Segmentation in urban labor markets: a machine learning application and a contracting perspective

Michael Kaiser[1]

[1]) Munich Graduate School of Economics
michael.kaiser@econ.lmu.de

Most recent version
This version: April 30, 2018

A significant fraction of the labor force in developing countries is "informally" employed outside the "formal" sector. Much of academic thinking uses this dichotomous view to study (urban) labor markets in developing countries, although in the presence of imperfectly enforced institutions employers and employees could choose any contract that satisfies their preferences and meets their constraints. I employ labor force survey data from Tanzania and unsupervised Bayesian machine learning to estimate the latent structure of observed contracts in the urban private sector of Dar-es-Salaam. The results suggest that around 30% of the relevant population cannot be adequately captured using a dichotomous view of the market. Controlling for employees' observable characteristics, I estimate wage distributions that suggest that workers are willing to trade off formal protections against higher pecuniary remuneration. Taken together the results suggest that a non-negligible fraction voluntarily chooses non-formal employment. An economic contracting framework is presented that formalizes the argument's underlying principles.

Keywords: Informal Labor Markets, Machine Learning, Labor Compensation, Labor Contracts
JEL codes: J46, J30, C10, J41

1 Introduction

The notion of the "informal sector" or "informal economy"[1] has featured extensively in academic and policy-focused analysis of economic strategies for developing countries and emerging markets. A widely cited number puts the size of the informal economy at 25-60% in Latin America and 13-50% in Asia (Schneider and Enste, 2000). The International Labor Organization (ILO) (2013) suggets that in some African countries the informal economy could account for as much as 70% of output. The informal economy generally features a low capital-to-labor ratio, thus offering employment for a large share of the population. While the terminology "formal vs informal" suggests a binary split, the ILO[2](2017) acknowledges significant heterogeneity in the forms of employment that can be characterized as *informal*.

The focus of this paper is employment relationships that exist within the informal economy - *informal employment*. There does not exist an universally accepted definition of the term but the ILO (2013) puts forward principles that are commonly applied in most definitions: informal employment encompasses a broad range of vulnerabilities, such as limited access to social protection, denial of labor rights, lack of organization, representation and lack of protection from economic risks. These concepts are operationalized in different ways depending on the context. In academic work it usually translates into a binary split, although the defining principles are multidimensional. This fails to incorporate heterogeneity among those for whom some dimensions but not all are met. Section 3.2 shows that this is true for about 30% of the salaried prime-aged males in Dar-es-Salaam, the capital of Tanzania. Nonetheless, academic research, both theoretical and empirical, has generated important insights relying on a binary[3] separation of employment relationships.

The empirical literature has largely focused on the extensive margin of formal employment, that is, entry and exit. Of immediate interest to policymakers is the evaluation of policy reforms intended to push people into the covered sector. Kugler and Kugler (2009) use data from payroll tax reforms in Colombia from the 1980s and 1990s to show that the elasticity of formal employment with respect to payroll taxes is .4%. More recent data from Colombia

[1] The former may be misleading as the unregulated activity is not limited to one sector of the economy.

[2] "[D]iversity and heterogeneity in the informal economy [mean that] measures to promote transition to formalization should respond to the diverse needs and situations across countries, economic sectors, contractual and occupational status and other criteria."

[3] Sometimes also referred to as the "covered" vs "uncovered" sector because of the absence of formal protections for workers.

confirms that the incidence of formal employment responds to payroll tax reform (Kugler et al., 2017). A different strand of the literature finds evidence that formal employment responds at the extensive margin to movements in the overall economic environment. Lower tariffs abroad for domestically produced tradeables decreased the share of informal employment in Brazil (Paz, 2014). Contradicting this result Goldberg and Pavcnik (2003) find that the informal sector does not respond to trade liberalization in Brazil but does in Colombia. They point to the importance of labor market institutions to understand the structure of employment relationships.

A growing number of theoretical papers study models of labor market interactions. Extended search models are used to explain the incidence of the informal economy and match stylized facts about developing country labor markets. Ulyssea (2010) fits a model with undirected search and separated markets to Brazilian data and shows that entry costs into the formal sector are a significant source of informal employment. Also using Brazilian[4] data Bosch and Esteban-Pretel (2012) calibrate a search model, in which firms can choose to contract formally or informally, to worker flows and emphasize the importance of entry costs, too. Considering the dimensions and spotty infrastructure of developing country metropolis, spatial search models have been developed to explain the presence of the informal sector through commuting costs (Moreno-Monroy and Posada, 2018). Contractual incompleteness in the sense of unverifiable effort can also lead to dual labor markets, i.e, situations in which "good" and "bad" jobs are offered to observationally similar workers (Altmann et al., 2013).

Neither the empirical nor the theoretical literature zooms into the structure of the relationships. In a seminal contribution, Maloney (2004) questioned the practice of classifying employment relationships with a binary indicator. He argues that informality may represent a choice, especially in weak state capacity environments where individuals may not trust the enforcement of provisions mandated by law. Fields (1990) was the first to argue that entry into the informal economy may represent a voluntary choice. There is some evidence that when formal jobs become more attractive, employees are willing to sacrifice wage payments to receive more secure benefits (Almeida and Carneiro, 2012). Guenter and Launov (Günther and Launov, 2012) advanced the literature by not relying on assumptions about the structure

[4]Brazil and other Latin American nations (see previous paragraph) conduct careful household panel studies which offer detailed data to study labor market structures.

of the labor market. They argue that informal employment can represent *latent segments* which comprise of voluntary and involuntary employees. Their analysis suggests that the informal sector is best described by two distinct earnings processes.

This research builds on these insights by not relying on any [5] assumptions about the structure of an urban labor market in a developing country. Fundamentally, the question that the analysis tries to answer is whether principals and agents trade off job quality and monetary remuneration against each other, i.e., high quality - low pay vs low quality - high pay. If that was the case, it would lend further support to the notion of voluntary entry into the uncovered economy. In the process of analyzing this question, the empirical methodology offers a description of labor arrangements in multidimensional space. These results suggest that there is profound heterogeneity in what could be called *informal* employment if it were coerced to a binary structure. This calls the reliance on binary splits in empirical work into question.

This work uses a Bayesian mixed-membership unsupervised learning algorithm to deduce the structure of employment relationships in a setting where the labor code is imperfectly enforced. I estimate the latent structure of employment contracts[6] in the urban labor of Dar-es-Salaam. Using labor force survey data from 2014 for prime-aged males in wage employment, the analysis shows that a simple dualistic split of employment relationships is insufficient to capture the heterogeneity in labor arrangements. The analysis suggests that somebody's labor arrangement may meet some dimensions of the aforementioned notion of formal employment but not all. This data-driven approach shows that the employment relationships are best represented by three latent structure. Incorporating the uncertainty inherent to classifying a labor arrangement, I estimate wage distributions across the three latent segments. These results suggest that there are contracts that seem to be result of trading off job-quality and wage payments. This lends further support to the notion of uncovered employment as a voluntary choice.

The next section will describe the data and illustrate issues with existing approaches.

[5] See the discussion in Section 3.1 about the structure of the segments as well as the distributions of contracts over latent segments. The estimation parameters in the main part are (currently) chosen such that the distributions of interest tend to have unbalanced mass across the support.

[6] Throughtout the text, the term "contract" is used in its economic rather than the legal meaning. A contract simply means an agreement between a principal (employer) and an agent (employee) for the purpose of employment. The notion does not imply that the agent may enforce the contract in court. A contract specifies the terms of the work environment.

The data is not free of concerns which are also discussed. Section 3 presents Latent Dirichlet Allocations, an unsupervised learning algorithm and applies it to the data. A criterion for out-of-sample performance is presented which suggests three latent segments best describe the data. Section 4 presents a simple economic framework to illustrate the mechanisms underlying the results. The last section concludes and offers implications for research and policy.

2 Data

2.1 Description

The empirical part is based on the "Integrated Labor Force Survey 2014" conducted by the Tanzanian National Bureau of Statistics. It elicits a large set of information on household demographics, wealth, socio-economic status and most importantly labor arrangements. While there exist earlier rounds of this survey, those are less extensive and are ill-suited to be collated to a repeated cross section. The sampling unit of the survey is the household but most questions are elicited at the individual level. There are 27510 adults between 15 and 65 years of age in the survey, out of which 13033 are male. For the analysis below, the sample will always be restricted to this group. The survey is nationally representative of Tanzania which is achieved by a two-stage sampling design that divides the country into 480 clusters and samples household within those. As the focus of this paper rests on (informal) salaried economic activity, the analysis will be restricted to urban areas since rural areas predominantly rely on agricultural activity. This further reduces the sample of working age adults to 9671, of which 5615 are located in the metropolitan area of Dar-es-Salaam.

In addition to those sampling restrictions all individuals who still attend school are dropped. This leave the sample at 8377 in urban areas and 4888 in Dar-es-Salaam. In order to be able to zoom in on contract configurations only individuals working for a wage are considered. This excludes those who are primarily self-employed which constitutes a large fraction of the informal sector. As noted by Bosch and Esteban-Pretel (2012) and others, registered and unregistered workers can coexist within the same establishment. Of course some establishments may only rely on unregistered workers. Therefore the sample is restricted to those individuals having indicated that their "main economic activity" is that

of a "paid employee". The final estimation sample then reduces to N=2096 in Dar-es-Salaam (and would contain N=3287 in all urban areas).

The survey elicits a large number of variables regarding somebody's work environment, such as whether employers pay income tax or contribute to social security, whether individuals have been injured at the work place or what their work hours arrangement is. A useful thought experiment is to imagine a face-to-face interview with an individual and asking her to describe her labor arrangement in great detail while coding her answers as binaries. The entire set of characteristics then mentioned by *all* individuals would determine the dimensions of the contract space (in the economic rather than legal sense). Every contract can be coded as a binary variable along each dimension. These variables that capture characteristics of a labor arrangement will be referred to as *features*. While the survey does ask a wide set of questions, it is certainly not exhaustive and does not capture all factors an individual would assess if she were to hypothetically rank jobs. In Table A.1, all 76[7] variables are reported along with their mean and their standard deviation.

2.2 Formal sector wage premiums

A recurring issue in the literature of urban labor markets in developing countries is the difficulty of defining what constitutes informal employment. While the definition given the by ILO (2013) makes intuitive sense, it is hard to operationalize the concept in empirical work. Table A.2 surveys definitions of formal employment as they have been used in academic work. Note that cross-country differences in definitions are not necessarily undesirable since institutional context does matter in defining informality. Defining informality differently in the same country may make it more difficult to compare results across studies. More importantly however, it implies a homogenous informal sector which is unlikely to be the case (Fields, 2004).

The labor force survey described in the previous section can be used to illustrate one first-order problem stemming from the use of several definitions. The conditional wage gap is defined as the wage premium a *formal* worker commands over an *informal* worker, conditional on observables. Table 1 estimates the conditional wage gap for four different

[7] More variables were initially in the set but those which have means below $.01 * n_{sample}$ and above $.99 * n_{sample}$ were dropped to limit the influence of dimensions that occur extremely irregularly or are ubiquitous.

definitions of formal employment. In addition to an indicator for formal employment, I control for education, literacy, citizenship status and seasonal wage variation by including quarter-of-the-year fixed effects. Odd columns always show results for private and public sector employees while even columns drop the latter.

Defining *formal* employment as cases when employees make social security contributions, the results indicate an extremely high wage premium of about 68% which may be as high as 89% or as low as 46% at the mean. Adding the deduction of income tax as an additional defining feature, the formal employment wage premium remains unchanged. Adding indicators for business registration and licensing to the definition would imply an imprecisely estimated wage premium between 11% and 64%. Magnitudes of around 60% at the mean are similar to other work which has found profound "formality premiums". Bargain & Kwenda (2011) find a 63%[8] premium at the mean for South Africa which is far greater than the ones they estimate for Mexico and Brazil ($\sim$ 20%). Results from 1998 in South Africa indicate a 55% "formality" premium at the mean and similar absence of heterogeneity along the distribution (Cichello et al., 2005).

Figure A.2 suggests that the above-mentioned OLS estimates do not mask heterogeneity along the distribution. The effects at the median are near the lower end of the 95% confidence interval of the estimated effect at the mean. While the fact that that mean effects do not mask heterogeneous effects at different points in the distribution, the finding that penalties tend to be higher towards the right tail of the conditional earnings distribution is at odds with existing research for South Africa (Bargain and Kwenda, 2011). However, Gong and van Soest (2002) find that the wage differentials in Mexico are positively correlated with human capital. Although most results do find that uncovered employment carries a wage penalty conditional on observeables, researches have found cases where informal workers command a wage premium (Mexico, (Marcouiller et al., 1997)). Finally, Pratap and Quintin (2006) report findings that suggest no conditional wage differentials between the two groups of workers.

