

Dimensionality Reduction for Structural Auction Estimation: A Johnson–Lindenstrauss Approach

Seth Kushniryk

Department of Economics, Queen’s University

October 19, 2026

Abstract

Public tenders come with pages of descriptive text about the work being procured. Structural auction econometrics makes no use of that text: applied work controls for observed differences across auctions with a handful of hand-coded covariates and treats whatever the covariates miss as unobserved heterogeneity. This paper brings the text into structural cost estimation directly, to my knowledge for the first time. A language model converts each tender’s text into a vector of numbers, called a sentence embedding, built so that tenders describing similar work receive similar vectors. The embedding has hundreds of coordinates, far more than the kernel density estimator behind the Guerre–Perrigne–Vuong (GPV) cost recovery can smooth over, so it must first be compressed to a manageable dimension. The central question of this paper is how. The natural default, principal components analysis (PCA), keeps the directions along which tenders vary most. I show that this criterion buys nothing here: the kernel’s bandwidth rule rescales every coordinate by its spread before smoothing, whereby the estimator is indifferent to the variance a compressed coordinate carries, and only the choice of subspace matters. Random projection compresses by a random matrix and offers a guarantee of a different kind, namely that pairwise distances between tenders are approximately preserved wherever the cost-relevant signal happens to sit. Averaging over several independent projections removes the noise the randomness introduces. On public-procurement data from British Columbia and Québec, the random-projection ensemble improves bid-prediction accuracy over a text-free baseline by eight to fourteen percentage points more than PCA does, and it never fails the density-estimation stage on which the GPV inversion rests. I recommend it as the default compression for text covariates in structural auction estimation.

1 Introduction

Public procurement makes up roughly 13% of Canadian GDP (OECD, 2023). A central empirical question about procurement auctions is how far winning bids sit above the cost of doing the work, and the difficulty is that costs are never observed. The standard tool for recovering them is the estimator of Guerre, Perrigne, and Vuong (2000), hereafter GPV, which works backwards from equilibrium bidding: the margin a bidder builds into its bid depends on the competition the bidder expects to face, so once the distribution of bids is estimated, each observed bid can be inverted into the cost that rationalises it. The inversion needs the distribution and density of bids conditional on the auction’s observable characteristics. Both are estimated with a kernel, a local-averaging procedure that estimates the bid distribution at a given auction by averaging over auctions with similar covariates.

Auctions differ in many more ways than any covariate list captures. Applied work conditions on a small hand-coded set, typically project category, region, year, the engineer’s estimate where one exists, and, in markets where hauling matters, each bidder’s distance to the work site. The restraint is forced by the estimator: the kernel that does the local averaging begins to fail the moment it is asked to condition on much of anything, so that past a handful of covariates no two auctions sit close enough to be usefully compared and the estimate degenerates into whichever lone nearest neighbour happens to fall closest to the query point, which is the curse of dimensionality in its plainest and least forgiving form. Heterogeneity the covariate list misses is treated as unobserved, and a considerable literature works around it, for instance the deconvolution approach of Krasnokutskaya (2011).

Much of that unobserved heterogeneity is nonetheless sitting in plain sight. Nearly every public tender carries a page or more of descriptive text about what is to be built, where, and to what specification, and the text speaks directly to cost. Tools from natural-language processing make the text usable in estimation. A sentence-transformer encoder (Reimers & Gurevych, 2019) converts a passage of text into a vector of numbers, called a sentence embedding, built so that texts with similar meaning receive nearby vectors. The first question of this paper is whether these embeddings can enter the GPV inversion as ordinary covariates, whereupon project heterogeneity the literature treats as unobservable becomes something the analyst conditions on.

The obstacle is dimension. The embeddings used here have 384 or 1,024 coordinates, the output widths of the two sentence encoders used below, and a kernel cannot average locally over so many: every auction is far from every other in a space that large, and the

estimated density is worthless; Section 3 makes this precise. The failure is gradual. Each added coordinate shrinks the number of auctions the kernel can usefully compare, so a few dozen coordinates leave the estimator strained at a generous bandwidth, and a thousand leave it with nothing to average over. Compression to somewhere between ten and a hundred dimensions keeps the local sample workable, and estimating the bid density that the cost inversion divides by, a harder task for the kernel than predicting bids, forces a further retreat, to ten (Section 8). How to compress is the paper’s second question, and the two questions, whether the text helps and how to compress it, are not separable: the contribution is joint. To my knowledge this is the first use of text embeddings in structural auction estimation, and the empirics show the novelty is usable only if the compression is chosen well, because the natural choice, principal components analysis (PCA), can leave the estimates worse than discarding the text altogether.

Two compression methods are the natural candidates. PCA (Hotelling, 1933; Pearson, 1901) finds the k directions along which the embeddings vary most and discards the rest. Johnson–Lindenstrauss random projection (JL; Johnson & Lindenstrauss, 1984) multiplies the embedding by a $k \times d$ matrix of random draws; the Johnson–Lindenstrauss lemma guarantees that with high probability the compressed vectors preserve the pairwise distances between auctions to within a small factor. PCA keeps what varies most. JL keeps the geometry, and requires no fitting whatsoever.

Intuition might say that PCA should win. Its coordinates arrive ordered by variance, and one might expect the kernel to read that ordering as an importance ranking, smoothing least along the directions that matter most. Lemma 1 shows the intuition fails for the estimator GPV uses. The standard bandwidth rule, Silverman’s rule of thumb (Silverman, 1986), sets the amount of smoothing on each coordinate in proportion to that coordinate’s own standard deviation; and once the smoothing scales with the spread in this fashion the kernel can no longer tell a high-variance coordinate from a low-variance one, because the very stretching by which PCA arranges its coordinates in order of variance is undone, coordinate by coordinate, by the bandwidth that divides each of them back out again before the first comparison between two auctions is ever made. Variance is rescaled away before smoothing begins. What survives is the choice of subspace and nothing else: PCA helps precisely when its top k directions contain the cost-relevant signal, and nothing in the construction guarantees they do, whereas JL’s distance guarantee holds wherever the signal sits.

Two further results complete the theory. The GPV estimator on JL-compressed covariates converges at the rate a k -dimensional problem would enjoy, with no penalty from the original dimension (Theorem 2). A single random projection does inject noise, inasmuch

as a different draw gives a different estimate; averaging the estimator over M independent draws removes that noise at rate $1/M$ (Proposition 3), and $M = 10$ is enough in practice.

I then take the theory to the data: 667 auctions from BC Bid, British Columbia’s public-tendering platform, and 22,663 tenders from the Système électronique d’appel d’offres (SEAO), Québec’s public-tendering registry. On both datasets, adding compressed embeddings to the hand-coded covariates improves out-of-sample prediction of bids, and the choice of compression matters in the way the theory says it should. At the default bandwidth, the JL ensemble improves prediction accuracy by ten percentage points more than PCA on both BC and Québec; on BC, PCA with text predicts worse than no text at all. Roughly ten percentage points of what earlier work would label unobserved heterogeneity is instead conditioned on. Widening the bandwidth beyond its default rehabilitates PCA, and on Québec it promotes a hybrid, exact top principal components plus a random projection of the remainder, to the top of the prediction ranking. The GPV inversion asks more of the covariates than good prediction, however. It divides by an estimated bid density, and on Québec the hybrid that wins the prediction horse race produces the worst density estimates and the largest share of recovered markups outside the economically sensible range. The JL ensemble sits within noise of the best density fit on both datasets and clear of the Québec failure. The recommendation therefore does not go to the winner of any single horse race, since each race has a different one; it goes to the only method that improves prediction at every bandwidth and passes the density stage on both datasets, and that method is the JL ensemble.

Section 2 relates the argument to prior work. Sections 3 through 6 set out the model, the compression theory, and the hybrid construction. Section 7 checks the mechanism in simulation, where the true costs are known. Sections 8 and 9 contain the BC and Québec applications, together with the density diagnostics behind the recommendation. Section 10 concludes.

2 Related work

The empirical-auctions literature has handled observed heterogeneity mostly in three ways: homogenising bids before estimation (surveyed in Athey & Haile, 2007, Section 3.6), modelling the dependence between bids and covariates semiparametrically (Hubbard, Li, & Paarsch, 2012), or treating the heterogeneity as unobserved and recovering its distribution from the bids themselves (Krasnokutskaya, 2011). The identification apparatus is Athey and Haile (2002), extended to interdependent costs by Somaini (2020). Empirical applications of GPV use small hand-picked covariate sets (e.g., Athey, Levin, & Seira,

2011; Hendricks, Pinkse, & Porter, 2003), and the affiliated-private-values extension (T. Li, Perrigne, & Vuong, 2002) retains the same conditioning structure. Sentence-transformer embeddings sit an order of magnitude above what any of these covariate schemes was designed for. The conditional-independence tests of Bajari and Ye (2003), for instance, presume a covariate space small enough that bids are conditionally independent given the covariates; a 1,024-dimensional embedding does not deliver that. The most direct empirical precedent for a text embedding entering a structural estimator at all is Compiani, Morozov, and Seiler (2026), whose demand-side application feeds PCA-compressed sentence and image embeddings into a mixed-logit random-coefficient model. The compression step there is the point of contact with this paper: Lemma 1 shows that under adaptive-bandwidth kernel estimation (which the GPV inversion of Section 3 requires) PCA delivers no importance ordering to the estimator, and Theorem 2 shows a random-projection alternative attains the same k -dimensional rate at negligible fit cost. Adjacent demand-side work learns the low-dimensional representation itself: Armona, Lewis, and Zervas (2021) recover latent product characteristics from consumer search data, and Magnolfi, McClure, and Sorensen (2025) estimate demand on product embeddings built from crowdsourced similarity comparisons.

The only prior structural application of JL in economics I am aware of is Chiong and Shum (2019), who compress large choice sets in a semiparametric multinomial-choice model. Their theory establishes that the projected estimator is consistent as the projection dimension grows. They also set PCA and the LASSO (least absolute shrinkage and selection operator) aside, on the ground that their estimator is built from moment inequalities involving inner products between covariate vectors, and neither method preserves inner products. My results differ on both counts. Theorem 2 is a finite-sample convergence rate rather than a consistency limit, so nothing in it deteriorates as the ambient dimension grows from 384 to 1,024. And my case against PCA rests on the bandwidth-invariance argument of Lemma 1, which applies to kernel estimators that are not built from inner products at all, the GPV estimator among them.

The theoretical scaffolding for the ensembled version of my estimator comes from the random-projection-ensembles framework of Cannings and Samworth (2017), who establish a $1/M$ convergence of the M -projection ensemble to its infinite-ensemble limit, with excess risk measured against the Bayes classifier. The same $1/M$ averaging carries over to GPV density estimation; the finite ensemble size I use is pinned down not by their Bayes-risk target but by the bias-variance trade-off of this paper’s Proposition 3. The variance bounds on the underlying random-projection matrix come from Achlioptas (2003), with a sparse extension by P. Li, Hastie, and Church (2006). Random Fourier

Features (Rahimi & Recht, 2007) is the standard kernel-aware alternative to JL; on BC it performs indistinguishably from a single JL draw at every k I tried and does not match the ensemble, so it makes no further appearance.

The sentence embeddings themselves are outputs of the transformer architectures of Devlin, Chang, Lee, and Toutanova (2019) and Reimers and Gurevych (2019). The topic-model alternative to sentence embeddings begins with Blei, Ng, and Jordan (2003) and would work under the same compression argument. Using text as a covariate rather than as an observable falls into the research programme surveyed by Gentzkow, Kelly, and Taddy (2019) and updated by Ash and Hansen (2023).

Double/debiased machine learning (ML) (Chernozhukov et al., 2018) shares the projection-and-plug-in shape but not the identification apparatus: here a linear random projection stands in for the fitted ML first stage, and a GPV density inversion for the moment condition.

3 Setup

Index auctions by $\ell = 1, \dots, L$ with n_ℓ bidders each. Bidder i in auction ℓ draws private cost $c_{i\ell} \sim F(\cdot | x_\ell)$ independently, with covariates $x_\ell \in \mathbb{R}^d$. Costs are private values, as opposed to interdependent ones, in the sense of Milgrom and Weber (1982). Under standard GPV regularity (Guerre et al., 2000), the symmetric Bayes-Nash equilibrium maps cost to bid one-to-one and the cost is recovered from the observed bid $b_{i\ell}$ via

$$\hat{c}_{i\ell} = b_{i\ell} - \frac{1}{n_\ell - 1} \cdot \frac{1 - G(b_{i\ell} | x_\ell)}{g(b_{i\ell} | x_\ell)}, \quad (1)$$

with G and g the conditional cumulative distribution function (CDF) and density of bids given x . The $1 - G$ term is the procurement form of the inversion: the lowest bid wins, so a bidder's margin is disciplined by the probability that every rival bids higher. Both objects are estimated by Nadaraya–Watson product kernels (Nadaraya, 1964; Watson, 1964) at Silverman bandwidths (Silverman, 1986):

$$\hat{g}(b | x) = \frac{\sum_j K_h(x - x_j) \phi_{h_b}(b - b_j)}{\sum_j K_h(x - x_j)}, \quad \hat{G}(b | x) = \frac{\sum_j K_h(x - x_j) \Phi((b - b_j)/h_b)}{\sum_j K_h(x - x_j)},$$

where j runs over bids in auctions other than the one being evaluated, K_h is the product Gaussian kernel over the covariate coordinates defined below, ϕ_{h_b} is a Gaussian density with bandwidth h_b on the bid coordinate, and Φ is the standard normal CDF.

Split $x_\ell = (s_\ell, t_\ell)$: $s_\ell \in \mathbb{R}^{d_s}$ is a low-dimensional structural block of hand-coded auction covariates, at $d_s = 37$ on BC Bid and 43 on Québec SEAO. The counts are indicator expansions rather than distinct measurements: roughly half a dozen underlying fields, entered as category, month, year, region, and issuer dummies, the same coding that appears as time, region, and asphalt-type fixed effects in Clark, Gauthier, Coviello, and Shneyerov (2018), account for nearly all of the columns (Sections 8 and 9 list the fields). The text component $t_\ell \in \mathbb{R}^d$ is a sentence-transformer embedding at $d = 384$ (multilingual small model) or $d = 1,024$ (French-specialised large model) depending on the encoder.

At large d the product kernel concentrates. The pair kernel weight,

$$K_h(x_\ell - x_{\ell'}) = \prod_{j=1}^d \phi(h_j^{-1}(x_{\ell,j} - x_{\ell',j})),$$

has expected log-magnitude of order $-d/2$ once distances are typical. In words: as d grows, the weight any auction places on any other shrinks exponentially, and the estimator ends up comparing each auction to its single nearest neighbour. The effective sample size at a query point scales as $N^{4/(d+4)}$ under Silverman bandwidths; at $d = 384$ the exponent is 0.010, and raising the Québec sample of 22,663 auctions to that exponent leaves an effective local sample of about 1.1 observations. Compression of t_ℓ before the kernel is therefore a precondition for running GPV on this class of data. Section 4 shows first that the bandwidth rule neutralises variance ranking, which frames the choice among compressions. Section 5 then develops JL and its ensemble, Section 6 the PCA-head hybrid, and Appendix A.2 a kernel-matched anisotropic variant.

4 Silverman equivalence and PCA's residual edge

Which compression to choose turns on what the kernel does with the compressed coordinates, and the bandwidth rule settles much of that question by itself. Under a product Gaussian kernel with per-coordinate Silverman bandwidths, the estimator is invariant to per-coordinate rescaling of its input, so any linear compression of t to $\mathbb{R}^k$ produces an output whose coordinate variances are absorbed into the bandwidth at the kernel step.

Lemma 1 (Silverman equivalence). *Let $z, \tilde{z} \in \mathbb{R}^k$ be two linear compressions of the same input, related by $\tilde{z}_j = c_j z_j$ for some positive constants $c_1, \dots, c_k$. Let $\hat{g}(b | z)$ and $\hat{g}(b | \tilde{z})$ denote the product-Gaussian kernel estimators applied to z and $\tilde{z}$, each with its own Silverman bandwidth vector. Then*

$$\hat{g}(b | z) = \hat{g}(b | \tilde{z}).$$

The proof is a direct calculation: Silverman gives $h_j \propto \sigma_j$, and the kernel weight for a pair of observations equals $\prod_j \exp(-(z_j - z'_j)^2 / (2h_j^2)) = \prod_j \exp(-(\tilde{z}_j - \tilde{z}'_j)^2 / (2\tilde{h}_j^2))$ because both numerator and denominator scale by c_j^2 .

Comparing PCA and JL under this equivalence produces this paper’s central mechanism. PCA’s coordinates arrive with decreasing variances $\sigma_1 > \sigma_2 > \dots > \sigma_k$; JL’s coordinates (Section 5) carry roughly equal variance. Lemma 1 says the kernel cannot tell the difference: rescale any coordinate of either output and the kernel weights do not move, so neither variance pattern reaches the estimator. What the estimator does respond to is the subspace the compression keeps, since bias depends on which directions of t survive. PCA’s usual advantage therefore rests entirely on alignment: it beats JL when the top- k eigenspace of t contains the cost-relevant signal, and not otherwise.

The anisotropic construction (Appendix A.2) samples the projection matrix with per-coordinate variances w_j , so that the distances it preserves are weighted ones, $\|u\|_W^2 = \sum_j w_j u_j^2$, rather than the ordinary Euclidean $\sum_j u_j^2$. Setting $w_j = 1/\sigma_{t_j}^2$ measures each coordinate in units of its own standard deviation, and a distance measured that way, called a diagonal Mahalanobis metric, is exactly the distance the product Gaussian kernel with Silverman bandwidths computes when it compares two auctions. The construction requires d scalars but no fitted basis. Its rate (Theorem 2) matches Theorem 2, with the approximation-error constant defined under the weighted metric. The empirical edge over ensembled JL is between 0.02 and 0.12 percentage points on the two datasets, inside every wild-cluster standard-error band the paper reports, and the anisotropic variant appears in the empirical tables and in the recovered-cost consensus without a separate main-text treatment.

The next two sections develop the two responses. Section 5 gives the random projection, whose distance guarantee does not depend on alignment, together with its convergence rate and its ensemble. Section 6 gives a hybrid that hard-codes the top- p eigendirections into the compressed feature, for the case where alignment does hold.

