cemmap
centre for microdata methods and practice

Semi-nonparametric models of multidimensional matching: an optimal transport approach

Dongwoo Kim
Young Jun Lee

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP12/24

UKRI Economic and Social Research Council

SEMI-NONPARAMETRIC MODELS OF MULTIDIMENSIONAL MATCHING: AN OPTIMAL TRANSPORT APPROACH

DONGWOO KIM
Department of Economics, Simon Fraser University

YOUNG JUN LEE
Korea Institute for International Economic Policy

This paper proposes empirically tractable multidimensional matching models, focusing on worker-job matching. We generalize the parametric model proposed by Lindenlaub (2017), which relies on the assumption of joint normality of observed characteristics of workers and jobs. In our paper, we allow unrestricted distributions of characteristics and show identification of the production technology, and equilibrium wage and matching functions using tools from optimal transport theory. Given identification, we propose efficient, consistent, asymptotically normal sieve estimators. We revisit Lindenlaub's empirical application and show that, between 1990 and 2010, the U.S. economy experienced much larger technological progress favoring cognitive abilities than the original findings suggest. Furthermore, our flexible model specifications provide a significantly better fit for patterns in the evolution of wage inequality.

Keywords: Multidimensional matching, transferable utility, optimal transport, sieve extremum estimation, technological progress, wage polarization.

Dongwoo Kim: dongwook@sfu.ca

Young Jun Lee: y.lee@kiep.go.kr

This paper is based on the third chapter of Lee's doctoral dissertation at University College London. We thank Dennis Kristensen, Krishna Pendakur, Martin Weidner, and Daniel Wilhelm for their helpful suggestions. We also greatly benefited from the seminar participants at UCL, Bocconi University, KIEP, and the KAEA micro virtual seminar. All errors are our own. The authors gratefully acknowledge support from the Social Sciences and Humanities Research Council of Canada under the Insight Grant (435-2024-0322) and PRIN PROJECT 2017 (prot.2017TMFPSH).

1. INTRODUCTION

In two-sided markets, agents form optimal matches based on their preferences and characteristics, generating a joint surplus sharable between partners according to their relative bargaining power. Matching models are widely used to analyze these dynamics, as seen in the labor market (workers matching with jobs) and the marriage market (spouses matching with each other). However, existing models often fail to capture real-world patterns due to restrictive assumptions. They often assume that agents match exclusively based on a single attribute or a scalar index aggregating multiple characteristics.[1] These assumptions are not innocuous. In the labor market, for instance, workers develop highly specialized skills – cognitive for mathematicians, and manual for gymnasts. The single index model fails to capture this specialization. To address these limitations, we need matching models that can directly accommodate multidimensional heterogeneity.[2]

A seminal paper, Lindenlaub (2017), proposes a parametric model for worker-job matching. Her model imposes joint normality of characteristics and does not allow three or more attributes. While the normality assumption enables tractable closed-form solutions for equilibrium assignment and wage functions, it may lead to misleading implications if the true distributions deviate from Gaussian. Moreover, applying the model requires data to conform to normality. Lindenlaub transformed data to standard normal and employed a Gaussian copula to introduce dependence. However, this transformation can distort the underlying relationships between attributes and their post-transformation joint distributions may remain non-normal. Hence, the estimated assignment mechanism may not accurately reflect the true matching process. Another challenge is that the theoretical model predicts

[1] Assortative spousal matching on income, wages, education, risk aversion, and preference for childbearing are investigated by Becker (1991), Grossbard-Schechtman (1993), Pencavel (1998), Choo and Siow (2006), Chiappori and Reny (2016), Legros and Newman (2007) and Chiappori and Oreffice (2008) among many others. Becker (1973) and Chiappori et al. (2012) investigate spousal matching that hinges on "ability indices".

[2] Studies such as Willis and Rosen (1979) and Papageorgiou (2014) favor the multidimensional setup over a single index model in the labor market context. Spousal choices are also based on a variety of attributes as shown in Becker (1991), Weiss and Willis (1997), Qian (1998), Silventoinen et al. (2003), Hitsch et al. (2010), and Oreffice and Quintana-Domeque (2010).

deterministic matching patterns, whereas real-world matching processes involve randomness. Lindenlaub introduced error terms to bridge this gap, relying on another strong assumption: the errors are normally distributed and uncorrelated, which, if violated, makes parameters estimated by maximum likelihood unreliable.

We overcome these challenges by developing empirically tractable semi-nonparametric multidimensional matching models. Our theoretical contributions are four-fold. First, we generalize Lindenlaub (2017)'s model by accommodating any arbitrary distributions of attributes. Our general framework is not limited to two-dimensional cases. Second, we introduce an optimal transport approach (Villani, 2003, 2008, De Philippis and Figalli, 2014) to derive unique solutions for the equilibrium assignment and wage functions. Applying optimal transport theory provides an elegant solution to our problem.[3] Third, we relax the Gaussian error assumption and allow for arbitrary correlation between errors. Lastly, we derive conditional moments from which production technology parameters, and the equilibrium assignment and wage functions are computed via efficient sieve estimators. While we focus on worker-job matching, our method can be more generally applied to other matching problems such as couple matching in the marriage market.

In the worker-job matching context, each worker possesses distinct skills, and each job requires specific skills to produce output according to production technology. The social planner's problem is to optimally assign workers to jobs to maximize total output in the economy. This problem can be formulated as an optimal transport problem, which yields the unique equilibrium wage and matching functions, accommodating multidimensional heterogeneity with arbitrary distributions. As optimal transport-based matching models predict deterministic matching patterns, following Lindenlaub (2017), we introduce error terms into the equilibrium assignment and wage functions to maintain the empirical model consistent with optimal transport theory. We estimate production technology, equilibrium assignment, and wage using the semiparametric M-estimation techniques proposed by Ai

[3] Applications of optimal transport have been proven very successful in multiple fields of economics (e.g., Ekeland (2010), Chiappori et al. (2010), Chiong et al. (2016), Lindenlaub (2017), Galichon and Salanié (2022), and many more). Galichon (2017) provides a comprehensive survey of the literature.

and Chen (2003) and Chen (2007). To our knowledge, this paper is the first in the literature to introduce semiparametric M-estimators to multidimensional matching models. Depending on assumptions on error terms, the model is estimated by sieve maximum likelihood (SML), least squares (SLS), or generalized least squares (SGLS). These estimators are efficient, asymptotically normal, and easy to implement. Our estimators perform very well in extensive simulation experiments for a wide class of data generating processes.

We apply our models to Lindenlaub (2017)'s matched worker-job data from the U.S. and estimate production technology, equilibrium assignment, and wage functions to investigate the technological shift and its effects on wage inequality between 1990 and 2010. We first estimate the models using the Gaussian transformed data. Our results show a larger technological progress favoring cognitive skills than Lindenlaub's estimates. Furthermore, more flexibility introduced in our models provides a much greater explanation power for the evolution of wage inequality, particularly the '*wage polarization*' phenomenon featuring stronger wage growth in the bottom and upper tails of the wage distribution relative to the median. The Gaussian model fails to predict wage polarization because, on top of misspecification bias, it restricts the equilibrium wage function to a quadratic form. We also conduct the same analysis on the original data that sharply differ from Gaussian. A virtue of our methods is that they can be directly applied to data without any transformation. The results from the original data show an even more significant technological shift in favor of cognitive abilities than those from the transformed data.

This paper is organized as follows. The remainder of this section discusses the related literature. Section 2 proposes the optimal transport approach for multidimensional matching. Section 3 proposes the empirical matching models and establishes the identification. Section 4 presents the sieve estimators. Section 5 derives the asymptotic properties of our sieve GLS estimator. Section 6 conducts simulation experiments. Section 7 revisits Lindenlaub (2017)'s empirical analysis. Section 8 concludes. Technical proofs and additional theoretical details are provided in the appendix.

1.1. *Related literature*

Choo and Siow (2006) (CS henceforth) introduces an empirical transferable utility (TU) model considering discrete characteristics and multidimensional unobserved heterogeneity.[4] Their discrete choice framework assumes that unobserved heterogeneity follows the extreme value type I distribution, under which the systemic match surplus is identified by logit formulae. Dupuy and Galichon (2014) extend this framework to continuous types. Galichon and Salanié (2022) allow for non-logit parametric distributions of unobserved heterogeneity and Gualdani and Sinha (2023) show partial identification of the systemic surplus under nonparametric assumptions. These papers rely on the separability assumption that unobserved heterogeneity does not have interactions in generating the surplus.

Our paper takes a different approach closely related to Lindenlaub (2017) and Bojilov and Galichon (2016). Unlike the CS framework, both papers focus on models where agents form matches given their multidimensional *continuous* attributes that are assumed to be joint normally distributed.[5] We further extend their models by dispensing with distributional assumptions, thereby offering a more flexible and robust framework for multidimensional matching. We propose efficient econometric procedures that jointly estimate both finite-dimensional parameters and infinite-dimensional functions using conditional moments implied by the model equilibrium, leveraging on the huge literature on the sieve M-estimation.[6] In particular, we employ the sieve GLS estimator proposed in Chen (2007) for our most flexible model specification. This estimator is efficient and computationally simpler than the SMD estimator.

---

[4] See Galichon et al. (2019) for the imperfectly transferable utility model with unobserved heterogeneity.

[5] Alternatively, Lise and Postel-Vinay (2020) consider a search-theoretic model in which workers are matched to firms in a dynamic setup.

[6] Shen (1997) establishes asymptotic properties of smooth functionals of sieve MLE. Newey and Powell (2003), Ai and Chen (2003, 2007), and Blundell et al. (2007) propose efficient sieve IV and sieve minimum distance (SMD) estimators. Chen and Pouzo (2009) further show that the SMD estimator under proper penalization is consistent and efficient when residuals are potentially nonsmooth. Chen (2007) provides an extensive overview of sieve estimation of semi-nonparametric models.

We establish the convergence, efficiency, and asymptotic normality of our sieve estimators relying on the smoothness of unknown nonparametric components (optimal transport maps). This smoothness condition can be verified by applying the results in the mathematical literature on optimal transport maps. Caffarelli (1992a,b, 1996) show the smoothness of transport maps when the distributions of characteristics on both sides are compactly supported. Cordero-Erausquin and Figalli (2019) further extend the earlier result to the cases where the distributions may have unbounded supports. The degree of the smoothness of a transport map depends on how smooth the densities are.

2. OPTIMAL TRANSPORT APPROACH FOR MULTIDIMENSIONAL MATCHING

We consider an environment where every worker with a bundle of skills sorts into a job demanding specific combinations of those skills. Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^d$ be spaces of worker and job characteristics endowed with probability measures P and Q respectively. Workers and jobs are described by the corresponding vectors of characteristics $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Every matched pair produces a single homogeneous good measured in money value according to production technology, $s(x, y)$. They share the produced quantity through a negotiation process that allows both parties to exploit mutually beneficial outcomes, resulting in wages and profits becoming endogenous at equilibrium. The model shares similarities with frameworks used in family economics,[7] where matches between individuals create a surplus that involves unobservable utility transfers. Unlike the marriage market, transfers between workers and jobs are observed in the labor market through wages and profits.

At the individual level, workers and jobs have many potential partners. Their choice to form a match depends on the entire set of opportunities to maximize their wages and profits, as stated below:

$$w(x) = \sup_{y \in \mathcal{Y}} \{s(x, y) - v(y)\}, \quad v(y) = \sup_{x \in \mathcal{X}} \{s(x, y) - w(x)\}. \tag{1}$$

[7] For recent reference books on theoretical matching with transferable utility, see Browning et al. (2014) and Chiappori (2017).

$w(x)$ can be interpreted as the price that the firm offering job y must pay to match with worker x. This price is specific to the worker but not to the job. Similarly, $v(y)$ can be interpreted as the price of job y. The second equation in (1) represents the firm's profit maximization problem given the wage schedule $w(x)$. We analyze such choices and the sharing of surplus from matching, within a market framework that relies on the equilibrium concept of stability. A matching is stable if no individual worker or job, nor any pair of them, prefers to deviate. Formally, if $w(x) + v(y) < s(x,y)$ for some $(x,y) \in \mathcal{X} \times \mathcal{Y}$, the matching is not stable.

As a part of the stable matching, $w^* : \mathcal{X} \to \mathbb{R}$ and $v^* : \mathcal{Y} \to \mathbb{R}$ are the solution to the following cost-minimization problem subject to stability constraints:

$$\inf_{w \in \mathcal{W}, v \in \mathcal{V}} \left\{ \mathbb{E}_P [w(X)] + \mathbb{E}_Q [v(Y)] \right\}, \text{ s.t. } w(x) + v(y) \geq s(x,y), \ \forall (x,y) \in \mathcal{X} \times \mathcal{Y}. \quad (2)$$

Here, $\mathcal{W}$ and $\mathcal{V}$ are function spaces that contain all integrable functions with respect to P and Q, respectively. A solution to (2), (w^*, v^*), also satisfies (1), meaning that we can interpret $w^*(x)$ as the equilibrium wage function for worker x and $v^*(y)$ as the equilibrium profit function for firm y in terms of Walrasian equilibrium (Galichon, 2017). Using the expression of $v(Y)$, we reformulate the optimization problem as:

$$\inf_{w \in \mathcal{W}} \left(\mathbb{E}_P [w(X)] + \mathbb{E}_Q \left[\sup_{x \in \mathcal{X}} \{ s(x,Y) - w(x) \} \right] \right). \quad (3)$$

If $w^*(x)$ is a solution to (3), then $w^*(x) + c$ is also a solution for any constant c. Thus some appropriate normalization is required to obtain a unique solution. We may impose a location constraint, $w(x_0) = 0$, for some $x_0 \in \mathcal{X}$ or a zero integration constraint, $\int_{\mathcal{X}} w(X)\, dX = 0$.

The total masses of workers and jobs are normalized to one with distributions P and Q, respectively. We define a matching as a probability measure π on $\mathcal{X} \times \mathcal{Y}$. If worker x is matched with job y, $\pi(x,y) > 0$ and the stability constraint holds with equality. Furthermore, if we sum up $\pi(x,y)$ for all y, it should be the total mass of worker x. Technically, π should satisfy the following feasibility constraints:

$$\int_{\mathcal{Y}} d\pi(x,y) = P(x), \ \forall x \in \mathcal{X}, \quad \int_{\mathcal{X}} d\pi(x,y) = Q(y), \ \forall y \in \mathcal{Y}. \quad (4)$$

The above optimization problem is a linear programming problem in w and v since the minimand and constraints are linear in these two functions. We apply the duality to this problem.[8] Indeed, (2) is the dual problem of the well-known Monge-Kantorovich optimal transport problem:

$$\sup_{\pi \in \mathcal{M}(P,Q)} \mathbb{E}_{\pi}\left[s\left(X,Y\right)\right] := \int_{\mathcal{X}\times\mathcal{Y}} s\left(x,y\right) d\pi\left(x,y\right), \tag{5}$$

where $\mathcal{M}\left(P,Q\right)$ is the set of all probability measures on $\mathcal{X}\times\mathcal{Y}$ satisfying feasibility constraints (4). We can see w and v as dual variables for feasibility constraints (4). Thus, the equilibrium wage, $w^{*}\left(x\right)$, should clear the market for worker x, and every worker x is employed with $w^{*}\left(x\right)$.

The primal optimal transport problem (5) can be understood as a social planner's problem whose solution, π^{*}, associates each x to y with a measurable function T^{*} such that $y = T^{*}(x)$. If there exists such a function T^{*}, then we can reformulate (5) using a deterministic matching function, $T : \mathcal{X} \to \mathcal{Y}$ as follows:

$$\max_{T(\cdot)} \mathbb{E}_{P}[s(X,T(X))], \quad \text{s.t. } T(P) = Q. \tag{6}$$

This is called the Monge problem proposed by Monge (1781). The solution to this problem, T^{*}, that maximizes the average overall surplus is called an optimal transport map. The primal problem (5) always has a solution that is not necessarily deterministic, while the Monge problem Monge (1781) can be ill-posed, meaning that there exists no solution T^{*} satisfying the constraint $T^{*}(P) = Q$ in some cases.[9]

Even when a deterministic solution exists, identifying the solution π^{*} in (5) is computationally very challenging except in a few cases where analytically tractable solutions exist

[8] For the duality, it is assumed that (i) $s : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{-\infty\}$ is an upper-semicontinuous function, and (ii) there are two lower semicontinuous functions $a \in \mathcal{W}$ and $b \in \mathcal{V}$ such that $s\left(x,y\right) \leq a\left(x\right) + b\left(y\right)$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. See Theorem 5.10 in Villani (2008) or Theorem 1 in Chiappori et al. (2010) for details.

[9] For example, let $\mathcal{X} = \{0\}\times[-1,1]$ and $\mathcal{Y} = \{-1,1\}\times[-1,1]$ with uniform marginal distributions $\mathcal{U}(\mathcal{X})$ and $\mathcal{U}(\mathcal{Y})$. Given the surplus function $s(x,y) = x'y$, there exists a unique optimal coupling that assigns one half of the mass at $(0,c)$ matches with $(-1,c)$ and the other half with $(1,c)$ for all $c \in [-1,1]$.

(Peyré et al., 2019). In contrast, the dual problem (2) involves optimizing over measurable and integrable functions. Under the following conditions, we can establish the unique existence and differentiability of the solution to the dual problem, say $w^* : \mathcal{X} \to \mathbb{R}$, which is closely related to T^*.

ASSUMPTION 1: *(i)* $s(x,y)$ *is differentiable as a function of* x*, for all* y*; (ii) For any fixed* $x \in \mathcal{X}$ *and* $y_1 \neq y_2 \in \mathcal{Y}$, $\nabla_x s(x,y_1) \neq \nabla_x s(x,y_2)$.

ASSUMPTION 2: *For any* $v : \mathcal{Y} \mapsto \mathbb{R} \cup \{\pm\infty\}$, $w(x) = \sup_{y\in\mathcal{Y}} \{s(x,y) - v(y)\}$ *is differentiable almost surely on*

$$\{x | \exists y \in \mathcal{Y},\ s(x,y) - w(x) \geq s(z,y) - w(z),\ \forall z \in \mathcal{X}\}.$$

Assumption 1(ii) is the twist condition, equivalent to the injectivity of $\nabla_x s(x,y)$ for each fixed x. The classical Spence-Mirrlees condition can be viewed as a twist condition in a one-dimensional context (see Carlier (2003) for more discussions). Assumption 2 is a smoothness condition on the conjugate function. There are several ways to ensure this assumption, including some restrictions on the supports of x and y, the function s, or the absolute continuity of P (see Theorem 10.28 and the following remarks in Villani (2008)). Under the above assumptions, the following proposition holds.

PROPOSITION 1: *Let Assumptions 1–2 hold. Then, there exists a unique (up to a constant) equilibrium wage function,* $w^*(x)$*, solving the dual problem* (3)*. Furthermore, the function,* $T^*(x)$*, satisfying* $\nabla w^*(x) := \nabla_x s(x,y)|_{y=T^*(x)}$*, is the unique equilibrium assignment for the Monge problem* (6)*.*

This proposition is a direct application of Theorem 10.28 in Villani (2008). Given the surplus function that satisfies Assumption 2, one can pin down the unique equilibrium wage function and hence the matching function. For instance, for the surplus function $s(x,y) = x'y$, the matching function T^* is obtained by the gradient of w^* i.e. $T^*(x) = \nabla w^*(x)$.

