Emerging Topics in Statistics and Biostatistics

Yichuan Zhao
Ding-Geng (Din) Chen
Editors

Statistical Modeling in Biomedical Research

Contemporary Topics and Voices in the Field

EXTRAS ONLINE

Springer

Emerging Topics in Statistics and Biostatistics

More information about this series at http://www.springer.com/series/16213

Yichuan Zhao • Ding-Geng (Din) Chen
Editors

Statistical Modeling in Biomedical Research

Contemporary Topics and Voices in the Field

Springer

Editors
Yichuan Zhao
Math and Statistics, 1342
Georgia State University
Atlanta, GA, USA

Ding-Geng (Din) Chen
School of Social Work
University of North Carolina
Chapel Hill, NC, USA

ISSN 2524-7735 ISSN 2524-7743 (electronic)
Emerging Topics in Statistics and Biostatistics
ISBN 978-3-030-33415-4 ISBN 978-3-030-33416-1 (eBook)
https://doi.org/10.1007/978-3-030-33416-1

Preface

The purpose of this book is to reflect the frontiers of statistical modeling in biomedical research, stimulate new research, and provide great opportunities for further collaborations. We received high-quality papers from distinguished researchers in biostatistics and biomedical research and have invited them to prepare book chapters. Finally, we selected 19 excellent papers for this book. All of the book chapters have been thoroughly peer-reviewed and revised several times before the final publication. This timely volume presents new developments in biomedical research, introduces innovative procedures, presents interesting applications in statistics and biomedical investigations, and contains the potential to impact statistics and biomedical research. This book makes contributions to biomedical studies in the data science era and provides new insights for biomedical researchers, postdocs, graduate students, applied investigators, and industry practitioners.

The 19 chapters are organized into five parts: Part I includes four chapters, which present next-generation sequence data analysis. Part II consists of three chapters on deep learning, precision medicine, and applications. Part III is composed of four chapters that present large-scale data analysis and its applications. Part IV outlines the biomedical research and the modeling. Part V consists of three chapters on survival analysis with complex data structures and its applications. The chapters are organized as self-contained units. In addition, we have included references at the end of each chapter. Furthermore, readers can easily request from us or the chapter authors computer programs or data sets used to facilitate the application of these statistical approaches in practice.

Part I: Next-Generation Sequence Data Analysis (Chapters 1–4)

The chapter, "Modeling Species-Specific Gene Expression Across Multiple Regions in the Brain," presents a new statistical approach for identifying genes with species-specific expression. This new approach avoids multiple pairwise comparisons and

can be susceptible to small changes in expression as well as intransitivity. In this chapter, Diao, Zhu, Sestan, and Zhao show that the proposed model can better identify human-specific genes than the naive approach. The authors also show that the new approach produces more robust gene classifications across regions and greatly reduces the number of human-specific genes.

In the chapter, "Classification of EEG Motion Artifact Signals Using Spatial ICA," Huang, Condor, and Huang proposed a new procedure, which reduces dimension by applying spatial independent component analysis (SICA) and classifies the gait speed for a given subject by the projected EEG motion artifact signals. The authors use SICA and principal component analysis for the dimensionality reduction before applying classifiers such as support vector machines, naïve Bayes, and multinomial logistic regression.

In the chapter, "Weighted K-means Clustering with Observation Weight for Single-Cell Epigenomic Data," Zhang, Wangwu, and Lin develop a weighted K-means algorithm. By down-weighting cells, the authors show that the new algorithm can lead to the improved detection of rare cell types. The authors finally investigated the proposed methods using extensive simulation studies.

In the chapter, "Discrete Multiple Testing in Detecting Differential Methylation Using Sequencing Data," Hao and Lin present the multiple testing issue in detecting differential methylation in next-generation sequencing studies. The existing FDR control procedures are often underpowered in methylation sequencing data analysis due to the discreteness. In this chapter, the authors also discussed several FDR control methods that can accommodate such discreteness.

Part II: Deep Learning, Precision Medicine, and Applications (Chapters 5–7)

The chapter, "Prediction of Functional Markers of Mass Cytometry Data via Deep Learning," presents a novel deep learning architecture for predicting functional markers in the cells given data on surface markers. The proposed approach can automate measurements of functional markers across cell samples, and the proposed procedure demonstrates the improved prediction performance of the deep learning architecture.

In the chapter, "Building Health Application Recommender System Using Partially Penalized Regression," the authors proposed to estimate the optimal policy, which maximizes the expected utility by partial regularization via orthogonality using the adaptive Lasso (PRO-aLasso). The chapter also shows that PRO-aLasso estimators share the same oracle properties as the adaptive Lasso.

In the chapter, "Hierarchical Continuous-Time Hidden Markov Model, with Application in Zero-Inflated Accelerometer Data," Xu, Laber, and Staicu propose a flexible continuous-time hidden Markov model to extract meaningful activity patterns from human accelerometer data and derive an efficient learning algorithm for the proposed model. In this chapter, the authors also develop a bootstrap procedure for interval estimation.

Part III: Large-Scale Data Analysis and Its Applications (Chapters 8–11)

In the chapter, "Privacy-Preserving Feature Selection via Voted Wrapper Method for Horizontally Distributed Medical Data," Lu and Zhang propose a privacy-preserving feature selection method, "privacy-preserving feature selection algorithm via voted wrapper methods (PPFSVW)." The experimental results show that the new algorithm workflow can work effectively to improve classification performance by selecting informative features and genes and can also make the classifier achieve the higher classification accuracy.