5 Johnson–Lindenstrauss projection

5.1 Definition

Sample $A \in \mathbb{R}^{k \times d}$ with independent and identically distributed (iid) entries $A_{ij} \sim \mathcal{N}(0, 1/k)$, fix A once, and set $z_\ell = At_\ell$; GPV runs in the projected covariate space $\mathbb{R}^{d_s+k}$ with s_ℓ passed through uncompressed. The output dimension k is the only parameter, and no data-fit step enters the construction.

The Johnson–Lindenstrauss lemma (Johnson & Lindenstrauss, 1984) guarantees that a random projection of this kind preserves every pairwise distance in a fixed set of N points to within a factor of $(1 \pm \varepsilon)$ once the target dimension k is of order $\log N / \varepsilon^2$. The sharp distributional form, that a Gaussian or sign projection achieves this with probability at least $1 - 2 \exp(-c \varepsilon^2 k)$ and an explicit constant c , is due to Dasgupta and Gupta (2003) and Achlioptas (2003). A fast Hadamard-preconditioned variant of Ailon and Chazelle (2009) cuts projection cost from $O(dk)$ to $O(d \log d + k^3)$. I use the dense construction: at $k \leq 100$ any projection-side savings are dwarfed by the $O(N^2 kd)$ downstream kernel summation.

5.2 Convergence rate

Write the projected conditional density as $g^A(b \mid s, z) = \mathbb{E}[g(b \mid s, t) \mid At = z]$, and define $\hat{g}^A$ as the kernel estimator of g^A in the $(d_s + k)$ -dimensional space.

Theorem 2 (JL convergence rate). *Define the projection approximation error*

$$\text{bias}_{\text{JL}}(b, s, t; A) = |g^A(b \mid s, At) - g(b \mid s, t)|.$$

Under standard GPV regularity (Guerre et al., 2000), the projected kernel density estimator $\hat{g}^A(b \mid s, z)$ satisfies

$$|\hat{g}^A(b \mid s, z) - g(b \mid s, t)| = O_P\left(L^{-R/(2R+d_s+k+3)}\right) + \text{bias}_{\text{JL}}(b, s, t; A),$$

where R is the order of smoothness assumed for $g(\cdot \mid \cdot)$. The first term is the standard $(d_s + k)$ -dimensional GPV rate. The second is a fixed constant of the projection: it depends on neither the sample size, nor the bandwidth, nor the kernel.

Proof in Appendix A.1. The theorem leaves bias_{JL} as a defined quantity rather than bounding it; a bound would require an assumption on how the embedding distribution concentrates around a low-dimensional set, which I do not impose. The distance-preservation guarantee is nonetheless the reason to expect the constant to be small: the projection preserves the between-auction geometry the kernel reads, so auctions nearby in t remain nearby in z . bias_{JL} is exactly the bias floor of Proposition 3, and the ensemble-size sweeps of the Monte Carlo and both empirical sections measure it.

Numerically, the exponent is $R/(2R + d_s + k + 3)$ rather than $R/(2R + d_s + d + 3)$. At $R = 1$, $d_s = 43$, and the paper’s working $k = 50$, the projected exponent is 0.0102; at $d = 1,024$, the dimension of the largest Québec embedding (Section 9), the ambient exponent is 0.0009, eleven times smaller. An exponent that small means the unprojected bound improves by

well under one per cent for every tenfold increase in the sample, in effect not converging at any attainable scale. Standard Silverman on $\mathbb{R}^{d_s+k}$ is therefore the right bandwidth for the projected space; no JL-aware modification is needed.

The closest prior use of JL in a structural estimator, Chiong and Shum (2019), obtains consistency of the projected estimator as the projection dimension grows but no explicit rate; Theorem 2 gives a finite-sample rate at fixed k . Section 2 details the contrast, including why their inner-product argument against PCA has no purchase on a kernel estimator.

5.3 Ensembled JL

A single JL projection is a random object and its estimator has draw-to-draw variance. Averaging M independent estimators, each on an independent draw $A^{(m)}$, reduces that variance at rate $1/M$.

Proposition 3 (Ensemble bias-variance). *Let $\hat{g}^{(m)}$ denote the kernel estimator using projection $A^{(m)}$, and define the ensemble $\hat{g}_M^{\text{ens}} = M^{-1} \sum_m \hat{g}^{(m)}$. Then*

$$\text{Bias}(\hat{g}_M^{\text{ens}}) = \mathbb{E}_A[\hat{g}^{(m)}] - g, \quad \text{Var}(\hat{g}_M^{\text{ens}}) = \frac{1}{M} \text{Var}_A(\hat{g}^{(m)}).$$

The proof follows from linearity and iid draws. Ensembling leaves the bias unchanged and shrinks the draw-attributable variance. Tolerating draw variance up to 2δ times the squared bias gives an optimum

$$M_{\text{stat}}^* = \frac{\sigma_{\text{draw}}^2}{2\delta \cdot \text{bias}_{\text{JL}}^2},$$

where δ is the tolerance that the analyst allows for tracking the asymptotic floor. On BC and Québec the empirical elbow lands at $M^* \in [10, 25]$ across every embedding, dataset, and k I test. Cannings and Samworth (2017) report the same $1/M$ scaling in a classification setting.

6 PCA-head, JL-tail hybrid

6.1 Motivation and construction

Lemma 1 rules out a variance-ranking advantage for PCA, but alignment remains a possibility. If the top eigenspace of t happens to contain the cost-relevant signal, PCA captures that signal exactly, where JL captures only a random shadow of it; in that case PCA carries

less bias, and in the opposite case more. The hybrid keeps the top p principal components exactly and compresses the residual subspace by JL, so it captures the alignment where it exists while spending the same total dimension as pure PCA or pure JL.

Let $V_p \in \mathbb{R}^{p \times d}$ hold the top- p principal-component loadings of t , computed from training data, and define the head $h_\ell = V_p t_\ell \in \mathbb{R}^p$ and the residual $r_\ell = (I - V_p^\top V_p)t_\ell$. Draw $A \in \mathbb{R}^{q \times d}$ with entries iid $\mathcal{N}(0, 1/q)$ and set the tail $\tilde{z}_\ell = Ar_\ell \in \mathbb{R}^q$; the kernel takes $(s_\ell, h_\ell, \tilde{z}_\ell) \in \mathbb{R}^{d_s+p+q}$, with the head deterministic given the training data and the tail random and independent of the head, at total compression dimension $k = p + q$.

6.2 Convergence rate

Theorem 4 (Hybrid convergence rate). *Define $\text{bias}_{\text{hyb}}(b, s, t; V_p, A) = |g^{\text{hyb}}(b | s, V_p t, Ar) - g(b | s, t)|$, with $r = (I - V_p^\top V_p)t$. Under standard GPV regularity (Guerre et al., 2000), the hybrid kernel density estimator satisfies*

$$|\hat{g}^{\text{hyb}}(b | s, h, \tilde{z}) - g(b | s, t)| = O_P\left(L^{-R/(2R+d_s+p+q+3)}\right) + \text{bias}_{\text{hyb}}(b, s, t; V_p, A).$$

If $g(b | s, t)$ depends on t only through $V_p t$, then $\text{bias}_{\text{hyb}} = 0$.

The first term is the $(d_s + p + q)$ -dimensional GPV rate. The second is the approximation error from compressing the residual: signal captured by the head contributes nothing to it (the zero case in the statement), and signal on the tail contributes through the compression of the $(d - p)$ -dimensional residual rather than of the full d . Proof in Appendix A.1, adapting Theorem 2 to the two-part feature vector.

6.3 Bias decomposition

The hybrid's total bias decomposes into head and tail contributions:

$$\underbrace{g(b | s, t) - g^{\text{hyb}}(b | s, h, \tilde{z})}_{\text{total bias}} = \underbrace{g(b | s, t) - g^{V_p}(b | s, h)}_{\text{head bias (data-fitted)}} + \underbrace{g^{V_p}(b | s, h) - g^{\text{hyb}}(b | s, h, \tilde{z})}_{\text{tail bias (JL, random)}}$$

g^{V_p} is the conditional density that observes only the head. Head bias vanishes if the true signal lies in the span of V_p , and tail bias is the JL approximation error on $V_p^\perp$. Against pure PCA at dimension p , the hybrid adds a JL guarantee on the tail. Against pure JL at dimension $p + q$, it fixes p exact directions and confines the q random draws to a smaller subspace. At matched total dimension it therefore weakly improves on both special cases in approximation bias. Estimation error is another matter: the finite-sample density fit

can still rank the hybrid below either special case, and Section 9.5 records a case where it does.

6.4 Data-dependence of the head

The head requires an $O(\min(N, d)^3)$ singular value decomposition (SVD) of the training embedding, and V_p inherits the data-fitting properties PCA has: head coordinates depend on the training sample, and V_p shifts under resampling. How far it shifts is measured by principal angles, the natural notion of rotation between two subspaces: an angle of zero means one span contains the other, and a right angle means some direction of the first is orthogonal to all of the second. On the Québec embedding, the principal angles between V_p fit on the full sample and V_p fit on random subsamples of $m = 15,000$ auctions stay under 0.06 radians, about three and a half degrees, for $p \leq 50$ across twenty replications; the fitted head barely moves under resampling, so bootstrapping V_p inside the resampling loop is not needed. The one-time SVD cost is dominated by downstream kernel-regression compute at both dataset scales.

6.5 Ensembling the tail

Only the tail is ensembled. Fix V_p , draw $A^{(1)}, \dots, A^{(M)}$ iid, and average the M tail predictions. The head contributes no draw-to-draw variance. The tail variance falls at $1/M$ with the same optimal $M^* = \sigma_{\text{draw}}^2 / (2\delta \cdot \text{bias}_{\text{JL, tail}})$ as pure JL. Empirical elbows on BC and Québec sit at $M \in [10, 25]$.

Pre-whitening the embedding, that is, rotating it to its principal axes and rescaling each axis to unit variance before projecting, and then applying JL might look like a way to combine the hybrid with the anisotropic variant, but the chain reduces to a single operation.¹

¹Let V be the full PCA basis of t and $\Lambda = \text{diag}(\lambda_j)$ its eigenvalues. Setting $\tilde{t} = \Lambda^{-1/2} V^\top t$ and $z = A\tilde{t}$ with $A_{ij} \sim \mathcal{N}(0, 1/k)$, the isotropic draw (equal variance in every input direction), $z = A't$ has row covariance Σ_t^{-1}/k . Whitened-plus-isotropic-JL is therefore a full-covariance anisotropic JL on the original coordinates. Lemma 1 then undoes the per-coordinate scaling at the Silverman kernel step, and what remains is a random projection in the rotated basis under $\|\cdot\|_{\Sigma_t^{-1}}$. The product kernel does not implement that metric (it treats coordinates independently and does not correct off-diagonal covariance), so whitened-JL preserves distances the kernel partially discards and loses both the diagonal-Mahalanobis alignment the anisotropic variant delivers and the head alignment the hybrid delivers. Empirically it is null on both datasets.

7 Monte Carlo

7.1 Design

The Monte Carlo (MC) exercise checks Lemma 1’s mechanism on simulated data, where the truth is known. Six latent factors $z_\ell \sim \mathcal{N}(0, I_6)$ drive everything: three visible factors, which the analyst observes through the structural block, and three hidden factors, which the analyst observes only through the high-dimensional covariate. The outcome is a linear index of all six,

$$y_\ell = z_\ell^\top \beta + 0.3 \varepsilon_\ell, \quad \beta = (1, 0.7, 0.5 \mid 3, 1, \tfrac{1}{2}), \quad \varepsilon_\ell \sim \mathcal{N}(0, 1).$$

The structural block $s_\ell \in \mathbb{R}^5$ mixes the three visible factors through a standard-normal loading matrix, plus noise of standard deviation 0.05. The text-like covariate $t_\ell \in \mathbb{R}^{100}$ mixes the three hidden factors through standard-normal loadings scaled by their outcome coefficients, and then adds η times a mixture of twenty nuisance factors that play no role in y , plus the same 0.05 noise. The parameter η therefore sets the nuisance share of t ’s variance: at low η the outcome-relevant factors dominate the covariance of t , the case most favourable to PCA, and at high η the nuisance factors dominate.

Each method compresses t_ℓ to $k = 10$ coordinates. The estimator is the leave-one-out (LOO) Nadaraya–Watson prediction of y_ℓ given s_ℓ and the compressed t_ℓ , at per-coordinate Silverman bandwidths with multiplier 2.5 (the calibrated value the empirical sections use), and the root-mean-squared error (RMSE) reported in the tables is the prediction error of that regression: thirty replications per cell, $N = 5,000$ observations each, $\eta \in \{0, 0.3, 1, 3, 10\}$. The main grid therefore tests the regression channel of the mechanism; the first-price design of Section 7.4 embeds a scaled-down version of the same covariate structure in an auction model and tests the cost-recovery channel. Ambient dimension is $d_t = 100$ rather than the Québec empirical $d = 1,024$ because Lemma 1 is scale-free in d , and running at $d = 1,024$ would recover the same qualitative gaps at ten times the wall-clock and no additional mechanism content.² The standard error of each cell mean stays under 0.05 everywhere on the grid, so the qualitative gaps reported below (all above 0.20 for JL versus PCA, above 1.0 at low η) are well outside Monte Carlo noise.

²The MC’s $O(N^2kd)$ kernel-summation cost scales linearly in d : five nuisance cells $\times$ four methods $\times$ thirty replications at $d = 100$, $N = 5,000$, $k = 10$ ran in roughly 2.6 CPU-hours on my machine, so a $d = 1,024$ analogue would cost about a CPU-day.

7.2 Adaptive per-coordinate bandwidth

The hidden-factor loading vector is $\lambda = (3, 1, \frac{1}{2})$, which concentrates signal in the top eigendirection of t and gives PCA the data-generating process (DGP) most favourable to its variance-ranking argument (a signal-orthogonal-to-top-eigenspace DGP would trivially favour JL and would not test the paper’s mechanism). The cleanest test of Lemma 1 on this DGP is at $\eta = 1$, where PCA delivers RMSE 2.63 against 2.30 for ensembled JL: PCA is beaten by thirteen per cent even in the regime where its variance ranking is best-aligned with the signal, and Lemma 1’s prediction that the ranking has no purchase under adaptive Silverman is confirmed.

At $\eta = 0$ the DGP itself is degenerate for PCA in a way that is not a clean mechanism test: the population covariance of t has rank $L_{\text{hid}} = 3$, sample eigenvectors 4-through- k carry near-zero variance, and adaptive Silverman scales those into near-zero bandwidths that drive the kernel weights on those coordinates towards Kronecker deltas. The estimator refuses to smooth along the noise-only coordinates, and the LOO evaluation degenerates. Lemma 1 still holds at the boundary but the estimator becomes numerically fragile there. JL avoids the fragility because its projected coordinates all have similar variance and get similar bandwidths. The $\eta = 0$ result is included in Table 1 for completeness rather than as the mechanism confirmation; the $\eta \in \{0.3, 1\}$ columns are where the ranking predicted by Lemma 1 obtains under well-conditioned PCA.

At $\eta \geq 3$ the nuisance directions dominate the signal and all four methods converge inside 0.06 RMSE, so the grid is uninformative there. The hybrid at $p = 3, q = 7$ tracks PCA at low η because its head captures the same top eigendirections and its tail is too narrow to help; widening the tail brings the hybrid towards the JL end (Table 1).

Table 1: Monte Carlo mean RMSE by nuisance η and method under adaptive per-coordinate Silverman; thirty replications per cell, standard error of each cell mean under 0.05. Bold marks methods within 0.1 RMSE of the column minimum. The $\eta \leq 1$ columns identify the mechanism: PCA sits far outside the tied JL and anisotropic cluster. At $\eta \geq 3$ every method ties.

Method	$\eta = 0$	$\eta = 0.3$	$\eta = 1$	$\eta = 3$	$\eta = 10$
PCA $k = 10$	2.550	2.555	2.634	3.017	3.326
JL $k = 10$, single	1.489	1.746	2.358	3.050	3.299
JL $k = 10$, ens $M = 10$	1.492	1.712	2.300	3.026	3.295
Anisotropic $k = 10$, single	1.540	1.845	2.447	3.070	3.300
Anisotropic $k = 10$, ens $M = 10$	1.512	1.780	2.354	3.024	3.293
Hybrid $p = 3, q = 7$, single	2.550	2.553	2.635	3.012	3.297
Hybrid $p = 3, q = 7$, ens $M = 10$	2.550	2.554	2.632	3.016	3.296

7.3 Fixed scalar bandwidth

Switching the adaptive per-coordinate bandwidth for a fixed scalar (the Silverman formula evaluated at the mean coordinate standard deviation, same 2.5 multiplier) turns Lemma 1’s mechanism off and gives PCA back its variance advantage. This is not a recommended bandwidth and does not appear in the empirical sections; it is a diagnostic for the mechanism. High-variance coordinates now enter the kernel exponent with strictly more weight, and PCA’s variance ranking picks up feature-importance content. At $\eta = 0$ the ranking flips: PCA delivers RMSE 1.28 against ensembled JL’s 1.54. The hybrid inherits PCA’s head advantage and stays within 0.01 of PCA at every $\eta \leq 1$. At $\eta = 3$ the four methods converge inside 0.03 RMSE, and at $\eta = 10$ the anisotropic variant edges the others by 0.06. Table 2 shows the full grid.

Table 2: Monte Carlo mean RMSE under a fixed scalar bandwidth (Silverman-mean across coordinates); same DGP as Table 1. Bold marks methods within 0.1 RMSE of the column minimum. PCA and hybrid tie at low η ; non-PCA methods converge inside 0.03 RMSE at $\eta \geq 3$.

