Estimating the Effect of a Mis-measured, Binary, Endogenous Regressor*

Francis J. DiTraglia[1] and Camilo García-Jimeno[1,2]

[1]Department of Economics, University of Pennsylvania
[2]NBER

This Version: August 15, 2017, First Version: October 31, 2015

Abstract

This paper studies identification and estimation of the effect of a binary, endogenous regressor subject to non-differential measurement error when a discrete-valued instrumental variable is available. We begin by showing that the only existing identification result for this case is incorrect. As such, identification in this model remains an open question. We then show that the usual mean independence assumption for the instrument and measurement error fail to identify the effect of interest. This motivates us to consider alternative and slightly stronger assumptions. We show that adding second and third moment independence assumptions suffices to identify the model and leads to an explicit method of moments estimator. Because our setting has features of a mixture model, however, this estimator suffers from a weak identification problem when the binary regressor has only a small effect on the outcome of interest. To address this difficulty, we derive a number of auxiliary moment inequalities that remain informative regardless of the magnitude of the treatment effect. Combining these with the moment equalities that emerge from our identification result, we propose a robust inference procedure based on generalized moment selection.

Keywords: Instrumental variables, Measurement error, Endogeneity, Binary regressor, Partial identification, Weak identification, Non-standard Inference

JEL Codes: C10, C18, C25, C26

*We thank Daron Acemoglu, Manuel Arellano, Kristy Buzard, Xu Cheng, Bernardo da Silveira, Bo Honoré, Sophocles Mavroeidis, Yuya Takahashi, and seminar participants at Chicago Booth, Princeton, Penn State, CEMFI, Manchester, Cambridge, UCL, and Oxford for valuable comments and suggestions.

1 Introduction

Many treatments of interest in applied work are binary. To take a particularly prominent example, consider treatment status in a randomized controlled trial. Even if the randomization is pristine, which yields a valid binary instrument (the offer of treatment), subjects may select into treatment based on unobservables, and given the many real-world complications that arise in the field, measurement error may be an important concern. This paper studies the use of a discrete instrumental variable to identify the causal effect of an endogenous, mis-measured, binary treatment in a model with additively separable errors. Specifically, we consider the following model

$$y = h(T^*, \mathbf{x}) + \varepsilon \tag{1}$$

where $T^* \in \{0, 1\}$ is a mis-measured, endogenous treatment, $\mathbf{x}$ is a vector of exogenous controls, and ε is a mean-zero error. Since T^* is potentially endogenous, $\mathbb{E}[\varepsilon|T^*, \mathbf{x}]$ may not be zero. Our goal is to non-parametrically estimate the average treatment effect (ATE) function

$$\tau(\mathbf{x}) = h(1, \mathbf{x}) - h(0, \mathbf{x}). \tag{2}$$

using a single discrete instrumental variable $z \in \{z_k\}_{k=1}^K$. We assume throughout that z is a relevant instrument for T^*, in other words

$$\mathbb{P}(T^* = 1|z_j, \mathbf{x}) \neq \mathbb{P}(T^* = 1|z_k, \mathbf{x}), \quad \forall k \neq j. \tag{3}$$

While the structural relationship involves T^*, we observe only a noisy measure T, polluted by non-differential measurement error. In particular, we assume that

$$\mathbb{P}(T = 1|T^* = 0, z, \mathbf{x}) = \alpha_0(\mathbf{x}) \tag{4}$$

$$\mathbb{P}(T = 0|T^* = 1, z, \mathbf{x}) = \alpha_1(\mathbf{x}) \tag{5}$$

where the mis-classification error rates $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ can depend on $\mathbf{x}$ but not z, and additionally that, conditional on true treatment status, observed treatment status provides no additional information about the error term. In other words, we assume that

$$\mathbb{E}[\varepsilon|T^*, T, z, \mathbf{x}] = \mathbb{E}[\varepsilon|T^*, z, \mathbf{x}]. \tag{6}$$

Although a relevant case for applied work, the setting we consider here has received little attention in the literature. The only existing result for the case of an endogenous treatment appears in an important paper by Mahajan (2006), who is primarily concerned with the

case of an exogenous treatment. As we show below, Mahajan's identification result for the endogenous treatment case is incorrect. As far as we are aware, this leaves the problem considered in this paper completely unsolved.

We begin by showing that the proof in Appendix A.2 of Mahajan (2006) leads to a contradiction. Throughout his paper, Mahajan (2006) maintains an assumption (Assumption 4) which he calls the "Dependency Condition." This assumption requires that the instrumental variable be relevant, namely that it generates variation in true treatment status. When extending his result for an exogenous treatment to the more general case of an endogenous one, however, he must impose an additional condition on the model (Equation 11), which turns out to violate the Dependency Condition. Since one cannot impose the condition in Equation 11 of Mahajan (2006), we go on to study the prospects for identification in this model more broadly. We consider two possibilities. First, since Mahajan's identification results require only a binary instrument, we borrow an idea from Lewbel (2007) and explore whether expanding the support of the instrument yields identification based on moment equations similar to those used by Mahajan (2006). While allowing the instrument to take on additional values does increase the number of available moment conditions, we show that these moments cannot point identify the treatment effect, regardless of how many (finite) values the instrument takes on.

We then consider a new source of identifying information that arises from imposing stronger assumptions on the instrumental variable. If the instrument is not merely mean independent but in fact *statistically independent* of the regression error term, as in a randomized controlled trial or a true natural experiment, additional moment conditions become available. We show that adding a conditional second moment independence assumption on the instrument identifies the *difference* of mis-classification rates $\alpha_1(\mathbf{x}) - \alpha_0(\mathbf{x})$. Because these rates must equal each other when there is no mis-classification error, our result can be used to test a necessary condition for the absence of measurement error. It can also be used to construct simple and informative partial identification bounds for the treatment effect. When one of the mis-classification rates is known, this identifies the treatment effect. More generally, however, this is not the case. We go on to show that a conditional third moment independence assumption on the instrument point identifies both $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ and hence the ATE function $\tau(\mathbf{x})$. Both our point identification and partial identification results require only a binary instrument, and lead to simple, closed-form method of moments estimators.

This project is still in progress. The present draft focuses on establishing identification in the simplest possible way: by directly solving a set of equations implied by conditional moment restrictions on ε. Additional results regarding efficient estimation and sharp bounds

for α_0, α_1 under weaker conditions on the instrumental variable are currently in progress. For some additional discussion of these results, see our conclusion below in Section 4.

The remainder of this paper is organized as follows. In section 2 we discuss the literature in relation to the problem considered here. Section 3 introduces notation and assumptions, and presents our main results. Section 4 concludes. All proofs appear in the Appendix.

2 Related Literature

Measurement error is a pervasive feature of economic data, motivating a long tradition of measurement error modelling in econometrics. The textbook case considers a continuous regressor (treatment) subject to classical measurement error in a linear model. In this setting, the measurement error is assumed to be unrelated to the true, unobserved, value of the treatment of interest. Regardless of whether this unobserved treatment is exogenous or endogenous, a single valid instrument suffices to identify its effect. When an instrument is unavailable, Lewbel (1997) shows that higher moment assumptions can be used to construct one, provided that the mis-measured treatment is exogenous. When it is endogenous, Lewbel (2012) uses a heteroskedasticity assumption to obtain identification.