From now on, we specify the surplus function in the form of bi-linear technology following Lindenlaub (2017):

$$s(x,y) := s(x,y;A,b) = x'Ay + x'b, \tag{7}$$

where A is a $d \times d$ matrix and b is a $d \times 1$ vector. Here the elements of A represent worker-job complementabilities and substitutabilities. The diagonal elements capture within-task complementarities and the off-diagonal elements indicate between-task complementarities. $x'b$ represents non-interaction skill terms. Matching assortativity depends crucially on the properties of the surplus function. To fix ideas, as the simplest example of Proposition 1, consider Becker (1973)'s spousal matching model where men and women are endowed with "ability indices", $x \in \mathcal{X} \subset \mathbb{R}$ and $y \in \mathcal{Y} \subset \mathbb{R}$, respectively. If $\partial^2 s(x,y)/\partial x \partial y \geq 0$, then the stable matching function $T^* : \mathcal{X} \to \mathcal{Y}$ is defined by $T^*(x) = F_y^{-1}(F_x(x))$ where F_x and F_y are the cumulative distribution functions of x and y. Furthermore, $w^*(x) = \int_{\min(\mathcal{X})}^{x} \nabla_x s(x,y)|_{y=T^*(x)} dx + c$ is unique up to a constant c. T^* having this property is defined as positive assortative matching (PAM) in the sense that high-type males match high-type females i.e., the matching function is strictly increasing in x. Negative assortative matching (NAM) is the opposite.

In our specification, the properties of A are pivotal to the assortativity of the equilibrium assignment. Proposition 2 of Lindenlaub (2017) implies that the equilibrium matching function T^* satisfies PAM if A is a diagonal matrix with all positive principal minors. The assignment is unaffected by non-interaction terms because $\mathbb{E}_\pi[X'AY + X'b] = \mathbb{E}_\pi[X'AY] + \mathbb{E}_P[X'b]$ and the latter does not depend on the choice of π. The original problem (5) and its dual problem (3) with $s(x,y)$ can be rewritten in terms of $s^o(x,y) = x'Ay$ as follows

$$\begin{aligned}
&\inf_{w \in \mathcal{W}} \left\{ \mathbb{E}_P[w(X)] + \mathbb{E}_Q\left[\sup_{x \in \mathcal{X}} \{ s(x,Y) - w(x) \} \right] \right\} \\
&= \sup_{\pi \in \mathcal{M}(P,Q)} \mathbb{E}_\pi[s(X,Y)] = \sup_{\pi \in \mathcal{M}(P,Q)} \mathbb{E}_\pi[s^o(X,Y)] + \mathbb{E}_P[X]'b \\
&= \inf_{w^o \in \mathcal{W}} \left\{ \mathbb{E}_P[w^o(X)] + \mathbb{E}_Q\left[\sup_{x \in \mathcal{X}} \{ s^o(x,Y) - w^o(x) \} \right] \right\} + \mathbb{E}_P[X]'b.
\end{aligned} \tag{8}$$

Note that the solution w^* to the problem with original s is obtained by $w^*(x) = w^{o*}(x) + x'b + c$ with any constant c where w^{o*} is the solution to the problem with s^o.

Now we further impose conditions on two probability measures, P and Q as well as A.

ASSUMPTION 3: *(i) P and Q have finite second moments, and (ii) P is absolutely continuous with respect to the Lebesgue measure.*

ASSUMPTION 4: *The matrix A in the production technology (7) is invertible.*

Assumptions 3–4 serve as primitive conditions to satisfy Assumptions 1–2, given our production technology (7). The following statement derives the equilibrium assignment and wage in terms of the solution to the dual Monge-Kantorovich problem (8).

PROPOSITION 2: *Let Assumption 3 holds. Then, there exists the unique (up to constant) convex solution, $w^{o*}(x)$, to the second dual problem in (8), and the equilibrium wage (w^*) and assignment ($y^* = T^*(x)$ where $y^* = (y_1^*, \ldots, y_d^*)'$) are given by*

$$w^*(x) = w^{o*}(x) + x'b + c, \quad A\left(y_1^*(x) \cdots y_d^*(x)\right)' = \nabla w^{o*}(x),$$

where c is the constant of integration. In addition, if Assumption 4 holds,

$$\left(y_1^*(x) \cdots y_d^*(x)\right)' = A^{-1}\nabla w^{o*}(x).$$

We can interpret this problem as assigning from $\mathcal{X}$ to $A\mathcal{Y} := \{AY : Y \in \mathcal{Y}\}$. Assumption 3 guarantees the existence of the convex solution to the dual problem (8).[10] $w^{o*}(x)$ (and so $w^*(x)$) implicitly depends on A.

We now consider the case where attributes of firms and workers are two-dimensional ($d = 2$), tailoring the model to our data. Every worker is endowed with a bundle of cognitive and manual skills, $x = (x_C, x_M) \in \mathcal{X} \subset \mathbb{R}^2$. In turn, each firm is endowed with both cognitive and manual skill demands, $y = (y_C, y_M) \in \mathcal{Y} \subset \mathbb{R}^2$. y_C (y_M) corresponds to the productivity or skill requirement of cognitive task C (manual task M).

[10] Proposition 1 presents the unique transformation of x to y without requiring this assumption. This proposition implicitly requires conditions in footnote 7 for duality, which are all satisfied under Assumption 3(i).

With $A = \left((\alpha_{CC}, \alpha_{MC})', (\alpha_{CM}, \alpha_{MM})'\right)$ and $b = (\beta_C, \beta_M)'$, define $\delta := \frac{\alpha_{MM}}{\alpha_{CC}}$ that represents the relative level of complementarities across cognitive and manual tasks. When both complementarity parameters are positive, the value of δ smaller than 1 indicates whether worker-firm complementary in the cognitive task is stronger than in the manual task.

When x and y follow standard joint Gaussian distributions, the nonparametric component of the wage function, $w^{o*}(x)$, becomes a quadratic function of x. Lindenlaub (2017) derives T^* and w^* in closed-form assuming joint normality of x and y and estimates the production technology to investigate how the technology in the U.S. has evolved. In practice, however, x and y are non-normal as natural skills tend to have skewed distributions. To align the data with the model, Lindenlaub (2017) converts each element of x and y into a standard normal variable using inverse transform. Their dependence is then modeled using a Gaussian copula. Figure 1 illustrates how she derives the equilibrium assignment and wage function from transformed data. If the transformed data, $\tilde{x}$ and $\tilde{y}$, follow the bivariate normal distribution, this transformation provides a way of studying dependence independent of the marginals by removing the marginal characteristics. However, the joint distribution of two Gaussian random variables is not in general normal, and hence, the model based on the closed form expression for T^* and w^* can be misspecified. To avoid such an potential misspecification, we allow x and y to have any arbitrary distributions in our model.

$$
\begin{array}{ccccc}
x \sim P & \longrightarrow & \begin{array}{l} \tilde{x}_C = \Phi^{-1}\left(F_{x_C}\left(x_C\right)\right) \sim \mathrm{N}\left(0,1\right) \\ \tilde{x}_M = \Phi^{-1}\left(F_{x_M}\left(x_M\right)\right) \sim \mathrm{N}\left(0,1\right) \end{array} & \xrightarrow{\ ?\ } & \begin{pmatrix} \tilde{x}_C \\ \tilde{x}_M \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \rho_{\tilde{x}} \\ \rho_{\tilde{x}} & 1 \end{pmatrix}\right) \\
\Big\downarrow T^*\left(x\right) & & & & \Big\downarrow \tilde{T}^*\left(\tilde{x}\right) \\
y \sim Q & \longrightarrow & \begin{array}{l} \tilde{y}_C = \Phi^{-1}\left(F_{y_C}\left(y_C\right)\right) \sim \mathrm{N}\left(0,1\right) \\ \tilde{y}_M = \Phi^{-1}\left(F_{y_M}\left(y_M\right)\right) \sim \mathrm{N}\left(0,1\right) \end{array} & \xrightarrow{\ ?\ } & \begin{pmatrix} \tilde{y}_C \\ \tilde{y}_M \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \rho_{\tilde{y}} \\ \rho_{\tilde{y}} & 1 \end{pmatrix}\right)
\end{array}
$$

FIGURE 1.—Lindenlaub (2017)'s transformation. Φ denotes the standard normal c.d.f. F_{x_C} and F_{x_M} denote the c.d.f. for x_C and x_M, respectively.

The actual impact of technological shifts on wage distribution may differ from the prediction based on the Gaussian model. Lindenlaub (2017) shows that (i) wage distributions are positively skewed for any pairs of α_{CC} and α_{MM}, (ii) the variance of wage distribution increases as cognitive or manual skill complementarities increases, and (iii) wage skewness is minimized when $\alpha_{CC} = \alpha_{MM}$. However, in our simulations using non-normal distributions (detailed in Appendix C), the obtained wage distribution's skewness does not reach its minimum when $\alpha_{CC} = \alpha_{MM}$.

3. EMPIRICAL MODEL AND IDENTIFICATION

This section describes an empirical model using the theoretical results in the previous section. Let $\{(w_i, x_i', y_i')\}_{i=1}^n$ represent an independent and identically distributed (i.i.d.) sequence of n matched observations on the worker i's wage w_i, her bundle of skills x_i, and the matched job's skill demands y_i. Optimal transport-based matching models are not directly applicable to empirical analysis because they produce deterministic predictions. To address this, the models are regularized by introducing unobserved heterogeneity, search frictions, or measurement errors.[11] Here we introduce measurement error in the equilibrium functions to keep the model in line with optimal transport theory.

Lindenlaub (2017) also introduces normally distributed classical measurement errors in her equilibrium solutions and estimates the model using maximum likelihood. Based on the joint normality of x and y, the closed-form expression for $w^*(x)$ involves the productivity correlation. However, Lindenlaub (2017) uses a correlation of error contaminated $y = (y_C, y_M) \in \mathbb{R}^2$, which is different from the actual productivity correlation. Her estimation results show that the estimated variances of measurement errors are greater than one, which is not desirable in the setting where y_C and y_M are assumed standard normal. If measurement errors are introduced in the assignment equation, then the productivity correlation should be reformulated. Furthermore, if measurement errors are neither normally distributed nor homoskedastic, such a reformulation of correlation does not work.

[11] Notice that we could avoid a situation in which unobserved heterogeneity affects the assignment by assuming that it involves non-interaction terms only.

Our methods do not rely on the closed-form solution under bivariate normality, thereby free from this problem.

The introduction of measurement errors can be motivated by the construction of skill measures in data. For instance, Sanders (2014), Lindenlaub (2017) and Lise and Postel-Vinay (2020), use the U.S. Department of Labour Occupational Characteristics Database (O*NET) to determine the levels of skills required to perform each categorical task. O*NET data provides rich information (more than 270 descriptors) on skill requirements for a large number of occupations. There could be measurement errors in three possible ways. First, the researchers conventionally classify the descriptors into predetermined skill categories e.g., "cognitive", "manual", and "interpersonal". However, this decision may be far from clear-cut for many descriptors. Second, the descriptors are aggregated within each category using principal component analysis. This procedure produces inevitable measurement errors even if the descriptors are correctly classified. Lastly, there may be unobserved factors not included in O*NET for skill requirements. Hence measurement errors in skill requirements can be interpreted as (additively separable) unobserved heterogeneity.

The empirical model with measurement errors is defined by:

$$\begin{aligned} w_i &= w^*(x_i) + \varepsilon_{wi} = w^{o*}(x_i) + x_i'b + c + \varepsilon_{wi}, \\ y_i &= y_i^* + \varepsilon_{yi} = A^{-1}\nabla w^{o*}(x_i) + \varepsilon_{yi}. \end{aligned} \tag{9}$$

Here, ε_{wi} is a scalar measurement error in the observed wage but it could be also understood as the match-specific idiosyncratic shock added to the equilibrium wage. ε_{yi} is a $d \times 1$ vector of measurement errors in the firm's skill demands. This may arise due to search friction or asymmetric information. Unlike Lindenlaub (2017)'s approach, we do not have to impose distributional assumptions on measurement errors, which are vulnerable to misspecification. Instead, we impose the following moment conditions assuming the exogeneity of x_i:

$$\mathbb{E}[\varepsilon_{wi}|x_i] = 0, \quad \mathbb{E}[\varepsilon_{yi}|x_i] = 0. \tag{10}$$

Let $\theta = \left(\text{vec}\left(A^{-1}\right)', b'\right)'$ denote a vector of unknown finite-dimensional parameters and $\theta \in \Theta$ where Θ is a compact subset of $\mathbb{R}^{d^2+d}$. The normalizing constant is not in

our parameters of interest, and henceforth, we refer to $w(x) := w^{o*}(x) + c$ as the constant added infinite-dimensional parameter. We denote $z_i = (w_i, x_i', y_i')'$ and $\rho(z_i; \theta, w) = \left(\rho_w(w_i, x_i; \theta, w), \rho_y(y_i, x_i; \theta, w)'\right)'$, where

$$\rho_w(w_i, x_i; \theta, w) = w_i - \left(w(x_i) + x_i'b\right), \quad \rho_y(y_i, x_i; \theta, w) = y_i - A^{-1}\nabla w(x_i).$$

For each observation i, the model (9) satisfies the moment conditions (10). This implies that the following conditional moments hold:

$$\mathbb{E}\left[\rho(z_i; \theta, w) | x_i\right] = 0, \tag{11}$$

at a true parameter (θ_0, w_0). Then (θ_0, w_0) are identified via the model (11) by Proposition 2 and the exogeneity of x_i as well as the following assumption on $\mathcal{Y}$.

ASSUMPTION 5: *There exist $y_1, \ldots, y_d, y_{d+1} \in \mathcal{Y}$ such that $\{y_1 - y_2, \ldots, y_d - y_{d+1}\}$ is linearly independent.*

Assumptions 3 and 4, combined with Proposition 2, imply the existence of a deterministic equilibrium characterized by a unique convex function w_0. When there is no non-interaction term with $b_0 = 0$, it follows that $\nabla w_0^* = \nabla w_0$. The strict convexity of w_0 further implies that $\mathbb{E}\left[\nabla w_0(x_i) \nabla w_0(x_i)'\right]$ has full rank, thus identifying A_0. Additionally, Assumption 5 is sufficient to identify the nonzero vector b_0, as stated in the following theorem.

THEOREM 1: *Let Assumptions 3-5 hold and the moment conditions (11) be satisfied. Then, θ_0 and $w_0 = w_0^{o*} + c_0$ are identified.*

We can further identify w_0^{o*} and c_0 separately under the normalization such as $w_0^{o*}(x_0) = 0$ for some $x_0 \in \mathcal{X}$ or $\int_{\mathcal{X}} w_0^{o*}(x)dx = 0$ when $\mathcal{X}$ is bounded. With the former constraint, c_0 and $w_0^{o*}(x)$ are identified with $w_0^*(x) = w_0^{o*}(x) + x'b$ since $c_0 = w_0^{o*}(x_0) + c_0 = w_0^*(x_0) - x_0'b$.

4. SIEVE-BASED SEMIPARAMETRIC ESTIMATION

The model parameters are identified by the semiparametric conditional moment restrictions (11). If the function w is parametrically specified, these moment conditions lead to standard GMM estimation. As w is infinite-dimensional in our specification, we approximate it using sieves. The unknown function $w \in \mathcal{W}$ is approximated by $w_n \in \mathcal{W}_n$ where $\mathcal{W}_n$ is an approximating multivariate function space becoming dense in $\mathcal{W}$ as $n \to \infty$. We generate $\mathcal{W}_n$ via tensor-product construction.

From now on, we assume that there are sets of firms and workers with $d = 2$ without loss of generality. Every worker is endowed with a bundle of cognitive and manual skills, $x = (x_C, x_M)$. In turn, each firm is endowed with both cognitive and manual skill demands, $y = (y_C, y_M)$. y_C (y_M) corresponds to the productivity or skill requirement of cognitive task C (manual task M). In our case,

$$\mathcal{W}_n = \left\{ w_n : \mathcal{X} \to \mathbb{R}, w_n(x;\gamma) = \sum_{j_C=0}^{k_{Cn}} \sum_{j_M=0}^{k_{Mn}} \gamma_{j_C j_M} p_{j_C}(x_C)\, p_{j_M}(x_M), \gamma_{j_C j_M} \in \mathbb{R} \right\}, \tag{12}$$

where $\{p_{j_C}(x_C)\}_{j_C=0}^{k_{Cn}}$ and $\{p_{j_M}(x_M)\}_{j_M=0}^{k_{Mn}}$ are known basis functions of x_C and x_M. Tensor-product space is simple to extend with higher dimensions and easy to implement. For our second and third conditional moment restrictions, we approximate $\partial w_0(x)/\partial x_C$ and $\partial w_0(x)/\partial x_M$ with same parameter values $\{\gamma_{j_C j_M}\}$ used to approximate $w_0(x)$ in $\mathcal{W}_n$:

$$\partial w_n(x;\gamma)/\partial x_C = \sum_{j_C=0}^{k_{Cn}} \sum_{j_M=0}^{k_{Mn}} \gamma_{j_C j_M} \left(\partial p_{j_C}(x_C)/\partial x_C\right) p_{j_M}(x_M),$$

$$\partial w_n(x;\gamma)/\partial x_M = \sum_{j_C=0}^{k_{Cn}} \sum_{j_M=0}^{k_{Mn}} \gamma_{j_C j_M} p_{j_C}(x_C) \left(\partial p_{j_M}(x_M)/\partial x_M\right).$$

We first consider the model (9) with normally distributed mean-zero measurement errors that are uncorrelated with each other and a diagonal matrix $A = \operatorname{diag}(\alpha_{CC}, \alpha_{MM})$ which rules out between-task complementarities. Then this model is effectively the Lindenlaub (2017) model without joint normality of x_i and y_i. We can estimate the parameters using sieve maximum likelihood (SML). Assuming $\varepsilon_i \sim N(0, \Sigma)$, we write the log-likelihood

function of model (9) as

$$L^{*}\left(\theta,\Sigma,w\left(\cdot\right)\right)=-\frac{n}{2}\log\det\left(\Sigma\right)+\sum_{i=1}^{n}\log\left|\det\left(\partial\rho_{i}/\partial\left(w_{i},y_{Ci},y_{Mi}\right)\right)\right|-\frac{1}{2}\sum_{i=1}^{n}\rho_{i}^{\prime}\Sigma^{-1}\rho_{i},$$

where $\rho_{i}=\rho\left(z_{i};\theta,w\right)$. Solving $\partial L^{*}/\partial\Sigma=0$ for Σ, we get $\Sigma=\frac{1}{n}\sum_{i=1}^{n}\rho_{i}\rho_{i}^{\prime}$, which yields the concentrated log-likelihood function of our model

$$L\left(\theta,w\left(\cdot\right)\right)=-\frac{n}{2}\log\det\left(\frac{1}{n}\sum_{i=1}^{n}\rho_{i}\rho_{i}^{\prime}\right). \tag{13}$$

The value of (θ,w) maximizing (13) is the sieve nonlinear full information maximum likelihood estimator of (θ,w).

The normality assumption on measurement errors has no theoretical or empirical ground. Without any distributional assumptions on measurement errors, we can still estimate (θ,w) using several sieve M-estimators. As $\rho\left(z;\theta,w\right)-\rho\left(z;\theta_{0},w_{0}\right)$ does not depend on y under Assumption 4, we can apply the sieve generalized least squares (GLS) procedure (Chen, 2007) that minimizes the following objective function with respect to (θ,w):

$$\min_{(\theta,w)}\sum_{i=1}^{n}\rho\left(z_{i};\theta,w\right)^{\prime}\left[\hat{\Sigma}_{0}\left(x_{i}\right)\right]^{-1}\rho\left(z_{i};\theta,w\right),$$

where $\hat{\Sigma}_{0}\left(x\right)$ is a consistent estimator of the optimal weighting matrix $\Sigma_{0}\left(x\right):=\mathrm{Var}\left[\rho\left(z_{i};\theta,w\right)|\,x_{i}=x\right]$. In addition, if A is diagonal, we can rewrite the last two moment conditions as $\rho_{C}\left(y_{C},x;\kappa_{C},w\right)=y_{C}-\kappa_{C}\nabla_{C}w\left(x\right)$ and $\rho_{M}\left(y_{M},x;\kappa_{M},w\right)=y_{M}-\kappa_{M}\nabla_{M}w\left(x\right)$, where $\kappa_{C}=\alpha_{CC}^{-1}$ and $\kappa_{M}=\alpha_{MM}^{-1}$. Table I outlines the three-step procedure to compute the SGLS estimator.

