Approximating Unobserved Heterogeneity with Finite Mixtures

John Rust, Georgetown University
Tiemen Woutersen, University of Arizona

DSE 2024

August 8, 2024

Outline of lecture

- Terminology: "random effects" vs "fixed effects"
- Limits to fixed effects: the Incidental Parameters problem
- Heterogeneous preferences in structural models: how does "randomness" enter preferences?
- The Random Coefficients Logit Model is Identified
- Theoretical vs "Practical" Identification of heterogeneity
- Finite mixtures of logit models are neural networks
- Value of Panel Data for Identification
- How to handle observed vs unobserved heterogeneity
- Empirical applications

Empirical Applications

- "Are People Bayesian?" El-Gamal and Grether
- Can the "Big 5" Personality Traints Explain Preference Heterogeneity? Tomas Jagelka
- Consumer heterogeneity in online shopping for smartphones, Chengjun Zhang
- Deep Learning of Heterogenous Consumer Aesthetics in Retail Fashion, Pranjal Rawat
- Bidding strategies in online ascending bid auctions, Cho, Paarsch, Rust

Fixed or Random Effects?

- I find these terms extremely confusing and prefer to avoid using them whenever possible.
- Statisticians and reduced-form econometricians love these terms.
- In a 2005 *Annals of Statistics* article, Andrew Gelman noted that "A persistent point of conflict in the ANOVA literature is the appropriate use of fixed or random effect" and he outlines 5 different definitions of fixed vs random effects that he found in the literature.
- There is not much more clarity about this distinction in structural eonometrics when it comes to estimation of preference parameters. We see the term "random coefficients logit models" suggesting these coefficients are *random* but actually the coefficients may be *fixed* at the individual level (i.e. "fixed effects") but random from the standpoint of their distribution in the overall population.
- So it is important to be clear what we are talking about and not use terminology loosely and carelessly. *Is a "fixed effect" a parameter or a random variable or neither?*

Simplest Regression Example

- Suppose we have panel data and are interested in estimating the regression model

$$y_{it} = \tau_i + \epsilon_{it}, \quad t = 1, \ldots, T, \ i = 1, \ldots, N \tag{1}$$

  where the unknown parameters are $\theta = (\tau_1, \ldots, \tau_N, \sigma^2)$ where $\epsilon_{it} \sim N(0, \sigma^2)$. Further assume for each i that $\{\epsilon_{it}\}$ are *IID*.
- *Fixed effects* would be an estimator that attempts to estimate all $N + 1$ elements of θ as fixed *parameters*.. However it is clear that if T is fixed, we cannot consistently estimate θ since there are $N + 1$ parameters and TN observations, so the number of observations per parameter is $TN/(N + 1)$ which has the finite limit $\lim_{N\to\infty} TN/(N + 1) = T$ as $N \to \infty$.
- Kiefer and Wolfowitz 1956 used this example "due to Neyman and Scott [1], who used it to prove that the m.l. estimator need not be consistent when there are infinitely many incidental parameters (constants)."

Two solutions to the incidental parameters problem

- **Solution 1** treat the incidental parameters $\{\tau_i\}$ as "nuisance parameters" and "difference them out" resulting in the following regression equation

$$\Delta y_{it} = y_{it} - y_{it-1} = \epsilon_{it} - \epsilon_{it-1}, \quad t = 2, \ldots, T, \ i = 1, \ldots, N \quad (2)$$

- Then using the differenced data we can consistently estimate σ^2 since we have 1 parameter and $N(T-1)$ observations.
- Note that Kiefer and Wolfowitz refer to σ^2 as the "structural parameter" though it is not necessarily "structural" relative to how we think of this term today.
- The drawback of the differencing is that is "throws the baby out with the bathwater" — though we can estimate σ^2, we cannot predict the values of y_{it} without knowing τ_i, and indeed, we cannot even determine the population distribution of y_{it} unless we make use of the un-differenced observations $\{y_{it}\}$.
- Kiefer and Wolfowitz's solution to the incidental parameters problem may be termed *random effects* — we treat the τ_i as realizations from a population distribution $G(\tau)$.

Kiefer and Wolfowitz's solution: random effects

- **Solution 2** Kiefer and Wolfowitz proposed direct maximum likelihood estimation of the unknown parameter $\theta = (\sigma^2, G)$ by maximum likelihood, i.e. maximizing $L(\sigma^2, G)$ given by

$$L(\sigma^2, G) = \prod_{i=1}^{N} \prod_{t=1} \int f(y_{it}|\tau, \sigma^2) G(d\tau), \tag{3}$$

  assuming the model is *identified*, i.e. if $f(y) = \int f(y|\tau, \sigma^2)$ is the true population distribution, then only parameter $(\omega^2, G(\tau))$ that satisfies $\int f(y|\tau, \omega^2) G(d\tau) = \int f(y|\tau, \sigma^2) G(d\tau)$ for all $y \in R$ is $\omega^2 = \sigma^2$ and $H = G$.
- Note that the parameter space is *infinite-dimensional* since there is no finite set of distributions or "vectors" that "span" the space of all probability distributions over τ.
- However if we are willing to restrict G to a parametric family and write it as $G(\tau|\beta)$ for some $k \times 1$ vector of parameters β, then we can estimate the model using ordinary parametric maximum likelihood.

Random coefficients maximum likelihood

- For example if $G(\tau|\beta) \sim N(\beta_0, \beta_1^2)$ then the intercept terms can be regarded as *random coefficients* and we can rewrite the original regression as a new regression involving just the 3 parameters $\theta = (\sigma^2, \beta_0, \beta_1)$

$$y_{it} = \beta_0 + \epsilon_{it} + u_i \quad \text{where } \epsilon_{it} \sim N(0, \sigma^2) \text{ and } u_i \sim N(0, \beta_1^2) \qquad (4)$$

- Note that *conditional on* τ_i the observations $\{y_{it}\}$ are IID $N(\tau_i, \sigma^2)$, the unobserved variation in τ_i leads to an inference of *serial dependence* in $\{y_{it}\}$ if we do not condition on τ_i:

$$\text{cov}(y_{it}, y_{it-1}) = E\{(\epsilon_{it} + u_i)(\epsilon_{it-1} + u_i)\} = \beta_1^2 > 0. \qquad (5)$$

- But we would like to avoid making too many parametric assumptions, so is direct non-parametric estimation of G feasible? Kiefer and Wolfowitz established the theoretical consistency of their maximum liklihood estimator but did not address how it could be calculated.

Estimating the model using *finite mixtures*

- Lindsay *Annals of Statistics* 1983, using *convex duality theory,* showed that the maximum likelihood estimator $\hat{G}$ will be a *finite mixture,* i.e. a discrete random variable taking the values $(\tau_1, \ldots, \tau_J)$ with probabilities $(p_1, \ldots, p_J)$ where J is at most the total number of distinct observations in the data set.
- Thus, let $\theta = (\sigma^2, \tau_1, \ldots, \tau_J, p_1, \ldots, p_J)$, then Kiefer and Wolfowitz's maximum likelihood estimator reduces to

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} \sum_{j=1}^{J} p_j \prod_{t=1}^{T} f(y_{it}|\tau_j, \sigma^2) \quad \text{subject to:} \sum_{j=1}^{J} p_j = 1. \quad (6)$$

- Notice that for any finite sample size, the maximum likelihood problem is *fully parametric* since it depends only on the $2J - 1$ free parameters in θ.
- However if we allow $J \to \infty$ as $N \to \infty$, the finite mixture family constitutes a *sieve* of expanding finite mixture models that is *dense* in the space of all probability distributions over τ. That is, given any $G(\tau)$, there is a finite mixture model that can approximate it arbitrarily closely.

Recall convergence concepts for CDFs

- A sequence of CDFs $\{G_j\}$ *converges weakly* (or *converges in distribution* to a CDF G if and only if

$$\lim_{j\to\infty} \int g(\tau) G_j(\tau) = \int g(\tau) G(d\tau), \quad \forall \text{ continuous, bounded } g \tag{7}$$

- We use the shorthand expression $G_j \Longrightarrow G$ is the sequence $\{G_j\}$ converges weakly to G.
- **Note:** The topology of weak convergence can be *metrized* by the *Prokhorov metric* which we denote by $\rho(G_j, G)$. Thus $G_g \Longrightarrow G$ if and only if $\lim_{j\to\infty} \rho(G_j, G) = 0$.
- If G is a continuous CDF, then weak convergence also implies uniform convergence of CDFs, i.e. $G_j \Longrightarrow G$ if and only if $\lim_{j\to\infty} \|G_j - G\| = 0$ where $\|G_j - G\|$ is given by

$$\|G_j - G\| = \sup_{\tau} |G_j(\tau) - G(\tau)|. \tag{8}$$

- We will say that maximum likelihood estimation of the mixture model is *consistent* if $G_j \Longrightarrow G$ with probability 1.

Bottom line of Kiefer and Wolfwitz 1956

- *Maximum likeihood is consistent* i.e. if $\hat{G}_N$ and σ^2_N are the maximum likelihood estimators of G and σ^2, respectively, then with probability 1 we have

$$\hat{G}_N \Longrightarrow G \quad \text{and} \quad \hat{\sigma}^2_N \to \sigma^2 \tag{9}$$

- Though Lindsay's results suggests that the MLE $\hat{G}_N$ will be a mixture over N points of support we will show that we can achieve very good fits to the data with MLEs involving many fewer point $J < N$.
- But to prove consistency of the MLE, we need $J \to \infty$ at the right rate as $N \to \infty$.
- Getting the right rates to establish consistency is a delicate question.
- Establishing the *rate of convergence* of $\hat{G}_N$ to G and whether the MLE is *asymptotically normal* is an unsolved problem.
- We argue that $f(y|\hat{G}_N, \hat{\sigma}^2) \to f(y|G, \sigma^2)$ "rapidly" (potentially at $O_p(1/\sqrt{N})$) but $\hat{G}_N \Longrightarrow G$ "slowly" i.e. at rates much slower than $O_p(1/\sqrt{N})$. That is, G will typically be *poorly identified* because the likelihood function is *nearly flat in a neighborhood of* G.

Heckman and Singer, 1984

- "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data" *Econometrica*

From their abstract

Conventional analyses of single spell duration models control for unobservables using a random effect estimator with the distribution of unobservables selected by ad hoc criteria. Both theoretical and empirical examples indicate that estimates of structural parameters obtained from conventional procedures are very sensitive to the choice of mixing distribu- tion. Conventional procedures overparameterize duration models. We develop a consistent nonparametric maximum likelihood estimator for the distribution of unobservables and a computational strategy for implementing it. For a sample of unemployed workers our estimator produces estimates in concordance with standard search theory while conventional estimators do not.

Heckman and Singer, 1984

- **Problem** Given cross sectional data on durations of single spells $\{t_1, \ldots, t_N\}$, can we infer unobserved heterogeneity in conditional distributions of spell length?
- **Identification problem** Let $F(t)$ be the marginal distribution of spells in the population (which can be non-parametrically estimated and thus is identified), and a parametric *structural distribution of spell length* $F(t|\tau)$ that depends on a "structural parameter" τ, can we infer the distribution $G(\tau)$ of structural parameters in the population?
- That is, can we invert this equation to uniquely determine $G(\tau)$ given knowledge of $F(t)$?

$$F(t) = \int F(t|\tau)G(d\tau). \tag{10}$$

- Can we also relax parametric assumption on $F(t|\tau)$ and recover *both* $G(\tau)$ from $F(t)$?

The identification problem

- **Answer:** generally not. We cannot generally identify both $F(t|\tau)$ and $G(\tau)$ given knowledge of $F(t)$.
- They present an example where $F(t) = 1 - \exp\{-\eta t\}$, i.e. exponentially distributed durations.
- **Explanation 1:** No heterogeneity, i.e. $G(\tau)$ is a unit mass on η.
- **Explanation 2:** Non-degenerate heterogeneity,

$$\begin{aligned} F(t|\tau) &= 1 - \int_{t/\sqrt{2\tau}}^{\infty} \frac{2}{2\sqrt{\pi}} \exp\{-x^2/2\} dx \\ G(\tau) &= 1 - \exp\{-\eta^2 \tau\} \end{aligned} \tag{11}$$

  Both models imply the same exponentially distributed population distribution $F(t)$.
- On the other hand, once we restrict $F(t|\tau)$ to a parametric family, we can often uniquely identify $G(\tau)$ from $F(t)$. That is, we can "invert" $F(t)$ to recover $G(\tau)$.

Parametric vs Non-parametric identification of G

- **Parametric identification** Assume G is a member of a parametric family, $G(\tau|\theta)$ that depends on a $K \times 1$ vector of parameters θ to be estimated. Then under fairly weak conditions we can show that there is a unique θ^* satisfying $F(t) = \int F(t|\tau)G(d\tau|\theta^*)$, so G is identified under these assumptions.
- **Non-parametric identification** Now, following Kiefer and Wolfowitz, drop the assumption that G is a member of a parametric family. Let G be *any* CDF over τ.
- Heckman and Singer apply the general approach of Kiefer and Wolfowitz 1956 to show that the non-parametric maximum likelihood estimator of G is consistent.
- But in practice, how do they estimate G? Using a parametric family of finite mixture models! That is $G(\tau)$ is a discrete distribution that takes the values $(\tau_1, \ldots, \tau_K)$ with probabilities $(p_1, \ldots, p_K)$. This is a member of a parametric family, $G(\tau|\theta)$ where $\theta = (\tau_1, \ldots, \tau_K, p_1, \ldots, p_K)$! The key difference this parametric family is a *sieve* i.e. $G(\tau|\theta)$ can approximate any CDF $G(\tau)$ arbitrarily closely for values of θ when K is sufficiently large.

Parametric estimation using finite mixtures

- Thus, in practice the "Heckman-Singer method" (NPMLE) is to estimate $F(t)$ using

$$\hat{F}(t) = \sum_{k=1}^{K} G(t|\hat{\tau}_k)\hat{p}_k, \tag{12}$$

where $\hat{\theta} = (\hat{\tau}_1, \ldots, \hat{\tau}_K, \hat{p}_1, \ldots, \hat{p}_K)$ is estimated by maximum likelihood

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) \equiv \prod_{i=1}^{N} \left[\sum_{k=1}^{K} g(t_i|\tau_k)p_k \right], \text{ subject to: } \sum_{k=1}^{K} p_k = 1, \tag{13}$$

where $g(t|\tau) = G'(t|\tau)$ is the density of the conditional CDF $G(t|\tau)$.

- One can use *model selection* to find the number of types K (more on this later).

Three key findings from applications of Heckman-Singer

Finding 1

The NPMLE typically estimates $F(t)$ very well. Or if there are auxiliary (non-heterogeneous) parameters α entering $G(t|\tau, \alpha)$ then "the NPMLE estimates the structural parameters (α) very well."

Finding 2

The number of types "discovered" by the "Heckman-Singer estimator" is often quite small: $K \in \{2, 3, 4\}$ is quite typical.

Finding 3

$G(\tau)$ is poorly estimated: "for both finite and continuous mixing measures, the NPMLE *never* estimates jt very well. This is true even when samples of size 5000 are used to generate NPMLE estimates."

The multinomial logit model (MNL)

- You have already been heavily exposed to the multinomial logit model, where choice probabilities (CCPs) are given by

$$P(j|x,\tau) = \frac{1}{1+\sum_{j=1}^{J}\exp\{x_j\tau\}}, \tag{14}$$

where τ is an $d \times 1$ vector of coefficients on the *attributes* x_j of J alternatives, and where we have normalized the utility of the *outside good* $j = 0$ to 0, so $x_j\tau$ can be viewed as the gain (or loss) in utility relative to the utility of the outside good.

- Key problem with the logit model is the *Independence of Irrelevant Alternatives* (IIA) property that stems from the McFadden's derivation of the MNL model as a *random utility model* (RUM)

$$P(j|x,\tau) = \Pr\{u_j \geq u_{j'},\ j' \neq j | x_1,\ldots,x_J\}, \tag{15}$$

where

$$u_j = x_j\tau + \epsilon_j, \quad u_0 = \epsilon_0, \tag{16}$$

where $\epsilon = (\epsilon_0, \epsilon_1, \ldots, \epsilon_J)$ is a vector of *IID* Type 3 extreme value random variables.

Limitations of the logit model

- BLP and others have noticed that in IO applications the IIA implies restrictive and unrealistic patterns of cross-price elasticities. For example if η_{ij} is the elasticity of demand (market share) of good i with respect to an increase in the price of good j, we have

$$\eta_{ij} = \frac{\partial}{\partial p_j} log(P(i|x,\tau)) = -\tau_p p_j P(j|x,\tau) > 0, \tag{17}$$

  where $\tau_p < 0$ is the "price coefficient". Notice that this only depends on j but not on i: *IIA implies that an increase in the the price of good j results in a proportionate increase in the market shares of all other goods.* Empirically, the cross-price elasticity is typically higher for products j that are more similar to good i.
- Suppose τ is a random vector with mean $\overline{\tau}$, so we can write $\tau = \overline{\tau} + \nu$. Then we can express the random utility u_j

$$u_j = x_j\overline{\tau} + \epsilon_j + x_j\nu, \tag{18}$$

  so now the error terms are no longer independent, so the correlation between u_j and u_i depends on x_j and x_i.

RUM vs RPM Models

- Notice that in the standard logit model, τ is treated as a *fixed structural/preference parameter* (and hence is estimated), wheres ϵ is treated as a vector of random variables, so the randomness comes only via ϵ not via τ.
- But what about cases where τ is treated as a random variable, as in the random coefficients logit model? Then is τ no longer a parameter or "fixed effect" but rather a "random effect"?
- However if we have panel data, with repeated observations of choices for the same individual, $\{j_1, \ldots, j_T\}$, the typical interpretation is that τ is *fixed for any individual* and only ϵ is random across successive choices, so τ is only *random across different individuals in the population.*
- The *random preference model* (RPM) discussed in Apesteguia and Ballester (2018) treats τ as an *IID* random vector *even for successive choices by the same person.*
- In effect, this makes τ a "random effect" even at the individual level: people's preferences are assumed to be very unstable and can change from choice to choice, or even second to second.

The random coefficients logit model (mixed logit)

- Let $G(\tau)$ be the (multivariate) distribution of the random coefficients. Then the CCPs $P(x) \equiv \{P(j|x)|j = 0, 1, \ldots, J\}$ are given by

$$P(j|x) = \int \left[\frac{1}{1 + \sum_{j=1}^{J} \exp\{x_j \tau\}} \right] G(d\tau). \tag{19}$$

- $P(j|x)$ is also called the *mixed logit* model since it is a mixture of logits with G as the "mixing distribution".
- McFadden and Train (2000) showed that mixed logits are "dense" in the sense that if $P(j|x)$ is the CCP implied by any random utility model, then $P(j|x)$ can be approximated aribtrarily closely by a mixed logit for some appropriate mixing distribution G.
- So can we apply the Kiefer-Wolfowitz/Heckman Singer approach and estimate very flexible discrete choice models with arbitrary CCPs $P(x)$ by maximum likelihood, directly maximizing over the mixing distribution G as an infinite-dimensional "parameter"?
- This depends crucially on whether G is identified from the CCPs P.

The random coefficients logit model is identified

- Title of 2012 *Journal of Econometrics* paper by Fox, Kim, Ryan and Bajari They provide both a non-constructive and a constructive proof of identification of G from P.
- **Non-constructive proof.** (proof by contradiction). Suppose there are two distinct distributions G and G' that imply the same CCP $P(x)$. Then we have

$$0 = \int P(x\tau)[G(d\tau) - G'(d\tau)], \tag{20}$$

  where $P(x\tau)$ denotes the MNL CCPs.
- $\Gamma(\tau) = G(\tau) - G'(\tau)$ is a *signed measure.* A theorem of Cybenko (1989) shows that if P has the property of being *discriminatory* then the only solution of the equation above is for $G(\tau) = G'(\tau)$ for all τ. Cybenko proved that *any bounded, measurable sigmoidal function of the form* $\sigma(x\tau)$ *is discriminatory.* But the MNL is a bounded, measurable sigmoidal function and hence is discriminatory. Hence assuming G is not identified contradicts Cybenko's result on discriminatory functions.

The random coefficients logit model is identified

- **Constructive proof** Consider the special case of the binary logit model where there are two choice, $j \in \{0, 1\}$ where

$$P(1|x, \tau, \theta) \equiv P(x|\tau, \theta) = \frac{1}{1 + \exp\{\theta + x\tau\}}, \tag{21}$$

  so the intercept is assumed to be common for different people and only the slope coefficient τ is a random coefficient.
- Notice that $P(x|\tau, \theta)$ has derivatives with respect that are continuous and bounded by 1 for all orders.
- Now use Fubini's Theorem to find the derivative of $P(x|\theta)$ the mixed CCP

$$\frac{d}{dx} P(x|\theta) = \frac{d}{dx} \left[\int P(x|\tau, \theta) G(d\tau) \right] = \int \left[\frac{d}{dx} P(x|\tau, \theta) G(d\tau) \right] \tag{22}$$

The random coefficients logit model is identified

- Using the fact that $\frac{d}{dx}P(x|\tau,\theta) = -\tau P(x|\tau,\theta)[1 - P(x|\tau,\theta)]$ we get

$$\frac{d}{dx}P(x|\theta)\bigg|_{x=0} = -P(0|\theta)[1 - P(0|\theta)]\int \tau G(d\tau), \qquad (23)$$

  where $P(0|\theta) = 1/(1 + \exp\{\theta\})$. Thus, if we know $P(x|\theta)$, then we back out $E\{\tau\}$ using its known derivative at $x = 0$.
- We can continue this way, taking higher order derivatives of $P(x|\theta)$ with respect to x and evaluate at $x = 0$ to get formulas of the form

$$\frac{d^n}{dx^n}P(x|\theta)\bigg|_{x=0} = \mathcal{P}(P(0|\theta))\int \tau^n G(d\tau), \qquad (24)$$

  where $\mathcal{P}(P(0|\theta))$ denotes a polynomial evaluated at $P(0|\theta)$.
- Observing that form almost all θ we have $\mathcal{P}(P(0|\theta)) \neq 0$, it follows that we can recover moments of all orders of the distribution $G(\tau)$ from the corresponding derivatives of $P(x|\theta)$ evaluated at $x = 0$.
- Since a distribution is uniquely determined by all of its moments, it follows that $G(\tau)$ is identified.

Does the constructive proof yield a practical estimator of G

- **NO!** Not possible to get good non-parametric estimates of the derivatives of all orders of the CCP $P(x|\theta)$ at the specific point $x = 0$ to produce reliable estimates of the moments of all orders of G, and invert these moments to produce a reliable/robust estimator.
- Are there better non-parametric estimators of G?
- Fox, Kim and Yang *Journal of Econometrics* 2016 developed a "A simple nonparametric approach to estimating the distribution of random coefficients in structural models".
- This is often called the "fixed grid" estimator since it depends on fixing "a grid of heterogeneous parameters and estimate only the weights on the grid points, an approach that is computationally attractive compared to alternative nonparametric estimators. "
- That is, they propose fixing a number of grid points R and choosing (pre-specifying) a grid of R points $(\tau_1, \ldots, \tau_R)$ and estimating G by $\hat{G}$ given by

$$\hat{G}(\tau) = \sum_{r=1}^{R} \hat{p}_r I\{\tau \leq \tau_r\}. \tag{25}$$

Implementing the fixed grid estimator by linear regression

- How do Fox, Kim and Yang estimate the probabilities $\hat{p}_r$ of the grid point? By (constrained) linear regression.
- Let the binary outcome y_j be defined as 1 if the person chooses alternative j and 0 otherwise. If $P(j|x)$ is the mixture of the MNL choice probabilities (representing the "average" probability of choosing j) then we have the following regression equation

$$y_j = P(j|x) + \nu_j \tag{26}$$

- Now approximate $P(j|x)$ using the the fixed grid over τ

$$P(j|x) = \int P(j|x,\tau)G(d\tau) \simeq \sum_{r=1}^{R} p_r P(j|x,\tau_r), \tag{27}$$

where $P(j|x,\tau)$ is the MNL choice probability or "kernel", $P(j|x,\tau) = \exp\{x_j\tau\}/(1+\sum_{j'=1}^{J}\exp\{x_{j'}\tau\})$.

Implementing the fixed grid estimator by linear regression

- Suppose we have cross-sectional data on N individuals and define the $J+1$ vector of choice indicators by $y_i = (y_{i0}, y_{i1}, \ldots, y_{iJ})$ were $y_{ij} = 1$ if person i chooses item j and 0 otherwise. Let $(x_1, \ldots, x_N)$ be the corresponding covariates entering the logit model.
- Then we have the constrained linear regression given by

$$\hat{p} = (\hat{p}_1, \ldots, \hat{p}_R) = \underset{p_1, \ldots, p_R}{\mathit{argmin}} \sum_{i=1}^{N} \sum_{j=0}^{J} \left(y_{ij} - \sum_{r=1}^{R} p_r P(j|x_i, \tau_r) \right)^2 \quad (28)$$

  subject to the constraint that $\sum_{r=1}^{R} p_r = 1$.
- This fixed grid regression estimator is appealing in its simplicity, but how do we choose the fixed grid $(\tau_1, \ldots, \tau_R)$ and number of grid points R?
- We do not know the support of the unknown distribution G, thus the researcher will have little clue about how to choose the fixed grid.

Is there a better way?

- Nevo (2016 *Econometrica*) used a modified GMM version of the Fox-Kim-Yang fixed grid estimator to estimate a structural model of residential demand for broadband, using a fixed grid with $R = 16807$ types, i.e. seven points of support for each of the five parameters. But he finds that only a very few fixed grid points get positive weight: "No plan has more than 20 types receiving positive weights, while the average number of types across plans is only 6.6.".
- What about estimating a finite mixture model to approximate unobserved heterogeneity? Instead of choosing a fixed grid for τ *we let the data tell us what the best mass points should be!*
- That is, estimate a mixture model with $Rd + R - 1$ free parameters given by $\theta = (\tau_1, \ldots, \tau_R, p_1, \ldots, p_R)$ and solve the following *nonlinear least squares problem*

$$\hat{\theta} = \underset{\tau_1,\ldots,\tau_R,p_1,\ldots,p_R}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{j=0}^{J}\left(y_{ij} - \sum_{r=1}^{R} p_r P(j|x_i,\tau_r)\right)^2 \tag{29}$$

subject to the constraint that $\sum_{r=1}^{R} p_r = 1$.

Back to the random coefficients logit model

- Further, for concreteness let's focus on the *binary logit model* where $P(x, \tau)$ is the probability of choosing an item given by

$$P(x, \tau) = \frac{1}{1 + \exp\{\tau_0 + x\tau_1\}} \tag{30}$$

  where x is the "attribute" of the item being chosen (think "price"), and we have the usual normalization that the utility of the "outside good" (i.e. not choosing the item) is normalized to 0.
- Unlike Fox *et. al.* we allow random coefficients for both intercept and slope. So $\tau = (\tau_0, \tau_1)$ is a 2×1 random vector that varies over individuals but is fixed for successive choices by the same individual.
- We approximate the CDF $G(\tau)$ for the random coefficients by a finite mixture

$$P(x, \theta) \equiv \int P(x, \tau) G(d\tau) \simeq \sum_{r=1}^{R} p_r P(x, \tau_r), \tag{31}$$

  where $\theta = (\tau_1, \ldots, \tau_R, p_1, \ldots, p_R)$.

Back to the random coefficients logit model

- For convenience I restrict $\{p_r\}$ to the $R-1$ dimensional simplex using the logit transformation to convert the estimation to an *unconstrained optimization* over $\gamma = (\gamma_1, \ldots, \gamma_{R-1})$ where

$$p_r(\gamma) = \exp\{\gamma_r\}/(1 + \sum_{r=1}^{R-1} \exp\{\gamma_r\}), \quad r = 1, \ldots, R \tag{32}$$

  and $p_R(\gamma) = 1/(1 + \sum_{r=1}^{R-1} \exp\{\gamma_r\})$.
- Given data $(y_1, \ldots, y_N, x_1, \ldots, x_N)$ we can estimate the parameters $\theta = (\tau_1, \ldots, \tau_R, \gamma_1, \ldots, \gamma_{R-1})$ by either *nonlinear least squares*

$$\hat{\theta}_N = \underset{\theta}{\mathit{argmin}} \sum_{i=1}^{N} (y_i - P(x_i, \theta))^2 , \tag{33}$$

  or *maximum likelihood*

$$\hat{\theta}_N = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} [y_i \log(P(x_i, \theta)) + (1 - y_i) \log(1 - P(x_i, \theta))] . \tag{34}$$

Asymptotic properites of NLLS estimator of G

- Let $\hat{G}_N(\tau) = \sum_{r=1}^{R} p_r(\hat{\gamma}) I\{\tau \leq \hat{\tau}_r\}$ be the NLLS estimator of G based on N *IID* observations. By the usual bias-variance decomposition, we can show that with probability 1 $\hat{G}_N \Longrightarrow G_R(G)$, where $G_R(G)$ is the *best approximation* of G using a finite mixture with R points, G_R, given by

$$G_R(G) = \underset{G_R}{argmin} \int \left[P(x, G_R) - P(x, G)\right]^2 F(dx) \equiv \delta(G, G_R), \quad (35)$$

  where $F(x)$ is the CDF for the covariates x.
- To see this, for any fixed G_R we can use the Law of Large Numbers to show that with probability 1

$$\frac{1}{N}\sum_{i=1}^{N}\left[y_i - P(x_i, G_R)\right]^2 = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - P(x_i, G) + P(x_i, G) - P(x_i, G_R)\right]^2$$

$$\rightarrow \int P(x, G)[1 - P(x, G)]F(dx) + \int \left[P(x, G) - P(x, G_R)\right]^2 F(dx),$$

  (i.e. "variance" plus "bias squared").

Asymptotic properites of MLE estimator of G

- Similarly, following White (1982) "Maximum likelihood estimation of misspecified models" the MLE $\hat{G}_N$ converges to $G_R(G)$ that minimizes the *Kullback-Leibler distance* between the true conditional probability $f(y|x, G) = yP(x, G) + (1 - y)[1 - P(x, G))]$ and its approximation via a finite mixture of G, $f(y|x, G_R) = yP(x, G_R) + (1 - y)[1 - P(x, G_R)]$. Denote this distance as $KL(G, G_R)$, given by

$$KL(G, G_R) = -\int \left[P(x, G) \log \left(\frac{P(x, G_R)}{P(x, G)} \right) \right] F(dx)$$
$$- \int \left[[1 - P(x, G)] \log \left(\frac{1 - P(x, G)}{1 - P(x, G_R)} \right) \right] F(dx).$$

- Thus, asymptotically, maximum likelihood converges to a $G_R(G)$ that minimizes the Kullback Liebler distance $KL(G, G_R)$ between G and G_R and nonlinear least squares converges to a $G_R(G)$ that minimizes the "L^2 distance" $\delta(G, G_R)$ between G and G_R, i.e. the mean squared error between the true CCP $P(x, G)$ and the finite mixture approximation to it, $P(x, G_R)$.

Identification conditions for G

- Under weak assumptions similar to Fox *et. al.* we can show identification of G using these "metrics", i.e.

$$\begin{aligned} KL(G, G') &= 0 \quad \text{implies} \quad G = G' \\ \delta(G, G') &= 0 \quad \text{implies} \quad G = G' \end{aligned} \tag{36}$$

- Thus, by finding a finite mixture model $G_R(G)$ that minimizesthe L^2 norm $\delta(G, G_R)$ or the Kullback-Leibler distance $KL(G, G_R)$ we can obtain a consistent estimator of G.
- We now show, using recently derived error bounds for approximation of certain classes of functions by neural networks, that the approximation errors $\delta(G, G_R)$ and $KL(G, G_R)$ decrease to zero *very rapidly* with the number of types, R.
- This means we can approximate the true CCP $P(x, G)$ by a finite mixture model $P(x, G_R)$ with a small number of types R, thereby explaining the "Heckman Singer puzzle" that in most empirical work, model selection results in very small values of R typically $R \in \{2, 3, 4\}$.

Strong identification of P, weak identification of G

- This implies that we can achieve *strong identification* of the CCP $P(x, G)$ in the sense that we can estimate it at nearly $O_p(1/\sqrt{N})$ rates, and the estimator *breaks the curse of dimensionality* in the sense that this rate of convergence does not depend on the dimension k of the vector τ of random coefficients.
- At the same time, we will show that G_R itself converges *very slowly* to G, explaining the the other paradoxical finding from Heckman-Singer, 1984. There is likely to be a curse of dimensionality for G, so that the rate of convergence decreases with k, though at this point this is still just a conjecture on our part.
- So in this sense, one can say that G is poorly identified, or weakly identified. "Weak identification commonly refers to the failure of classical asymptotics to provide a good approximation to the finite sample distribution of estimates" (Andrews and Mikusheva). Weak identification can also be defined as *very slow convergence of an estimator to the true parameter.*
- I will demonstrate these results both with theory and some numerical calculations.

Finite mixture models are neural networks

- Recall the definition of a *single layer feedforward neural network.* It is a parametric function $f(x, \theta)$ defined by

$$f(x, \theta) = w_0 + \sum_{r=1}^{R} w_r \sigma(\omega_{0,r} + x\omega_{1,r}), \tag{37}$$

  where σ is a sigmoidal *squashing function,* R is the number of *hidden units,* $\theta = (w_0, w_1, \ldots, w_R, \omega_{O,1}, \ldots, \omega_{0,R}, \omega_{1,1}, \ldots, \omega_{1,R})$, where w_0 is an *output bias* and $\{w_r\}$ are *output weights* and $\{\omega_{0,r}\}$ are *input biases* and $\{\omega_{1,r}\}$ are *input weights.*
- The finite mixture approximation $P(x, G_R)$ to $P(x, G)$ is a neural network where $w_0 = 0$, the input weights $\{w_r\}$ are restricted to the $R-1$-dimensional simplex (i.e. they are positive and sum to 1), and $\sigma(y) = 1/(1 + \exp\{y\})$ is the *logistic squashing function.*

Universal approximation theorem

For any continuous function $f(x)$ and $\epsilon > 0$, there exists an R and corresponding parameter θ such that $\sup_x |f(x, \theta) - f(x)| < \epsilon$.

When do neural networks break the curse of dimensionality?

- For the most general class of continuous functions on a compact domain X, neural networks exhibit a *curse of dimensionality* that is, the number of hidden units R needed to product an ϵ-approximation to any f in the worst case grows like $R(\epsilon) = O(\epsilon^{-d})$. Thus the number of hidden units (and thus total parameters of the neural network) "blows up" at an exponential rate in the dimension of x, d.
- However starting with Barron (1993) approximation bounds have been derived for special subclasses of functions $f \in \mathcal{F}$ that break the curse of dimensionality in the sense that the approximation error is $O(R^{-M})$ for some integer M independent of d.
- Mhaskar (2020) provides such a bound relevant for our case in "Dimension Independent Bounds for General Shallow Networks" *Neural Networks*
- Mhaskar derives bounds for the class of functions $\mathcal{F}$ of the form

$$f(x) = \int P(x, \tau)\mu(d\tau), \tag{38}$$

where μ is a signed measure over a compact subset of R^d.

Mhaskar's approximation bound

- Consider approximating $f \in \mathcal{F}$ by neural networks of the form

$$f(x, \theta) = \sum_{r=1}^{R} w_r G(x, \tau_r), \tag{39}$$

  where $\theta = (w_1, \ldots, w_R, \tau_1, \ldots, \tau_R)$, but he does not assume the weights are probabilities.
- Note that approximating mixed logits is a special case of Mhaskar's bound where $\mu = G$ is a *probability measure* and $\{w_r\}$ are *probability weights* and $G(x, \tau) = 1/(1 + exp(x\tau))$.