Departures from the linear, classical measurement error setting pose serious identification challenges. One strand of the literature considers relaxing the assumption of linearity while maintaining that of classical measurement error. Schennach (2004), for example, uses repeated measures of each mis-measured treatment to obtain identification, while Schennach (2007) uses an instrumental variable. Both papers consider the case of exogenous treatments.[1] More recently, Song et al. (2015) rely on a repeated measure of the mis-measured treatment and the existence of a set of additional regressors, conditional upon which the treatment of interest is unrelated to the unobservables, to obtain identification. Another strand of the literature considers relaxing the assumption of classical measurement error, by allowing the measurement error to be related to the true value of the unobserved treatment. Chen et al. (2005) obtain identification in a general class of moment condition models with mis-measured data by relying on the existence of an auxiliary dataset from which they can estimate the measurement error process. In contrast, Hu and Shennach (2008) and Song (2015) rely on an instrumental variable and an additional conditional location assumption on the measurement error distribution. More recently, Hu et al. (2015) use a continuous instrument to identify the ratio of partial effects of two continuous regressors, one measured with error, in a linear single index model.

[1] For comprehensive reviews of the challenges of addressing measurement error in non-linear models, see Chen et al. (2011) and Schennach (2013).

Many treatments of interest in economics, however, are binary, and in this case classical measurement error is impossible. Because a true 1 can only be mis-measured as a 0 and a true 0 can only be mis-measured as a 1, the measurement error must be *negatively* correlated with the true treatment status (Aigner, 1973; Bollinger, 1996). For this reason, even in a textbook linear model, the instrumental variables estimator can only remove the effect of endogeneity, not that of measurement error (Frazis and Loewenstein, 2003). Measurement error in a discrete variable is usually called mis-classification.[2] The simplest form of mis-classification is so-called *non-differential* measurement error. In this case, conditional on true treatment status, and possibly a set of exogenous covariates, the measurement error is assumed to be unrelated to all other variables in the system.

A number of papers have studied this problem without the use of instrumental variables under the assumption that the mis-measured binary treatment is exogenous. The first to address this problem was Aigner (1973), who characterized the asymptotic bias of the OLS estimator in this setting, and proposed a technique for correcting it using outside information on the mis-classification process. Another early contribution by Bollinger (1996) provides partial identification bounds. More recently, Chen et al. (2008a) use higher moment assumptions to obtain identification in a linear regression model, and Chen et al. (2008b) extend these results to the non-parametric setting. van Hasselt and Bollinger (2012) and Bollinger and van Hasselt (2015) provide additional partial identification results.

Continuing under the assumption of an exogenous treatment, a number of other papers in the literature have considered the identifying power of an instrumental variable, or something like one. Black et al. (2000) and Kane et al. (1999) more-or-less simultaneously pointed out that when *two* alternative measures of treatment are available, both subject to non-differential measurement error, a non-linear GMM estimator can be used to recover the treatment effect. In essence, one measure serves as an instrument for the other although the estimator is quite different from IV.[3] Subsequently, Frazis and Loewenstein (2003) correctly note that an instrumental variable can take the place of one of the measures of treatment in a linear model with an exogenous treatment, allowing one to implement a variant of the GMM estimator proposed by Black et al. (2000) and Kane et al. (1999). However, as we will show below, the assumptions required to obtain this result are stronger than Frazis and Loewenstein (2003) appear to realize: the usual IV assumption that the instrument is mean

[2]For general results on the partial identification of discrete probability distributions using mis-classified observations, see Molinari (2008).

[3]Ignoring covariates, the observable moments in this case are the joint probability distribution of the two binary treatment measures and the conditional means of the outcome variable given the two measures. Although the system is highly non-linear, it can be manipulated to yield an explicit solution for the treatment effect provided that the true treatment is exogenous.

independent of the regression error is insufficient for identification.

Mahajan (2006) extends the results of Black et al. (2000) and Kane et al. (1999) to a more general nonparametric regression setting using a binary instrument in place of one of the treatment measures. Although unaware of Frazis and Loewenstein (2003), Mahajan (2006) makes the correct assumption over the instrument and treatment to guarantee identification of the conditional mean function. When the treatment is in fact exogenous, this coincides with the treatment effect. Hu (2008) derives related results when the mis-classified discrete regressor may take on more than two values. Lewbel (2007) provides an identification result for the same model as Mahajan (2006) under different assumptions. In particular, the variable that plays the role of the "instrument" need not satisfy the exclusion restriction provided that it does not interact with the treatment and takes on at least three distinct values.

Much less is known about the case in which a binary, or discrete, treatment is not only mis-measured but endogenous. Frazis and Loewenstein (2003) briefly discuss the prospects for identification in this setting. Although they do not provide a formal proof they argue, in the context of their parametric linear model, that the treatment effect is unlikely to be identified unless one is willing to impose strong and somewhat unnatural conditions.[4] The first paper to provide a formal result for this case is Mahajan (2006). He extends his main result to the case of an endogenous treatment, providing an explicit proof of identification under the usual IV assumption in a model with additively separable errors. As we show below, however, Mahajan's proof is incorrect.

The results we derive here most closely relate to the setting considered in Mahajan (2006) in that we study non-parametric identification of the effect of a binary, endogenous treatment, using a discrete instrument. Unlike Mahajan (2006) we consider and indeed show the necessity of using higher-moment information to identify the causal effect of interest. Unlike Kreider et al. (2012), who partially identify the effects of food stamps on health outcomes of children under weak measurement error assumptions, we do not rely on auxiliary data. Unlike Shiu (2015), who considers a sample selection model with a discrete, mis-measured, endogenous regressor, we do not rely on a parametric assumption about the form of the first-stage. Finally unlike Ura (2015), who studies local average treatment effects under very general forms of mis-classification but presents only partial identification results, we point identify an average treatment effect under non-differential measurement error. Moreover, unlike the identification strategies from the existing literature described

[4]For example, one could consider using the results of Hausman et al. (1998), who study regressions with a mis-classified, discrete *outcome* variable, as a first-stage in an IV setting. In principle, this approach would fully identify the mis-classification error process. Using these results, however, requires either an explicit, nonlinear, parametric model for the first stage, or an identification at infinity argument.

above, we do not rely upon continuity of the instrument, a large support condition, or restrictions on the relationship between the true, unmeasured treatment and its observed surrogate, subject to the condition that the measurement error process is non-differential.

