AN ABSTRACT OF THE THESIS OF

Liqiang He for the degree of Master of Science in Computer Science presented on June 9, 2017.

Title: Species Distribution Modeling of Citizen Science Data as A Classification Problem with Class-Conditional Label Noise

Abstract approved: _____

Rebecca A. Hutchinson

Species distribution models (SDM), which quantify the correlation between the distribution of a species and environmental factors, are increasingly used to map and monitor animal and plant distributions in the context of awareness of environmental change and its ecological consequence. For perfect data, this is a straightforward classification problem from environmental features to presence or absence labels. But for imperfect data, such as the citizen science data from eBird, in which volunteers report locations where they observed or failed to observe sets of species, mistakes will cause label noise. In this case, both the class features and the observation features would be sources of false positive noise and false negative noise. However, few common modeling approaches for this task address these sources of noise explicitly. In this work, I explore the idea of treating this problem as a classification problem with class-conditional label noise. By leveraging additional information about observation features, this model outperforms other candidates significantly when sufficient data is available. I describe the conditions under which the parameters of my proposed model are identifiable, explore the impact of model misspecification, and apply this model to simulated data and real data from the eBird citizen science project.

Species Distribution Modeling of Citizen Science Data as A Classification Problem with Class-Conditional Label Noise

by

Liqiang He

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented June 9, 2017
Commencement June 2017

Master of Science thesis of Liqiang He presented on June 9, 2017.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Liqiang He, Author

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Rebecca A. Hutchinson for her guidance, understanding, encouragement, her friendship and everything she provided to me throughout this work. In addition, my sincere gratitude to Dr. Sarah C. Emerson for the guidance and collaboration for my research.

I would like to thank my thesis committee: Prof. Thomas G. Dietterish, Prof. Amir Nayyeri and Prof. John P. Bolte, for their encouragement, insightful comments, and hard questions.

Also, I would like to thank to the OSU Writing Center and Dr. Adam D. Haley for the sincere help during the thesis writing. I thank my research fellows in Dr. Hutchinson's group: Eugene Seo and Laurel Hopkins, for the research discussions and feedback.

Last but not the least, I would like to thank my wife Xiaoqin Bai for her understanding and love during the past years.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

LIST OF TABLES (Continued)

Table Page

LIST OF ALGORITHMS

Algorithm Page

LIST OF APPENDIX FIGURES

Figure Page

LIST OF APPENDIX FIGURES (Continued)

Figure Page

LIST OF APPENDIX FIGURES (Continued)

Figure Page

Chapter 1: Introduction

Species distribution models (SDMs) integrate environmental variables and field data to produce a spatially explicit, statistically derived response surface [4]. They are increasingly used to improve the understanding of species habitat factors and interactions between species and environment, to predict the response of species to climate change, and to identify and manage conservation areas. In this work, I build a species distribution model using binary labels to represent the detection or non-detection of one species at a set of locations. In other words, the datasets I use include "presence-absence" data, but not "presence-background" or "occupancy-detection" data.

In this work, I explore SDMs in the context of citizen science projects, which engage volunteers in data collection. Citizen scientists now participate in projects in a variety of areas, including climate change, conservation biology, invasive species, etc. In the eBird project, bird watchers can report their observations to a database at the Cornell Lab of Ornithology using an online checklist system on their mobile devices [25]. Citizen scientists begin their observation tasks by selecting any location on the map, while the mobile app helps them to record information like the bird species they observed, the time, the duration, traveling distance, and the number of observers. After the user submits the checklists, the eBird project system combines with GIS layers of the environmental variables to produce a geographic distribution of suitable species' habitats. If the participants report all the species they find, we can infer the absence of the complement of this set of species. Using the records in the dataset provides access to a set of instances of a species with detection or non-detection labels at each location, linked to habitat features and observation features.

I propose to formulate the task of species distribution modeling using citizen science data as a classification problem with class conditional label noise. In the case of species *presence*, label noise comes from *imperfect detection* of secretive or cryptic species [14]. In the case of species *absence*, label noise comes from mistaken identifications by

observers [22]. Many common modeling approaches assume no false positive noise in their models, because in many field studies, the data are collected by highly trained observers. However, given the variety of volunteers' expertise levels in the citizen science project context, this assumption cannot always hold. In this proposed model, the species distribution, represented by binary labels for each instance, suffers from both false positive and false negative noise. The two kinds of class conditional label noise are not necessary symmetric, which indicates that this is a challenging learning problem [23].

In the species distribution modeling process, model selection is also a crucial problem that involves selecting the features to be included and the functional form of their relationship to probabilities in the model. In reality, a species' distribution is impacted by many factors, such as climate, environment, and human activities. It is challenging to fully understand one species' habitat, which means that there are potentials for missing or adding class features when specifying SDMs. Moreover, the noise source of the detection is complicated. Some factors, such as the number of observers, the eyesight of birders, and the distance of the observation, are easily overlooked during the model construction. Therefore, missing covariates or adding irrelevant covariates becomes inevitable.

In the graphical model I will present, there are three probabilities: one class probability and two noise probabilities. Thus, in addition to the typical task of identifying relevant features, we must also partition those features into the class, false negative, and false positive models. I assume that in many cases, some features can easily be identified as class features based on ecologically relevant knowledge [5]. However, some features, such as the surrounding landscape type, impact both the class probability and noise probabilities. One typical case is the camouflage phenomenon of some species. Camouflage is conferred by background matching and disruption, which also reflects the importance of surrounding habitat features to this class. For example, as shown in Figure 1.1, a nighthawk is well hidden in the lower left corner. It is difficult for the observer to detect the bird due to the similarity between the color of the bird's feather and the color of the background. Hence, the class features, such as the landscape type, are also noise features that impact the detection of the species, which implies that there are some "overlaps" between the class features and noise features in the true model. As a result, if

we failed to understand the overlaps, our model would suffer from another kind of model misspecification, the mis-specifying of overlaps.

Figure 1.1: Nighthawk

In addition to feature misspecification, we have to identify the functional form used in the simulation. In the true model, the functional form can be any possible functional form, such as linear, exponential, quadratic, or a combination of them, depending on how complex the model is. In this work, I focus on linear functions of the features in each probability model, leaving other functional forms for future work. Additionally, the link function used in the data generation and fitting process can be logistic, probit, scale, or any other valid link function type. Hence, the link function misspecification, which is related to the identifiability problem that I will discuss, is also a crucial source of model misspecification.

In this work, I construct simulations to test the sample complexity for the proposed model, comparing with candidate models. I evaluate the model performance of the proposed model without considering model misspecification, but varying the class balance, noise rates, and feature distribution. Additionally, I design experiments to explore the influence of link function misspecification, missing or adding covariates and mis-specifying overlaps. Then I report on an application of our proposed model to the eBird data to compare against alternative approaches, from work with collaborators [6]. I use held-out log likelihood on validation data to select among candidate models.

The goals of this work are:

⋆ to connect literature addressing this problem across several research communities (crowdsourcing, ecology, machine learning);

⋆ to propose a generalization of the existing approaches;

⋆ to describe conditions for identifiability of the proposed model;

⋆ to explore performance of the proposed approach in simulation as compared with existing approaches;

⋆ to evaluate whether model selection correctly identifies the features associated with each sub-model of the approach;

⋆ to investigate model performance under three kinds of model misspecification;

⋆ to compare predictions from the proposed approach to existing approaches.

The thesis is organized as follows:

- The related work is reviewed in Chapter 2. Four topics of related work are discussed here, including classification problems with class-conditional label noise, statistical ecology models, crowdsourcing models, and model selection.
- I present our model framework in Chapter 3.
- Five simulation experiments are presented in Chapter 4. Experiments of sample complexity, model performance without model misspecification, link function misspecification, and two kinds of feature misspecification are introduced here.
- We implement our model framework on the eBird Reference Dataset in Chapter 5.
- I conclude the thesis in Chapter 6.

This thesis is an extension of our paper "Species Distribution Modeling of Citizen Science Data as a Classification Problem with Class-conditional Noise". Sarah C. Emerson and Rebecca A. Hutchinson contributed significantly on the identifiability and empirical sections, respectively. The sections 2, 3.3 and 5 are cited from this paper. I appreciate their effort on the contributions and enjoy the collaboration with them.

Chapter 2: Related Work

This chapter is taken directly from our paper [6]. For the coherence and comprehension, we place them here.

2.1 Classification Problems with Label Noise

Our work builds on research into classification settings with label noise [17, 20, 16, 13, 10]. Frénay and Verleysen (2014) provide a helpful review of types of label noise and methods for dealing with them [3]. As shown in Figure 2.1, three label noise models are summarized as Noise Completely at Random Model (NCAR), Noisy at Random Model (NAR), Noisy Not at Random Model (NNAR).

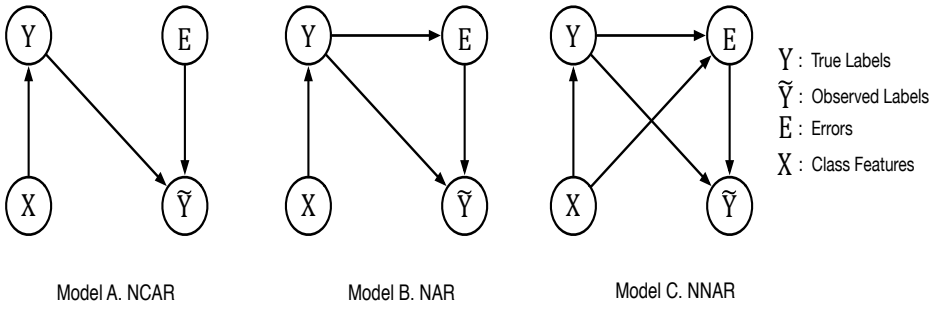

Figure 2.1: Model Framework summarized by Frénay and Verleysen

We focus on asymmetric label noise [23], as opposed to methods that assume equal rates of false positives and false negatives. Our framework is most closely related to the 'noisy not at random' (NNAR) model since we treat label noise as feature-dependent [3]. In contrast to the typical view of NNAR noise, in our work, the set of features on which the noise depends are fully or partially disjoint from the features on which the true class label depends.

2.2 Related work about statistical models in ecology

Our approach builds on statistical models in ecology. MacKenzie *et al.* [15] popularized *occupancy models* to account for imperfect detection in ecological studies. This approach requires multiple observations during a period when the true status of the species (the true class label) is unchanging, and it incorporates an explicit model of the observation process to correct for underreporting. However, it assumes that there are no false positives in the data. Further work on this family of models added the possibility of false positives [22], but the resulting model had limited application in practice due to identifiability issues [18]. Extensions of it improved model behavior, but they required additional information (e.g. multiple survey methods; [18]). More recent work has also explored variants of occupancy models requiring only a single observation at each location [11]. There is also discussion in the literature about the identifiability of single-visit occupancy models (without false positives) [8, 24]; in particular, identifiability for these models relies on the fidelity of the link function.

2.3 Related work about crowdsourcing models

Citizen science has both similarities to and differences from other crowdsourcing paradigms for generating labeled data, in which workers are presented with a set of instances to label. Some similarities are obvious; for example, the labels are being generated by a pool of workers with variable skill levels. Errors are likely to occur in both settings, causing both false negatives and false positives. In both paradigms, the difficulty of the instance as well as the skill and effort levels of the labeler may affect the probability of mislabeling an instance [1]. However, there are some important distinctions between citizen science and crowdsourcing as well. In crowdsourcing, the set of instances to be labeled and their assignments to labelers are controlled by the task setting. In contrast, citizen scientists effectively choose their own instances to label from an infinite set of possibilities, since volunteers choose the times and places to make their observations. A consequence of this freedom is that we do not in general have multiple labels of each instance. Some analyses of citizen science data have grouped together checklists that are close in space and time to construct multiple label structure [26, 7, 27], but this requires assumptions about the granularity with which to aggregate. In addition, in citizen science, there are

sources of variability in mislabeling probabilities that go beyond labeler effort and task difficulty: the observation conditions under which the labels were generated (e.g. time of day, weather). Work by Raykar *et al.* takes a similar approach to crowdsourced data as the model we propose herein, except that it relies on multiple labels per instance and does not include features in the noise models [2].

2.4 Relation to our approach

Our proposed approach unifies several of the related works discussed above. It can be seen as a novel use of the model from Royle and Link [21], applied to single-observation data rather than multiple-observation data. Alternatively, it can be seen as an extension of the single-observation occupancy model of [12] that allows for false positives. In the crowdsourcing context, our approach can be viewed as an extension of the Raykar *et al.* model that uses only a single label per instance and introduces features to describe the noise processes [2]. Instead of giving suggestions on model selection when building a model, in this work we focus more on the impact of three kinds of model misspecification, since in most of the species distribution modeling processes, the true model is unknown.

Chapter 3: Model Framework and Our Approach

We consider the problem of learning an SDM from citizen science data. As with previous formulations of the problem, we are given environmental features and species labels, and we use a latent variable for the true label of each instance. The challenge we face in citizen science data is that both false positive and false negative noise may corrupt the labels. To deal with this, we treat the problem as a classification problem with class conditional label noise. We have an additional source of information to bring to bear on the problem (beyond what is usually provided for classification with label noise), which is a set of observation or noise features that help determine when instances have had their labels 'flipped' one way or the other. Below, we detail our model framework, parameter learning approach, and challenges to be addressed.

3.1 Model Framework

As mentioned in Chapter 2, Frénay et al. summarized three possible statistical models of label noise, named as Noise Completely at Random Model (NCAR), Noisy at Random Model (NAR), Noisy Not at Random Model (NNAR) [3].

In this work, we propose to formulate the species distribution modeling task as a classification problem with class-conditional noise, and we name it as Noise with Noise-specific Features Model (NNSF), as shown in Figure 3.1.

The key difference between our framework and previous work is that we allow errors to depend on an additional set of features W, distinct from the features X that influence the true class. Hence, the errors E, which flip the true label $Y = 1$ to the observation label $\tilde{Y} = 0$ or flip the true label $Y = 0$ to the observation label $\tilde{Y} = 1$, depend on the classes, the class features, and the noise features. For clarity, we will refer to X as the

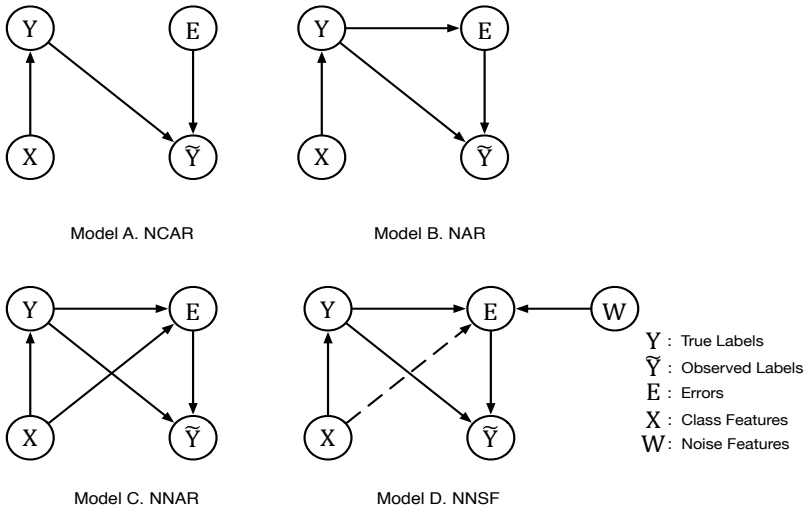

Figure 3.1: Model Framework

class features and W as the *noise features*. We define the following quantities:

$$\begin{aligned}
\psi &:= P(Y = 1|X) = \sigma_f(f(X;\alpha)) \\
\rho &:= P(E = 1|Y = 0, W) = P(\tilde{Y} = 1|Y = 0, W) = \sigma_g(g(W^{(\rho)};\beta)) \\
\eta &:= P(E = 1|Y = 1, W) = P(\tilde{Y} = 0|Y = 1, W) = \sigma_h(h(W^{(\eta)};\gamma)),
\end{aligned}$$

where ψ, ρ, η are used to denote the true class probability, false positive probability, and false negative probability, respectively.
Therefore, the probability of a positive observation for instance i is

$$p_i = P(\tilde{Y}_i = 1|X_i, W_i) = \psi_i(1 - \eta_i) + (1 - \psi_i)\rho_i,$$

and the likelihood function for the model is

$$L = \prod_{i=1}^{N} p_i^{Y_i}(1 - p_i)^{1-Y_i}.$$

3.2 Problems and Approaches

We use maximum likelihood estimation to learn the parameters. The log likelihood function is shown as below.

$$l = \log(L) = \log(\prod_{i=1}^{N} p_i^{Y_i}(1-p_i)^{1-Y_i}.)$$

Clearly, there is a summation of log in the above log likelihood. Therefore, the function $l = 0$ is not convex in general. To handle this non-convex property, in this work we use random restarts within each model fitting process. After we learned the parameters, we can use this retrieved model to predict the species distribution.

3.3 Challenge: Identifiability

This section is cited from our paper [6], contributed by Sarah C. Emerson. For the consideration of comprehension, I place it here.

The Identifiability problem arises when comparing to some baseline condition, say X_0. Will the occupancy probability under condition X_{new} be increased or decreased? It is difficult to answer without making some assumptions. For instance, suppose that we observe that for larger values of X, $\tilde{Y}$ is more likely to be one. Unless we make additional constraining assumptions, we cannot tell whether this is because Y is more likely to be one in this setting, because E is more likely to be one when Y is zero, or because E is more likely to be zero when Y is one in this setting (or some combination of these possibilities).

Without significantly constraining the problem, we are unable to distinguish between these different explanations. This is the same concern that was discussed by Knape and Korner-Nievergelt [9], and addressed further with some constraint recommendations by Sólymos and Lele [24]. Specifically, we are concerned with the identifiability of the fundamental parameters ψ, ρ, and η, and also with the identifiability of the coefficient parameters α, β, and γ. Clearly the coefficient parameters are not identifiable if the fundamental parameters are non-identifiable. If the fundamental parameters *are* identifiable, then the identifiability of the coefficient parameters depends on the form and

parameterization of the covariate functions. Our concern is how to make both the the fundamental parameters *and* the coefficient parameters identifiable, since it is the coefficient parameters that address the primary scientific questions of interest: namely, if a covariate value changes, what are the effects on the probability that the species of interest is present and, secondarily, on the probabilities of observation errors.

Starting with complete generality, letting $Z = (X, W)$ and $\theta = (\psi, \eta, \rho)$, we need:

$$P(\tilde{Y} = 1|Z = z, \theta) = \psi(z)\,[1 - \eta(z)] + [1 - \psi(z)]\,\rho(z)$$

to satisfy

$$\sup_z \left| P(\tilde{Y} = 1|z, \theta) - P(\tilde{Y} = 1|z, \theta^*) \right| > 0$$

whenever $\theta \neq \theta^*$, so that every distinct value of the parameters produces distinct value of the likelihood. First note that there are an infinite number of solutions since there are three unknowns in this single equation. The solutions are naturally constrained by the fact that ψ, η, and ρ are all probabilities, and therefore must be in the range $[0, 1]$. The solution space is further reduced if we require that ψ, η, and ρ have certain functional forms as a function of Z, but of course then the solution is dependent upon the form chosen. We focus here on the logistic link function, as one of many link functions that would work to assist identifiability. In some settings, it can be further useful to require that at least one of the active covariates is continuous, since in the case of discrete covariates handled with indicator functions the functional form may not provide as much structure. For example, if a different value of $\psi(z), \rho(z)$, and $\eta(z)$ is allowed for every possible discrete value of $Z = z$, we are back to the most general case and have lost identifiability. However, in more restricted settings such as when distinct non- or minimally-overlapping sets of covariates affect the different fundamental parameters ψ, ρ, and η, we have identifiability without requiring that any of the covariates be continuous.

Note that even specifying the functional form (link function) is not enough to provide identifiability if the model families for η and ρ are the same and include $1 - \sigma$ whenever σ is a valid link. This is because the following two complementary models give identical

probabilities: $\theta_1 = (\psi, \eta, \rho)$ and $\theta_2 = (1-\psi, 1-\rho, 1-\eta)$. To address this complementary model source of non-identifiability, we must constrain the problem to allow only one of these results (i.e., define a way to select between these two equal-likelihood solutions). This can be accomplished in several ways: either by selecting the solution that gives values of ψ, η, or ρ in a certain range (e.g., choose the solution with the minimum value of η), or by specifying that, for instance, the function η depends on a certain set of covariates, while ρ depends on a distinct set. Constraining the covariates that appear in the different noise models to be distinct sets will work to eliminate the issue of complementary model solutions since if we know that $\rho = \sigma_g(g(W^{(\rho)}; \beta))$ and $\eta = \sigma_h(h(W^{(\eta)}; \gamma))$ with distinct covariate sets $W^{(\rho)}$ and $W^{(\eta)}$, then $\rho_2 = 1-\eta = 1-\sigma_h(h(W^{(\eta)}; \gamma))$ is not a valid solution because it depends on the wrong covariates (and likewise for $\eta_2 = 1 - \rho$).

