y \log(\mu) - Y + Y \log(Y)$$

generated by $q(\mu) = \mu - a - \mu\{\log(\mu) - \log(a)\}$, and link function $F(\cdot) = \log(\cdot)$.

Conditions. We have the following conditions in which B_1 and B_2 are sufficiently large constants. Those are not the weakest possible, but serve to facilitate the technical derivations.

B1. The parameter space Θ of $\widetilde{\boldsymbol{\theta}}_c$ is compact in $\mathbb{R}^{d+1}$ and $\|\widetilde{\boldsymbol{\theta}}_c^*\|_1 < B_1$;

B2. For all bins, $N_c(\tau_k)$ are uniformly bounded by B_2;

B3. The minimum eigenvalue of $\Sigma_c = \frac{1}{n}\widetilde{\mathcal{X}_c}\widetilde{\mathcal{X}_c}^T$ are uniformly bounded away from 0, i.e. $\lambda_{\min}(\Sigma_c) > \delta$ for some positive constant $\delta > 0$;

B4. $\sqrt{n/\{C \log(n)\}} \liminf_{n\to\infty} \min_{j\in\mathcal{M}^*} |\theta_{c,j}^*| \to \infty$.

Proof of Theorem 3.1. Since $d = P + (C-1)Q + M$ and P, Q, M are all fixed constants, d and C have the same diverging rate. Let $r_n = \sqrt{d/n}$ and $\widetilde{\mathbf{u}} = (u_0, u_1, \ldots, u_d)^T \in \mathbb{R}^{d+1}$. Similar to Zhang *et al.* (2010), it suffices to show that for any given $\epsilon > 0$, there is a large constant U_ϵ such that, for large n,

$$\mathrm{P}\left\{ \inf_{\|\widetilde{\mathbf{u}}\|=U_\epsilon} L_c(\widetilde{\boldsymbol{\theta}}_c^* + r_n\widetilde{\mathbf{u}}) > L_c(\widetilde{\boldsymbol{\theta}}_c^*) \right\} \geq 1 - \epsilon. \tag{A.2.1}$$

Now we separate the loss function and penalty term by

$$\begin{aligned} L_c(\widetilde{\boldsymbol{\theta}}_c^* + r_n\widetilde{\mathbf{u}}) - L_c(\widetilde{\boldsymbol{\theta}}_c^*) &= \{\ell_c(\widetilde{\boldsymbol{\theta}}_c^* + r_n\widetilde{\mathbf{u}}) - \ell_c(\widetilde{\boldsymbol{\theta}}_c^*)\} \\ &\quad + \{\mathcal{P}_{(\eta_c,\alpha_c)}(\widetilde{\boldsymbol{\theta}}_c^* + r_n\widetilde{\mathbf{u}}) - \mathcal{P}_{(\eta_c,\alpha_c)}(\widetilde{\boldsymbol{\theta}}_c^*)\} \equiv I_1 + I_2. \end{aligned}$$

Further by Taylor expansion of I_1,

$$\begin{aligned} I_1 &= \frac{r_n}{n}\sum_{k=1}^{n}\{\exp(\widetilde{\mathsf{X}}_{c,k}^T\widetilde{\boldsymbol{\theta}}_c^*) - \mathsf{Y}_{c,k}\}\widetilde{\mathsf{X}}_{c,k}^T\widetilde{\mathbf{u}} + \frac{r_n^2}{2n}\sum_{k=1}^{n}\exp(\widetilde{\mathsf{X}}_{c,k}^T\widetilde{\boldsymbol{\theta}}_c^*)(\widetilde{\mathsf{X}}_{c,k}^T\widetilde{\mathbf{u}})^2 \\ &\quad + \frac{r_n^3}{6n}\sum_{k=1}^{n}\exp(\widetilde{\mathsf{X}}_{c,k}^T\widetilde{\boldsymbol{\theta}}_c')(\widetilde{\mathsf{X}}_{c,k}^T\widetilde{\mathbf{u}})^3 = I_{1,1} + I_{1,2} + I_{1,3} \end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}_c'$ is located between $\widetilde{\boldsymbol{\theta}}_c^*$ and $\widetilde{\boldsymbol{\theta}}_c^* + r_n\widetilde{\mathbf{u}}$. Then those terms can be bounded as