Mhaskar's Bound

For any $f \in \mathcal{F}$ there is a number of hidden units $R < \infty$ and parameter θ (which potentially can depend on f) and an absolute constant $c > 0$ such that for any positive integer $S > 0$ we have

$$\|f - \sum_{r=1}^{R} w_r G(x, \tau_r)\| \leq c \frac{\sqrt{\log(R)}}{R^S} \|\mu\|_{TV}. \tag{40}$$

Implication of Mhaskar's approximation bound

- It solves the puzzle noted widely in empirical work that *we can get very good fits for mixture models using only a small number of types, R.* In our case, we are trying to approximate $f(x) = P(x, G)$, the mixed CCP of the random coefficients logit model.
- We think Mhaskar's bound may apply to a much wider class of structural models where the parameters τ do not interact with the covariates x in a "linear-in-parameters" fashion.
- But for now, note that in Mhaskar's bound, $\mu = G$ is a probability measure, so $\|\mu\|_{TV} = \|G\|_{TV} = 1$. Thus, there is a finite mixture model with only a small number of types R that provides a close uniform approximation to the mixed CCP $P(x, G)$.
- However Mhaskar did not study the question of whether/how fast the approximate distribution of types G_R approaches the true one G in the uniform (or total variation norm).
- Using numerical calculations we show that $P(x, G_R)$ closely approximates $P(x, G)$ for small values of R even though the implied G_R is quite far from G in the total variation norm.

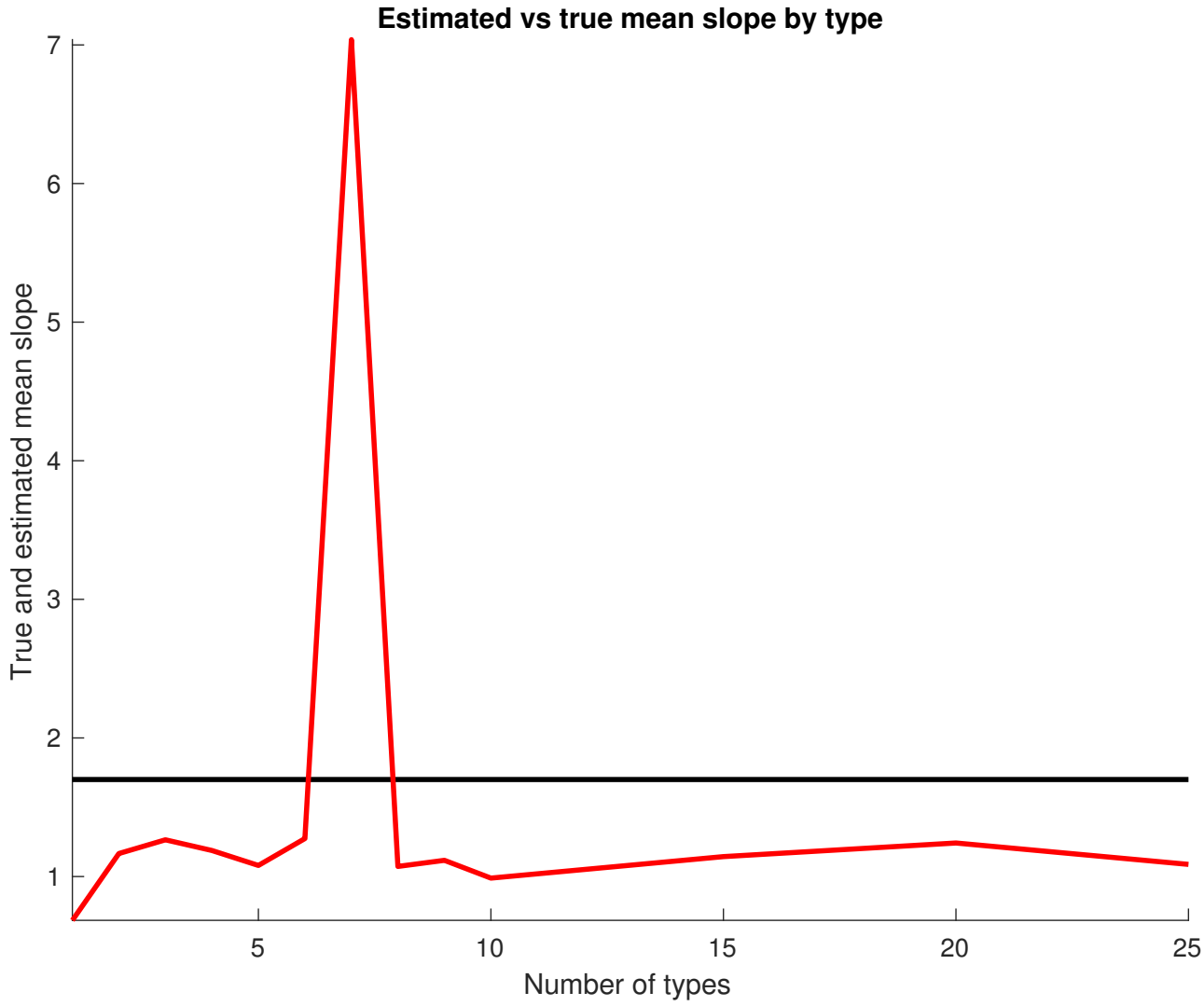
Intuition for Mhaskar's approximation bound

- The approximation $P(x, G_R)$ to $P(x, G)$ can be viewed as a convex combination of "basis functions" $(P(x, \tau_1), \ldots, P(x, \tau_R))$ which are each logistic.
- The logistic function is quite flexible in the shapes it can take as we vary the coefficients $\tau \in R^k$.
- In estimating the finite mixture model, the estimation algorithm is essentially *endogenously choosing a "best basis" for approximating* $P(x, G)$
- Then the resulting estimate $P(x, \hat{G}_R)$ is just a linear combination (actually a *convex combination*) of these optimally chosen set of basis functions $\{P(x, \hat{\tau}_1), \ldots, P(x, \hat{\tau}_R)$,

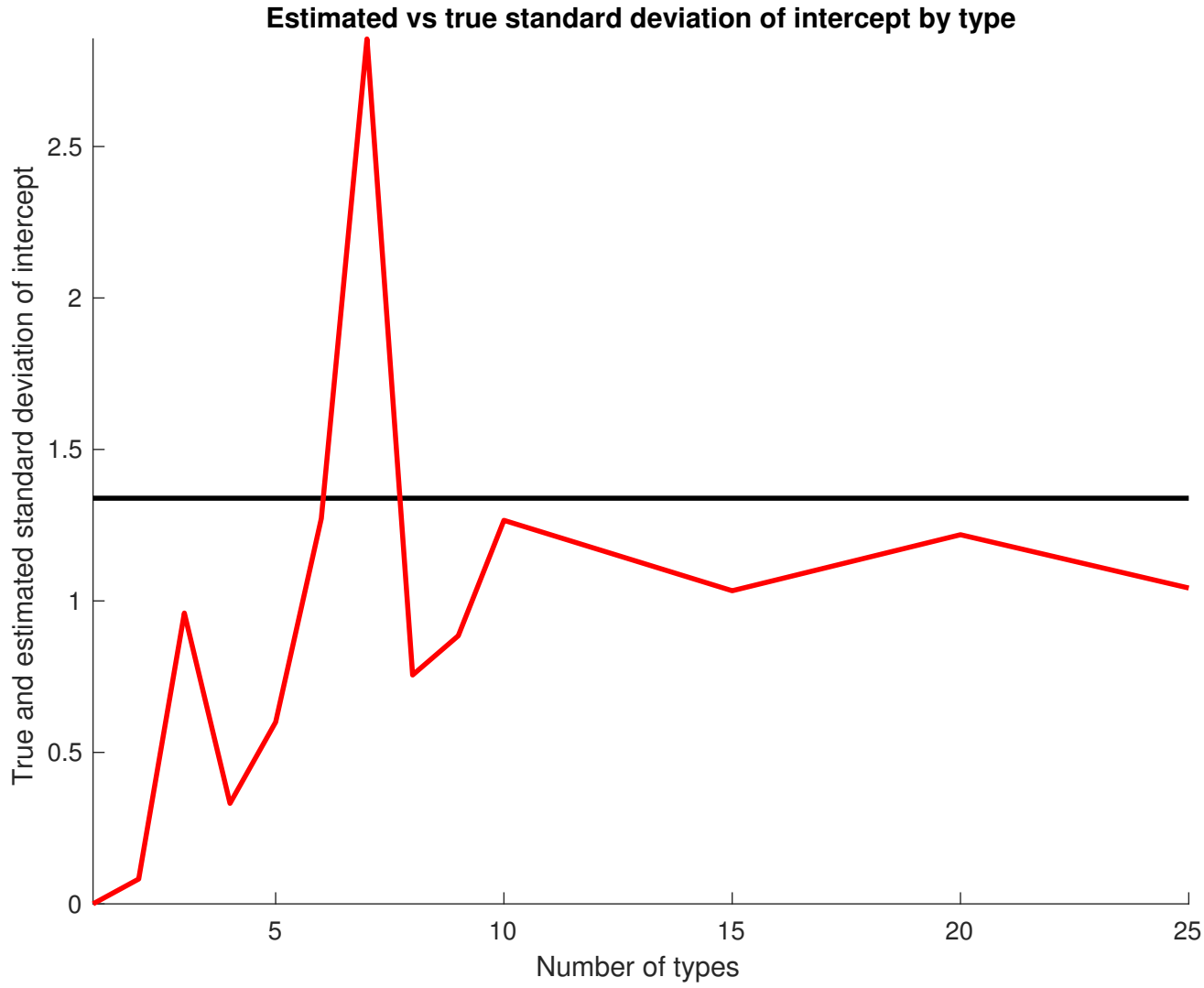
$$P(x, \hat{G}_R) = \sum_{r=1}^{R} \hat{p}_r P(x, \hat{\tau}_r). \tag{41}$$

- Because this basis is so well chosen, we only need a small number R of these "basis functions" to approximate $P(x, G)$ very well, uniformly over x.

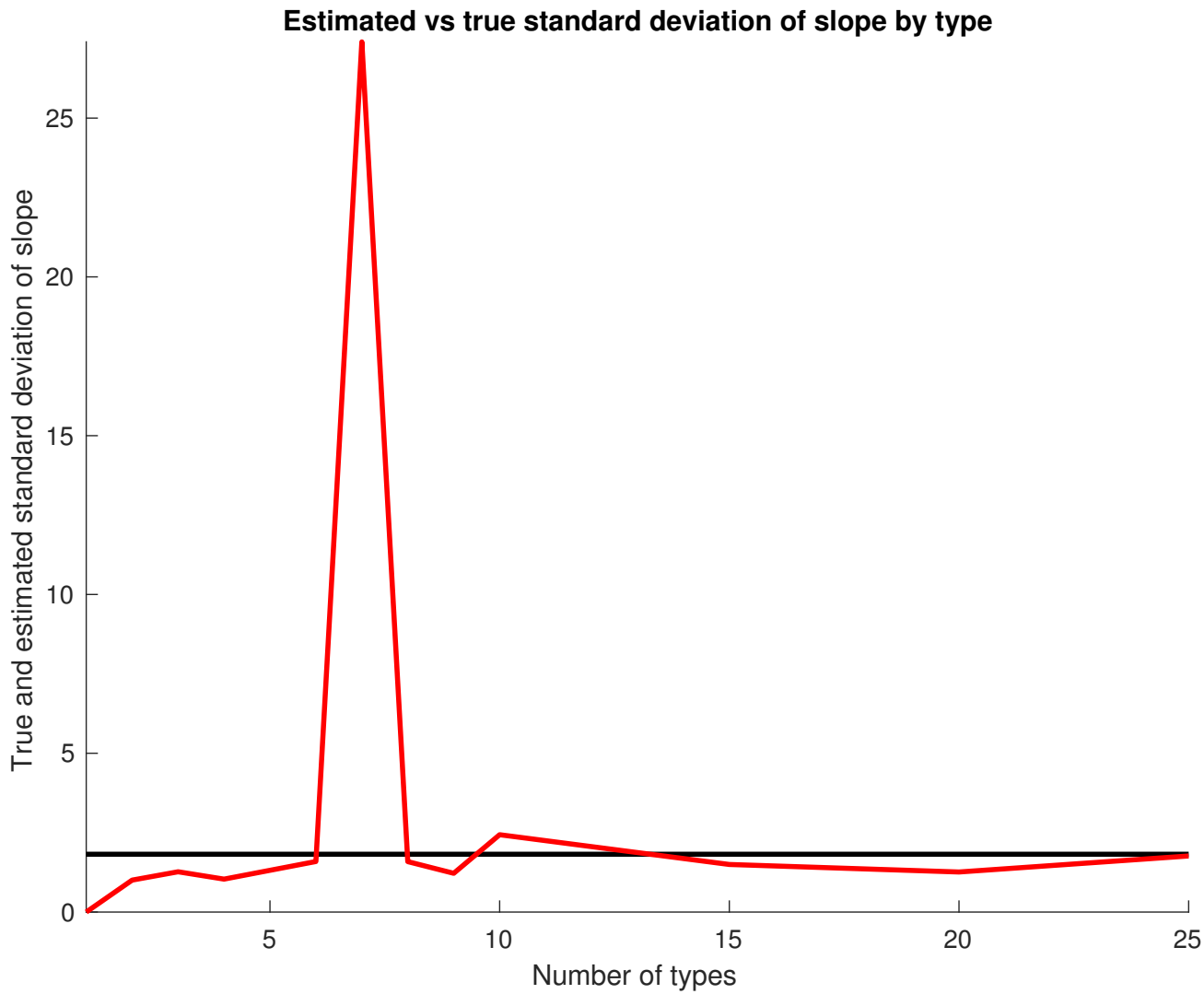
True vs approximated Mean x-coefficient by number of types



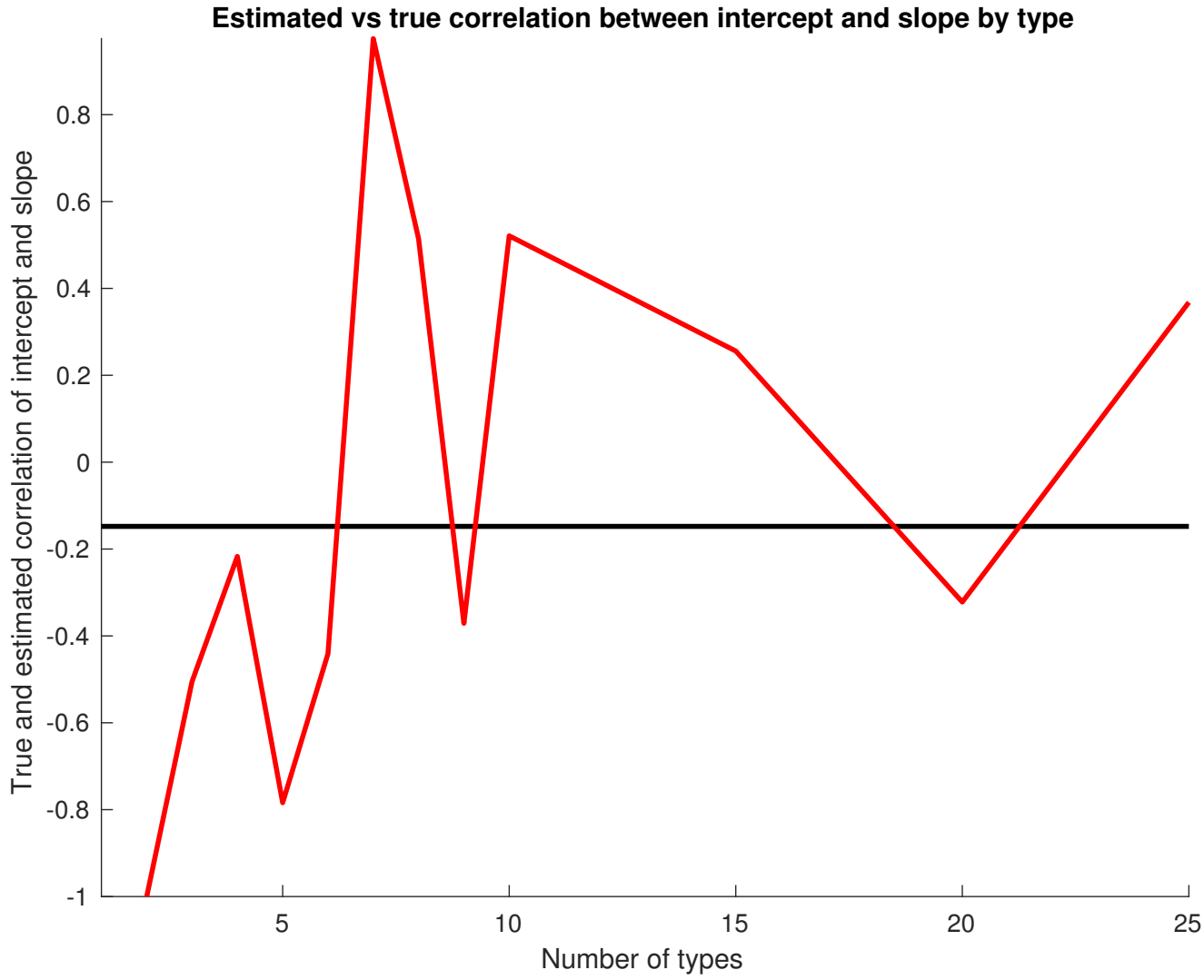
True vs approximated Std of Intercept by number of types



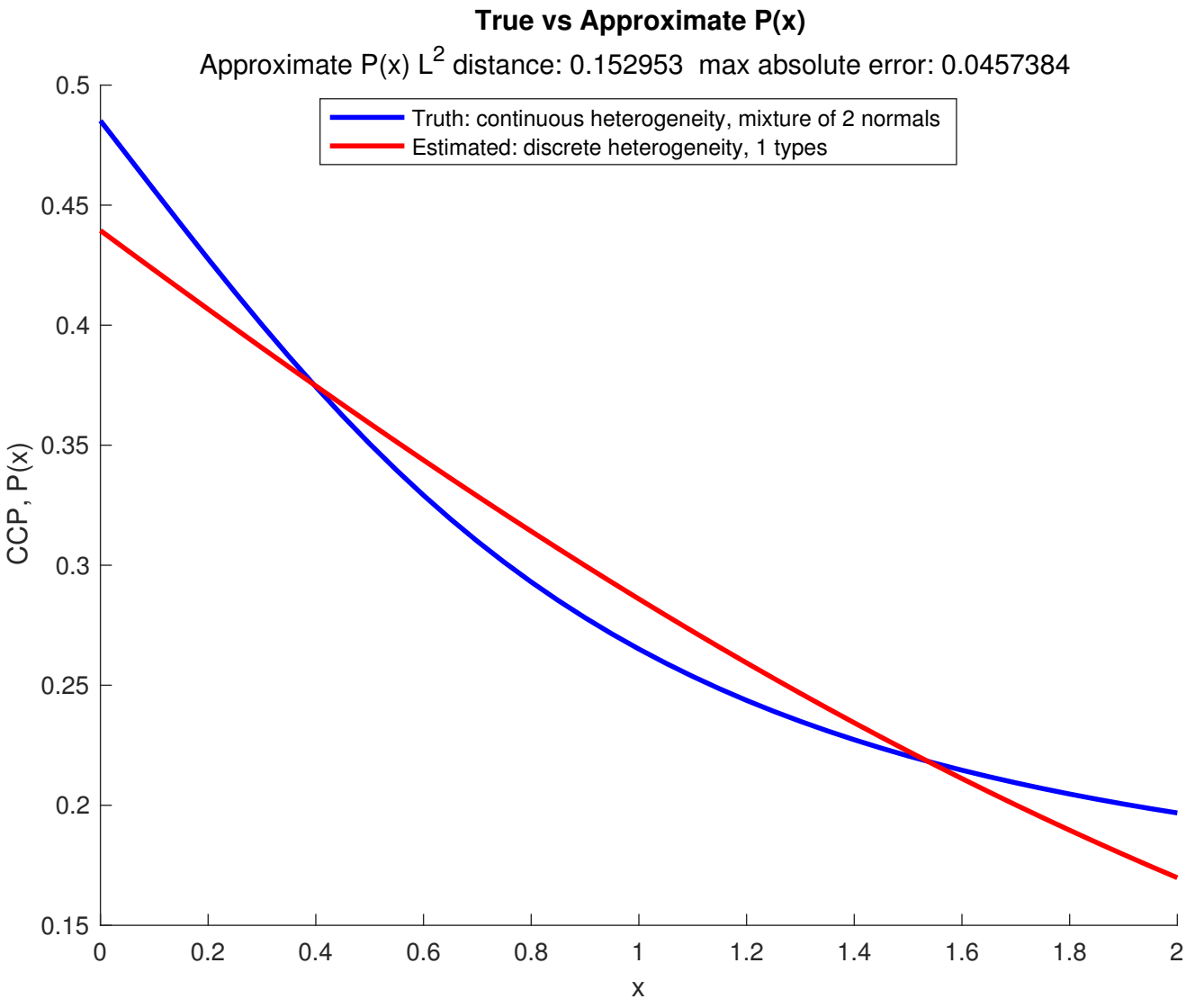
True vs approximated Std of x-coefficient by number of types



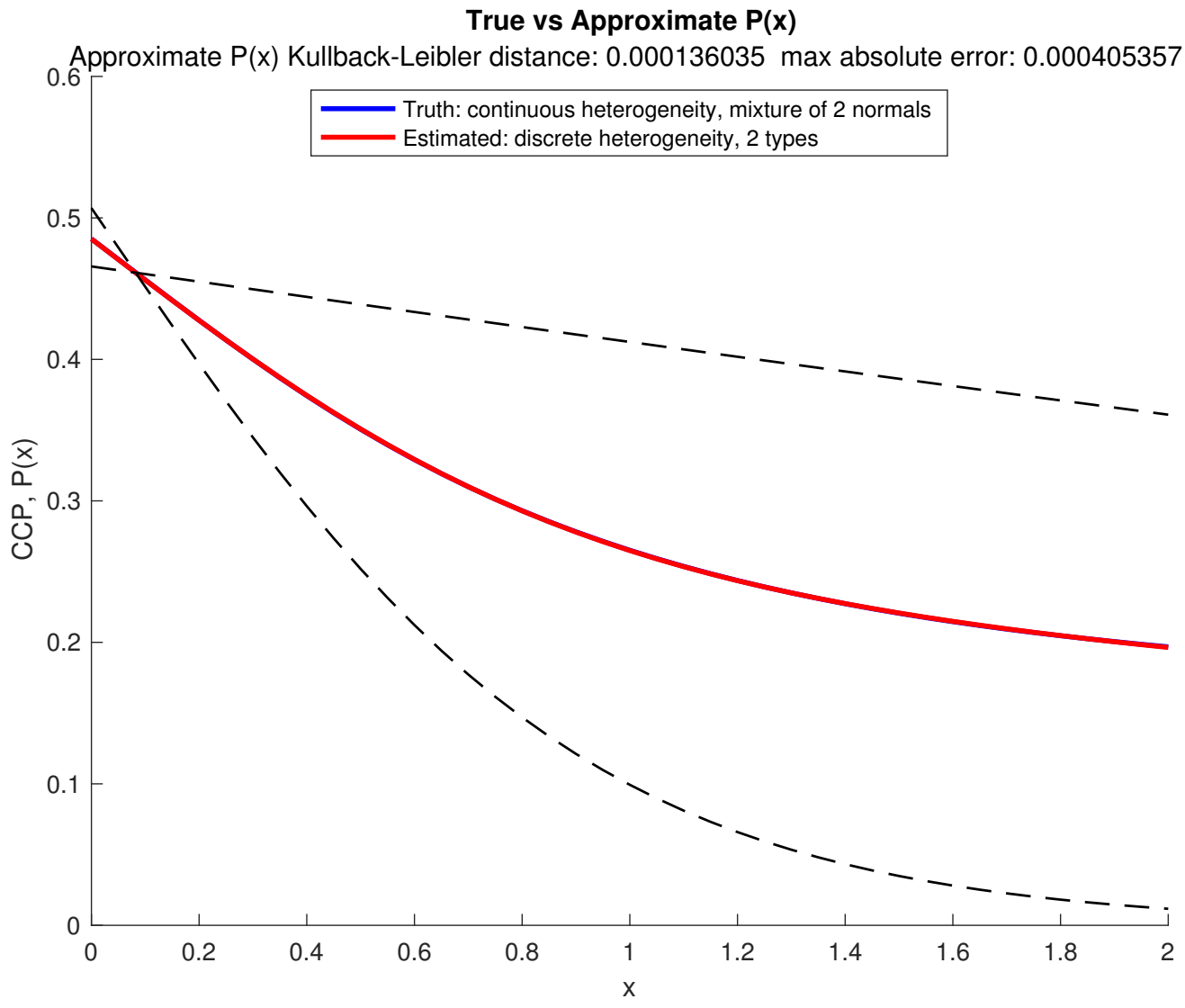
True vs approximated intercept/slope correlation by types