The chapter, "Improving Maize Trait Through Modifying Combination of Genes," proposes a computational method for detecting complex traits associated with gene interactions using a combination of gene expression and trait data across a set of maize hybrids. This new method represents changes of expression patterns in a gene pair and employs network topology to describe the inherent genotype–phenotype associations. In this chapter, the authors also investigate the proposed method on several phenotypic traits and achieved consistent results.

In the chapter, "Molecular Basis of Food Classification in Traditional Chinese Medicine," the authors used machine learning methods by using food molecular composition to predict the hot, neutral, or cold label of food, and achieved more than 80% accuracy, which indicated that TCM labels have a significant molecular basis. This research is the first study to quantitatively investigate the relationship between TCM labels and the molecular composition of food.

The chapter, "Discovery Among Binary Biomarkers in Heterogeneous Populations," presents jointly modeled binary and continuous disease outcomes when the association between predictors and these outcomes exhibits heterogeneity. In this chapter, Geng and Slate use ideas from logic regression to find Boolean combinations of these biomarkers and adopt a mixture-of-finite-mixtures fully Bayesian approach to simultaneously estimate the number of subgroups, the subgroup membership structure, and the subgroup-specific relationships between outcomes and predictors.

Part IV: Biomedical Research and the Modeling (Chapters 12–16)

In the chapter, "Heat Kernel Smoothing on Manifolds and Its Application to Hyoid Bone Growth Modeling," Chung, Adluru, and Vorperian propose a unified heat kernel smoothing framework for modeling 3D anatomical surface data extracted from medical images. In this chapter, the authors apply the proposed method in characterizing the 3D growth pattern of human hyoid bone between ages 0 and 20 obtained from CT images. A significant age effect is detected on localized parts of the hyoid bone.

In the chapter, "Optimal Projections in the Distance-Based Statistical Methods," Yu and Huo propose a new way to calculate distance-based statistics, particularly when the data are multivariate. The main idea is to pre-calculate the optimal projection directions given the variable dimension and to project multidimensional variables onto these pre-specified projection directions. In this chapter, the authors also show that the exact solution of the nonconvex optimization problem can be derived in two special cases and propose an algorithm to find some approximate solutions.

The chapter "Kernel Tests for One, Two, and K-Sample Goodness-of-Fit: State of the Art and Implementation Considerations," discusses statistical distances in the goodness-of-fit and reviewed multivariate two-sample goodness-of-fit tests from machine learning point of view. In this chapter, Chen and Markatou introduce a class of one- and two-sample tests constructed using the kernel-based quadratic distance. The implementation of these tests, with emphasis on the kernel-based two-sample test, is provided.

The chapter, "Hierarchical Modeling of the Effect of Pre-exposure Prophylaxis on HIV in the US," centers on the effectiveness of chemical prophylaxis on the populations involved in the HIV epidemic in the US. In this chapter, the authors use a system of nonlinear differential equations to represent the system of populations involved in the HIV epidemic and define model parameters for both the national and the urban case, representing low and high sexual network densities. These results indicate that the undiagnosed high-risk infected group is the largest contributor to the epidemic under both national and urban conditions.

The chapter, "Mathematical Model of Mouse Ventricular Myocytes Overexpressing Adenylyl Cyclase Type 5," studies a new model of transgenic (TG) mouse ventricular myocytes overexpressing adenylyl cyclase type 5. The proposed model describes β_1- and β_2-adrenergic signaling systems very well. In this chapter, Bondarenko finds that the overexpression of AC5 results in an increased basal cAMP production.

Part V: Survival Analysis with Complex Data Structure and Its Applications (Chapters 17–19)

The chapter, "Non-parametric Maximum Likelihood Estimator for Case-Cohort and Nested Case–Control Designs with Competing Risks Data," assumed cause-specific hazards given by the Cox's regression model and provided non-parametric maximum likelihood estimators (NPMLEs) in the nested case–control or case-cohort design with competing risks data. In this chapter, the authors propose an iterative algorithm based on self-consistency equations to compute the NPMLE and established the consistency and asymptotic normality of the proposed estimators.

In the chapter, "Variable Selection in Partially Linear Proportional Hazards Model with Grouped Covariates and a Diverging Number of Parameters," Afzal

and Lu proposed a hierarchical bi-level variable selection approach for censored survival data in the linear part of a partially linear proportional hazards model. The benefit of the proposed method is that it enables us to conduct a simultaneous group selection and individual variable selection within selected groups. The chapter also develops computational algorithms and establishes the selection consistency, and asymptotic normality of the proposed estimators.

The chapter, "Inference of Transition Probabilities in Multi-State Models using Adaptive Inverse Probability Censoring Weighting Technique," develops a model-specific, state-dependent adaptive IPCW (AIPCW) technique for estimating transition probabilities in multi-state models. In this chapter, Zhang and Zhang conduct intensive simulation studies and the results show that the proposed AIPCW procedure improves the accuracy of transition probability estimates compared to the existing SIPCW approach.

The two editors are so grateful to all of the people who have provided the great support for the publication of this book. We deeply thank all the chapter authors (in the "Contributors") for their excellent contributions to this book. We sincerely thank all the reviewers (in the "List of Chapter Reviewers") for their insightful and helpful reviews, which significantly improved the presentation of the book. Moreover, our deep appreciations go to the organizers of the 6th Workshop on Biostatistics and Bioinformatics since several book chapters are based on the presentations in this workshop. Last but not least, our sincere acknowledgments go to the wonderful support of Laura Aileen Briskman (Editor, Statistics, Springer Nature) from Springer New York and Gerlinde Schuster (Editorial Assistant, Statistics, Springer), who made this interesting book publish on time. We look forward to readers' comments on further improvements for the book. Please contact us: Dr. Yichuan Zhao (email: yichuan@gsu.edu) and Dr. Ding-Geng (Din) Chen (email: dinchen@email.unc.edu).