$$|I_{1,1}| \leq \left\| \frac{r_n}{n}\sum_{k=1}^{n}\{\exp(\widetilde{\mathsf{X}}_{c,k}^T\widetilde{\boldsymbol{\theta}}_c^*) - \mathsf{Y}_{c,k}\}\widetilde{\mathsf{X}}_{c,k}^T \right\| \|\widetilde{\mathbf{u}}\| = O_P(r_n\sqrt{d/n})U_\epsilon; \tag{A.2.2}$$

$$I_{1,2} = \frac{r_n^2}{2}\widetilde{\mathbf{u}}^T(\frac{1}{n}\mathcal{X}_c^T\Omega\mathcal{X}_c)\widetilde{\mathbf{u}} \geq c_1 r_n^2 U_\epsilon^2; \tag{A.2.3}$$

$$|I_{1,3}| \leq O_P(r_n^3 d^{3/2})U_\epsilon^3, \tag{A.2.4}$$

where $\Omega = \text{diag}(e^{\mathsf{X}_{c,1}^T\widetilde{\boldsymbol{\theta}}_c^*}, \ldots, e^{\mathsf{X}_{c,n}^T\widetilde{\boldsymbol{\theta}}_c^*})$ and c_1 is some positive constant. Here the data are not i.i.d., so we need apply the martingale version of central limit theorem (Theorem 7.4 in Durrett, 2004) to obtain the rate of $I_{1,1}$ in (A.2.2), in which each element of $\sum_{k=1}^n\{\exp(\widetilde{\mathsf{X}}_{c,k}^T\widetilde{\boldsymbol{\theta}}_c^*) - Y_{c,k}\}\widetilde{\mathsf{X}}_{c,k}^T$ is a martingale with bounded increments. (A.2.3) and (A.2.4) follow conditions B1-B3.

For the penalty term,

$$\begin{aligned} I_2 \;\geq\; & -(1-\alpha_c)\eta_c\left\{\sqrt{P}\|r_n\mathbf{u}_{\boldsymbol{\gamma}^{(c)}}\| + \sqrt{Q}\sum_{i\neq c}\|r_n\mathbf{u}_{\boldsymbol{\gamma}^{(c,i)}}\|\right\} \\ & -\alpha_c\eta_c\left\{\|r_n\mathbf{u}_{\boldsymbol{\gamma}^{(c)}}\|_1 + \sum_{i\neq c}\|r_n\mathbf{u}_{\boldsymbol{\gamma}^{(c,i)}}\|_1\right\} - \eta_c\|r_n\mathbf{u}_{\boldsymbol{\beta}_c}\|_1 \end{aligned}$$

where $\mathbf{u}_{\boldsymbol{\gamma}^{(c)}}$, $\mathbf{u}_{\boldsymbol{\gamma}^{(c,i)}}$ and $\mathbf{u}_{\boldsymbol{\beta}_c}$ are components of $\tilde{\mathbf{u}}$ corresponding to $\boldsymbol{\gamma}^{(c)}$, $\boldsymbol{\gamma}^{(c,i)}$ and $\mathbf{u}_{\boldsymbol{\beta}_c}$ respectively. Then,

$$\begin{aligned} |I_2| \;\leq\; & \eta_c r_n\left\{\sqrt{P}\|\mathbf{u}_{\boldsymbol{\gamma}^{(c)}}\| + \sqrt{Q}\sum_{i\neq c}\|\mathbf{u}_{\boldsymbol{\gamma}^{(c,i)}}\| + \sqrt{M}\|\mathbf{u}_{\boldsymbol{\beta}_c}\|\right\} \\ \leq\; & \eta_c r_n\sqrt{\max(P,Q,M)}\sqrt{C+1}\|\tilde{\mathbf{u}}\| = O(\eta_c r_n\sqrt{d})U_\epsilon. \end{aligned}$$

Since $d^4/n = O(C^4/n) = o(1)$, we can choose some large C_ϵ such that $I_{1,1}$, $I_{1,3}$ and I_2 are all dominated by $I_{1,2}$, which is positive. This implies (A.2.1).