The SGLS estimator allows for arbitrary correlation between measurement errors and heteroskedasticity. We can impose homoskedasticity by assuming $\Sigma_{0}\left(x\right)=\Sigma_{0}$ so that the optimal weighting matrix does not vary with x. If we further assume that the measurement errors are uncorrelated i.e., Σ_{0} is diagonal, we can use the sieve least squares (SLS) estimator from step 1 of the three-step procedure. We summarize the key differences in assumptions imposed in different estimation procedures in Table II.

TABLE I

THREE-STEP PROCEDURE FOR SIEVE GLS ESTIMATION (CHEN, 2007)

ALGORITHM: Computing the Sieve GLS Estimator of θ and w

1. Obtain an initial consistent sieve LS estimator $\left(\tilde{\theta}_n, \tilde{w}_n\right)$ by $\min_{(\theta,w)} \sum_{i=1}^{n} \rho(z_i;\theta,w)' \rho(z_i;\theta,w)$,
2. Obtain a consistent estimator $\hat{\Sigma}_0(x)$ of $\Sigma_0(x) = \mathrm{Var}[\rho(Z;\theta,w)|X=x]$ using $\left(\tilde{\theta}_n, \tilde{w}_n\right)$ and sieve LS estimation.
3. Obtain the optimally weighted sieve GLS estimator $\left(\hat{\theta}_n, \hat{w}_n\right)$ by $\min_{(\theta,w)} \sum_{i=1}^{n} \rho(z_i;\theta,w)' \left[\hat{\Sigma}_0(x_i)\right]^{-1} \rho(z_i;\theta,w)$.

TABLE II

COMPARISON OF KEY ASSUMPTIONS IN ESTIMATION PROCEDURES

Assumptions	Lindenlaub (2017)	Sieve Estimators		
	ML	SML	SLS	SGLS
Joint normality of x and y	✓			
Normality of measurement errors	✓	✓		
Uncorrelated measurement errors	✓	✓		
Homoskedasticity of measurement errors	✓	✓	✓	

We implement the sieve estimators using finite-dimensional Bernstein polynomials to construct the approximating space $\mathcal{W}_n$ of $\mathcal{W}$ on $[0,1]^2$. The basis functions are $p_{j_C}(x_C) = \binom{k_{C_n}}{j_C} (x_C)^{j_C} (1-x_C)^{k_{C_n}-j_C}$ and $p_{j_M}(x_M) = \binom{k_{M_n}}{j_M} (x_M)^{j_M} (1-x_M)^{k_{M_n}-j_M}$, where $j_C = 0,1,\ldots,k_{C_n}$, $j_M = 0,1,\ldots,k_{M_n}$, and $\binom{k}{j}$ is a binomial coefficient.[12]

If $\gamma_{j_C j_M} = w(j_C/k_{C_n}, j_M/k_{M_n})$, the Bernstein polynomial $w_n(x;\gamma)$ converges uniformly to $w(x)$ by the Stone-Weierstrass approximation theorem (see, e.g., Lorentz (1986)).

[12] x does not lie in $[0,1]^2$ in many applications. To satisfy the domain restriction for our simulation studies and empirical application, we use the following linear transformation when $(x_C, x_M) \in [\underline{x}_C, \overline{x}_C] \times [\underline{x}_M, \overline{x}_M]$: $p_{j_C}(x_C) = \binom{k_{C_n}}{j_C} x_1^{j_C} (1-x_1)^{k_{C_n}-j_C}$ and $p_{j_M}(x_M) = \binom{k_{M_n}}{j_M} x_2^{j_M} (1-x_2)^{k_{M_n}-j_M}$, where $x_1 = (x_C - \underline{x}_C)/(\overline{x}_C - \underline{x}_C)$ and $x_2 = (x_M - \underline{x}_M)/(\overline{x}_M - \underline{x}_M)$.

This provides an approach to imposing shape restrictions on the sieve estimator with a linear constraint which can be solved easily.[13] Without any constraint, the equilibrium wage function, $w(x)$, is unique and convex (up to constant). To obtain a more stable estimator, without loss of generality, we impose linear constraints on the Bernstein polynomials, which are necessary for the function to be convex. A detailed description of implementing this convexity constraint in the estimation procedures is provided in Appendix B.

To understand technological changes in the production function, the parametric components of the model are of primary interest. The SGLS estimator is ideal in this case because for θ (i) it is easy to use, $\sqrt{n}$-consistent, and asymptotically normal; (ii) it is semiparametrically efficient; and (iii) the asymptotic variance estimator of $\hat{\theta}$ is consistent and easy-to-compute.[14] We formally derive its asymptotic properties in the following section.

5. ASYMPTOTIC THEORY FOR THE SGLS ESTIMATOR

We establish consistency, convergence rate, asymptotic normality, and semiparametric efficiency of our SGLS estimator using results in Chen and Shen (1998), Ai and Chen (2003), and Chen (2007). Define $\lambda := (\theta, w(\cdot))$. Let $\hat{\lambda}_n$ and λ_0 denote our sieve GLS estimator and the true parameter values, respectively. We first show that $\hat{\lambda}_n$ converges to λ_0 at a rate faster than $n^{-1/4}$ under a pseudo norm $\|\cdot\|$. For any $\lambda_1 = (\theta_1, w_1(\cdot)), \lambda_2 = (\theta_2, w_2(\cdot)) \in \Lambda$, $\|\cdot\|$ is defined as

$$\|\lambda_1 - \lambda_2\|^2 = \mathbb{E}\left[\left(\frac{d\rho(z_i;\lambda_0)}{d\lambda}[\lambda_1 - \lambda_2]\right)' \Sigma(x_i)^{-1} \left(\frac{d\rho(z_i;\lambda_0)}{d\lambda}[\lambda_1 - \lambda_2]\right)\right],$$

where

$$\frac{d\rho(z;\lambda_0)}{d\lambda}[\lambda_1 - \lambda_2] = \begin{pmatrix} w_1(x) - w_2(x) + x'(b_1 - b_2) \\ \nabla_C w_0(x)(\kappa_{C1} - \kappa_{C2}) + \kappa_{C0}(\nabla_C w_1(x) - \nabla_C w_2(x)) \\ \nabla_M w_0(x)(\kappa_{M1} - \kappa_{M2}) + \kappa_{M0}(\nabla_M w_1(x) - \nabla_M w_2(x)) \end{pmatrix}.$$

[13] Compiani (2022) uses linear constraints for the function w to impose monotonicity restrictions and a so-called "diagonal dominance" constraint.

[14] The sieve minimum distance estimator can be considered. However, when $\rho(z;\theta,w) - \rho(z;\theta_0,w_0)$ does not depend on y, the SGLS estimator is simpler to implement and computationally faster.

The pseudo metric is comparatively weaker than the standard sup or L_2 metric, wherein convergence of $\hat{\lambda}_n$ to λ_0 based on the standard metric implies convergence of $\hat{\lambda}_n$ using the pseudo metric. Ai and Chen (2003) show that $\hat{\lambda}_n$ converging at a rate faster than $n^{-1/4}$ under the weaker metric $||\cdot||$ suffices to derive the $\sqrt{n}$-asymptotic normality of the parametric component, $\hat{\theta}_n$.

Let $\Lambda = \Theta \times \mathcal{W}$ be equipped with a norm $\|\lambda\|_s = |\theta|_e + \|w\|_\infty + \|\nabla_C w\|_\infty + \|\nabla_M w\|_\infty$, where $|\cdot|_e$ denotes the Euclidean norm and $\|w\|_\infty = \sup_{x \in \mathcal{X}} |w(x)|$ is the supremum norm. We introduce the Hölder class of functions. Let $[m]$ be the largest nonnegative integer such that $[m] < m$. A real-valued function w on $\mathcal{X}$ is said to be in Hölder space $\Lambda^m(\mathcal{X})$ if it is $[m]$ times continuously differentiable on $\mathcal{X}$ and

$$\max_{\ell_1+\ell_2 \leq [m]} \sup_x \left| \frac{\partial^{\ell_1+\ell_2} w(x)}{\partial x_C^{\ell_1} \partial x_M^{\ell_2}} \right| + \sup_{m_1+m_2=[m]} \sup_{x,x'} \left| \frac{\partial^{[m]} w(x)}{\partial x_C^{m_1} \partial x_M^{m_2}} - \frac{\partial^{[m]} w(x')}{\partial x_C^{m_1} \partial x_M^{m_2}} \right| / \left| x - x' \right|_e^{m-[m]}$$

is finite. We provide the following assumptions for convergence.

ASSUMPTION 6: *(i)* $\{w_i, y_i', x_i'\}_{i=1}^n$ *are i.i.d.; (ii)* $\mathcal{X}$ *is compact and a Cartesian product of compact intervals* $\mathcal{X}_C$ *and* $\mathcal{X}_M$.

ASSUMPTION 7: $\Sigma(x)$ *and* $\Sigma_0(x) \equiv \mathrm{Var}(\rho(z_i, \lambda_0) | x_i = x)$ *are positive definite and bounded uniform over* $x \in \mathcal{X}$.

ASSUMPTION 8: $\Lambda \equiv \Theta \times \mathcal{W}$ *is compact under* $\|\cdot\|_s$.

ASSUMPTION 9: *(i)* $w \in \Lambda^m(\mathcal{X})$ *with* $m > 2$*; (ii)* $\forall w \in \Lambda^m(\mathcal{X}), \exists w_n(x;\gamma) \in \mathcal{W}_n$ *such that* $\|w_n - w\|_\infty = O((k_{Cn} k_{Mn})^{-m/2})$ *with* $k_{Cn}, k_{Mn} = O\left(n^{1/2(m+1)}\right)$.

Assumptions 6–8 are typical conditions imposed in the estimation of conditional mean functions with the tensor product of finite-dimensional linear sieves. We do not explicitly require identification of λ here as Assumptions 3–4 guarantee it by Theorem 1. Assumption 9 quantifies the deterministic approximation error of functions in $\Lambda^m(\mathcal{X})$ by the linear sieve basis functions. Most papers in the literature require $m > d_{\mathcal{X}}/2$, where $d_{\mathcal{X}}$ is the dimension of $\mathcal{X}$. However, our objective function involves $\nabla_C w(x)$ and $\nabla_M w(x)$, so we

need a higher order of m. Note that the smoothness of w can be verified using the theory of optimal transport. The degree of the smoothness of the solution function w depends on how smooth the densities of x and y^* are.[15] The following proposition establishes the convergence rate of $\hat{\lambda}_n$.

PROPOSITION 3: *If Assumptions 3-9 hold, then* $\|\hat{\lambda}_n - \lambda_0\| = o_p\left(n^{-1/4}\right)$.

We now derive the asymptotic normality of the parametric components of the SGLS estimator, $\hat{\theta}_n$. Define $D_v(x) := (D_{v_1}(x), D_{v_2}(x), D_{v_3}(x), D_{v_4}(x))$ where

$$D_{v_1}(x) = \begin{pmatrix} v_1(x) \\ \kappa_{C0}\nabla_C v_1(x) - \nabla_C w_0(x) \\ \kappa_{M0}\nabla_M v_1(x) \end{pmatrix}, \quad D_{v_2}(x) = \begin{pmatrix} v_2(x) \\ \kappa_{C0}\nabla_C v_2(x) \\ \kappa_{M0}\nabla_M v_2(x) - \nabla_M w_0(x) \end{pmatrix},$$

$$D_{v_3}(x) = \begin{pmatrix} v_3(x) - x_C \\ \kappa_{C0}\nabla_C v_3(x) \\ \kappa_{M0}\nabla_M v_3(x) \end{pmatrix}, \qquad D_{v_4}(x) = \begin{pmatrix} v_4(x) - x_M \\ \kappa_{C0}\nabla_C v_4(x) \\ \kappa_{M0}\nabla_M v_4(x) \end{pmatrix}.$$

Let $v^* = (v_1^*, v_2^*, v_3^*, v_4^*)$, where v_j^* solves

$$\inf_{v_j} \mathbb{E}\left[D_{v_j}(x_i)' \Sigma(x_i)^{-1} D_{v_j}(x_i)\right]. \tag{14}$$

ASSUMPTION 10: *(i)* $\mathbb{E}\left[D_{v^*}(x_i)' D_{v^*}(x_i)\right]$ *is positive definite; (ii) Each element of* v^* *belongs to the Hölder space* $\Lambda^m(\mathcal{X})$ *with* $m > 2$.

ASSUMPTION 11: $\theta_0 \in \text{int}(\Theta)$.

Under Assumptions 3–10, it is clear to see from Lemma B.1 in Ai and Chen (2003) that $|\hat{\theta}_n - \theta_0|_e = o_p\left(n^{-1/4}\right)$, $\|\hat{w}_n - w_0\|_2 = \left(\mathbb{E}\left[\left(\hat{w}_n(x_i) - w_0(x_i)\right)^2\right]\right)^{1/2} = o_p\left(n^{-1/3}\right)$,

[15] Assuming the densities are bounded away from zero and infinity, if the densities of variables x and y^* belong to the space Λ^{m-2}, the function w_0 is a member of $\Lambda^m(\mathcal{X})$. For a more comprehensive understanding, refer to Caffarelli (1992a,b, 1996) which covers the case of compactly supported $\mathcal{X}$ and $\mathcal{Y}^*$. Cordero-Erausquin and Figalli (2019) provides an extended result for distributions with unbounded supports.

and $\|\nabla_C \hat{w}_n - \nabla_C w_0\|_2, \|\nabla_M \hat{w}_n - \nabla_M w_0\|_2 = o_p\left(n^{-1/4}\right)$. Now the following theorem provides the asymptotic normality of $\hat{\theta}_n$.

THEOREM 2: *Let Assumptions 3–11 hold. Then,* $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N\left(0, V_1^{-1} V_2 V_1^{-1}\right)$, *where*

$$\begin{aligned} V_1 &= \mathbb{E}\left[D_{v^*}(x_i)' \Sigma(x_i)^{-1} D_{v^*}(x_i)\right], \\ V_2 &= \mathbb{E}\left[D_{v^*}(x_i)' \Sigma(x_i)^{-1} \Sigma_0(x_i) \Sigma(x_i)^{-1} D_{v^*}(x_i)\right]. \end{aligned} \tag{15}$$

The asymptotic variance $V_1^{-1} V_2 V_1^{-1}$ can be consistently estimated (see, Remark 4.2 in Chen (2007)) and the standard errors of $(\hat{\alpha}_{CC}, \hat{\alpha}_{MM}) = (1/\hat{\kappa}_C, 1/\hat{\kappa}_M)$ are obtained by using the delta method. Furthermore, if all conditions of Theorem 2 are satisfied with $\Sigma(x) = \Sigma_0(x)$, $\hat{\theta}_n$ achieves semiparametric efficiency with a consistent estimator $\hat{\Sigma}_0(x)$ of $\Sigma_0(x)$. The estimation of $\Sigma_0(x)$ is straightforward through series least square estimation, using the initial consistent SLS estimator $(\tilde{\theta}_n, \tilde{w}_n)$. To ensure the efficiency of the SGLS estimator, $\hat{\Sigma}_0(x)$ is required to exhibit the following uniform convergence rate.

ASSUMPTION 12: $\hat{\Sigma}(x) = \Sigma_0(x) + o_p\left(n^{-1/4}\right)$ *uniformly over* $x \in \mathcal{X}$.

Let $v_0 = (v_{01}, v_{02}, v_{03}, v_{04})$, where v_{0j} solves (14) with $\Sigma(x)$ replaced by $\Sigma_0(x)$. Now the following theorem establishes the semiparametric efficiency of $\hat{\theta}_n$.

THEOREM 3: *Suppose that all conditions of Theorem 2 with* $\Sigma(x) = \Sigma_0(x)$ *and* $v^* = v_0$ *hold, and Assumption 12 is satisfied. Then,* $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N\left(0, V_0^{-1}\right)$, *with* $V_0 = \mathbb{E}\left[D_{v_0}(x_i)' \Sigma_0(x_i)^{-1} D_{v_0}(x_i)\right]$.

6. MONTE CARLO SIMULATIONS

This section evaluates the finite sample performances of our sieve estimators using known data-generating processes (DGPs). We first generate Monte Carlo samples from Lindenlaub (2017)'s quadratic-Gaussian model. Workers' skill bundle, x, and occupations'

skill requirements, y, follow joint Gaussian distributions:

$$\begin{pmatrix} x_C \\ x_M \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_x \\ \rho_x & 1 \end{pmatrix} \right), \quad \begin{pmatrix} y_C \\ y_M \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_y \\ \rho_y & 1 \end{pmatrix} \right),$$

from which $\{x_i\}_{i=1}^n$ are drawn with sample size $n = 3000$. Given the production technology (7) assuming A is diagonal, the equilibrium assignment y^* is provided by

$$y^* = \begin{pmatrix} y_C^* \\ y_M^* \end{pmatrix} = \underbrace{\begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}}_{:=J} \begin{pmatrix} x_C \\ x_M \end{pmatrix},$$

where

$$J = \frac{1}{\sqrt{1 + 2\delta(\rho_x\rho_y + \sqrt{1-\rho_y^2}\sqrt{1-\rho_x^2}) + \delta^2}} \begin{pmatrix} 1 + \delta\frac{\sqrt{1-\rho_y^2}}{\sqrt{1-\rho_x^2}} & \delta\left(\rho_y - \rho_x\frac{\sqrt{1-\rho_y^2}}{\sqrt{1-\rho_x^2}}\right) \\ \rho_y - \rho_x\frac{\sqrt{1-\rho_y^2}}{\sqrt{1-\rho_x^2}} & \delta + \frac{\sqrt{1-\rho_y^2}}{\sqrt{1-\rho_x^2}} \end{pmatrix}.$$

Recall that $\delta = \frac{\alpha_{MM}}{\alpha_{CC}}$. We generate $\{y_i^*\}_{i=1}^n$ using the equilibrium assignment function. Then we generate the equilibrium wage $\{w_i^*\}_{i=1}^n$ using

$$w_0^*(x) = \frac{\alpha_{CC}}{2}(J_{11}x_C^2 + 2J_{12}x_Cx_M + \delta J_{22}x_M^2) + \beta_C x_C + \beta_M x_M + c.$$

Lastly, we draw measurement errors from mean-zero Gaussian distributions:

$$\varepsilon_w \sim N(0, \sigma_w^2), \quad \varepsilon_C \sim N(0, \sigma_C^2), \quad \varepsilon_M \sim N(0, \sigma_M^2),$$

and add them to w_i^*, y_{Ci}^*, and y_{Mi}^* respectively to generate the observable data $(w_i, y_i, x_i)_{i=1}^n$ following (9). The true parameter values used in simulations are:

$$(\alpha_{CC}, \alpha_{MM}, \beta_C, \beta_M, c, \rho_x, \rho_y, \sigma_w, \sigma_C, \sigma_M) = (0.5, 0.2, 1.7, -0.4, 30, -0.4, -0.5, 2, 1, 1),$$

which are close to the ML estimates in Lindenlaub (2017).

We estimate the production technology parameters using Lindenlaub (2017)'s parametric ML estimator and our sieve estimators (SML, SLS, and SGLS) across 1000 Monte Carlo samples. As we mentioned earlier, the ML estimator in Lindenlaub (2017) suffers

from bias because the solution uses measurement error contaminated productivity correlation $\tilde{\rho}_y = corr(y)$ which is different from the true productivity correlation $\rho_y = corr(y^*)$. If the measurement errors in y are negligible e.g. (σ_C, σ_M) are close to 0, the ML estimator works well for this DGP. However, given the current parameter specification, the measurement errors are substantial so the ML estimator can be highly inconsistent. To address this issue, we define a corrected ML estimator (referred to as 'ML*') that uses the corrected productivity correlation:

$$\rho_y = \frac{\tilde{\rho}_y \sqrt{var(y_1) var(y_2)}}{\sqrt{var(y_1) - \sigma_C^2} \sqrt{var(y_2) - \sigma_M^2}}.$$

This correction in turn yields much more precise estimates than the original ML estimator. The sieve estimators do not share this problem.