The two complementary equally optimal solutions are also the reason that this model does not perform well when the noise models do not depend on covariates (the 'constant noise' settings in the experiments below): the resulting lack of identifiability means that there are two complementary solutions that fit the observed data equally well, so the parameter estimates are not unique (and thus have very large variances). Likewise, if the covariates that affect the noise models are effectively close to constant, identifiability will be very fragile as we will be very close to the constant noise setting. Thus we have a spectrum of identifiability that depends on the chosen form of the model and the roles and distributions of the relevant covariates.

In summary, in the fitting process, we have to correctly specify the fundamental parameters (ψ, ρ, and η), the coefficient parameters (α, β, and γ), and the form and parameterization of the covariate functions ($f(x)$, $g(w)$, and $h(u)$). Identifiability depends on the link function, and also relates to the feature distributions. Due to the equality of likelihood between $\theta_1 = (\psi, \eta, \rho)$ and $\theta_2 = (1 - \psi, 1 - \rho, 1 - \eta)$, we suggest to allow only one of these two results.

3.4 Model Specification

When modeling the species distribution in reality, the true model is completely hidden. In other words, we do not know the exact functional form for each probability, or the

features involved in it. Therefore, when proposing a model with probabilities ψ, ρ, and η, model misspecification will arise.

In this work, we use linear functions to represent the relationship between probabilities and features, and leave other functional forms for future work. We focus on the influence of the incorrect arrangement of features on each probability, for example, extra or missing covariates and mis-specified overlap between features in the models.

3.4.1 Feature Misspecification I: Extra or Missing Covariates

The first kind of feature misspecification is extra or missing covariates in each probability function.

During model construction, we can select class features based on relevant ecology knowledge. However, it is not completely clear whether some features should be considered in the class probability ψ or the two noise probabilities (ρ and η) when building the statistical model. Given the true model is hidden, there is a high potential that missing or extra covariates will arise. Hence, it is crucial to explore the performance impact that caused by this kind of feature misspecification.

In this work, I construct simulation experiments with a given true model. By adding or removing covariates from that true model, I will measure the performance change, compared with other candidates. In eBird data, we have a model selection problem as well. Hence, we apply this proposed model to eBird data, and study the model performance by varying the feature participation in the three functions.

3.4.2 Feature Misspecification II: Mis-specifying overlaps

In the last section, we discussed the model misspecification of extra or missing features. However, some features can affect both the class probabilities and noise probabilities, such as weather conditions. We define these features appearing in two or three probabilities as "overlaps" in our model. Overlap is interesting because of the identifiability challenge. A perfectly overlapping model is completely non-identifiable.

Since we have three probabilities ψ, ρ and η, we denote below three types of overlaps.

Type I: Overlaps between class model and one noise model.

Type II: Overlaps between class model and two noise models.

Type III: Overlaps between the two noise models.

The Type I and II overlaps represent the features that not only can be attributed to class features, but also can be considered as noise features. For example, the distance to the stream is an important factor to many bird species' distribution. However, the sound of the stream can affect the detection, causing either false positive or false negative records.

The Type III overlap represents the observation features that impact both false positive noise and false negative noise. These features include observers' characteristics (such as age, expertise level), the method of observation, and any other features that cause variation in detection. For instance, The House Finches and the House Sparrows look similar to each other, which makes it difficulty for observers, especially non-expert volunteers, to distinguishing them. For example, as shown in Figure 3.2, one observer might record all the birds as House Finches, and mark zero for the House Sparrows. Thus, he introduced false negative noise to the dataset of House Sparrows, and introduced false positive noise into the dataset of House Finches. Thus, the expertise level

Figure 3.2: House Finches and House Sparrows.

Figure 3.3: A house female finch that looks similar to house sparrows.

of the observer appears as a noise feature to both the false positive noise rate and the false negative noise rate.

Since the true model of the species distribution can be very complex, all three types of overlap may exist. However, as we discussed before, the true model is unknown, which means that there is a high potential to miss one or more overlaps when building statistical models.

In this work, we construct simulation experiments to evaluate the influence of this type of feature misspecification. The true model is designed to have all three types of overlaps. By removing one or two overlaps, we can investigate the influence of missing of each overlap. In the experiment, we only consider type III overlap, and leave the other two kinds of overlaps for future work.

Chapter 4: Simulation Experiment

4.1 Background and General Settings

In Chapter 3, we introduced a classification model with class conditional label noise, including both false positive noise and false negative noise. In this chapter, we designed simulation experiments to explore model performance under scenarios with different purposes.

In order to study the sensitivity of the sample size to the model performance, we designed experiment I to compare the MSE (mean square of error) of class probabilities among sample sizes of 200, 400, 800, 1600, 3200, 6400, and 12800. We found that the sample size of 3200 was sufficient to compare model performance with other competitors, and we selected it for the other experiments. In experiment II, we studied the model performance comparing to other competitors without considering model misspecification. In experiment III, we introduced model misspecification of link functions, and explored the impact to model performance. Finally, we evaluated two other types of model misspecification in experiment IV and V.

4.1.1 Method Comparison

We denote our proposed model that has noise features for both false positives and false negatives as the "FP" model. In several of the following experiments, we compare our proposed models with:

1. *LR1*: Logistic regression ignoring label noise (using just the class features to predict the noisy labels).

2. *LR2*: Logistic regression ignoring the label structure but using the noise features (using the class and noise features all together to predict the noisy labels).

3. *OCC* (for OCCupancy): Single-visit occupancy model, using all noise features for false negatives and ignoring false positives [14].

4. *CN* (for Constant Noise models): Proposed model with false positives and false negatives but without noise features (using a constant/intercept-only model for both); similar to [2].

4.1.2 General Set-up and Evaluation Methods

In all of these experiments, for each scenario, we set the non-intercept coefficients to 1 and vary the intercepts to explore the effects of class balance and noise rates on performance. For the class balance, the low, medium, and high values are $\bar{\psi} \in \{0.25, 0.5, 0.75\}$. For the noise levels, the low, medium, and high levels are $\bar{\eta} \in \{0.1, 0.25, 0.4\}$ and $\bar{\rho} \in \{0.1, 0.25, 0.4\}$. The continuous features have N(0,1) distribution, while the categorical features have a Bern(0.5) distribution.
For all the methods, we measure the quality of predictions of the true class labels using mean squared error (MSE), and compare model performance among different models.

4.2 Experiment I: How does the sample complexity compare among these five methods?

4.2.1 Experiment Introduction

An important question is, how much data will be sufficient if we want to model this problem using the proposed approach? In this experiment, we construct simulations to explore model performance with varied sample sizes.

4.2.2 Experiment Set-up and Data Generation Process

We constructed models of the following form,

$$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$$
$$g = \beta_0 + \beta_1 W_1 + \beta_2 W_2$$
$$h = \gamma_0 + \gamma_1 W_3 + \gamma_2 W_4$$

where all the features have continuous distributions, from $N(0, 1)$. We simulated 30 training sets each with varied sizes of 200, 400, 800, 1600, 3200, 6400, and 12800, and used the logistic link function to generate datasets.

4.2.3 Simulation Result

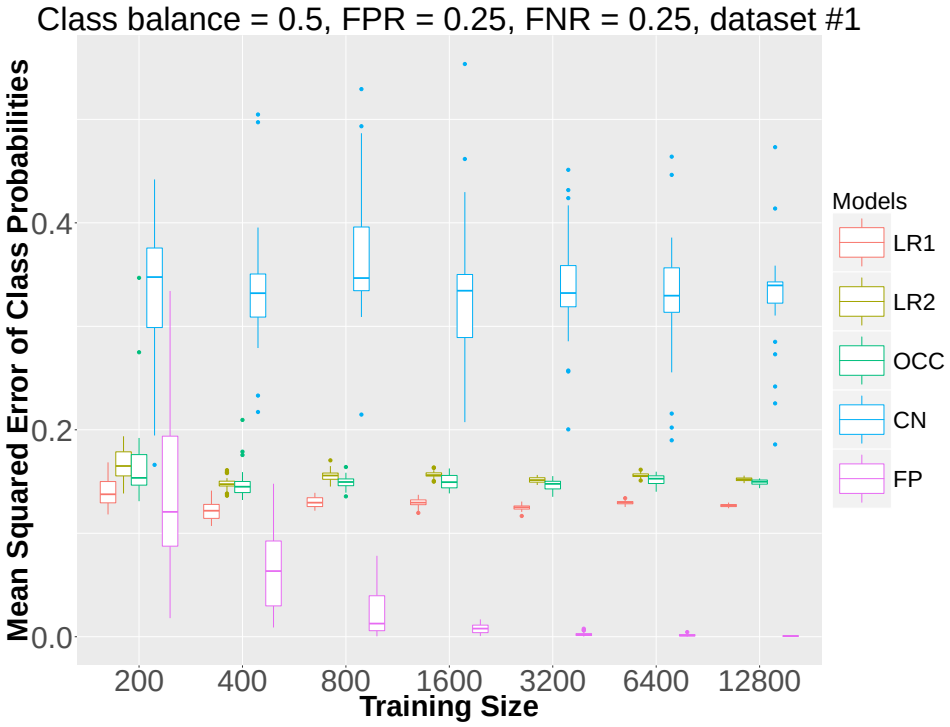

Figure 4.1: Mean squared error in the class probabilities for dataset #1 with mid-level class balance and noise rates. Each boxplot represents 30 simulated datasets.

Figure 4.1 shows clearly that the sample complexity requirements of the *FP* model are greater than those of simpler models. Hence, for small datasets, even the simple logistic regressions predict class probabilities better (Figure 4.1). The specific threshold for enough data to support the *FP* approach will vary depending on other characteristics of the problem (e.g. number of features). In the following experiments, we selected 3200 as the sample size to simulate data.

4.3 Experiment II: How does the proposed model compare to other competitors in the absence of model misspecification?

4.3.1 Experiment Introduction

In this experiment, we designed simulations to explore model performance and identifiability under scenarios with varying feature overlap between feature functions (i.e. f, g, and h), feature distributions, class balance, and noise rates.

We constructed datasets from the FP model and fit with the FP model and the other four competitors. Therefore, the FP model is specified correctly. That is, every time, we know exactly what the structure of the true model is. However, there is model misspecification for the four competitors. We hypothesize that the FP model should outperform all other models across all the setting combinations.

4.3.2 Experiment Set-up and Data Generation Process

In this simulation, the form of the model is the same as we used in experiment I, as shown below:

$$
\begin{aligned}
f &= \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 \\
g &= \beta_0 + \beta_1 W_1 + \beta_2 W_2 \\
h &= \gamma_0 + \gamma_1 W_3 + \gamma_2 W_4
\end{aligned}
$$

All seven features were generated independently (called *none*, for no overlap). We created three variants of this setting (#1-3 in Table 4.1): 1) *cont*: all continuous features,

Number	Name	Feature overlap	Feature distributions
1	*none-cont-logistic*	none	all $N(0,1)$
2	*none-mix-logistic*	none	odd $N(0,1)$, even $Bern(0.5)$
3	*none-cat-logistic*	none	all $Bern(0.5)$
4	*noise-cont-logistic*	noise: $W_2 = W_4$	all $N(0,1)$
5	*noise-mix-logistic*	noise: $W_2 = W_4$	odd $N(0,1)$, even $Bern(0.5)$
6	*class-cont-logistic*	class&noise: $X_1 = W_1$; $X_3 = W_3$	all $N(0,1)$
7	*class-mix-logistic*	class&noise: $X_1 = W_1$; $X_3 = W_3$	odd $N(0,1)$, even $Bern(0.5)$

Table 4.1: Simulated data settings. Names reflect the overlap setting, feature distributions, and link functions. Feature overlap refers to the extent to which the features of each feature function were shared. Feature distributions were either Gaussian or Bernoulli, in some cases depending on whether the index of the feature was odd or even. The logistic link function was used to generate data and fit the models.

from $N(0,1)$, 2) *mix*: odd-indexed continuous and even-indexed categorical features, from $N(0,1)$ and $Bern(0.5)$, and 3) *cat*: all categorical features, from $Bern(0.5)$. In the second set of simulations, we set $W_2 = W_4$ to explore the effect of overlapping noise features (called *noise*, for overlap in the noise features; #4-5 in Table 4.1). In this setting, we created variants with both continuous and mixed features. In the third set of simulations, we generated overlap between the class and noise models by setting $X_1 = W_1$ and $X_3 = W_3$ (called *class*, for overlap in the class and noise features; #6-7 in Table 4.1). Similarly to the *noise* setting, we used the *cont* and *mix* feature settings.

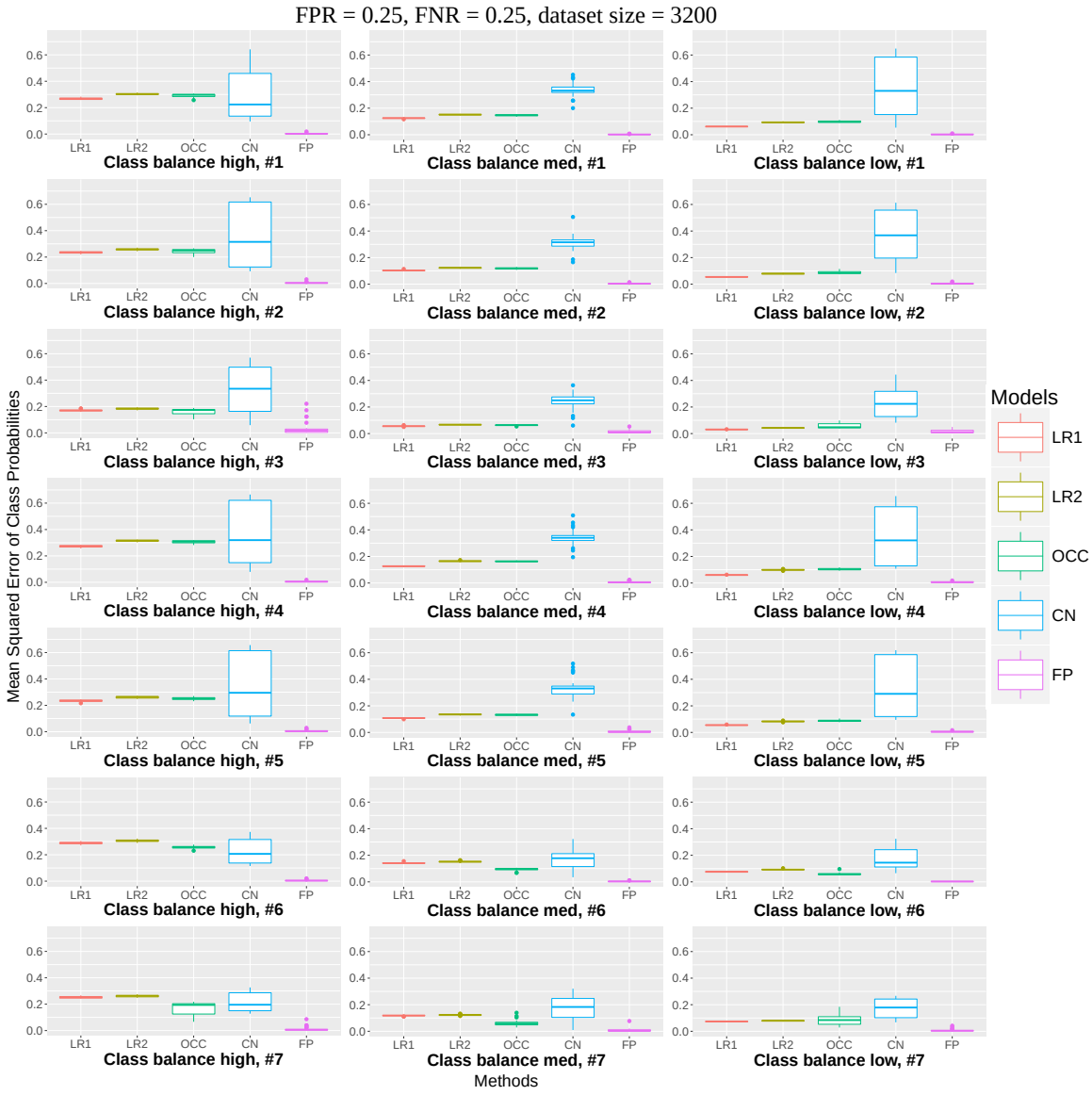

Figure 4.2: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.25 and FNR = 0.25. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

4.3.3 Simulation Result

As stated before, we also fit four other competitors (LR1, LR2, OCCU, and CN model) for comparison with our proposed model. The results from an intermediate setting for the noise rates are presented in Figure 4.2 as a representative example, since the results are relatively consistent across settings. The results of the nine noise rate combinations are available in the appendix Figures A.1 to A.9.

In this ideal setting with identifiable, well-specified models and enough data, the types of features and their overlap in the feature functions are not critical (Figure 4.2, #1-7). The *FP* model predicts the class probabilities better than the alternative methods, as expected. The CN model performs worse than other models, and has high variationce. The reason is that model CN considers the noise structures but ignores the noise features, and is therefore not identifiable.

4.4 Experiment III: How does the proposed model compare to the four competitors when the link function is misspecified?

4.4.1 Experiment Introduction

In experiment II, we used the logistic link function in both the data generation process and the model fitting process. However, if the data were generated by other link functions and fit by the logistic link function, we would expect that the result would be different.

From the discussion in the Identifiability section, we concluded that identifiability depends on not only the features, but also the link functions. Therefore, in this experiment, we designed simulations to explore this question. By comparing the model performance with other candidates, we can evaluate the impact of link function misspecification.

Number	Name	Feature overlap	Feature distributions	Link
5	*noise-mix-logistic*	noise: $W_2 = W_4$	odd $N(0,1)$, even $Bern(0.5)$	*logistic*
8	*noise-mix-probit*	noise: $W_2 = W_4$	odd $N(0,1)$, even $Bern(0.5)$	*probit*
9	*noise-mix-scale*	noise: $W_2 = W_4$	odd $N(0,1)$, even $Bern(0.5)$	*scale*

Table 4.2: Simulated data settings 2. Link functions *probit* and *scale* were used to generate data, though the *logit* link was always used in fitting.

4.4.2 Experiment Set-up and Data Generation Process

To explore the sensitivity of misspecification of link functions, we constructed the experiment with the function settings used in the previous two experiments, as shown below:

$$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$$
$$g = \beta_0 + \beta_1 W_1 + \beta_2 W_2$$
$$h = \gamma_0 + \gamma_1 W_3 + \gamma_2 W_4$$

We generated datasets from the *noise-mix* setting (#5 in Table 4.2) using two other link functions (and the model was always fit with a *logit* link): the *probit* link and a *scale* link that scaled and shifted the real values linearly to transform them into probabilities (#8-9 in Table 4.2). The feature overlap, distribution, and link settings are also summarized in Table 4.2. Note that all settings have at least one distinct feature in each feature function, so the coefficient parameters are identifiable.

4.4.3 Simulation Result

Figure 4.3 shows the MSE on the data probabilities of the FP model, along with the other competitors. As expected, the proposed FP model performs worse in this experiment compared to in Experiment I. This means that misspecification of the link function impacts the model performance significantly. Again, the *CN* model tends to perform worst, which is likely due to lack of identifiability in the noise parameters, since the *CN* model does not have features specified for g and h to distinguish between the two symmetric solutions discussed above.

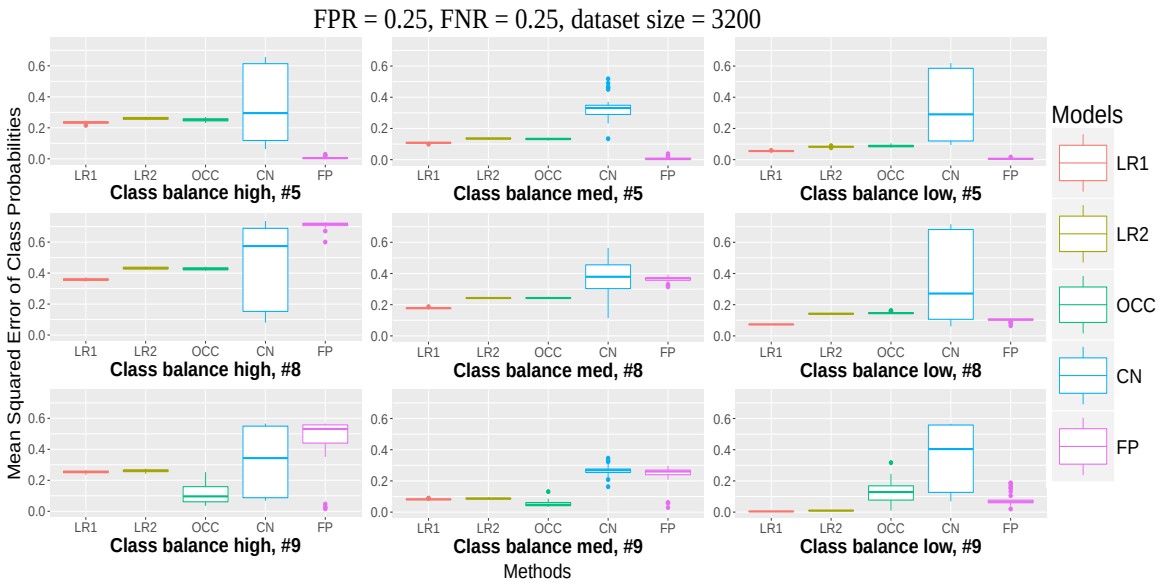

Figure 4.3: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.2). All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

4.5 Experiment IV: How does missing or extra covariates impact the performance?

4.5.1 Experiment Introduction

In previous simulations, we constructed the experiments with knowledge of the true model in the fitting process. As discussed in Chapter 3, however, the true model is usually unknown in reality. Therefore, there is a high risk of incorrectly specifying the model, which will cause the model identifiability problem and inaccuracy in the predictions.