Method	$\eta = 0$	$\eta = 0.3$	$\eta = 1$	$\eta = 3$	$\eta = 10$
PCA $k = 10$	1.284	1.445	1.860	3.017	3.473
JL $k = 10$, single	1.554	1.697	2.250	3.082	3.433
JL $k = 10$, ens $M = 10$	1.538	1.640	2.123	3.020	3.421
Anisotropic $k = 10$, single	1.522	1.753	2.351	3.111	3.433
Anisotropic $k = 10$, ens $M = 10$	1.487	1.653	2.171	3.014	3.418
Hybrid $p = 3, q = 7$, single	1.291	1.485	1.943	3.032	3.431
Hybrid $p = 3, q = 7$, ens $M = 10$	1.291	1.486	1.933	3.000	3.423

7.4 First-price cost recovery

A separate specification adapts the Monte Carlo to first-price procurement auctions, at a smaller scale than the regression grid: the text-like covariate has $d_t = 60$ coordinates, ten nuisance factors stand in for twenty, and the outcome coefficients are $\beta = (1, 0.6, 0.3 \mid 1.5, 0.8, 0.4)$, with the same standard-normal loadings and 0.05 noise as before. Costs are $c_{i\ell} \sim \text{Unif}[0, \mu(z_\ell)]$ with $\log \mu(z_\ell) = z_\ell^\top \beta$, three iid bidders per auction, Bayes-Nash bids $b_{i\ell} = c_{i\ell}(n-1)/n + \mu(z_\ell)/n$. Each method compresses the text-like covariate to $k = 8$ coordinates (the hybrid at $p = 3, q = 5$). Method choice enters through the estimated conditional bid density $\hat{g}(b \mid x)$ and CDF $\hat{G}(b \mid x)$: the Section 3 estimators, evaluated leave-one-bid-out at 2.5 times the per-coordinate Silverman bandwidth on the covariates and 1.5 times on the bid coordinate. (Excluding a bid’s own auction-mates, as the empirical

sections do, is unnecessary here: simulated costs are iid within an auction.) The GPV inversion $\hat{c} = b - (1 - \hat{G})/((n - 1)\hat{g})$ produces cost estimates from those, with $\hat{g}$ floored at its second percentile so the inversion cannot explode where the estimated density is near zero. Because the costs are simulated, the reported cost-recovery RMSE is computed against the true draws c_{it} . Across 200 replications at $N_{\text{auctions}} = 250$, every method beats the structural baseline (the same estimator run on the structural block alone, with no text component), by seven per cent RMSE on recovered costs at $\eta = 0$ and tapering to about two and a half per cent at $\eta \geq 3$ as the nuisance dilutes what the text adds. The four methods stay within one per cent of one another at every $\eta \geq 1$ and about two per cent at $\eta = 0$, with the internal ranking swapping across the grid (the anisotropic variant and single JL give the tightest costs at $\eta = 0$; hybrid and PCA are marginally better at $\eta = 3$). The same relative insensitivity of the density-based inversion to projection choice, against a 10-point regression gap, recurs at much larger scale on real data in Section 9.

An ensemble-size sweep on the $\hat{c}$ Monte Carlo (one hundred replications, $M \in \{1, 5, 10, 25, 50\}$) confirms Proposition 3's $1/M$ variance decay extends from conditional-mean estimation to conditional-density estimation. The ensemble averages $\hat{g}$ and $\hat{G}$ across the M draws and then applies the inversion once to the averages. At $M = 1$ the JL $\hat{c}$ RMSE is 12.05; at $M = 10$ it is 11.98; the elbow lies between $M = 10$ and $M = 25$, matching the elbow observed on both real datasets. Figure 1 shows that no method uniformly dominates on cost recovery in the first-price MC: methods stay within one to two per cent of one another across the grid, and the ranking swaps as η moves. Figure 2 shows the two bandwidth regimes as separate panels, and Figure 3 the ensemble decay for both $\hat{c}$ and $\hat{y}$.

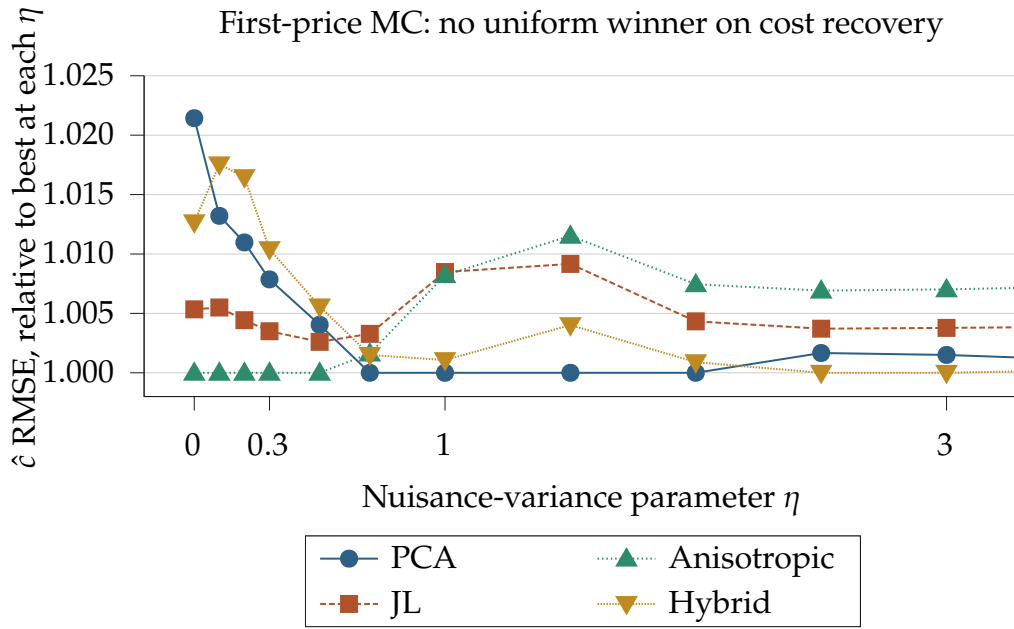


Figure 1: First-price MC: $\hat{c}$ RMSE by method and nuisance η , relative to the best method at each η . Method spread stays inside one per cent at every $\eta \geq 1$ and just over two per cent at $\eta = 0$, and the ranking is not stable: the anisotropic variant leads at $\eta = 0$, the hybrid at $\eta = 3$. Density-based inference is more forgiving of projection choice than conditional-mean regression.

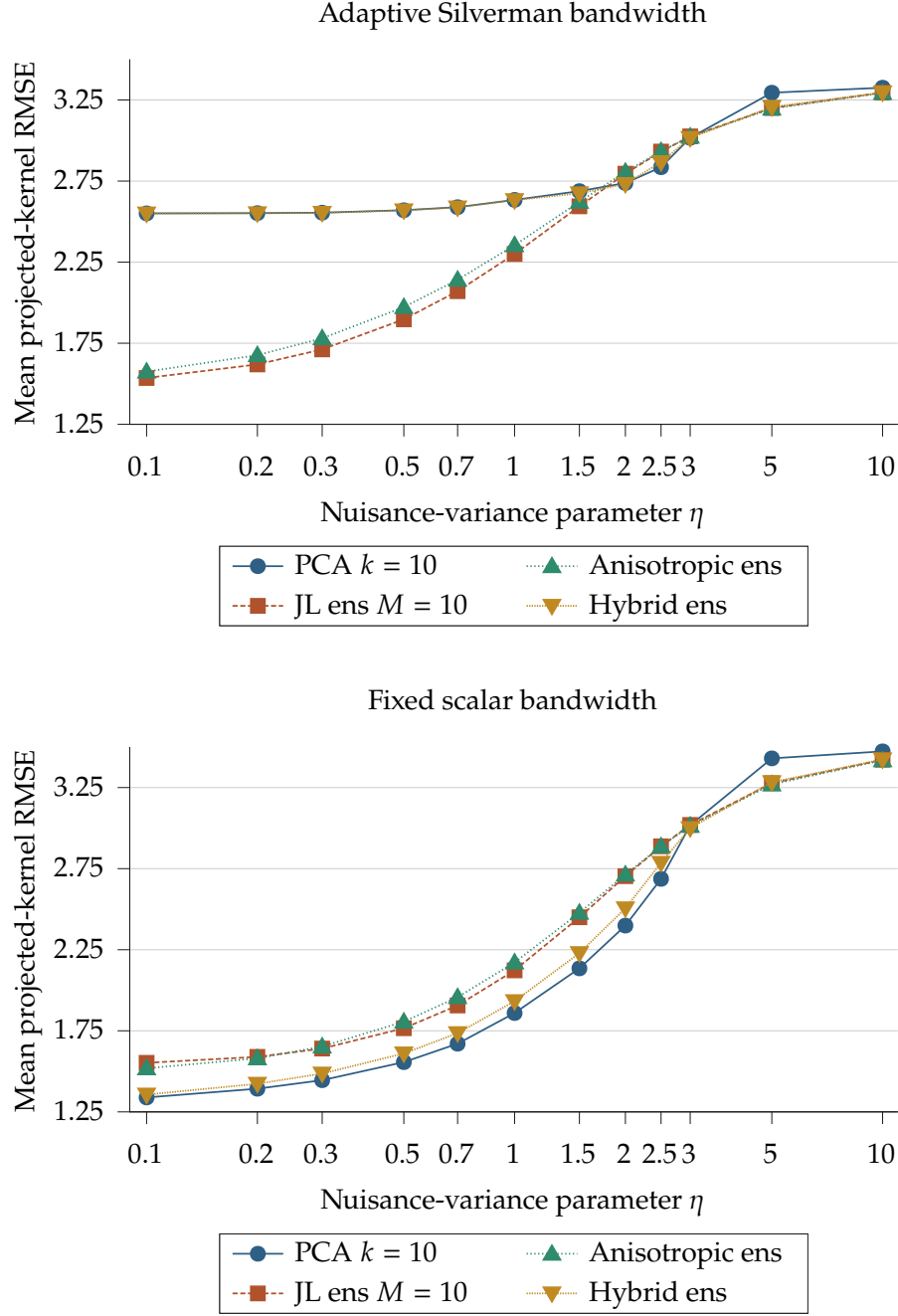


Figure 2: Monte Carlo mean RMSE by nuisance η under two bandwidth regimes; $\eta = 0$ baseline omitted by the log- x axis. Top panel: adaptive per-coordinate Silverman, where PCA underperforms across the range while the non-PCA methods share the lead. Bottom panel: fixed scalar bandwidth, where the ranking flips at low η and PCA wins by margin while the hybrid tracks its head advantage. Same DGP in both panels, only the bandwidth rule changes. Illustrates Lemma 1 empirically.

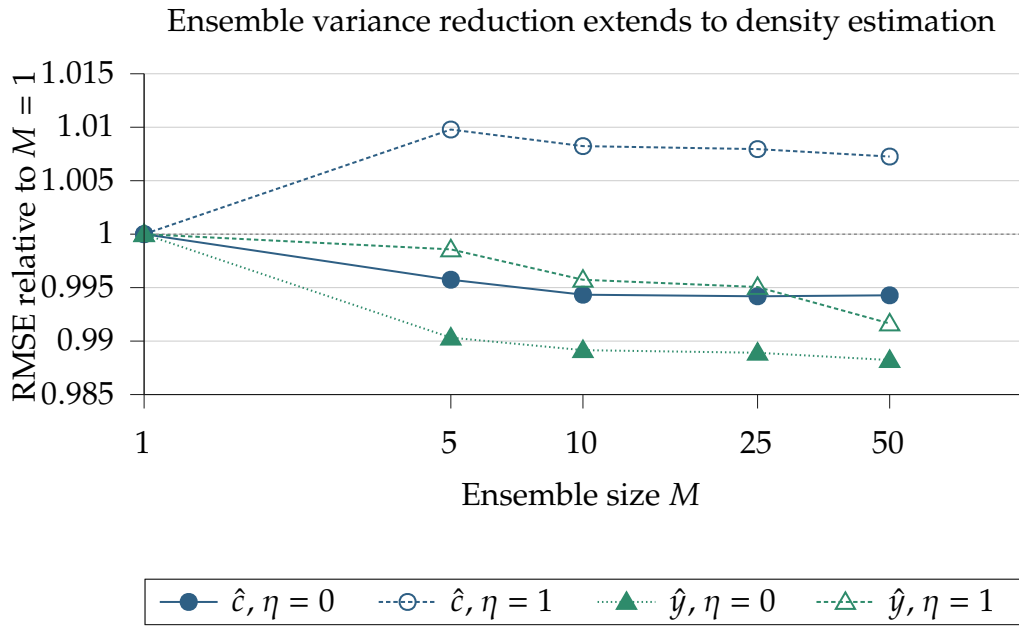


Figure 3: Ensemble-size sweep on the first-price MC. RMSE relative to the $M = 1$ estimate falls in M for both $\hat{c}$ (cost recovery) and $\hat{y}$ (bid regression). The elbow at $M \approx 10$ agrees with the elbow on both real datasets. Proposition 3 extended empirically from conditional-mean to density-based inference.

8 BC Bid

This section and the next run the same two-stage design on the two datasets: BC Bid here, Québec’s SEAO in Section 9. The first stage is a prediction horse race: each compression method enters the kernel regression of log bids on the structural block and the compressed text, and its score is the leave-one-auction-out RMSE reduction against the structural baseline. Prediction is not the end goal, but it is the sharpest available measure of how much cost-relevant signal each compression delivers to the kernel, and it can be measured precisely enough to rank the methods. The second stage runs the machinery the markups need: held-out fit of the estimated bid density, and the economic sensibility of the markups the GPV inversion recovers from it. On real data the true costs are never observed, so there is no direct measure of whether one method’s recovered costs land closer to the truth than another’s. The two stages are observable substitutes: the leave-one-auction-out RMSE reduction scores the prediction stage, the held-out log-likelihood measures how well the estimated bid density $\hat{g}$, the object the GPV inversion divides by, describes bids it was not fit on, and the feasibility and monotonicity of the recovered markups check the output against restrictions the theory imposes. The first-price Monte Carlo (Section 7.4), where the true costs are known, supplies the check thereanent: adding compressed text there improves recovered-cost accuracy directly.

The recommendation is settled by one rule, stated now so that no later result can be read as picking a favourite after the fact: a compression is recommended only if it improves on the structural baseline at the prediction stage at both the default and the calibrated bandwidth, and then passes the density stage the inversion divides by. Winning a single race at a single tuning does not meet the rule. Both possible failures occur in the data: PCA’s prediction gains collapse at the default bandwidth on both datasets, and the hybrid’s calibrated-bandwidth prediction win on Québec does not survive the density stage (Section 9.5). Figure 8 collects the verdicts once both datasets are in.

8.1 Data

BC Bid is British Columbia’s public-sector procurement platform. My sample is 667 auctions and 3,240 bids under the standard filter $n_\ell \geq 3$, with a 37-dimensional structural block s_ℓ of hand-coded auction covariates (number of bidders, project-category indicators over the ten most common contract categories, log estimated contract value, month and year indicators, geographic region indicators, and issuing-agency indicators over the eight largest issuers) and a 384-dimensional sentence embedding of the concatenated tender text

produced by the multilingual MiniLM-L12 encoder of Reimers and Gurevych (2019).³ The regression outcome is $y_\ell = \log(\text{bid amount})$. RMSE is computed leave-one-auction-out: when predicting a bid, the kernel is barred from using any bid in the same auction, not merely the bid itself. The distinction matters, because bids within an auction are highly correlated; letting the kernel see a bidder’s own auction-mates makes the prediction problem artificially easy and inflates measured RMSE reductions on BC by roughly a factor of eight. The ratio of auctions to covariates is admittedly thin, at eighteen to one before the compressed text enters. The wide calibrated bandwidth compensates in part, since the kernel pools across much of the sample rather than fitting cell by cell, and the uncertainty that remains shows up in the wild-cluster standard errors below, at one and a half to two percentage points on every BC reduction.

8.2 Bandwidth and dimension choices

Throughout the empirics the bandwidth on covariate coordinate j is

$$h_j = m \cdot 1.06 \sigma_j n^{-1/(d_s+k+4)},$$

where σ_j is the training-sample standard deviation of that coordinate, n the number of training bids, and m a scalar multiplier. The Silverman default is $m = 1$; the calibrated choice is $m = 2.5$. The sweep tables and the calibration grid of Table 3 index their rows by this multiplier m , the only bandwidth dial in the paper besides the bid coordinate’s own. Density estimation adds the bid coordinate, whose bandwidth is $h_b = m_b \cdot 1.06 \sigma_b n^{-1/5}$ with its own multiplier m_b .

The primary empirical result runs at the default Silverman bandwidth $m = 1$, which is the bandwidth against which Lemma 1’s mechanism is stated and the only bandwidth that carries a theoretical guarantee. Table 4 reports the calibrated bandwidth $m = 2.5 \times$ Silverman as a robustness check to show that all methods still improve on the structural baseline once the bandwidth is widened enough for the kernel to smooth. The calibrated bandwidth is picked by a coarse held-out log-likelihood (LL) grid over $m \in \{0.5, 1.0, 2.5, 5.0\} \times$ Silverman on a single five-fold split after the projection; the LL maximum lands at $m = 2.5$ on that grid for both datasets (the fourteen-point grid below refines the BC optimum to 2.0).

Proposition 3 predicts that the Silverman scale minimising the asymptotic mean integrated squared error (AMISE), the standard bandwidth-selection criterion, is projection-

³Model card: paraphrase-multilingual-MiniLM-L12-v2.

specific, so I calibrate it per method. Table 3 reports a fourteen-point held-out log-likelihood grid over $m \in \{0.3, 0.5, 0.7, 1.0, 1.25, 1.5, 1.75, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 6.0\}$, computed per (dataset, method) at auction-partitioned five-fold cross-validation (CV). The grid is run once for both datasets and reported together here, ahead of the Québec analysis, so that Section 9 can refer back to a single calibration table. On BC every method calibrates to $m^* = 2.0$ within a 0.05-log-likelihood band, so the projection-specific optimum coincides on this dataset. On Québec three of the four methods (PCA, ensembled JL, anisotropic ensemble) also share $m^* = 2.5$, matching the earlier coarse-grid calibration; the hybrid alone prefers $m^* = 3.0$, consistent with its higher compressed dimension ($p + q = 100$ against $k = 50$ elsewhere) pulling the AMISE optimum wider through the Silverman rate exponent. The empirical tables below use $m = 2.5$ across all methods on Québec for cross-method comparability at a single bandwidth, and the BC tables stay at 2.5 likewise, the 2.0-versus-2.5 difference there being worth under 0.05 log-likelihood; sensitivity to the hybrid’s projection-specific $m^* = 3.0$ is a 0.03-log-likelihood improvement over the shared calibration on the density diagnostic and does not shift the primary regression ranking. Each empirical claim below that depends on the calibrated bandwidth also carries a companion result at the default $m = 1$, so the recommendation is not conditional on this single choice.