■

Proof of Theorem 3.2. Let $\widehat{\widetilde{\boldsymbol{\theta}}}_c^{(\eta_c,\alpha_c)} = (\widehat{\theta}_{c,0}^{(\eta_c,\alpha_c)}, \widehat{\boldsymbol{\theta}}_c^{(\eta_c,\alpha_c)T})^T$ denote the minimizer of (3.4) with tuning parameters (η_c, α_c), and $\mathcal{M}(\widehat{\boldsymbol{\theta}}_c^{(\eta_c,\alpha_c)}) = \{1 \leq j \leq$

$d : \widehat{\theta}_{c,j}^{(\eta_c,\alpha_c)} \neq 0\}$ denote the index set of nonzero elements. Define

$$\begin{aligned}\Lambda_{\mathcal{C}} &= \{(\eta_c,\alpha_c) : \eta_c > 0, \alpha_c \in [0,1], \mathcal{M}(\widehat{\boldsymbol{\theta}}_c^{(\eta_c,\alpha_c)}) \in \mathcal{C}\},\\ \Lambda_{\mathcal{W}} &= \{(\eta_c,\alpha_c) : \eta_c > 0, \alpha_c \in [0,1], \mathcal{M}(\widehat{\boldsymbol{\theta}}_c^{(\eta_c,\alpha_c)}) \in \mathcal{W}\},\end{aligned}$$

which collect all pairs of (η_c, α_c) which produce correct and wrong models respectively. Also define the un-penalized estimator for a given model $\mathcal{M}$ as

$$\widehat{\widetilde{\boldsymbol{\theta}}}_{c,\mathcal{M}}^{(u)} = \underset{\widetilde{\boldsymbol{\theta}}_c:\theta_{c,j}=0,\forall j\notin\mathcal{M}}{\arg\min} \ell_c(\widetilde{\boldsymbol{\theta}}).$$

Remember the associated BIC value is defined as

$$\mathrm{BIC}_c(\eta_c,\alpha_c) = 2\ell_c(\widehat{\widetilde{\boldsymbol{\theta}}}_c^{(\eta_c,\alpha_c)}) + \frac{\log(n)}{n}\mathrm{df}(\widehat{\boldsymbol{\theta}}_c^{(\eta_c,\alpha_c)}).$$

To prove Theorem 3.2, the following lemmas are needed.

Lemma A.2 (Theorem 1 of Zhang (2010))**.** *Assume conditions* B1–B3*. If* $C^4/n \to 0$*, as* $n \to \infty$*, then* $\|\widehat{\widetilde{\boldsymbol{\theta}}}_{c,\mathcal{M}}^{(u)} - \widetilde{\boldsymbol{\theta}}_c^*\| = O_P(\sqrt{C/n})$ *for any* $\mathcal{M} \in \mathcal{C}$*.*

Lemma A.3. *Assume conditions* B1–B3*. If* $C^4/n \to 0$*, as* $n \to \infty$*, then for any* $\sqrt{n/C}$*-consistent estimate* $\widehat{\widetilde{\boldsymbol{\theta}}}_c$*, we have*

$$\ell_c(\widehat{\widetilde{\boldsymbol{\theta}}}_c) - \ell_c(\widetilde{\boldsymbol{\theta}}_c^*) = O_P(C/n). \tag{A.2.5}$$

Proof: Let $\tilde{\mathbf{u}} = \widehat{\widetilde{\boldsymbol{\theta}}}_c - \widetilde{\boldsymbol{\theta}}_c^* = O_P\{\sqrt{(C/n)}\}$. an use the similar Taylor

expansion as in the proof of Theorem 3.1,

$$\ell_c(\widehat{\widetilde{\boldsymbol{\theta}}}_c) - \ell_c(\widetilde{\boldsymbol{\theta}}_c^*) = I_{1,1} + I_{1,2} + I_{1,3}$$

and

$$|I_{1,1}| \leq O_P(d/n),\ I_{1,2} = O_P(d/n),\ |I_{1,3}| \leq O_P(d^3/n^{3/2}).$$

Then (A.2.5) follows. ■

The following proof of Theorem 3.2 includes two parts for handling wrong models and correct models separately.