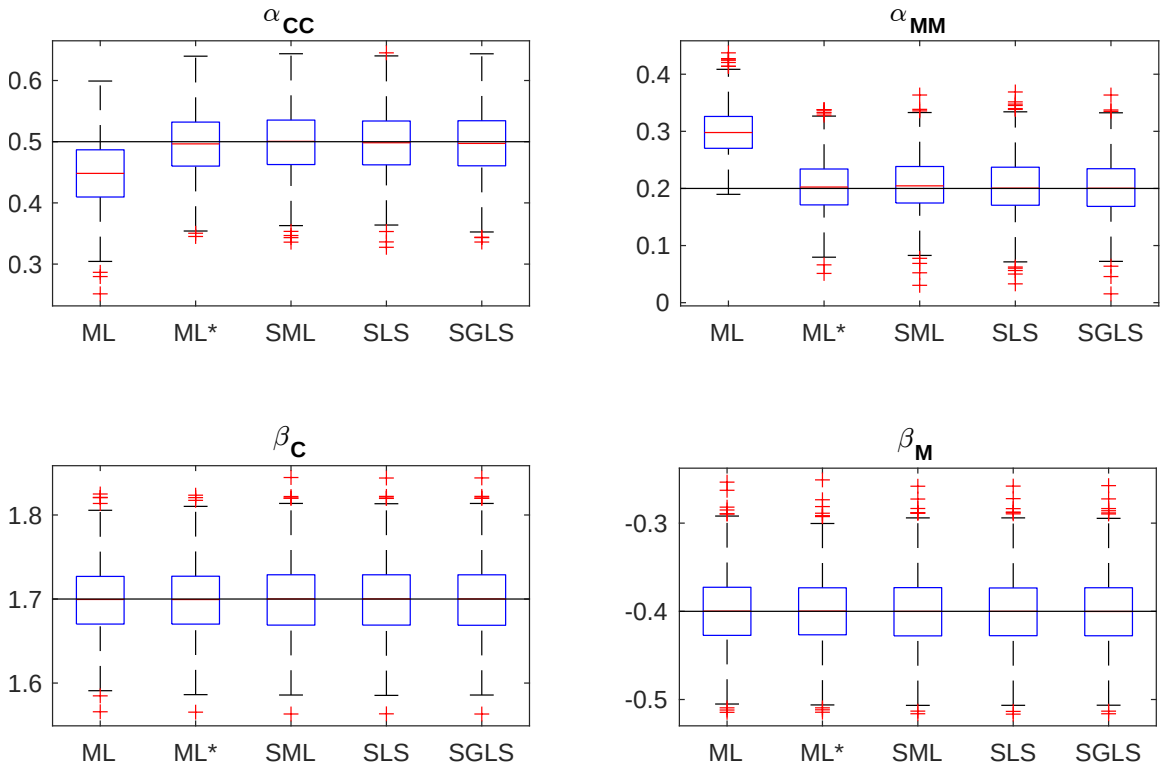

FIGURE 2.—Box plots of parameter estimates (Gaussian DGP)

The box plots in Figure 2 summarize the distributions of parameter estimates delivered by the 5 estimators we consider. For the linear coefficients (β_C, β_M), all the estimators equally work well. On the other hand, it is clear that the original ML estimator is heavily biased for the complementarity parameters $(\alpha_{CC}, \alpha_{MM})$ as expected. The other 4 estimators perform very well for $(\alpha_{CC}, \alpha_{MM})$ as the distributions of their parameter estimates are centered around the true parameter values. We document the estimators' bias and root-mean-squared errors (RMSE) in Table III. It is surprising that the sieve estimators, while more robust than ML and ML* estimators, tend to be not less efficient than the parametric ML estimators. The SML estimator's root-mean-squared errors (RMSE) are a tad larger than the ML* estimator for most parameters. The SLS and SGLS estimators, perform very similarly as the measurement errors are uncorrelated, are slightly less efficient than the SML estimator as expected.

TABLE III

FINITE SAMPLE PERFORMANCES OF THE ESTIMATORS (GAUSSIAN DGP)

		ML	ML*	SML	SLS	SGLS
α_{CC}	Bias	-0.0536	-0.0041	-0.0007	-0.0018	-0.0027
	RMSE	0.0775	0.0513	0.0520	0.0523	0.0523
α_{MM}	Bias	0.0992	0.0044	0.0065	0.0031	0.0022
	RMSE	0.1077	0.0473	0.0491	0.0502	0.0489
β_C	Bias	-0.0006	-0.0007	-0.0009	-0.0009	-0.0009
	RMSE	0.0416	0.0414	0.0427	0.0427	0.0427
β_M	Bias	-0.0000	-0.0001	-0.0000	-0.0000	-0.0000
	RMSE	0.0398	0.0395	0.0397	0.0397	0.0397

Now we consider DGPs for which the Gaussian model is moderately misspecified. We first generate $\{x_{1i}, x_{2i}\}_{i=1}^n$ and $\{y_{1i}, y_{2i}\}_{j=1}^n$ separately from the Gumbel copula with the shape parameter values 1.3 and 1.4 respectively. Then we transform them into standard

normally distributed variables:

$$x_{Ci} = \Phi^{-1}(x_{1i}), \quad x_{Mi} = \Phi^{-1}(1 - x_{2i}), \quad y_{Cj} = \Phi^{-1}(y_{1j}), \quad y_{Mj} = \Phi^{-1}(1 - y_{2j}),$$

so that (x_{Ci}, x_{Mi}) and (y_{Cj}, y_{Mj}) are negatively correlated. Define matrices x and y by

$$x := \begin{bmatrix} x_{C1} & x_{M1} \\ \vdots & \vdots \\ x_{Cn} & x_{Mn} \end{bmatrix}, \quad y := \begin{bmatrix} y_{C1} & y_{M1} \\ \vdots & \vdots \\ y_{Cn} & y_{Mn} \end{bmatrix}.$$

The skill demand and supply bundles are standard normally distributed in this DGP but their joint distributions are not Gaussian. The Gumbel copula exhibits asymmetric tail dependence (the upper tail has stronger dependence than the lower tail), whereas the Gaussian copula has symmetric dependence. However, given the current parameter setup, the Gumbel copula does not drastically differ from the Gaussian copula. The production technology is specified the same as before.

There exists no closed-form solution for the equilibrium assignment in this case. We, therefore, numerically solve the equilibrium matching through linear programming for each Monte Carlo sample. To do so, we first compute the pairwise surplus of each possible match between x and y and construct the surplus matrix S whose ij entry is the surplus generated by worker i and firm j. Let $\mathbb{I}_n$ denote a $n \times n$ identity matrix and $\mathbf{1}_n$ be a $n \times 1$ one vector. Let f be a vector generated by flattening S by column. Then the solution (x^*) to the following linear programming problem provides the equilibrium assignment:

$$\max_x f'x, \quad \text{s.t.} \begin{bmatrix} \underbrace{A_1}_{n \times n^2} \\ \underbrace{A_2}_{n \times n^2} \end{bmatrix} x \leq \underbrace{b}_{2n \times 1}, \tag{16}$$

where $A_1 := \mathbb{I}_n \bigotimes \mathbf{1}_n'$, $A_2 := \mathbf{1}_n' \bigotimes \mathbb{I}_n$, and $b := \mathbf{1}_{2n}$. Reshaping x^* into a $n \times n$ matrix gives the optimal transportation matrix, T. Then the optimal matching for x is given by $y^* := Tx$. The wage w^* is computed by the solution to the dual problem (16).[16]

For measurement errors, we consider two different specifications. In the first case, the errors are independently drawn from a gamma distribution, $\Gamma(a, b)$, with $a = 1$, $b = 2$. They are demeaned and scaled to have the same means and variances specified in the Gaussian DGP. Under this specification, the ML* and SML estimators are further misspecified as measurement errors are non-normal. In the second case, we generate the errors from a joint normal distribution in which the errors are correlated as follows:

$$\begin{pmatrix} \varepsilon_w \\ \varepsilon_C \\ \varepsilon_M \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0.5 \\ 1 & 0.5 & 1 \end{pmatrix} \right).$$

The ML* and SML estimators are still misspecified as measurement errors are correlated. The SLS estimator is consistent but not as efficient as the SGLS estimator as it does not take the correlation structure of errors into account. The observable data $(y_i, x_i, w_i)_{i=1}^n$ are generated by adding the measurement errors to w_i^*, y_{Ci}^*, and y_{Mi}^* respectively.

The estimation results are provided in Table IV. In both cases, the ML* estimator is misspecified for both the distributions of X, Y, and measurement errors so that it performs the worst for all the parameters. It exhibits especially large biases and RMSE for complementarity parameters. Even for the linear productivity parameters, the ML* estimator shows much larger RMSEs than the sieve estimators. In contrast, all the sieve estimators equally work well for the linear coefficients. The SML estimator is misspecified for the distributions of measurement errors but it produces accurate estimates for the complementary parameters $(\alpha_{CC}, \alpha_{MM})$ in both cases. The SLS and SGLS estimators perform similarly to the SML estimator when the measurement errors are drawn from the Gamma distributions. In the case of correlated errors, the SLS estimator performs similarly to the SML estima-

[16] Even with the moderate sample size $n = 3000$, the constraint matrix is enormous $(6000 \times 9,000,000)$. Solving the linear program (16) over many Monte Carlo samples is computationally demanding. We employ GUROBI OPTIMIZER 10.0 to solve it efficiently.

TABLE IV

FINITE SAMPLE PERFORMANCES OF THE ESTIMATORS (GUMBEL DGP)

		Gamma errors				Joint Gaussian errors			
		ML*	SML	SLS	SGLS	ML*	SML	SLS	SGLS
α_{CC}	Bias	-0.0608	-0.0021	-0.0015	-0.0020	-0.2032	-0.0023	0.0008	0.0112
	RMSE	0.3780	0.0538	0.0535	0.0538	0.2971	0.0914	0.0925	0.0809
α_{MM}	Bias	-0.0550	0.0019	0.0024	0.0020	-0.3339	0.0018	0.0034	-0.0123
	RMSE	0.4209	0.0512	0.0527	0.0513	1.6626	0.0886	0.0898	0.0827
β_C	Bias	0.0055	0.0003	0.0003	0.0003	-0.0015	-0.0054	-0.0054	-0.0024
	RMSE	0.1138	0.0406	0.0405	0.0406	0.1013	0.0762	0.0757	0.0760
β_M	Bias	-0.0056	0.0011	0.0011	0.0011	-0.0327	-0.0058	-0.0056	-0.0032
	RMSE	0.1033	0.0398	0.0398	0.0399	0.1137	0.0735	0.0728	0.0677

tor. The SGLS outperforms the other estimators as it takes into account the correlations between measurement errors, resulting in more efficient estimation.

Lastly, we consider a DGP in which Lindenlaub (2017)'s model is more severely misspecified. Specifically, we draw x and y from finite Gaussian mixture distributions. Each mixture distribution has two Gaussian components with one-half weight for each. For both x and y, the Gaussian components, K_1 and K_2, are specified as follows.

$$K_1 \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right), \quad K_2 \sim N\left(\begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}1 & -\rho\\ -\rho & 1\end{pmatrix}\right).$$

We set ρ equal to 0.4 for x and 0.5 for y. The equilibrium assignments and wages are solved via linear programming as before. We also generate the measurement errors from a joint Gaussian mixture distribution which has two components:

$$M_1 \sim N\left(\begin{pmatrix}1\\1\\1\end{pmatrix}, \begin{pmatrix}1 & 0.7 & 0.7\\ 0.7 & 1 & 0.3\\ 0.7 & 0.3 & 1\end{pmatrix}\right), \quad M_2 \sim N\left(\begin{pmatrix}-3\\-3\\-3\end{pmatrix}, \begin{pmatrix}1 & 0.7 & 0.7\\ 0.7 & 1 & 0.3\\ 0.7 & 0.3 & 1\end{pmatrix}\right).$$

In this case, both (x, y) and measurement errors have bi-modal distributions that are far from a normal distribution.

TABLE V

FINITE SAMPLE PERFORMANCES OF THE ESTIMATORS (GAUSSIAN MIXTURE DGP)

		ML*	SML	SLS	SGLS
α_{CC}	Bias	0.2896	-0.0014	-0.0016	-0.0017
	RMSE	0.3653	0.0429	0.0425	0.0385
α_{MM}	Bias	0.2446	-0.0017	0.0006	-0.0027
	RMSE	0.3584	0.0454	0.0431	0.0337
β_C	Bias	0.7123	-0.0007	-0.0013	0.0008
	RMSE	0.7204	0.0428	0.0426	0.0364
β_M	Bias	-0.1288	-0.0001	0.0002	-0.0011
	RMSE	0.1592	0.0421	0.0419	0.0305

As the margins of x and y are not standard normal, the ML estimator is not directly applicable. Therefore, we use the inverse transform method to convert x and y to standard normal variables for the ML* estimator. Our sieve estimators can be applied without this transformation so we use untransformed data for the sieve-based estimators. The estimates of technology parameters are reported in Table V. Not surprisingly, the ML* estimator delivers parameter estimates that are very different from the true values. On top of misspecification, the transformation procedure introduces additional bias as the actual assignments are determined on the original data. On the contrary, the sieve estimators still perform extremely well in this case. Estimators relying on fewer assumptions deliver more accurate estimates. The SML estimator produces the least precise estimates among the sieve estimators. The SGLS estimator incorporates the correlation structure among measurement errors so it possesses substantial efficiency gains compared to the SLS estimator.

Our simulation exercises provide evidence that the Gaussian model can be misleading when the model is misspecified. Even with moderate misspecification, the ML estimator does not produce reliable estimates. Furthermore, Gaussian transformation is necessary if

the marginal distributions of skill supply and requirements are not standard normal. This transformation may not precisely recover the underlying assignment mechanism between workers and jobs even if the model is correctly specified after transformation. On the contrary, our sieve estimators do not suffer from these problems. Therefore, our estimators can be more generally applicable regardless of underlying distributions of x, y, and measurement errors with no need for transformation. They also show excellent finite sample performances under correct specifications.

7. EMPIRICAL APPLICATION TO U.S. WORKER-JOB MATCHING

We revisit the dataset constructed by Lindenlaub (2017) to learn how production technology in the US has evolved. We estimate the production technology parameters in the model using the dataset and sieve estimators. The National Longitudinal Survey of Youth (NLSY) data and U.S. Department of Labor Occupational Characteristics Database (O*NET) are used to construct workers' cognitive and manual skills as well as the occupational skill requirements of firms. To assess the effect of technological changes on wage inequality, we compare estimation results based on two cohorts: the first cohort commencing in 1979 (referred to as NLSY79) and the second commencing in 1997 (referred to as NLSY97). Following Lindenlaub's main specification, we focus on employed workers aged 27 to 29 during the years 1990-91 and 2009-10, sourced from the NLSY79 and NLSY97 cohorts, respectively.[17] The wage, w, is defined as the CPI-adjusted hourly rate.

Firms' skill demands, (y_C, y_M), is constructed from the O*NET, which contains information on skill requirements for each occupation. Sanders (2014) classifies occupational skill requirements into cognitive and manual categories and constructs two task scores for over 400 occupations. These scores are employed to obtain skill demands for individuals' currently matched jobs.[18] To construct the skill supply bundle (x_C, x_M), survey responses in the NLSY on education and training are used. Given college education, apprenticeships,

[17] The dataset excludes military samples and oversamples of special demographic/racial groups to give primary focus on the core sample of the NLSY.

[18] For instance, the occupation 'dancer' has a normalized cognitive score (y_C) of 0.34 and a normalized manual score (y_M) of 1, indicating the job is highly manual. On the contrary, the highly cognitive job 'physicist' has a skill

TABLE VI

SUMMARY STATISTICS OF WORKER SKILLS (x) AND FIRMS' SKILL DEMANDS (y)

	1990/91 ($n = 2984$)				2009/10 ($n = 4495$)			
	x_C	x_M	y_C	y_M	x_C	x_M	y_C	y_M
Mean	0.3596	-0.2912	0.0135	-0.1189	0.5667	-0.6601	0.0468	-0.2509
SD	0.7423	0.9923	0.8490	1.0240	0.7556	0.8358	0.9280	0.9656
Min	-2.0595	-1.7004	-2.0622	-1.6949	-2.3019	-1.8116	-2.5200	-1.6597
Max	2.1649	2.1855	2.0925	2.1895	1.9160	2.1838	3.0504	2.1351

government training degrees, and occupational training history, workers are qualified for specific occupations. Lindenlaub (2017) matches individuals to their qualified occupations and obtains the value of (x_C, x_M) using the normalized skill requirements (y_C, y_M) of the matched jobs.[19] It is important to note that the workers' skills are independent of their current occupation because the skill bundles are constructed using qualified occupations.

Table VI presents summary statistics of workers' skills and firms' skill demands. In 1990/91, workers had higher cognitive skills on average than manual skills, and firms also required more cognitive skills than manual skills. Two decades later, workers had increased cognitive skills and decreased manual skills on average compared to 1990/91, with firms also showing a similar trend. The skill correlation (ρ_x) shifted from -0.40 to -0.52, indicating increased worker specialization. In contrast, the productivity correlation (ρ_y) remained stable at -0.49. Initially, jobs were more specialized than workers, but skill supply caught up, resulting in slightly greater worker specialization in 2009/10.

Lindenlaub (2017) transforms each element of x and y into a standard Gaussian variable, and their dependence is modeled using the Gaussian copula. However, while their margins are standard normally distributed, the transformed variables are not guaranteed to be joint

demand bundle of $(y_C = 1, y_M = 0.11)$. We use the normalized task scores for illustration purposes following Lindenlaub (2017). The original supports of worker skills and firms' skill demands are provided in Table VI.

[19]For example, a worker who studied economics at a university is qualified for the 'economist' job. Then the worker possesses a normalized skill bundle of $(x_C = 0.615, x_M = 0.034)$, that is required to be an economist.

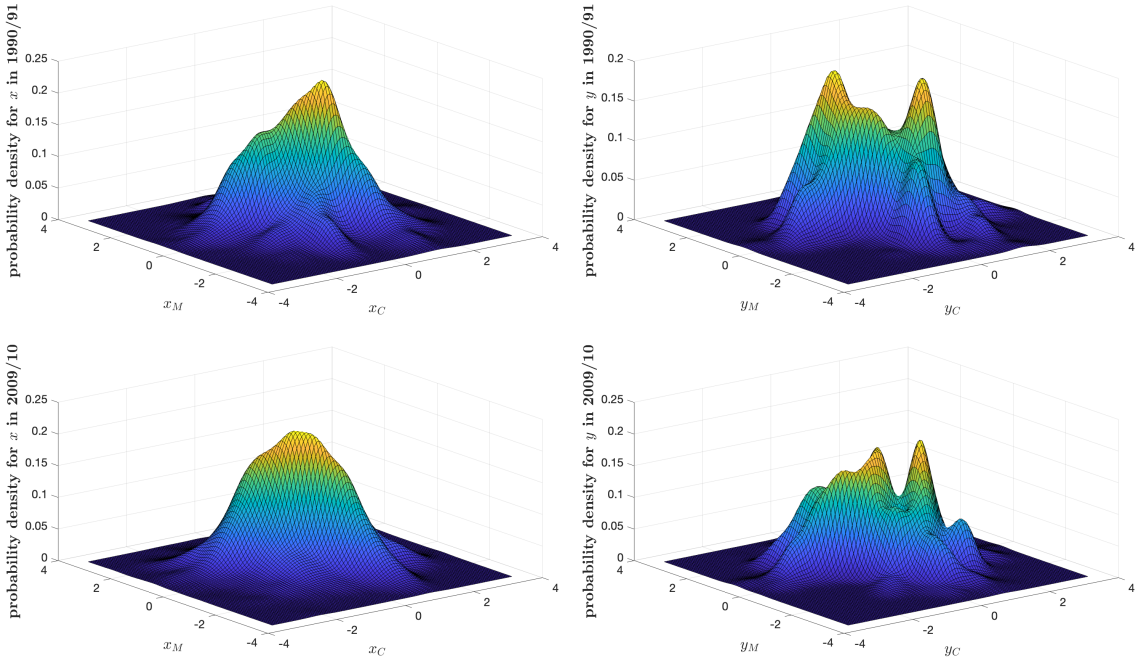

FIGURE 3.—Joint densities of $\tilde{x}$ and $\tilde{y}$ (transformed data)

normally distributed. Let $\tilde{x}$ and $\tilde{y}$ be Gaussian-transformed x and y respectively. As shown in Figure 3, the joint distributions of $\tilde{x}$ and $\tilde{y}$ are not normal. Especially, the joint density of $\tilde{y}$ is multi-modal in both periods.