Table 3: Held-out log-likelihood-optimal Silverman multiplier m^* per (dataset, method) from a fourteen-point CV grid $m \in \{0.3, 0.5, 0.7, 1.0, 1.25, 1.5, 1.75, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 6.0\}$, auction-partitioned five-fold; the reported LL is the auction-averaged log-density at the CV-optimal m^* . Both datasets are calibrated here; the Québec analysis follows in Section 9.

Dataset	PCA $k = 50$	JL ens $M = 10$	Aniso ens	Hybrid $p = q = 50$
BC	2.0 (−1.52)	2.0 (−1.49)	2.0 (−1.49)	2.0 (−1.53)
Québec	2.5 (−1.45)	2.5 (−1.46)	2.5 (−1.46)	3.0 (−1.43)

Projection dimension $k = 50$ is the working choice for the regression stage on both datasets. The exponent controlling the Nadaraya–Watson estimation error at feature dimension $d_s + k$ is $R/(2R + d_s + k + 3)$; at $R = 1$, $d_s = 37$ (BC) or 43 (QC), and $k = 50$, the rate constant is $L^{-0.010}$ and the estimator has effective sample $N^{4/(d_s+k+4)} \approx N^{0.045}$. The density stage adds a bid coordinate that pushes the exponent past its usable range at $k = 50$; the density diagnostic therefore runs at $k = 10$ (Section 9.5). Ensembling ($M = 10$) is enough to sit inside the Proposition 3 elbow on both datasets and both stages.

8.3 Results

At the calibrated bandwidth the top three methods on BC deliver reductions between 8.67% and 8.85% at $k = 50$ (and $M = 10$ for the ensembles). The hybrid at $p = 50, q = 50$ trails at 6.62% because adding a large head does not pay on BC: the leading eigendirections of the concatenated embedding do not concentrate the cost signal. Bootstrap standard errors on the individual reductions run between 1.4 and 1.9 percentage points, so the top three methods are indistinguishable at the calibrated bandwidth.

The BC evidence that discriminates between methods therefore sits at the default bandwidth, where PCA turns negative and ensembled JL remains stable (Figure 4 and Table 5), rather than at the calibrated one. The calibrated-bandwidth reductions establish that all methods still improve on the structural baseline once the bandwidth is widened enough for the kernel to smooth. Appendix A.2 argues for the anisotropic variant on kernel-metric grounds, but its empirical edge over ensembled JL is inside the standard-error band on both datasets; I default to ensembled JL absent an external argument for the anisotropic weights.

Table 4: BC Bid: RMSE reduction vs structural baseline at the calibrated bandwidth $m = 2.5 \times$ Silverman. Bold marks methods within one bootstrap standard error of the column best: the top three tie at this bandwidth, and the evidence that separates them sits at the default bandwidth (Table 5).

Method	RMSE	Reduction
Structural only	1.197	0.00%
Hybrid $p = 50, q = 50$, ensembled $M = 10$	1.117	+6.62%
Anisotropic $k = 50$, ensembled $M = 10$	1.093	+8.67%
JL $k = 50$, ensembled $M = 10$	1.093	+8.69%
PCA $k = 50$	1.091	+8.85%

The ranking depends on the bandwidth. Sweeping $m \in \{0.5, 1.0, 2.5, 5.0\} \times$ Silverman, ensembled JL wins or ties everywhere except $m = 2.5$; PCA is competitive only at $m = 2.5$ and turns negative at $m \leq 1.0$. At the default bandwidth $m = 1$, ensembled JL delivers +1.56% against PCA’s −8.58%, a gap of 10.1 percentage points. At $m = 0.5$ the gap widens to 12.44 points.⁴ The anisotropic and hybrid variants track ensembled JL across the curve. Single-shot JL is a null result at every bandwidth, as Proposition 3 suggests. Figure 4 shows both datasets and Table 5 gives the numbers.

⁴Treating Silverman’s 1.06 constant itself as the multiplier ($m = 1.06$) gives +2.16% against −5.92% on BC and +11.42% against +2.14% on Québec, gaps of 8.1 and 9.3 points. Nothing turns on the distinction between the two defaults.

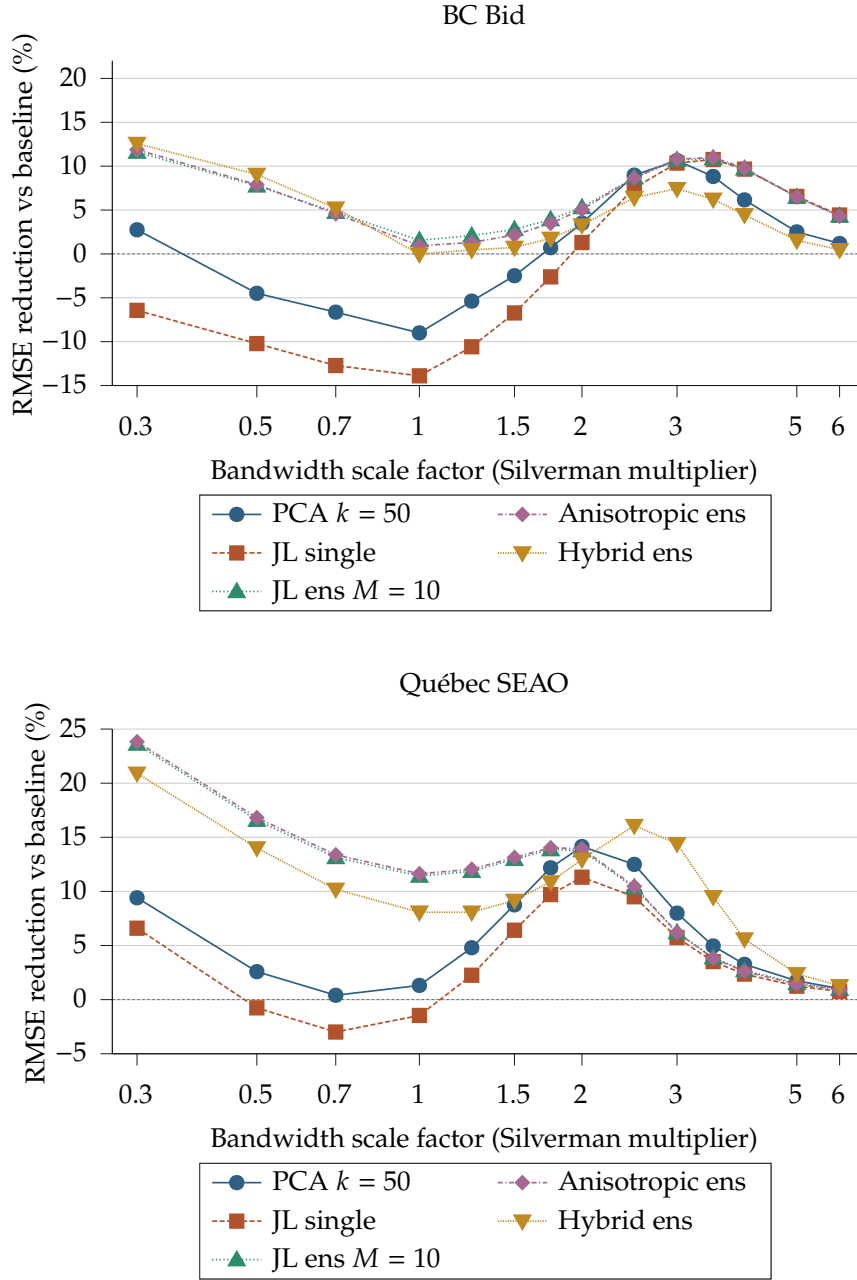


Figure 4: RMSE reduction against structural baseline as a function of the Silverman bandwidth multiplier, five methods per panel. Top: BC Bid. Bottom: Québec SEO on Solon-large embedding. Ensembled JL (navy solid) is the winner or within noise of the winner at every bandwidth on BC and at every bandwidth below the calibrated $m = 2.5$ on Québec, where the hybrid leads from $m = 2.5$ on; PCA is competitive only at $m = 2.5$ and turns negative on BC at lower bandwidths. The curves come from a finer fourteen-point sweep run separately from Table 5’s grid; at shared bandwidths the two runs differ by up to half a percentage point through re-drawn projections and data preparation.

Table 5: RMSE reduction (%) vs structural baseline by bandwidth multiplier and method, both datasets. Bold marks methods within one wild-cluster standard error of the row maximum (≈ 2 percentage points on BC, ≈ 1 on Québec), computed at the calibrated bandwidth in Tables 6 and 9. A single boldface indicates the winner sits outside noise from the runner-up; multiple boldface indicates a statistical tie for the row.

m	PCA $k = 50$	JL single	JL ens $M = 10$	Aniso ens $M = 10$	Hybrid ens $M = 10$
<i>BC Bid, MiniLM-L12-384</i>					
0.5	-4.68	-10.22	+7.76	+7.90	+8.30
1.0	-8.58	-13.90	+1.56	+0.93	+0.24
2.5	+8.85	+7.53	+8.69	+8.67	+6.62
5.0	+2.54	+6.55	+6.47	+6.49	+1.56
<i>Québec SEAO, Solon-large-1024</i>					
0.5	+2.57	-0.76	+16.58	+16.81	+14.02
1.0	+1.45	-1.46	+11.42	+11.63	+8.06
2.5	+12.51	+9.50	+10.35	+10.47	+16.07
5.0	+1.71	+1.24	+1.51	+1.42	+2.38

8.4 Inference

The small BC sample complicates inference. I run an auction-level wild-cluster bootstrap (Cameron, Gelbach, & Miller, 2008; MacKinnon & Webb, 2018) at $R = 9,999$ replications, holding the $\{V_p, A^{(m)}\}$ projections fixed and reweighting the auction-level residuals by Rademacher signs.

The standard errors (SEs) and p -values below rest on two different bootstrap draws. Standard errors use the *unrestricted* bootstrap, which reweights residuals from the fitted projection model and gives sampling variability around the point estimate. The p -values use the *restricted* bootstrap, which reweights residuals from the null (structural-only) fit and gives the sampling distribution of the reduction statistic under H_0 of no improvement over the structural baseline; reported p -values are one-sided fractions of $R + 1 = 10,000$ restricted draws whose reduction reaches the observed one.⁵

Table 6 reports the point estimate, wild-cluster standard error, and 95% confidence interval (CI) point $\pm 1.96 \cdot \text{SE}$ for all four methods. Every reduction is significant at $p < 10^{-4}$, the smallest achievable at $R = 9,999$. The top three methods have standard errors between 1.4 and 1.9 percentage points. The hybrid’s standard error is 3.75, and its 95% interval narrowly includes zero even though the p -value under the null rejects at

⁵Mammen’s two-point weights give the same SEs on the unrestricted variant to within 0.1 percentage points on every cell, so I report Rademacher only. Kernel matrices are precomputed once per projection because K is a fixed function of $(X, h, \text{auction ids})$ under the wild bootstrap; each replication reduces to a matrix-vector product against K , mathematically identical to a full LOO kernel pass.

the same 10^{-4} threshold (all bootstrap draws under H_0 land below the observed 6.62%). A placebo permutation of the outcome across auction identifiers, with covariates and projections held fixed, gives a mean placebo reduction of -3.52% with standard deviation 1.18, twelve percentage points from the real number and far beyond anything chance fitting of the replicate structure could produce.

Table 6: BC Bid: wild-cluster bootstrap inference on RMSE reduction at the calibrated bandwidth $m = 2.5 \times \text{Silverman}$. Rademacher weights, unrestricted residuals for SEs, restricted (null-model) residuals for p -values; $R = 9,999$ replications. Every method rejects H_0 (no reduction over structural baseline) at the smallest achievable one-sided p .

Method	Reduction (%)	SE	95% CI	p (H_0)
PCA $k = 50$	+8.85	1.94	[+5.05, +12.65]	$< 10^{-4}$
JL $k = 50$, ensembled $M = 10$	+8.69	1.40	[+5.95, +11.43]	$< 10^{-4}$
Anisotropic $k = 50$, ensembled $M = 10$	+8.67	1.50	[+5.73, +11.61]	$< 10^{-4}$
Hybrid $p = q = 50$, ensembled $M = 10$	+6.62	3.75	[−0.73, +13.97]	$< 10^{-4}$

8.5 Density diagnostic

Held-out log-likelihood of the bid density (the average log-density the estimator assigns to bids it did not train on) is computed across five auction-level folds, at $k = 10$ and bandwidth multipliers $(m, m_b) = (2.5, 1.5)$. On BC it is -1.546 for PCA, -1.563 for a single JL draw and for the anisotropic variant (ensembling moves JL to -1.559), and -1.548 for the hybrid at its regression configuration $p = q = 50$; redrawing the random projections moves the non-PCA numbers by roughly 0.01. All four methods land inside 0.02 log-likelihood of one another, and the feasibility fraction, the share of bids whose recovered markup $\hat{m} = (b - \hat{c})/b$ lands in the economically sensible range $[0, 1]$, sits between 80.8% and 84.2% with the hybrid best (Table 11). Density fit on BC is therefore approximately projection-invariant, and nothing in it penalises the hybrid; the hybrid’s density troubles in Section 9 are a Québec phenomenon. Regression at any bandwidth other than the calibrated one, meanwhile, gives ensembled JL an 8-to-12 percentage-point lead. Section 9 runs the same two stages at much larger scale, where their disagreement decides the recommendation.

9 Québec SEAO

9.1 Data and encoders

The Système électronique d’appel d’offres (SEAO) is Québec’s public-sector procurement registry; the tables abbreviate the Québec sample as QC. Clark et al. (2018) used municipal-clerk records from the Montreal and Quebec City asphalt markets in their bid-rigging investigation, whereas SEAO covers a broader panel of provincial and municipal tenders across every construction category. My sample is 22,663 public-construction tenders under $n_\ell \geq 3$ with 110,646 associated bids, after filtering to \$-denominated lump-sum submissions; nine further tenders clear the bidder filter but lack a usable winning contract amount and are dropped. Because the contract data are one row per tender, the auction-level outcome $y_\ell = \log(\text{winning contract amount})$ is the only regression target available (no bid-level analogue exists), and RMSE is auction-level LOO; scoring at the auction level also removes the within-auction leakage that a bid-level target must otherwise be masked against (Section 8). Covariates are a 43-dimensional structural block (number of bidders, log estimated value, UNSPSC (United Nations Standard Products and Services Code) family indicators aggregated to the six most common families, contract-type indicators for lump-sum versus unit-price, month and year indicators, and issuer indicators for the twelve largest issuing organisations) together with a text embedding of the tender’s descriptive page.

I compare three encoders on the Québec data.⁶ Multilingual MiniLM-L12 gives a 384-dimensional embedding, French masked-language Camembert-large gives 1,024 dimensions, and the French-specialised sentence encoder Solon-large gives 1,024 dimensions. Under the hybrid ensemble at $p = 50$, $q = 50$, $M = 10$ at the calibrated bandwidth, MiniLM returns an 8.87% RMSE reduction, Camembert 12.71%, and Solon 16.0%. The encoder gap dwarfs the within-encoder variation across the p -sweep, with all three encoders peaking at $p = 50$ (Figure 5). Appendix A.5 shows that continuing each encoder’s pretraining on the tender texts themselves (masked-language-model fine-tuning, hereafter MLM) does not close this gap; raw Solon returns 16.01% at the winning cell against 13.48% for MLM-tuned Solon and 13.88% for MLM-tuned MiniLM. All subsequent Québec results use raw Solon.

⁶Model cards: paraphrase-multilingual-MiniLM-L12-v2, dangvantuan/sentence-camembert-large, OrdalieTech/Solon-embeddings-large-0.1.

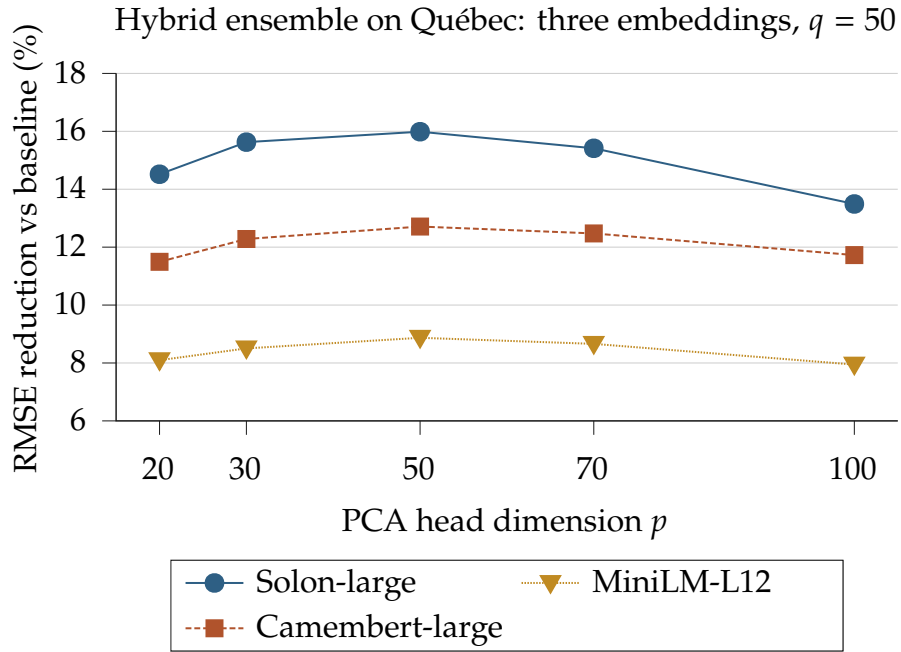


Figure 5: Québec SEAO hybrid ensemble ($q = 50, M = 10$) across three sentence encoders. RMSE reduction against structural baseline as a function of PCA head dimension p . Solon-large dominates at every p by margin larger than any within-embedding variation. All three embeddings peak at $p = 50$.