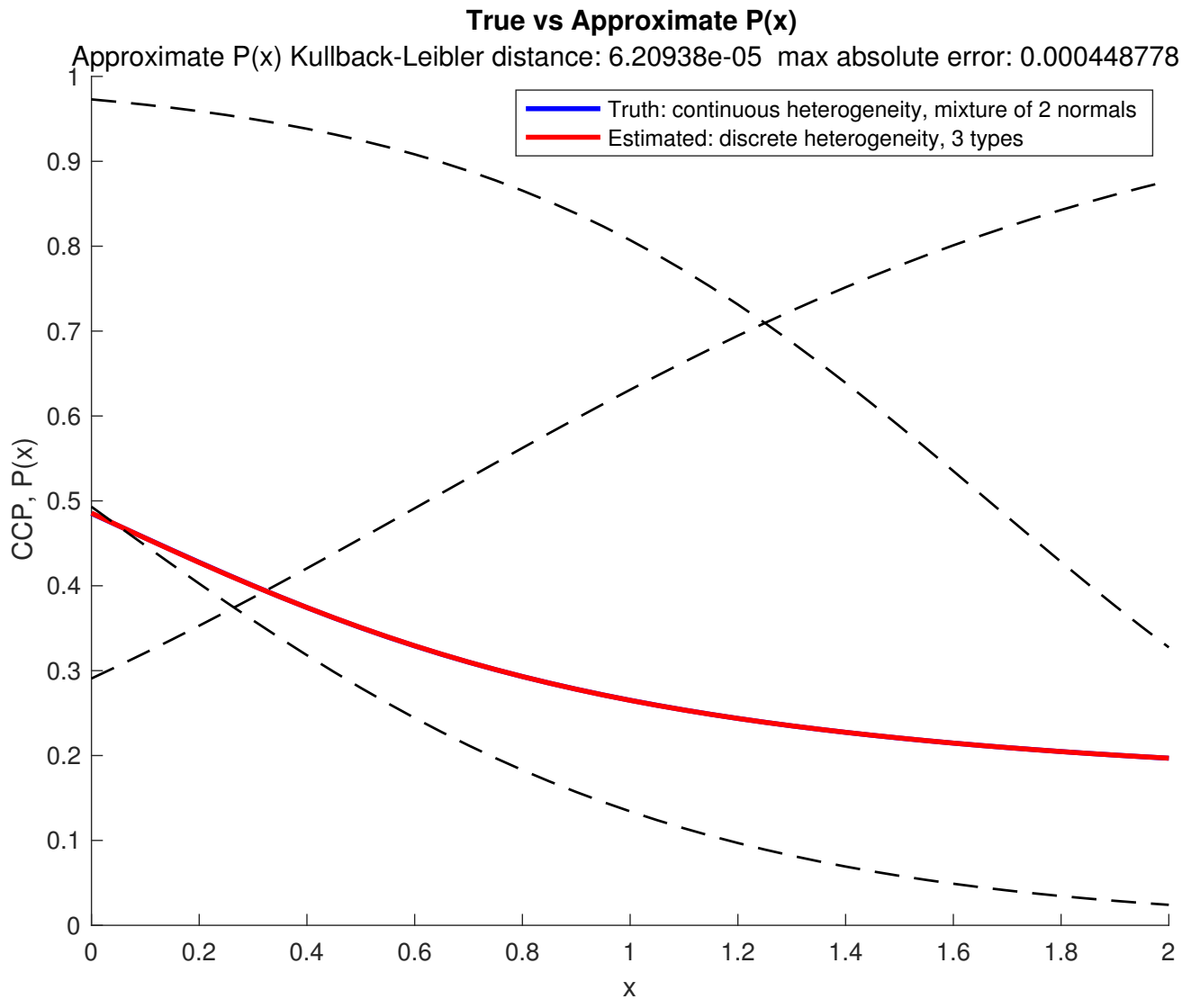
True vs approximate CCP, 1 type



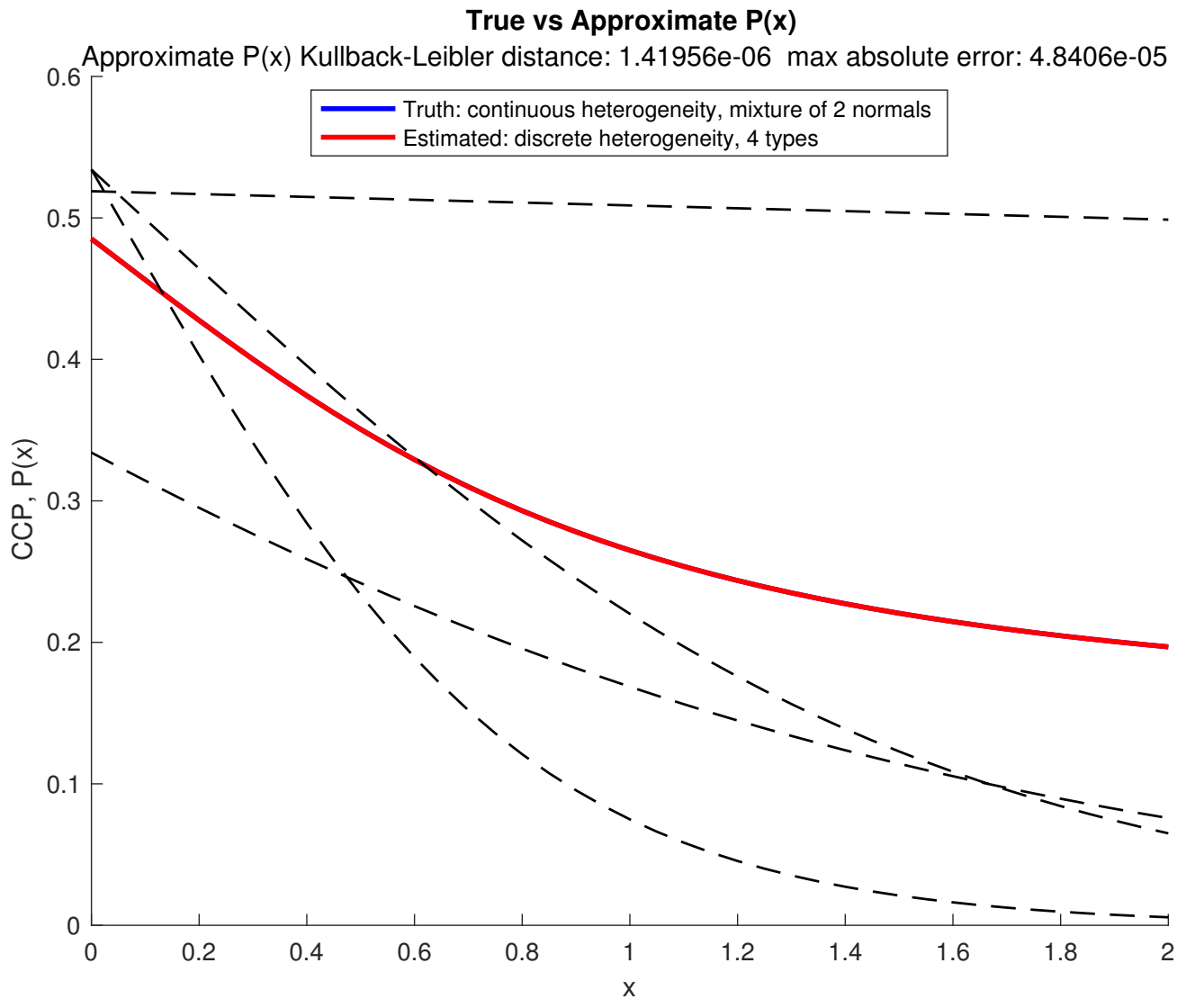
True vs approximate CCP, 2 types



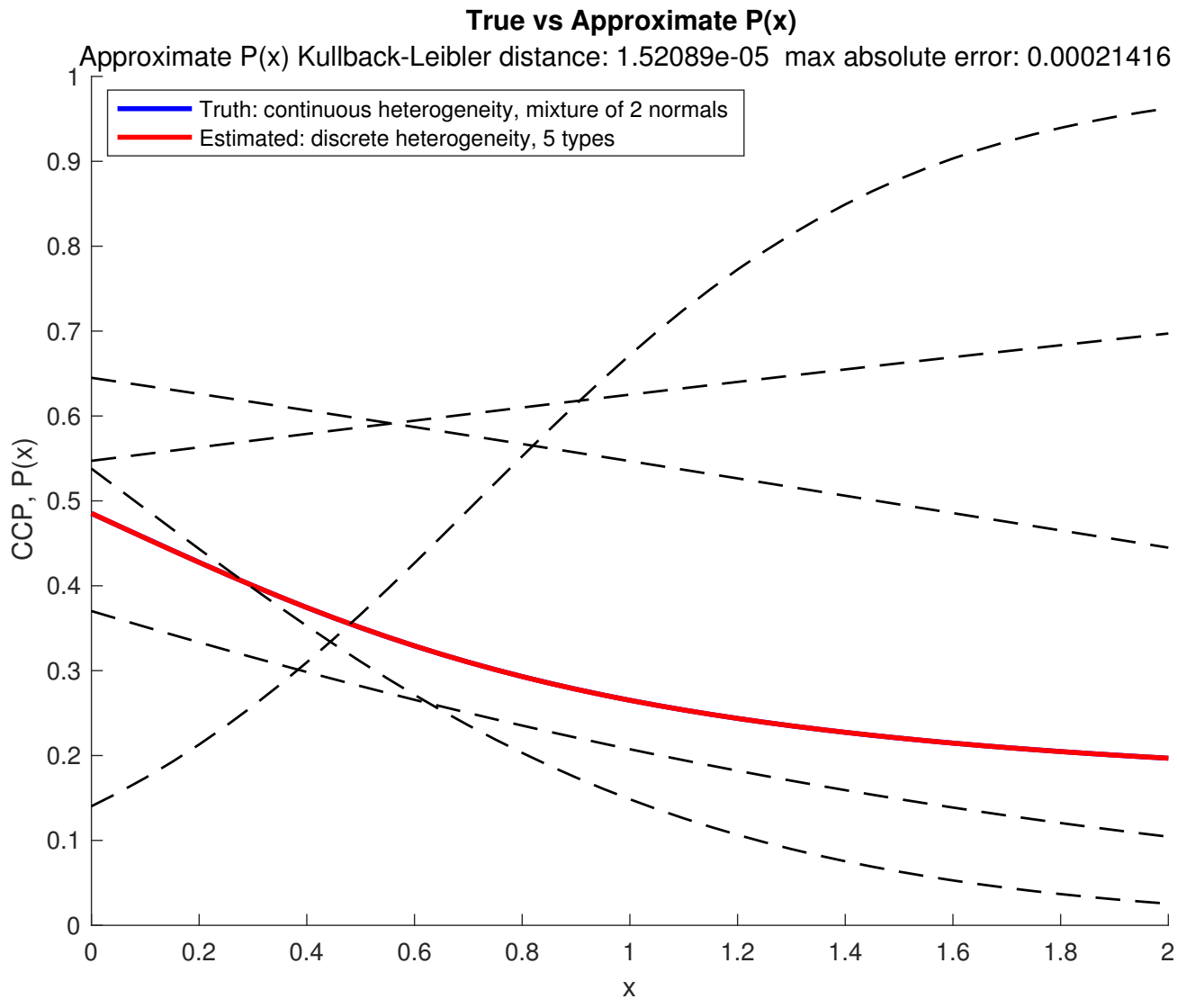
True vs approximate CCP, 3 types



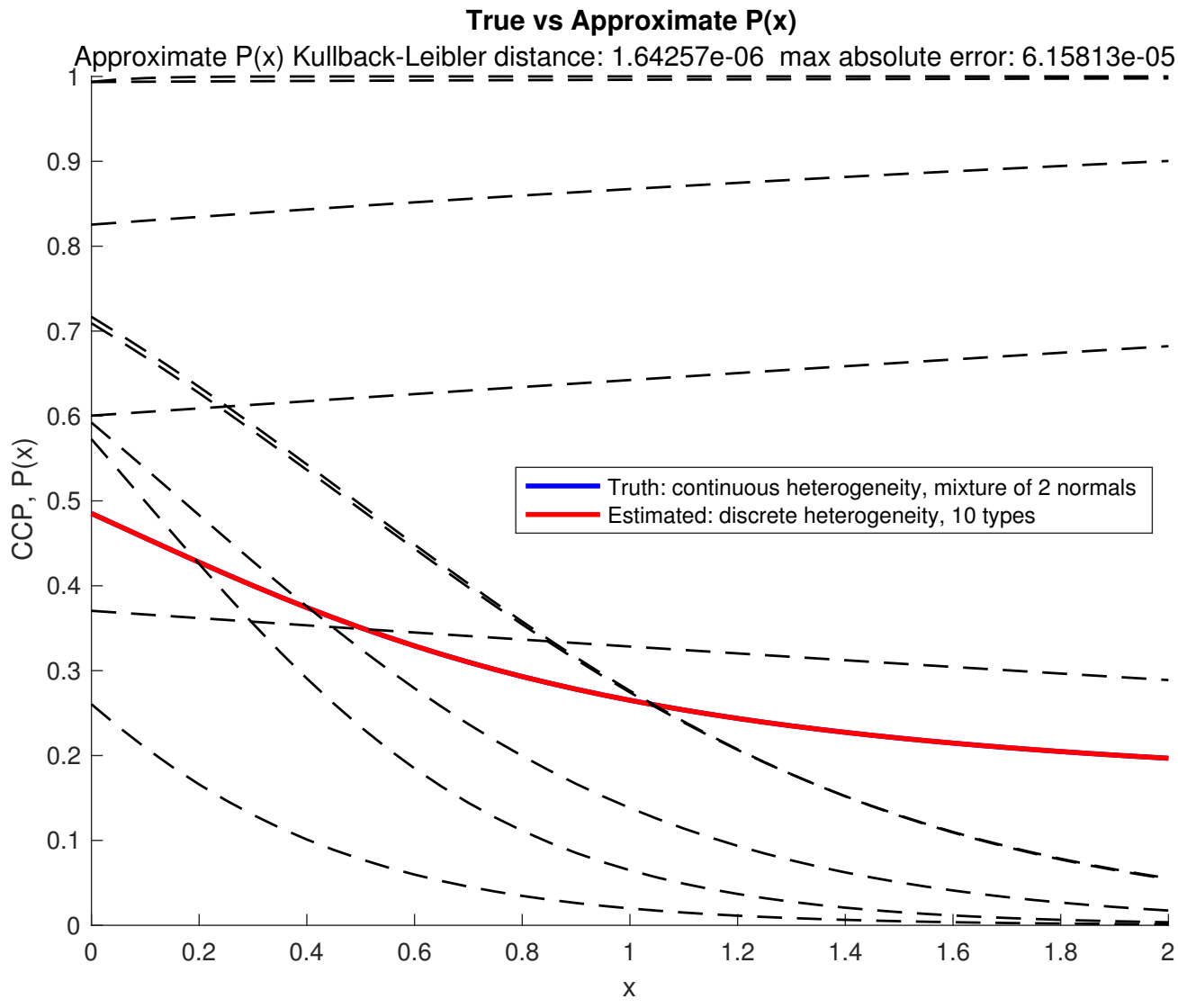
True vs approximate CCP, 4 types



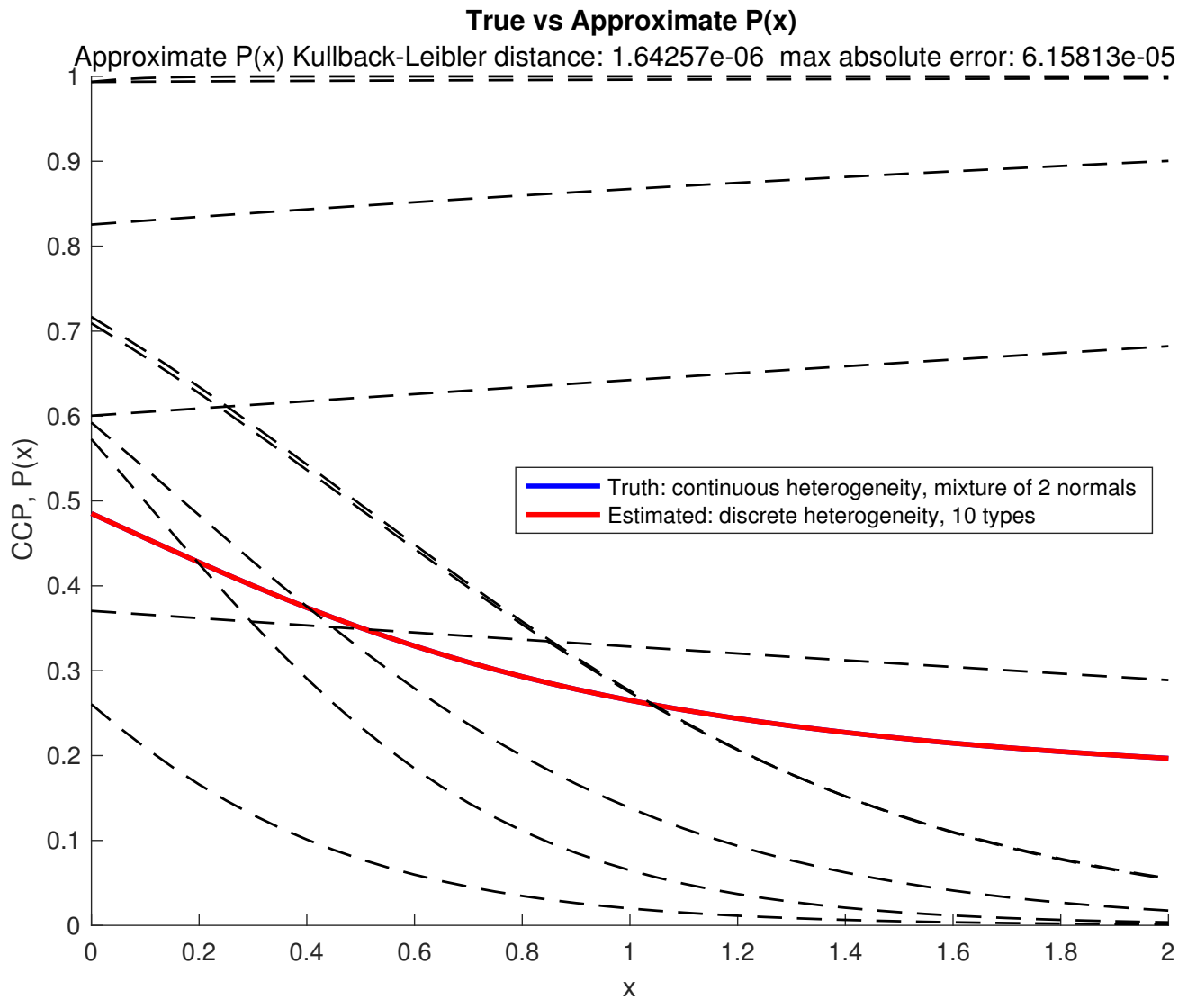
True vs approximate CCP, 5 types



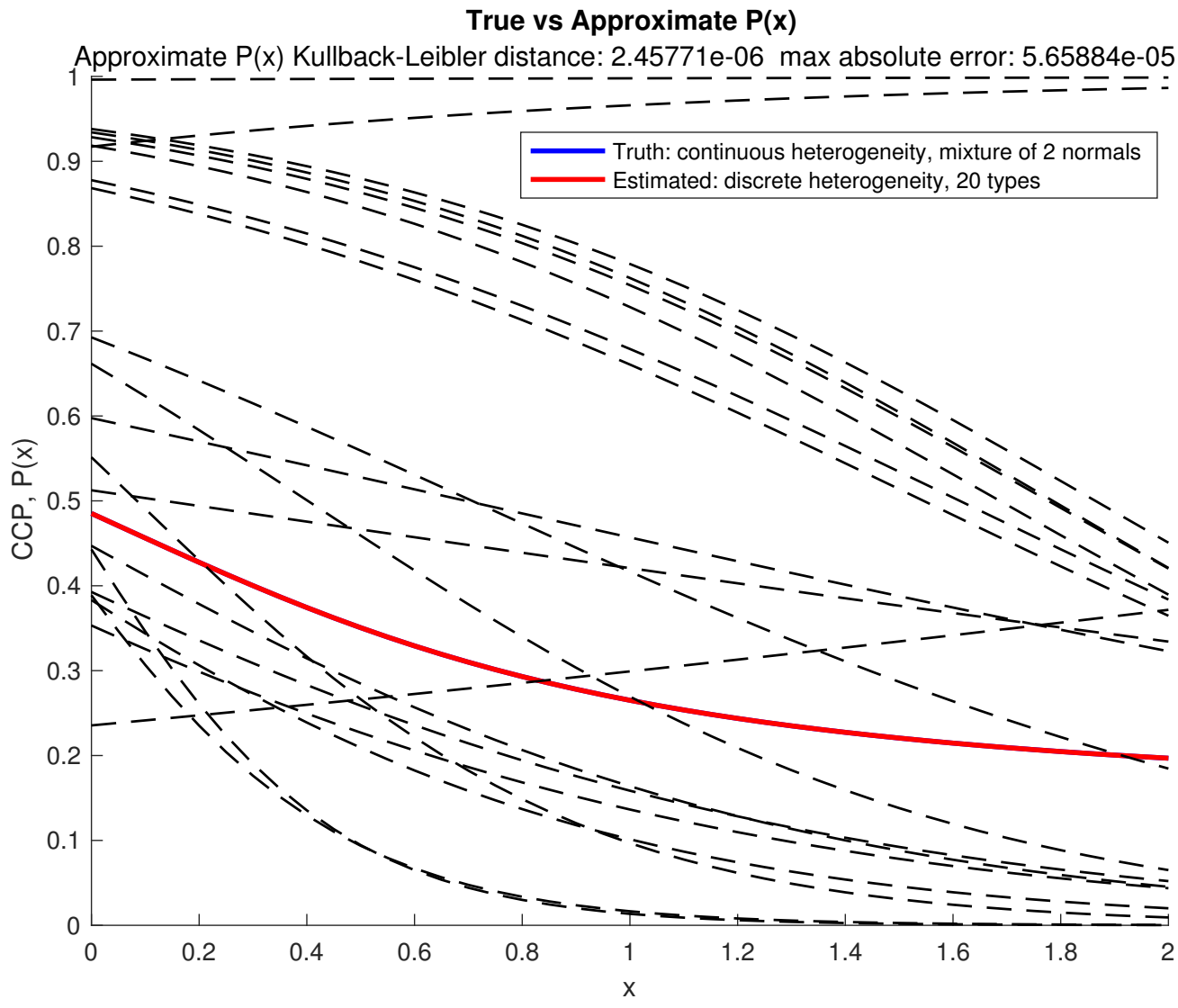
True vs approximate CCP, 10 types



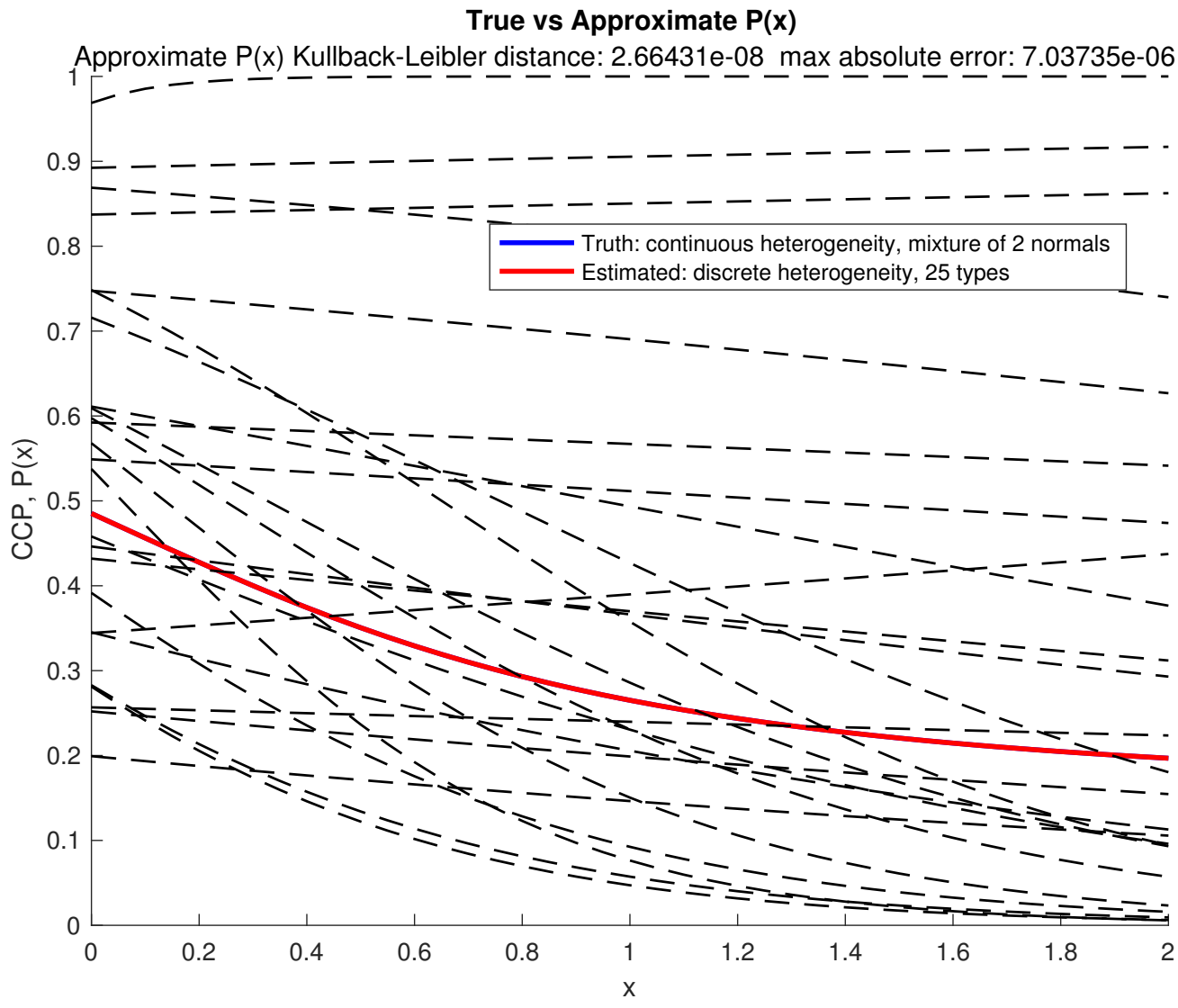
True vs approximate CCP, 15 types



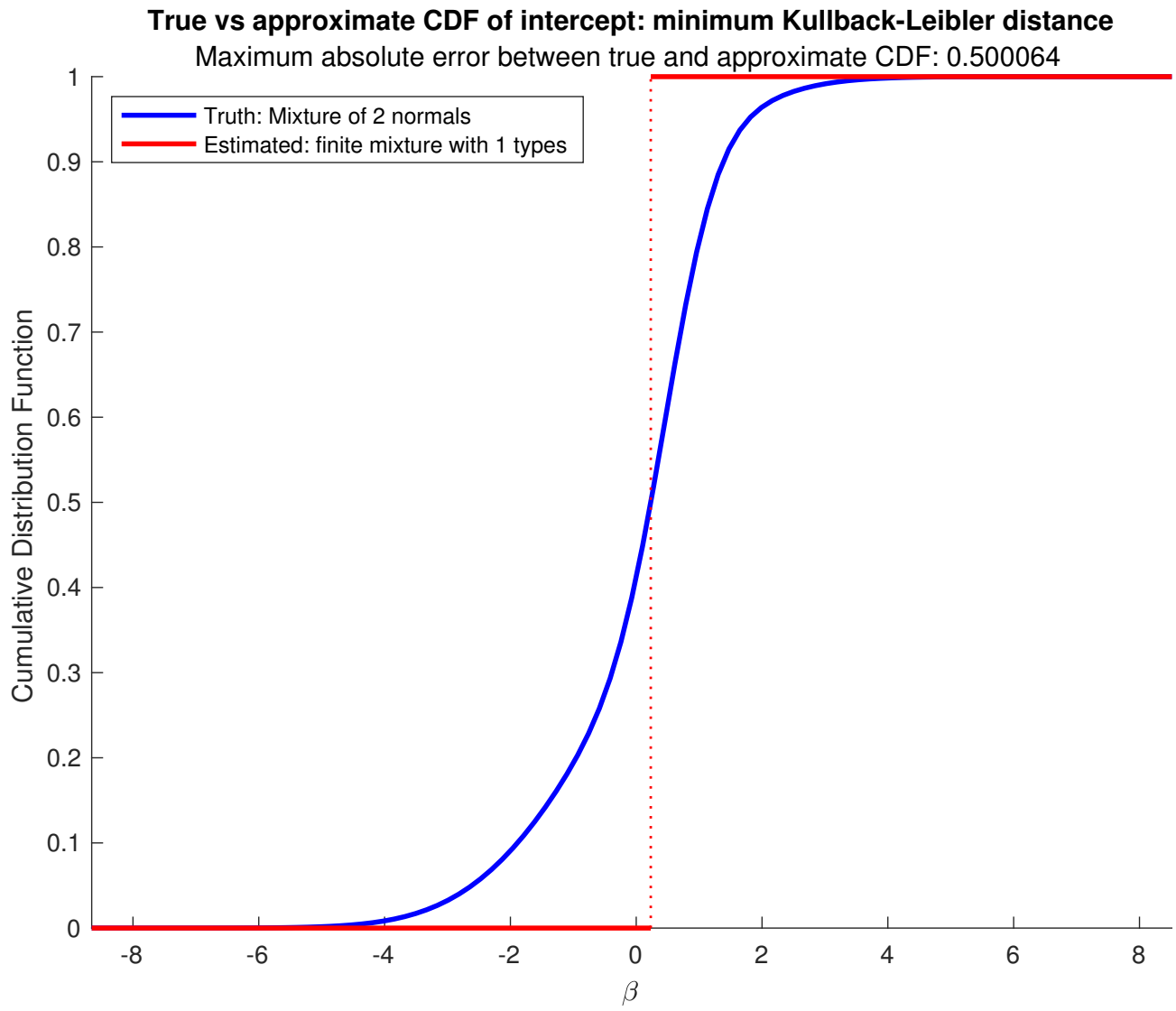
True vs approximate CCP, 20 types



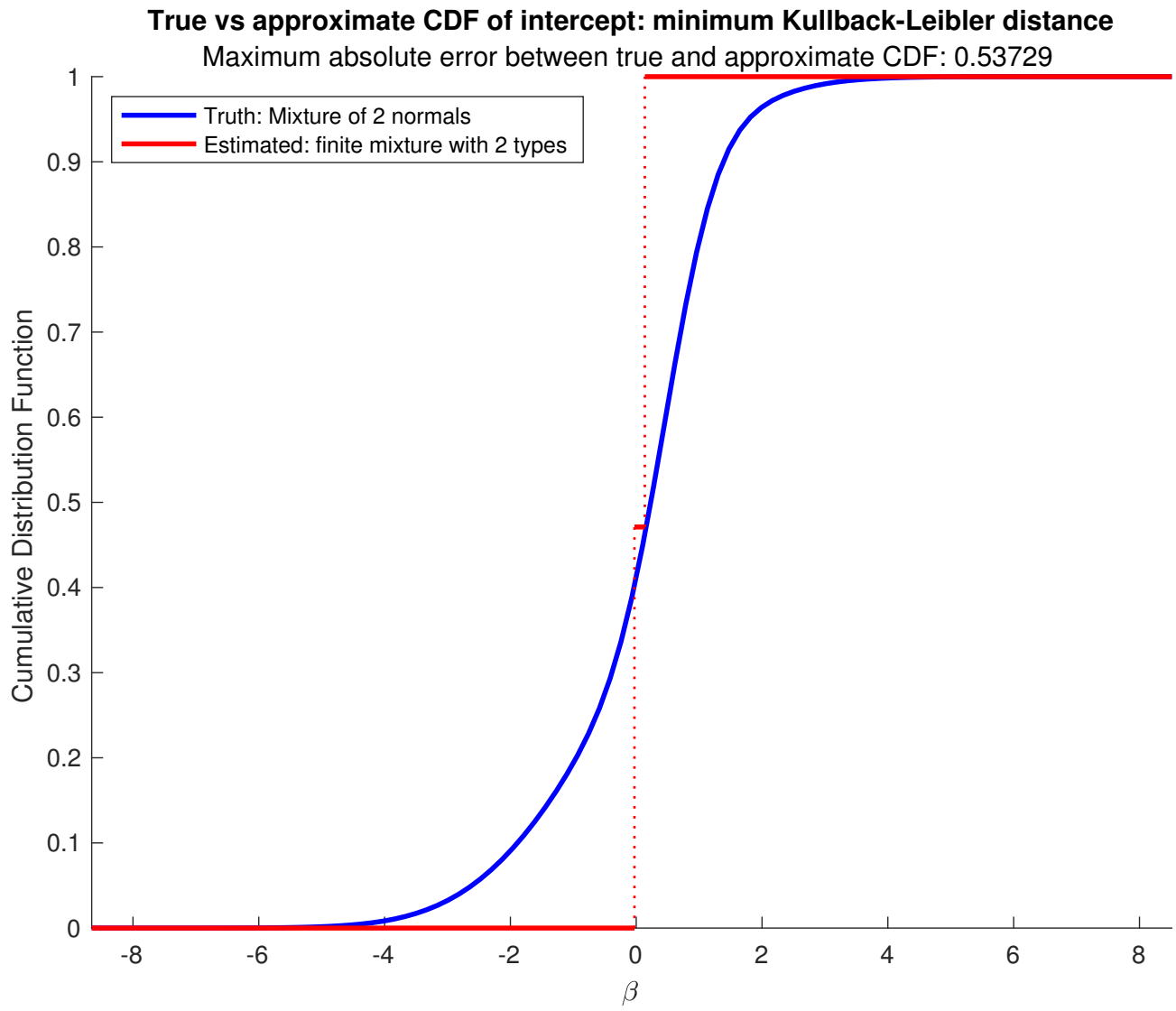
True vs approximate CCP, 25 types



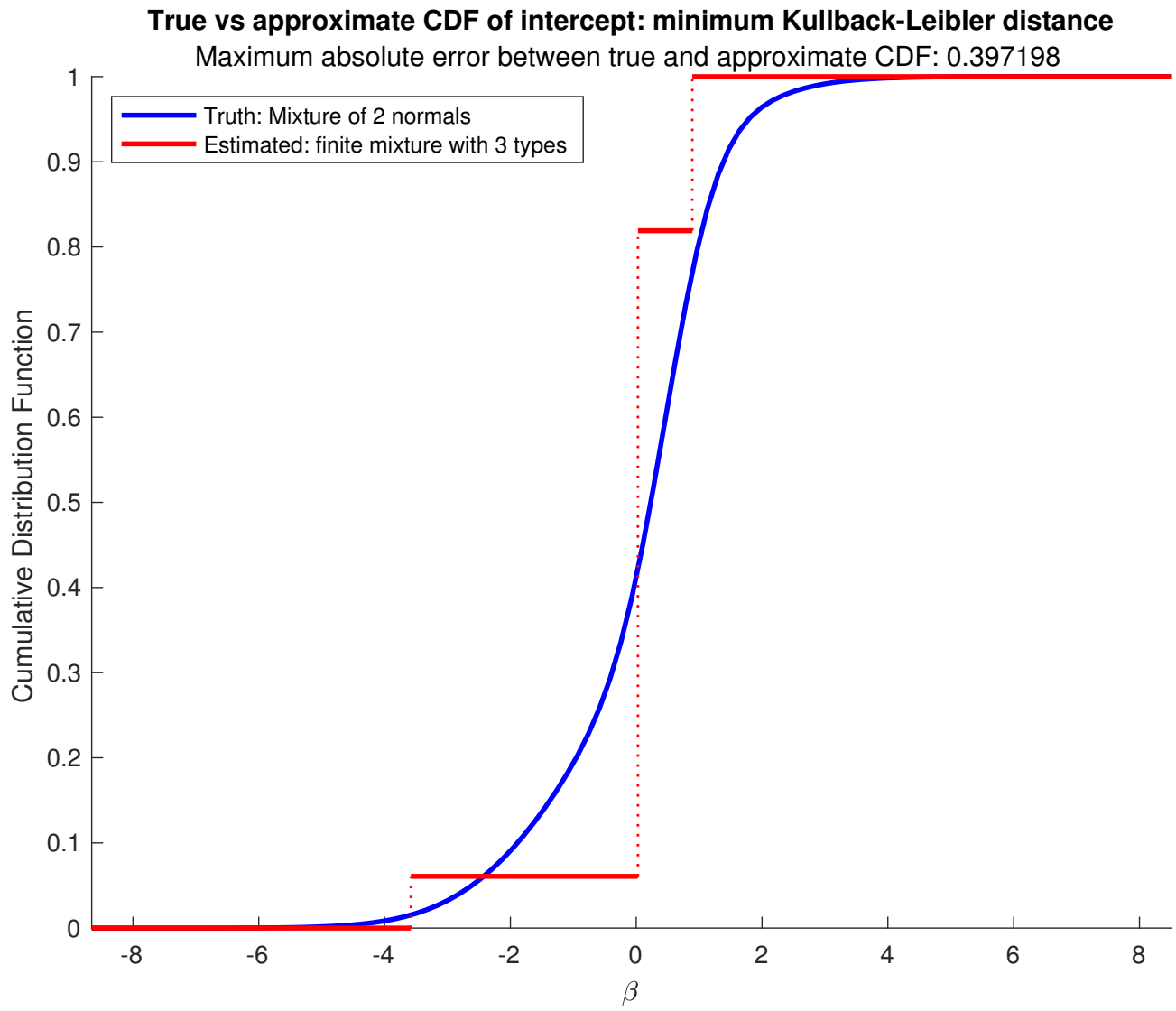
True vs approximate intercept CDF, 1 type



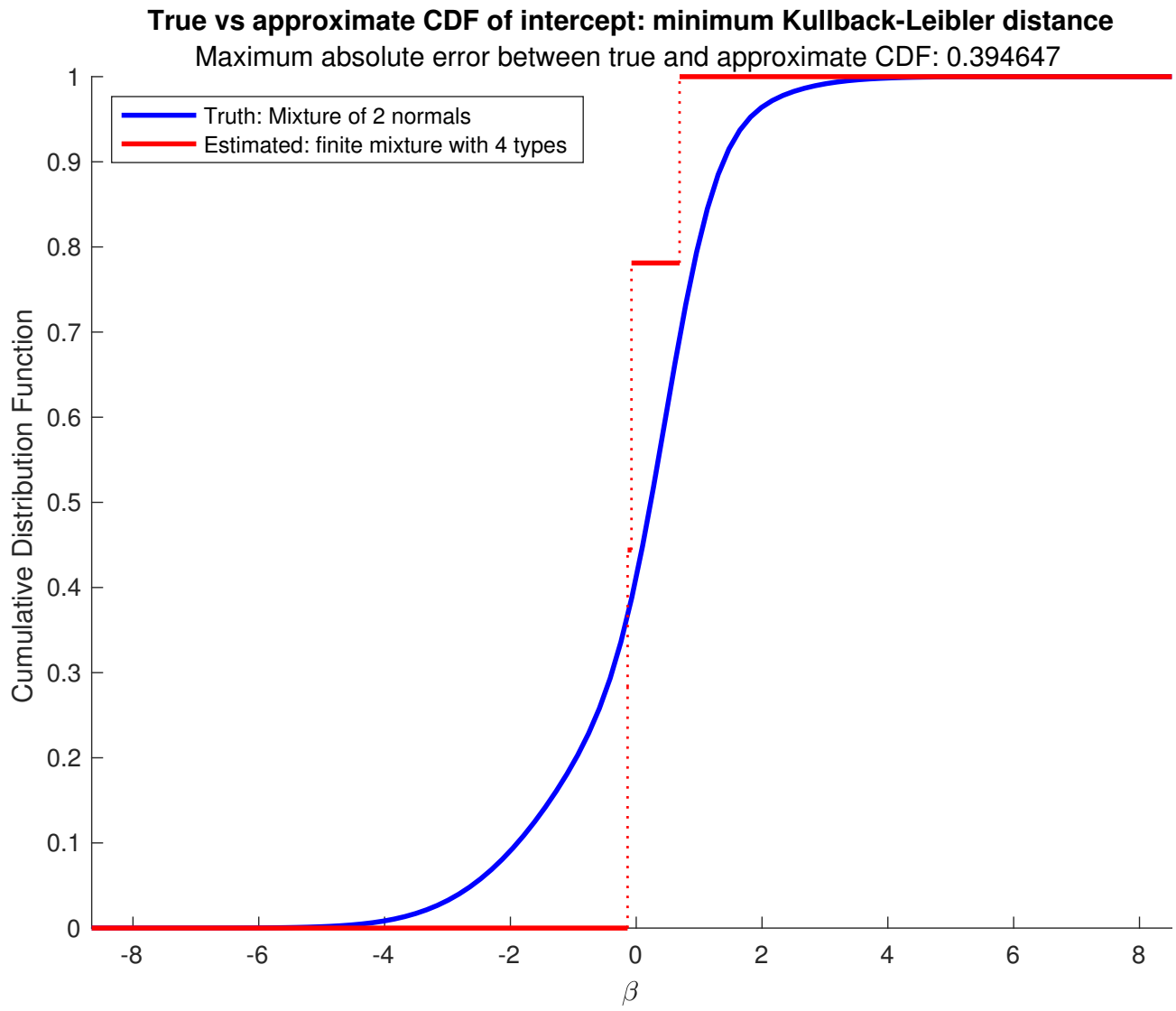
True vs approximate intercept CDF, 2 types



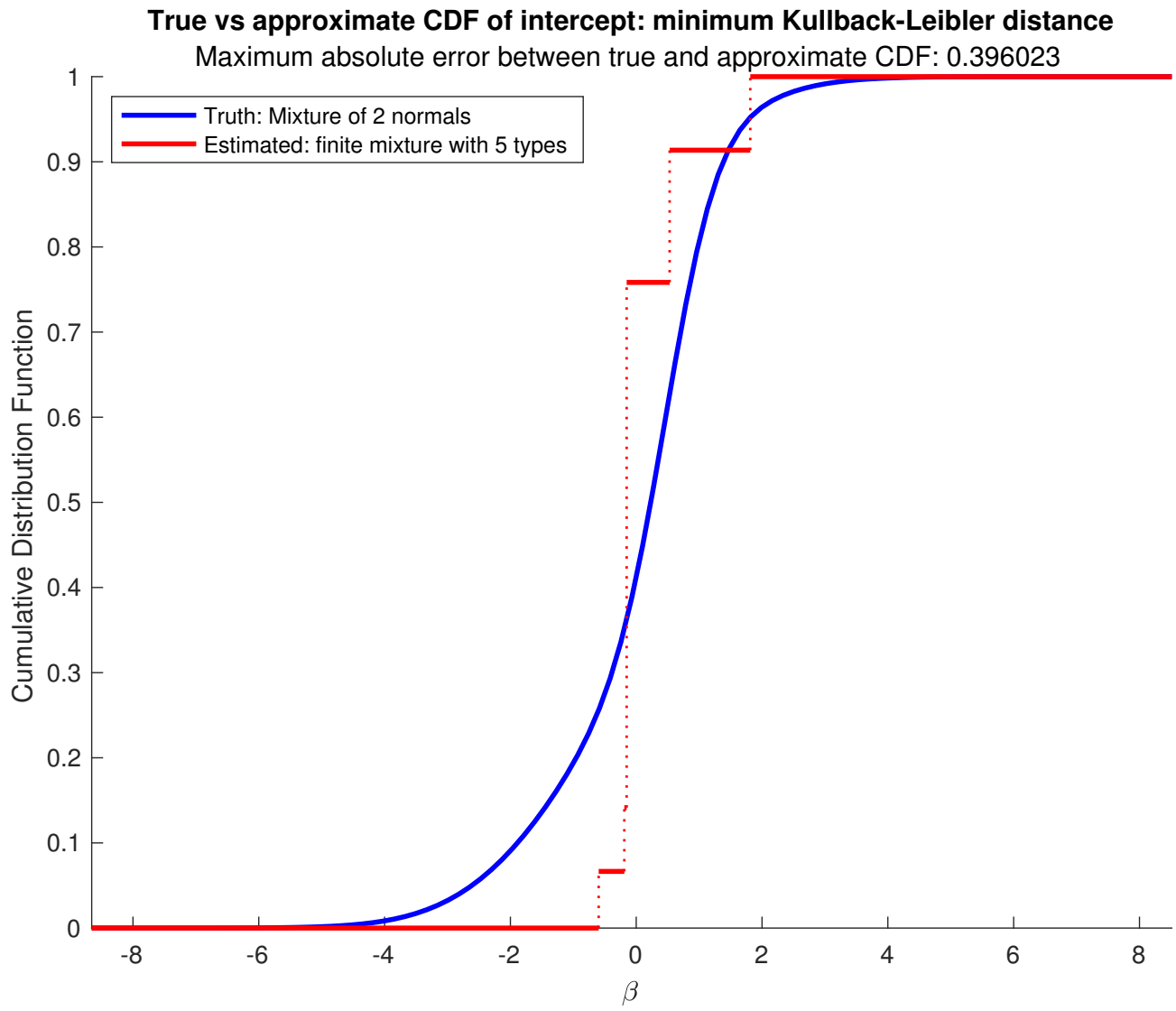
True vs approximate intercept CDF, 3 types