Atlanta, GA, USA — Yichuan Zhao
Chapel Hill, NC, USA — Ding-Geng (Din) Chen

Contents

Contributors

Nagesh Adluru Waisman Laboratory for Brain Imaging and Behavior, University of Wisconsin, Madison, WI, USA

Arfan Raheen Afzal Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada
Tom Baker Cancer Centre, Alberta Health Services, Calgary, AB, Canada

Vladimir E. Bondarenko Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, USA
Neuroscience Institute, Georgia State University, Atlanta, GA, USA

I-Shou Chang Institute of Cancer Research and Division of Biostatistics and Bioinformatics, Institute of Population Health Science, National Health Research Institutes, Zhunan Town, Miaoli County, Taiwan

Yang Chen Department of Biostatistics, University at Buffalo, Buffalo, NY, USA

Yi-Hau Chen Institute of Statistical Science, Academia Sinica, Nankang, Taipei, Taiwan

Yingqing Chen Biostatistics Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Yu Chen Bayer U.S. Crop Science, Monsanto Legal Entity, Chesterfield, MO, USA
Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN, USA

Ken Cheung Columbia University, New York, NY, USA

Moo K. Chung Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

Aubrey Condor Department of Statistics and Data Science, University of Central Florida, Orlando, FL, USA

Renee Dale Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA

Liyang Diao Department of Biostatistics, Yale University, New Haven, CT, USA

Junxian Geng Boehringer Ingelheim Pharmaceuticals Inc., Ridgefield, CT, USA

Xiaosong Han Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China

Guanshengrui Hao Washington University in St. Louis, St. Louis, MO, USA

Hongyu He Math Department, Louisiana State University, Baton Rouge, LA, USA
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Maxwell Hostetter II Georgia State University, Atlanta, GA, USA

Helen J. Huang Department of Mechanical and Aerospace Engineering, University of Central Florida, Orlando, FL, USA

Hsin-Hsiung Huang Department of Statistics and Data Science, University of Central Florida, Orlando, FL, USA

Xiaoming Huo School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Suprateek Kundu Emory University, Atlanta, GA, USA

Eric B. Laber Department of Statistics, North Carolina State University, Raleigh, NC, USA

Yanchun Liang College of Computer Science and Technology, Jilin University, Changchun, China
Zhuhai Laboratory of Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Department of Computer Science and Technology, Zhuhai College of Jilin University, Zhuhai, China

Nan Lin Washington University in St. Louis, St. Louis, MO, USA

Zhixiang Lin Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong

Jingdong Liu Bayer U.S. Crop Science, Monsanto Legal Entity, Chesterfield, MO, USA

Xuewen Lu Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada

Yunmei Lu Department of Computer Science, Georgia State University, Atlanta, GA, USA

Marianthi Markatou Department of Biostatistics, University at Buffalo, Buffalo, NY, USA

Xin Ma Emory University, Atlanta, GA, USA

David C. Mohr Northwestern University, Evanston, IL, USA

Eun Jeong Oh Columbia University, New York, NY, USA

Chun-Hao Pan Unimicron Technology Corporation, Guishan, Taoyuan, Taiwan

Daniel Pimentel-Alarcón Georgia State University, Atlanta, GA, USA

Min Qian Columbia University, New York, NY, USA

Peng Qiu Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Nenad Sestan Department of Neuroscience, Yale University, New Haven, CT, USA

Elizabeth H. Slate Department of Statistics, Florida State University, Tallahassee, FL, USA

Claudia Solís-Lemus Emory University, Atlanta, GA, USA

Ana-Maria Staicu Department of Statistics, North Carolina State University, Raleigh, NC, USA

Houri K. Vorperian Vocal Tract Development Laboratory, Waisman Center, University of Wisconsin, Madison, WI, USA

Duolin Wang Department of Electric Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

Jie-Huei Wang Department of Statistics, Feng Chia University, Taichung, Taiwan

Juexin Wang Department of Electric Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

Jiaxuan Wangwu Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong

Dong Xu Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA
Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

Hao Xu Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China

Zekun Xu Department of Statistics, North Carolina State University, Raleigh, NC, USA

Sean Yang Bayer U.S. Crop Science, Monsanto Legal Entity, Chesterfield, MO, USA

Yun Yang Philocafe, San Jose, CA, USA

Chuanping Yu School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Qin Zeng Bayer U.S. Crop Science, Monsanto Legal Entity, Chesterfield, MO, USA

Mei-Jie Zhang Medical College of Wisconsin, Milwaukee, WI, USA

Wenyu Zhang Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong

Yanqing Zhang Department of Computer Science, Georgia State University, Atlanta, GA, USA

Ying Zhang Merck & Co., Kenilworth, NJ, USA

Haiyan Zhao Centre for Artificial Intelligence, FEIT, University of Technology Sydney (UTS), Broadway, NSW, Australia

Hongyu Zhao Department of Biostatistics, Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

Ying Zhu Department of Biostatistics, Yale University, New Haven, CT, USA
Department of Neuroscience, Yale University, New Haven, CT, USA

Part I
Next Generation Sequence Data Analysis

Modeling Species Specific Gene Expression Across Multiple Regions in the Brain

Liyang Diao, Ying Zhu, Nenad Sestan, and Hongyu Zhao

Abstract *Motivation*: The question of what makes the human brain functionally different from that of other closely related primates, such as the chimpanzee, has both philosophical as well as practical implications. One of the challenges faced with such studies, however, is the small sample size available. Furthermore, expression values for multiple brain regions have an inherent structure that is generally ignored in published studies.