This brief discussion of existing research on wage dynamics highlights the fact that cross-country comparisons have to made with care. External validity of wage pattern analysis based on a dualistic split of the workforce in "informal" and "formal" is limited at best. The

[8]This value comes from their specification that is most closely replicated in Table 1. When these authors employ panel data to account for unobserved heterogeneity, the effect for South Africa drops to about 30%.

ILO (2002) has acknowledged this point and stresses that differences in the legal framework across countries render comparisons virtually impossible. Therefore this analysis should not be taken as a universal analysis of employment configurations in urban labor markets. Rather, this analysis intends to be internally valid with respect to the labor market of the Dar-es-Salaam metropolitan area and concedes that one is unlikely to be able to extrapolate the substantive findings.

2.3 Limitations of the data

There is reason to believe that these estimates in this analysis suffer from measurement error. If we assume that public sector employees do always make social security contributions, then the exclusion of 498 public sector employees between columns 1 and 2 in Table 1 should not affect the number of individuals classified as informal. However, the number of formal employees only changes by 320, implying that dropping public sector employees also led to the exclusion of some employees classified as informal according to the respective definitions. This hints at measurement error in the sense that some individuals may not be aware of the exact provisions in their labor arrangement. If that is true for public sector employees there is reason to suspect that private sector employees may be unaware of their employer contributed benefits, too.

The ideal dataset for this analysis would be both vastly longer and wider. The algorithm was originally introduced for text data analysis (see next section) and speech tends to span high dimensional space. An ideal dataset would contain language snippets of workers who describe their labor arrangement in free language. This data would be worked into a discrete matrix of word frequencies. One could think of the binary matrix above as coding somebody's language purely as having mentioned a particular fact at most once.[9]

The Labor Force Survey 2014 was conducted on a nationally representative sample. This includes vast rural areas whose industrial structure is heavily dominated by (subsistence)-farming. Other urban areas (Arusha) are available as well but industrial activity tends to be centered in the captial Dar-es-Salaam. A variety of sample restrictions had to be imposed to obtain a set of individuals facing a comparable institutional framework

[9] If I were to have actual text data, it would be possible that somebody stresses a particular work feature by mentioning it repeatedly. The data matrix would then no longer be a matrix of indicators but rather of discrete variables counting feature (word) occurrences across contracts.

and cost of living[10].

3 Empirical analysis

3.1 Latent Dirichlet Allocations

The goal of the analysis is to describe labor contract configurations in low dimensional space in order to understand trade-offs resulting from principal-agent interactions. Recall that the data is informative of 76 characteristics of an employee's contract configuration. In order to reduce the dimensionality of the contract space, an unsupervised learning algorithm is employed. A Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a Bayesian mixture model which provides a generative description of uncorrelated latent dimensions[11]. The algorithm is initially developed for text data and is used to find underlying "topics" across text snippets. In principle the algorithm can be used for any kind of discrete data. While widely used in linguistic and statistical research, applications have been sparse in Economics. Recently the procedure has been used to classify CEO behavior (Bandiera et al., 2017) or study the effect of central bank committee deliberations (Hansen et al., 2014).

Throughout the analysis the following notation and generative process for a contract (the unit of observation) is assumed. Each contract $w_i = (w_1, w_2, ..., w_F)$ of the universe of contracts N is a collection of F different (contractual) features. This is to say that each observation is a contract indexed by i which is described by collection of contractual features F index by f. Denoting the number of contracts by N, the resulting data matrix has dimensions $NxF = 2096x76$.

The two distributions that LDA tries to estimate are i) a distribution over all features for each latent segment and ii) a distribution over all latent segments for each contract. The former can be thought of as describing probabilities that a given latent segment generates a particular contract feature f. The latter can intuitively be thought of giving shares for each latent segment in a contract. Note that LDA is a mixture model and hence contracts

[10]Ideally, I would like to use a dataset such as "Informal Sector Survey" conducted in the Philippines in 2008 which asks even more questions about contractual arrangements. More importantly, the number of observations is vastly higher ($\sim$ 20000). However, data acquisition has stalled and therefore I have been analyzing the data at hand in order to assess the feasibility of the approach

[11]The models in the following sections are estimated using the *topicmodels*-package implemented in R (Hornik and Grün, 2011).

will not be deterministically assigned to latent segments. The probabilistic nature reflects uncertainty in the contract-segment assignment. The data generating proceeds by specifying the following two priors:

$$\beta \sim Dirichlet(\delta)$$
$$\theta \sim Dirichlet(\alpha)$$

β describes the feature distribution for each segment while θ describe the segment distribution for each contract. The DGP for a contract i then proceeds by sampling each of the F features w_f by

1. Sample a segment s_f according to $z_f \sim multinomial(\theta)$
2. Sample a feature w_f from a multinomial distribution conditioned on the previously sampled topic s_f and β which gives the probability of feature occuring in a segment; $p(w_f|s_f, \beta)$.

The values for δ and α are referred to as hyperparameters and represent the researcher's prior beliefs about the structure of the segment-over-feature distribution β (which is FxK) and the contract-over-feature distribution θ (which is NxK). The Dirichlet-hyperparameters govern the dispersion of the distribution in space with lower values leading to a distribution with fewer points of great mass as opposed to the many points with similar mass. For instance, a low α represents the belief that that each contract is more likely to be generated by fewer rather than more segments. Figure A.1 shows how mass points tend to vary in a two-dimensional Dirichlet distribution given different values of the hyperparameter.

For this analysis, the hyperparameters reflect the prior beliefs that both distributions tend to have their mass on a few points rather than similar mass on all points. This is to say that the priors reflect the belief that contracts are generated by a few segments rather than many. Similarly, segments are characterized by a limited number of features. Therefore, the hyperparameters were set to $\alpha = \delta = .1$. While $\delta = .1$ has been suggested in the literature (Griffiths and Steyvers, 2004), $\alpha = .1$ is lower than what the same authors suggest. They propose as a rule of thumb $\alpha = 50/K$ (K being the number of latent segments) which would imply $\alpha = 16\frac{2}{3}$ with three latent segments.[12]. Do note that this Bayesian procedure never

[12]Their analysis concerns text data for which the data matrix is significantly wider than in this application.

puts zero mass anywhere in the distribution. A non-zero prior ensures that no distribution has zero mass anywhere. While the distributions of interest may be polarized, they have non-zero mass on all points (e.g., each segments' share in the generation of a contract is non-zero) and they add up to one.

Estimation of the distributions β and θ is the goal of the analysis. Estimation and inference is done via a Gibbs-sampling procedure which essentially inverts the aforementioned data generating process. By starting from the random prior distributions, Gibbs-sampling is a Markov-Chain-Monte-Carlo method which attempts to maximize the likelihood of the data generating process over the distributions β and θ. The MCMC-procedure relies on a "burnin"-period in order to allow the sampling chain some time to stabilize around the "true" value. Having reached this point, a larger number of additional chains is sampled. Each of these chains has a likelihood associated with it but in order to reduce autocorrelation, the chains are "thinned" by evaluating only - for instance - every 200^{th} chain[13]. This procedure is then repeated for a number of random starting points. For the analysis below, the "burnin" period is usually set to 1000 iterations after which another 2000 iterations are completed where only every 400^{th} is taken (Griffiths and Steyvers, 2004). 40 random start points are evaluated.

Arguably the most important input for the estimation of the Latent Dirichlet Allocation is the number of latent segments denoted by K that are believed to have generated the observed data. This is a first-order choice the researcher makes. However, there are criteria which can be employed in order to inform the choice. Firstly, economic theory can guide the choice of K. If for instance we were to believe in the classic "queuing hypothesis", i.e, that workers queue in the informal economy until they are matched with a formal job, we may want to estimate a model with two latent segments (see Section 3.2). Likelihood based criteria also exist which try to find the "optimal" number of latent segments using a training dataset such that the out-of-sample performance of the model is maximized. Section 3.3 will present one widely used criteria in more detail.

Therefore, it is reasonable to assume that that latent segments in those cases represent mixture distributions in which more features (i.e., words) carry loadings.

[13]These steps is described in more detail in Griffiths and Steyvers (2004)

3.2 Dualistic view

The dualistic view[14] suggests a labor market that is divided into a formal sector offering jobs in compliance with labor regulation and an informal sector offering inferior jobs for those not participating in the formal economy. This view has been criticized as informal employment sometimes constitutes a deliberate choice and informal jobs come in a variety of configurations (cf Maloney (2004) or Fields (2004)). However, if we were to take the dualistic view at face value, we could use LDA to learn about the systematic configuration of labor contracts across the two segments.

Figure A.3 plots the the distribution of $\beta^f_{k\in\{1,2\}}$ which describes the makeup of a segment as a distribution of contractual features. Within one segment, the distribution of loadings will sum to one and features that have higher loadings indicate that contracts generated from that segment are more likely to contain these features. Note that the labels that the algorithm assigns to the two segments are not ordinal, i.e., there is no objective criterion by which the two segments can be ranked. However, the segment-feature probabilities can be used to assess the composition of contracts representative of that segment. For instance, contracts generated by segment two tend to be written, stipulate work for all months of the year and include deductions for income and social security. Segment one tends to generate contracts that do not include the possibility of maternity leave, no social security deductions and involve businesses older than 10 years. In order to assess the key differences between the two segments, it is informative to analyze the per feature segment difference, i.e., the probability that a certain feature is generated by segment one or two. Figure 1 plots $\frac{log_2(\beta^f_{k=1})}{log_2(\beta^f_{k=2})}$, i.e, the log-2 difference between feature-segment distributions, with values larger than zero indicating that a feature is more likely to be generated by segment one and vice versa. A value of zero indicates that a feature is equally likely in either segment. By analyzing the right tail of Figure 1 informs about contract features that a more likely to be generated by segment one. These include verbal (or casual) contracts in firms with less than five employees. Employees tend to rate their job security as unreliable. On the other hand, the left tail of Figure 1 indicates that the social security contributions and permanent contracts are much more likely to be contractual arrangements in segment two. This segment also

[14] Maloney (1999) for instance describes this view as a formal sector with good jobs and informal workers in jobs with no quality standards who receive fewer benefits, earn lower wages, and endure worse working conditions

captures public sector jobs in local and central government. Subjectively, Figures 1 and A.3 suggest that segment two captures what we would term the *formal* segment while segment one corresponds to the *informal* sector.

Another key quantity that is estimated by the LDA algorithm is the matrix $\gamma^i_{k\in\{1,2\}}$. i.e., the contract-over-segment distribution. This Nxk matrix provides the probability that a given contract is generated from the respective segment. In the case of two latent segments, the probability can be interpreted as a share since the two probabilities will sum to one (Bandiera et al., 2017). Building on the (subjective) judgment that segment two corresponds to a notion of better-quality jobs (see Figure 1), the probability that a contract is generated by segment two measures the degree of formality of an employment relationship on a continuous scale between 0 and 1.

Figure 2 plots both the cumulative density of the probabilities that a given contract is generated by segment 2. Additionally, the plot shows the smoothed distribution of residualized and 99%-winsorized wages[15].

Firstly, the left and right portions of the cumulative density (dark blue) in Figure 2 indicate that for 42% of observed contracts in the sample the likelihood that is was generated by segment two is extremely low ($<\sim .007$) while 31% of all contracts are extremely likely ($>\sim .993$) to have come from the feature distribution of segment 2. These are the two portions separated by the dashed black lines. Under the assumption that $\gamma^i_{k=2}$ does in fact provide a measure of formality assuming a labor market with two latent segments, the cumulative distribution's left tail in Figure 2 is an estimate of the share of informally employed but salaried individuals. The estimate would suggest a size of salaried informal sector of roughly 40%.

Given the sampling restrictions (Dar-es-Salaam, work for pay, male, 15-65), it is challenging to find comparable estimates of the prevalence of informal employment. Using data from 2006, the ILO (2012) puts the percentage of male informal employment in Tanzania at roughly 70%. This is comparable to the proportion of contracts that have a segment two

[15] The raw hourly wage is 99% winsorized, meaning that values below the first percentile are replaced by the wage at the first percentile. Similarly, wages above the 99^{th} percentile are replaced by the wage at the 99^{th} percentile. These wages are then regressed on indicators for literacy in Kiswahili, English or both, an indicator for being a citizen of a foreign country, indicators for secondary or university education separately, quarter-of-year interview date fixed effects and a constant. The set of control variables mirrors those used in Table 1. Figure 2 plots then the residuals from this regression to which the mean of the initial dependent variable is added back.

share below .99. Charmes (2012) reports that the share of informal employment in total non-agricultural employment in Tanzania was 46% for the period 2005-2010. The difference in these estimates may arise from measurement issues conditional on having a common definition, or as Table A.2 suggests, may stem from different definitions being used. Both of these would imply the need for novel measurement and flexible country-specific definitions. Finally, estimates such as those do overlook significant heterogeneity in what they term "informal employment".