3 Bonferroni Plots

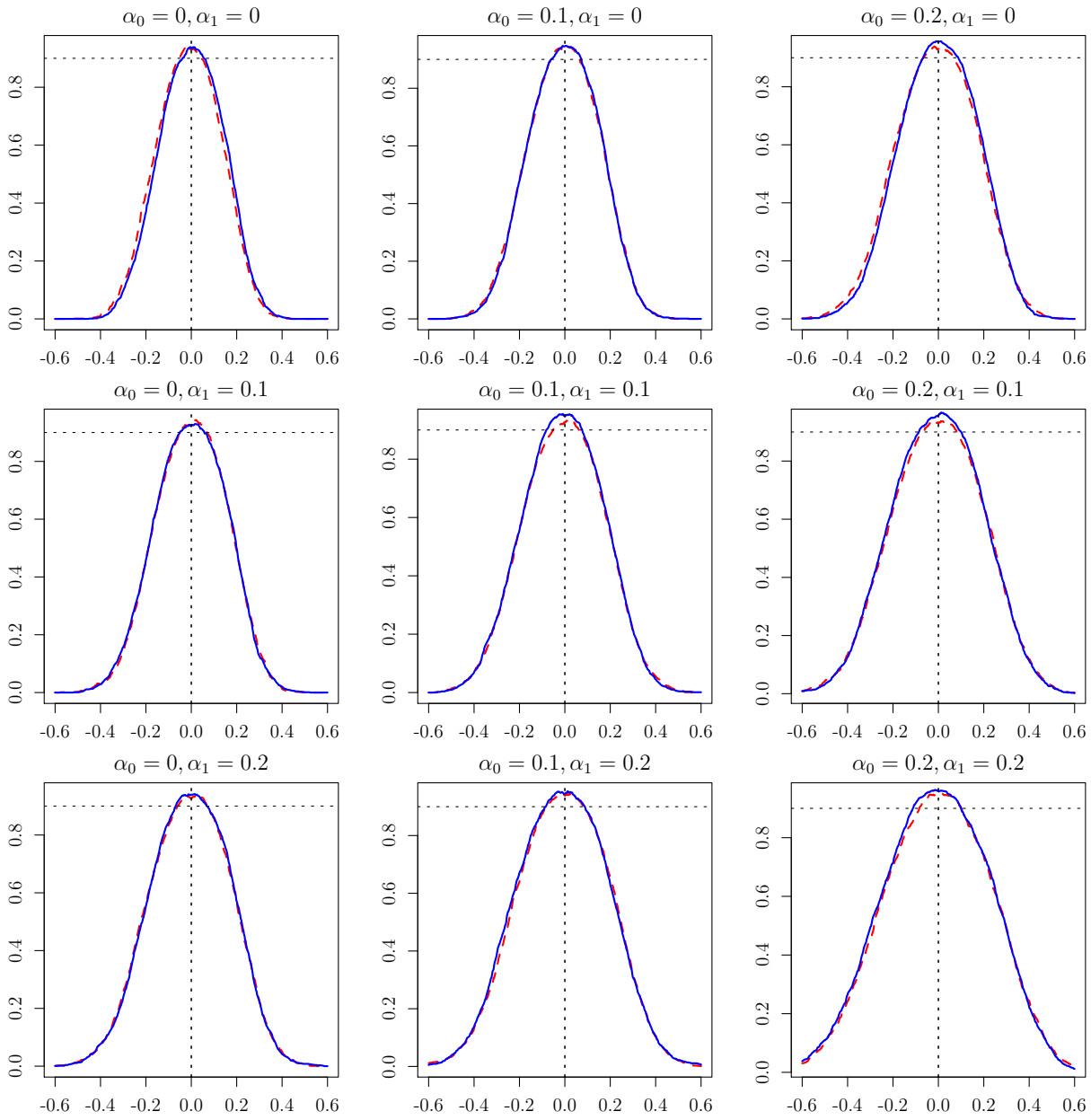

Figure 1: $\beta = 0, n = 1000$

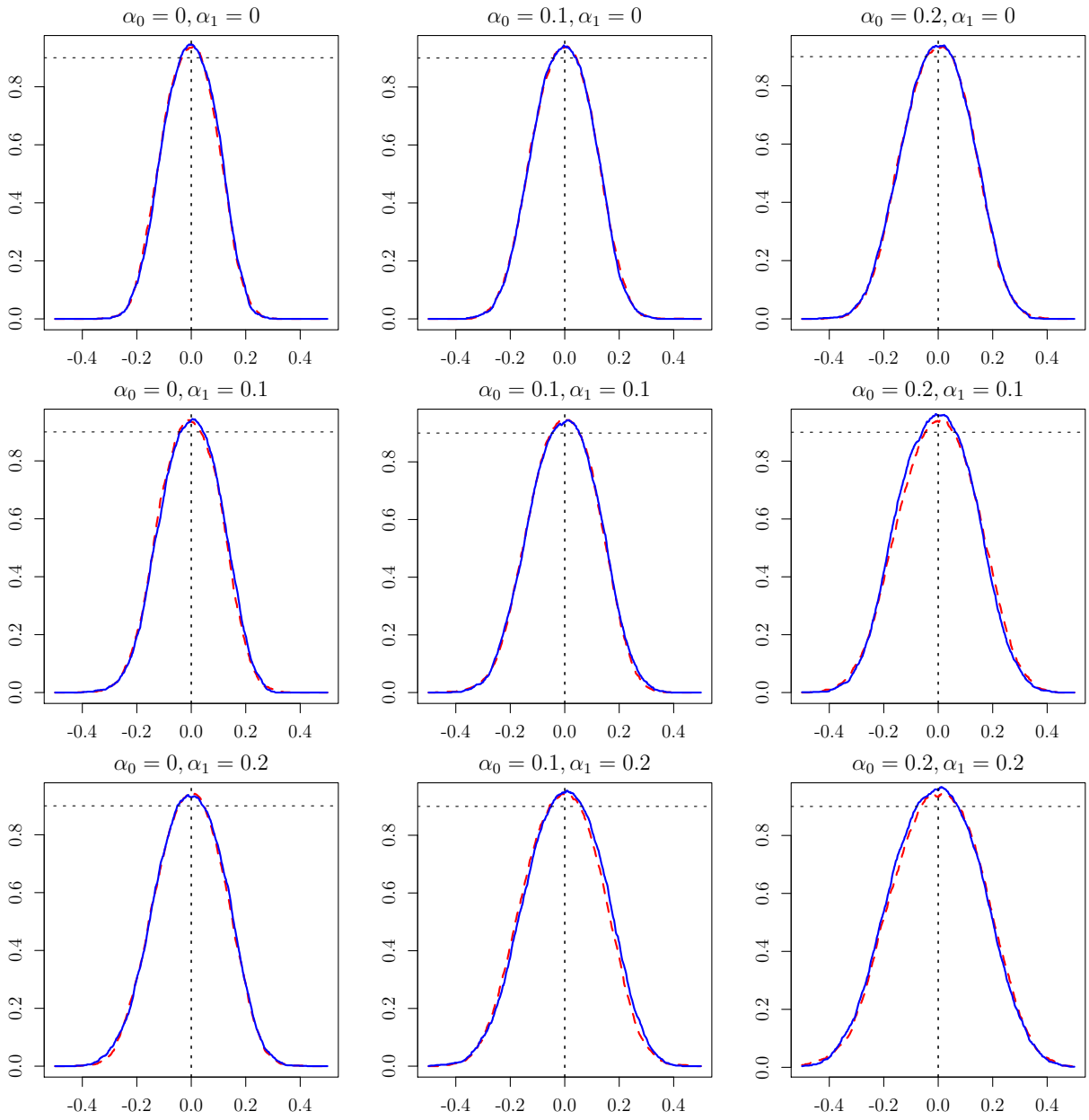

Figure 2: $\beta = 0, n = 2000$

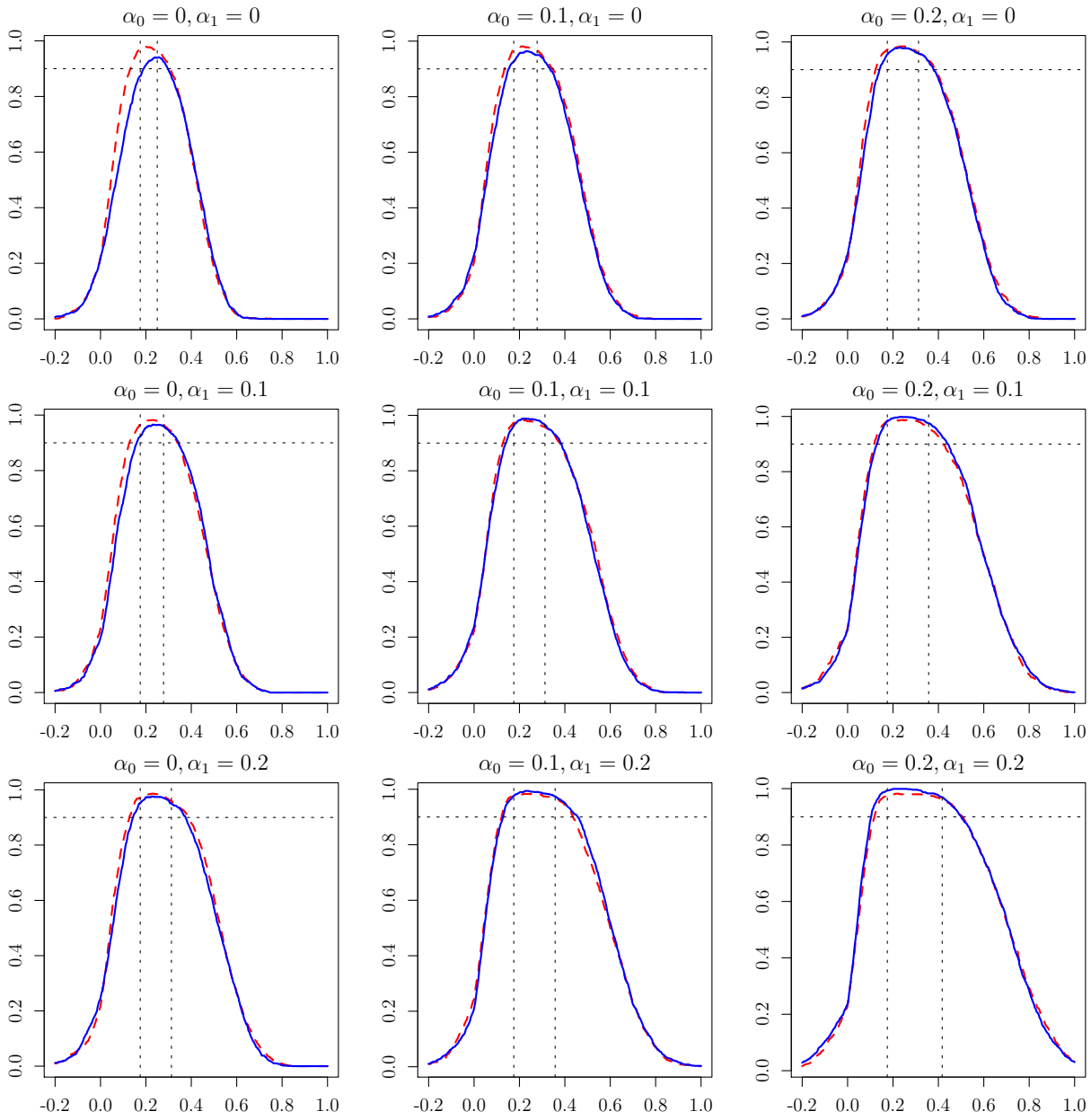

Figure 3: $\beta = 0.25, n = 1000$

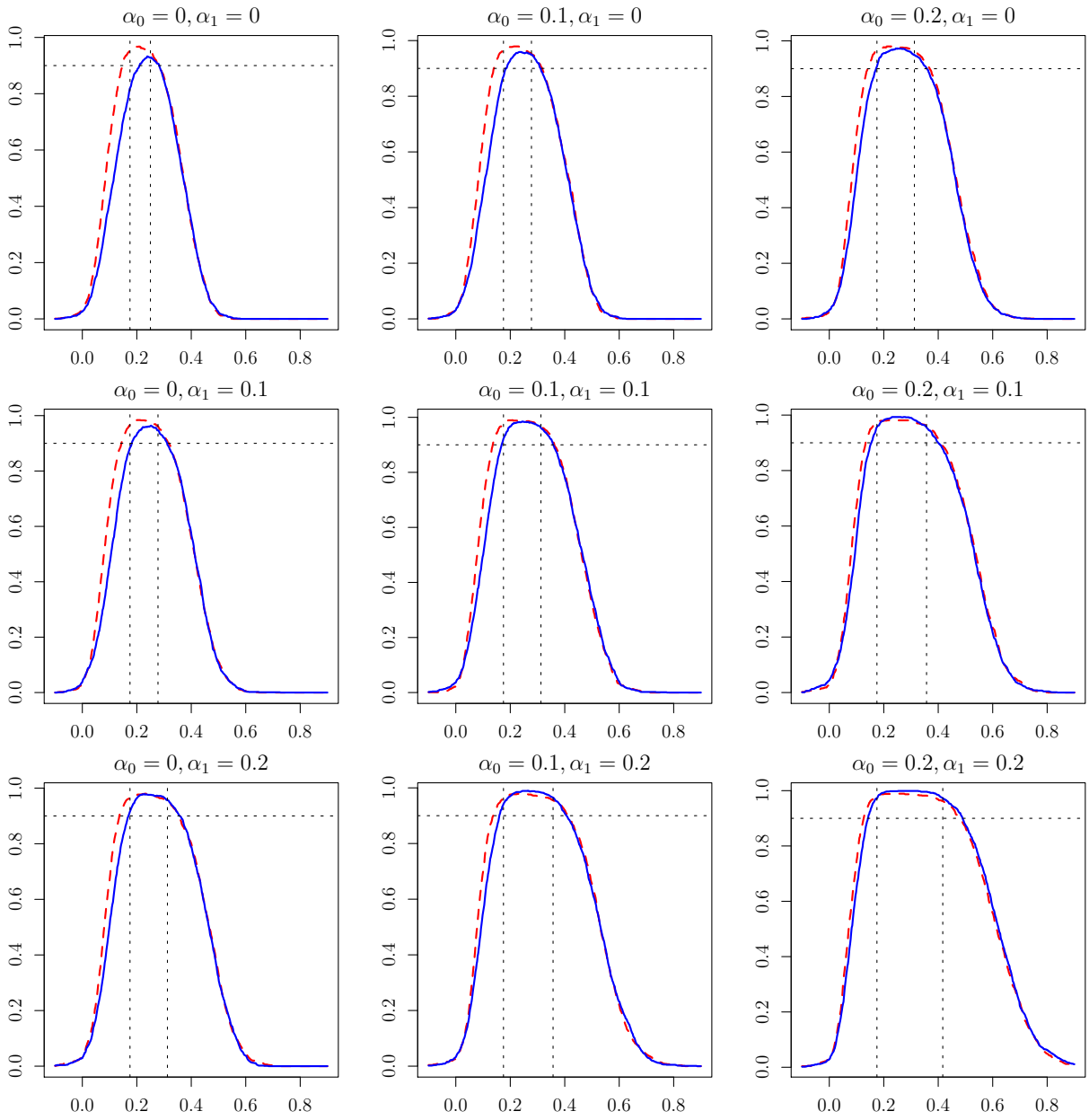

Figure 4: $\beta = 0.25, n = 2000$

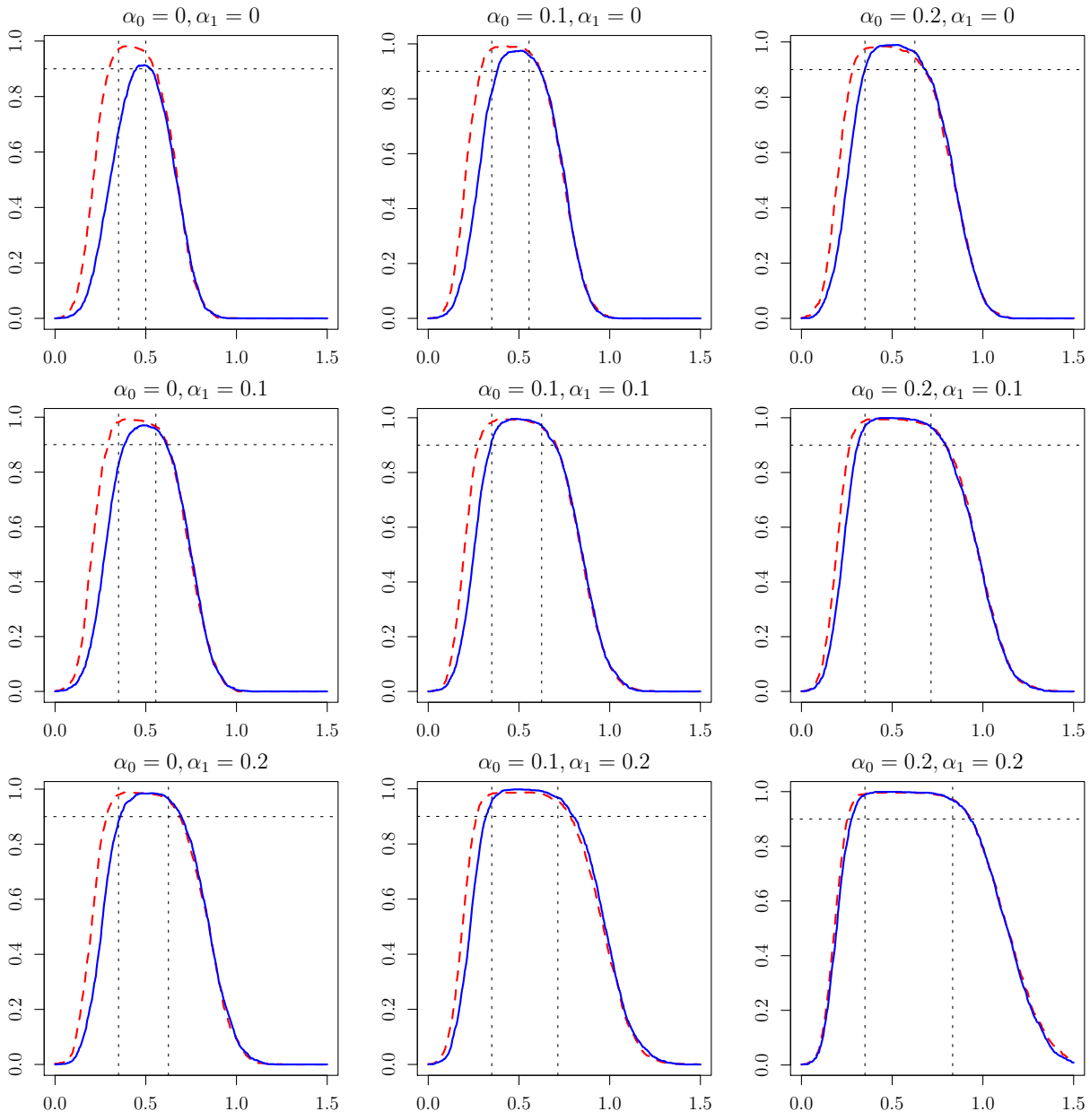

Figure 5: $\beta = 0.5, n = 1000$

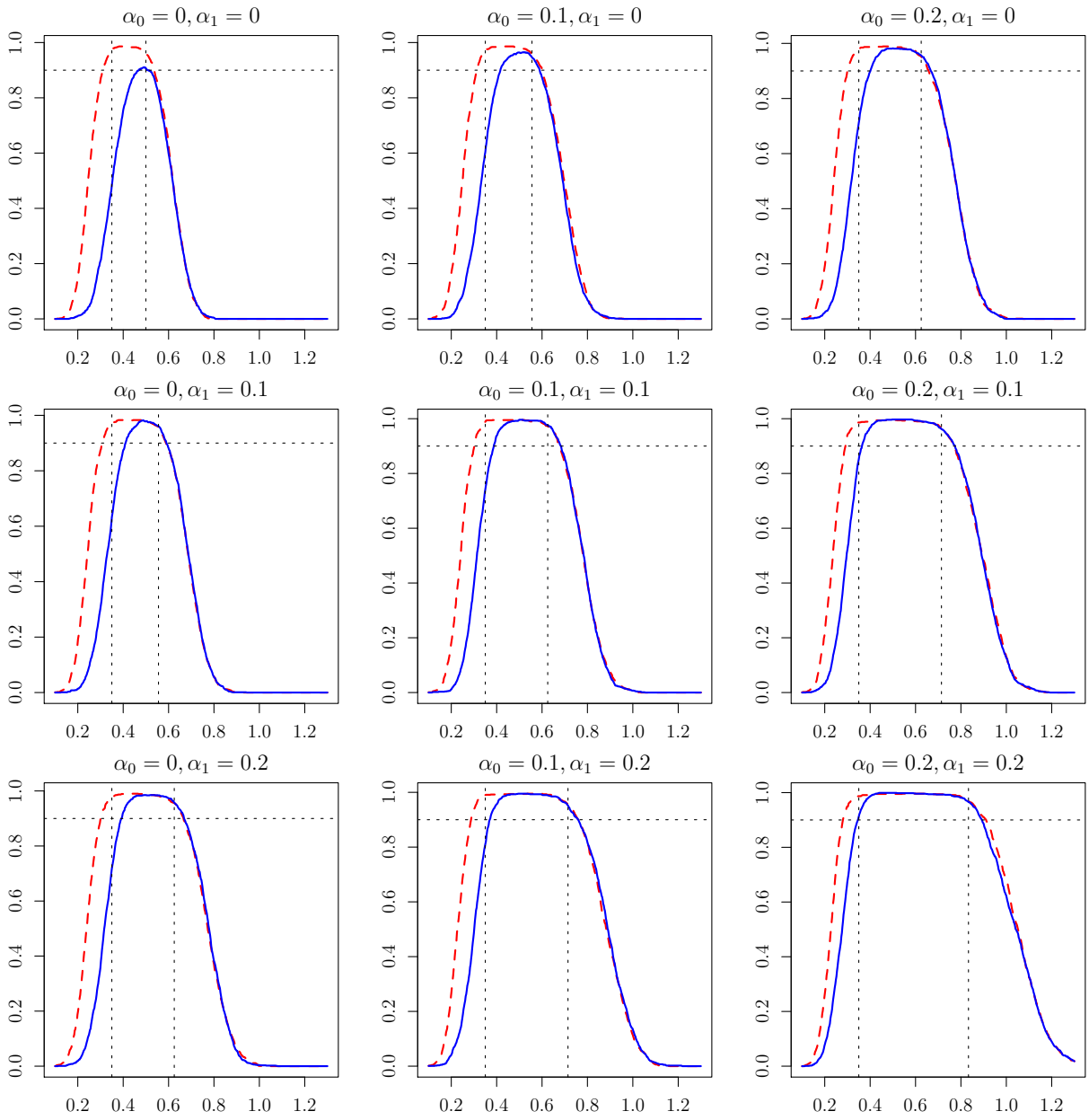

Figure 6: $\beta = 0.5, n = 2000$

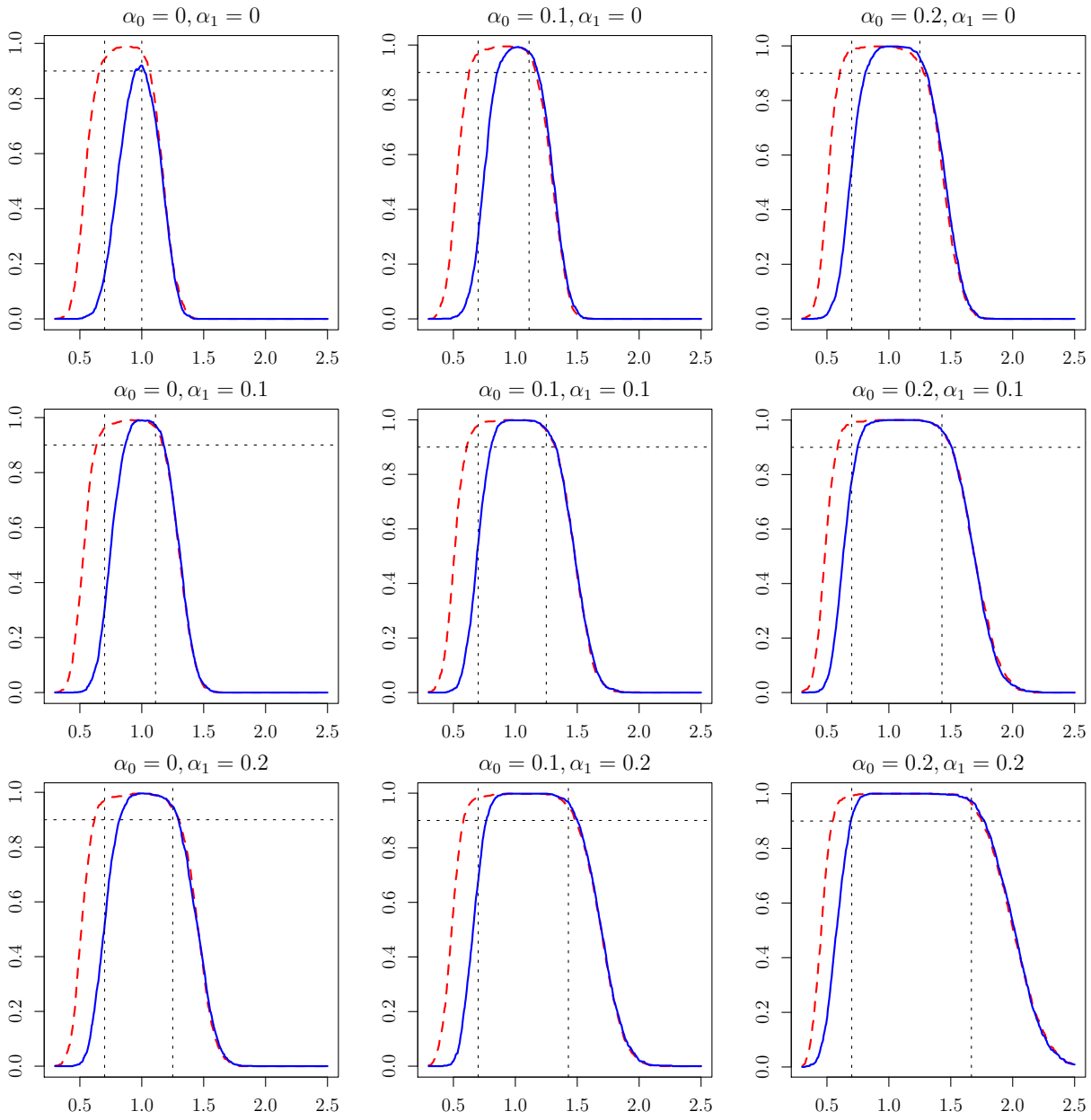

Figure 7: $\beta = 1, n = 1000$

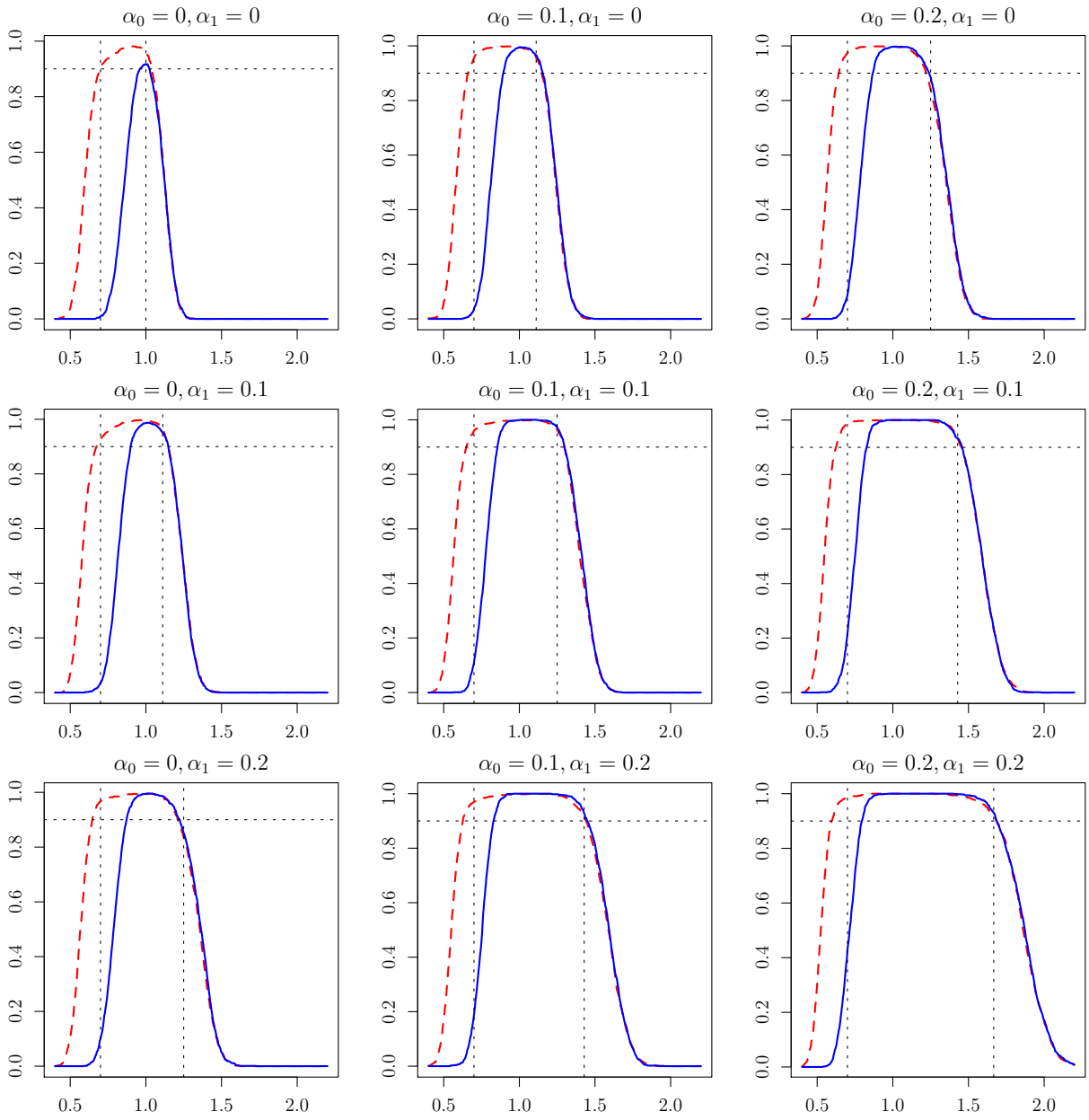

Figure 8: $\beta = 1, n = 2000$

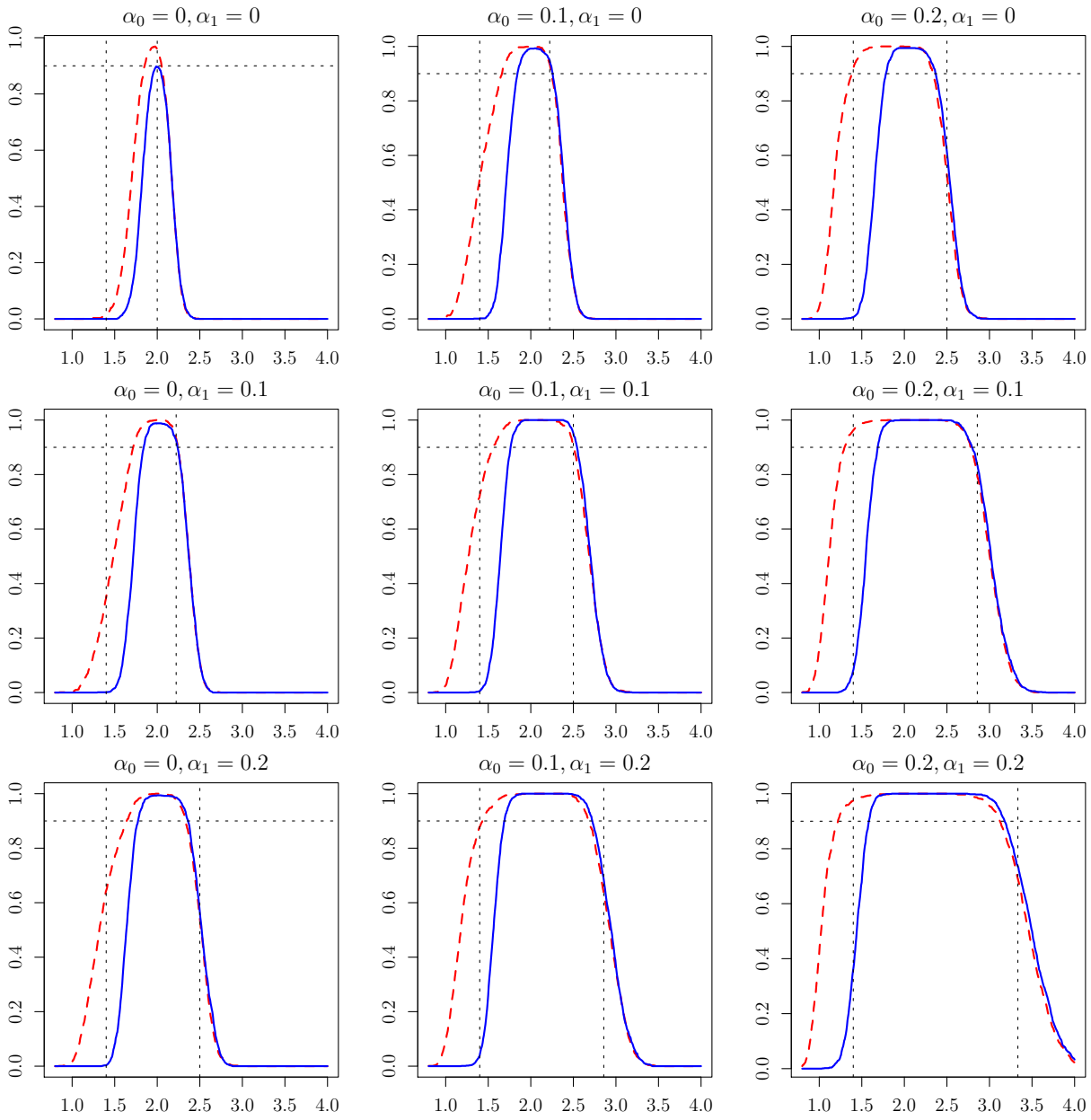

Figure 9: $\beta = 2, n = 1000$

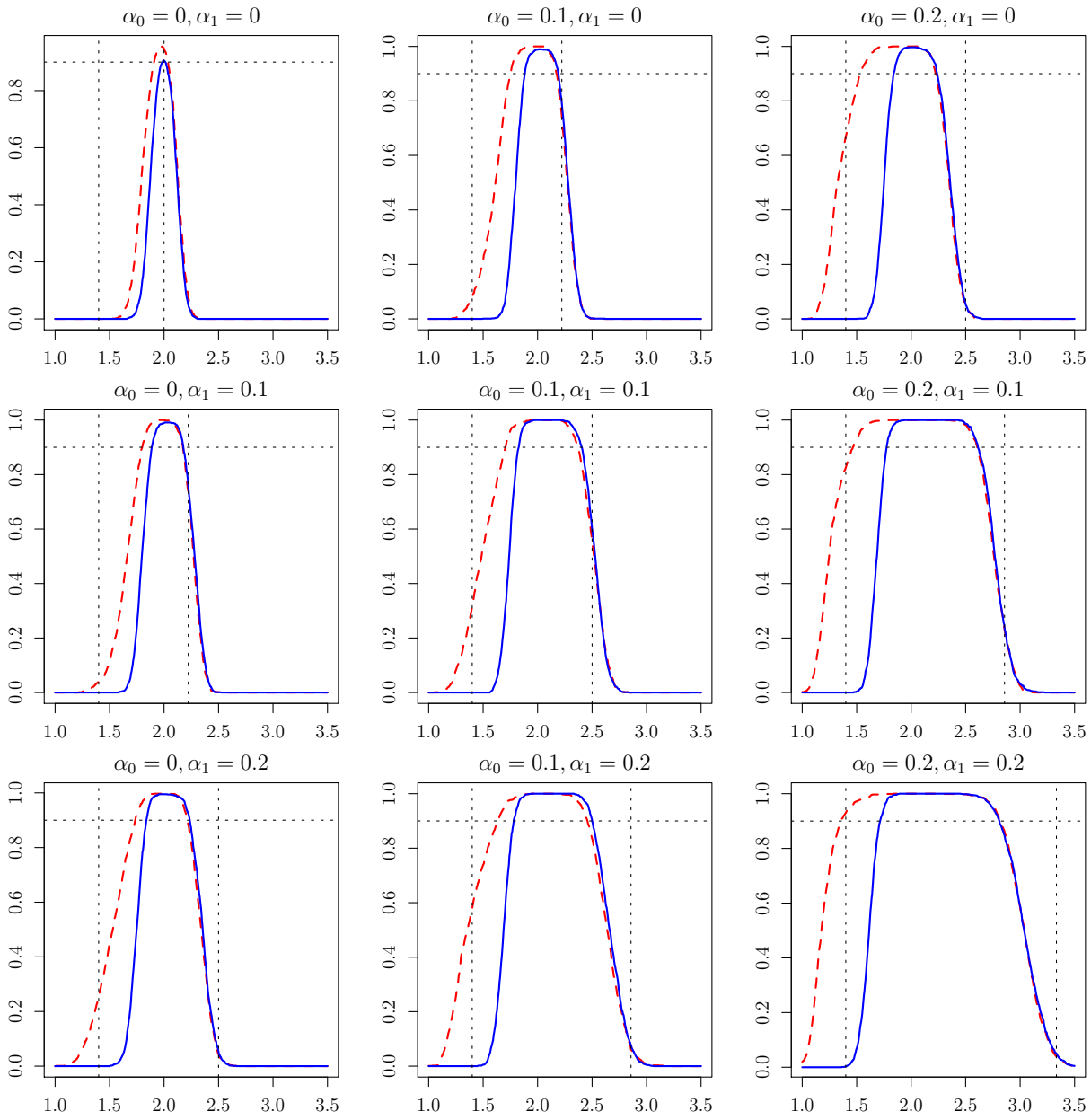

Figure 10: $\beta = 2, n = 2000$

References