In this simulation, to explore the difference in performance between the true model and the proposed models, we designed the feature misspecification scenario of missing or adding covariates.

4.5.2 Experiment Set-up and Data Generation Process

In this experiment, we constructed the true model used to generate datasets as below.

$$
\begin{aligned}
f &= \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 \\
g &= \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4 \\
h &= \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4
\end{aligned}
$$

Compared with previous models, the true model used here not only has more features involved in each probability function, but also exhibits more complicated overlaps. For instance, the class features X_2 and X_3 appear in the false positive probability and false negative probability functions, respectively, and the class feature X_4 appears in both noise probability functions. Hence, the true model intrinsically contains all three overlap types.

To explore the impact of this kind of feature misspecification, we created six variants (FP1 - FP6) of the true model by simply removing one feature or adding one irrelevant feature, as listed in Table 4.3. Since these added/removed features do not have an overlap counterpart in other probability functions, we can evaluate the impact of missing or irrelevant features that do not overlap with other models.

Similar to the settings in previous experiments, all features were generated independently, and have continuous distributions from $N(0, 1)$. We leave categorical features for future study. The sample size is 3200 for each model, following the discussion in experiment I. The link function is logistic function, used in both data generating and model fitting processes.

Model	Comments	Three Model Probability Functions
FP	true model	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4$
FP_ALL	complete overlap	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 W_4 + \beta_3 U_1 + \beta_4 X_1 + \beta_5 X_2$ $+\beta_6 X_3 + \beta_7 X_4 + \beta_8 X_5$ $h = \gamma_0 + \gamma_1 W_1 + \gamma_2 W_4 + \gamma_3 U_1 + \gamma_4 X_1$ $+\gamma_5 X_2 + \gamma_6 X_3 + \gamma_7 X_4 + \gamma_8 X_5$
FP1	miss class feature X_1	$f = \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4$
FP2	add class feature X_6	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \boldsymbol{\alpha_6 X_6}$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4$
FP3	miss FP noise feature W_1	$f = \alpha_0 + \alpha_1 X_1 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4$
FP4	add FP noise feature W_5	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4 + \boldsymbol{\beta_5 W_5}$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4$
FP5	miss FN noise feature U_1	$f = \alpha_0 + \alpha_1 X_1 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4$ $h = \gamma_0 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4$
FP6	add FN noise feature U_5	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4 + \boldsymbol{\gamma_5 U_5}$

Table 4.3: Simulated data settings for Type II feature misspecification.

4.5.3 Simulation Result

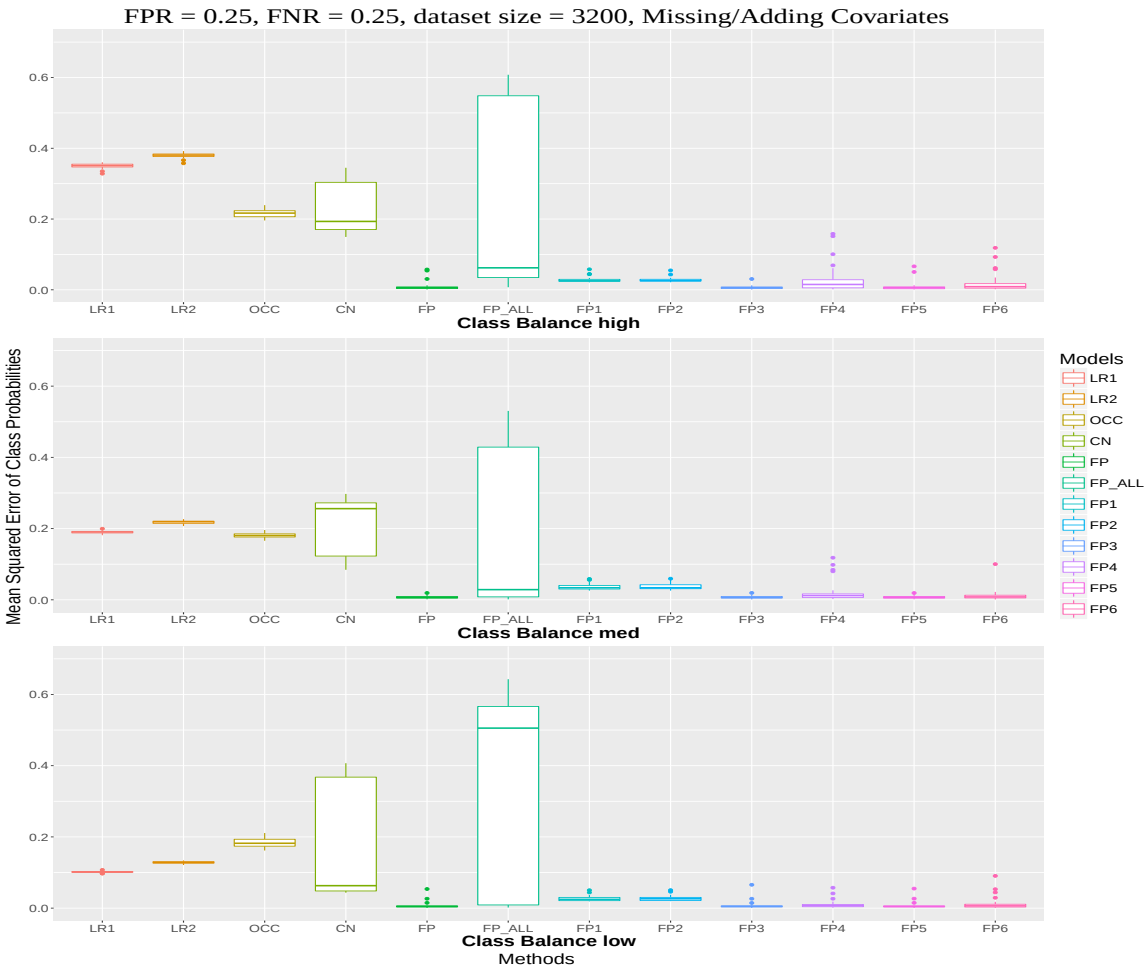

Figure 4.4: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.25$ and $FNR = 0.25$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

The results across different combinations of class balance and noise rates are relatively consistent, as shown in appendix Figure A.10 to Figure A.18. Here, we present results from the intermediate setting for noise rates as a representative example, as shown in Figure 4.4. In this experiment setting, there are overlaps within the models. Given the

identifiability concerns, this model complication makes it more challenging to fit than in the setting in Experiment II. However, the correctly-specified FP model always performs the best on MSE, which is consistent with our conclusion from Experiment II.

Generally speaking, missing or adding one feature, no matter whether it is a class feature or a noise feature, does not have a critical impact on the model performance. Models FP1 to FP6, all of which have the model misspecification of adding or missing covariates, exhibit much lower MSE of class probabilities than the four competitors. This means that when building a statistical model to study the species distribution, this kind of model misspecification, if only related to one covariate, might be not a serious factor in the prediction inaccuracy. Additionally, we find that missing/adding class features (FP1 and FP2) will result in a small decline of model performance (higher MSE). However, the impact of missing or adding noise features is even smaller. The reason is that class features affect class probability directly, while noise features impact the class balance by changing the noise rates in the label flipping process.

One thing to note is that the above conclusion may not be valid while missing more than one feature, since all the models explored in this experiment only involve one feature missed or added. Moreover, another hidden assumption is that this added/ removed feature does not have overlap counterpart in other probability functions. We will discuss the overlap misspecification in next experiment.

Additionally, we simulated the performance of model FP_ALL, which includes all the features in all three probability functions. As we can see from the Figure 4.4, its performance is not good, sometimes even worse than the four competitors. This result is as expected, because the FP_ALL model is the fully non-identifiable model. In other words, it is equivalent to the CN model, which considers the noise structures but ignores the contribution of noise features. Therefore, it is not a wise option to introduce all the features to all of the probability function models.

4.6 Experiment V: How do the missing overlaps impact the performance?

4.6.1 Experiment Introduction

In the previous experiment, the true model includes complicated overlaps, and all proposed FP models correctly specify the overlaps. We understand that the complicated overlaps make identifiability more difficult, but we do not know how significant the impact will be if we mis-specify the overlap. To explore the influence of missing overlaps, we designed this experiment.

4.6.2 Experiment Set-up and Data Generation Process

In this experiment, we constructed the same true model to generate the true value dataset, as shown below:

$$
\begin{aligned}
f &= \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 \\
g &= \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 W_4 \\
h &= \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \gamma_4 W_4
\end{aligned}
$$

These three functions are exactly the same as the previous experiment. All the features were generated independently, and have continuous distributions from N (0, 1). The sample size was 3200 for each model, following the discussion in experiment I. Same to the previous experiment's setting, we used the logistic function as the link function in both the data generating and model fitting processes.

Different from the previous experiments, we varied the fitted models by manipulating the overlap combinations. As shown in Table 4.4, models FP7 - FP9 each miss one type of overlap, and models FP10 - FP12 miss two overlaps. For comparison, we also fit model FP_ALL, which includes all the features in all of the three probability functions. Hence, by evaluating the model performance (MSE) between these models and the true model, we can better understand the impact of overlap misspecification.

Model	Comments	Three Model Probability Functions
FP	true model	$f = \alpha_0 + \alpha_1 X_1 + \boldsymbol{\alpha_2 X_2} + \alpha_3 X_3 + \boldsymbol{\alpha_4 X_4} + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \boldsymbol{\beta_2 X_2} + \boldsymbol{\beta_3 X_4} + \boldsymbol{\beta_4 W_4}$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \boldsymbol{\gamma_3 X_4} + \boldsymbol{\gamma_4 W_4}$
FP_ALL	complete overlap	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 W_4 + \beta_3 U_1 + \beta_4 X_1 + \beta_5 X_2$ $+\beta_6 X_3 + \beta_7 X_4 + \beta_8 X_5$ $h = \gamma_0 + \gamma_1 W_1 + \gamma_2 W_4 + \gamma_3 U_1 + \gamma_4 X_1$ $+\gamma_5 X_2 + \gamma_6 X_3 + \gamma_7 X_4 + \gamma_8 X_5$
FP7	no type I overlap	$f = \alpha_0 + \alpha_1 X_1 + \boldsymbol{\alpha_2 X_2} + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \boldsymbol{\beta_2 W_2} + \beta_3 X_4 + \beta_4 W_4$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 U_2 + \gamma_3 X_4 + \gamma_4 W_4$
FP8	no type II overlap	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \boldsymbol{\alpha_4 X_4} + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \boldsymbol{\beta_3 W_3} + \beta_4 W_4$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \boldsymbol{\gamma_3 U_3} + \gamma_4 W_4$
FP9	no type III overlap	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \beta_3 X_4 + \boldsymbol{\beta_4 W_4}$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \gamma_3 X_4 + \boldsymbol{\gamma_4 U_4}$
FP10	only type I overlap	$f = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \boldsymbol{\alpha_4 X_4} + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \beta_2 X_2 + \boldsymbol{\beta_3 W_3} + \boldsymbol{\beta_4 W_4}$ $h = \gamma_0 + \gamma_1 U_1 + \gamma_2 X_3 + \boldsymbol{\gamma_3 U_3} + \boldsymbol{\gamma_4 U_4}$
FP11	only type II overlap	$f = \alpha_0 + \alpha_1 X_1 + \boldsymbol{\alpha_2 X_2} + \boldsymbol{\alpha_3 X_3} + \alpha_4 X_4 + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \boldsymbol{\beta_2 W_2} + \beta_3 X_4 + \boldsymbol{\beta_4 W_4}$ $h = \gamma_0 + \gamma_1 U_1 + \boldsymbol{\gamma_2 U_2} + \gamma_3 X_4 + \boldsymbol{\gamma_4 U_4}$
FP12	only type III overlap	$f = \alpha_0 + \alpha_1 X_1 + \boldsymbol{\alpha_2 X_2} + \boldsymbol{\alpha_3 X_3} + \boldsymbol{\alpha_4 X_4} + \alpha_5 X_5$ $g = \beta_0 + \beta_1 W_1 + \boldsymbol{\beta_2 W_2} + \boldsymbol{\beta_3 W_3} + \beta_4 W_4$ $h = \gamma_0 + \gamma_1 U_1 + \boldsymbol{\gamma_2 U_2} + \boldsymbol{\gamma_3 U_3} + \gamma_4 W_4$

Table 4.4: Simulated data settings for Type II feature misspecification.

4.6.3 Simulation Result

From the simulation, we generated results across settings with different class balance and noise rates. Figures 4.5, 4.6 and 4.7 show the MSEs of all 12 models with different noise rates when class balances are low ($\psi = 0.25$), medium ($\psi = 0.50$), and high ($\psi = 0.75$), respectively. Within each of the above figures, the two noise rates are varied among 0.10, 0.25 and 0.40. The FP model represents the true model fitting, which has been discussed in the last experiment. Here we mostly focus on the model performance of models FP7 to FP12, of which the model constructions are detailed in Table 4.4.

Generally, the performances of our proposed models are very good. In 19 of 27 subplots, all of our proposed models (FP7 to FP12) outperform all candidate methods. And in 26 of 27 subplots, at least one FP model outperforms all competitors. The model performance varies across class balance and noise rate settings.

From Figure 4.5, which represents the low class balance ($\psi = 0.25$), we find that our proposed models outperform the four competitors in 6 of 9 combinations. When the false negative noise rate is low (FNR = 0.1) or the false positive noise rate is high (FPR = 0.4), the differences of MSE among these models are not significant, and our proposed models perform better than the competitors. When increasing the false negative noise rate and decreasing the false positive noise rate at the same time, however, the MSEs of some of our proposed models become worse than the simpler competitors. On the other hand, even under these severe conditions, some of our proposed models, such as model 9, still perform well. The reason for the performance decrease is the low class balance condition, which corresponds to a small amount of presence labels in the dataset. In other words, the data is sparse because the species is rare. Under this condition, the high false negative noise rate, which flips the 'presence' true labels to the 'absence' observed labels, causes further reduction of presences in the observed labels. At the same time, the low false positive noise rate means a low chance to flip 'absence' labels to 'presence' labels, which does not increase the presence observed labels. Hence, the total number of 'presence' observed labels becomes smaller than the number of the true 'presence' labels, which causes further sparsity in the data. We also find that models FP7 to FP12 have higher variance in these three subfigures.

In Figure 4.7, where the class balance is set to high ($\psi = 0.40$), our proposed models perform mostly better than the competitors, except model FP12, which is missing two overlaps. We find that the MSE of every simulated model decreases as the noise rates increase. For instance, when the false positive noise rate increases from 0.1 to 0.4 and the false negative rate is fixed at 0.1 (the first column), the MSE of the LR1 model decreases from approximately 0.5 to 0.3. This means that given the high class balance condition, the model retrieved is much closer to the true model when the noise rates are higher (such as FPR = 0.4 and FNR = 0.4). Comparatively, in the less noisy models, the retrieved model is significantly different from the true models, resulting in high MSE in the class probabilities. We still don't fully understand the mechanism behind these patterns.

Under this high class balance condition, our proposed models cannot always outperform the competitors, except model FP9, which performs best among FP7 to FP12 across these nine noise rate combinations. This may mean that type III overlap (overlaps between two noises) is less important than the other overlaps. The choice between setting this feature to be a false positive noise or a false negative noise can be achieved by the fitting process, assuming sufficient data is available. Moreover, model FP12 mostly performs worst among FP7 to FP12, which indicates that overlap type I and overlap type II are important. If we ignore them, the accuracy of the prediction will be reduced.

Figure 4.6 shows the model performance at medium class balance (0.25). Compared to Figures 4.5 and 4.7, under this class balance condition, our proposed models all perform better than the four competitors. Again, model FP9 (missing Type III overlap) performs closest to the true FP model, which indicates the overlaps between two noise probability functions are not very important. But for the overlaps involved in both the class balance and the noise rate, ignoring them will reduce prediction accuracy.

In short, the experiments above show that our proposed model performs better than the competing methods most of the time. We have to acknowledge that the missing overlaps impact the model performance in several class balance and noise rate combinations. This confirms that the overlap condition is a key factor in identifiability, which is consistent with our discussions in the identifiability section. Among these three types

of overlap, the overlaps between the two noise models appear to be less important than the other two, since this kind of model misspecification can be recovered by the fitting process. Therefore, if we can correctly specify the overlaps between class and noise, like in model FP9, we can achieve a good performance.

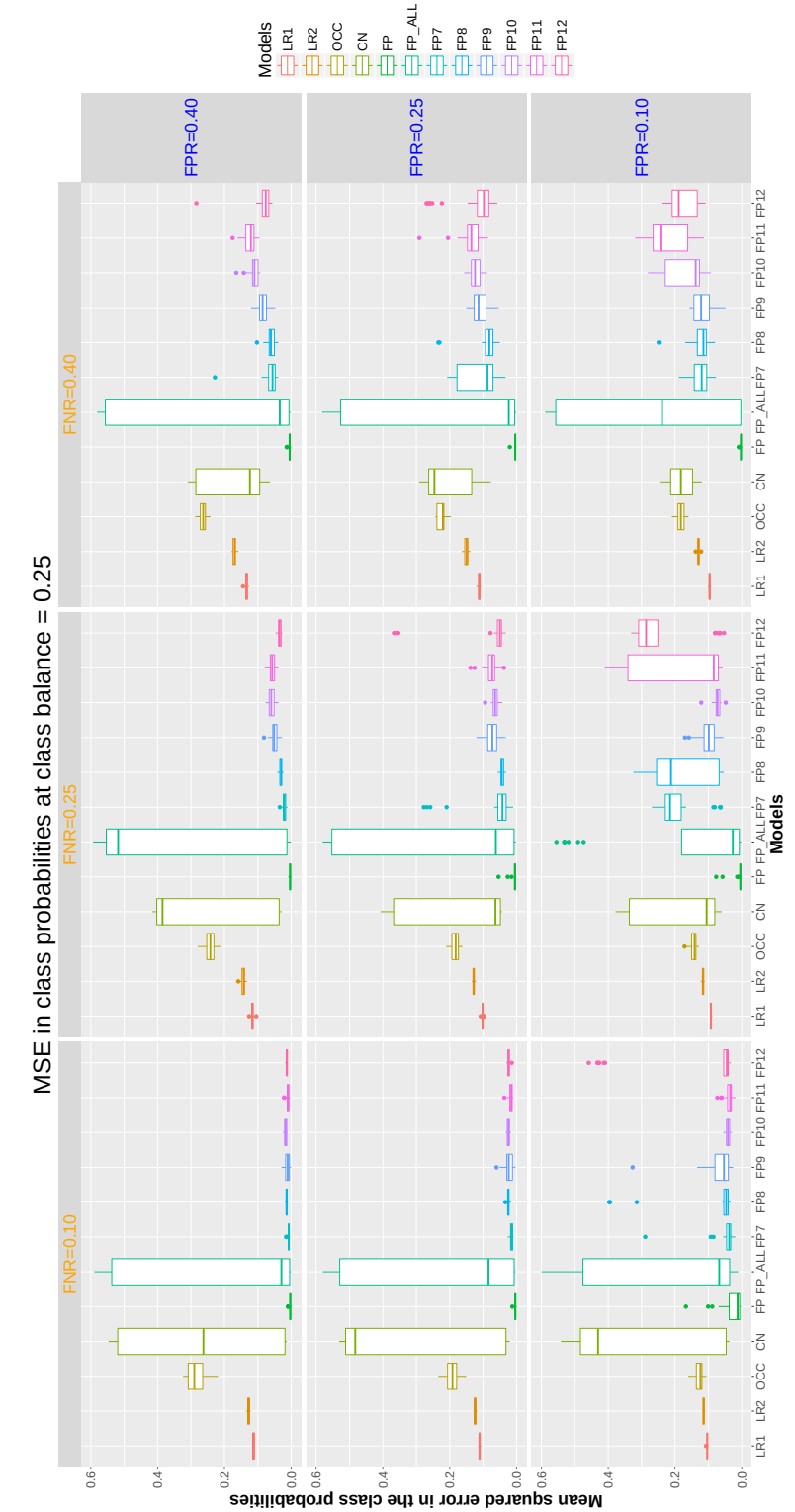

Figure 4.5: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.4) at class balance = 0.25. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