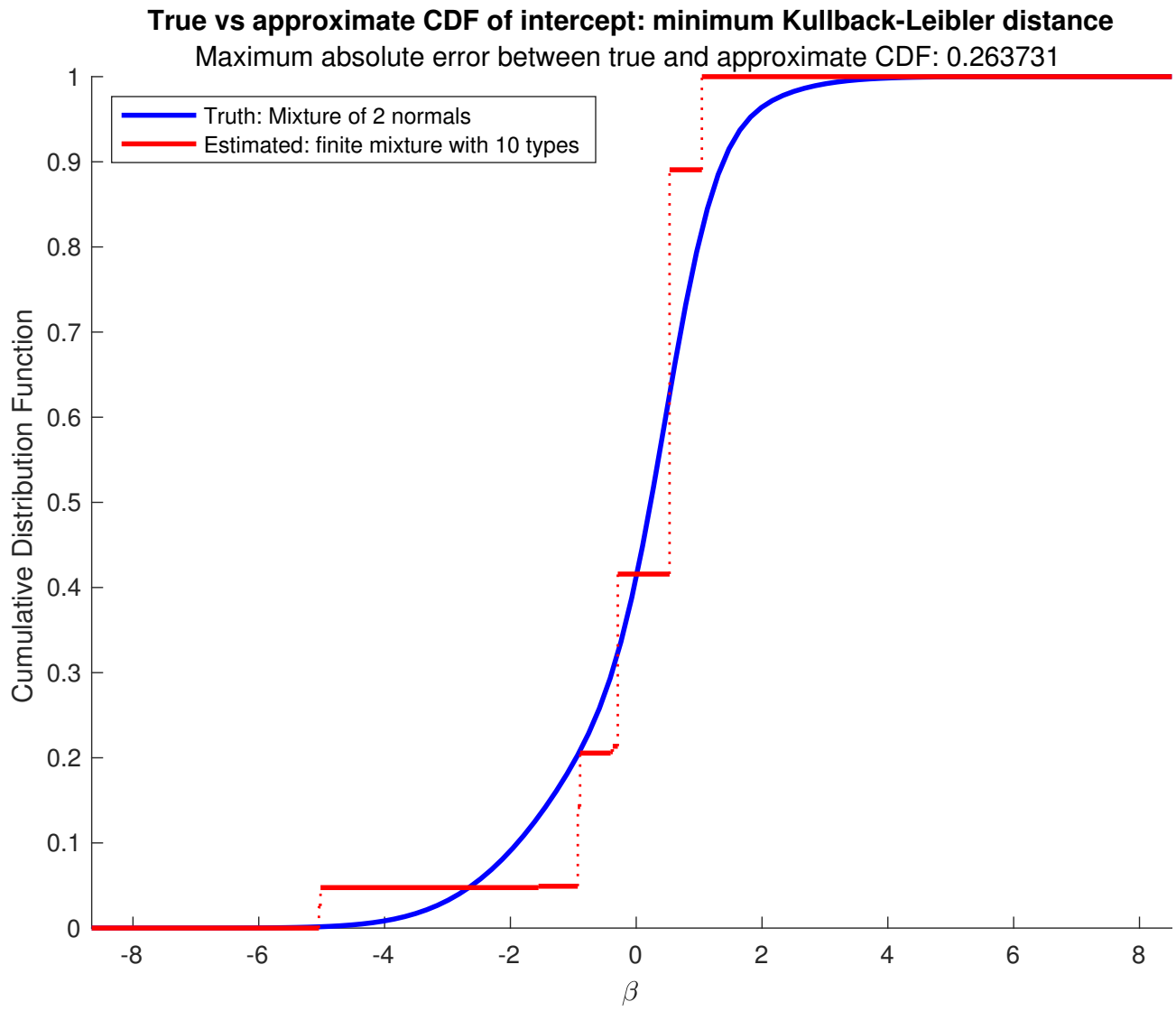
True vs approximate intercept CDF, 4 types



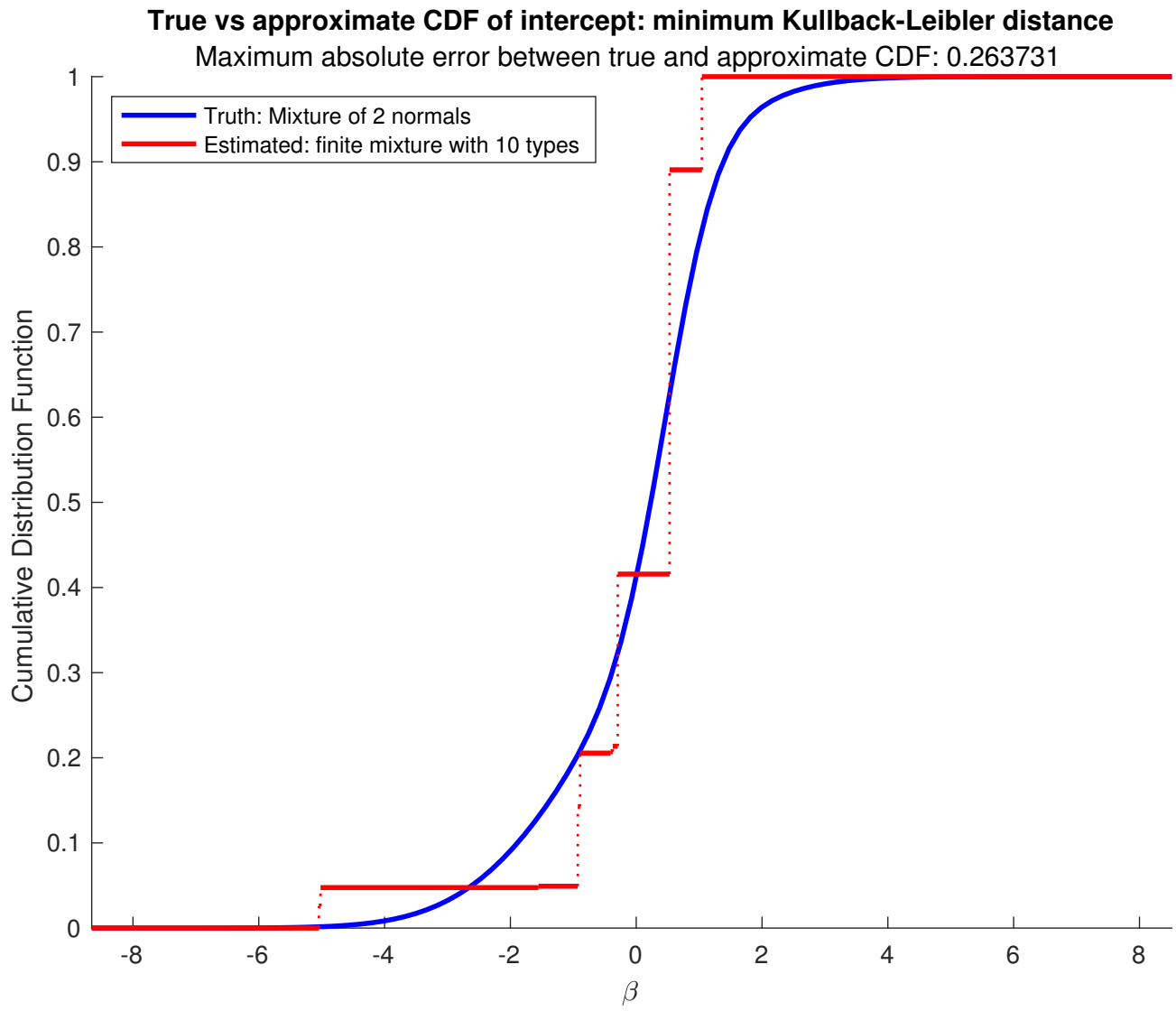
True vs approximate intercept CDF, 5 types



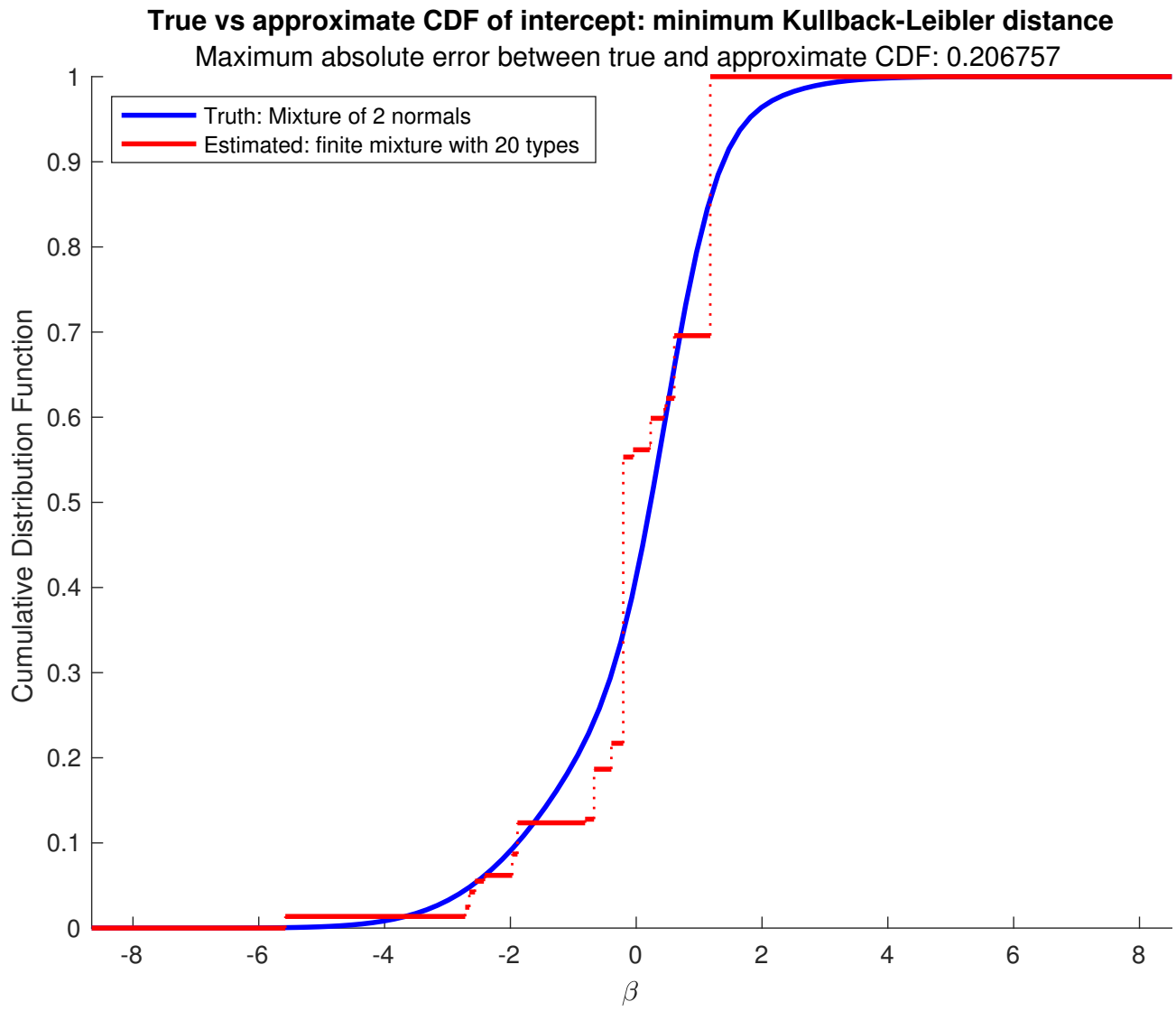
True vs approximate intercept CDF, 10 types



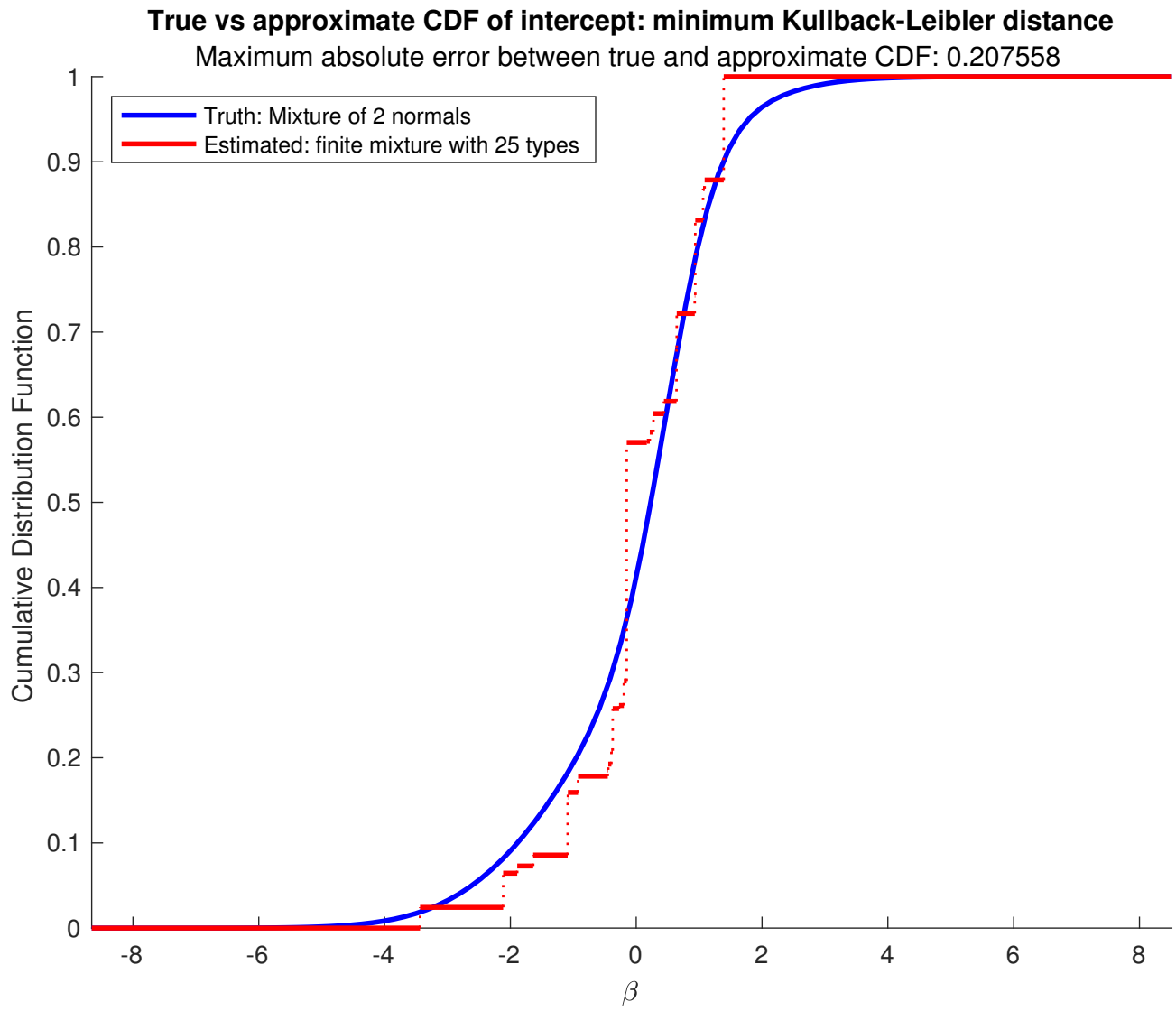
True vs approximate intercept CDF, 15 types



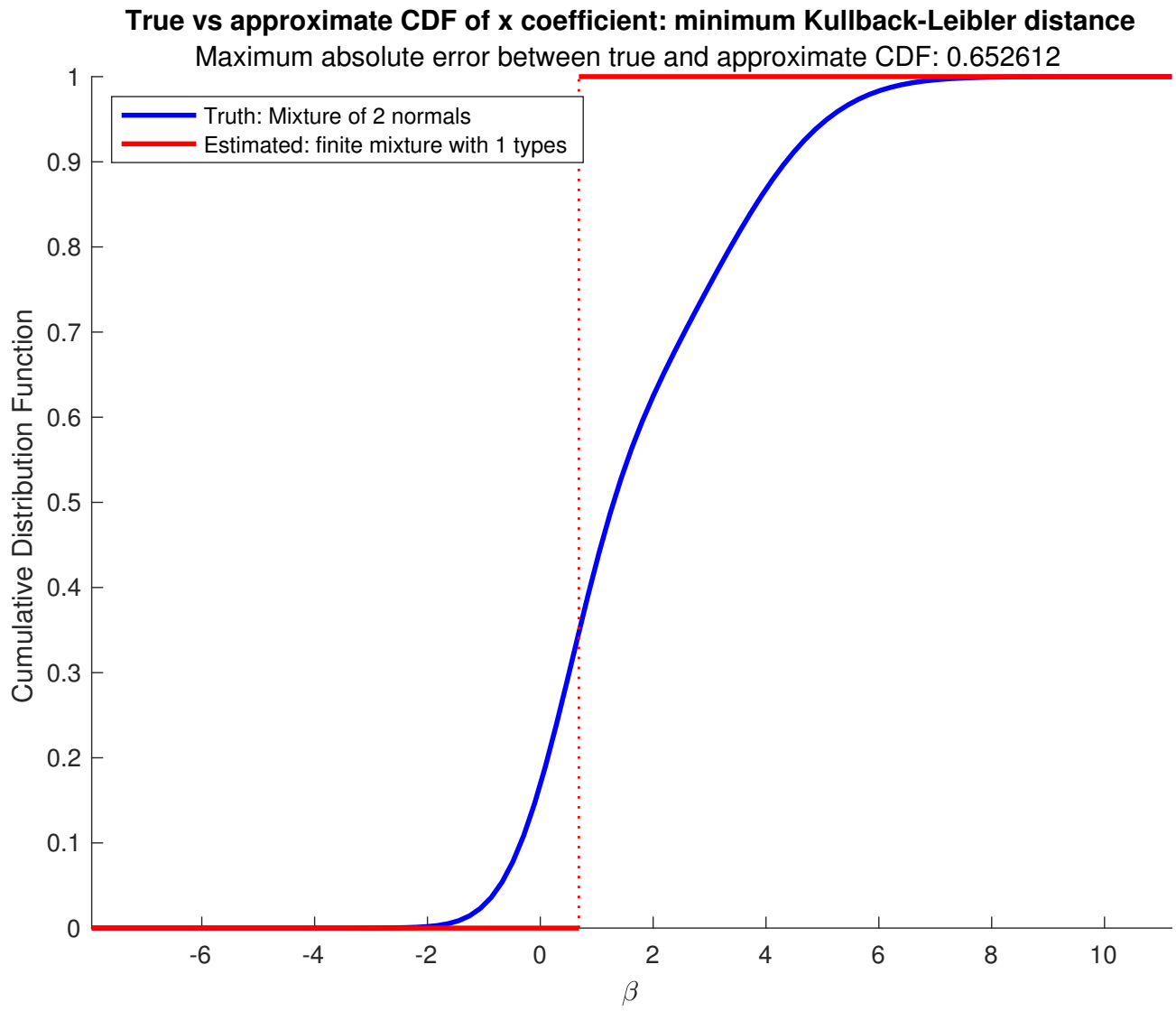
True vs approximate intercept CDF, 20 types



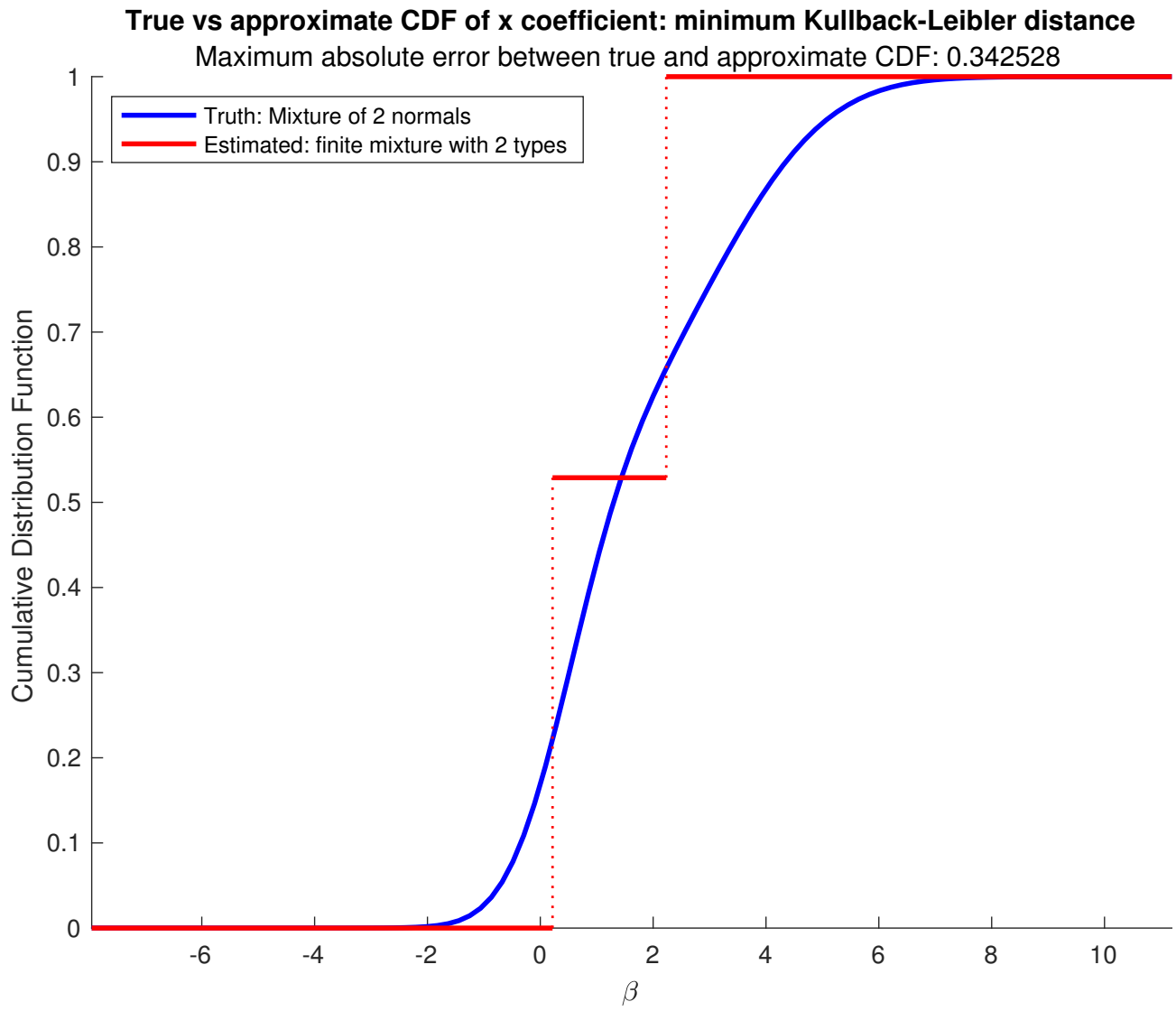
True vs approximate intercept CDF, 25 types