9.2 Results

At the default Silverman bandwidth $m = 1$ ensembled JL delivers an 11.42% RMSE reduction against 1.45% for PCA at $k = 50$, a 10.0-percentage-point gap consistent with Lemma 1’s prediction that PCA loses its variance-ranking edge under adaptive Silverman. The default-bandwidth result is Table 5’s Québec row.

Table 7 reports the calibrated bandwidth $m = 2.5\times$ Silverman as a robustness check. There the hybrid at $p = 50, q = 50, M = 10$ delivers 16.07% reduction against 12.51% for PCA at $k = 50$ and 10.35% for ensembled JL. Head-tail decomposition pays on Québec at this bandwidth because the top 50 eigendirections of the Solon embedding concentrate the cost-relevant signal, which Section 9.3’s principal-components diagnostic verifies against tender metadata. The hybrid’s calibrated-bandwidth win is therefore conditional on both the bandwidth widening and the encoder-alignment diagnostic returning positive; neither condition holds automatically on a target dataset. Sub-sample bootstrap at $m = 5,000, R = 20$ puts a standard error of 1.44 percentage points on the hybrid reduction and gives a [14.09%, 18.43%] interval that clears both zero and PCA’s point estimate at the calibrated bandwidth.

Table 7: Québec SEAO: RMSE reduction vs structural baseline at the calibrated bandwidth $m = 2.5\times$ Silverman, Solon-large embedding. Bold marks the stage-one winner at this bandwidth, the hybrid; under the rule of Section 8 the win is provisional until the density stage, which the hybrid fails (Section 9.5).

Method	RMSE	Reduction
Structural only	1.313	0.00%
PCA $k = 50$	1.149	+12.51%
JL $k = 50$, ensembled $M = 10$	1.177	+10.35%
Anisotropic $k = 50$, ensembled $M = 10$	1.176	+10.47%
Hybrid $p = 50, q = 50$, ensembled $M = 10$	1.102	+16.07%

The Québec bandwidth curve looks like BC’s. At $m = 1$ ensembled JL returns +11.42% against PCA’s +1.45%, a 10-point gap. At $m = 0.5$ the gap opens to 14 points (+16.58% against +2.57%). The hybrid tracks ensembled JL at $m \leq 1.0$ and only pulls ahead at $m = 2.5$. The magnitudes deserve comparison. At every bandwidth a practitioner might plausibly default to, the gap between the hybrid and ensembled JL amounts to two or three percentage points, while the gap between ensembled JL and PCA runs to ten or fourteen. The bandwidth multiplier thus moves the headline number by more than the projection choice that is this paper’s subject, a matter whereanent the literature primarily addresses bandwidth rates and boundary conditions (Hickman & Hubbard, 2015; Marmer

& Shneyerov, 2012) rather than guidelines for practitioners.

The hybrid’s (p, q) grid peaks at $(50, 50)$ on every encoder. On Solon, the p -sweep at $q = 50$ returns 14.5%, 15.6%, 16.0%, 15.4%, and 13.5% at $p = 20, 30, 50, 70, 100$ (the headline table’s 16.07% is the same $p = q = 50$ configuration in the bandwidth-sweep run; the 0.1-point difference is tail-draw noise); the companion q -sweep at $p = 50$ returns 15.13%, 15.73%, 15.96%, 15.49%, 14.22% at $q = 20, 30, 50, 70, 100$ (Figure 6). Both sweeps decline symmetrically away from 50, so the peak is not a boundary artefact.

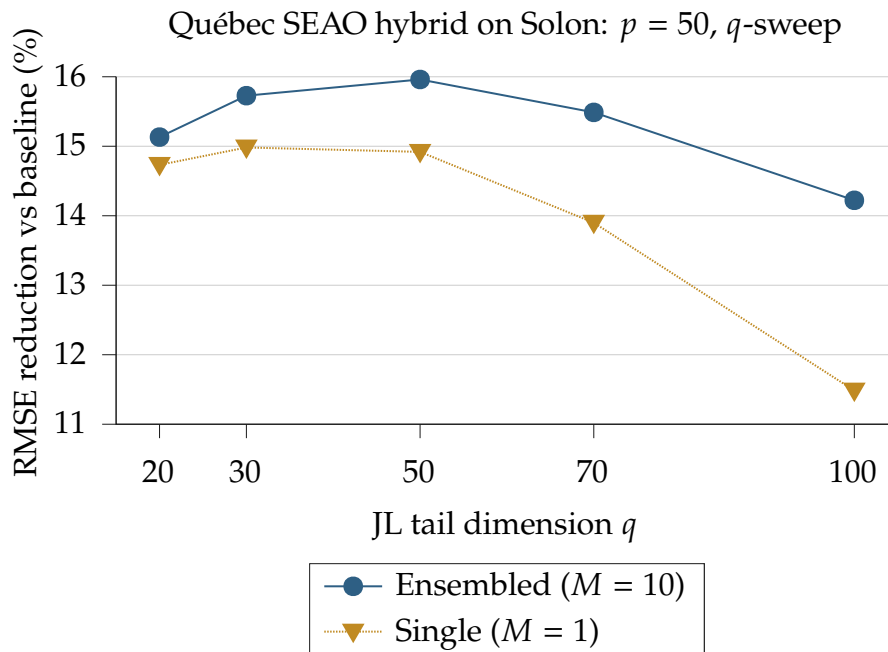


Figure 6: Companion q -sweep at $p = 50$ on Solon. Peak at $q = 50$ matches the p -sweep of Figure 5. The (p, q) surface therefore has a well-defined interior maximum, not a boundary artefact.

9.3 Principal-components diagnostic

The hybrid wins on Québec and near-ties on BC. To see why, I compare Solon’s top five principal components (PCs) against two sources of tender metadata unused elsewhere in the paper: UNSPSC family and the identity of the issuing organisation.

Table 8 reports mean PC scores by UNSPSC family for the six families with at least one hundred tenders. Solon’s PC1 alone accounts for 7.8% of embedding variance and separates specialty-trade construction (family 7213, mean PC1 = +0.14) and structural-materials procurement (family 3022, +0.05) from general building construction (family 7200, −0.13) and building maintenance (family 7210, −0.08). PC2 and PC3 give finer

within-cluster distinctions but with weaker signal. On the issuer dimension, PC1 and PC2 cluster tenders from the Ministère des Transports (across three name variants over the sample period) tightly together and away from tenders issued by the Ville de Québec.

Solon’s leading eigendirections therefore track project type and issuer identity rather than language style, and both are standard cost drivers on which structural procurement models condition. A compression that captures them at $p = 50$ has in effect recovered the structural covariates from the text. The hybrid’s head does this exactly, by construction; JL captures the same signal, but spread over a random subspace with draw-to-draw variance on top.

On BC the same diagnostic returned no clean cost interpretation for any of the top three MiniLM eigendirections; leading MiniLM variance on BC seems to track language style, document format, and attachment count rather than project type. That matches the BC empirical result: neither PCA nor the hybrid lifts BC by the Québec margin because there is no cost-relevant structure in BC’s leading variance to exploit.

Table 8: Mean Solon PC score by UNSPSC family, tenders with matching UNSPSC assignment ($N \geq 100$). Positive PC1 indicates specialty trade or materials procurement; negative PC1 indicates general building construction or maintenance.

UNSPSC family	Description	N	PC1	PC2
3022	Structural materials	2,736	+0.05	+0.01
7116	Utility infrastructure	214	−0.10	−0.01
7200	General building construction	5,099	−0.13	−0.02
7210	Building support and maintenance	6,648	−0.08	+0.01
7213	Specialty-trade construction	4,719	+0.14	+0.05
7219	Specialty maintenance services	2,935	+0.15	−0.08

9.4 Inference

Inference on Québec uses the same auction-level wild-cluster bootstrap as BC (Section 8), scaled to $L = 22,663$ clusters, $R = 9,999$ replications, Rademacher weights and unrestricted residuals. At Québec’s cluster count the standard errors are an order of magnitude tighter than on BC: 0.30 percentage points on PCA, 0.26 on ensembled JL, 0.25 on anisotropic JL, and 0.67 on the hybrid. Table 9 reports the intervals. Every method again rejects the no-improvement null at the smallest achievable one-sided $p < 10^{-4}$. A placebo permutation of the outcome across auctions returns a mean reduction of -9.0% with standard deviation 0.5, twenty-five percentage points below the real number and far outside any bootstrap interval. The hybrid’s confidence interval $[+14.75\%, +17.39\%]$ sits fully above every other

method's, so its regression-stage lead over PCA and JL at the calibrated bandwidth is not attributable to sampling variation.

Table 9: Québec SEAO: wild-cluster bootstrap inference on RMSE reduction at the calibrated bandwidth $m = 2.5 \times$ Silverman, Solon-large embedding. Rademacher weights, unrestricted residuals for SEs, restricted residuals for p -values; $R = 9,999$ replications.

Method	Reduction (%)	SE	95% CI	p (H_0)
PCA $k = 50$	+12.51	0.30	[+11.92, +13.10]	$< 10^{-4}$
JL $k = 50$, ensembled $M = 10$	+10.35	0.26	[+9.85, +10.85]	$< 10^{-4}$
Anisotropic $k = 50$, ensembled $M = 10$	+10.47	0.25	[+9.98, +10.96]	$< 10^{-4}$
Hybrid $p = q = 50$, ensembled $M = 10$	+16.07	0.67	[+14.75, +17.39]	$< 10^{-4}$

Table 10: Bootstrap summary statistics on RMSE reduction (%). Québec uses $m = 5,000$ auctions sampled with replacement, twenty replications. BC uses $m = 400$ auctions sampled without replacement, thirty replications. Each dataset shows the method pair its own comparison turns on: the hybrid-versus-PCA race across encoders on Québec, the PCA-versus-anisotropic tie on BC.

Dataset	Embedding	Method	Mean	Std	Q05	Q95
QC	MiniLM-L12-384	PCA $k = 50$	+10.60	0.99	+9.11	+12.09
QC	MiniLM-L12-384	Hybrid ens	+12.09	1.22	+10.30	+14.12
QC	Camembert-large	PCA $k = 50$	+12.62	1.09	+11.20	+14.47
QC	Camembert-large	Hybrid ens	+12.60	1.51	+10.89	+15.53
QC	Solon-large	PCA $k = 50$	+15.65	1.10	+14.30	+17.06
QC	Solon-large	Hybrid ens	+15.68	1.44	+14.09	+18.43
BC	MiniLM-L12-384	PCA $k = 50$	+7.38	1.97	+4.24	+10.24
BC	MiniLM-L12-384	Aniso ens	+7.38	1.64	+4.89	+9.95

9.5 Density-stage failure of the Québec regression winner

At the calibrated bandwidth the hybrid takes the regression stage with a 95% wild-cluster interval [+14.75%, +17.39%] that clears every other method's interval. Under the rule of Section 8 that win is provisional: the GPV inversion needs more than the regression, because it divides by the estimated bid density, and at the density stage the hybrid falls behind. The density gap between the hybrid and ensembled JL is small in log-likelihood terms (0.029 at the density stage), far smaller in relative terms than the regression gap the hybrid wins, so the call rests on three judgements: that standard GPV does go through the full joint density; that the density ranking is not the artefact of a single bandwidth, which Figure 9 confirms; and that the hybrid's feasibility shortfall carries real economic cost,

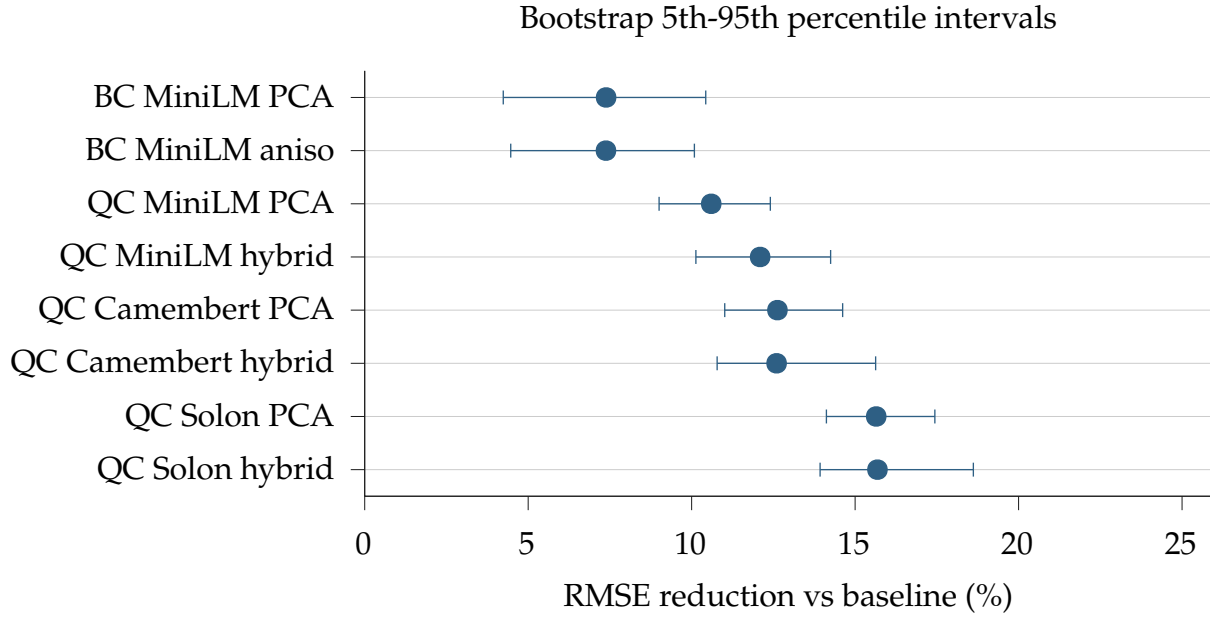


Figure 7: Bootstrap 5th to 95th percentile intervals on headline reductions. Québec uses $m = 5,000$ subsample with replacement; BC uses $m = 400$ without replacement (BC’s smaller auction pool inflates with-replacement estimates through duplication). PCA and hybrid intervals overlap at every panel on Québec. On BC, PCA and anisotropic bootstrap intervals cover the same region.

which Table 13 quantifies. A GPV variant that recovered markups without estimating the full joint density would decide the choice in the hybrid’s favour on Québec. Figure 8 lays the logic out.

The density diagnostic runs PCA, JL, and the anisotropic variant at $k = 10$ rather than the regression-stage $k = 50$, at bandwidth multipliers $(m, m_b) = (2.5, 1.5)$ in the notation of Section 8; the hybrid keeps its regression configuration $p = q = 50$. Because that leaves the hybrid with a hundred compressed coordinates against the others’ ten, Figure 9 adds a dimension-matched hybrid at $p = q = 5$ and sweeps m from 1 to 5, so neither the dimension gap nor the bandwidth pair can account for the ranking on its own. The joint density $\hat{g}(b \mid s, z)$ requires an integration over the projected-text coordinate that becomes infeasible at $k = 50$ under Silverman bandwidths: the effective sample scales as $N^{4/(d_s+k+5)}$, which at $d_s = 43$ and $k = 50$ leaves fewer than one effective observation per query point on the Québec panel. At $k = 10$ the effective sample is workable while the JL rate of Theorem 2 still applies to the projected estimator (the rate is $L^{-R/(2R+d_s+k+3)}$, so the $k = 50 \rightarrow k = 10$ substitution only changes the polynomial rate constant).

Held-out log-likelihood of the bid density across five auction-level folds is -1.582 for PCA at $k = 10$, -1.590 for ensembled JL at the same k , -1.600 for the anisotropic

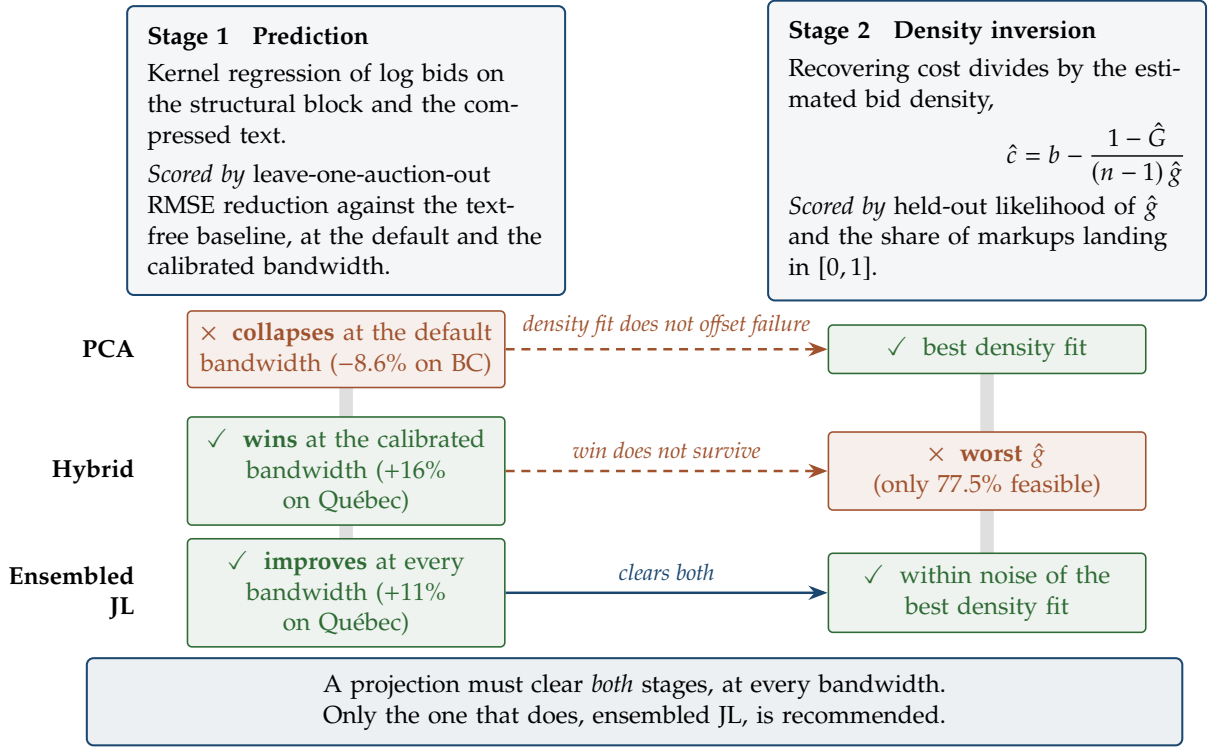


Figure 8: The two stages a projection must clear, and why winning one race does not settle the choice. Stage 1 scores how well the compressed text predicts bids, at both the default and the calibrated bandwidth; Stage 2 scores the estimated bid density that the GPV inversion divides by. PCA posts the best density fit but collapses at the default prediction bandwidth; the hybrid wins Stage 1 on Québec at the calibrated bandwidth yet posts the worst density fit and the lowest feasibility at Stage 2. Ensembled JL is the one projection that clears both stages, and the recommendation of Section 8’s rule rests on that.

variant, and -1.619 for the hybrid at $p = q = 50$. Feasibility rates (fraction of bids with recovered markup $\hat{m}$ in the economically-meaningful range $[0, 1]$) are 83.7%, 83.5%, 83.4%, and 77.5%. Cost-recovery monotonicity (fraction of auctions where recovered $\log \hat{c}$ is weakly increasing in bid rank, a first-order rationality check the LOO estimator is not guaranteed to satisfy at every kernel-evaluation point) is 97.9%, 97.8%, 97.8%, and 89.2%.