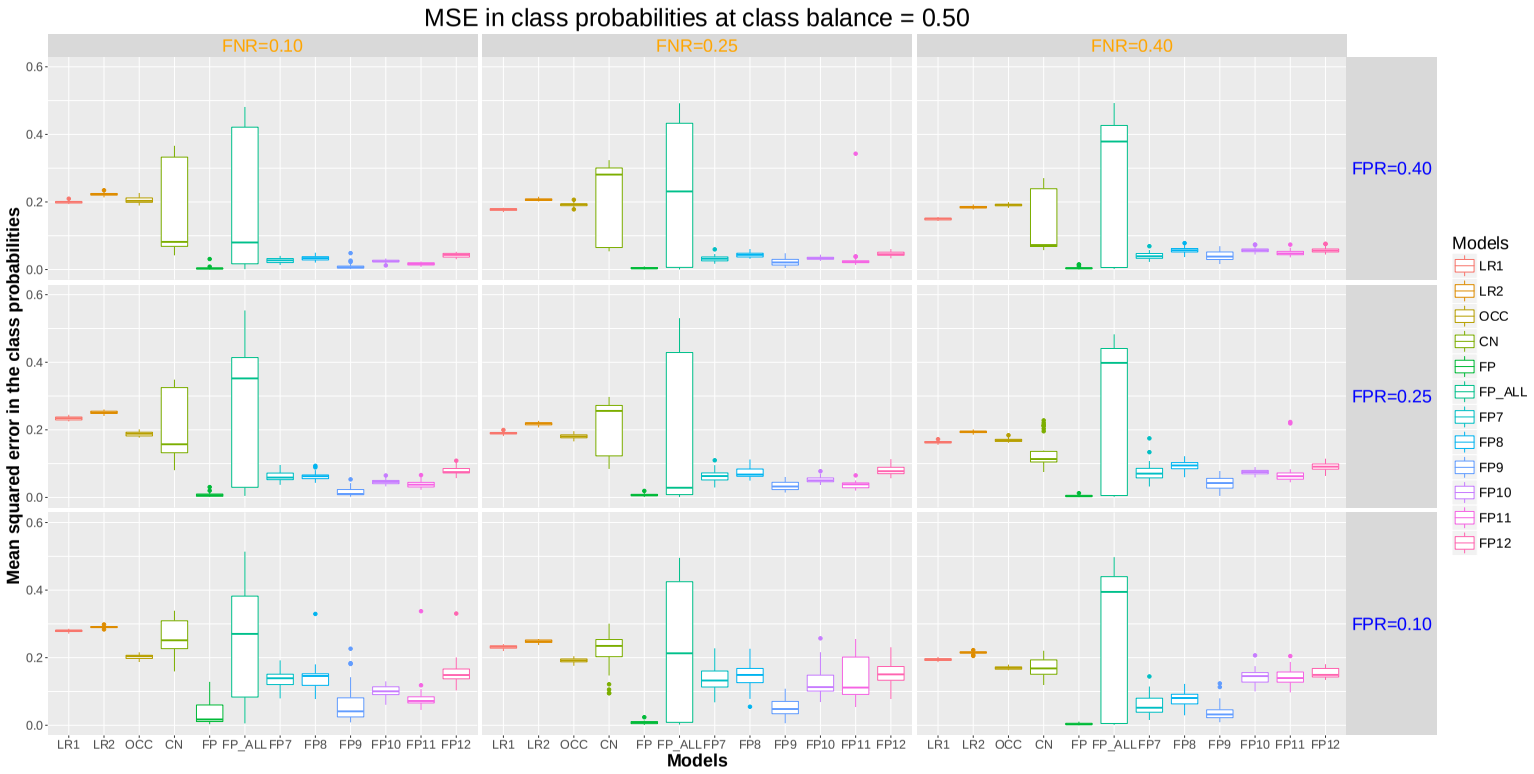

Figure 4.6: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.4) at class balance = 0.50. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

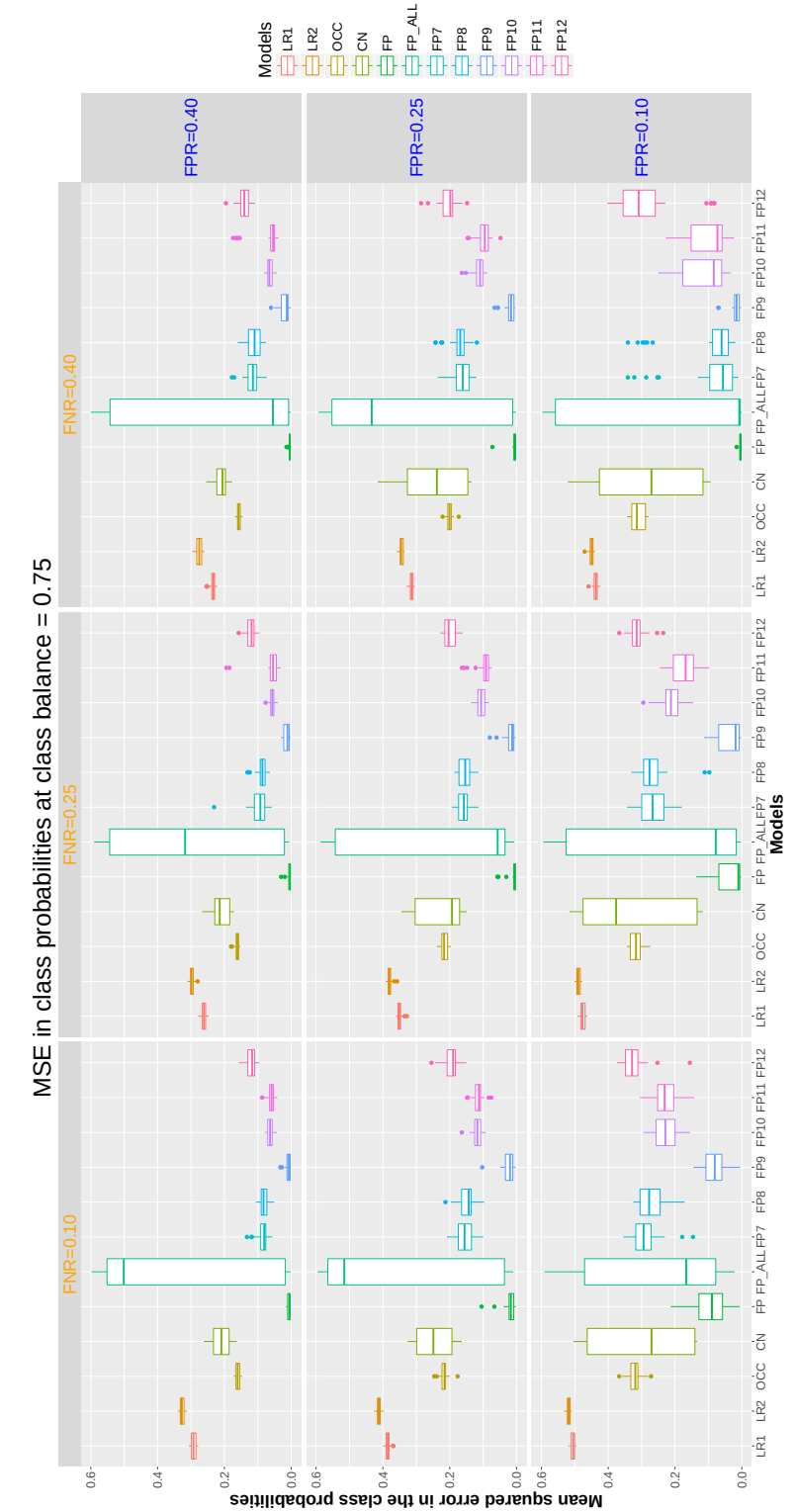

Figure 4.7: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.4) at class balance = 0.75. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

Chapter 5: Empirical Experiment

5.1 Acknowledgement

This chapter comes from our paper [6], contributed by Rebecca A. Hutchinson. I place it here for the consideration of comprehension.

5.2 Background

The eBird Reference Dataset consists of checklists of bird species observed during birding events in which citizen scientists report all of the species they observed [19]. We used data collected in 2012. Following previous analyses of eBird data [27, 7, 26], we chose to focus on stationary and traveling counts from California and New York in May, June, and July, during which time habitat associations are relatively stable.

Habitat Features	
Feature	Type
Human population from 2000 census	real
Housing percent vacant	real
Elevation	real
Average temperature	categorical
Precipitation	categorical
Percent of surrounding area covered in 15 land cover types	real
Observation Features	
Feature	Type
Day of year	real
Time of day	real
Effort in hours	real
Effort in distance travelled	real
Number of observers	real

Table 5.1: Features of models fit to the eBird data, taken from the eBird Reference Dataset.

The eBird Reference Dataset is distributed with both environmental (class) features and observation (noise) features. The features we considered included 11 class features (3 real, 2 categorical, 6 principal components of land cover measurements) and 5 noise features (all real-valued) and are listed in Table 5.1. We scaled continuous features to $N(0, 1)$. After removing records with missing values and outliers for the features, the California data contained 16,742 checklists and the New York data contained 11,982 checklists. For each state, we randomly selected 4000 checklists as a test set, 4000 checklists as a validation set, and used the remaining checklists for training.

The eBird Reference Dataset is also distributed with information about the species it contains, including whether or not one species is often confused with another. We selected species with this property for this analysis, since species confusions are a potential source of false positives in the eBird data [26]. In addition, we limited the species pool to species observed in at least 10% of the checklists in the state. The pool of species is listed in the Table 5.3.

Simulated Species	Variables in Submodels	Average rates
1	$\psi = f(ELEVATION, HUMAN.POPULATION)$	0.55
	$\rho = g(DAY)$	0.05
	$\eta = h(EFFORT.HRS, TIME)$	0.2
2	$\psi = f(ELEVATION, HUMAN.POPULATION)$	0.55
	$\rho = g(DAY, TIME)$	0.1
	$\eta = h(EFFORT.HRS, TIME)$	0.4

Table 5.2: Model forms and average rates of class balance/occupancy, false positives, and false negatives for the species simulated from eBird features.

California						
Common Name	Scientific Name	Report frequency	Model selected	Est. avg. occ. prob	Est. avg. false. neg. prob	Est. avg. false. pos. prob.
American crow	Corvus brachyrhynchos	0.34	12	0.52	0.40	0.037
Song sparrow	Melospiza melodia	0.32	7	0.52	0.37	0.013
Red-winged blackbird	Agelaius phoeniceus	0.23	17	0.32	0.28	0.029
Nuttalls woodpecker	Picoides nuttallii	0.18	11	0.57	0.42	0.0058
Western kingbird	Tyrannus verticalis	0.15	18	0.38	0.36	0.031
Western wood pewee	Contopus sordidulus	0.13	13	0.43	0.39	0.038
Northern rough-winged swallow	Stelgidopteryx serripennis	0.13	9	0.46	0.50	0.016
Sim1	Simulus primus	0.48	7	0.56	0.17	0.054
Sim2	Simulus secondus	0.38	20	0.53	0.40	0.099
New York						
Common Name	Scientific Name	Report frequency	Model selected	Est. avg. occ. prob	Est. avg. false. neg. prob	Est. avg. false. pos. prob.
Red-winged blackbird	Agelaius phoeniceus	0.59	18	0.31	0.30	0.077
Song sparrow	Melospiza melodia	0.56	5	0.50	0.15	0.021
American crow	Corvus brachyrhynchos	0.49	17	0.72	0.28	0.035
Red-eyed vireo	Vireo olivaceus	0.26	20	0.28	0.22	0.095
Wood thrush	Hylocichla mustelina	0.21	5	0.42	0.46	0.012
Eastern wood pewee	Contopus virens	0.14	2	0.40	0.32	0.051
Indigo bunting	Passerina cyanea	0.14	1	0.30	0.0	0.061
Veery	Catharus fuscescens	0.13	13	0.47	0.30	0.024
Sim1	Simulus primus	0.48	15	0.62	0.20	0.048
Sim2	Simulus secondus	0.40	20	0.54	0.40	0.11

Table 5.3: Species modeled with the eBird Reference Dataset. The table also indicates the overall freqency of positive reports of the species in the data, the model selected for the FP method, and the estimated average occupancy (class balance), false negative, and false positive rates. Note that these results deserve further evaluation from an ecological perspecitve; for example, a false negative rate of 0 for the Indigo bunting may be a sign of model overfitting and not a realistic estimate.

Model	Variables in Noise Models
1	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h()$
2	$\rho = g(EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY)$
3	$\rho = g(DAY, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(EFFORT.HRS)$
4	$\rho = g(DAY, EFFORT.HRS, N.OBS, TIME)$ $\eta = h(EFFORT.DIST)$
5	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, TIME)$ $\eta = h(N.OBS)$
6	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS)$ $\eta = h(TIME)$
7	$\rho = g(EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.HRS)$
8	$\rho = g(EFFORT.HRS, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.DIST)$
9	$\rho = g(EFFORT.HRS, EFFORT.DIST, TIME)$ $\eta = h(DAY, N.OBS)$
10	$\rho = g(EFFORT.HRS, EFFORT.DIST, N.OBS)$ $\eta = h(DAY, TIME)$
11	$\rho = g(DAY, N.OBS, TIME)$ $\eta = h(EFFORT.HRS, EFFORT.DIST)$
12	$\rho = g(DAY, EFFORT.DIST, TIME)$ $\eta = h(EFFORT.HRS, N.OBS)$
13	$\rho = g(DAY, EFFORT.DIST, N.OBS)$ $\eta = h(EFFORT.HRS, TIME)$
14	$\rho = g(DAY, EFFORT.HRS, TIME)$ $\eta = h(EFFORT.DIST, N.OBS)$
15	$\rho = g(DAY, EFFORT.HRS, N.OBS)$ $\eta = h(EFFORT.DIST, TIME)$
16	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST)$ $\eta = h(N.OBS, TIME)$
17	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$
18	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.HRS, EFFORT.DIST, TIME)$
19	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS)$
20	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.DIST, N.OBS, TIME)$
21	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.HRS, N.OBS, TIME)$

Table 5.4: Models considered for the eBird species. Models 1-16 assign each of the five noise features to exactly one of the two submodels. Models 17-21 assign all five noise features to one submodel and all except one feature to the other submodel.

We also created two simulated species using the eBird features, which are more realistic than standard normal or binary features (see Table 5.2). The class models for both species depended on two real-valued class features. The noise models for the first simulated species (*Sim1*) had disjoint sets of features (two for false negatives and one for false positives). The noise models for the second simulated species (*Sim2*) shared one feature (two features for each model with one overlapping). Both species had true occupancy rates near 55%. The average false negative rates were 20% and 40%, and the average false positive rates were 5% and 10%, for *Sim1* and *Sim2* respectively.

5.3 Experiment Setting

In the simulated experiments, we achieved identifiability by specifying each feature function in accordance with the data-generating mechanisms. In the eBird data, we faced a model selection problem in assigning features to feature functions, in particular for the noise models. We fit 21 different models to each of the real and simulated eBird species, all of which met the identifiability conditions discussed above (at least one unique feature in each feature function). In each model, the class feature function included all of the habitat features. The false positive and false negative feature functions partitioned the noise features differently. Models 1-16 assigned each noise feature to exactly one of the noise models. Models 17-21 included all noise features in one noise model and all noise features except one in the other noise model (see Table 5.4). Each of these 21 models has a symmetric analog, so in each case we chose between the pair of symmetric models by selecting the one in which the false positive rate (ρ) was less than the detection rate ($1 - \eta$). That is, we used the constraint that observers are more likely to detect the correct species than to misidentify it.

We compared against the same set of methods as in the simulated data experiments. For the simulated species, we can refer to the data-generating models to evaluate the models. For the real species, we do not have access to 'ground truth' about the species true presence or absence while a checklist was collected, but we can examine the differences in the class probabilities predicted by different methods.

5.3.1 Result and Analysis

For *Sim1*, the models selected in each state based on the validation sets were not congruent with the data-generating mechanisms. None of the models were fully correct, so by 'congruent' we mean that the correct features were included in the noise models in combinations such that either the model as written or the symmetric analog could represent the generating model if the irrelevant features were given coefficients of 0. Interestingly though, on the test sets, the selected models had lower MSE on ψ compared with most other *FP* models and the *OCC* model. The selected models also had lower MSE on ψ, ρ, and η than the *CN* model (see Table 5.5 and Table 5.7). Therefore, despite lack of a perfect representation of the data-generating model, the predictions were superior to alternative methods. For *Sim2*, the models selected in each state were congruent with the data-generating mechanisms (see Table 5.6 and Table 5.8). Again, many of the 21 *FP* models, including those unable to represent the data-generating mechanism, outperformed *OCC* and *CN*. In California, the vast majority of the *FP* models outperform the *OCC* and *CN* alternatives, whereas only a subset of the *FP* models in New York are clearly superior to the alternatives. This may be due to greater sample size; the training dataset for California is roughly twice as large as that for New York.

For the real species, estimated false positive rates ranged from 0.0058 to 0.095 (see supplement, Table S4). These low rates are likely due to eBird's quality control measures. Given the low rates, some *FP* models made predictions similar to the *OCC* models that ignored false positives, like *Picoides nuttallii* in California. For other species, like *Vireo olivaceus* in New York, the predictions from *FP* and *OCC* models differ more, suggesting overprediction of the species by the *OCC* model (Figure 5.1). Some models, like the *OCC* model for *Vireo olivaceus* in New York, also showed signs of overfitting in the form of boundary estimates for the class probabilities. Figure 5.1 also compares the *FP* predictions to the *CN* predictions for the class probabilities; the degree of correlation between these predictions varied across species. We expect this variation is due to differences in the importance of the noise features and to identifiability issues with the *CN* method.

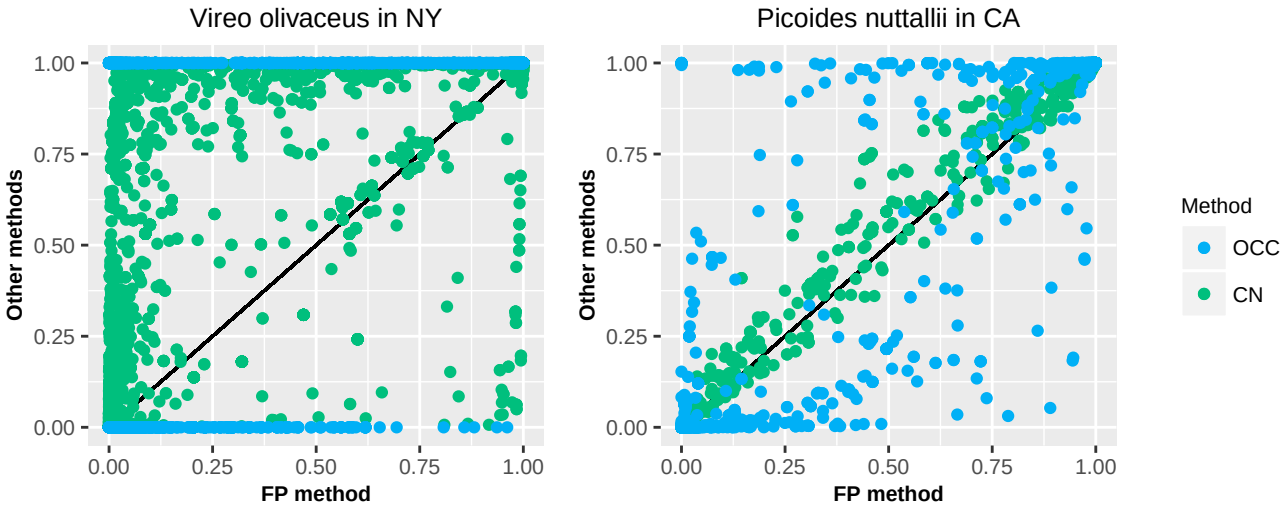

Figure 5.1: Comparison of predicted class probabilities from the *FP* method on the x-axis versus the *OCC* and *CN* methods on the y-axis. The estimated average false positive rates are 0.095 and 0.0058 for *Vireo olivaceus* and *Picoides nuttallii* respectively.