3.3 Out-of-sample fit and number of latent segments

The previous section presents evidence that unsupervised learning can be employed to detect meaningful dimensions of urban labor markets in which employers can shirk on at least some dimensions of the labor code. Figure 2 suggests that significant shares of contracts appear to be very likely to be generated by only one of two latent segments. This leaves a share of about 30% of contracts which are balanced mixture of the two segments. Section 3.1 puts forward that Latent Dirichlet are estimated for a fixed number of latent segments. By employing cross-validation, one can gauge the out-of-sample performance of several different models vis-à-vis the number of latent segments estimated.

The out-of-sample performance of a given model is evaluated using the quantity *perplexity* which is a commonly employed measure in unsupervised learning (Hornik and Grün, 2011). The perplexity is given by the geometric mean per-contract likelihood, that is, the likelihood that a given contract is generated by segment distribution that were previously estimated. The aforementioned authors provide a mathematical formulation of the quantity.

Using ten-fold cross validation, perplexity is calculated for models where the number of latent segments is varied from the set $k \in \{2, 3, ..., 10\}$[16], holding all other parameters of the estimation procedure constant. The results are shows in Figure 3 and further broken down whether private and public sector employees are pooled or only the former is considered. It appears that within these numbers of latent segments, a local optimum is given by seven latent segments in the pooled sample and ten latent segments in the private sector sample.

[16]In principle one can estimate as many latent segments as there are dimensions (columns) in the contract feature space. However, the computational burden grows as the number of latent segments increases. Additionally, it is not clear whether a model with $k > 10$ latent segments is useful in understanding labor markets. Blei (2012) notes that interpretability of latent models is valid concern when deciding on the number of latent segments to be estimated.

Both lines in Figure 3 do suggest that three latent segments outperform two latent segments in terms of their out-of-sample performance. Afterwards, more segments do not seem to do much better than three segments. As noted before, when choosing the number of latent segments one should take into account the task at hand (Blei, 2012). Proceeding by estimating the number of latent segments as suggested by the local minimum would not improve the fit significantly as compared to three latent segments. Moreover, as Figure 2 suggests the two latent segment model does well in describing 70% of the observed contracts. Therefore, I opt for the estimation of a model with three latent segments. This is broadly in line with Guenter and Launov (2012) who also find evidence that three latent segments best describe their data.

3.4 A model with three latent segments

The discussion and the results from Section 3.3 suggest that a model with three latent segments does well in describing the structure of the labor market of Dar-es-Salaam. Recall that the two segment model suggested that around 30% of contracts fell somewhere in the middle of a two-segment continuum (see Figure 2) and could therefore be said to contain elements of both, (subjectively) formal and informal, contractual features.

The three segment model is estimated in the same fashion as the dual segment model. Applying the same reasoning and arguments, Figure A.4 would suggest that segment one is one most likely to generate contracts in compliance with labor regulations[17]. The segment-over-feature distribution has mass points for written contracts, paid income tax as well as social security deductions. The latter two features occupy ranks 37 and 47 in segment two. Their respective ranks in segment three are 62 and 71. Similarly, the rank of the probability mass of having written contracts drops from one to 19 to 38 as one moves in ascending order of the segment *labels*.

Figure 4 provides further evidence that contracts generated from segment one are more likely to contain features that are commonly stipulated in labor codes. Firstly, the right hand side of the top panel indicates that segment one absorbs most of the public sector contracts but also is significantly more likely to offer maternity leave and union representation, compared

[17]Note that the numbering of segments is not ordinal. The segment numbering should be thought of as segment labels. Any ordinal ranking of these segments would have to come from the researcher and would therefore by subjective.

to segment two. The left tail of the top panel shows contracts in segment two are comparatively more often verbal or casual in nature while not including contributions to the welfare state. The center panel of Figure 4 suggests similar trends for the comparison between segments one and three. More interesting is the comparison between segments two and three as Figure A.4 is not indicative of a (subjective) hierarchy. The left tail would suggest that segment three is profoundly more likely to contain contracts of individuals with establishments of less than five employees[18]. Segment three also differentiates itself from segment two in that business records are more unlikely to be kept. On the other hand, segment two is more likely to generate contracts that contain some element of welfare state participation. Contracts in segment two tend to be with business of five or more employees[19].

The three-segment model therefore describes one segment of the labor market offering high-quality jobs that are aligned with the notion of "formal" employment. Additionally, there are two segments of the market which do differ in their contractual configuration but appear to offer similar jobs. These segments suggest that jobs are of heterogeneous quality and question the notion of homogeneous *informal* employment.

While segment two and three appear similar in terms of their feature configuration, Figure A.5 indicates that the three-segment model profoundly reduces the fraction of contracts with a balanced segment distribution. Recall that contract-segment distribution sums to one for each contract. The product of the segment shares is then maximized at $(\frac{1}{2})^2 = \frac{1}{4}$ in the two-segment model and $(\frac{1}{3})^3 = \frac{1}{27}$ in the three-segment model. Figure A.5 shows the cumulative densities of the models. The two-segment model in panel a) has around 25% of contracts with (strongly) ambiguous contract-feature distributions. This number is reduced (to about 5%). This shows that the additional segment is necessary in describing other configurations of observed contracts.

[18]Note that some authors have used the number of employees as a criterion to define a binary indicator for informal employment (see Table A.2). This suggests that such a definition would mask significant heterogeneity across employment relationship.

[19]The discussion of the differences between segments two and three omits some features that appear on the far ends of the right and left tails. For instance, segment three is more likely to contain contracts that are not exhausting an individuals labor supply ("avail.more.work", less than 40h of work because of a lack thereof). These and the ones on the right tail which are not included in the discussion do not carry significant mass within their respective segments as shown in Figure A.4. Though they do matter in comparative terms, those features lack meaningfulness in absolute terms.

3.5 Implications of the three segment model

The previous section provides suggestive evidence that assigning the binary labels *formal* and *informal* to employment relationships in developing country is unlikely to capture the universe of contracts. Jobs come in various configurations, and if enforcement of regulations is limited, each contractual dimension can be used to create incentives[20].

When dimensions other than the wage can be used to create unique "packages" of incentives, do principals trade pecuniary and non-pecuniary incentives off against each other? In other words, conditional on observeable characteristics of an agent, are there low-pay but high-quality as well as high-pay but low-quality jobs? Note that there is no reason to suspect that there are only these two configurations but the trade-off may well be continuous. If this trade-off actually exists, one would expect to see that there are jobs which contain features such as welfare state contributions, regulated hours and other protections with a wage below the wage level in jobs that contain none (not all) of these features.

Using the model with three latent segments in the labor market, one can estimate the wage distribution across all three segments. A contract's distribution over segments can be thought of as incorporating the uncertainty when trying to assess whether a job is "(in)formal". One can sample repeatedly from this distribution and each instance results in a vector of contract-to-segment assignments. Figure 5 plots the resulting wage distribution from drawing 2000 random samples. The wage distribution is represented by taking the median (as well as the 2.5^{th} and 97.5^{th} percentile) across the 2000 draws at each percentile of the wage distribution. Note that the 95% confidence bounds are extremely tight which is a consequence of the three-segment model resulting in very few spread-out contract-over-feature instances (see Figure A.5).

In Figure 5a, the distribution of the 99% winsorized and residualized wage is shown[21]. The distribution for segments two and three are almost indistinguishable. This likely captures the fact that segments two and three describe similar jobs which come in different contractual configuration. Figure 5a clearly establishes that the lowest 35% of the segment one distribution is below the other other two segments. Hence, the lowest wages paid for higher-quality jobs are below wages for (heterogeneous) low-quality jobs. But the distributions also confirm the commonly found stylized fact that that the highest paid wages in the high-quality job sector

[20] Section 4 [which is work in progress] will formally establish this point.

[21] For a description of the construction of that variable please refer back to Footnote 15

are (significantly) higher than the highest wages for low-quality ("informal") jobs (cf (Basu et al., 2015)). Hence it appears as if the aforementioned remuneration trade-off is at work for at least a subset of the labor market. Some principal-agent interactions appear to be resulting in a bargain where job quality and pecuniary benefits are traded off against each other.

Note the profound differences in the wage distributions[22] in panels a) and b) of Figure 5. Selection of higher-ability and better-educated individuals into better-quality jobs likely accounts for the difference. These individuals are more productive and therefore command higher wages. As shirking on job quality is more difficult in high-value activities of the economy, wages become the sole incentive instrument. In terms of raw wages the highest wages in the uncovered segments of the labor market are about as high as the 55^{th} percentile of the covered segments. The respective figure is significantly higher once selection effects are accounted for.

Finally, Table A.3 lists the 15 most frequently occurring occupations across the three segments. These are obtained by the above resampling procedure. Each occupation titles frequency across all iterations is then divided by 2000 to calculate each occupation title's average frequency. Segment one contains jobs that are intuitively "high-quality", such as accountants, teachers and doctors. Note that also drivers and guards feature in that list. Segment two and three contain similar jobs such as craftsmen and low-skill service sector jobs as well as drivers in various capacities. Segment three additionally features domestic helpers and food vendors. The appearance of drivers and guards across all segments may suggest that these are occupations where aforementioned incentive trade-off is at work.

4 An economic contracting framework

This section will offer and formalize an economic framework[23] within which the above methodology and results can be understood. As already alluded to in the introduction, one commonly cited argument against the idea of an inherently disadvantaged informal sector is the fact that it may offer more flexibility than regulated employment. Additionally, if enforcement of standards or provision of social safety goods is weak, agents (employees) may

[22] Note that - while not residualized - the wage variable in panel b) is still 99%-winsorized.

[23] This is still very much work in progress!

decrease their valuation of benefits commonly associated with formal employment (Maloney, 2004).

At the beginning of an employment relationship a principal and an agent have to enter a contract, that is, an agreement over what the principal offers the agent. Denote this contract by Γ which is a $1xF$ vector stipulating which benefits the principal offers. The dimension of F is the set of possible benefits that the principal may offer or withhold from the agent, i.e., $\Gamma_{1,f} \in \{0,1\}\ \forall f \in \{1,...,F\}$. Additionally, the agreement between principal and agent sets a scalar wage w. The agent provides a monetary output $q(\Gamma, \overline{q_i}); q : R^{F+1} \mapsto R$[24] to the principal. That is, the value of the output depends on the contract agreed upon. In particular, I assume that positive but diminishing returns along each dimension of the contract. The principal profits are hence given by

$$\pi = q(\Gamma, \overline{q_i}) - w - \underline{c}\Gamma^T$$

where $\underline{c}$ is a $1xF$ vector with the exogenous (and constant) cost of providing the contract configuration Γ. All contract dimensions are assumed to be costly for the principal ($c_{1,f} > 0\ \forall f \in F$). The agent in turn has a utility function $u(w, \Gamma); u : R^{F+1} \mapsto R$ with the domain $[w, \Gamma]$. I assume positive and diminishing marginal utility in all dimensions. Moreover, agents have an outside option with utility level u_0. Finally, an agent would never accept a job with zero wage payments $u(w = 0, \Gamma) < u_0$.