Aigner, D. J., 1973. Regression with a binary independent variable subject to errors of observation. Journal of Econometrics 1, 49–60.

Black, D. A., Berger, M. C., Scott, F. A., 2000. Bounding parameter estimates with nonclassical measurement error. Journal of the American Statistical Association 95 (451), 739–748.

Bollinger, C. R., 1996. Bounding mean regressions when a binary regressor is mismeasured. Journal of Econometrics 73, 387–399.

Bollinger, C. R., van Hasselt, M., 2015. Bayesian moment-based inference in a regression models with misclassification error, working Paper.

Chen, X., Hong, H., Nekipelov, D., 2011. Nonlinear models of measurement errors. Journal of Economic Literature 49 (4), 901–937.

Chen, X., Hong, H., Tamer, E., 2005. Measurement error models with auxiliary data. The Review of Economic Studies 72 (2), 343–366.

Chen, X., Hu, Y., Lewbel, A., 2008a. Nonparametric identification of regression models containing a misclassified dichotomous regressor with instruments. Economics Letters 100, 381–384.

Chen, X., Hu, Y., Lewbel, A., 2008b. A note on the closed-form identification of regression models with a mismeasured binary regressor. Statistics & Probability Letters 78 (12), 1473–1479.

Frazis, H., Loewenstein, M. A., 2003. Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables. Journal of Econometrics 117, 151–178.

Hausman, J., Abrevaya, J., Scott-Morton, F., 1998. Misclassification of the dependent variable in a discrete-response setting. Journal of Econometrics 87, 239–269.

Hu, Y., 2008. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. Journal of Econometrics 144 (1), 27–61.

Hu, Y., Shennach, S. M., January 2008. Instrumental variable treatment of nonclassical measurement error models. Econometrica 76 (1), 195–216.