Results: We present a new statistical approach to identify genes with species specific expression, that (1) avoids multiple pairwise comparisons, which can be susceptible to small changes in expression as well as intransitivity, and (2) pools information across related data sets when available to produce more robust results,

Electronic Supplementary Material The online version of this chapter (https://doi.org/10.1007/978-3-030-33416-1_1) contains supplementary material, which is available to authorized users.

Availability and Implementation: Code for estimating the Markov random field parameters and obtaining posterior probabilities for the MRF can be found in the data package attached. All code is written in R.

L. Diao
Department of Biostatistics, Yale University, New Haven, CT, USA

Y. Zhu
Department of Biostatistics, Yale University, New Haven, CT, USA

Department of Neuroscience, Yale University, New Haven, CT, USA
e-mail: ying.zhu@yale.edu

N. Sestan
Department of Neuroscience, Yale University, New Haven, CT, USA
e-mail: nenad.sestan@yale.edu

H. Zhao (✉)
Department of Biostatistics, Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
e-mail: hongyu.zhao@yale.edu

Y. Zhao, D.-G. (Din) Chen (eds.), *Statistical Modeling in Biomedical Research*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-030-33416-1_1

such as in the case of gene expression across multiple brain regions. We demonstrate through simulations that our model can much better identify human specific genes than the naive approach. Applications of the model to two previously published data sets, one microarray and one RNA-Seq, suggest a moderately large benefit from our model. We show that our approach produces more robust gene classifications across regions, and greatly reduces the number of human specific genes previously reported, which we show were primarily due to the noise in the underlying data.

Keywords Gene expression · R code · Posterior probabilities · Markov random field · RNA sequencing · Akaike · Bayes

1 Introduction

As the human genome was first being sequenced, a natural question began to emerge: Can we determine what parts of our genomes differentiate us from our closest primate relatives? The origins of characteristic human abilities, such as speech, social behaviors, and abstract thinking, might be uncovered by comparing the genomes of humans, chimpanzees, and outgroups such as gorillas and macaques. Beyond the questions of our innate "human-ness", comparisons of the genomic differences between humans and other great apes have potentially wide-ranging practical effects: see [27] for an extensive collection of possible phenotypic comparisons of interest, ranging from differences in female reproductive biology, to brain size, to control of fire, and to usage of toys and weapons. In addition to observational phenotypic differences, the authors also note widely different incidence rates for certain diseases in humans and chimpanzees which have long been known. Diseases such as the progression from HIV to AIDS, infection by *P. falciparum* malaria, and occurrence of myocardial infarction, for example, are common in humans yet very rare in the great apes. Alzheimer's disease is a neurodegenerative disease characterized by the presence of amyloid plaques and neurofibrillary tangles in the brain, resulting in memory loss, dementia, and eventually death. While the diagnosis of these symptoms in primates may be difficult, one comparison that can be made is in age-matched dissections of human and primate brains. In such studies, human brains show development of these signature plaques as well as the neurofibrillary tangles, whereas chimpanzee brains show neither [26].

There are many approaches to finding the differences between human and primate genomics, several of which are delineated in [27]. We could analyze various kinds of genomic differences, such as indels, chromosomal changes, gene duplications, and repetitive element insertions. In this manuscript we will focus on differences in gene expression as measured by microarray and RNA-Seq technologies, which have been used in several studies [4, 9, 13, 14]. In the first study [9], only a single region of the brain, the left prefrontal lobe (Brodmann area 9) was analyzed. However, all subsequent studies have sampled at least three regions of the brain. In these studies, the analysis was conducted by performing pairwise comparisons and setting a cutoff for whether a gene has human specific gene expression or not in each region.

There are some issues with this straightforward approach, one of which is due to issues arising with pairwise comparisons: When all three pairwise comparisons are performed among three species, for example, intransitive results can easily result. When pooled pairwise comparisons are performed, i.e. by pooling two species and comparing the pooled species against the remainder, results are highly subject to slight changes in expression, as demonstrated in the following. The second issue is related to the structure of the data. Namely, while there are sure to be genes with differential expression patterns across brain regions, we expect that most genes do not. Thus, instead of analyzing each region separately, we should be able to use the gene expression in other regions to inform the analysis of a given region. Particularly in the case of primate studies, samples are difficult to attain, so sample sizes tend to be very limited. By pooling information, we can obtain more robust estimates of differential gene expression.

In this manuscript, we propose a method which overcomes the two issues described above. The problem of intransitivity in pairwise comparisons is well known, and examples are detailed in [6, 7]. We follow the author's suggestion here and propose an information criterion based model selection approach, testing various information criteria for which performs best for small sample sizes. An additional benefit of using an information criterion is that it produces a relative class membership probability for each gene, for each class. This enables us to use a Markov random field (MRF) to "smooth" assigned class memberships across brain regions, so that in regions with less certainty, we can use information from neighboring regions to inform the decision.

We demonstrate through simulations that the Bayesian information criterion (BIC) performs best for small sample sizes, and that the addition of the MRF can significantly reduce the number of classification errors when the neighbor effect is moderate, particularly for those genes with high variance. We then apply our method to two recently published brain expression data sets, one microarray and one RNA-Seq [14]. In these data, three brain regions were sampled: the caudate nucleus (CN), frontal pole (FP), and hippocampus (HP). We find evidence of a moderate neighbor effect among the three regions, and demonstrate that the Markov random field is able to reduce the number of incorrect classifications compared to the naive approach. Among the top genes we identified as being human specific include those associated with various neurological disorders and neural function, which we did not find using the naive ANOVA approach described in the original paper.