The institutional framework in which this principal-agent interaction takes place stipulates a labor code which is imperfectly enforced. In order to comply with the labor code, the contract between the principal and the agent has to meet a set of exogenous conditions. That is, a contract in compliance with the labor code has restrictions on some $c < F$ or all $c = F$ dimensions of the contracting space. Assume that the contracting dimension j captures severance pay (in case of firing, the agent is to receive a lump sum payment) and that the labor code prescribes such severance pay. A contract in compliance with the labor code must then have $\Gamma_{1,j} = 1$, which is costly since $c_{1,j} > 0$. Define a contract that is in compliance with the labor code as

[24] This is similar to the notion of efficiency wages. Intuitively, this means that in addition to a idiosyncratic productivity level $\overline{q_i}$, agents are more productive in better work environments. I do assume that higher wages *per se* do not lead to higher output.

$$\Gamma : [\Gamma_{1,c}\ \Gamma_{1,n}]$$
$$\Gamma_{1,c} = 1\ \forall\ \ 0 \leq c \leq F$$
$$\Gamma_{1,n} \in \{0,1\}\ \forall\ \ c < n \leq F$$

Enforcement of the labor code is limited in the sense that each principal-agent interaction has a probability $\lambda(q)$ of being monitored. The probability of being monitored is a function of the output. High-value added interactions are more likely to be monitored $\frac{\partial \lambda(q)}{\partial q} > 0$[25]. If a principal-agent interaction is monitored and found to not be in compliance with the labor code denoted by the set of restrictions $\underline{C}_{1,c} = 1\ \forall c$, a fine F is assessed which is fixed but multiplies in the number of dimensions violated. That multiplicative nature captures the fact that the analysis in Sections 3.2 and 3.4 suggests that shirking on compliance is not binary but that contracts may comply to some but not the full extent. This gives rise to the principal's optimization problem [26]

$$max_{w,\Gamma}\ \ \pi = q(\Gamma, \overline{q_i}) - w - \underline{c}\Gamma^T - \lambda(q)([\underline{C} - \Gamma_c]^T) * F$$
$$\text{subject to:}\quad u(w,\Gamma) \geq u_0$$

This setup gives rise to a number of trade-offs that the principal has to take into account[27]. There is an interplay in the agent's utility function between wage and non-pecuniary work benefits (Γ). To the extent that the agent values the latter, it may be cost-effective to offer a mix rather than a pure-wage contract. More productive agents produce higher output for the principal but at the same time, they render they principal-agent interaction more prone to monitoring. The principal may have to balance the latter by fully complying with $\underline{C}$, thereby avoiding penalties if being monitored. One could imagine different combinations of (w, Γ) which meet the agent's participation constraint. The one combination that is desirable from the principal's point of view then depends on the interplay of λ, F and $\underline{C}$.

[25] There is a second order effect coming from the dependency between q and Γ. Through increasing a workers productivity, a more compliant work contract would also increase the probability of being monitored. This may seem counter intuitive as one would expect worse conditions to lead to more monitoring. This is an issue that the model will have to incorporate as I move forward.

[26] Recall that Γ_c are those contracting dimensions subject to the labor code. If $c = F$, then the labor code puts forward conditions along all contractible dimensions, i.e, then $\Gamma = \Gamma_c$

[27] This will be subject to further analysis in the course of this project.

A number of extension are possible to this framework. Rather than considering a single interaction, one may analyze an organization with the possibility to offer separate contracts to agents. This would reflect the fact that covered and uncovered employment exists within the same organization. Introducing endogeneity to the utility function may account for the fact that more productive (better educated) individuals demand high-quality contracts while still earning high wages. Agency problems may be another promising avenue. A intuitive plausible structure would introduce a negative dependency between the cost of effort and the quality of the contracts. Finally, agents may reciprocate if principals have to incur costs to comply with the labor code.

5 Concluding remarks

This analysis tries to address questions that have been receiving ample attention by using new methodology and doing away with assumptions that have been made in the past. Fundamentally, the paper tries to provide an answer to whether employees in urban labor markets with imperfect enforcement of labor codes trade-off job-quality and wage. If so, this would provide evidence for a view held by an increasing number of researchers and policymakers that employment outside the covered sector may represent a choice.

Existing research that has studied labor markets in developing countries often splits employment relationships into "formal" and "informal" based on criteria such as whether income tax and/or social security are paid. The analysis in Section 3.2 presents two issues with this approach. Firstly, it ignores substantive heterogeneity among those for whom the criterion is not true. Secondly, if the criterion is based on two more dimensions, it assumes that those for whom no criterion is true are similar to those for whom at least one but all are true. The framework presented in Section 4 attempts to rationalize the latter as a contracting outcome under imperfect monitoring.

The empirical analysis based on labor force survey data from Dar-es-Salaam in 2014 is largely based on an unsupervised learning algorithm which attempts to describe latent segments in the labor market which (probabilistically) generate the observed labor contracts in the market. Each contract therefore is a mixture of these segments which are distributions over contractual features. Some segments are more likely to "generate" bad-quality jobs, and vice versa. Section 3.2 shows that - while capturing meaningful variations - the dualistic view

is an incomplete tool to describe all observed labor relationships. A model with three latent segments has better out-of-sample performance, although the improvement is rather small.

Incorporating the uncertainty inherent into classifying labor contracts, wage distributions are estimated in Section 3.5. The results suggest that at least within a certain range, agents do seem to trade job-quality versus monetary remuneration. This is true even after trying to account for selection into low-quality jobs. This provides suggestive evidence for the view that some poor quality -"informal" - jobs constitute a choice rather than a last resort.

In keeping with the cautious remarks from Section 2.2, one should be careful to extrapolate to other contexts. Labor relationships present an equilibrium from a bargaining process which is the result of the parties internalizing the particular environment. Furthermore, Section 2.3 presents some concerns with the data. Certainly the sensitivity of the learning process with respect to estimation parameters is another source of concern for the results. Further analysis will have to be provided to establish that the results are not the artifact of selective parameter choice.

References

Almeida, R. and Carneiro, P. (2012). Enforcement of labor regulation and informality. *American Economic Journal: Applied Economics*, 4(3):64–89.

Altmann, S., Falk, A., Grunewald, A., and Huffman, D. (2013). Contractual incompleteness, unemployment, and labour market segmentation. *Review of Economic Studies*, 81(1):30–56.

Bandiera, O., Hansen, S., Prat, A., and Sadun, R. (2017). Ceo behavior and firm performance. Technical report, National Bureau of Economic Research.

Bargain, O. and Kwenda, P. (2011). Earnings structures, informal employment, and self-employment: New evidence from brazil, mexico, and south africa. *Review of income and wealth*, 57(s1).

Basu, A., Chau, N., and Kanbur, R. (2015). Contractual dualism, market power and informality. *The Economic Journal*, 125(589):1534–1573.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bosch, M. and Esteban-Pretel, J. (2012). Job creation and job destruction in the presence of informal markets. *Journal of Development Economics*, 98(2):270–286.

Charmes, J. (2012). The informal economy worldwide: Trends and characteristics. *Margin: The Journal of Applied Economic Research*, 6(2):103–132.

Cichello, P. L., Fields, G. S., and Leibbrandt, M. (2005). Earnings and employment dynamics for africans in post-apartheid south africa: A panel study of kwazulu-natal. *Journal of African Economies*, 14(2):143–190.

Fields, G. S. (1990). Labour market modelling and the urban informal sector: Theory and evidence.

Fields, G. S. (2004). A guide to multisector labor market models. *World Bank Working Papers*, page 86.

Garganta, S. and Gasparini, L. (2015). The impact of a social program on labor informality: The case of auh in argentina. *Journal of Development Economics*, 115:99–110.

Goldberg, P. K. and Pavcnik, N. (2003). The response of the informal sector to trade liberalization. *Journal of development Economics*, 72(2):463–496.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Günther, I. and Launov, A. (2012). Informal employment in developing countries: Opportunity or last resort? *Journal of development economics*, 97(1):88–98.

Hansen, S., McMahon, M., and Prat, A. (2014). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*.

Hornik, K. and Grün, B. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

International Labour Organization (2002). Decent work and the informal economy. Technical report, International Labour Conference 90th Session.

International Labour Organization (2012). Statistical update on employment in the informal economy. Technical report.

International Labour Organization (2013). Women and men in the informal economy: a statistical picture. Technical report, International Labour Organization.

International Labour Organization (2017). Informal economy. Retrieved April 4, 2017, 2:34pm.

Kugler, A. and Kugler, M. (2009). Labor market effects of payroll taxes in developing countries: evidence from colombia. *Economic development and cultural change*, 57(2):335–358.

Kugler, A., Kugler, M., and Prada, L. O. H. (2017). Do payroll tax breaks stimulate formality? evidence from colombia's reform. Technical report, National Bureau of Economic Research.

Maloney, W. F. (1999). Does informality imply segmentation in urban labor markets? evidence from sectoral transitions in mexico. *The World Bank Economic Review*, 13(2):275–302.

Maloney, W. F. (2004). Informality revisited. *World Development*, 32(7):1159–1178.

Marcouiller, D., de Castilla, V. R., and Woodruff, C. (1997). Formal measures of the informal-sector wage gap in mexico, el salvador, and peru. *Economic development and cultural change*, 45(2):367–392.

Moreno-Monroy, A. I. and Posada, H. M. (2018). The effect of commuting costs and transport subsidies on informality rates. *Journal of Development Economics*, 130(1):99–112.

Paz, L. S. (2014). The impacts of trade liberalization on informal labor markets: A theoretical and empirical evaluation of the brazilian case. *Journal of International Economics*, 92(2):330–348.

Pratap, S. and Quintin, E. (2006). Are labor markets segmented in developing countries? a semiparametric approach. *European Economic Review*, 50(7):1817–1841.

Schneider, F. and Enste, D. (2000). Shadow economies: size, causes, and consequences. *Journal of Economic Literature*, 38(1):77–114.

Ulyssea, G. (2010). Regulation of entry, labor market institutions and the informal sector. *Journal of Development Economics*, 91(1):87–99.

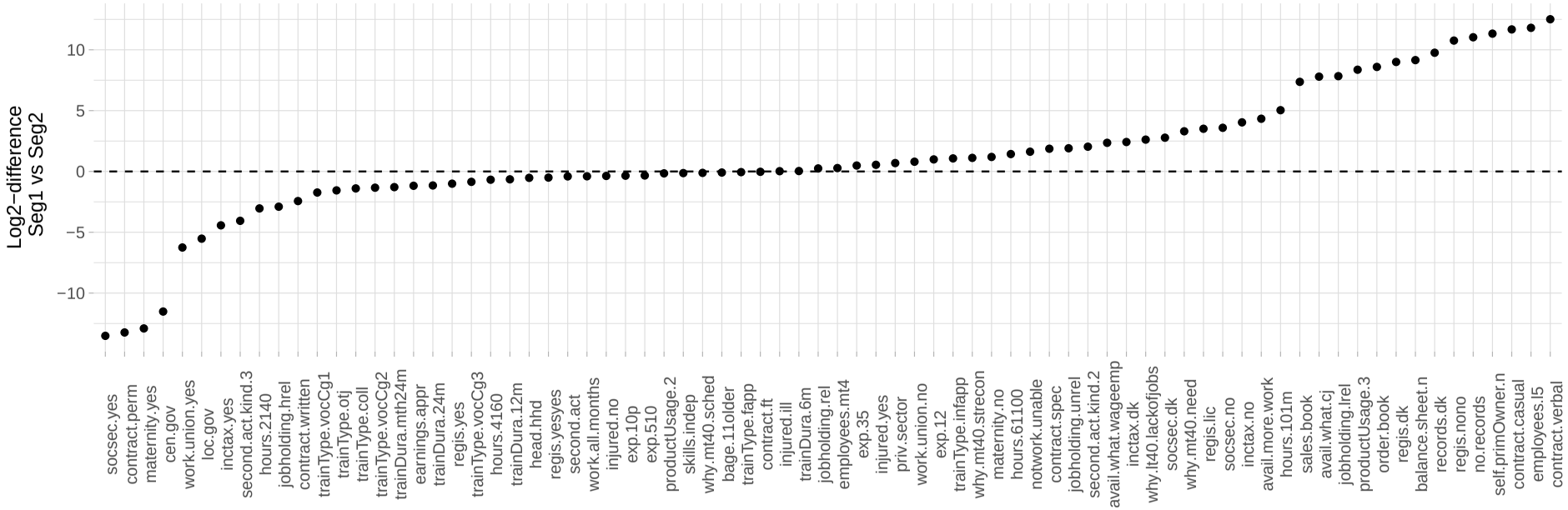
van Soest, A. and Gong, X. (2002). Wage differentials and mobility in the urban labour market: a panel data analysis for mexico. *Labor Economics*, 9(4):413–549.