Model	Validation NLL	Test NLL	Test MSE on ψ	Test MSE on η	Test MSE on ρ
1	2406.6	2278.1	0.055	0.00042	0.0076
2	2375.1	**2039.6**	**0.023**	0.000079	**0.000073**
3	2449.1	2438.1	0.078	0.018	0.0078
4	2398.8	2281.3	0.056	0.00037	0.0076
5	2391	2264	0.053	0.00036	0.0076
6	2602.6	2264.2	0.043	0.03	0.00074
7	**2238.6**	2065.4	0.023	0.018	0.00023
8	2394.6	2046.8	0.024	0.000046	0.00011
9	2383.1	2045.3	0.024	0.000054	0.000098
10	2478	2215.4	0.034	0.038	0.0022
11	2432.3	2300.1	0.048	0.039	0.0025
12	2441.3	2409.8	0.072	0.019	0.0078
13	2404	2053	0.025	**0.000021**	0.00014
14	2383.6	2268.8	0.055	0.00031	0.0076
15	2329.6	2089.9	0.026	0.019	0.00025
16	2430.5	2154	0.028	0.028	0.0007
17	2458	2071.2	0.027	0.00012	0.00031
18	2345.7	2091.7	0.03	0.00011	0.00032
19	2396.1	2058	0.025	0.00011	0.00018
20	2445	2079.7	0.028	0.00014	0.00032
21	2425.3	2064.3	0.026	0.00077	0.0003
OCC	-	-	0.093	0.00089	-
CN	-	-	0.07	0.086	0.0078

Table 5.5: Model selection results for *Sim1* in CA. Bold-numbered models (or their symmetric analogs) are consistent with the true data-generating model. Bold values indicate the best value in each column. Here, model 7 is chosen using the validation set even though it is not consistent with the data-generating model. On the test set, model 2 performs best on the class model, but model 7 is nearly tied with it. All 21 FP models outperform the OCC model, and all but two outperform the CN model on ψ.

Model	Validation NLL	Test NLL	Test MSE on ψ	Test MSE on η	Test MSE on ρ
1	2174.7	2170.9	0.018	0.0028	0.034
2	2124.3	2218.2	0.048	0.0007	0.019
3	2400.7	2452.7	0.09	0.044	0.036
4	2179.4	2169.2	**0.017**	0.0028	0.034
5	2178.6	2173	0.019	0.0028	0.034
6	2549.5	2821.9	0.16	0.061	0.02
7	2338.7	2545.4	0.1	0.04	0.019
8	2128.4	2223.8	0.05	0.00067	0.019
9	2121.8	2204.2	0.045	0.00064	0.019
10	2984.8	3146.4	0.12	0.084	0.001
11	2979.9	3143.8	0.12	0.085	0.0013
12	2413.4	2467.1	0.086	0.046	0.036
13	2125.6	2209.5	0.046	0.0006	0.019
14	2181.9	2170.6	0.018	0.0027	0.034
15	2371.7	2639.1	0.12	0.042	0.019
16	2514.8	2705	0.14	0.057	0.02
17	2106.4	2119.2	0.024	**0.00027**	0.000096
18	2100.3	2117.4	0.023	0.00043	0.000088
19	2157.4	2261.5	0.059	0.001	0.018
20	**2099.4**	**2116.5**	0.023	0.00044	0.0001
21	2100.3	2116.7	0.023	0.00043	**0.000064**
OCC	-	-	0.15	0.011	-
CN	-	-	0.17	0.11	0.034

Table 5.6: Model selection results for *Sim2* in CA. Bold-numbered models (or their symmetric analogs) are consistent with the true data-generating model. Bold values indicate the best value in each column. Here, model 20 is chosen using the validation set, which is consistent with the data-generating model. It also has the best negative log-likelihood on the test set, though model 4 does slightly better on MSE of the class probabilities. All but one of the FP models outperform the OCC model, and they all outperform the CN model on ψ.

Model	Validation NLL	Test NLL	Test MSE on ψ	Test MSE on η	Test MSE on ρ
1	2113	2435.6	0.061	0.0011	0.009
2	2113.7	2454.4	0.065	0.0003	**0.00024**
3	2061.4	2360.7	0.045	0.021	0.0094
4	2115.9	2438.1	0.061	0.0011	0.0089
5	2129.8	2444.6	0.062	0.001	0.009
6	2095.9	2469.3	0.062	0.027	0.00092
7	2043.8	**2317.4**	**0.04**	0.021	0.00046
8	2116.7	2456.3	0.066	0.0003	0.00028
9	2122.7	2456.5	0.066	0.00025	0.00056
10	2111.4	2507.3	0.067	0.028	0.0099
11	2107.7	2475.1	0.062	0.027	0.0098
12	2058.4	2374.7	0.047	0.021	0.0095
13	2124.6	2456.3	0.066	**0.00025**	0.00055
14	2129.7	2443.8	0.062	0.001	0.009
15	**2041.7**	2327.9	0.042	0.021	0.00083
16	2093.2	2440.4	0.058	0.027	0.00056
17	2143.9	2484.8	0.07	0.00036	0.001
18	2144.6	2462.6	0.067	0.00048	0.0011
19	2121.8	2448.1	0.065	0.00051	0.00062
20	2143.8	2467.8	0.068	0.00051	0.001
21	2144.2	2467.1	0.068	0.00051	0.001
OCC	-	-	0.05	0.0014	-
CN	-	-	0.055	0.057	0.0095

Table 5.7: Model selection results for *Sim1* in NY. Bold-numbered models (or their symmetric analogs) are consistent with the true data-generating model. Bold values indicate the best value in each column. Here, model 15 is chosen using the validation set even though it is not consistent with the data-generating model. On the test set, model 7 performs best on the class model, even though it is not consistent with the data-generating model. Four of the FP models outperform the OCC and CN models on ψ.

Model	Validation NLL	Test NLL	Test MSE on ψ	Test MSE on η	Test MSE on ρ
1	2223.6	2132.6	0.037	0.0025	0.045
2	2072.1	2037.8	0.029	0.00017	0.023
3	2448.2	2367	0.11	0.042	0.046
4	2227.9	2141.3	0.039	0.0025	0.045
5	2223.6	2128.2	0.036	0.0025	0.044
6	2378.9	2310.5	0.076	0.053	0.025
7	2342	2265.4	0.08	0.039	0.023
8	2072	2042.6	0.03	0.00016	0.023
9	2076	2029.5	0.026	0.00012	0.022
10	2496.4	2424.7	0.036	0.085	0.0017
11	2482.7	2425.4	0.036	0.085	0.0017
12	2462.7	2378.5	0.11	0.042	0.046
13	2072.1	2033.9	0.028	**0.00011**	0.022
14	2225.5	2136.9	0.039	0.0024	0.045
15	2358.1	2272.1	0.08	0.04	0.023
16	2370.7	2305	0.076	0.051	0.025
17	2071.5	1957.9	0.0093	0.00013	0.00025
18	2071.8	**1957.7**	0.0091	0.00014	0.00024
19	2075.3	2037.4	0.029	0.00021	0.022
20	**2067.2**	1959.1	0.094	0.00016	**0.00015**
21	2071.4	1957.8	**0.0091**	0.00013	0.00025
OCC		- -	0.034	0.016	-
CN		- -	0.085	0.11	0.049

Table 5.8: Model selection results for *Sim2* in NY. Bold-numbered models (or their symmetric analogs) are consistent with the true data-generating model. Bold values indicate the best value in each column. Here, model 20 is chosen using the validation set, which is consistent with the data-generating model. Model 18 has the best negative log-likelihood on the test set, and model 21 performs best in terms of MSE on $\psi$for the test set, but all four consistent models have very similar performance. Eight of the FP models outperform the OCC model, and 19 of them outperform the CN model onψ.

Chapter 6: Conclusion and Future Works

In this work, I have discussed the approach of treating species distribution models as a classification problem with class conditional label noise in the context of citizen science projects. I have analyzed the model performance of this approach without considering model misspecification first, in the scenarios of varying class balance, two noise rates, feature distributions, and the sample complexity. In all of these settings, the approach I presented outperforms all candidate methods. I have discussed the impact on model performance of introducing three kinds of model misspecification (link function misspecification, adding or missing covariates, and missing overlaps), and also addressed several important considerations for the model selection process. In addition, we have applied this approach to eBird data and compared model performance with other candidate methods. The results show that the model I proposed performs best among all simulated methods.

In contrast to other approaches, the NNSF (Noise with Noise-Specific Features) model includes additional noise features, involving both false positives and false negatives, separated from class features. This model is different from other methods that ignore labeling noise or only investigate false negatives but ignore false positives, which causes the underestimation of the species distribution. Moreover, since the noise rate setting is not necessary to be symmetric in the model, my approach is more challenging than methods that only consider symmetric noise rates.

Due to the complexity of this proposed model, it is important that enough data be available for fitting. For projects where only small datasets are available, simpler methods will privide a better solution. In order to handle the identifiability problem, we suggest that there should be at least one unique feature in each feature function. As with similar models in the ecology literature, identifiability also relies on the fidelity of the link functions. The result of the experiments show that link function misspecification is a significant source of prediction inaccuracy.

I have extended the study of our paper to model misspecification. I found that missing or adding one feature will not cause a large decline in model performance. But if we include all features in all probabilities, the proposed model is actually reduced to the CN model, which considers the noise structures but ignores the contribution of noise features. From the overlap misspecification experiments, I found that it is safe to consider noise features as only false positive or false negative. But if a class feature is also a noise feature, mis-specifying this overlap will lead to performance declines and fluctuation in some noise level and class balance conditions. Based on the results from the simulation experiments, I suggest that users implement our model if sufficient data is available, arranging features carefully with relevant ecological knowledge and selecting proper link functions.

The empirical experiments explored model selection among a variety of feature combinations for the noise feature functions. For the simulated species with low false positive rates, the selected model, which gave better predictions of class probabilities on the test than most candidates, was not the same as the one used in the data-generating process. We hypothesize that the low false positive rates ($\rho = 5\%$) contributed to selecting an inconsistent model, though perhaps these models could be recovered with more data. For the simulated species with higher false positive rates ($\rho = 10\%$), the selected model was consistent with the data-generating process. This selected model outperformed other *FP* models as well as *OCC* and *CN*. For the real species, due to the low false positive rates of the data, our proposed model performs consistently with the *OCC* models and in some cases the *CN* models. Given the uncertainty around model selection for the simulated species with the lowest false positive rates, caution is warranted in interpreting the selected models for the real species.

In this work, I considered only binary labels for only one particular species during each of the modeling processes. Hence, I cannot get information about the impact from other species, especially predators and competitors, since information about them is not available in the dataset as features. The approach I presented treats the information about other species as a hidden feature that was missed in the model, to simulate the impact of model misspecification. In future work, I will extend our approach to multi-

class classification models with class-conditional label noise, embedding the interaction between species into a deep neural network to model the species distribution.

Additionally, I plan to investigate the model performance with other functional types in the class model and two noise models. I will consider using different link functions in both data generating and model fitting processes. In addition to Gaussian distribution and Bernoulli distribution, I will focus more on other feature distributions, such as Poisson distribution, which is widely used in ecology and statistics.

Bibliography

[1] Wei Bi, Liwei Wang, James T Kwok, Zhuowen Tu, Hong Kong, United States, and United States. Learning to Predict from Crowdsourced Data. In *Uncertainty in Artificial Intelligence: Proceedings of the Thirtieth Conference*, 2014.

[2] Vikas C. Raykar et al. Learning from crowds. *Journal of Machine Learning Research*, 11(1297-1322), April 2010.

[3] Benoît Frénay and Michel Verleysen. Classification in the Presence of Label Noise: a Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[4] Antoine Guisan and Wilfried Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 89:993–1009, 2005.

[5] Robert J. Hijmans and Jane Elith. Species distribution modeling with r. *R package version*, 0.8-11, 2011.

[6] Rebecca A. Hutchinson, Liqiang He, and Sarah C. Emerson. Species distribution modeling of citizen science data as a classification problem with class-conditional noise. *AAAI*, 2017.

[7] Rebecca A. Hutchinson, Li-Ping Liu, and Thomas G. Dietterich. Incorporating Boosted Regression Trees into Ecological Latent Variable Models. In *Proceedings of the Twenty-fifth Conference on Artificial Intelligence*, 2011.

[8] Jonas Knape and Fränzi Korner-Nievergelt. Estimates from non-replicated population surveys rely on critical assumptions. *Methods in Ecology and Evolution*, 6:298–306, dec 2014.

[9] Jonas Knape and Franzi Korner-Nievergelt. Estimates from nonreplicated population surveys rely on critical assumptions. *Methods in Ecology and Evolution*, 6.3(298-306), 2015.

[10] Neil D. Lawrence and Bernhard Scholkopf. Estimating a Kernel Fisher Discrimant in the Presence of Label Noise. *Proceedings of the 18th International Conference on Machine Learning*, pages 306–313, 2001.

[11] S. R. Lele, M. Moreno, and E. Bayne. Dealing with detection error in site occupancy surveys: what can we do with a single survey? *Journal of Plant Ecology*, 5(1):22–31, January 2012.

[12] Subhash R. Lele, Monica Moreno, and Erin Bayne. Dealing with detection error in site occupancy surveys: what can we do with a single survey? *Journal of Plant Ecology*, 5.1(22-31), 2012.

[13] Yunlei Li, Lodewyk F.A. Wessels, Dick de Ridder, and Marcel J.T. Reinders. Classification in the presence of class noise using a probabilistic Kernel Fisher method. *Pattern Recognition*, 40(12):3349–3357, 2007.

[14] Darryl I. MacKenzie, James D. Nichols, Gideon B. Lachman, Sam Droege, J. Andrew Royle, and Catherine A. Langtimm. Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One. *Ecology*, 83(8):2248–2255, August 2002.

[15] et al MacKenzie, Darryl I. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83.8:2248–2255, 2002.

[16] Naresh Manwani, P S Sastry, and Senior Member. Noise Tolerance Under Risk Minimization. *IEEE TRANSACTIONS ON CYBERNETICS*, 43(3):1146–1151, 2013.

[17] Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong, and Robert C. Williamson. Learning from Corrupted Binary Labels via Class-Probability Estimation. *Journal of Machine Learning Research*, 37, 2015.

[18] David A. Miller, James D. Nichols, Brett T. McClintock, Evan H. Campbell Grant, Larissa L. Bailey, and Linda A. Weir. Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, 92(7):1422–1428, 2011.

[19] M. Arthur Munson, Kevin Webb, Daniel Sheldon, Daniel Fink, Wesley M. Hochachka, Marshall Iliff, Mirek Riedewald, Daria Sorokina, Brian Sullivan, Christopher Wood, and Steve Kelling. The ebird reference dataset, version 4.0. Technical report, 2012.

[20] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with Noisy Labels. *Advances in neural information processing systems*, pages 1196–1204, 2013.

[21] Andrew J. Royle and William A. Link. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87.4(835-841), 2006.

[22] J. Andrew Royle and William A. Link. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841, 2006.

[23] Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with Asymmetric Label Noise : Consistency and Maximal Denoising. *JMLR: Workshop and Conference Proceedings*, 30:1–23, 2013.

[24] Péter Sólymos and Subhash R. Lele. Revisiting resource selection probability functions and single-visit methods: Clarification and extensions. *Methods in Ecology and Evolution*, 7:196–205, 2015.

[25] Brian L. Sullivan, Jocelyn L. Aycrigg, Jessie H. Barry, Rick E. Bonney, Nicholas Bruns, Caren B. Cooper, Theo Damoulas, André A. Dhondt, Tom Dietterich, Andrew Farnsworth, Daniel Fink, John W. Fitzpatrick, Thomas Fredericks, Jeff Gerbracht, Carla Gomes, Wesley M. Hochachka, Marshall J. Iliff, Carl Lagoze, Frank A. La Sorte, Matthew Merrifield, Will Morris, Tina B. Phillips, Mark Reynolds, Amanda D. Rodewald, Kenneth V. Rosenberg, Nancy M. Trautmann, Andrea Wiggins, David W. Winkler, Weng Keen Wong, Christopher L. Wood, Jun Yu, and Steve Kelling. The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, 169:31–40, 2014.

[26] Jun Yu, Rebecca A. Hutchinson, and Weng-keen Wong. A Latent Variable Model for Discovering Bird Species Commonly Misidentified by Citizen Scientists. *Proceedings of the 29th National Conference on Artificial Intelligence*, pages 500–506, 2014.

[27] Jun Yu, Weng-Keen Wong, and Rebecca A. Hutchinson. Modeling Experts and Novices in Citizen Science data for Species Distribution Modeling. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010)*, 2010.

APPENDICES

Appendix A: Supplementary Experiment Results

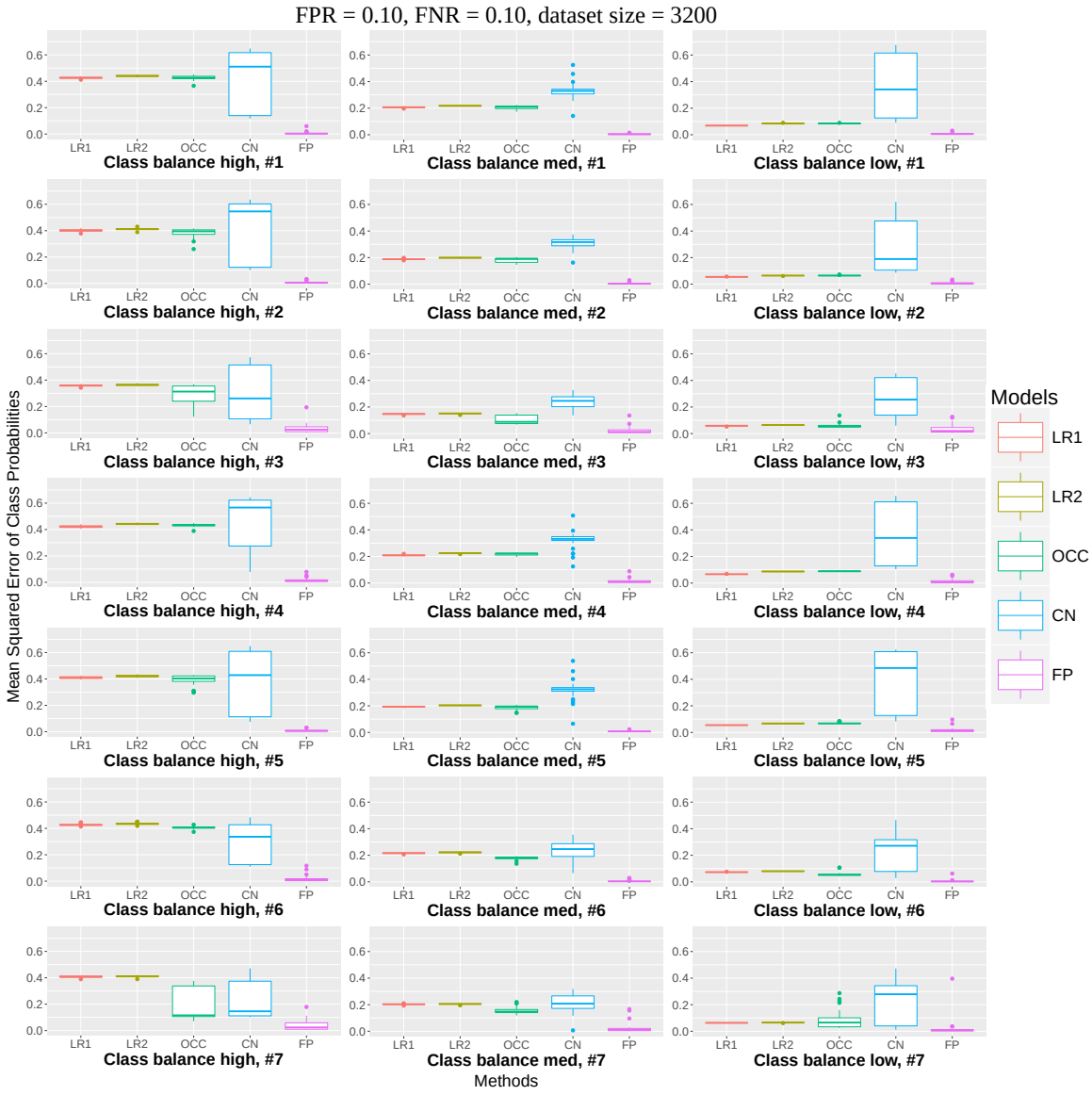

Figure A.1: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.10 and FNR = 0.10. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