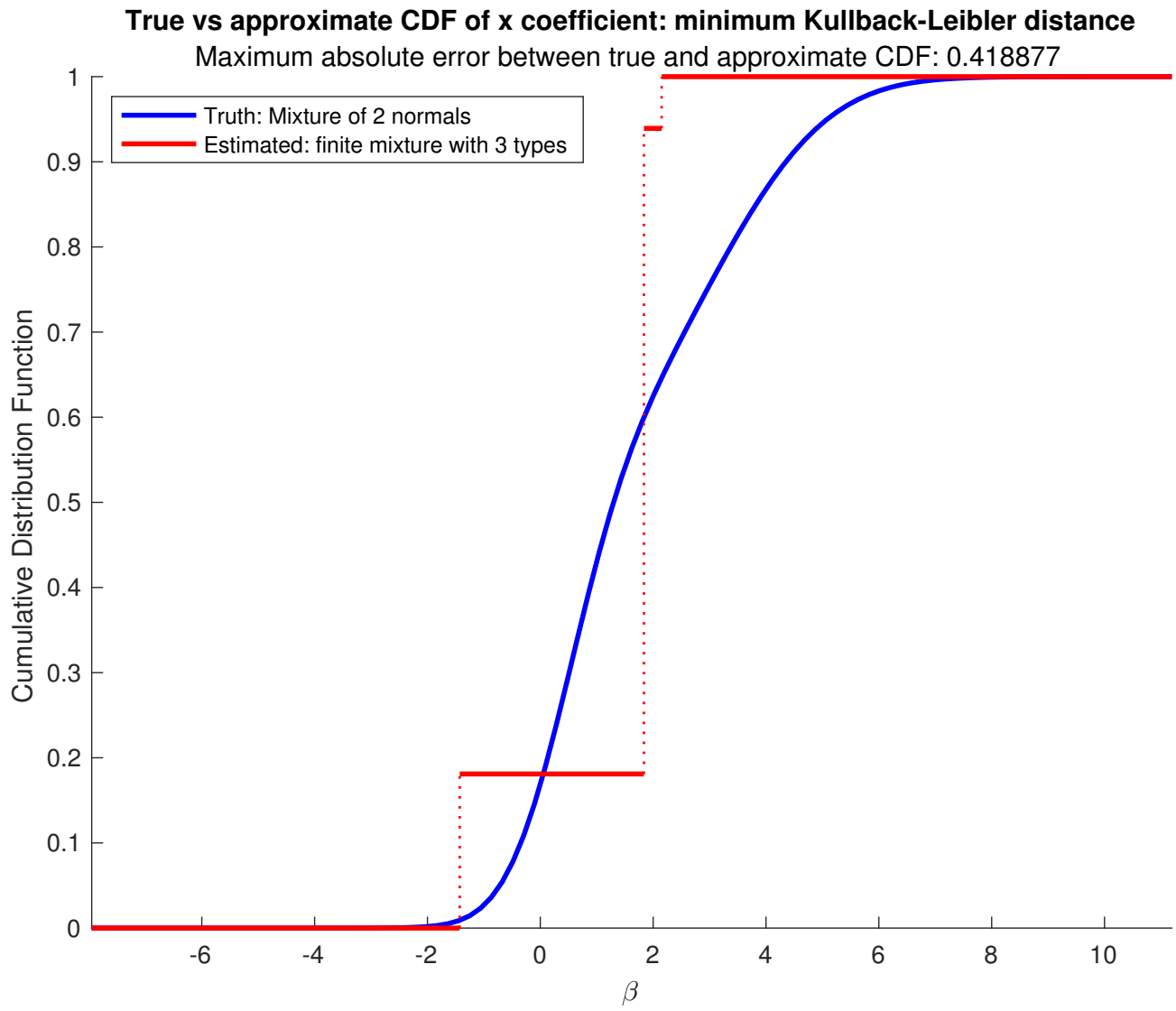
True vs approximate slope CDF, 1 type



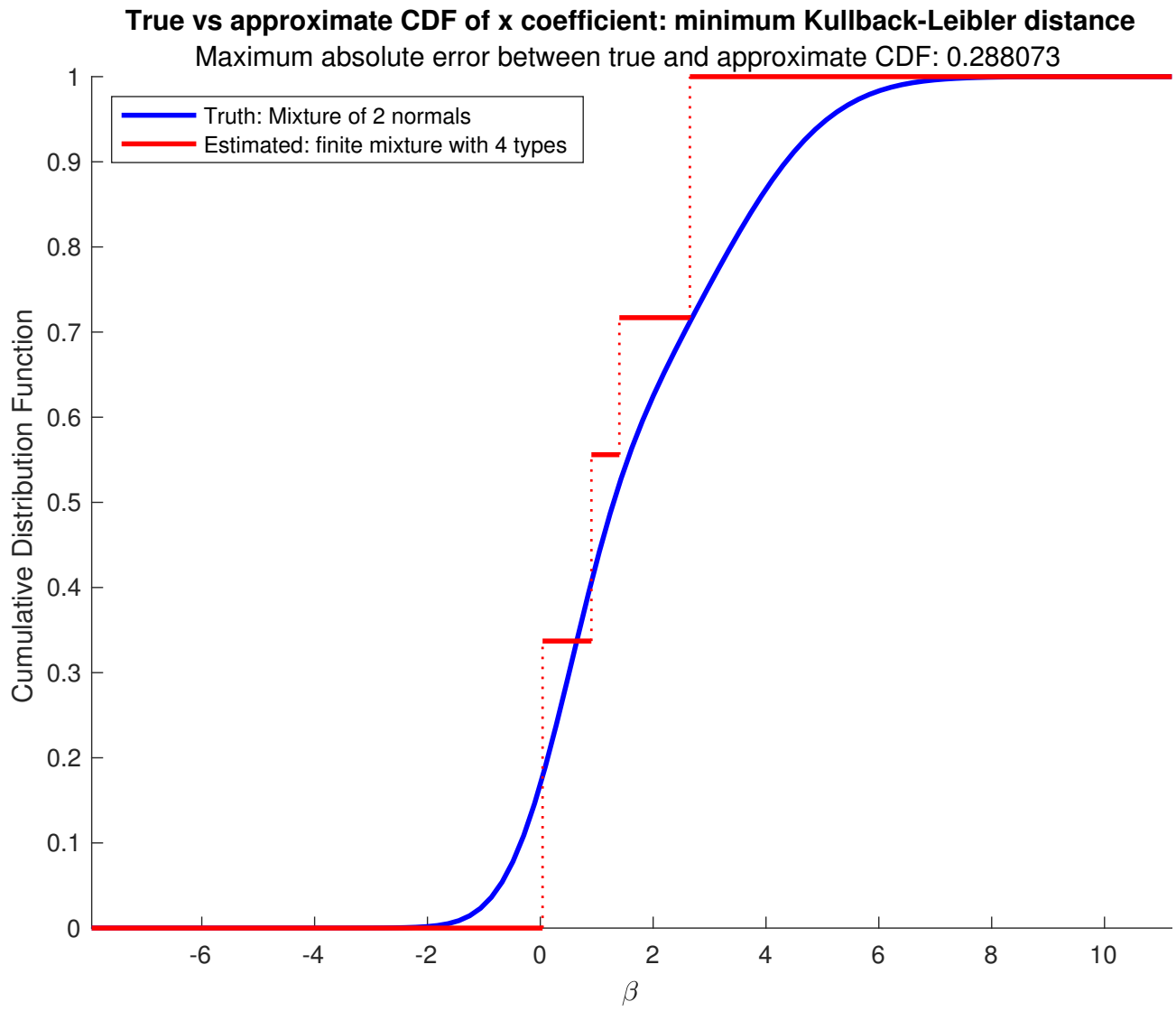
True vs approximate slope CDF, 2 types



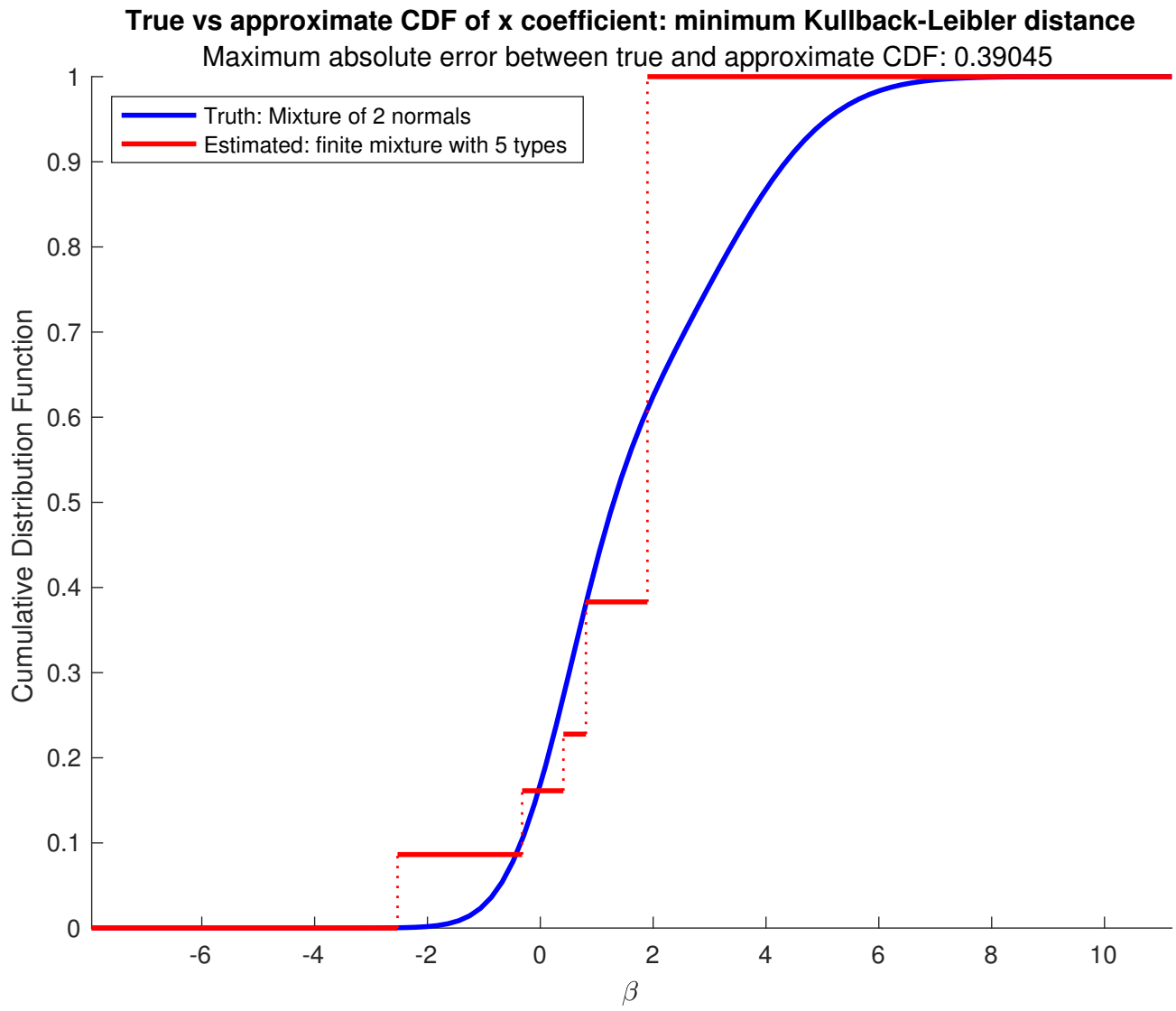
True vs approximate slope CDF, 3 types



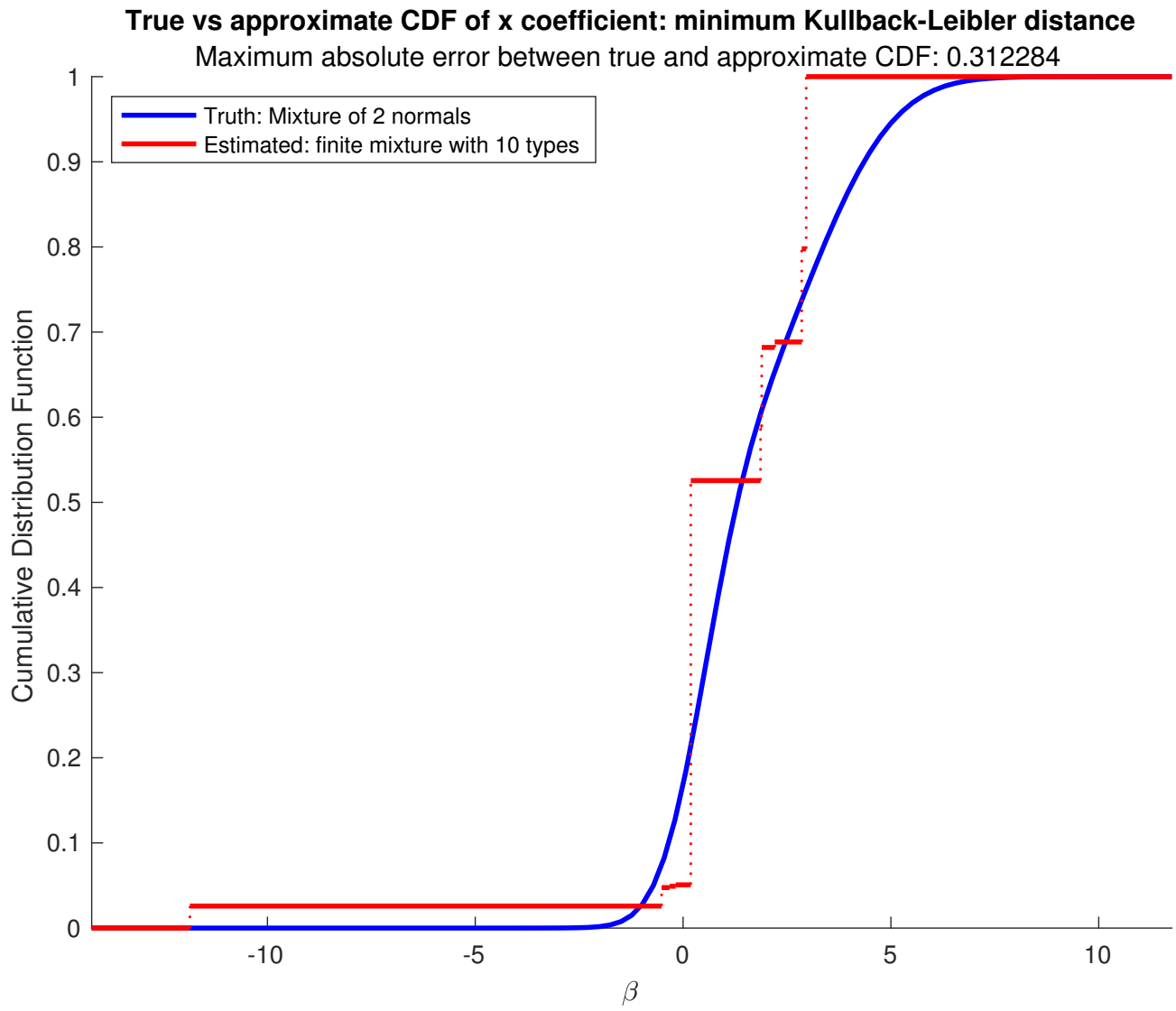
True vs approximate slope CDF, 4 types



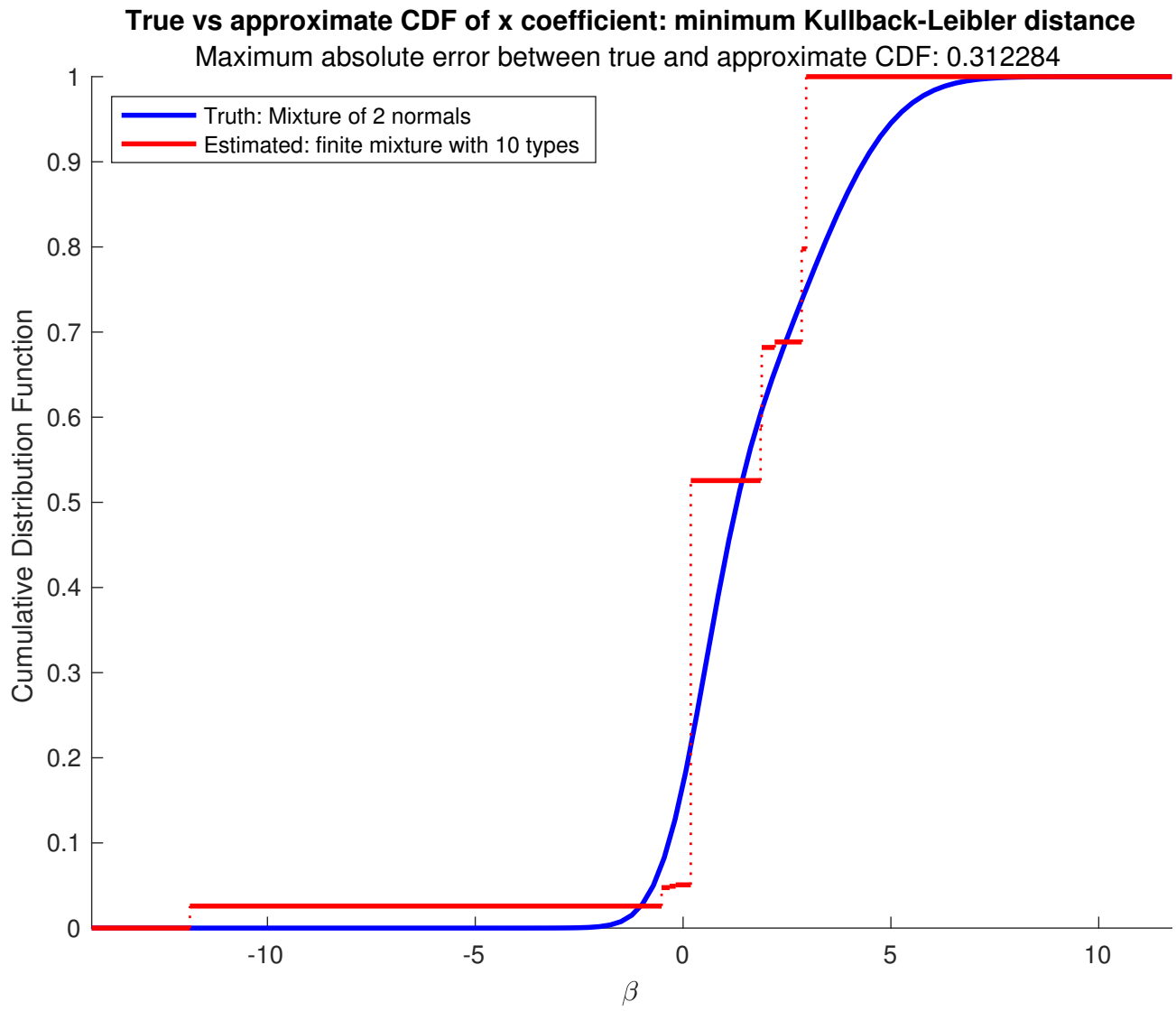
True vs approximate slope CDF, 5 types



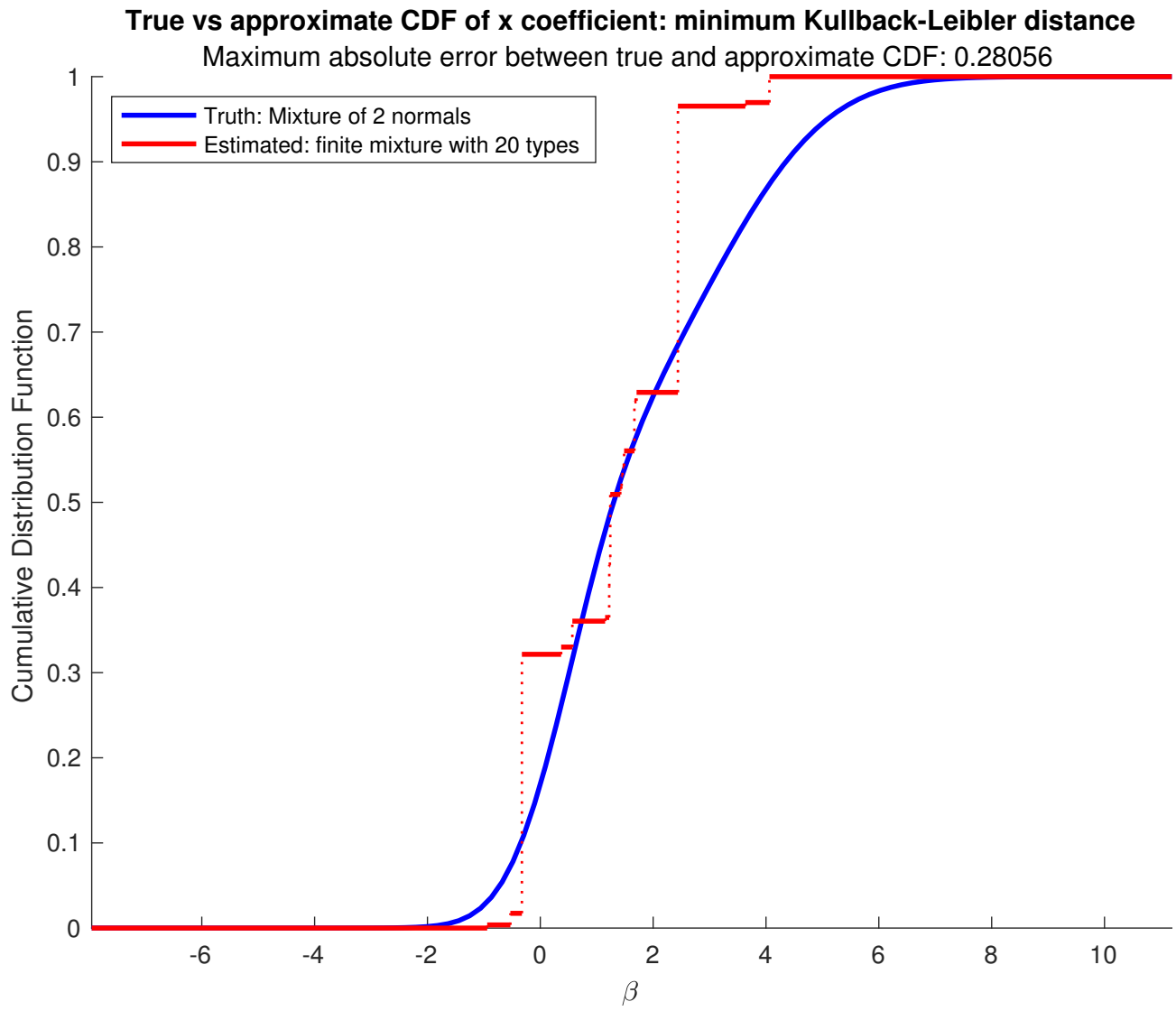
True vs approximate slope CDF, 10 types



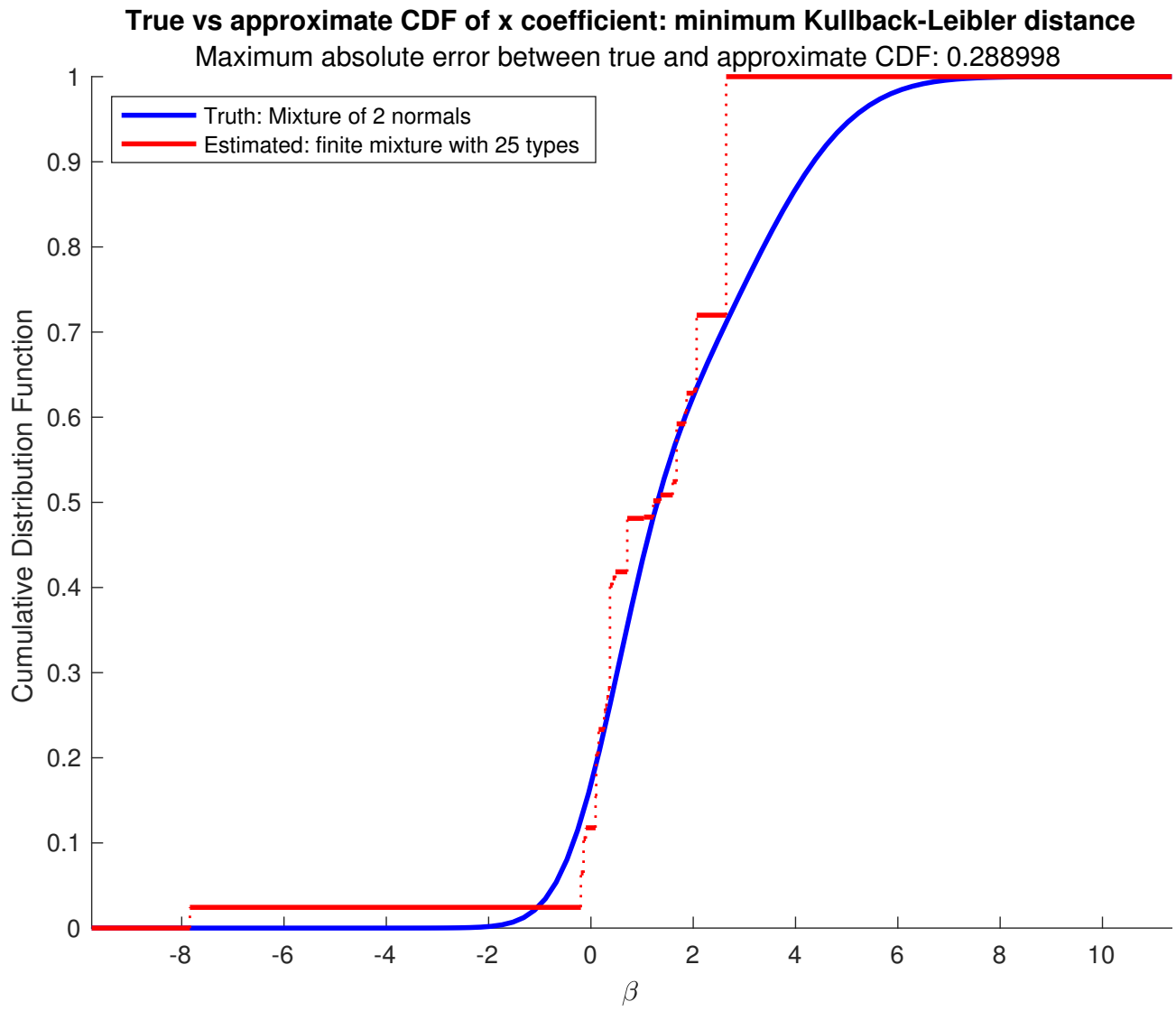
True vs approximate slope CDF, 15 types



True vs approximate slope CDF, 20 types



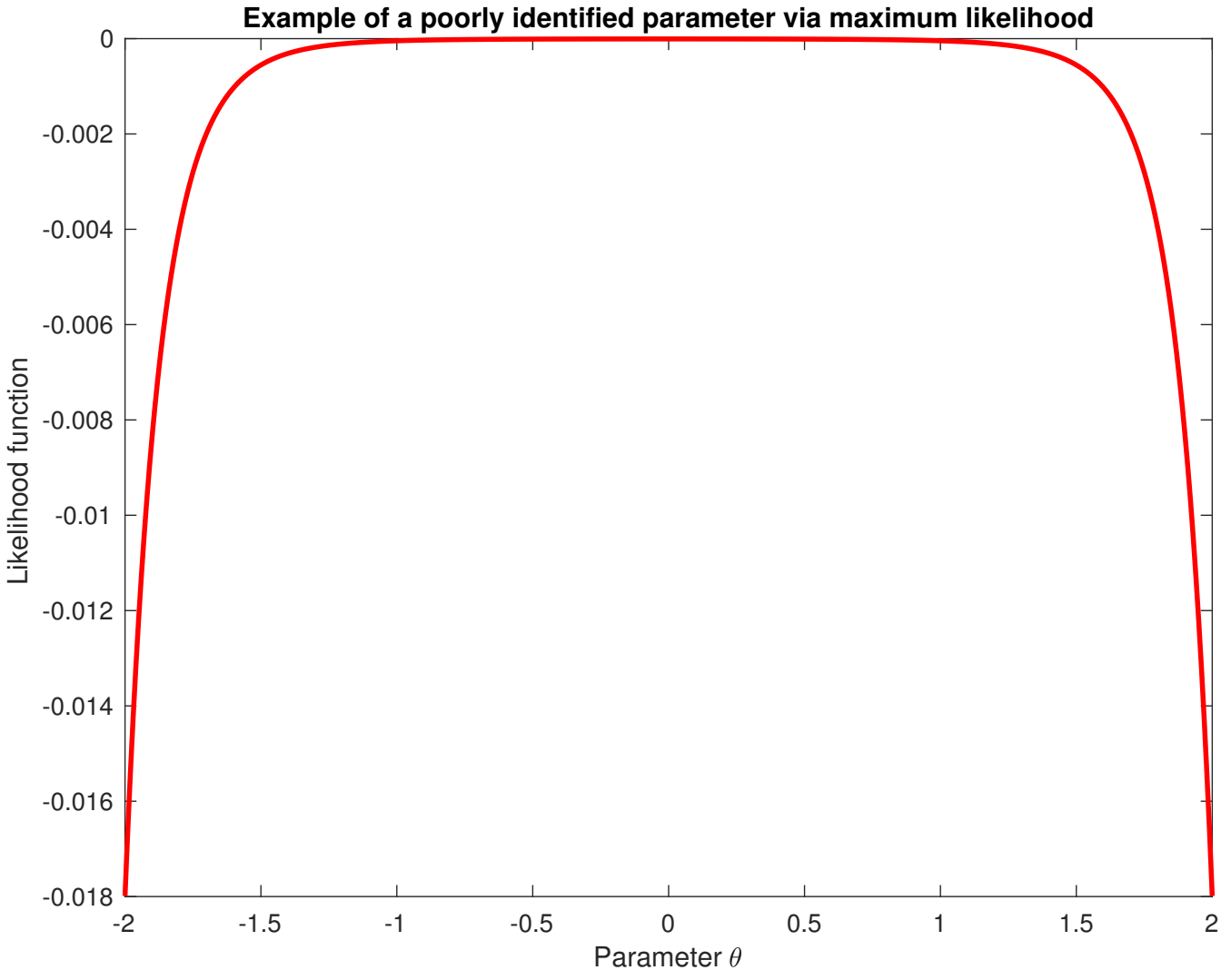
True vs approximate slope CDF, 25 types



What's going on here?

- By computing the Kullback-Leilber distance from finite mixture approximation to CCP $P(x, G_R(G))$ and two continuous random coefficients CCP $P(x, G)$, we are essentially going to the asymptotic limit with an infinite number of observations. Thus the slow convergence of $G_R(G)$ to G reflects inherent *approximation error* not *sampling error.*
- But the good news is that we do verify the super-fast convergence of $P(x, G_R(G))$ to $P(x, G)$ as R increases. We can basically "nail" $P(x, G)$ uniformly over the domain of x when R is as small as 2.
- Why the super fast convergence of $P(x, G_R(G))$ to $P(x, G)$ but the super slow convergence of $G_R(G)$ to G as R increases?
- **Intuition** The Kullback-Liebler distance $KL(G_R, G)$ (and the L^2 distance $\delta(G_R, G)$) is extremely *flat* in G_R near G. Thus, there is a wide range of distributions G_R that nearly minimize $KL(G_R, G)$ as R increases, and lacking infinite precision computers, we encounter a situation of *huge multiplicity of local optima* due to to the fact that $KL(G_R, G)$ is nearly zero for G_R in a *very big* neighborhood of G.

Illustration of a poorly identified parameter



Flat Likelihood $\Longrightarrow$ Singular Information Matrix

- Recall the hessian matrix of the Kullback-Leibler distance equals the *Information Matrix* when the model is correctly specified (i.e. then $G_R = G$).
- The inverse of the information matrix gives the asymptotic distribution of the maximum likelihood estimator.
- But if the information matrix is nearly singular, then the variances of many of the parameters are *huge!*
- So near flatness of the KL distance $\Longrightarrow$ near singularity of the Information Matix $\Longrightarrow$ weak identification of G using a sieve of finite mixtures to estimate G by maximum likelihood.
- However by Delta Theorem we have

$$\sqrt{N}[p(x, \hat{G}_R) - p(x, G)] \Longrightarrow N(0, \nabla'_\theta P(x|\hat{\theta}_R)\mathcal{I}^{-1}\nabla_\theta P(x|\hat{\theta}_R) \quad (42)$$

where $\mathcal{I}$ is the information matrix for $\hat{\theta}_R$ which is asymptotically equal to the hessian of $LK(G_R(\theta_R), G)$, $\nabla^2_\theta KL(G_R(\theta_R), G)$.

Our ability to identify G improves with panel data

- Suppose we observe each individual i for T periods and we assume that preference parameters τ_i remain *fixed* across successive choices $t = 1, \ldots, T$ for each $i = 1, \ldots, N$. Intuitively this extra information should help to identify G.
- Take the most extreme case where we assume $T \to \infty$ and $N \to \infty$ (i.e. "big data"). Then it is pretty clear we can estimate G at $\sqrt{N}$ rates via this *2-step fixed effects estimator.*
- **Step 1** For each person i estimate *person-specific preference parameters* using $(y_{i1}, \ldots, y_{iT}, x_{i1}, \ldots, x_{iT})$, $\hat{\tau}_i$ given by

$$\hat{\tau}_i = \operatorname*{argmax}_{\tau} L(\tau) = \frac{1}{T} \sum_{t=1}^{T} \log(f(y_{it}|x_{it}, \tau)), \tag{43}$$

  where $f(y|x, \tau) = P(x, \tau)^y (1 - P(x, \tau))^{(1-y)}$.
- **Step 2** Now estimate G using the *empirical CDF* $\hat{G}_N$ using the estimated fixed effects, $(\hat{\tau}_1, \ldots, \hat{\tau}_N)$,

$$\hat{G}_N(\tau) = \frac{1}{N} \sum_{i=1}^{N} I\{\tau \leq \hat{\tau}_i\}. \tag{44}$$

Large T asymptotics for unobserved heterogeneity

- For large T we have $\hat{\tau}_i = \tau_i + O_p(1/\sqrt{T})$ where the estimation "noise" in the types is disappearing at rate $1/\sqrt{T}$.
- It follows that each estimated "fixed effect" $\hat{\tau}_i$ equals the true value τ_i put a small amount of noise so we can also write

$$I\{\tau \leq \hat{\tau}_i\} = I\{\tau \leq \tau_i\} + O_p(1/\sqrt{T}). \tag{45}$$

- Since this estimation noise is independent across observations i we can write

$$\hat{G}_N(\tau) = G_N(\tau) \equiv \frac{1}{N}\sum_{i=1}^{N} I\{\tau \leq \tau_i\} + O_p(1/\sqrt{T}), \tag{46}$$

where $G_N(\tau)$ is the ordinary *empirical CDF* based on 'uncontaminated" observations $(\tau_i, \ldots, \tau_N)$.

- Thus, as $T \to \infty$ and $N \to \infty$ we can show that

$$\sqrt{N}[\hat{G}_N(\tau) - G(\tau)] \Longrightarrow N(0, G(\tau)[1 - G(\tau)]), \tag{47}$$

and in fact $\hat{G}_N(\tau)$ converges *uniformly* to $G(\tau)$ using standard results on *empirical processes.*

Fixed T asymptotics for unobserved heterogeneity

- Between the cross sectional case $T = 1$ and large T asymptotics above, there are a range of cases depending on assumptions about how fast T increases as a function of N.
- Let's consider the *fixed* T case where T does not depend on N but $N \to \infty$. Does this improve our ability to infer G?
- **Yes!** There is more information using T successive observations per individual so if $\mathcal{I}_{R1}$ is the information matrix a finite mixture approximation with R types in the cross sectional case, $T = 1$, and $\mathcal{I}_{RT}$ is the information matrix in the panel case with $T > 1$, then

$$\mathcal{I}_{RT} > \mathcal{I}_{R1}, \tag{48}$$

  where the inequality should be interpreted as $\mathcal{I}_{RT} - \mathcal{I}_{R1}$ is a positive semi-definite matrix.
- Another way to say this is that in the fixed panel case, the Kullback-Leibler distance $KL(G_R, G)$ is less "flat" for G_R in a neightborhood of G, and this implies that the maximum likelihood estimator $\hat{G}_R$ should converge at a faster rate to G and/or have smaller asymptotic variance.

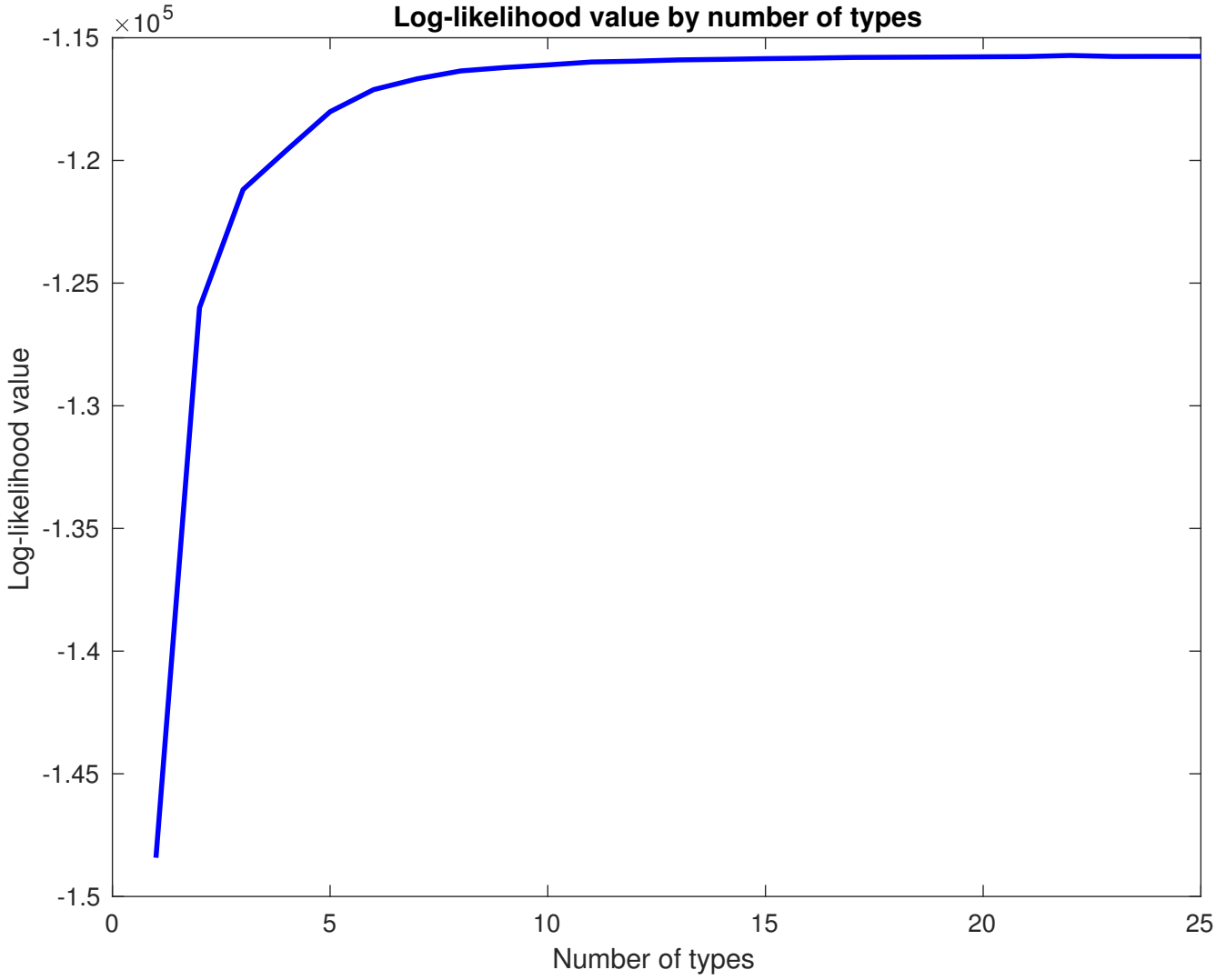
A limited Monte Carlo example

- We simulated data from $N = 5000$ individuals for $T = 50$ periods each, where the "true model" is a binary logit wheres G equal to a mixture of 2 bivariate normal distributions over (τ_0, τ_1).
- We maximized the likelihood function using hand-written `Matlab` code with analytic hessians of the likelihood and using the White robust misspecification-consistent covariance matrix for the parameters, since for any fixed R G_R is misspecified.
- We used AIC and BIC information criteria to select the "best" model, i.e. the number of types R.

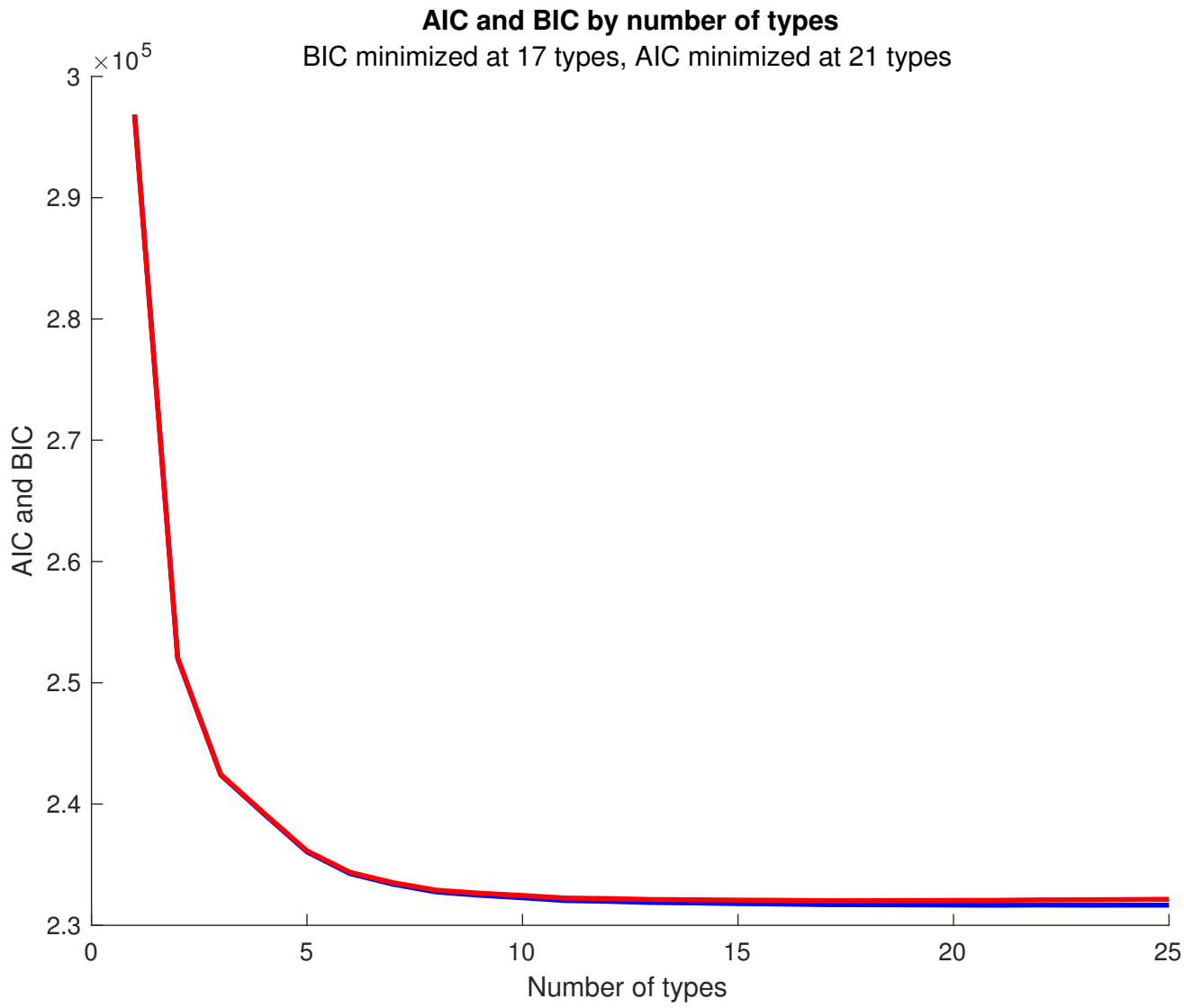
$$\begin{aligned} \text{AIC} &= -2L_N(\hat{\theta}_R) + 2(3R-1) \\ \text{BIC} &= -2L_N(\hat{\theta}_R) + \log(N)(3R-1) \end{aligned}$$

where $3R-1$ is the total number of parameters in θ_R when there are R types. The likelihood function obviously increases monotonically in R but the AIC and BIC penalize models with too many parameters, avoiding "overfitting". The "preferred" or "selected" model is the number of types R that minimizes AIC and BIC, respectively.