Table 1: Formality premiums for several definitions of formal employment

	Hourly total wage							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Formal (def 1)	711	731						
	(490, 933)	(502, 960)						
Formal (def 2)			849	796				
			(598, 1,100)	(517, 1,075)				
Formal (def 3)					1,182	1,427		
					(797, 1,567)	(669, 2,184)		
Formal (def 4)							396	490
							(50, 742)	(139, 840)
Priv sector	−273		−251		−138		−471	
	(−554, 7)		(−530, 29)		(−433, 157)		(−731, −210)	
Weekly hours	−22	−22	−21	−22	−22	−22	−24	−23
	(−26, −18)	(−26, −18)	(−26, −17)	(−26, −18)	(−26, −18)	(−27, −17)	(−28, −20)	(−28, −19)
Informal mean	1140	1073	1181	1103	1389	1246	1822	1306
# formal	739	419	657	358	394	161	235	224
Sample	All	Private	All	Private	All	Private	All	Private
Quarter of int FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Add controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj R-squared	0.37	0.32	0.37	0.32	0.38	0.33	0.36	0.3
Observations	2,096	1,598	2,096	1,598	2,096	1,598	2,096	1,598

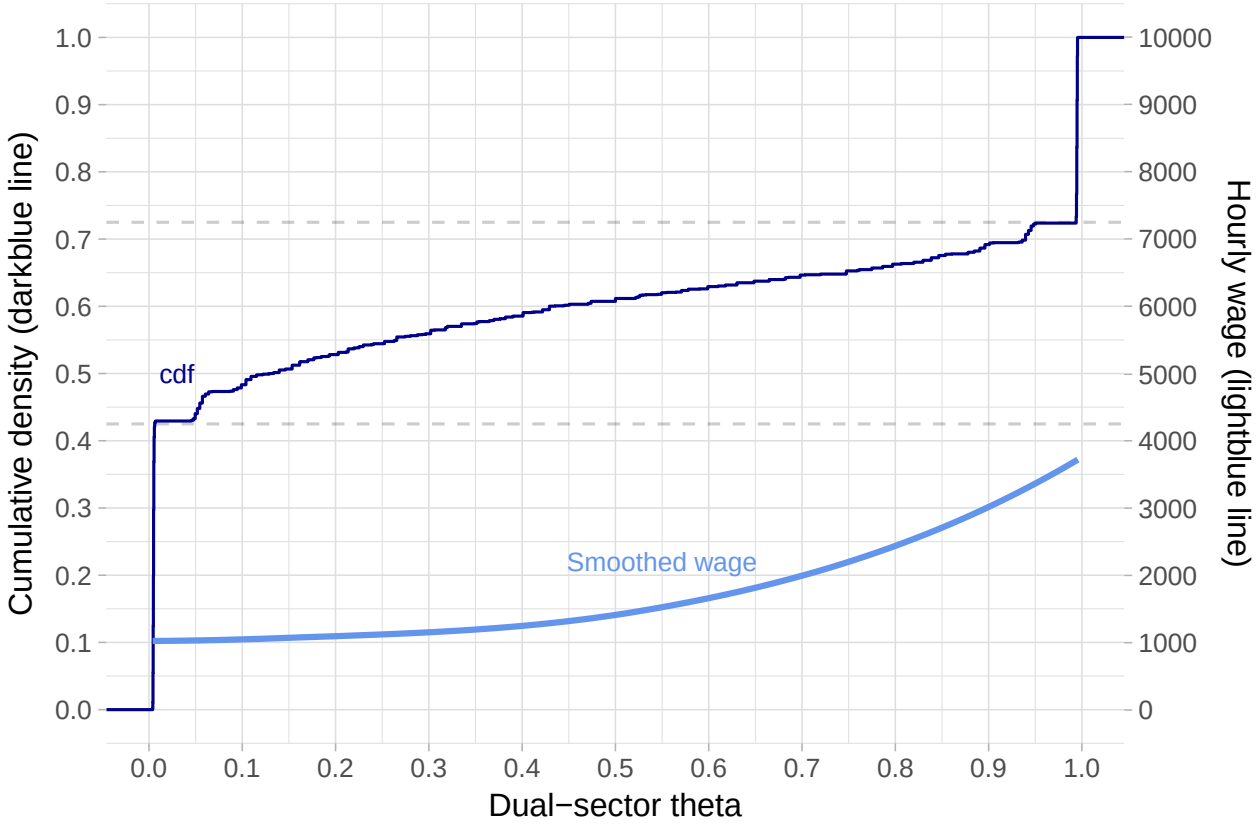
This table estimates formal sector wage premiums based on several commonly found definitions of *formal employment*. The dependent variable is the hourly wage after applying a 99% winsorization. Note that the sample is restricted to males from Dar-es-Salaam working for pay. In columns 1 and 2, formal is defined as an employee who makes social security contributions. Columns 3-4 define formal employment as making social security contributions and income tax being deducted. The definition of columns 3 and 4 is kept in the remaining columns, and columns 5 and 6 add the possibility of maternity leave as an additional criterion while columns 7 and 8 add the registration and licensing of the business as an additional defining variable. The row "# formal" indicates the number of observation for which the criterion is true. The additional control variables include separate indicators for literacy in Kiswahili, English or both, an indicator for being a citizen of a foreign country, indicators for secondary or university education separately and quarter-of-year interview date fixed effects. 95% confidence interval based on Eicker-Huber-White robust standard errors are reported in brackets. This and all other tables in this text are created using: Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.R package version 5.2.1. https://CRAN.R-project.org/package=stargazer

Figure 1: Dualistic view - differences between latent segments



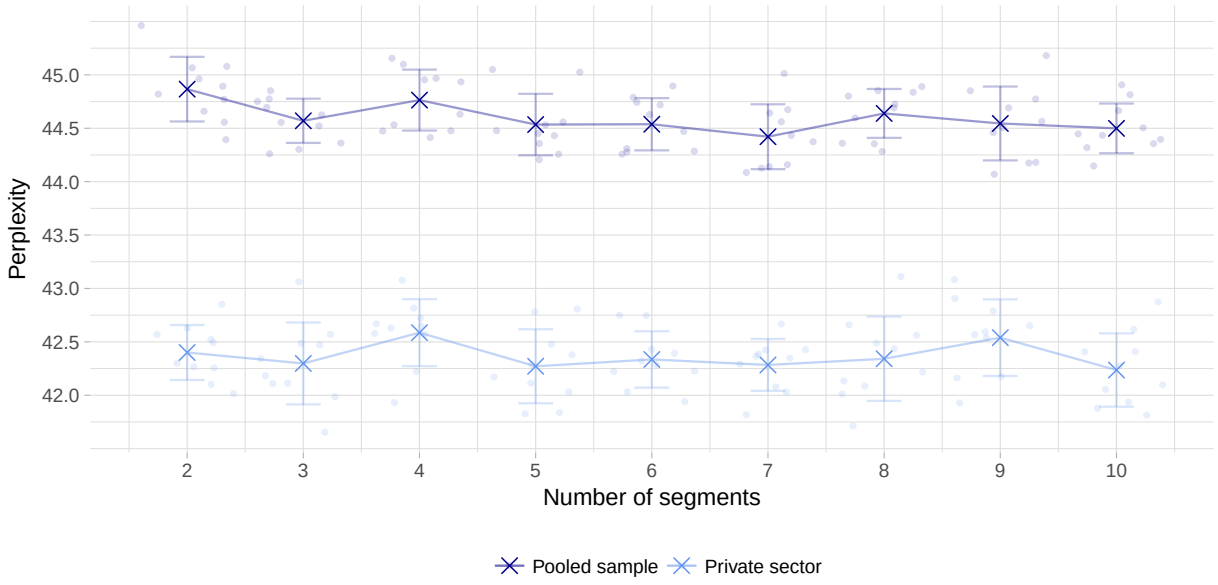
This figure shows illustrates the differences across the two segments given the dualistic view of the labor market. For each contractual feature, the figure shows the $\log_2$-difference between the probability that a contract generated by segment one contains the respective feature and the probability that a contract generated by segment two contains the same segment. If the the two probabilities are equal, the $\log_2$-difference will be zero, which is indicated by the dashed black line. Values larger than zero mean that the probability for a given feature is higher in segment one, and vice versa. The segment-feature probabilities are obtained from estimating a Latent Dirichlet Allocation using Gibbs-Sampling with a 1000 "burn-in"-iterations after which every 400^{th} draw from the next 2000 iterations is taken. The former tries to account for the fact that the Gibbs-sampling will tend to converge only over time while the latter tries to avoid evaluating auto-correlated draws. This procedure is repeated for 40 randomly chosen starting points. Finally, the hyperparameters governing the prior distribution are chosen to be $\alpha = \delta = .1$, reflecting the the fact that contracts tend to have mass on one segment rather than both and segments are unlikely to have feature mass spread equally.

Figure 2: Dualistic view - $\gamma^i_{k=2}$ and smoothed wage distribution



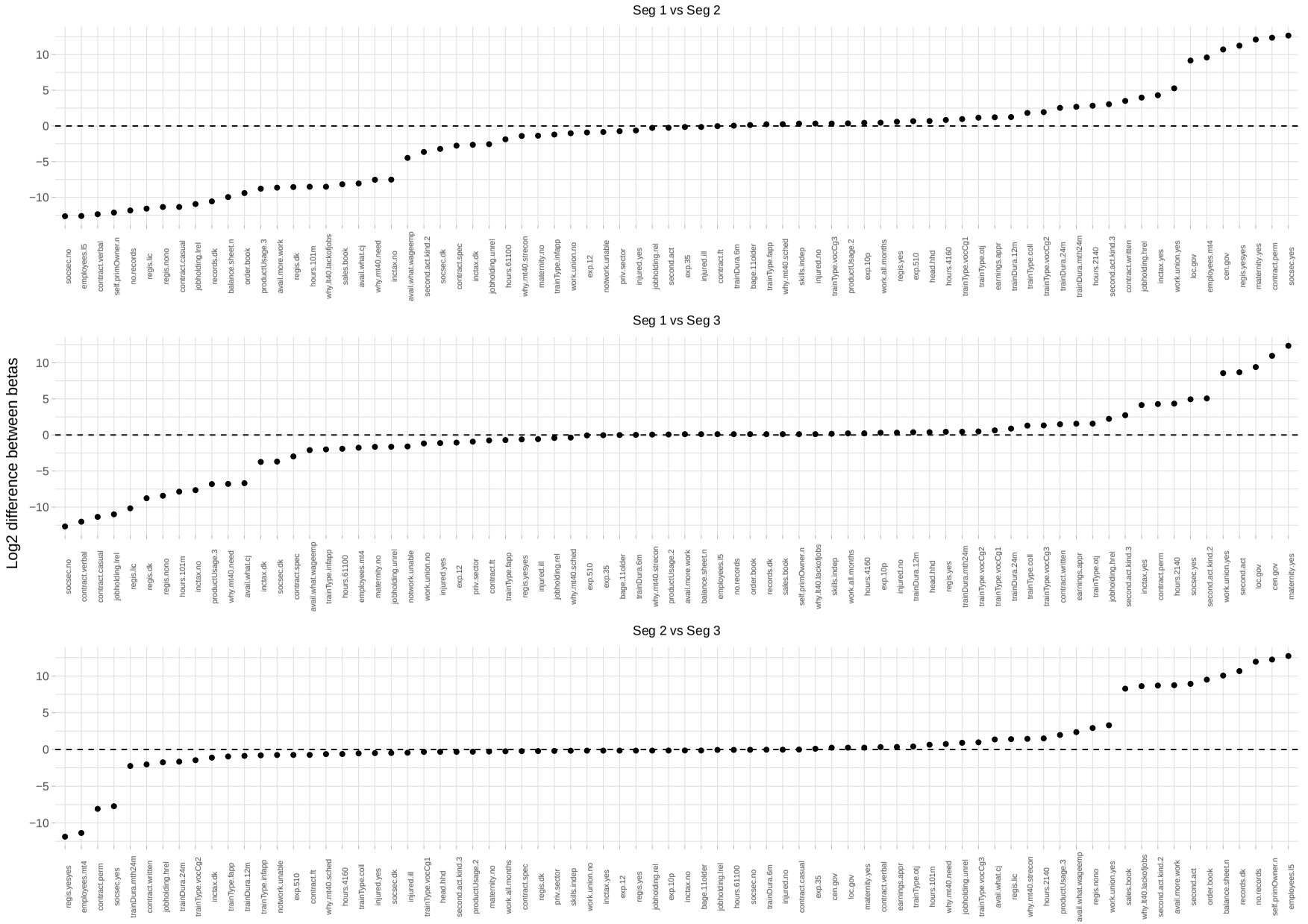
This figure reports two distinct quantities. The darkblue line (labelled "cdf") shows the cumulative density of the sample probabilities that a given contract was generated by segment two. Given the dual-segment setup, this can be interpreted as a "formality" index (see Section 3.2). The relevant y-axis for the cumulative density is the left-hand side one. The contract-segment probabilities were estimated from a Latent Dirichlet Allocation. The notes to Tables 1 and A.3 provide detailed information on the parameters of the estimation process. The lightblue line ("Smoothed wage") shows a smooth local linear regression of residualized and 99%-winsorized total hourly wage on the probability that a contract was generated by segment two. The winsorization process along with the model that was used to compute residuals is described in the main text in Footnote 15. The robust local linear regression used a second degree polynomial. The relevant scale for the wage variable is one the right hand side. Latent Dirichlet Allocation using Gibbs-Sampling with a 1000 "burn-in"-iterations after which every 400^{th} draw from the next 2000 iterations is taken. The former tries to account for the fact that the Gibbs-sampling will tend to converge only over time while the latter tries to avoid evaluating auto-correlated draws. This procedure is repeated for 40 randomly chosen starting points. Finally, the hyperparameters governing the prior distribution are chosen to be $\alpha = \delta = .1$