We employ Mardia's test (Mardia, 1970) to formally test the joint normality of $\tilde{x}$ and $\tilde{y}$. We first create a $n \times n$ matrix for $\tilde{x}$:

$$C = (c_{ij}) = x^* S^{-1} (x^*)',$$

where the i-th row of x^* is $x_i^* = \tilde{x}_i - \sum_{i=1}^n \tilde{x}_i/n$, and define multivariate measures of skewness and kurtosis as follows:

$$b_1 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n c_{ij}^3, \quad b_2 = \frac{1}{n} \sum_{i=1}^n c_{ii}^2.$$

Under bivariate normality, the limiting distribution of $\frac{nb_1}{6}$ is a chi-square distribution with $d(d+1)(d+2)/6$ degrees of freedom and the limiting distribution of $\frac{\sqrt{n}(b_2 - d(d+2))}{\sqrt{8d(d+2)}}$ is the standard normal distribution where d is the dimensionality of $\tilde{x}$. We conduct the same

procedure for $\tilde{y}$. Table VII shows the test statistics. The test strongly rejects the bivariate normality of $\tilde{x}$ and $\tilde{y}$ in both periods. The normality assumption imposed in the Lindenlaub model is not satisfied even after transforming x and y. Therefore, our semiparametric approach is more appropriate in this case.

TABLE VII

MARDIA'S MULTIVARIATE NORMALITY TEST STATISTICS (P-VALUES IN PARENTHESES)

	1990/91 ($n = 2984$)		2009/10 ($n = 4495$)	
	$\tilde{x}$	$\tilde{y}$	$\tilde{x}$	$\tilde{y}$
Skewness	4.58 (0.333)	100.09 (0.000)	16.34 (0.003)	145.14 (0.000)
Kurtosis	4.44 (0.000)	0.29 (0.774)	14.42 (0.000)	1.98 (0.048)

We estimate the production technology in each period separately. First, we modify Lindenlaub (2017)'s MLE procedure to accommodate the cases where the true skill requirements $\tilde{y}^*$ have variances not equal to 1. As the inverse transform method converts the measurement error contaminated $\tilde{y}$, not $\tilde{y}^*$, it is essentially the case that $var(\tilde{y}^*) \neq 1$. We also allow the Gaussian measurement errors to be correlated with each other. We refer to this generalized and corrected ML procedure as 'ML*'. Then, we relax the normality of X and Y and estimate the technology parameters using SML. Next, we further relax the normality of the measurement errors and estimate the parameters using SLS. Finally, we allow measurement errors to be correlated with each other and estimate the parameters through SGLS. Sieve estimation is conducted using Bernstein polynomial basis functions of degree 3, which performs the best in terms of information criteria and model fit. We compare our semiparametric estimates of the technology parameters to Lindenlaub's original estimates and the 'ML*' estimates.

The estimation results are provided in Table VIII. All the models clearly show a huge shift in the relative importance between manual and cognitive tasks over the two decades. Our results are qualitatively consistent with Lindenlaub's but quantitatively very different. In 1990/91, the estimated complementarity in manual tasks was much larger than in cogni-

TABLE VIII

ESTIMATES OF PRODUCTION TECHNOLOGY PARAMETERS ON TRANSFORMED DATA

	1990/91					2009/10				
	ML	ML*	SML	SLS	SGLS	ML	ML*	SML	SLS	SGLS
α_{CC}	0.203	0.765	0.454	0.000	0.000	0.739	1.119	2.048	2.293	2.290
	(0.342)	(0.574)	(0.009)	(0.000)	(0.000)	(0.198)	(0.408)	(0.036)	(0.256)	(0.265)
α_{MM}	0.479	1.270	1.422	1.084	0.856	0.055	0.486	0.237	1.033	0.291
	(0.175)	(0.148)	(0.031)	(0.113)	(0.098)	(0.154)	(0.633)	(0.006)	(0.215)	(0.059)
β_C	1.686	1.711	1.692	1.719	1.585	2.203	2.208	2.115	2.063	2.198
	(0.143)	(0.589)	(0.068)	(0.434)	(0.416)	(0.152)	(0.540)	(0.068)	(0.589)	(0.534)
β_M	-0.421	-0.392	-0.374	-0.382	-0.388	0.210	0.243	0.198	0.180	0.327
	(0.141)	(0.406)	(0.068)	(0.329)	(0.309)	(0.152)	(0.731)	(0.076)	(0.545)	(0.532)

Standard errors in the parentheses. ML indicates the original estimates in Lindenlaub (2017).

tive tasks. The Gaussian model (ML*) indicates that the complementarity in manual tasks is roughly 1.7 times as large as that of cognitive tasks. Dispensing with the normality of skill demand and supply, the ratio becomes larger than 3. When we further generalize the model by removing the Gaussian assumption on measurement errors, the complementarity in cognitive tasks shrinks close to 0 (but significantly larger than 0), whereas that of manual tasks is still close to 1. The estimates of linear productivity coefficients β_C and β_M are similar across all the specifications as shown in simulations.

This pattern becomes the opposite in the 20 years. All the complementarity estimates for 2009/10 indicate a substantial increase in the complementarity between cognitive worker and job attributes, whereas the complementarity in manual tasks heavily decreased. In the Gaussian model (ML*), the ratio of estimated complementarities in manual tasks to cognitive tasks ($\frac{\alpha_{MM}}{\alpha_{CC}}$) is around 0.43, which is similar to the estimated value by SLS. However, SML and SGLS deliver much smaller values close to 0.1. The linear coefficients are similar across specifications. These patterns imply substantial changes in the relative complementarities across tasks because of technological advances. Lindenlaub describes this phenomenon as "task-biased technological change in favor of cognitive tasks". The cogni-

tive dimension became much more important in labor market sorting. Our semiparametric models suggest that the "task-biased technological change" favoring cognitive tasks in the last two decades may have been much larger than previously found. The increases in β_C and β_M indicate that both cognitive and manual skill productivity have risen. However, the estimated manual skill productivity in both periods is insignificant in most specifications.

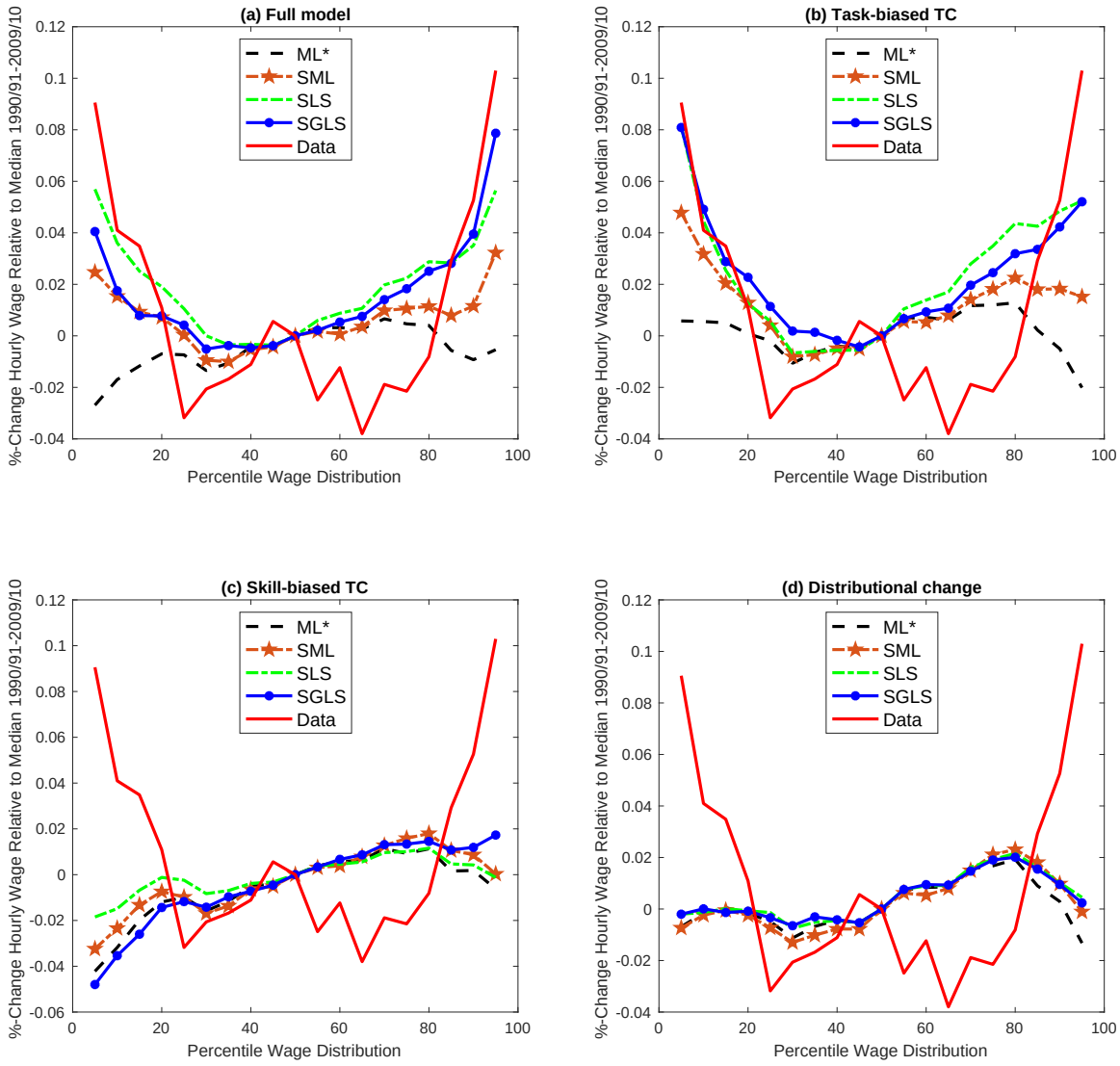

FIGURE 4.—Actual and model predicted wage polarization (transformed data)

Therefore, following Lindenlaub's description, we can conclude that the U.S. economy has also experienced "skill-biased technological change" in favor of cognitive skills.

We now investigate the effect of estimated technological changes on wage inequality. Wage inequality in the U.S. labor market until the late 2000s is well characterized by *wage polarization* that is defined as stronger wage growth in the bottom and upper tails relative to the median. The red solid line in Figure 4 plots how wages changed relative to the median wage between 1990/91 and 2009/10 by wage percentile, implying that the U.S. labor market experienced a spike in the upper-tail wage inequality, while the lower-tail inequality declined. This phenomenon is surprisingly well-predicted in our models as shown in Figure 4 (a). All the semiparametric models predict substantial wage polarization once the estimated parameter values are fed in. On the contrary, the Gaussian model fails to account for wage polarization in both tails. We can also observe that the model fit improves as the model becomes more flexible.

To further explore why the matching model requires greater flexibility to account for wage polarization, we compare the curvature of estimated wage functions across different models in Figures 5–6. In 1990/91, all models produce almost linear wage functions, indicating a relatively uniform relationship between wages and cognitive skills. However, in 2009/10, our semiparametric models predict a significantly steeper curvature, particularly at high cognitive skill levels, whereas the Gaussian model still generates a more linear wage function. The Gaussian model constrains the wage function to a quadratic form of standard normal variables, limiting its shape to a low-degree polynomial. In contrast, our models do not impose such constraints, allowing for greater flexibility in curvature that better fits the data. Notably, our most flexible model predicts a sharply increasing slope in the wage function for 2009/10, which is relatively flat at low cognitive skill levels and very steep at high skill levels, generating substantial wage polarization.

To understand the driving forces behind wage polarization, we isolate the effects of technological and distributional changes in Figure 4. We only keep task-biased technological change (shutting down changes in linear productivity coefficients) in panel (b), skill-biased technological change (shutting down changes in complementarity parameters) in (c), and

distributional change (shutting down both changes in linear productivity and complementarity parameters) in (d). We find that task-biased technological change explains wage polarization remarkably well. Especially, all three semiparametric models exhibit an excellent fit in the lower tail in panel (b), while the Gaussian model shows only a slight decline in lower tail inequality. In contrast, skill-biased technological change exacerbates wage inequality in the lower tail as shown in panel (c). The distributional change has a negligible impact on wage inequality. In summary, despite their parsimony, our matching models ef-

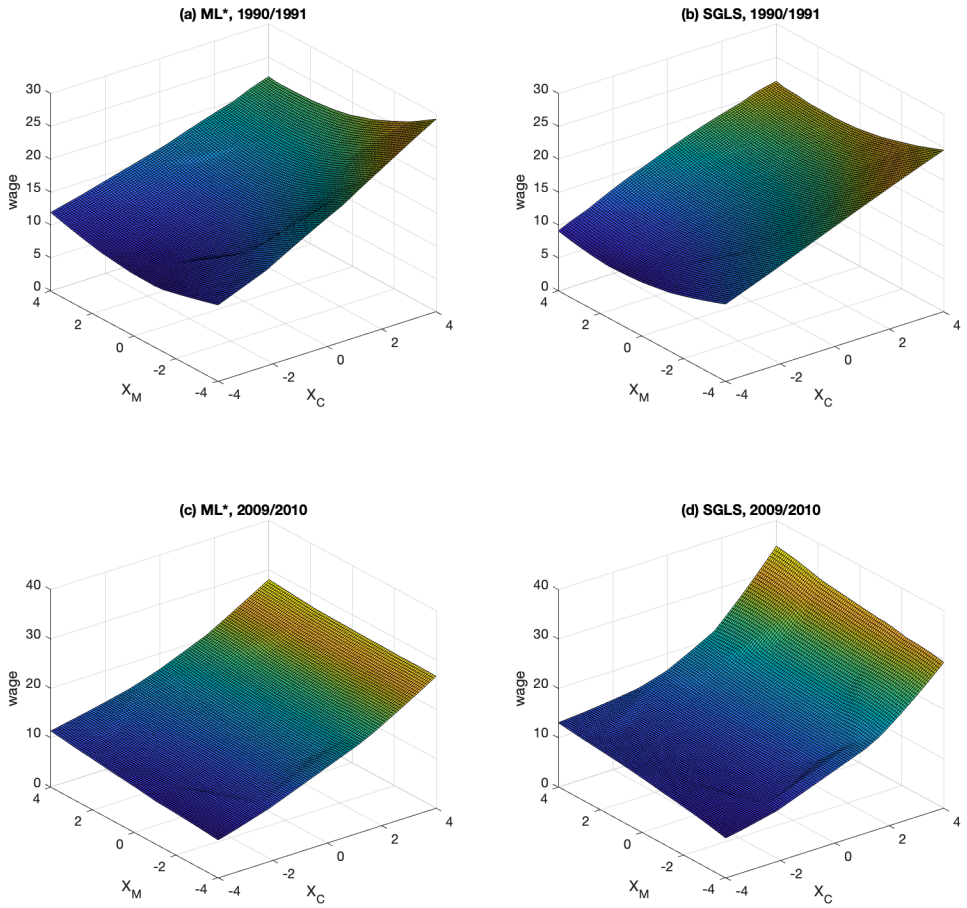

FIGURE 5.—Estimated wage functions (transformed data)

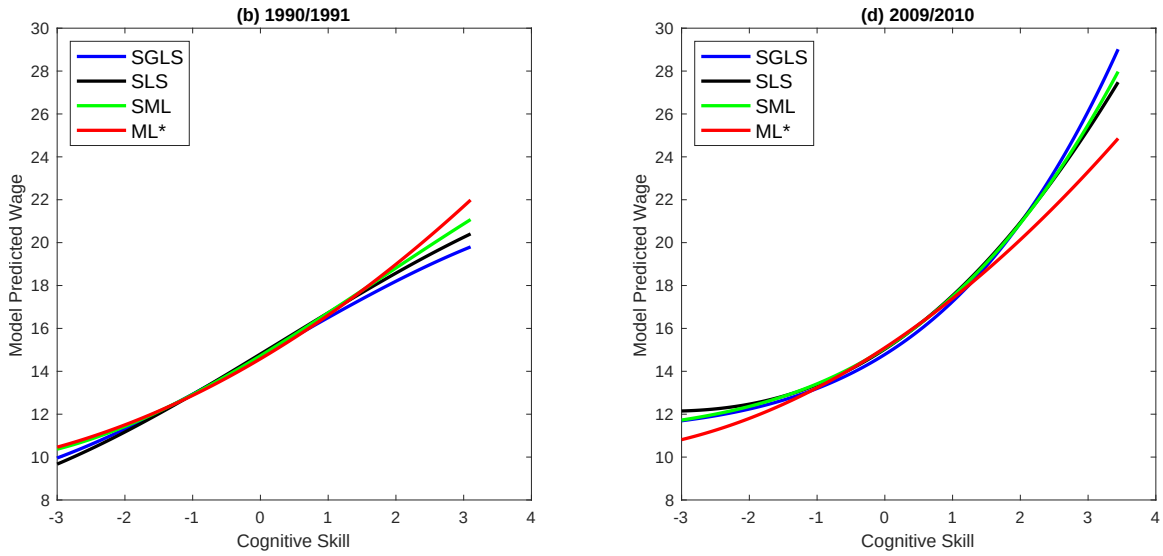

FIGURE 6.—Predicted wage with respect to cognitive skill (transformed data)

fectively account for wage inequality's evolution over the past 20 years in the U.S., with task-biased technological change being the primary driver of the observed pattern.

Lastly, we estimate our semiparametric models on the original data. Unlike the Gaussian model, our models can be applied directly to the data without any transformation. As we described in Figure 1, the matching occurs based on original distributions. Hence transforming marginal distributions to standard normal can lead to a different solution from $T(x)$. The marginal distributions of workers' skill supply x and firms' skill demand y in Figure 7 are skewed and multi-modal, quite different from standard normal. Therefore, it is crucial to investigate the robustness of the Gaussian model using the original data. We report the estimated parameters in Table IX, using Bernstein polynomial basis functions of degree 4 to accommodate the less well-behaved original data. The estimated parameters reveal similar patterns across our models. Notably, the complementarity in cognitive tasks increased significantly from near 0 in 1990/91 to around 6 in 2009/10, while the complementarity in manual tasks decreased from near 2 to almost 0. Additionally, linear skill productivity improved, with a larger increase in cognitive skill productivity. These findings confirm that the U.S. economy experienced substantial task-biased and skill-biased technological changes favoring cognitive skills over the two decades.

The estimated models on the original data effectively capture the patterns of wage polarization, particularly in the upper tail as in Figure 8. While the models slightly under-predict wage polarization in the lower tail, they confirm that task-biased technological change was the primary driver of wage polarization. In the absence of skill-biased technological change, the model shows a significant relative wage increase in the lower tail. However, skill-biased technological change had a negative impact on wage inequality, exacerbating lower tail inequality. Distributional change improved upper tail inequality but worsened lower tail inequality.