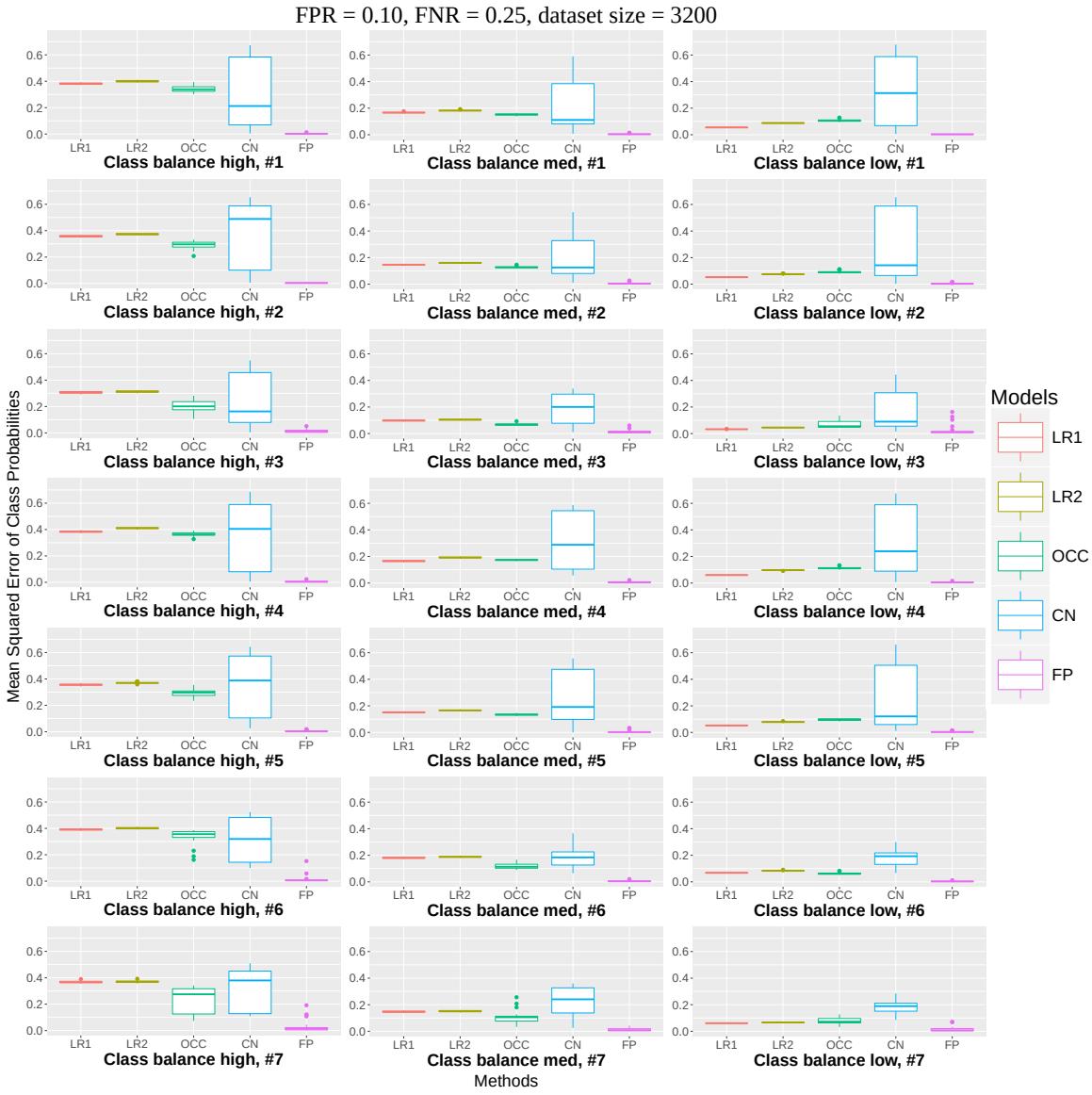

Figure A.2: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.10 and FNR = 0.25. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

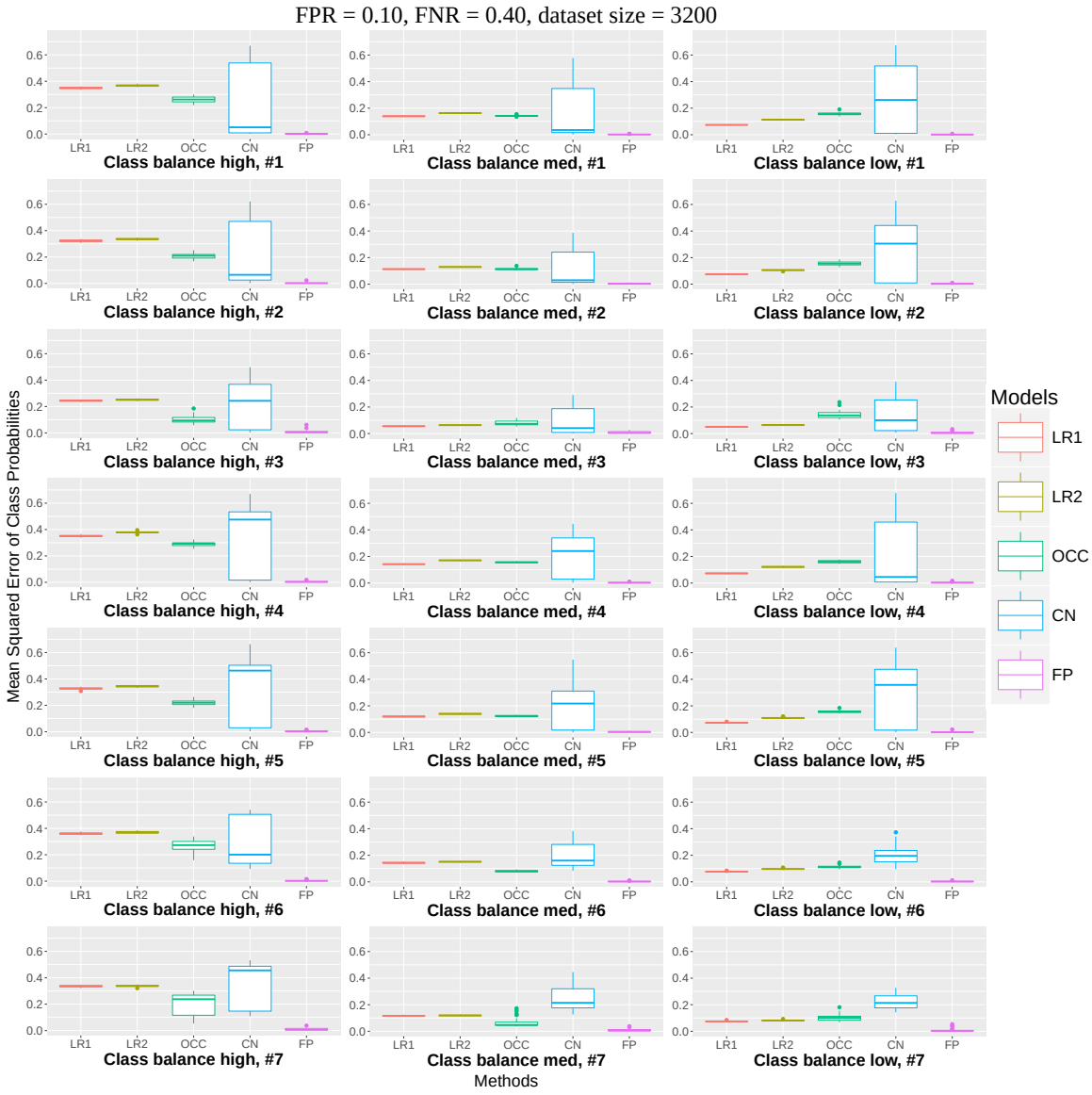

Figure A.3: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.10 and FNR = 0.40. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

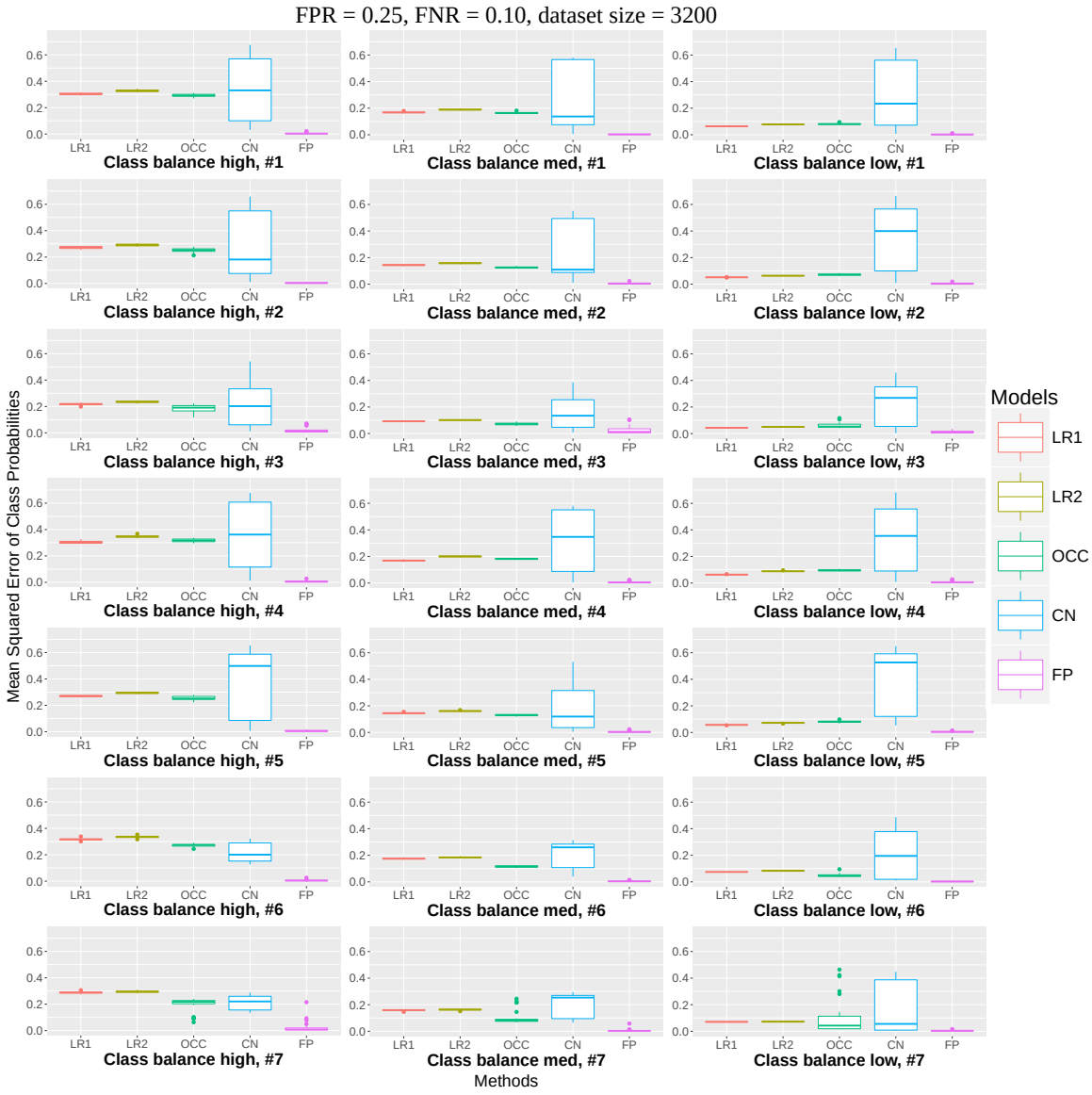

Figure A.4: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.25 and FNR = 0.10. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

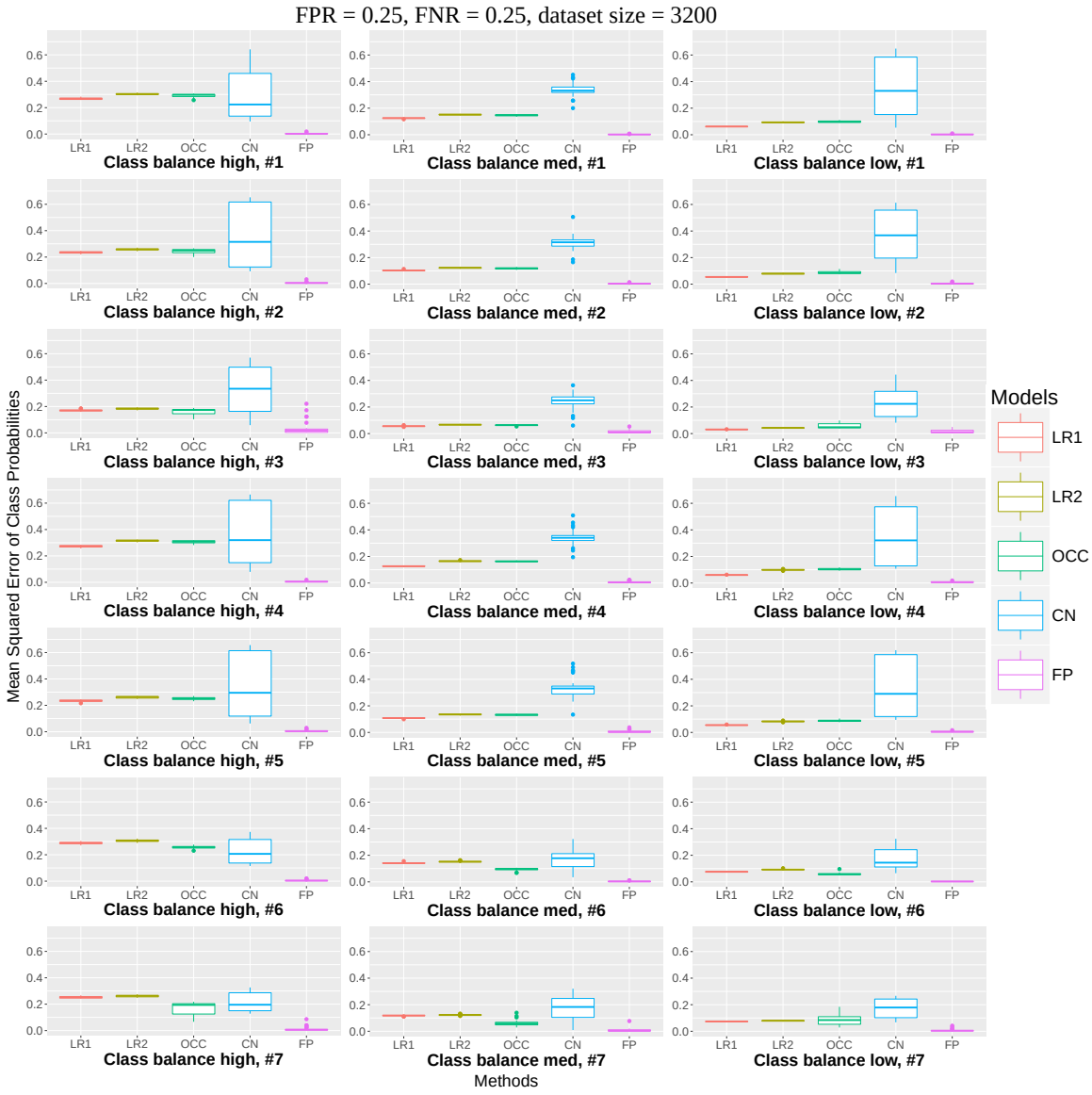

Figure A.5: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.25 and FNR = 0.25. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

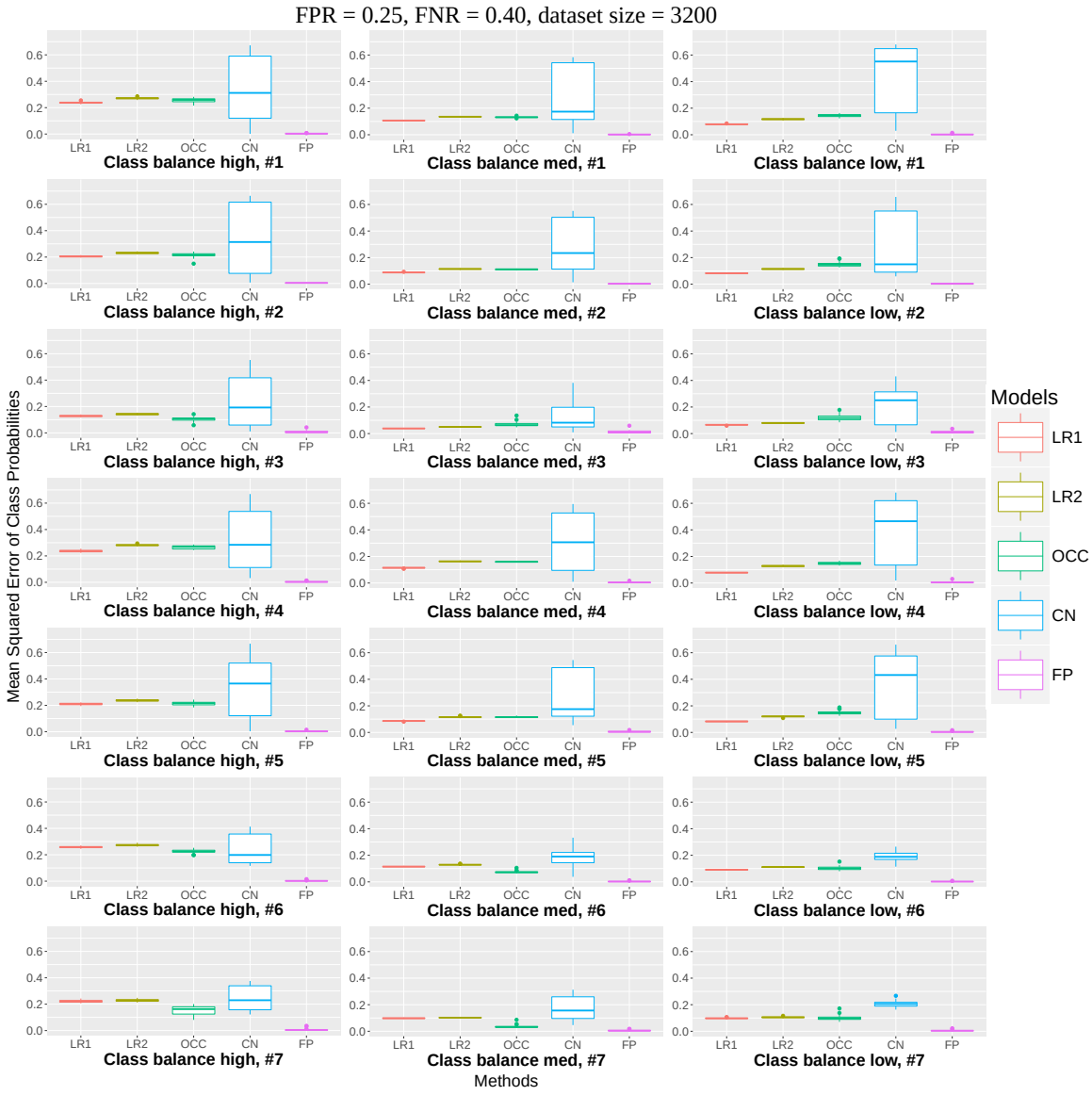

Figure A.6: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.25 and FNR = 0.40. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

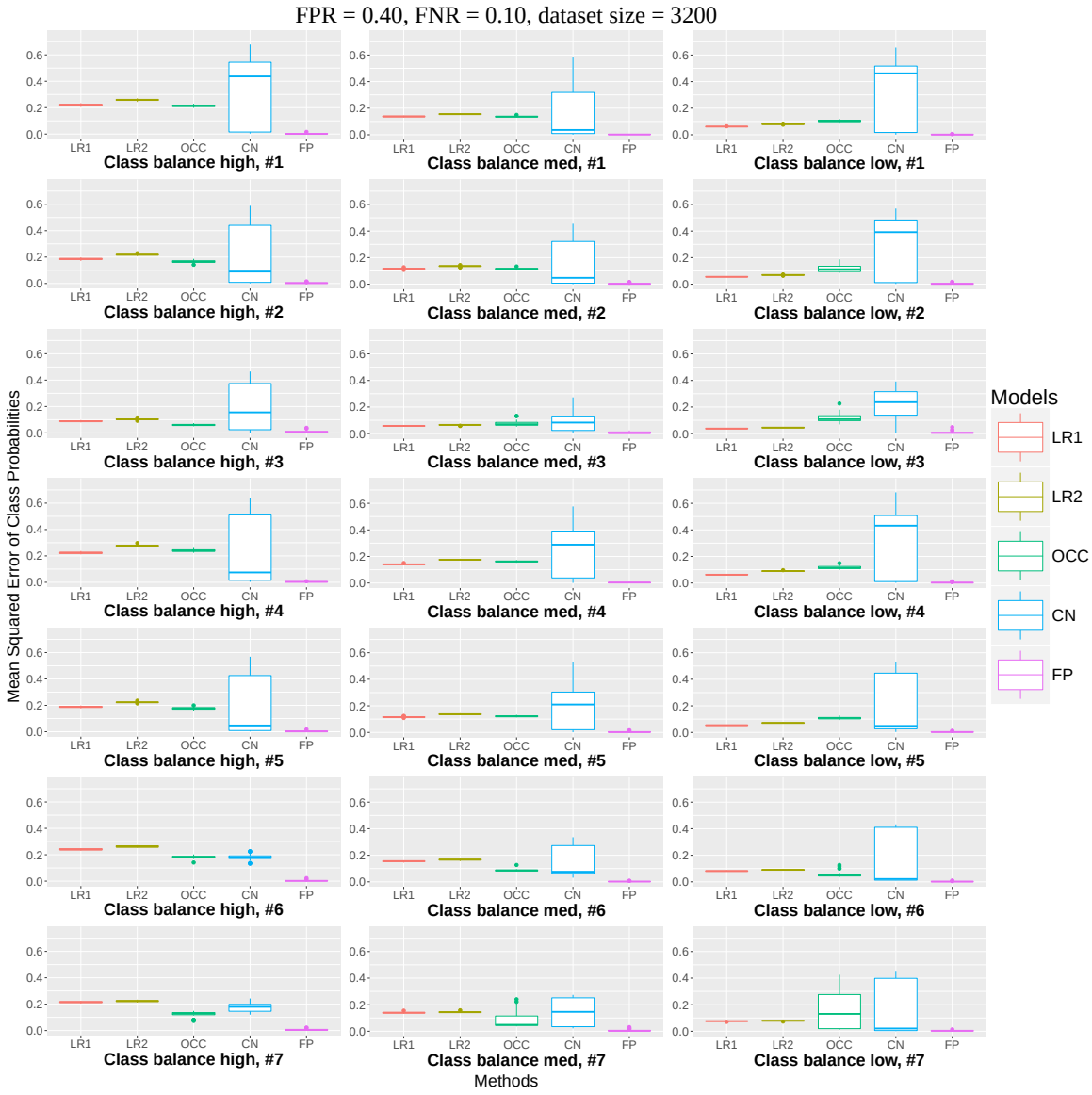

Figure A.7: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.40 and FNR = 0.10. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