2 Methods

2.1 Overview

2.2 Use of Information Criterion for Model Selection

For each gene, we determine the appropriate latent model based on its expression levels. The latent models are described in Table 1. For this model selection

Table 1 Description of latent classes for a three species comparison

Model	Description
M_1	$Hu = Ch = Ma$
M_2	$Ma \neq Hu = Ch$
M_3	$Hu \neq Ma = Ch$
M_4	$Ch \neq Hu = Ma$
M_5	$Hu \neq Ch \neq Ma$

step, we evaluate the performance of three types of information criteria: Akaike's information criterion (AIC), the small sample size corrected version of the AIC (AICc), and the Bayesian information criterion (BIC) [1, 11, 22]. These are given in Eqs. (1)–(3), where L is the likelihood of the model, k is the number of parameters in the model, and n is the sample size.

$$AIC = -2 \cdot \ln(L) + 2 \cdot k \tag{1}$$

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{2}$$

$$BIC = -2 \cdot \ln L + k \cdot \ln(n) \tag{3}$$

Let $\mathbf{I} = (I_1, \ldots, I_5)$ be the vector of information criteria calculated for each of the 5 models, for a particular gene g. Then the probability of g belonging to model i is given by

$$p_i = \frac{1}{W} \exp\left(0.5\,(min(\mathbf{I}) - I_i)\right) \tag{4}$$

where W is the normalizing constant $W = \sum_i I_i$.

We choose to use information criteria as a natural approach to performing model selection. In particular, we choose to perform model selection in lieu of multiple pairwise comparisons because the latter can often result in intransitivity. For example, we may find that $B > A$, $A > C$, and yet $B = C$. With model selection, such a nonsensical result is not possible. Use of model selection was advocated in [6, 7], which also extensively pointed out the problem of intransitive decisions.

2.3 Estimating Prior Probabilties of Class Membership

For the information criteria described above, we must first determine which models to use for microarray and RNA-Seq data types. Differential expression testing for microarray data has often been carried out using the t-test, but this can be problematic particularly when sample sizes are small, as variance estimates become unstable. Several methods have attempted to pool information across multiple genes in order to better model the variance [8, 10, 24]. Jeanmougin et al. [12] found in

a comparison of multiple methods that appropriate modeling of the variance can significantly reduce the number of false positives found.

Here we assume that the data follow a Gaussian distribution after appropriate normalization just as in the standard t-test, and do not perform any moderated estimates of variance. We therefore simply estimate standard deviations of the microarray data. In simulations we find that both the unmoderated t-test and the information criteria perform relatively well under reasonable variances.

RNA-Seq data are widely modeled according to the negative binomial model [18, 19]. We use `DESeq2` [18] to estimate the mean and dispersion parameters of the negative binomial model for the RNA-Seq data.

Let y_g be the vector of normalized microarray expression values for a given gene g, where we drop the subscript g for clarity. In the remainder of this manuscript, we adopt the following shorthand when referencing the three species, human, chimp, and macaque: Hu, Ch, and Ma, respectively.

Then let y_{Hu}, y_{Ch}, and y_{Ma} denote vectors of expression values for human, chimp, and macaque samples, respectively, for the given gene. We estimate the means μ and standard deviations σ (likewise dispersions ϕ) for each of seven species groupings separately. i.e., μ_{Ma} is the mean expression value for macaque samples, σ_{Hu} the standard deviation of human samples, $\mu_{Hu,Ch}$ the mean of the pooled human and chimp samples, and so on. The seven species groupings are Ma, Hu, Ch, $\{Ma, Hu\}$, $\{Ma, Ch\}$, $\{Hu, Ch\}$, $\{Ma, Hu, Ch\}$.

Then the model likelihood can be computed as:

$$
\begin{aligned}
P(y|M_1) &= P(y|\mu, \sigma) \\
P(y|M_2) &= P(y_{Ma}|\mu_{Ma}, \sigma_{Ma})\, P(y_{Hu,Ch}|\mu_{Hu,Ch}, \sigma_{Hu,Ch}) \\
P(y|M_3) &= P(y_{Hu}|\mu_{Hu}, \sigma_{Hu})\, P(y_{Ch,Ma}|\mu_{Ch,Ma}, \sigma_{Ch,Ma}) \\
P(y|M_4) &= P(y_{Ch}|\mu_{Ch}, \sigma_{Ch})\, P(y_{Hu,Ma}|\mu_{Hu,Ma}, \sigma_{Hu,Ma}) \\
P(y|M_5) &= P(y_{Ma}|\mu_{Ma}, \sigma_{Ma})\, P(y_{Hu}|\mu_{Hu}, \sigma_{Hu})\, P(y_{Ch}|\mu_{Ch}, \sigma_{Ch}) \qquad (5)
\end{aligned}
$$

Here P indicates Gaussian densities, e.g., the probability of observing microarray values y_{Ma} for macaque samples, given mean and standard deviations μ_{Ma} and σ_{Ma}, respectively. We can obtain similar model likelihoods for the RNA-Seq data, with mean and dispersion parameters estimated using `DESeq2`, and based on negative binomial probability densities. The model is specified as:

$$y \sim NB(\text{mean} = \mu, \text{dispersion} = \alpha) \qquad (6)$$

$$\mu = sq \qquad (7)$$

$$\log(q) = \sum_r x_j \beta_j \qquad (8)$$

s is a factor unique to each sample that accounts for differences in library size among samples, x_j are covariates (species, batch effects, etc.), and β_j are the corresponding

coefficients. For the [14] RNA-Seq data, no batch effects were identified, so that $\log(q)$ simply equals β for the species grouping.