In summary, the estimated semiparametric models on the original data exhibit similar patterns to those on the Gaussian transformed data, with task- and skill-biased technological changes being more pronounced. Our models demonstrate a remarkable fit for wage polarization, highlighting the substantial changes in production technology in the U.S. over the past two decades. This exercise showcases the versatility and effectiveness of our semi-

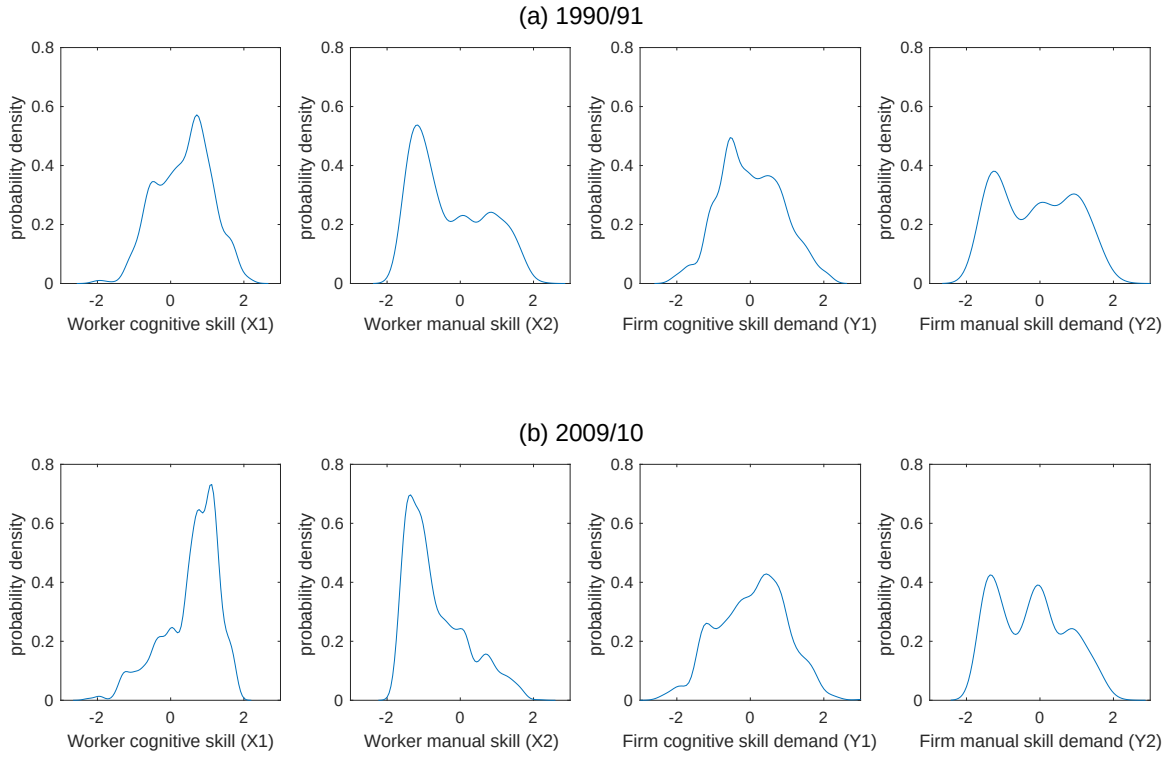

FIGURE 7.—Marginal distributions of skill supply and demand

TABLE IX

ESTIMATES OF PRODUCTION TECHNOLOGY PARAMETERS ON ORIGINAL DATA

	1990/91			2009/10		
	SML	SLS	SGLS	SML	SLS	SGLS
α_{CC}	0.000	0.000	0.000	6.032	5.784	6.471
	(0.000)	(0.000)	(0.000)	(0.110)	(0.184)	(0.196)
α_{MM}	1.947	2.292	2.234	0.000	1.020	0.003
	(0.060)	(0.081)	(0.079)	(0.000)	(0.024)	(0.000)
β_C	2.238	2.246	2.108	3.395	3.299	3.707
	(0.083)	(0.225)	(0.220)	(0.101)	(0.396)	(0.292)
β_M	-0.341	-0.441	-0.317	0.254	0.384	0.440
	(0.073)	(0.211)	(0.213)	(0.097)	(0.333)	(0.239)

Standard errors in the parentheses

parametric models and sieve-based estimators, which can accommodate any underlying joint distributions of skill supply and demand without requiring data transformation or distributional assumptions. Moreover, our approach builds upon standard sieve-based estimators, which have been proven to achieve semiparametric efficiency and are easy to implement in practice.

8. CONCLUSION

Theoretical matching models often face empirical challenges due to discrepancies between model assumptions and real-world data. Lindenlaub (2017) presents a tractable theoretical model suitable for comparative statics and qualitative analysis of multidimensional matching. However, its empirical application is limited by restrictive distributional assumptions on observed characteristics and measurement errors. We generalize this model by relaxing these key distributional restrictions, enabling our models to accommodate datasets with matched pairs, regardless of the underlying characteristic and error distributions. Our simulation results demonstrate the accuracy of our semi-nonparametric estimators across various data generating processes. Moreover, our flexible models generate significant wage polarization, aligning with U.S. data patterns, whereas the parametric Gaussian model falls

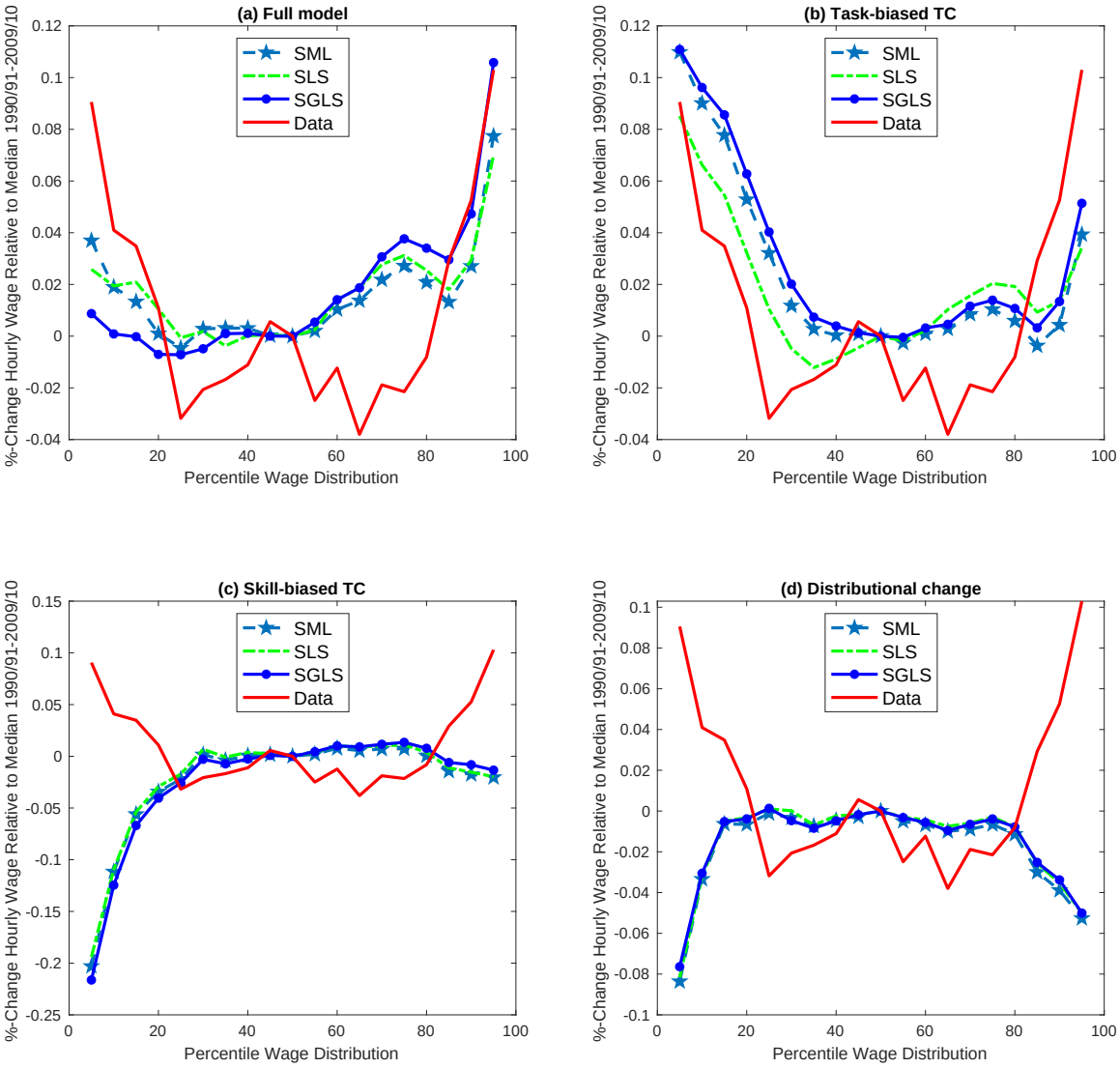

FIGURE 8.—Actual and model predicted wage polarization (original data)

short. Our estimated models indicate that task-biased technological progress, favoring cognitive abilities over manual skills, is the primary driver of wage polarization.

Our study opens up promising research avenues. First, incorporating additional dimensions like interpersonal and digital skills into our model holds great potential. By employing advanced techniques like artificial neural networks (Chen et al., 2023) to approximate equilibrium functions in high-dimensional spaces, one could enhance the model's accuracy in elucidating intricate matching patterns and their impact on wage inequality. Sec-

ond, addressing measurement errors in assessing worker skills is crucial. Overcoming this challenge requires innovative econometric approaches that can accurately estimate models amidst multidimensional measurement errors. Finally, our framework's application extends beyond the worker-job matching problem, offering insights into matching problems in diverse contexts like the marriage market.

While our approach offers valuable insights, it is essential to acknowledge its limitations. Frictionless matching models, grounded in optimal transport, assume efficient equilibrium, eliminating concerns of unemployment or skill mismatch. Introducing randomness in assignment through factors like search frictions and unobserved heterogeneity poses a fruitful challenge, especially when extending existing one-dimensional theories e.g., in Eeckhout and Kircher (2011), to multidimensional settings.

REFERENCES

AI, CHUNRONG AND XIAOHONG CHEN (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71 (6), 1795–1843. [3, 5, 19, 20, 21, 4]

——— (2007): "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables," *Journal of Econometrics*, 141 (1), 5–43. [5]

BECKER, GARY S. (1973): "A Theory of Marriage: Part I," *Journal of Political Economy*, 81 (4), 813–846. [2, 10]

BECKER, GARY S (1991): *A treatise on the family: Enlarged edition*, Harvard university press. [2]

BLUNDELL, RICHARD, XIAOHONG CHEN, AND DENNIS KRISTENSEN (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75 (6), 1613–1669. [5]

BOJILOV, RAICHO AND ALFRED GALICHON (2016): "Matching in closed-form: equilibrium, identification, and comparative statics," *Economic Theory*, 61, 587–609. [5]

BROWNING, MARTIN, PIERRE-ANDRÉ CHIAPPORI, AND YORAM WEISS (2014): *Economics of the Family*, Cambridge University Press. [6]

CAFFARELLI, LUIS A. (1992a): "Boundary regularity of maps with convex potentials," *Communications on Pure and Applied Mathematics*, 45 (9), 1141–1151. [6, 21]

CAFFARELLI, LUIS A (1992b): "The regularity of mappings with a convex potential," *Journal of the American Mathematical Society*, 5 (1), 99–104. [6, 21]

——— (1996): "Boundary regularity of maps with convex potentials–II," *Annals of mathematics*, 144 (3), 453–496. [6, 21]

CARLIER, GUILLAUME (2003): "Duality and existence for a class of mass transportation problems and economic applications," *Advances in mathematical economics*, 1–21. [9]

CHEN, JIAFENG, XIAOHONG CHEN, AND ELIE TAMER (2023): "Efficient estimation of average derivatives in NPIV models: Simulation comparisons of neural network estimators," *Journal of Econometrics*. [41]

CHEN, XIAOHONG (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," Elsevier, vol. 6 of *Handbook of Econometrics*, 5549–5632. [4, 5, 17, 18, 19, 22, 1, 2, 3]

CHEN, XIAOHONG AND DEMIAN POUZO (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152 (1), 46–60. [5]

CHEN, XIAOHONG AND XIAOTONG SHEN (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66 (2), 289–314. [19, 2]

CHIAPPORI, PIERRE-ANDRÉ (2017): *Matching with Transfers: The Economics of Love and Marriage*, Princeton University Press. [6]

CHIAPPORI, PIERRE-ANDRÉ, ROBERT J MCCANN, AND LARS P NESHEIM (2010): "Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness," *Economic Theory*, 42, 317–354. [3, 8]

CHIAPPORI, PIERRE-ANDRÉ AND SONIA OREFFICE (2008): "Birth control and female empowerment: An equilibrium analysis," *Journal of Political Economy*, 116 (1), 113–140. [2]

CHIAPPORI, PIERRE-ANDRÉ, SONIA OREFFICE, AND CLIMENT QUINTANA-DOMEQUE (2012): "Fatter attraction: anthropometric and socioeconomic matching on the marriage market," *Journal of Political Economy*, 120 (4), 659–695. [2]

CHIAPPORI, PIERRE-ANDRÉ AND PHILIP J. RENY (2016): "Matching to share risk," *Theoretical Economics*, 11 (1), 227–251. [2]

CHIONG, KHAI XIANG, ALFRED GALICHON, AND MATT SHUM (2016): "Duality in dynamic discrete-choice models," *Quantitative Economics*, 7 (1), 83–115. [3]

CHOO, EUGENE AND ALOYSIUS SIOW (2006): "Who marries whom and why," *Journal of political Economy*, 114 (1), 175–201. [2, 5]

COMPIANI, GIOVANNI (2022): "Market counterfactuals and the specification of multiproduct demand: A nonparametric approach," *Quantitative Economics*, 13 (2), 545–591. [19]

CORDERO-ERAUSQUIN, DARIO AND ALESSIO FIGALLI (2019): "Regularity of monotone transport maps between unbounded domains," *Discrete and Continuous Dynamical Systems*, 39 (12), 7101–7112. [6, 21]

DE PHILIPPIS, GUIDO AND ALESSIO FIGALLI (2014): "The Monge-Ampère equation and its link to optimal transportation," *Bull. Amer. Math. Soc. (N.S.)*, 51 (4), 527–580. [3]

DUPUY, ARNAUD AND ALFRED GALICHON (2014): "Personality Traits and the Marriage Market," *Journal of Political Economy*, 122 (6), 1271–1319. [5]

EECKHOUT, JAN AND PHILIPP KIRCHER (2011): "Identifying Sorting-In Theory," *The Review of Economic Studies*, 78 (3), 872–906. [42]

EKELAND, IVAR (2010): "Notes on optimal transportation," *Economic Theory*, 437–459. [3]

FLOATER, MICHAEL S (1994): "A weak condition for the convexity of tensor-product Bézier and B-spline surfaces," *Advances in Computational Mathematics*, 2, 67–80. [9]

GALICHON, ALFRED (2017): "A survey of some recent applications of optimal transport methods to econometrics," *The Econometrics Journal*, 20 (2), C1–C11. [3, 7]

GALICHON, ALFRED, SCOTT DUKE KOMINERS, AND SIMON WEBER (2019): "Costly Concessions: An Empirical Framework for Matching with Imperfectly Transferable Utility," *Journal of Political Economy*, 127 (6), 2875–2925. [5]

GALICHON, ALFRED AND BERNARD SALANIÉ (2022): "Cupid's invisible hand: Social surplus and identification in matching models," *The Review of Economic Studies*, 89 (5), 2600–2629. [3, 5]

GROSSBARD-SCHECHTMAN, SHOSHANA (1993): "A Theory of Marriage, Labor and Divorce," . [2]

GUALDANI, CRISTINA AND SHRUTI SINHA (2023): "Partial identification in matching models for the marriage market," *Journal of Political Economy*, 131 (5), 1109–1171. [5]

HITSCH, GÜNTER J, ALI HORTAÇSU, AND DAN ARIELY (2010): "Matching and sorting in online dating," *American Economic Review*, 100 (1), 130–163. [2]

LEGROS, PATRICK AND ANDREW F NEWMAN (2007): "Beauty is a beast, frog is a prince: Assortative matching with nontransferabilities," *Econometrica*, 75 (4), 1073–1102. [2]

LINDENLAUB, ILSE (2017): "Sorting Multidimensional Types: Theory and Application," *The Review of Economic Studies*, 84 (2), 718–789. [2, 3, 4, 5, 10, 12, 13, 14, 16, 18, 22, 23, 28, 30, 31, 33, 34, 40, 9]

LISE, JEREMY AND FABIEN POSTEL-VINAY (2020): "Multidimensional Skills, Sorting, and Human Capital Accumulation," *American Economic Review*, 110 (8), 2328–76. [5, 14]

LORENTZ, GEORGE G (1986): *Bernstein polynomials*, American Mathematical Soc. [18]

MARDIA, K. V. (1970): "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, 57 (3), 519–530. [32]

MONGE, G (1781): "Histoire de l'Académie Royale des Sciences de Paris," *De L'imprimerie Royale*. [8]

NEWEY, WHITNEY K. AND JAMES L. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71 (5), 1565–1578. [5]

OREFFICE, SONIA AND CLIMENT QUINTANA-DOMEQUE (2010): "Anthropometry and socioeconomics among couples: Evidence in the United States," *Economics & Human Biology*, 8 (3), 373–384. [2]

PAPAGEORGIOU, THEODORE (2014): "Learning your comparative advantages," *Review of Economic Studies*, 81 (3), 1263–1295. [2]

PENCAVEL, JOHN (1998): "Assortative mating by schooling and the work behavior of wives and husbands," *The American Economic Review*, 88 (2), 326–329. [2]

PEYRÉ, GABRIEL, MARCO CUTURI, ET AL. (2019): "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, 11 (5-6), 355–607. [9]

QIAN, ZHENCHAO (1998): "Changes in assortative mating: The impact of age and education, 1970–1890," *Demography*, 35 (3), 279–292. [2]

SANDERS, CARL (2014): "Skill accumulation, skill uncertainty, and occupational choice," *Working Paper*. [14, 30]

SHEN, XIAOTONG (1997): "On methods of sieves and penalization," *The Annals of Statistics*, 25 (6), 2555 – 2591. [5]

SILVENTOINEN, KARRI, JAAKKO KAPRIO, EERO LAHELMA, RICHARD J VIKEN, AND RICHARD J ROSE (2003): "Assortative mating by body height and BMI: Finnish twins and their spouses," *American Journal of Human Biology*, 15 (5), 620–627. [2]

VILLANI, CÉDRIC (2003): *Topics in optimal transportation*, vol. 58, American Mathematical Soc. [3, 1]

VILLANI, CÉDRIC (2008): *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften, Springer. [3, 8, 9]

WANG, J. AND S.K. GHOSH (2012): "Shape restricted nonparametric regression with Bernstein polynomials," *Computational Statistics & Data Analysis*, 56 (9), 2729–2741. [8]

WEISS, YORAM AND ROBERT J WILLIS (1997): "Match quality, new information, and marital dissolution," *Journal of labor Economics*, 15 (1, Part 2), S293–S329. [2]

WILLIS, ROBERT J AND SHERWIN ROSEN (1979): "Education and self-selection," *Journal of political Economy*, 87 (5, Part 2), S7–S36. [2]

APPENDIX A: TECHNICAL PROOFS

PROOF OF PROPOSITION 2: In the equilibrium, the firm maximizes its profit so the first-order condition of the firm's maximization problem is satisfied:

$$\nabla w^{*}(x) - b = \nabla_{x} x'\tilde{y}\big|_{\tilde{y}=T(x)}.$$

By Theorem 2.12 in Villani (2003), $\nabla_{x} x'\tilde{y}|_{\tilde{y}=T(x)} = \nabla w^{o*}(x)$ and therefore,

$$\nabla w^{*}(x) = \nabla w^{o*}(x) + b.$$

This implies that $w^{*}(x) = w^{o*}(x) + x'b + c$ where c is the constant of integration. *Q.E.D.*