Figure 3: Out-of-sample performance of models with different numbers of latent segments



This figure plots the perplexity of Latent Dirichlet Allocation models that differ in the number of latent segments that are estimated. For a description of the perplexity measure, see Section 3.1. The darkblue line describes models ran on the pooled sample of private and public employees while the lightblue line stems from the sample restricted to the former. The perplexity is calculated from a ten-fold cross validation procedure in which the same is split into a 2/3 training and 1/3 test sample each time. Perplexity then measures the out-of-sample performance of the model estimated in the training sample in the test sample. The dots indicate the mean perplexity across the ten folds and the error bars indicate the standard deviation. The dots report the perplexity measures of all ten folds. Note that lower values for perplexity suggest better out-of-sample performance. Latent Dirichlet Allocation were estimated using Gibbs-Sampling with a 1000 "burn-in"-iterations after which every 400^{th} draw from the next 2000 iterations is taken. The former tries to account for the fact that the Gibbs-sampling will tend to converge only over time while the latter tries to avoid evaluating auto-correlated draws. This procedure is repeated for two randomly chosen starting points (to reduce the computational burden). Finally, the hyperparameters governing the prior distribution are chosen to be $\alpha = \delta = .1$

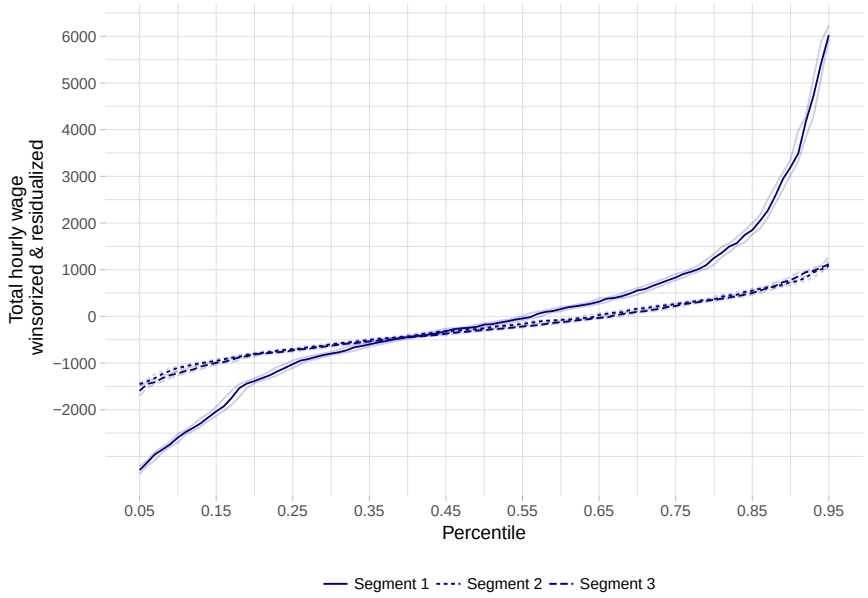
Figure 4: Differences between latent segments in the three segment model



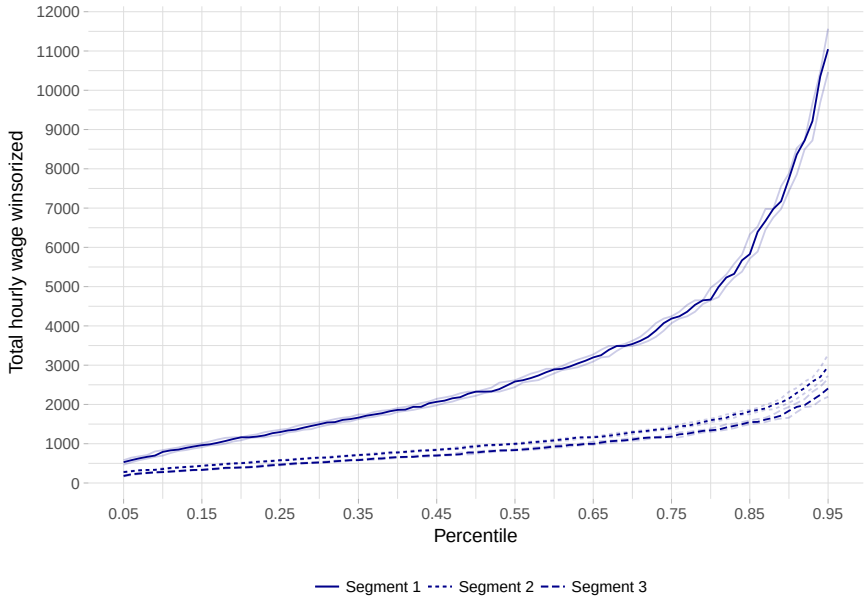
This figure shows illustrates the differences across the two segments given the dualistic view of the labor market. For each contractual feature, the figu re shows the $\log_2$-difference between the probability that a contract generated by segment one contains the respective feature and the probability that a contract generated by segment two contains the same segment. If the the two probabilities are equal, the $\log_2$-difference will be zero, which is indicated by the dashed black line. Values larger than zero mean that the probability for a given feature is higher in segment one, and vice versa. The segment-feature probabilities are obtained from estimating a Latent Dirichlet Allocation using Gibbs-Sampling with a 1000 "burn-in"-iterations after which every 400^{th} draw from the next 2000 iterations is taken. The former tries to account for the fact that the Gibbs-sampling will tend to converge only over time while the latter tries to avoid evaluating auto-correlated draws. This procedure is repeated for 40 randomly chosen starting points. Finally, the hyperparameters governing the prior distribution are chosen to be $\alpha = \delta = .1$, reflecting the the fact that contracts tend to have mass on one segment rather than both and segments are unlikely to have feature mass spread equally.

Figure 5: (Probabilistic) wage percentiles of the three-segment model

(a) Residualized wage



(b) Raw wage



These figures show the wage distributions across the three segments. Each distribution is represented by the median, the 2.5^{th} the and 97.5^{th} percentile of 2000 repetitions of the following procedure. Sample a segment assignment for each contract based on the estimated contract-over-segment distribution. For each sample compute the percentiles of the wage distribution across all segments. This results in 2000 values for each percentile in each segment which are the basis for the estimate of the wage distribution shown here. In panel a), the residualized and 99%-winsorized wage is used (see Footnote 15 for a detailed description. In panel b), the raw but 99% winsorized wage is used.

APPENDIX

Table A.1: Variables and summary statistics for Dar-Es-Salaam sample

Variable	Mean	Standard Deviation	Variable	Mean	Standard Deviation
Avail more work	0.02	0.13	Unable to work all time	0.03	0.17
Avail more work current job	0.02	0.13	Order books	0.03	0.17
Avail more wage work	0.01	0.11	Private sector	0.76	0.43
Business age >11	0.99	0.11	Products only for sale	0.97	0.17
No balance sheet	0.04	0.20	Products for sale and consump	0.02	0.16
Central gov	0.09	0.28	Don't know about records	0.07	0.25
Casual cntract	0.25	0.43	Don't know about registr	0.04	0.19
Fixed-time contract	0.21	0.41	Licensed	0.19	0.39
Permanent contract	0.29	0.45	Neither registered nor licensed	0.13	0.34
Specific contract	0.24	0.43	Registered	0.10	0.30
Verbal contract	0.44	0.50	Registered and licensed	0.34	0.47
Written contract	0.55	0.50	Sales book	0.01	0.11
Earnings appropriate	0.20	0.40	Engages in 2nd activity	0.04	0.19
<5 employees	0.27	0.44	2nd activity: employee	0.02	0.13
>=5 employees	0.19	0.39	2nd activity: self-employed	0.01	0.11
Experience >10a	0.62	0.49	Not primary owner	0.19	0.40
1-2a experience	0.12	0.33	Use skills independently	0.98	0.13
3-5a experience	0.13	0.33	Social security don't know	0.03	0.18
5-10a experience	0.17	0.38	No social security	0.61	0.49
Head hhd	0.75	0.43	Social security	0.35	0.48
Hours >101h	0.03	0.16	Training duration (6m,12m]	0.09	0.29
Hours 21-40h	0.16	0.37	Training duration (12m,24m]	0.05	0.21
Hours 41-60a	0.37	0.48	Training duration <6m	0.82	0.38
Hours 61-100a	0.43	0.50	Training duration >24m	0.03	0.16
Inc tax don't know	0.04	0.20	Training: college	0.18	0.38
No income tax	0.57	0.50	Training: formal apprenc	0.05	0.22
Income tax	0.39	0.49	Training: informal apprenc	0.07	0.26
Injured/fallen ill	0.02	0.15	Taining: on the job	0.04	0.20
Not injured	0.76	0.43	Training: vocational G1	0.01	0.11
Injured	0.21	0.41	Training: vocational G2	0.01	0.11
Highly reliable job	0.27	0.44	Training: vocational G3	0.01	0.12
Reliable job	0.19	0.39	<40h: lack of jobs	0.02	0.12
Unreliable job	0.52	0.50	>40h: necessary	0.15	0.36
Highly unreliable job	0.02	0.15	>40h: scheduled	0.64	0.48
Local gov	0.03	0.17	>40h: strong econ	0.05	0.22
No maternity leave	0.76	0.43	Worked all months	0.88	0.33
Maternity leave	0.23	0.42	No work union	0.80	0.40
No business records	0.16	0.36	Work union	0.19	0.39

This table reports basic summary statistics for all contractual features that are used in the analysis in alphabetic order. Since all variables are binaries, the means in the second and fourth column are the sample shares for the respective variable. The standard deviation is computed as $sd = \sqrt{p(1-p)}$ with p being the mean of the respective variable.

Table A.2: Overview of formality vs informality definitions in the literature

Garganta and Gasparini (2015)	JDE	Argentina	informal = workers who have no deductions for pensions in their job
Ulyssea (2010)	JDE	Brazil	informal = workers engaged in legal activities who do not contribute to social security
Bosch and Esteban-Pretel (2012)	JDE	Brazil	formal = possession of "carteira de trabalho", i.e, a working permit
Goldberg and Pavcnik (2003)	JDE	Brazil	formal = possession of "carteira de trabalho" [Brazil]; formal = employer pays social security contributions [Colombia]
Paz (2014)	JIntEcon	Brazil	formal = employer complies with payroll tax law
Kugler et al. (2017)	NBER	Colombia	multiple definitions of formality: written contract, health contributions, pension contributions, both, workers compensation insurance
Bargain and Kwenda (2011)	RIncWeal	BRA, MEX, SA	formal = possession of "carteira de trablho" [Brazil]; formal = employee contributes to social security [Mexico]; formal = firm provides medical and deducts unemployment insurance contributions (SA)

This table puts forward definitions used by various authors in academic literature analyzing urban labor markets in developing countries. Column one reports the citation handle, column 2 the commonly used accronym of the journal in which the research was/will be published and column three indicates the country(ies) which the research concerns. Column 4 paraphrases and summarizes the definition which the authors use for (in)formal employment.