Part 1. In this part, we will show that with probability tending to 1, any wrong model produced by $(\eta_c, \alpha_c) \in \Lambda_{\mathcal{W}}$ is less preferable than the the un-penalized estimate from full model in terms of BIC values, i.e.

$$\mathrm{P}\left\{\inf_{(\eta_c,\alpha_c)\in\Lambda_{\mathcal{W}}} \mathrm{BIC}_c(\eta_c, \alpha_c) - \mathrm{BIC}_c(0,0) > 0\right\} \to 1. \tag{A.2.6}$$

By Lemma A.3, we have

$$\begin{aligned}
& \inf_{(\eta_c,\alpha_c)\in\Lambda_{\mathcal{W}}} \mathrm{BIC}_c(\eta_c, \alpha_c) - \mathrm{BIC}_c(0,0) \\
= & \inf_{(\eta_c,\alpha_c)\in\Lambda_{\mathcal{W}}} \mathrm{BIC}_c(\eta_c, \alpha_c) - \ell_c(\widehat{\widetilde{\boldsymbol{\theta}}}^{(u)}_{c,\mathcal{M}_F}) - \frac{d\log(n)}{n} \\
= & \min_{\mathcal{M}\in\mathcal{W}} \ell_c(\widehat{\widetilde{\boldsymbol{\beta}}}^{(u)}_{c,\mathcal{M}}) - \ell_c(\widetilde{\boldsymbol{\theta}}_c^*) - \frac{d\log(n)}{n} + O_P(C/n).
\end{aligned}$$

Then again by Taylor's expansion,

$$\ell_c(\widehat{\widetilde{\boldsymbol{\beta}}}^{(u)}_{c,\mathcal{M}}) - \ell_c(\widetilde{\boldsymbol{\theta}}^*_c) \;\; > \;\; c_1 \|\widehat{\widetilde{\boldsymbol{\beta}}}^{(u)}_{c,\mathcal{M}} - \widetilde{\boldsymbol{\theta}}^*_c\|^2 + O_P(\sqrt{C/n}) \|\widehat{\widetilde{\boldsymbol{\beta}}}^{(u)}_{c,\mathcal{M}} - \widetilde{\boldsymbol{\theta}}^*_c\|.$$

By definition, $\min_{\mathcal{M}\in\mathcal{W}} \|\widehat{\widetilde{\boldsymbol{\beta}}}^{(u)}_{c,\mathcal{M}} - \widetilde{\boldsymbol{\theta}}^*_c\| \geq \min_{j\in\mathcal{M}^*} |\theta^*_{c,j}|$, thus

$$\begin{aligned} & \inf_{(\eta_c,\alpha_c)\in\Lambda_{\mathcal{W}}} \mathrm{BIC}_c(\eta_c,\alpha_c) - \mathrm{BIC}_c(0,0) \\ \geq \;\; & c_1 \min_{j\in\mathcal{M}^*} |\theta^*_{c,j}|^2 + O_P(\sqrt{C/n}) \min_{j\in\mathcal{M}^*} |\theta^*_{c,j}| - \frac{d\log(n)}{n} + O_P(C/n) \end{aligned}$$

which is guaranteed to be positive asymptotically as long as condition $\mathrm{B}4$ holds. This implies (A.2.6).