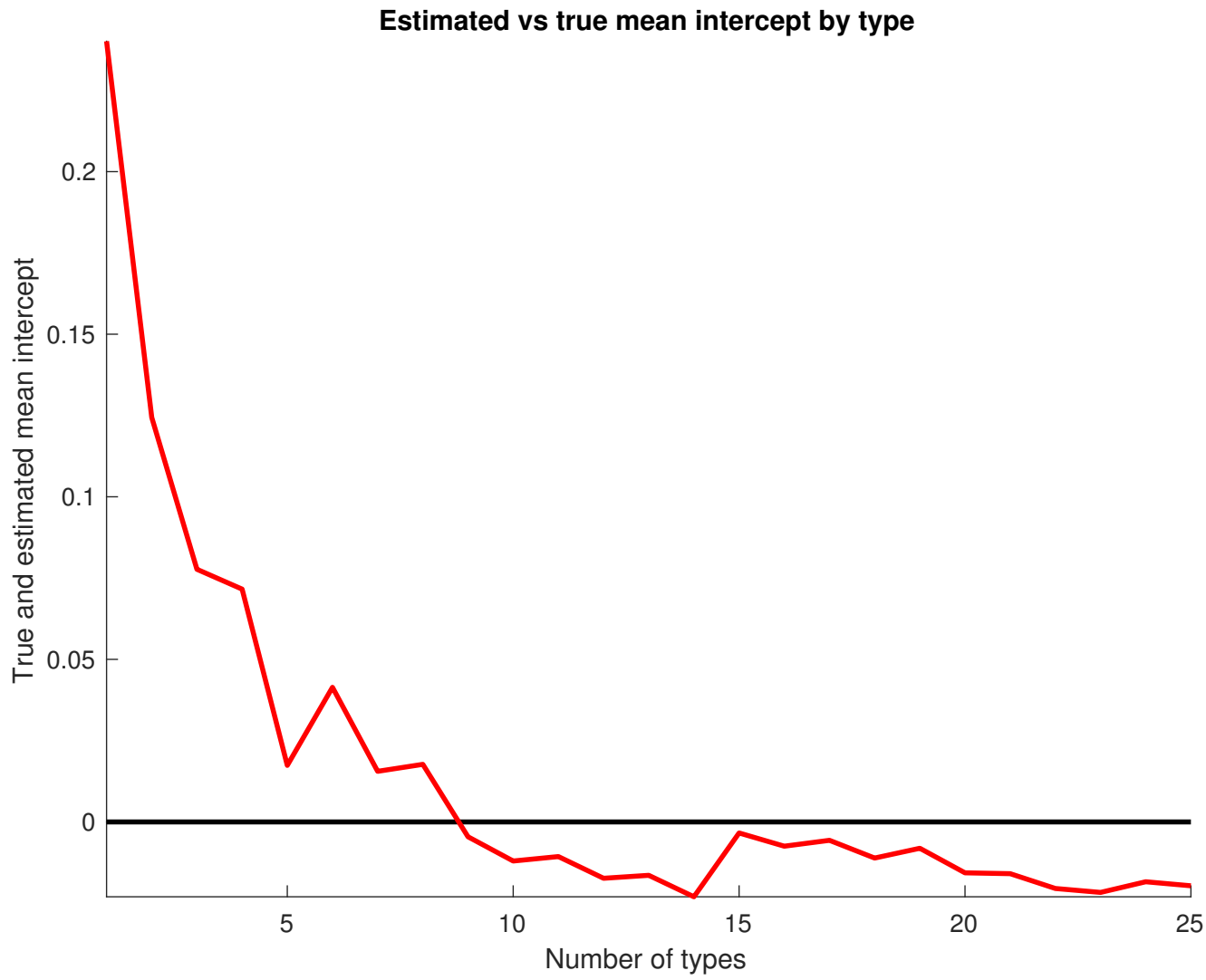
Log-likelihood by number of types



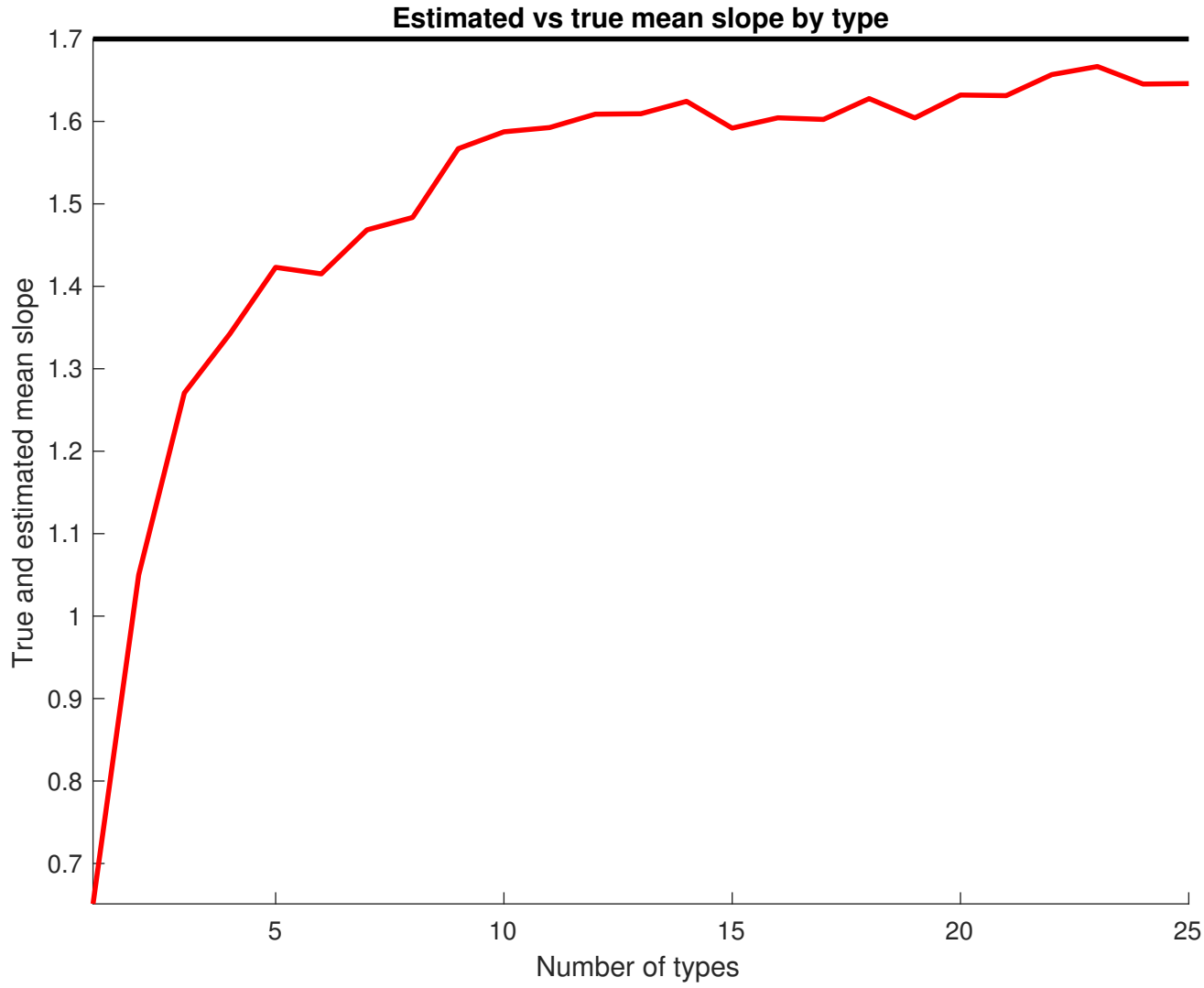
AIC and BIC by number of types



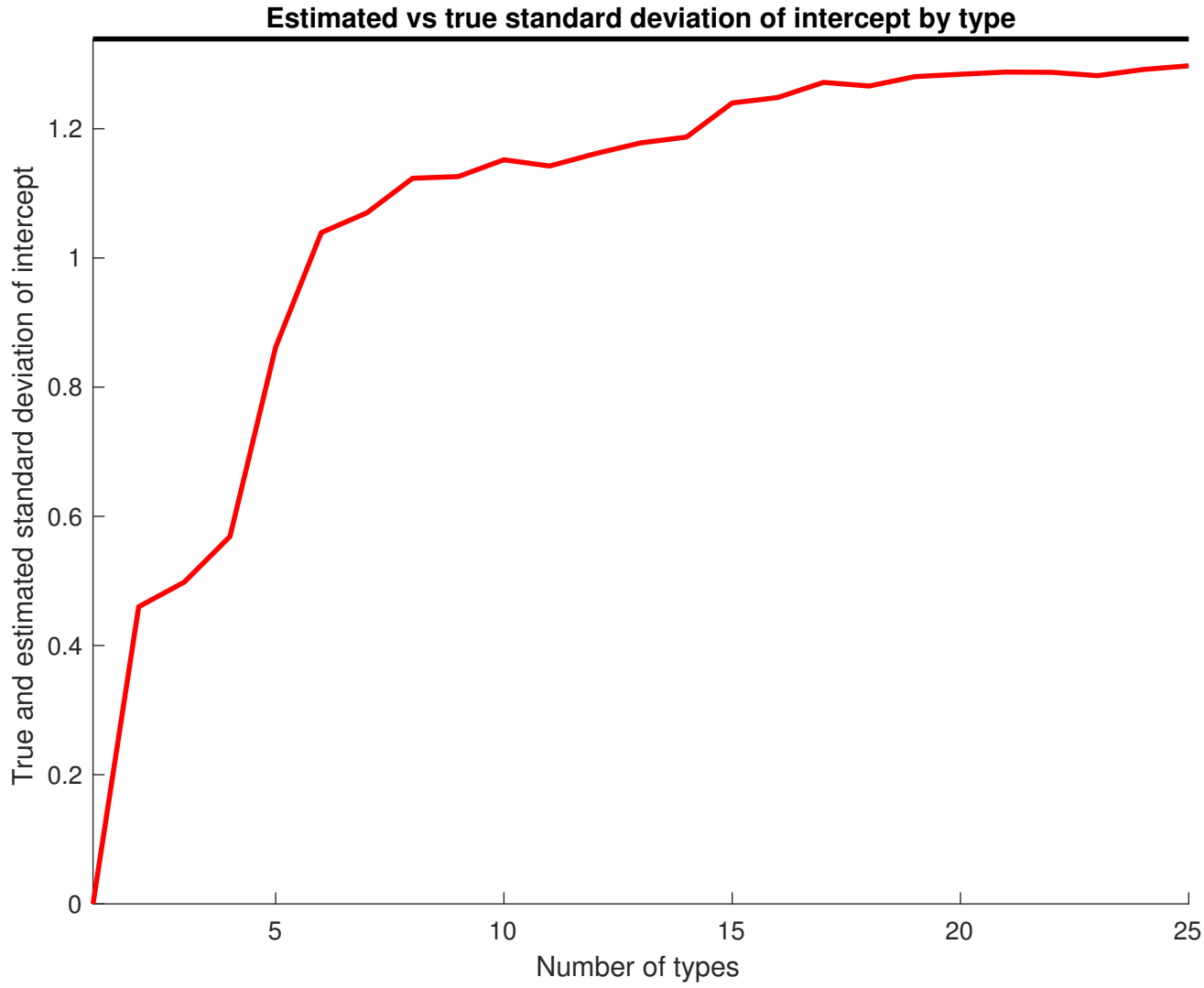
True vs Estimated Mean Intercept by number of types



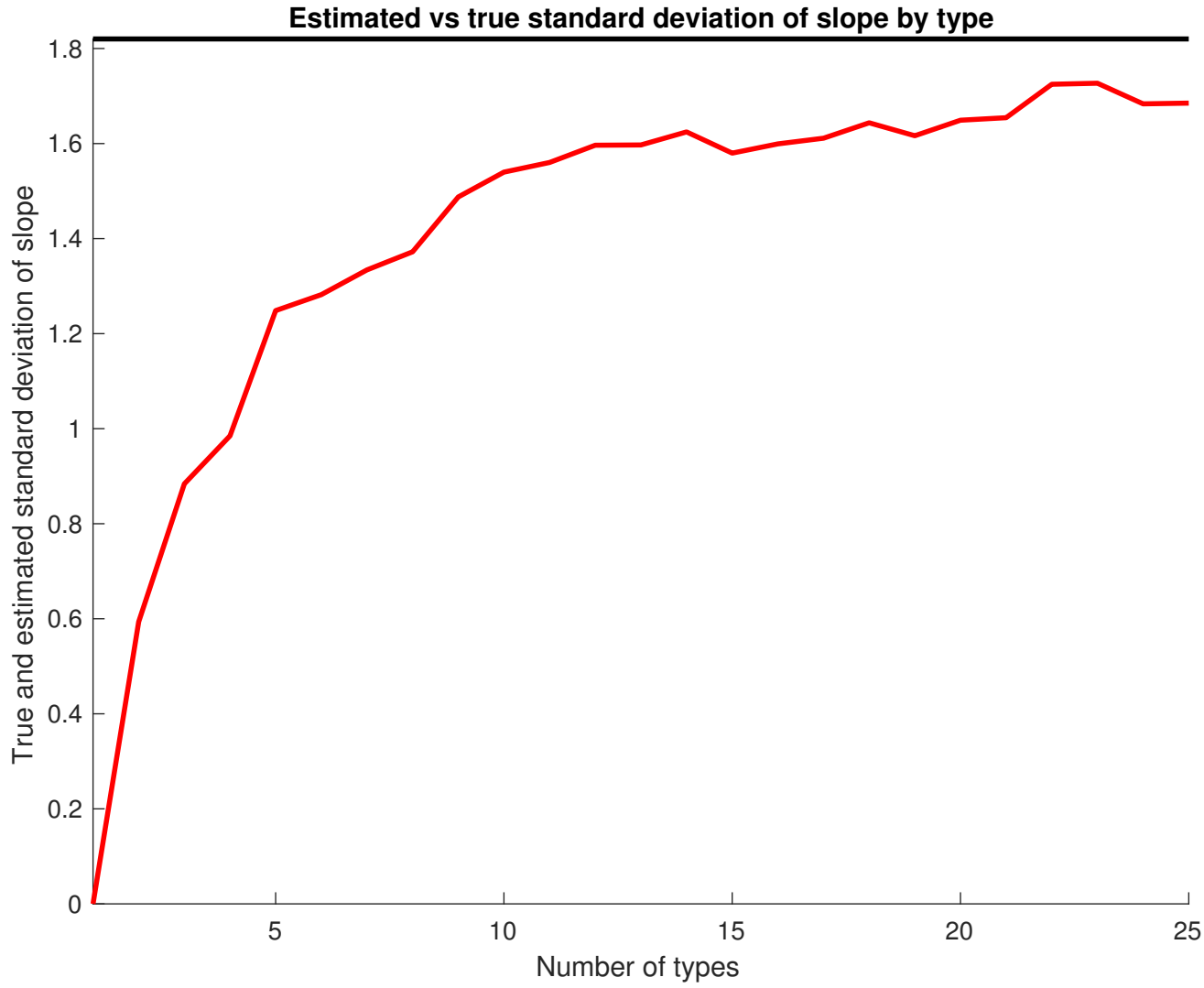
True vs Estimated Mean x-coefficient by number of types



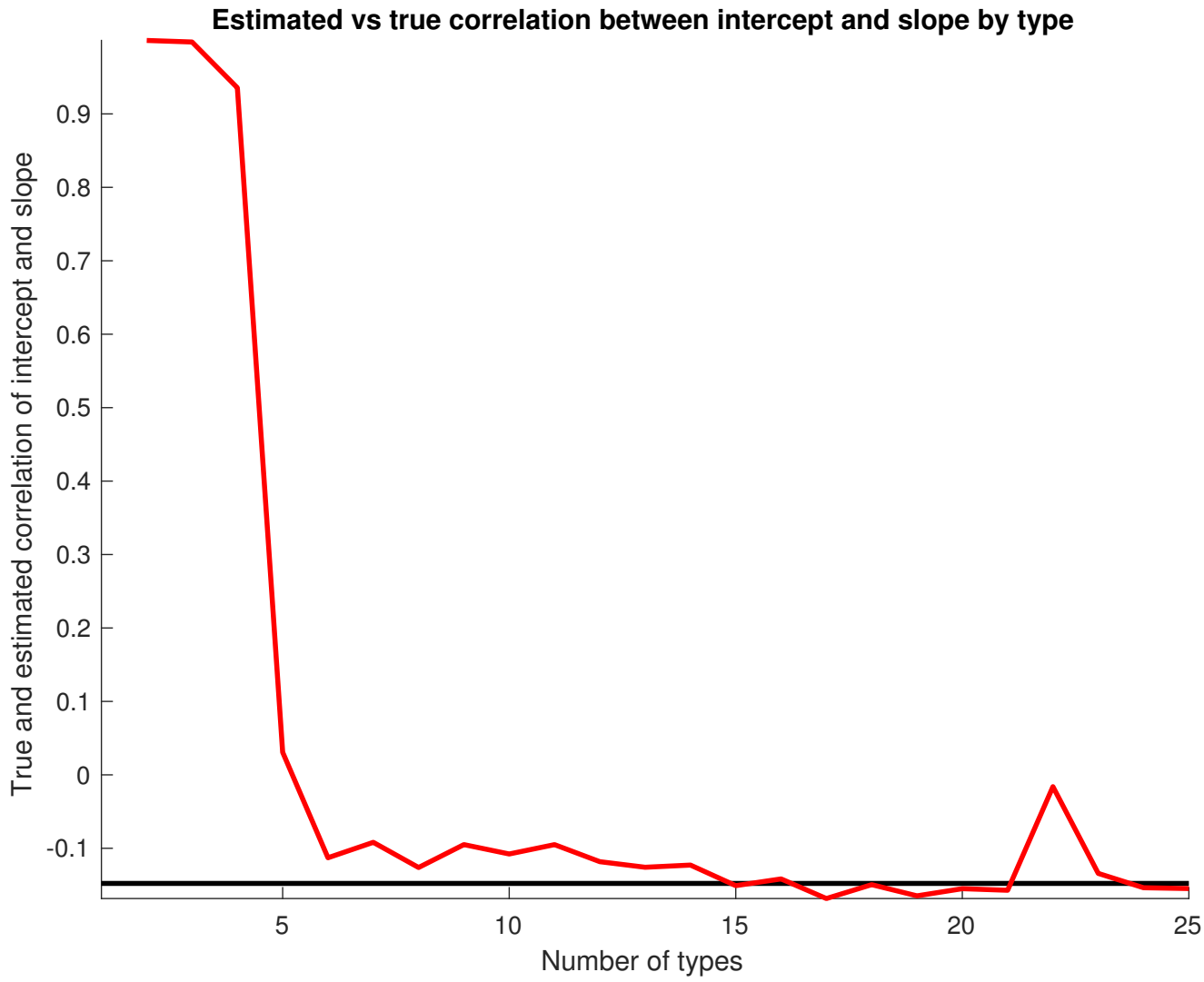
True vs Estimated Std of Intercept by number of types



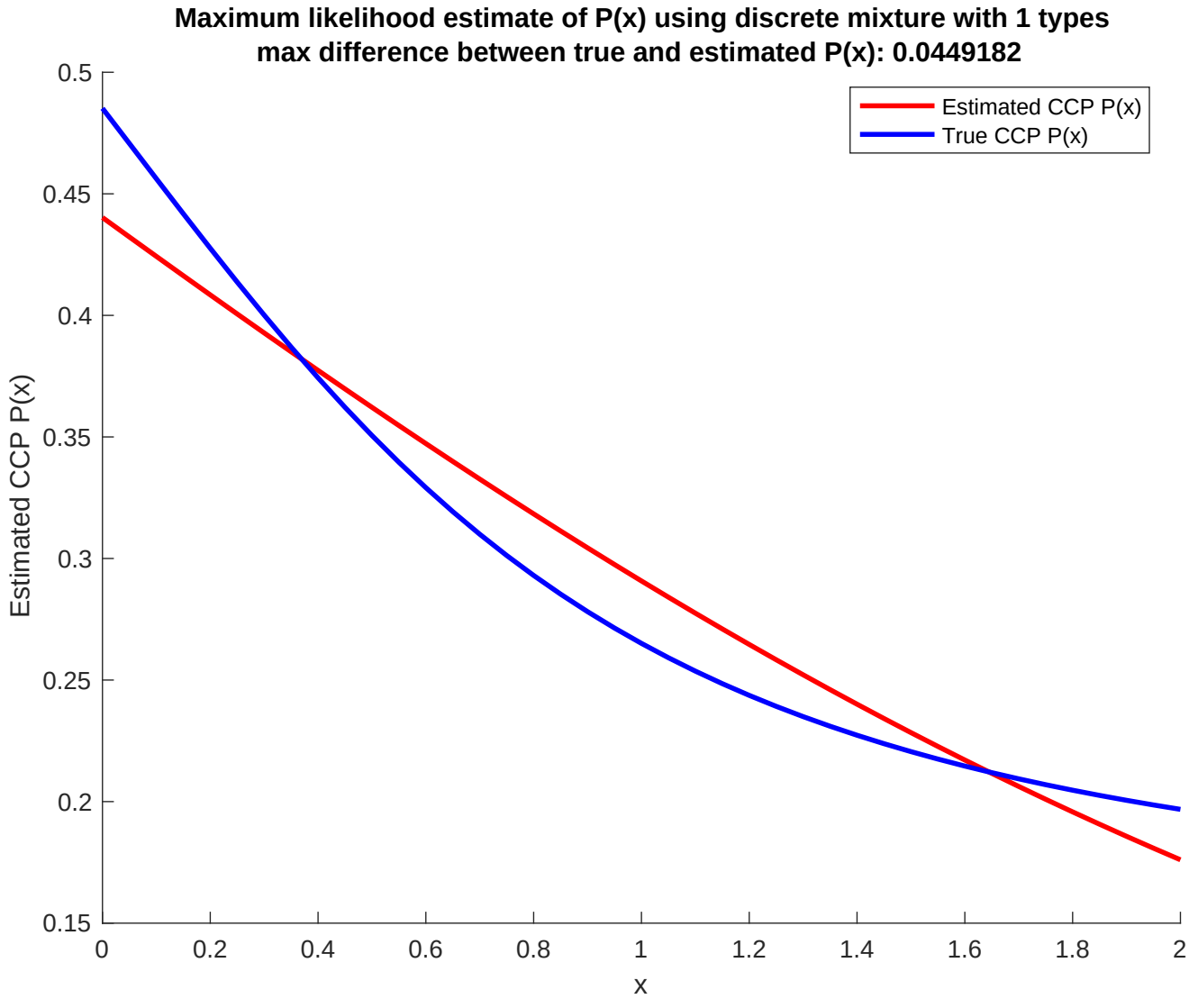
True vs Estimated Std of x-coefficient by number of types



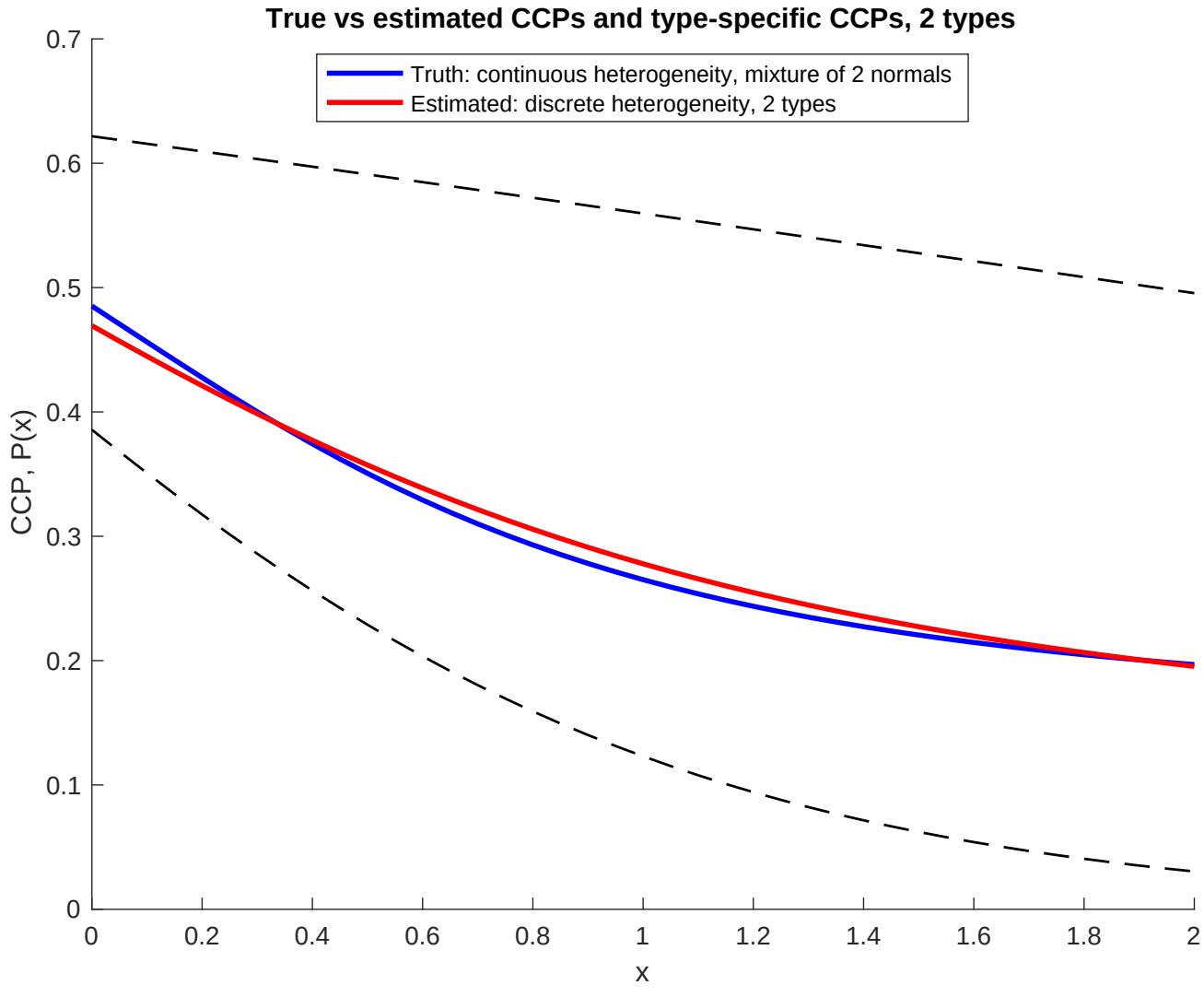
True vs Estimated intercept/slope correlation by types