On every metric but the structural R^2 , where its regression advantage carries over, the hybrid does worst by a clear margin, and PCA and the two JL variants land inside 0.02 log-likelihood, half a percentage point of feasibility, and 0.1 percentage points of monotonicity of one another. The GPV inversion output on the Québec sample is therefore approximately invariant to projection choice among the three non-hybrid methods, and the hybrid is the only projection whose regression lead does not translate into a better density fit.

Signal concentration at the density stage is the likely mechanism.⁷

Table 11 reports the full density diagnostic for both datasets across all four methods.

The hybrid’s regression advantage does not survive the density stage. PCA’s held-out-likelihood lead over the dimension-matched hybrid is 0.013 to 0.027 at $h_x = 1$ depending on the bid-coordinate bandwidth, about 0.009 at the calibration $h_x = 2.5$, and under 0.004 by $h_x = 5$ (Figure 9). The lead never changes sign, so the hybrid’s density deficit is bandwidth-robust; its size, largest exactly where the kernel is narrowest, is not.

9.6 Recovered markups

The markup exercise runs the inversion end to end: estimate $\hat{g}$ and $\hat{G}$ on the structural block plus the $k = 10$ compressed text (the hybrid again at $p = q = 50$), at the same $(m, m_b) = (2.5, 1.5)$ bandwidth pair as the density diagnostic, evaluate both at each observed bid excluding the bid’s own auction, and invert bid by bid. Estimation is on the

⁷Dimension alone does not explain the deficit: the dimension-matched hybrid at $p = q = 5$, ten coordinates against ensembled JL’s ten, still trails PCA at every bandwidth pair in Figure 9. JL distributes cost-signal information across all its coordinates; the hybrid concentrates it on the head and leaves the tail statistically noisy. Kernel density estimation integrates over all feature dimensions, so noise in the tail hurts the density fit more than signal in the head helps it. The Monte Carlo of Section 7 does not reproduce the reversal because its factor structure delivers the same signal to the conditional mean and the joint density. A companion simulation separates the two channels: one direction of the covariate drives the mean of the outcome, an orthogonal direction drives only its dispersion, and a parameter moves the share of covariate variance each direction carries. The simulation confirms that PCA’s top component locks onto whichever direction carries more variance, exactly as the variance-ranking logic predicts. It does not, however, reproduce the hybrid’s density failure: the hybrid’s held-out likelihood stays flat across regimes at the simulation’s modest scale (3,000 observations, 100 ambient dimensions). The failure observed on Québec may need the full scale of a real text embedding to appear, or the mechanism proposed here may not be the operative one. The empirical result stands either way.

Table 11: Density diagnostics per method per dataset, at bandwidth multipliers $(m, m_b) = (2.5, 1.5)$. Held-out log-likelihood (LL) is mean across five auction-level cross-validation folds. Feasibility is the fraction of bids with recovered markup rate in $[0, 1]$. R^2 is from regressing log recovered cost on the structural block. Monotonicity is the fraction of auctions with $n_\ell \geq 3$ where recovered log-cost increases weakly in bid rank. JL, anisotropic, and hybrid rows are single random draws; redrawing the projections moves their BC log-likelihoods by roughly 0.01, and the ensembled figures quoted in the text differ by a similar margin. The QC rows are stable across redraws.

Dataset	Method	LL	Feasible	R^2	Monotone
BC	PCA $k = 10$	-1.546	0.813	0.551	1.000
BC	JL $k = 10$, single	-1.563	0.808	0.571	0.998
BC	Anisotropic $k = 10$, single	-1.563	0.812	0.550	0.998
BC	Hybrid $p = 50, q = 50$	-1.548	0.842	0.601	0.991
QC	PCA $k = 10$	-1.582	0.837	0.114	0.979
QC	JL $k = 10$, single	-1.600	0.834	0.114	0.978
QC	Anisotropic $k = 10$, single	-1.600	0.834	0.115	0.978
QC	Hybrid $p = 50, q = 50$	-1.619	0.775	0.120	0.892

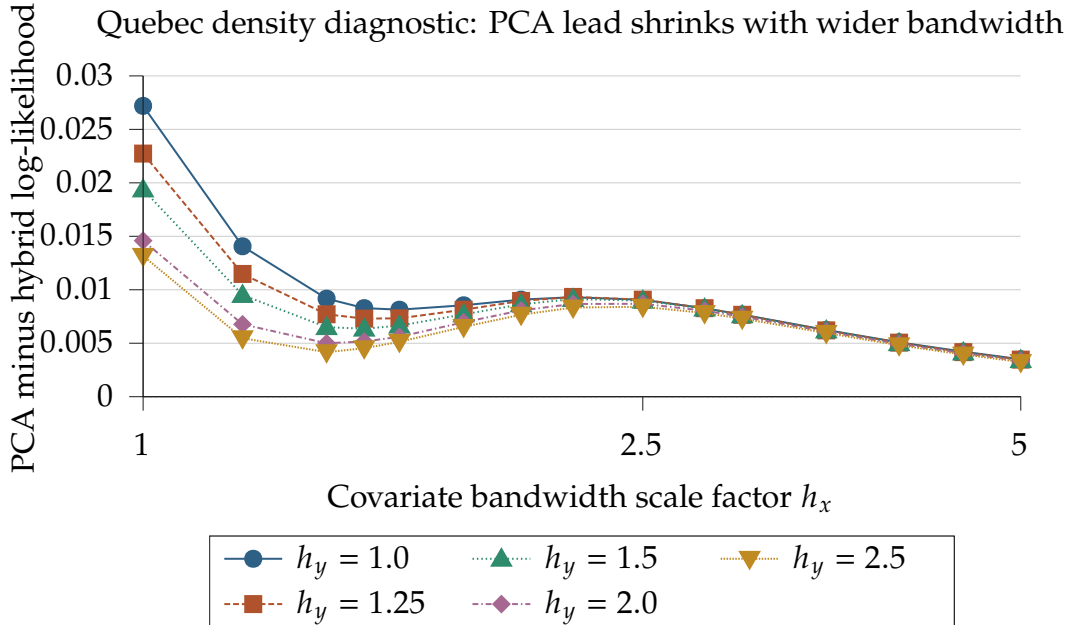


Figure 9: Québec density diagnostic: PCA $k = 10$ minus hybrid $p = 5, q = 5$ held-out log-likelihood as a function of the covariate-bandwidth scale factor h_x , one curve per bid-coordinate bandwidth h_y . PCA's lead is largest at the narrowest bandwidth and falls by an order of magnitude by $h_x = 5$, staying positive throughout.

log-bid scale, and there the inversion delivers the markup rate directly. Dividing equation (1) by the bid gives $\hat{m} = (b - \hat{c})/b = (1 - \hat{G})/((n - 1) \hat{g}_b)$, and the change of variables to log bids absorbs the factor of b into the density, leaving $\hat{m} = (1 - \hat{G})/((n - 1) \hat{g})$ with $\hat{g}$ now the estimated density of log bids; recovered cost is $\hat{c} = b(1 - \hat{m})$. Table 13 shows recovered median markup by number of bidders, over feasible bids ($\hat{m} \in [0, 1]$). The median falls steeply in n , from 59% at $n = 3$ to 44% at $n = 4$ and 26% at $n \geq 5$, the direction the $1/(n - 1)$ term predicts: fewer rivals mean weaker discipline on the margin, and the estimator amplifies bid dispersion at low n . The full-sample median is 33.0–33.1% on PCA, ensembled JL, and the anisotropic variant. These levels sit above the asphalt benchmark. Clark et al. (2018) recover asphalt markups that fall from about eleven dollars per metric ton with two bidders to under a dollar with seven; against their per-ton bids of roughly seventy dollars, that is at most fifteen per cent of the bid, and only a few per cent once several bidders compete, so the recovered SEAO rates run well above it. Composition explains part of the gap: their market is a single commodity trade with per-ton pricing, while the SEAO panel is dominated by low- n specialty-trade contracts whose fixed-price terms carry quotation cushions for change-orders. Under-conditioning explains the rest. A model-free check puts the median money left on the table (runner-up bid over winning bid, minus one) at 6.6% in the $n \geq 5$ regime, so directly observable margins are benchmark-sized; the GPV level is higher because the inversion reads the dispersion of the conditional bid density as margin, and the residual cross-auction dispersion the covariates fail to absorb is large even after the text enters. A text-free run of the same inversion, structural block only, returns 33.3% full-sample and 26.1% at $n \geq 5$, within 0.3 points of the text-conditioned numbers, so the level is a property of the inversion and the observable heterogeneity, not of the text or the projection. I read it with caution accordingly. The projection comparison, which is the paper’s object, does not depend on it: what matters below is that three of the four methods agree on the distribution and one does not.

The hybrid returns a lower full-sample median (24.1%) and lower feasibility (77.6% against 83.3–83.7% on the other three), consistent with the density fit it delivered above. The 6-percentage-point feasibility gap is a real economic loss in this class of application: the analyst loses one bid in seventeen to an out-of-range recovered markup relative to ensembled JL.

Table 12 decomposes feasibility by projection method $\times n_\ell$ bucket. The hybrid’s deficit against the three-method minimum is +1.1 percentage points at $n_\ell = 3$, jumps to +6.6 at $n_\ell = 4$, and stays in the +6–+8 range at every larger bucket up through $n_\ell \geq 11$; the naive prediction from $1/(n - 1)$ scaling that the loss would concentrate at $n_\ell = 3$ does not survive the decomposition. At $n_\ell = 3$ all four methods post feasibility near 61%, so the

low- n noise floor on the density estimator dominates the method-level differences. As n grows the non-hybrid methods reach feasibility above 95% that the hybrid cannot match, and the deficit widens with better sample support rather than shrinking. The largest absolute count of infeasible bids attributable to the hybrid (1,701) lands at $n_\ell = 6-7$, the modal bucket, not at $n_\ell = 3$. The pattern is consistent with the density-mis-estimation mechanism of Section 9.5: the hybrid’s poorer joint density fit surfaces as infeasibility bias uniformly across the n_ℓ distribution, not as a low- n tail concentration.

Table 12: Québec SEAO recovered-markup feasibility (fraction of bids with $\hat{m} \in [0, 1]$) by projection method and n_ℓ bucket. All 110,646 bids covered; buckets partition the n_ℓ -support. Hybrid deficit, in percentage points (pp), is against the minimum feasibility of the three non-hybrid methods within each bucket.

Method	$n = 3$	$n = 4$	$n = 5$	$n = 6-7$	$n = 8-10$	$n \geq 11$
Bids in bucket	20,118	21,968	18,975	27,524	16,753	5,308
PCA $k = 10$	62.8	78.8	86.4	91.5	95.2	96.9
Ensembled JL $M = 10$	61.5	78.5	86.1	91.3	95.2	96.5
Anisotropic ensemble	61.9	78.7	86.2	91.3	95.2	96.9
Hybrid $p = q = 50, M = 10$	60.4	71.9	79.4	85.1	87.6	90.5
Hybrid deficit (pp)	+1.1	+6.6	+6.8	+6.2	+7.6	+6.1

Figure 10 shows the recovered-markup densities on both datasets.

Table 13: Recovered median markup rate on Québec SEAO by number of bidders, feasible bids only, PCA projection (ensembled JL and the anisotropic variant agree within 0.4 percentage points in every cell). Contract shares are the fraction of sample auctions in each regime. The $1/(n-1)$ term in the GPV inversion drives the decline in n ; the Clark et al. (2018) Montreal-asphalt markups run from about \$11 per metric ton at two bidders down to under \$1 at seven.

Bidder regime	Share of contracts	Median $\hat{m}$
$n = 3$	30%	59%
$n = 4$	24%	44%
$n \geq 5$	46%	26%
Full sample	100%	33%

PCA, ensembled JL, and anisotropic JL agree within two-tenths of a percentage point on the full-sample median markup. The hybrid disagrees, and the disagreement lines up with the density-fit and feasibility evidence. Without ground-truth cost, the hybrid’s regression advantage does not translate into a different recovered cost distribution.

Table 14 decomposes the hybrid feasibility deficit across UNSPSC segments and years.

Across the three largest UNSPSC segments, the hybrid’s shortfall against the three-way minimum ranges from +2.2 percentage points on segment 71 (specialty extraction, $n = 1,118$) to +8.4 on segment 30 (structural components and basic-shape materials procurement, $n = 13,882$), with the dominant segment 72 (general building construction and facility maintenance, $n = 94,304$, 85% of QC bids) at +5.4. The segment where the hybrid loses most is the one whose text descriptions centre on material-specification language, which the top- p Solon eigendirections align to (Section 9.3); this is the regime where the hybrid concentrates signal into the head and leaves the tail statistically noisy at the density stage. Across the main-sample years 2018–2024 (107,000 bids, 97% of the panel), the deficit varies between +4.2 and +7.9 percentage points with no monotone trend, so the feasibility loss is not a rare-vintage artefact.

Table 14: Québec SEAO hybrid feasibility deficit (percentage points below the minimum of the three non-hybrid methods) by UNSPSC segment and by year. Top-three UNSPSC segments cover 99% of the panel; the main-sample years 2018–2024 cover 97%. Full-sample deficit is +5.7 percentage points.

Slice	Deficit (pp)	Bids
<i>By UNSPSC segment</i>		
Seg. 72 (building construction / maintenance)	+5.4	94,304
Seg. 30 (structural components / materials)	+8.4	13,882
Seg. 71 (specialty extraction)	+2.2	1,118
<i>By year, main sample</i>		
2018	+5.5	21,608
2019	+5.8	16,868
2020	+7.9	4,647
2021	+5.8	20,300
2022	+4.2	15,693
2023	+5.9	20,001
2024	+4.8	7,793

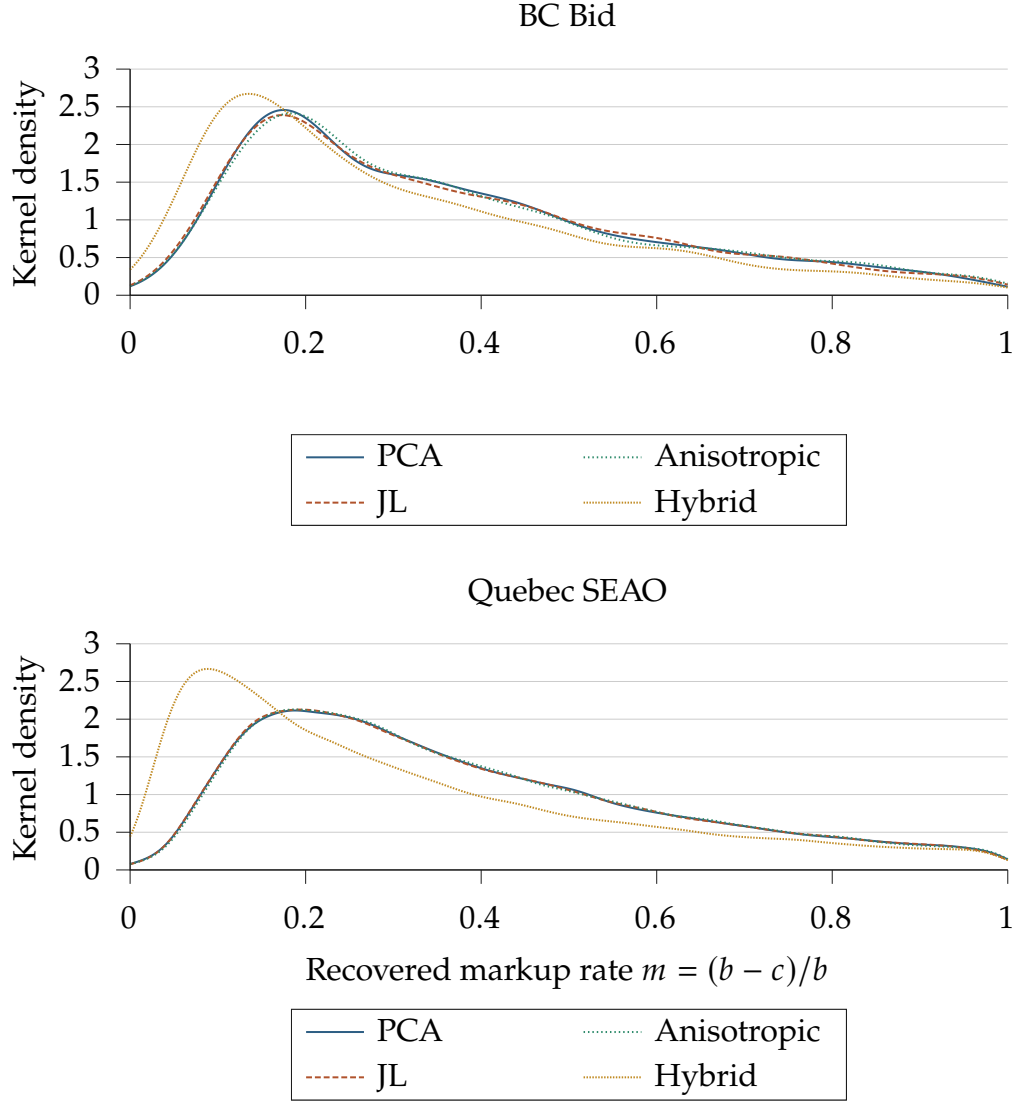


Figure 10: Kernel density of recovered markup rate $\hat{m} = (b - \hat{c})/b$ across bids, by projection method, per dataset. The three non-hybrid methods overlap almost exactly on both BC (top) and Québec (bottom). The hybrid density is shifted towards lower markups by about nine percentage points on Québec and six on BC. On Québec the shift arrives together with the feasibility gap and the log-likelihood deficit, which marks it as density mis-estimation rather than a different recovered cost function; on BC the hybrid's density fit and feasibility are both fine, so the smaller BC shift is not by itself evidence of failure.