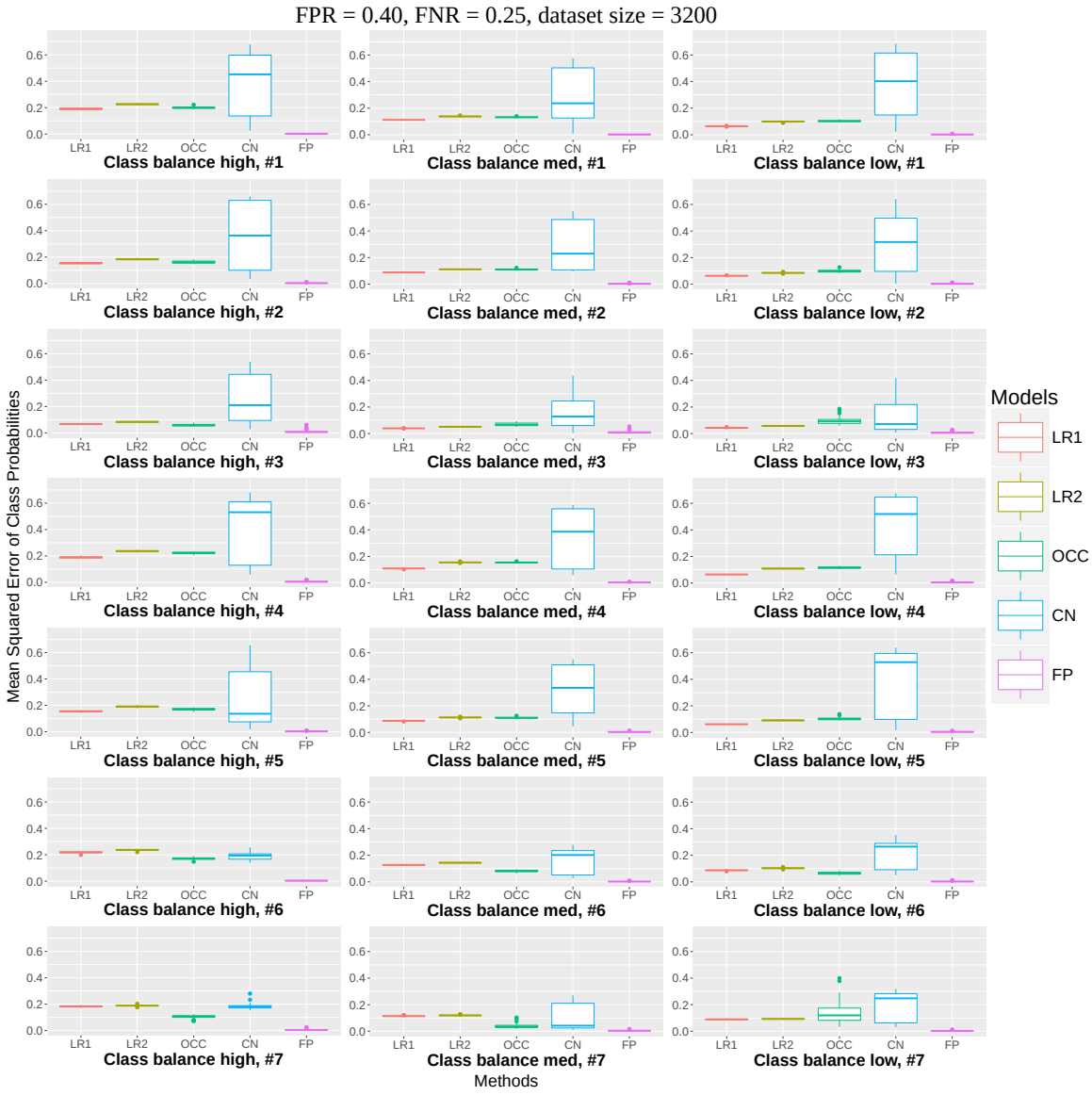

Figure A.8: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.40 and FNR = 0.25. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

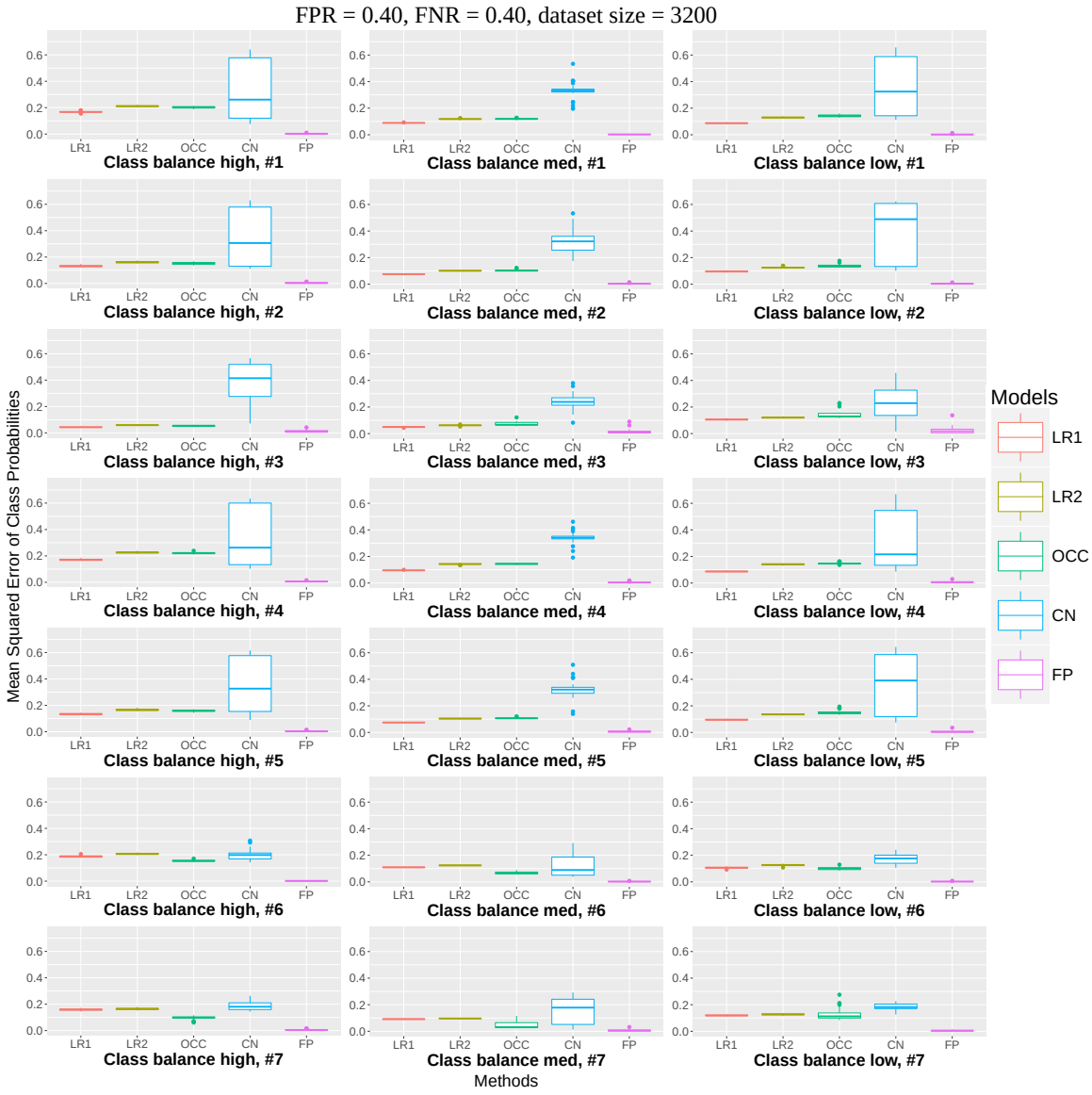

Figure A.9: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.1). Noise rate combination is FPR = 0.40 and FNR = 0.40. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

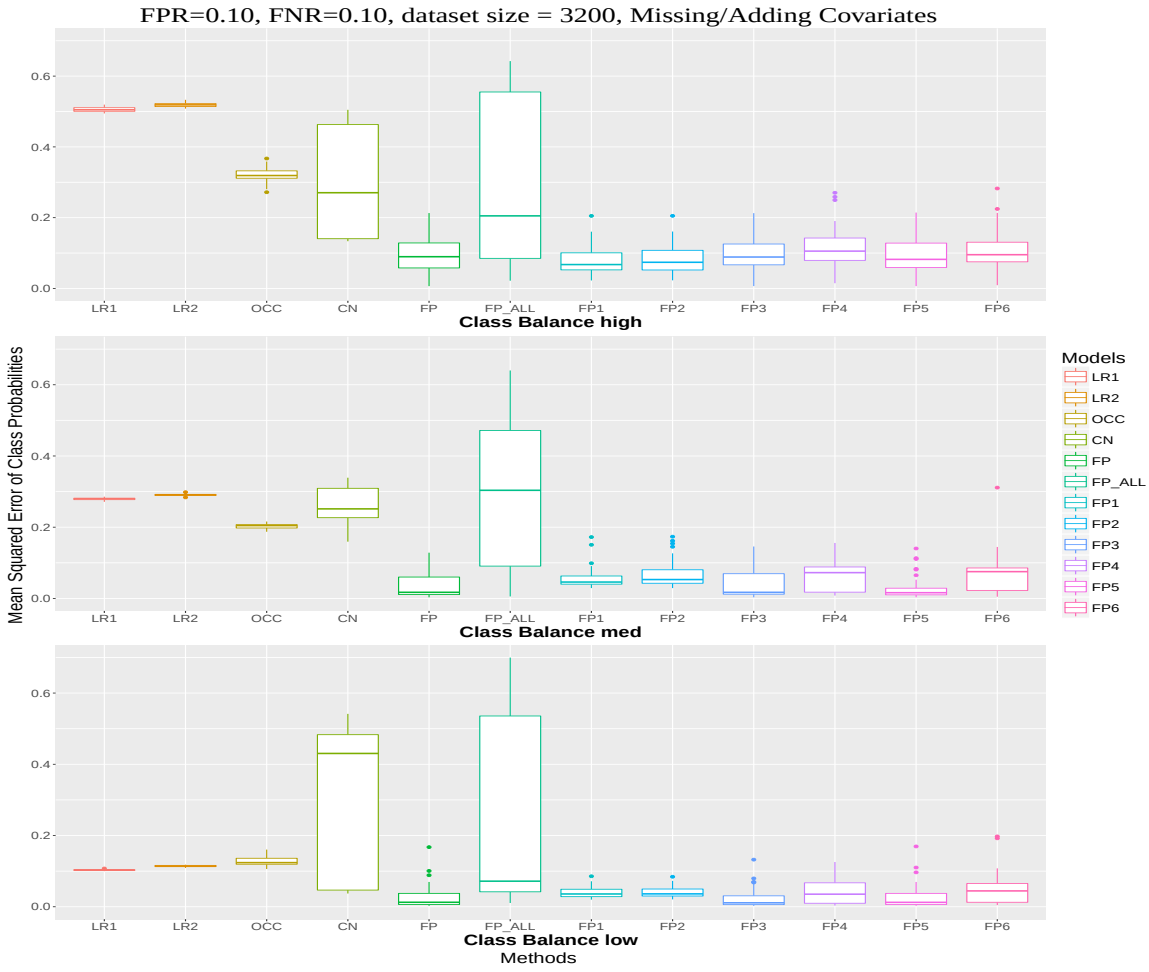

Figure A.10: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.1$ and $FNR = 0.1$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

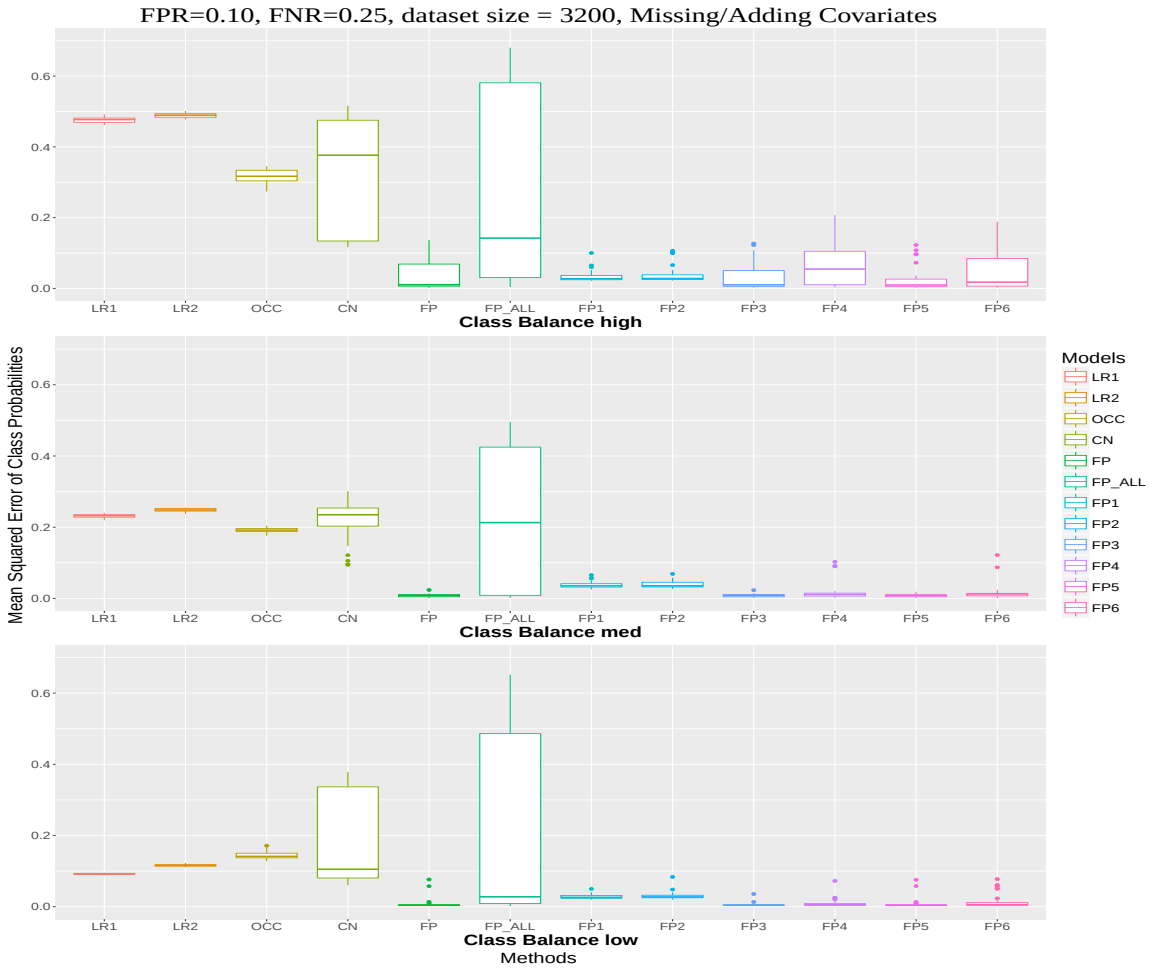

Figure A.11: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.1$ and $FNR = 0.25$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

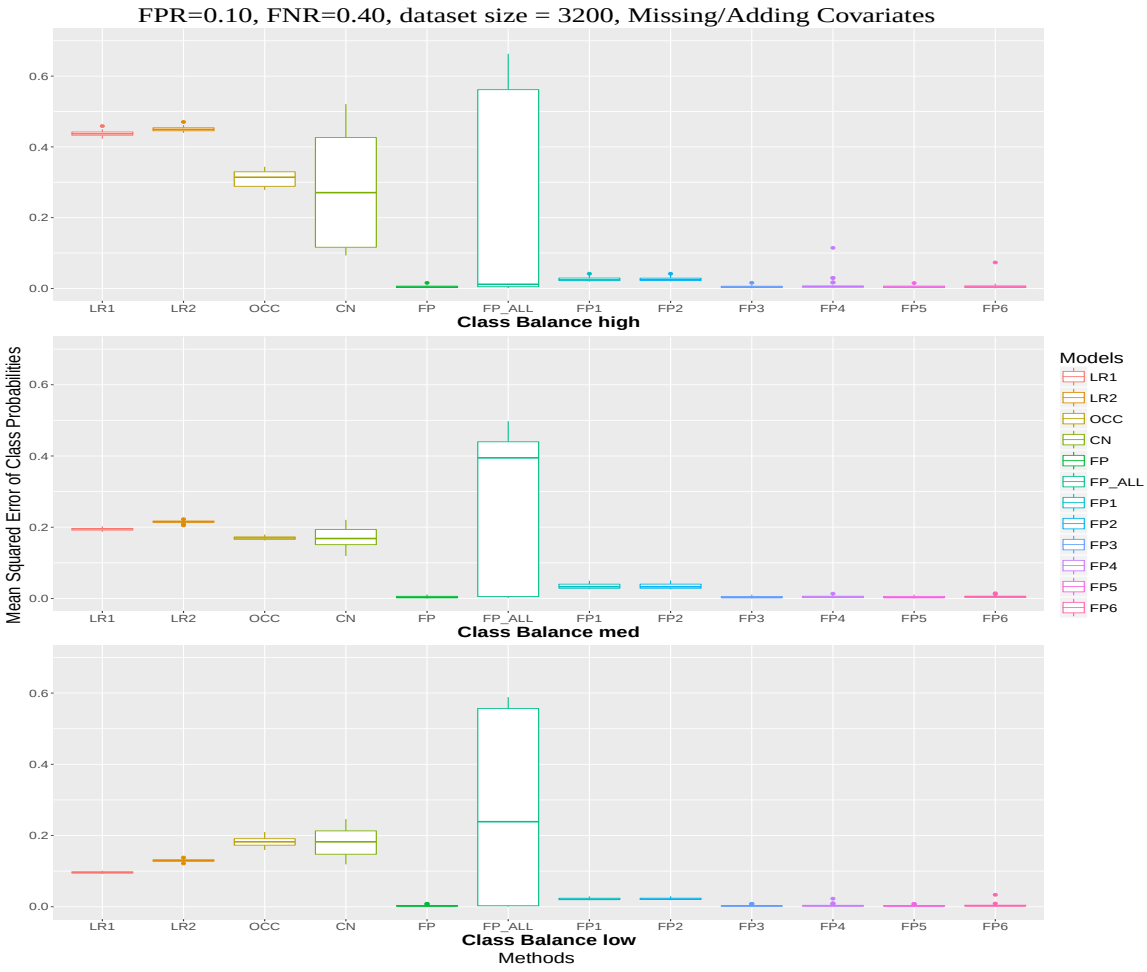

Figure A.12: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.1$ and $FNR = 0.40$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