PROOF OF THEOREM 1: We first note from $\mathbb{E}[\varepsilon_{wi}|x_i] = 0$ that the convex function $w_0^{*}(x) = \mathbb{E}[w_i|x_i = x] = w_0(x) + x'b_0$ is identified. Since $\nabla w_0^{*}(x)$ is also identified and $\mathbb{E}[y_i|x_i = x] = A_0^{-1}(\nabla w_0^{*}(x) - b_0)$, the invertibility of A_0 implies that A_0 and b_0 are identified: We consider $y_1, \ldots, y_d, y_{d+1} \in \mathcal{Y}$ satisfying Assumption 5. Then, there are corresponding $d+1$ distinct points $x_1, \ldots, x_d, x_{d+1} \in \mathcal{X}$ such that $A_0 y_1 = \nabla w_0(x_1), \ldots, A_0 y_d = \nabla w_0(x_d), A_0 y_{d+1}^{*} = \nabla w_0(x_{d+1})$. It follows from the invertibility of A_0 that $\nabla w_0(x_1) - \nabla w_0(x_2) = A_0(y_1 - y_2), \ldots, \nabla w_0(x_d) - \nabla w_0(x_{d+1}) = A_0(y_d - y_{d+1})$ are independent. It follows from $\nabla w_0^{*}(x) = \nabla w_0(x) + b_0$ that $\nabla w_0^{*}(x_1) - \nabla w_0^{*}(x_2), \ldots, \nabla w_0^{*}(x_d) - \nabla w_0^{*}(x_{d+1})$ are also linearly independent. Then, A_0 is identified from

$$\begin{pmatrix} \mathbb{E}[y_i|x_i = x_1] - \mathbb{E}[y_i|x_i = x_2] \\ \vdots \\ \mathbb{E}[y_i|x_i = x_d] - \mathbb{E}[y_i|x_i = x_{d+1}] \end{pmatrix}' = A_0^{-1} \begin{pmatrix} \nabla w_0^{*}(x_1) - \nabla w_0^{*}(x_2) \\ \vdots \\ \nabla w_0^{*}(x_d) - \nabla w_0^{*}(x_{d+1}) \end{pmatrix}',$$

in turn, b_0 and $w_0(x) = w_0^{*}(x) - x'b_0$ are also identified. *Q.E.D.*

PROOF OF PROPOSITION 3: We obtain the result by applying Theorem 3.2 in Chen (2007). We note that

$$\begin{aligned} &\rho'(z;\lambda)\Sigma(x)^{-1}\rho(z;\lambda) - \rho'(z;\lambda_0)\Sigma(x)^{-1}\rho(z;\lambda_0) \\ &= [\rho(z;\lambda) - \rho(z;\lambda_0)]'\Sigma(x)^{-1}[\rho(z;\lambda) - \rho(z;\lambda_0)] - 2\varepsilon'\Sigma(x)^{-1}[\rho(z;\lambda) - \rho(z;\lambda_0)]. \end{aligned}$$

and

$$\rho(z;\lambda)-\rho(z;\lambda_0)=\begin{pmatrix} w(x)-w_0(x)+x'(b-b_0) \\ \nabla_C w_0(x)(\kappa_C-\kappa_{C0})+\kappa_C\nabla_C\{w(x)-w_0(x)\} \\ \nabla_M w_0(x)(\kappa_M-\kappa_{M0})+\kappa_M\nabla_M\{w(x)-w_0(x)\}\end{pmatrix}.$$

Then, it is easy to see from Assumption 7 that

$$\|\lambda-\lambda_0\|^2 \asymp \mathbb{E}\left[\rho'(z_i;\lambda)\Sigma(x_i)^{-1}\rho(z_i;\lambda)-\rho'(z_i;\lambda_0)\Sigma(x_i)^{-1}\rho(z_i;\lambda_0)\right],$$

i.e., there exists a finite $C_1>0$ such that

$$C_1^{-1}\|\lambda-\lambda_0\|^2 \le \mathbb{E}\left[\rho'(z_i;\lambda)\Sigma(x_i)^{-1}\rho(z_i;\lambda)-\rho'(z_i;\lambda_0)\Sigma(x_i)^{-1}\rho(z_i;\lambda_0)\right] \le C_1\|\lambda-\lambda_0\|^2.$$

Condition 3.6 in Chen (2007) is assumed with Assumption 6. Now we check Conditions 3.7 and 3.8 in Chen (2007). Again, Assumption 7 implies that there exists a $C_2>0$ such that

$$\begin{aligned}&\mathbb{E}\left[\left(\rho'(z_i;\lambda_0)\Sigma(x_i)^{-1}\rho(z_i;\lambda_0)-\rho'(z_i;\lambda)\Sigma(x_i)^{-1}\rho(z_i;\lambda)\right)^2\right] \\ &\le C_2\mathbb{E}\left[|\rho(z_i;\lambda_0)-\rho(z_i;\lambda)|_e^4\right].\end{aligned}$$

By Lemma 2 in Chen and Shen (1998), we have $\|w-w_0\|_\infty \le c\|w-w_0\|_2^{m/(m+1)}$ and $\|\nabla_C\{w-w_0\}\|_\infty, \|\nabla_M\{w-w_0\}\|_\infty \le c\|w-w_0\|_2^{(m-1)/m}$ for some finite $c>0$, where $\|w\|_2^2=\mathbb{E}\left[w^2(x_i)\right]$. Since $\|\cdot\|_2 \asymp \|\cdot\|$,

$$\mathbb{E}\left[\left(\rho'(z_i;\lambda_0)\Sigma(x_i)^{-1}\rho(z_i;\lambda_0)-\rho'(z_i;\lambda)\Sigma(x_i)^{-1}\rho(z_i;\lambda)\right)^2\right] \le C_3\|\lambda-\lambda_0\|^{2[1+(m-1)/m]}.$$

So Condition 3.7 is satisfied for all $\varepsilon \le 1$. On the other hand,

$$\begin{aligned}&\left|\rho'(z_i;\lambda_0)\Sigma(x_i)^{-1}\rho(z_i;\lambda_0)-\rho'(z_i;\lambda)\Sigma(x_i)^{-1}\rho(z_i;\lambda)\right| \\ &\le \|\lambda-\lambda_0\|_\infty|\Sigma(x_i)|^{-1}\left(2|\varepsilon|_e+\|\lambda\|_\infty+\|\lambda_0\|_\infty\right),\end{aligned}$$

almost surely. Using Lemma 2 in Chen and Shen (1998) again, Condition 3.8 is satisfied.

To apply Theorem 3.2 in Chen (2007), It remains to compute the deterministic approximation error rate $\inf_{\lambda\in\Theta\times\mathcal{W}_n}\|\lambda-\lambda_0\|$ and the metric entropy with bracketing. By the same proof as that for Proposition 3.3 in Chen (2007), they are also computed, and then the result follows. *Q.E.D.*

PROOF OF THEOREM 2: We obtain the limiting distribution of $\hat{\theta}_n$ by verifying that Assumptions 4.1 and 4.2 of Proposition 4.4 in Chen (2007) are satisfied. It is easy to see that Assumptions 4.2.(ii) and (iv) are satisfied with the expression for ρ. Assumptions 4.1.(i) and 4.2.(i) are our assumption 11 and 7. Assumption 4.1.(ii) is implied by our assumption 7 and 10. Assumption 4.1.(iii) is implied by our Proposition 3 and Assumption 10: there is $\pi_n u^* \in \mathcal{W}_n$ such that $\|\pi_n u^* - u^*\| \times \|\hat{\lambda}_n - \lambda_0\| = o_p\left(n^{-1/2}\right)$. Let $u^*_\theta = \left(u^*_{\theta 1}, u^*_{\theta 2}, u^*_{\theta 3}, u^*_{\theta 4}\right)' = \left(\mathbb{E}\left[D_{v^*}(x_i)' \Sigma(x_i)^{-1} D_{v^*}(x_i)\right]\right)^{-1} \eta$, $u^*_w = -v^* u^*_\theta$ and $u^* = \left(u^*_\theta, u^*_w\right)$, where $\eta \in \mathbb{R}^4$ is an arbitrary unit vector. It remains to show that Assumption 4.2.(iii) (Conditions 4.2' and 4.3') in Chen (2007) are satisfied with

$$\frac{\partial \ell\left(z;\bar{\lambda}\right)}{\partial \lambda}[\pi_n u^*] = \begin{pmatrix} \bar{w}(x) + x'\bar{b} - w \\ \bar{\kappa}_C \nabla_C \bar{w}(x) - y_C \\ \bar{\kappa}_M \nabla_M \bar{w}(x) - y_M \end{pmatrix}' \Sigma(x)^{-1} \begin{pmatrix} \pi_n u^*_w + u^*_{\theta 3} x_C + u^*_{\theta 4} x_M \\ u^*_{\theta 1} \nabla_C \bar{w}(x) + \bar{\kappa}_C \nabla_C (\pi_n u^*_w(x)) \\ u^*_{\theta 2} \nabla_M \bar{w}(x) + \bar{\kappa}_M \nabla_M (\pi_n u^*_w(x)) \end{pmatrix}$$

$$= \begin{pmatrix} \bar{w}(x) - w_0(x) + x'\left(\bar{b} - b_0\right) - \varepsilon_w \\ \bar{\kappa}_C \nabla_C \bar{w}(x) - \kappa_{C0} \nabla_C w_0(x) - \varepsilon_C \\ \bar{\kappa}_M \nabla_M \bar{w}(x) - \kappa_{M0} \nabla_M w_0(x) - \varepsilon_M \end{pmatrix}' \Sigma(x)^{-1} \begin{pmatrix} \pi_n u^*_w + u^*_{\theta 3} x_C + u^*_{\theta 4} x_M \\ u^*_{\theta 1} \nabla_C \bar{w}(x) + \bar{\kappa}_C \nabla_C (\pi_n u^*_w(x)) \\ u^*_{\theta 2} \nabla_M \bar{w}(x) + \bar{\kappa}_M \nabla_M (\pi_n u^*_w(x)) \end{pmatrix}$$

for all $\bar{\lambda} \in \Theta \times \mathcal{W}_n$ with $\|\bar{\lambda} - \lambda_0\| = o(1)$. Since

$$\begin{pmatrix} \hat{w}(x) - w_0(x) + x'\left(\hat{b} - b_0\right) \\ \hat{\kappa}_C \nabla_C \hat{w}(x) - \kappa_{C0} \nabla_C w_0(x) \\ \hat{\kappa}_M \nabla_M \hat{w}(x) - \kappa_{M0} \nabla_M w_0(x) \end{pmatrix} = \frac{d\rho(z;\lambda_0)}{d\lambda}\left[\hat{\lambda}_n - \lambda_0\right] + \begin{pmatrix} 0 \\ \left(\hat{\kappa}_C - \kappa_{C0}\right)\left(\nabla_C \hat{w}(x) - \nabla_C w_0(x)\right) \\ \left(\hat{\kappa}_M - \kappa_{M0}\right)\left(\nabla_M \hat{w}(x) - \nabla_M w_0(x)\right) \end{pmatrix},$$

and

$$\begin{pmatrix} \pi_n u^*_w + u^*_{\theta 3} x_C + u^*_{\theta 4} x_M \\ u^*_{\theta 1} \nabla_C \hat{w}(x) + \hat{\kappa}_C \nabla_C (\pi_n u^*_w(x)) \\ u^*_{\theta 2} \nabla_M \hat{w}(x) + \hat{\kappa}_M \nabla_M (\pi_n u^*_w(x)) \end{pmatrix}$$

$$= \frac{d\rho(Z;\lambda_0)}{d\lambda}[\pi_n u^*] + \begin{pmatrix} 0 \\ u^*_{\theta 1} \nabla_C \left(\hat{w}(x) - w_0(x)\right) + \left(\hat{\kappa}_C - \kappa_{C0}\right) \nabla_C (\pi_n u^*_w(x)) \\ u^*_{\theta 2} \nabla_M \left(\hat{w}(x) - w_0(x)\right) + \left(\hat{\kappa}_M - \kappa_{M0}\right) \nabla_M (\pi_n u^*_w(x)) \end{pmatrix},$$

Condition 4.3' is satisfied given the definition of $\|\cdot\|$ and

$$\langle \hat{\lambda}_n - \lambda_0, \pi_n u^* \rangle = \mathbb{E}\left[\left(\frac{d\rho(z_i;\lambda_0)}{d\lambda}\left[\hat{\lambda}_n - \lambda_0\right]\right)' \Sigma(x_i)^{-1} \frac{d\rho(z_i;\lambda_0)}{d\lambda}[\pi_n u^*]\right].$$

Condition 4.2.' can be verified by applying Lemma 4.2 in Chen (2007). Condition on the metric entropy with bracketing for Lemma 4.2 is satisfied with Assumption 9. *Q.E.D.*

PROOF OF THEOREM 3: We follow the proofs of Theorem 4.1 and 6.2 in Ai and Chen (2003). Let $\mathcal{N}_{on} \equiv \left\{\lambda \in \Theta \times \mathcal{W}_n : \|\lambda - \lambda_0\|_s = o(1),\ \|\lambda - \lambda_0\| = o\left(n^{-1/4}\right)\right\}$. By Proposition 3, the sieve GLS estimator $\tilde{\lambda}_n$ in Step 1 satisfies $\|\tilde{\lambda}_n - \lambda_0\|_s = o_p(1)$ and $\|\tilde{\lambda}_n - \lambda_0\| = o_p\left(n^{-1/4}\right)$. Hence $\tilde{\lambda}_n \in \mathcal{N}_{on}$. Using the proof similar to those of Proposition 3, we can also show that $\hat{\lambda}_n \in \mathcal{N}_{on}$. Let $u_{0\theta} = (u_{0\theta 1}, u_{0\theta 2}, u_{0\theta 3}, u_{0\theta 4})' = \left(\mathbb{E}\left[D_{v_0}(x_i)' \Sigma(x_i)^{-1} D_{v_0}(x_i)\right]\right)^{-1} \eta$, $u_{0w} = -v_0 u_{0\theta}$ and $u_0 = (u_{0\theta}, u_{0w})$, where $\eta \in \mathbb{R}^4$ is an arbitrary unit vector. Then, we have

$$(\theta - \theta_0)'\eta = \langle \lambda - \lambda_0, u_0 \rangle = \mathbb{E}\left[\left(\frac{d\rho(z_i;\lambda_0)}{d\lambda}[\lambda - \lambda_0]\right)' \Sigma_0(x_i)^{-1} \frac{d\rho(z_i;\lambda_0)}{d\lambda}[u_0]\right]$$

for all $\lambda \in \Lambda$. Let $\varepsilon_n = o\left(n^{-1/2}\right) > 0$. Denote $u_{n0} := (u_{n0\theta}, u_{n0w}) = \pi_n u_0$ to simplify notation. We take a continuous path $\lambda(t) = \hat{\lambda}_n \pm t\varepsilon_n u_{n0}$. Then $\{\lambda(t) : t \in [0,1]\}$ in $\mathcal{N}_{on}$. Let

$$\hat{Q}_n(\lambda(t)) = -\frac{1}{n}\sum_{i=1}^{n} \rho(z_i;\lambda(t))' \hat{\Sigma}_0(x_i)^{-1} \rho(z_i;\lambda(t)).$$

By definition of $\hat{\lambda}_n$, and a Taylor expansion around $t = 0$ up to second order, we obtain

$$0 \le \hat{Q}_n\left(\hat{\lambda}_n\right) - \hat{Q}_n\left(\hat{\lambda}_n \pm \varepsilon_n u_{n0}\right) = -\left.\frac{d\hat{Q}_n(\lambda(t))}{dt}\right|_{t=0} - \frac{1}{2}\left.\frac{d^2\hat{Q}_n(\lambda(t))}{dt^2}\right|_{t=s},$$

for some $s \in [0,1]$, where

$$-\left.\frac{d\hat{Q}_n(\lambda(t))}{dt}\right|_{t=0} = \frac{\pm 2\varepsilon_n}{n}\sum_{i=1}^{n} \rho\left(z_i;\hat{\lambda}_n\right)' \hat{\Sigma}_0(x_i)^{-1} \frac{d\rho\left(z_i;\hat{\lambda}_n\right)}{d\lambda}[u_{n0}]$$

$$-\left.\frac{d^2\hat{Q}_n(\lambda(t))}{dt^2}\right|_{t=s} = \underbrace{\frac{2\varepsilon_n^2}{n}\sum_{i=1}^{n} \rho(z_i;\lambda(s))' \hat{\Sigma}_0(x_i)^{-1} \frac{d^2\rho(z_i;\lambda(s))}{d\lambda d\lambda}[u_{n0}, u_{n0}]}_{:=A_1}$$

$$+ \underbrace{\frac{2\varepsilon_n^2}{n}\sum_{i=1}^{n} \left(\frac{d\rho(z_i;\lambda(s))}{d\lambda}[u_{n0}]\right)' \hat{\Sigma}_0(x_i)^{-1} \frac{d\rho(z_i;\lambda(s))}{d\lambda}[u_{n0}]}_{:=A_2},$$

with

$$\frac{d\rho\left(z;\lambda\left(s\right)\right)}{d\lambda}\left[\varepsilon_n u_{n0}\right] \equiv \left.\frac{d\rho\left(z;\lambda\left(t\right)\right)}{dt}\right|_{t=s}, \quad \frac{d^2\rho\left(z;\lambda\left(s\right)\right)}{d\lambda d\lambda}\left[\varepsilon_n u_{n0}, \varepsilon_n u_{n0}\right] \equiv \left.\frac{d^2\rho\left(z;\lambda\left(t\right)\right)}{dt^2}\right|_{t=s}.$$

We note from Assumption 10.(ii) that each element of

$$\frac{d^2\rho\left(z;\lambda\left(s\right)\right)}{d\lambda d\lambda}\left[u_{n0}, u_{n0}\right] = \left.\frac{d^2\rho\left(z;\lambda + t u_{n0}\right)}{dt^2}\right|_{t=s} = \begin{pmatrix} 0 \\ -2u_{0\theta 1}\nabla_C u_{n0w}\left(x\right) \\ -2u_{0\theta 2}\nabla_M u_{n0w}\left(x\right) \end{pmatrix}$$

is uniformly bounded over $x \in \mathcal{X}$. For A_1, we write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\lambda\left(s\right)\right)'\hat{\Sigma}_0\left(x_i\right)^{-1}\frac{d^2\rho\left(z_i;\lambda\left(s\right)\right)}{d\lambda d\lambda}\left[u_{n0}, u_{n0}\right] \\
=&\frac{1}{n}\sum_{i=1}^{n}\left(\rho\left(z_i;\lambda\left(s\right)\right) - \rho\left(z_i;\lambda_0\right)\right)'\hat{\Sigma}_0\left(x_i\right)^{-1}\frac{d^2\rho\left(z_i;\lambda\left(s\right)\right)}{d\lambda d\lambda}\left[u_{n0}, u_{n0}\right] \\
&+\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\lambda_0\right)'\left(\hat{\Sigma}_0\left(x_i\right)^{-1} - \Sigma_0\left(x_i\right)^{-1}\right)\frac{d^2\rho\left(z_i;\lambda\left(s\right)\right)}{d\lambda d\lambda}\left[u_{n0}, u_{n0}\right] \\
&+\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\lambda_0\right)'\Sigma_0\left(x_i\right)^{-1}\frac{d^2\rho\left(z_i;\lambda\left(s\right)\right)}{d\lambda d\lambda}\left[u_{n0}, u_{n0}\right].
\end{aligned}$$