Part 2. For any two $\sqrt{n/C}$-consistent choices of (η_c, α_c) and $(\eta'_c, \alpha'_c) \in \Lambda_{\mathcal{C}}$,3 if $\mathrm{df}(\widehat{\boldsymbol{\theta}}_c^{(\eta_c,\alpha_c)}) < \mathrm{df}(\widehat{\boldsymbol{\theta}}_c^{(\eta'_c,\alpha'_c)})$, then

$$\mathrm{BIC}(\eta'_c,\alpha'_c) - \mathrm{BIC}(\eta_c,\alpha_c) \geq \ell_c(\widehat{\boldsymbol{\theta}}_c^{(\eta_c,\alpha_c)}) - \ell_c(\widehat{\boldsymbol{\theta}}_c^{(\eta'_c,\alpha'_c)}) + \frac{\log(n)}{n} = O_P(C/n) + \frac{\log(n)}{n}.$$

The right-hand side is positive if $C = o\{\log(n)\}$ as $n \to \infty$, which implies

$$\mathrm{P}\{\mathrm{BIC}(\eta'_c,\alpha'_c) - \mathrm{BIC}(\eta_c,\alpha_c) > 0\} \to 1.$$

Thus the model with smaller number of covariates will be selected. ■

Bibliography

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, 96, 6745–6750.

Beck, A. and Teboulle, M. (2009), "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, 2, 183–202.

Berry, T., Hamilton, F., Peixoto, N., and Sauer, T. (2012), "Detecting connectivity changes in neuronal networks," *Journal of Neuroscience Methods*, 209, 388–397.

Bickel, P. J. and Li, B. (2006), "Regularization in statistics," *Test*, 15, 271–303.

Brègman, L. M. (1967), "The relaxation method of finding the common point of convex sets and its application to the solution of problems in con-

vex programming," *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

Brillinger, D. R. (1992), "Nerve cell spike train data analysis: A progression of technique," *Journal of the American Statistical Association*, 87, 260–271.

Candes, E. and Tao, T. (2007), "The Dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, 35, 2313–2351.

Chatterjee, S., Steinhaeuser, K., Banerjee, A., Chatterjee, S., and Ganguly, A. R. (2012), "Sparse group lasso: Consistency and climate applications," *Proceedings of the SIAM International Conference on Data Mining*, 47–58.

Chen, Z., Putrino, D. F., Ghosh, S., Barbieri, R., and Brown, E. N. (2011), "Statistical inference for assessing functional connectivity of neuronal ensembles with sparse spiking data," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 19, 121–135.

Cox, D. R. and Isham, V. (1980), *Point processes*, vol. 12, Chapman and Hall/CRC.

Devilbiss, D. M., Jenison, R. L., and Berridge, C. W. (2012), "Stress-induced impairment of a working memory task: Role of spiking rate and spiking history predicted discharge," *PLoS Computational Biology*, 8, e1002681.

Devilbiss, D. M., Page, M. E., and Waterhouse, B. D. (2006), "Locus ceruleus regulates sensory encoding by neurons and networks in waking animals," *The Journal of Neuroscience*, 26, 9860–9872.

Devilbiss, D. M. and Waterhouse, B. D. (2004), "The effects of tonic locus ceruleus output on sensory-evoked responses of ventral posterior medial thalamic and barrel field cortical neurons in the awake rat," *The Journal of Neuroscience*, 24, 10773–10785.

Durrett, R. (2010), *Probability: theory and examples*, vol. 3, Cambridge university press.

Eldawlatly, S., Zhou, Y., Jin, R., and Oweiss, K. G. (2010), "On the use of dynamic Bayesian networks in reconstructing functional neuronal networks from spike train ensembles," *Neural Computation*, 22, 158–189.

Fan, J. (1997), "Comments on "Wavelets in statistics: A review" by A. Antoniadis," *Journal of the Italian Statistical Society*, 6, 131–138.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Fan, J. and Song, R. (2010), "Sure independence screening in generalized linear models with NP-dimensionality," *The Annals of Statistics*, 38, 3567–3604.

Frank, L. E. and Friedman, J. H. (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109–135.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise coordinate optimization," *The Annals of Applied Statistics*, 1, 302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "A note on the group lasso and a sparse group lasso," *arXiv preprint*, arXiv:1001.0736.

Geng, Z., Wang, S., Yu, M., Monahan, P. O., Champion, V., and Wahba, G. (2013), "Group variable selection via convex Log-Exp-Sum penalty with application to a breast cancer survivor study," *arXiv preprint*, arXiv:1306.4397.