10 Conclusion

The text that accompanies public tenders can serve as a covariate in GPV cost recovery, provided the embedding is compressed before it reaches the kernel. For the compression I recommend the JL ensemble, the one method that meets the rule of Section 8 on both datasets: $M = 10$ draws, $k = 50$ for the regression stage and $k = 10$ where a joint density is the target, run at the default per-coordinate Silverman bandwidth. On BC and Québec this choice beats PCA by ten percentage points of RMSE reduction at that default, requires no fitting, and matches PCA on recovered markups without the hybrid’s density-stage failure. A widened bandwidth can promote the hybrid to first place on the regression stage where the encoder’s leading eigendirections happen to align with observable cost drivers (Section 9.3). But the hybrid then fails the density diagnostic on Québec, and its win depends on a bandwidth choice that is itself a judgement call (Section 8). Ensembled JL depends on neither condition.

Lemma 1 rests only on product-kernel structure and carries over to any product-kernel estimator, such as quantile regression or conditional-independence testing on high-dimensional covariates. GPV inversion is the sharp case: the formula divides two independently estimated density objects, so the recovered cost is unusually sensitive to their shared compression. Proposition 3’s AMISE decomposition motivates the projection-specific bandwidth calibration adopted here, and extending that recipe to the other product-kernel estimators just named is a natural next step.

The empirical base is two datasets, and neither the Monte Carlo nor the empirics deliver a clean mechanism test on their own. Applying the recommendation to a new dataset therefore requires two diagnostics before any commitment to the hybrid: the principal-components-versus-metadata comparison of Section 9.3 and the density diagnostic of Section 9.5. Replication on a procurement platform in a third language and on a private-sector business-to-business tender platform would separate the boundary conditions of the mechanism from artefacts of the two datasets tested here. The natural companion estimator is the unobserved-heterogeneity deconvolution of Krasnokutskaya (2011): Appendix A.6 shows it repairs the markup level that under-conditioning inflates, while inheriting the same compression question this paper answers, since the covariates it conditions on face the same kernel and the same dimensional budget.

References

- Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4), 671–687.
- Ailon, N., & Chazelle, B. (2009). The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1), 302–322.
- Armona, L., Lewis, G., & Zervas, G. (2021). *Learning product characteristics and consumer preferences from search data* (Working Paper No. 3858377). Rochester, NY: SSRN.
- Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15, 659–688.
- Athey, S., & Haile, P. A. (2002). Identification of standard auction models. *Econometrica*, 70(6), 2107–2140.
- Athey, S., & Haile, P. A. (2007). Nonparametric approaches to auctions. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6A, pp. 3847–3965). Elsevier.
- Athey, S., Levin, J., & Seira, E. (2011). Comparing open and sealed bid auctions: Evidence from timber auctions. *Quarterly Journal of Economics*, 126(1), 207–257.
- Bajari, P., & Ye, L. (2003). Deciding between competition and collusion. *Review of Economics and Statistics*, 85(4), 971–989.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.
- Cannings, T. I., & Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society, Series B*, 79(4), 959–1035.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1), C1–C68.
- Chiong, K. X., & Shum, M. (2019). Random projection estimation of discrete-choice models with large choice sets. *Management Science*, 65(1), 256–271.
- Clark, R., Gauthier, J.-F., Coviello, D., & Shneyerov, A. (2018). Bid rigging and entry deterrence in public procurement: Evidence from an investigation into collusion and corruption in Québec. *Journal of Law, Economics, and Organization*, 34(3), 301–363.
- Compiani, G., Morozov, I., & Seiler, S. (2026). Demand estimation with text and image data. *RAND Journal of Economics*. doi: 10.1111/1756-2171.70052

- Dasgupta, S., & Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1), 60–65.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 4171–4186).
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall.
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574.
- Guerre, E., Perrigne, I., & Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68(3), 525–574.
- Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2), 217–288.
- Hendricks, K., Pinkse, J., & Porter, R. H. (2003). Empirical implications of equilibrium bidding in first-price, symmetric, common value auctions. *Review of Economic Studies*, 70(1), 115–145.
- Hickman, B. R., & Hubbard, T. P. (2015). Replacing sample trimming with boundary correction in nonparametric estimation of first-price auctions. *Journal of Applied Econometrics*, 30(5), 739–762.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441 and 498–520.
- Hubbard, T. P., Li, T., & Paarsch, H. J. (2012). Semiparametric estimation in models of first-price, sealed-bid auctions with affiliation. *Journal of Econometrics*, 168(1), 4–16.
- Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability* (Vol. 26, pp. 189–206). American Mathematical Society.
- Kotlarski, I. (1967). On characterizing the gamma and the normal distribution. *Pacific Journal of Mathematics*, 20(1), 69–76.
- Krasnokutskaya, E. (2011). Identification and estimation of auction models with unobserved heterogeneity. *Review of Economic Studies*, 78(1), 293–327.
- Li, P., Hastie, T. J., & Church, K. W. (2006). Very sparse random projections. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 287–296).

- Li, T., Perrigne, I., & Vuong, Q. (2002). Structural estimation of the affiliated private value auction model. *RAND Journal of Economics*, 33(2), 171–193.
- MacKinnon, J. G., & Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *Econometrics Journal*, 21(2), 114–135.
- Magnolfi, L., McClure, J., & Sorensen, A. (2025). Triplet embeddings for demand estimation. *American Economic Journal: Microeconomics*, 17(1), 282–307.
- Marmer, V., & Shneyerov, A. (2012). Quantile-based nonparametric inference for first-price auctions. *Journal of Econometrics*, 167(2), 345–357.
- Milgrom, P. R., & Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica*, 50(5), 1089–1122.
- Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9(1), 141–142.
- OECD. (2023). *Government at a glance 2023* (Tech. Rep.). Paris: Organisation for Economic Co-operation and Development. doi: 10.1787/3d5c5d31-en
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11), 559–572.
- Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in neural information processing systems*.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing* (pp. 3982–3992).
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall.
- Somaini, P. (2020). Identification in auction models with interdependent costs. *Journal of Political Economy*, 128(10), 3820–3871.
- Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. Chapman and Hall.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4), 359–372.

Appendix

A.1 Proofs

Throughout the appendix the standard GPV regularity conditions of Guerre et al. (2000) are assumed: the conditional bid density $g(b | s, t)$ is R times continuously differentiable on the interior of its support, bounded away from zero on the interior, and has bounded R -th derivatives; the covariate marginal of t has a density f_t bounded away from zero and infinity on the compact support; the joint distribution of (b, s, t) has finite moments up to a sufficient order to invoke the standard multivariate kernel-density mean-integrated-squared-error (MISE) argument of Wand and Jones (1995), Chapter 4. For results on local polynomial extensions of the kernel regression estimator relevant to the mean bid function, see Fan and Gijbels (1996).

A.1.1 Proof of Theorem 2

Fix a query point (b, s, t) and a projected point $z = At$. Write $g^A(b | s, z) = \mathbb{E}[g(b | s, T) | AT = z]$, the population conditional density of the bid given the structural block and the JL-projected text. The estimator $\hat{g}^A$ is the product-kernel density estimator with per-coordinate Silverman bandwidths applied to the $(d_s + k)$ -dimensional feature (s, z) . Write

$$\hat{g}^A(b | s, z) - g(b | s, t) = [\hat{g}^A - g^A] + [g^A - g].$$

The first term is the standard nonparametric estimation error at feature dimension $d_s + k$. The one-step kernel-density error follows from the asymptotic MISE approximations of Wand and Jones (1995), Section 4.3, applied to the $(1 + d_s + k)$ -dimensional joint density of (b, s, z) , and is $O_P(L^{-R/(2R+d_s+k+1)})$. The theorem states the slower two-step rate of Guerre et al. (2000), $O_P(L^{-R/(2R+d_s+k+3)})$, the uniform rate for the value density recovered from the estimated $\hat{g}$ and $\hat{G}$ at covariate dimension $d_s + k$. The two-step rate is the rate at which the chapter's downstream objects converge, and the one-step error is faster, so the stated bound holds. For $R > 2$ both rates require a kernel of order R with the matching bandwidth exponent; the Gaussian-kernel, Silverman-bandwidth implementation attains them for $R \leq 2$, and the chapter's numerical evaluations set $R = 1$.

The second term is $\text{bias}_{\text{JL}}(b, s, t; A) = |g^A(b | s, At) - g(b | s, t)|$ by definition, and three properties are asserted of it. It is fixed in L : g^A and g are population objects, so no sample quantity enters. It is free of the bandwidth and the kernel, since neither appears in either population object. It is finite: g^A is an average of values of g , and g is bounded under the regularity conditions. Combining the two terms gives the statement.

No bound on the magnitude of bias_{JL} is claimed. The conditioning set $\{t' : At' = z\}$ is the affine subspace $t + \text{null}(A)$, and the size of $g^A - g$ is governed by how much g varies along $\text{null}(A)$ under the conditional law of T , a property of the embedding distribution rather than of the projection alone. In practice the constant is small inasmuch as sentence embeddings concentrate near a set of low intrinsic dimension, whence the conditional law of T given $AT = At$ concentrates near t ; formalising this would require an explicit concentration assumption, and the paper instead measures the constant. Proposition 3 identifies bias_{JL} as the ensemble's bias floor, and the ensemble-size sweeps estimate it empirically.

A.1.2 Proof of Proposition 3

By linearity of expectation,

$$\mathbb{E}[\hat{g}_M^{\text{ens}}] = \frac{1}{M} \sum_{m=1}^M \mathbb{E}[\hat{g}^{(m)}] = \mathbb{E}_A[\hat{g}^{(m)}],$$

since the $\hat{g}^{(m)}$ have identical marginal distributions. This gives the bias identity.

For the variance, the projections $A^{(1)}, \dots, A^{(M)}$ are iid, so the resulting estimators $\hat{g}^{(1)}, \dots, \hat{g}^{(M)}$ are iid conditional on the data. Then

$$\text{Var}(\hat{g}_M^{\text{ens}}) = \text{Var}\left(\frac{1}{M} \sum_m \hat{g}^{(m)}\right) = \frac{1}{M^2} \sum_m \text{Var}(\hat{g}^{(m)}) = \frac{1}{M} \text{Var}_A(\hat{g}^{(m)}).$$

The variance-optimal M minimises squared bias plus variance. Squared bias is constant in M ; variance decays as $1/M$. If the practitioner tolerates draw-attributable variance up to $2\delta \cdot \text{bias}_{\text{JL}}^2$, then

$$\frac{1}{M_{\text{stat}}^*} \sigma_{\text{draw}}^2 = 2\delta \cdot \text{bias}_{\text{JL}}^2,$$

which rearranges to $M_{\text{stat}}^* = \sigma_{\text{draw}}^2 / (2\delta \cdot \text{bias}_{\text{JL}}^2)$.

A.1.3 Proof of Lemma 1

Fix two observations $z, z' \in \mathbb{R}^k$ and their rescalings $\tilde{z}_j = c_j z_j$, $\tilde{z}'_j = c_j z'_j$. Silverman's bandwidth on the j -th coordinate is

$$h_j = c_S \hat{\sigma}_{z_j} N^{-1/(k+4)}, \quad \tilde{h}_j = c_S \hat{\sigma}_{\tilde{z}_j} N^{-1/(k+4)} = c_j h_j,$$

because $\hat{\sigma}_{\tilde{z}_j} = c_j \hat{\sigma}_{z_j}$. The Gaussian kernel weight on the pair is

$$\prod_{j=1}^k \exp\left(-\frac{1}{2}(z_j - z'_j)^2 / h_j^2\right) = \prod_{j=1}^k \exp\left(-\frac{1}{2}c_j^2(z_j - z'_j)^2 / (c_j h_j)^2\right) = \prod_{j=1}^k \exp\left(-\frac{1}{2}(\tilde{z}_j - \tilde{z}'_j)^2 / \tilde{h}_j^2\right),$$

since the numerator scales by c_j^2 and the denominator by c_j^2 identically. The product kernel is therefore pointwise identical on the rescaled coordinates up to its normalising constant $\prod_j \tilde{h}_j^{-1} = (\prod_j c_j^{-1}) \prod_j h_j^{-1}$, which is common to every observation. Both estimators are ratios of sums that agree up to this common factor, which cancels in the ratio, so $\hat{g}(b | z) = \hat{g}(b | \tilde{z})$ observation-by-observation.

The lemma has an immediate consequence for finite- N estimation: two compressions of t that differ only by per-coordinate scaling produce numerically identical estimators, not merely asymptotically-equivalent ones. This is why PCA's variance ranking does not survive under standard Silverman.

A.1.4 Proof of Theorem 4

Decompose $t = V_p^\top V_p t + (I - V_p^\top V_p)t = V_p^\top h + r$, where $h = V_p t \in \mathbb{R}^p$ is the head and $r \in \text{range}(V_p)^\perp$ the residual. The hybrid feature is $(s, h, \tilde{z})$ with $\tilde{z} = Ar$. The population conditional density in the hybrid feature space is $g^{\text{hyb}}(b | s, h, \tilde{z}) = \mathbb{E}[g(b | s, T) | V_p T = h, A(I - V_p^\top V_p)T = \tilde{z}]$.

Decompose the error as before:

$$\hat{g}^{\text{hyb}}(b | s, h, \tilde{z}) - g(b | s, t) = [\hat{g}^{\text{hyb}} - g^{\text{hyb}}] + [g^{\text{hyb}} - g].$$

The first term is the kernel-density estimation error at feature dimension $d_s + p + q$. Standard rates give $O_P(L^{-R/(2R+d_s+p+q+3)})$.

The second term is bias_{hyb} by definition; it is fixed in L and free of the bandwidth and kernel for the same reasons as in Appendix A.1.1. The alignment claim is direct. Conditioning on $V_p T = h$ fixes the head coordinates of T exactly at t 's head; only the residual is uncertain. If $g(b \mid s, t)$ depends on t only through $V_p t$, the conditioning determines every argument g responds to, so $g^{\text{hyb}}(b \mid s, h, \tilde{z}) = g(b \mid s, t)$ pointwise and $\text{bias}_{\text{hyb}} = 0$. This is the exact-recovery limit of the head. When the signal has both head and tail components, the head contribution is captured exactly, and bias_{hyb} is driven by the tail alone, through the compression of the $(d - p)$ -dimensional residual rather than of the full d .

The bias decomposition of Section 6 follows from adding and subtracting $g^{V_p}(b \mid s, h) = \mathbb{E}[g(b \mid s, T) \mid V_p T = h]$ inside the total-bias expression.

A.2 Anisotropic JL: construction, metric, and rate

Standard JL preserves Euclidean distances on $\mathbb{R}^d$. The product Gaussian kernel with Silverman bandwidths $h_j \propto \sigma_{t_j}$ evaluates

$$K(t, t') = \exp\left(-\frac{1}{2} \sum_{j=1}^d (t_j - t'_j)^2 / \sigma_{t_j}^2\right), \quad (\text{A.1})$$

a Gaussian kernel in the diagonal Mahalanobis metric $\|t - t'\|_{\Sigma_{\text{diag}}^{-1}}^2 = \sum_j (t_j - t'_j)^2 / \sigma_{t_j}^2$, so the projection metric and the kernel metric diverge whenever the coordinate variances of t are unequal. The anisotropic construction matches the two. Sample $A \in \mathbb{R}^{k \times d}$ with $A_{ij} \sim \mathcal{N}(0, w_j/k)$ for positive per-coordinate weights $w_1, \dots, w_d$; setting $\tilde{A} = AW^{-1/2}$ with $W = \text{diag}(w_j)$ makes $\tilde{A}$ a standard JL matrix on the rescaled input $W^{1/2}t$, so anisotropic JL amounts to standard JL after a coordinate rescaling.

Lemma 1 (Weighted distance preservation). *For any $\varepsilon \in (0, 1/2)$ and any finite set of N points $t_1, \dots, t_N \in \mathbb{R}^d$, with probability at least $1 - 2N^2 \exp(-c\varepsilon^2 k)$,*

$$(1 - \varepsilon) \|t_i - t_j\|_W^2 \leq \|A(t_i - t_j)\|^2 \leq (1 + \varepsilon) \|t_i - t_j\|_W^2 \quad \forall i, j,$$

where $\|u\|_W^2 = \sum_j w_j u_j^2$.

Define $\tilde{t}_i = W^{1/2}t_i$. Then $At_i = \tilde{A}\tilde{t}_i$ with $\tilde{A}_{ij} \sim \mathcal{N}(0, 1/k)$ iid, $\|A(t_i - t_j)\|^2 = \|\tilde{A}(\tilde{t}_i - \tilde{t}_j)\|^2$, and $\|\tilde{t}_i - \tilde{t}_j\|^2 = \|t_i - t_j\|_W^2$. The standard JL lemma applied to the $\tilde{t}$ -points and matrix $\tilde{A}$ gives the stated bound with probability at least $1 - 2N^2 \exp(-c\varepsilon^2 k)$.

Setting $w_j = 1/\sigma_{t_j}^2$ makes $\|\cdot\|_W^2$ the diagonal Mahalanobis metric (A.1) the kernel uses, and the anisotropic projection preserves the exact metric determining the kernel weights. The empirical implementation uses $\sqrt{w_j} \propto 1/\sigma_{t_j}$ from the training-sample standard deviation; the construction requires d scalars but no fitted basis.