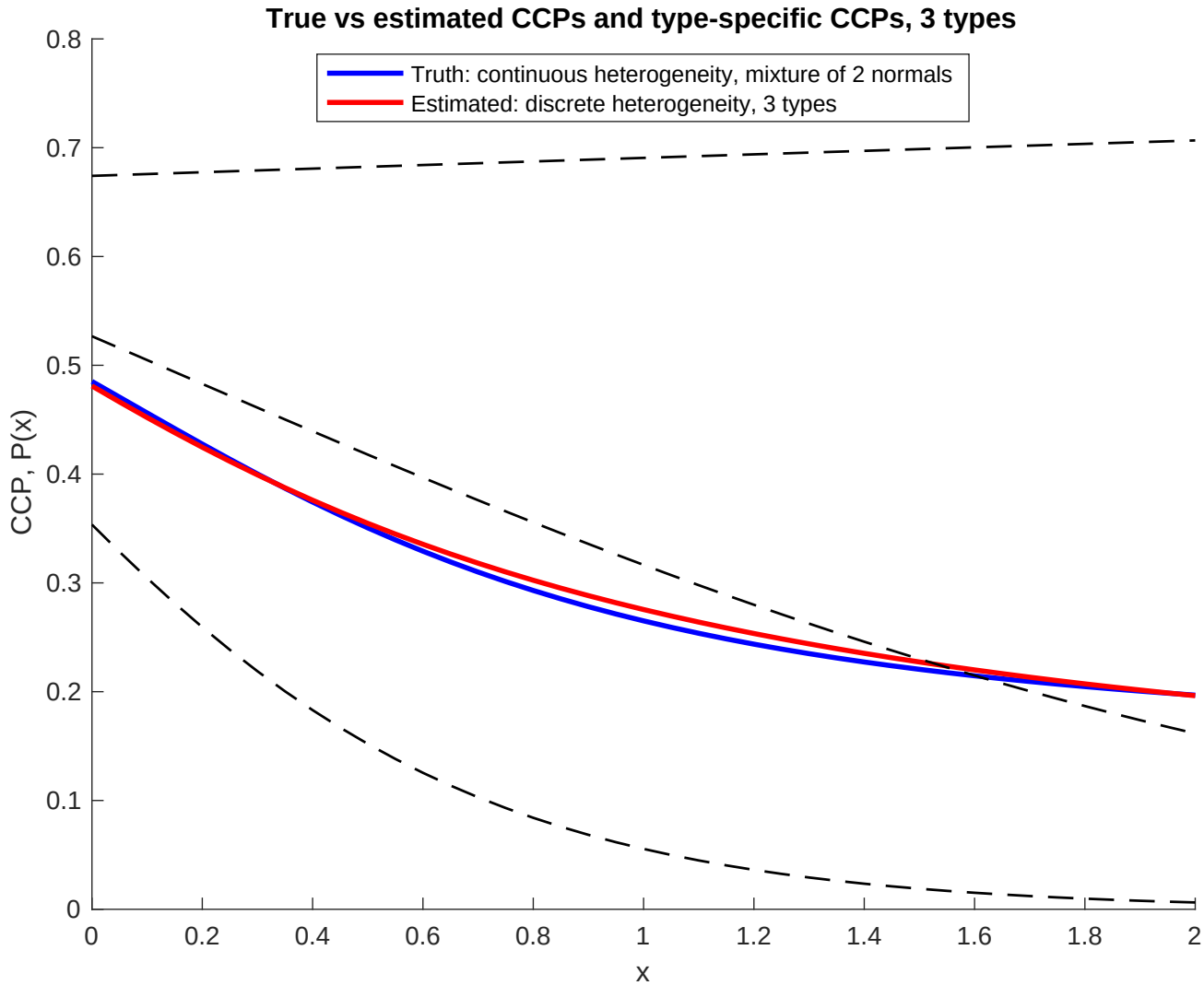
True vs estimated CCP, 1 type



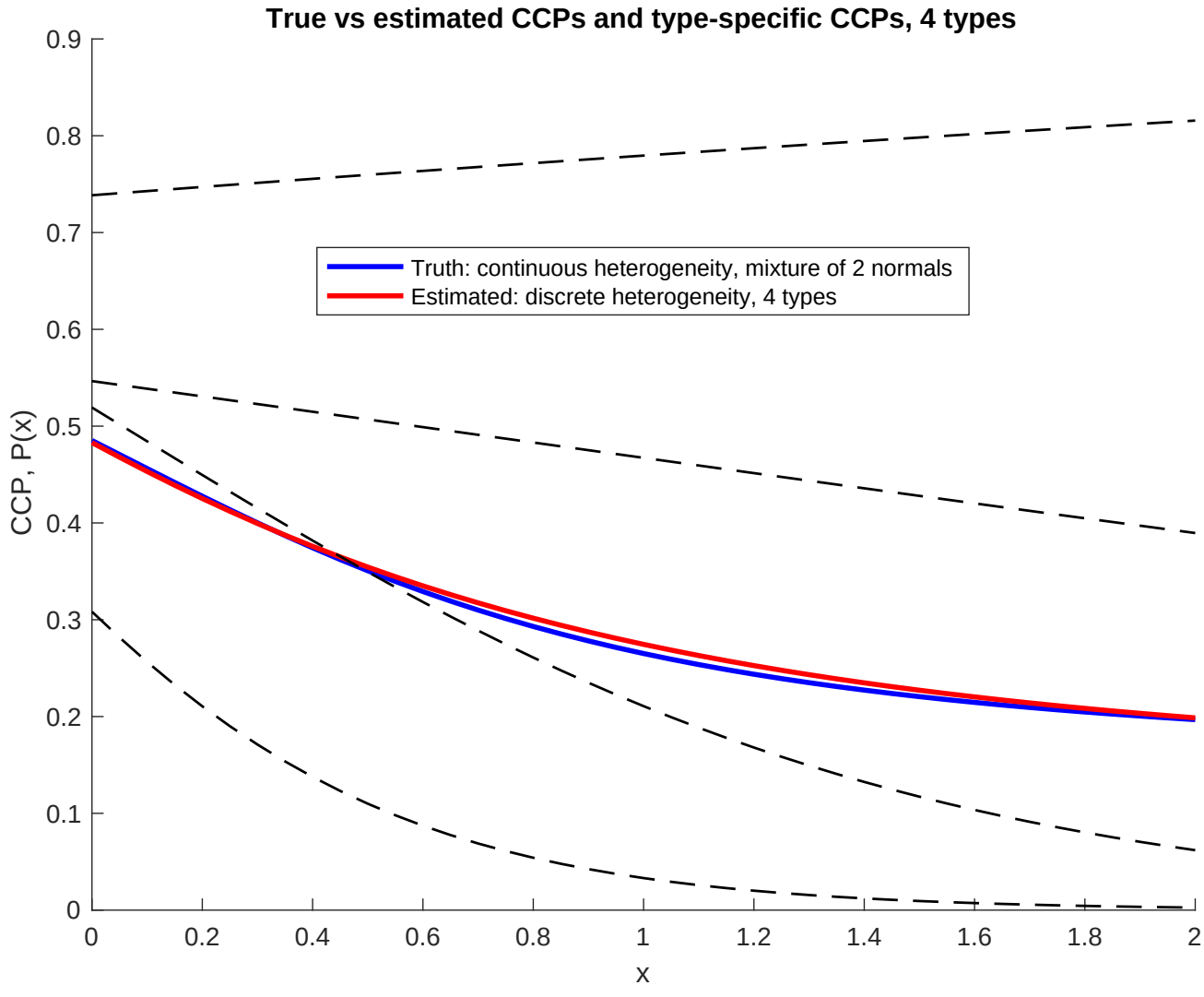
True vs estimated CCP, 2 types



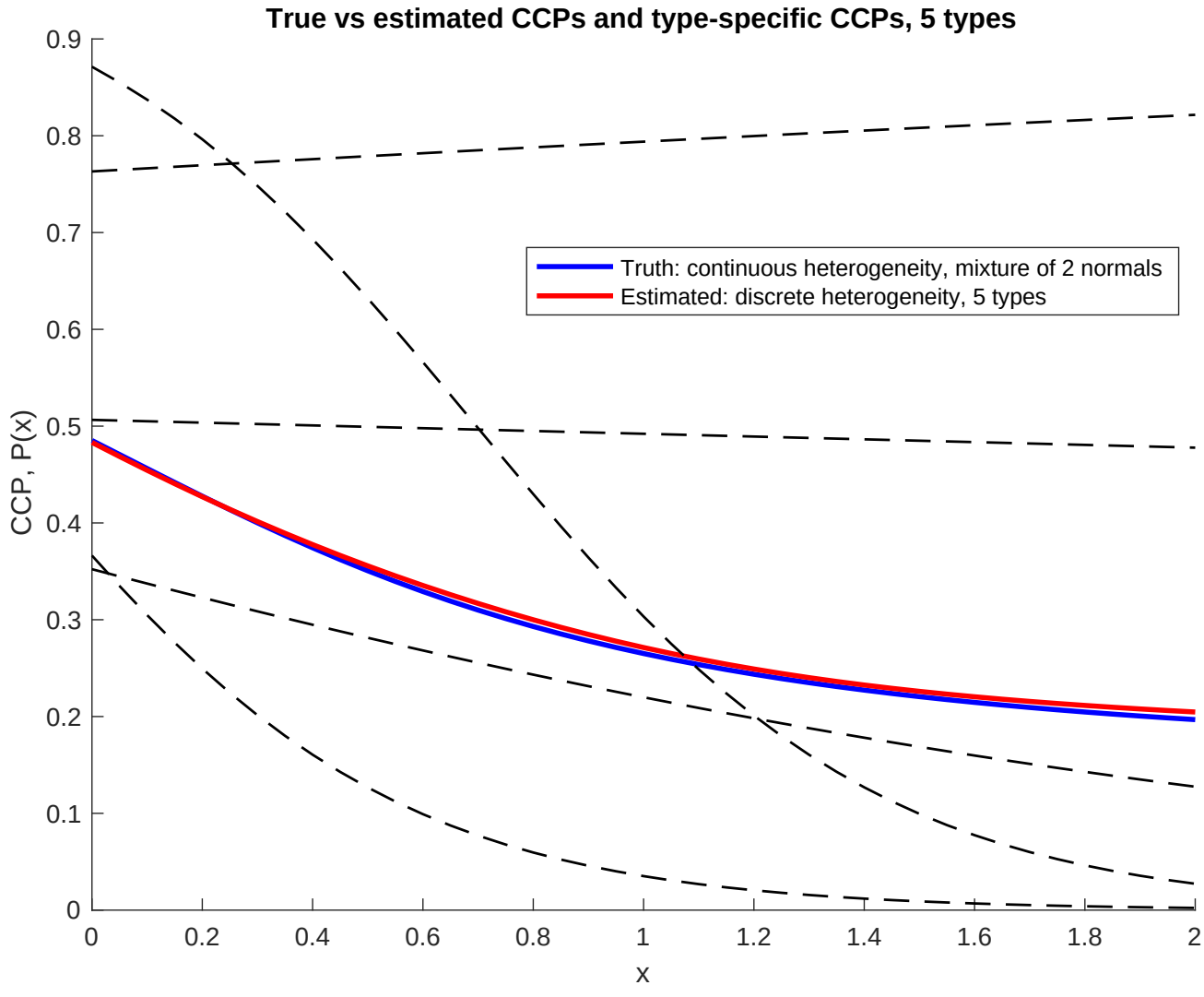
True vs estimated CCP, 3 types



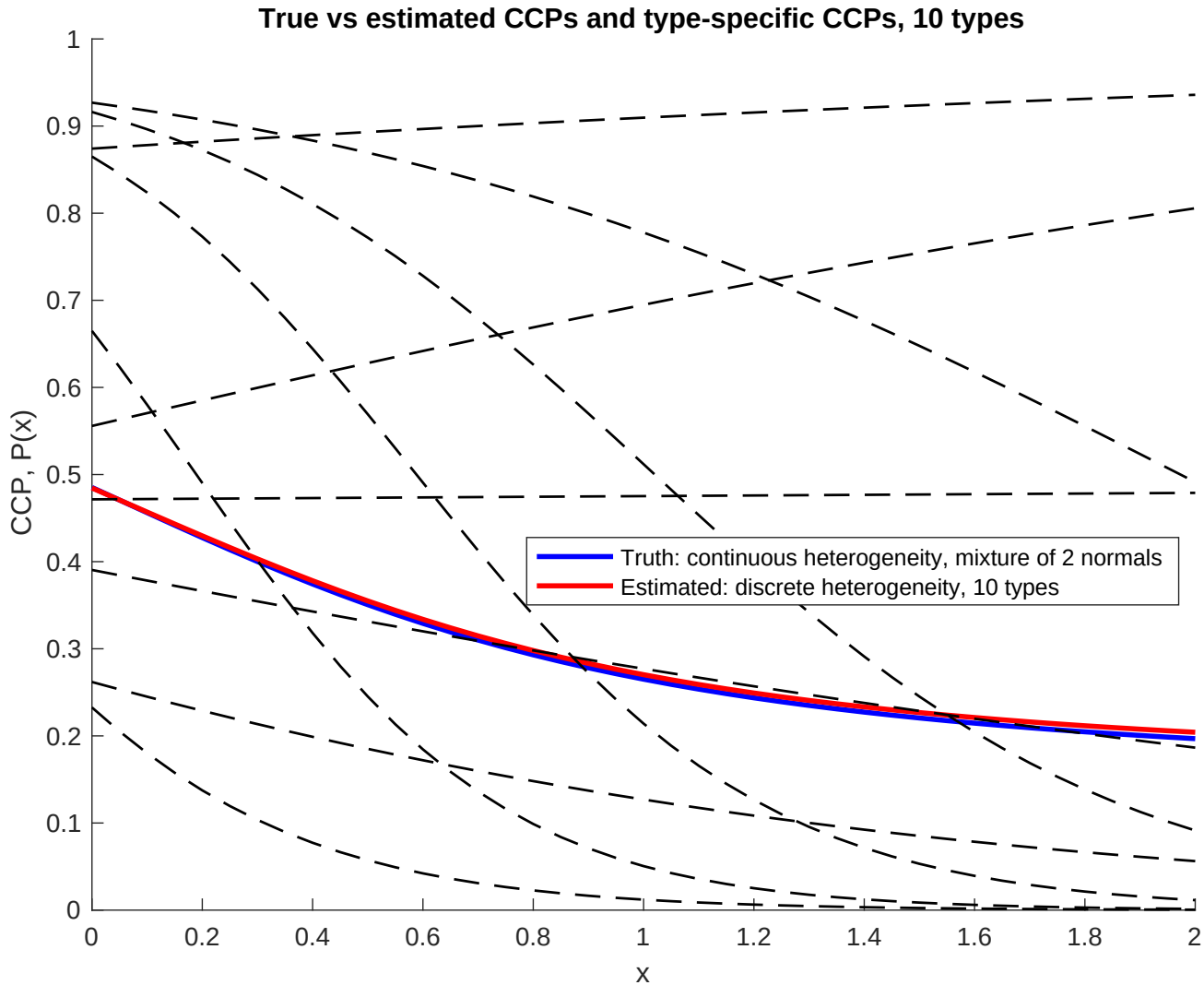
True vs estimated CCP, 4 types



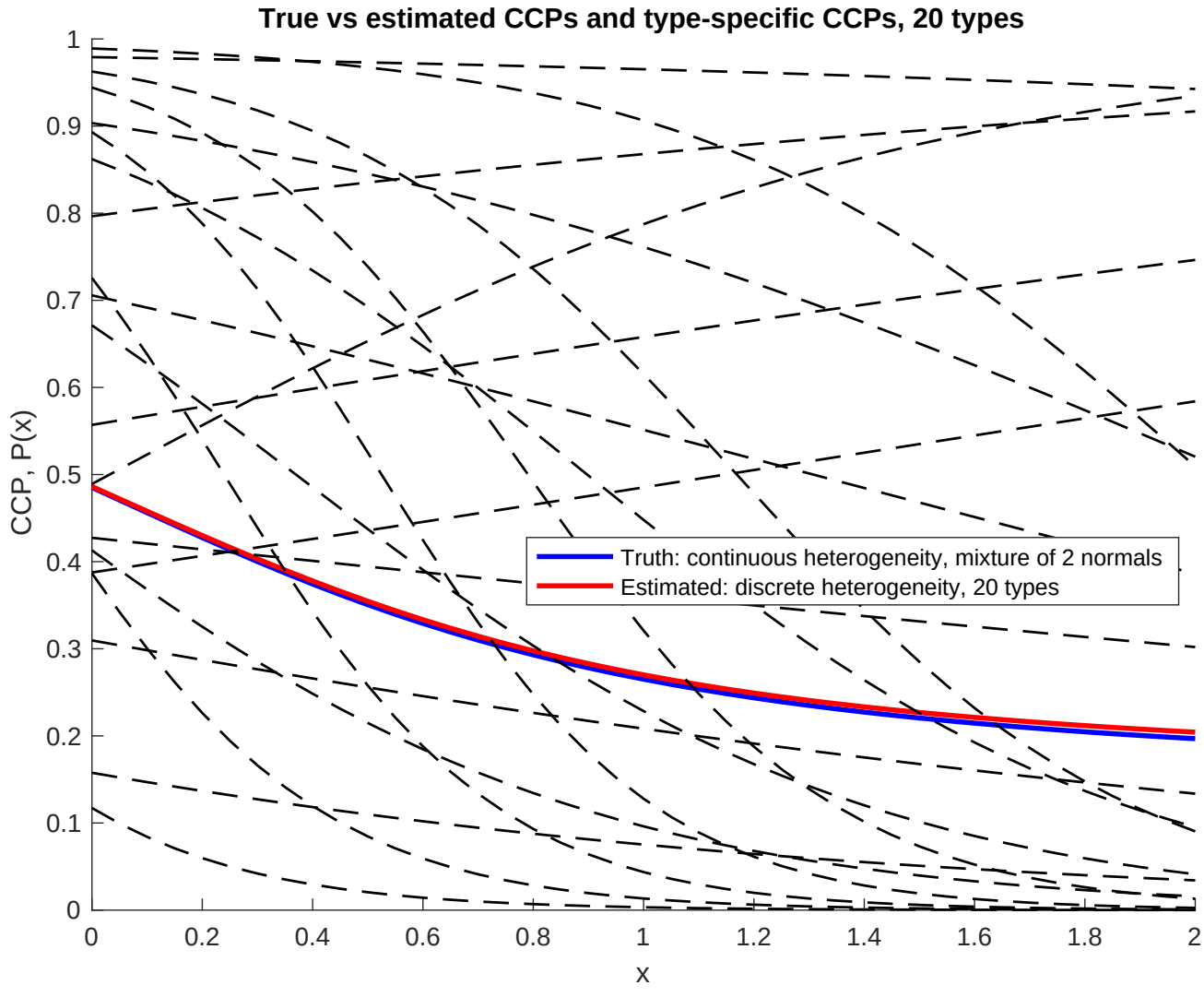
True vs estimated CCP, 5 types



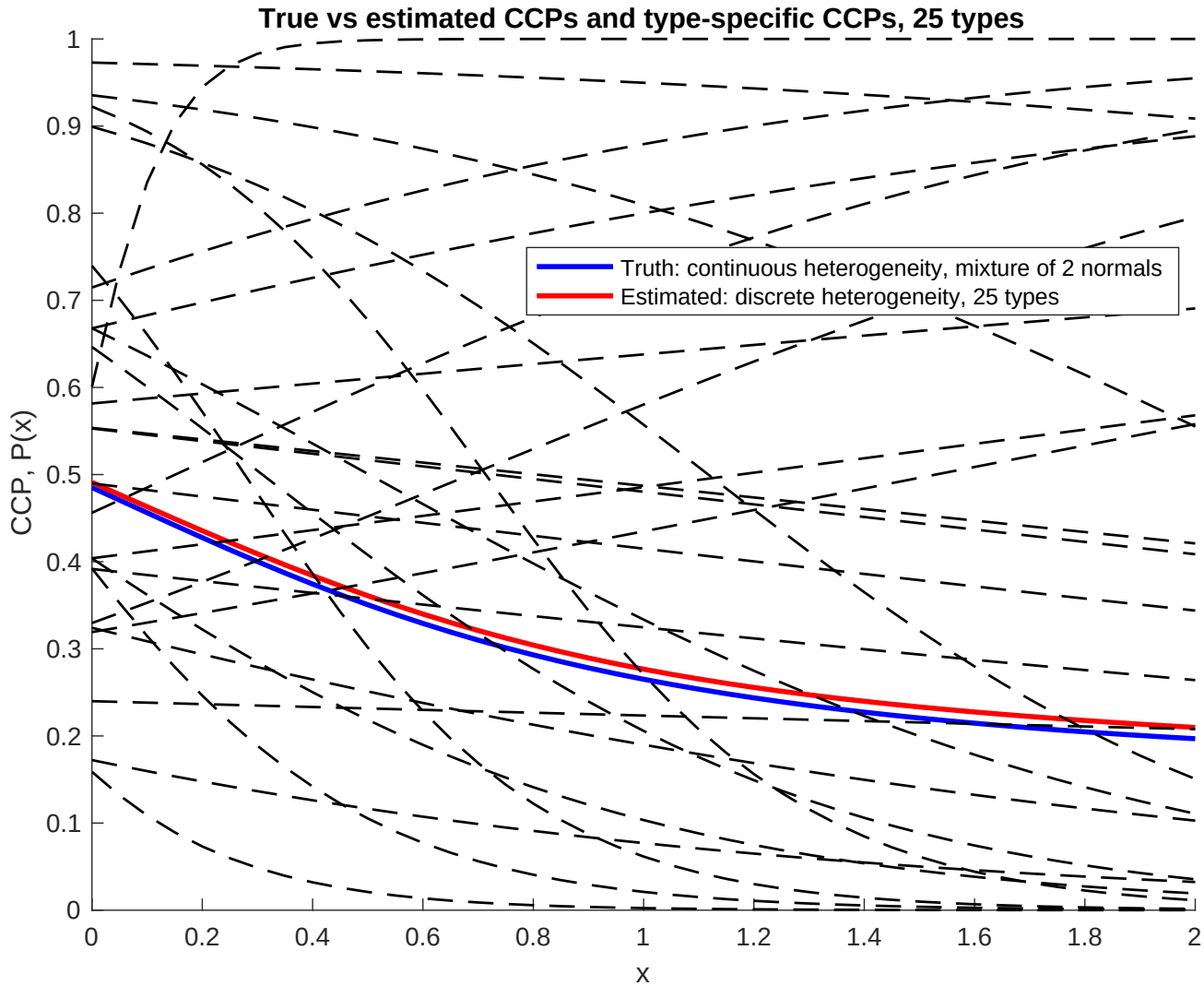
True vs estimated CCP, 10 types



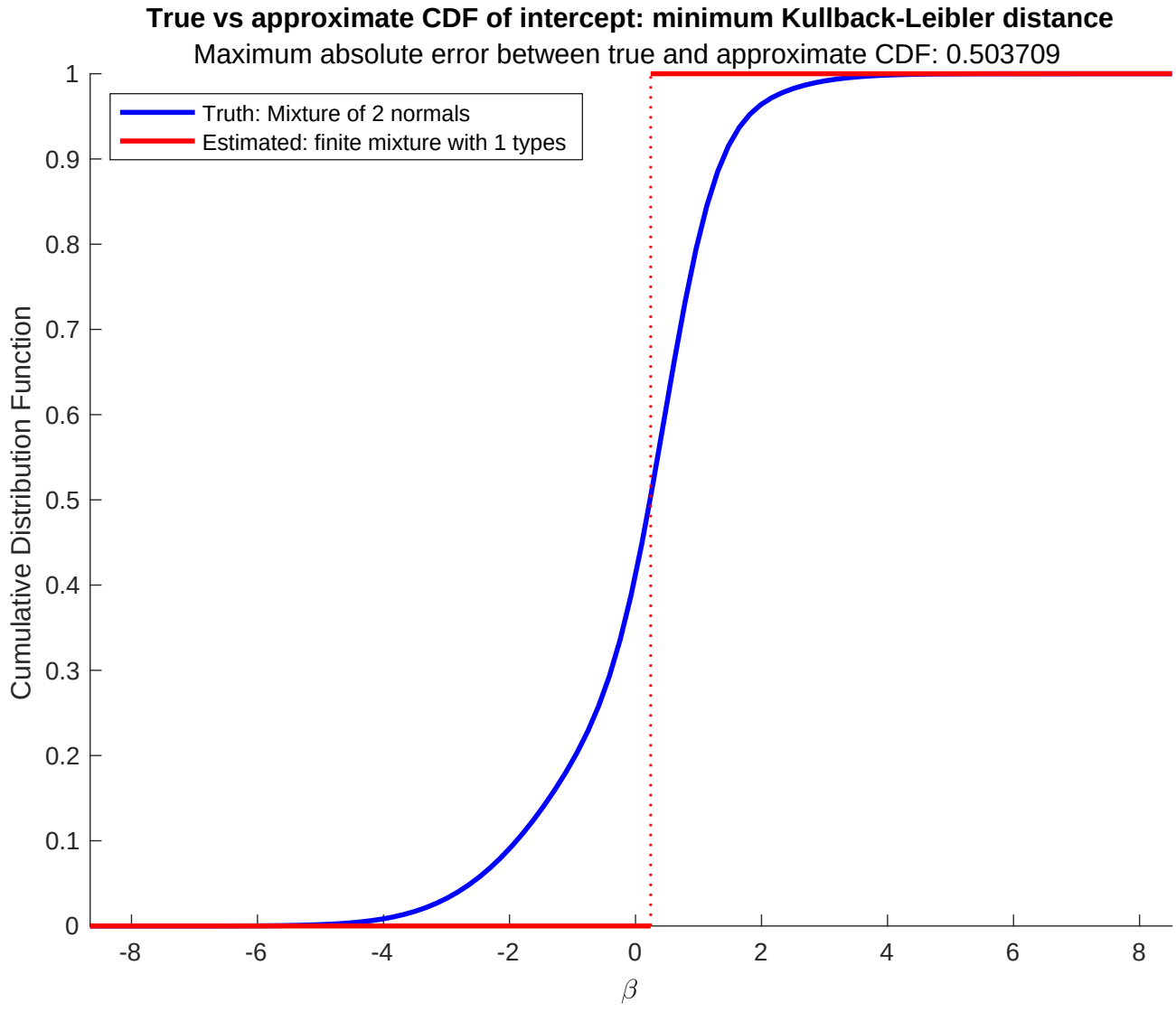
True vs estimated CCP, 20 types



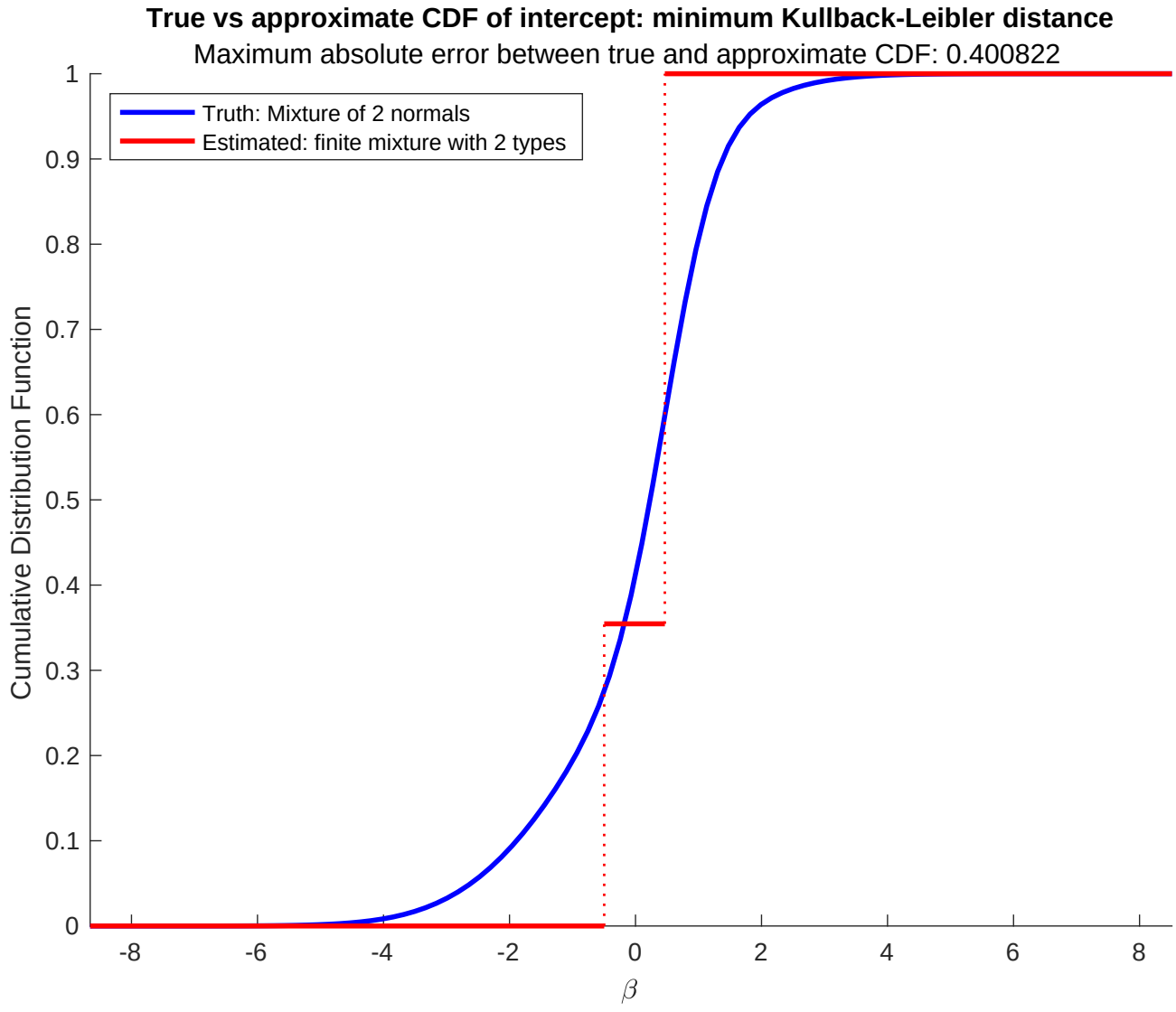
True vs estimated CCP, 25 types



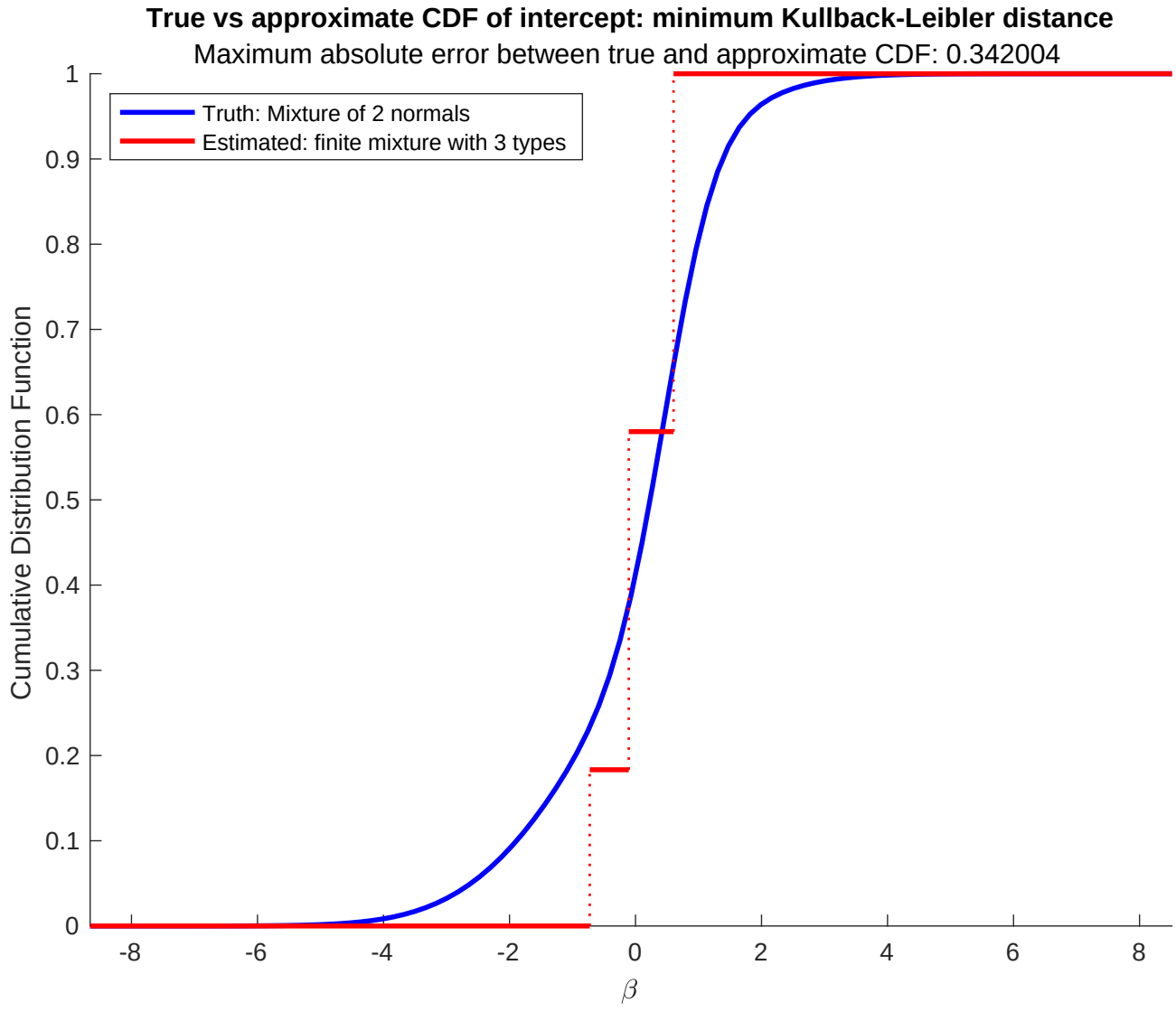
Actual vs Estimated Distribution on intercepts, 1 type



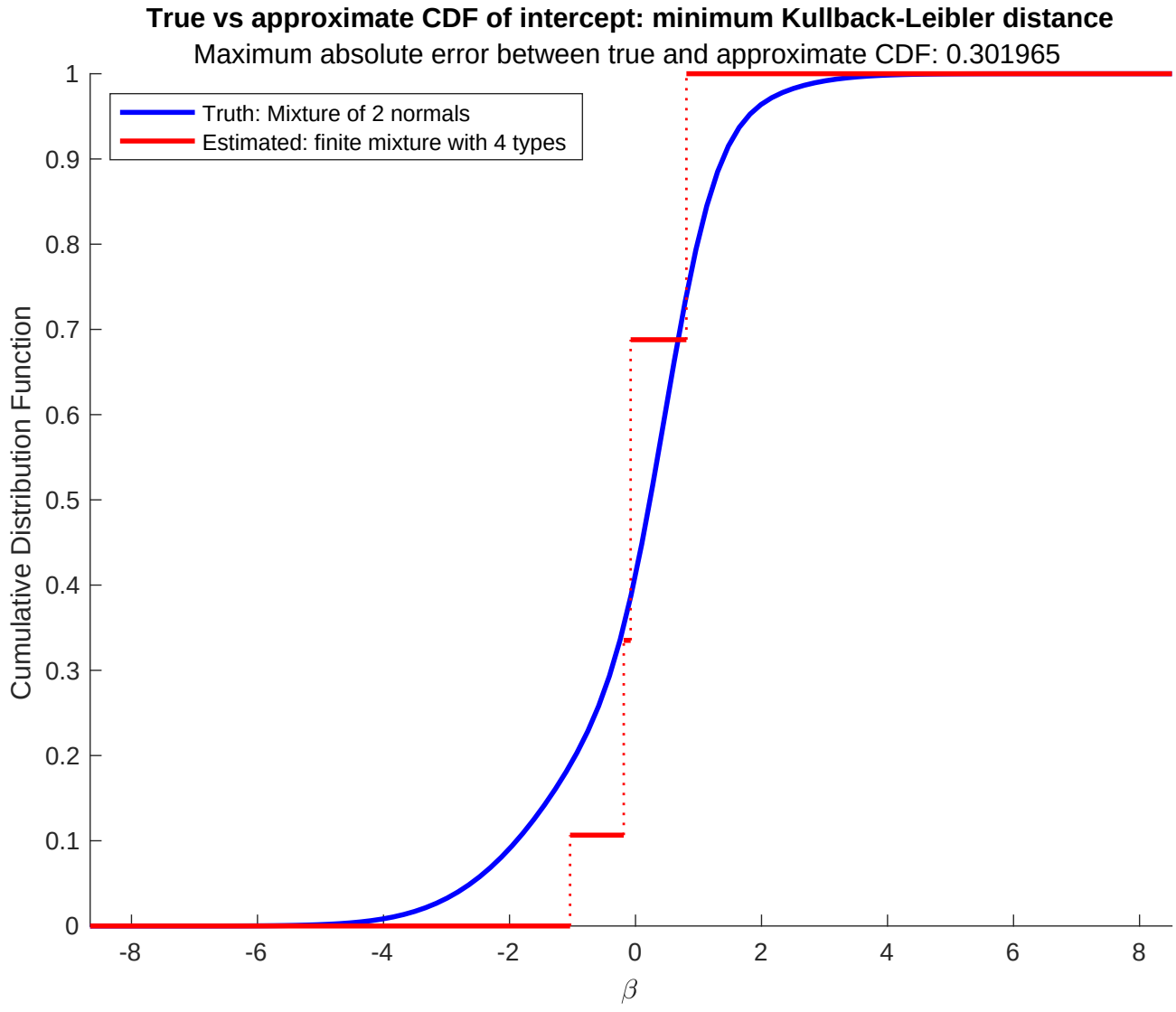
Actual vs Estimated Distribution on intercepts, 2 types



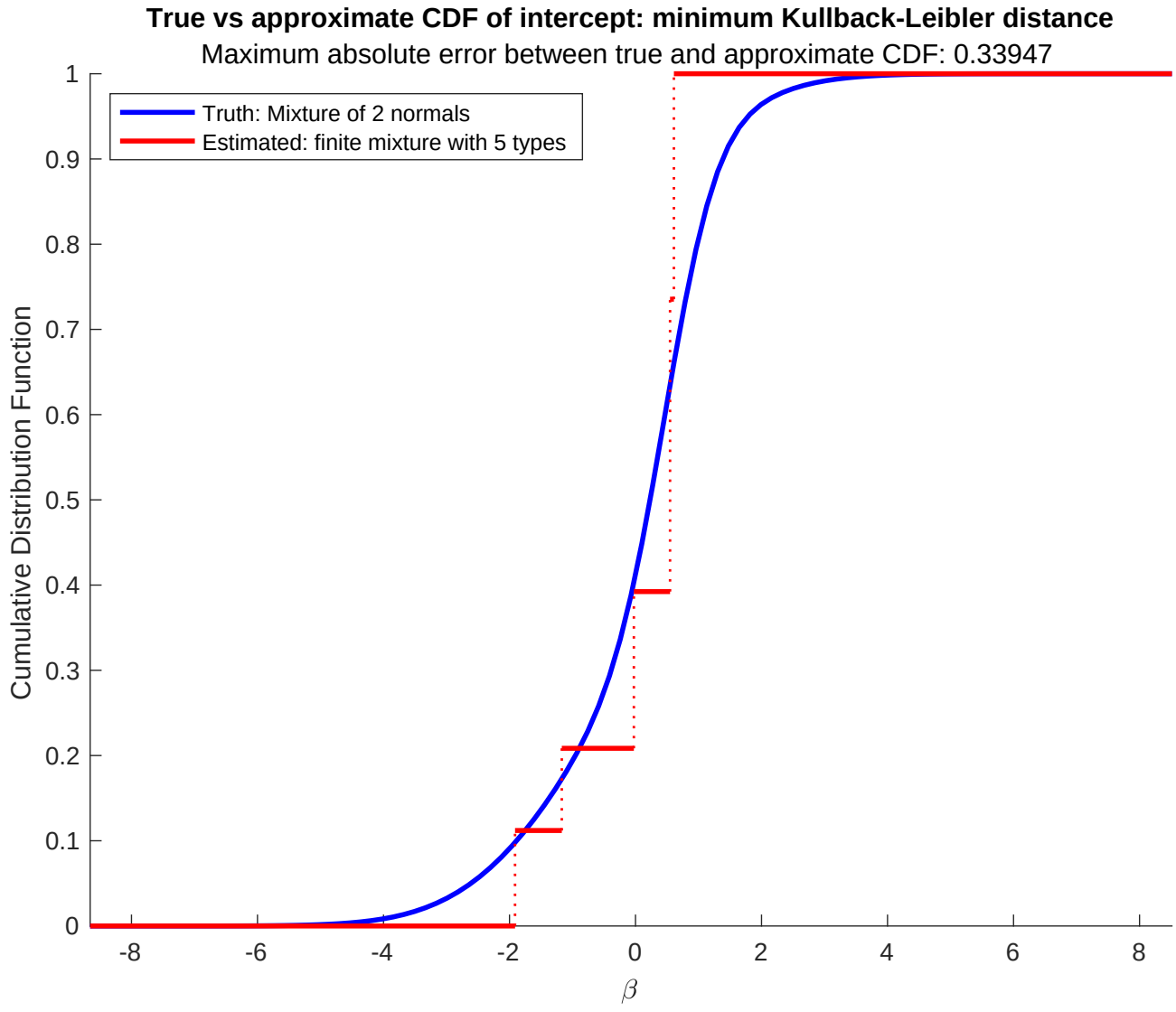
Actual vs Estimated Distribution on intercepts, 3 types



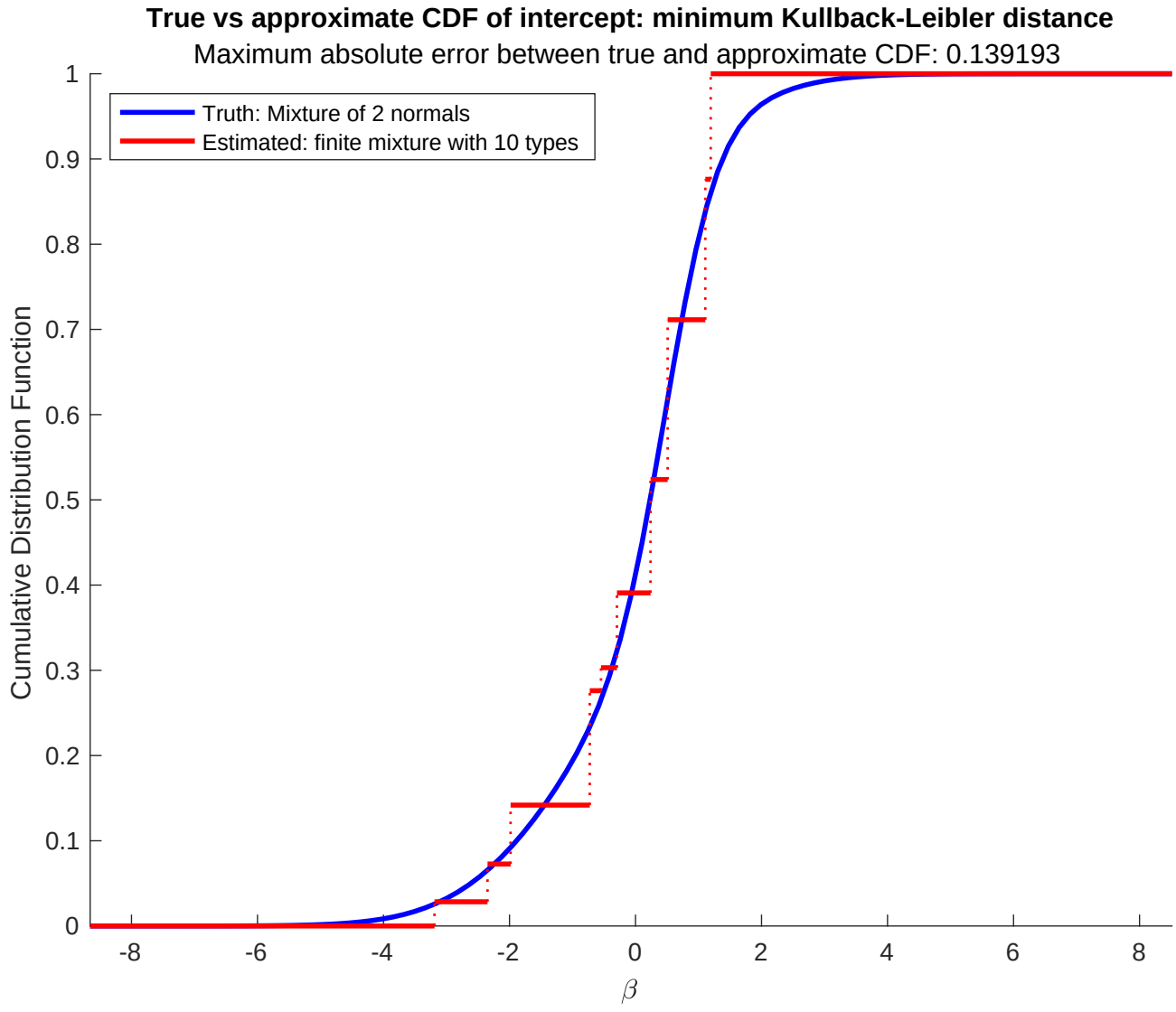
Actual vs Estimated Distribution on intercepts, 4 types



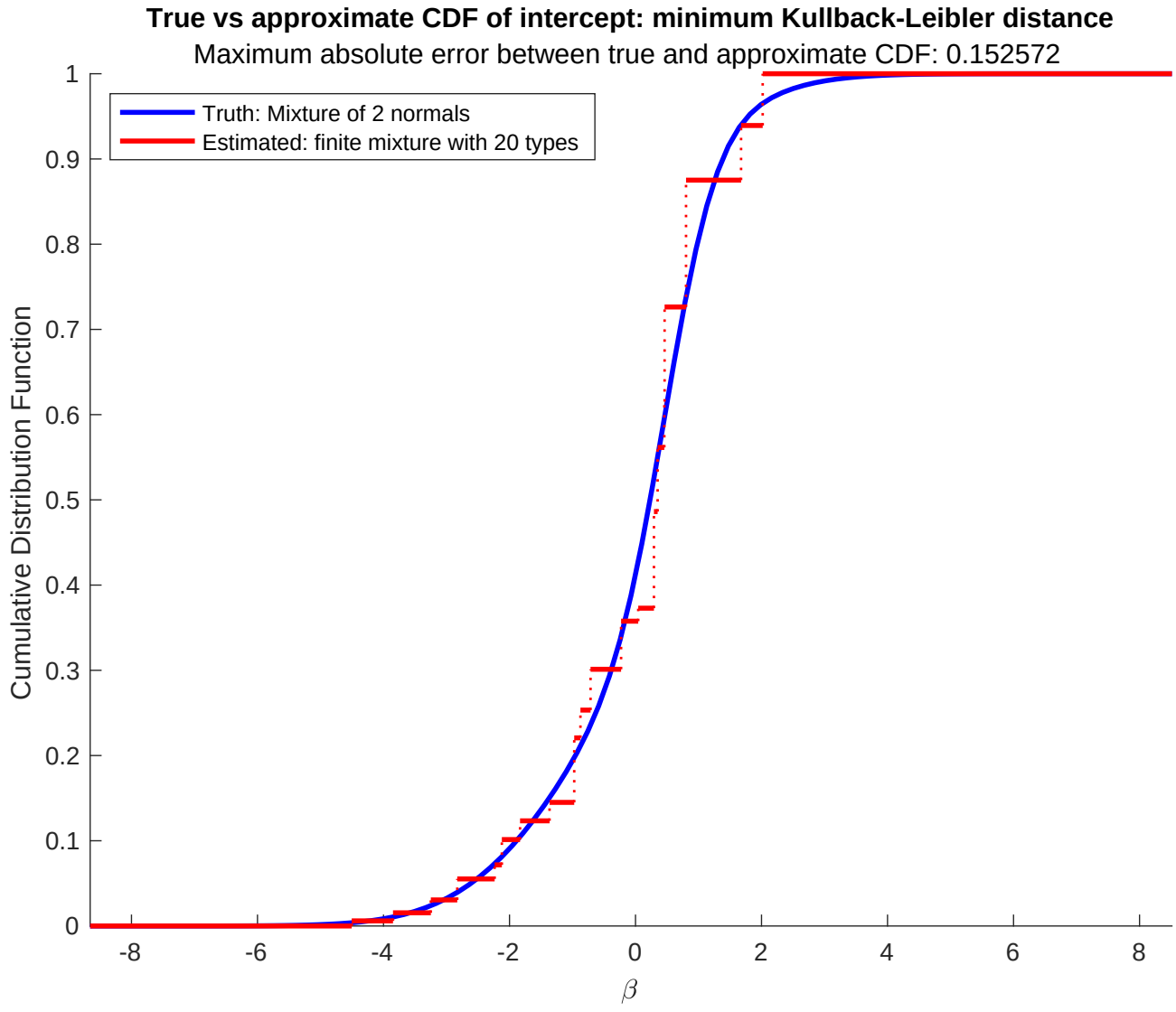
Actual vs Estimated Distribution on intercepts, 5 types