Table A.3: Frequently occurring occupation titles in the three segment model

(a) Segment 1

Occupation title	Avg. frequency
Accountants	44.31
Car, Taxi and Light Van Drivers	42.61
Security Guards	40.69
Heavy Truck Drivers	38.31
Other Protective Service Workers	26.82
Primary Education Teachers	22.85
Bus Drivers and Driver-Conductors	20.95
Policemen and Policewomen	19.69
Medical Doctors	17.87
Secondary Education Teaching Professionals	16.67
Motor Vehicle Mechanics and Fitters	15.67
Shop Salespersons and Demonstrators	14.98
Accounting and Bookkeeping Clerks	12.47
Secondary Education Teachers, Associate Professionals	10.51
Other Business and Administrative Professionals	10.11

Table A.4: Segment 1

(a) Segment 2

Occupation title	Avg. frequency
Bus Drivers and Driver-Conductors	87.92
Heavy Truck Drivers	68.54
Security Guards	64.67
Shop Salespersons and Demonstrators	49.58
Motor Vehicle Mechanics and Fitters	47.47
Bricklayers, Masons and Tile Setters	33.01
Car, Taxi and Light Van Drivers	27.12
Other Sales and Services Elementary Occupations	23.53
Builders, Traditional Material	23.41
Motorcycle Riders	18.75
Carpenters	16.52
Cooks	14.13
Transport Conductors	12.52
Other Protective Service Workers	10.58
Messengers, Package and Luggage Porters and Deliverers	10.49

Table A.5: Segment 2

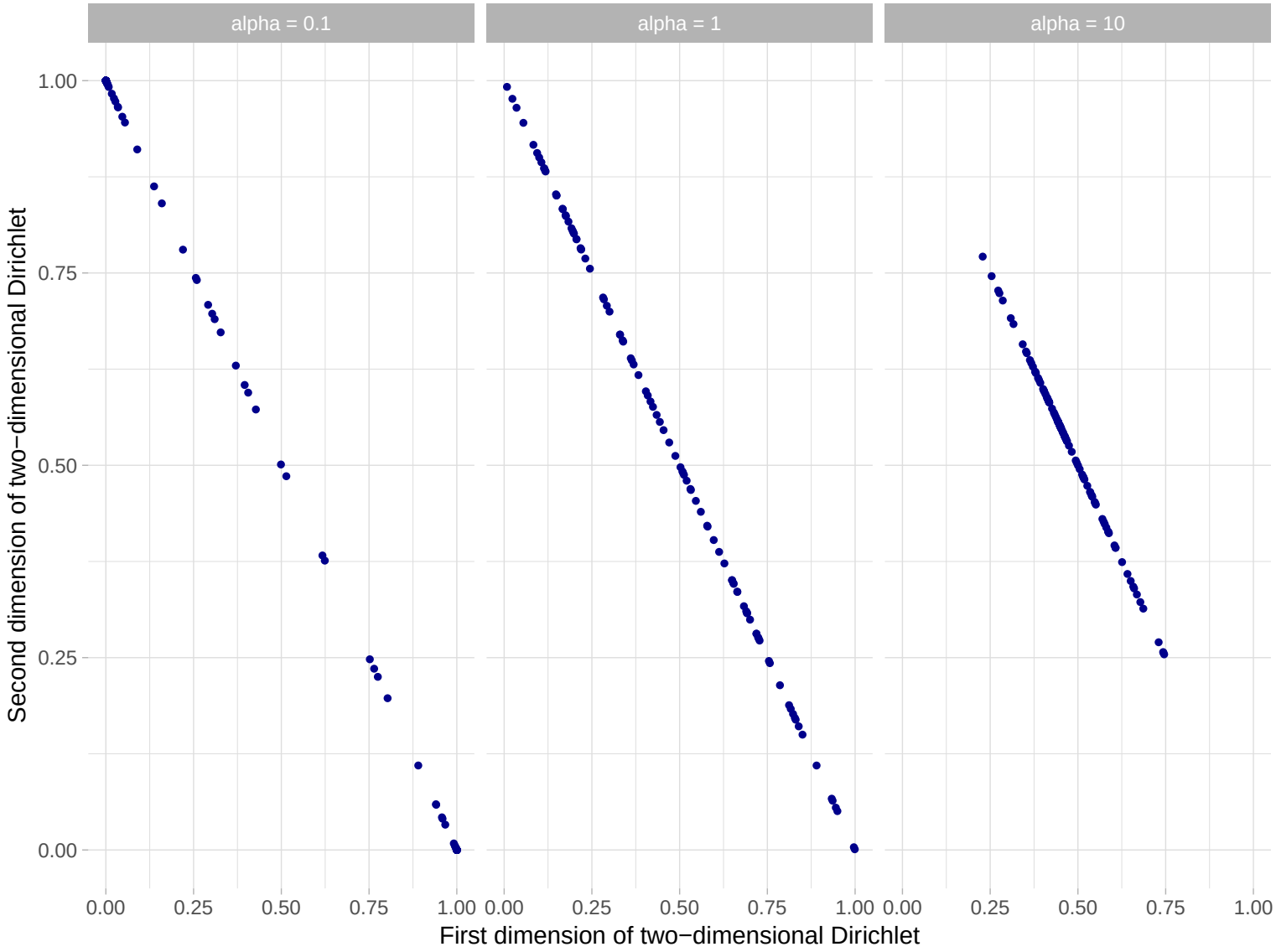
(a) Segment 3

Occupation title	Avg. frequency
Shop Salespersons and Demonstrators	79.44
Bus Drivers and Driver-Conductors	71.14
Motorcycle Riders	70.48
Bricklayers, Masons and Tile Setters	32.15
Other Sales and Services Elementary Occupations	31.87
Car, Taxi and Light Van Drivers	30.27
Heavy Truck Drivers	28.15
Security Guards	23.64
Domestic Helpers and Cleaners	22.28
Builders, Traditional Material	17.95
Transport Conductors	17.15
Hairdressers, Barbers, Beauticians and Related Workers	13.28
Motor Vehicle Mechanics and Fitters	10.85
Street Food Vendors	8.26
Building Construction Labourers	8.24

Table A.6: Segment 3

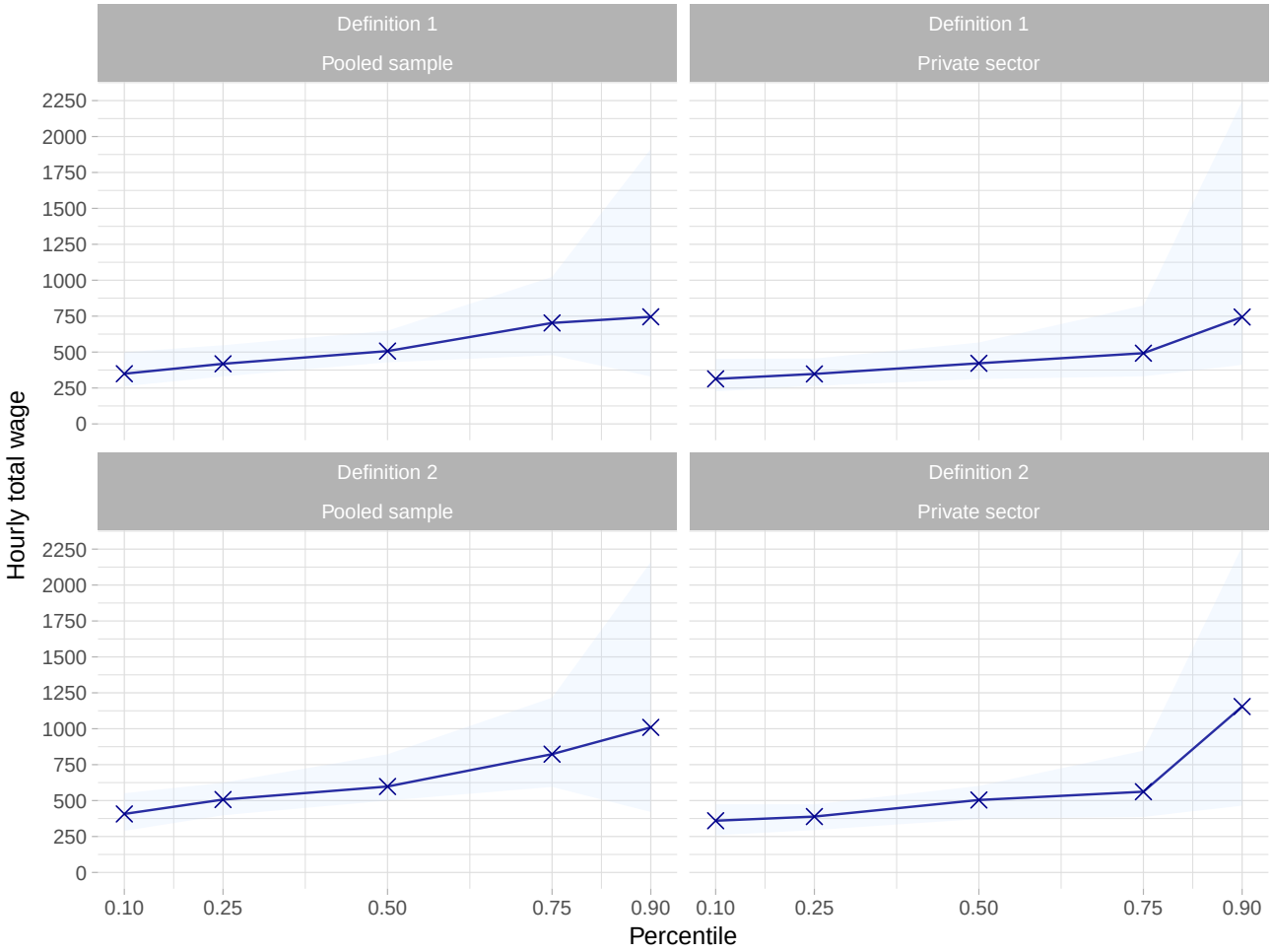
These tables show the 15 most frequently occuring occupation titles across the three segments. They are obtained via the same

Figure A.1: Dispersion of Dirichlet(α) for $\alpha \in \{.1, 1, 10\}$



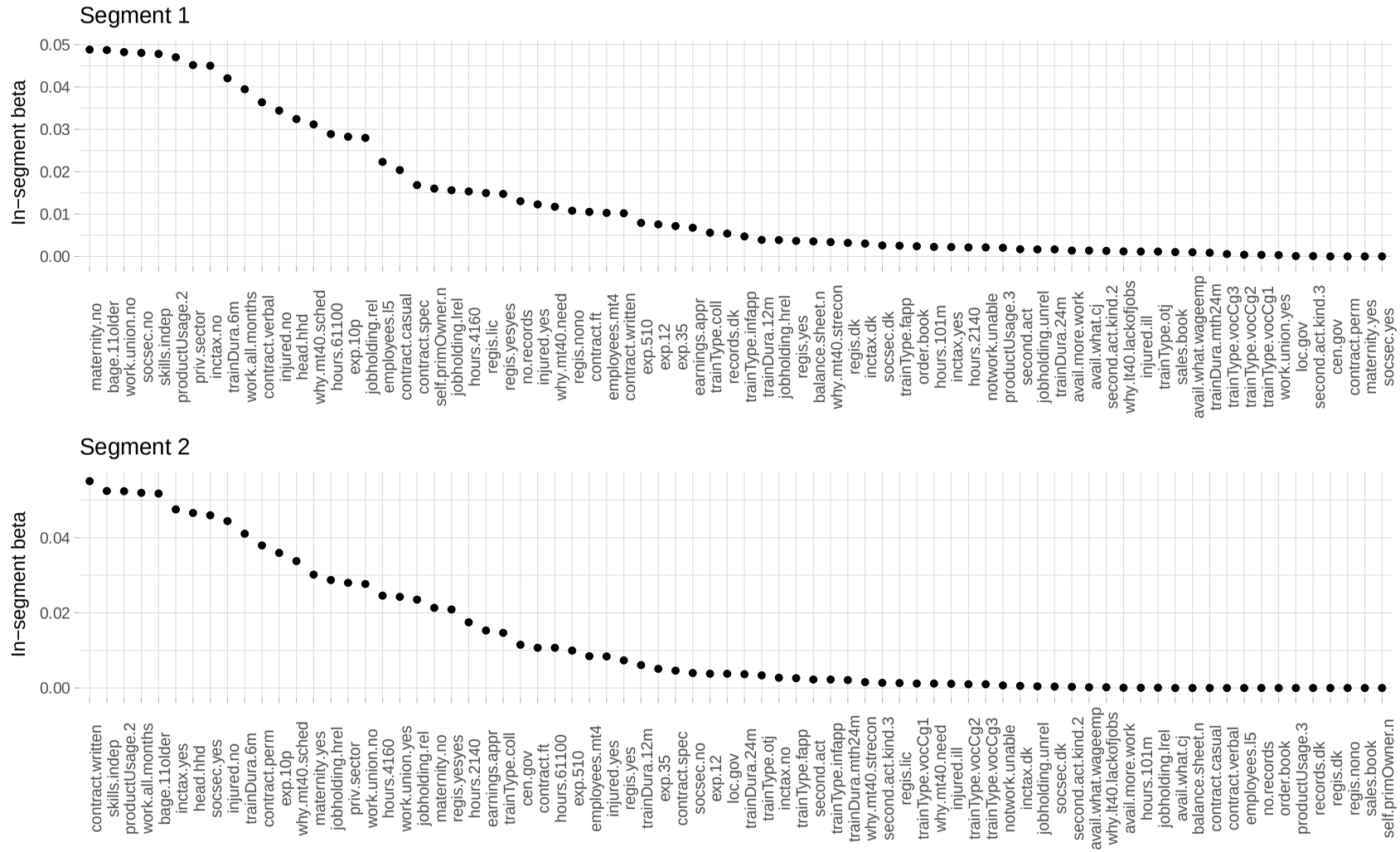
This is figure intends to visualize the effect that different prior beliefs have on the dispersion of the to-be-estimated distribution. Each panel plots the two dimension of a randomly generated two-dimensional Dirichlet distribution with 100 realizations each against each other. The shape paramater ("hyperparameter") was varied between 0.1 (left), 1 (middle) and 10 (right). The dispersion of the points along the line (each pair adds up to 1) indicates the spread of the distribution. Lower shape parameters lead to more polarized distribution while higher ones results in distributions with less mass at the corner points.