2.4 Empirical Bayes Shrinkage Priors for Negative Binomial Model

For the negative binomial model, we find that using the `DESeq2` estimate of μ can result in overfitting when the absolute count values are small, thus leading to genes with low expression and/or high variability being ranked as highly species specific. To avoid this, we propose a shrunken mean model, which is derived from the interpretation of the negative binomial distribution as a hierarchical gamma-Poisson mixture.

In particular, we assume that the counts for each gene g arise from a Poisson distribution, whose mean itself is gamma distributed:

$$p(\mathbf{k_g}; \lambda_g) = \prod_{i=1}^{n} \frac{e^{-\lambda_g} \lambda_g^{k_{gi}}}{k_{gi}!} \tag{9}$$

$$\lambda_g \sim \Gamma(\alpha_g, \beta_g) \tag{10}$$

Here n is the number of samples. We drop the subscript g in the following for clarity, understanding that we are calculating the posterior mean for a particular gene g. Then the posterior mean of $\lambda = \lambda_g$ takes on the form

$$\begin{aligned} \hat{\lambda} &= \frac{n}{n+\beta}\left(\frac{\sum k_i}{n}\right) + \frac{\beta}{n+\beta}\left(\frac{\alpha}{\beta}\right) \\ &= \frac{n\mu}{n\mu+\alpha}\left(\frac{\sum k_i}{n}\right) + \frac{\alpha}{n\mu+\alpha}\mu \end{aligned} \tag{11}$$

We get Eq. (11) by noting that the mean of $\Gamma(\alpha, \beta)$ is $\mu = \alpha/\beta$ and the dispersion parameter of the negative binomial is the same α in $\Gamma(\alpha, \beta)$. Thus, when the mean μ and/or sample size n is large with respect to the dispersion α, $\hat{\lambda}$ is shrunken towards the average count value, whereas if the dispersion parameter is large, it is shrunken towards the mean of the underlying Γ.

In practice, we must obtain $\hat{\lambda}$ while taking into consideration differences in library size among samples. To do this, we take the k_i above to be $k_i^* = k_i/s_i$, where s_i is the size factor for sample i.

2.5 Leveraging Gene Expression Profiles over Several Brain Regions Using a Markov Random Field

While different regions of the brain may have different expression patterns, in general we find that the correlation of gene expressions across different regions is very high (see Fig. S2). Further, when sample sizes are small, robust estimates of model parameters can be difficult to obtain, even when borrowing information across genes as both `limma` and `DESeq2` do. We propose to address this issue by utilizing prior model probabilities in neighboring regions to obtain more stable posterior model probabilities.

To do so, assume that the underlying "true" states of the genes are an instantiation of a locally dependent Markov random field (MRF) [3]. Let $z_{g,r}$ denote the unknown true model membership of gene g in region r, $z_{g,r} \in \{M_1, \cdots, M_5\}$. Intuitively, if $z_{g,r_1} = M_2$, then we are more likely to believe that $z_{g,r_2} = M_2$ as well, for regions $r_1, r_2 \in R$. Under this model, only the neighboring regions of R, $R \setminus \{r\}$, have an effect on $z_{g,r}$. We will assume that all brain regions are thus neighbors of each other.

Generally speaking, the issue of finding the most likely Z,

$$Pr(Z|Y) \propto l(Y|Z)Pr(Z) \tag{12}$$

is extremely difficult. We use the simulated field approximation proposed in [5], which produces a solution via the expectation-maximization (EM) algorithm, and which the authors showed performed favorably compared to other approaches.

Let the conditional probability $p(z_{gr} = M_i|R \setminus \{r\})$ be

$$p(z_{g,r} = M_i|V \setminus \{gr\}) \propto exp\left\{\alpha_i + \beta \sum_{r' \in R\setminus\{r\}} I_{M_i}(z_{g,r'})\right\} \tag{13}$$

where $I_{M_i}(z_i)$ is an indicator variable, such that $I_{M_i}(z_i) = 1$ if $z_i = M_i$ and 0 otherwise.

Thus we see that the probability of model membership is proportional to the number of neighbors belonging to the same model. The strength of this "neighbor effect" is given by β. In total we have five parameters that need to be estimated, denoted by $\Theta = \{\alpha_1^*, \alpha_2^*, \alpha_4^*, \alpha_5^*, \beta\}$ (here we have taken $\alpha_i^* = \alpha_i - \alpha_3$ to avoid identifiability issues).

The steps of the simulated field algorithm are as follows:

1. Initialization:
   (a) Set the initial parameters Θ.
   (b) Obtain an initial estimate of the models Z. These are the states corresponding to maximum relative BIC prior probabilities.

2. For each gene g and region r:
   (a) Calculate the model probability, for each model, of $z_{g,r}$.
   (b) Sample $z^*_{g,r}$ accordingly.
   (c) Move to the next region and/or gene and repeat.
3. Once we have obtained Z^*, an updated state matrix for all genes and regions, re-estimate the parameters Θ.
4. Repeat 2–3 until convergence.

We update the parameters Θ using the Newton Raphson method.

After obtaining estimates of the MRF parameters, we can obtain the posterior model membership probabilities using Markov chain Monte Carlo (MCMC).