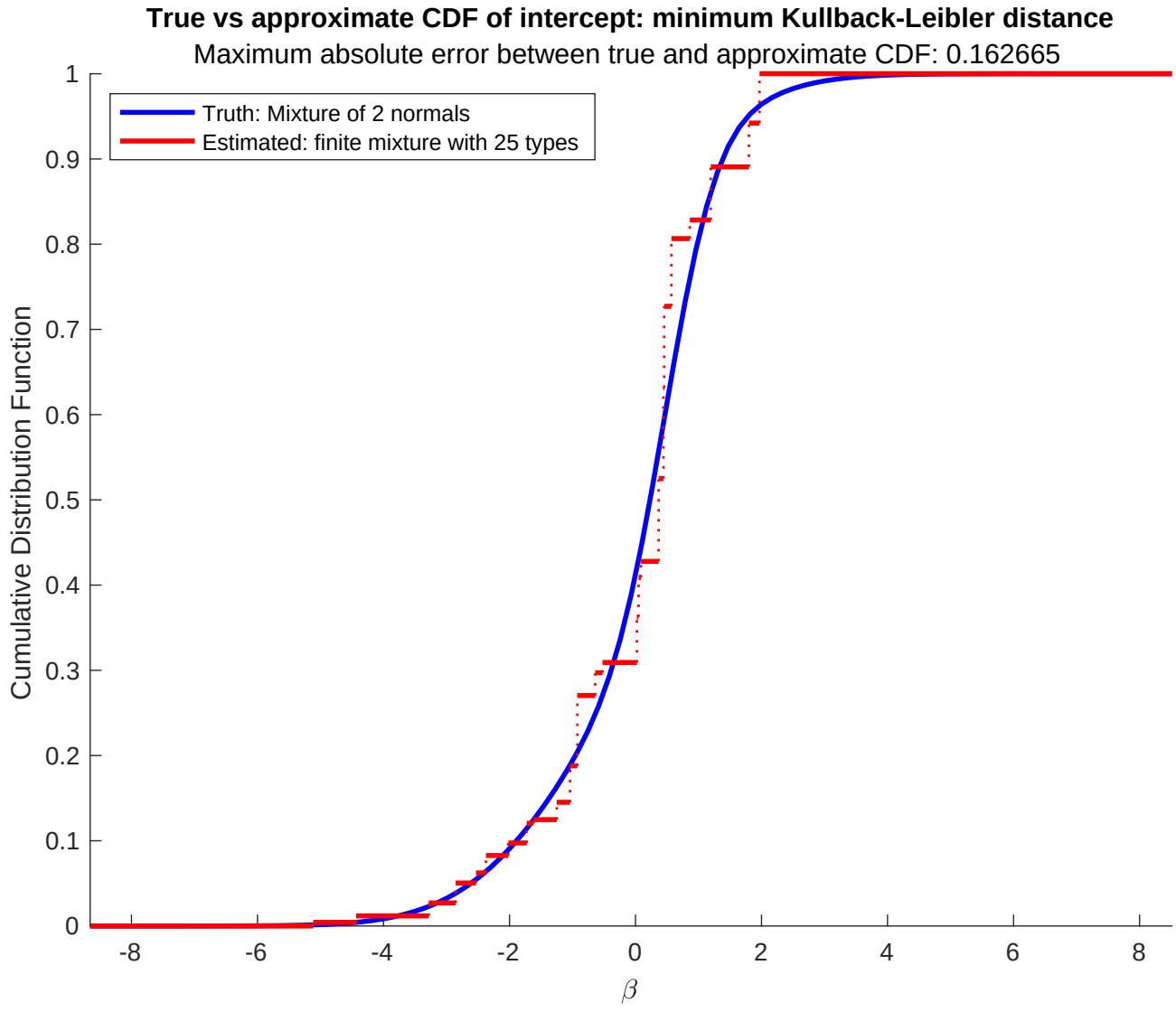
Actual vs Estimated Distribution on intercepts, 10 types



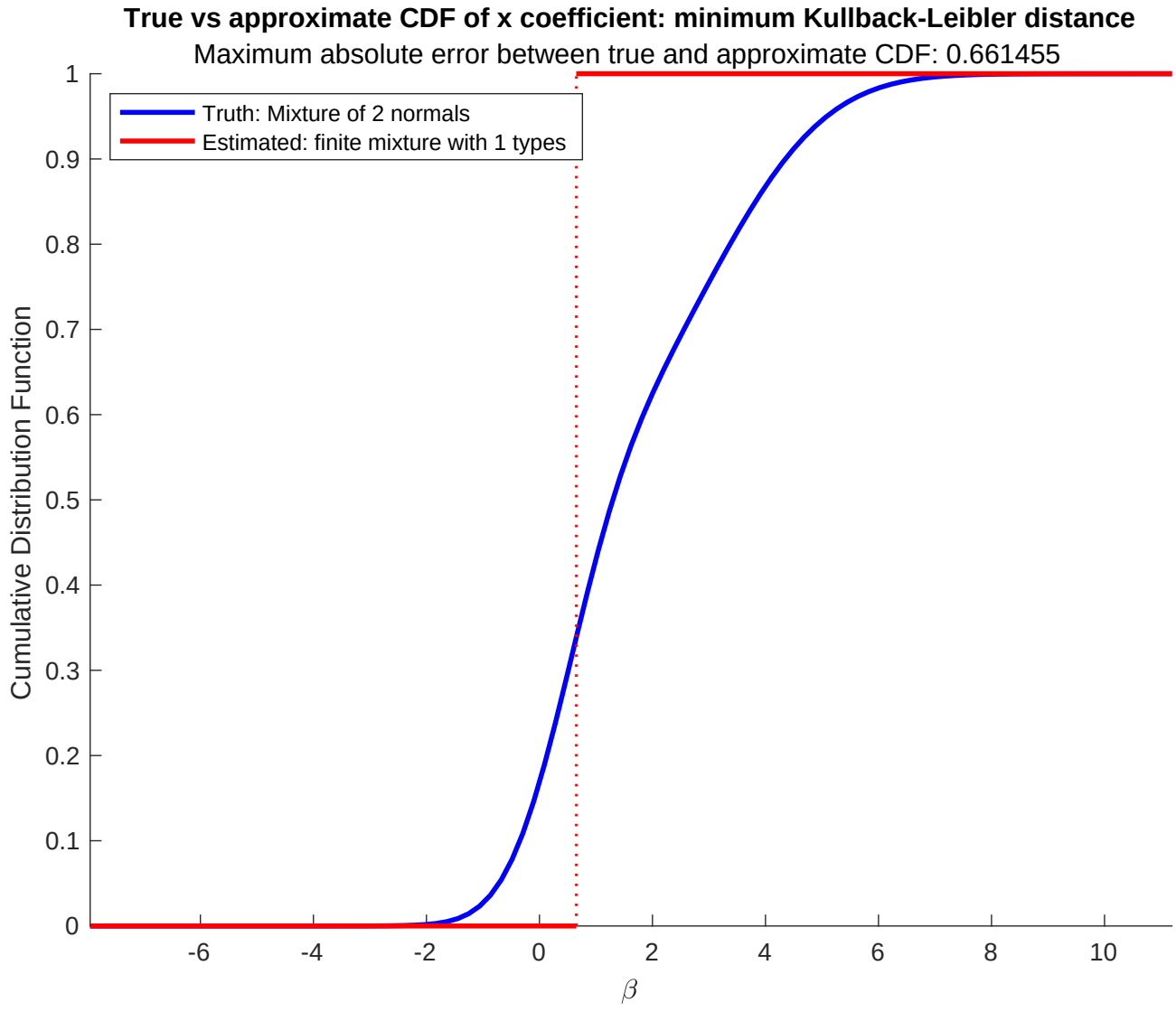
Actual vs Estimated Distribution on intercepts, 20 types



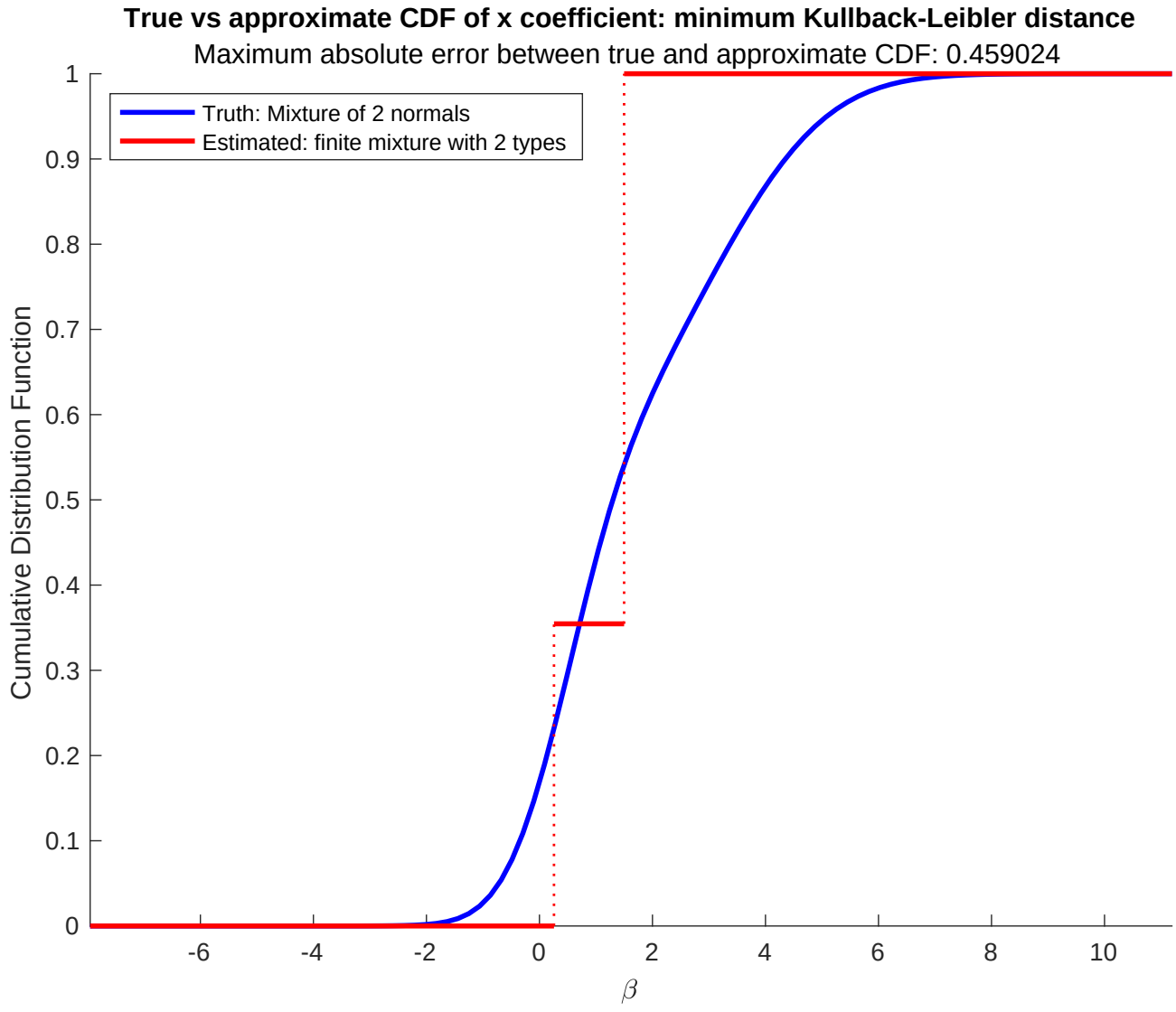
Actual vs Estimated Distribution on intercepts, 25 types



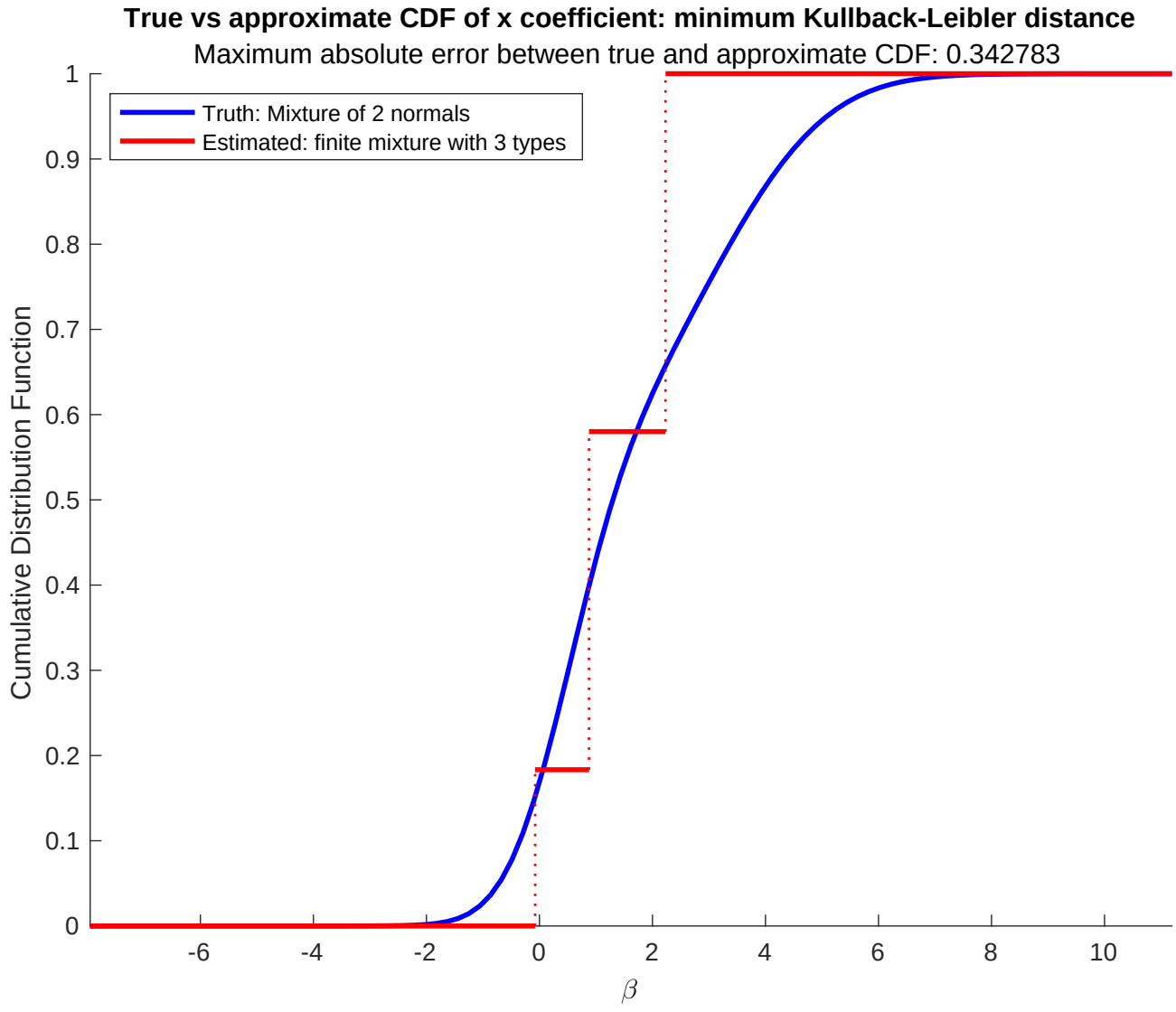
Actual vs Estimated Distribution on slopes, 1 type



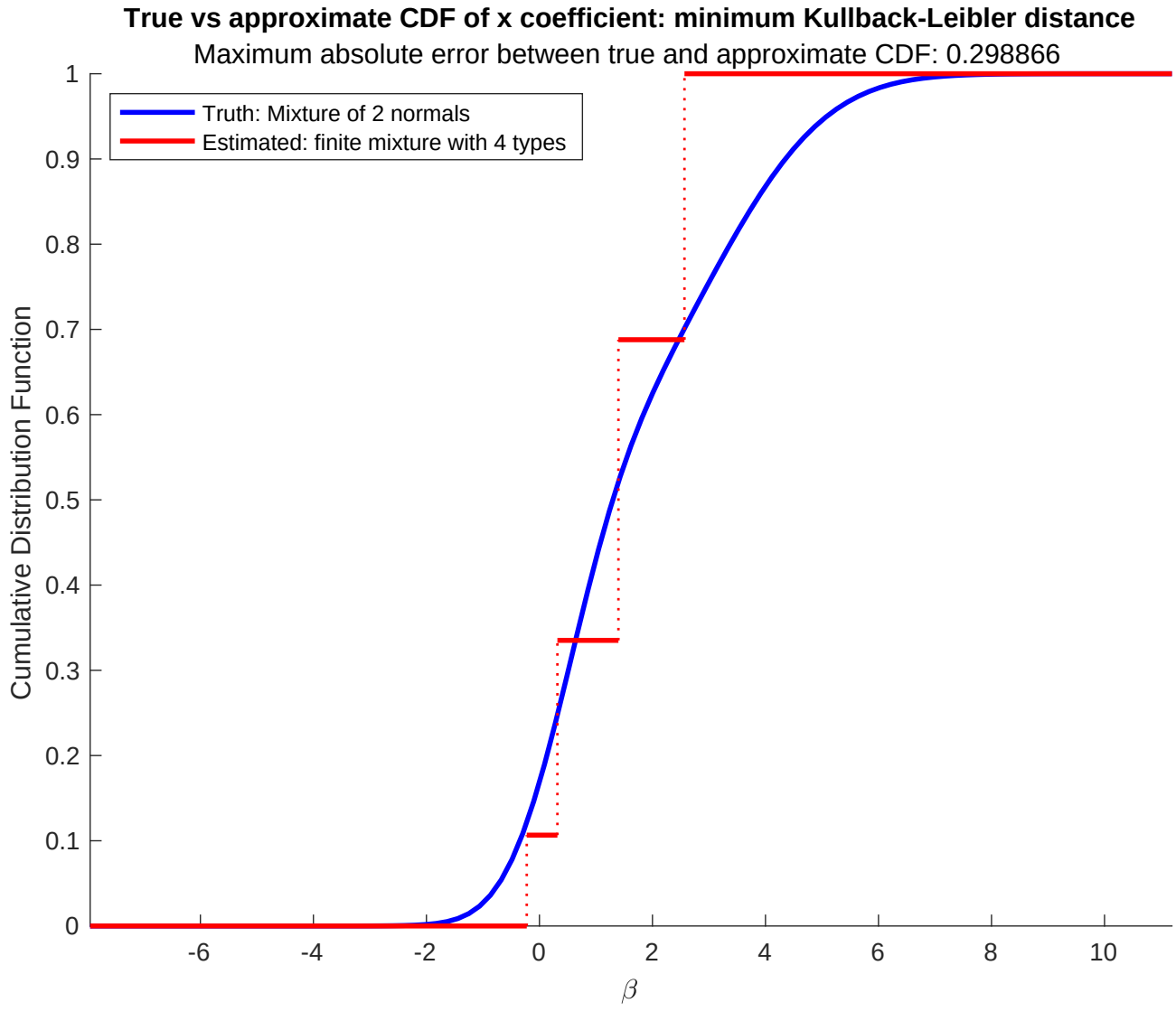
Actual vs Estimated Distribution on slopes, 2 types



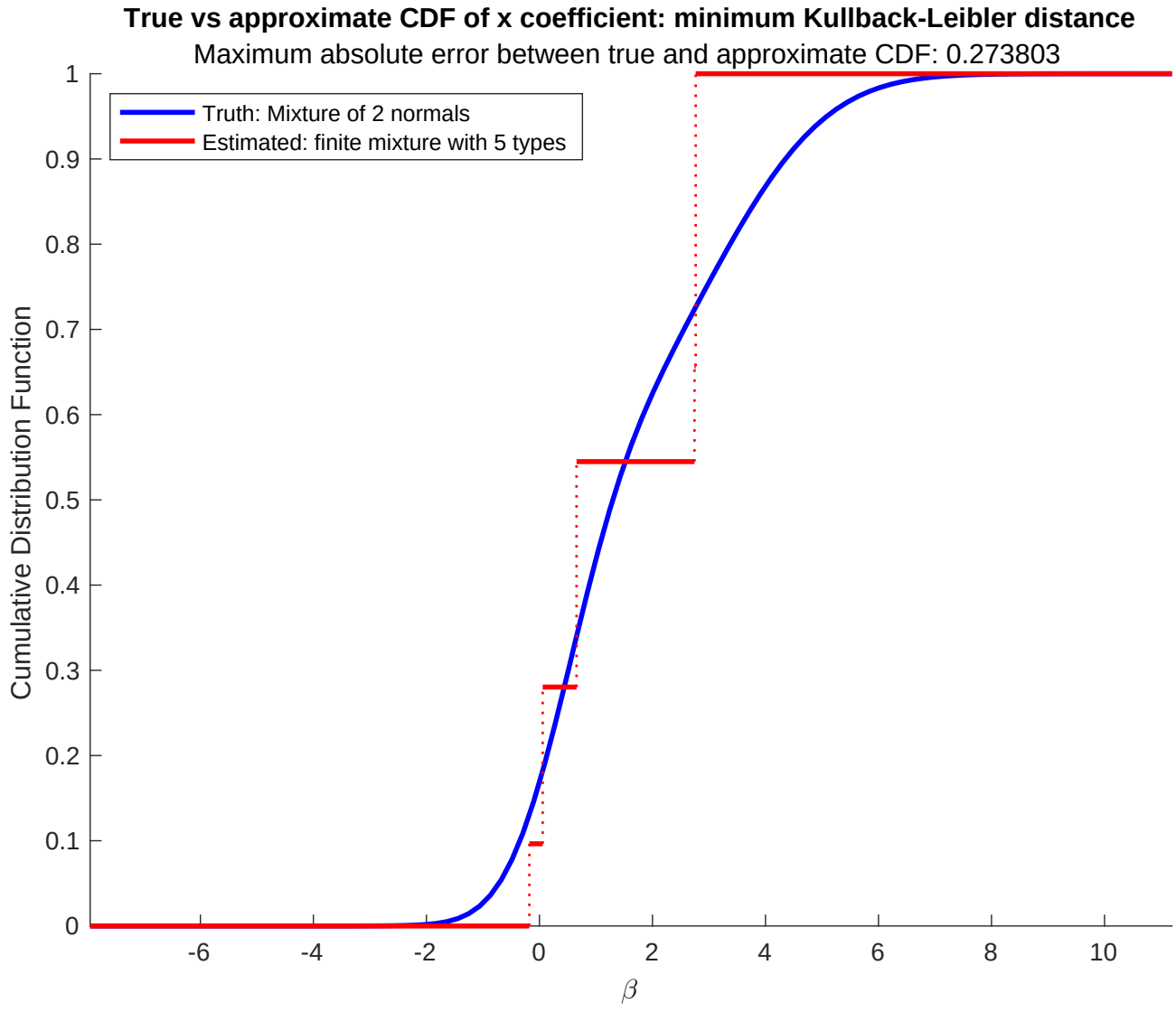
Actual vs Estimated Distribution on slopes, 3 types



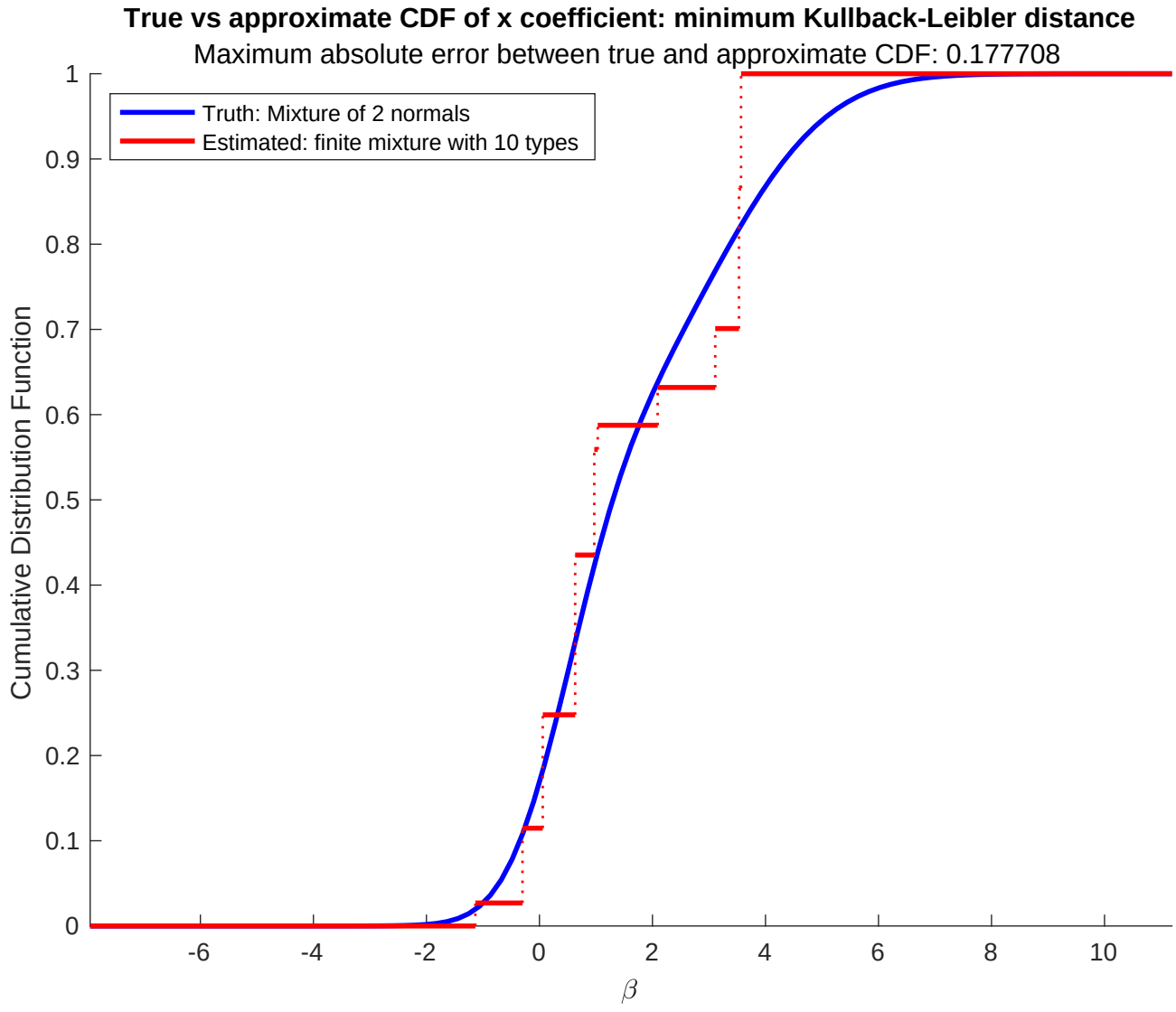
Actual vs Estimated Distribution on slopes, 4 types



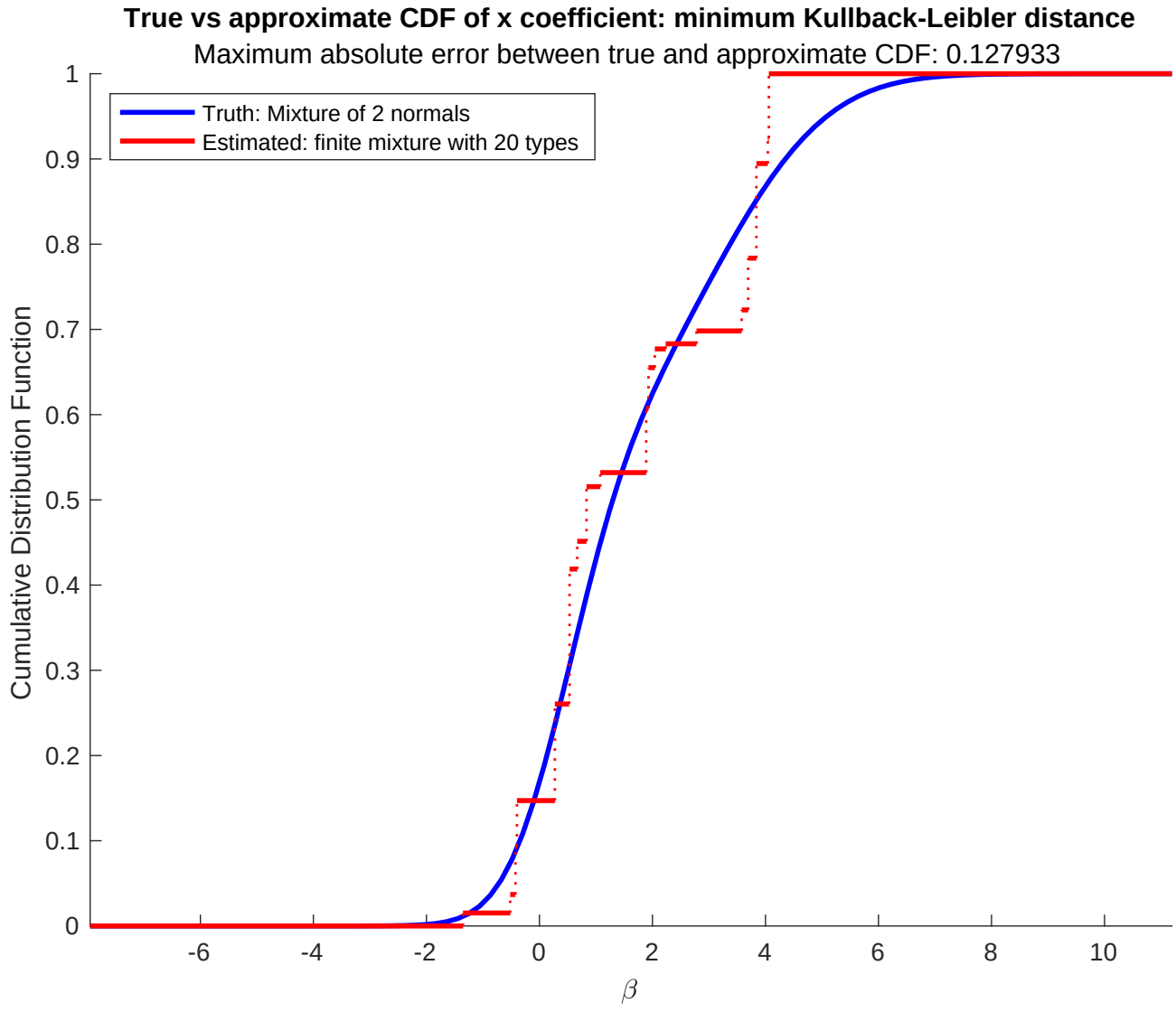
Actual vs Estimated Distribution on slopes, 5 types



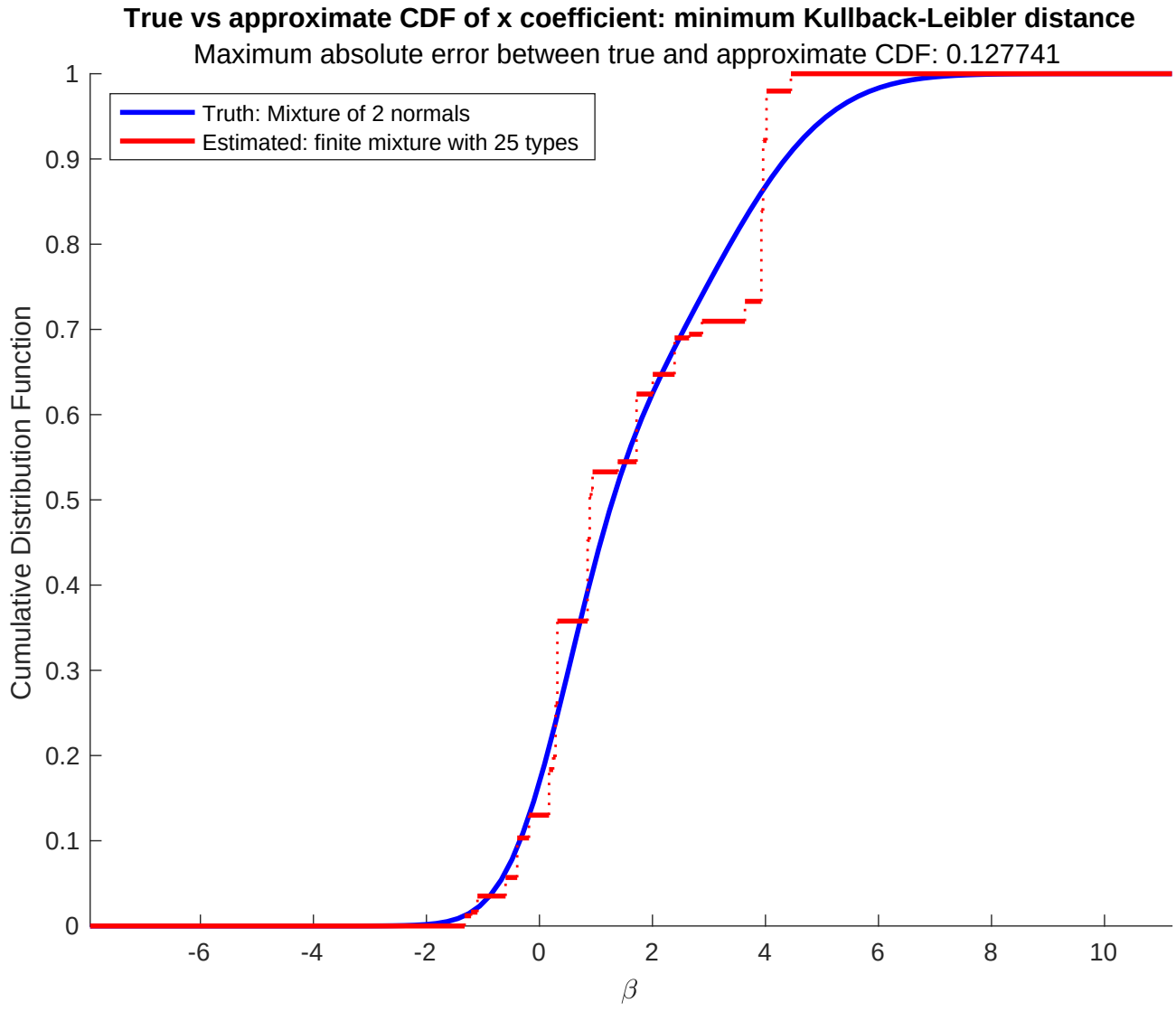
Actual vs Estimated Distribution on slopes, 10 types



Actual vs Estimated Distribution on slopes, 20 types



Actual vs Estimated Distribution on slopes, 25 types



Observed vs Unobserved Heterogeneity

- So far we have considered discrete choice models where the x vector represents *attributes* of items in the choice set, but not *observed characteristics of consumers.* Let z be a $k \times 1$ vector of observed characteristics.
- By ignoring z we are effectively assuming all heterogeneity is unobserved where all consumers have preference coefficients drawn from the same distribution $G(\tau)$.
- One way to incorporate observed heterogeneity is to allow G to depend on z, so we could estimate $G(\tau|z)$. Note that $G(\tau|z)$ is a *conditional CDF* which is much more challenging to estimate non-parametrically than a single common CDF $G(\tau)$. The methods for how to incorporate observed and unobserved heterogenity is a "frontier topic" in econometrics.
- One simplistic approach: if z are categorical variables taking M possible values $\{z_1, \ldots, z_M\}$, we can simply partition the data into the M observed groups/types and then apply the methods above resulting in different CDFs of unobserved types for each observed type m, $G_R(\tau|z_m)$, $m = 1, \ldots, M$.

Incorporating observed but not unobserved heterogeneity

- If z include continuous covariates such as income y etc, we must either "discretize" or find new ways to estimate $G(\tau|z)$ when z has both continuous and discrete components.
- Farrell, Liang and Misra (2021 *Econometrica* "Deep Neural Networks for Estimation and Inference" ignore unobserved heterogeneity and focus on flexibly and non-parametrically accounting for observed heterogeneity using deep neural networks.
- Consider the binary logit model with coefficients τ. Farrell *et. al.* allows for observed heterogeneity only by assuming that τ is some unknown function of z, we have

$$P(1|x,z) = \frac{1}{1+\exp\{x\tau(z)\}}, \tag{49}$$

and proposed "discovering" the form of this heterogeneity by specifying the $\tau(z)$ as the output of a deep neural network with inputs z. This estimator suffers from a curse of dimensionality in the sense that the rate of convergence of the DNN estimate $\hat{\tau}(z)$ of $\tau(z)$ converges at a rate that slows as the dimension of the continuous components of z increases.

Conditional finite mixture models

- We can extend the DNN approach of Farrell *et. al.* to allow both observed and unobserved heterogeneity by approximating $G(\tau|z)$ by a *conditional finite mixture* $G_R(\tau|z)$ given by

$$G_R(\tau|z) = \sum_{r=1}^{R} p_r(\gamma(z)) I\{\tau \leq \tau_r(z)\}, \tag{50}$$

  where DNNs are used to approximate the "parameters" $\gamma(z)$ and $(\tau_1(z), \ldots, \tau_R(z))$.
- The interpretation of this specification is that for each "observed type" z there are also R "unobserved types" $(\tau_1(z), \ldots, R(z))$.
- Does adding this extra flexiblity essentially gives us the best of both worlds: ability to deal with unobserved heterogeneity while incorporating observed heterogeneity in a flexible and computationally convient way?
- Intuitively, for any fixed z, we should have $G_R(\tau|z) \Longrightarrow G(\tau|z)$ as $R \to \infty$, but the question is: how fast must the parameters of the DNN grow to guarantee that this convergence is uniform over z?