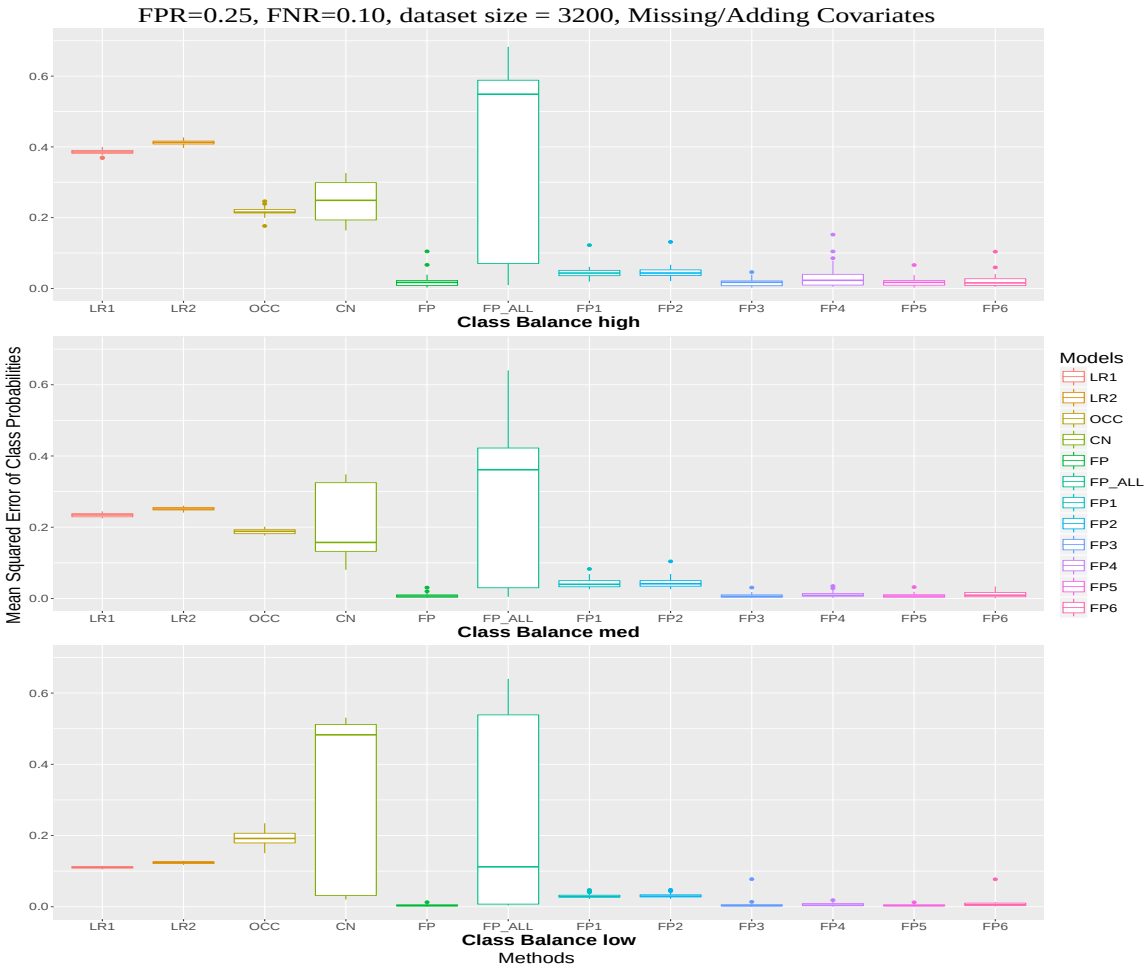

Figure A.13: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.25$ and $FNR = 0.10$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

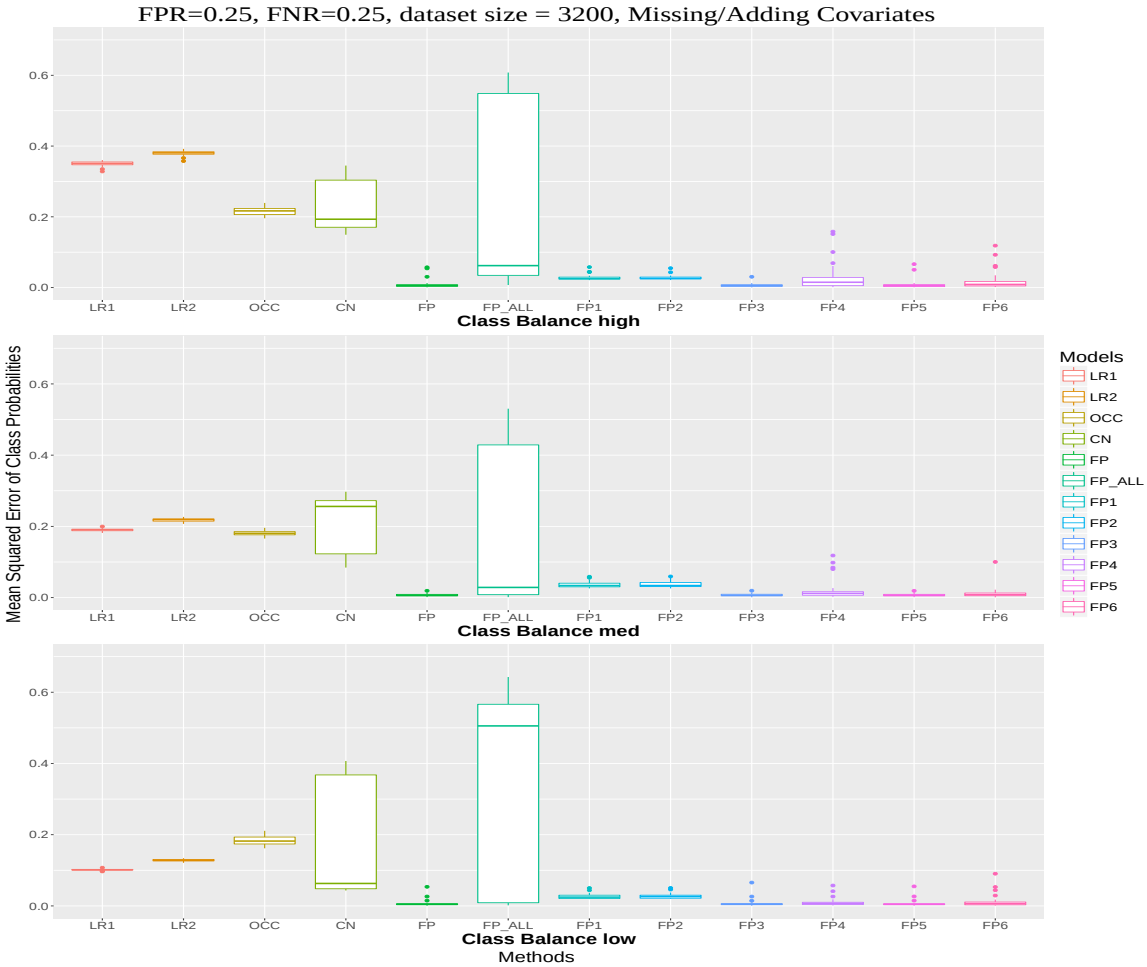

Figure A.14: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.25$ and $FNR = 0.25$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets. This figure is identical to Figure 4.4.

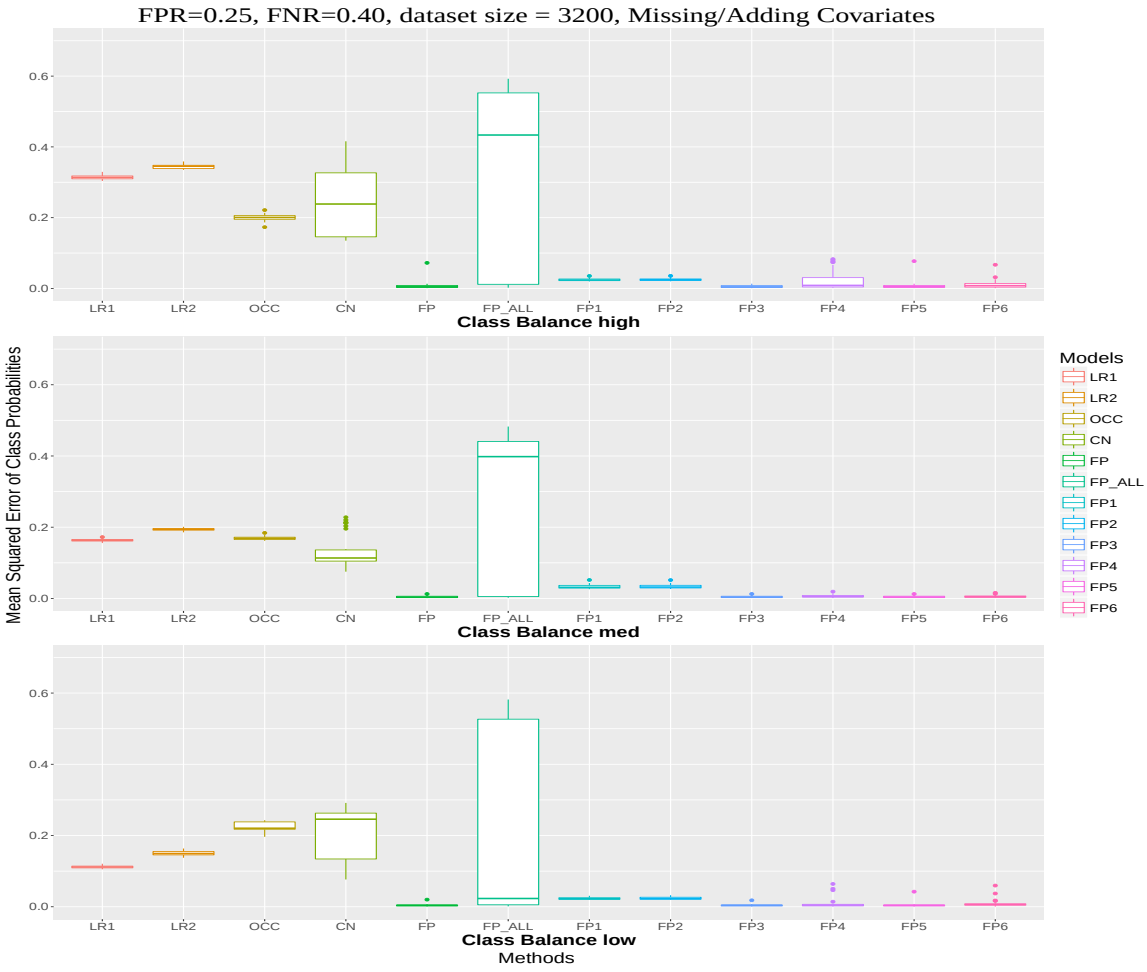

Figure A.15: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.25$ and $FNR = 0.40$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

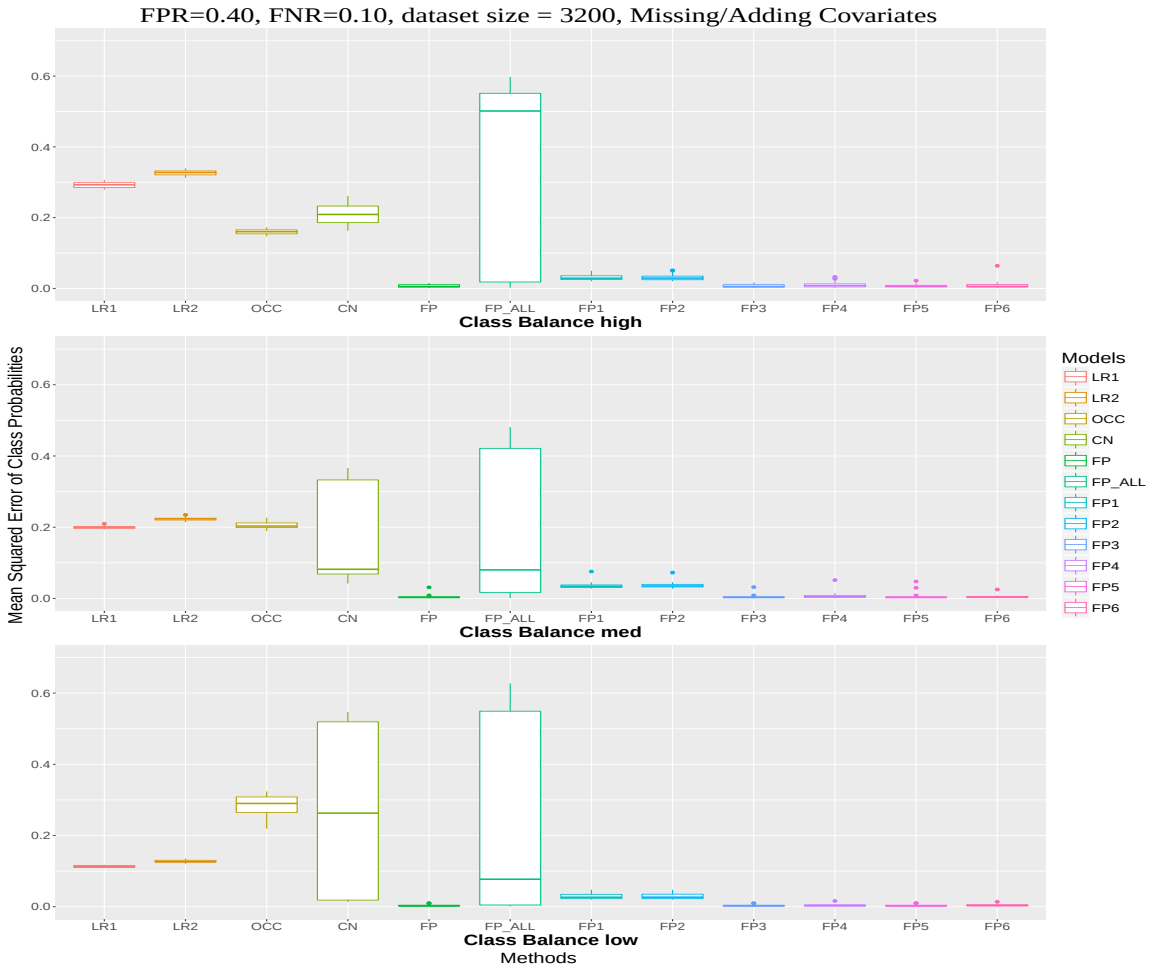

Figure A.16: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.40$ and $FNR = 0.10$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

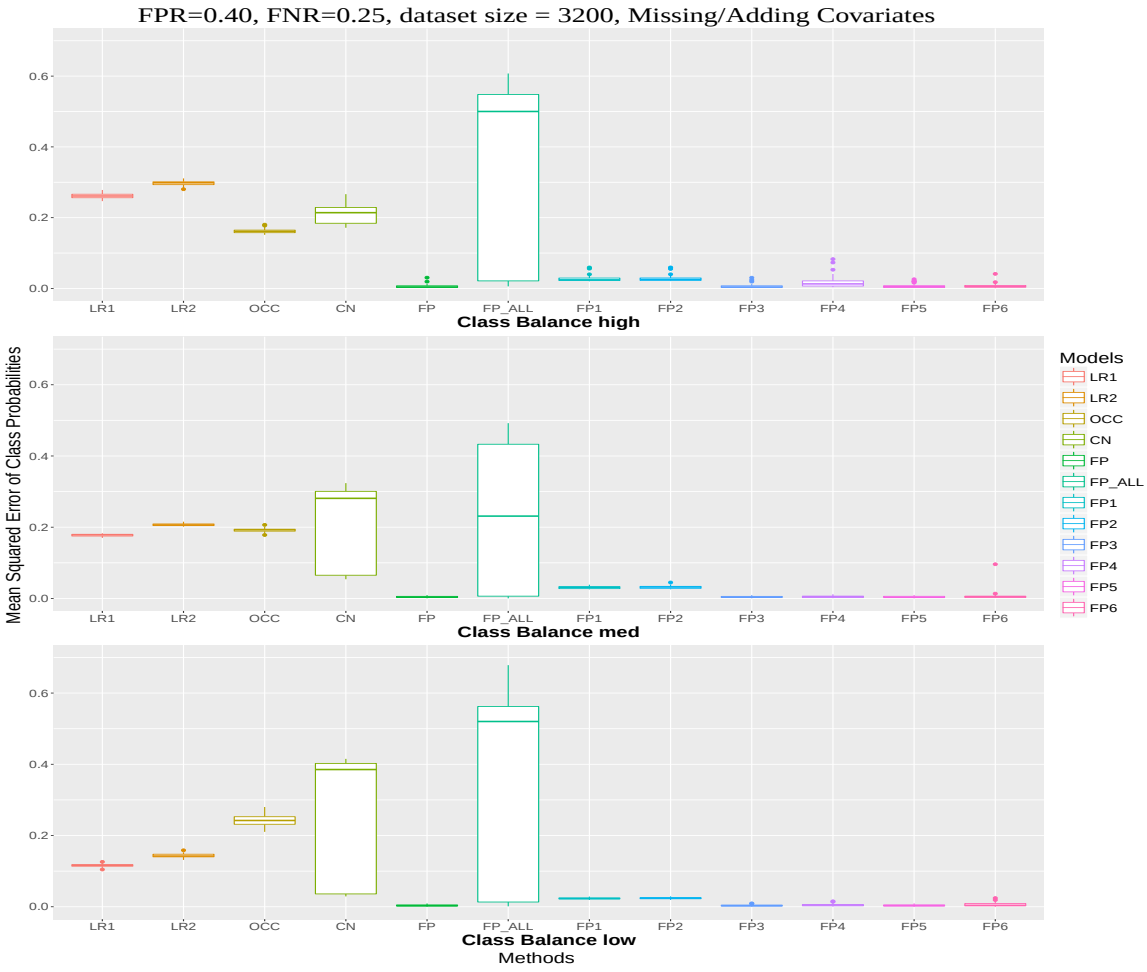

Figure A.17: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.40$ and $FNR = 0.25$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.

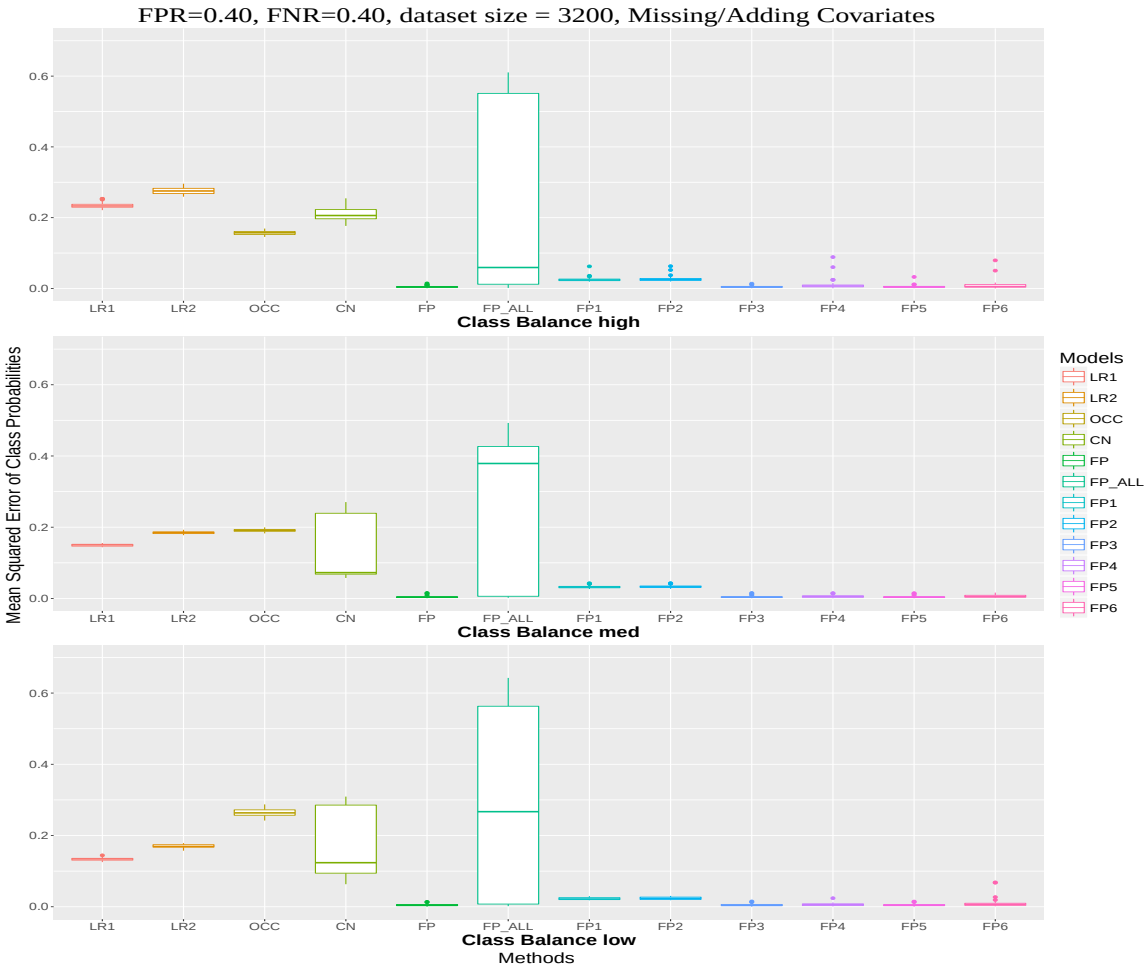

Figure A.18: Mean squared error in the class probabilities (ψ) for each method on each of the data-generating models (Table 4.3). Noise rate combination is $FPR = 0.40$ and $FNR = 0.40$. All datasets had 3200 training instances, and each boxplot represents 40 simulated datasets.