2.6 Simulations

We perform three types of simulations to evaluate each of the following:

1. Best information criterion to use for model selection
2. Accuracy of estimation of Markov random field parameters using the simulated field algorithm
3. Reduction of classification errors due to implementation of Markov random field

In each set of simulations, we generate gene expression values based on the gene's classification into one of five latent models, listed in Table 1. We generate expression values for three simulated "species" according to a Gaussian distribution, with species means given in Table 2. We tested three values of σ: 0.15, 0.25, and 0.5. For each species, we simulate five samples, comparable to the number of samples present in the Konopka experimental data. Simulations by both the DESeq [2], DESeq2 [18], and a similar method edgeR [21] have shown that these negative binomial approaches model the variances well. Thus, we will assume that the parameter estimates produced by DESeq2 of the means and dispersions are reasonable, and so evaluation of the information criteria on Gaussian simulated data is sufficient.

Table 2 Gaussian simulation parameters

Model	μ_{Ch}	μ_{Hu}	μ_{Ma}
M_1	2	2	2
M_2	2	2	2.5
M_3	2	2.5	2
M_4	2.5	2	2
M_5	1.5	2	2.5

We test σ = 0.15, 0.25, and 0.5

2.6.1 Selection of Information Criterion

We test three different information criteria to see how well they classify genes belonging to each of the five models M_i: the AIC, AICc, and BIC. Additionally, although they are not directly comparable, we use two types of t-tests as a benchmark against which to compare the information criteria: the pairwise t-test as well as the pooled t-test. In the pairwise t-test, a gene is determined to be human specific if the comparisons Hu vs. Ch and Hu vs. Ma are both significant, while Ch vs. Ma is not significant (at $p = 0.05$). Species specificity for the other two species is similarly defined. Note that in some cases a gene will not be classifiable by this method.

In the pooled t-test, a gene is determined to be human specific if the comparison Hu vs. Ch, Ma is significant. If more than one such comparison is significant, then the mean difference between Hu and Ch, Ma must be larger than the difference in means of the other comparison in order for the gene to be declared species specific. This is similar to the approach taken by [14].

Performance is assessed according to the percentage of misclassified genes, which we call the classification error. For the information criterion approaches, we choose M_i with the highest probability. We perform 100 such simulations.

2.6.2 Estimation of Markov Random Field Parameters

To determine how well we are able to estimate the true parameters Θ of an MRF, we simulate the latent models according to an MRF model, then generate gene expression values as before, and see if we can recover Θ.

The latent class matrix Z of $G = 1000$ genes by $R = 3$ regions is generated as follows: first, we randomly assign to each z_{gr} one of the five classes $M_1, \cdots, M_5$. Then, given MRF parameters $\Theta = \{\alpha_1^*, \alpha_2^*, \alpha_4^*, \alpha_5^*, \beta\}$, we update each element of the Z matrix according to the probability given in Eq. (13). We perform five complete steps of updating to obtain the final Z matrix. In practice, very few steps are required for the Z matrix to converge.

We take $\boldsymbol{\alpha} = (0.8, 0.3, 0.1, -0.1)$ and let β vary as one of $(1, 1.5, 2)$.

From these simulations, we can determine how well Θ is estimated, given (a) the underlying Gaussians are known, and (b) the prior probabilities are determined using the BIC. The former gives us an indication of how well the simulated field algorithm is able to estimate the MRF parameters when the prior probabilities are "exact"; the latter introduces noise from "inexact" priors. We calculate "exact" priors by taking, for each M_i, the probability of observing the values x given the known μ_i and σ_i corresponding to M_i. i.e.,

$$p(M_i) = \frac{1}{W} \prod_x p(x|\mu_i, \sigma_i) \tag{14}$$

where W is a normalizing constant. In theory it is possible that $I_i = I_j$ for $i \neq j$, though we did not observe this in practice. In such cases of tied values of information criteria, we can divide

We run 20 steps of the simulated field algorithm for parameter estimation, and 50 steps of burnin and 100 steps of sampling for posterior probability estimation. We perform 20 simulations for each value of β.

2.6.3 Improvement in Performance Due to Markov Random Field

In the previous section all expression values are generated from Gaussians with the same variance, and thus all are equally noisy. Conceivably, when data are very noisy to begin with, borrowing information from neighbors does not improve predicted classification. However, in the case where some genes are less noisy than others, we may expect to observe an improvement.

To evaluate this effect, we perform a similar set of simulations as in Sect. 2.6.2, but this time randomly selected genes to have different variances. Out of a total of $G \times R = 1000 \times 3 = 3000$ genes, we let 50% be generated from a Gaussian with means given in Table 2 and $\sigma = 0.15$, 30% be generated with $\sigma = 0.25$, and the remaining 20% be generated with $\sigma = 0.5$. All other simulation parameters follow those in Sect. 2.6.2.

2.7 Experimental Data

We analyze two data sets published in [14]. The authors collected samples from three regions in human, chimp, and macaque brains, and compared their expression patterns using two microarray platforms (Affymetrix and Illumina) as well as next-generation sequencing (NGS). Since the authors found in their original manuscript that the Affymetrix arrays were able to capture more genes than the Illumina arrays, here we focus our analysis on the Affymetrix microarray and Illumina NGS data.

We downloaded the log transformed and quantile normalized microarray data deposited at the NCBI Gene Expression Omnibus (GEO) under accession number GSE33010. For genes corresponding to more than one probe, we took the maximum value over all probes. We mapped probes to their appropriate gene symbols from the downloaded `.soft` file. The RNA-Seq expression counts table was downloaded from GEO under accession number GSE33587. Only genes that were uniquely aligned to the genome were retained.