Gerhard, F., Pipa, G., Lima, B., Neuenschwander, S., and Gerstner, W. (2011), "Extraction of network topology from multi-electrode recordings: is there a small-world effect?" *Frontiers in Computational Neuroscience*, 5, 4.

Gerstein, G. L. and Perkel, D. H. (1969), "Simultaneously recorded trains of action potentials: analysis and functional interpretation," *Science*, 164, 828–830.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999), "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286, 531–537.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The elements of statistical learning*, vol. 1, Springer.

Kelly, R. C., Kass, R. E., Smith, M. A., and Lee, T. S. (2010), "Accounting for network effects in neuronal responses using L1 regularized point

process models," *Advance in Neural Information Processing Systems*, 23, 1099–1107.

Kim, J., Kim, Y., and Kim, Y. (2008), "A gradient-based optimization algorithm for LASSO," *Journal of Computational and Graphical Statistics*, 17, 994–1009.

Kim, S., Putrino, D., Ghosh, S., and Brown, E. N. (2011), "A Granger causality measure for point process models of ensemble neural spiking activity," *PLoS Computational Biology*, 7, e1001110.

Liu, J. and Ye, J. (2010), "Moreau-Yosida regularization for grouped tree structure learning," in *Advance in Neural Information Processing Systems*, pp. 1459–1467.

McCullagh, P. (1983), "Quasi-likelihood functions," *The Annals of Statistics*, 11, 59–67.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, vol. 37, Chapman and Hall/CRC.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008), "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 53–71.

Mishchenko, Y., Vogelstein, J. T., and Paninski, L. (2011), "A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data," *The Annals of Applied Statistics*, 5, 1229–1261.

Okatan, M., Wilson, M. A., and Brown, E. N. (2005), "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity," *Neural Computation*, 17, 1927–1961.

Perkel, D. H., Gerstein, G. L., and Moore, G. P. (1967), "Neuronal spike trains and stochastic point processes: II. Simultaneous spike trains," *Biophysical Journal*, 7, 419–440.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. (2008), "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, 454, 995–999.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245.

Stevenson, I. H., Rebesco, J. M., Hatsopoulos, N. G., Haga, Z., Miller, L. E., and Kording, K. P. (2009), "Bayesian inference of functional connectivity and network structure from spikes," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 17, 203–213.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.

Truccolo, W. (2010), "Stochastic models for multivariate neural point pro-

cesses: Collective dynamics and neural decoding," in *Analysis of Parallel Spike Trains*, Springer, pp. 321–341.

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., and Brown, E. N. (2005), "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," *Journal of Neurophysiology*, 93, 1074–1089.

Vapnik, V. (2000), *The nature of statistical learning theory*, Springer.

Wang, S., Nan, B., Zhu, N., and Zhu, J. (2009), "Hierarchically penalized Cox regression with grouped variables," *Biometrika*, 96, 307–322.

Wasserman, L. and Roeder, K. (2009), "High dimensional variable selection," *The Annals of Statistics*, 37, 2178–2201.

Wright, S. J. (2012), "Accelerated block-coordinate relaxation for regularized optimization," *SIAM Journal on Optimization*, 22, 159–186.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zhang, C. (2010), "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, 38, 894–942.

Zhang, C., Jiang, Y., and Chai, Y. (2010), "Penalized Bregman divergence for large-dimensional regression and classification," *Biometrika*, 97, 551–566.

Zhang, C., Jiang, Y., and Shang, Z. (2009), "New aspects of Bregman divergence in regression and classification with parametric and non-parametric estimation," *Canadian Journal of Statistics*, 37, 119–139.

Zhao, M., Batista, A., Cunningham, J. P., Chestek, C., Rivera-Alvidrez, Z., Kalmar, R., Ryu, S., Shenoy, K., and Iyengar, S. (2012), "An L_1-regularized logistic model for detecting short-term neuronal interactions," *Journal of Computational Neuroscience*, 32, 479–497.

Zhu, L., Li, L., Li, R., and Zhu, L. (2011), "Model-Free feature screening for ultrahigh-dimensional data," *Journal of the American Statistical Association*, 106, 1464–1475.