Since

$$\begin{aligned}
\rho\left(z;\lambda\left(s\right)\right) - \rho\left(z;\lambda_0\right) =& \frac{d\rho\left(z;\lambda_0\right)}{d\lambda}\left[\hat{\lambda}_n - \lambda_0\right] + \begin{pmatrix} 0 \\ \left(\hat{\kappa}_C - \kappa_{C0}\right)\left(\nabla_C\hat{w}\left(x\right) - \nabla_C w_0\left(x\right)\right) \\ \left(\hat{\kappa}_M - \kappa_{M0}\right)\left(\nabla_M\hat{w}\left(x\right) - \nabla_M w_0\left(x\right)\right) \end{pmatrix} \\
&\mp s\varepsilon_n\begin{pmatrix} u_{n0w}\left(x\right) + x_C u_{0\theta 1} + x_M u_{0\theta 2} \\ \hat{\kappa}_C\nabla_C u_{n0w}\left(x\right) + u_{0\theta 3}\nabla_C\hat{w}\left(x\right) + u_{0\theta 3}\nabla_C u_{n0w}\left(x\right) \\ \hat{\kappa}_M\nabla_M u_{n0w}\left(x\right) + u_{0\theta 4}\nabla_M\hat{w}\left(x\right) + u_{0\theta 4}\nabla_M u_{n0w}\left(x\right) \end{pmatrix},
\end{aligned}$$

the first term of the right-hand side is $o_p\left(n^{-1/4}\right)$ uniformly over $\lambda(s) \in \mathcal{N}_{on}$, which implies that A_1 is $o_p\left(\varepsilon_n^2\right)$. For A_2, we write

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\rho(z_i;\lambda(s))}{d\lambda}[u_{n0}]\right)'\hat{\Sigma}_0(x_i)^{-1}\frac{d\rho(z_i;\lambda(s))}{d\lambda}[u_{n0}] \\
=&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\rho(z_i;\lambda(s))}{d\lambda}[u_{n0}]-\frac{d\rho(z_i;\lambda_0)}{d\lambda}[u_{n0}]\right)'\hat{\Sigma}_0(x_i)^{-1}\frac{d\rho(z_i;\lambda(s))}{d\lambda}[u_{n0}] \\
&+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\rho(z_i;\lambda_0)}{d\lambda}[u_{n0}]\right)'\hat{\Sigma}_0(x_i)^{-1}\left(\frac{d\rho(z_i;\lambda(s))}{d\lambda}[u_{n0}]-\frac{d\rho(z_i;\lambda_0)}{d\lambda}[u_{n0}]\right) \\
&+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\rho(z_i;\lambda_0)}{d\lambda}[u_{n0}]\right)'\left(\hat{\Sigma}_0(x_i)^{-1}-\Sigma_0(x_i)^{-1}\right)\frac{d\rho(z_i;\lambda_0)}{d\lambda}[u_{n0}] \\
&+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d\rho(z_i;\lambda_0)}{d\lambda}[u_{n0}]\right)'\Sigma_0(x_i)^{-1}\frac{d\rho(z_i;\lambda_0)}{d\lambda}[u_{n0}]
\end{aligned}
$$

Since

$$
\begin{aligned}
\frac{d\rho(z;\lambda(s))}{d\lambda}[u_{n0}]-\frac{d\rho(z;\lambda_0)}{d\lambda}[u_{n0}] =& \begin{pmatrix} 0 \\ u_{0\theta 1}\nabla_C(\hat{w}(x)-w_0(x))+(\hat{\kappa}_C-\kappa_{C0})\nabla_C u_{n0w}(x) \\ u_{0\theta 2}\nabla_M(\hat{w}(x)-w_0(x))+(\hat{\kappa}_M-\kappa_{M0})\nabla_M u_{n0w}(x) \end{pmatrix} \\
&\pm s\varepsilon_n\begin{pmatrix} 0 \\ u_{0\theta 1}\nabla_C u_{n0w}(x)+u_{0\theta 1}\nabla_C(\pi_n u_w^*(x)) \\ u_{0\theta 2}\nabla_M u_{n0w}(x)+u_{0\theta 2}\nabla_M(\pi_n u_w^*(x)) \end{pmatrix},
\end{aligned}
$$

the first two terms on the right-hand side are $o_p\left(n^{-1/4}\right)$ and the third term is $o_p(1)$ uniformly over $\lambda(s) \in \mathcal{N}_{on}$, which implies that A_2 is $O_p\left(\varepsilon_n^2\right)$. Moreover, since $\varepsilon_n = o\left(n^{-1/2}\right) > 0$, we obtain uniformly over $\lambda(s) \in \mathcal{N}_{on}$:

$$
\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\hat{\lambda}_n\right)'\hat{\Sigma}_0(x_i)^{-1}\frac{d\rho\left(z_i;\hat{\lambda}_n\right)}{d\lambda}[u_{n0}] = o_p\left(n^{-1/2}\right).
$$

Write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\hat{\lambda}_n\right)'\hat{\Sigma}_0\left(x_i\right)^{-1}\frac{d\rho\left(z_i;\hat{\lambda}_n\right)}{d\lambda}\left[u_{n0}\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\hat{\lambda}_n\right)'\hat{\Sigma}_0\left(x_i\right)^{-1}\left(\frac{d\rho\left(z_i;\hat{\lambda}_n\right)}{d\lambda}\left[u_{n0}\right]-\frac{d\rho\left(z_i;\lambda_0\right)}{d\lambda}\left[u_0\right]\right)\\
&+\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\hat{\lambda}_n\right)'\left[\hat{\Sigma}_0\left(x_i\right)^{-1}-\Sigma_0\left(x_i\right)^{-1}\right]\frac{d\rho\left(z_i;\lambda_0\right)}{d\lambda}\left[u_0\right]\\
&+\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\hat{\lambda}_n\right)'\Sigma_0\left(x_i\right)^{-1}\frac{d\rho\left(z_i;\lambda_0\right)}{d\lambda}\left[u_0\right].
\end{aligned}$$

Using proofs for second derivative terms together with Assumption 12, the first two terms on the right-hand side are $o_p\left(n^{-1/2}\right)$. Since $\left\{\rho\left(z_i;\lambda\right)'\Sigma_0\left(x_i\right)^{-1}\frac{d\rho(z_i;\lambda_0)}{d\lambda}\left[u_0\right]:\lambda\in\mathcal{N}_{on}\right\}$ is a Donsker class,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\rho\left(z_i;\hat{\lambda}_n\right)-\rho\left(z_i;\lambda_0\right)\right)'\Sigma_0\left(x_i\right)^{-1}\frac{d\rho\left(z_i;\lambda_0\right)}{d\lambda}\left[u_0\right]\\
&=\mathbb{E}\left[\left(\rho\left(z_i;\hat{\lambda}_n\right)-\rho\left(z_i;\lambda_0\right)\right)'\Sigma_0\left(x_i\right)^{-1}\frac{d\rho\left(z_i;\lambda_0\right)}{d\lambda}\left[u_0\right]\right]+o_p\left(n^{-1/2}\right).
\end{aligned}$$

With $\rho\left(z_i;\hat{\lambda}_n\right)-\rho\left(z_i;\lambda_0\right)-\frac{d\rho(z_i;\lambda_0)}{d\lambda}\left[\hat{\lambda}_n-\lambda_0\right]=o_p\left(n^{-1/2}\right)$ uniformly over $x_i\in\mathcal{X}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\hat{\lambda}_n\right)'\Sigma_0\left(x_i\right)^{-1}\frac{d\rho\left(z_i;\lambda_0\right)}{d\lambda}\left[u_0\right]\\
&=\langle\hat{\lambda}_n-\lambda_0,u_0\rangle+\frac{1}{n}\sum_{i=1}^{n}\rho\left(z_i;\lambda_0\right)'\Sigma_0\left(x_i\right)^{-1}\frac{d\rho\left(z_i;\lambda_0\right)}{d\lambda}\left[u_0\right]+o_p\left(n^{-1/2}\right).
\end{aligned}$$

Then,

$$\sqrt{n}\left(\theta-\theta_0\right)'\eta=\sqrt{n}\langle\hat{\lambda}_n-\lambda_0,u_0\rangle=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho\left(z_i;\lambda_0\right)'\Sigma_0\left(x_i\right)^{-1}\frac{d\rho\left(z_i;\lambda_0\right)}{d\lambda}\left[u_0\right]+o_p\left(1\right),$$

and Theorem 3 follows from applying a standard CLT for i.i.d. data. *Q.E.D.*

APPENDIX B: BERNSTEIN POLYNOMIALS WITH CONVEX CONSTRAINTS

The equilibrium wage function, $w\left(x\right)$, obtained by the optimal transport theory is unique and convex (up to constant). However, finite sample estimators of $w\left(x\right)$, $\hat{w}_n\left(x;\gamma\right)$, might

be nonconvex at the values close to the boundary of $\mathcal{X}$. To obtain a more stable estimator for $w(x)$, we impose a convexity restriction in the sieve-based estimation procedures without loss of generality. Among many possible linear approximating spaces, we particularly consider the following Bernstein polynomial sieve space:

$$\mathcal{W}_n = \left\{ w_n : \mathcal{X} \to \mathbb{R} : w_n(x;\gamma) = \sum_{j_1,\ldots,j_d=0}^{k_n} \gamma_{j_1\cdots j_d} \left[\prod_{\ell=1}^{d} p_{j_\ell}(x_\ell) \right] : \right.$$
$$\left. p_{j_\ell}(x_\ell) = \binom{k_n}{j_\ell} \left(\frac{x_\ell - \underline{x}_\ell}{\overline{x}_\ell - \underline{x}_\ell} \right)^{j_\ell} \left(\frac{\overline{x}_\ell - x_\ell}{\overline{x}_\ell - \underline{x}_\ell} \right)^{k_n - j_\ell} \right\},$$

for $j_\ell = 1, 2, \ldots, k_n$ where p_{j_ℓ} is the Bernstein basis polynomial.

Let $\mathcal{W}^{cvx}$ be the set of midpoint convex functions:

$$\mathcal{W}^{cvx} = \left\{ w \in C(\mathcal{X}) : 2w\left(\frac{x_1 + x_2}{2} \right) \leq w(x_1) + w(x_2),\ \forall x_1, x_2 \in \mathcal{X} \right\},$$

where $C(\mathcal{X})$ is the class of all continuous functions on $\mathcal{X}$. We do not assume that the true function $w(x)$ has derivatives of any order. In fact, $\mathcal{W}^{cvx}$ is the class of all continuous convex functions because a continuous function that is midpoint convex is convex.

For the first, we consider the one-dimensional ($d = 1$) constrained Bernstein polynomial sieve space, $\mathcal{W}_n^{cvx} = \{ w_n(x;\gamma) \in \mathcal{W}_n : A\gamma \geq 0 \}$, where

$$A\gamma \equiv \underbrace{\begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ & & \ddots & & & \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}}_{(k_n-1)\times(k_n+1)} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{k_n} \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Since the second derivatives of $w_n(x;\gamma)$ can be written as

$$w_n^{(2)}(x;\gamma) = \frac{k_n(k_n-1)}{(\overline{x} - \underline{x})^2} \sum_{j=0}^{k_n-2} (\gamma_{j+2} - 2\gamma_{j+1} + \gamma_j) \binom{k_n - 2}{j} \left(\frac{x - \underline{x}}{\overline{x} - \underline{x}} \right)^j \left(\frac{\overline{x} - x}{\overline{x} - \underline{x}} \right)^{k_n - 2 - j},$$

the above restriction ensures $w_n^{(2)}(\cdot) \geq 0$ for all n. Wang and Ghosh (2012) show that $\{\mathcal{W}_n^{cvx}\}$ is nested and dense in $\mathcal{W}^{cvx}$ with respect to sup-norm.

For the two-dimensional sieve in eq.(12), we consider the following linear constraints:

$$\begin{aligned}
&\gamma_{j_C+2,j_M} - 2\gamma_{j_C+1,j_M} + \gamma_{j_C,j_M} \geq 0, \quad \forall j_C = 0,\ldots,k_{C_n}-2, j_M = 0,\ldots,k_{M_n}, \\
&\gamma_{j_C,j_M+2} - 2\gamma_{j_C,j_M+1} + \gamma_{j_C,j_M} \geq 0, \quad \forall j_C = 0,\ldots,k_{C_n}, j_M = 0,\ldots,k_{M_n}-2.
\end{aligned} \tag{17}$$

Then, the Bernstein polynomial sieve space with linear constraints (17) is nested and dense in

$$\begin{aligned}
\widetilde{\mathcal{W}}^{cvx} = \Bigg\{ w \in C(\mathcal{X}) : 2w\left(\frac{x_{C1}+x_{C2}}{2}, x_M\right) \leq w(x_{C1},x_M) + w(x_{C2},x_M) \ \& \\
2w\left(x_C, \frac{x_{M1}+x_{M2}}{2}\right) \leq w(x_C,x_{M1}) + w(x_C,x_{M2}), \\
\forall (x_{C1},x_M),(x_{C2},x_M),(x_C,x_{M1}),(x_C,x_{M2}) \in \mathcal{X} \Bigg\},
\end{aligned}$$

which is larger than

$$\begin{aligned}
\mathcal{W}^{cvx} = \Bigg\{ w \in C(\mathcal{X}) : 2w\left(\frac{x_{C1}+x_{C2}}{2}, \frac{x_{M1}+x_{M2}}{2}\right) \leq w(x_{C1},x_{M1}) + w(x_{C2},x_{M2}), \\
\forall (x_{C1},x_{M1}),(x_{C2},x_{M2}) \in \mathcal{X} \Bigg\}.
\end{aligned}$$

Note that Floater (1994) provides sufficient conditions for the two-dimensional Bernstein polynomial $w_n(x;\gamma)$ to be convex, which includes linear inequalities (17) as well as additional nonlinear constraints. We use (17) for our estimation because (i) they are easy to impose in the optimization procedure and be extended to higher-dimensional functions, and (ii) $w(x) \in \mathcal{W}^{cvx} \subset \widetilde{\mathcal{W}}^{cvx}$.

APPENDIX C: EFFECTS OF PRODUCTION TECHNOLOGY ON WAGE DISTRIBUTION

In this section, we check whether the Lindenlaub (2017) model's predictions on the effects of technological changes on wage distribution hold for the transformed data and other distributions. We consider three DGP's with the same quadratic production function, $s(x,y) = \alpha_{CC} x_C y_C + \alpha_{MM} x_M y_M$, and three different skill distributions, respectively: (1) bivariate normal distribution (Lindenlaub, 2017), (2) transformed Gumbel copula, and (3) untransformed Gumbel copula.

For the first DGP, we set workers' skill bundle, x, and jobs' skill requirements, y, to follow standard joint normal distributions with $\rho_x = -0.2$ and $\rho_y = -0.6$, respectively. For both nonnormal transformed and untransformed DGPs, we set $x = (x_1, x_2)$ and $y = (y_1, y_2)$ to follow the Gumbel copula with the shape parameter values 1.25 and 2.5 respectively. Then Kendall's correlation coefficients of x and x are 0.2 and 0.6. For the transformed data, we convert x and y into standard normally distributed variables:

$$x_{Ci} = \Phi^{-1}(x_{1i}),\ x_{Mi} = \Phi^{-1}(1 - x_{2i}),\ y_{Cj} = \Phi^{-1}(y_{1j}),\ y_{Mj} = \Phi^{-1}(1 - y_{2j}),$$

so that (x_{Ci}, x_{Mi}) and (y_{Cj}, y_{Mj}) are negatively correlated. For the untransformed data,

$$x_{Ci} = x_{1i},\ x_{Mi} = 1 - x_{2i},\ y_{Cj} = y_{1j},\ y_{Mj} = 1 - y_{2j}.$$

As we mentioned in the main text, there is no guarantee that the transformed data follows a joint normal distribution, implying that no closed-form solution exists for the equilibrium wage and assignment function. Hence, we numerically solve the equilibrium matching through linear programming for each Monte Carlo sample and different parameter values of $(\alpha_{CC}, \alpha_{MM})$.

Figure C.1 plots the skewness and variance of the distribution of wages for each DGP. As Lindenlaub (2017) derived, for all three DGPs, wage distributions are positively skewed for different pairs of α_{CC} and α_{MM}, and the wage dispersion increases as cognitive or manual skill complementarity increases. However, for the transformed Gumbel copula, the skewness decreases as α_{CC} decreases and α_{MM} increases. This simulation result is inconsistent with the theoretical result for normally distributed X and Y, in which the skewness is minimized when $\alpha_{CC} = \alpha_{MM}$. It implies that estimating the Gaussian model in Lindenlaub (2017) with transformed data can mislead the effects of technological changes on wages and inequality.

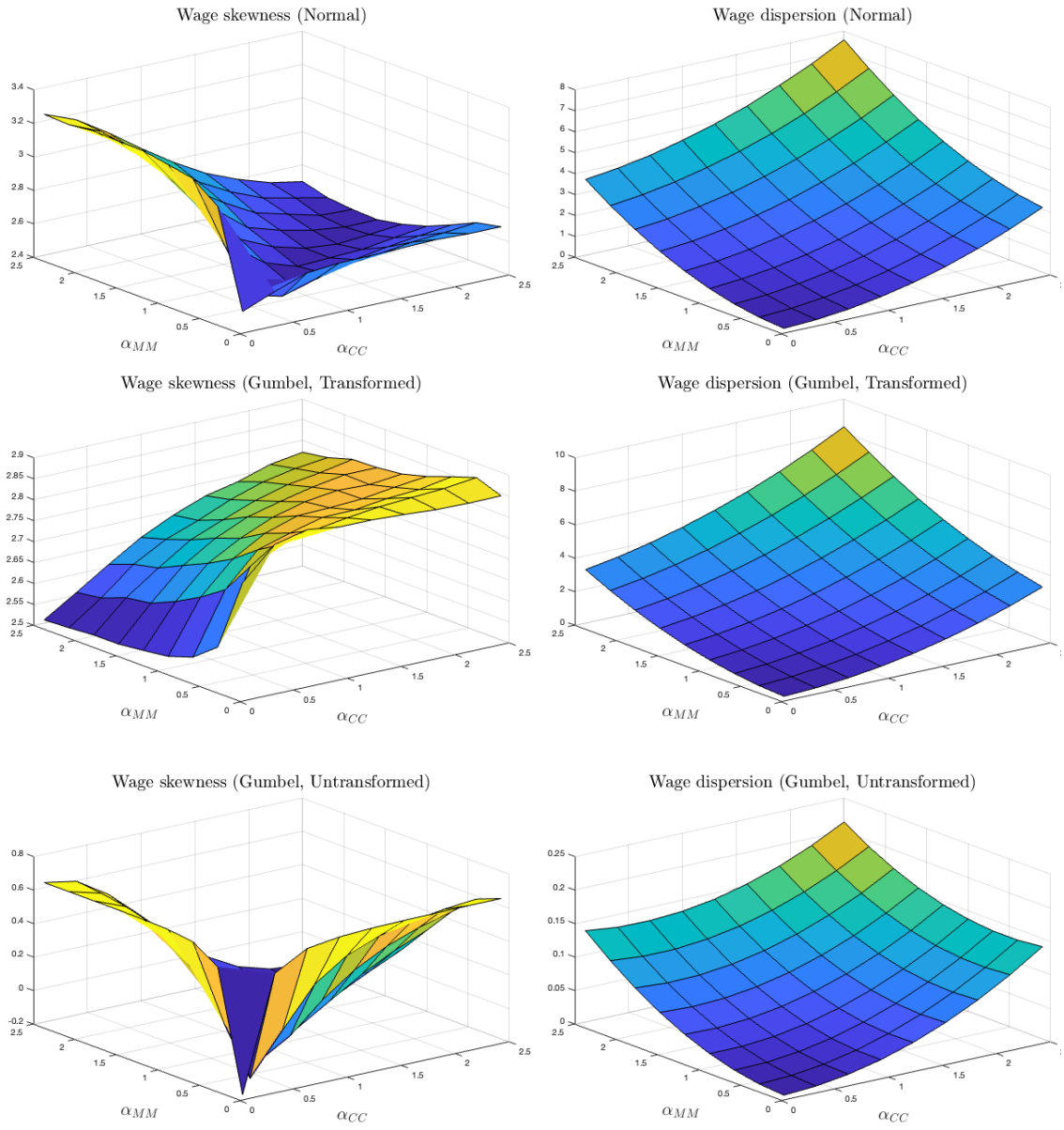

FIGURE C.1.—Effects of changes in production technology on wages and inequality