Konopka et al. [14] filtered both microarray and RNA-Seq data to retain only those genes which they deemed "present". For the Affymetrix microarray data, they defined such genes to be those which had a detection score of 0.05 or less in all samples, for each region and species. For RNA-Seq data, a gene was considered "present" if for each individual of a species and in a brain region, at least 2 reads

aligned to the gene. Additional details of the processing steps can be found in the Supplementary Experimental Procedures of the original manuscript.

Here we did not filter the RNA-Seq genes for "presence", or perform additional filtering of the microarray data. In total, we retained 18,458 genes in the microarray data and 16,036 in the RNA-Seq data. We used the same set of genes for analysis of each brain region.

3 Results

3.1 Simulation Results

3.1.1 BIC Produces Best Classifications Overall Under a Variety of Different Scenarios and Parameters

We evaluate five different criteria for model selection: the three information criteria (AIC, AICc, and BIC), as well as the pooled and pairwise t tests. The t tests we use here only as a benchmark with which to compare the information criteria, because such tests do not produce relative model likelihoods and thus are not useful for the Markov random field part of our model.

We noticed marked differences in classification error depending on the criterion used and σ (see Figs. S3, S4, and S5). Additionally, some criteria are better at distinguishing particular classes of M_i than others.

Unsurprisingly, all classifiers perform best for $\sigma = 0.15$, and poorly for $\sigma = 0.5$. In comparison, half of the estimated σ over all genes and all models fitted in Table 1 were less than or equal to 0.2 for each of the three regions, with 95% of the estimated σ being less than 0.61, suggesting that the classifiers should perform relatively well. In particular, we find that the pooled and pairwise t test classifiers make virtually no errors for $\sigma = 0.15$ for any M_i. However, as σ increases, we find that the pairwise t test classifier performs substantially worse than the others at differentiating genes which are species specific to at least one species. The pooled t test classifier has among the lowest classification error rates across all σ tested and for all M_i except M_5, which it is unable to distinguish.

Of the three information criteria, the AIC incurs the most errors at detecting nonspecific genes, and AICc is significantly worse at detecting genes of class M_5. Overall, the BIC model selection criterion maintains relatively low error across the five classes compared to the other methods under the various σ, with error rates similar to that of the pooled t test classifier. Thus, we choose to use the BIC to determine prior probabilities for use in our downstream analyses. However, the best choice may change on a case to case basis, with considerations for sample size. In our particular case where the sample size is very small for each region and species, the BIC appears to perform the best.

3.1.2 Estimation of Markov Random Field Parameters is Precise for Exact Priors

We perform extensive simulations to study how well we are able to correctly estimate the parameters of the MRF using the simulated field algorithm, both for exact and BIC priors (see Sect. 2.6.2).

The value of β and most values of $\boldsymbol{\alpha}$ are well estimated for exact priors (Fig. S6). For larger values of β, we observe higher estimation error for α_i. If we fix $\boldsymbol{\alpha} = \mathbf{0}$, the estimate of β does not suffer. In fact, we surprisingly observe a greater decrease in classification error of the MRF model compared to the naive prior model when we hold $\boldsymbol{\alpha}$ fixed, compared to using the estimated $\boldsymbol{\alpha}$. This is reasonable if we consider that, in the context of Eq. (13), the magnitude of α_i indicates the prior likelihood a gene is classified as class M_i in the absence of any other information. Thus, incorrectly estimated values of α_i may skew the results. However, if we fix $\boldsymbol{\alpha} = \mathbf{0}$, we are essentially only using the classifications of neighboring regions and not these prior beliefs to influence $p(z_{g,r} = M_i)$.

Overall the classification errors under the exact model are extremely low for $\sigma = 0.15$ and $\sigma = 0.25$, being less than 5%. However, when σ is increased to 0.5, the classification error jumps to nearly 30%. In the latter case, the MRF model can significantly reduce the error, even for moderate values of β (Fig. S9). For $\beta = 1$, 1.5, and 2, the reduction in classification error from nearly 30% is to 21%, 16%, and 11%, respectively.

The picture is less clear when using BIC priors. Notably, parameter estimates become significantly worse at $\sigma = 0.5$. This is unsurprising, as we saw in our previous simulations comparing the different information criteria and the pairwise classifier that all methods make many classification errors when σ is large. Thus, when priors are very noisy, MRF parameter estimation is poor. In both cases of $\boldsymbol{\alpha}$ free and fixed, we find a small reduction in classification errors for $\sigma = 0.25$ (Fig. S10). However, when $\sigma = 0.5$, we find no improvement for $\boldsymbol{\alpha}$ fixed, and in fact find that the MRF model performs slightly worse for $\boldsymbol{\alpha}$ free.

Parameter estimation and classification results for AICc priors are given in Figs. S8 and S11, respectively, to demonstrate that noisier priors produce poorer results: Here we find no improvement from the MRF model for $\sigma = 0.5$.

3.1.3 Markov Random Field Can Significantly Improve Classification Errors When Some Neighboring Genes Have Smaller Variance

In the previous section, we find that when the priors are noisy, the MRF model yields little improvement. However, that is under the assumption that all gene expression values arise from distributions with the same high variance. In reality this is not always the case, as a few randomly selected genes demonstrate (Fig. S12). It is thus reasonable to ask if the MRF can improve the classification of such genes, which have high variability in one region but lower variability in others.