Figure A.2: Formality premiums - quantile regressions



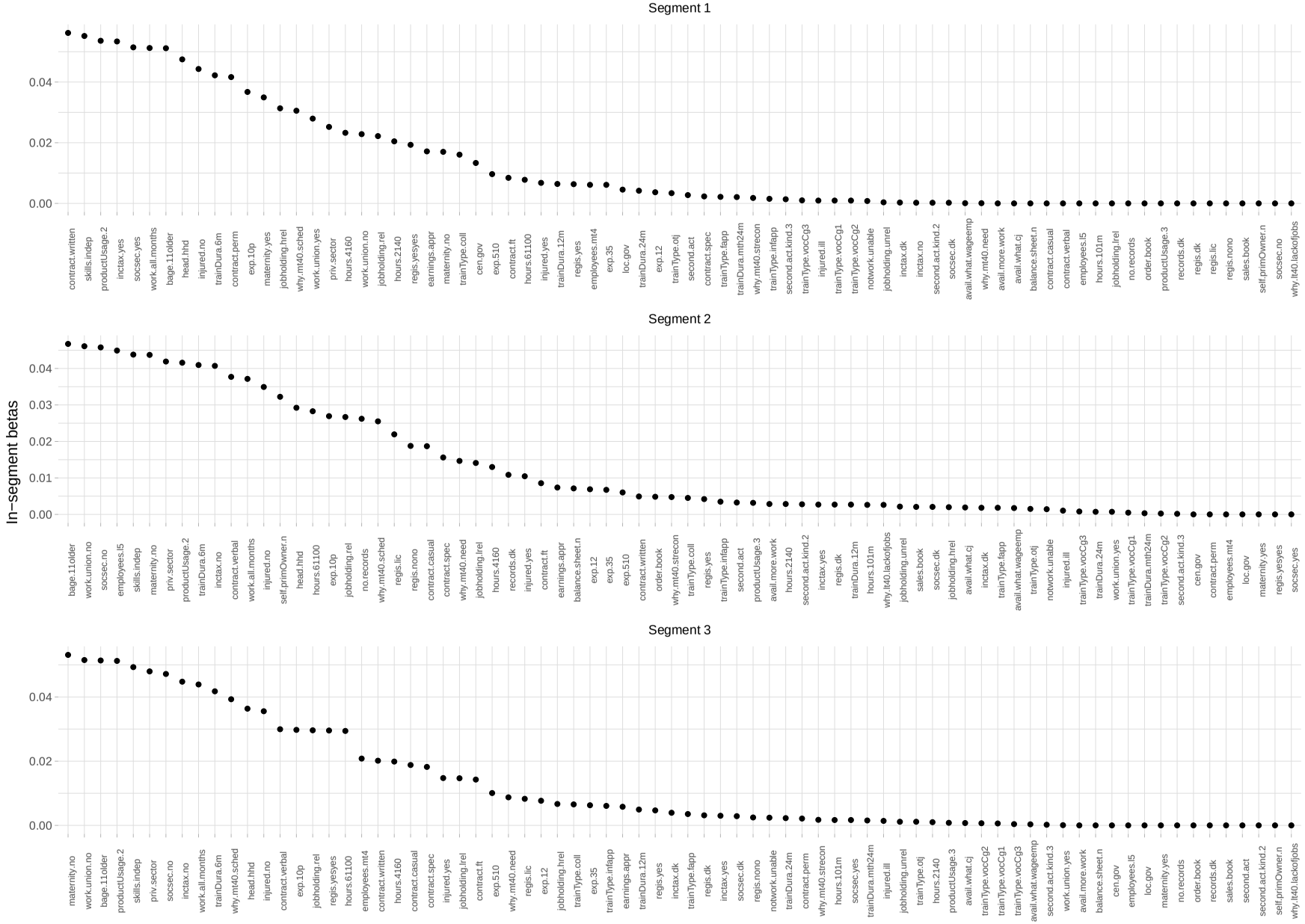
This figure presents estimates of conditional percentiles of an individuals total (cash and in-kind payments) hourly wage. The blue line shows the point estimates for being a "formal" employee according to one of two definitions: "Definition 1" defines formality as making social security contributions, while "Definition 2" defines formality as making social security contributions and paying income tax. Separate point estimates were estimated for the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentile. The regressions include separate indicators for literacy in Kiswahili, English or both, an indicator for being a citizen of a foreign country, indicators for secondary or university education separately and quarter-of-year interview date fixed effects. The lightblue shaded areas indicate the lower and upper bound of a 95% confidence interval based on 2000 bootstrap samples. Results are further broken down by pooling private and public sector employees or restricting the sample to just the former.

Figure A.3: Dualistic view - two latent segments



This figure shows illustrates the segment-feature distributions across the two segments given the dualistic view of the labor market. For each segment, the dots indicate the probability that a contract generated from that segment contains the respective feature. Note that the distributions will sum to one in both panels. The segment-feature probabilities are obtained from estimating a Latent Dirichlet Allocation using Gibbs-Sampling with a 1000 "burn-in"-iterations after which every 400^{th} draw from the next 2000 iterations is taken. The former tries to account for the fact that the Gibbs-sampling will tend to converge only over time while the latter tries to avoid evaluating auto-correlated draws. This procedure is repeated for 40 randomly chosen starting points. Finally, the hyperparameters governing the prior distribution are chosen to be $\alpha = \delta = .1$, reflecting the the fact that contracts tend to have mass on one segment rather than both and segments are unlikely to have feature mass spread equally.

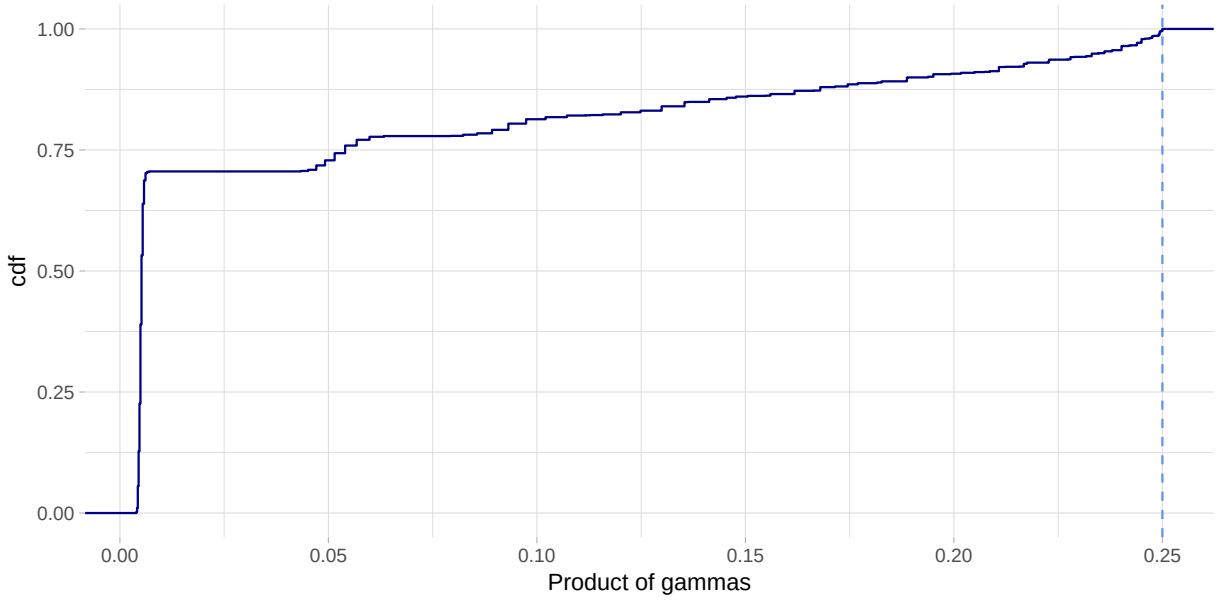
Figure A.4: Three latent segments in the pooled sample - segment-over-feature distribution



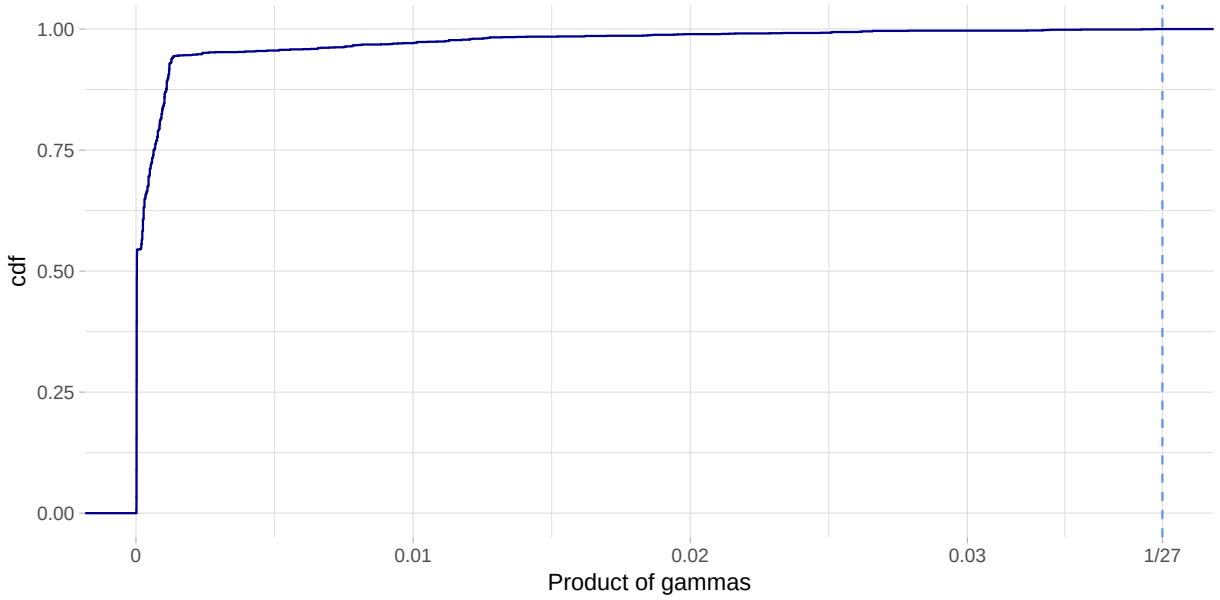
This figure shows illustrates the segment-feature distributions across the two segments given three latent segments in the pooled sample. For each segment, the dots indicate the probability that a contract generated from that segment contains the respective feature. Note that the distributions will sum to one in both panels. The segment-feature probabilities are obtained from estimating a Latent Dirichlet Allocation using Gibbs-Sampling with a 1000 "burn-in"-iterations after which every 400^{th} draw from the next 2000 iterations is taken. The former tries to account for the fact that the Gibbs-sampling will tend to converge only over time while the latter tries to avoid evaluating auto-correlated draws. This procedure is repeated for 40 randomly chosen starting points. Finally, the hyperparameters governing the prior distribution are chosen to be $\alpha = \delta = .1$, reflecting the the fact that contracts tend to have mass on one segment rather than both and segments are unlikely to have feature mass spread equally.

Figure A.5: Cumulative density of of $\Pi_{k=1}^{K}\gamma_k^i$ for the dual ($K = 2$) and tertiary ($K = 3$) model

(a) Two-segment model



(b) Three-segment model



These figures show the cumulative density of the product of the contract γ-distribution, i.e., the "share" of a specific segment in a given contract. Note that regardless of the number of segments estimated, the γ-values for a contract will sum to one. The product of the values for a contract can be thought of as a measure of dispersion of a contract in K-dimensional segment space. In the two-segment model the most ambiguous contract "shares" are $\gamma_{k=1}^i = \gamma_{k=2}^i = 0.5$ which would result in a product of 1/4. This value is indicated by the vertical dashed line in panel a). In the three-segment model ambiguity is maximized at $\gamma_k^i = 1/3$ for a product value of 1/27, indicated by the vertical dashed line in panel b). Note that in both panels the cdf curve appears to span beyond the theoretical maximum. This is a plotting anomaly and there are no actual values beyond the theoretical maximum.