Hu, Y., Shiu, J.-L., Woutersen, T., 2015. Identification and estimation of single index models with measurement error and endogeneity. The Econometrics Journal (Forthcoming).

Kane, T. J., Rouse, C. E., Staiger, D., July 1999. Estimating returns to schooling when schooling is misreported. Tech. rep., National Bureau of Economic Research, NBER Working Paper 7235.

Kreider, B., Pepper, J. V., Gundersen, C., Jolliffe, D., 2012. Identifying the effects of SNAP (food stamps) on child health outcomes when participation is endogenous and misreported. Journal of the American Statistical Association 107 (499), 958–975.

Lewbel, A., 1997. Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D. Econometrica, 1201–1213.

Lewbel, A., March 2007. Estimation of average treatment effects with misclassification. Econometrica 75 (2), 537–551.

Lewbel, A., 2012. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. Journal of Business & Economic Statistics 30 (1), 67–80.

Mahajan, A., 2006. Identification and estimation of regression models with misclassification. Econometrica 74 (3), 631–665.

Molinari, F., 2008. Partial identification of probability distributions with misclassified data. Journal of Econometrics 144 (1), 81–117.

Schennach, S. M., 2004. Estimation of nonlinear models with measurement error. Econometrica 72 (1), 33–75.

Schennach, S. M., 2007. Instrumental variable estimation of nonlinear errors-in-variables models. Econometrica 75 (1), 201–239.

Schennach, S. M., 2013. Measurement error in nonlinear models – a review. In: Acemoglu, D., Arellano, M., Dekel, E. (Eds.), Advances in Economics and Econometrics. Vol. 3. Cambridge University Press, pp. 296–337.

Shiu, J.-L., 2015. Identification and estimation of endogenous selection models in the presence of misclassification errors. Economic Modelling (Forthcoming).

Song, S., 2015. Semiparametric estimation of models with conditional moment restrictions in the presence of nonclassical measurement errors. Journal of Econometrics 185 (1), 95–109.

Song, S., Schennach, S. M., White, H., 2015. Semiparametric estimation of models with conditional moment restrictions in the presence of nonclassical measurement errors. Quantitative Economics (Forthcoming).

Ura, T., November 2015. Heterogeneous treatment effects with mismeasured endogenous treatment. Tech. rep., Duke University Department of Economics.

van Hasselt, M., Bollinger, C. R., 2012. Binary misclassification and identification in regression models. Economics Letters 115, 81–84.