Theorem 2 (Anisotropic JL rate). *Under the assumptions of Theorem 2, the anisotropic JL estimator $\hat{g}^{A_W}(b \mid s, z)$ satisfies*

$$|\hat{g}^{A_W}(b \mid s, z) - g(b \mid s, t)| = O_P\left(L^{-R/(2R+d_s+k+3)}\right) + \text{bias}_{\text{aniso}}(b, s, t; A_W),$$

where $\text{bias}_{\text{aniso}}(b, s, t; A_W) = |g^{A_W}(b \mid s, A_W t) - g(b \mid s, t)|$. The first term is the Nadaraya–Watson rate at feature dimension $d_s + k$, unchanged from Theorem 2. The second is the anisotropic analogue of bias_{JL} , defined with g^{A_W} in place of g^A ; the weighted metric changes which directions the projection preserves, and hence the constant, but not its character.

Adapt Appendix A.1.1’s argument with A replaced by $A_W = AW^{-1/2}$ acting on t . The first-term rate is unchanged because it depends only on the projected feature dimension. The second term is a defined constant exactly as before; the conditioning subspace is now $t + \text{null}(A_W)$, so the constant is governed by the variation of g along a different null space, the one selected by the weighted draw. Ensembling works identically to Proposition 3, with bias $\mathbb{E}_A[\hat{g}^{A_W}] - g$, variance $\text{Var}_{A_W}(\hat{g}^{A_W})/M$, and the same functional form for M^* . The empirical edge over ensembled JL is between 0.02 and 0.12 percentage points on the BC and Québec panels, inside every wild-cluster standard-error band; the anisotropic variant is retained in the empirical tables (Sections 8 and 9) but does not enter the recommendation, because an edge that never clears the noise band does not justify the d estimated coordinate scales the construction adds over plain JL.

A.3 Projection-aware bandwidth

Lemma 1 normalises the estimator across projections whose compressed coordinates share the same variance pattern up to per-axis rescaling. It does not pin down the scalar multiplier m in $h_j = m \cdot 1.06 \sigma_j n^{-1/(d+4)}$. Silverman’s canonical constant 1.06, that is $m = 1$, is the value that minimises the asymptotic mean integrated squared error (AMISE), the standard bandwidth-selection criterion, when the target density is Gaussian. The projected joint density $g_\Pi(b, z)$ that a given projection Π induces need not be Gaussian, and the value of m that minimises AMISE for it depends on the projected density’s curvature.

Proposition 3 (Projection-dependent AMISE constant). *Consider two linear projections $\Pi, \Pi': \mathbb{R}^d \rightarrow \mathbb{R}^k$ applied to a common source distribution, inducing projected joint densities $g_\Pi, g_{\Pi'}$ on (b, z) . Under the standard AMISE bandwidth derivation for a product-Gaussian kernel estimator with per-coordinate bandwidths $h_j = m \cdot 1.06 \sigma_j n^{-1/(d+4)}$, the AMISE-optimal multipliers satisfy $m_\Pi^* = m_{\Pi'}^*$ if and only if*

$$\int_z [\Delta_z g_\Pi(b | z)]^2 f_\Pi(z) dz = \int_z [\Delta_z g_{\Pi'}(b | z)]^2 f_{\Pi'}(z) dz,$$

where $\Delta_z = \sum_j \partial_{z_j}^2$ is the Laplacian in the projected coordinates.

The proposition follows from the standard AMISE bandwidth calculation (Silverman, 1986; Wand & Jones, 1995): the AMISE-optimal overall scale $1.06 m$ satisfies $(1.06 m)^{d+4} \propto R(K)/V(g)$ where $R(K)$ is a kernel-only constant and $V(g)$ is the integrated squared Laplacian of the target density. Silverman’s plug-in $m = 1$ arises by substituting a Gaussian target. Different projections induce projected densities with different Laplacians even when the marginal coordinate variances are matched, so the plug-in constant is not projection-invariant even where Lemma 1’s coordinate-rescaling invariance holds. The two invariances address different objects: Lemma 1 normalises across variance patterns at any fixed m ; Proposition 3 says the optimal m still depends on the projection map through curvature.

The empirical implication is that the calibrated bandwidth should be picked per projection method. Section 8 implements this: a fine held-out log-likelihood grid over m selects the optimum separately for each of PCA, ensembled JL, anisotropic ensemble, and hybrid on each dataset (Table 3).

A.4 Wall-clock runtime

Compute is not a binding constraint at these dataset scales. The slower projection (hybrid) fits the full Québec embedding in three seconds, and the downstream $O(N^2 k d)$ kernel-summation cost dwarfs the projection step at either PCA or JL. Table A.1 and Figure A.1 give the fit-and-transform wall-clock on Solon-large. JL runs the projection in fifteen milliseconds against PCA’s 1.2 seconds (88×). Anisotropic JL is 5× PCA. The hybrid is slower than PCA because it pays the PCA fit plus the JL projection. Randomised SVD (Halko, Martinsson, & Tropp, 2011) would close some of the JL-versus-PCA fit gap, but it would still be a fitted method; JL requires no fit at all and carries the distance guarantee of the Johnson–Lindenstrauss lemma besides.

Table A.1: Fit and transform wall-clock on the Solon-large embedding ($22,663 \times 1,024$). Summary over five runs; speed-up relative to PCA. The hybrid is timed at its regression configuration $p = q = 50$; the other methods at $k = 10$.

Method	Mean (s)	Median (s)	Min (s)	Speedup
PCA $k = 10$	1.185	1.220	0.980	1.0
Hybrid $p = 50, q = 50$	3.104	3.181	2.708	0.4
Anisotropic $k = 10$	0.261	0.261	0.261	4.7
JL $k = 10$	0.016	0.014	0.014	88.2

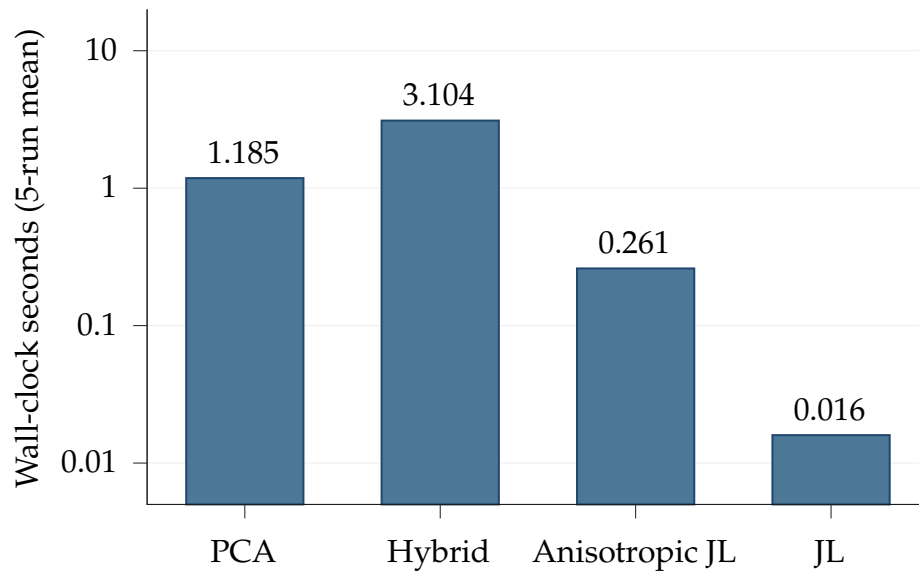


Figure A.1: Fit-and-transform wall-clock on the Solon-large embedding ($22,663 \times 1024$), five-run mean, log scale. JL is the fastest and hybrid the slowest.

A.5 MLM continued pretraining of the sentence encoders

I ran masked-language-model (MLM) continued pretraining on the three Québec encoders and on MiniLM for BC. The objective is the standard masked-language-modelling one at 15% mask probability with cross-entropy loss on masked tokens. Training used learning rate 5×10^{-5} at batch size 16 and sequence length 384, in half-precision (fp16) arithmetic, updating all encoder weights rather than low-rank adapters (LoRA). It ran one epoch on the 22,663-document Québec dataset and three epochs on the 904-document BC dataset. Total compute was 2158 seconds on Modal L40S GPUs at under one US dollar. Each fine-tuned encoder was swapped back into the sentence-transformer wrapper and re-encoded, and the ($k = 50, M = 10$) calibrated-bandwidth cell was rerun across the four projections.

Table A.2 has the Québec results. The direction of the MLM effect depends on how much prior French exposure the encoder had. On multilingual MiniLM, fine-tuning adds three to five percentage points to every projection’s reduction; on French-specialised Solon, it subtracts three-quarters to two-and-a-half. Camembert is an intermediate case and moves by less than a point in either direction. Raw Solon still dominates every tuned configuration on every projection, since the largest MLM lift (MiniLM under the hybrid) reaches only 13.88% against raw Solon’s 16.01% at the same cell.

On BC the fine-tuning outcome is worse. Continuing MLM pretraining of MiniLM on the 904-document BC dataset drives every projection below the raw baseline by between 4.8 percentage points (JL) and 14.3 (hybrid); the token-level MLM objective overwrites the sentence-embedding geometry when the dataset is too small to fine-tune without overfitting.

On both datasets the sign of the raw-versus-tuned difference is uniform across the four projections, so the JL-versus-PCA comparison in the main text is unaffected. The practical recommendation for a French dataset at Québec’s scale is to pick a French-specialised encoder over a multilingual generic one and leave it untuned. On a dataset the size of BC’s, MLM continued pretraining is not an option.

Table A.2: Québec SEAO: RMSE reduction (%) at calibrated bandwidth for each encoder before and after MLM continued pretraining.

Encoder	Variant	PCA $k = 50$	JL ens	Aniso ens	Hybrid ens
MiniLM (multilingual, 384d)	raw	6.45	6.11	6.02	8.82
	MLM-tuned	11.18	8.30	8.64	13.88
Camembert-large (French, 1024d)	raw	8.70	8.69	8.64	12.64
	MLM-tuned	9.84	9.08	9.10	12.16
Solon-large (French-specialised, 1024d)	raw	12.50	10.35	10.47	16.01
	MLM-tuned	11.01	9.59	9.72	13.48

A.6 Unobserved auction-level heterogeneity and the markup level

Section 9.6 attributes the level of the recovered markups to under-conditioning: cross-auction dispersion that the covariates fail to absorb reaches the inversion and enters the estimate as margin. That attribution is directly testable. A deconvolution estimator in the spirit of Krasnokutskaya (2011) removes the auction-level component of the residual, and the markup level that remains afterwards is informative about how much of the original level was attributable to auction-level heterogeneity.

Model and identification

Write the log bid as

$$\log b_{i\ell} = m(x_\ell) + u_\ell + e_{i\ell},$$

with $m(x_\ell)$ the covariate-explained part, u_ℓ a factor at the auction level, shared by every bidder in auction ℓ , visible to the bidders but not to the econometrician, and $e_{i\ell}$ an idiosyncratic term, mean zero and independent of u_ℓ . Heterogeneity of this multiplicative form travels through the equilibrium bid function without distortion, whereupon the markup *rate* turns on the idiosyncratic term alone. Every bid in the auction carries the same scale factor e^{u_ℓ} , so $(b-c)/b$ is free of it, and on the log-bid scale $\hat{m} = (1-F_e)/((n-1)f_e)$, where F_e and f_e denote the distribution and density of e . The naive inversion of Section 9.6 divides instead by the density of $u + e$; wherever $\text{Var}(u) \gg \text{Var}(e)$, the dispersion of u enters the estimate as margin.

The split is estimated three ways, under successively stronger identifying conditions. The least demanding is a moment calculation: two bids from the same auction satisfy $\text{Cov}(r_1, r_2) = \sigma_u^2$, so the pair covariance of homogenised residuals within auctions delivers the auction-level variance, and what is left over is idiosyncratic. Next comes the characteristic-function argument of Kotlarski (1967), whereby within-auction pairs identify the complete distributions of both u and e . The most heavily parameterised is a mixture fit. Demeaning within the auction eliminates u exactly; given the auction size, the demeaned residual is distributed as a known convolution of f_e with itself; a normal mixture for f_e , mean-zero imposed and components bounded away from degeneracy, is then matched to the empirical characteristic function of that demeaned residual. Conditioning separates the last two. The Kotlarski construction has the characteristic function of the *total* residual in its denominator, an object decaying on the scale of $\sigma_u + \sigma_e$, so a large auction-level variance share drives the denominator to zero at frequencies too low to identify more than the coarse shape of f_e . No trace of u enters the demeaned residual, and the mixture fit is thereby conditioned on the scale of σ_e alone, however large the share.

Implementation

The homogenisation step borrows the machinery of Section 9.5 unchanged: log bids are regressed by a cross-fitted kernel mean on the structural block together with the $k = 10$ ensembled-JL text, ten members with predictions averaged, at the calibrated bandwidth $m = 2.5$; the cross-fitting is partitioned by auction, so no bid is ever predicted using its own auction. With text the residual standard deviations come out at 1.240 on Québec and 1.159 on BC; without, 1.262 and 1.185. Bidder-count buckets matching Table 13 then partition the residuals, and every estimator is run bucket by bucket.

All three estimators are calibrated on a Monte Carlo in which the costs are known. Two departures from the first-price specification of Section 7.4 define the design: costs and bids acquire a common auction-level multiplier $u_\ell = \exp(\sigma_u \zeta_\ell - \sigma_u^2/2)$, and the idiosyncratic cost draw becomes lognormal in place of uniform, the equilibrium bid function being computed by numerical integration on a scale-invariant grid. The change of cost family is necessary rather than cosmetic: under a uniform draw the idiosyncratic log-bid density inherits a support edge, and the characteristic function of such a density decays too slowly for deconvolution of any kind, a defect of the simulation rather than of bid data. The simulation runs 200 replications per cell at $\sigma_u \in \{0, 0.3, 0.6\}$, with a true median markup rate of 0.181. The naive pipeline returns 1.36–1.42, inflation by a factor of seven to eight that widens as σ_u rises. The Kotlarski estimator returns 0.65, with a replication standard deviation of 0.21, which is the conditioning failure the previous subsection predicts. The mixture returns 0.139 with a standard deviation of 0.028, at every σ_u alike, and regardless of whether the homogenisation includes text. Its downward bias at this scale is a small-sample matter; enlarging the simulated sample closes it. With $\sigma_u = 0.3$ held fixed, the recovered median moves from 0.140 at 250 auctions through 0.158 at 1,000 and 0.179 at 5,000 to 0.192 at 20,000, the truth sitting at 0.181 throughout and the replication standard deviation shrinking from 0.032 to 0.011. At Québec-bucket sample sizes the estimator is accordingly unbiased to within roughly six per cent either way. Sample size does far less for the naive pipeline, whose inflation only edges down from 7.7× at 250 auctions to 6.7× at 20,000; a conditioning failure is not cured by more data.

Results

The variance split and the corrected markups appear in Table A.3. In every dataset and every bucket the auction level accounts for 89–97% of residual variance; the under-conditioning claim of Section 9.6 now has a number attached. On Québec the corrected medians are 8.8/5.6/3.1% where Table 13’s naive figures were 59/44/26%, a reduction of roughly the sevenfold factor the Monte Carlo attributes to the naive pipeline, and at $n \geq 5$ the corrected figure sits next to the model-free money-on-the-table median (runner-up bid over winning bid, minus one) of 6.6% and next to the competitive margins of a few per cent found by Clark et al. (2018). The decline of markups in n is preserved. Feasibility exceeds 93% everywhere. Nor do the tuning choices move the answer: taking the mixture order through $K = 2, 3, 4$, and varying the fitted frequency range, the noise down-weighting scale, and the component-scale floor one parameter at a time, no reported median shifts by more than 1.2 percentage points on BC or 1.0 on Québec, and the Québec $n \geq 5$ median, the cell nearest the competitive benchmark, shifts by at most 0.2.

Table A.3: Markup rates after deconvolution, by dataset and bidder count. The variance split comes from the within-pair moment estimator, the markup quantiles from the mixture-CF estimator at $K = 3$ under text-inclusive homogenisation; homogenising on the structural block alone alters no quantile by more than 0.1 percentage points. The inversion uses bidder count n^* , the bucket value, taken as the within-bucket median for $n \geq 5$. At these sample sizes the Monte Carlo scale sweep bounds the mixture’s bias within roughly six per cent of the median.

Dataset	Bucket	n^*	σ_u^2	σ_e^2	UH share	q_{25}	q_{50}	q_{75}
QC	$n = 3$	3	1.620	0.045	0.973	6.6%	8.8%	12.2%
QC	$n = 4$	4	1.662	0.046	0.973	4.0%	5.6%	8.3%
QC	$n \geq 5$	6	1.370	0.074	0.949	2.5%	3.1%	5.3%
BC	$n = 3$	3	1.143	0.094	0.924	6.0%	9.9%	17.1%
BC	$n = 4$	4	1.446	0.090	0.942	4.9%	7.5%	11.3%
BC	$n \geq 5$	7	1.158	0.138	0.894	3.5%	3.8%	7.1%

What the correction says about the text

The correction also clarifies the division of labour between the deconvolution and the text. The moment split records the text absorbing variance at the auction level and nowhere else: when the compressed text enters the Québec homogenisation at $n = 3$, σ_u^2 falls from 1.676 to 1.620 while σ_ϵ^2 holds to four decimals, inasmuch as an auction-level covariate can do nothing else. For the *level* of the markup, then, text and deconvolution substitute for one another; the corrected medians agree with and without text, since auction-level variation the covariates fail to catch lands in u and is stripped either way. No such substitution holds for conditional objects. Deconvolution removes u while learning nothing about which auctions drew it high, wherefore cost prediction, comparison across segments, and any counterfactual conditioned on project characteristics still need the covariates; nothing in this appendix touches the ten-point prediction gap of Sections 8 and 9. Withal the text shrinks the very factor the deconvolution must assume independent of the covariates, so the correction and the conditioning reinforce one another rather than compete.

Two caveats thereanent. The multiplicative-separable form, symmetric within the auction, is assumed rather than tested; and the mixture family, flexible though it is, remains parametric. What is claimed is therefore an order of magnitude rather than a point estimate: strip the auction-level heterogeneity and the markup level that remains runs single-digit to low-teens, whence the distance between these figures and the naive 33% measures the heterogeneity still outwith the covariate list.